Profiling patterns of interhelical associations in membrane proteins.

Cabrera, Gorka Lasso

PROFILING PATTERNS OF INTERHELICAL ASSOCIATIONS IN MEMBRANE PROTEINS

by

Gorka Lasso Cabrera

A thesis submitted for the degree of Doctor of Philosophy of the

University of Wales Swansea

January 2007

To Marta, Marian and Juan

This thesis would not have been possible without your

support, advice and love

"Imagination is more important than knowledge"

Albert Einstein (1879 - 1955)

Profiling patterns of interhelical associations in polytopic membrane proteins

Summary

A novel set of methods has been developed to characterize polytopic membrane proteins at the topological, organellar and functional level, in order to reduce the existing functional gap in the membrane proteome.

Firstly, a novel clustering tool was implemented, named PROCLASS, to facilitate the manual curation of large sets of proteins, in readiness for feature extraction.

TMLOOP and TMLOOP writer were implemented to refine current topological models by predicting membrane dipping loops. TMLOOP applies weighted predictive rules in a collective motif method, to overcome the inherent limitations of single motif methods. The approach achieved 92.4% accuracy in sensitivity and 100% reliability in specificity and 1,392 topological models described in the Swiss-Prot database were refined.

The subcellular location (TMLOCATE) and molecular function (TMFUN) prediction methods rely on the TMDEPTH feature extraction method along data mining techniques. TMDEPTH uses refined topological models and amino acid sequences to calculate pairs of residues located at a similar depth in the membrane. Evaluation of TMLOCATE showed a normalized accuracy of 75% in discriminating between proteins belonging to the main organelles.

At a sequence similarity threshold of 40%, TMFUN predicted main functional classes with a sensitivity of 64.1-71.4% and 70% of the olfactory GPCRs were correctly

predicted. At a sequence similarity threshold of 90%, main functional classes were predicted with a sensitivity of 75.6-92.8% and class A GPCRs were sub-classified with a sensitivity of 84.5%-92.9%. These results reflect a direct association between the spatial arrangement of residues in the transmembrane regions and the capacity for polytopic membrane proteins to carry out their functions.

The developed methods have for the first time categorically shown that the transmembrane regions hold essential information associated with a wide range of functional properties such as filtering and gating processes, subcellular location and molecular function.

DECLARATION

This work has not previously been accepted in substance for any degree and is not being concurrently submitted in candidature for any degree.

Signed .... (candidate)

Date .....2/07/07.....

STATEMENT 1

This thesis is the result of my own investigations, except where otherwise stated. Where correction services have been used, the extent and nature of the correction is clearly marked in a footnote(s).

Other sources are acknowledged by footnotes giving explicit references. A bibliography is appended.

Signed .... (candidate)

Date .....2/07/07.....

STATEMENT 2

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed .... (candidate)

Date .....2/07/07.....

Table of contents

List of tables

List of figures

Acknowledgements

After four long years, the time has come to write these lines. There has been so much happening during these last four years, so many people that is difficult not to forget someone while I am looking into the past. This project was born as a product of coincidences when I landed in Luton to do the last year of my degree as Erasmus student.

Returning to Luton after finishing the degree, leaving my friends and family behind, was not easy. However, thanks to great friends such as Paqui, Carlos, Ron, Anthony and Pierre I did not find that difficult to start this new journey, close to its final destination now. Before settling down in Swansea, my second stop of this adventure was Tregaron. I wonder what the locals thought when they discovered that a lost foreigner was renting a small cottage in their village (specially considering that they do not even like people from the next village). Thanks to Lewis and Laura and her family I made myself at home in the Welsh valleys. The following stop of my journey was Swansea, where I have spent the last three years. I have met so many great people here that I would not have space to write their names. Thanks to my officemates Julia, Rachel, Trish, Sarah, Ruth, Claire, Wendy, Mary, Maria, Ned, Steve and Mark for their company during those long days at University. Special thanks to Alberto, Joy and Dev who have always be there to help me every time I needed. Thanks to my old and current flat mates Lorraine and Ayisha who with I have enjoyed every single minute of their company, thanks Kevin for being such a great guy. This last year and half would not have been the same without the Swansea Salsa crew, to all of them, thank you so much, you do not realize how much you have given me. Special thanks to Nicky, who has always be there for the last two years. Thank you for all your support and being such a good friend. Thanks Rachel for everything, for your friendship, your advice, for being a great flat mate over the last year. Together, we have been through ups and down in our PhDs but we finally did it. I wish you the best of lucks in the future. I could not finish this paragraph without thanking my friend Nicolas and his invaluable help. I will never forget how much you helped me when I was stuck while programming, all our "scientific converstations" while having a pint, lunch or coffee. I told you once you were a great person I have not change my mind. I hope our friendship never ends and that sometime in the future we both could work in the same institution.

Thanks Vasilis for all your good suggestions, I do not think I am going to meet many people as great as you in future conferences. Special thanks to my great Brazilian friend Roberto. Thank you for all the support and being such a good host while I was in Brazil. Thank you for teaching me the basis of object oriented programming despite being busy with your PhD. Thanks to my secondary supervisor John Antoniw. John, it has been a pleasure working with you. Thank you for all your help and comments, there has no been a single meeting where I have not learned anything new (always leaving with a big smile). I still have loads to learn from you, so I hope we can work together for many years.

Despite the distance, many friends have been supporting me over these last for years. Thanks Jose Luis for always being there every time I needed (even if it is a late hours in the evening). Thanks Maria for being the way you are, please never change. Special thanks to my best mate Javi. Javitxu, 16 years of supporting each other... and counting.

Special thanks to my supervisor. Jonathan, before doing the honours project with you I already had a plan for a different following years. However, your passion for bioinformatics and membrane proteins dragged me into this research field which I am not planning to leave. Thank you for everything you have done for me, thank you (and Roisin) for your kindness while I was lost in the valleys. Thanks for those long conversations about our project and so many other things. I hope we can start a new journey together after this one ends.

The biggest acknowledgements are obviously to my family. It has been eight years since I left Bilbao but I have never felt alone thanks to your constant advice and support. I would not have done it without you. I know it has been difficult for you, so it has been for me, not being without you. Special thanks to my grandmothers Visi and Adela. I wish my grandfather could also read these lines to make him proud. Special thanks to my sister Marta, my mother Marian and my father Juan for everything. This thesis is dedicated to you. You have given me so much unconditional love during these last years that I do not known how to express how much you mean to me. Despite the distance, I have always carried you with me, deep in my heart.

Finally I would like to thank the Basque Government for their economic support during the four years this PhD took.

Abbreviations

ABC	-	ATP Binding Cassette
AC	-	Accession Code
AQP	-	Aquaglyceroporin
ATP	-	Adenosine 5'-triphosphate
BPDPFECD	-	Binding Protein Dependent Permease FeCD subfamily
CGI	-	Common Gateway Interface
CLC	-	ClC chloride channel (only applicable when describing membrane dipping loops)
DE	-	Definition
DNA	-	Deoxyribonucleic acid
DMD	-	Duchenne Muscular Dystrophy
EC	-	Enzyme Classification
ER	-	Endoplasmic Reticulum
FT	-	Feature
GO	-	Gene Ontology
GPCR	-	G protein-coupled receptor
GPI	-	Glycosylphosphatidylinositol
HMM	-	Hidden Markov Model
ID	-	Identifier
IIS	-	Internet Information Service
IUBMB	-	International Union of Biochemistry and Molecular Biology
MEMBLOOP	-	Membrane dipping loop
MDL	-	Membrane dipping loop (in supplementary information)
MVC	-	Model View Controller
NAD^+	-	Nicotinamide adenine dinucleotide
$NADP^+$	-	Nicotinamide adenine dinucleotide phosphate
NCBI	-	National Center for Biotecnology Information
NMR	-	Nuclear Magnetic Resonance

OS	-	Organism Species
OC	-	Organism Classification
PDB	-	Protein Data Bank
PS	-	Photosystem
SVM	-	Support Vector Machine
RAD	-	Rapid Application Development
RBF	-	Radial Basis Function
SDF	-	Sodium : Dicarboxylate symporter
SQ	-	Sequence
TC	-	Transport Classification
TM	-	Transmembrane region
TRANSMEM	-	Transmembrane region

Membrane dipping loops structural classification

HIHO	-	Helix-In-turn-Helix-out
HILO	-	Helix-In-turn-Loop-out
LIHO	-	Loop-In-turn-Helix-out

Accuracy predictive scores

Q	-	Overall accuracy
nQ	-	Normalized accuracy
GAv	-	Geometric Average
MCC	-	Matthews Correlation coefficient
GC	-	Generalized Coefficient

Amino acid one and three leter code

A	-	Ala	-	Alanine
C	-	Cys	-	Cysteine
D	-	Asp	-	Aspartic acid
E	-	Glu	-	Glutamic acid
F	-	Phe	-	Phenylalanine
G	-	Gly	-	Glycine
H	-	His	-	Histidine
I	-	Ile	-	Isoleucine
K	-	Lys	-	Lysine
L	-	Leu	-	Leucine
M	-	Met	-	Methionine
N	-	Asn	-	Asparagine
P	-	Pro	-	Proline
Q	-	Gln	-	Glutamine
R	-	Arg	-	Arginine
S	-	Ser	-	Serine
T	-	Thr	-	Threonine
V	-	Val	-	Valine
Y	-	Tyr	-	Tyrosine
W	-	Trp	-	Tryptophan

CHAPTER 1

General Introduction

1.1 Membrane proteins

1.1.1 The abundance and importance of membrane proteins

The extensive number of different genome sequencing projects has provided the scientific community with a massive amount of genomic information. Although DNA contains all the required information for any cell, all cellular processes are mediated by proteins. Therefore, the post-genomic era has focused on the identification and characterization of all proteins encoded by entire genomes in the lifetime of a cell. Among cellular proteins, membrane proteins have been shown to be an abundant and important group of proteins. Previous studies have shown that membrane proteins account for 20-30% of the entire cellular proteome (Boyd et al., 1998, Wallin and von Heijne, 1998). Such abundance is reflected by the diverse cellular functions carried out by membrane proteins, where most of these functions are crucial for cellular development and maintenance. Membrane proteins are essential mediators of transfer of material and information between compartments within cells, between cells and their environment, and between tissues of different organ systems. These proteins are responsible for creating and maintaining the particular and finely regulated composition of the cell interior relative to the outside. Likewise, membrane proteins sense signals from the environment and mediate neurotransmission and other communication processes. Furthermore, membrane proteins intervene in energy transformation processes such as photosynthesis, respiration and ATP production.

The importance of membrane proteins is also reflected at the pharmacological level, where aberrant function of these types of molecules leads to many disease states such as Alzheimer's disease, cancer, heart disease, cystic diseases, Duchenne Muscular Dystrophy (DMD) and neurological disorders. Membrane proteins are the target of many pharmacologically and toxicologically active substances and are responsible, in part, for the uptake, metabolism and clearance of these substances. Moreover, these proteins have a great importance in drug discovery as they account for approximately 60-70% of all known pharmaceutical drugs targets (Wu and Yates, 2003).

1.1.2 The lipid bilayer

All membrane proteins are either non-covalently bound or embedded in the lipid bilayer, which defines the boundaries of the cellular organelles and the entire cell. Cell membranes play a critical role in cellular structure and maintain the essential differences between the cytosol and the extracellular environment. Although the basic structure and function of the cellular membrane is provided by the lipid bilayer, membrane proteins confer unique compartment-specific functions and communication between separated environments (Wu and Yates, 2003).

Lipid molecules in the membrane are amphipathic, that is, they have a polar or hydrophilic end and a non-polar or hydrophobic end. The most abundant lipids in the membrane are phospholipids. Phospholipids have a polar head group and two hydrophobic hydrocarbon tails. The polar head group is composed by a phosphatidylcholine moiety and the tails are usually fatty acids. Fatty acids consist of a carboxylic group attached to a single 14-24 CH_2 chain. These fatty acids may have one or more cis-double bonds creating a kink in the tail. The differences in length and saturation of fatty acids influence the packing of contiguous phospholipids, which consequently affects the fluidity of the membrane. However, phospholipids are not the only lipid components of the membrane. Cholesterol and glycolipids are also components of the eukaryotic cell membrane. Likewise, other molecules such as sphingomyelin or phosphatidylserine, are components of biological membranes. The lipid and protein composition is not the same for different biological

membranes, and confers different physicochemical and functional properties upon the corresponding membranes. The lipid bilayer of bacteria is often composed of phospholipids and does not contain cholesterol whereas in eukaryotes the lipid composition is more diverse containing different types of phospholipids and large amounts of cholesterol (Alberts et al., 1994).

The accepted model used to describe the lipid membrane, is the fluid mosaic model proposed by Singer and Nicholson (Singer and Nicolson, 1972). This model describes the biological membrane as a two-dimensional fluid environment composed of two lipid layers. According to this model, the outer layer is composed mainly of head groups of lipids, where the highly polar nature of these groups allows the lipid bilayer to interact with the aqueous solution. On the other hand, the inner area of the membrane is composed of hydrophobic hydrocarbon chains of fatty acids. The hydrocarbon chains are aligned parallel to each other and their non-polar ends contact with each other in the middle of the cell membrane. These contacts create a non-polar barrier, which is impermeable to most polar molecules and ions but allows small, non-polar molecules to pass through. This hydrophobic region also provides the most distinctive region of solvation for membrane proteins due to the absence of the hydrophobic effect and the strength of ionic interactions over long distances in the low dielectric field (Popot and Engelman, 2000).

1.1.3 Structural types of membrane proteins

Membrane proteins possess a wide variety of shapes and sizes. Interestingly, the basic architectural principles of these proteins are quite different to the structural basis of soluble proteins. Different regions of membrane proteins are located within environments of different composition. These environments correspond to the lipid environment, the membrane-water interface and the aqueous medium. On the contrary, soluble proteins are only located in aqueous environments. These different environments impose different physicochemical constraints, which lead to different principles governing the assembly of their secondary structures.

Membrane proteins are either bound to one side of the membrane or completely span through the cell membrane. Peripheral anchored membrane proteins are non-covalently bound to the membrane either by lipids (lipid chain-anchored membrane proteins) or by a glycosylphosphatidylinositol molecule (GPI-anchored membrane proteins). Peripheral anchored membrane proteins can be considered as tethered soluble proteins that do not seem to be influenced by the cell membrane (von Heijne, 1996). On the other hand, integral membrane proteins completely span the lipid bilayer at least once.

1.1.3.1 Integral membrane proteins

Integral membrane proteins can be further sub-classified into two categories according to the secondary structures that compose these proteins: β-barrel membrane proteins and α-helical membrane proteins.

1.1.3.1.1 β-barrel membrane proteins

The transmembrane domain of this protein type is composed by antiparallel β-strands that form a β-barrel structure **(figure 1.1)**. An array of β-strands (in β-barrel conformation) constitutes the backbone of the structure and the loop region connecting the β-polypeptides contains the helices, which surround the β-barrel structure. β-sheet structure is formed when at least two almost fully extended polypeptide chains are brought together side by side so that regular hydrogen bonds can be formed between the peptide backbone amide NH and carbonyl oxygen of adjacent chains (Zubay et al., 1998). Due to the trans orientation of the NH and carbonyl groups, multi-stranded structures are obtained when successive chains are added to the sheet. According to the alignment of the different polypeptide chains, β-sheets can be classified as parallel β-sheets (all the chains arranged with the same N-to-C polypeptide sense) or antiparallel β-sheets (chains arranged in opposite N-to-C polypeptide sense). In the case of β-barrel membrane proteins, all crystallized membrane proteins display a β-barrel structure composed by antiparallel β-sheets. The β-sheet that forms the barrel has an hourglass-shaped surface with cylindrical

curvature. Twisted β-strands with a staggered hydrogen-bond pattern produce a cylindrical curvature (Zubay et al., 1998). This structure is formed as a consequence of a competition between the natural tendency of chains to be twisted in a right-handed sense and the inter hydrogen bonds (between adjacent β-strands).

Porins are by far the best studied protein family belonging to this structural type. Porins are found in the outer membrane of many bacteria. These proteins allow the passive diffusion of small molecules across the membrane through the water filled pore formed by the antiparallel β-strands (Schulz, 1996). A polypeptide loop lining the inner wall of the pore determines the functional properties of the pore (Garavito, 1998).

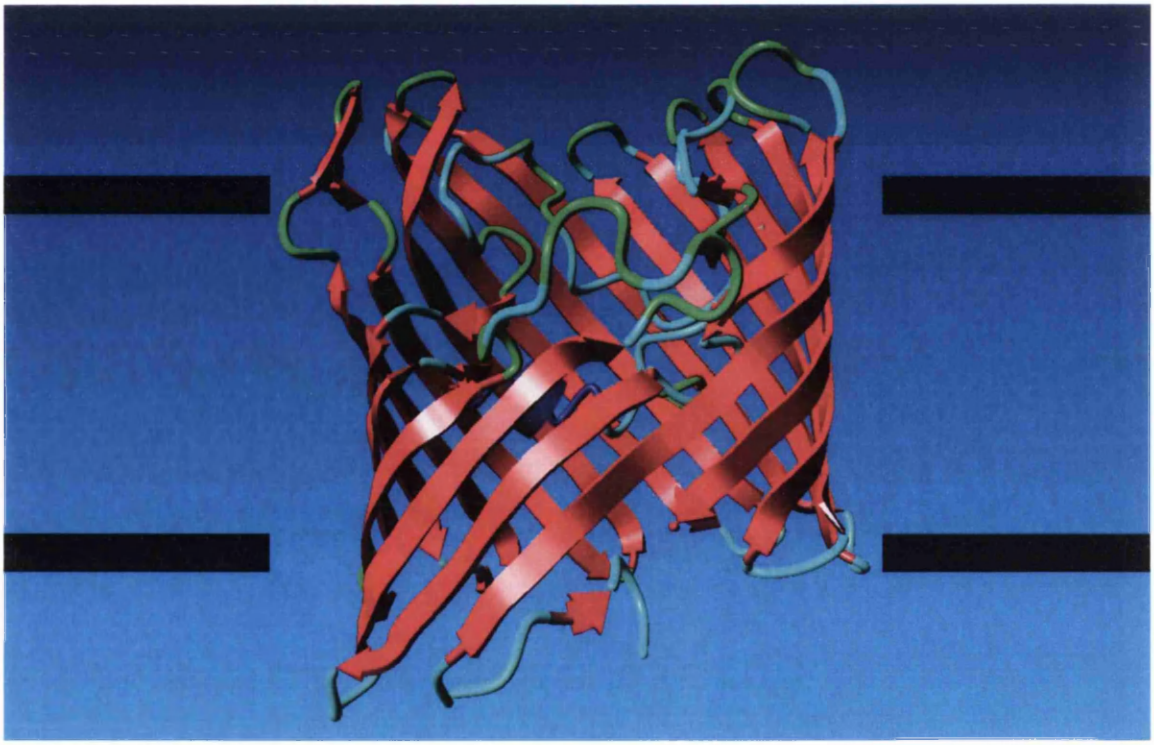

Figure 1.1. Structure of the the anion-selective porin Omp32 (PDB code 1E54) (Zeth et al., 2000). Image generated with Yasara (Yasara biosciences, www.yasara.org).

1.1.3.1.2 α-helical membrane proteins

The transmembrane domain of this protein type is composed of α-helices that completely traverse the lipid bilayer. These proteins can be further sub-classified into single

α-helix transmembrane proteins, which only span the membrane once, and polytopic membrane proteins, which contain two or more α-helices that completely span the membrane. The classical concept of the topology of polytopic membrane proteins was that of a bundle of hydrophobic α-helices, each composed by 15-25 residues, orientated approximately perpendicularly to the membrane plane **(figure 1.2)**. However, the different crystallized structures have shown that it is not that simple **(figure 1.3)**. The transmembrane regions of polytopic membrane proteins can include not only ordered secondary structures, but also unfolded secondary structures, and α-helices can vary in length and composition. Furthermore, the angle formed by the helical axis and the membrane plane can vary significantly from helix to helix, membrane dipping loops (re-entrant loops) dip to a certain depth in the membrane and then turn back, and many helices located at the water-lipid interface are orientated parallel to the membrane plane.

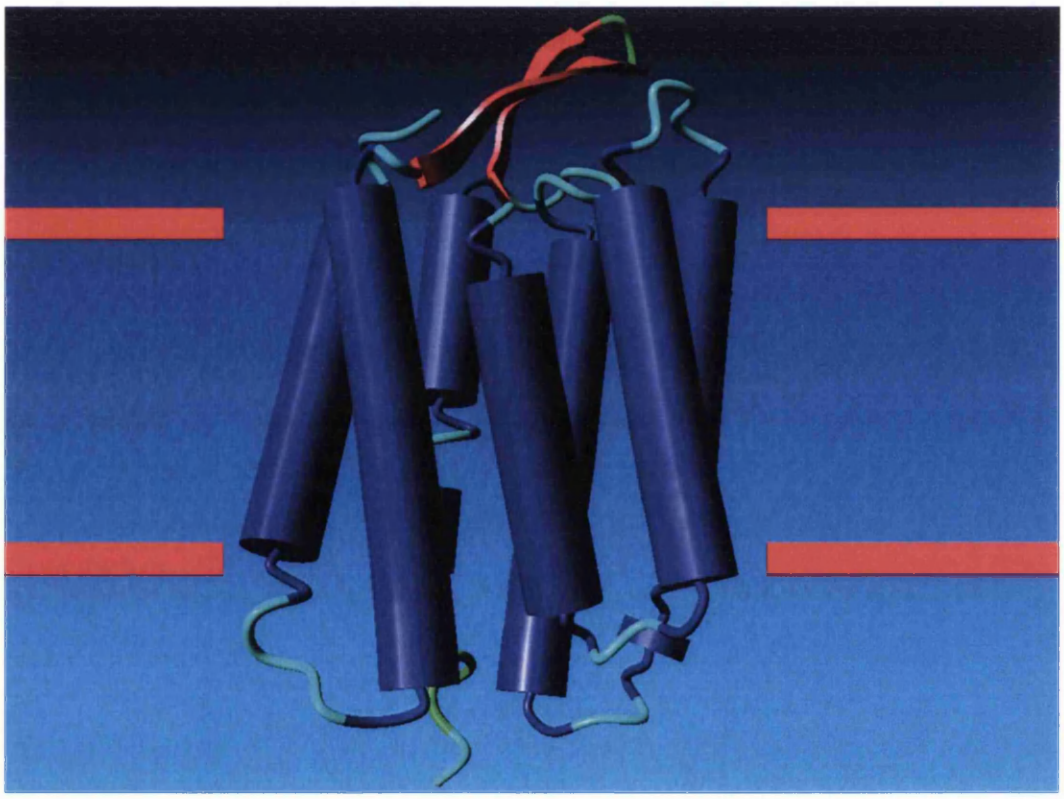

Figure 1.2 Structure of bacteriorhodopsin (PDB code: 1QHJ) (Belrhali et al., 1999). Image generated with Yasara (Yasara biosciences, www.yasara.org).

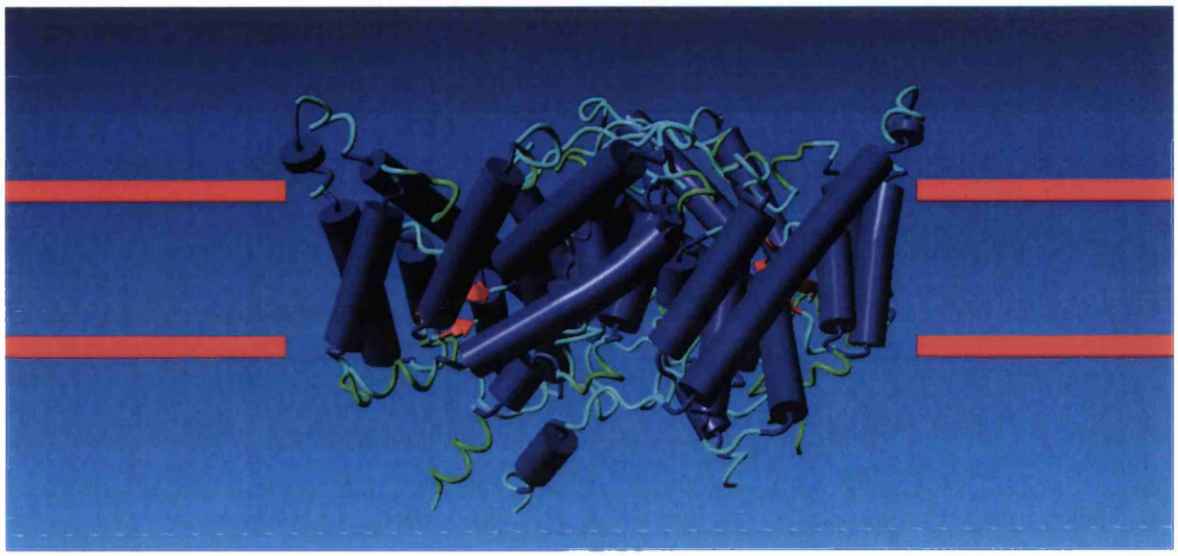

Figure 1.3. Structure of the ClC Chloride channel (PDB code: 1KPK) (Dutzler et al., 2002). Image generated with Yasara (Yasara biosciences, www.yasara.org).

In an α-helix, the polypeptide backbone follows the path of a right-handed helical spring to form an arrangement in which each residue's carbonyl group forms a hydrogen bond with the amide NH group of the residue four amino acids further along the polypeptide chain (Zubay et al., 1998). All residues have identical conformation, keeping the N-Cα (Φ) and Cα-C (Ψ) rotation angles similar along the polypeptide main chain (Φ= -60°, Ψ= -45° to –50°). This leads to a 100° rotation along the helical axis from one residue to the following residue and the approximate distance along the helical axis between contiguous residues is 1.5Å **(figure 1.4)**. Therefore, within a regular structure each 360° of helical turn incorporates approximately 3.6 amino acids. The stability of this conformation is a consequence of the formation of stable hydrogen bonds between all the backbone NH groups and carbonyl groups and the tight packing achieved. The α-helix bundle domain is based on the helix-loop-helix motif creating clusters of interacting α-helices. The segments of α-helix are held together in one polypeptide chain by interconnecting loops of extended polypeptide chain (Zubay et al., 1998).

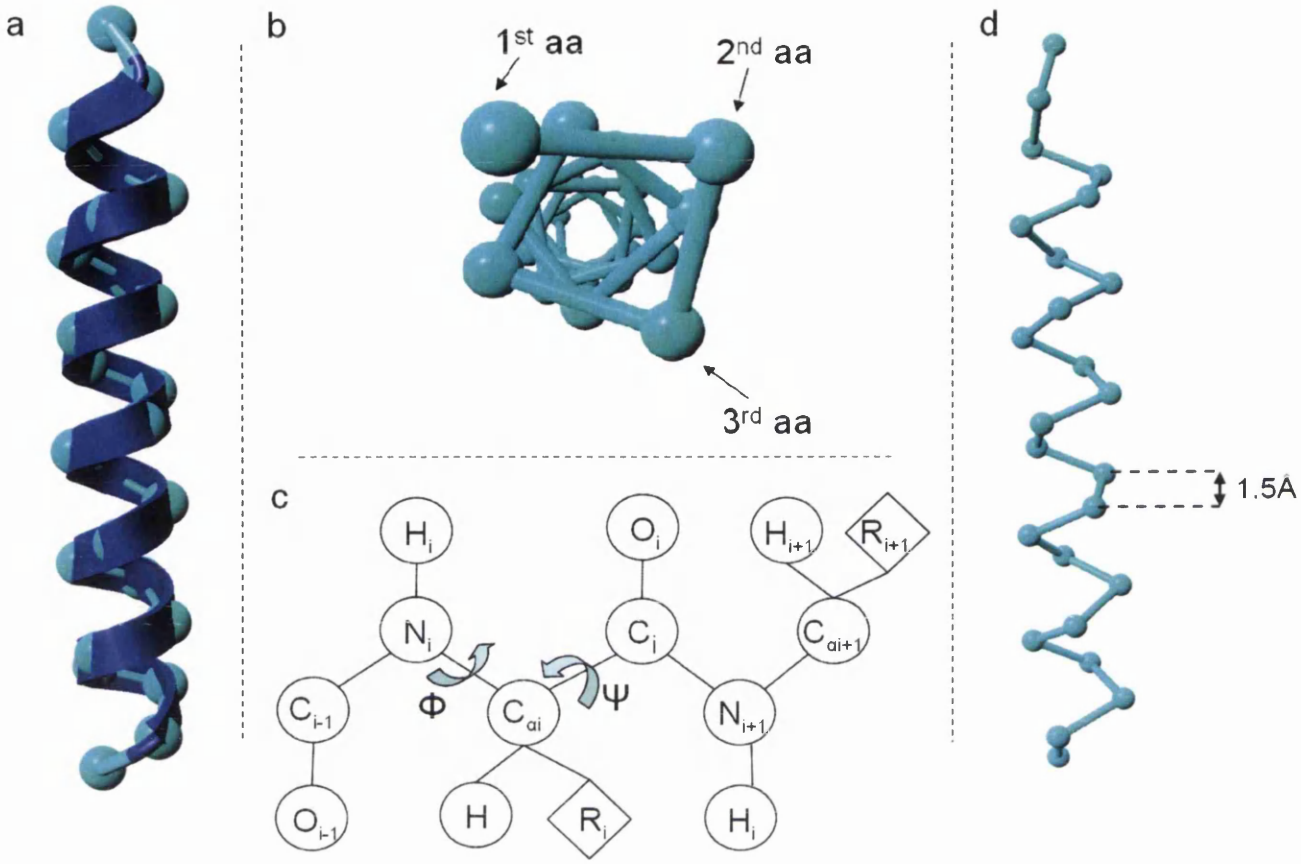

Figure 1.4. Spatial conformation of the α-helix. a) helical structure with the corresponding α-carbon for each amino acid. b) The helical backbone , the figure illustrates the rotated characteristic of helical structure. c) The rotation angles along the helical backbone. d) The helical backbone, the figure illustrates the distance along the helical axis between two contiguous residues. Image generated with Yasara (Yasara biosciences, www.yasara.org).

A study carried out by Ulmschenider and Sansom (Ulmschneider and Sansom, 2001) reflects the amino acid distributions in integral membrane proteins. The hydrophobic residues Alanine, Isoleucine, Leucine and Valine make up to 34% of all residues in α-helical proteins. Leucine has the highest propensity for being in the transmembrane region and might contribute strongly to the formation of helices. Due to the highly hydrophobic nature of this residue, it was expected that Leucine would prefer the lipid exposed area of the transmembrane regions. However, strongly hydrophobic residues (Phenylalanine, Leucine, Isoleucine and Valine) do not show any preference for location within the buried lipid bilayer. By contrast, small hydrophobic residues (Glycine and Alanine) prefer to be located within the lipid bilayer rather than on the surface of the transmembrane domain. Glycine has a high frequency in transmembrane regions. This residue has been considered a

helix breaker for globular proteins, however, in membrane proteins it promotes, with alanine, the interhelical associations facilitating the closer packing of helices due to its short side chain. Aromatic residues show a high propensity to face the lipid head-groups at either or both transmembrane termini forming the "aromatic belt" (Adamian and Liang, 2001, Pilpel et al., 1999). These residues play an important role in membrane protein structure determination. They are believed to anchor the proteins into the membrane through an interaction of their aromatic rings with the lipid head groups. Likewise, histidine is frequently found in active sites and often plays an important functional role (Adamian and Liang, 2001). Proline is abundant in the loop regions outside the membrane, but it is also detected towards the centre of the bilayer. This residue may increase the stability of the transmembrane regions by 'interlocking' the helices, or by providing molecular hinges that enable conformational transitions in more complex membrane proteins (Ulmschneider and Sansom, 2001). Polar residues are poorly represented in transmembrane regions. In fact, with the exception of threonine and serine, which account for ~7% of the residues, the remaining polar residues constitute only 1-3% (Curran and Engelman, 2003). Serine and threonine side chains in a helix can form H-bonds to the carbonyl oxygen of the preceding turn of the helix (Ulmschneider and Sansom, 2001), minimising the energetic penalty. These percentages can be rationalized in view of the energetic penalty of locating polar amino acids in the low dielectric medium of the lipid bilayer (White and Wimley, 1999). However, polar and charged amino acid residues are structurally and functionally important as they appear to be less mutable when they occur in transmembrane regions. The positively charged residues, arginine, histidine and lysine show a preference to face the lipid only when located at the cytoplasmic end. These residues might facilitate the anchoring of the transmembrane helices to the membrane by means of polar interactions with the negatively charged head groups of the lipids (Pilpel et al., 1999) **(for a more detailed review describing the folding process of polytopic membrane proteins please see chapter 6)**.

In terms of functional diversity, polytopic membrane proteins are by far the most significant membrane protein type, yet the most challenging in terms of protein structure prediction. Therefore, the majority of recent experimental and computational approaches

have been focused on the elucidation of the folding processes **(Chapter 6)**, structures and functional properties (**Chapter 7** and **Chapter 8**) pertaining to polytopic membrane proteins.

1.1.4 Structural and functional gap in the membrane proteome

Despite the natural abundance and medical significance of membrane proteins, many membrane proteins remain structurally unknown. By January 2007, only 277 membrane protein structures were found in the Protein Data Bank (PDB) (Berman et al., 2000), corresponding to 93 or 121 unique proteins found in the Hartmut Michel's database of crystallized membrane proteins, (http://www.mpibp-frankfurt.mpg.de/michel/public/memprotstruct.html) and the Stephen White laboratory, at University of California, Irvine, (http://blanco.biomol.uci.edu/Membrane_Proteins_xtal.html) respectively. This accounts for less than 1% of all protein structures contained in the PDB. Experimental studies to elucidate the structure of proteins are mainly based on X-ray crystallography, Nuclear Magnetic Resonance (NMR) and electron microscopy. Although these methods have proven to be very successful for structure determination in soluble proteins, these techniques are severely hampered by the lipid environment. Lipid molecules complicate the preparation of quality X-ray crystals and samples for multidimensional solution NMR studies. Likewise, when membrane proteins are extracted from the membrane, proteins lose their natural folding pattern unless they are maintained in a non-polar environment similar to that found in the membrane. Therefore, it might be possible that artefactual results are obtained by experimental analysis. The extreme hydrophobic nature of most membrane proteins have made them difficult targets for X-ray crystallography and NMR techniques (Deber et al., 2001). Furthermore, quite often it is difficult to obtain sufficient amounts, and of sufficient quality to apply X-ray crystallography or other experimental approaches (Saidijam et al., 2003). Experimental analyses to elucidate functional properties of newly identified membrane proteins are also difficult due to the nature of the lipid membrane. This situation contrasts with exponential growth in the number of sequences deposited in biomolecular databases due to the substantial number of different sequencing projects being

carried out. Therefore, the structural and functional gap observed in the membrane proteome is constantly widening.

In order to alleviate this situation, computational methods are being developed to predict the structure and function of newly identified proteins. Structural prediction of membrane proteins is mostly focussed on the prediction of topology of membrane proteins **(please see chapter 6 for a more detailed review on topology prediction methods).** However, efforts are also being made to predict the three-dimensional structure of polytopic membrane proteins. Structural prediction methods can broadly be divided into two main groups, comparative methods and *ab initio* methods. Comparative methods rely on detectable similarity spanning most of the modelled sequence and at least one known structure. The sequence to be modelled is aligned with the known structure, the modelled protein including either sequential or simultaneous modelling of the core of the protein, loops and side chains. *Ab initio* methods attempt to predict the structure of the protein directly from the sequence. These methods are based on the assumption that the native state of a protein is at the global free energy minimum, and carry out a large scale search of conformational space for protein tertiary structures that are particularly low in free energy for the given amino acid sequence (Baker and Sali, 2001).

Several methods have been implemented for the functional prediction of newly identified proteins, such as sequence similarity-based methods, genomic context methods and data mining methods. These predictions not only consider the molecular process involved but also functional features such as the subcellular location, identification of functionally important residues, post-translational modifications and protein-protein interactions (**please see chapter 7 and chapter 8 for a more detailed review**). However, these methods are mostly applied to soluble proteins and only a few methods have been implemented to predict functional properties of membrane proteins.

1.2 Bioinformatics and data mining

Over the last decade, bioinformatics has established itself as a discipline within the biological and biomedical research field. This discipline certainly stands out for its multidisciplinary character, where experts from a wide range of backgrounds (e.g. biology, chemistry, physics, medicine, mathematics, computer science and statistics) meet to achieve a common goal. Therefore, techniques first developed for completely unrelated tasks have been successfully applied in computational biology. For instance, Hidden Markov Models (Baum and Petrie, 1966) were first applied to speech recognition (Jelinek, 1969) and were first applied in computational biology in 1989 (Churchill, 1989). Further applications and improvements have led into the implementation of popular tools such as PSI-BLAST (Altschul et al., 1997) for remote homologue detection and TMHMM (Krogh et al., 2001) for the topology prediction of membrane proteins. Bioinformatics tools have been applied to diverse tasks such as gene detection, splice variant prediction, sequence comparison, phylogenesis, protein structure prediction, functional prediction, post-translational modification predictions and microarray data analysis. Although there are still many problems that have not been unravelled (e.g. sequence-to-structure-to-function paradigm), bioinformatics is advancing into areas for which little experimental data is currently available. Systems biology, for example, aims to predict the behaviour of biological systems under particular conditions, which involve not only interactions between different pathways (e.g. gene relation networks and enzymatic pathways) in a cell but also interactions between the different biological units within a system. Whereas this discipline is still contentious for some scientists, who argue that there is not yet enough data to support those predictions, others believe that the computational infrastructure should be developed in preparation for the more enriched data of the future. Examples of such developments are the EcoCyc database (Karp et al., 2000), which contains a literature-based curation of the entire genome, transcriptional regulation, transporters and metabolic pathways in *E. coli*, and the BRENDA database (Schomburg et al., 2002), which contains enzyme and metabolic information.

Many of the questions involved in bioinformatics research can also be regarded as a machine learning task (Frank et al., 2004): i) Prediction of an output given a particular feature vector, ii) clustering instances sharing similar features and iii) selection of features to predict a particular outcome. Data mining techniques such as support vector machines or Bayesian methods, have already been used for tasks such as functional characterization (Karchin et al., 2002), probe selection for gene-expression arrays (Tobler et al., 2002), subcellular location prediction (Nair and Rost, 2005) and classifying gene expression profiles (Li et al., 2003). Weka (Witten and Frank, 2005) is a data mining platform, which provides a wide suite of algorithms for the different machine learning tasks. The different data mining methods implemented within the Weka platform vary in complexity from simple and straightforward methods such as the k-nearest neighbour or decision trees, to more sophisticated methods such as support vector machines and neural networks. Increased complexity does not necessarily mean better performance and it is often observed that the simpler classifiers perform better than sophisticated ones. The reason for this is that the nature of real data sets varies, so specific data mining methods perform better depending on the nature of particular data sets. This platform allows rapid comparison of different data mining techniques in order to identify the data mining method that best fits a given dataset. Below, different data mining techniques used in the project are briefly described.

1.2.1 Bayesian methods

Bayesian methods are statistical based methods, where predictions are made considering different probabilities and costs associated with such predictions (Duda et al., 2001). The basic theory underlying these methods assumes that all attributes used to make such decisions are equally important and independent from each other. Likewise, it is assumed that numeric attributes follow a normal distribution. Although these assumptions do not reflect the empirical properties of the data sets, these methods have proven very useful for the data classification problem (Witten and Frank, 2005).

The basic principle of these methods is based on the Bayes formula:

$$P(\omega_j \mid \chi) = \frac{p(\chi \mid \omega_j)P(\omega_j)}{p\chi} \qquad (1.1)$$

where $P(\omega_j| \chi)$ is the posterior probability of the state of nature belonging to class ω_j given the observed values for the set of features χ, $p(\chi \mid \omega_j)$ is the likelihood of the class ω_j with respect to the set of observed values for the set of features features χ, $P(\omega_j)$ is the prior probability of the class ω_j and $p(\chi)$ is the evidence factor, which guarantees that all possible posterior probabilities (as many as the different classes involved in the classification process) sum to 1. In a hypothetical case, where the set of features is composed by three different features, and assuming that each feature is independent from each other, equation 1.1 can be extended to

$$P(\omega_j \mid \chi) = \frac{p(\chi_1 \mid \omega_j)p(\chi_2 \mid \omega_j)p(\chi_3 \mid \omega_j)P(\omega_j)}{p\chi} \qquad (1.2)$$

where the likelihood of the class ω_j with respect to each observed feature is computed individually. Therefore, equation 1.1 and 1.2 can be paraphrased into:

$$posterior\ probability = \frac{likelihood \cdot prior\ probability}{evidence} \qquad (1.3)$$

The Bayesian method using equation 1.1 (or 1.2) is called the naïve Bayesian method as it assumes independency between attributes (Witten and Frank, 2005). The main limitation of this method is that if the probability of the class ω_j with respect to a particular feature is found to be zero then the posterior probability will also be zero. This limitation can easily be overcome if particular features that occur zero times receive a likelihood value higher than zero. One strategy is to add 1 to each count while computing the likelihood of class ω_j with respect to each feature, and this is also known as the *Laplace estimator*. Another possibility is to include a prior constant value and probabilities to each extracted feature where the prior probabilities for all extracted features sum to 1.

A more sophisticated Bayesian method is the Bayesian network. This method is represented by a directed acyclic graph, where each node corresponds to a particular feature and the edges represent conditional dependencies **(figure 1.5)**. Bayesian networks can be used to represent causal dependencies where particular features can influence others. Nodes immediately before a particular node are considered the parents of that node and the nodes immediately after are considered the children of the node. Through a direct application of Bayes rule, it is possible to determine the probability of any configuration of features in the joint distribution. In order to calculate this value, it is necessary beforehand to compute the conditional probability tables, which give the probability of any feature at a node for each conditioning event (that is, for the values of the variables in the parent nodes) (Duda et al., 2001). In order to compute the likelihood at a particular node, only the information contained in the current node and the corresponding parental nodes is considered. The assumption made in Bayesian networks is that ancestor nodes do not provide additional information about the likelihood value of a particular node in light of the information provided by the parental nodes. In statistics, this property is called conditional independence (Witten and Frank, 2005).

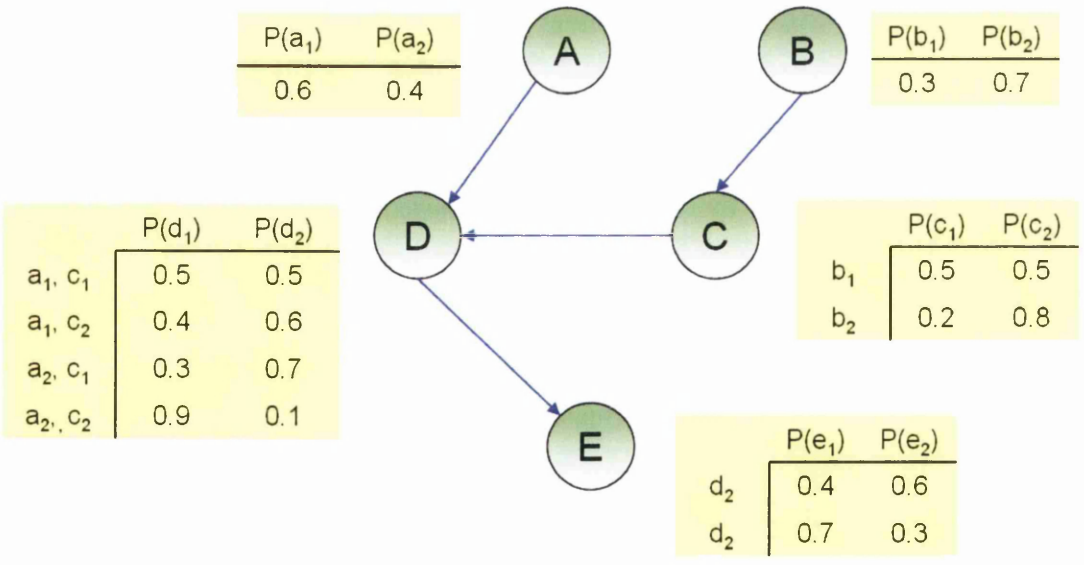

Figure 1.5. Example of a Bayesian network diagram. The nodes correspond to different features and the edges reflect the relationships between the nodes. The tables describe the combined probability for each feature.

Employing the Bayesian network architecture involves two different tasks: i) searching through the space of different networks (different arrangements of nodes) and ii) calculation of the log-likelihood of the given network, in order to measure the quality of the network (Witten and Frank, 2005).

1.2.2 Linear regression and logistic regression

Regression analysis is a statistical method used to develop a mathematical equation that shows how variables are related. The variable predicted by the equation is known as the dependent variable (which corresponds to the different classes used in a training set) whereas the variables used to predict the value of the dependent variable are known as independent variables (Anderson et al., 1994). Linear regression models assume a direct and linear dependency of the dependent variable (class) with respect to the independent variables (features). Further, the class can be expressed as a linear combination of the features, with predetermined weights (Witten and Frank, 2005):

$$\omega = z_0 + z_1 \chi_1 + z_2 \chi_2 + \ldots + z_k \chi_k \qquad (1.4)$$

where ω is the class; $\chi_1, \ldots, \chi_k$ are the different features; and $z_0, z_1, \ldots, z_k$ are the different weights.

The weights are computed based on the training data. Therefore, for a particular instance, which belongs to class $\omega^{(1)}$ and contains the features $\chi_0^{(1)}, \ldots, \chi_k^{(1)}$, the expected value corresponds to the sum of the product of the observed values (features) and the corresponding weights:

$$z_0 \chi_0^{(1)} + z_1 \chi_1^{(1)} + z_2 \chi_2^{(1)} + \ldots + z_k \chi_k^{(1)} = \sum_{j=0}^{k} w_j \chi_j^1 \qquad (1.5)$$

Linear regression methods compute weighted values in order to minimize the sum of the squares of the differences between the observed value $\omega^{(1)}$ and the expected value

(1.5) for each instance. This optimization is performed considering all instances within a training set, composed of n instances, in order to minimize overall sum of the squares of the differences for the given training set:

$$\sum_{i=1}^{n}\left(\omega^{(i)} - \sum_{j=0}^{n} z_j \chi_j^i\right)^2 \qquad (1.6)$$

In the multiclass problem, a regression analysis for each class is performed for each class and the output is set to 1 for training instances that belong to the class considered and *vice versa* (Witten and Frank, 2005). In order to classify an unknown instance, it is necessary to calculate the value of each linear expression based on the observed features of the unknown instance and select the class whose corresponding linear expression reports the largest value. This approach in also known as the multi-response linear regression (Witten and Frank, 2005).

The main limitation of linear regression methods is that the computed scores do not relate to probability scores. Furthermore, by minimizing the differences between the observed and the expected values using the square difference formula **(1.6),** it is assumed that errors are statistically independent and normally distributed with the same standard deviation. In order to overcome these limitations, the logistic regression method constructs a linear model based on a transformed vector of observed features. According to this method, the original target independent variables (features) for a specific instance are transformed to

$$\log(\Pr[1 \mid \chi_1,\ldots,\chi_k])/(1-\Pr[1 \mid \chi_1,\ldots,\chi_k]) \qquad (1.7)$$

Following the feature transformation, the obtained values are then approximated using a linear function. The obtained model is:

$$\Pr[1 \mid \chi_1,\ldots,\chi_k] = 1/(1+\exp(-z_0 - z_1\chi_1 - \ldots - z_k\chi_k)) \qquad (1.8)$$

This data transformation process and linear approximation is known as the logit transformation, which can be also written as:

$$P = \frac{e^{z_0+z_1\chi_1+\ldots+z_k\chi_k}}{1+e^{z_0+z_1\chi_1+\ldots+z_k\chi_k}} \tag{1.9}$$

where $P = \Pr[1|\chi_1, \ldots, \chi_k]$.

Unlike linear regression, logistic regression uses the log-likelihood to evaluate the differences between the observed values and the expected values. The chosen weights need to maximize the log-likelihood:

$$\sum_{i=0}^{n}\left(1-\omega^{(i)}\right)\log\left(1-\Pr\left[1 \mid \chi_1^{(i)},\ldots,\chi_k^{(i)}\right]\right)+\omega^{(i)}\log\left(\Pr\left[1 \mid \chi_1^{(i)},\ldots,\chi_k^{(i)}\right]\right) \tag{1.10}$$

When considering more than two classes, a similar approach to that defined for linear regression can be performed. However, the sum of the probabilities estimated will not be 1 (several efficient solutions have been implemented to optimize this problem). An alternative approach is the pairwise classification, where a two-class classifier is implemented for every possible combination of classes. The final prediction for an unknown instance corresponds to the most supported class.

1.2.3 K-star

The K-star method (Cleary and Leonard, 1995) is an instance-based classifier. This type of method predicts the class for an unknown instance based upon the class of those training instances similar to it, according to a similarity function. The underlying assumption is that similar instances will have similar classifications. The challenge of instance-based classifiers is to define "similar instance" and "similar classification". The majority of these methods use the Euclidean distance as the similarity function:

$$\sqrt{\left(\chi_1^{(1)}-\chi_1^{(2)}\right)^2+\left(\chi_2^{(1)}-\chi_2^{(2)}\right)^2+\ldots+\left(\chi_k^{(1)}-\chi_k^{(2)}\right)^2} \tag{1.11}$$

The K-star method, influenced by information theory, differs from the majority of these methods as it uses an entropic distance measure based on transformations. According to this distance measure, the unknown instance is transformed into a different instance by means of predefined operations. Such a distance is proportional to the absolute difference between two instances. Then the probability of the transformed instance occurring is computed, assuming that such operations took place randomly. The probability is defined as the probability of all paths from instance *i* to instance *a*. When the attributes of the instances are real values, this probability is determined by the product of the probability of the individual transformations. The probability of an instance *i* belonging to a certain class z_j is computed by summing the probabilities from *i* to each instance in the training set belonging to the class z_j. The class with the highest probability is then chosen as the appropriate class for the unknown instance *i*. The robustness of the method relies upon consideration of all possible transformation paths being considered, weighted probabilities and the generalized distance between a given set and the unknown sequence by considering transformations to all instances in the set.

1.2.4 Decision trees and random forest

Decision trees follow the "divide-and-conquer" approach to classify a given data set considering several features. According to this approach, the data set used for training is iteratively divided at nodes where a particular feature is tested. This process continues until a node is reached that cannot be further sub-classified (also known as a leaf). The observed architecture of a basic classifier is reminiscent of that of an inverted tree **(figure 1.6)**. Once a decision tree has been built, unknown instances are routed down the tree evaluating the values of the corresponding features tested at the appropriate nodes.

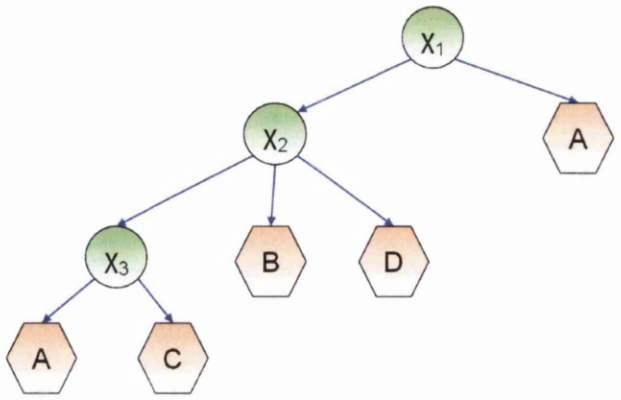

Figure 1.6. A decision tree. According to this hypothetical tree, three features are used to discriminate between four different classes.

The challenge of building effective trees is essentially to decide which feature should be tested at each node. Small trees should be preferred over long trees in order to arrive at a leaf by the shortest path. Each additional node normally introduces an error rate unless it has been proven to correctly classify all test instances. Such error is accumulative unless the subsequent nodes can "rescue" a misclassified instance. One of the most widely used algorithms based on the decision tree theory is C4.5 (Quinlan, 1993), whose corresponding implementation within the Weka platform is known as J48.

One of the drawbacks of using decision trees is the instability of the data mining technique during the training process. If small changes are introduced in the training set, a different feature can be selected at a particular node leading to different ramifications after that node. This drawback can be minimized if different decision trees are implemented from a given training set and the final prediction corresponds to the most supported class in the voting process. This is the principle underlying the random forest method (Breiman, 2001). Given a particular training set with N instances and M features, the random forest technique randomly samples the original training set into different training sets of size $n < N$. Each training set is altered by replacement, where the training set is altered by deleting and replicating randomly chosen instances (this process is also known as bagging). Once the different training sets have been assembled (yet not independent as all of them were derived from the same original training set), a decision tree is built for each training set

using a random subset of features m (where m < M) at each node. Combination of these trees decreases the generalized expected error as the variance component is reduced (theoretically, each training set used is finite and thus not fully representative and will therefore introduce an error factor while training). Therefore, by using an "out of the bag" error it is possible to estimate the generalized error of the method.

1.2.5 Support Vector Machines

Support vector machines (Cortes and Vapnik, 1995) have been implemented to apply linear solutions to non-linear problems. The principle underlying this method is to convert each instance vector (containing all extracted features) of the training set into a vector of computed components (h-dimensional space) by a non-linear function, also known as the kernel function. Each instance is then mapped onto the h-dimensional space of computed features and a maximum margin hyperplane is calculated to classify the training and unknown instances **(figure 1.7)**.

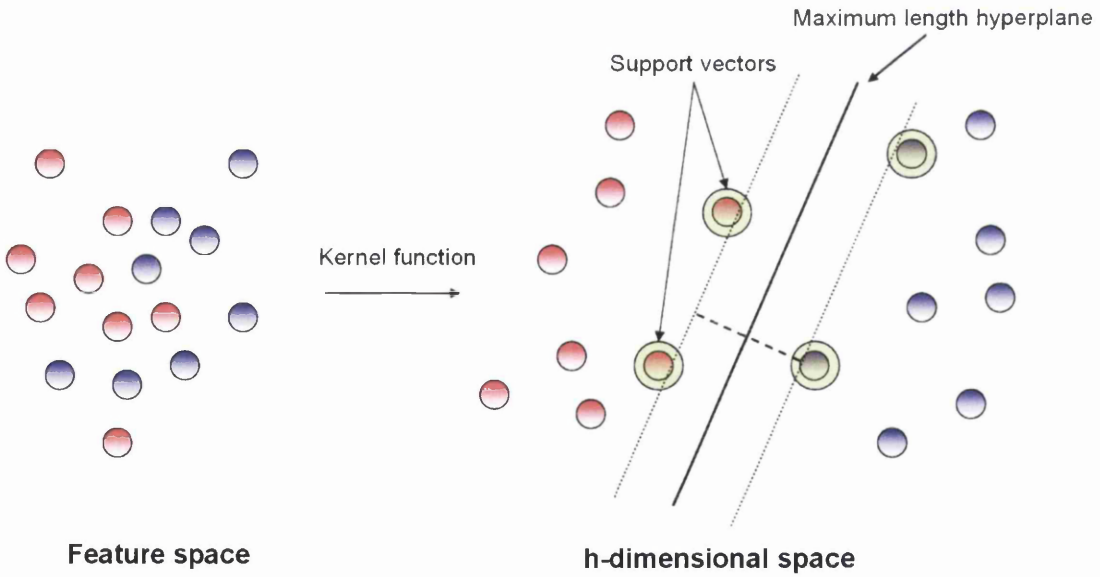

Figure 1.7. Support vector machine. Spatial representation of the feature space and the h-dimensional space created by the kernel function.

The maximum margin hyperplane is the plane within the h-dimensional space that achieves the largest distance between classes. The instances closest to the maximum margin hyperplane are called support vectors. The maximum hyperplane vector is computed with the formula:

$$\chi = b + \sum \alpha_i y_i a(i) \cdot a \qquad (1.12)$$

where i is the support vector, y_i is the class value of the training instance a(i), b and α_i are numeric values to be computed by the algorithm, a is the vector representing a test instance and a(i) corresponds to the vectors representing support vectors.

1.2.6 The radial basis function

The RBF network is a special type of neural network composed of an input layer, a hidden layer composed by units that apply the radial basis function and an output layer whose output describes the class for the test instance **(figure 1.8)**. Each hidden unit corresponds to a particular point in the input space and its output relies on the distance between the unit and the test instance. A non-linear transformation is used to calculate such distances, mainly using the Gaussian activation function (Radial basis function). The output from the hidden units is then converted through a sigmoid function as with classical neural networks. The parameters to be learned by these networks are the centres and widths of the radial basis function units and the weights used to form the linear combination of the outputs obtained from the hidden units (Witten and Frank, 2005).

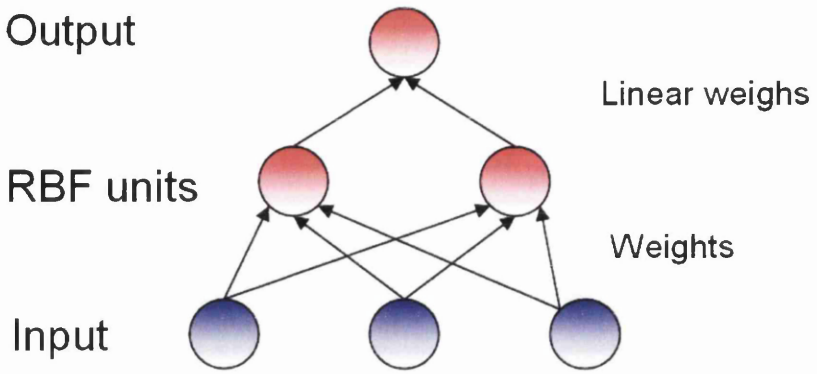

Figure 1.8. Structure of a simple RBF network. The input is composed by instances with three features, which are analyzed by two hidden units using different weights. The output from the hidden layer is used in the output layer to predict the class of the test instance.

1.2.7 MultiboostAB

The MultiboostAB (Webb, 2000) classifier belongs to the classifier committee learning method. This method generates multiple classifiers to create a committee by repeated application of a single base learning method. Each classifier within the committee votes for a particular class j in order to reach a consensus prediction (either the most voted class or the class with the highest support if assuming weighted classifiers). MultiboostAB combines two different committee learning methods, namely AdaBoost and Wagging (Webb, 2000). This combination is performed to achieve maximal reduction of: i) the bias of the base learning method and ii) the variance of the training set. The bias measures the intrinsic tendency or error rate of the classifier to predict an instance given a particular training set. On the other hand, the variance measures the expected error generated by the training set, which by nature is finite and not fully representative of the actual population of instances (Witten and Frank, 2005). The AdaBoost method is used to apply the base learning method to a set of training sets t derived from the training set T by sampling. Each instance i in the training set t_j is given a weight which is a function of the performance achieved by previous classifiers upon that particular instance i. The Wagging method works in a similar fashion but rather than weight each instance i of the training set t_j based on the performance of previous classifiers, it assigns random instance weights based on the

continuous Poisson distribution (Webb, 2000). This method is implemented by applying the Wagging method to a set of committees previously created using the AdaBoost method. By combining both AdaBoost and Wagging, MultiBoostAB inherits the ability of the AdaBoost method to reduce both the bias and the variance while further reducing the variance by means of the Wagging method.

1.2.8 Evaluation methods in data mining

Evaluating data mining methods is a key step in analyzing the classifying performance of a particular data mining method and comparing different data mining methods in order to identify the best performing data mining technique for a given training set.

1.2.8.1 Ten fold cross-validation

Evaluation of any data mining technique should be performed upon a test set rather than upon the training set. The error rate obtained using the training set is not a good indicator of the performance of the method as it tends to overestimate the classifying performance. Classifiers are "trained" using the training set so they are bound to perform better than if using a test set. Therefore, a test set, where no instances are included in the training set, is required to evaluate each data mining method. If a large set is to be used, the set can be split into two different data sets where one set can be used as the training set and the remaining set can be used as the test set. However, quite often it is not possible to assemble large sets of instances. Alternatively, methods have been implemented to create a training set and test set out of a single data set. Undoubtedly, the most common method is the *k* fold cross-validation. This method divides a given set in *k* subsets where each class is represented by the k^{th} part of the instances belonging to the corresponding class (stratification). The training set is implemented using *k*-1 subsets and the remaining subset is used as the test set. This procedure is iteratively repeated *k* times so all instances have been tested by the method. The obtained results for each fold of the validation (true positives, true negatives, false positives and false negatives) are then summed up and the overall accuracy of the data mining method is computed. The standard *k* fold cross-

validation method applied is the ten fold cross-validation as numerous analyses have shown that by using ten fold cross-validation the most accurate error estimates are obtained (Witten and Frank, 2005).

1.2.8.2 Evaluation parameters

The predictive accuracy of each data mining method is estimated based on a confusion matrix Z, which reports the results obtained from the ten fold cross-validation, that is, the number of data points belonging to the class *i* and predicted as members of the class *j*. The accuracy in predicting a particular class is defined by its sensitivity, specificity and geometric average (GAv). Sensitivity is defined as the percentage of proteins, which belong to i^{th} class and are correctly predicted, whereas the specificity is the percentage of proteins predicted as members of the i^{th} class that are correctly predicted as such. To calculate these parameters, it is necessary first to compute the number of proteins belonging to the i^{th} class (x_i), the number of proteins predicted as members of the i^{th} class (y_i), and the total number of proteins N contained in the matrix Z:

$$x_i = \sum_j z_{ij} \tag{1.13}$$

$$y_i = \sum_j z_{ji} \tag{1.14}$$

$$N = \sum_{ij} z_{ij} \tag{1.15}$$

where z_{ij} and z_{ji} belongs to a particular cell within the confusion matrix Z given 2 coordinates. Given the parameters x_i, y_i and N, sensitivity and specificity are defined as:

$$Sensitivity_{(i)} = 100 z_{ii} / N \tag{1.16}$$

$$Specificity_{(i)} = 100 z_{ii} / N \tag{1.17}$$

Both the sensitivity and specificity can be combined by calculating its geometric average (GAv), which is a useful indicator of the accuracy of the method to predict the i^{th} class.

$$GAv_{(i)} = \sqrt{Sensitivity_{(i)} \cdot Specificity_{(i)}} \tag{1.18}$$

The evaluation of the performace of the method is given by the accuracy (Q), the normalized accuracy (nQ) and the Matthews correlation coefficient (for data sets with two classes) or the generalized correlation (for data sets with three or more classes). The accuracy of the method is defined as the percentage of correctly predicted data points within the data set:

$$Q = 100\frac{\sum_i z_{ii}}{N} \tag{1.19}$$

However, the accuracy Q is often not representative of the accuracy of the method when unbalanced sets containing classes with different sizes are to be mined as it is biased towards the performance of the largest class. The normalized accuracy nQ, assuming the equiprobability for each class, has proven to be a more accurate value:

$$nQ = 100\frac{\sum_i \frac{z_{ii}}{x_i}}{K} \tag{1.20}$$

where K is the number of classes contained in the data set. The Matthews correlation coefficient is also used to measure the accuracy of the predictive method. This parameter, also called the Pearson correlation, was first applied to the prediction of protein secondary structure by Matthews (Matthews, 1975). This parameter measures linear relationships between two variables. Its values range from -1 to 1 where -1 indicates a perfect negative linear relationship and vice versa. A value of 0 shows no linear relationship, that is, a complete random relationship between two independent variables:

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FN)(TP+FP)(TN+FP)(TN+FN)}} \tag{1.21}$$

where TP, TN, FP and FN correspond to True Positive, True Negative, False Positive and False Negative. The Matthews correlation coefficient is only applicable to data sets

containing two classes. The equivalent coefficient to measure the predictive accuracy over a data set with more than two classes is the Generalized Correlation (GC) (Baldi et al., 2000):

$$GC = \sqrt{\frac{\sum_{ij}\left(z_{ij} - e_{ij}\right)^2}{\frac{e_{ij}}{N(k-1)}}} \qquad (1.22)$$

where $e_{ij} = \frac{x_i y_i}{N}$

1.2.8.3 ROC curves

The Receiver Operating Characteristic curves are used in signal detection to analyze the trade-off between the sensitivity and the specificity of a given data mining method (Witten and Frank, 2005). ROC curves are defined by the true positive rate or sensitivity on the x axis and the false positive rate or 1-specificity on the y axis **(figure 1.9)**. The graph uses a diagonal threshold, which represents the accuracy of the random classification, to distinguish between good and bad classifiers. Curves above the diagonal threshold correspond to good classifiers and *vice versa*. Following this principle, ROC curves closer to the left-hand border of the plot correspond to better classifiers where the dependency between sensitivity and specificity is minimized (an increased sensitivity does not involve a reduction in specificity). Additionally, the area under the curve (AUC) can also be computed to evaluate the accuracy of a given data mining method. An AUC of 0.5 corresponds to a random classification. The greater the area, the more accurate the classifier.

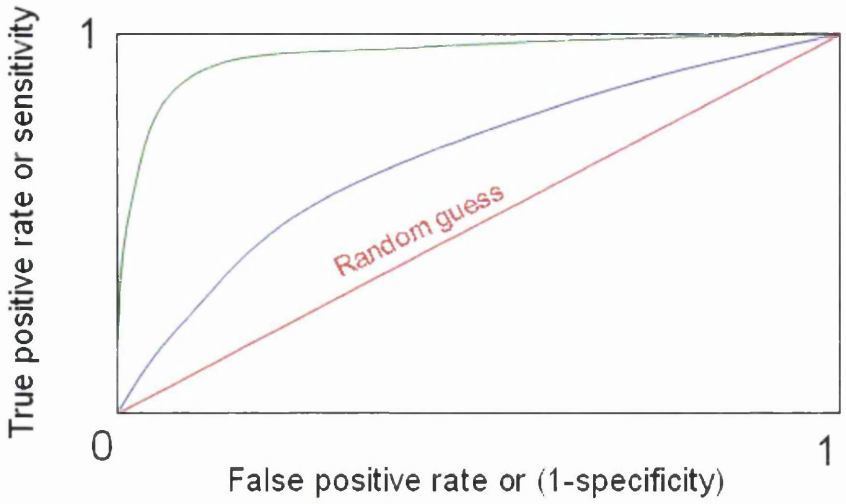

Figure. 1.9. Example of the ROC curve plot. The diagonal threshold (in red) represents the classification of instances achievable by random guessing. Classifiers above this threshold show better performance accuracy than random guessing. The figure shows two different ROC curves above the diagonal threshold. The green ROC curve shows a greater area below the curve, which represents an accurate classifier with a low trade-off between specificity and sensitivity.

1.3 References

ADAMIAN, L. & LIANG, J. (2001) Helix-helix packing and interfacial pairwise interactions of residues in membrane proteins. *J Mol Biol,* 311, 891-907.

ALBERTS, B., BRAY, D., LEWIS, J., RAFF, M., ROBERTS, K. & WATSON, J. (1994) *Molecular biology of the cell,* New York, Garland Publishing, Inc.

ALTSCHUL, S. F., MADDEN, T. L., SCHAFFER, A. A., ZHANG, J., ZHANG, Z., MILLER, W. & LIPMAN, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res,* 25, 3389-402.

ANDERSON, D. R., SWEENEY, D. J. & WILLIAMS, T. A. (1994) *Introduction of statistics. Concepts and applications,* St. Paul, West Publishing Company.

BAKER, D. & SALI, A. (2001) Protein structure prediction and structural genomics. *Science,* 294, 93-6.

BALDI, P., BRUNAK, S., CHAUVIN, Y., ANDERSEN, C. A. & NIELSEN, H. (2000) Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics,* 16, 412-24.

BAUM, L. & PETRIE, T. (1966) Statistical inference for probabilistic functions of finite state Markov chains. *Annals of Mathematical Statistics,* 37, 1554-1563.

BELRHALI, H., NOLLERT, P., ROYANT, A., MENZEL, C., ROSENBUSCH, J. P., LANDAU, E. M. & PEBAY-PEYROULA, E. (1999) Protein, lipid and water organization in bacteriorhodopsin crystals: a molecular view of the purple membrane at 1.9 A resolution. *Structure,* 7, 909-17.

BERMAN, H. M., WESTBROOK, J., FENG, Z., GILLILAND, G., BHAT, T. N., WEISSIG, H., SHINDYALOV, I. N. & BOURNE, P. E. (2000) The Protein Data Bank. *Nucleic Acids Res,* 28, 235-42.

BOYD, D., SCHIERLE, C. & BECKWITH, J. (1998) How many membrane proteins are there? *Protein Sci,* 7, 201-5.

BREIMAN, L. (2001) Random forests. *Machine Learning,* 45, 5-32.

CHURCHILL, G. A. (1989) Stochastic models for heterogeneous DNA sequences. *Bull Math Biol,* 51, 79-94.

CLEARY, J. G. & LEONARD, E. T. (1995) K*: an instance-based learner using an entropic distance measure. *12th International Conference on Machine Learning.*

CORTES, C. & VAPNIK, V. (1995) Support vector networks. *Machine Learning,* 20, 273-293.

CURRAN, A. R. & ENGELMAN, D. M. (2003) Sequence motifs, polar interactions and conformational changes in helical membrane proteins. *Curr Opin Struct Biol,* 13, 412-7.

DEBER, C. M., WANG, C., LIU, L. P., PRIOR, A. S., AGRAWAL, S., MUSKAT, B. L. & CUTICCHIA, A. J. (2001) TM Finder: a prediction program for transmembrane protein segments using a combination of hydrophobicity and nonpolar phase helicity scales. *Protein Sci,* 10, 212-9.

DUDA, R. O., HART, P. E. & STORK, D. G. (2001) *Pattern Classification,* New York, A Wiley-Interscience Publication.

DUTZLER, R., CAMPBELL, E. B., CADENE, M., CHAIT, B. T. & MACKINNON, R. (2002) X-ray structure of a ClC chloride channel at 3.0 A reveals the molecular basis of anion selectivity. *Nature,* 415, 287-94.

FRANK, E., HALL, M., TRIGG, L., HOLMES, G. & WITTEN, I. H. (2004) Data mining in bioinformatics using Weka. *Bioinformatics,* 20, 2479-81.

GARAVITO, R. M. (1998) Membrane protein structures: the known world expands. *Curr Opin Biotechnol,* 9, 344-349.

JELINEK, F. (1969) A fast sequential decoding algorithm using a stack. *IBM Journal of Research and Development,* 13, 675.

KARCHIN, R., KARPLUS, K. & HAUSSLER, D. (2002) Classifying G-protein coupled receptors with support vector machines. *Bioinformatics,* 18, 147-59.

KARP, P. D., RILEY, M., SAIER, M., PAULSEN, I. T., PALEY, S. M. & PELLEGRINI-TOOLE, A. (2000) The EcoCyc and MetaCyc databases. *Nucleic Acids Res,* 28, 56-9.

KROGH, A., LARSSON, B., VON HEIJNE, G. & SONNHAMMER, E. L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol,* 305, 567-80.

LI, J., LIU, H., DOWNING, J. R., YEOH, A. E. & WONG, L. (2003) Simple rules underlying gene expression profiles of more than six subtypes of acute lymphoblastic leukemia (ALL) patients. *Bioinformatics,* 19, 71-8.

MATTHEWS, B. W. (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta,* 405, 442-51.

NAIR, R. & ROST, B. (2005) Mimicking cellular sorting improves prediction of subcellular localization. *J Mol Biol,* 348, 85-100.

PILPEL, Y., BEN-TAL, N. & LANCET, D. (1999) kPROT: a knowledge-based scale for the propensity of residue orientation in transmembrane segments. Application to membrane protein structure prediction. *J Mol Biol,* 294, 921-35.

POPOT, J. L. & ENGELMAN, D. M. (2000) Helical membrane protein folding, stability, and evolution. *Annu Rev Biochem,* 69, 881-922.

QUINLAN, R. (1993) *C4.5: Programs for Machine Learning,* San Mateo, CA, Morgan Kaufmann Publishers.

SAIDIJAM, M., PSAKIS, G., CLOUGH, J. L., MEULLER, J., SUZUKI, S., HOYLE, C. J., PALMER, S. L., MORRISON, S. M., POS, M. K., ESSENBERG, R. C., MAIDEN, M. C., ABU-BAKR, A., BAUMBERG, S. G., NEYFAKH, A. A., GRIFFITH, J. K., STARK, M. J., WARD, A., O'REILLY, J., RUTHERFORD, N. G., PHILLIPS-JONES, M. K. & HENDERSON, P. J. (2003) Collection and characterisation of bacterial membrane proteins. *FEBS Lett,* 555, 170-5.

SCHOMBURG, I., CHANG, A. & SCHOMBURG, D. (2002) BRENDA, enzyme data and metabolic information. *Nucleic Acids Res,* 30, 47-9.

SCHULZ, G. E. (1996) Porins: general to specific, native to engineered passive pores. *Curr Opin Struct Biol,* 6, 485-90.

SINGER, S. J. & NICOLSON, G. L. (1972) The fluid mosaic model of the structure of cell membranes. *Science,* 175, 720-31.

TOBLER, J. B., MOLLA, M. N., NUWAYSIR, E. F., GREEN, R. D. & SHAVLIK, J. W. (2002) Evaluating machine learning approaches for aiding probe selection for gene-expression arrays. *Bioinformatics,* 18 Suppl 1, S164-71.

ULMSCHNEIDER, M. B. & SANSOM, M. S. (2001) Amino acid distributions in integral membrane protein structures. *Biochim Biophys Acta,* 1512, 1-14.

VON HEIJNE, G. (1996) Principles of membrane protein assembly and structure. *Prog Biophys Mol Biol,* 66, 113-39.

WALLIN, E. & VON HEIJNE, G. (1998) Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. *Protein Sci,* 7, 1029-38.

WEBB, G. (2000) MultiBoosting: A Technique to Combining Boosting and Wagging. *Machine Learning,* 40, 159-197.

WHITE, S. H. & WIMLEY, W. C. (1999) Membrane protein folding and stability: physical principles. *Annu Rev Biophys Biomol Struct,* 28, 319-65.

WITTEN, I. H. & FRANK, E. (2005) *Data Mining: Practical machine learning tools and techniques,* San Francisco, Morgan Kaufmann.

WU, C. C. & YATES, J. R., 3RD (2003) The application of mass spectrometry to membrane proteomics. *Nat Biotechnol,* 21, 262-7.

ZETH, K., DIEDERICHS, K., WELTE, W. & ENGELHARDT, H. (2000) Crystal structure of Omp32, the anion-selective porin from Comamonas acidovorans, in complex with a periplasmic peptide at 2.1 A resolution. *Structure,* 8, 981-92.

ZUBAY, G. L., PARSON, W. W. & VANCE, D. E. (1998) *Principles of Biochemistry,* Iowa, Mcgraw-Hill College Division.

CHAPTER 2

Aims and Objectives

Considering the current functional gap observed in the membrane proteome and the important roles of polytopic membrane proteins, this project has been focused on the development of computational tools for the characterization of polytopic membrane proteins. Such characterization was to be performed at three different levels: i) at the topological level, ii) at the subcellular location level and iii) at the molecular function level.

Characterization at the topological level was aimed at refining existing topological models of polytopic membrane proteins contained in the Swiss-Prot database. This refinement was based upon prediction of membrane dipping loops (so called re-entrant loops), which current topology prediction methods often fail to predict.

Subcellular location signals are mostly believed to be located in the extra-membraneous domain of polytopic membrane proteins. Similarly, the role of the transmembrane domain has often been thought to be restricted to the less function-specific roles of maintaining structure and facilitating conformational changes (with the exception, of course, of transport proteins and channels), while the extra-membranous loops of polytopic membrane proteins have been considered to play the major protein-specific functional roles such as ligand binding, chemical catalysis and signal transduction.

Our hypothesis was that the transmembrane domain must also play an important role in directing not only the subcellular location but also the molecular function of polytopic membrane proteins. Therefore, the primary aim of this study was to investigate the notion that the transmembrane domain of polytopic membrane proteins do contain important topological, organellar and functional signatures, and that these can be exploited in the development of reliable predictive bioinformatics tools, of value to the wider scientific community, particularly for those engaged in the annotation and molecular characterisation of membrane proteins.

CHAPTER 3

PROCLASS, a tool for the supervised assembly of sets of proteins: exploiting the user's molecular expertise to cluster the annotation space of proteins

3.1 Introduction

3.1.1 Data explosion

For the last two decades biological sequence databases have been growing exponentially. Since the first genome was sequenced in 1995 (Fleischmann et al., 1995) 401 new genomes have been completed (21 eukaryotes and 380 prokaryotes), 351 genomes have been drafted and another 506 genome projects are in progress (the National Center for Biotechnology Information, http://www.ncbi.nlm.nih.gov/). Although the gap between characterized protein and DNA sequence databases is constantly increasing (Janssen et al., 2003) due to the inability of high-throughput experimental techniques to cope with the information explosion obtained from current genome sequencing projects, protein databases are also growing exponentially **(figure 3.1)**.

Computational methods have been developed as a preliminary resource to help laboratory scientists in designing appropriate experiments and therefore increase the rate of functional characterization of gene products (Chapters 7 and 8 review different approaches to predict the subcellular location and functional properties of uncharacterized gene products respectively). The most popular computational approach used to annotate new gene products is the sequence similarity method, which is based on the detection of homologues and assumes that there is a strong correlation between sequence similarity and molecular function. Sequence similarity methods have proven to be useful but have limitations, which have led to the development of complementary methods for the annotation of genes and gene products based on a wide range of different techniques. The main methods can be classified into the following categories: phylogenomics, structural

genomics, orthology detection, gene context methods (i. gene fusion, ii. genomic neighbourhood, iii. similar phylogenetic pattern and iv. conserved co-expression), data mining and pattern discovery.

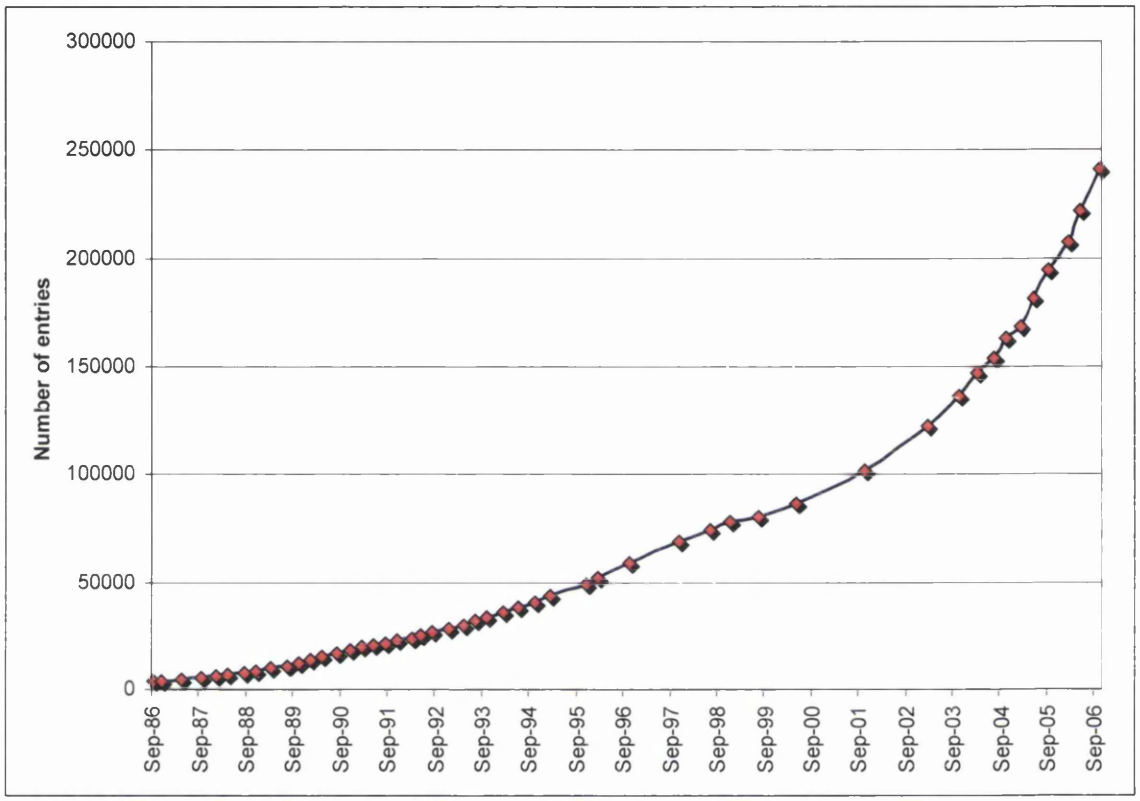

Figure 3.1. Exponential growth in number of sequences in the UniProtKB/Swiss-Prot database. The rapid increase of the database reflects the necessity of a tool to facilitate the development of manual curation of data sets to be used by different data mining and pattern discovery techniques and extract relevant biological information.

3.1.2 Text mining applied to the annotation space of proteins

Data mining and pattern discovery methods have been shown to be very successful in extracting relevant biological information from large datasets and have been succesfully applied to predicting structure, function and subcellular localization (the corresponding methods are described in chapters 5, 6, 7 and 8). These techniques are being constantly improved to increase their sensitivity and specificity and also to handle larger training sets without exponentially increasing the processing time. As the size of protein databases

increases, the chances of constructing more representative training sets also increase, which in turn can improve the accuracy of current methods **(figure 3.2)**. However, the bottleneck in this iterative loop is the development of manually curated training sets. As the size of protein databases increases, manual curation of training sets becomes a more difficult and tedious task increasing the probability of incorporating errors in a given training set.

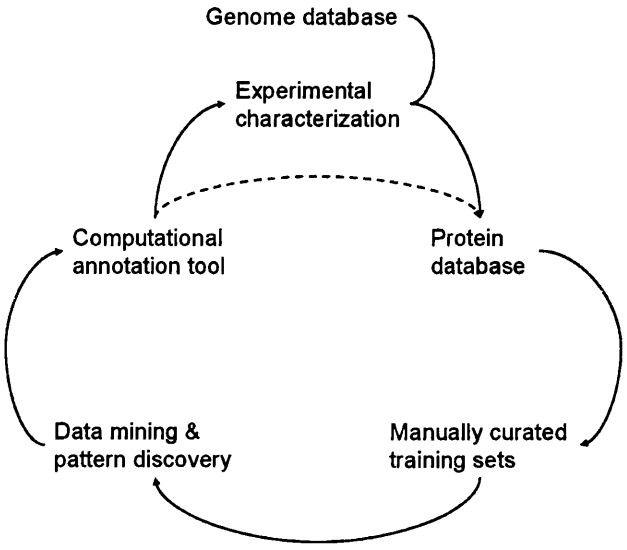

Figure 3.2. Iterative annotation process in protein databases. The information contained in protein databases is retrieved to create manually curated training sets, which lead to the implementation of predictive tools. These tools can be used to annotate or guide experimental characterization of new gene products. As the protein databases increase in size, more representative training sets can be built, which will improve the predictive accuracy of current predictive tools.

The text mining approach was primarily applied to the biomedical field in order to effectively manage and extract information from the increasing wealth of the scientific literature. This approach is still at an early stage but it has already shown promising results. The majority of these methods have been implemented to automatically extract protein-protein interaction from the rich formatted literature. A classical approach in text mining is based on the computation of co-occurrence of biological entities such as genes and proteins. Following this principle, pairs of proteins found to be similarly distributed among the scientific literature are believed to functionally interact (which does not necessarily imply physical interaction). Text mining approaches have also been used for diverse tasks such as extracting functional properties of genes and proteins, extracting gene-drugs or gene-

disease relationships and building gene association networks. In order to assess different text mining approaches, the BioCreAtIvE evaluation was used (Hirschman et al., 2005, Yeh et al., 2005). This evaluation is composed of two tasks: i) the first task is based on the extraction of gene or protein names from text and identification of their corresponding standardized gene identifiers for three model organisms (fly, mouse, yeast). ii) The second task is based on the extraction of functional annotation by requiring identification of short text passages that support Gene Ontology (Ashburner et al., 2000) annotations for particular proteins (based on full text articles).

Protein databases can be defined as collections of textual information, usually restricted by a controlled vocabulary, which can be exploited by natural language processing (Kunin and Ouzounis, 2005). Text mining approaches have already been applied to different protein databases to automatically classify proteins based on their annotation space and to detect false annotations contained in databases (Kaplan and Linial, 2005, Kunin and Ouzounis, 2005, Levy et al., 2005). The annotation space of proteins can be defined as the set of words used to describe the functional properties of proteins. A recent paper described a novel text-based method that uses concepts from statistical information retrieval (IR) and the annotation space of proteins to collect relevant information of a set of genes associated with a particular biological event (such as a specific cancer type) and provide information about properties common and unique to subsets of the gene list (Semeiks et al., 2006). Despite improvements that have been made to extract biological information from the annotation space of biological sequence databases, very little has been done to facilitate the development of manually curated data sets. The closest approach is the CLAN approach (Kunin and Ouzounis, 2005), which clusters function and sequence specific protein families that can be used to characterize a new protein sequence. At the first clustering level, protein sequences are automatically clustered using the annotation space of proteins (analyzing the description statement and the gene name statement contained in the Swiss-Prot database), and at the second clustering level these clusters are subsequently refined by clustering the sequence space of proteins using BLASTp (Altschul et al., 1997) and GeneRage (Enright and Ouzounis, 2000). This approach can be used then to characterize new gene products by multiple sequence alignment methods (sequence-to-

profile or HMM derived multiple sequence alignments). Automatic clustering methods are not suitable for the development of training sets as the probability of including misclassified proteins within the training set increases with the size of the database to be analyzed. Training sets to be used by data mining techniques should be manually curated and should contain non-redundant sets of proteins, which should be experimentally characterized when possible. Clustering sequences by similarity in order to construct sets of proteins to be mined is not a desirable option when predictive tools independent of sequence similarity are to be implemented. Recent work has shown that pattern discovery and data mining techniques can be used for the implementation of predictive tools to detect distantly related motifs (Lasso et al., 2006) and reliable prediction of subcellular location under low sequence similarity (Nair and Rost, 2005).

PROCLASS has been implemented to facilitate the development of large manually curated data sets by using the annotation space of proteins, the molecular expertise of the user and an attribute profile clustering technique (exact binary pattern matching). A protein attribute profile is a binary vector where each element describes a molecular or biological property (Semeiks et al., 2006). PROCLASS can be used to cluster proteins based on a set of terms selected by the user (elements of the protein vector). These terms have previously been extracted from the functional statements of the Swiss-Prot database. Subsequently, the obtained clusters can be individually analyzed and merged by the user through the interface if found to describe a similar biological property but using a different annotation space. PROCLASS was used to develop two manually curated data sets, a data set composed of polytopic membrane proteins located at different subcellular locations and a data set composed of polytopic membrane proteins with different molecular functions. Protein attribute clustering by exact pattern/vector matching was found to be the most reliable clustering method to guarantee the manual curation of large datasets reporting a minimum number of clusters with misclassified proteins. When the appropriate terms are to be selected this clustering method significantly reduces the protein space (all proteins with the same binary vectors are reduced into a single cluster), thus facilitating the manual curation.

PROCLASS analysis during the development of sets of proteins located at different subcellular locations showed that the annotation of this biological property in the Swiss-Prot database is not precise enough. Only 14.42% of all polytopic membrane proteins belonging to eukaryotes could be clustered. The remaining proteins were found not to be experimentally characterized, or vaguely described without specifying the corresponding organellar location. During the development of the functional dataset, PROCLASS reduced the protein space from 15,279 proteins to 1,622 clusters (where all proteins within a cluster are identical in terms of the binary vector) facilitating the manual curation of the dataset. The majority of the protein clusters obtained (97.9%) did not contain proteins belonging to different functional types (according to our definition of molecular function, which combines the molecular activity of the protein with the ligand binding specificity) validating the clustering method implemented in PROCLASS.

3.2 Methods

3.2.1 PROCLASS implementation

PROCLASS has been implemented in a Microsoft Windows environment. The programming language used for such implementation was Borland Delphi 7.0. This programming language belongs to the object oriented class of programming languages. Borland Delphi has proven to be a very useful language for Rapid Application Development (RAD). The basic architecture of PROCLASS was implemented following the model-view-controller (MVC) fashion where the interface and the functionality of the program are to be considered as different layers that are indirectly linked by a cross-linking layer represented by the TController class. This allows the programmer to carry out important changes in the interface with minor corrections in the TController class, therefore preventing the need to adapt the functionality layer to the modified interface. Other computational tools described in following chapters, namely TMLOOP **(Chapter 5)**, TMLOOP writer **(Chapter 5)**, TMDEPTH **(Chapter 6)**, TMLOCATE **(Chapter 7)** and TMFUN **(Chapter 8)**, have been implemented in a similar fashion.

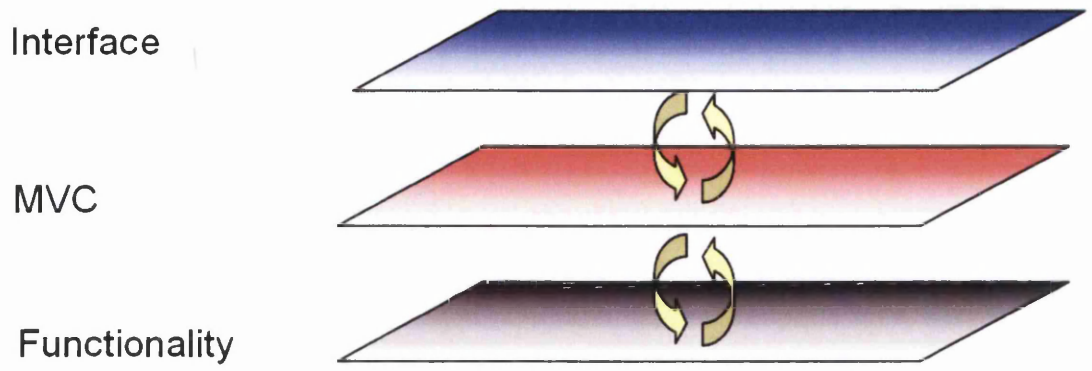

Figure 3.3. Basic architecture of a program that describes the model-view-controller (MVC). The algorithm implementation is composed of three layers. While the top layer contains the code corresponding to the interface of the program, the bottom layer contains code that refers to the functionality of the program. Both layers are linked through an intermediate layer, also known as the MVC. This layer avoids the necessity of updating of the bottom layer if the top layer is modified. All algorithms described in this thesis have been implemented using the model-view-controller.

The development of PROCLASS has been oriented towards exploiting the user's biological knowledge to build sets of proteins contained in a local version of the Swiss-Prot database or in a user-defined database of Swiss-Prot text files.

Clustering the annotation space of proteins using PROCLASS is based on five different stages:

a. The training stage
b. The terminology search stage
c. The clustering stage
d. The curation stage
e. The data set construction stage

3.2.1.1 The training stage

The training stage is a crucial stage that relies on the user's classification scheme. Therefore, in order to classify proteins contained in the same version of the Swiss-Prot database based on two different classifications schemes (i.e. subcellular location and molecular function), it is necessary to perform two separate analyses. During this stage both

the user and PROCLASS "train" each other in order to select the most appropriate terms to be searched during the terminology search stage. The user trains PROCLASS to detect equivalent terms and ignore terms not important for the classification purposes. Conversely, PROCLASS shows the user a list of terms, not previously identified by the user that could be relevant to the classification being undertaken.

The software manages two different sets of terms: i) terms specified by the user and contained in the mined statement of the database (so-called cross-linked terms) and ii) terms not specified by the user but also contained in the mined statement of the database. First PROCLASS runs a preliminary terminology search and displays, in separate "list-view" objects, the list of cross-linked and not cross-linked terms. PROCLASS is then trained to convert equivalent terms into a single term previously specified by the user (e.g. "Potassium" and "Potassium-inducible") and to include terms that are not relevant in the non-significant term list (e.g. "post-translational" and "genomic"). Protein databases often use synonyms when describing a particular feature, one of the challenges of text mining is to detect these synonyms and treat them as a single term instead of treating each synonym as a separate term. A user might find that synonyms of a particular cross-linked term are described in the list-view object corresponding to not cross-linked terms (if the synonym word has not yet been described by the user) and can easily set up a new equivalency where all synonyms will appear under a single term in the cross-linked term list-view object (e.g. "Sodium" and "Na^+"). The non-significant term list describes the terms that are not relevant for the purposes of the classification and therefore should not be considered during the terminology search stage. This dramatically decreases the processing time during the following stage.

Finally, PROCLASS displays a list of terms not considered by the user but that could be important for classification purposes. A clear example is the development of a set of functionally related protein clusters found in the membrane. No human expertise can list all terms related to all molecular functions found in the membrane and PROCLASS has proven to be an exceptional tool in facilitating the manual identification of terms important for protein classification purposes.

During the training stage, the user-defined information such as equivalencies or non-significant terms will be saved in the corresponding configuration files whose format is a conventional text format that makes it easier for the user to modify and back up files if necessary.

3.2.1.2 The terminology search stage

During this stage, PROCLASS analyzes the Swiss-Prot statements specified by the user and performs a terminology search, based on the equivalencies and non-significant terms identified during the training stage. PROCLASS can analyze the definition and/or keyword and/or subcellular location statement contained in the Swiss-Prot database according to the user's requirements. As in the training stage during the preliminary terminology search, cross-linked terms (all terms manually described by the user prior to the training stage plus those terms identified during the training stage) and terms that are not cross-linked (terms that have neither been selected as non-significant nor are relevant for classification purposes) are listed in separate list-view objects. Each list-view object contains three columns, the first column lists the term string, the second column lists the total number of times that the corresponding term is present in the database (M) and the third column lists the number of proteins containing the corresponding term (N). The term list-views can also be sorted alphabetically or numerically when clicking on each column header in order to make the identification of terms or groups of terms and selection of the terms to be used in the clustering stage easier. The user is required to delete those terms not to be used during the clustering stage, and the different clustering methods developed use all the terms listed in both the cross-linked and the non cross-linked term list-view. Terms to be deleted are normally those that will not improve the accuracy of the clustering but decrease it (it is recommended to delete those terms only present in less than 20 proteins, as groups of this size are not viable in terms of the later pattern recognition methods, unless the obtained cluster is to be clustered with other clusters, and also to delete terms that are widely used in the annotation space of proteins).

3.2.1.3 The clustering stage

PROCLASS can be used to perform three different types of clustering.

3.2.1.3.1 Manually defined protein list

By selecting the required terms (Ctrl + term to be included) and clicking on the "Make Protein List" the user creates a list of proteins that contains all the terms selected. The protein list will then be described in the protein list/cluster list view by its user-defined name, the number of proteins contained in the corresponding protein list and the list of terms used to construct the corresponding protein list.

3.2.1.3.2 Protein clustering by exact matching

Using the annotation space each protein is represented by a computed binary vector. Proteins are then clustered if represented by identical vectors. This clustering method reflects the importance of previous stages and the selection of terms to be used for protein clustering based on annotation strings. If only highly abundant terms are to be considered the obtained clusters might be too general. By contrast, if highly specific terms are considered the clustering method would lead to highly specific clusters and the risk of the number of proteins with unique binary vectors (and therefore not clustered) increases.

3.2.1.3.3 Protein clustering allowing mismatches

As in the previous clustering method, each protein is converted into a binary vector using its annotation space. First, this clustering method runs an all-against-all vector pairwise comparison and proteins are paired when the number of mismatches between vectors is equal or lower than the user-specified maximum number of mismatches (the list or pairs will be subsequently kept in memory for further processing). Once all pairs of proteins have been obtained, these pairs of proteins are clustered together only if both

members of the first pair are paired with both members of the second pair, or with all members of a cluster of pairs already made in a previous iteration. At this stage of the process, redundancy is minimized to ensure that each protein is only present in one cluster. When one particular protein is present in more than one cluster, only the cluster with the highest average term-match (total number of term matches with the remaining members of the cluster / number of proteins in the cluster) will keep the redundant term whereas the remaining clusters will no longer contain the redundant term. If two clusters show the same average term-match, both clusters will be then merged into a single cluster.

After this iterative process it is possible that proteins, previously paired with other proteins, are not included in a particular cluster. Subsequently, non-clustered proteins are compared against the obtained clusters. If the non-clustered protein has been found to be paired with all members of a cluster (checked in the list of protein pairs computed earlier) the non-clustered protein will be listed as a new member of the given cluster. Finally, a cluster refinement is achieved where clusters are merged together if each member of the first cluster is paired with all members of the second cluster (also checked in the list of protein pairs computed earlier). After the processes of recovery of non-clustered proteins and the cluster refinement, protein redundancy is minimized following the same principle explained earlier.

As with proteins lists, the obtained clusters (obtained either by exact matching or clustering allowing mismatches) will be described in the protein list/cluster list view by its user-defined name, the number of proteins contained in the corresponding protein list and the list of terms used to construct the corresponding protein list.

3.2.1.4 The curation stage

As explained earlier, the aim of PROCLASS is not to provide an automatic clustering method for the development of training sets but to facilitate the manual curation of sets of proteins. The curation stage relies entirely on the user's expertise to merge

biologically-related clusters and delete those proteins that despite being related in the annotation space do not belong to the given cluster for the purposes of that classification. After selecting a particular protein list or cluster from the protein list/cluster list-view, the user can visualize all proteins belonging to the given list or cluster, the Swiss-Prot list-view will then show all proteins belonging to the selected protein list/cluster describing their Swiss-Prot accession codes and the numbers of terms obtained during the terminology search stage. After selecting the desired proteins in the Swiss-Prot list-view, the user can read information contained in the Swiss-Prot database about each of the selected proteins in a batch mode. If a clustered protein does not follow the principles governing the protein classification and needs to be removed from the cluster, this can be easily achieved by selecting the corresponding protein in the protein list/cluster list-view and deleting it through an interface command.

A common process of the curation stage is to merge clusters that belong to the same class following the user's classification criteria. As explained above, the terms to be selected will constrain the outcome of the clustering stage. It is recommended to design the terminology search stage to obtain a higher number of clusters where it is likely that more than one cluster belongs to the same protein class, instead of obtaining a lower number of clusters where it is possible that one cluster overlaps with more than one protein class according to the user's classification criteria. PROCLASS has been designed to construct protein classes following a "bottom-up" classification approach where protein clusters are merged together to construct bigger protein classes. Protein lists or clusters can be clustered by selecting the corresponding protein list/clusters from the protein list/cluster list-view (Ctrl + protein list/cluster) and clicking on the "Merge protein lists/clusters" button. If a cluster is found to belong to more than one protein class, PROCLASS can be used to perform a "top-down" classification approach but this will result in a more time-consuming classification approach. In order to sub-classify a given cluster, the user will be required to build the cluster set (explained in the data set construction stage) and start a new classification analysis using PROCLASS.

Additionally, the user is given the option of writing a separate text file describing the name of the class and the clusters to be included within the corresponding class **(figure 3.4)**. After loading the text file (Tools > Load manually curated clusters) written by the user, PROCLASS will automatically merge the corresponding clusters and name the new class after the name specified by the user.

```
275 GROUP--->·EC·3.6.1.1·inorganic·diphosphatase¶
276 ProteinList·    0-AutoCl-838·
277 ProteinList·    0-AutoCl-908·
278 ProteinList·    0-AutoCl-508
279 ProteinList·    0-AutoCl-509
280 #¶
281 ¶
282 GROUP--->·EC·3.6.1.27·Undecaprenyl·diphosphatase¶
283 ProteinList·    0-AutoCl-40·
284 ProteinList·    0-AutoCl-39·
285 #¶
```

Figure 3.4 Example of manual curation of clusters in a separate text file. The text file must specify the class name and the clusters contained in the class in a specific format to be processed by PROCLASS.

3.2.1.5 The data set construction stage

Once the protein clusters have been merged into the protein classes designed following the user's classification criteria, PROCLASS can be used to construct a data set containing the different classes designed by the user. At this stage, a protein class is defined as a protein cluster that has been manually curated and modified if necessary (either by merging other clusters or removing undesired proteins). Firstly, the user is required to select the desired protein classes listed in the protein list/cluster list-view and then click on Tools > Build data set. PROCLASS will automatically retrieve the information contained in the database loaded in PROCLASS and save it in a text file for each protein. All proteins belonging to a particular class will be saved in a folder named after the name of the corresponding class.

PROCLASS also gives the user the option of saving the annotation analysis and clustering results in a text file format that can be re-loaded by PROCLASS in order to continue the analysis. It is recommended that the user performs several back-ups at different stages to facilitate the recovery of the analysis carried out in a previous stage.

3.2.2 PROCLASS clustering performance evaluation

PROCLASS clustering performance was evaluated with a set of 1,291 membrane proteins belonging to different functional types **(table 3.1)**. The dataset contained proteins that had been both experimentally and non-experimentally characterized. In the training stage, only the terms contained in at least 20 proteins, regardless of their relevance to describing molecular function, were selected to be searched during the terminology search stage where both the definition ('DE') and the keyword ('KW') statements were analyzed. Clustering was performed by exact pattern matching and by clustering allowing increasing numbers of mismatches. Each cluster was individually analyzed and clusters containing different protein types resulted.

Protein type	Number of proteins
Acetylcholine receptor	29
Amino acid transporter	59
Dopamine receptor	32
Glucose transporter	47
Opsin	85
Aquaglyceroporin	57
Olfactory receptor	12
Oxidoreductase	269
P-type ATPase	69
Serotonin receptor	59
Serotonin transporter	13
Binding protein dependent permease	229
ClC chloride channel	63
Sodium dicarboxylate symporter	60
Potassium channel	131
Protein channel	104
PsaF family	16

Table 3.1. Data sets of different functional types used to evaluate PROCLASS clustering methods. The data set was composed of proteins whose function was experimentally tested and proteins whose function has been elucidated by sequence similarity and *in-silico* prediction methods.

3.2.3 Development of sets of membrane proteins located at a particular subcellular compartment

The Swiss-Prot release 50.2 of 27-06-2006 was used for the development of the manually curated set of membrane proteins located at various subcellular locations. Two different sets of proteins were assembled, one for prokaryotes (archaea and eubacteria) and one for eukaryotes, based on the two empire system scientific classifications.

The Swiss-Prot database contained 97,023 eukaryotic proteins where 18,622 were transmembrane proteins. The TMDEPTH approach **(Chapter 6)** works on polytopic membrane proteins in order to compute the percentage of pairs of residues located at a similar depth in the membrane. Therefore, only transmembrane proteins with two or more transmembrane regions were isolated from the database. Using an in-house software, the Membrane Protein Data Set Creator **(figure 3.5)**, 10,896 membrane proteins containing two or more transmembrane regions were identified.

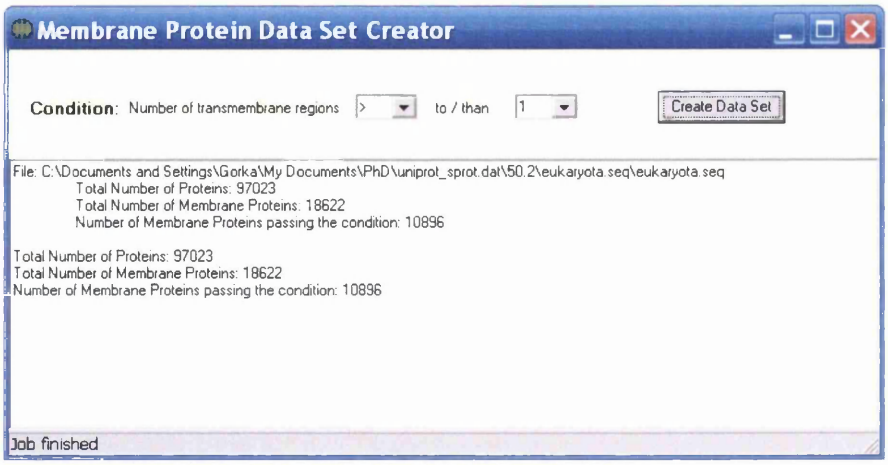

Figure 3.5. Screenshot of the Membrane Protein Data Set Creator, which filters a local copy of the Swiss-Prot database and creates a new file that contains only the proteins passing the condition defined by the user regarding the number of transmembrane regions.

During the PROCLASS training stage, the different terms were analyzed and equivalencies were set up by analyzing only the subcellular location statement of the Swiss-Prot database. The preliminary terms manually assigned were obtained from the literature

(Alberts et al., 1994): "nucleus", "mitochondria", "endoplasmic", "reticulum", "peroxisome", "golgi", "apparatus", "lysosome", "chloroplast", "vacuole", "vesicle", "microsome" and "membrane". After the analysis using PROCLASS, new terms were added to the cross-linked term list, namely "inner", "outer", "thylakoid" and "cell". "Inner" and "outer" were used to discriminate between the outer and inner membranes of the nucleus, mitochondria and chloroplast. By using the "thylakoid" term the thylakoidal membrane could be distinguished from the chloroplast membrane. The Swiss-Prot database only contained 72 eukaryotic proteins with the terms "plasma" and "membrane" **(Please see appendix A table A.2 on CD),** which did not reflect the abundance of membrane proteins in the plasma membrane. Further examination showed that the terms "cell" and "membrane" were also used to define the plasma membrane. PROCLASS was trained and equivalencies were set up for the following:

- Nucleus = nuclear
- Inner = inner-membrane
- Mitochondria = mitochondrion, mitochondrial
- Peroxisome = peroxisomal
- Lysosome = lysosomal
- Cell = cellular

The terminology search stage was carried out by analyzing the subcellular location and the keyword statement for each eukaryotic membrane protein with two or more transmembrane regions contained in the Swiss-Prot database. PROCLASS discarded those proteins containing the term "preliminary" in the accession code statement ("AC") and speculative terms (not confirmed by experimentation) in the subcellular location statement ("probable", "potential", "similarity").

3.2.4 Development of sets of membrane proteins with specific molecular functions

The Swiss-Prot release 49.7 of 16-05-2006 was used for the development of the manually curated set of functional classes of polytopic membrane proteins regardless of their taxonomic classification. The database contained 219,361 proteins where 30,469 were transmembrane proteins. As with the subcellular location data set, only α-helical membrane proteins containing two or more transmembrane regions were to be considered, due to the approach being based upon potential interhelical associations. Using the Membrane Protein Data Set Creator, 20,149 membrane proteins containing two or more transmembrane regions were isolated.

Two main challenges arose during the development of the functional set. The first challenge was to define protein function and to identify terms that would allow distinction between different functional classes of membrane proteins according to the given definition. The term "function" is a vague term that can be applied at different levels - at the biochemical level, the chemical reaction and the substrate specificity can be used to define the biochemical function of a protein, whereas at the cellular level the definition of protein function involves its interacting partners and the function carried out by the complex by virtue of its subcellular location (Skolnick and Fetrow, 2000). Further, at the physiological level, function involves the corresponding metabolic pathway or physiological role. Finally, at the phenotypic level the function of the protein accounts for its role within the totality of the organism observed by deletion or mutation of the gene encoding the protein.

For the purposes of our classification and further analysis **(Chapter 8),** protein function was defined at the biochemical level. However, within the biochemical level there are also various sub-levels of protein function. At the broadest level, membrane proteins could be classified as enzymes, transporters or receptors while at the most specific level a membrane protein (e.g. a potassium channel) can be classified into several categories such as "voltage gated potassium channel", "mechanosensitive potassium channel" or "calcium gated potassium channel". In this work, molecular function was defined as the biochemical

activity (including specific binding to ligands or structures) carried out by a protein without describing the location or the turnover of the event (Ashburner et al., 2000).

The second challenge was to identify functional terms covering most (if not all) of the known functions carried out by integral α-helical membrane proteins and that can also be used by PROCLASS to accurately cluster proteins into functional types based on the annotation space. These terms will include ligands (e.g. "calcium", "ATP"), biochemical activites (e.g. "oxidoreductase", "binding", "hydrolase", "transport") and protein types (e.g. "ATPase"). During the PROCLASS training stage, the different functional terms were analyzed and equivalencies were set up by analyzing the definition ('DE') and keyword ('KW') statement of the Swiss-Prot database. The preliminary terms manually assigned corresponded to ligands contained in the PDBsum database, version of 7th of April 2006, (Laskowski et al., 2005), which contains a list of 7,595 ligands described in the Protein Data Bank (Berman et al., 2000). PROCLASS listed, in the cross-linked term list-view, the ligand terms mentioned in the filtered Swiss-Prot database while in the not cross-linked term list-view it showed other terms included in the definition and keyword statements of the filtered Swiss-Prot database. The ligand terms listed in at least 20 different membrane proteins were listed in the final cross-linked term list. Those not cross-linked terms, believed to be important to achieve an accurate clustering based on the functional annotation space, were also listed in the final version of the cross-linked term list (to be used in the subsequent stages of the PROCLASS analysis). As with the development of the subcellular location data set, PROCLASS was trained to detect equivalent terms and ignore those not functionally relevant terms, decreasing the processing time during the terminology search stage. Finally, 344 terms were listed in the cross-linked term list during the training stage (including ligands, biochemical activities and protein types) **(Please see appendix A table A.2 on CD)**.

The terminology search stage was carried out by analyzing the definition ('DE') and the keyword ('KW') statement for each eukaryotic membrane protein with two or more transmembrane regions contained in the Swiss-Prot database. PROCLASS discarded those

proteins containing the term "preliminary" in the accession code statement ("AC") and non-experimental terms in the definition statement ("probable", "potential", and "similarity").

3.3 Results and discussion

3.3.1 PROCLASS clustering performance evaluation

PROCLASS clustering performance was evaluated by analyzing a set of membrane proteins of various functional types **(table 3.1)**. The given set contained 1,291 membrane proteins where proteins were both experimentally and computationally annotated. As explained in the method section, during the training stage, only those terms contained in at least 20 proteins were selected to be analyzed in the following stages. Terms not directly relevant to describing molecular function (e.g. "endoplasmic", "reticulum" and "postsynaptic") but contained in the description or keyword statement of at least 20 proteins were also selected.

PROCLASS analyzed 1,089 membrane proteins, which corresponded to the experimentally annotated proteins **(table 3.2)**. Clustering using exact pattern matching reported 102 clusters where only one of the clusters contained proteins belonging to different functional types. This cluster contained three proteins, one threonine/serine symporter, one proline/betaine symporter and one proton/glucose symporter, that were described by the terms "transport" and "symport". The terms "proton/glucose", "threonine", "serine" and "proline" were not found to be common enough and subsequently were not analyzed during the terminology search stage. This clustering method reduced the protein space by 7 fold as the protein space was reduced from 1,089 data points to 161 data points (102 clusters and 58 non-clustered proteins). This systematic reduction provides virtually the same confidence as if manually clusteing the original set composed by 1,089 data points. The 102 clusters were manually merged if the corresponding terms contained in each cluster were found to describe the same molecular function.

	Number of proteins	Clusters	Clusters with different protein types	Clustered proteins
Database	1291	-	-	-
PROCLASS terminology stage	1089	-	-	-
PROCLASS clustering by exact matching	1089	102	1 (0.98%)	1031
PROCLASS clustering allowing 1 mismatch	1089	57	9 (15.8%)	1087
PROCLASS clustering allowing 2 mismatches	1089	48	15 (31.25%)	1087
PROCLASS clustering allowing 3 mismatches	1089	33	10 (30.3%)	1087
PROCLASS clustering allowing 4 mismatches	1089	21	16 (76.19%)	1088

Table 3.2. Summary of the evaluation of the different clustering methods implemented in PROCLASS. Clustering by exact pattern/vector matching reported minimum error rates and proved to be the best method to facilitate the manual curation of data sets.

PROCLASS was also used to cluster proteins according to their functional annotation space allowing increasing number of mismatches between protein vectors. The number of mismatches allowed was increased gradually from one to four, but no further analyses were carried out at a higher number of mismatches as the processing time of the clustering was found to increase exponentially with the number of mismatches allowed **(figure 3.6)**. As the number of allowed mismatches increased, the number of clusters decreased but the proportion of clusters containing proteins belonging to different functional types increased **(table 3.2)**. In order to successfully cluster functionally related proteins according to the annotation space allowing mismatches, the user needs to lower the threshold of minimum number of proteins containing a given term. If the threshold is set to 20 the terms more likely to be used are those related to general functional properties (e.g. ligand and the molecular activity) of the protein and therefore clustering allowing mismatches will lead to clusters of proteins that are not functionally related. However, due to the nature of clustering allowing mismatches, it cannot be expected that this clustering method will perform as well as clustering by exact pattern or vector matching. The latter will minimize the number of clusters containing unrelated proteins and is *per se* the best method available to facilitate the manual curation of data sets.

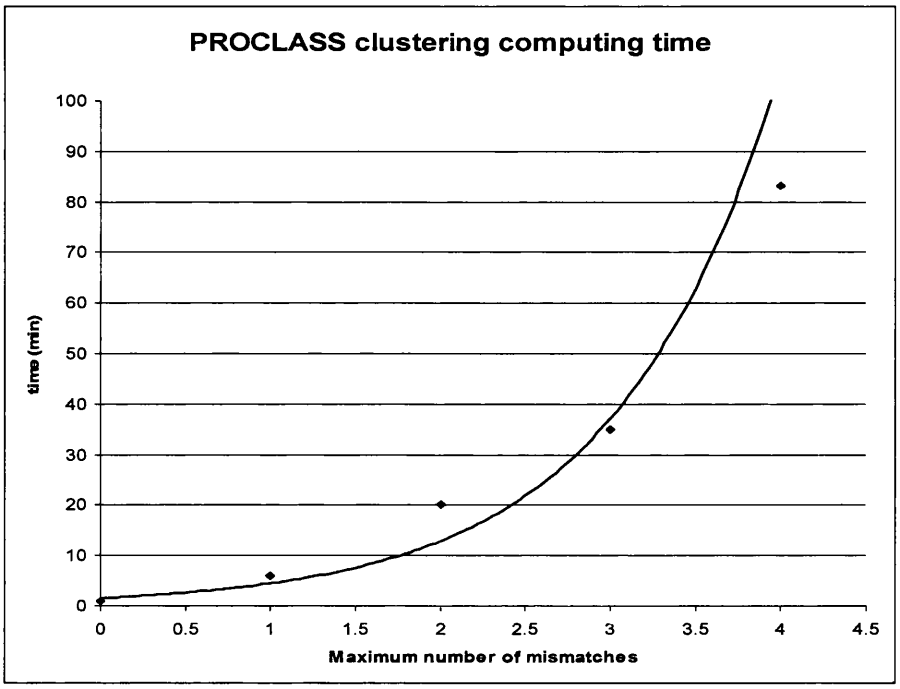

Figure 3.6. Exponential dependence of the processing time and the number of mismatches allowed in PROCLASS. An increasing number of mismatches allowed leads to an exponential increase in processing time.

3.3.2 Development of sets of membrane proteins located in a particular subcellular compartment

As explained in the method section, during the training stage, 19 different terms were identified (excluding equivalent terms) in the subcellular location statement of the Swiss-Prot database. These terms were analyzed during the terminology search stage **(Please see appendix A table A.1 on CD)** and used to construct vectors to describe the subcellular location of membrane proteins. As expected, the most common term was found to be the "membrane" term (contained in 5,374 membrane proteins) whereas the least abundant term was the term "lysosome", with only 19 proteins containing the given term. During the clustering stage, 99.4% of the proteins analyzed by PROCLASS were clustered by exact pattern or vector matching and none of the 38 clusters obtained contained proteins belonging to different subcellular locations **(Please see figure 3.7 and appendix A table A.2)**. The clustering method using exact pattern/vector matching achieved a 76-fold

reduction in the protein space showing its potential when appropriate terms are used to construct protein vectors. During the curation stage, the obtained clusters were merged if found to belong to the same subcellular location. Surprisingly, it was found that the majority of the clustered proteins belonged to clusters with undefined subcellular location **(Please see appendix A table A.3 on CD)**, 72 % of the 5,619 membrane clustered proteins clustered **(figure 3.8)** did not contain a specific subcellular location. The proteins without defined subcellular location were clustered in three different clusters **(Please see appendix A table A.3 on CD),** which were described by the terms: i) "membrane" (4,044 proteins), ii) "inner" and "membrane" (2 proteins) and iii) "outer" and "membrane" (2 proteins). Further analysis of the undefined clusters revealed that many of these proteins belong to the plasma membrane as their corresponding protein functional types are known to be essentially in that membrane (e.g. serotonin receptor). However, the subcellular location statement often refers to these proteins as "polytopic membrane proteins" without specifying their corresponding subcellular location.

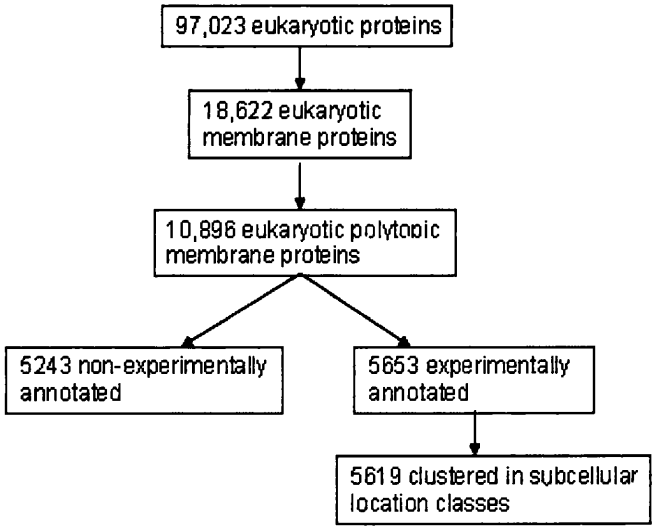

Figure 3.7. Summary of the subcellular location data collection process. The subcellular location of 48.12% of the polytopic membrane proteins isolated was found to be annotated using non-experimental methods.

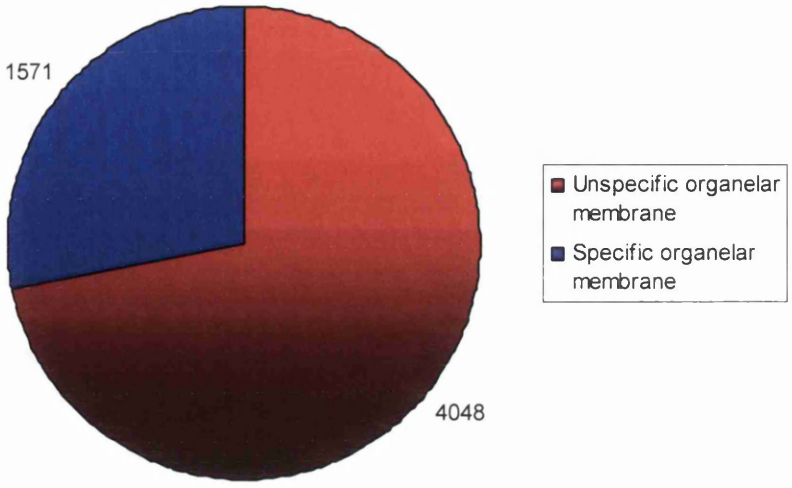

Figure 3.8. Proportion of the proteins with an undefined subcellular location annotation (red), and proteins with a defined subcellular location (blue). 72% of the proteins analyzed by PROCLASS were clustered in the "undefined membrane" cluster.

Other problems lie behind the low number of plasma membrane proteins found with PROCLASS. The cellular membrane contains a wide variety of membrane proteins performing essential tasks such as transport of ions and molecules across the lipid bilayer, cell signaling and cell to cell communication, however the number of membrane proteins found to be in the plasma membrane accounts only for 18% of the data set **(figure 3.9)**. This is a noted challenge for the generally well annotated Swiss-Prot database and it is believed that annotators working at Geneva are currently working on a new subcellular location nomenclature, which will be more precise than that of the current version (Personal communication, Swiss-Prot curators, Swiss-Prot).

Similarly, the number of polytopic nuclear membrane proteins found in the Swiss-Prot database **(Please see appendix A table A.3 on CD)** was not as high as expected considering the importance of the nuclear membrane. Nuclear membrane proteins were found to account for 1% of the membrane proteome in eukaryotes **(figure 3.9)**. Unlike plasma membrane proteins, it was not possible to find nuclear membrane proteins in the

undefined clusters obtained in the curation stage. Further research using the Sequence Retrieval System and the gene ontology database revealed that it was not possible to retrieve new sequences from the Swiss-Prot database unless proteins containing non-experimental terms ("probable", "potential", "putative", "hypothetical" or "similarity") were included.

The information contained in the Swiss-Prot database concerning subcellular organization was also quantified **(figure 3.9)**. However, this quantification does not represent the organellar organization of the membrane proteome in eukaryotes for various reasons: i) monotopic membrane proteins have not been considered in this study, ii) as explained above, organelle-specific membrane proteomes (e.g. the membrane proteome of the nucleus or plasma membrane) might be underrepresented in the database (either due to a lack of experimental data or annotation irregularities), iii) particular species might be underrepresented (as academic and industrial efforts might be biased towards specific species). Despite these problems, such quantification is still useful to assess the current informative content in the Swiss-Prot database. An interesting relationship is that found between membrane proteins belonging to the chloroplast membrane (either outer or inner) and the membrane proteins located in the thylakoid. According to the analysis carried out with PROCLASS, the thylakoidal membrane contains 8.5 times more proteins than the chloroplast membrane. These results concur with the functionality of the thylakoid as proteins involved in the electron-transport chain as well as in the photosynthetic light-absorbing system and ATP production are located in the thylakoid (Alberts et al., 1994).

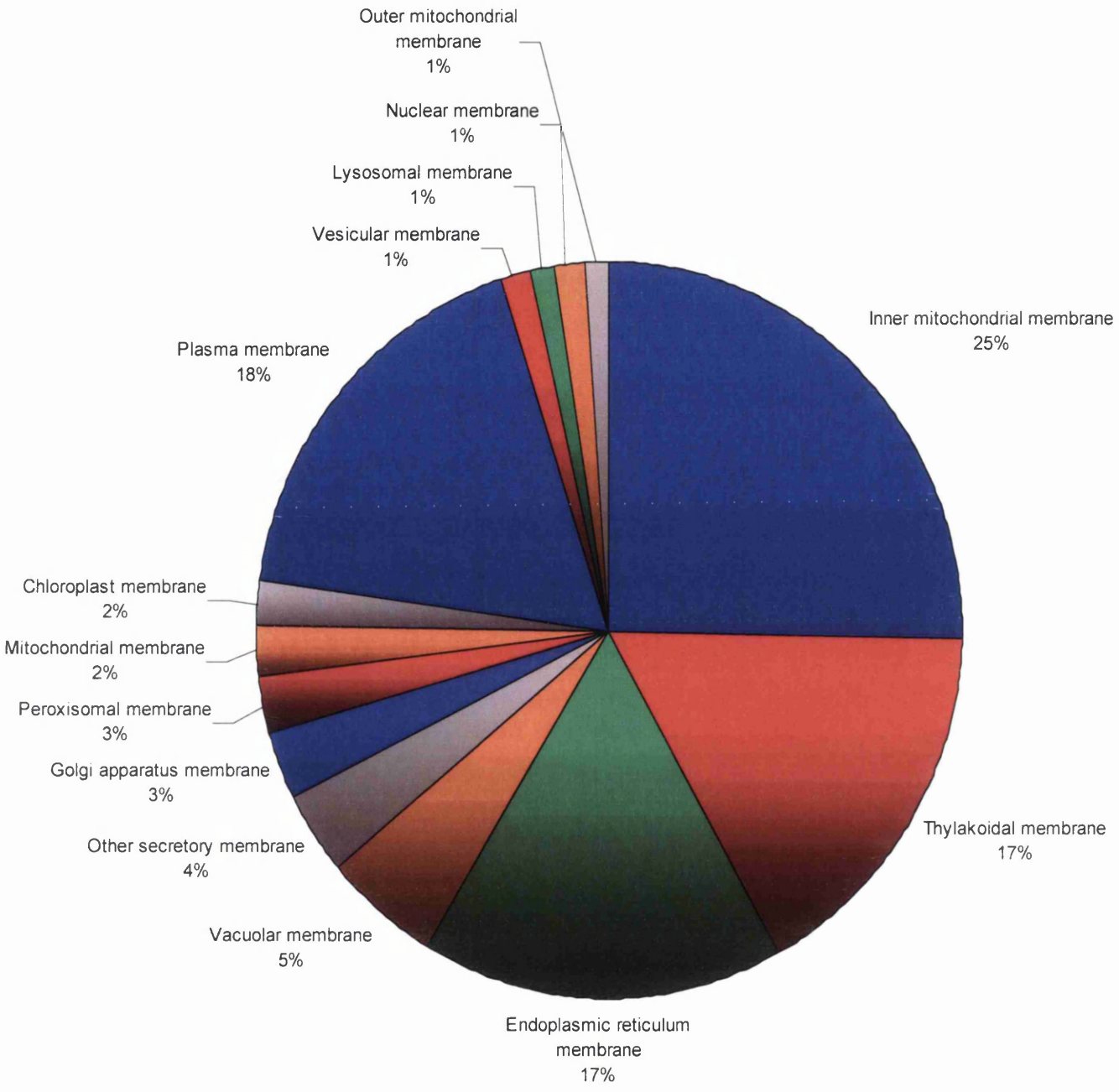

Figure 3.9. The different organelles considered in the data set. Proteins belonging to more than one subcellular location were not included in the data set. Plasma membrane proteins only accounted for 18% of the data set due to annotation irregularities in the Swiss-Prot database. The number of nuclear polytopic membrane proteins was expected to be higher but further research showed that no more proteins could be retrieved unless proteins whose subcellular location that has not been experimentally tested were included.

The annotation quality of subcellular location of membrane proteins was not as good as expected. Considering the 10,896 eukaryotic polytopic membrane proteins **(figure 3.8)**, only 5,653 membrane proteins (51.88%) contained reliable annotations (not containing non-experimental terms). PROCLASS successfully clustered 5,619 membrane proteins **(table 3.3)** according to their subcellular location statement, and further analysis with PROCLASS showed that 72 % of the 5,619 membrane proteins clustered together **(figure 3.6)** did not contain a specific subcellular location. This gap in the subcellular annotation of membrane proteins resulted in only 1,571 membrane proteins out of 10,896 being included in the data set.

The terminology search stage	
Number of searched terms	19
Number of selected terms	19
Number of proteins analyzed	5,653
Maximum number of proteins with the same term	5,374
Minimum number of proteins with the same term	19
Average number of proteins per term	436
Standard deviation	1205
The clustering stage	
Number of clusters	38
Number of proteins clustered	5,619
Maximum number of proteins within a cluster	4,044
Minimum number of proteins within a cluster	2
Average number of proteins per cluster	148
Standard deviation	655
The curation stage	
Number of manually curated clusters	29
Number of manually curated clusters belonging to single organelles	14
Number of proteins belonging to single organelles	1,532
Number of manually curated clusters belonging to multiple organelles	7
Number of proteins belonging to multiple organelles	39
Number of clusters belonging to unspecific organelles	2
Number of proteins belonging to unspecific organelles	4,048

Table 3.3. Summary of the different stages of the clustering process using PROCLASS. Clustering was achieved by exact vector/pattern matching.

3.3.3 Development of sets of membrane proteins with specific molecular functions

Constructing a manually curated data set of the main functions carried out in the membrane by polytopic membrane proteins can be a tedious task when no computational support is available. Querying databases for data set construction often requires prior knowledge of the protein type being searched and it is not guaranteed that all related proteins will be reported by the search engine implemented in the database. If general terms are to be queried, the number of proteins reported that need manual curation will make the task of creating large manually curated data sets very complicated.

During the training stage with PROCLASS, the definition ('DE') and keyword ('KW') statements were analyzed and all terms contained in these statements were reported. Because it is difficult to manually name all functions carried out in the membrane (and with the appropriate nomenclature), this is an essential stage, where the user is required to select those terms related to molecular function definition. In the terminology search stage, out of the 344 functionally related terms, 252 terms were found to be present in at least 20 or more proteins **(Please see appendix A table A.4 on CD)**. The threshold imposed to select the terms to be used to construct the protein vectors (minimum support of 20) was set up considering the clustering by exact pattern/vector matching to be used in the following PROCLASS stage. PROCLASS analyzed 15,279 proteins **(table 3.4)** out of the 20,149 listed in the data set built with the Membrane Protein Data Set Creator meaning that 75.83% of the proteins were experimentally annotated. This result contrasts with that found during the development of the subcellular location data set where only 51.88% of the proteins were experimentally annotated. In the clustering stage, proteins were clustered by exact pattern/vector matching, and the clustering method produced 1,079 clusters **(Please see appendix A table A.5 on CD)** covering 96.45% of the proteins analyzed by PROCLASS. As in previous examples, clustering by exact pattern/vector matching proved to be a particularly useful approach for the development of manually curated sets of proteins. Using this clustering method, the protein space was reduced from 15,279 to 1,622 data points (including the 1,079 clusters and the 543 proteins with unique vectors and therefore not clustered) making the manual curation of the protein space manageable.

During the curation stage, 808 clusters (comprising 9,907 proteins) were used to build the manually curated data set **(table 3.4)**. Out of the 808 clusters, 17 needed to be individually refined. The remaining 271 clusters (comprising 4829 proteins) could not be used for the final data set as they contained uncommon protein types (less than 20 proteins) such as "fatty acid transporter" or proteins whose functional definition was not precise enough (e.g. "chloroplast envelope membrane protein" or "ASC1-like protein 1, Alternaria stem canker resistance-like protein 1"). During the clustering stage, 543 proteins were not clustered as they were found to have unique vectors that could not be clustered by the clustering method performed. These proteins were manually checked and 115 proteins were successfully retrieved and included in the final manually curated data set.

The terminology search stage	
Number of terms searched	344
Number of selected terms	252
Number of proteins analyzed	15,279
Maximum number of proteins with the same term	8,148
Minimum number of proteins with the same term	20
Average number of proteins per term	275
Standard deviation	787
The clustering stage	
Number of clusters	1,079
Number of proteins clustered	14,736
Maximum number of proteins within a cluster	2,124
Minimum number of proteins within a cluster	2
Average number of proteins per cluster	14
Standard deviation	71
The curation stage	
Number of clusters included in the manually curated set	808
Number of cluster required to be refined	17
Number of clustered proteins included in the manually curated set	9,907
Number of non-clustered proteins included in the manually curated set	115
Number of clusters not included in the manually curated set	271
Number of proteins not included in the manually curated set	5,257

Table 3.4. Protein clustering summary (clustering was achieved by exact pattern/vector matching) using PROCLASS during the development of a manually curated data set of different functions carried out by polytopic membrane proteins.

The 808 clusters selected during the curation stage were cross-linked against other curated databases when possible. The clusters containing enzymes were cross-linked against the Enzyme Database (International Union of Biochemistry and Molecular Biology, http://www.chem.qmul.ac.uk/iubmb/enzyme/), which facilitates searching the IUBMB Enzyme Nomenclature List. Clusters containing enzymes were not found to contain non-enzymatic proteins or proteins with different EC numbers although it was common to find more than one cluster containing proteins with the same EC number **(Please see appendix A table A.6 on CD)**. These clusters were subsequently merged following a bottom-up approach into the six main EC numbers describing the principal biochemical activities of proteins: oxidoreductase, transferase, hydrolase, lyase, isomerase and ligase. During the bottom-up clustering approach, clusters containing the same protein sub-types (accounting for 20 or more proteins) were merged before merging all clusters belonging to a particular group, for instance the EC 1.1 group **(Please see appendix A table A.6)** was composed of nine clusters obtained by PROCLASS and within the EC 1.1 the 3-hydroxy-3methylglutaryl-coenzyme reductase was distinct as the only protein type composed of 20 or more proteins.

Similarly, G protein coupled receptors were cross-linked against the GPCRDB database (Horn et al., 1998) **(Please see appendix A table A.7 on CD)**. Out of the 808 clusters selected, 149 clusters contained GPCR proteins, 16 of those clusters contained proteins corresponding to different types of GPCR whose corresponding ligands were not common enough to be considered. These clusters were analyzed with PROCLASS with a new set of terms developed considering their definition statement during the training stage and the corresponding proteins were successfully clustered. In the case of molecular transporters, ion channels and other receptors (non-GPCR), clusters were merged when they were found to contain proteins with the same or similar ligands (e.g. sugar transporters contained not only glucose transporters but other monosaccarides such as fructose or lactose). No cluster was found with proteins binding different ligands **(Please see appendix A tables A.8-A.13 on CD)** with the exception of one cluster that contained five light driven proton pumps, one light driven chloride pump and two light driven anion pumps. The

proteins belonging to this cluster were manually analyzed and included in the corresponding cluster.

These results showed that PROCLASS (using the exact vector/pattern matching clustering method) is a suitable method for constructing manually curated data sets, as the clustering method reduces the protein space to a reasonable size and the number of proteins not correctly clustered is kept to a minimum.

The enzymatic function is the most common type of function carried out in the membrane by polytopic membrane proteins, accounting for 43% of all functions in the membrane **(figure 3.10)**. As expected, the transport activity in the membrane is the second most important function accounting for 29% of all functions in the membrane (15% for molecular transporters and 14% for ion channels). The third most common function in the membrane is the receptor function (23%) where 87% of all receptors are G protein coupled receptors reflecting the importance of these receptors in eukaryotic cells. Photosynthesis, including chlorophyll-binding proteins, is the fourth most common function (5%) followed by protein networking (0.45%) and adhesion (0.38%). These results do not reflect the distribution of protein functions in the membrane for various reasons. Firstly, analysis with PROCLASS has not been designed to distinguish between subunits, therefore functional groups in the membrane might be overrepresented when a significant number of proteins contained in the cluster are individual subunits of protein complexes. A representative example is in enzymatic function, where complexes such as cytochrome c oxidase are composed of four transmembrane polytopic subunits **(Please see appendix A table A.6 on CD)**. Secondly, membrane proteins with one transmembrane region have not been considered in this study, as the TMFUN approach is only applicable to polytopic membrane proteins. Monotopic membrane proteins are known to act as linkers in the membrane and as adhesion and receptor proteins. Thirdly, this membrane proteome cannot be considered to be representative considering the functional gap described in the membrane proteome **(Chapter 1).**

Within the enzymatic functional group, hydrolase (EC 3) and transferase (EC 2) are the most common enzymatic functions accounting for 38% and 37% of all enzymatic

functions carried out in the membrane respectively **(figure 3.11)**. The main functional hydrolase subtypes are ATPases (EC 3.6.3.1 to 3.6.3.16), ABC transporters (EC 3.6.3.17 to 3.6.3.49) and peptidases (EC 3.4). In the transferase group, the most common functional subtype is the transferases transferring phosphorus-containing groups (EC 2.7) where kinases are the best representative example **(Please see appendix A table A.1-A.5 on CD for detailed enzyme subclassification)**. Sugar transporters, amino acid transporters and protein transporters account for 20%, 19% and 17% of all molecular transporters in the membrane respectively **(figure 3.12)**. Within the ion channels group, the most common ion channel type found in the membrane are the cation channels, which account for 69% of all ion channels in the membrane **(figure 3.13)**. The potassium channel was found to be the most common ion channel found (21%) followed by sodium (9%), iron (9%), calcium (6%) and chloride (6%). These results are consistent with the most important ions for the cell. The class A of GPCR proteins was found to be the most common class (84% of all GPCR proteins) in the Swiss-Prot database **(figure 3.13)** where olfactory and peptide GPCR proteins were the most common subtypes (25% each subtype) followed by the amine GPCR (14%) **(Please see appendix A table A.6 on CD for detailed subclassification of class A GPCR)**. Within the photosynthesis function **(figure 3.14)**, chlorophyll binding proteins were the most common protein type found (24%) followed by the photosystem I P700 (22%) and the photosystem II D1 proteins (15%).

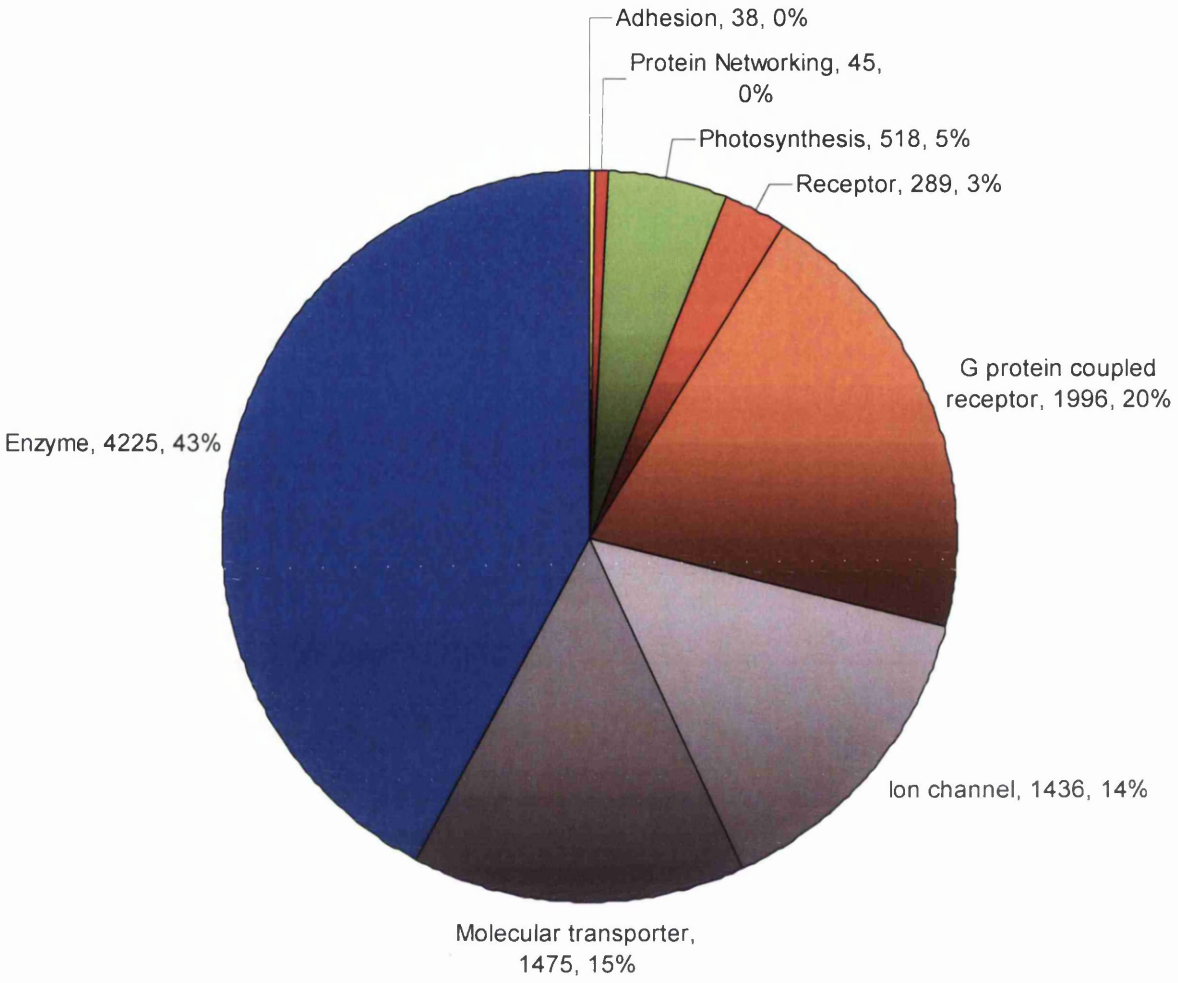

Figure 3.10. Proportions of functions carried out by polytopic membrane proteins.

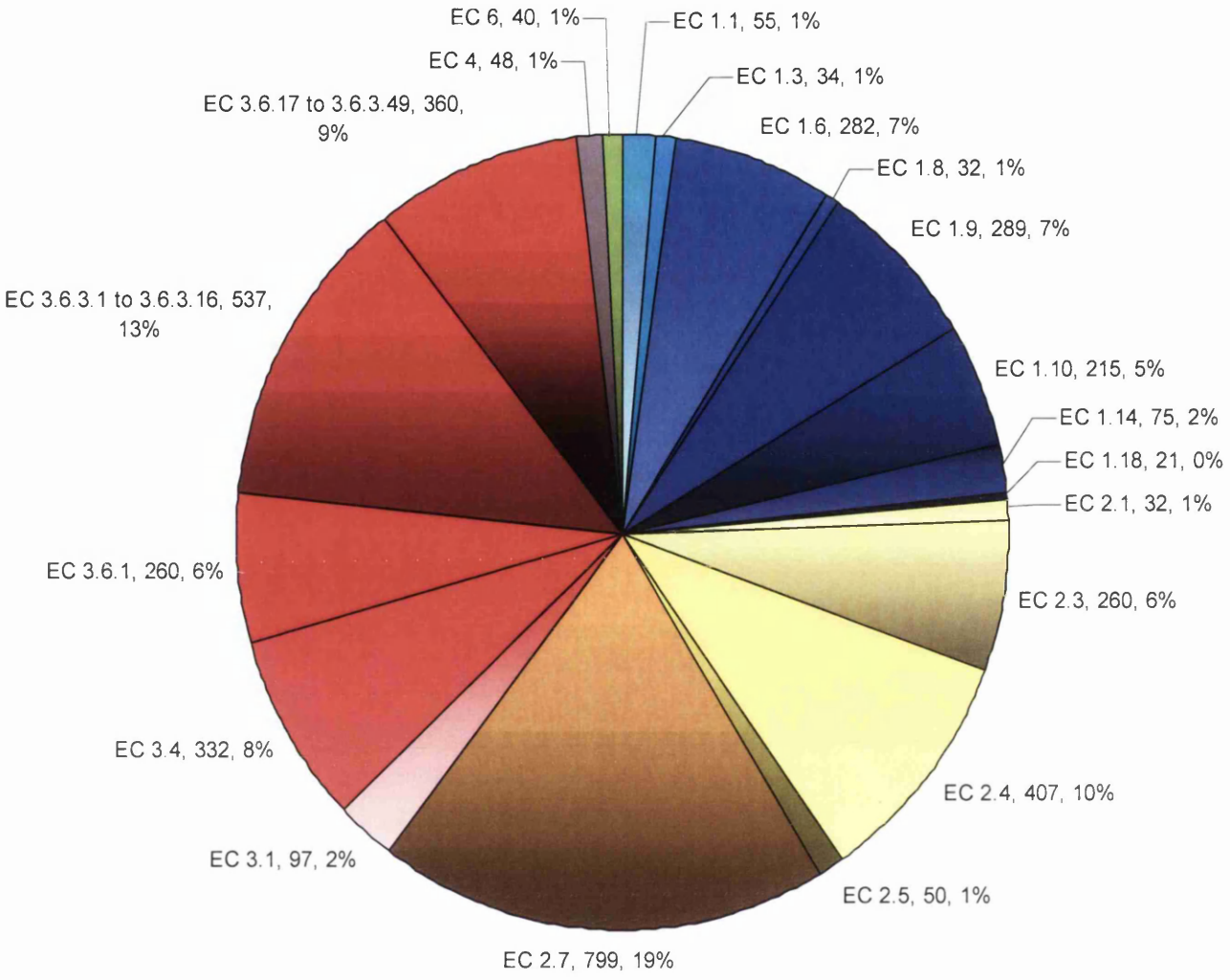

Figure 3.11. Proportions of the different enzyme polytopic membrane proteins found in the Swiss-Prot database. Sections in blue, yellow, red, grey and green correspond to oxidoreductases, transferases, hydrolases, lyases and ligases respectively.

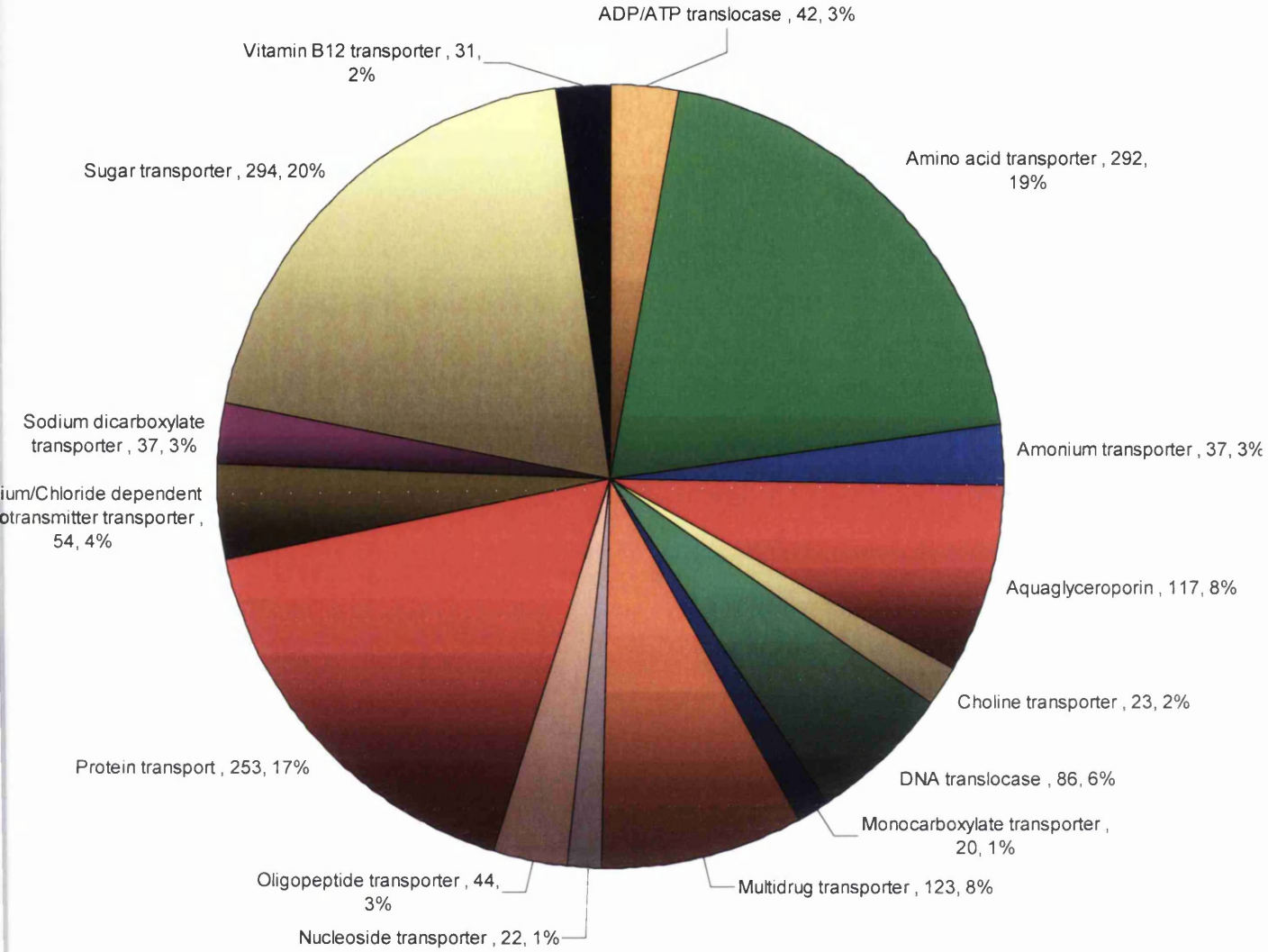

Figure 3.12. Proportions of the different transmembrane molecular transporters found in the Swiss-Prot database.

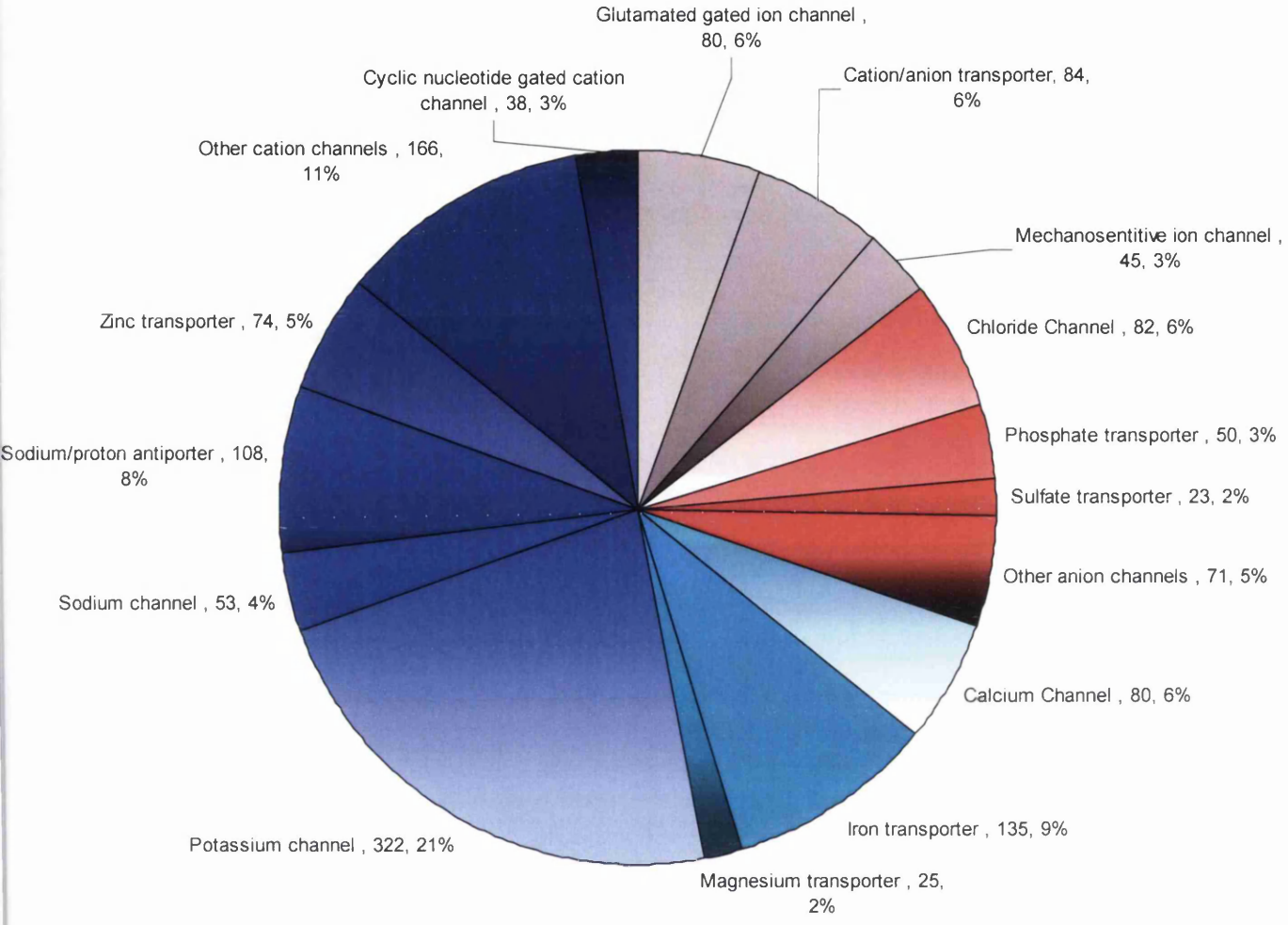

Figure 3.13. Proportions of the different ion channels found in the Swiss-Prot database. Sections in blue, red and grey correspond to cation channels, anion channels and general ion channels respectively.

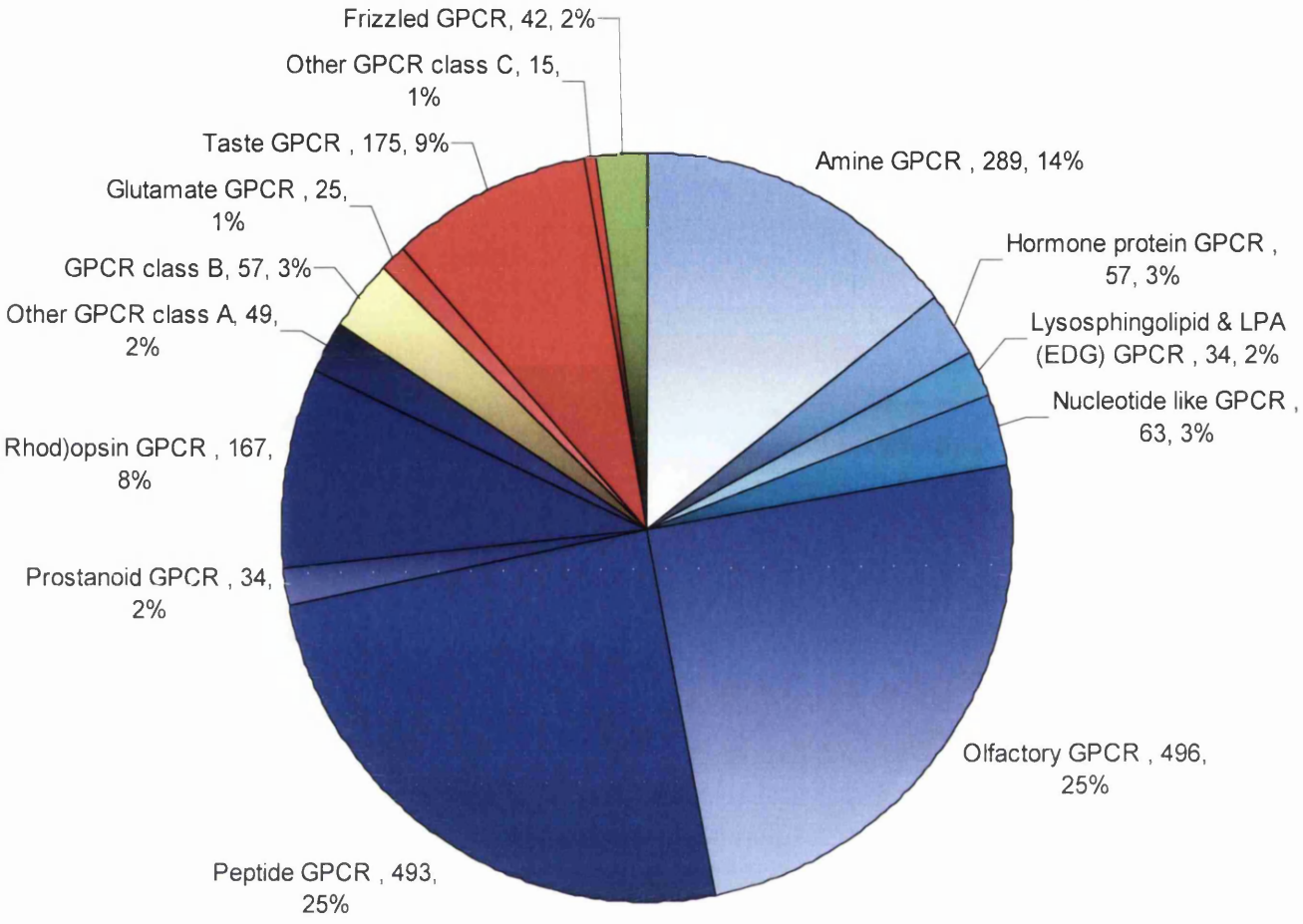

Figure 3.14. Proportions of the different classes of G protein coupled receptors found in the Swiss-Prot database. Sections coloured in blue, yellow, red and green correspond to the GPCRa, GPCRb, GPCRc and Frizzled GPCR class respectively.

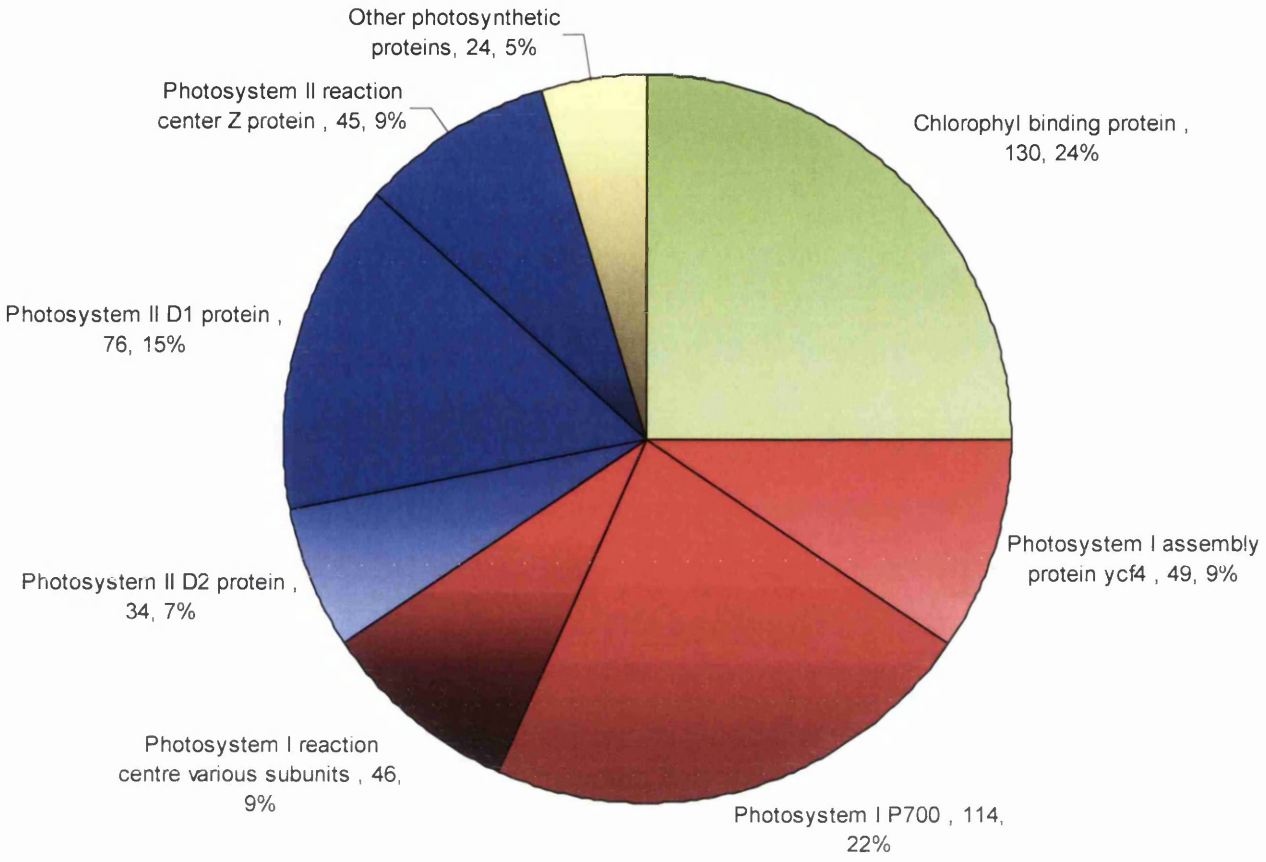

Figure 3.15. Proportions of polytopic membrane proteins related to photosynthesis found in the Swiss-Prot database. Sections coloured in red and blue correspond to photosystem I and II respectively.

3.4 Conclusions

Current experimental technologies cannot cope with the rate at, which new genomes are being sequenced. Therefore, novel computational approaches based on pattern discovery and data mining are being designed to facilitate the experimental characterization of unknown genes. However, the exponential rate at which gene and protein databases are growing also complicates the assembly of manually curated training sets to be used in computational biology. Automatic methods to assemble these training sets are usually error prone and manual curation is desired. Therefore, PROCLASS has been implemented to

facilitate the manual curation of large sets of proteins. The developed method significantly reduces the number of data points by clustering the annotation space of proteins based on exact pattern matching, thus facilitating the manual curation of the dataset. Using PROCLASS, two different data sets of polytopic membrane proteins have been assembled based on different classification schemes: i) subcellular location and ii) molecular function. Exploration of the obtained clusters showed that specific annotations contained in the Swiss-Prot database needed further refinement. Likewise, the obtained results were used to identify and classify the main functions carried out by polytopic membrane proteins without the need for prior knowledge. PROCLASS has proven to be a novel tool that should be a significant aid to those involved in the manual assembly of large sets of proteins.

3.5 References

ALBERTS, B., BRAY, D., LEWIS, J., RAFF, M., ROBERTS, K. & WATSON, J. (1994) *Molecular biology of the cell,* New York, Garland Publishing, Inc.

ALTSCHUL, S. F., MADDEN, T. L., SCHAFFER, A. A., ZHANG, J., ZHANG, Z., MILLER, W. & LIPMAN, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res,* 25, 3389-402.

ASHBURNER, M., BALL, C. A., BLAKE, J. A., BOTSTEIN, D., BUTLER, H., CHERRY, J. M., DAVIS, A. P., DOLINSKI, K., DWIGHT, S. S., EPPIG, J. T., HARRIS, M. A., HILL, D. P., ISSEL-TARVER, L., KASARSKIS, A., LEWIS, S., MATESE, J. C., RICHARDSON, J. E., RINGWALD, M., RUBIN, G. M. & SHERLOCK, G. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet,* 25, 25-9.

BERMAN, H. M., WESTBROOK, J., FENG, Z., GILLILAND, G., BHAT, T. N., WEISSIG, H., SHINDYALOV, I. N. & BOURNE, P. E. (2000) The Protein Data Bank. *Nucleic Acids Res,* 28, 235-42.

ENRIGHT, A. J. & OUZOUNIS, C. A. (2000) GeneRAGE: a robust algorithm for sequence clustering and domain detection. *Bioinformatics,* 16, 451-7.

FLEISCHMANN, R. D., ADAMS, M. D., WHITE, O., CLAYTON, R. A., KIRKNESS, E. F., KERLAVAGE, A. R., BULT, C. J., TOMB, J. F., DOUGHERTY, B. A., MERRICK, J. M. & ET AL. (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science,* 269, 496-512.

HIRSCHMAN, L., COLOSIMO, M., MORGAN, A. & YEH, A. (2005) Overview of BioCreAtIvE task 1B: normalized gene lists. *BMC Bioinformatics,* 6 Suppl 1, S11.

HORN, F., WEARE, J., BEUKERS, M. W., HORSCH, S., BAIROCH, A., CHEN, W., EDVARDSEN, O., CAMPAGNE, F. & VRIEND, G. (1998) GPCRDB: an information system for G protein-coupled receptors. *Nucleic Acids Res,* 26, 275-9.

JANSSEN, P., AUDIT, B., CASES, I., DARZENTAS, N., GOLDOVSKY, L., KUNIN, V., LOPEZ-BIGAS, N., PEREGRIN-ALVAREZ, J. M., PEREIRA-LEAL, J. B., TSOKA, S. & OUZOUNIS, C. A. (2003) Beyond 100 genomes. *Genome Biol,* 4, 402.

KAPLAN, N. & LINIAL, M. (2005) Automatic detection of false annotations via binary property clustering. *BMC Bioinformatics,* 6, 46.

KUNIN, V. & OUZOUNIS, C. A. (2005) Clustering the annotation space of proteins. *BMC Bioinformatics,* 6, 24.

LASKOWSKI, R. A., CHISTYAKOV, V. V. & THORNTON, J. M. (2005) PDBsum more: new summaries and analyses of the known 3D structures of proteins and nucleic acids. *Nucleic Acids Res,* 33, D266-8.

LASSO, G., ANTONIW, J. F. & MULLINS, J. G. (2006) A combinatorial pattern discovery approach for the prediction of membrane dipping (re-entrant) loops. *Bioinformatics,* 22, e290-7.

LEVY, E. D., OUZOUNIS, C. A., GILKS, W. R. & AUDIT, B. (2005) Probabilistic annotation of protein sequences based on functional classifications. *BMC Bioinformatics,* 6, 302.

NAIR, R. & ROST, B. (2005) Mimicking cellular sorting improves prediction of subcellular localization. *J Mol Biol,* 348, 85-100.

SEMEIKS, J. R., RIZKI, A., BISSELL, M. J. & MIAN, I. S. (2006) Ensemble attribute profile clustering: discovering and characterizing groups of genes with similar patterns of biological features. *BMC Bioinformatics,* 7, 147.

SKOLNICK, J. & FETROW, J. S. (2000) From genes to protein structure and function: novel applications of computational approaches in the genomic era. *Trends Biotechnol,* 18, 34-9.

YEH, A., MORGAN, A., COLOSIMO, M. & HIRSCHMAN, L. (2005) BioCreAtIvE task 1A: gene mention finding evaluation. *BMC Bioinformatics,* 6 Suppl 1, S2.

CHAPTER 4

Pattern discovery applied to membrane dipping loop amino acid sequences

4.1 Introduction

4.1.1 Membrane dipping (re-entrant) loops

Polytopic membrane proteins are embedded membrane proteins composed of a bundle of α-helices that completely span the membrane. These transmembrane α-helices are generally connected by extramembraneous loops of various lengths. However, crystallized structures of membrane proteins such us those belonging to the aquaglyceroporin family (Gonen et al., 2004, Harries et al., 2004, Murata et al., 2000, Ren et al., 2000, Ren et al., 2001, Savage et al., 2003, Stroud et al., 2003a, Sui et al., 2001) or potassium channels (Doyle et al., 1998, Jiang et al., 2002, Jiang et al., 2003, Kuo et al., 2003, Long et al., 2005, Nishida and MacKinnon, 2002, Zhou et al., 2001) have shown that membrane dipping loops can also interconnect α-helical transmembrane regions. These loops are characterised by their particular structure: the N-terminal section of the loop partially transverses the lipid bilayer but with the C-terminal section then returning to the same side of the membrane as the N-terminal section of the loop. These membrane dipping loops have been proposed to play key roles in the functionality of membrane proteins and it has been suggested that they play mayor roles as selectivity filters in the aquaglyceroporin family (Gonen et al., 2004, Harries et al., 2004, Murata et al., 2000, Ren et al., 2000, Ren et al., 2001, Savage et al., 2003, Stroud et al., 2003a, Sui et al., 2001), potassium channels (Doyle et al., 1998, Jiang et al., 2002, Jiang et al., 2003, Kuo et al., 2003, Long et al., 2005, Nishida and MacKinnon, 2002, Zhou et al., 2001) and chloride channels (Dutzler et al., 2002, Dutzler et al., 2003) and also act as gates of molecular pores in the glutamate homolog transporter (Locher et al., 2002) and in the protein conducting channel (Mitra et al., 2005, Van den Berg et al., 2004). Membrane dipping loops in crystallized membrane

proteins have been shown to contain a α-helical structure located either in one half of the loop or in both halves. Following these observations, membrane dipping loops can be classified in three structural categories: helix-in-turn-loop-out, loop-in-turn-helix-out, helix-in-turn-helix-out.

Detection of membrane dipping loops in the current crystallized membrane proteins has also proven to be a challenge due to the lack of information regarding the lipidic environment where the membrane proteins are embedded. Crystallographic, electron microscopy and nuclear magnetic resonance (NMR) analyses to elucidate the structure of membrane proteins have proved to be a difficult task due to the difficulties of membrane protein expression and purification (Byrne and Iwata, 2002). Membrane proteins are usually extracted from their lipidic environment and crystallized using detergent molecules that cover the hydrophobic surface of the membrane proteins (Ostermeier and Michel, 1997). Therefore, crystallized structures do not show the localization of the lipid molecules surrounding the protein (except for a few highly ordered lipid molecules) and the boundaries of the membrane can not be located (Lee, 2003).

4.1.2 The PDB_TM database

TMDET (Tusnady et al., 2004) has been implemented to predict the membrane localization based on the coordinates of crystallographic data obtained from membrane proteins. TMDET assesses first the chain type and length and builds the oligomer structure. Different membrane orientations are then calculated and the biomolecule is cut in 1Å wide slices along the normal vector. An objective function is then applied to each of the 1Å slices of the biomolecule to measure the fitness of the membrane position. This function combines 2 factors, a hydrophobic factor, which measures the relative hydrophobic membrane-exposed surface area and a structural factor, which combines three factors: the straightness factor, the turn factor and the end-chain factor (Tusnady et al., 2004).

Using TMDET, the Protein Data Bank (Berman et al., 2000) was analysed and a new database called the Protein Data Bank of Transmembrane Proteins (PDB_TM) was

created. Crystallized structures were classified into 3 categories: globular proteins, globular fragments of transmembrane proteins, and transmembrane protein (the latter was sub-classified into alpha, beta and coil proteins); and the position of the membrane boundaries as well as the transmembrane regions was also specified (Tusnady et al., 2004).

4.1.3 Sequence similarity detection methods and Pattern discovery methods

By evolution, conserved nucleotides and residues are often indicative of a common structural or functional role either at the gene or protein level. Sequence similarity detection methods have been successfully applied in fields such as gene discovery, splicing prediction, phylogenesis, protein structure and functional prediction, or gene expression analysis. Multiple sequence alignment techniques have become the routine approach to measuring sequence similarity and identifying important residues (Altschul et al., 1990, Pearson and Lipman, 1988). These alignments can be used to develop different motif representation techniques such as single (Falquet et al., 2002) or multiple motif methods (Attwood et al., 1999, Henikoff et al., 1999, Wu and Brutlag, 1995), profiles (Bucher et al., 1996) and hidden Markov models (Baldi et al., 1994, Eddy, 1996, Krogh et al., 1994). However, multiple sequence alignment methods have proved to be computationally very expensive (Wang and Jiang, 1994), and the accuracy of the alignment diminishes when distantly related sequences need to be aligned. An alternative approach is based on pattern discovery methods using an unaligned set of sequences. The problem of detecting all possible patterns in a set of sequences has also proven to be computationally expensive but heuristics and restrictions in the architecture of patterns (e.g. maximum length, number of non-wild elements) (Jonassen et al., 1995, Rigoutsos and Floratos, 1998, Sagot et al., 1995) have made it possible to analyse large set of biological sequences and discover structurally and functionally important patterns (Darzentas et al., 2005).

The TEIRESIAS algorithm (Rigoutsos and Floratos, 1998) has been implemented to discover patterns in an unaligned set of nucleotide and/or amino acid sequences. This software performs an unsupervised pattern discovery and reports

maximal patterns without enumerating the entire solution. The algorithm restricts the pattern discovery process by limiting the search to patterns with user-defined parameters: the minimum number of literals in any pattern, the maximum extent of an elementary pattern and the minimum support required for a pattern (L, W and K respectively).

This algorithm performs pattern discovery in two stages: a scanning stage and a convolution stage. During the scanning stage, TEIRESIAS identifies elementary patterns (<L,W> patterns containing exactly L residues) with sufficient support K. The convolution stage is an iteration stage where the stack of the elementary patterns discovered are sorted according to their "prefix-wise less than" character; the current top pattern is then fused with a subsequent elementary pattern that is suffix-wise less than the top pattern (patterns are fused only if the support for the fused pattern is $\geq$ K), then the fused pattern becomes the current top pattern of the stack and the process starts again. When the current top pattern can no longer be extended more to the right, the same process is applied trying to extend the current top pattern to the left (prefix-wise). When the extension in both directions of the current top pattern has finished, it is removed from the stack and reported if found to be maximal (the most informative pattern at a given support K' $\geq$ K). The convolution process starts again with the new current top pattern until no more patterns remain in the stack.

4.1.4 Prediction of membrane dipping loops using sequence pattern discovery

Despite the acknowledged importance of the membrane dipping loops in the different mechanisms of action of several crystallized membrane proteins, to our knowledge no extensive analyses have been carried out to determine conserved patterns in these motifs and identify potential functionally important residues. Membrane dipping loops contained in crystallized membrane proteins were identified using three different sources: the PDB_TM database, crystallographic literature and manual identification of proteins listed in the database of membrane proteins of known 3D structure (the Stephen White laboratory at University of California, Irvine, http://blanco.biomol.uci.edu/Membrane_Proteins_xtal.html). TEIRESIAS was applied

to sets of sequences (obtained from the Swiss-Prot database) assumed to contain similar membrane dipping loops to those previously identified and the discovered patterns were subsequently evaluated using PATTERNTEST. These patterns were then mapped onto the literature of crystallographic studies and it was found that some of the residues contained in these patterns were already reported by experimental analyses to be important residues for the functionality of the corresponding membrane protein, thus validating our approach. Additional discovered patterns also described other residues whose functional roles have not yet been fully characterized leading to the potential for targeting further experimental research aimed at understanding the exact functional roles of these residues and understanding of the mechanism of action of membrane proteins with membrane dipping loops.

4.2 Methods

4.2.1 Selection of crystallized structures with dipping loops

Crystallized membrane proteins containing membrane dipping loops were primarily identified by use of the PDB_TM database (Tusnady et al., 2004); update 24/10/05. The given database version contained 34125 PDB structures, 539 out of all the PDB structures contained in the PDB_TM database were classified as transmembrane proteins. The proteins to be used for the analysis of dipping loops were required to be alpha helical membrane proteins and PDB_TM classified 436 out of 539 membrane proteins as alpha helical membrane proteins.

All protein chains in the PDB_TM database have a record with 3 attributes: CHAINID (the chain identifier), NUM_TM (number of transmembrane regions) and TYPE (type of transmembrane regions). There are 8 different types of transmembrane regions listed in the PDB_TM database: Side1, Side2, Beta-strand, alpha-helix, coil, membrane-inside, membrane-loop and unknown localizations; listed respectively as 1, 2, B, H, C, I, L and U.

Structures with dipping loops were detected by searching in the PDB_TM database for the keyword “type="L"/>”, which corresponds to membrane dipping loops.

The structures referred to in the PDB_TM database as having dipping loops motifs are listed in **table 4.1**.

PDB acc. code	Protein definition	Loop location in sequence	
		begin	end
1BL8	KcsA K+ Channel (Chain A, B, C, D)	63	78
1BGY	Cytochrome BC1 (Chain C, O)	270	286
1EZV	Cytochrome BC1 (Chain C, L)	274	285
1F6G	KcsA K+ Channel (Chain A, B, C, D)	65	78
1FQY	Aquaporin 1 (Chain A, B, C, D)	Loop 1: 73 Loop 2: 189	Loop 1: 85 Loop 2: 201
1FX8	Glycerol facilitator glpF (Chain A, B, C, D)	Loop 1: 64 Loop 2: 199	Loop 1: 78 Loop 2: 213
1H6I	Aquaporin 1 (Chain A)	Loop 1: 72 Loop 2: 188	Loop 1: 86 Loop 2: 202
1IH5	Aquaporin 1 (Chain A, B, C, D)	Loop 1: 73 Loop 2: 189	Loop 1: 86 Loop 2: 201
1J4N	Aquaporin 1 (Chain A, B, C, D)	Loop 1: 73 Loop 2: 191	Loop 1: 88 Loop 2: 204
1J95	KcsA K+ Channel (Chain A, B, C)	66	79
1JBO	Photosystem I (Chain F)	106	122
1JVM	KcsA K+ Channel (Chain A, B, C)	66	79
1K4C	KcsA K+ Channel (Chain C, F, I, L)	63	79
1K4D	KcsA K+ Channel (Chain C, F, I, L)	63	79
1KB9	Cytochrome BC1 (Chain C)	272	288
1KPK	CLCa Chloride channel (Chain A, B)	Loop 1: 135* Loop 2: 181 Loop 3: 343* Loop 4: 392	Loop 1: 155* Loop 2: 200 Loop 3: 365* Loop 4: 415
1KPL	CLCa Chloride channel (Chain A, B)	Loop 1: 135* Loop 2: 181 Loop 3: 343* Loop 4: 392	Loop 1: 155* Loop 2: 201 Loop 3: 365* Loop 4: 416
1L7V	Vitamin B12 transport system permease protein btuC (Chain A, B)	175	186
1LDA	Glycerol facilitator glpF (Chain A, B, C, D)	Loop 1: 64 Loop 2: 199	Loop 1: 77 Loop 2: 213
1LDF	Glycerol facilitator glpF (Chain A, B, C, D)	Loop 1: 64 Loop 2: 200	Loop 1: 77 Loop 2: 213
1LDI	Glycerol facilitator glpF (Chain A, B, C, D)	Loop 1: 64 Loop 2: 200	Loop 1: 78 Loop 2: 217
1LNQ	Calcium gated K+ channel MthK (Chain A, B, C, D)	47	62
1ORQ	Voltage dependent K+ channel KvAP (Chain C, D, E, F)	183	200
1ORS	Voltage dependent K+ channel KvAP (Chain C, D, E, F)	183	200
1OTS	CLCa Chloride channel (Chain A, B)	Loop 1: 181 Loop 2: 393	Loop 1: 203 Loop 2: 415
1OTT	CLCa Chloride channel (Chain A, B)	Loop 1: 181 Loop 2: 393	Loop 1: 203 Loop 2: 415
1OTU	CLCa Chloride channel (Chain A, B)	Loop 1: 181 Loop 2: 393	Loop 1: 203 Loop 2: 415
1P7B	Inward Rectifier K+ channel KirBac1.1 (Chain A, B, C, D)	98	114

1R3I	KcsA K+ channel (Chain C, D, G, K)	63	79
1R3J	KcsA K+ channel (Chain C, F, I, L)	63	79
1R3K	KcsA K+ channel (Chain C, F, I, L)	63	78
1R3L	KcsA K+ channel (Chain C, F, I, L)	63	79
1RC2	Aquaporin Z (Chain B)	Loop 1: 60 Loop 2: 183	Loop 1: 73 Loop 2: 195
1RH5	Protein conducting channel (Chain A)	56	67
1RHZ	Protein conducting channel (Chain A)	56	68
1S33	KcsA K+ channel (Chain A, B, C, D)	63	79
1S5H	KcsA K+ channel (Chain C, F, I, L)	63	79
1S6E	Aquaporin 6 theoretical model (Chain A)	Loop 1: 79 Loop 2: 193	Loop 1: 92 Loop 2: 206
1SOR	Aquaporin 0 (Chain A, B, C, D)	Loop 1: 65 Loop 2: 182	Loop 1: 77 Loop 2: 194
1UFD	Aquaporin (Chain A, B, C, D)	Loop 1: 95 Loop 2: 205	Loop 1: 106 Loop 2: 218
1XFH	Glutamate transporter homolog (Chain A, B, C)	Loop 1: 262 Loop 2: 338*	Loop 1: 289 Loop 2: 369*
1XL4	NAD(P) Transhydrogenase (Chain A, B, C, D)	84	99
1XL6	Inward Rectifier K+ channel (Chain A, B, C, D)	84	100
1YMG	Aquaporin 0 (Chain A, B, C, D)	Loop 1: 63 Loop 2: 180	Loop 1: 78 Loop 2: 194
1YO9	Photosystem I, theoretical model psaF (Chain F)	126	141
2A79	Voltage gated K+ channel Kv1.2 (Chain B, F, J, N)	364	377
2ABM	Aquaporin Z (Chain A, B, C, D)	Loop 1: 61 Loop 2: 182	Loop 1: 70 Loop 2: 195
2AFL	KcsA K+ channel, theoretical model (Chain A, B, C, D)	38	50
2BOB	KcsA K+ channel (Chain C, F, N)	63	79
2BOC	KcsA K+ channel (Chain C, F, N)	63	78

Table 4.1. List of PDB structures (50 structures) predicted to have membrane dipping loops according to the PDB_TM database and the literature. The table also describes the function definition contained in the PDB database and the beginning and ending position for each membrane dipping loop predicted in the PDB_TM database.Loops marked with an asterisk (*) were not detected in the PDB_TM database but identified as membrane dipping loop in the corresponding literature and visually confirmed using RasMol (Sayle and Bissel, 1992).

The 50 structures that were predicted to have membrane dipping loops were classified into 12 different protein types according to the protein definition listed in the PDB database:

1. KcsA potassium channel
2. Voltage gated (KvAP) potassium channel
3. Calcium gated potassium channel
4. Inward rectifier potassium channel
5. Glutamate transporter
6. Vitamin B12 transport system permease protein

7. Cytochrome BC1
8. Photosystem I
9. Aquaporin
10. Glycerol facilitator
11. Transhydrogenase
12. Protein conducting channel

The predicted membrane dipping loops belonging to each of the membrane proteins listed in the PDB_TM database were cross-referenced to the literature corresponding to the crystallized structures. Although these papers accurately describe the three-dimensional structure of membrane proteins, the boundaries of the lipid bilayer can only be approximated, as membrane proteins need to be extracted from the membrane to elucidate their structure. Therefore, although most of the loops predicted in the PDB_TM database as membrane dipping loops were found to be described in the literature some loops were not identified and were considered as potential membrane dipping loops. Crystallographic structures described in the literature can also contain additional membrane dipping loops that were not listed in the PDB_TM database and these loops were also considered as potential loops. The latter was the case for the glutamate transporter homologue (PDB accession code 1XFH) and the ClC chloride channel (PDB accession code 1KPK, 1KPL). In addition, a manual identification of membrane dipping loops of PDB structures contained in the database "Membrane Proteins of Known 3D Structure" (the Stephen White laboratory at University of California, Irvine, http://blanco.biomol.uci.edu/Membrane_Proteins_xtal.html) was carried out to ensure that all PDB structures containing a membrane dipping loop were considered. Ultimately all the crystallized structures with membrane dipping loops were visually confirmed using the RasMol visualization software (Sayle and Bissel, 1992). The 50 structures discovered in the PDB_TM database contained 69 listed membrane dipping loops. In the literature, five additional membrane dipping loops belonging to three of the determined structures were described. No additional membrane dipping loops or structures containing this motif were manually identified. Of the 50 PDB structures considered, 46 were used in this study as membrane proteins containing membrane dipping loops. Membrane dipping loops found in the NAD(P) Transhydrogenase (PDB accession code 1XL4) and the cytochrome BC1 (PDB accession code 1BGY, 1EZV, 1KB9) could not be visually confirmed (see discussion).

The remaining membrane dipping loops were classified according to their structural arrangement into 3 different structural categories: helix-in-turn-helix-out, helix-in-turn-loop-out and loop-in-turn-helix-out.

The following table summarizes the protein types confirmed as proteins with membrane dipping loops, clusters of the different protein types with membrane dipping loops (according to the conservation of residues in the dipping loop region, and conservation of the three dimensional conformation of the membrane dipping loops) and the structural classification of the loops belonging to the different clusters.

Protein type	Protein cluster	Dipping loop structural type
Voltage gated KvAP K+ channel KcsA K+ channel Calcium gated K+ channel MthK Indward rectifier K+ channel	Potassium channel	Helix–in-loop-out
Protein Conducting Channel secY	secY / SEC61 alpha family	Helix-in-turn-loop-out
Aquaporin Glycerol facilitator glpF	Major Intrinsic Protein family	L1: Loop-in-turn-helix-out L2: Loop-in-turn-helix-out
Vitamin B12 transport system permease protein BtuC	Binding protein dependent transport system permease family	Helix-in-turn-helix-out
Clc Chloride Channel Clca	Chloride channel family	L1: Helix-in-turn-helix-out L2: Helix-in-turn-helix-out L3: Helix-in-turn-helix-out L4: Helix-in-turn-helix-out
Photosystem I psaF	psaF family	Helix-in-turn-helix-out
Glutamate transporter homolog	Sodium : dicarboxylate (SDF) symporter family. FecCD subfamily.	L1: Helix-in-turn-helix-out L2: Helix-in-turn-loop-out

Table 4.2. List of protein types containing membrane dipping loops. According to the residue conservation in the loop region and the 3-D structural similarity of the membrane dipping loops different protein types were clustered: a) Voltage gated K^+ KvAP channel, KcsA K^+ channel, Calcium gated K^+ channel MthK and Indward rectifier K^+ channel were clustered together with the Potassium channel group; b) Aquaporins and Glycerol facilitator glpF proteins were classified as aquaglyceroporins. Each loop type contained in the different dipping loop protein groups were structurally classified into 3 categories: helix-in-turn-helix-out, helix-in-turn-loop-out and loop-in-turn-helix-out.

4.2.2 Detection in the Swiss-Prot database of sequences belonging to different dipping loop protein groups and identification and isolation of sequence regions belonging to the corresponding dipping loop motif

Proteins belonging to the same families as those with membrane dipping loops were identified in the Swiss-Prot database (Boeckmann et al., 2003) regardless of their taxonomic group using the Uniprot/Swiss-Prot family/domain classification database

and by keyword search. Therefore, the 7 sets of proteins containing membrane dipping loops were assembled using the Swiss-Prot database and composed of eukaryote proteins and/or bacterial proteins and/or archaea proteins. The different sets of membrane proteins assembled were filtered at 2 different levels, the first level was based on the quality of the functional annotation and the second level was based on sequence redundancy avoidance. In the first level the Swiss-Prot annotation of retrieved membrane proteins was manually analysed in order to remove fragments and proteins with inappropriate or insufficient functional annotation. Swiss-Prot annotation often includes keywords such us "hypothetical", "probable", "potential", "by similarity" and "fragment", which are indicative of proteins whose function has been predicted by non-experimental approaches or whose sequence is not complete. Swiss-Prot files containing these keywords were not included in the set used to discover patterns.

At the second filtering level, sets of sequences belonging to particular protein groups containing membrane dipping loops were filtered based on the sequence identity of the members composing the set (Hobohm et al., 1992). A bioinformatics tool, Non-Red (Liakopoulos and colleagues, Dept. of Cell Biology and Biophysics, University of Athens, http://athina.biol.uoa.gr/bioinformatics/NON-RED), was used to avoid redundant sets of sequences. Non-Red removes pairs of sequences with similarity/homology higher than a user-defined level. In order to remove highly identical protein sequences the minimum alignment length was set to 80 (default value) and the minimum identity level was set to 95%. Therefore, pairs of sequences sharing a sequence identity of 0.95 or higher were avoided by removing the protein sequence of the given pair more similar to the remaining proteins in the set. The filtered set was defined as the gold standard set for the study **(table 4.3)**. Where the protein containing a membrane dipping loop belonged to a particular subfamily, it was important to ascertain whether the structural motif was conserved only in that particular subfamily or instead was a common feature present in other subfamilies or in the entire protein family. ClustalW (Chenna et al., 2003) was used to analyze the residue conservation in the sequence region pertaining to the dipping loop motifs across the entire protein family set. When no clear differences in residue conservation were observed between subfamilies it was taken that the membrane dipping loop was a structural motif conserved across the entire protein family. By contrast, when there was little or no conservation across the different protein subfamilies, loops were included in the pattern

discovery process as members of the particular subfamily only, as there was no evidence that the given membrane dipping loop was conserved throughout the entire protein family.

In order to detect and define the membrane dipping loop regions in sequences corresponding to proteins not yet crystallized, sequences belonging to crystallized membrane proteins and listed in the PDB_TM database were used as reference sequences. These reference sequences were found in the Swiss-Prot database by searching in the PDB files of the corresponding structure for the "DBREF" tag, which cross-links the PDB and the Swiss-Prot database. The exact localization of these motifs in sequences belonging to the same membrane dipping loop protein group as the reference proteins was ascertained by multiple sequence alignment using ClustalW (Chenna et al., 2003). Once all sequences belonging to the same dipping loop group were aligned, the dipping loop region was identified using the reference protein and the equivalent structural motif located in the remaining sequences in the alignment. To isolate the regions of sequence belonging to dipping loops, the limits of the dipping loops were obtained from the information contained in the PDB_TM database and manually verified by three dimensional visualisation using Rasmol (Sayle and Bissel, 1992). In order to minimise the error in identifying the ends of each membrane dipping loop, or possibly missing the appropriate section, 5 residues before the predicted starting position and 5 residues after the predicted ending position were also considered. Within each set, all sequences were then reduced to the region corresponding to the particular membrane dipping loop detected in the crystallized membrane protein.

Protein type	Swiss-Prot accession codes
Binding-protein-dependent-permease family, FeCD subfamily	O87656, P06609, P06972, P15029, P15030, P23876, P23877, P37737, P37738, P40410, P40411, P49936, P49937, Q6D656, Q6LQ76, Q7N3Q3, Q8D927, Q8X4L7, Q8ZDX4, Q8ZPS8, Q9KSL2, Q87Q39, Q47085, Q47086, Q56992
ClC Chloride channel	P35522, P35523, P37020, P51788, P51789, P51790, P51800, P51801, P51802, P51804, P92941, P92942, P92943, Q8D6J0, Q8FL15, Q8GX93, Q8RXR2, Q8X794, Q8ZBM0, Q8ZPK5, Q8ZRP8, Q9AGD5, Q9KM62, Q9MZT1, Q9R0A1, Q9R279, Q9TT16, Q9WU45, Q9WUB6, Q9WVD4, Q87GZ9, Q06393, Q61418, Q64347, Q96282
PsaF family family	Q9X7I4, O31127, O78457, P0A401, P0A402, P12355, P12356, P13192, P29256, P31083, P46486, P48115, P49483, P51193, P58564, Q7NH05, Q9TLW6
Sodium:dicarboxylate symporter family	O00341, O19105, O35874, O35921, O57321, O59010, P20672, P21345, P24944, P31597, P31601, P39817, P43005, P43006, P43007, P51907, P51912, P56564, P58734, P96603, Q8P5J5, Q8PGZ1, Q8X5M2, Q8XR66, Q8XUB6, Q8Y2K5, Q8Z287, Q8ZA28, Q9AAH2, Q9ANR2, Q9I4F5, Q9I713, Q9N1R2, Q9PEQ2, Q9RRG7, Q9X7K6, Q88NL9, Q95JC7, Q98AV2, Q848I3, Q885Z9, Q986R8, Q01857, Q15758, Q25605, Q95135
Potassium channels	AAA95997, AAC26099, gi2570854, O02670, O08962, O18965, O27564, O43525, O43526, O54853, O70339, O70507, O70617, O73606, O73925, O88758, O88943, P0A334, P08510, P16388, P17658, P17970, P17971, P17972, P19024, P22001, P22459, P22460, P22462, P22739, P25122, P35560, P48048, P48544, P48547, P48548, P50638, P51787, P52186, P52187, P52192, P56696, P58126, P63251, P79197, P92960, P97414, Q4TZY1, Q5JUK3, Q5NVJ6, Q8AYS8, Q8GXE6, Q8JZN3, Q8K3F6, Q8QFV0, Q8TAE7, Q8TDN1, Q8TDN2, Q9ER47, Q9H252, Q9JKA8, Q9JM63, Q9M8S6, Q9MZS1, Q9NPI9, Q9NR82, Q9NS40, Q9NSA2, Q9NZV8, Q9P1Z3, Q9QWS8, Q9QYU3, Q9QZ65, Q9R1T9, Q9TSZ3, Q9TT17, Q9TV66, Q9UIX4, Q9UJ96, Q9UL51, Q9ULD8, Q9ULS6, Q9UNX9, Q9UQ05, Q9WVJ0, Q9Y3Q4, Q9YDF8, Q9Z0V1, Q9Z0V2, Q9Z258, Q9Z307, Q9Z351, Q90ZC7, Q91ZF1, Q94A76, Q95L11, Q95V25, Q96KK3, Q96L42, Q920E3, Q02280, Q03717, Q03719, Q03720, Q03721, Q05037, Q09470, Q12791, Q14003, Q14500, Q14721, Q15756, Q28293, Q38849, Q38898, Q38998, Q39128, Q57603, Q61423, Q61743, Q61762, Q61923, Q62897, Q62976, Q63099, Q63472, Q63664, Q63734, Q63959, Q64198, Q64273, Q90854, Q92806, Q92953
Aquaglyceroporin	O14520, O43315, O54794, O62735, O94778, P06624, P11244, P18156, P22094, P23900, P25818, P26587, P29972, P30301, P34080, P37451, P41181, P42767, P43286, P43287, P47862, P47863, P47864, P47865, P48838, P50501, P51180, P53386, P55064, P55087, P56401, P56403, P56404, P56405, P56627, P60844, P93004, Q8LFP7, Q8VZW1, Q9C4Z5, Q9LA79, Q9WTY0, Q96PS8, Q02013, Q06019, Q06611, Q08733, Q23808, Q51389
secY / SEC61 alpha family	O51451, Q9ZJS9, O08387, Q99S39, Q8CNF3, Q05217, P16336, Q05207, P38375, P58118, P47416, Q59548, O52351, P10250, O33006, P0A5Z3, P38376, P43416, P49977, Q59912, Q59916, P33108, Q9PJN1, Q9Z7S5, P49976, P78283, P43804, P57571, Q8K969, Q89A85, Q9ZCS5, O66491, O63066, Q9XQU4, P93690, Q38885, P0A4H1, P77964, P28527, P38397, P51297, Q37143, P46249 P25014, P28540, P49461, Q60175, P28541, O59442. Q9V1V8, Q8U019, O26134, P28542, Q977V3, Q9HPB1, O28377, P32915, Q6FRY3, Q6CPY9, Q752H7, Q6BN08, Q9P8E3, Q96TW8, P78979, Q870W0, P79088, Q5R5L5, Q8AY33, Q9JLR1, Q25147, P38379, P49978, Q9UX84, Q9YDD0, P38353

Table 4.3. Gold standard set for each protein type with at least one protein member containing a membrane dipping loop in its crystallized structure.

4.2.3 TEIRESIAS analysis

Using the different alignments, as described above, the membrane dipping loop regions in sequence were identified and isolated. Each set of sequences belonging to a membrane dipping loop protein group was composed of partial sequences that corresponded to a particular structural motif. These sets of sequences were used as the input for the analysis using TEIRESIAS (Rigoutsos and Floratos, 1998).

The TEIRESIAS algorithm uses 3 different parameters to detect patterns within sequences:

- L: This parameter controls the minimum number of literals in any pattern. Following the recommendations by Rigoutsos and Floratos (1998), this parameter was set to 3 for all the analyses as it was found to be the smallest value for which the pattern recognition engine benefits by the convolution step.
- W: This parameter controls the maximum extent of an elementary pattern. This pattern was set up to a maximum value for each membrane dipping loop analysis corresponding to the length of the membrane dipping loop sequence (including the five additional residues included at both sides). Therefore the pattern discovery search was maximized to the length of the membrane dipping loop detecting conserved pairs of residues located in different halves of the structural motif but that may be spatially associated in the membrane.
- K: This parameter describes the support for a pattern, that is, the minimum number of sequences (if seq_version parameter enabled) / times (if seq_version parameter disabled) in which a pattern appears. It was believed that conserved patterns, eligible to be used as predictive rules of a given membrane dipping loop **(Chapter 5)** should have at least a support of 70%. The seq_version parameter was enabled for all the different analyses carried out and the K parameter was set up to the number of sequences corresponding to 70% of the sequences contained in the gold standard set used for the TEIRESIAS analysis. Therefore, all the patterns listed by TEIRESIAS were found to be present in at least 70% of all the sequences used as an input for the TEIRESIAS analysis.

The TEIRESIAS software can perform exact discovery of pattern within a set of sequences and/or discovery of patterns within a set of sequences using equivalency sets. The Bioinformatics & Pattern Discovery group working at IBM Watson Research Centre (www.research.ibm.com/**bioinformatics**/) have implemented 2 different equivalency sets of amino acids for the analysis using TEIRESIAS. One equivalency set was implemented based on the chemical nature of amino acids **(table 4.4)** and a second equivalency set was developed based on the structural nature of amino acids **(table 4.5)**.

For each set containing a particular membrane dipping loop, 3 analyses were carried out using the TEIRESIAS algorithm: exact pattern discovery, pattern discovery using chemical equivalencies and pattern discovery using structural equivalencies **(table 4.6)**.

Alanine ↔ Glycine
Aspartic acid ↔ Glutamic acid
Phenylalanine ↔ Tyrosine
Lysine ↔ Arginine
Isoleucine ↔ Leucine ↔ Methionine ↔ Valine
Glutamine ↔ Asparagine
Serine ↔ Threonine

Table 4.4. Equivalency set based on the chemical nature of amino acids (the Bioinformatics & Pattern Discovery group, http://cbcsrv.watson.ibm.com/Tspd.html).

Cysteine ↔ Serine
Aspartic acid ↔ Leucine ↔ Asparagine
Glutamic acid ↔ Glutamine
Phenylalanine ↔ Histidine ↔ Tryptophan ↔ Tyrosine
Isoleucine ↔ Threonine ↔ Valine
Lysine ↔ Methionine ↔ Arginine

Table 4.5. Equivalency set based on the structural nature of amino acids (the Bioinformatics & Pattern Discovery group, http://cbcsrv.watson.ibm.com/Tspd.html).

Gold standard sets			TEIREISIAS pattern discovery parameters			
Protein group	**Loop type**	**Number of sequences**	**L**	**W**	**K (70%)**	**Pattern discovery method**
Potassium channel	Helix-in-turn-loop-out	134	3	25	94	Exact discovery Chemical equivalency Structural equivalency
secY / SEC61 alpha family	Helix-in-turn-loop-out	75	3	30	52	Exact discovery Chemical equivalency Structural equivalency
Major Intrinsic Protein family	L1: Loop-in-turn-helix-out L2: Loop-in-turn-helix-out	49 49	3 3	20 20	34 34	Exact discovery Chemical equivalency Structural equivalency
Binding protein dependent transport system permease family	Helix-in-turn-helix-out	25	3	30	17	Exact discovery Chemical equivalency Structural equivalency
Chloride channel family	L1: Helix-in-turn-helix-out L2: Helix-in-turn-helix-out L3: Helix-in-turn-helix-out L4: Helix-in-turn-helix-out	35 35 35 35	3 3 3 3	30 30 30 30	24 24 24 24	Exact discovery Chemical equivalency Structural equivalency
psaF family	Helix in helix out	16	3	25	11	Exact discovery Chemical equivalency Structural equivalency
Sodium : dicarboxylate (SDF) symporter family	L1: Helix in helix out L2: Helix-in-turn-loop-out	46 46	3 3	30 30	37 37	Exact discovery Chemical equivalency Structural equivalency

Table 4.6. Summary of the different pattern discovery analyses carried out using TEIREASIAS. Columns one to three summarize the different gold standard sets used. Columns four to seven summarize the pattern discovery process carried out using TEIRESIAS.

4.2.4 Validating Patterns obtained by TEIRESIAS using PATTERNTEST

Patterns detected by TEIRESIAS were not guaranteed to be specific to the corresponding loop type belonging to a particular gold standard set as it was not possible to include negative control sets in the pattern discovery process. Therefore, it might be possible to discover patterns from one particular dipping loop set in other sets of membrane proteins, whose structure does not actually contain a membrane dipping loop, leading to patterns with poor specificity and the incorrect prediction of false positives. To validate the patterns, an additional tool was implemented, named PATTERNTEST, whose function was to validate the patterns obtained using TEIRESIAS against positive and negative control sets assembled by the user. Two separate negative control sets were assembled, the first negative control set was composed of protein sequences belonging to the remaining sets of membrane dipping

loop motifs, whereas the second negative control set was composed of 363 membrane proteins, the positive control set was composed of protein sequences, which were used to discover the patterns in the first place. The patterns discovered by TEIRESIAS, but found to be present in any of the negative control sets and/or present in the positive control set but located in a sequential region that did not correspond to the given membrane dipping loop were eliminated. Additionally, PATTERNTEST was also used to detect common patterns in sets of dipping loops sharing a structural similarity or assembled from the same protein family to find common patterns in structurally related membrane dipping loop motifs (for the discovery of residues important for the folding and stabilization of the structural motif) and common patterns in membrane dipping loops possibly caused by ancestral gene duplication events.

PATTERNTEST was implemented using Borland Delphi 7. The software loads text files with a specific format **(figure 4.1)** where the set of sequences used and the patterns found by TEIRESIAS are described. PATTERNTEST **(figure 4.2)** can perform a series of different pattern (regular expression type) handling tasks: 1. Sum patterns saved in different files, 2. If more than one file is loaded, it can display the patterns that specifically belong to a particular file or if required, display common patterns present in the required files but not present in the remaining files loaded, 3. It can sort patterns according to their "N" parameter (see TEIRESIAS analysis), and 4. It can evaluate the selectivity of the patterns against Swiss-Prot-like text files, text files that contain information (e.g. organism, function, subcellular location and structural features) for a given protein written according to the nomenclature use by the Swiss-Prot database annotators. Most of the operations cited above were implemented using the Systools library.

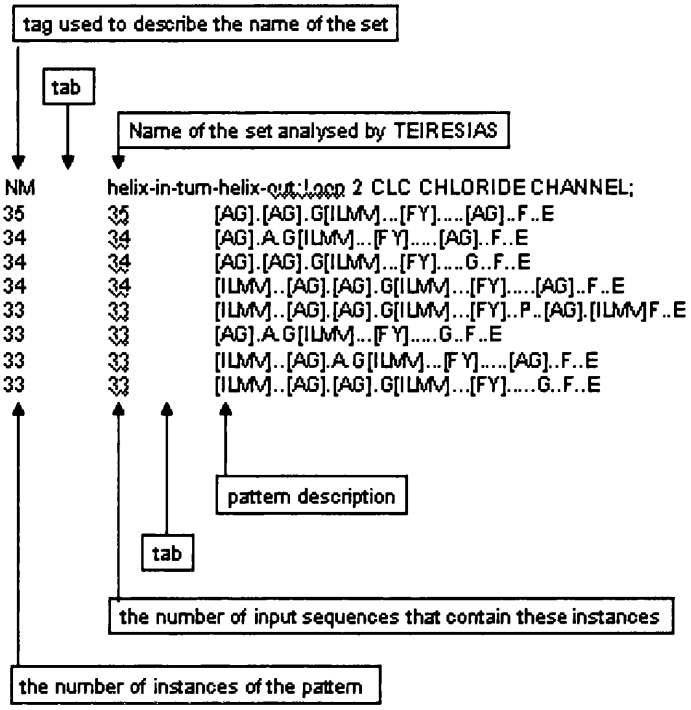

Figure 4.1. Example of the input format required by PATTERNTEST. The first row of each loaded file corresponds to the name of the set analysed by TEIRESIAS, this row is tagged with the "NM" tag. The following rows correspond to the patterns discovered by TEIRESIAS where the first number , "N", corresponds to the number of the instances of the given pattern in the set, the second number, "M", corresponds to the number of input sequences that contain those instances and the regular expression corresponds to the pattern discovered in the corresponding membrane dipping loop region.

The different evaluations carried out using PATTERNTEST were:

1. Detection of patterns specific to the helix-in-turn-loop-out motif in potassium channels.
2. Detection of patterns specific to the helix-in-turn-loop-out motif in the SecY/SEC61 alpha family.
3. Detection of patterns specific to the loop-in-turn-helix-out (L1) motif in the major intrinsic protein family.
4. Detection of patterns specific to the loop-in-turn-helix-out (L2) motif in the major intrinsic protein family.
5. Detection of patterns specific to the helix-in-turn-helix-out motif in the binding protein dependent permease system family.
6. Detection of patterns specific to the helix-in-turn-helix-out (L1) motif in the chloride channel family.
7. Detection of patterns specific to the helix-in-turn-helix-out (L2) motif in the chloride channel family.

8. Detection of patterns specific to the helix-in-turn-helix-out (L3) motif in the chloride channel family.
9. Detection of patterns specific to the helix-in-turn-helix-out (L4) motif in the chloride channel family.
10. Detection of patterns specific to the helix-in-turn-helix-out motif in the psaF family.
11. Detection of patterns specific to the helix-in-turn-helix-out motif (L1) in the sodium : dicarboxylate symporter family.
12. Detection of patterns specific to the helix-in-turn-loop-out (L2) in the sodium : dicarboxylate symporter family.
13. Detection of common and specific patterns in the L1 and L2 loop in the major intrinsic protein family.
14. Detection of common and specific patterns in the L1 and L3 loop in the chloride channel family.
15. Detection of common and specific patterns in the L2 and L4 loop in the chloride channel family.
16. Detection of common and specific patterns in the L1, L2, L3 and L4 loop in the chloride channel family.
17. Detection of common and specific patterns in the L1 and L2 loop in the sodium : dicarboxylate symporter family.
18. Detection of common and specific patterns in all helix-in-turn-loop-out like dipping loops.
19. Detection of common and specific patterns in all loop-in-turn-helix-out like dipping loops.
20. Detection of common and specific patterns in all helix-in-turn-helix-out like dipping loops.
21. Detection of common patterns in all dipping loops (independently of the dipping loop type).

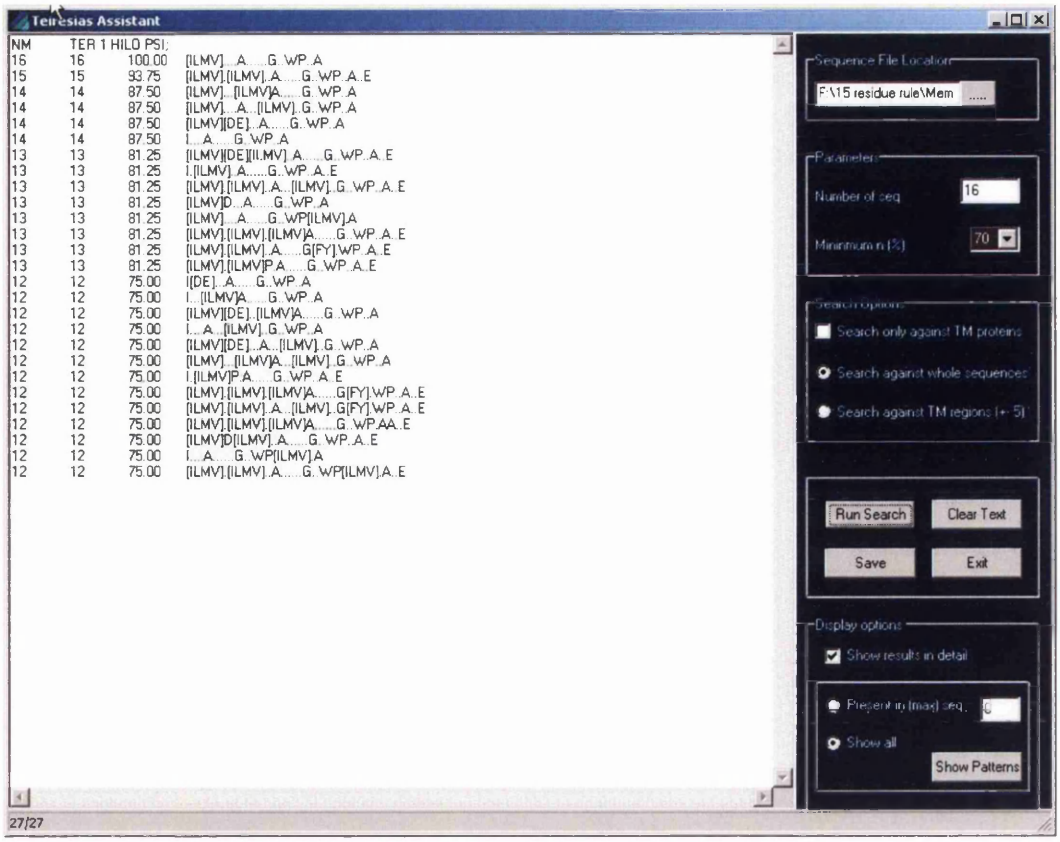

Figure 4.2. Screenshot of the pattern evaluation task performed by PATTERNTEST. In order to perform the evaluation, the user needs to input the required parameter and select the output type. The first box is used to indicate the Swiss-Prot-like files that would be used to evaluate the patterns described in a previously loaded file. In the parameters panel, the user is asked to enumerate the number of sequences used during the pattern discovery process and the minimum required support of the patterns to be evaluated. In the search options panel, the software could be set up to perform the evaluation only against transmembrane proteins (with 2 of more transmembrane regions) specifying if the evaluation should be performed against the whole protein sequence or just against the transmembrane regions (plus 5 residues upstream and downstream to minimise the error in predicting the limits of the transmembrane regions). In the display options panel, the user can set up the program to display only the non-matching patterns or obtain a detailed report of the evaluation process where the accession code and the definition of matching Swiss-Prot-like text files is displayed.

The three-dimensional structure of the ClC Chloride channel showed that the N-terminal half of the protein was structurally related to the C-terminal half (Dutzler et al., 2002). Therefore, loop 1 and loop 3 together with loop 2 and loop 4 were evolutionarily related. According to these finding the evaluation using PATTERNTEST was designed to discover common and specific patterns between evolutionarily related loops (evaluation 14-15) and between all the loops belonging to the chloride channel family (evaluation 16).

To detect patterns specific to a particular loop in a dipping loop protein group (evaluations 1 to 12) a filtering protocol was designed. The filtering protocol was composed of 4 steps:

1. Filtering the patterns obtained from TEIRESIAS against patterns, also detected by TEIRESIAS, belonging to other dipping loops. Patterns common to 2 or more dipping loops (therefore independent of the dipping loop protein group) were removed at this point (e.g. common patterns in loop 1 and loop 2 of the major intrinsic protein family were removed). After this filtering, the remaining patterns were guaranteed not to be common (not present in at least 70% of the dipping loop sequences) in other dipping loops. Likewise, processing time was saved during step 2 as common patterns were removed and subsequently not analysed in the following step, which required more processing time for each pattern.
2. Filtering the patterns obtained from step 1 against sequences belonging to other dipping loop protein groups. In this step, patterns that were present in sequences belonging to different dipping loop protein groups were removed. At this stage, patterns that survived this filter were considered to be specific for the dipping loop being analysed, in terms of the different dipping loops identified using the PDB_TM database.
3. Filtering the patterns obtained from step 2 against membrane protein sequences known not to have dipping loops in their structures. A dataset composed of 363 membrane proteins known not to have dipping loops in their structure was assembled (the development of this database is explained later). In this step, patterns that were present in other membrane proteins without dipping loops were removed. It was likely that some of the patterns detected by TEIRESIAS were not specific to a particular dipping loop but to the general structure of polytopic membrane proteins. Therefore, these patterns were not indicative of a dipping loop and needed to be removed.
4. Filtering the patterns obtained from step 3 against the same sequences used to detect the patterns being analysed. By applying this filter it was guaranteed that patterns were only present in the dipping loop region and not outside this motif. These patterns could be specific to a particular membrane dipping loop, but the same sequence pattern might also be found in a non-membrane dipping loop

region of the given membrane protein, leading to the prediction of false positives.

Patterns that passed the 4 filtering steps were guaranteed to be thoroughly specific for the dipping loop analysed, that is, not present either in other dipping loops or in membrane proteins without dipping loops. Likewise, the remaining loop patterns were guaranteed to be present specifically in the region of the sequence corresponding to the dipping loop.

As mentioned above, PATTERNTEST was also implemented to detect common patterns in different membrane dipping loops belonging to: the same gold standard set, the same structural type **(table 4.2)** or all the different gold standard sets assembled (evaluation 13 to 21). Discovery of patterns belonging to different membrane dipping loops (either sharing a structural similarity or involving all membrane dipping loops identified) might be indicative of important residues during the folding stage. To find these patterns a protocol composed of 5 steps was designed:

1. Using the patterns detected by TEIRESIAS, this step was designed to detect common patterns in different loops belonging to: the same dipping loop protein group (evaluation 11 to 13) or the same structural type of dipping loop (evaluation 14 to 16) or all the dipping loop groups detected in the PDB_TM database (evaluation 17). After this step, the following steps included here were similar to those applied in the filtering protocol described above (evaluation 1 to 10).
2. Filtering the patterns obtained from step 1 against patterns, also detected by TEIRESIAS, belonging to: other dipping loop protein groups (evaluation 11 to 13) and other structural types of dipping loops (evaluation 14 to 16).
3. Filtering the patterns obtained from step 2 against sequences belonging to: other dipping loop protein groups (evaluation 11 to 13) and other structural types of dipping loops (evaluation 14 to 16). Step 2 and step 3 worked together to detect common and specific patterns in: different loops belonging to the same dipping loop protein group or in different loops belonging to the same structural type of dipping loop.

4. Filtering the patterns obtained from step 3 against membrane protein sequences known not to have dipping loops in their structures.
5. Filtering the patterns obtained from step 4 against the same sequences used to detect the patterns being analysed. By applying this filter, it was guaranteed that patterns were only present in the dipping loop regions and not outside the motif region.

The patterns validated by PATTERNTEST were ensured to be specific to: dipping loops belonging to: a particular membrane dipping loop, a structural type of membrane dipping loop or the general structure of membrane dipping loops.

The patterns obtained using PATTERNTEST were not validated against a set of globular proteins sequences. The spectrum of 3D motifs in globular proteins is known to be larger than the spectrum of 3D motifs in membrane proteins. The reason for this is that the lipid bilayer constrains the variety of 3D motifs in the membrane due to its molecular and physicochemical properties whereas the aqueous environment allows globular proteins to have a wider range of 3D motifs. Therefore, it was likely that specific sequence patterns of dipping loops, detected by TEIRESIAS and evaluated by PATTERNTEST, would be present in globular protein sequences as well, but these could not be considered as false positives as membrane proteins and globular proteins are regarded as two different sets of proteins whose structure is stabilized under very different conditions.

The patterns that passed all different filtering stages applied using PATTERNTEST (evaluation 1 to 17) were eligible to be selected as rules and used in a predictive algorithm, named TMLOOP, capable of predicting the dipping loops that were characterised using TEIRESIAS and PATTERNTEST in a query amino acid sequence **(Chapter 5)**.

4.2.5 Development of a dataset of membrane proteins known not to have dipping loops

The development of a dataset of membrane proteins without dipping loops was necessary to validate the patterns obtained from TEIRESIAS as it was possible that some of the patterns detected were not specific to a particular membrane dipping loop but to the general structure of polytopic membrane proteins.

A set composed by 363 different membrane proteins without dipping loops was developed. Proteins were identified using 2 different sources. The first source was the Stephen's White database of "Membrane Proteins of Known Structure" (the Stephen White laboratory at University of California, Irvine, http://blanco.biomo.uci.edu/Membrane_Proteins_xtal.html), which is a manually curated database, which contains a selection of crystallized membrane proteins listed in the PDB database. This database contains a functional classification of integral membrane proteins whose structure has been determined (by crystallography, or sometimes NMR) to a resolution sufficient to identify transmembrane regions. It also maintains a low level of redundancy listing representative crystallographic analyses. These structures were visualised using RasMol visualisation software (Sayle and Bissel, 1992) in order to corroborate the prediction described in the PDB_TM database and therefore identify crystallized structures without dipping loops in their structure **(table 4.7)**. The second source was the Swiss-Prot database, through which sequences belonging to protein families known not to have membrane dipping loops in their structures were obtained **(table 4.8)**.

PDB code	Definition	Corresponding Swiss-Prot files with transmembrane regions
1AP9	Bacteriorhodopsin	P02945
1AR1	Cytochrome c oxidase	Chain A P08305 Chain B P98002
1AT9	Bacteriorhodopsin	P02945
1BCC	Ubiquinol cytochrome c oxidoreductase	Chain C P18946 Chain D P00125 Chain E P13272 Chain G P13271 Chain F P00130
1BGY	Cytochrome BC1 complex	Chain C & O P00157 Chain E, I, Q & V P13272 Chain F & R P00129 Chain G & S P13271 Chain H & T P00126 Chain J & V P00130 Chain K & W P07552 Chain N P23004 Chain R P00129
1BRR	Bacteriorhodopsin	P02945
1BRX	Bacteriorhodopsin	P02945
1C3W	Bacteriorhodopsin	P02945
1C8R	Bacteriorhodopsin	P02945
1C8S	Bacteriorhodopsin	P02945
1E12	Halorhodopsin	P16102
1EHK	BA3-type cytochrome c oxidase	Chain A Q56408 Chain B P98052 Chain C P82543
1EUL	Hydrolase	P11719
1EYS	Photosynthetic reaction centre	Chain h Q93RD8
1F88	Rhodopsin	P02699
1FE1	Photosystem II	P35876
1FFT	Ubiquinol reductase	Chain A & F P18401 Chain B & G P18400 Chain C & H P18402
1GMZ	Rhodopsin	P02699
1H2S	Sensory rhodopsin II transducer complex	Chain A P25896 Chain B P42196
1H68	Sensory rhodopsin II	P42196
1IWG	Bacterial multidrug efflux transporter	P31224
1IWO	Sarcoplasmic reticulum calcium ATPase	P04191
1JBO (psaF excluded)	Photosystem I	Chain A P25896 Chain B P25897 Chain I P25900 Chain J P25901 Chain K P20453 Chain L P25902 Chain M P25903
1JGJ	Sensory rhodopsin II	P42196
1KQF	Formate dehydrogenase	Chain B P24184 Chain C P24185
1KQG	Formate dehydrogenase	Chain B P24184 Chain C P24185
1KZU	Formate dehydrogenase	Chain A, D & G P26789 Chain B, E, & H P26790
1L0V	Fumarate reductase	Chain C & O P03805

		Chain D & P P03806
1L9H	Rhodopsin	P02699
1LGH	Light harvesting complex II	Chain A, D, J & G P97253 Chain B, E, K & H P95673
1MSL	Mechanosensitive ion channel	O53898
1NEK	Succinate dehydrogenase II	Chain A P10444 Chain C P10446 Chain D P10445
1NEN	Succinate dehydrogenase II	Chain A P10444 Chain C P10446 Chain D P10445
1NKZ	Formate dehydrogenase	Chain A, D & G P26789 Chain B, E, & H P26790
1OCC	Cytochrome c oxidase	Chain A & N P00396 Chain B & O P00404 Chain C & P P00415 Chain D & Q P00423 Chain G & T P07471 Chain I & V P04038 Chain J & W P07470 Chain K & X P13183 Chain L & Y P00430 Chain M & Z P10175
1OED	Nicotinic acetylcholine receptor	Chain A P02710 Chain B P02712 Chain C P02718 Chain E P02714
1OGV	Photosynthetic reaction centre	Chain H P11846 Chain L P02954 Chain M P02953
1OKC	ADP/ATP translocase I	P02722
1OY6	Bacterial multidrug efflux transporter	P31224
1PF4	MsbA lipid transporter ("flippase")	Q9KQW9
1PRC	Photosynthetic reaction centre	Chain H P06008 Chain L P06009 Chain M P06010
1PSS	Photosynthetic reaction centre	Chain H P11846 Chain L P02954 Chain M P02953
1PV6	Lactose permease	P02920
1PV7	Lactose permease	P02920
1Q16	NarGHI nitrate reductase A	Chain C P11350
1Q90	Cytochrome b6f	Chain A P23577 Chain B Q00471 Chain D P23230 Chain R P49728 Chain G Q08362 Chain L P50369 Chain M Q42496 Chain N Q9SBM8
1QCR	Mitochondrial Cytochrome BC1	Chain C P00157 Chain D P00125 Chain E P13272 Chain G P13271 Chain J P00130 Chain K P07552
1QHJ	Bacteriorhodopsin	P02945
1QKO	Bacteriorhodopsin	P02945

1QKP	Bacteriorhodopsin	P02945
1QLA	Respiratory complex II-like fumarate reductase	Chain C & F P17413
1QLB	Respiratory complex II-like fumarate reductase	Chain C & F P17413
1RWT	Major light harvesting complex	P12333
1S5L	Photosystem II	Chain A P35876 Chain B Q8DIQ1 Chain C Q8DIF8 Chain D Q8CM25 Chain E Q8DIP0 Chain F Q9DIN9 Chain H Q8DJ43 Chain I Q8DJZ6 Chain J P59087 Chain M Q8DHA7 Chain T Q8DIQ0 Chain X Q8DHE6 Chain Z Q8DHJ2
1S7B	Multidrug resistance efflux transporter	P23895
1SU4	Sarcoplasmic reticulum calcium ATPase	P04191
1T5S	Sarcoplasmic reticulum calcium ATPase	P04191
1T5T	Sarcoplasmic reticulum calcium ATPase	P04191
1U19	Rhodopsin	P02699
1U77	AmtB ammonia channel	P37905
1U7C	AmtB ammonia channel	P37905
1U7G	AmtB ammonia channel	P37905
1VF5	Cytochrome b6f complex	Not listed in Swiss-Prot database
1VFP	Sarcoplasmic reticulum calcium ATPase	P04191
1WPG	Sarcoplasmic reticulum calcium ATPase	P04191
1XIO	Sensory rhodopsin II	Not listed in Swiss-Prot database
1XP5	Sarcoplasmic reticulum calcium ATPase	P04191
1XQF	AmtB ammonia channel	P37905
1YCE	Rotor of F-type Sodium ATPase	Not listed in Swiss-Prot database
1Z2R	MsbA lipid transporter ("flippase")	P63359
2BRD	Bacteriorhodopsin	P02945
2BLZ	Rotor of V-type Sodium ATPase	Not listed in Swiss-Prot database
2BG9	Nicotinic acetylcholine receptor	Not listed in Swiss-Prot database
2PPS (psaF excluded)	Photosystem I	Chain A P25896 Chain B P25897 Chain I P25900 Chain J P25901 Chain K P20453 Chain L P25902 Chain M P25903
2RCR	Photosynthetic reaction centre	Chain H P11846 Chain L P02954 Chain MP02953

Table 4.7. Dataset of PDB structures known not to have membrane dipping loops in their structure. Each PDB structure is described using the functional annotation listed in the PDB database and the corresponding Swiss-Prot links.

Protein family	Corresponding Swiss-Prot accession codes
Acetylcholine receptor	P08911, P16395, P17200, P30372, P30544, P32211, P41984, P49578, P56489, P56490, Q9ERZ3, Q9ERZ4, Q9N2A4, Q9U7D5
Dopamine receptor	O73810, P21728, P21917, P24628, P30728, P30729, P35406, P41596, P42288, P42289, P42290, P42291, P47800, P50130, P51436, P52703, P53452, P53453, P53454, P61168, Q24563, Q61616, P25115, P21918
Glucose transporter	O43826, O74713, O74969, P10870, P11168, P11170, P13866, P15686, P15729, P18631, P21906, P23585, P28568, P31636, P31639, P31675, P32465, P32466, P32467, P33026, P39003, P46896, P49374, P53791, P78831, Q06222, Q12300, Q27115, Q27994, O07881, P12336, P14246, P23586, P27674, P32037, P47842, P53790, P53792, P58352, Q07647, Q90592, Q9P3U6
Opsins	O13227, O15973, O15974, O62796, O93441, P23820, P28681, P28683, P28684, P32312, P35357, P35403, P51474, P79898, P87366, P87369, Q17053, Q90214, Q90215, Q9YGZ1, O16005, O18766, O42604, O62793, O62798, O93459, P02699, P09241, P22328, P22671, P24603, P28682, P29403, P31355, P31356, P32311, P35356, P35359, P35362, P41590, P41591, P49912, P51471, P51488, P51489, P52202, P56514, P79812, P79848, P79863, Q17292, Q17296, Q8HY69, Q90245, Q98980, Q9DGG4, Q9YGZ0, Q9YGZ2, Q9YGZ3, Q9YGZ4, Q9YGZ5, Q9YGZ9, Q9YH01, Q9YH05
P-type ATPase	P07038, P22036, P28876, P35670, P36640, P37617, P39168, P54210, P98204, Q00804, Q03194, Q43128, Q59385, Q59998, Q9MA0, Q9X5X3, Q9ZL53, O22218, O23087, O43108, O81108, P04191, P09627, P11506, P13585, P13586, P19657, P22700, P23220, P23634, P24545, P28877, P35316, P38929, P54211, P98194, Q00779, Q01896, Q03669, Q04656, Q07421, Q08436, Q08853, Q16720, Q64518, Q64542, Q92105, Q93084, Q9LF79, Q9LV11, Q9S7J8, Q9SY55, Q9X5V3, Q9XES1, Q9YGL9, P25169, P28774, P50992, Q64436, P51165, Q92030
Serotonin receptor	O08890, O42385, P18599, P20905, P28285, P28335, P28566, P30966, P31387, P32304, P34969, P35364, P41595, P47898, P49145, P50406, P60020, P97288, Q16951, Q25414, Q91559, Q9R1C8, O08892, O42384, O70528, P11614, P19327, P28221, P28286, P28565, P30939, P30994, P34968, P35363, P35404, P46636, P49144, P50128, P56496, P79748, Q02152, Q02284, Q16950, Q60484, Q64264, Q9N298
Archaeal fungal-bacterial opsin	O74631, P02945, P16102, P33743, P42196, P42197, P71411, Q48334, Q9AF7, Q9F7P4
Olfactory receptor	Q78PE1, Q78PE2, Q78PE3, Q7TMF7, Q7TMF8, Q7TMF9, Q920Y6, Q920Y7, Q920Y8, Q920Z0, Q9R0Z2

Table 4.8. Dataset of protein accession codes listed in the Swiss-Prot database known to belong to proteins families whose structures do not posses membrane dipping loops.

4.3 Results and discussion

4.3.1 Clustering of proteins with dipping loops found in the PDB_TM database

In the case of potassium channels and aquaglyceroporins, membrane dipping loops are known to be conserved through evolution and to work as selective filters to transport specific substrates and avoid transport through the membrane of other ions or molecules (Agre and Kozono, 2003, Doyle et al., 1998, Fujiyoshi et al., 2002, Gonen et al., 2004, Harries et al., 2004, Jiang et al., 2002, Jiang et al., 2003, Kuo et al., 2003, Long et al., 2005, MacKinnon, 2003, Murata et al., 2000, Pao et al., 1991, Ren et al., 2000, Savage et al., 2003, Stroud et al., 2003a, Stroud et al., 2003b, Sui et al., 2001, Zhou et al., 2001). The different crystallized potassium channels showed a similar spatial arrangement of the membrane dipping loop motif. Likewise the multiple sequence alignment of sequences of the different potassium channels **(Please see supplementary information S4.Alignment potassium channel MDL on CD)** showed that most of the sequences contained conserved residues in the equivalent region (including residues belonging to the GYGD motif) to the dipping loop motif mapped onto the reference sequence. Therefore, all different potassium channel sequences, those of voltage gated KvAP potassium channels, KcsA potassium channels, calcium gated potassium channels and inward rectifier potassium channels, were clustered together in a single set. Aquaporins and glyceroporins also showed a similar three dimensional arrangement of their motifs (both loops) and the multiple sequence alignment revealed conserved residues (as with potassium channels the NPA motif was highly conserved, but containing small variations) in the region pertaining to membrane dipping loops one and two. Accordingly, aquaporins and glyceroporins were clustered together in the aquaglyceroporin set.

4.3.2 Visual confirmation of dipping loops in the structures listed in the PDB_TM database

In order to confirm the predictions listed in the PDB_TM database, the dipping loops of each PDB structure were manually confirmed by using RasMol (Sayle and Bissel, 1992). The NAD(P) transhydrogenase is predicted in the Swiss-Prot database

(accession code: P07001) to have 4 α-helical transmembrane regions, the sequence positions of these transmembrane regions are: 402-422, 423-443, 453-473, 477-497. These segments could not be highlighted in **figure 4.3** because the corresponding PDB file did not include the C-terminal sequence of the protein. Due to the fact that the transmembrane domain could not be visualised and that the structure of the predicted membrane dipping loop in the PDB_TM database did not match any of the observed structural patterns in membrane dipping loops (namely loop-in-turn-helix-out, helix-in-loop out and helix-in-turn-helix-out), the NAD(P) transhydrogenase predicted dipping loop was considered to be a false positive in the PDB_TM database.

Figure 4.3 Structure of the NAD(P) transhydrogenase (PDB code: 1XL4). The predicted membrane dipping loop listed in the PDB_TM database is highlighted in red. The predicted loop shows a beta strand-in-helix-out structure, which has not been observed in other proteins containing membrane dipping loops. Likewise, it is not possible to locate the membrane as no transmembrane regions have been included in the structure.

The transmembrane domain of the Cytochrome BC1 is composed of 4 different subunits. As can be seen in **figure 4.4**, Cytochrome b (coloured in blue) has 8 transmembrane regions and 4 horizontal helices on the intermembrane side (Iwata et al., 1998). The predicted dipping loop boundaries in the PDB_TM database related to one of the horizontal helices located on the intermembrane side. This predicted membrane dipping loop was also considered a false positive as in the literature (Iwata et al., 1998)

the corresponding region is referred to as being horizontal and located on the intermembrane side of the mitochondrion.

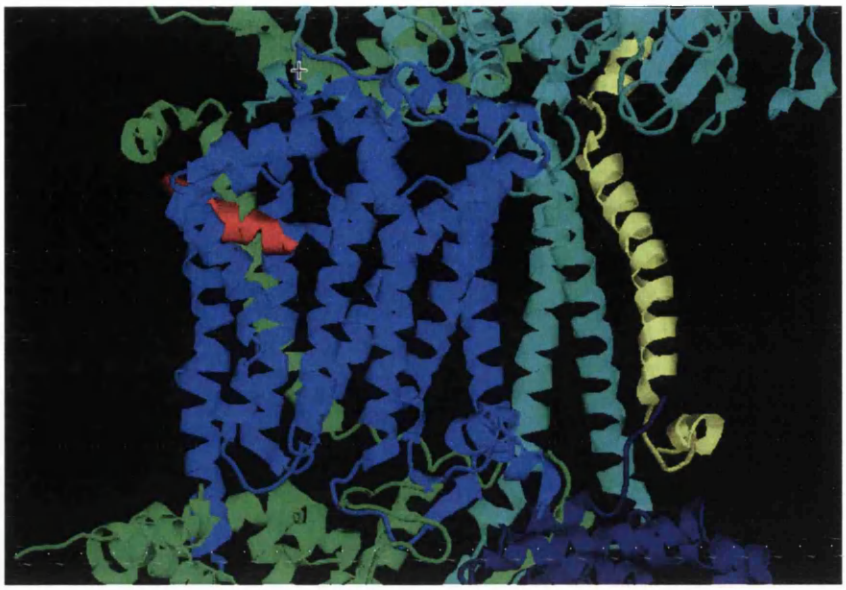

Figure 4.4. Structure of the Cytochrome BC1 (PDB accession codes: 1BGY, 1EZV, 1KB9). Subunit b has been highlighted in blue and the predicted dipping loop related to the subunit b in red.

The glutamate transporter homologue (PDB accession code 1XFH) was predicted to have a single dipping loop motif (residues 262-289) according to the PDB_TM database. However, the crystallised structure was described by its elucidators (Yernool et al., 2004) as a polytopic membrane protein containing 2 membrane dipping loops, named HP1 (residues 265-288) and HP2 (residues 338-369) **(figure 4.5)**. Both dipping loops were described as key elements of the transport machinery of the protein and the spatial arrangement of these motifs were described as helix–in-helix-out motifs.

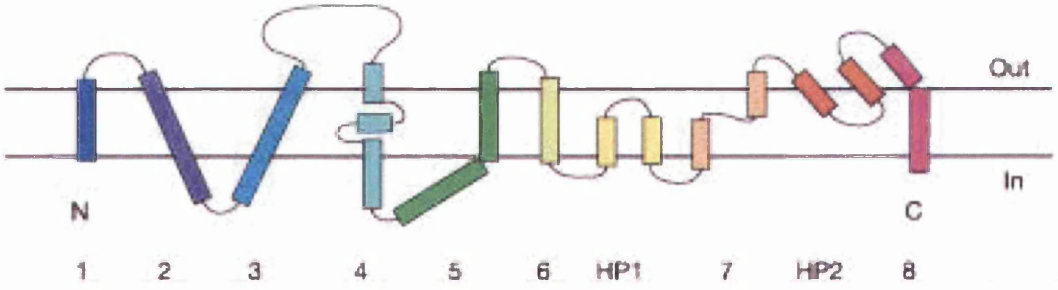

Figure 4.5. Schematic view of the topology of the glutamate transporter homologue (PDB accession code: 1XFH) (Yernool et al., 2004).

After visualising the structure of the glutamate transporter homologue, both membrane dipping loops were confirmed **(figure 4.6)**. HP1 (highlighted in red) was also

described as a helix-in-turn-helix-out motif, whereas HP2 (highlighted in green and cyan) was described as a helix-in-turn-loop-out dipping loop motif. It was believed that the beginning of the helix HP2b was not situated in the membrane but close to the predicted outer boundary of the lipidic bilayer. In **figure 4.5**, residues believed to be in the membrane were coloured in green (residues 338-360) and residues believed to be outside the membrane were coloured in cyan (residues 361-369). However, as a precaution it was decided better to analyse the region between residues 338 and 369 during the pattern discovery process.

Figure 4.6. Structure of the glutamate transporter homologue (PDB accession code: 1XFH). Only dipping loops belonging to chain A were highlighted. HP1 was coloured in red and HP2 was coloured in green (believed to be inside the membrane) and cyan (believed to be outside the membrane).

The dipping loops listed in the PDB_TM database (L2, L4) belonging to the ClC chloride channel (PDB accession code 1KPK, 1KPL) were visually confirmed using RasMol. However, according to the literature, two more dipping loops were found in the structure (L1, L3), that involved residues 135-155 and 345-365 (Dutzler et al., 2002). These regions were analysed visually using RasMol (Sayle and Bissel, 1992) and although they were not detected in the PDB_TM database, both loops were confirmed as helix-in-turn-helix-out loops **(figure 4.7)**. Therefore the ClC chloride channel structure was assumed to contain 4 different helix-in-turn-helix-out loops, which were eligible to be analysed by TEIRESIAS **(table 4.1)**.

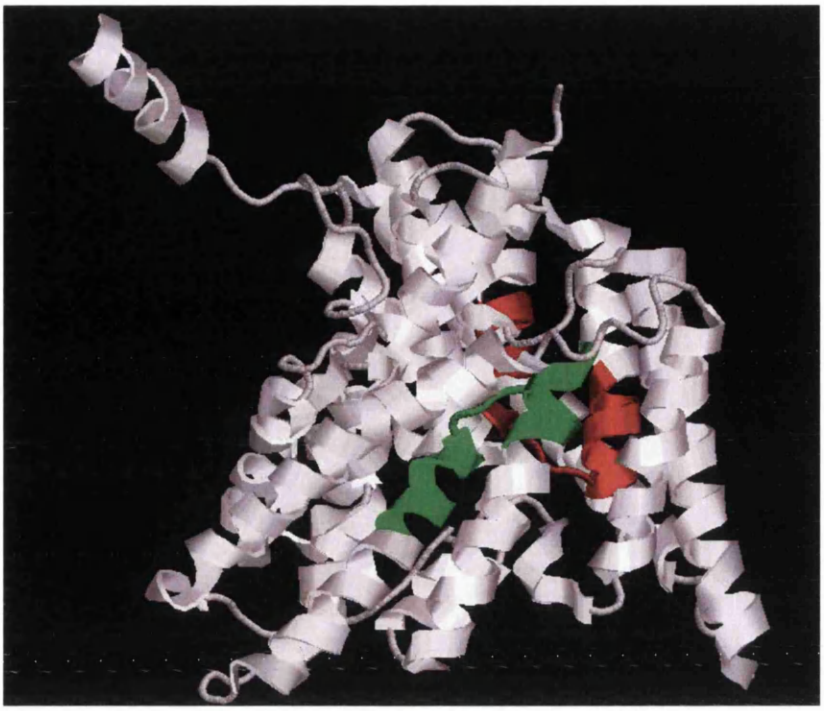

Figure 4.7. Structure of the ClCa monomer chloride channel (PDB accession code: 1KPK). Dipping loops reported by Dutzler et al. (Dutzler et al., 2002) were highlighted using Rasmol. Residues belonging to the first loop (135-155) were highlighted in red whereas residues belonging to the third loop (345-365) were highlighted in green.

The remaining protein types were visually confirmed by using RasMol (Sayle and Bissel, 1992) and their dipping loops were classified into 3 structural categories (see methods section).

4.3.3 Training set development and membrane dipping loop identification

The potassium channel group was composed of 3 different families of Potassium channels listed in the Swiss-Prot database: the potassium channel family, the potassium channel (TC 1.A.1.4) family and the inward rectifier-type potassium channel. 222 potassium channels were obtained from the Swiss-Prot database (excluding non-reliable proteins, filtered out according to their functional annotation quality as described in the methods section), after filtering this dataset using Non-Red, a non-redundant set of potassium channels was obtained composed of 134 sequences. The multiple sequence alignment of these proteins showed a good alignment **(Please see supplementary information S4.Alignment potassium channel MDL on CD)** in the region belonging

to the helix-in-turn-loop-out type membrane dipping loop, which supports the statement that the three potassium channel families share a common three dimensional motif in this region (Jiang et al., 2003, MacKinnon, 2003).

The protein conducting channel family was composed of 105 sequences from the Swiss-Prot database. The filter applied using Non-Red selected 77 out of 105 sequences as non-redundant sequences. The multiple sequence alignment of the remaining 77 sequences using ClustalW showed a generally good alignment in the transmembrane regions however, unconserved residues were observed in the region belonging to the helix-in-turn-loop-out type dipping loop **(Please see supplementary information S4.Alignment protein channel MDL on CD)**. This multiple sequence alignment showed a seven residue gap in the region corresponding to the structural motif. Based on the alignment results, it was assumed that the membrane dipping loop motif predicted in the PDB structures 1RH5 and 1RHZ might not be conserved throughout the evolution of the secY/SEC61-alpha family. However, the analysis of the region belonging to the dipping loop motif using TEIRESIAS software was continued as patterns of residues, using amino acid structural and/or chemical equivalencies, could still be found.

The aquaglyceroporin family was composed of 69 different sequences obtained from the Swiss-Prot database. A non-redundant set of 49 sequences was obtained after applying Non-Red. The multiple sequence alignment obtained for both loops **(Please see supplementary information S4.Alignment aquaglyceroporin MDL1 and 2 on CD)** showed that most of the sequences contained the NPA motif in both dipping loops. The NPA motif is known to be well conserved among this family and forms the basis of the selectivity filter. The presence of conserved residues in both membrane dipping loop regions during alignment and the conserved three dimensional motifs found in the different crystallised structures (Stroud et al., 2003a, Gonen et al., 2004, Harries et al., 2004, Savage et al., 2003, Murata et al., 2000, Sui et al., 2001, Ren et al., 2001, Ren et al., 2000) belonging to this family confirmed that both membrane dipping loop motifs were universally conserved in the aquaglyceroporin family.

According to the Swiss-Prot database, the binding-protein-dependent transport system permease family was composed by 252 proteins and 8 different subfamilies: the AraH/rbsC subfamily, the CysTW subfamily, the FbpB subfamily, the FecCD subfamily, the HisMQ subfamily, the LivHM subfamily, the MalFG subfamily and the OppBC subfamily. The Swiss-Prot accession code of the protein corresponding to the PDB structure 1L7V was P06609. According to the annotation of P06609 in the Swiss-Prot database, the crystallized protein was found to be a member of the FecCD subfamily and it was the only available structure for this family. No evidence was found in the literature to infer a conserved membrane dipping loop among the different subfamilies. Likewise, a multiple sequence alignment of all the protein sequences belonging to this family **(Please see supplementary information S4.Alignment bindin protein dependent permease MDL on CD)** revealed a poor alignment in the dipping loop region. Therefore, it could not be inferred that a common dipping loop motif is shared by different subfamilies of the binding-protein-dependent transport system permease family. However, aligned protein sequences belonging to the FecCD subfamily showed conserved residues in the membrane dipping loop region. This alignment of sequences of the FecCD subfamily suggested that the dipping loop found to be in the PDB structure 1L7V could be a motif that is conserved across the FecCD subfamily. Of the 31 proteins belonging to the FecCD subfamily, obtained from the Swiss-Prot database, a non-redundant set of 25 proteins was obtained by applying Non-Red as described in the methods section.

The chloride channel family was composed of 63 membrane proteins, which were subclassified according to the Swiss-Prot database in: the Chloride family (TC 1.A.11), the ClcA subfamily (TC 1.A.11) and the ClcB subfamily (TC 1.A.11). According to the annotation of the reference protein in the Swiss-Prot database (accession code: P37019), the protein corresponding to the PDB structure 1OTS belonged to the ClcA subfamily, however as in the previous case there was no evidence suggesting that the four dipping loops found in the PDB structure were conserved among members of the family. The set of 63 chloride channels were filtered using Non-Red (as described in the methods section). The non-redundant set obtained, composed of 35 sequences, was aligned using ClustalW to analyse the residue conservation in the regions belonging to the four dipping loops found using the PDB_TM database and the

literature (Dutzler et al., 2002). The multiple sequence alignment of the 4 helix-in-turn-helix-out dipping loop sequences showed conserved residues among the chloride channel family **(Please see alignments in enclosed CD)**. However, the corresponding alignment showed a better alignment within the ClcA subfamily (7 members) than within the chloride channel family (35 members) in all the dipping loops. This was expected as the sub-classification of chloride channels was probably achieved based on the significant sequence identity between proteins belonging to the ClcA and ClcB subfamilies. The alignment for loop 4 was poor within the chloride channel family, however it was reported that loop 2 and loop 4 faced each other at the interface between monomers linking the two repeated halves within each monomer (Estevez and Jentsch, 2002). The fact that the alignment for loop 2 showed conserved residues among the chloride channel family, suggesting a conserved membrane dipping loop motif, together with the fact that loop 2 and loop 4 have been suggested to be associated, encouraged us to continue with the discovery of patterns in the loop 4 region.

The psaF family in the Swiss-Prot database was composed of 18 proteins. Filtering using Non-Red revealed a non-redundant set of 16 membrane proteins. The multiple sequence alignment **(Please see supplementary information S4.Alignment psaF MDL on CD)** carried out using ClustalW showed conserved residues in the region corresponding to the membrane dipping loop. Therefore, the membrane dipping loop observed in the PDB structure 1JBO was proposed to be conserved among members of the psaF family.

Based on the Swiss-Prot database, 67 proteins were found to belong to the sodium : dicarboxylate (SDF) symporter family. After filtering this protein set using Non-Red, a non-redundant set of 53 proteins was obtained. The crystallized structure listed in the PDB database (1XFH) was defined as a glutamate transporter homologue, the same protein was defined as a hypothetical proton glutamate symport protein in the Swiss-Prot database (TrEMBL accession code: O59010). Despite this protein being defined as hypothetical and as a homologue, it was the only structure available for the sodium : dicarboxylate symporter family. Therefore it was considered as the reference sequence in order to identify and isolate the region corresponding to the dipping loop motif. The 53 proteins belonging to the non-redundant set and the reference sequence that belonged to the TrEMBL database were then aligned using ClustalW. The

alignment **(Please see supplementary information S4.Alignment Sodium:dicarboxylate symporter MDL 1 & 2 on CD)** showed conserved residues in the regions corresponding to both dipping loops, these results supported the primary assumption of using a hypothetical protein as a reference to identify and isolate the regions belonging to both dipping loops. Based on the amino acid conservation, both dipping loops were considered to be common 3D motifs among sequences belonging to the sodium : dicarboxylate symporter family.

4.3.4 Discovered patterns and their functional role

4.3.4.1 Patterns specific to particular loops (evaluation 1 to 12)

The following two tables **(tables 4.9-4.10)** summarize the results of the pattern discovery process. As mentioned in the methods section, the TEIRESIAS software does not guarantee that the patterns discovered were specific to the given input sequences. Therefore PATTERNTEST was used to detect: a) Patterns only present in a particular dipping loop motif **(table 4.8)** b) Patterns specific to membrane dipping loop motifs belonging to a particular set, a structural type of membrane dipping loop or to the general structure of membrane dipping loops.

Ev	Dipping loop type and protein group	TEIRESIAS analysis	Patterns found by TEIRESIAS	Pattern validation using PATTERNTEST			
				step 1	step 2	step 3	step 4
1	Helix-in-turn-loop-out Potassium channel	Chemical equivalency	382	352	43	37	35
		Structural equivalency	103	99	14	10	10
		Exact discovery	5	5	0	0	0
2	Helix-in-turn-loop-out secY / SEC61 alpha family	Chemical equivalency	12	3	0	0	0
		Structural equivalency	0	0	0	0	0
		Exact discovery	0	0	0	0	0
3	L1: Loop-in-turn-helix-out Aquaglyceroporin family	Chemical equivalency	863	700	186	182	167
		Structural equivalency	73	70	21	21	21
		Exact discovery	22	21	10	7	6
4	L2: Loop-in-turn-helix-out Aquaglyceroporin family	Chemical equivalency	249	212	54	29	24
		Structural equivalency	32	31	4	1	1
		Exact discovery	19	18	4	1	1
5	Helix-in-turn-helix-out Binding protein dependent transport system permease	Chemical equivalency	7506	6263	89	82	82
		Structural equivalency	479	460	46	43	43
		Exact discovery	31	31	11	11	11
6	L1: Helix-in-turn-helix-out Chloride channel family	Chemical equivalency	936	631	32	29	29
		Structural equivalency	46	44	7	5	5
		Exact discovery	14	14	3	3	3
7	L2: Helix-in-turn-helix-out Chloride channel family	Chemical equivalency	2419	1944	105	101	98
		Structural equivalency	97	96	39	35	35
		Exact discovery	63	62	31	28	28
8	L3: Helix-in-turn-helix-out Chloride channel family	Chemical equivalency	610	368	11	11	9
		Structural equivalency	45	45	11	7	7
		Exact discovery	10	10	2	2	2
9	L4: Helix-in-turn-helix-out Chloride channel family	Chemical equivalency	3751	3126	89	72	66
		Structural equivalency	137	129	0	0	0
		Exact discovery	26	26	0	0	0
10	Helix-in-turn-helix-out psaF family	Chemical equivalency	182	128	27	27	27
		Structural equivalency	41	41	16	16	16
		Exact discovery	13	13	12	9	9
11	L1: Helix-in-turn-helix-out Sodium : dicarboxylate (SDF) symporter family	Chemical equivalency	3613	3298	164	135	134
		Structural equivalency	347	344	29	21	19
		Exact discovery	63	63	17	12	11
12	L2: Helix-in-turn-loop-out Sodium : dicarboxylate (SDF) symporter family	Chemical equivalency	14887	13504	537	410	324
		Structural equivalency	1327	1302	157	149	142
		Exact discovery	103	102	22	18	17

Table 4.9. Summary of the filtering process carried out using PATTERNTEST. The number of patterns specific to particular dipping loops and discovered by TEIRESIAS is specified at each filtering stage. The secY / SEC61 alpha family was found to be the only dipping loop protein group without conserved patterns in the dipping loop region.

4.3.4.1.1 Potassium channels

Potassium channels have diverged throughout evolution by developing different gating domains that have been found to be attached in a modular fashion to a conserved pore unit (MacKinnon, 2003). The conserved pore unit has been found to be composed of 2 α-helical transmembrane regions and a pore forming helix. This pore forming helix is part of a dipping loop motif, which is highly conserved among the different potassium channels and is well known to be the selectivity filter that allows transport of K+, Rb+ and Cs+ through the membrane but not Na+ or Li+ (Doyle et al., 1998, Jiang et al., 2002, Jiang et al., 2003, Kuo et al., 2003, Long et al., 2005, Nishida and MacKinnon, 2002, Zhou et al., 2001). As expected, patterns found by TEIRESIAS and evaluated by PATTERNTEST were composed of residues known to be involved in the selectivity filter. Only 5 patterns were discovered by TEIRESIAS using an exact discovery text mining approach, but these patterns were ruled out by PATTERNTEST during the evaluation process. The most widely conserved patterns within the potassium channels were found to be "[ST]..[ST].**G[FY]G**" (support 0.89) and "[ST][ILMV]**G[FY]G**" (support 0.87). These patterns include (elements in bold in the patterns) part of the sequence also known as the potassium channel signature (GYGD). The selectivity filter contained four K^+ ion binding sites by means of layers of carbonyl oxygen atoms and a single layer of threonine (corresponding to the second element in the first pattern and the first element in the second pattern) hydroxyl oxygen atoms, which mimic the hydration shell of K^+ ions (MacKinnon, 2003). The tyrosine included in both patterns has also been proposed to participate in a sheet of aromatic residues located around the selectivity filter, which stabilizes the structure and avoids the carbonyl oxygen atoms from approaching close enough to allow the dehydration of a Na^+ ion (Doyle et al., 1998). The last glycine found in both patterns has been proposed to assist in the hydration and dehydration of a K^+ ion at the extracellular entry (Zhou et al., 2001). Potassium channels conduct ions with high conduction rates using the electrostatic repulsion between contiguous K^+ ions in the filter and by coupling a conformational change in the selectivity filter, which reduces the strength of the binding between ions and the filter (MacKinnon, 2003). This conformational change has been proposed to mainly involve two residues, a valine and a glycine (Zhou et al., 2001, Berneche and Roux, 2000). Both residues have been identified in the second most supported pattern discovered by TEIRESIAS (corresponding to the second and third element in the

pattern) whereas only the glycine was reported by the most supported pattern (corresponding to the sixth element in the pattern).

4.3.4.1.2 Aquaglyceroporins

The aquaglyceroporin family is a well-known protein family whose structure and function has been studied in depth by a variety of techniques such as mutagenesis, molecular dynamics or crystallographic analyses. Aquaglyceroporins are membrane proteins that transport water and glycerol molecules across the membrane. This protein family has been the focus of attention of many scientists because of their capacity to selectively transport glycerol and water molecules but not other small molecules or ions, and still perform such transport at very high diffusion rates. The different crystallized proteins belonging to the aquaglyceroporin family (Gonen et al., 2004, Harries et al., 2004, Murata et al., 2000, Ren et al., 2000, Ren et al., 2001, Savage et al., 2003, Stroud et al., 2003a, Sui et al., 2001) showed a conserved homotetramer structure, with each monomer being composed of two tandem repeats (Preston and Agre, 1991), each containing three α-helical transmembrane regions forming a right-handed bundle and a membrane dipping loop between the second and third transmembrane region. Both dipping loops contain highly conserved residues and a signature known as the NPA motif (asparagine-proline-alanine). The pattern with the highest support for the loop 1, "SG...**N.**.[ILMV][ST]", and for the loop2, [ILMV]**NP.**R.....[ILMV], contained part of the NPA motif as expected (elements in bold in the patterns). The multiple sequence alignment of sequences belonging to Loop 1 and Loop 2 in aquaglyceroporins **(Please see alignments in enclosed CD)** revealed that the proline and the alanine contained in the NPA motif were not universally conserved. The glycerol facilitator in *S. cerevisiase* (Swiss-Prot accession code: P23900) was found to have a A(354)S and a P(481)L substitution in the NPA motif corresponding to loop 1 and loop 2 respectively. Likewise, all sequences belonging to the aquaporin 7 group (Swiss-Prot accession code: O14520, O54794 and P56403) were found to have a P-A and a A-S substitution in the NPA motif belonging to loop 1 and loop 2 respectively. Both motifs were known to be associated with each other, through head-to-tail sidechain interactions between the alanine and proline of one domain with the proline and alanine of the other domain (Savage et al., 2003), bringing together the dipoles caused by the two short α-helices, which resulted in partial positive charges surrounding the highly conserved asparagines

(Agre and Kozono, 2003). Therefore, as was shown for the glycerol facilitator of *S. cerevisiae* and the aquaporin 7 proteins, amino acid changes within the NPA motif in one loop should be compensated by changes in the other loop to maintain the constriction imposed within the protein channel and high specificity of the filter. The asparagines from both NPA motifs oriented their side-chains to the pore and were believed to bind a transient central water molecule, which undergoes a dipole reorientation. The pattern discovered for the loop 2 in aquaglyceroporin proteins also highlighted an arginine, this residue was believed to be critical as it provided a strong positive charge at the narrowest region of the channel (Fujiyoshi et al., 2002) as a mechanism to repel protons and other cations. This selectivity against ions was also believed to be achieved by the polarization of the central water (through the ND2 groups of the asparagines and the positive dipoles of the short helices), which prevents adjacent water molecules from conducting protons through the channel, breaking the "proton wire" produced by the linear network of water molecules in the membrane (de Grotthuss, 1809).

4.3.4.1.3 ClC Chloride channels

ClC chloride channels are voltage gated ion channels that transport chloride ions across the membrane outside the cell. As explained in the methodology section, ClC chloride channels were composed of 4 different helix-in-turn-helix-out loops. Loops 2 and 4 were identified through the PDB_TM database whereas Loops 1 and 3 were reported in the literature (Dutzler et al., 2003, Estevez and Jentsch, 2002). Loops 2 and 4 were proposed to link the two repeated halves within each monomer and make contacts with each other at the interface between monomers (Estevez and Jentsch, 2002). On the other hand, loops 1 and 3 were reported to be part of the selectivity filter in ClC chloride channels, the selectivity filter was suggested to be composed by highly conserved residues that were brought together near the membrane centre and belonged to N-terminal positive end charges of helices creating a constriction in the pore (this spatial arrangement is reminiscent of that found in the aquaglyceroporin family discussed earlier). The most supported patterns found in loops involved in the selectivity filter were "[ILMV]**G[KR].GP**.[ILMV]", "[ILMV]**G..GP**.V" and "[ILMV]**G..GP**.[ILMV].....[AG]" (the support for all three was 0.86) for loop 1 and

"P.**G...P**....G...G" (support 0.91) and "P.**G.F.P**....G...G" (support 0.89) for loop 3. Dutzler et al., (2002) previously discovered part of these patterns (elements in bold in the patterns), the patterns G[KR]EGP and G.F.P were described as highly conserved regions in ClC chloride channels. The detection of these enhanced patterns validate the approach taken. Both loops seemed to be associated in creating a constriction in the pore. A glutamate belonging to dipping loop 1 was proposed to act as a gating residue (Dutzler et al., 2002, Dutzler et al., 2003, Estevez and Jentsch, 2002). This amino acid oriented its side chain towards the constriction imposed by loop 1 and 3 obstructing the pore in the closed state. This glutamate was not included in the most supported pattern but it was detected in a subsequent patterns with lower support value (e.g. "K.....G...G.**E**G..[ITV]", support 0.83). From the multiple sequence alignment corresponding to the region belonging to loop 1 **(Please see supplementary information S4.Alignment chloride channel MDL 1 on CD)** it could be seen that a cluster of six chloride channels developed a mutation in this position and glutamate was replaced by either valine or leucine. Interestingly, these proteins also developed a 2 amino acid insertion in the loop 1 region that was not found in other sequences used for the discovery of patterns. Further analysis, showed that these mutations were only present in ClCKa and ClCKb chloride channels. Two anion sites, namely Scen and Sint, have been suggested to be present in the selectivity filter, one (Sint) involved in the conduction selectivity (closer to the intracellular solution) and the remaining ion (located in Scen) involved in gating selectivity (closer to the extracellular solution) (Dutzler et al., 2002). An arginine included in the pattern with the highest support in loop 1 (corresponding to the 3rd element in "[ILMV]G[K**R**].GP.[ILMV]") has also been proposed to contribute to an electrostatic potential that probably attracts Cl^- ions into the pore extracellular entry (Dutzler et al., 2002). Residues belonging to loop 3, mainly an isoleucine residue and a phenylalanine residue (phenylalanine residue corresponds to the 5th element in the pattern with the second highest support), were also known to possess nitrogen atoms that coordinate the chloride channel in the Scen binding site (Dutzler et al., 2003, Estevez and Jentsch, 2002). The patterns with the highest support found in loops involved in dimerization were "[AG].[AG].G[ILMV]...[FY].....[AG]..F..E" (support 1.0) for loop 2 and "[AG].....[ILMV]...[ILMV][ILMV]..E[ILMV]T", "[AG]....[AG].....[ILMV]...[ILMV][ILMV]..E[ILMV][ST]" and [AG].....[ILMV][ST]..[ILMV][ILMV]..E.[ST] (the support for all three was 0.91).

The high support for the patterns corresponding to loop 2 and loop 4 suggests that these residues play an essential role in ClC Chloride channels and support the theory of Estevez and colleagues (Estevez and Jentsch, 2002).

4.3.4.1.4 Sodium : dicarboxylate symporters

The sodium : dicarboxylate (SDF) symporter family is composed of proteins that catalyze sodium and/or proton ions together with: 1) a Krebs cycle dicarboxylate (malate, succinate, or fumarate); 2) a dicarboxylic amino acid (glutamate or aspartate); 3) a small, semipolar, neutral amino acid (alanine, serine, cysteine or threonine); 4) neutral and acidic amino acids; 5) most zwitterionic and dibasic amino acids (Barabote R. D. et al., 2006). The crystallized structure of the glutamate transporter homologue showed a homotetramer with a central extracellular cavity located approximately halfway through the membrane that allowed the substrate access to the binding site in the protein (Yernool et al., 2004). Although the structure of this protein contained two membrane dipping loops motifs named HP1 and HP2, only HP1 was predicted by the PDB_TM database. Both dipping loops were visually confirmed using RasMol **(figure 4.6)** and subsequently analysed. The pattern with the highest support found in the HP1 region was found to be "[ILMV].....T.**S[ST]**...[ILMV]P" (support 0.89) and the pattern with the highest support found in the HP2 region was "[ILMV].....[ILMV].....S.G..[AG][ILMV].....[ILMV].[ILMV].....[ILMV]" (support 0.96). It has been suggested that the binding site is formed mainly by a motif located in TM7 and conserved residues in TM8. The binding site is flanked by both membrane dipping loops, HP1 flanking the intracellular side and HP2 flanking the extracellular side, acting as gates in the membrane (Yernool et al., 2004). In the closed state, the HP1 loop has been proposed to interact with HP2 loop through a serine-rich motif that is located in HP1, interacting with a conserved proline, which belongs to the HP2 loop. The serine-rich motif was composed of 3 to 4 serine residues but only the last two serines were included in the pattern (elements in bold in the patterns) with the highest support found in the HP1 region (elements 16 and 17), the complete motif was however detected by subsequent patterns with lower support (e.g. "[ILMV].....**SSS**.....E", support 0.8). This suggested that Ser 277 (corresponding to element 16 in the pattern with the highest support) followed by Ser 278 was the most important residue of the

serine-rich motif. The conserved proline located in the HP2 loop was not detected in any of the patterns discovered using TEIRESIAS. The multiple sequence alignment obtained using sequences belonging to the sodium : dicarboxylate symporter family **(Please see supplementary information S4.Alignment sodium dicarboxylate symporter MDL 1 on CD)** showed that the corresponding proline was conserved in 23 out of 46 sequences used in the training set, whereas the remaining proteins contained a threonine instead. These results suggested that proline might not be as functionally important as proposed, instead other residues, which were listed in the pattern with the highest support corresponding to HP2, might act as an anchor together with the serine-rich motif corresponding to HP1. These observations might lead to further experimental research in understanding the mechanism of glutamate transporters and sodium : dicarboxylate symporters.

4.3.4.1.5 Binding-protein-dependent permease family. FeCD subfamily

The binding-protein-dependent permease family belongs to the ATP-binding cassette (ABC) superfamily. These transporters have been suggested to catalyze both the uptake and efflux of a wide variety of substrates in all species (Davidson, 2002). The binding-protein-dependent permease system consists of a periplasmic binding protein, two membrane spanning domains forming a translocation pore and two ATP-binding cassettes located in the cytoplasm (Horlacher et al., 1998). The pattern with the highest support found in the dipping loop was found to be "[AG].[ILMV].F[ILMV][AG]L[ILMV].P.[ILMV]" (support 0.96). The region corresponding to the membrane dipping loop in the crystallized structure was suggested to be important for binding the periplasmic binding protein BtuF (Locher et al., 2002). Unfortunately, to our knowledge, no experimental approaches have been carried out to corroborate this suggestion. However, the highly conserved residues found in the dipping loop area emphasize the importance of the dipping loop for the functionality of the protein and supports the primary suggestion of Locher and colleagues.

4.3.4.1.6 SecY / SEC61 alpha family

Members of this family belong to a heterotrimer complex that is known to transport soluble proteins, such as secretory proteins, across the membrane and passage membrane proteins into the membrane. The α-subunit (SecY in archaea and eubacteria and Sec61α in mammals) forms the channel pore, the β-subunit (Secβ in archaea, SecE in eubacteria and Sec61β in mammals) contains a single transmembrane helix partially associated with the α-subunit (the β-subunit is the only subunit known to be non-essential for the function of the complex) and the γ-subunit (SecE in eubacteria and archaea and Sec61γ in mammals) clamps together the 2 halves of the α-subunit (TM 1-5 and TM 6-10) (Van den Berg et al., 2004, Mitra et al., 2005). The membrane dipping loop detected in the PDB_TM database has been reported as the TM2a, also known as the channel plug (Van den Berg et al., 2004). It has been suggested that in the closed state the plug is placed in the pore blocking the translocation of polypeptide chains and the channel opens by displacement of the plug, which moves away from the pore probably to a new position close to the C-terminus of the γ-subunit known as the plug-pocket (Van den Berg et al., 2004, Collinson, 2005). In accordance with other multiple sequence alignment analyses (Collinson, 2005, Van den Berg et al., 2004), the multiple sequence alignment attained using 75 α-subunits of the protein conducting channel showed high sequence conservation. However, the multiple sequence alignment **(Please see supplementary information S4.Alignment protein channel MDL on CD)** did not show conserved residues in the dipping loop region known as the plug. Likewise, the pattern discovery process using TEIRESIAS and PATTERNTEST did not detect any specific pattern in this region either. These results were also confirmed in recent work by Junne and colleages (Junne et al., 2006) that describes the structural domain as conserved despite its primary sequence not being well conserved. Mutation or deletion of the channel plug did not affect viability or growth of the yeast construct but reduced the corresponding translocation efficiency and the formation of the polymer. Junne and colleagues proposed that the channel plug played an important role in stabilizing Sec61p during the formation of the translocon rather than acting as a sealing gate in yeast. These recent results do not reflect the importance of the TM2a helix suggested during the gating process of the pore in bacteria and leads to further research on the gating process of the protein conducting complex.

4.3.4.1.7 PsaF family

Proteins belonging to the psaF family form part of an oligomeric complex known as the Photosystem I (PSI), which is a light-driven plastocyanin:ferredoxin oxidoreductase mediating electron transfer from reduced plastocyanin in the thylakoid lumen to oxidized ferredoxin in the stroma (Haldrup et al., 2000). The PsaF subfamily has been suggested to mediate plastocyanin docking and fast electron transport kinetics in eukaryotic PSI (Hippler et al., 1999, Haldrup et al., 2000). By contrast, in cyanobacteria psaF proteins have been suggested not to bind plastocyanin but to contribute to structural features on the surface of PSI and bind carotenoids, which serve as a light harvesting and photo-protecting molecule (Jordan et al., 2001). The proteins used for the pattern recognition belonged to both cyanobacteria and eukaryotes. Three similar patterns with support of 1.0 ("[ILMV]....A.....G..WP..A", "A.....G..WP..A..[EQ]", "A.....G..WP..A" were discovered using the chemical equivalency set, the structural equivalency set and by exact discovery respectively) using a training set that comprised both cyanobacteria and eukaryotes. Therefore, these patterns might indicate residues with a common functional role in cyanobacteria and eukaryote cells.

4.3.4.2 Common patterns in different loops belonging to the same protein group, the same structural type or all the protein groups used (evaluation 13 to 21)

Ev	Dipping loop type and protein group	TEIRESIAS analysis	Patterns found by TEIRESIAS	Pattern validation using PATTERNTEST			
				step 1	step 2	step 3	step 4
13	L1: Loop-in-turn-helix-out	Chemical equivalency	428	312	42	27	27
19	L2: Loop-in-turn-helix-out	Structural equivalency	28	6	0	0	0
	Aquaglyceroporin	Exact discovery	14	5	0	0	0
14	L1: Helix-in-turn-helix-out	Chemical equivalency	106	106	0	0	0
	L3: Helix-in-turn-helix-out	Structural equivalency	0	0	0	0	0
	Chloride channel family	Exact discovery	0	0	0	0	0
15	L2 Helix-in-turn-helix-out	Chemical equivalency	244	15	0	0	0
	L4: Helix-in-turn-helix-out	Structural equivalency	0	0	0	0	0
	Chloride channel family	Exact discovery	0	0	0	0	0
16	L1: Helix-in-turn-helix-out L2: Helix-in-turn-helix-out L3: Helix-in-turn-helix-out L4: Helix-in-turn-helix-out Chloride channel family	Chemical equivalency Structural equivalency Exact discovery	0 0 0	0 0 0	0 0 0	0 0 0	0 0 0
17	L1: Helix-in-turn-helix-out L2: Helix-in-turn-loop-out Sodium : dicarboxylate (SDF) symporter family	Chemical equivalency Structural equivalency Exact discovery	124 2 0	124 0 0	0 0 0	0 0 0	0 0 0
18	Helix-in-turn-loop-out loops	Chemical equivalency Structural equivalency Exact discovery	0 1 0	0 0 0	0 0 0	0 0 0	0 0 0
20	Helix-in-turn-helix-out loops	Chemical equivalency Structural equivalency Exact discovery	0 0 0	0 0 0	0 0 0	0 0 0	0 0 0
21	All dipping loops	Chemical equivalency Structural equivalency Exact discovery	0 0 0	0 0 0	0 0 0	0 0 0	0 0 0

Table 4.10. Summary of the pattern discovery process and filtering process carried out using PATTERNTEST. The number of patterns common to: a) membrane dipping loop belonging to the same dipping loop protein group; b) membrane dipping loop structural type; c) all the membrane dipping loops listed. Evaluations 13 and 19 were based on the same sequences, as there were no other loop-in-helix-out domains outside the aquaglycerol family. Only the loops belonging to the major intrinsic protein family were found to share common amino acid patterns.

Crystallization analyses of the membrane proteins containing membrane dipping loops have shown that the general structure of the aquaglyceroporin family, the chloride channel family and the secY/SEC61 α family is composed of an internal tandem repetition. According to these findings, it would be likely that common patterns were found in the corresponding loops belonging to both repetitions in the protein. However, this internal repetition could not be identified at the sequential level in the case of chloride channels and protein conducting channels and only the corresponding solved structures revealed an early gene duplication event. The solved structure of the chloride channels maintained the topological symmetry between both halves and loops were

conserved in both internal repetitions. On the other hand, secY/SEC61 α-subunit proteins only conserved the dipping loop in the amino-terminal half of the protein. The results obtained for the discovery of common patterns in membrane dipping loops belonging to different proteins families **(table 4.10)** were in accordance with the degree of similarity between both halves of these protein families. Our approach discovered 27 specific patterns in loop 1 and loop 2 in the aquaglyceroporin family (this analysis corresponds to evaluation 13 and 19 as there were no other loop-in-helix-out domains found outside the aquaglyceroporin family), and the patterns with the top 2 highest scores were "[ST]G...**NP[AG]**" (support 0.86) and "[ST]G...**NPA**" (support 0.85). These common patterns, not present in membrane dipping loops belonging to other protein families, emphasize the importance of the NPA motif as a functional motif for aquaglyceroporin family. In the case of chloride channels, common non-specific patterns were found in loops 1-3 and 2-4 (evaluation 14 and 15), whereas no patterns were found in all 4 membrane dipping loops. The latter was expected as the reported functional roles of loops 1-3 and loops 2-4 are quite different. Comparing the results regarding common patterns in loops involved in selectivity belonging to the aquaglycerol protein family with the common patterns in loops belonging to the chloride channel family, it could be suggested that the dipping loops belonging to the repeated halves in both protein families emerged by a similar gene duplication event, but in evolution these proteins have been exposed to different pressures. These different pressures have lead to different selectivity mechanisms probably motivated by the nature of the different substrates transported across the membrane and the different needs of the cell for the transported molecules. Cells have a permanent need for water molecules that requires a quasi-continuous transport of water across the membrane, however the known functional roles that involved transport of chloride ions, e.g. transfer of electrical signals across the cells and stabilization of the membrane potential, require a short and rapid transport of these ions. These different needs have lead to a similar structural arrangement bringing together the polar ends of short α-helices approximately half way through the membrane but at the same time have developed a fast gating mechanism for the unidirectional transport of chloride ions through the membrane that might explain the divergence in sequence of both halves in the chloride channel family, whereas aquaglyceroporins have maintained a highly selective open pore, which would not have required divergent evolution of the protein halves. In the case of the secY/SEC61 α family, it was not possible to analyse the presence of

common patterns in the dipping loops belonging to both halves because one of the halves has not maintained the membrane dipping loop. In terms of evolution, this could be explained by the unidirectional transport of polypeptide chains and the need for the protein to bind the ribosome on the intracellular side of the membrane. In order to bind the ribosome, these proteins might have been forced to suppress the dipping loop located in the intracellular side of the membrane as it might have imposed a steric impedance for the transport of the polypeptide, and therefore only the membrane dipping loop on the extracellular side of the membrane was conserved to act as a gate.

No common patterns were found in loops belonging to: the sodium:dicarboxylate symporter family, helix-in-turn-loop-out motif, helix-in-turn-helix-out motif or the general dipping loop motif **(table 4.10)**. This reflects that the residues contained in membrane dipping loops respond to a local functional need of the protein rather than to a similar folding mechanism. However, the absence of common patterns in different membrane dipping loops should not be used to infer different folding mechanisms events for different membrane dipping loops. Neighbouring residues located in different transmembrane regions may play an important role in stabilising the motif and minimizing the energy penalty imposed when locating polar or charged residues in the lipid bilayer and therefore a hypothetical common folding event might also be dictated by neighbouring residues and the common surrounding environment.

4.4 Conclusions

Membrane dipping loops often play essential functional roles in the mechanism of action of polytopic α-helical membrane proteins. However, no extensive analyses have been carried out to determine conserved patterns in these motifs and identify potential functionally important residues. Therefore, a rigorous pattern discovery protocol has been applied in order to identify conserved residues in the structural domains of protein families (where at least one crystallized structure contained a membrane dipping loop). Additionally, the developed method was also applied to detect common patterns in membrane dipping loops belonging to different protein families, but with a similar arrangement of secondary structures (helix-in-turn-helix-out, helix-in-turn-loop-out, loop-in-turn-helix-out). Interestingly, some of the residues contained in discovered

patterns were already described as functionally important by experimental analysis, thus validating our approach. Furthermore, discovered patterns also contained residues whose functional roles have not yet been fully characterized, leading to the potential for targeting further experimental research aimed at understanding the exact functional roles of these residues and understanding of the mechanism of action of membrane proteins with membrane dipping loops. No patterns were discovered in membrane dipping loops belonging to different protein families but sharing a similar arrangement of secondary structure. These results highlight the important functional role of the discovered sequence patterns rather than that of different folding events. The discovered patterns are a reliable resource to be used for the prediction of membrane dipping loops and functional characterization of proteins with unknown function.

4.5 References

AGRE, P. & KOZONO, D. (2003) Aquaporin water channels: molecular mechanisms for human diseases. *FEBS Lett,* 555, 72-8.

ALTSCHUL, S. F., GISH, W., MILLER, W., MYERS, E. W. & LIPMAN, D. J. (1990) Basic local alignment search tool. *J Mol Biol,* 215, 403-10.

ATTWOOD, T. K., FLOWER, D. R., LEWIS, A. P., MABEY, J. E., MORGAN, S. R., SCORDIS, P., SELLEY, J. N. & WRIGHT, W. (1999) PRINTS prepares for the new millennium. *Nucleic Acids Res,* 27, 220-5.

BALDI, P., CHAUVIN, Y., HUNKAPILLER, T. & MCCLURE, M. A. (1994) Hidden Markov models of biological primary sequence information. *Proc Natl Acad Sci U S A,* 91, 1059-63.

BARABOTE R. D., HILLER M. B. & SAIER M. H. (2006) Transport Classification Database.

BERMAN, H. M., WESTBROOK, J., FENG, Z., GILLILAND, G., BHAT, T. N., WEISSIG, H., SHINDYALOV, I. N. & BOURNE, P. E. (2000) The Protein Data Bank. *Nucleic Acids Res,* 28, 235-42.

BERNECHE, S. & ROUX, B. (2000) Molecular dynamics of the KcsA K(+) channel in a bilayer membrane. *Biophys J,* 78, 2900-17.

BOECKMANN, B., BAIROCH, A., APWEILER, R., BLATTER, M. C., ESTREICHER, A., GASTEIGER, E., MARTIN, M. J., MICHOUD, K., O'DONOVAN, C., PHAN, I., PILBOUT, S. & SCHNEIDER, M. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res,* 31, 365-70.

BUCHER, P., KARPLUS, K., MOERI, N. & HOFMANN, K. (1996) A flexible motif search technique based on generalized profiles. *Comput Chem,* 20, 3-23.

BYRNE, B. & IWATA, S. (2002) Membrane protein complexes. *Curr Opin Struct Biol,* 12, 239-43.

CHENNA, R., SUGAWARA, H., KOIKE, T., LOPEZ, R., GIBSON, T. J., HIGGINS, D. G. & THOMPSON, J. D. (2003) Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res,* 31, 3497-500.

COLLINSON, I. (2005) The structure of the bacterial protein translocation complex SecYEG. *Biochem Soc Trans,* 33, 1225-30.

DARZENTAS, N., RIGOUTSOS, I. & OUZOUNIS, C. A. (2005) Sensitive detection of sequence similarity using combinatorial pattern discovery: a challenging study of two distantly related protein families. *Proteins,* 61, 926-37.

DAVIDSON, A. L. (2002) Structural biology. Not just another ABC transporter. *Science,* 296, 1038-40.

DE GROTTHUSS, D. (1809) Sur la deÂcomposition de l'eau et des corps qu'elle tient en dissolution aÁ l'aide de. l'eÂlectriciteÂ galvanique. *Ann Chim,* 58, 54-74.

DOYLE, D. A., MORAIS CABRAL, J., PFUETZNER, R. A., KUO, A., GULBIS, J. M., COHEN, S. L., CHAIT, B. T. & MACKINNON, R. (1998) The structure of the potassium channel: molecular basis of K+ conduction and selectivity. *Science,* 280, 69-77.

DUTZLER, R., CAMPBELL, E. B., CADENE, M., CHAIT, B. T. & MACKINNON, R. (2002) X-ray structure of a ClC chloride channel at 3.0 A reveals the molecular basis of anion selectivity. *Nature,* 415, 287-94.

DUTZLER, R., CAMPBELL, E. B. & MACKINNQN, R. (2003) Gating the selectivity filter in ClC chloride channels. *Science,* 300, 108-12.

EDDY, S. R. (1996) Hidden Markov models. *Curr Opin Struct Biol,* 6, 361-5.

ESTEVEZ, R. & JENTSCH, T. J. (2002) CLC chloride channels: correlating structure with function. *Curr Opin Struct Biol,* 12, 531-9.

FALQUET, L., PAGNI, M., BUCHER, P., HULO, N., SIGRIST, C. J., HOFMANN, K. & BAIROCH, A. (2002) The PROSITE database, its status in 2002. *Nucleic Acids Res,* 30, 235-8.

FUJIYOSHI, Y., MITSUOKA, K., DE GROOT, B. L., PHILIPPSEN, A., GRUBMULLER, H., AGRE; P. & ENGEL, A. (2002) Structure and function of water channels. *Curr Opin Struct Biol,* 12, 509-15.

GONEN, T., SLIZ, P., KISTLER, J., CHENG, Y. & WALZ, T. (2004) Aquaporin-0 membrane junctions reveal the structure of a closed water pore. *Nature,* 429, 193-7.

HALDRUP, A., SIMPSON, D. J. & SCHELLER, H. V. (2000) Down-regulation of the PSI-F subunit of photosystem I (PSI) in Arabidopsis thaliana. The PSI-F subunit is essential for photoautotrophic growth and contributes to antenna function. *J Biol Chem,* 275, 31211-8.

HARRIES, W. E., AKHAVAN, D., MIERCKE, L. J., KHADEMI, S. & STROUD, R. M. (2004) The channel architecture of aquaporin 0 at a 2.2-A resolution. *Proc Natl Acad Sci U S A,* 101, 14045-50.

HENIKOFF, J. G., HENIKOFF, S. & PIETROKOVSKI, S. (1999) New features of the Blocks Database servers. *Nucleic Acids Res,* 27, 226-8.

HIPPLER, M., DREPPER, F., ROCHAIX, J. D. & MUHLENHOFF, U. (1999) Insertion of the N-terminal part of PsaF from Chlamydomonas reinhardtii into photosystem I from Synechococcus elongatus enables efficient binding of algal plastocyanin and cytochrome c6. *J Biol Chem,* 274, 4180-8.

HOBOHM, U., SCHARF, M., SCHNEIDER, R. & SANDER, C. (1992) Selection of representative protein data sets. *Protein Sci,* 1, 409-17.

HORLACHER, R., XAVIER, K. B., SANTOS, H., DIRUGGIERO, J., KOSSMANN, M. & BOOS, W. (1998) Archaeal binding protein-dependent ABC transporter: molecular and biochemical analysis of the trehalose/maltose transport system of the hyperthermophilic archaeon Thermococcus litoralis. *J Bacteriol,* 180, 680-9.

IWATA, S., LEE, J. W., OKADA, K., LEE, J. K., IWATA, M., RASMUSSEN, B., LINK, T. A., RAMASWAMY, S. & JAP, B. K. (1998) Complete structure of the 11-subunit bovine mitochondrial cytochrome bc1 complex. *Science,* 281, 64-71.

JIANG, Y., LEE, A., CHEN, J., CADENE, M., CHAIT, B. T. & MACKINNON, R. (2002) Crystal structure and mechanism of a calcium-gated potassium channel. *Nature,* 417, 515-22.

JIANG, Y., LEE, A., CHEN, J., RUTA, V., CADENE, M., CHAIT, B. T. & MACKINNON, R. (2003) X-ray structure of a voltage-dependent K+ channel. *Nature,* 423, 33-41.

JONASSEN, I., COLLINS, J. F. & HIGGINS, D. G. (1995) Finding flexible patterns in unaligned protein sequences. *Protein Sci,* 4, 1587-95.

JORDAN, P., FROMME, P., WITT, H. T., KLUKAS, O., SAENGER, W. & KRAUSS, N. (2001) Three-dimensional structure of cyanobacterial photosystem I at 2.5 A resolution. *Nature,* 411, 909-17.

JUNNE, T., SCHWEDE, T., GODER, V. & SPIESS, M. (2006) The plug domain of yeast Sec61p is important for efficient protein translocation, but is not essential for cell viability. *Mol Biol Cell,* 17, 4063-8.

KROGH, A., BROWN, M., MIAN, I. S., SJOLANDER, K. & HAUSSLER, D. (1994) Hidden Markov models in computational biology. Applications to protein modeling. *J Mol Biol,* 235, 1501-31.

KUO, A., GULBIS, J. M., ANTCLIFF, J. F., RAHMAN, T., LOWE, E. D., ZIMMER, J., CUTHBERTSON, J., ASHCROFT, F. M., EZAKI, T. & DOYLE, D. A. (2003) Crystal structure of the potassium channel KirBac1.1 in the closed state. *Science,* 300, 1922-6.

LEE, A. G. (2003) Lipid-protein interactions in biological membranes: a structural perspective. *Biochim Biophys Acta,* 1612, 1-40.

LOCHER, K. P., LEE, A. T. & REES, D. C. (2002) The E. coli BtuCD structure: a framework for ABC transporter architecture and mechanism. *Science,* 296, 1091-8.

LONG, S. B., CAMPBELL, E. B. & MACKINNON, R. (2005) Crystal structure of a mammalian voltage-dependent Shaker family K+ channel. *Science,* 309, 897-903.

MACKINNON, R. (2003) Potassium channels. *FEBS Lett,* 555, 62-5.

MITRA, K., SCHAFFITZEL, C., SHAIKH, T., TAMA, F., JENNI, S., BROOKS, C. L., 3RD, BAN, N. & FRANK, J. (2005) Structure of the E. coli protein-conducting channel bound to a translating ribosome. *Nature,* 438, 318-24.

MURATA, K., MITSUOKA, K., HIRAI, T., WALZ, T., AGRE, P., HEYMANN, J. B., ENGEL, A. & FUJIYOSHI, Y. (2000) Structural determinants of water permeation through aquaporin-1. *Nature,* 407, 599-605.

NISHIDA, M. & MACKINNON, R. (2002) Structural basis of inward rectification: cytoplasmic pore of the G protein-gated inward rectifier GIRK1 at 1.8 A resolution. *Cell,* 111, 957-65.

OSTERMEIER, C. & MICHEL, H. (1997) Crystallization of membrane proteins. *Curr Opin Struct Biol,* 7, 697-701.

PAO, G. M., WU, L. F., JOHNSON, K. D., HOFTE, H., CHRISPEELS, M. J., SWEET, G., SANDAL, N. N. & SAIER, M. H., JR. (1991) Evolution of the MIP family of integral membrane transport proteins. *Mol Microbiol,* 5, 33-7.

PEARSON, W. R. & LIPMAN, D. J. (1988) Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A,* 85, 2444-8.

PRESTON, G. M. & AGRE, P. (1991) Isolation of the cDNA for erythrocyte integral membrane protein of 28 kilodaltons: member of an ancient channel family. *Proc Natl Acad Sci U S A,* 88, 11110-4.

REN, G., CHENG, A., REDDY, V., MELNYK, P. & MITRA, A. K. (2000) Three-dimensional fold of the human AQP1 water channel determined at 4 A resolution by electron crystallography of two-dimensional crystals embedded in ice. *J Mol Biol,* 301, 369-87.

REN, G., REDDY, V. S., CHENG, A., MELNYK, P. & MITRA, A. K. (2001) Visualization of a water-selective pore by electron crystallography in vitreous ice. *Proc Natl Acad Sci U S A,* 98, 1398-403.

RIGOUTSOS, I. & FLORATOS, A. (1998) Combinatorial pattern discovery in biological sequences: The TEIRESIAS algorithm. *Bioinformatics,* 14, 55-67.

SAGOT, M. F., VIARI, A. & SOLDANO, H. (1995) A distance-based block searching algorithm. *Proc Int Conf Intell Syst Mol Biol,* 3, 322-31.

SAVAGE, D. F., EGEA, P. F., ROBLES-COLMENARES, Y., O'CONNELL, J. D., 3RD & STROUD, R. M. (2003) Architecture and selectivity in aquaporins: 2.5 a X-ray structure of aquaporin Z. *PLoS Biol,* 1, E72.

SAYLE, R. & BISSEL, A. (1992) RasMol: A Program for Fast Realistic Rendering of Molecular Structures with Shadows. *Proccedings of the 10th Eurographics UK 1992.* University of Edinburgh, UK.

STROUD, R. M., MIERCKE, L. J., O'CONNELL, J., KHADEMI, S., LEE, J. K., REMIS, J., HARRIES, W., ROBLES, Y. & AKHAVAN, D. (2003a) Glycerol facilitator GlpF and the associated aquaporin family of channels. *Curr Opin Struct Biol,* 13, 424-31.

STROUD, R. M., SAVAGE, D., MIERCKE, L. J., LEE, J. K., KHADEMI, S. & HARRIES, W. (2003b) Selectivity and conductance among the glycerol and water conducting aquaporin family of channels. *FEBS Lett,* 555, 79-84.

SUI, H., HAN, B. G., LEE, J. K., WALIAN, P. & JAP, B. K. (2001) Structural basis of water-specific transport through the AQP1 water channel. *Nature,* 414, 872-8.

TUSNADY, G. E., DOSZTANYI, Z. & SIMON, I. (2004) Transmembrane proteins in the Protein Data Bank: identification and classification. *Bioinformatics,* 20, 2964-72.

VAN DEN BERG, B., CLEMONS, W. M., JR., COLLINSON, I., MODIS, Y., HARTMANN, E., HARRISON, S. C. & RAPOPORT, T. A. (2004) X-ray structure of a protein-conducting channel. *Nature,* 427, 36-44.

WANG, L. & JIANG, T. (1994) On the complexity of multiple sequence alignment. *J Comput Biol,* 1, 337-48.

WU, T. D. & BRUTLAG, D. L. (1995) Identification of protein motifs using conserved amino acid properties and partitioning techniques. *Proc Int Conf Intell Syst Mol Biol,* 3, 402-10.

YERNOOL, D., BOUDKER, O., JIN, Y. & GOUAUX, E. (2004) Structure of a glutamate transporter homologue from Pyrococcus horikoshii. *Nature,* 431, 811-8.

ZHOU, Y., MORAIS-CABRAL, J. H., KAUFMAN, A. & MACKINNON, R. (2001) Chemistry of ion coordination and hydration revealed by a K+ channel-Fab complex at 2.0 A resolution. *Nature,* 414, 43-8.

CHAPTER 5

TMLOOP, a bioinformatics tool to predict membrane dipping loops

5.1 Introduction

Membrane dipping loops play essential functional roles within membrane proteins. As discussed in **chapter 4**, membrane dipping loops have been proposed to act as selectivity filters in potassium channels (Doyle et al., 1998, Jiang et al., 2002, Jiang et al., 2003, Kuo et al., 2003, Long et al., 2005, Nishida and MacKinnon, 2002, Zhou et al., 2001), aquaglyceroporins (Gonen et al., 2004, Harries et al., 2004, Murata et al., 2000, Ren et al., 2000, Ren et al., 2001, Savage et al., 2003, Stroud et al., 2003, Sui et al., 2001), and chloride channels (Dutzler et al., 2002, Dutzler et al., 2003), as molecular gates of membrane pores, such as in the glutamate homolog transporter and the protein conducting channel (Van den Berg et al., 2004) and also as linking structural motifs between subunits in chloride channels (Estevez and Jentsch, 2002). Prediction of membrane dipping loops from protein sequence has proved difficult as such regions are frequently amphiphilic, containing hydrophobic sections that are too intermittent to be identified as membrane regions. Membrane dipping loops require interactions with adjacent highly hydrophobic helices to become inserted in the membrane and minimise the energy penalty imposed by the location of polar or charge residues in a low dielectric environment. *In-silico* topology prediction approaches often fail to predict membrane dipping loops in polytopic α-helical membrane proteins due to their residue composition differing with that membrane spanning segments. To date, the bioinformatics approaches of our group, working on the membrane dipping loops of glycerol channels, in collaboration with Stefan Hohmann and colleagues, have relied upon homology modelling (Bill et al., 2001) and comparison of test sequences with those of known loops in terms of secondary structure and the propensity scoring of successive residues to reside in α or β conformation (Hedfalk et al., 2004, Karlgren et

al., 2004, Tamas et al., 2003), underpinned by extensive laboratory work including measuring channel efflux, mutagenesis and genetic screening.

In this chapter, we describe the development of a novel and reliable approach to the difficult problem of predicting membrane dipping loops directly from sequence that may be generically applied to membrane proteins. The pattern discovery approach carried out using TEIRESIAS (Rigoutsos and Floratos, 1998) and PATTERNTEST **(Chapter 4)** led to the development of a bioinformatics tool, named TMLOOP, to predict membrane dipping loops using discovered patterns as weighted predictive rules. This software was used to explore the performance of a single motif method compared to a variation of this approach, called the collective motif method. Single motif methods, such as the PROSITE database (Falquet et al., 2002), describe a structural motif, catalytic site or protein family using a unique motif. Therefore, this method requires exact pattern matching to find structural or functional relatedness and can miss distant relatives, which contain small variations of the pattern (Scordis et al., 1999). By evolution, the constraints imposed in the outside world have been reflected in changes at the molecular level to ensure the adaptability of an organism to a changing environment. In **chapter 4**, the functional importance of many of the residues contained in the membrane dipping loops and their direct relevance to the molecular function of the protein was described. Evolution of membrane dipping loops might reflect evolutionary pressures for changes in the gating process of a membrane protein, the re-adjustment of specificity for the corresponding ligand according to new needs of the cell, or even binding of a different ligand in a similar fashion. All these pressures, are ultimately translated into small variations of the residues associated with these motifs and rearrangements with the interacting helices. Divergent evolution would generate 2 membrane dipping loops from a common ancestral structural motif with small variations in sequence, which allowed the binding of the same ligand with different specificity or binding different ligands in a similar fashion. On the other hand, phylogenetically unrelated proteins might generate a similar three-dimensional membrane dipping loop to bind the same or a similar ligand, which would involve two different membrane dipping loops containing a similar but not identical selective filter or binding site. The collective method is based on the use of different patterns discovered using TEIRESIAS and PATTERNTEST **(Chapter 4)**, these patterns belong to the same structural motif and contain small variations of the discovered pattern with

the highest support. Following this principle, the collective motif method appears to be a more flexible approach than the single motif method and distantly related membrane dipping loops containing similar but not identical sequences in the corresponding structural motif can be co-detected.

TMLOOP had been shown to successfully predict membrane dipping loops but it had not been implemented to describe the starting and ending position of the structural domain, instead it reports the average pattern starting position and the average length of the pattern, which might not be related to the actual starting and ending position of a particular membrane dipping loop. In order to approximate the boundaries of the structural domain an extension, named TMLOOP writer, has been implemented. This extension of the TMLOOP algorithm represents the final stage of an iterative loop **(figure 5.1)**. This iterative loop involves the pattern discovery approach applied to membrane dipping loops **(Chapter 4)**, implementation of a predictive tool (TMLOOP) for prediction of these particular structural domains, evaluation of predicted loops, and description of such domains in the Swiss-Prot database (TMLOOP writer).

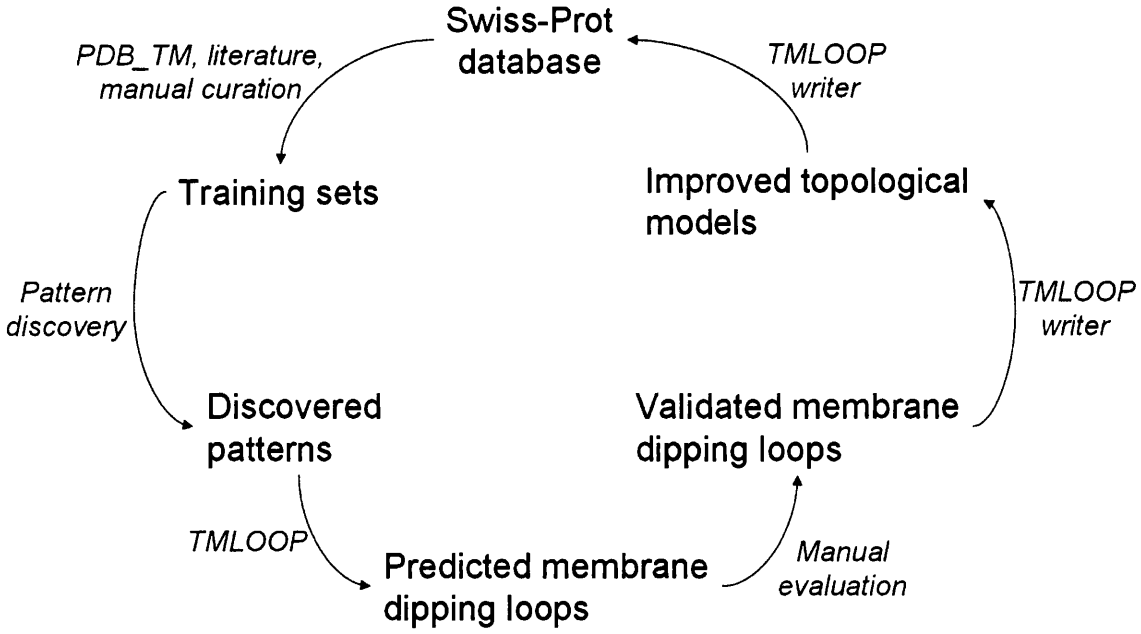

Figure 5.1. Iterative loop describing the prediction and annotation of membrane dipping loops. The training sets assembled from the Swiss-Prot database were analyzed by Teiresias (Rigoutsos and Floratos, 1998) in order to discover patterns of residues present in these domains. Validated patterns (using PATTERNTEST, **Chapter 4**) were used as predictive rules by TMLOOP. TMLOOP was applied to the Swiss-Prot database in order to predict new membrane dipping loops. TMLOOPwriter annotated the manually validated membrane dipping loops in the corresponding transmembrane sections of the Swiss-Prot database.

5.2 Methods

5.2.1 TMLOOP implementation

5.2.1.1 Description

TMLOOP was implemented to predict membrane dipping loops in polytopic membrane proteins. TMLOOP uses patterns, using a regular expression format, discovered by TEIRESIAS and validated by PATTERNTEST **(Chapter 4)** as weighted predictive rules where the weight was calculated by dividing the number of sequences in the training set containing a particular pattern by the total number of sequences in the training set. TMLOOP requires a set of user-defined parameters and allows the user to decide whether to perform a single motif search or a collective motif search. The single motif method describes a structurally or functionally important site in sequence using a unique motif. Therefore this approach requires exact pattern matching to find structural or functional relatedness and can miss distant relatives, which contain small variations of the pattern (Scordis et al., 1999). By contrast, the collective motif method is based on the use of different partially overlapping patterns, which belong to the same structural or functional motif and therefore distant relative proteins containing small variations of the most common patterns can be co-detected.

There are three user-defined parameters required to run the prediction: i) I is the minimum inter-loop length required between two contiguous loops, where two pattern matches would point to the same loop only if the distance of both matches in the sequence is lower than I; ii) S, the minimum pattern support, which restricts the patterns used for the prediction such that only the patterns whose support is equal or higher than S would be used as a predictive rule (only applicable in the collective motif method); and iii) C, the minimum prediction confidence, which restricts the reporting of protein matches to those predictions with a score equal or higher than C (only applicable in the collective motif method).

5.2.1.2 The algorithm

Two versions of TMLOOP have been implemented, a web version (http://membraneproteins.swan.ac.uk/TMLOOP) and an executable desktop version. Both versions perform load sequences in FASTA format, the basic output is a text-format-like output where the predicted membrane dipping loop is described by the following parameters: i) loop type, ii) loop score, iii) average pattern starting point and iv) average pattern length.

5.2.1.3 Software development

The algorithm was implemented using Borland Delphi, an object oriented programming language used for rapid application development (RAD). The web version of TMLOOP uses a Common Gateway Interface (CGI), with a Microsoft IIS (Internet Information Server) web server. The code generated by Delphi is a console application, which is placed in the IIS web server and called by the input at theTMLOOP interface by HTML code.

The basic architecture of TMLOOP was implemented following the model-view-controller (MVC) fashion where the interface and the functionality of the program are to be considered as different layers that are indirectly linked by a cross-linking layer represented by the class TController **(figure 3.3)**. The functionality of TMLOOP is performed by five different classes: TTmLoop, TClassifier, TDippingGroup, TPattern and TPredLoop. TMLOOP first loads the query sequences input into memory and analyzes the format input by the user. If the input is correct, the software loads into memory the different user-required parameters (prediction method, I, S, C and output format) and checks that no incorrect values have been input. If parameters were correctly input TMLOOP loads into memory the corresponding pattern files (either files belonging to the single motif method or to the collective motif method). TMLOOP then analyzes the query sequences by pattern matching using the Microsoft VBScript_RegExp_55_TLB library.

The pattern files were previously assembled using the TEIRESIAS and PATTERNTEST software and must follow a specific format **(figure 5.2)**.

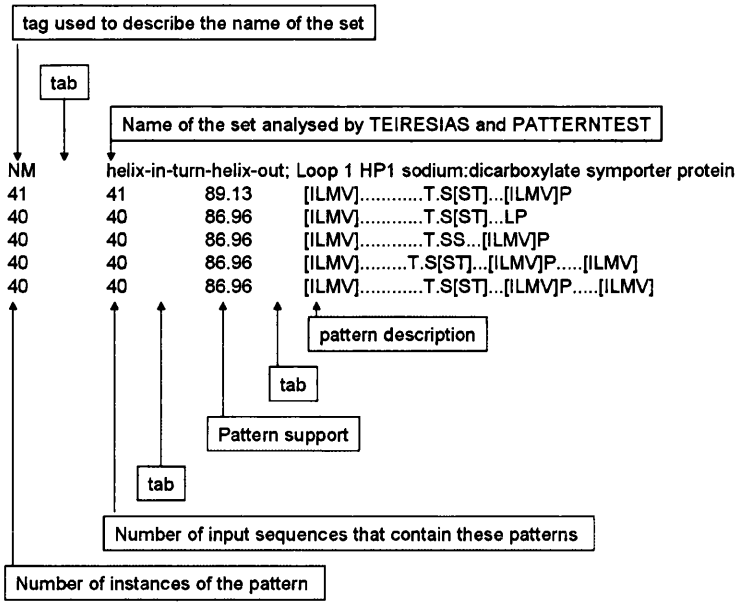

Figure 5.2. Example of a pattern file loaded by TMLOOP.

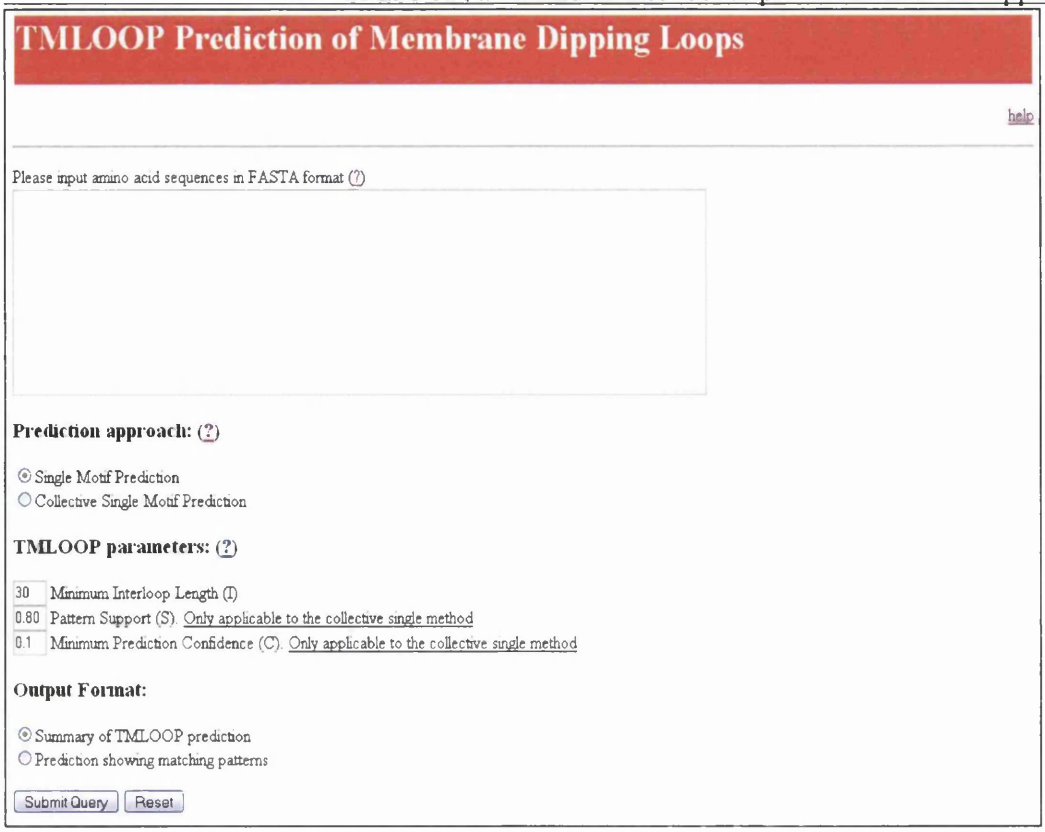

Figure 5.3. Input interface of the web version of TMLOOP

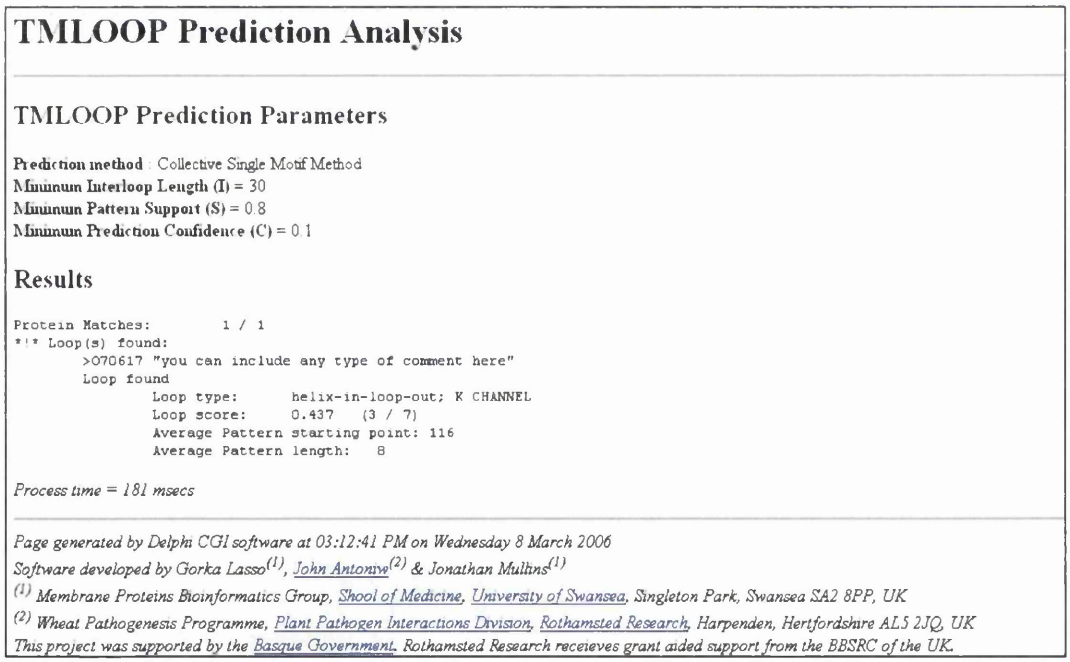

Figure 5.4. Output interface of the web version of TMLOOP

5.2.2 Collection of patterns from membrane dipping loops

Patterns discovered by TEIRESIAS using an amino acid chemical equivalency set, structural equivalency set and exact discovery (identity) and validated by PATTERNTEST **(Chapter 4)** were brought together into a single text file for each membrane dipping loop analysed. Included patterns were listed following the format required by TMLOOP **(figure 5.2)**. In order to prevent pattern redundancy, each assembled set of patterns was manually checked and pairs of identical patterns were removed. This was necessary as TEIRESIAS was found to report identical patterns (which do not contain any ambiguity) using different pattern discovery analyses (chemical equivalence, structural equivalence and exact discovery). The remaining set of patterns for each of the membrane dipping loops analysed was considered in the set to be used in the collective motif method where partially overlapping patterns describe the same membrane dipping loop.

For the single motif method, the pattern with the highest support listed in each of the set of patterns assembled for the collective motif method was selected. If the highest support value found in any of these sets was shared by more than one pattern an empirical selection of the pattern was achieved by evaluating the predictive power of each pattern against a test set composed of proteins ruled out during the redundancy filtering process using the Non-Red software, when assembling gold standard sets for the pattern discovery process **(Chapter 4)**.

5.2.3 TMLOOP evaluation

TMLOOP was evaluated by tenfold cross-validation. The sensitivity and specificity of TMLOOP was calculated at different parameters of I, S and C and the advantages and disadvantages of both the single motif method and the collective motif method were assessed. The 2 pore domain potassium channel family was used as a test set to evaluate TMLOOP and the collective motif method; this protein family has been predicted to contain two membrane dipping loops in its structure and it is believed to form homodimers (mimicking the homotetramer assembled in the structure of single pore potassium channels). This family showed itself to be a good candidate for the

validation as none of its family members has been included in the potassium channel gold standard set and the regions in sequence belonging to both membrane dipping loops are not identical and therefore the second membrane dipping loop would be missed by TMLOOP using the single motif method. The evaluation was also used to set up the default parameters of TMLOOP in order to maximize the accuracy of the prediction.

5.2.4 Membrane dipping loop prediction in the Swiss-Prot database

TMLOOP was applied to the Swiss-Prot database (version 48.0). Only proteins containing two or more α-helical transmembrane regions (polytopic membrane proteins) were analysed. Both the single motif method and the collective motif method were applied and the parameters I, S and C were optimized to maximize the sensitivity and specificity of the method according to the evaluation of TMLOOP using a test set of 172 α-helical membrane proteins. Predicted membrane dipping loops were classified as true positives, false positives or possible membrane dipping loops that have not yet been experimentally tested. In order to identify possible hitherto undesignated membrane dipping loops, it was necessary to identify structural or functional relatedness to the corresponding crystallized protein type known to have a similar membrane dipping loop. Membrane dipping loops have been shown to play essential roles in protein function, most of the membrane dipping loops characterized act either as selectivity filters or pore gates **(Chapter 4)**. Therefore when no structural evidence is available it is desirable to look for functional relatedness. A pipeline of verification approaches was designed to identify possible membrane dipping loops that may merit experimental testing:

- Database annotation: The Swiss-Prot database contains links to other family and domain databases, such as InterPro and the PRINTS database, that can be used to identify structural or functional relatedness. Additionally the International Union of Biochemistry and Molecular Biology (IUBMB, http://www.chem.qmul.ac.uk/iubmb/) and the Transport Classification Database (http://www.tcdb.org/) were also used to infer structural or functional relatedness.
- BLASTP analysis: This tool (http://www.expasy.org/tools/blast/) was used to

find remote homology to the corresponding protein type containing the predicted membrane dipping loop. BLASTP was used in a similar fashion as Darzentas and colleagues (Darzentas et al., 2005). Query sequences were checked against the Swiss-Prot database of curated sequences with an E-value cutoff of e-100.

- Analysis of local residue conservation: If no structural or functional evidence was discovered by using methods described above, the residue conservation in the predicted membrane dipping loop region was analyzed using ClustalW. The matching query sequence predicted to have a particular membrane dipping loop was aligned with the sequences belonging to the corresponding gold standard set containing the predicted membrane dipping loop. When significant residue conservation was observed, the predicted membrane dipping loop was considered as a possible loop to be experimentally tested.
- Analysis of the predicted membrane dipping loop in sequence: Usually membrane dipping loops are not hydrophobic enough to insert in the membrane by themselves and often place polar or charge residues in the membrane. Therefore these loops require interactions with highly hydrophobic helices to become inserted in the membrane and minimise the energy penalty imposed when locating polar or charge residues in a low dielectric environment. Following this principle, crystallized membrane proteins with membrane dipping loops have been shown to contain these structural motifs between alpha transmembrane helices connected by short extramembraneous loops. Therefore the relative position of the predicted membrane dipping loops with respect to the position of transmembrane helices in the sequence was also analyzed. Membrane dipping loops located far away from transmembrane region were unlikely but possible. On the other hand membrane dipping loops located close to the amino- or carboxy-terminus were considered as false positives.

5.2.5 TMLOOP writer implementation

5.2.5.1 Calculation of the boundaries of predicted membrane dipping loops

TMLOOP successfully predicts membrane dipping loops but it does not report

the boundaries of the structural domain, the only information regarding the localization of the membrane dipping loop refers to the average pattern starting point and the average pattern length. The principles of homology modeling were followed in order to approximate the boundaries of the predicted membrane dipping loops. Both the membrane dipping loops found in the corresponding crystallized structure and the average pattern starting point for each method (single and collective motif method) were mapped onto the alignments used to isolate the sequences pertaining to membrane dipping loops. The distance between the membrane dipping loop starting point (by homology) and the average pattern starting point was then measured for each of the sequences contained in the training set used by TMLOOP and an average distance between the structural domain starting point and the pattern starting point was then computed **(figure 5.5)**. This average distance was different for each loop depending on the prediction method used by TMLOOP, therefore for each loop the average distance was computed for the i) single motif method ii) collective motif method and iii) single and collective motif method (average of the distances computed for the single and collective motif methods) **(table 5.1)**. The membrane dipping loop length was calculated by measuring the length of the membrane dipping loops found in the corresponding crystallized structures. Using these two computed parameters it was then possible to approximate the limits of the predicted structural domain considering only the average pattern starting point and the prediction method used by TMLOOP.

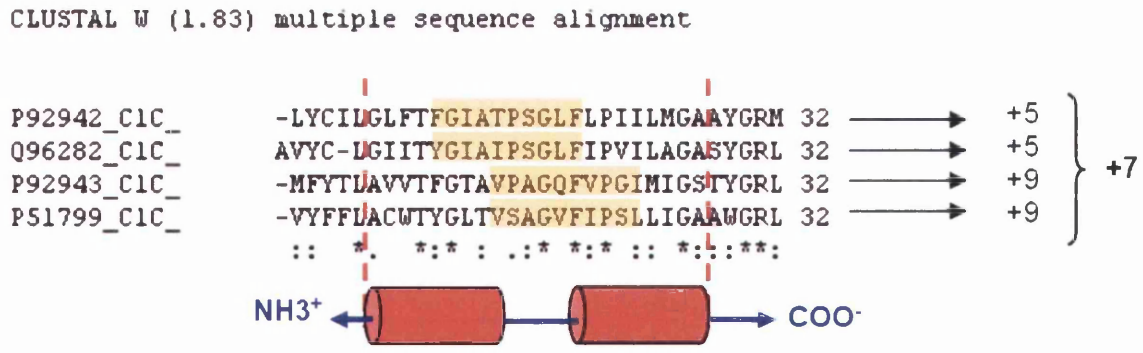

Figure 5.5. Calculation of the average distance between a hypothetical membrane dipping loop and the average starting points of a hypothetical set of patterns predicted by TMLOOP using the consensus motif. The multiple sequence alignment is used to infer the structural domain to the aligned sequences. For each sequence the distances between the structural domain and the average pattern starting position is computed. As the consensus motif uses diverse patterns, the average pattern starting position might not be the same for each sequence. In order to approximate the position of the membrane dipping loop these distances were averaged.

Membrane dipping loop type	Prediction method	Membrane dipping loop	
		Starting point	Length
HIHO-SDF1	S	+3	24
	C	-1	
	S&C	+1	
HILO-SDF2	S	+5	21
	C	-1	
	S&C	+2	
HIHO-CLC1	S	-10	21
	C	-6	
	S&C	-8	
HIHO-CLC2	S	0	21
	C	0	
	S&C	0	
HIHO-CLC3	S	-9	21
	C	-9	
	S&C	-9	
HIHO-CLC4	S	-3	24
	C	0	
	S&C	-2	
HIHO-PSAF	S	0	17
	C	+4	
	S&C	+2	
HILO-K^+	S	-7	17
	C	-8	
	S&C	-8	
HIHO-BPDPFECD	S	-11	26
	C	-5	
	S&C	-8	
LIHO-AQP1	S	+1	15
	C	+2	
	S&C	+2	
LIHO-AQP2	S	-2	15
	C	+1	
	S&C	-1	
LIHO-AQP1&2	S	+1	15
	C	+1	
	S&C	+1	

Table 5.1. Calculation of the localization parameters of the characterized membrane dipping loops. The second column specifies the prediction methods used by TMLOOP ("S" for single motif method, "C" for collective motif method and "S&C" for both). The third column refers to the average distance between the starting position of the structural domain and the average starting position of the pattern considering a particular prediction method. Positive numbers indicate downstream positions (e.g. +2 indicates that the structural domain starts two positions after the average pattern starting position) and vice versa.

5.2.5.2 Inclusion of membrane dipping loops in the transmembrane statement

The main purpose of the TMLOOP writer was to include the predicted membrane dipping loops in a SwissProt-like text file and rewrite the statement of those predicted dipping loop regions found to overlap with transmembrane. There are five possible scenarios when including a predicted membrane dipping loop in the transmembrane statement of Swiss-Prot like text files **(figure 5.6)**: i) the beginning of the membrane dipping loop overlaps with a transmembrane region, ii) the end of the membrane dipping loop overlaps with a transmembrane region, iii) the beginning of the membrane dipping loop overlaps with a transmembrane region and the end of the membrane dipping loop overlaps with the next transmembrane region, iv) the membrane dipping loop completely overlaps with a transmembrane region and v) the membrane dipping loop does not overlap with any transmembrane region. The latter case is the simplest case as no transmembrane statement is affected by the prediction and inclusion of the membrane dipping loop, however in the first four cases it is necessary to recalculate the limits of the transmembrane regions affected by the inclusion of the membrane dipping loop. In these cases the region of the transmembrane region overlapping with the membrane dipping loops was deleted and it was necessary to determine whether the remaining transmembrane region belonged to a separate transmembrane region. Since it is theoretically possible that the interhelical loop could be composed of as few as one residue, one extra residue was considered to allow for the interhelical loop between transmembrane regions when deleting the overlapping region of the affected transmembrane region. Research carried out by Monné and coworkers (Monne et al., 1999) suggested that the minimal helical hairpin was composed of two 14 residues transmembrane helices. These results were in accordance with a comparative analysis of crystallized α-helical membrane proteins carried out where the minimum helical length of a helix that completely transversed the membrane was found to be 15 residues. Following these results it was agreed to consider as separate transmembrane regions those remaining regions whose length was found to be 14 or higher. The segments whose length after deletion of the overlapping region (plus one residue as interhelical loop) was lower than 14 residues were deleted from the corresponding transmembrane statement.

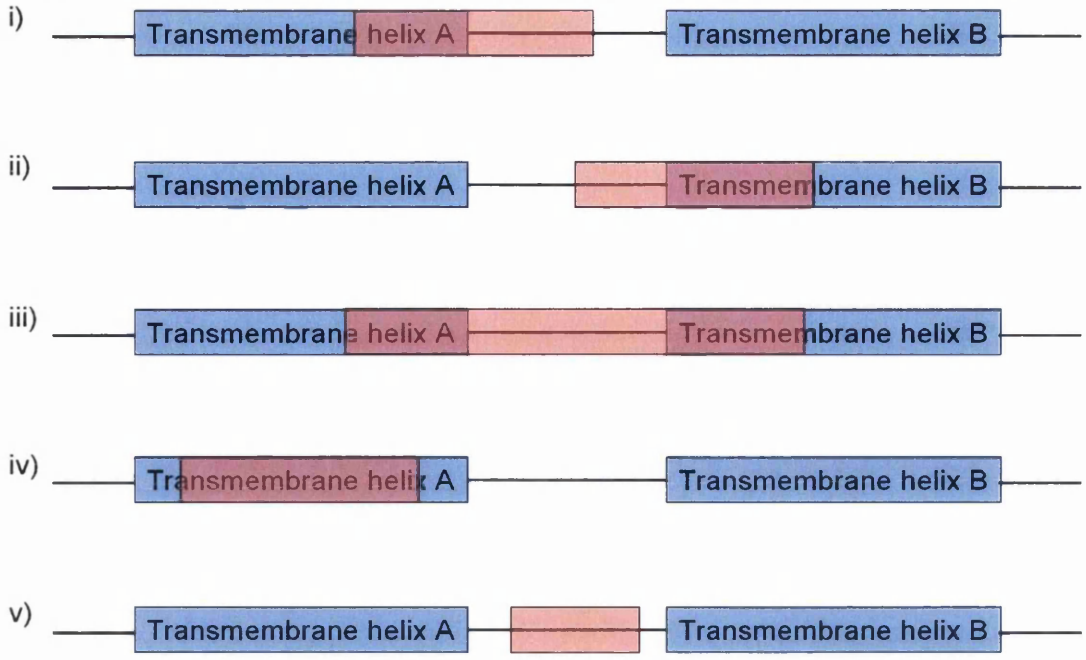

Figure 5.6. The possible overlapping scenarios when including a predicted membrane dipping loop (in red) in a transmembrane statement describing transmembrane regions (in blue) predicted by different topology prediction softwares.

Following the format of the Swiss-Prot statements, the membrane dipping loop was included in the corresponding Swiss-Prot like text file following the description of the transmembrane region previous to the membrane dipping loop. Keywords were also included in a similar fashion to allow further bioinformatics tools to exploit the information generated by TMLOOP and TMLOOP writer **(figure 5.7)**. The membrane dipping loop statement begins with the Swiss-Prot keyword "FT", which corresponds to the feature table (this statement usually describes posttranslational modifications, binding sites, enzyme active sites and local secondary structures), the second term in the statement corresponds to the term "MEMBLOOP", which refers to the membrane dipping loop domain, the third and fourth element in the statement corresponds to the beginning and ending of the structural domain, the fifth element corresponds to the membrane dipping loop type **(table 5.2)** according to the TMLOOP membrane dipping loop nomenclature, the last element of the statement refers to the tool used to predict the membrane dipping loop. If the boundaries of a transmembrane region are needed to be re-calculated due to overlap between the annotated transmembrane region and the membrane dipping loop, a notification of the update will be included at the end of the corresponding transmembrane statement.

```
FT   TRANSMEM      88      108       1 (Potential).
FT   TRANSMEM     131      151       2 (Potential).
FT   MEMBLOOP     188      208       HIHO-CLC1    Predicted by TMLOOP.
FT   TRANSMEM     210      226       TMLimits refined by TMLOOP writer.
FT   MEMBLOOP     247      267       HIHO-CLC2    Predicted by TMLOOP.
FT   TRANSMEM     278      298       6 (Potential).
```

Figure 5.7. Example of the transmembrane statement section modified by TMLOOP writer. The statements highlighted in red correspond to the membrane dipping loops predicted by TMLOOP whereas the statement in blue corresponds to the transmembrane region whose limits were refined by the TMLOOP writer.

5.2.5.3 TMLOOP input file development

The input file **(figure 5.8)** contains a description of all true positives found by TMLOOP in the Swiss-Prot database **(Please see S5.True MDL input file on CD)**. Each membrane dipping loop is described by the corresponding Swiss-Prot accession code of the protein containing the structural domain, the membrane dipping loop type found, the prediction method used to detect the corresponding loop (single motif method, collective motif method or both) and the average starting position of the matching patterns. The TMLOOP prediction method is specified by the characters "S" and "C" (corresponding to the single and collective motif method respectively).

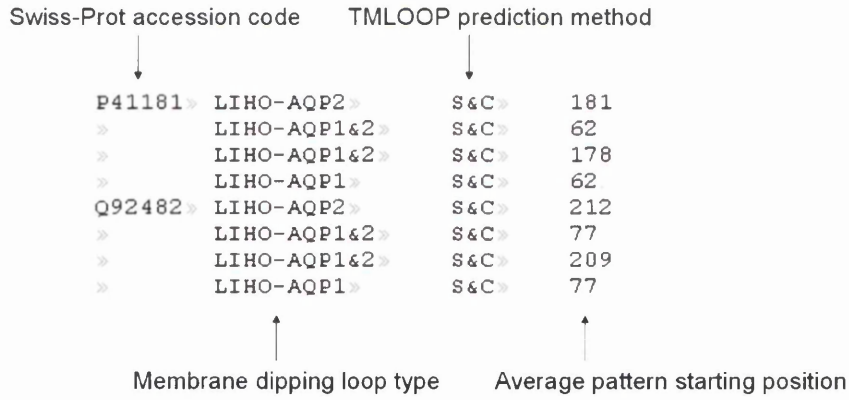

Figure 5.8. Example of the input text file to be loaded by TMLOOP writer. TMLOOP writer searches for the corresponding Swiss-Prot like text file and includes the corresponding membrane dipping loop information in the transmembrane statement. In order to include this information TMLOOP writer needs first to calculate the membrane dipping loop starting point, which is computed by considering the prediction method used by TMLOOP, the average pattern starting position and the average distance between the average pattern starting position and the structural domain starting position.

The membrane dipping loop nomenclature is composed of two sections, the first section refers to the structural type of the membrane dipping loop and the second section refers to the characterized membrane dipping loop used to construct the matching patterns **(table 5.2)**.

Characterized membrane dipping loop	TMLOOP writer code
Loop 1 aquaglyceroporin	LIHO-AQP1
Loop 2 aquaglyceroporin	LIHO-AQP2
Loop 1 & 2 aquaglyceroporin	LIHO-AQP1&2
Binding protein dependent permase FeCD subfamily	HIHO-BPDPFECD
Loop 1 ClC chloride channel	HIHO-CLC1
Loop 2 ClC chloride channel	HIHO-CLC2
Loop 3 ClC chloride channel	HIHO-CLC3
Loop 4 ClC chloride channel	HIHO-CLC4
Potassium channel	HILO-K+
Loop 1 Sodium/dicarboxylate symporter	HIHO-SDF1
Loop 2 Sodium/dicarboxylate symporter	HILO-SDF2
psaF family	HIHO-PSAF

Table 5.2. Abbreviation for each of the membrane dipping loops characterized by TMLOOP. The first term refers to the membrane dipping loop structural type: Loop-in-turn-helix-out (LIHO), helix-in-turn-helix-out (HIHO) and helix-in-turn-loop-out (HILO). The second term refers to the protein type found to contain the corresponding loop and if more than one membrane dipping loop is present, the corresponding membrane dipping loop number of the structural domain.

5.3 Results & Discussion

5.3.1 Single motif predictive rule selection

Table 5.3 lists the patterns with the highest support discovered using TEIRESIAS for each membrane dipping loop **(Chapter 4)**. As described in the method section, if for a particular membrane dipping loop more than one pattern is found with the highest support, an evaluation was carried out against a test set composed of proteins ruled out during the redundancy filtering process using Non-Red software when assembling gold standard sets for the pattern discovery process **(Chapter 4)**. This was the case for loop 1 in aquaglyceroporins, loop 1 and loop 4 in chloride channels and the loop found in psaF proteins.

If the predictive score against a test set was found to be identical for these patterns (as was the case of the loop 4 in chloride channels and the loop belonging to the psaF family), the pattern with highest number of elements shared by all the patterns sharing the same support and performance score was selected .

Membrane dipping loop	Support	Top pattern	Predictive score against a test set
L1: Loop-in-turn-helix-out Aquaglyceroporin family	0.96	SG...N..[ILMV][ST]	0.88
		SG.H.N...[ST]	0.91
		[ITV]SG.H.N	0.86
		[ITV]SG...N.A	0.87
L2: Loop-in-turn-helix-out Aquaglyceroporin family	0.94	[ILMV]NP.R.....[ILMV]	-
L1 & L2: Loop-in-turn-helix-out Aquaglyceroporin family	0.86	[ST]G...NP[AG]	
Helix-in-turn-helix-out Binding protein dependent transport system permease family	0.96	[AG].[ILMV].F[ILMV][AG] L[ILMV].P.[ILMV]	
L1: Helix-in-turn-helix-out Sodium : dicarboxylate (SDF) symporter family	0.89	[ILMV].....T.S[ST]...[IL MV]P	
L2: Helix-in-turn-loop-out Sodium : dicarboxylate (SDF) symporter family	0.96	[ILMV].....[ILMV].....S.G. .[AG][ILMV].....[ILMV].[IL MV].....[ILMV]	
L1: Helix-in-turn-helix-out Chloride channel family	0.86	**[ILMV]G[KR].GP.[ILMV]**	0.82
		[ILMV]G..GP.V	0.64
		[ILMV]G..GP.[ILMV].....[A G]	0.75
L2: Helix-in-turn-helix-out Chloride channel family	1.0	[AG].[AG].G[ILMV]...[FY]. ....[AG]..F..E	-
L3: Helix-in-turn-helix-out Chloride channel family	0.91	P.G...P....G...G	
L4: Helix-in-turn-helix-out Chloride channel family	0.96	**[AG].....[ILMV]...[ILM V][ILMV]..E[ILMV]T**	0.82
		[AG]....[AG].....[ILMV]...[IL MV][ILMV]..E[ILMV][ST]	0.82
		[AG].....[ILMV][ST]..[IL MV][ILMV]..E.[ST]	0.82
Helix-in-turn-loop-out Potassium channel	0.89	[ST]..[ST].G[FY]G	-
Helix-in-turn-helix-out psaF family	1.0	[ILMV]....A.....G..WP..A	1.0
		A.....G..WP..A..[EQ]	1.0
		A.....G..WP..A	1.0

Table 5.3 Selection of patterns for the single motif method. Patterns with the highest support found for each gold standard set analysed with TEIRESIAS. When more than one pattern with the highest support was found for a given membrane dipping loop (as in the case of the loop 1 of aquaglyceroporins) an empirical selection of the pattern was carried out by evaluating each pattern against a test set. The pattern with the highest predictive score was selected (in bold). If the predictive score against a test set was shown to be identical for these patterns (as in the case of the loop belonging to the psaF family), the pattern with the most number of elements shared by all the patterns sharing the same support and performance score was selected (in bold).

5.3.2 TMLOOP evaluation

The main problem of single motif methods, is that prediction of a structural motif or functional category depends upon exact matching with a single pattern. Therefore distantly related proteins containing small variations of the pattern can not be detected. With TMLOOP, a single motif method (using the single pattern with the highest support found for each membrane dipping loop) can be employed to predict a particular membrane dipping loop, or alternatively a set of partially overlapping patterns, can be used as weighted predictive rules (collective motif method). The single motif approach and the collective motif approach were evaluated by tenfold cross-validation using various combinations of C and S validation **(table 5.4)**. The minimum interloop length was set to 30, which was observed to be the maximum length observed for a membrane dipping loop **(table 4.6)**.

				S			Top score
				70	80	90	pattern
C	0.01	Sensitivity	Av	95.64	92.87	43.24	87.05
			Sd	2.58	2.98	5.57	3.2
		Specificity	Av	95.93	98.18	98.88	100
			Sd	6.93	4.42	0.46	0
	0.1	Sensitivity	Av	90.57	**92.43**	43.09	87.05
			Sd	3.05	**2.48**	5.56	3.2
		Specificity	Av	95.25	**100**	100	100
			Sd	2.36	**0**	0	0
	0.3	Sensitivity	Av	85.93	87.09	41.36	87.05
			Sd	3.85	3.73	5.69	3.2
		Specificity	Av	100	100	100	100
			Sd	0	0	0	0
	0.5	Sensitivity	Av	77.35	84.76	41.07	87.05
			Sd	5.19	4.28	5.04	3.2
		Specificity	Av	100	100	100	100
			Sd	0	0	0	0
	0.7	Sensitivity	Av	63.55	76.78	38.45	87.05
			Sd	5.83	4.16	5,14	3.2
		Specificity	Av	100	100	100	100
			Sd	0	0	0	0

Table 5.4. Evaluation of TMLOOP by ten fold cross-validation. Two different approaches were carried out : i) using the pattern for each membrane dipping loop set with the highest score (top score approach, which is a single motif approach, in orange) using an I value of 30; and ii) using various values of S (minimum pattern support) and C (minimum prediction confidence) with a fixed I value (minimum inter-loop length) of 30 (collective motif approach, in blue). The results shown with the white background are the data relating to the optimal performance of TMLOOP.The top score approach, which proved to be a conservative approach, gave a confidence of 1.0 for each prediction since here TMLOOP uses just one rule per membrane dipping loop considered and therefore the prediction is based upon exact single pattern matching (either yes or no). The sensitivity (overall percentage of true positives) and specificity (overall percentage of true negatives) computed for the ten evaluations have been averaged (Av), the corresponding standard deviation (Sd) has also computed.

			S			Top score
			70	80	90	pattern
C	0.01	Sensitivity	90.32	82.26	-	37.01
		Specificity	100	91.94	91.94	100
	0.1	Sensitivity	82.26	82.26	-	37.01
		Specificity	100	100	91.94	100
	0.3	Sensitivity	45.16	40.32	-	37.01
		Specificity	100	100	100	100
	0.5	Sensitivity	8.06	40.32	-	37.01
		Specificity	100	100	100	100
	0.7	Sensitivity	3.26	33.87	-	37.01
		Specificity	100	100	100	100

Table 5.5. Prediction of membrane dipping loops in the two pore domain potassium family. Two different approaches were carried out: i) using the pattern for each membrane dipping loop set with the highest score (top score approach, which is a single motif approach, in orange) using an I value of 30; and ii) using various values of S (minimum pattern support) and C (minimum prediction confidence) with a fixed I value (minimum inter-loop length) of 30 (collective motif approach, in blue). The results shown with the white background are the data relating to the optimal performance of TMLOOP.

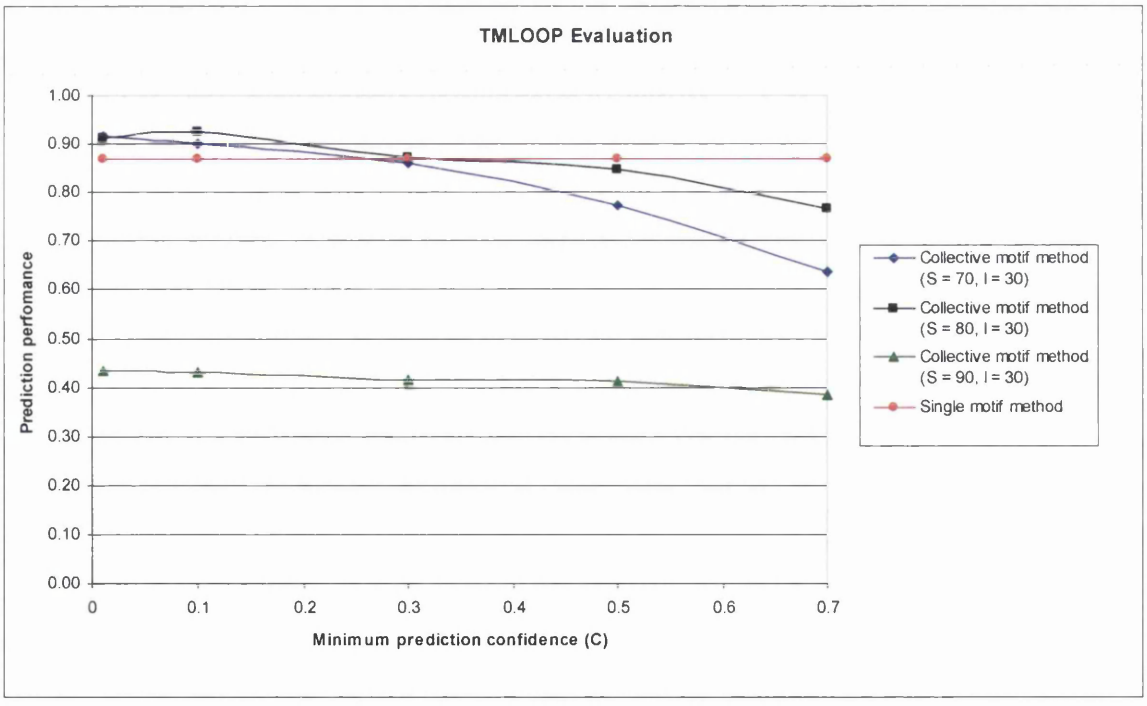

Figure 5.9. Comparison of the performance of single and collective motif methods tested by tenfold cross-validation. This graph shows the prediction performance (considering both the sensitivity and specificity) of each TMLOOP analysis (i. the single motif method in red, ii. the collective method –S (minimum pattern support) = 70, I (minimum inter-loop length) = 30- in blue, iii. the collective motif method –S = 80, I = 30- in black and iv. the collective motif method –S = 90, I = 30 in green) carried out at various levels of minimum prediction confidence (C). The collective method (S = 80, I = 30) showed the highest predictive score at a minimum confidence value C of 0.1. The C value of 0.3 is considered to be the threshold, below which the most accurate prediction method is the collective motif method and above which the single motif method performs better.

As expected, both approaches performed well against the training set, although the single motif method approach was found to be more accurate as the confidence parameter (C) increased **(table 5.4, figure 5.9)**. The single motif method predicts membrane dipping loops with a confidence of 1. This is because this method only uses one pattern to predict the structural motif and therefore prediction depends upon exact matching with the pattern used for each gold standard set. By contrast, in the case of the collective single motif method, to obtain a prediction with a confidence value of 1 it is necessary that all patterns considered by TMLOOP match the sequence. The reason why the prediction accuracy of TMLOOP dropped significantly when the minimum pattern support parameter (S) was set to 90 in the collective motif approach **(table 5.4, figure 5.9)** was simply because some of the sets of patterns did not have a single pattern whose support was 0.90 or higher **(table 5.3)** and therefore no patterns were considered for the prediction of the given membrane dipping loop. It was intended to use ROC curves to evaluate the prediction accuracy of TMLOOP by measuring the area under the curve (AUC). However, the obtained results **(Please see S5.Evaluation on CD)** showed that the specificity in the majority of the different evaluation folds is 100% which would generate a plot where the majority of the data points of the curve are located on the *x* axis (1-specificity equal to 0 for most of the data points). Consequently, the obtained AUC value will be 0 for many of the different set of parameters (S and C), thus not reflecting the real predictive accuracy of the method. Therefore, in this case, the ROC curve is not an informative evaluation method due to the nature of the evaluation results.

The evaluation showed that the collective approach (S = 80, C = 0.1, I = 30) was the most accurate method where TMLOOP achieved a sensitivity of 91.4% and a specificity of 100% (predictive score = 0.92, **table 5.4** and **figure 5.9**). Although the single motif method was found to be a better approach with higher values of C, it also proved to be a conservative prediction as expected. The flexibility of the collective motif approach allowed TMLOOP to detect 90.32% of the dipping loops contained in the two pore domain potassium channel family, in contrast to the 37.01% obtained by the single motif approach **(table 5.5)**. As discussed above, the two pore domain potassium channel family has been predicted to contain 2 membrane dipping loops, and unlike other potassium channels this family does not have identical membrane dipping

loops in its structure and the second loop contains small variations in sequence compared to the first membrane dipping loop. The 37.01% of predicted loops by the single motif method corresponded to the 74.02% of the first loops contained in the test set using only 2 pore domain potassium channels. This fact reflects that the single motif method is not suitable for predicting novel membrane dipping loops not yet experimentally discovered whereas the collective single motif method has shown to be capable of detecting membrane dipping loops distantly related to membrane dipping loops used in the gold standard sets. The comparison of the single motif method and the collective motif method **(figure 5.9)** showed that at confidence values higher than 0.3 the single motif method is more accurate than the collective motif method and *vice versa*.

5.3.3 Membrane dipping loop prediction across the Swiss-Prot database

TMLOOP was applied to the Swiss-Prot database to predict membrane dipping loops in polytopic membrane proteins. Prediction was carried out by the single motif method approach using only the pattern with the highest support for each membrane dipping loop analyzed (I = 30) and the collective motif approach using TMLOOP with S, C and I set to 80, 0.1 and 30 respectively. Version 48 of the Swiss-Prot database contained 194,317 sequence entries where 29,127 sequence entries corresponded to α-helical membrane proteins. Table 5.6 and table 5.7 summarize the prediction of membrane dipping loops. TMLOOP identified 1637 membrane dipping loops in 850 membrane proteins where 85.28% of those loops corresponded to true positives, 10.08% corresponded to false positives and 4.64% corresponded to possible loops. As expected the single motif method was shown to be a more conservative approach where 1209 loops (89.03%) were identified as true positives, 117 loops (8.62%) were identified as false positives and 32 loops (2.36%) were identified as possible loops. The collective single motif method detected 1595 membrane dipping loops, 1392 loops (87.27%) were identified as true positives, 128 loops (8.03%) were considered false positives and 75 loops (4.7%) were considered as possible loops. Table 5.7 lists all possible loops identified by both methods not yet experimentally tested. These loops were validated as described in the methods section.

		True positives	False positives	Potential loops
Single motif method	Membrane dipping loops	1209	117	32
	Proteins	581	115	32
Collective motif method	Membrane dipping loops	1392	128	75
	Proteins	605	128	75
Consensus prediction	Membrane dipping loops	1204	78	31
	Proteins	576	78	31

Table 5.6. Membrane dipping loop prediction in the Swiss-Prot database. The table summarises the analysis of the SwissProt database using TMLOOP (a) when only the pattern with the highest support is used (single motif approach) and (b) when all patterns whose support is $\geq$ 80 are used and only predictions with score $\geq$ 0.1 are reported (collective motif approach). The I value (minimum inter-loop length) was set to 30 for both methods. The last two rows (in red) show the consensus prediction considering both approaches.

Swiss-Prot accession code	Definition	Predicted membrane dipping loop
Q9H2Y9	Solute carrier organic anion transporter family, member 5A1	helix-in-turn-helix-out ClC choride channel loop-1 like loop
Q8KWT2, Q8KWS7, P39642	Putative bacilysin exporter bacE	Loop-in-turn-helix-out Loop 1 & 2 aquaporin like
Q9NRA2, Q8BN82	Sialin (Solute carrier family 17 member 5)	helix-in-turn-helix-out ClC choride channel loop-1 like loop
Q58902	Hypothetical protein MJ1507	helix-in-turn-loop-out K^+ channel like
Q64SU9	Hypothetical transport protein BF2680	helix-in-turn-helix-out ClC choride channel loop-1 like loop
Q7UH36	Hypothetical transport protein RB4869	helix-in-turn-loop-out K^+ channel like
Q8AAG5	Hypothetical transport protein BT0500	helix-in-turn-helix-out ClC choride channel loop-1 like loop
Q57943	Hypothetical protein MJ0523	helix-in-turn-helix-out ClC choride channel loop-4 like loop
Q8NSS8	Hypothetical transport protein Cgl0590/cg0683	helix-in-turn-loop-out K^+ channel like
P74635	Hypothetical protein slr0753	helix-in-turn-helix-out ClC choride channel loop-1 like loop
P0AAC6, P0AAC7	Inner membrane protein yccA	helix-in-turn-helix-out Na^+:dicarboxylate symporter loop-1 like loop
P38745	Hypothetical 61.2 kDa protein in APM2-DUR3 intergenic region precursor	helix-in-turn-loop-out K^+ channel like
P37643	Inner membrane metabolite transport protein yhjE	helix-in-turn-loop-out K^+ channel like
P54181	Hypothetical protein ypnP	helix-in-turn-loop-out K^+ channel like
Q9V7S5	Putative inorganic phosphate cotransporter	helix-in-turn-helix-out ClC choride channel loop-1 like loop
P0A629, P0A628	Phosphate transport system permease protein pstC-1	helix-in-turn-helix-out ClC choride channel loop-1 like loop
P10603, P27182	ATP synthase C chain	helix-in-turn-helix-out ClC choride channel loop-1 like loop
P0A304, P0A305	ATP synthase C chain	helix-in-turn-loop-out K^+ channel like
Q8YGH4, Q8G1E6	Pyrophosphate-energized proton pump	helix-in-turn-loop-out K^+ channel like
P34299, Q8LGN0, Q9C5V5, O81078, Q9ULK0, Q61627, Q62640	Glutamate receptor precursor (glutamate-gated ion channel)	helix-in-turn-loop-out K^+ channel like
Q58671	Probable Na(+)/H(+) antiporter 3 (MjNapA)	helix-in-turn-loop-out K^+ channel like
Q15629, Q01685, Q15629, Q91V04, Q9GKZ4	Translocation associated membrane protein 1	helix-in-turn-loop-out K^+ channel like
Q8XED4, Q8FCT7, P33650, Q57IW8, Q5PLZ1, Q83ST5, P74884, Q57986, P73182	Ferrous iron transport protein B	helix-in-turn-loop-out K^+ channel like
Q97QP7, Q54875, Q59947, Q59986	Immunoglobulin A1 protease precursor	helix-in-turn-loop-out K^+ channel like
Q09917	Hypothetical protein C1F7.03 in chromosome I	helix-in-turn-loop-out K^+ channel like
Q8IZK6, Q8K595	Mucolipin-2	helix-in-turn-loop-out K^+ channel like
P91645, Q13936, Q01815, P15381, P22002, Q24270, Q01668, Q99244, P27732, O60840, Q02789, P07293, Q9JIS7, Q13698, O57483, Q02485, O73700, Q25452, P22316	Voltage-dependent calcium channel alpha-1 subunit	helix-in-turn-loop-out K^+ channel like
O28069 (top score pattern approach)	Hypothetical protein AF2214	helix-in-turn-helix-out ClC choride channel loop-4 like loop

Table 5.7. List of proteins, including the corresponding Swiss-Prot accession codes, containing plausible membrane dipping loops according to TMLOOP. Proteins listed were predicted by using either the single motif approach (I = 30) and/or the collective motif approach (S = 80, C = 0.1 and I = 30).

The following section is a detailed discussion of the loop types examined.

Solute carrier organic anion transporter 5A1

TMLOOP has predicted a helix-in-helix-out ClC chloride channel loop 1 like loop. This loop has been proposed to act as a selectivity filter for chloride ions **(Chapter 4)**. The predicted loop overlaps with the predicted TM5 of the protein. Although no link has been found between this protein and chloride channels it might be possible that this protein uses a similar structural motif to filter organic anions transported through the membrane.

Putative bacilysin exporter bacE

This protein belongs to the major facilitator superfamily, which has a broad specificity of ligands. The predicted loop is a loop-in-turn-loop-out loop 1 & 2 aquaglyceroporin like loop, which overlaps with the predicted TM4. The major facilitator superfamily catalyzes solute : cation symport and/or solute : proton or solute : solute antiport. There is no clear ligand for this protein but the detected NPA motif in the membrane might indicate a selectivity filter located in the membrane.

Sialin

This protein belongs to the major facilitator superfamily and the sodium/anion cotransporter family. The loop predicted in its structure is a helix-in-turn-helix-out ClC chloride channel loop 1 like loop. This protein has been proposed in the Swiss-Prot database (its function has not been experimentally tested) to transport anionic substances through the membrane and its function as a free sialic acid transporter in the lysosomes is highlighted. BLASTP analysis links this protein to other sodium/anion cotransporters but not chloride channels. The predicted structural motif overlaps with the predicted TM9 of the protein. As in the case of the solute carrier organic anion transporter 5A1 the Sialin protein might have developed a loop similar to the loop predicted by TMLOOP to be used as a selectivity filter for anionic substances.

Hypothetical protein MJ1507

TMLOOP predicted a helix-in-turn-helix-out potassium channel like loop at position 40, which coincides with the end of the predicted TM1. BLASTP analysis showed distant relationships with an aquaporin and other transporters (e.g. macrilide-specific ABC-type efflux carrier, lipoprotein-releasing system transmembrane proteins and bacitracin export permease). Therefore, it might be possible for this protein to act as a transporter. The predicted loop then might act as a filter for cationic substances (ions or molecules) as it has been shown that the predicted loop acts as a potassium selective filter in these ion channels.

Hypothetical transport proteins BF2680, RB4869, BT0500, CgI0590/cg0683.

According to the Swiss-Prot database these proteins belong to the aspartate:alanine exchanger family. This protein family catalyzes a negative charge movement through the membrane. The loop predicted by TMLOOP in BF2680 and BT0500 is a a helix-in-turn-helix-out ClC chloride channel loop 1 like loop, which overlaps with the predicted TM8. The loop predicted in RB4869 and CgI0590/cg0683 is a helix-in-turn-loop-out potassium channel like loop, which overlaps with TM5 and TM7 respectively. BLASTP analyses of these proteins showed distant relationships with ion channels such as the potassium channel and the cyclic nucleotide-gated cation channel, and also proteinS such as the aspartate/alanine antiporter and the aerobic C4-dicarboxylate transport protein (which contains two membrane dipping loops in its strucure). All these distant relationships might indicate the possible charged substance (ions or molecules) transport capacity of these proteins. It is believed that these proteins might have 2 membrane dipping loops in their structure, the first loop would correspond to that predicted in RB4869 and CgI0590/cg0683 and the second loop would correspond to the loop predicted in BF2680 and BT0500. Both loops would be three-dimensionally associated in a similar fashion as the selectivity filters for ClC chloride channels, however it is not possible to elucidate what type of ligands these proteins would transport.

Hypothetical protein MJ0523

TMLOOP predicted a helix-in-turn-helix-out loop 4 ClC chloride channel like loop. This loop has been suggested to link the repeated halves of ClC chloride channel within each monomer and make contacts with each other at the interface between monomers (Estevez and Jentsch, 2002). The predicted loop overlaps with the predicted TM3 and BLASTP analysis shows a distant relationship with ClC Chloride channels. As the predicted loop is not linked directly to the function of ClC chloride channels but indirectly by stabilizing the protein, it is difficult to validate this membrane dipping loop. BLASTP analyses support the prediction made by TMLOOP. According to the InterPro database (Mulder et al., 2005), this protein belongs to the IPR011311, which contains small membrane proteins that are predicted to be transmembrane subunits of multi-subunit membrane-bound [NiFe]-hydroenase Eha complexes, the predicted loop might then be important for the assembly of the protein complex by linking different subunits in the membrane.

Hypothetical protein s1r0753

This protein contains a predicted helix-in-turn-helix-out loop 1 ClC chloride channel like loop, which ovelaps with predicted TM11. According to the Swiss-Prot database, this protein belongs to the divalent anion:sodium symporter family. Members of this family transport organic di- and tricarboxylates of the Kreb's cycle, dicarboxylate amino acid, inorganic sulphate and phosphate through the membrane (Saier and colleages, Division of Biolgical Science at UC San Diego, http://www.tcdb.org/). BLASTP analyses showed links to other divalent anion:sodium symporters, therefore this protein might use the predicted membrane dipping loop as a selective filter to transport negatively charged substances (ions or molecules) through the membrane.

Inner membrane protein yccA

TMLOOP has detected a helix-in-turn-helix-out loop 1 sodium:dicarboxylate symporter protein in this protein, which corresponds to TM1 according to the topological model of the protein in the Swiss-Prot database. No function has been

determined yet for this protein but BLASTP analyses link this protein to transporters such as metal transporters, lipoprotein transporters, permeases, amino acid transporters, ion channels and the sodium:dicarboxylate symporter supporting the prediction made by TMLOOP and the assumption made by Saier and colleagues (Saier and colleages, Division of Biolgical Science at UC San Diego, http://www.tcdb.org/) that postulate that this protein could act either as a transporter or a receptor. The finding made by TMLOOP also supports this functional prediction and emphasizes its transport properties.

Hypothetical 61.2kDa protein in APM2-DUR3 intergenic region precursor

TMLOOP predicted a helix-in-loop-out potassium channel like membrane dipping loop in the region between predicted TM3 and TM4. BLASTP analysis showed distant relations with cationic transporters such as ammonium transporters, mucolipin and calcium channels. Therefore, it might be possible that this hypothetical protein contains a membrane dipping loop similar to the membrane dipping loop found in potassium channels whose function would be to act as a selectivity filter for cationic substances (ions or molecules).

Inner membrane metabolite transport protein yhjE

This protein belongs to the major facilitator superfamily and the sugar transporter family. TMLOOP predicted a helix-in-loop-out potassium channel like membrane dipping loop in the region corresponding to TM8. BLASTP analyses showed links to proton cotransporters and organic cation transporters. Therefore, it might be possible that the predicted membrane dipping loop acts as a selectivity filter for cationic substances (ion or molecule) using a similar mechanism as the potassium channel.

Hypothetical protein ypnP

TMLOOP predicted a helix-in-loop-out potassium channel like membrane dipping loop in the region corresponding to TM3. The Swiss-Prot database linked this protein to the InterPro domain IPR002528, which belongs to the multi antimicrobial extrusion family whose function is to mediate resistance (drug/sodium antiporters) to a

wide range of cationic dyes, fluroquinolones, aminoglycosides and other structurally diverse antibodies and drugs. BLASTP analyses showed that the protein is distantly related to cation cotransporters and potassium channels and therefore it could be possible that this protein contains a membrane dipping loop similar to that found in potassium channels.

Putative inorganic phosphate cotransporter

According to the Swiss-Prot database this protein belongs to the major facilitator superfamily and the sodium/anion cotransporter family. TMLOOP predicted a helix-in-turn-helix-out loop 1 ClC chloride channel like loop in the region corresponding to the extramembraneous loop between TM6 and TM7. BLASTP analysis revealed distant relationships with other sodium:anion cotransporters such as sialin, which was previously predicted by TMLOOP as a protein containing a similar membrane dipping loop.

Phosphate transport system permease protein pstC-1

According to the Swiss-Prot database this protein is part of a complex and its function is probably related to the translocation of the substrate across the membrane. TMLOOP predicted a helix-in-turn-helix-out loop 1 ClC chloride channel like loop in the region corresponding to the amino-terminus. Considering the transport activity of the protein, the essential role of membrane dipping loops as selectivity filters in the ClC chloride channel, and the anionic character of the ions transporter, it is possible that this protein type contains a membrane dipping loop.

ATP synthase C chain

TMLOOP has predicted two different membrane dipping loops in the region corresponding to the first transmembrane region in different Swiss-Prot files. These loops are the helix-in-turn-helix-out loop 1 ClC chloride channel like loop and the helix-in-turn-loop-out potassium channel like loop. Although the predicted loops are contradictory in terms of the nature of the ligand transported through the membrane, this might indicate a novel membrane dipping loop. This protein catalyzes the transport of

protons across the membrane therefore this possible "novel" membrane dipping loop might be involved in filtering the protons in a similar way as ClC chloride channels and potassium channels.

Pyrophosphate-energized proton pump

This protein belongs to the proton translocating pyrophosphatase family. Phylogenetic studies led to a subclassification of this protein family (Belogurov et al., 2002). The first protein subfamily is composed of potassium independent proton pyrophosphatases, which contain a conserved cysteine (Cys 222) whereas the members belonging to the second subfamily are potassium dependent and contain a conserved cysteine in position 573 instead of position 222. It is not known whether potassium is transported through the membrane (Saier and colleages, Division of Biolgical Science at UC San Diego, http://www.tcdb.org/). According to the Swiss-Prot file this protein belongs to the second protein subfamily and therefore is potassium dependent. TMLOOP predicted a helix-in-turn-loop-out potassium channel like loop, this prediction is supported by the fact that the protein has been annotated as potassium dependent and the BLASTP analysis, which showed a distant relationship with a probable potassium channel. Further experimental analyses should focus on the validation of the predicted loop and a possible interaction between the residue in position 222, Cys573 and the predicted structural motif located in predicted TM5.

Glutamate receptor precursor

TMLOOP predicted a helix-in-turn-loop-out potassium channel like loop in the region between predicted transmembrane TM2 and TM3. This protein belongs to the glutamate-gated ion channel family of neurotransmitters receptors. The transport classification database (Saier and colleages, Division of Biolgical Science at UC San Diego, http://www.tcdb.org/) postulates a distant relationship between this family and the ligand-gated ion channel family (TC 1.A.10). These proteins are highly permeable to monovalent cations and show differential permeability for calcium ions. The Swiss-Prot database contains a InterPro link to the IRR001622, which points to the potassium channel pore validating the prediction made by TMLOOP.

Probable sodium:proton antiporter 3

This protein belongs to the sodium : proton exchanger family and the Swiss-Prot database links this protein to the low-affinity sodium (potassium, lithium and caesium) : proton antiporter. TMLOOP predicted a helix-in-turn-loop-out potassium channel like loop in the region corresponding to the predicted TM9. Saier and colleagues (Saier and colleages, Division of Biolgical Science at UC San Diego, http://www.tcdb.org/) suggest that TM4 and TM9 contain essential residues for the binding of both drugs and cations. BLASTP analyses also supported the prediction made by TMLOOP as it was found that potassium : proton antiporters were distantly related to this protein.

Translocation associated membrane protein

This protein is required for the translocation of secretory proteins across the membrane of the endoplasmic reticulum. TMLOOP predicted a helix-in-turn-loop-out potassium channel like loop in the region corresponding to predicted TM7. BLASTP analyses showed distant relationships with a sodium and chloride dependent neurotransmitter transporter and a probable potassium transporter, therefore it might be possible for this protein to contain a membrane dipping loop similar to that found in potassium channels.

Ferrous iron transport protein B

This protein catalyzes transport of ferrous iron through the membrane using energy obtained from ATP hydrolysis. TMLOOP predicted a helix-in-turn-loop-out potassium channel like loop in the region corresponding to TM5. Although this protein type topology resembles that of an α-helical membrane protein containing 13 predicted transmembrane regions, its closest crystallized membrane protein relative is the iron (III) dicitrate transport protein fecA, which belongs to the porin structural type (PDB accession code: 1KMO) where a β-barrel structure crosses the lipidic bilayer acting as a pore. As a general mechanism for the selectivity of porin proteins, a structural motif (either composed of α-helices or β-sheet) is usually placed within the β-barrel pore. A multiple alignment was performed using the sequence belonging to the predicted membrane dipping loop and the whole sequence belonging to the iron(III) dicitrate

transport protein fecA. The alignment showed that the predicted dipping loop motif aligned with a part of the motif located within the protein pore **(figure 5.10)**. Therefore, it could be possible that the region predicted by TMLOOP is involved in ligand binding and considering that membrane dipping loops act as selectivity filters it is also eminently possible that the ferrous iron transport protein B contains a membrane dipping loop in the predicted location.

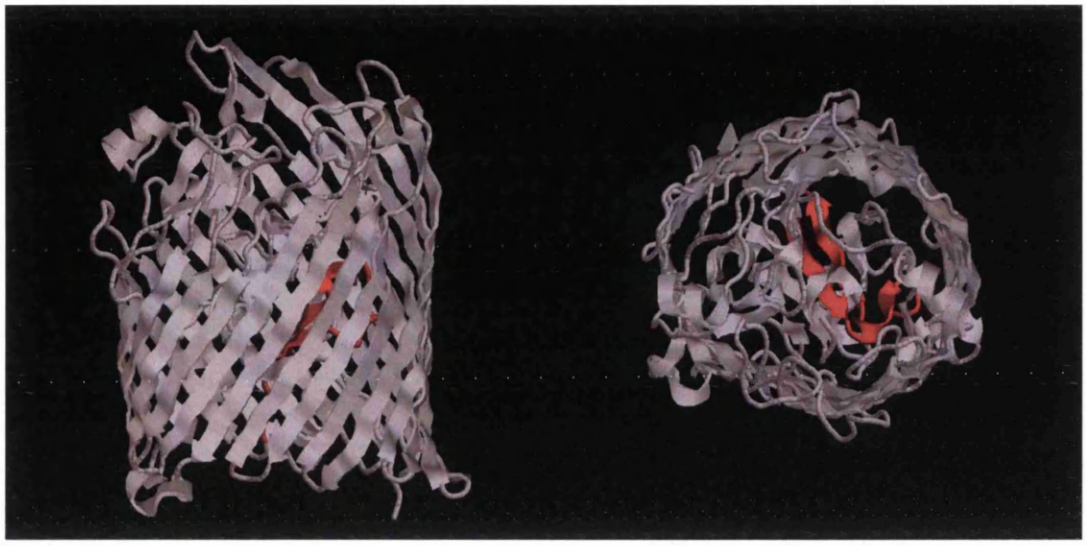

Figure 5.10. Structure of the iron (III) dicitrate transport protein fecA. The region highlited in red corresponds to the region aligned to the predicted membrane dipping loop in the ferrous iron transport protein B.

Immunoglobulin A1 protease precursor

According to functional annotation in the Swiss-Prot database, this protein is a zinc metalloprotease that cleaves immunoglobulin A1. TMLOOP predicted a helix-in-turn-loop-out potassium channel like loop in a region (average pattern starting point: 1510) far from the predicted transmembrane regions according to the Swiss-Prot database (106-154). However, the topology prediction software IRENE (Bologna Biocomputing Group, http://www.biocomp.unibo.it/) and TMPRED (Hofmann and Stoffel, 1993) classified the region corresponding to the predicted membrane dipping loop as a transmembrane region. BLASTP analyses showed distant relationships with a sodium/potassium/calcium exchanger 1 therefore it might be possible that this protein type uses a membrane dipping loop similar to that found in potassium channels to act as a selectivity filter for zinc or other cations.

Hypothetical protein C1F7.03 in chromosome I

According to the Swiss-Prot database this protein is a polycystic kidney disease-related ion channel 2. TMLOOP predicted a helix-in-turn-loop-out potassium channel like loop in the region corresponding to the transmembrane region 4. The GO (gene ontology) database (Ashburner et al., 2000) contained three annotation matches to calcium channels (0005262, 0006816, 0019722), likewise BLASTP analyses showed distant relationships to other cation transporters. As in previous cases, it could be possible that this protein developed a structural motif similar to the membrane dipping loop found in potassium channels.

Mucolipin-2

This proteins belongs to the polycystin subfamily and the voltage gated ion channel superfamily according to the Swiss-Prot database. TMLOOP predicted a helix-in-turn-loop-out potassium channel like loop, which could be possible as both the InterPro database (IPR002111 and IPR005821) and BLASTP analyses showed relationships with cation channels.

Voltage-dependent calcium channel alpha-1 subunit

According to TMLOOP, 19 members belonging to this protein type contain a helix-in-turn-loop-out potassium channel like loop in the region corresponding to a predicted transmembrane region in the Swiss-Prot database. BLASTP analyses showed distant relationships to other cationic channels (calcium and sodium). Likewise the skeletal muscle calcium channel alpha-1 subunit (Swiss-Prot accession code P22316) showed a link to the InterPro database potassium channel group (IPR003091). These facts indicate a plausible membrane dipping loop acting as a selectivity filter in the voltage-dependent calcium channel alpha-1 subunit.

Hypothetical protein AF2214

This is the only protein not predicted by the collective motif method, but the single motif method. This protein was predicted by TMLOOP as a protein containing a

helix-in-turn-helix-out loop 4 ClC chloride channel like loop, which has been proposed to link the repeated halves of ClC chloride channel within each monomer and make contacts with each other at the interface between monomers (Estevez and Jentsch, 2002). BLASTP analysis showed a distant relationship with a proton : chloride exchange transporter ClCa. Therefore, it might be possible that this hypothetical membrane protein contains a similar membrane dipping loop.

5.4 Conclusions

TMLOOP has proven to be a useful bioinformatics tool for the prediction of membrane dipping loops. This software was used to explore the disadvantages of the single motif method and a variation of this method, named the collective motif method, was designed to avoid the inherent limitation of the single motif method to detect distantly related structural motifs. Evaluation of the TMLOOP software using both the single motif method and the collective motif method showed that the collective motif method was a more flexible approach where structural motifs containing small variations of the discovered pattern with the highest support can be co-detected. Further evaluation showed that TMLOOP predicted membrane dipping loops with high sensitivity and specificity, and when applied to the Swiss-Prot database using a set of default parameters the collective motif method was found to maximize the sensitivity and specificity of the predictive tool. TMLOOP predicted 76 possible membrane dipping loops not detected previously by other methods leading to the potential for further experimental research on these possible membrane dipping loops.

5.5 References

ASHBURNER, M., BALL, C. A., BLAKE, J. A., BOTSTEIN, D., BUTLER, H., CHERRY, J. M., DAVIS, A. P., DOLINSKI, K., DWIGHT, S. S., EPPIG, J. T., HARRIS, M. A., HILL, D. P., ISSEL-TARVER, L., KASARSKIS, A., LEWIS, S., MATESE, J. C., RICHARDSON, J. E., RINGWALD, M., RUBIN, G. M. & SHERLOCK, G. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet,* 25, 25-9.

BELOGUROV, G. A., TURKINA, M. V., PENTTINEN, A., HUOPALAHTI, S., BAYKOV, A. A. & LAHTI, R. (2002) H+-pyrophosphatase of Rhodospirillum rubrum. High yield expression in Escherichia coli and identification of the Cys residues responsible for inactivation my mersalyl. *J Biol Chem,* 277, 22209-14.

BILL, R. M., HEDFALK, K., KARLGREN, S., MULLINS, J. G., RYDSTROM, J. & HOHMANN, S. (2001) Analysis of the pore of the unusual major intrinsic protein channel, yeast Fps1p. *J Biol Chem,* 276, 36543-9.

DARZENTAS, N., RIGOUTSOS, I. & OUZOUNIS, C. A. (2005) Sensitive detection of sequence similarity using combinatorial pattern discovery: a challenging study of two distantly related protein families. *Proteins,* 61, 926-37.

DOYLE, D. A., MORAIS CABRAL, J., PFUETZNER, R. A., KUO, A., GULBIS, J. M., COHEN, S. L., CHAIT, B. T. & MACKINNON, R. (1998) The structure of the potassium channel: molecular basis of K+ conduction and selectivity. *Science,* 280, 69-77.

DUTZLER, R., CAMPBELL, E. B., CADENE, M., CHAIT, B. T. & MACKINNON, R. (2002) X-ray structure of a ClC chloride channel at 3.0 A reveals the molecular basis of anion selectivity. *Nature,* 415, 287-94.

DUTZLER, R., CAMPBELL, E. B. & MACKINNON, R. (2003) Gating the selectivity filter in ClC chloride channels. *Science,* 300, 108-12.

ESTEVEZ, R. & JENTSCH, T. J. (2002) CLC chloride channels: correlating structure with function. *Curr Opin Struct Biol,* 12, 531-9.

FALQUET, L., PAGNI, M., BUCHER, P., HULO, N., SIGRIST, C. J., HOFMANN, K. & BAIROCH, A. (2002) The PROSITE database, its status in 2002. *Nucleic Acids Res,* 30, 235-8.

GONEN, T., SLIZ, P., KISTLER, J., CHENG, Y. & WALZ, T. (2004) Aquaporin-0 membrane junctions reveal the structure of a closed water pore. *Nature,* 429, 193-7.

HARRIES, W. E., AKHAVAN, D., MIERCKE, L. J., KHADEMI, S. & STROUD, R. M. (2004) The channel architecture of aquaporin 0 at a 2.2-A resolution. *Proc Natl Acad Sci U S A,* 101, 14045-50.

HEDFALK, K., BILL, R. M., MULLINS, J. G., KARLGREN, S., FILIPSSON, C., BERGSTROM, J., TAMAS, M. J., RYDSTROM, J. & HOHMANN, S. (2004) A regulatory domain in the C-terminal extension of the yeast glycerol channel Fps1p. *J Biol Chem,* 279, 14954-60.

HOFMANN, K. & STOFFEL, W. (1993) TMbase - A database of membrane spanning protein segments. *Biol Chem,* 374, 199.

JIANG, Y., LEE, A., CHEN, J., CADENE, M., CHAIT, B. T. & MACKINNON, R. (2002) Crystal structure and mechanism of a calcium-gated potassium channel. *Nature,* 417, 515-22.

JIANG, Y., LEE, A., CHEN, J., RUTA, V., CADENE, M., CHAIT, B. T. & MACKINNON, R. (2003) X-ray structure of a voltage-dependent K+ channel.

Nature, 423, 33-41.

KARLGREN, S., FILIPSSON, C., MULLINS, J. G., BILL, R. M., TAMAS, M. J. & HOHMANN, S. (2004) Identification of residues controlling transport through the yeast aquaglyceroporin Fps1 using a genetic screen. *Eur J Biochem,* 271, 771-9.

KUO, A., GULBIS, J. M., ANTCLIFF, J. F., RAHMAN, T., LOWE, E. D., ZIMMER, J., CUTHBERTSON, J., ASHCROFT, F. M., EZAKI, T. & DOYLE, D. A. (2003) Crystal structure of the potassium channel KirBac1.1 in the closed state. *Science,* 300, 1922-6.

LONG, S. B., CAMPBELL, E. B. & MACKINNON, R. (2005) Crystal structure of a mammalian voltage-dependent Shaker family K+ channel. *Science,* 309, 897-903.

MONNE, M., NILSSON, I., ELOFSSON, A. & VON HEIJNE, G. (1999) Turns in transmembrane helices: determination of the minimal length of a "helical hairpin" and derivation of a fine-grained turn propensity scale. *J Mol Biol,* 293, 807-14.

MULDER, N. J., APWEILER, R., ATTWOOD, T. K., BAIROCH, A., BATEMAN, A., BINNS, D., BRADLEY, P., BORK, P., BUCHER, P., CERUTTI, L., COPLEY, R., COURCELLE, E., DAS, U., DURBIN, R., FLEISCHMANN, W., GOUGH, J., HAFT, D., HARTE, N., HULO, N., KAHN, D., KANAPIN, A., KRESTYANINOVA, M., LONSDALE, D., LOPEZ, R., LETUNIC, I., MADERA, M., MASLEN, J., MCDOWALL, J., MITCHELL, A., NIKOLSKAYA, A. N., ORCHARD, S., PAGNI, M., PONTING, C. P., QUEVILLON, E., SELENGUT, J., SIGRIST, C. J., SILVENTOINEN, V., STUDHOLME, D. J., VAUGHAN, R. & WU, C. H. (2005) InterPro, progress and status in 2005. *Nucleic Acids Res,* 33, D201-5.

MURATA, K., MITSUOKA, K., HIRAI, T., WALZ, T., AGRE, P., HEYMANN, J. B., ENGEL, A. & FUJIYOSHI, Y. (2000) Structural determinants of water permeation through aquaporin-1. *Nature,* 407, 599-605.

NISHIDA, M. & MACKINNON, R. (2002) Structural basis of inward rectification: cytoplasmic pore of the G protein-gated inward rectifier GIRK1 at 1.8 A resolution. *Cell,* 111, 957-65.

REN, G., CHENG, A., REDDY, V., MELNYK, P. & MITRA, A. K. (2000) Three-dimensional fold of the human AQP1 water channel determined at 4 A resolution by electron crystallography of two-dimensional crystals embedded in ice. *J Mol Biol,* 301, 369-87.

REN, G., REDDY, V. S., CHENG, A., MELNYK, P. & MITRA, A. K. (2001) Visualization of a water-selective pore by electron crystallography in vitreous ice. *Proc Natl Acad Sci U S A,* 98, 1398-403.

RIGOUTSOS, I. & FLORATOS, A. (1998) Combinatorial pattern discovery in biological sequences: The TEIRESIAS algorithm. *Bioinformatics,* 14, 55-67.

SAVAGE, D. F., EGEA, P. F., ROBLES-COLMENARES, Y., O'CONNELL, J. D., 3RD & STROUD, R. M. (2003) Architecture and selectivity in aquaporins: 2.5 a X-ray structure of aquaporin Z. *PLoS Biol,* 1, E72.

SCORDIS, P., FLOWER, D. R. & ATTWOOD, T. K. (1999) FingerPRINTScan: intelligent searching of the PRINTS motif database. *Bioinformatics,* 15, 799-806.

STROUD, R. M., MIERCKE, L. J., O'CONNELL, J., KHADEMI, S., LEE, J. K., REMIS, J., HARRIES, W., ROBLES, Y. & AKHAVAN, D. (2003) Glycerol facilitator GlpF and the associated aquaporin family of channels. *Curr Opin*

Struct Biol, 13, 424-31.
SUI, H., HAN, B. G., LEE, J. K., WALIAN, P. & JAP, B. K. (2001) Structural basis of water-specific transport through the AQP1 water channel. *Nature,* 414, 872-8.
TAMAS, M. J., KARLGREN, S., BILL, R. M., HEDFALK, K., ALLEGRI, L., FERREIRA, M., THEVELEIN, J. M., RYDSTROM, J., MULLINS, J. G. & HOHMANN, S. (2003) A short regulatory domain restricts glycerol transport through yeast Fps1p. *J Biol Chem,* 278, 6337-45.
VAN DEN BERG, B., CLEMONS, W. M., JR., COLLINSON, I., MODIS, Y., HARTMANN, E., HARRISON, S. C. & RAPOPORT, T. A. (2004) X-ray structure of a protein-conducting channel. *Nature,* 427, 36-44.
ZHOU, Y., MORAIS-CABRAL, J. H., KAUFMAN, A. & MACKINNON, R. (2001) Chemistry of ion coordination and hydration revealed by a K+ channel-Fab complex at 2.0 A resolution. *Nature,* 414, 43-8.

CHAPTER 6

TMDEPTH, combining sequence and topological information to extract features of polytopic membrane proteins

6.1 Introduction

6.1.1 The folding process of α-helical membrane proteins

The lipid environment surrounding integral membrane proteins is characterized by particular physicochemical properties, which constrain the spatial arrangement of membrane proteins and restrict their structural diversity. The basic structure of polytopic membrane proteins is characterized by a bundle of hydrophobic α-helices that traverse the membrane, from one extramembraneous side to the opposite side. Polytopic membrane proteins are directly involved in the majority of the functions carried out in the membrane yet these proteins remain broadly similar in terms of their basic structure. Therefore, extensive research has been carried out to elucidate the principles governing the folding process of α-helical membrane proteins and how a constrained arrangement of hydrophobic α-helices support the wide range of biochemical activities carried out by helical membrane proteins.

The folding process of α-helical membrane proteins has been conceptualized as a two stage process (Popot and Engelman, 1990) where first hydrophobic polypeptide segments form independently stable transmembrane α-helices across the membrane and second, the helical segments assemble laterally to form the native structure of the protein. Different factors have been found to promote the association of helices within the lipid bilayer: i) prosthetic groups located in the membrane (e.g. retinal, chlorophyll, heme groups and carotenoids) (Moriki et al., 2001, Schlessinger, 2002), ii) differential effects of lipids, such us membrane lipid composition, lipid packing or interactions between specific lipid

groups and the membrane protein (Popot and Engelman, 2000), iii) the length and folding of extramembraneous loops, which might constrain the location of contiguous helices and their orientation (Allen et al., 2001, Kim et al., 2001), iv) helix packing residues and motifs, which promote interhelical contacts (Adamian and Liang, 2001, Choma et al., 2000, Dawson et al., 2003, Gratkowski et al., 2001, Langosch et al., 1996, Zhou et al., 2000, Zhou et al., 2001, Dawson et al., 2002, Russ and Engelman, 2000, Senes et al., 2000, Adamian and Liang, 2002, Liu et al., 2002, Eilers et al., 2002, Adamian et al., 2003, Walters and DeGrado, 2006, Lemmon et al., 1992) and v) steric clashes at helix-helix interfaces and restrictions of side-chain rotamers can constrain the space of interacting helices (Popot and Engelman, 2000).

Transmembrane helices have been exhaustively studied in order to characterize these segments and create bioinformatics tools that could accurately predict the transmembrane regions of a polytopic membrane protein and interacting helices in the native state. Compositional analyses have shown that α-helices located in the membrane are mainly hydrophobic. The hydrophobic residues Ala, Ile, Leu and Val constitute 34% to 50% of all residues in α-helical membrane proteins (Ulmschneider and Sansom, 2001, Senes et al., 2000) whereas strong polar residues (Gln, Asn, His, Asp, Glu, Arg and Lys) are poorly represented (Eilers et al., 2000, Tourasse and Li, 2000). The polar residues Ser and Thr are the only exception as they have been found to account for approximately 7% of the residues (Curran and Engelman, 2003). Although Ser and Thr can not promote helix packing by themselves, motifs of four to five Ser and Thr residues can drive strong helix interactions by side-chain hydrogen bonds that function cooperatively to create a strong helix-helix association (Dawson et al., 2002). Aromatic residues have been shown to have a high propensity to face the lipid head-groups near the boundary between the hydrophilic and hydrophobic regions of the membrane forming the so-called "aromatic belt" (Pilpel et al., 1999, Monne et al., 1999). Gly is frequently found in the transmembrane regions, the interhelical associations of this residue are probably the best characterized as glycine has been shown to be an essential residue in promoting helix-helix packing. Mutagenesis analyses of the glycophorin A (GpA) and the ToxR proteins showed a right-handed motif composed of seven residues, LIxxGVxxGVxxT, which appeared as a dimerization motif

(Langosch et al., 1996, Lemmon et al., 1992). Further mutagenesis analyses in the major coat protein (MCP) from bacteriophage (Deber et al., 1993) featured a GxxxG dimerization motif, which was shown to mediate strong helix association for several different TM domains (Russ and Engelman, 2000) and was shown to be the most occurring helix motif found (Senes et al., 2000). Interestingly, this motif has also been associated with β-branched residues located at i+1 positions relative to the Gly and it was proposed that the groove and the ridge created by the Gly and the β-branched residues might facilitate the extensive contacts between helices in a "ridge into grooves" fashion (Senes et al., 2000, MacKenzie et al., 1997). The Gly residues found in GxxxG motif, also called GG4, can also be replaced by small residues such as Ala and Ser ([small]xxx[small] motif). These small residues were proposed to reduce the steric hindrance of helical backbones, which might facilitate the formation of hydrogen bonds (Senes et al., 2001). Interhelical hydrogen bonds have been proposed to be a major factor during the assembly of membrane proteins, the low dielectric constant of the membrane causes an increase in the density of the donor and acceptor atoms, which in turns increases coulombic interaction between the partial effective charges on the donor and the acceptor atoms (Shan and Herschlag, 1996). Therefore, hydrogen bonds in non-aqueous environments enhance the stability of interactions between helices and minimize the energy penalty imposed when polar and ionizable atoms are located in the membrane. Further research showed that nearly every helix in the membrane is connected by at least one hydrogen bond to its closest neighboring helix and that these pair of helices were packed tighter than those pairs of helices not connected by hydrogen bonds (Adamian and Liang, 2002).

Research (Agre and Kozono, 2003, Eilers et al., 2002, Adamian and Liang, 2001, Walters and DeGrado, 2006) has also been focused on the discovery of spatial interhelical motifs by analyzing the spatial coordinates of the crystallized membrane proteins deposited in the Protein Data Bank (Berman et al., 2000). Adamian and co-workers (Adamian and Liang, 2001) analyzed the propensity of residues for being located in voids and pockets, which might allocate ligands or prosthetic groups such as heme or water molecules or allow conformational changes under mechanical forces. Phe, Trp and His residues were found to have the highest propensity to be in such spatial domains whereas small residues (Ser, Gly,

Ala, and Thr) had the lowest propensity. The propensities of single residues and of pairs of residues for interhelical contacts was also computed and it was found that soluble and membrane proteins contained different propensities, which might reflect the constraints of the different environments surrounding soluble and membrane proteins and the different folding mechanisms used. Disulphide bonds are important for maintaining the stability of soluble proteins but are not common in membrane proteins. There is a lack of correlation between soluble and membrane proteins in the propensity scale of residues involved in interhelical contacts. Helices belonging to membrane and soluble proteins do not share common pairs of residues with a high degree of helical interfacial pairwise propensity. The atomic polar interactions found in membrane proteins are more diverse in membrane proteins than in soluble proteins, which include pairs between charged residues (salt bridges), polar residues and between charged and polar residues whereas in soluble proteins the only interhelical atomic contacts of polar atoms were found in pairs of charged residues (salt bridges). Helical backbone-backbone interactions were also found to be more common in membrane proteins than in soluble proteins. This conclusion was also supported by more recent research (Eilers et al., 2002), which postulated that the higher divergence of residues found in the helical interfaces might reflect the relationship between structure and function in membrane proteins where the functional sites are often found in the interior of the molecule, whereas in soluble proteins the functional sites are mostly found on the protein surface, either in clefts or grooves. This research also highlighted the high propensity of small residues for packing in helix interfaces and proposed that two general spatial motifs mediated the association between helices in the membrane. The first motif described as "knobs-into-holes", first described in soluble coiled coils (Langosch and Heringa, 1998), is a general motif used by both soluble and membrane proteins, which mainly relies on four residues (Leu, Ala, Ile and Val). This motif is exemplified by the heptad motif known as the leucine zipper LxxLxxxLxx. The second type of motif includes small and polar residues, which allow the backbones of interacting helices to closely approach each other, the GG4 motif explained earlier exemplifies this type of loop. However, further motifs have also been described: i) the "serine zipper", typical of cytochrome c oxidases, involves two hydrogen bonds between the side chain of a serine in one helix and the carbonyl oxygen or an amide hydrogen of the polypeptide backbone corresponding to another serine from a

different helix and ii) the "polar clamp", typical of GPCRs, is composed of three residues located in two interacting helices where the side chain of a residue is "clamped" by two hydrogen bonds with either two side chains, a side chain and a main chain oxygen or nitrogen, or two main chain oxygen or nitrogen atoms of residues located at positions i , i+4 in the opposite helix (Adamian and Liang, 2002). Further work was focused on the analysis of higher-order interhelical spatial interactions involving three residues and two transmembrane helices (Adamian et al., 2003), several triplets were found specifically in membrane proteins. Approximately one third of these triplets could potentially involve hydrogen bonds, which supports the importance of this type of interaction in helix packing. The triplets showed a preference for Gly and Ala residues and unexpectedly Met was also frequently found in these spatial motifs. Likewise, some of the motifs over-represented in membrane proteins such as the GG4 motif showed strong correlation with triplets. Recent work (Walters and DeGrado, 2006) analyzed a library of pairs of interacting transmembrane helices and clustered the helical pairs according to their three-dimensional similarity. Interestingly, five different structural clusters accounted for three quarters of the clustering space. The top four clusters were defined as "antiparallel GAS $_{left}$", "antiparallel GAS $_{right}$", "parallel GAS $_{left}$" and "antiparallel GAS $_{right}$" (where the first term refers to the orientation of one helix relative to the remaining helix, the second term refers to the residues involved and the third term refers to the crossing angle between helices). These motifs are in accordance with statements made in previous research, which highlighted the high propensity of small residues for being in the helix-helix interfaces and the importance of these residues to promote tight helix packing. As with Adamian and co-workers, motifs previously discovered were found to be part of the motifs clustered (e.g the GG4 motif was found to be part of the "parallel GAS $_{right}$" motif).

6.1.2 Topology prediction methods for α-helical membrane proteins

The different research carried out to understand the folding and packing mechanism of membrane proteins has the final aim of accurately predicting the transmembrane regions in membrane proteins and elucidate the helical relationships within the membrane. While the three dimensional structural prediction of membrane proteins still needs further

developments, the accuracy of predicting the transmembrane region is increasing with the development of new and more sophisticated predictive tools. Since the 1980's, the number of computational methods used to predict transmembrane regions has been gradually increasing and more than 35 different methods have been reported. In general terms, these methods could be classified as methods based on i) propensities such as hydrophobicity scales, ii) multiple sequence alignments and iii) consensus methods, however classification of particular methods into these three categories is not always straightforward. The first predictive tools were based on hydrophobic scales (Claros and von Heijne, 1994, Degli Esposti et al., 1990, Kyte and Doolittle, 1982, Ponnuswamy and Gromiha, 1993, von Heijne, 1992), simultaneously other methods were developed based on statistical analyses (Klein et al., 1985) and parameters and propensities such as positional propensities, physicochemical parameters (Efremov and Vergoten, 1996, Gromiha, 1999, Stoffel et al., 1993, Cserzo et al., 1997) or a combination of hydrophobicity and other propensities (Hirokawa et al., 1998, Lohmann et al., 1996). Other methods were based on multiple sequence alignments to predict transmembrane regions (Jones et al., 1994, Persson and Argos, 1994, Rost et al., 1996, Rost et al., 1995). Empirically designed rules (von Heijne and Gavel, 1988), compositional analyses (Fariselli and Casadio, 1996, Persson and Argos, 1997, Persson and Argos, 1996) were used to refine the prediction of transmembrane regions by predicting also the orientation of integral membrane proteins. Data mining techniques such as neural networks (Aloy et al., 1997, Casadio et al., 1996, Lohmann et al., 1996, Rost et al., 1995) and hidden Markov models (Kahsay et al., 2005, Kall et al., 2004, Krogh et al., 2001, Sonnhammer et al., 1998, Tusnady and Simon, 2001, Tusnady and Simon, 1998, Viklund and Elofsson, 2004, Xu et al., 2006, Zhou and Zhou, 2003, Martelli et al., 2003) have been successfully applied for the prediction of transmembrane regions. Evaluations of the predictive tools described above have shown that accuracy of prediction of the topology of helical membrane protein ranks between 60 to 70% (Chen et al., 2002, Ikeda et al., 2002, Kall and Sonnhammer, 2002, Moller et al., 2001) however no recent evaluations have been carried out to test the predictive performance of the latest algorithms. All the evaluations performed have described the better performance of predictive methods based on hidden markov models and TMHMM (Sonnhammer et al., 1998, Krogh et al., 2001) has been described as the best performing algorithm in two out of four evaluations

(Moller et al., 2001, Kall and Sonnhammer, 2002). Ikeda and colleagues (Ikeda et al., 2002) compared each single method evaluated against a consensus prediction method and showed that by performing consensus prediction the accuracy of the prediction could be increased by up to nine percentage points. Consensus prediction methods have become more popular in recent years and since the first consensus method was developed for the prediction of helical membrane proteins in 1994 (Parodi et al., 1994) and increasing number of tools have been developed with the same purpose (Amico et al., 2006, Arai et al., 2004a, Nikiforovich, 1998, Nilsson et al., 2000, Nilsson et al., 2002, Promponas et al., 1999, Taylor et al., 2003, Xia et al., 2004, Ikeda et al., 2002).

6.1.3 The transmembrane domain-to-function approach

The recent developments in topology predictions have opened a new protein space to be analyzed. Topology and function have been proposed to be directly related, Liu and colleagues (Liu et al., 2002) found that among transporters and channels a 12 transmembrane helix bundle is preferred and the 7TM receptor superfamily probably represents the best example of the relationship between topology and function. Shimizu and coworkers have used the topological space of membrane proteins to extract binary patterns and classify and identify function, reporting promising results (Inoue et al., 2004, Sugiyama et al., 2003). However, these results are based upon classification of the training set and no evaluation of their method (i.e jack-knife test or x-fold cross-validation) has been reported yet. Further research carried out by Shimizu and coworkers showed that clustering the topological space of membrane proteins could improve the annotation of uncharacterized membrane proteins compared to conventional clustering based on sequence similarity (Arai et al., 2004b). However, the clustering method was not evaluated using experimentally characterized membrane proteins. Despite the direct relationship between topology and membrane protein function, prediction of specific membrane protein functions by considering only topological information is a difficult task if the protein sequence is to be ignored. For instance, transporters are believed to require a minimum of approximately 6 transmembrane regions to form a pore through which molecules and ions can go through and as mentioned above the 12 transmembrane helix bundle is preferred by this

superfamily. Based on topological information it might be possible to predict these transporters but describing the type of transporter or the corresponding ligand is beyond the scope of methods based solely on topological information. It is believed that by combining sequence and topological information the resulting predictive tools should be more reliable and informative.

Transmembrane regions in membrane proteins often contain functionally important sites such as ligand binding sites (e.g. potassium channels and aquaglyceroporins) and sites promoting conformational changes (e.g. proline kinks in GPCRs). These motifs are often composed of various residues located in different transmembrane regions and the detection of these motifs in sequence has proven to be extremely difficult. Previous work showed that functional clusters of crystallized membrane proteins showed specific patterns of interhelical associations of residues located at a similar depth (Lasso, Honours Thesis, 2001). Motifs such as those discovered in membrane dipping loops showed residues involved in ligand binding sites and molecular gates whose interactions with other residues at a similar depth that have been found experimentally to be critical for the correct functionality of the membrane proteins **(Chapter 4)**. The crystallized structure of the calcium pump of the sarcoplasmic reticulum showed two calcium binding sites, the first calcium binding site **(figure 6.1, cluster in red)** was composed of Asn768, Glu771, Thr799, Asp800 and Glu908 and the second binding site **(figure 6.1, cluster in green)** was composed of Val304,Ala305, Ile307, Glu309, Asn796 and Asp800 (Toyoshima et al., 2000). As shown in **figure 6.1** residues forming both binding sites were clustered at a similar depth in the membrane and important associations were found between those residues involved in the binding site (e.g. Asn796-Ala305).

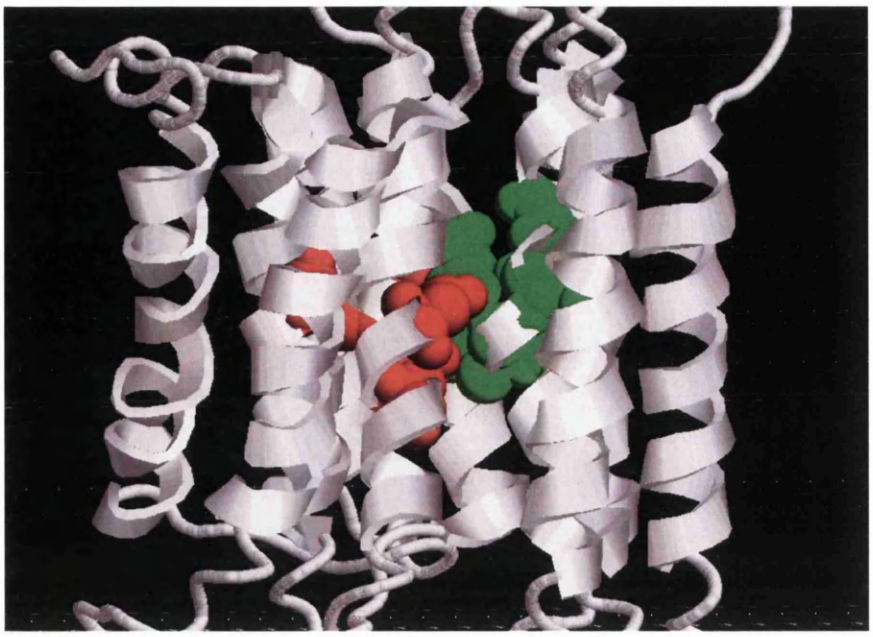

Figure 6.1 Structure of the calcium pump of the sarcoplasmic reticulum. Amino acids involved in the calcium binding site were highlighted in red for the binding site I and in green for the binding set II.

Likewise, these motifs not only can guide the molecular function of a membrane protein but also determine its subcellular location. Subcellular location of proteins can be directed by two different mechanisms: i) targeting signals, which are linear motifs that target a specific organelle and ii) signal patches, which are three-dimensional motifs composed of residues located in different positions in the amino acid sequence but interacting spatially. While different targeting signals have been reported signal patches still remain to be described.

These motifs, if found to be in the transmembrane space of the membrane protein, must be composed of residues located at a similar depth in the membrane. Following this principle, it is possible to approximate pairs of residues potentially interacting (if such a pair is composed of residues located at a similar depth) based on the outcome of topology predition methods and the amino acid sequence.

6.1.4 The TMDEPTH approach

TMDistance (Togawa, PhD Thesis, 2006) was implemented to compute the number of interhelical associations between residues located in different transmembrane regions in crystallized α-helical membrane proteins. This algorithm based its calculation upon the inter-atomic distances between residues and a certain threshold specified by the user. Comparison of interhelical associations between functional clusters of crystallized structures showed that these clusters contained specific patterns of interhelical associations, which involved pairs of residues located at a similar depth (Lasso, Honours Thesis, 2001). Based on these results, a new method was implemented to detect pairs of residues located at a similar depth in the membrane based solely on the amino acid sequence and the topology of the α-helical membrane protein. TMDEPTH extracts topological information contained in the SwissProt database and predicts the orientation of the N-terminus based on a variation of the positive-inside rule (von Heijne and Gavel, 1988). The membrane thickness is approximated based on the shortest helical structure, which transverses the membrane and each residue in the membrane is given a depth value based on the estimated membrane thickness, the length of the corresponding transmembrane helix and its position in the given segment. According to this method, two residues are associated (which does not imply a physical interaction) when their corresponding depth values are within a certain range (1.5 Å). Following this assumption, the depth space is then searched and pairs of residues with similar depth values and belonging to different transmembrane regions (and pair of residues belonging to a different half of a given membrane dipping loop) are computed **(figure 6.2)**. All computed interhelical associations are stored in a 20x20 matrix which describes all possible interhelical associations (210 associations) within the membrane.

The depth associations summarize the interhelical packing patterns within the transmembrane regions of polytopic α helical membrane proteins. These patterns describe conformational arrangements such as ridge-groove arrangements, spacer motifs, regions of high flexibility and salt bridges, and consequently describe a wide range of biologically important features. Such features can range from structural domains to sites involved in

triggering conformational changes, lipid binding, prosthetic group binding, ligand binding or active sites that catalyze a biochemical reaction.

Matrices obtained with TMDEPTH lead to novel research in the membrane protein field where sequence and topological information are combined. These matrices were mined using different data mining methods in order to develop predictive tools capable of predicting the subcellular location **(Chapter 7)** and function of membrane proteins **(Chapter 8)**.

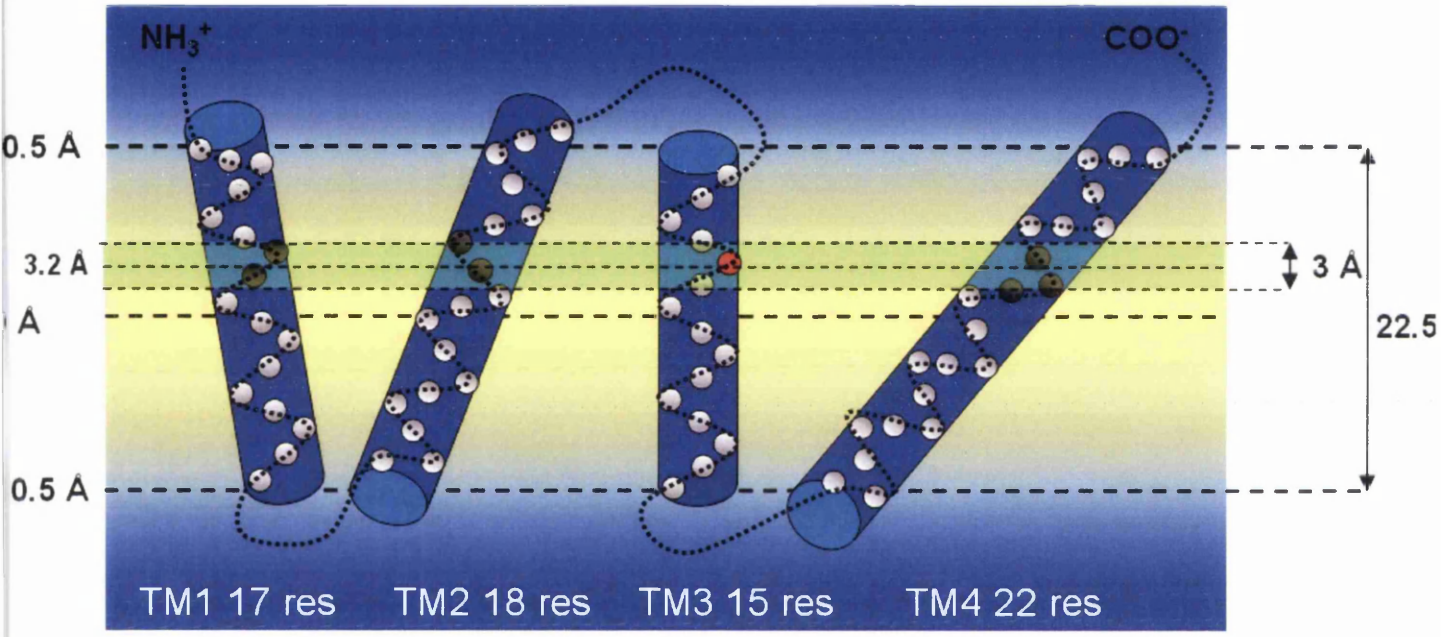

Figure 6.2. Schematic representation of the TMDEPTH algorithm. The figure shows a hypothetical protein composed of four transmembrane regions. In order to estimate the membrane thickness the shortest helix with length ≥ 14 residues is considered (in this example, TM3). This transmembrane region is considered to traverse the membrane perpendicular to the lipid face. The membrane thickness corresponds to the number of residues contained in the shortest helix (15 residues in TM3) multiplied by the intrahelical distance along the helical axis between contiguous residues (1.5Å). Longer transmembrane helices (TM1, TM2 and TM4) do not traverse the membrane perpendicular to the lipid faces but are tilted in order to place the longer hydrophobic region in the lipid bilayer. Therefore, the longer the helix the more acute the angle between the membrane normal axis and the helical axis of the transmembrane region. TM4 is the longest transmembrane region and needs to form a more acute angle with the membrane normal than TM1 and TM2 in order to accommodate the 22 residues contained. Based on the computed model each residue in the membrane is given a depth value which is estimated based on the membrane thickness, the length of the corresponding transmembrane region and the position of the corresponding residue in the helix. TMDEPTH computes the interhelical associations of pairs of residues located at a similar depth in the membrane. This figure illustrates the computed associations for the fifth residue located in TM3 (residue highlighted in red). This residue has a depth value of 3.2 Å, TMDEPTH sets an upper and a lower range (± 1.5Å, in green) and all residues located in other transmembrane regions (TM1, TM2 and TM4) whose depth values are within the depth range (1.7Å-4.7Å) (residues highlighted in black) are assumed to form a potential association with the fifth residue of TM3 (residue highlighted in red). This process is iteratively repeated for all residues in the membrane and all associations are stored in a 20x20 matrix.

6.2 Algorithm development

TMDEPTH was implemented to load local text files corresponding to proteins contained in the Swiss-Prot database and to report the percentage of interhelical associations in a specific format allowing further data mining analyses with different programs.

The algorithm contains seven different steps:

- Extraction of information from the Swiss-Prot like text file
- Orientation prediction
- Calculation of membrane depth
- Calculation of depth for each residue located in the membrane
- Calculation of interhelical associations
- Data standardization and extraction of other biological relevant information
- Report of the extracted features in a specific format required by the user

6.2.1 Extraction of information from the Swiss-Prot like text file.

Once a Swiss-Prot like text file has been loaded into TMDEPTH, the software extracts specific information contained in the file. Some of the extracted information is only used by TMDEPTH to report a general description of the protein analyzed whereas other information is essential for the reporting of potential interhelical associations of transmembrane regions. The protein descriptive information extracted by TMDEPTH is the protein ID, the protein accession number, the protein function description, the organism species and the classification of the organism (contained in the statements under the tags "ID", "AC", "DE", "OS" and "OC" respectively). The information required by TMDEPTH to calculate the interhelical associations are the amino acid sequence (contained in the statement with the tag "SQ") and details describing transmembrane regions, which are found in the feature statements under the tags "FT TRANSMEM" and "FT MEMBLOOP". The first tag corresponds to transmembrane regions consensually predicted

by different transmembrane topology prediction programs used by curators working at the Swiss Institute of Bioinformatics in Geneva, the second tag corresponds to membrane dipping loops **(figure 6.4)** predicted by TMLOOP and included in files by TMLOOP writer **(Chapter 5)**.

Using the principles of object oriented programming and the extracted information the protein is divided into different objects (an object is an instance of a class) based on four different classes **(figure 6.3)**: i) the $NH3^-$-terminus class, ii) the transmembrane class, iii) the loop class and iv) the COO^--terminus class. The $NH3^+$-terminus and the COO^--terminus class can only have one object each but the transmembrane class must have as many objects as transmembrane regions (including predicted membrane dipping loops) described in the Swiss-Prot like text file and the loop class must have as many objects as loops connecting the transmembrane regions (that is one object less than the number of transmembrane objects). The only transmembrane regions not considered by TMDEPTH as such are those with a length lower than 14 residues, which are instead included in the corresponding loop object. All annotated transmembrane regions (unless they are membrane dipping loops) are considered as helical structures that completely traverse the membrane (β–barrel membrane proteins needed to be filtered out at the protein annotation level when constructing the data set). Based on an experimental analysis (Monne et al., 1999) and a preliminary analysis carried out with crystallized membrane proteins **(Chapter 5)** the minimum length of a helix to completely tranverse the membrane was set to 14. Therefore short transmembrane regions were considered as false positives and included in the preceding loop object **(figure 6.3)**. The instances of the classes described above contained their corresponding fragments of the protein sequence, which is necessary to compute the orientation of the protein and the potential interhelical associations in the lipid bilayer.

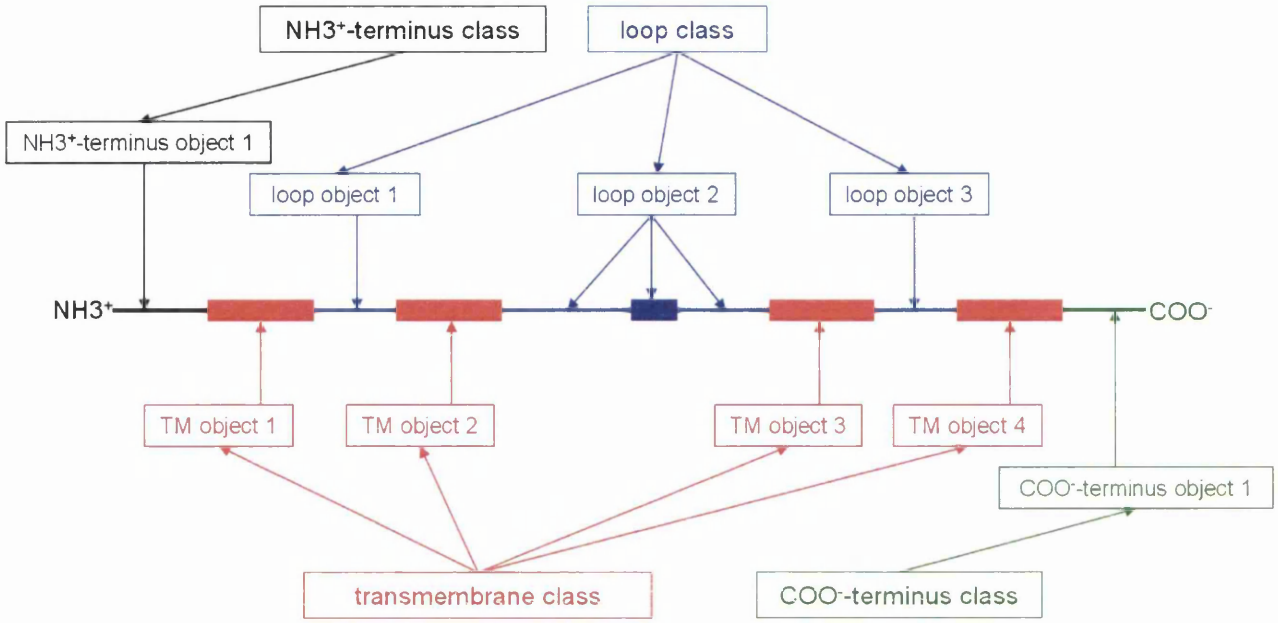

Figure 6.3. Schematic representation of the different structural domains of polytopic membrane proteins and how they are distributed into different classes. The transmembrane regions have been represented as rectangles whereas the loops connecting the transmembrane regions and the terminus regions have been represented as lines connected to the transmembrane regions. Each segment has been coloured based on the class type of the corresponding object. Segments colored in black, red, blue and green belong to the $NH3^+$-terminus, transmembrane, loop or COO^--terminus class respectively. One of the annotated transmembrane regions (blue rectangle) is not considered as such, since its corresponding length is lower than 14 and it is not believed to form a helical structure that completely traverses the membrane.

6.2.2 Orientation prediction

This step is based on the orientation prediction of each extramembraneous domain in respect of the $NH3^+$-terminus ($NH3^+$-terminus side and $NH3^+$-terminus opposite side) and the orientation of the molecule based on the positive-inside rule, which postulates that the extramembraneous domains located in the cytosol of the cell are more positively charged than those on the external side (von Heijne and Gavel, 1988).

In order to orientate each extramembraneous domain with respect to the $NH3^+$-terminus it was necessary to examine the precedent transmembrane region. If the corresponding transmembrane region was annotated as "FT TRANSMEM" and passed the minimum length condition the transmembrane region was assumed to completely traverse the

membrane and the extramembraneous domain to be orientated would be located at the opposite side of the previous extramembraneous domain. If the transmembrane region was annotated as "FT MEMBLOOP" then the transmembrane region was considered as a membrane dipping loop whose C-terminal section returns to the same extramembraneous side as the N-terminal section of the loop **(figure 6.4)**. In this case, the extramembraneous domain to be orientated would be located at the same side as the previous extramembraneous domain. Because the $NH3^+$-terminus is the first extramembraneous domain it can not be orientated and it is used as a reference to orientate the remaining extramembraneous domains including the COO^-terminus.

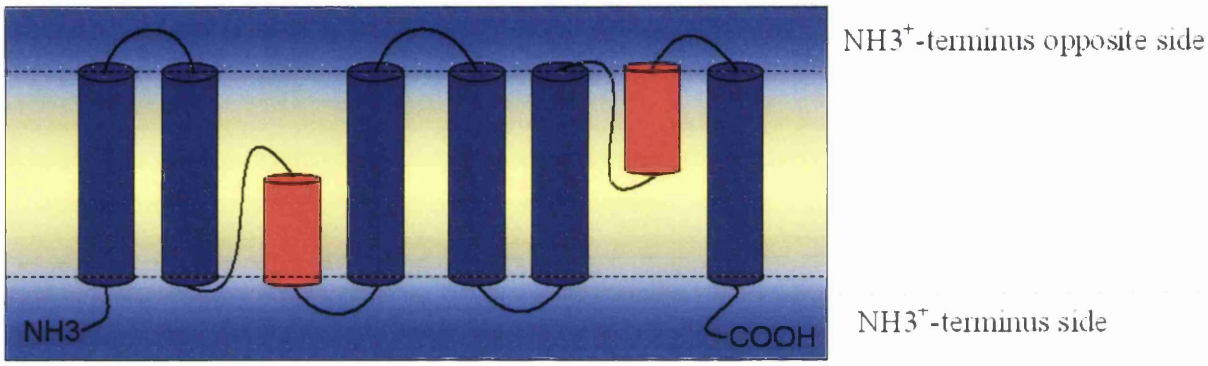

Figure 6.4. Transmembrane regions annotated as "FT TRANSMEM" (in blue) completely span the membrane whereas transmembrane regions annotated as "FT MEMBLOOP" (in red) partially traverse the membrane and return to the same extramembraneous side as the N-terminal half of the structural domain. Following this principle the orientation of each extramembraneous domain with respect to the $NH3^+$-terminus was predicted.

The protein orientation was then computed using the positive-inside rule (von Heijne and Gavel, 1988), which showed that intracellular extramembraneous domains contained a higher ratio of positively charged residues to negatively charged residues than the extracellular extramembraneous domains.

For each extramembraneous side of the membrane proteins ($NH3^+$-terminus side and $NH3^+$-terminus opposite side) the ratio of positively charged residues/negatively charged residues was calculated analyzing the corresponding sequences of each of the loop objects and the $NH3^+$-terminus and the COO^--terminus objects, built in the object oriented programming environment. The extramembraneous side found to have the highest ratio was

considered to be in the cytosol. When the ratios of both extramembraneous sides were identical a similar rule was applied to the number of positively charged residues. Therefore, the extramembraneous side with the highest number of positively charged residues was considered to be located on the intracellular side. If the number of positively charged residues was still found to be the same for both extramembraneous sides by default, TMDEPTH assigned the $NH3^+$-terminus side as the intracellular side.

Although the prediction of the molecular orientation through the positive-inside rule is an interesting task to be executed by TMDEPTH, it is not essential for the calculation of interhelical associations of residues located at a similar depth. Whether the extramembraneous sides are located outside or inside the cell is not relevant as it does not affect the associations of pairs of residues located at a similar depth. On the other hand, a proper orientation of the extramembraneous domains with respect of the $NH3^+$-terminus is important as a single mistake in the orientation of one extramembraneous loops would affect the orientation of all extramembraneous domains following the wrongly orientated loop **(figure 6.5)**. Subsequently, for a pair of residues predicted to be at a similar depth where one residue precedes the wrongly orientated loop and the other residue follows the wrongly orientated loop, the prediction would be a consequence of the extramembraneous domain orientation error and would not reflect the real situation.

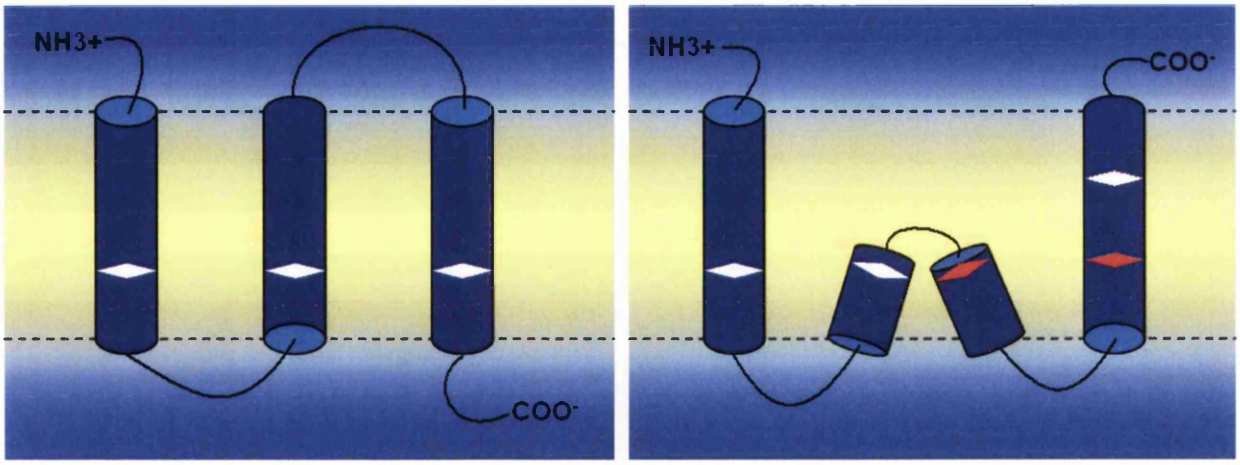

Figure 6.5. Example of an erroneous prediction of interhelical associations of residues located at a similar depth when a transmembrane region is wrongly annotated and subsequently all following loops are wrongly orientated. 6.4a (left) Correct topological model of a hypothetical membrane protein, the white rhombus represents residues from different transmembrane regions located at a similar depth in the membrane. Figure 6.5b (right) Incorrect topological model of a hypothetical membrane protein, due to an erroneous annotation of the second transmembrane region, the second loop is wrongly orientated. As a consequence two residues (red rhombus) were predicted to be at a similar depth whereas the appropriate residue at the corresponding depth belonging to the transmembrane region three was incorrectly located closer to the opposite side of the membrane.

6.2.3 Calculation of membrane thickness

The membrane thickness needs to be calculated in order to give depth values to all residues belonging to transmembrane regions. Transmembrane α-helices do not have similar length, instead the lengths of these helices can vary from approximately 15 to even 40 residues. In order to accommodate these hydrophobic helices in the membrane the longer helices lie at a more acute angle through the membrane, increasing the length of these helices that may be accommodated within the membrane. On the other hand shorter helices do not normally need to form acute angles with the membrane faces and they normally appear perpendicular to the membrane faces. TMDEPTH uses this fact to calculate the membrane thickness by using the shortest transmembrane region that completely spans the membrane (annotated as "FT TRANSMEM") and considering this segment running straight through the membrane. Therefore the membrane thickness can be calculated based on the number of residues found in the shortest transmembrane region spanning the membrane and the intrahelical distance between two contiguous residues (1.5 Å), i.e.

$$T = \left(n^{(h)} - 1\right) \cdot 1.5\text{Å} \tag{6.1}$$

where T is the thickness of the membrane, n corresponds to the number of residues and h corresponds to the shortest helix that completely traverse the membrane.

6.2.4 Calculation of depth for each residue located in the membrane

Once the membrane thickness has been estimated, TMDEPTH calculates the membrane depth of each residue located in a transmembrane region by considering the length of the corresponding α-helix, the estimated membrane thickness and the helical position of the corresponding residue. Although the intrahelical distance between contiguous residues remains 1.5 Å along the helical axis perpendicular to the faces of the membrane, the intrahelical distance between contiguous residues, perpendicular to the faces of the membrane, varies with the helical length. As the length of the helix increases, the angle between the membrane axis perpendicular to the faces of the membrane and the helical axis also increases **(figure 6.6),** which decreases the intrahelical distance, perpendicular to the faces of the membrane, between two contiguous residues, i.e.

$$d = \frac{T}{n^{(h)} - 1} \tag{6.2}$$

where d is the intrahelical distance, parallel to the membrane normal, between contiguous residues, T is the membrane thickness and n(H) is the number of residues of the corresponding helix. According to this formula and equation 6.1, the intrahelical distance d is equal to the intrahelical distance parallel to the helical axis (1.5 Å) for the shortest transmembrane helix used to calculate the membrane thickness.

Residues located closer to the extracellular side receive positive depth values whereas residues located closer to the intracellular side receive negative depth values, the geometrical center to the membrane, equidistant to both extramembraneous sides, is given a depth value of 0 Å, i.e.

$$e^{(r)} = \pm\left(\frac{T}{2} - d \cdot p^{(r)}\right) \qquad (6.3)$$

where $e^{(r)}$ corresponds to the depth value of a residue r, T is the membrane thickness, d is the intrahelical distance parallel to the membrane axis between contiguous residues , $p^{(r)}$ is the position in the helix of the residue r (the first residue in the helix is given the position 0). Positive values are applicable to those helices whose N-terminus is located at the intracellular side of the membrane, and vice versa.

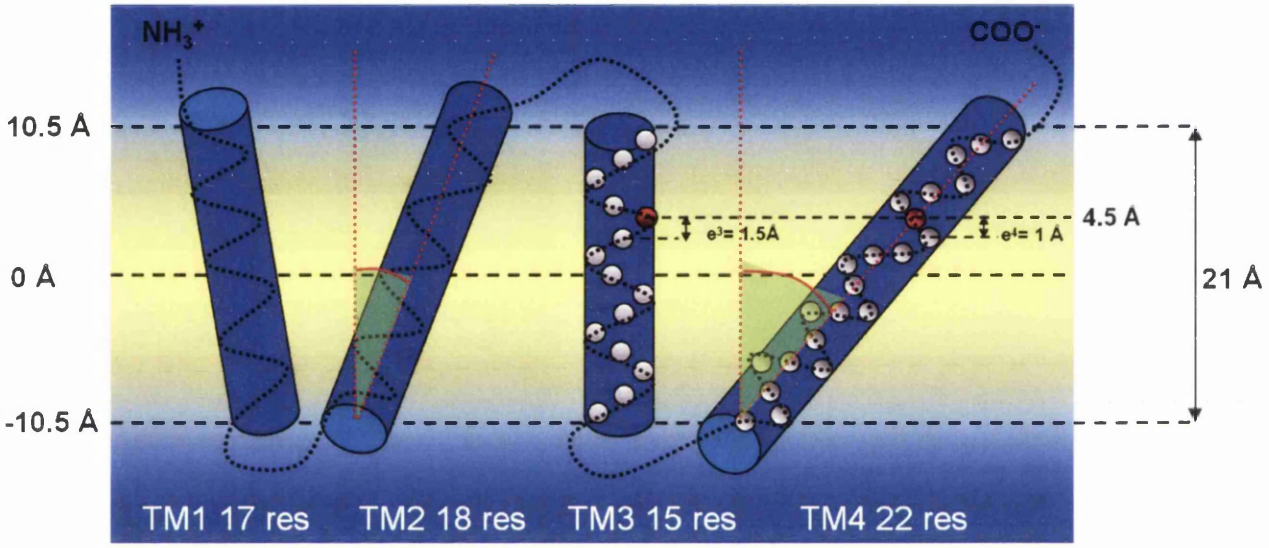

Figure 6.6. Example of membrane thickness calculation and fixing of the α-helices in the membrane. Transmembrane region 3 is the shortest helix and its length is used to calculate the membrane thickness ((15-1) x 1.5Å), this helix is allocated parallel to the membrane normal whereas the remaining helices need to form an angle with the membrane normal to locate the hydrophobic helical structure in the membrane, the longer the helix the higher the degrees of the angle between the membrane normal and the axis of the helix. Longer helices require smaller values of intrahelical distance, parallel to the membrane normal, between contiguous residues (d). As shown in the figure, the 5th residue of TM3 and the 16th residue of TM5 both have the same depth value (residues highlighted in red, e = 4.5Å). The intrahelical distance parallel to the membrane normal d for TM3 (d^3) is 1.5 Å whereas for TM4 (D^4) is 1 Å.

Equation 6.2 and **equation 6.3** are only applicable to those residues belonging to transmembrane regions that completely traverse the membrane (annotated as "FT TRANSMEM"). To calculate the membrane depth of residues belonging to membrane dipping loops (annotated as "FT MEMBLOOP"), a different model was needed. Membrane dipping loops were found to be composed of two different secondary structure types: an α-helix and an unstructured loop. Each membrane dipping loop characterized was

found to have a particular arrangement of α-helices and unstructured loops, which were classified into 3 different structural types, namely helix-in-turn-helix-out, helix-in-turn-loop-out and loop-in-turn-helix-out **(Chapter 4)**. The lengths of the secondary structures were not identical across the different membrane dipping loops classified within each structural category and each membrane dipping loop type needed to have its own particular model to calculate the depth of their corresponding residues. Using the principles of homology base modeling and making use of the crystallized membrane proteins containing membrane dipping loops, a model was constructed for each membrane dipping loop predicted by TMLOOP **(figure 6.7)**.

In order to calculate depth values for residues located in membrane dipping loops, it was necessary to place the corresponding model in a virtual lipidic environment whose thickness was dictated by the smallest helix spanning the membrane. It was necessary to set up some constraints to position membrane dipping loops in the membrane following empirical observations when visualizing membrane dipping loops in crystallized structures: i) How deep the membrane dipping loop projects into the membrane is dictated by the length of the α-helix; however the maximum depth a membrane dipping loop can achieve is the geometrical center of the membrane (0 Å); ii) if the α-helix length is longer than the membrane thickness/2 then the helix would form an angle with the membrane normal (in a similar fashion as longer α-helices that completely transverse the membrane) in order to position the helical structure in the membrane; iii) in the case of helix-in-turn-loop-out and loop-in-turn-helix-out the unstructured loops are of the corresponding length to reach the same depth as the intramembraneous terminal of the α-helix; iv) in helix-in-helix-out membrane dipping loops the maximum depth of the domain is dictated by the shortest helix and as before it can not be deeper than the geometrical centre of the membrane and v) the depth values of residues located in the intramembraneous loop in the helix-in-turn-helix-out are the same as the depth of the flanking residues located at the end of the first helix and at beginning of the second helix. **Figure 6.8** shows two examples of the positioning of membrane dipping loops based on the empirically observed constraints described above.

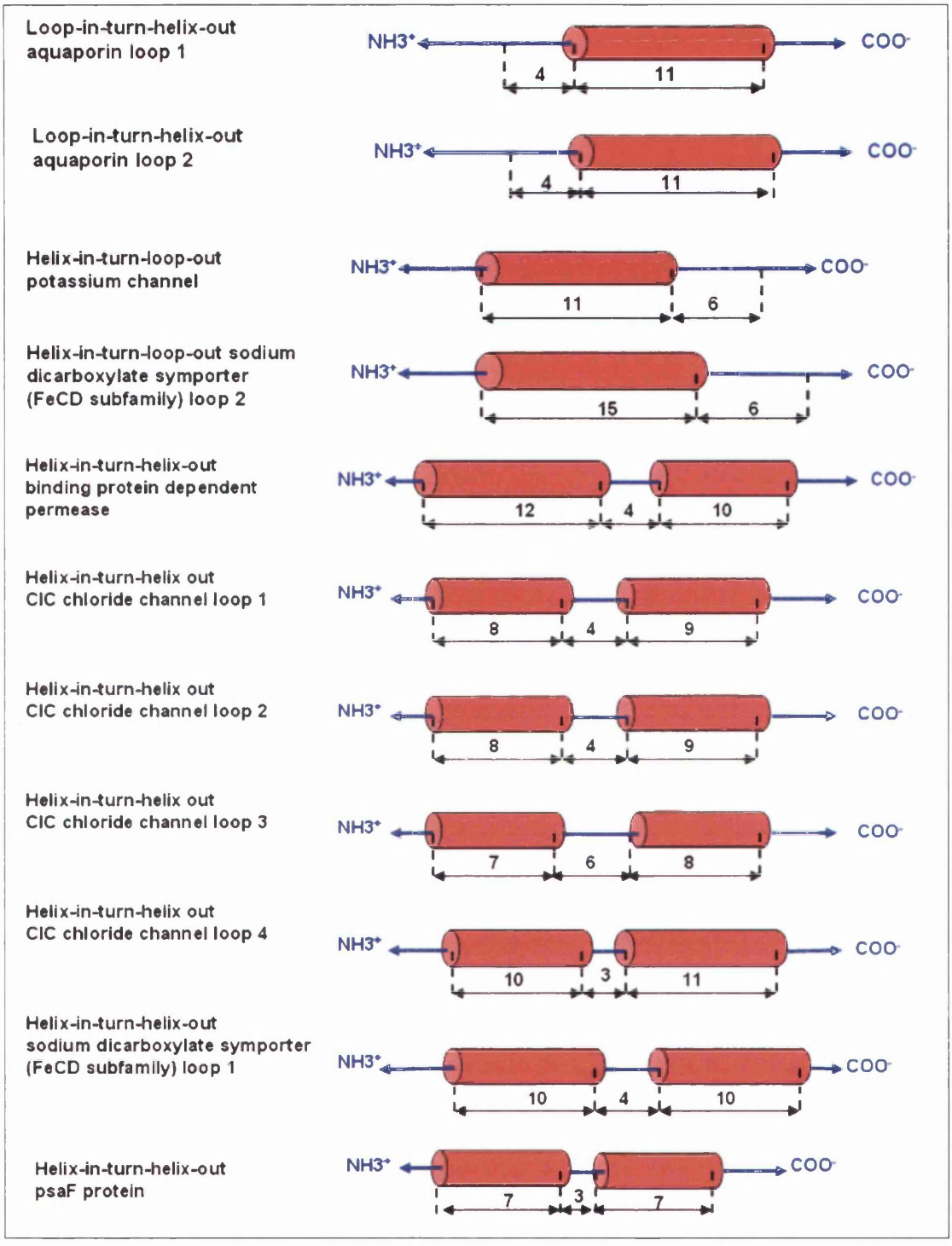

Figure 6.7. Structural models for each of the membrane dipping loops characterized. TMDEPTH applies the corresponding template to the membrane dipping loop predicted by TMLOOP and annotated by TMLOOP writer in order to predict the depths values for each residue belonging to the structural motif.

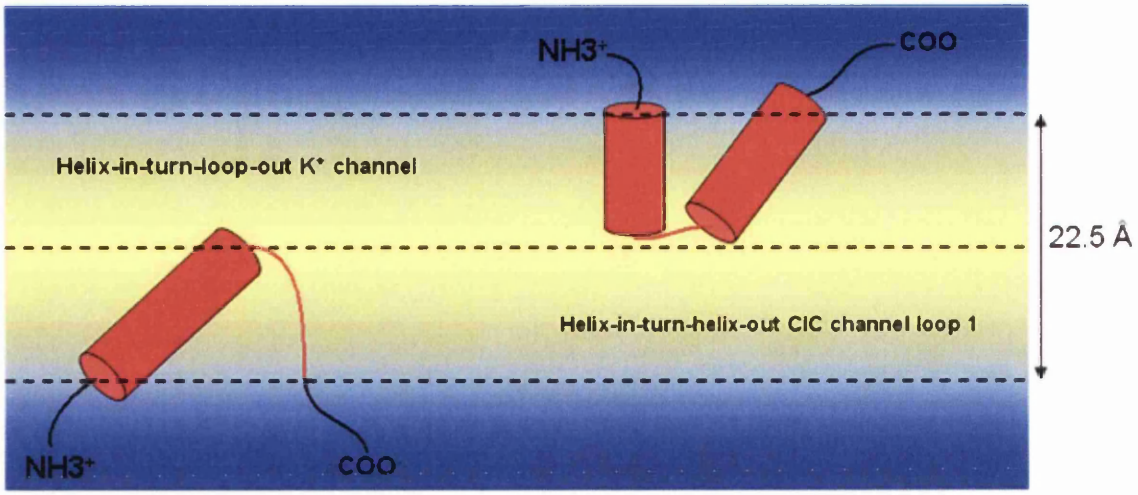

Figure 6.8. Positioning of membrane dipping loops in the membrane.

To calculate the depth of residues belonging to α-helices in membrane dipping loops the equations 6.1, 6.2 and 6.3 had to be adapted to the constrains explained above, i,e.

$$t = n^{(h)} \cdot 1.5\text{Å} \qquad \textit{only if} \qquad t \le T/2 \qquad (6.4.1)$$

$$t = T/2 \qquad \textit{only if} \qquad T/2 < n^{(h)} \cdot 1.5\text{Å} \qquad (6.4.2)$$

where t is the maximum depth achieved by the membrane dipping loop, n(h) is the number of residues h of the shortest helix in the given structural domain and T is the membrane thickness **(6.1)**.

$$d^{(H)} = \frac{t}{n^{(H)} - 1} \qquad \text{only if secondary structure belongs to a helix} \qquad (6.5.1)$$

$$d^{(L)} = \frac{t}{n^{(L)} - 1} \qquad \text{only if secondary structure belongs to a loop} \qquad (6.5.2)$$

$$d^{(L)} = 0 \qquad \text{only if helix-in-turn-helix-out} \qquad (6.5.3)$$

where d(H) is the intrahelical distance parallel to the membrane normal between contiguous residues, t is the maximum depth achieved by the membrane dipping loop, n(H) is the number of residues in the given helix, d(L) in the intraloop distance parallel to the

membrane normal between two contiguous residues and n(L) is the number of residues in the given loop.

$$e^{(r)} = \pm\left(t - d^{(H)} \cdot p^{(r)}\right) \quad \text{only if secondary structure belongs to a helix} \qquad (6.6.1)$$

$$e^{(r)} = \pm\left(t - d^{(L)} \cdot p^{(r)}\right) \quad \text{only if secondary structure belongs to a loop} \qquad (6.6.2)$$

where e(r) is the depth value of a given residue, t is the maximum depth in the membrane achieved the membrane dipping loop, d(H) is the intrahelical distance parallel to the membrane normal between contiguous residues, p(r) is the position of the given residue in the corresponding helix (starting position is 0) and d(L) is the intraloop distance parallel to the membrane normal between two contiguous residues.

6.2.5 Calculation of interhelical associations

After each residue in the membrane is given a depth value, TMDEPTH searches for pairs of residues located in different transmembrane regions whose depth values are within a range of ±1.5 Å. This depth range was believed to minimize potential errors by topology prediction method to accurately identify the first residue in the membrane for a given helix. The depth range was set to 1.5 Å, which corresponds to the intrahelical distance, along the helix axis, between contiguous residues. By using this depth range, only one residue error is allowed by topology prediction methods. However, the residue error allowed by topology prediction methods increases with longer transmembrane helices (as the helix tilt increases the number of residues located within the depth range increases).

Associations are stored in a 20 x 20 matrix where the x and y axis list the 20 known residues. TMDEPTH iteratively compares each transmembrane region with each of the others (but not against itself) and within each transmembrane region comparison, each residue in the given transmembrane region is compared to the residues of the other transmembrane regions **(figure 6.9)**.

20x20 Matrix calculated by depth similarity

	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL	TOTAL
ALA	12	1	0	0	0	0	0	3	0	3	3	0	0	2	0	3	2	2	2	5	38
ARG	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
ASN	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ASP	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CYS	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GLN	0	0	0	0	0	0	1	0	0	1	3	0	1	1	0	0	1	1	0	1	10
GLU	0	0	0	0	0	1	0	0	0	1	0	0	0	2	0	1	0	0	0	1	6
GLY	3	0	0	0	0	0	0	0	0	2	0	0	1	1	0	0	0	2	1	2	12
HIS	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	1	0	0	0	4
ILE	3	0	0	0	0	1	1	2	0	0	3	0	2	2	0	0	0	0	1	1	16
LEU	3	0	0	0	0	3	0	0	3	3	8	0	1	6	0	5	3	3	3	4	45
LYS	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
MET	0	0	0	0	0	1	0	1	0	2	1	0	0	1	0	1	3	1	0	0	11
PHE	2	0	0	0	0	1	2	1	0	2	6	0	1	0	0	0	0	1	1	2	19
PRO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SER	3	0	0	0	0	0	1	0	0	0	5	0	1	0	0	0	0	1	0	1	12
THR	2	0	0	0	0	1	0	0	1	0	3	0	3	0	0	0	0	1	0	3	14
TRP	2	0	0	0	0	1	0	2	0	0	3	0	1	1	0	1	1	0	0	0	12
TYR	2	0	0	0	0	0	0	1	0	1	3	0	0	1	0	0	0	0	0	1	9
VAL	5	0	0	0	0	1	1	2	0	1	4	0	0	2	0	1	3	0	1	0	21

Figure 6.9. Example of a 20x20 matrix calculated by TMDEPTH. The x and y axis show the 20 different amino acids and the last column shows the total number of associations for each residue.

6.2.6 Data standardization and extraction of other biological relevant information

In order to combine matrices corresponding to proteins belonging to the same cluster (to identify common patterns of interhelical associations) and compare matrices corresponding to proteins belonging to different clusters (to identify specific patterns of interhelical associations) it was necessary to normalize these matrices. Standardization was achieved by obtaining the percentage of each possible pair. i.e,

$$P_{(i,j)} = \frac{n_{(i,j)}}{\sum_{i,j=0}^{20} n_{(i,j)}} \cdot 100 \qquad (6.7)$$

where i and j refer to the x and y coordinates in the 20x20 matrix respectively.

By standardization of the 20x20 matrix a triangular 20x20 matrix was developed **(figure 6.10)**. This matrix shows the percentage of interhelical associations between residues located at a similar depth (±1.5Å) in the membrane. A triangular matrix was used because the values reflected in this matrix are mutual ($x_{ij} = x_{ji}$).

20 x 20 percentage triangle matrix

	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
ALA	5.22																			
ARG	0.87	0.00																		
ASN	0.00	0.00	0.00																	
ASP	0.00	0.00	0.00	0.00																
CYS	0.00	0.00	0.00	0.00	0.00															
GLN	0.00	0.00	0.00	0.00	0.00	0.00														
GLU	0.00	0.00	0.00	0.00	0.00	0.87	0.00													
GLY	2.61	0.00	0.00	0.00	0.00	0.00	0.00	0.00												
HIS	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00											
ILE	2.61	0.00	0.00	0.00	0.00	0.87	0.87	1.74	0.00	0.00										
LEU	2.61	0.00	0.00	0.00	0.00	2.61	0.00	0.00	2.61	2.61	3.48									
LYS	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00								
MET	0.00	0.00	0.00	0.00	0.00	0.87	0.00	0.87	0.00	1.74	0.87	0.00	0.00							
PHE	1.74	0.00	0.00	0.00	0.00	0.87	1.74	0.87	0.00	1.74	5.22	0.00	0.87	0.00						
PRO	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00					
SER	2.61	0.00	0.00	0.00	0.00	0.00	0.87	0.00	0.00	0.00	4.35	0.00	0.87	0.00	0.00	0.00				
THR	1.74	0.00	0.00	0.00	0.00	0.87	0.00	0.00	0.87	0.00	2.61	0.00	2.61	0.00	0.00	0.00	0.00			
TRP	1.74	0.00	0.00	0.00	0.00	0.87	0.00	1.74	0.00	0.00	2.61	0.00	0.87	0.87	0.00	0.87	0.87	0.00		
TYR	1.74	0.00	0.00	0.00	0.00	0.00	0.00	0.87	0.00	0.87	2.61	0.00	0.00	0.87	0.00	0.00	0.00	0.00	0.00	
VAL	4.35	0.00	0.00	0.00	0.00	0.87	0.87	1.74	0.00	0.87	3.48	0.00	0.00	1.74	0.00	0.87	2.61	0.00	0.87	0.00

Figure 6.10. Standardized triangle 20x20 matrix. As with the 20x20 matrix of interhelical association the x and y axis list the 20 different amino acids and each coordinate within the matrix represent the percentage of interhelical associations within the membrane of two given residues located at a similar depth.

Other possible relevant information was extracted from the standardized triangular 20x20 matrix. The percentage of the participation within the standardized matrix was obtained for each residue **(figure 6.11a)**. The standardized associations contained in the 20x20 triangle matrix were also clustered according to the biochemical behaviour of 20 residues listed (Non-polar, polar and charged) into a 3x3 standardized triangular matrix **(figure 6.11b)**.

Percentage of residue participation

ALA:	16.52	LEU:	19.57
ARG:	0.43	LYS:	0.00
ASN:	0.00	MET:	4.78
ASP:	0.00	PHE:	8.26
CYS:	0.00	PRO:	0.00
GLN:	4.35	SER:	5.22
GLU:	2.61	THR:	6.09
GLY:	5.22	TRP:	5.22
HIS:	1.74	TYR:	3.91
ILE:	6.96	VAL:	9.13

3 x 3 percentage triangle matrix based on the amino acid biochemical behaviour

	Non-Polar	Polar	Charged
Non-Polar	45.22		
Polar	43.48	1.74	
Charged	6.96	2.61	0.00

Sum of Percentages: 100.00

Figure 6.11. Figure 6.11a (left) shows the percentage participation for each amino acid in the formation of interhelical associations between residues located at a similar depth in the membrane. Figure 6.11b (right) shows the percentage of interhelical associations of clusters of residues of similar physicochemical property located at a similar depth in the membrane.

6.2.7 Report of the extracted features in a specific format required by the user

TMDEPTH has also been implemented to report the percentage of interhelical associations computed (the normalized triangular 20x20 matrix, the percentage of participation for each amino acid and the percentage of interhelical associations of clusters residues of similar physichochemical property) in a specific format required by the user. The user is given the option of saving the corresponding information in the specific format for Microsoft Excel, a data mining tool based on decision trees (Rulequest Research, http://www.rulequest.com/index.html), a data mining tool based on support vector machines (Joachims, 2002) and a combinatorial pattern discovery tool (Rigoutsos and Floratos, 1998). These data analyses, data mining and pattern recognition programs along with the Weka platform (Witten and Frank, 2005), which can also load input files specific for C4.5 (Quinlan, 1993), have proven to be very successful in the analysis of biological data **(Chapter 7, chapter 8)**.

6.2.8 Analysis of protein complexes with TMDEPTH

TMDEPTH can also apply the same principles to protein complexes. When a protein complex is loaded into TMDEPTH, the software first loads each subunit and then computes the corresponding parameters in a similar fashion as with monomers. In order to calculate the membrane thickness the shortest transmembrane region (whose length is 14 residues or higher) within the complex is used to calculate the membrane thickness. Depth values are given to each residue in the membrane based on the membrane thickness, the length and orientation of the corresponding helix and its position in the transmembrane region. When loading complexes, two different levels of interhelical associations are calculated: i) the first level is within each subunit using the same principles as with monomers, ii) the second level is between subunits where each helix in each subunit is paired with the helices of the remaining subunits and pairs of residues located at a similar depth are listed in the 20x20 matrix.

6.3 References

ADAMIAN, L., JACKUPS, R., JR., BINKOWSKI, T. A. & LIANG, J. (2003) Higher-order interhelical spatial interactions in membrane proteins. *J Mol Biol,* 327, 251-72.

ADAMIAN, L. & LIANG, J. (2001) Helix-helix packing and interfacial pairwise interactions of residues in membrane proteins. *J Mol Biol,* 311, 891-907.

ADAMIAN, L. & LIANG, J. (2002) Interhelical hydrogen bonds and spatial motifs in membrane proteins: polar clamps and serine zippers. *Proteins,* 47, 209-18.

AGRE, P. & KOZONO, D. (2003) Aquaporin water channels: molecular mechanisms for human diseases. *FEBS Lett,* 555, 72-8.

ALLEN, S. J., KIM, J. M., KHORANA, H. G., LU, H. & BOOTH, P. J. (2001) Structure and function in bacteriorhodopsin: the effect of the interhelical loops on the protein folding kinetics. *J Mol Biol,* 308, 423-35.

ALOY, P., CEDANO, J., OLIVA, B., AVILES, F. X. & QUEROL, E. (1997) 'TransMem': a neural network implemented in Excel spreadsheets for predicting transmembrane domains of proteins. *Comput Appl Biosci,* 13, 231-4.

AMICO, M., FINELLI, M., ROSSI, I., ZAULI, A., ELOFSSON, A., VIKLUND, H., VON HEIJNE, G., JONES, D., KROGH, A., FARISELLI, P., LUIGI MARTELLI, P. & CASADIO, R. (2006) PONGO: a web server for multiple predictions of all-alpha transmembrane proteins. *Nucleic Acids Res,* 34, W169-72.

ARAI, M., MITSUKE, H., IKEDA, M., XIA, J. X., KIKUCHI, T., SATAKE, M. & SHIMIZU, T. (2004a) ConPred II: a consensus prediction method for obtaining transmembrane topology models with high reliability. *Nucleic Acids Res,* 32, W390-3.

ARAI, M., OKUMURA, K., SATAKE, M. & SHIMIZU, T. (2004b) Proteome-wide functional classification and identification of prokaryotic transmembrane proteins by transmembrane topology similarity comparison. *Protein Sci,* 13, 2170-83.

BERMAN, H. M., WESTBROOK, J., FENG, Z., GILLILAND, G., BHAT, T. N., WEISSIG, H., SHINDYALOV, I. N. & BOURNE, P. E. (2000) The Protein Data Bank. *Nucleic Acids Res,* 28, 235-42.

CASADIO, R., FARISELLI, P., TARONI, C. & COMPIANI, M. (1996) A predictor of transmembrane alpha-helix domains of proteins based on neural networks. *Eur Biophys J,* 24, 165-78.

CHEN, C. P., KERNYTSKY, A. & ROST, B. (2002) Transmembrane helix predictions revisited. *Protein Sci,* 11, 2774-91.

CHOMA, C., GRATKOWSKI, H., LEAR, J. D. & DEGRADO, W. F. (2000) Asparagine-mediated self-association of a model transmembrane helix. *Nat Struct Biol,* 7, 161-6.

CLAROS, M. G. & VON HEIJNE, G. (1994) TopPred II: an improved software for membrane protein structure predictions. *Comput Appl Biosci,* 10, 685-6.

CSERZO, M., WALLIN, E., SIMON, I., VON HEIJNE, G. & ELOFSSON, A. (1997) Prediction of transmembrane alpha-helices in prokaryotic membrane proteins: the dense alignment surface method. *Protein Eng,* 10, 673-6.

CURRAN, A. R. & ENGELMAN, D. M. (2003) Sequence motifs, polar interactions and conformational changes in helical membrane proteins. *Curr Opin Struct Biol,* 13, 412-7.

DAWSON, J. P., MELNYK, R. A., DEBER, C. M. & ENGELMAN, D. M. (2003) Sequence context strongly modulates association of polar residues in transmembrane helices. *J Mol Biol,* 331, 255-62.

DAWSON, J. P., WEINGER, J. S. & ENGELMAN, D. M. (2002) Motifs of serine and threonine can drive association of transmembrane helices. *J Mol Biol,* 316, 799-805.

DEBER, C. M., KHAN, A. R., LI, Z., JOENSSON, C., GLIBOWICKA, M. & WANG, J. (1993) Val-->Ala mutations selectively alter helix-helix packing in the transmembrane segment of phage M13 coat protein. *Proc Natl Acad Sci U S A,* 90, 11648-52.

DEGLI ESPOSTI, M., CRIMI, M. & VENTUROLI, G. (1990) A critical evaluation of the hydropathy profile of membrane proteins. *Eur J Biochem,* 190, 207-19.

EFREMOV, R. G. & VERGOTEN, G. (1996) Recognition of transmembrane alpha-helical segments with environmental profiles. *Protein Eng,* 9, 253-63.

EILERS, M., PATEL, A. B., LIU, W. & SMITH, S. O. (2002) Comparison of helix interactions in membrane and soluble alpha-bundle proteins. *Biophys J,* 82, 2720-36.

EILERS, M., SHEKAR, S. C., SHIEH, T., SMITH, S. O. & FLEMING, P. J. (2000) Internal packing of helical membrane proteins. *Proc Natl Acad Sci U S A,* 97, 5796-801.

FARISELLI, P. & CASADIO, R. (1996) HTP: a neural network-based method for predicting the topology of helical transmembrane domains in proteins. *Comput Appl Biosci,* 12, 41-8.

GRATKOWSKI, H., LEAR, J. D. & DEGRADO, W. F. (2001) Polar side chains drive the association of model transmembrane peptides. *Proc Natl Acad Sci U S A,* 98, 880-5.

GROMIHA, M. M. (1999) A simple method for predicting transmembrane alpha helices with better accuracy. *Protein Eng,* 12, 557-61.

HIROKAWA, T., BOON-CHIENG, S. & MITAKU, S. (1998) SOSUI: classification and secondary structure prediction system for membrane proteins. *Bioinformatics,* 14, 378-9.

IKEDA, M., ARAI, M., LAO, D. M. & SHIMIZU, T. (2002) Transmembrane topology prediction methods: a re-assessment and improvement by a consensus method using a dataset of experimentally-characterized transmembrane topologies. *In Silico Biol,* 2, 19-33.

INOUE, Y., IKEDA, M. & SHIMIZU, T. (2004) Proteome-wide classification and identification of mammalian-type GPCRs by binary topology pattern. *Comput Biol Chem,* 28, 39-49.

JOACHIMS, T. (2002) *Learning to Classify Text Using Support Vector Machines: methods, theory and algorithms*, Kluwer Academic Publishers.

JONES, D. T., TAYLOR, W. R. & THORNTON, J. M. (1994) A model recognition approach to the prediction of all-helical membrane protein structure and topology. *Biochemistry,* 33, 3038-49.

KAHSAY, R. Y., GAO, G. & LIAO, L. (2005) An improved hidden Markov model for transmembrane protein detection and topology prediction and its applications to complete genomes. *Bioinformatics,* 21, 1853-8.

KALL, L., KROGH, A. & SONNHAMMER, E. L. (2004) A combined transmembrane topology and signal peptide prediction method. *J Mol Biol,* 338, 1027-36.

KALL, L. & SONNHAMMER, E. L. (2002) Reliability of transmembrane predictions in whole-genome data. *FEBS Lett,* 532, 415-8.

KIM, J. M., BOOTH, P. J., ALLEN, S. J. & KHORANA, H. G. (2001) Structure and function in bacteriorhodopsin: the role of the interhelical loops in the folding and stability of bacteriorhodopsin. *J Mol Biol,* 308, 409-22.

KLEIN, P., KANEHISA, M. & DELISI, C. (1985) The detection and classification of membrane-spanning proteins. *Biochim Biophys Acta,* 815, 468-76.

KROGH, A., LARSSON, B., VON HEIJNE, G. & SONNHAMMER, E. L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol,* 305, 567-80.

KYTE, J. & DOOLITTLE, R. F. (1982) A simple method for displaying the hydropathic character of a protein. *J Mol Biol,* 157, 105-32.

LANGOSCH, D., BROSIG, B., KOLMAR, H. & FRITZ, H. J. (1996) Dimerisation of the glycophorin A transmembrane segment in membranes probed with the ToxR transcription activator. *J Mol Biol,* 263, 525-30.

LANGOSCH, D. & HERINGA, J. (1998) Interaction of transmembrane helices by a knobs-into-holes packing characteristic of soluble coiled coils. *Proteins,* 31, 150-9.

LASSO, G. (2001) 20x20 matrix classification. *Membrane Protein Bioinformatics Group.* Luton, University of Bedfordshire.

LEMMON, M. A., FLANAGAN, J. M., TREUTLEIN, H. R., ZHANG, J. & ENGELMAN, D. M. (1992) Sequence specificity in the dimerization of transmembrane alpha-helices. *Biochemistry,* 31, 12719-25.

LIU, Y., ENGELMAN, D. M. & GERSTEIN, M. (2002) Genomic analysis of membrane protein families: abundance and conserved motifs. *Genome Biol,* 3, research0054.

LOHMANN, R., SCHNEIDER, G. & WREDE, P. (1996) Structure optimization of an artificial neural filter detecting membrane-spanning amino acid sequences. *Biopolymers,* 38, 13-29.

MACKENZIE, K. R., PRESTEGARD, J. H. & ENGELMAN, D. M. (1997) A transmembrane helix dimer: structure and implications. *Science,* 276, 131-3.

MARTELLI, P. L., FARISELLI, P. & CASADIO, R. (2003) An ENSEMBLE machine learning approach for the prediction of all-alpha membrane proteins. *Bioinformatics,* 19 Suppl 1, i205-11.

MOLLER, S., CRONING, M. D. & APWEILER, R. (2001) Evaluation of methods for the prediction of membrane spanning regions. *Bioinformatics,* 17, 646-53.

MONNE, M., NILSSON, I., ELOFSSON, A. & VON HEIJNE, G. (1999) Turns in transmembrane helices: determination of the minimal length of a "helical hairpin" and derivation of a fine-grained turn propensity scale. *J Mol Biol,* 293, 807-14.

MORIKI, T., MARUYAMA, H. & MARUYAMA, I. N. (2001) Activation of preformed EGF receptor dimers by ligand-induced rotation of the transmembrane domain. *J Mol Biol,* 311, 1011-26.

NIKIFOROVICH, G. V. (1998) A novel, non-statistical method for predicting breaks in transmembrane helices. *Protein Eng,* 11, 279-83.

NILSSON, J., PERSSON, B. & VON HEIJNE, G. (2000) Consensus predictions of membrane protein topology. *FEBS Lett,* 486, 267-9.

NILSSON, J., PERSSON, B. & VON HEIJNE, G. (2002) Prediction of partial membrane protein topologies using a consensus approach. *Protein Sci,* 11, 2974-80.

PARODI, L. A., GRANATIR, C. A. & MAGGIORA, G. M. (1994) A consensus procedure for predicting the location of alpha-helical transmembrane segments in proteins. *Comput Appl Biosci,* 10, 527-35.

PERSSON, B. & ARGOS, P. (1994) Prediction of transmembrane segments in proteins utilising multiple sequence alignments. *J Mol Biol,* 237, 182-92.

PERSSON, B. & ARGOS, P. (1996) Topology prediction of membrane proteins. *Protein Sci,* 5, 363-71.

PERSSON, B. & ARGOS, P. (1997) Prediction of membrane protein topology utilizing multiple sequence alignments. *J Protein Chem,* 16, 453-7.

PILPEL, Y., BEN-TAL, N. & LANCET, D. (1999) kPROT: a knowledge-based scale for the propensity of residue orientation in transmembrane segments. Application to membrane protein structure prediction. *J Mol Biol,* 294, 921-35.

PONNUSWAMY, P. K. & GROMIHA, M. M. (1993) Prediction of transmembrane helices from hydrophobic characteristics of proteins. *Int J Pept Protein Res,* 42, 326-41.

POPOT, J. L. & ENGELMAN, D. M. (1990) Membrane protein folding and oligomerization: the two-stage model. *Biochemistry,* 29, 4031-7.

POPOT, J. L. & ENGELMAN, D. M. (2000) Helical membrane protein folding, stability, and evolution. *Annu Rev Biochem,* 69, 881-922.

PROMPONAS, V. J., PALAIOS, G. A., PASQUIER, C. M., HAMODRAKAS, J. S. & HAMODRAKAS, S. J. (1999) CoPreTHi: a Web tool which combines transmembrane protein segment prediction methods. *In Silico Biol,* 1, 159-62.

QUINLAN, R. (1993) *C4.5: Programs for Machine Learning,* San Mateo, CA, Morgan Kaufmann Publishers.

RIGOUTSOS, I. & FLORATOS, A. (1998) Combinatorial pattern discovery in biological sequences: The TEIRESIAS algorithm. *Bioinformatics,* 14, 55-67.

ROST, B., CASADIO, R. & FARISELLI, P. (1996) Refining neural network predictions for helical transmembrane proteins by dynamic programming. *Proc Int Conf Intell Syst Mol Biol,* 4, 192-200.

ROST, B., CASADIO, R., FARISELLI, P. & SANDER, C. (1995) Transmembrane helices predicted at 95% accuracy. *Protein Sci,* 4, 521-33.

RUSS, W. P. & ENGELMAN, D. M. (2000) The GxxxG motif: a framework for transmembrane helix-helix association. *J Mol Biol,* 296, 911-9.

SCHLESSINGER, J. (2002) Ligand-induced, receptor-mediated dimerization and activation of EGF receptor. *Cell,* 110, 669-72.

SENES, A., GERSTEIN, M. & ENGELMAN, D. M. (2000) Statistical analysis of amino acid patterns in transmembrane helices: the GxxxG motif occurs frequently and in association with beta-branched residues at neighboring positions. *J Mol Biol,* 296, 921-36.

SENES, A., UBARRETXENA-BELANDIA, I. & ENGELMAN, D. M. (2001) The Calpha ---H...O hydrogen bond: a determinant of stability and specificity in transmembrane helix interactions. *Proc Natl Acad Sci U S A,* 98, 9056-61.

SHAN, S. O. & HERSCHLAG, D. (1996) The change in hydrogen bond strength accompanying charge rearrangement: implications for enzymatic catalysis. *Proc Natl Acad Sci U S A,* 93, 14474-9.

SONNHAMMER, E. L., VON HEIJNE, G. & KROGH, A. (1998) A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc Int Conf Intell Syst Mol Biol,* 6, 175-82.

STOFFEL, W., DUKER, M. & HOFMANN, K. (1993) Molecular cloning and gene organization of the mouse mitochondrial 3,2-trans-enoyl-CoA isomerase. *FEBS Lett,* 333, 119-22.

SUGIYAMA, Y., POLULYAKH, N. & SHIMIZU, T. (2003) Identification of transmembrane protein functions by binary topology patterns. *Protein Eng,* 16, 479-88.

TAYLOR, P. D., ATTWOOD, T. K. & FLOWER, D. R. (2003) BPROMPT: A consensus server for membrane protein prediction. *Nucleic Acids Res,* 31, 3698-700.

TOGAWA, R. (2006) Development of a suite of bioinformatics tools for the analysis and prediction of membrane protein structure. Luton, University of Bedfordshire.

TOURASSE, N. J. & LI, W. H. (2000) Selective constraints, amino acid composition, and the rate of protein evolution. *Mol Biol Evol,* 17, 656-64.

TOYOSHIMA, C., NAKASAKO, M., NOMURA, H. & OGAWA, H. (2000) Crystal structure of the calcium pump of sarcoplasmic reticulum at 2.6 A resolution. *Nature,* 405, 647-55.

TUSNADY, G. E. & SIMON, I. (1998) Principles governing amino acid composition of integral membrane proteins: application to topology prediction. *J Mol Biol,* 283, 489-506.

TUSNADY, G. E. & SIMON, I. (2001) The HMMTOP transmembrane topology prediction server. *Bioinformatics,* 17, 849-50.

ULMSCHNEIDER, M. B. & SANSOM, M. S. (2001) Amino acid distributions in integral membrane protein structures. *Biochim Biophys Acta,* 1512, 1-14.

VIKLUND, H. & ELOFSSON, A. (2004) Best alpha-helical transmembrane protein topology predictions are achieved using hidden Markov models and evolutionary information. *Protein Sci,* 13, 1908-17.

VON HEIJNE, G. (1992) Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule. *J Mol Biol,* 225, 487-94.

VON HEIJNE, G. & GAVEL, Y. (1988) Topogenic signals in integral membrane proteins. *Eur J Biochem,* 174, 671-8.

WALTERS, R. F. & DEGRADO, W. F. (2006) Helix-packing motifs in membrane proteins. *Proc Natl Acad Sci U S A,* 103, 13658-63.

WITTEN, I. H. & FRANK, E. (2005) *Data Mining: Practical machine learning tools and techniques,* San Francisco, Morgan Kaufmann.

XIA, J. X., IKEDA, M. & SHIMIZU, T. (2004) ConPred_elite: a highly reliable approach to transmembrane topology predication. *Comput Biol Chem,* 28, 51-60.

XU, E. W., KEARNEY, P. & BROWN, D. G. (2006) The use of functional domains to improve transmembrane protein topology prediction. *J Bioinform Comput Biol,* 4, 109-23.

ZHOU, F. X., COCCO, M. J., RUSS, W. P., BRUNGER, A. T. & ENGELMAN, D. M. (2000) Interhelical hydrogen bonding drives strong interactions in membrane proteins. *Nat Struct Biol,* 7, 154-60.

ZHOU, F. X., MERIANOS, H. J., BRUNGER, A. T. & ENGELMAN, D. M. (2001) Polar residues drive association of polyleucine transmembrane helices. *Proc Natl Acad Sci U S A,* 98, 2250-5.

ZHOU, H. & ZHOU, Y. (2003) Predicting the topology of transmembrane helical proteins using mean burial propensity and a hidden-Markov-model-based method. *Protein Sci,* 12, 1547-55.

CHAPTER 7

TMLOCATE: Prediction of subcellular location of eukaryotic membrane proteins based on sequence and topological information

7.1 Introduction

7.1.1 Subcellular location linked with function

The intracellular structures delimited by a lipid bilayer in eukaryotic cells are known to perform particular functions and be involved in biological pathways that are specific for a given organelle. The organelle-specific functions and biological pathways are carried out by proteins, located either in the surrounding membrane or inside the organelle, that in protein trafficking are targeted to that organelle in order to perform their specific biochemical activity. Therefore, identification of the subcellular location of a protein is considered as the first step towards elucidation of the function of the protein. The problem of subcellular identification has been tackled at the experimental level, with the development of high-throughput techniques, and at the computational level by attempting to predict the location of a given protein based upon selected features obtained from the amino acid sequence. High-throughput methods are often hampered by the low concentration of proteins and the dynamic environment of the cell where it is a continuous traffic of lipid and proteins between organelles is found (Aturaliya et al., 2006). On the other hand, prediction of subcellular location based solely on sorting signals or amino acid sequence has proven to be a challenge due to the apparent lack of a universal targeting mechanism.

7.1.2 Current methods *used* to predict subcellular location

Since the first methods to predict the subcellular location of proteins were published in the early 1990's, a variety of methods based on different features and using different data mining methods have been published. A common approach has been based on the detection of sorting signals to predict the subcellular location of proteins. Claros and colleagues discovered that 76-94% of the analyzed mitochondrial proteins contained mitochondrial targeting motifs (Claros and Vincens, 1996). Further work carried out by Fujiwara developed a hidden Markov model that detected various known mitochondrial targeting motifs obtaining a 86.9% prediction accuracy (Fujiwara et al., 1997). Emanuelsson and colleagues developed a neural network based method to identify chloroplast transit peptides and their cleavage sites and reported 88% prediction accuracy using a non-redundant set of proteins (Emanuelsson et al., 1999). Bickmore stated that motifs and domains are often shared by proteins located within the same nuclear compartment (Bickmore and Sutherland, 2002). Aiming to predict nuclear localization, Cokol and colleagues collected a set of experimentally checked nuclear localization signals (NLS) and extended the given set by "*in silico*" mutagenesis, the final set of NLSs was found to match 43% of all known nuclear proteins (Cokol et al., 2000, Nair et al., 2003).

Information contained in the N-terminus was exploited in conjunction with neural networks to discriminate between proteins located in the mitochondrion, chloroplast, the secretory pathway and "other" locations. The method reported a prediction accuracy of 85-90% (Emanuelsson et al., 2000). Petsalaki and co-authors extended this work by combining neural networks, profile hidden Markov models and scoring matrices (Petsalaki et al., 2006). Although neural networks and other "black box" methods such as hidden Markov models or support vector machines report accurate predictive values, it is difficult to understand the basis upon which the predictions are made, Bannai and colleagues (Bannai et al., 2002) extracted simple and interpretable rules obtained from the N-terminus to detect targeting signal. The authors combined different amino acid indexes, obtained both theoretically and experimentally (Kawashima and Kanehisa, 2000) to compute different physicochemical properties of the N-terminus. Likewise, the approximate patterns were combined with alphabet indexes (classification of characters of an alphabet -e.g. amino

acids- into a smaller set of characters –e.g. binary classification of hydrophobicity-). The developed method showed Matthews Correlation Coefficients (MCC) between 0.64 and 0.92 although it did not achieved higher predictive scores than the neural network based program TargetP (Emanuelsson et al., 2000). Boden and Hawkins (Boden and Hawkins, 2005) introduced a new model, namely the recurrent network model, to detect targeting peptides, which increased the prediction accuracy by 5-6% compared to TargetP. In more recent work the authors (Hawkins and Boden, 2006) combined targeting signals detection methods with a multilayer classifier system composed by neural networks and support vector machines further improving the accuracy by 2% compared to the previous approach.

Despite this approach providing promising results, prediction of subcellular location based solely on the presence of targeting signals has limitations: i) proteins targeted to the same organelle have been shown to contain different targeting signals or conversely some proteins lack a particular motif, which describes a given targeting mechanism in other proteins found in the same organelle, ii) high-throughput genome analyses can lead to the "calling" of unreliable 5' – regions, which may result in targeting signals being missed or partially included (Reinhardt and Hubbard, 1998) and iii) targeting signals, namely silent motifs, can be found in unrelated protein sequences due to neutral mutations (these signals are called silent because they do not promote subcellular localization as they are not accessible to the appropriate receptor) (Neuberger et al., 2004).

Another classical approach is the prediction of subcellular location based on the differential amino acid composition of proteins belonging to different organelles. The early work of Cedano and colleagues (Cedano et al., 1997), which discriminates between integral membrane proteins, anchored membrane proteins, extracellular proteins, intracellular proteins and nuclear proteins, solely based on the amino acid composition encouraged further research. Different data mining methods, such as neural networks (Cai et al., 2002a, Reinhardt and Hubbard, 1998), support vector machines (Cai et al., 2002b, Park and Kanehisa, 2003) and a covariant discriminant algorithm (Chou and Elrod, 1999b), were employed to exploit differential amino acid composition. Variations of this approach were also introduced by different authors: i) Feng combined amino acid composition and

hydrophobicity profiles to predict the subcellular location of prokaryotic proteins (Feng and Zhang, 2001); and ii) Matsuda and co-authors introduced the use of local (roughly N-terminal, middle and C-terminal) compositions of amino acids, twin amino acids and local frequencies of distance between successive amino acids, and reported prediction accuracy values of 87-91% (Matsuda et al., 2005).

Other research has focused on the development of predictive algorithms that combine sorting signals and amino acid compositions in order to account for different targeting mechanisms. The well established predictive algorithm, PSORT, successfully combined both sets of features (Nakai and Horton, 1999, Nakai and Kanehisa, 1991, Nakai and Kanehisa, 1992). Fugiwara applied neural networks to mine the differential amino acid composition, and hidden Markov models to sorting signals reporting an accuracy of 86-91% (Fujiwara and Asogawa, 2001) whereas the work of Reczko and colleagues was only based on neural networks and reported an accuracy of 91% (Reczko and Hatzigerrorgiou, 2004).

The use of sequence order has been introduced in later research in order to include contextual information. Huang and colleagues (Huang and Li, 2004), and Park and colleagues (Park and Kanehisa, 2003) extracted dipeptide information from the amino acid sequence whereas Yu and colleagues (Yu et al., 2004) used n-peptide compositions. Chou implemented the quasi-sequence-order effect based on the physicochemical distance between residues (Chou, 2000) and the pseudo-amino acid composition (Chou, 2001, Chou and Cai, 2003) to introduce the sequence order effect in the prediction of subcellular location of proteins. Pan and colleagues used the digital signal processing approach to translate the amino acid sequence into a numerical sequence in order to detect numerical signals that encapsulate the sequence order effect and can be used to predict the subcellular location (Pan et al., 2005, Pan et al., 2003). Cui and colleagues introduced the sequence order effect by dividing each protein sequences into two symmetrical halves and computing the amino acid composition of each half constructing a forty-dimensional vector (Cui et al., 2004). Finally, Markov chains have also been applied to make better use of the sequence order and contextual information (Bulashevska and Eils, 2006).

Another common approach is the use of homology-based methods to infer subcellular location under the assumption that highly identical protein sequences must be located in similar cellular compartments and it has been concluded that subcellular localization can indeed be inferred through homology search (Nair and Rost, 2002b). Marcotte and colleagues also studied the evolutionary relationships of proteins belonging to the same organelle using a phylogenetic approach and showed that proteins with the same subcellular location often show similar phylogenetic profiles (Marcotte et al., 2000). Xie and colleagues used position-specific scoring matrices extracted from the profile created by PSI-BLAST to generate a 400-dimension input vector, which was mined using support vector machines, and obtained an overall prediction accuracy of 90% (Xie et al., 2005).

Mott and colleagues used domains from the SMART database, the basic data of which are derived from hidden Markov models obtained from manually derived alignments of protein families (Letunic et al., 2006), clustering proteins on the basis of their domain co-occurrence to discriminate between the secretory, cytoplasmic and nuclear classes (Mott et al., 2002). Similarly, MITOPRED (Guda et al., 2004) was designed to discriminate between mitochondrial and non-mitochondrial proteins based on alignments and hidden Markov models contained in the Pfam database.

The five different categories mentioned above (based solely on sorting signals or amino acid composition, combining sorting signals with amino acid composition, introduction of sequence order and homology-based methods) describe the main approaches used to predict the subcellular location of proteins, though other methods have been developed, which employ different approaches to predict the subcellular location of proteins. Sarda and colleagues implemented a method based on multiple physicochemical properties of the residues composing the amino acid sequence and support vector machines reporting a prediction accuracy value of 93% (Sarda et al., 2005). Nair and colleagues developed a novel method to predict the subcellular location of proteins based on Swiss-Prot keywords and reported 82% accuracy showing that functional annotation can be used to infer subcellular location (Nair and Rost, 2002a), providing of course that the function is

known and the protein fully annotated in the database. Chou and colleagues developed a method to predict subcellular location based on the functional domain composition, where each protein is represented by a set of functional domains, and also considers partial sequence order (Chou and Cai, 2002). A similar approach, carried out by Scott and colleagues, is based on the combinatorial presence of InterPro and membrane domains (Scott et al., 2004). Further research was based on functional domains combined with the amino acid composition (Guda and Subramaniam, 2005) and with the pseudo-amino acid composition (Cai and Chou, 2003, Cai and Chou, 2004, Chou and Cai, 2004). Other research used the amino acid composition corresponding to the surface of crystallized proteins (Andrade et al., 1998) and combined this property with general amino acid composition, other structural features (e.g. predicted secondary structure) and sequence alignments (Nair and Rost, 2003a).

Other methods have integrated multiple features to predict subcellular location. Drawid and Gerstein combined a set of 30 features, including sorting signals, physico-chemical properties, amino acid composition on the surface of proteins and absolute mRNA expression levels, into a Bayesian system (Drawid and Gerstein, 2000). Nair and Rost combined nuclear localization signals (PredictNLS), keywords contained in the Swiss-Prot database (LOCkey), a homology-based method (LOChom) and an ab initio prediction method based on neural networks (LOC3Dini) to predict the subcellular location of proteins of known structure (Nair and Rost, 2003b). Although each method has been individually evaluated, the predictive accuracy of the consensus method was not reported, instead the method with the highest support was chosen. The predictions were stored in a database named LOC3D.

A specific version of PSORT for the subcellular prediction of proteins from Gram-negative bacteria combined amino acid composition, sequence similarity, sorting signals and the presence of signal peptides and transmembrane α-helices (Gardy et al., 2003). Bhasin and Raghava combined physicochemical properties, with amino acid composition, dipeptide composition and PSI-BLAST (Bhasin and Raghava, 2004). Pierleoni and colleagues designed a multilayer system based on different support vector machines that

based their prediction on the amino acid composition, the local amino acid composition contained in the N-terminus and C-terminus and sequence profiles obtained with BLAST (Pierleoni et al., 2006). Scott and colleagues created a Bayesian network predictor that combines InterPro motifs, targeting signals and protein-protein interaction data. This method was also used to discriminate between proteins located in the lumen of the organelle and proteins that are associated with these organelles on their cytosolic side (Scott et al., 2005).

7.1.3 Membrane proteins and prediction of subcellular location

To our knowledge, only a single previous attempt has been carried out to specifically predict the subcellular location of membrane proteins (Chou and Elrod, 1999a), which probably reflects the difficulty of this task. The research carried out by Chou and colleagues demonstrates that the subcellular location of membrane proteins is closely correlated with their amino acid composition. However, the assembled data set was not filtered at the sequence redundancy level, which may well have caused the over-fitting of the training set with the corresponding model and a subsequent overestimation of the performance. This was probably more obvious in the subsets of lysosome, nucleus and peroxisome where the imposed constraints at the protein name level were relaxed in order to obtain a significant set size. Likewise, the accuracy levels reported based on a jackknife evaluation considered only the non-normalized accuracy and the predictive accuracy for particular classes was omitted. Therefore, it is likely that the 65.9% accuracy reported by the jackknife validation was biased by the larger subsets (80% of the data set was composed by proteins located in the plasma membrane).

Mutagenesis analysis of human peroxin 2 protein (PEX2), a protein containing two transmembrane regions, showed that the minimum peroxisomal targeting signal involved the first transmembrane domain and that the second transmembrane region increased the targeting efficiency (Biermanns et al., 2003). Experimental analysis of the peroxisomal protein PEX16p (Honsho et al., 2002, Jones et al., 2004) showed two different regions essential for targeting the given protein into the peroxisome, which involve both of the

transmembrane regions found in this protein suggesting a possible role for interhelical associations between transmembrane regions in organellar targeting. This experimental evidence demonstrates that the transmembrane domain of at least some membrane proteins can play an important role in their organellar sorting.

Other experimental work has been focussed on the discovery of retention signals in the membrane. Rather than targeting to a specific organelle, retention signals maintain the location of a protein in a particular organelle. These types of signals seem to be particularly important within the secretory pathway where there is constant traffic of proteins between the different organelles involved. Previous work has concluded that the transmembrane regions of particular membrane proteins are essential retention domains (Aoki et al., 1992, Cocquerel et al., 1999, Colley, 1997, Hobman et al., 1997, Hobman et al., 1995, Ma et al., 2004, Op De Beeck et al., 2004). Several experimental analyses have shown that the presence of hydrophilic residues in the middle of transmembrane regions might be important for retention in the ER (Bonifacino et al., 1991, Cocquerel et al., 2000, Letourneur and Cosson, 1998, Yang et al., 1997). Furthermore, the length of the transmembrane regions has also been pinpointed as a form of retention signal. The plasma membrane has a higher content of cholesterol than the membrane found in the Golgi apparatus and consequently the plasma membrane is thicker than the membrane of the Golgi apparatus. Therefore, membrane proteins with longer transmembrane regions would not be retained in the Golgi apparatus whereas membrane proteins with shorter helices would be retained, as the protein would have attained its minimum potential energy (Munro, 1995). A similar retention property has been found in the ER (Pedrazzini et al., 1996, Yang et al., 1997, Szczesna-Skorupa and Kemper, 2000).

7.1.4 Our approach

The TMDEPTH approach **(Chapter 6)** has been applied to a data set of polytopic membrane proteins located in different cellular organelles assembled using PROCLASS **(Chapter 3)**. TMDEPTH combines sequence and a refined topological model obtained from the Swiss-Prot database in combination with TMLOOP and TMLOOP writer

(Chapter 5). TMDEPTH creates matrices of interhelical associations corresponding to residues located at a similar depth in the membrane. The created matrices were mined using different data mining methods (e.g. Bayesian methods and Support Vector Machines) within the Weka platform (Witten and Frank, 2005) in order to find the best classifier to predict the subcellular location of a given membrane protein using the information generated by TMDEPTH. The different data mining analyses carried out showed that using a set of classifiers arranged in a tree-based mode mimicking the processes and evolutionary relationship of cellular sorting achieved maximal predictive accuracy. A similar relationship has been described by Nair and Rost (2005). The designed architecture was set up to first distinguish between 4 classes, namely chloroplast, mitochondria, plasma membrane and secretory classes. The secretory class is further divided into subclasses (peroxisome, lysosome, Golgi apparatus, endoplasmic reticulum and the nucleus), which are predicted in the subsequent nodes of the tree-based set of classifiers. The method reported an approximate accuracy of 75% in correctly classifying a membrane protein into the four different classes, which reflects the importance of the transmembrane regions at this level. The performance of subsequent classifiers designed to sub-classify the secretory class suggested that other protein features, located outside the membrane, must play more important roles in locating a given membrane protein within the appropriate organelle within the secretory pathway. The obtained results are also consistent with evolutionary relationships between the different organelles, which associate a strong relationship between the mitochondrion and the peroxisome. This relationship probably reflects the recruitment of proteins originally targeted to the mitochondrion, first proposed by Gabaldón and colleagues (Gabaldon et al., 2006).

Considering that current topology prediction methods approximately achieve a 75% accuracy in correctly predicting the topology of a polytopic membrane protein and that some of the classes in the set (nucleus, lysosome and peroxisome) were composed by only a few proteins, it is believed the accuracy of the developed method can only increase with the improvement of topology prediction methods and high-throughput identification of membrane proteins. However, the longer term aim of the research carried out is not only to propose a method that can accurately predict the subcellular location of membrane proteins

based solely on transmembrane information but to formulate a method, that combines sequence and topology information, and can be used along with other protein features such as sorting signals and amino acid composition-based methods, to accurately predict the subcellular location of polytopic membrane proteins.

7.2 Methods

7.2.1 Data set development

The data set was retrieved from the Swiss-Prot database (release 50.2 of 27-06-2006) using PROCLASS **(Chapter 5)**. The obtained data set contained 5,619 eukaryotic membrane proteins with more than one transmembrane region clustered into 24 different categories including undefined membrane locations and multi-organellar locations **(Please see Appendix A, table A.1 on CD)**. The data set obtained using PROCLASS was filtered at the organellar and protein level in order to develop a data set suitable for TMDEPTH and the data mining process. At the organellar level, clusters of proteins corresponding to undefined membrane locations or/and multi-organellar locations were not included in the data set. Likewise, clusters of proteins belonging to the vacuole and the vesicle were not included in the data set as these organelles were considered as transient organelles whose main function is the storage and transport of proteins between organelles. At the protein level, both the structural annotation and sequence redundancy were analyzed. Membrane proteins whose structure was found to be composed of β-sheets forming a β-barrel structure were not included in the data set as TMDEPTH was implemented to detect associations of residues belonging to different α-helical structures located at a similar depth in the membrane. The estimation of the membrane thickness and depth values for the different amino acids is not applicable to β-barrel membrane proteins due to their very different structural properties. In order to remove β-barrel membrane proteins, Swiss-Prot like text files containing the term "porin" (used to describe β-barrel membrane proteins) were excluded from the data set. Based on two different empirical observations **(Chapter 6)**, TMDEPTH was implemented to consider transmembrane regions whose length is less than 14 residues long as false positives, so discarding these segments while retrieving

transmembrane information from the Swiss-Prot database. Although it is possible that those predicted segments are false positives, it is also possible that the topology prediction underestimated the length of the given segment or that the predicted segment corresponds not to a helical structure but to an unstructured loop or loop-helical structure that completely traverses the membrane. In order to minimize potential errors when calculating the matrices using TMDEPTH, those proteins containing at least one transmembrane region with less than 14 residues were excluded from the data set. Predicted membrane dipping loops **(Chapter 5)** were included in the transmembrane statement of corresponding Swiss-Prot like text files using the TMLOOP writer (only those membrane dipping loops found to be true positives). Therefore, improving the quality of the topological model found in the Swiss-Prot database. Sequence similarity was also analyzed within each cluster using CD-HIT (Li et al., 2001). CD-HIT is a clustering tool, which uses a "greedy" incremental algorithm (Holm and Sander, 1998) to cluster proteins above a certain threshold. Protein sequences are first sorted in order of decreasing length and the longest sequence is used as a representative sequence. The remaining sequences are then compared to the representative sequence based on a n-peptide comparison (from 2 to 5 residues peptides). If the number of identical n-peptides is higher than a user defined threshold an alignment is performed to confirm the sequence identity. If the sequence identity is above a given threshold sequences are clustered together. If the sequence identity (either at the n-peptide level or at the pairwise alignment level) is below the given threshold, the sequence becomes the representative sequence of a new cluster (if it is found not to cluster with other representative sequences belonging to other clusters created iteratively). CD-HIT also reports the representative sequences for each set and can therefore be used to develop representative sets. CD-HIT was set to report only the representative sequences of clusters sharing more than 90% identity based on pentapeptide comparison (n = 5, c = 0.9). By filtering the data set at 90% identity, sequence redundancy was avoided and possible bias was minimized when mining the obtained data set.

After filtering at the organellar level and the protein level, the assembled data set contained 895 membrane proteins classified by PROCLASS into 11 different organellar clusters **(table 7.1)**.

Organelles	N° of seq.
Inner mitochondrial membrane	205
Thylakoidal membrane	100
Endoplasmic Reticulum membrane	201
Golgi apparatus membrane	42
Peroxisomal membrane	27
Chloroplast membrane	24
Plasma membrane	219
Nuclear membrane	21
Outer mitochondrial membrane	12
Lysosomal membrane	16
Mitochondrial membrane	28

Table 7.1 Filtered data set obtained with PROCLASS. The data set was filtered at the organellar level and at the protein level. After the filtering process the data set suffered a reduction of 84%.

Data mining techniques tend to underestimate, or under-detect, small classes in unbalanced sets in order to increase the overall non-normalized accuracy of the prediction. During the data mining (data mining using five classes: secretory - including endoplasmic reticulum, Golgi apparatus, lysosome and peroxisome; mitochondria - including inner mitochondria membrane, outer mitochondria membrane and undefined mitochondrial membrane; and chloroplast – including chloroplast membrane and thylakoidal membrane; plasma membrane; nuclear membrane) process it was observed that the nuclear class size was significantly smaller than the remaining four classes. Following the statement made by Nair and colleagues (Nair and Rost, 2005): "More data with noise is better and less data with less noise", it was decided to manually include in the data set proteins whose nuclear localization has not yet been experimentally demonstrated. The protein families corresponding to the 21 filtered nuclear proteins retrieved using PROCLASS were analyzed and it was found that some of these families were specifically targeted to the nuclear membrane: the NDC1 family, the non-repetitive / WGA-negative nucleoporin family, the Sad1 interacting factor, the SAD/UND proteins, Lamin B receptors, ULP1-interacting protein and the nucleolar complex protein. Additionally, the Swiss-Prot and the Trembl database were searched for nurim proteins, which belong to a recently identified protein

family whose function has not been elucidated yet but it is known to specifically localize in the nuclear membrane (Rolls et al., 1999). No nurim protein was found in the Swiss-Prot database but eight proteins were obtained from the Trembl database. Because the Trembl database does not include topological information of the membrane proteins, TMHMM (Sonnhammer et al., 1998, Krogh et al., 2001) was used to predict the topology of the nurim protein retrieved from the Trembl database. The retrieved proteins belonging to the protein families described above were filtered at the protein level (both the structure and the sequence redundancy) as performed with the data set obtained with PROCLASS. After filtering the new nuclear proteins the size of the nuclear protein class could only be increased up to 40 non-redundant nuclear membrane proteins. The subcellular location clusters **(table 7.1)** belonging to different compartments of the same organelle were merged together into eight different classes **(table 7.2)**.

Organelles	N° of seq.
Mitochondria	245
Chloroplast	124
Endoplasmic Reticulum	201
Golgi apparatus	42
Peroxisome	27
Plasma membrane	219
Nuclear membrane	40
Lysosome	16

Table 7.2 Filtered data set obtained with PROCLASS including manually retrieved nuclear membrane proteins whose location has not been experimentally checked.

A separate data set of non plant eukaryotic membrane proteins was constructed using PROCLASS based on the same cross-linked term list used for the development of the previous data set **(Chapter 5)**. PROCLASS searched the annotation space of proteins contained in a local version of the Swiss-Prot database (release 50.6 of 5-9-2006), which contained proteins belonging to the *animalia* and *fungi* kingdom but not to the *plantae* kingdom. As with the data set described in **table 7.2**, retrieved proteins were filtered at the organellar and protein level based on the same constraints. **Table 7.3** describes the data set of non-plant eukaryotic polytopic membrane proteins clustered according to their subcellular location.

Organelles	N° of seq.
Mitochondria	202
Endoplasmic Reticulum	167
Golgi apparatus	37
Peroxisome	26
Plasma membrane	149
Nuclear membrane	40
Lysosome	16

Table 7.3. Filtered data set of non-Plant eukaryotic proteins obtained with PROCLASS.

TMDEPTH was used to calculate the corresponding matrices for each of the proteins contained in the different classes using the sequence and the topological information described in the local Swiss-Prot like text files. TMDEPTH was required to save the computed interhelical associations of residues located at a similar depth (percentage of amino acid participation in interhelical associations, the normalized 20x20 triangular matrix of interhelical associations and the normalized 3x3 triangular matrix of interhelical associations of clusters of biochemically equivalent residues) in the C4.5 format (Quinlan, 1993), which can be processed not only by C4.5 and its latest Windows version (See5) but also by the Weka platform (Witten and Frank, 2005). Therefore, the data set to be processed by Weka was composed of eight different classes, 914 and 637 data points (corresponding to the data set of eukaryotic membrane proteins and the data set of non-plant eukaryotic membrane proteins) and 236 attributes per data point (20 attributes correspond to the percentage of interhelical association participation for each residue, 210 attributes correspond to the normalized 20x20 triangular matrix of interhelical associations and 6 attributes correspond to the normalized 3x3 triangular matrix of interhelical associations of clusters of biochemically equivalent residues).

The obtained data sets were filtered using the supervised attribute selection filtering methods built within the Weka platform. This method is used to select attributes based on an evaluator (determines how attribute subsets are evaluated) and a search method. The default settings within Weka platform (Witten and Frank, 2005) were applied to this task. The evaluator was the CfsSubsetEval method, which evaluates the significance of a subset

of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them. This evaluator identifies locally predictive attributes and iteratively adds attributes with the highest correlation with the class (if there is not already an attribute in the subset with a higher correlation). On the other hand the BestFirst search method searches the space of attribute subsets by greedy hillclimbing augmented with a backtracking facility. The level of backtracking was controlled by the number of consecutive non-improving nodes, which by default was set to five. The different data mining methods and architectures (see below) were applied to both the non-filtered data sets and the filtered data sets to maximize the accuracy prediction of a given data mining tool.

7.2.2 Development of the data mining workflow

The Weka platform (Witten and Frank, 2005) was used to design and evaluate the different data mining analyses carried out. Two different data mining architectures were designed: i) the single-step data mining workflow and ii) the multi-step (or tree-based) data mining workflow. In the single-step data mining workflow, a single data mining method is used to predict the different classes in a single step whereas in the multi-step data mining workflow a tree-based workflow is designed where each node represents a particular data mining method trained to classify particular classes depending on the position of the corresponding node in the tree.

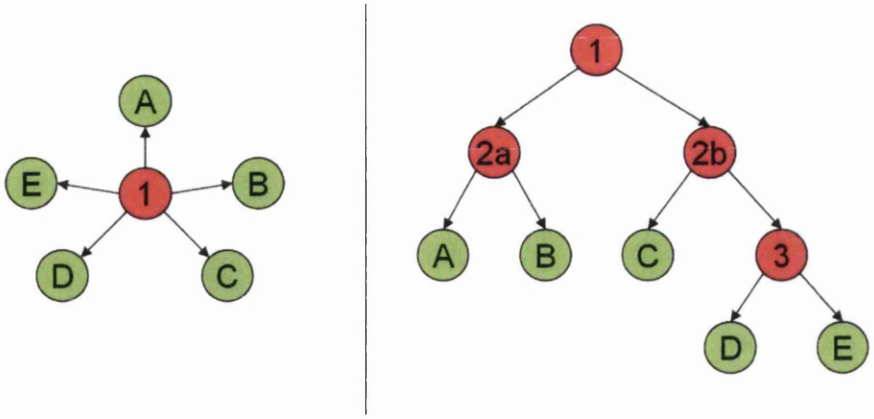

Figure 7.1. Example of the main data mining workflow architectures using a data set with five different classes (A-E, circles in green). Single-step data mining workflow (left), the single circle in red represents the single classifier method used to classify the different classes. Multi-step data mining workflow (right), four classifiers are used to classify the five different classes.

The Weka platform has the advantage of using a wide range of data mining techniques such as Bayesian methods, support vector machines, the closest neighbour or tree-based algorithms. No single data mining technique stands out in performance compared to other data mining techniques. In contrast, depending on the data set to be mined, a particular data mining method will perform better than the others, regardless of the complexity of the method. Therefore, both the single-step method and the multi-step method were tested using different data mining methods and different parameters **(table 7.3)** in order to maximize the predictive accuracy of the method. When performing multi-step data mining, different combinations, which also include the combination of the single-step and the multi-step architecture, were tested in order to find the tree-based architecture that best fits the given data set. Likewise, at each node of the different tree architectures designed, the different methods listed in **table 7.3** were compared to maximize the predictive accuracy of each tree. Selection of the most accurate data mining method at a given node within a tree-based set of classifiers was based upon comparison of the accuracy (Q), the normalized accuracy (nQ) and the Matthews correlation coefficient (MCC) or the Generalized coefficient (GC) of the different data mining methods. Highly selective data mining methods are required for the tree-base set of classifiers throughout the whole of the tree, as the percentage accuracy of the method is accumulative as the different nodes are reached along the tree. Therefore, if the very first node of the tree applies a data mining

method with high sensitivity but poor specificity the performance of the following nodes will be affected and subsequently the predictive accuracy of the tree will be decreased more severely by earlier nodes. Therefore, methods with fairly good sensitivity and high specificity are preferred over methods with high sensitivity and lower specificity.

Data mining technique	Parameters
Bayesian networks	Default
Naïve Bayesian	Default
Naïve Bayesian simple	Default
Logistic regression	Default
RBF Network	Default
KStar	Default
MultiBoostAB	Classifier: Random forest; Iterations: 30
J48	Default
Random forest	Default
Support vector machine (1)	Default
Support vector machine (2)	c = 30
Support vector machine (3)	c = 50
Support vector machine (4)	c = 1; exp = 5
Support vector machine (5)	c = 1;exp = 9
Support vector machine (6)	c = 50; exp = 5; feature space normalization = true
Support vector machine (7)	c = 50; exp = 9; feature space normalization = true
Support vector machine (8)	c = 50; exp = 5; feature space normalization = true; $\gamma = 0.001$
Support vector machine (9)	c = 50;exp = 5
Support vector machine (10)	c = 50; exp = 15; feature space normalization = true
Support vector machine (11)	c = 1; exp = 15; feature space normalization = true; $\gamma = 0.0001$

Table 7.4. List of the different data mining techniques applied during the data mining process using the Weka Knowledge Flow tool.

7.2.3 Evaluation

Evaluation was achieved by ten fold cross-validation. When evaluating data mining methods applied in a single-step fashion, the ten fold cross-validation of the single node corresponds to the evaluation of the entire method to predict the different subcellular locations. However, evaluation of a set of classifiers arranged in a tree-based fashion cannot be achieved by summing up independent evaluations of each node contained in a given tree. In order to evaluate a tree-based set of classifiers it was necessary to train the corresponding classifier at each node with the subset of the training set that had progressed to that point and then apply the trained tree-based set of classifiers to each data point in the test set. This process continues along the tree until reaching a leaf from where no further classification can be achieved.

The accuracy in predicting a particular class is defined by its sensitivity **(1.16)**, specificity **(1.17)** and geometric average **(1.18)**. The evaluation of the performance of the method is given by the accuracy **(1.19)**, the normalized accuracy **(1.20)** and the Matthews correlation coefficient for data sets with two classes **(1.21)** or the Generalized correlation for data sets with three or more classes **(1.22)**. These predictive values are compared between the different data mining techniques **(table 7.4)** in order to select the most accurate data mining technique for predicting a particular functional class.

7.3 Results and Discussion

7.3.1 Evolutionary relationships between organelles belonging to the secretory class and the nucleus

The assembled data set was classified using two different classification schemes. In the first classification scheme, eight classes were considered: endoplasmic reticulum, Golgi apparatus, lysosome, peroxisome, chloroplast, mitochondria, nucleus and plasma membrane; whereas in the second classification scheme, the subcellular organelles involved in the secretory pathway (endoplasmic reticulum, Golgi apparatus, lysosome and peroxisome) were clustered together following the classification scheme described in recent

papers (Nair and Rost, 2005, Pierleoni et al., 2006). Ten fold cross-validation of the single-step architecture showed that the predictive accuracy of the different data mining methods was higher if the given data set was classified into 5 categories (secretory, nucleus, plasma membrane, mitochondria and chloroplast) rather than 8 categories **(table 7.5)**. Likewise, **table 7.6** shows that members of the Golgi apparatus, lysosome and peroxisome tend to be predicted as members of the endoplasmic reticulum, possibly reflecting the evolutionary relationships between these organelles. Based on these results, and the fact that predictive architectures that mimic cellular sorting have been shown to improve the prediction of subcellular location (Nair and Rost, 2005), it was decided to cluster organelles involved in the secretory pathway and train specific data mining methods to further classify the members of the secretory class into their respective subcellular organelles. The accuracy in predicting the nucleus class is very low compared to the remaining classes regardless of the classification scheme applied **(table 7.8** and **table 7.9)**. The average confusion matrices corresponding to the single step architecture based on 8 and 5 class sets (**table 7.6** and **table 7.7** respectively) showed that nuclear membrane proteins tend to be classified as endoplasmic reticulum (37%) and secretory (51%). These results might also reflect an evolutionary relationship between the endoplasmic reticulum and the nucleus, however it is also possible that these results reflect the limitations of data mining methods to accurately predict classes with small size in unbalanced sets. The correlation values between sensitivity and size and between specificity and size was found to be 0.87 and 0.62 for the data set classified into 8 categories and 0.75 and 0.73 for the data set classified into 5 categories. These values indicate a dependency of the data mining methods upon class size.

	8 classes			5 classes		
Data mining method	**Q**	**nQ**	**GC**	**Q**	**nQ**	**GC**
Bayesian networks	46	30.5	0.3	48.4	41.6	0.37
Naïve Bayesian	48	38.3	0.37	53	45.7	0.4
Naïve Bayesian simple	-	-	-	-	-	-
Logistic regression	40	29.6	0.27	46.5	41.8	0.32
RBF Network	29	13.4	-	34	23.4	-
KStar	60	48	0.49	61.7	56.6	0.5
MultiBoostAB	63	42.3	0.49	70	57.6	0.57
J48	50	35.2	0.34	58	48.6	0.42
Random forest	59	39.1	0.46	66.2	54.4	0.53
Support vector machine (1)	59	37.6	0.42	64.7	53.7	0.51
Support vector machine (2)	53	41.1	0.39	56.8	48.9	0.43
Support vector machine (3)	53	41.6	0.39	56.8	48.9	0.43
Support vector machine (4)	60	42	0.43	64.8	54.9	0.5
Support vector machine (5)	55	37.2	0.38	59.3	50.6	0.45
Support vector machine (6)	68	44.1	0.52	72.6	59.7	0.6
Support vector machine (7)	70	49.3	0.56	74.1	60.8	0.61
Support vector machine (8)	70	49.3	0.56	74.1	60.8	0.61
Support vector machine (9)	60	42	0.43	64.8	54.9	0.5
Support vector machine (10)	62	39.6	0.48	68.6	55.7	0.57
Support vector machine (11)	59	32.3	-	66.6	52.9	-

Table 7.5. Comparison of two different models using the single-step architecture and ten fold cross-validation. The first model corresponds to columns 2-4 and is based on a data set composed by 8 classes: i) Endoplasmic reticulum, ii) Golgi, iii) Lysososme, iv) Peroxisome, v) Chloroplast, vi) Mitochondria, vii) Nucleus and viii) Plasma membrane. The second model corresponds to colums 5-7 and is based on a data set composed by 5 classes where the endoplasmic reticulum, Golgi, lysosome and peroxisome are merged to form the secretory class. The 5 class model reports higher levels of accuracy (Q), normalized accuracy (nQ) and generalized correlation (GC) for each data mining method used.

ER	Golgi	Lysosome	Peroxisome	Chloroplast	Mitochondrial	Nucleus	Plasma membrane	
54	6	1	2	3	14	3	18	ER
38	12	1	0	4	15	2	28	Golgi
13	8	31	1	4	17	1	26	Lysosome
25	4	1	23	7	28	1	10	Peroxisome
6	2	1	1	67	14	1	9	Chloroplast
9	4	1	3	5	66	1	12	Mitochondrial
37	5	3	2	4	26	6	18	Nucleus
14	5	1	1	3	10	1	66	Plasma membrane

Table 7.6 Average confusion matrix of the single-step architecture using the assembled data set classified into eight categories. The values listed in the matrix correspond to percentages. The highlighted cells indicate that 38%, 13% and 25% of the proteins belonging to the Golgi, lysosome, peroxisome respectively are being predicted as proteins belonging to the endoplasmic reticulum possibly reflecting the evolutionary relationship between these organelles.

Chloroplast	Secretory	Mitochondria	Nucleus	Plasma membrane	
67	11	14	1	7	Chloroplast
5	61	13	2	19	Secretory
6	18	64	1	11	Mitochondria
5	**51**	26	4	15	Nucleus
3	27	9	1	60	Plasma membrane

Table 7.7. Average confusion matrix (%) of the single-step model based on 5 classes and 21 nuclear membrane proteins. The most significant misclassification (highlighted cell) corresponded to the nuclear membrane proteins being predicted as membrane proteins of the secretory pathway.

Class	Sensitivity	Specificity	GAv
ER	53.99	55.07	54.08
Golgi	11.90	18.48	12.95
Lysosome	30.51	51.15	37.61
Peroxisome	23.09	52.55	31.70
Chloroplast	66.65	75.23	70.53
Mitochondria	65.64	67.22	65.35
Nucleus	5.60	15.92	8.66
Plasma membrane	65.67	60.63	62.86

Table 7.8. Average values of sensitivity, specificity and the geometric distance for each of the classes used in the 8-class model using the single-step architecture. The nucleus showed to be the class with the lowest predictive accuracy (highlighted cells).

Class	Sensitivity	Specificity	GAv
Chloroplast	70.11	72.28	70.76
Secretory	58.93	61.74	60.03
Mitochondrial	65.74	68.75	66.65
Nucleus	4.20	11.86	5.40
Plasma membrane	64.28	60.22	61.84

Table 7.9. Average values of sensitivity, specificity and the geometric distance for each of the classes used in the 5-class model using the single-step architecture. The nucleus showed to be the class with the lowest predictive accuracy (highlighted cells).

In order to minimize the observed bias towards larger classes new nuclear membrane proteins were retrieved from the Swiss-Prot database (as described in the method section). Unfortunately, the size of the nuclear class could only be increased to up to 40 sequences, which was not considered sufficient to balance the data set. In order to further evaluate the possible evolutionary relationships between the endoplasmic reticulum and the nucleus, two different data mining analyses based on the single step architecture were performed. The first analysis **(table 7.10)**, is similar to that shown in **table 7.7**, but using 40 nuclear

sequences rather than 21. Although the nuclear class remains the smallest class in the unbalanced set, the size of the class has increased by 2-fold and an improvement in predicting this class was expected. However, the percentage of nuclear proteins predicted as secretory proteins **(table 7.10)** was found to be similar to that found **(table 7.7)** when using a nuclear set composed by 21 sequences (50% and 51% respectively). In the second data mining analysis, the data set was composed by five classes but each class was filtered using CD-HIT (at the sequence similarity level) with different strength in order to create a more balanced data set. Therefore, larger classes were filtered more stringently than smaller classes. The chloroplast, peroxisome, Golgi, lysosome, outer mitochondria membrane, unspecified mitochondria membrane and nucleus were filtered using a minimum sequence identity of 90%, whereas the thylakoidal membrane was filtered at a minimum sequence identity of 50% and the endoplasmic reticulum and inner mitochondrial membrane were filtered at a minimum sequence identity of 40%. The newly filtered class sizes were reduced to 216, 113, 126, 49 and 40 for the secretory, mitochondrial, plasma membrane, chloroplast and nuclear classes respectively. The chloroplast and nuclear class are of roughly similar size (49 and 40 proteins respectively) and the percentage of proteins predicted as secretory should be similar for these two classes if no evolutionary relationships were involved. The average confusion matrix of this analysis is shown in **table 7.11**, which shows that the percentage of nuclear proteins predicted as secretory proteins is as high as 60% whereas 38% of the chloroplast membrane proteins are predicted as secretory. Comparison of the results listed in **tables 7.6, 7.7, 7.8 and 7.9** indicate an evolutionary relationship between the nucleus and the secretory pathway and more specifically between the nucleus and the endoplasmic reticulum **(table 7.6)**. These results are in accordance with the arrangement of organelles within the cell as it is well known that the outer nuclear membrane is contiguous with the endoplasmic reticulum. Based on the evolutionary relationships between lysosome, peroxisome, Golgi apparatus, endoplasmic reticulum and nucleus, a tree-based architecture was designed, which distinguishes first between four classes, namely chloroplast, mitochondria, plasma membrane and secretory, and further sub-classifies the secretory class into peroxisome, lysosome, golgi apparatus, endoplasmic reticulum and nucleus **(figure 7.2)**.

Chloroplast	Secretory	Mitochondria	Nucleus	Plasma membrane	
66	8	15	1	9	Chloroplast
4	59	14	3	20	Secretory
6	14	67	2	11	Mitochondria
2	**50**	21	14	13	Nucleus
4	23	9	1	63	Plasma membrane

Table 7.10. Average confusion matrix (%) of the single-step model based on 5 classes and 40 nuclear membrane proteins. The matrix shows that 50% of the nuclear membrane proteins tend to be predicted as membrane proteins belonging to the secretory class (highlighted cell.

Chloroplast	Secretory	Mitochondria	Nucleus	Plasma membrane	
25	**38**	19	2	16	Chloroplast
4	66	9	4	17	Secretory
7	33	44	3	13	Mitochondria
4	**60**	9	5	12	Nucleus
4	45	6	2	43	Plasma membrane

Table 7.11. Average confusion matrix (%) of the single-step model based on 5 classes and 70 nuclear membrane proteins. The different classes and subclasses of the given data set have been filtered at different stringency using CD-HIT in order to obtain a more balanced set and analyze the relationship between the nucleus and the secretory class. The highlighted cells correspond to the percentage of chloroplast and nuclear membrane proteins classified as membrane proteins belonging to the secretory class. Although both classes have similar size, the misclassifications concerning the nucleus class is nearly twice the percentage of chloroplast membrane proteins being classified as secretory membrane proteins.

7.3.2 Development of the predictive architecture

The results described above were used to design the general architecture of a predictive tool that combines multiple data mining methods to predict the subcellular location of membrane proteins. The evolutionary relationships between different organelles showed that the endoplasmic reticulum, Golgi apparatus, nucleus, peroxisome and lysosome were to be predicted together as secretory in the first level of the predictive architecture along with chloroplast, mitochondria and plasma membrane. These four classes were classified in a single-step mode using a wide range of data mining methods **(table 7.4)**. The ten fold cross-validation results **(table 7.12)** showed that the support vector machine was the most accurate technique for the given data set. Among the different sets of parameters used in SVM, four sets (highlighted in **table 7.12**) were found to maximize the evaluation parameters Q, nQ and GC, where two of these SVM used the attribute selection filter prior to the mining stage. The predictive accuracy for each class **(table 7.13)** showed

that the SVM(10)-attribute selection contained the highest predictive accuracy for the plasma membrane class and it was therefore selected as the data mining method to be used at the first level of the predictive tool.

Level 1 (1st and 2nd variant)	Attribute selection			Without Attribute selection		
Data mining method	**Q**	**nQ**	**GC**	**Q**	**nQ**	**GC**
Bayesian networks	59.1	59.5	0.49	49.6	52.8	0.42
Naïve Bayesian	53.4	54.9	0.44	56.5	59.6	0.5
Naïve Bayesian simple	52.6	54.7	0.44	-	-	-
Logistic regression	62.7	62	0.52	51.3	52.4	0.37
RBF Network	59.7	59.8	0.49	38.3	30.2	-
KStar	67.2	68.7	0.59	63.7	67.1	0.57
MultiBoostAB	73.4	71.8	0.66	71.1	69.9	0.65
J48	60.1	60.2	0.47	57.7	57.7	0.45
Random forest	68.8	67.5	0.6	64.6	64.1	0.56
Support vector machine (1)	62.5	60.4	0.51	66.5	66.3	0.57
Support vector machine (2)	62	61.6	0.51	61.8	63.4	0.51
Support vector machine (3)	61.8	61.7	0.51	60.7	62.7	0.49
Support vector machine (4)	73	73	0.65	66.3	67	0.57
Support vector machine (5)	69.4	69.9	0.6	63.3	64	0.54
Support vector machine (6)	75.2	75.3	0.69	75.2	73.5	0.69
Support vector machine (7)	73.5	74	0.66	76.3	76	0.7
Support vector machine (8)	73.5	74	0.66	76.3	76	0.7
Support vector machine (9)	67.4	68.4	0.58	66.3	67	0.57
Support vector machine (10)	**76.1**	**75.4**	**0.7**	66.8	62.9	0.62
Support vector machine (11)	76	74.8	0.7	64.4	60	0.6

Table 7.12. Ten fold cross-validation results of the different data mining methods used to classify membrane proteins into secretory, mitochondria, chloroplast and plasma membrane in a single-step mode. The highlighted cells correspond to the methods that maximize the predictive accuracy.

		Sensitivity	Specificity	GAv
SVM(10) Att sel	Chloroplast	73.4	91.0	81.7
	Secretory	78.8	73.2	76.0
	Mitochondria	79.2	72.4	75.7
	Plasma membrane	70.3	79.0	74.5
SVM(11) Att sel	Chloroplast	70.2	94.6	81.5
	Secretory	79.8	72.4	76.0
	Mitochondria	80.0	72.1	75.9
	Plasma membrane	69.4	79.6	74.3
SVM(7)	Chloroplast	75.8	89.5	82.4
	Secretory	76.4	75.9	76.1
	Mitochondria	82.0	71.5	76.6
	Plasma membrane	69.9	76.5	73.1
SVM(8)	Chloroplast	75.8	89.5	82.4
	Secretory	76.4	75.9	76.1
	Mitochondria	82.0	71.5	76.6
	Plasma membrane	69.9	76.5	73.1

Table 7.13. Predictive accuracy for each class after evaluation by ten fold cross-validation. Only the methods that showed maximal values of overall accuracy in predicting the chloroplast, mitochondria, secretory and plasma membrane are included. Plasma membrane is the class with the lowest predictive accuracy (highlighted cells). Among the four data mining methods SVM(10) with attribute selection showed the highest accuracy in predicting this class.

The following levels were designed to subclassify the proteins predicted as secretory into peroxisome, lysosome, Golgi apparatus, endoplasmic reticulum and nucleus. **Table 7.14** shows the average confusion matrix obtained using different data mining methods **(table 7.13)** with and without prior attribute selection, and this matrix describes an overprediction of endoplamic reticulum membrane proteins. It is believed that two different factors play a major role in the overprediction of the endoplasmic reticulum, the first factor is the evolutionary relationships of these organelles and the fact that the majority of these proteins are first synthesized in the membrane of the endoplasmic reticulum and then transported to their appropriate location. The second factor is the tendency of different data mining methods to overpredict the largest class within an unbalanced data set. As the single-step mode did not show that it could accurately distinguish between the different classes involved in the secretory pathway, a multi-step architecture mimicking the process of subcellular sorting was designed to further classify the secretory class. At this stage, different tree structures and combinations were tested in order to maximize the predictive accuracy of the method at each level. Two tree variants were found to maximize the predictive accuracy of the method. In the first variant **(figure 7.2)**, the second level was

designed to distinguish between Peroxisome/Lysosome and Golgi/ER/Nucleus, the level 3a distinguishes between Peroxisome and Lysosome and the level 3b between Golgi and ER/Nucleus. Finally, in level 4 ER and Nucleus were distinguished. The second tree variant **(figure 7.3)** distinguishes between Peroxisome and Lysosome/Golgi/ER/Nucleus at level 2, in level 3 the Lysosome is predicted and then level 4 and 5 corresponds to level 3b and 4 of the first tree variant.

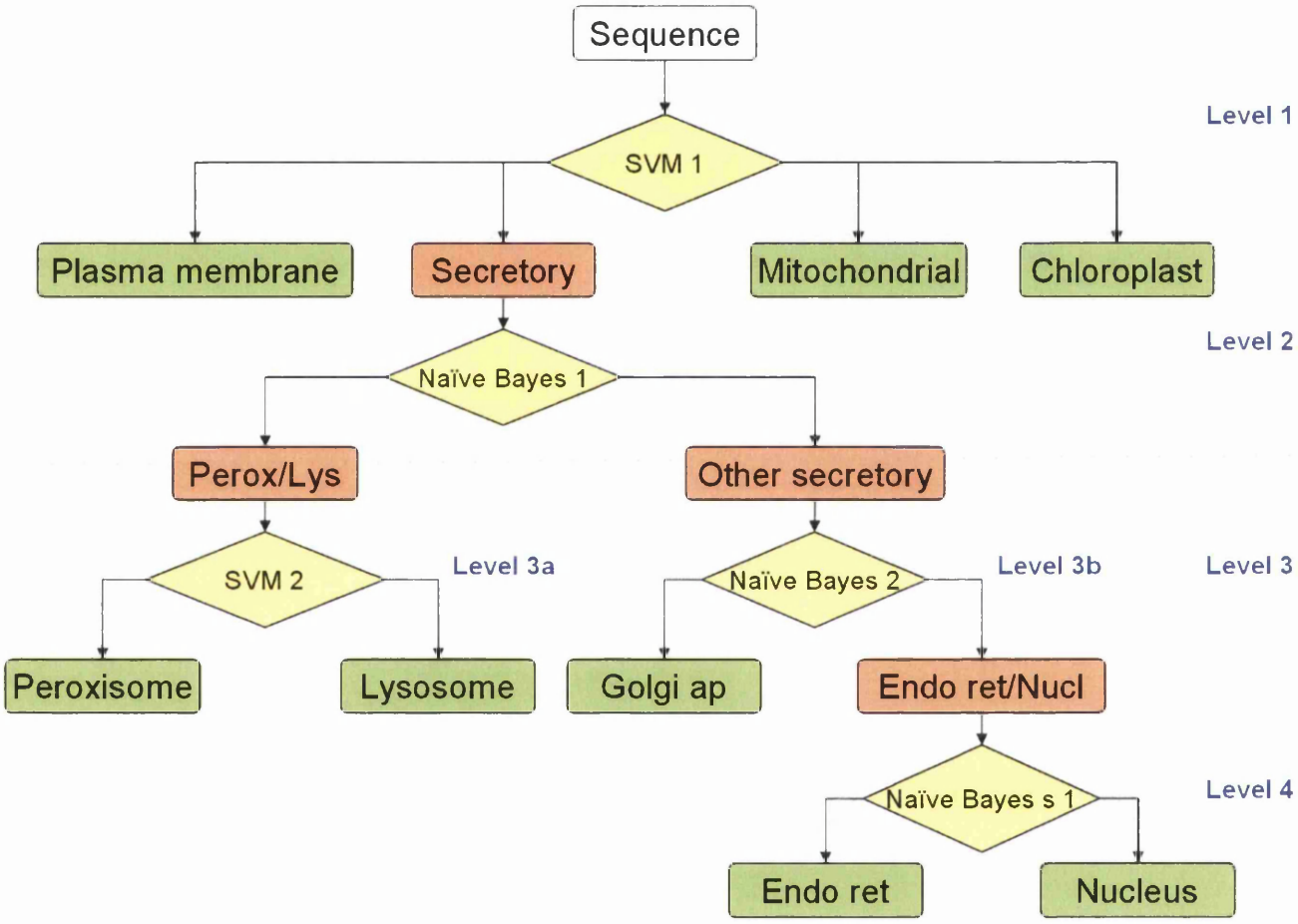

Figure 7.2 First variant of the predictive tree architecture

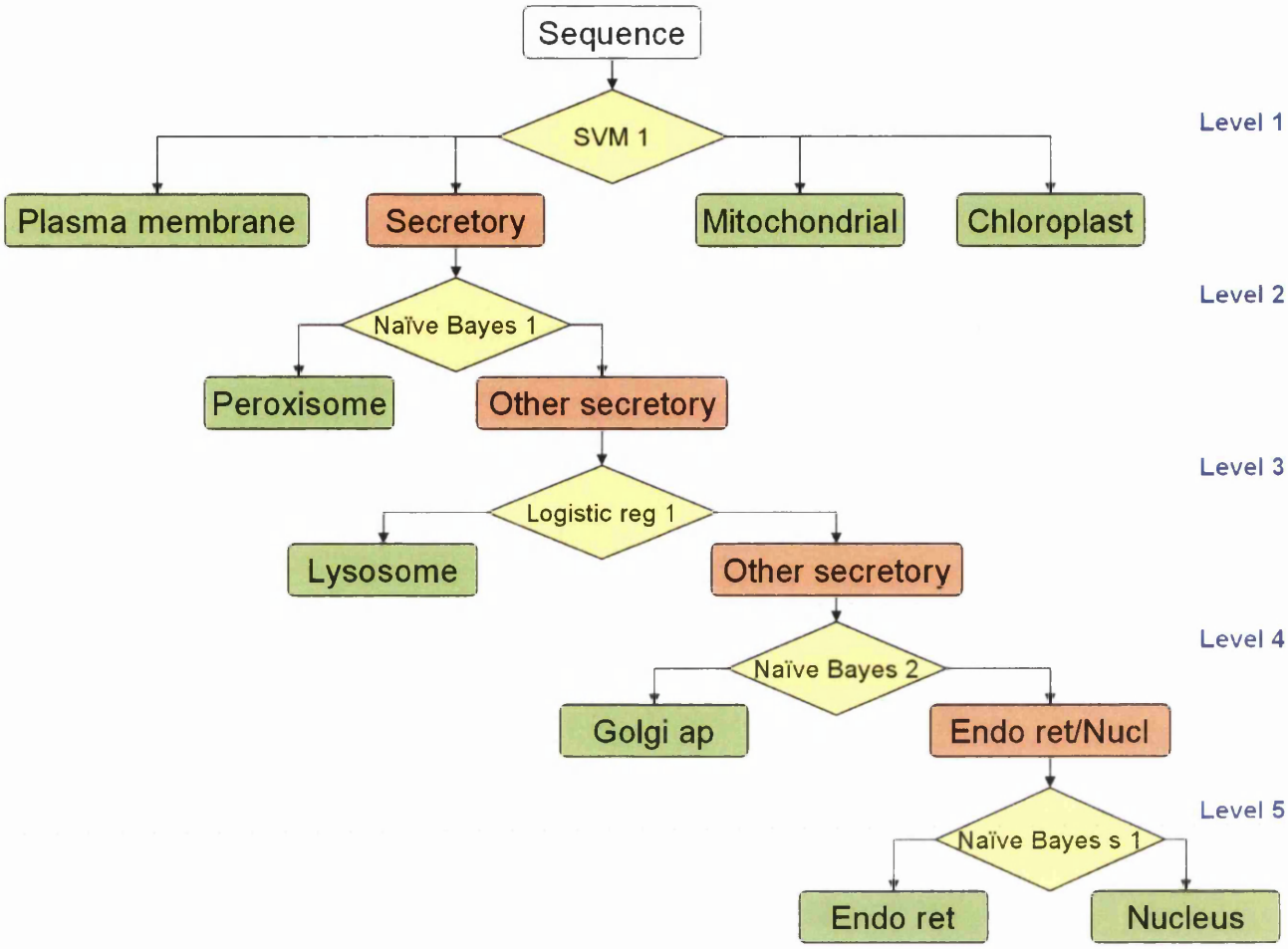

Figure 7.3. Second variant of the predictive architecture

In the first variant, the ten fold cross-validation results at level two **(table 7.15)** showed that the naïve Bayesian-Attribute selection and the SVM(10)-Attribute selection method reported maximum accuracy values. The prediction accuracy for both classes showed that the naïve Bayesian-Attribute selection method **(table 7.16)** was more consistent than the SVM(10)-Attribute selection method as with the latter only 47.6% of the Peroxisome/Lysosome class could be correctly predicted. Therefore the Naïve Bayesian method was chosen as the predictive method for level two of the first variant. At level 3a, data mining methods (with prior attribute selection) reported maximal accuracy values in distinguishing between Peroxisome and Lysosome: the logistic regression, the RBF network, the SVM(7) and the SVM(8) **(table 7.17)**. Further analysis showed that the

logistic regression **(table 7.18)** predicted the Peroxisome class with the lowest sensitivity and was therefore discarded as a candidate predictive method to be used. The RBF network, SVM(7) and the SVM(8) reported identical values for both classes. SVM(8) was chosen as the predictive method to distinguish between the Peroxisome and Lysosome classes.

ER	Nucleus	Golgi app	Peroxisome	Lysosome	
162	12	17	5	5	ER
28	7	2	2	2	Nucleus
29	2	8	2	2	Golgi app
14	1	2	9	1	Peroxisome
8	0	1	0	6	Lysosome

Table 7.14. Average confusion matrix of the single-step mode using different data mining techniques. Most of the membrane proteins tend to be predicted as proteins belonging to the endoplasmic reticulum. Considering the size of each subset this is probably due to a overprediction of the largest class in an unbalanced data set.

Level 2 (1st variant)	**Attribute selection**			**Without Attribute selection**		
Data mining method	**Q**	**nQ**	**MCC**	**Q**	**nQ**	**MCC**
Bayesian networks	82.5	67.2	0.31	76.4	62.7	0.21
Naïve Bayesian	**70.2**	**69.1**	**0.27**	79.1	67.3	0.29
Naïve Bayesian simple	-	-	-	-	-	-
Logistic regression	88	64.5	0.38	79.8	62.7	0.23
RBF Network	86.8	53.9	0.17	86.2	49.6	-0.03
KStar	82.5	64.3	0.27	85.9	62.3	0.29
MultiBoostAB	86.2	66.4	0.35	89.6	62.4	0.43
J48	86.2	64.4	0.33	79.1	53.5	0.07
Random forest	86.5	67.6	0.37	89.6	62.4	0.43
Support vector machine (1)	85.7	52.2	0.15	84.7	57.6	0.19
Support vector machine (2)	86.7	57.7	0.3	82.5	63.3	0.26
Support vector machine (3)	87.1	58.9	0.33	82.5	63.3	0.26
Support vector machine (4)	85.3	52	0.12	87.1	62	0.32
Support vector machine (5)	84.6	51.6	0.08	85.6	57.2	0.2
Support vector machine (6)	86	68.1	0.4	88.7	57	0.35
Support vector machine (7)	86	65.2	0.36	89.9	61.6	0.46
Support vector machine (8)	86	65.2	0.36	89.9	61.6	0.46
Support vector machine (9)	86.7	62.6	0.35	87.1	62	0.32
Support vector machine (10)	88.1	71.4	0.48	88.7	57	0.35
Support vector machine (11)	86.7	56.7	0.29	86.8	50	-

Table 7.15 Ten fold cross-validation results of different data mining methods used to distinguish between Peroxisome/Lysosome and Golgi/ER/Nucleus. The highlighted cells correspond to the data mining methods that maximized the prediction at this level.

		Sensitivity	Specificity	GAv
Naïve b Att sel	Perox/Lys	67.4	25.9	41.8
	Other	70.7	93.5	81.3
SVM(10) Att sel	Perox/Lys	47.6	62.5	54.6
	Other	95.1	91.3	93.2

Table 7.16. Predictive accuracy for each class distinguished at level 2 (the first tree variant). The best two predictive methods at this level were compared. The Perox/Lys class showed the lowest predictive accuracy in both data mining analyses (highlighted cells). The SVM(10)-attribute selection method only predicted 47.6% of the Perox/Lys class and was therefore discarded as a predictive method to be used at this level.

Level 3a (1st variant)	**Attribute selection**			**Without Attribute selection**		
Data mining method	**Q**	**nQ**	**MCC**	**Q**	**nQ**	**MCC**
Bayesian networks	83.7	83.2	0.66	76.7	72.6	0.49
Naïve Bayesian	90.7	91.3	0.81	83.7	81.9	0.65
Naïve Bayesian simple	-	-	-	-	-	-
Logistic regression	93	94.4	0.87	79.1	83.3	0.65
RBF Network	93.0	93.2	0.85	58.1	47.6	-0.08
KStar	86.0	86.3	0.71	69.8	73.4	0.46
MultiBoostAB	88.4	88.2	0.76	81.4	76.3	0.60
J48	81.4	80.1	0.60	60.5	58.3	0.16
Random forest	88.4	88.2	0.76	81.4	76.3	0.60
Support vector machine (1)	90.7	88.8	0.80	76.7	78.9	0.56
Support vector machine (2)	88.4	89.5	0.77	76.7	78.9	0.56
Support vector machine (3)	88.4	89.5	0.77	76.7	78.9	0.56
Support vector machine (4)	79.1	74.4	0.54	74.4	78.4	0.56
Support vector machine (5)	72.1	65.0	0.37	46.5	56.1	0.17
Support vector machine (6)	88.4	86.9	0.75	79.1	71.9	0.57
Support vector machine (7)	93.0	93.2	0.85	79.1	71.9	0.57
Support vector machine (8)	**93.0**	**93.2**	**0.85**	79.1	71.9	0.57
Support vector machine (9)	79.1	79.5	0.57	74.4	78.4	0.56
Support vector machine (10)	83.7	78.1	0.67	76.7	68.8	0.52
Support vector machine (11)	81.4	75.0	0.62	74.4	65.6	0.47

Table 7.17. Ten fold cross-validation results of the different data mining methods used to classify membrane proteins into Peroxisome and Lysosome (first tree variant). The highlighted cells correspond to the methods that maximize the predictive accuracy.

		Sensitivity	Specificity	GAv
Log Reg Att sel	Peroxisome	88.9	100.0	94.3
	Lysosome	100.0	84.2	91.8
RBF net Att sel	Peroxisome	92.6	96.2	94.4
	Lysosome	93.8	88.2	91.0
SVM(7) Att sel	Peroxisome	92.6	96.2	94.4
	Lysosome	93.8	88.2	91.0
SVM(8) Att sel	Peroxisome	92.6	96.2	94.4
	Lysosome	93.8	88.2	91.0

Table 7.18. Predictive accuracy for each class distinguished at the level 3a (first tree variant). The best four predictive methods at this level were compared. The logistic regression-attribute selection method predicted the peroxisome class with the lowest accuracy whereas the remaining three methods were more consistent. Because the RBF network, SVM(7) and SVM(8) were found to report identical predictive methods, SVM(8) was selected randomly as the predictive method to be used at this level.

In the second variant, the ten fold cross-validation results at level two **(table 7.19)** showed that the Naïve Bayesian-Attribute selection method was the most accurate predictive method to distinguish between Peroxisome and other secretory organelles (lysosome, Golgi apparatus, endoplasmic reticulum and nucleus). These results show the limitations of evaluating a predictive method based on the non-normalized accuracy Q. The unbalanced set classified at this level (27 peroxisome proteins and 299 proteins belonging to other secretory organelles) often causes the larger class to be over-predicted by different data mining methods. In the "worst case scenario" no proteins would be predicted as peroxisome proteins but as the negative class. Given the definition of Q and nQ (**1.19** and **1.20** respectively), the Q value of the method of this hypothetical case would be 91.7% whereas the nQ value would be 50%. Therefore, the Q value is not a good parameter to be considered while considering the most accurate method in light of the unbalanced data set. The same limitation is observed during the ten fold cross-validation of the different data mining methods used at level 3 of the second tree variant **(table 7.20)**. Data mining analyses using this unbalanced set (16 lysosome proteins and 283 proteins belonging to the Golgi apparatus, endoplasmic reticulum and nucleus) showed a tendency to overestimate the larger class. Interestingly, the method selected as the predictive tool at this level, the logistic regression method, was found to have the highest nQ score but not the highest MCC value. The method with the highest MCC score was the MultiboostAB method, which correctly predicted 50% of the lysosome proteins and 99.6% of the negative class (golgi/ER/nucleus). On the other hand, the logistic regression method correctly predicted

75% of the lysosome proteins and 92.9% of the negative class and was found to be a more consistent method. A similar situation was found with the different data mining methods possessing higher MCC values than the logistic regression method.

Level 2(2nd variant)	Attribute selection			Without Attribute selection		
Data mining method	**Q**	**nQ**	**MCC**	**Q**	**nQ**	**MCC**
Bayesian networks	89.6	65.7	0.31	85	68.2	0.28
Naïve Bayesian	**91.7**	**78.6**	**0.52**	85.3	71.8	0.33
Naïve Bayesian simple	91.4	76.8	0.49	-	-	-
Logistic regression	92.6	69	0.45	87.1	61	0.21
RBF Network	92.6	72.4	0.48	91.7	50	-
KStar	92.3	62.1	0.36	93.6	64.5	0.46
MultiBoostAB	92.9	70.9	0.48	94.2	66.5	0.53
J48	91.4	59.9	0.28	88	54.7	0.11
Random forest	92	67	0.4	93.9	64.6	0.49
Support vector machine (1)	91.4	49.8	-0.02	93.6	69.5	0.5
Support vector machine (2)	93.6	67.8	0.49	91.7	63.5	0.34
Support vector machine (3)	93.6	67.8	0.49	91.7	63.5	0.34
Support vector machine (4)	92.3	63.8	0.38	93.6	62.8	0.46
Support vector machine (5)	92.6	67.3	0.43	92	55.2	0.23
Support vector machine (6)	93.9	68	0.51	92.3	53.7	0.26
Support vector machine (7)	92.9	70.9	0.48	93.9	63	0.49
Support vector machine (8)	92.9	70.9	0.48	93.9	63	0.49
Support vector machine (9)	91.7	68.5	0.41	93.6	62.8	0.46
Support vector machine (10)	92	58.6	0.29	92.3	53.7	0.26
Support vector machine (11)	92.3	53.7	0.26	91.7	50	-

Table 7.19. Ten fold cross-validation results of the different data mining methods used to distinguish membrane proteins from Peroxisome and Lysosome/Golgi/ER/Nucleus (second tree variant). The highlighted cells correspond to the methods that maximize the predictive accuracy.

Level 3(2nd variant)	Attribute selection			Without Attribute selection		
Data mining method	**Q**	**nQ**	**MCC**	**Q**	**nQ**	**MCC**
Bayesian networks	94.3	49.8	-0.01	94.3	49.8	-0.01
Naïve Bayesian	71.9	79.3	0.28	93.6	70.1	0.39
Naïve Bayesian simple	-	-	-	-	-	-
Logistic regression	95	59	0.32	**92**	**84**	**0.49**
RBF Network	95	53.1	0.24	94.6	50	-
KStar	96	68.4	0.51	93.6	76	0.46
MultiBoostAB	97	74.8	0.65	96.7	68.8	0.6
J48	93.3	49.3	-0.03	94.6	61.8	0.33
Random forest	96.3	68.6	0.55	95	53.1	0.24
Support vector machine (1)	94.6	50	-	93.6	58.3	0.22
Support vector machine (2)	94.3	49.8	-0.01	93.3	78.8	0.48
Support vector machine (3)	94.3	49.8	-0.01	93.3	78.8	0.48
Support vector machine (4)	95.3	56.3	0.35	93.3	67	0.34
Support vector machine (5)	95.3	56.3	0.35	93.6	67.2	0.35
Support vector machine (6)	95.7	59.4	0.42	96.7	68.8	0.6
Support vector machine (7)	95.7	59.4	0.42	96.7	68.8	0.6
Support vector machine (8)	95.7	59.4	0.42	96.7	68.8	0.6
Support vector machine (9)	95.3	56.3	0.35	93.3	67	0.34
Support vector machine (10)	95.7	59.4	0.42	96.3	65.6	0.55
Support vector machine (11)	95	53.1	0.24	94.6	50	-

Table 7.20. Ten fold cross-validation results of the different data mining methods used to distinguish membrane proteins from Lysosome and Golgi/ER/Nucleus (second tree variant). The highlighted cells correspond to the method that maximizes the predictive accuracy.

As explained above, both tree-variants converge once peroxisome and lysosome proteins have been distinguished. The next level (level 3b for variant one and level 4 for variant two) distinguishes between Golgi apparatus and ER/Nucleus class. As in previous levels, different combinations were evaluated and it was found that distinguishing between Golgi apparatus and ER/Nucleus class achieved the highest accuracy. The ten fold cross-validation results for the different data mining methods evaluated **(table 7.21)** showed that the Naïve Bayesian-Attribute selection, Naïve Bayesian simple-Attribute selection, Logistic regression-Attribute selection and Naïve Bayesian method reported the highest predictive scores of nQ and MCC. Further analysis of the evaluation of these four data mining methods **(table 7.22)** showed that the Naïve Bayesian method (without prior attribute selection) was the most consistent method as the sensitivity value for the worst predicted class was still higher than in the other methods. The final level (corresponding to level 4 in variant one and level 5 in variant two) distinguishes between the endoplasmic reticulum

and the nucleus. The ten fold cross-validation results **(table 7.23)** clearly show the Naïve Bayesian simple-Attribute selection method as the most accurate method to distinguish between these two classes.

Level 3b (1st variant) Level 4 (2nd variant)	**Attribute selection**			**Without Attribute selection**		
Data mining method	**Q**	**nQ**	**MCC**	**Q**	**nQ**	**MCC**
Bayesian networks	82.0	53.0	0.09	82.0	53.0	0.09
Naïve Bayesian	62.9	72.3	0.32	**67.8**	**66.4**	**0.24**
Naïve Bayesian simple	63.8	73.2	0.33	-	-	-
Logistic regression	84.8	62.6	0.30	64.3	50.5	0.01
RBF Network	85.5	52.2	0.15	85.2	50.0	-
KStar	78.1	61.6	0.21	80.2	63.8	0.26
MultiBoostAB	84.8	58.6	0.25	84.1	51.3	0.06
J48	83.4	60.8	0.25	78.1	57.6	0.15
Random forest	84.8	54.7	0.17	83.7	50.2	0.01
Support vector machine (1)	85.2	50.0	-	79.5	49.6	-0.01
Support vector machine (2)	85.2	50.0	-	70.7	53.3	0.06
Support vector machine (3)	84.8	49.8	-0.02	70.7	53.3	0.06
Support vector machine (4)	86.2	53.6	0.25	79.9	57.7	0.16
Support vector machine (5)	85.5	52.2	0.15	79.2	57.3	0.15
Support vector machine (6)	81.3	63.4	0.27	84.8	50.8	0.05
Support vector machine (7)	80.9	64.2	0.27	83.7	54.1	0.13
Support vector machine (8)	80.9	64.2	0.27	83.7	54.1	0.13
Support vector machine (9)	86.2	56.5	0.26	79.9	57.7	0.16
Support vector machine (10)	80.2	57.9	0.17	85.2	51.0	0.08
Support vector machine (11)	85.5	52.2	0.15	85.2	50.0	-

Table 7.21. Ten fold cross-validation results of the different data mining methods used to distinguish membrane proteins from Golgi and ER/Nucleus (first and second tree variant). The highlighted cells correspond to the methods that maximize the predictive accuracy.

		Sensitivity	**Specificity**	**GAv**
Naïve b Att sel	ER/Nucleus	58.9	95.9	75.2
	Golgi	85.7	26.7	47.8
Naïve b s Att sel	ER/Nucleus	59.9	96.3	76.0
	Golgi	86.5	26.9	48.2
Log reg Att sel	ER/Nucleus	94.2	88.7	91.4
	Golgi	31.0	48.1	38.6
Naïve b	ER/Nucleus	68.5	91.7	79.2
	Golgi	64.3	26.2	41.1

Table 7.22. Predictive accuracy for each class distinguished at level 3b (first tree variant) and level 4 (second tree variant). The best four predictive methods at this level were compared. The highlighted cells correspond to the class predicted with the lowest sensitivity, the naïve Bayesian method reported the higher sensitivity among the highlighted classes and was therefore considered to be the more consistent method.

Level 4 (1st variant) Level 5 (2nd variant)	Attribute selection			Without Attribute selection		
Data mining method	**Q**	**nQ**	**MCC**	**Q**	**nQ**	**MCC**
Bayesian networks	78.4	54.0	0.10	76.8	54.0	0.09
Naïve Bayesian	78.4	79.1	0.47	79.3	61.5	0.24
Naïve Bayesian simple	**80.3**	**81.9**	**0.51**	-	-	-
Logistic regression	85.5	67.3	0.41	69.7	54.8	0.08
RBF Network	84.2	65.5	0.36	83.4	50.0	-
KStar	79.7	70.8	0.37	82.6	65.5	0.33
MultiBoostAB	85.9	67.5	0.42	84.6	54.8	0.25
J48	82.2	71.3	0.40	72.6	54.5	0.08
Random forest	85.1	65.0	0.37	84.2	53.5	0.20
Support vector machine (1)	85.5	56.3	0.33	80.9	55.5	0.15
Support vector machine (2)	85.9	61.5	0.37	71.8	52.0	0.04
Support vector machine (3)	86.3	62.8	0.39	71.8	52.0	0.04
Support vector machine (4)	85.9	57.5	0.36	79.3	53.5	0.09
Support vector machine (5)	84.6	55.8	0.25	80.9	53.5	0.11
Support vector machine (6)	78.0	65.8	0.29	83.0	49.8	-0.03
Support vector machine (7)	79.3	67.5	0.32	83.8	53.3	0.17
Support vector machine (8)	79.3	67.5	0.32	83.8	53.3	0.17
Support vector machine (9)	85.9	61.5	0.37	79.3	53.5	0.09
Support vector machine (10)	73.4	64.1	0.24	83.4	50.0	-
Support vector machine (11)	85.1	57.0	0.29	83.4	50.0	-

Table 7.23. Ten fold cross-validation results of the different data mining methods used to distinguish membrane proteins from endoplasmic reticulum and nucleus (first and second tree variant). The highlighted cells correspond to the method that maximizes the predictive accuracy.

7.3.3 Evaluation of the tree-based set of classifiers

During the development of the different variants of the tree-based set of classifiers, at each node the different data mining methods were evaluated by ten fold cross-validation independently of previous evaluations carried out at earlier nodes. Although bringing together the evaluations carried out at each node might give a rough estimation of the overall predictive accuracy of a given variant of the tree (based on probabilities), it was considered more accurate to evaluate the whole of the variant tree by ten fold cross-validation considering the set of classifiers as a single predictive algorithm. Following this principle, the sequences were tested by the tree until reaching a leaf. In this way, the variants of the two tree-based set of classifiers were evaluated by ten fold cross-validation (**table 7.24** and **table 7.25** respectively). Both variants showed similar accuracy values, however the second variant showed a slight improvement over the first variant as the Q

value improved from 57.7 to 60 and the nQ value improved from 45 to 46.2. The first variant demonstrated a higher sensitivity in predicting the peroxisome and lysosome class but with a lower specificity for those two classes. On the other hand, the prediction sensitivity for the Golgi apparatus and the endoplasmic reticulum class improved in the second variant and only the endoplasmic reticulum class showed a lower specificity. The nucleus class prediction accuracy was similar for each tree variant with the exception of a slight improvement in specificity in the second variant. When both the sensitivity and specificity are brought together using the geometric average GAv, the second variant showed higher values than the first variant.

		Sensitivity	Specificity	GAv	Q	nQ	GC
1	Chloroplast	72.6	92.8	82.1	75.5	74.8	0.7
	Mitochondria	79.2	71.9	75.4			
	Plasma membrane	69.9	78.9	74.2			
	Secretory	77.6	71.7	74.6			
2	Chloroplast	72.6	92.8	82.1	65.9	61.5	0.6
	Mitochondria	79.2	71.9	75.4			
	Plasma membrane	69.9	78.9	74.2			
	Perox-Lys	32.6	10.1	18.1			
	Other Secret	53.4	70.6	61.4			
3	Chloroplast	72.6	92.8	82.1	60.8	49.9	0.5
	Mitochondria	79.2	71.9	75.4			
	Plasma membrane	69.9	78.9	74.2			
	Peroxisome	22.2	7.1	12.5			
	Lysosome	43.8	13.0	23.8			
	Golgi app	21.4	12.2	16.1			
	Other Secret	40.2	69.3	52.8			
4	Chloroplast	72.6	92.8	82.1	57.5	45.0	0.5
	Mitochondria	79.2	71.9	75.4			
	Plasma membrane	69.9	78.9	74.2			
	Peroxisome	22.2	7.1	12.5			
	Lysosome	43.8	13.0	23.8			
	Golgi app	21.4	12.2	16.1			
	Endo ret	28.9	66.7	43.9			
	Nucleus	22.5	17.0	19.5			

Table 7.24 Analysis of the first variant predictive tree by ten fold cross-validation.

		Sensitivity	Specificity	GAv	Q	nQ	GC
1	Chloroplast	72.6	92.8	82.1	75.5	74.8	0.7
	Mitochondria	79.2	71.9	75.4			
	Plasma membrane	69.9	78.9	74.2			
	Secretory	77.6	71.7	74.6			
2	Chloroplast	72.6	92.8	82.1	71.8	62.34	0.6
	Mitochondria	79.2	71.9	75.4			
	Plasma membrane	69.9	78.9	74.2			
	Peroxisome	18.52	12	15			
	Other Secret	71.6	69	70.1			
3	Chloroplast	72.6	92.8	82.1	69.4	56.1	0.6
	Mitochondria	79.2	71.9	75.4			
	Plasma membrane	69.9	78.9	74.2			
	Peroxisome	18.52	12	15			
	Lysosome	30	23.1	26.3			
	Other Secret	66.7	65.7	66.2			
4	Chloroplast	72.6	92.8	82.1	62.6	50.7	0.5
	Mitochondria	79.2	71.9	75.4			
	Plasma membrane	69.9	78.9	74.2			
	Peroxisome	18.52	12	15			
	Lysosome	30	23.1	26.3			
	Golgi	34.1	12.5	20.7			
	Other secretory	45.6	64.3	54.2			
5	Chloroplast	72.6	92.8	82.1	60	46.2	0.5
	Mitochondria	79.2	71.9	75.4			
	Plasma membrane	69.9	78.9	74.2			
	Peroxisome	18.52	12	15			
	Lysosome	30	23.1	26.3			
	Golgi app	34.1	12.5	20.7			
	ER	36.3	61.3	47.2			
	Nucleus	22.5	17.31	19.7			

Table 7.25. Analysis of the second variant predictive tree by ten fold cross-validation.

Both tree variants have been empirically designed and found to mimic cellular sorting. These architectures support the statement made by Nair and colleagues (Nair and Rost, 2005) who postulated that mimicking cellular sorting improves the prediction of subcellular location.

The method itself showed promising results at the higher levels, while the accuracy of the method decreases as the secretory class is further classified. The developed method

distinguishes between chloroplast, mitochondria, plasma membrane and secretory with approximately 75% accuracy based solely on the information contained in the transmembrane region. These results demonstrate that polytopic α-helical membrane proteins contain important information located in their transmembrane regions and that particular residues located at a similar depth in the membrane may be important in the organellar localization of polytopic α-helical membrane proteins. Proteins belonging to organelles that correspond to the secretory pathway, if incorrectly predicted, tend to be predicted as other organelles involved in the secretory pathway, reflecting the evolutionary relationships between these organelles. **Table 7.26** and **table 7.27** show the percentage confusion matrices for the first and second variant of the tree respectively. All coloured cells correspond to cells where the percentage data points belonging to the class *i* and predicted as members of the class *j* is ≥ 15%. Apart from the cells corresponding to true positives (coloured in grey), the majority of these highlighted cells involved two organelles involved in the secretory pathway (including the plasma membrane). These prediction errors might indicate that within the secretory pathway other signaling features located outside the membrane, such as sorting signals and signal peptides, might be more important in the location of a protein to its appropriate organelle. An organellar distance d was calculated using the percentage values included in **table 7.26** where

$$d_{ij} = d_{ji} = \left(\frac{x_{ji} + x_{ij}}{2} \right)^2 \qquad (7.1)$$

Chloroplast	Mitochondria	Plasma membrane	Peroxisome	Lysosome	Golgi	Endopl Ret	Nucleus	
72.58	14.52	3.23	4.84	1.61	2.42	0.00	0.81	Chloroplast
1.22	79.18	2.86	7.76	2.04	1.22	2.86	2.86	Mitochondria
1.37	7.31	69.86	5.94	3.65	7.76	3.20	0.91	PlasmaMembrane
0.00	37.04	0.00	22.22	3.70	11.11	18.52	7.41	Peroxisome
0.00	6.25	12.50	0.00	43.75	18.75	0.00	18.75	Lysosome
0.00	14.29	16.67	16.67	9.52	21.43	16.67	4.76	Golgi
0.50	8.96	9.45	11.44	11.94	15.42	28.86	13.43	ER
0.00	17.50	5.00	27.50	7.50	12.50	7.50	22.50	Nucleus

Table 7.26. Percentage confusion matrix for the first variant of the tree based set of classifiers. Cells coloured in grey correspond to the percentage of true positives, cells coloured in yellow correspond to misclassifications where the percentage of erroneously predicted membrane proteins is ≥ 15% but < 30%. Cells coloured in red correspond to misclassifications where the percentage of erroneously predicted membrane proteins is ≥ 30%.

Chloroplast	Mitochondria	Plasma Membrane	Peroxisome	Lysosome	Golgi	Endopl Ret	Nucleus	
72.58	14.52	3.23	0.00	2.42	4.03	0.00	3.23	Chloroplast
1.22	79.18	2.86	3.67	1.22	2.04	6.12	3.67	Mitochondria
1.37	7.31	69.86	2.28	0.46	13.24	3.65	1.83	PlasmaMembrane
0.00	37.04	0.00	18.52	0.00	11.11	22.22	11.11	Peroxisome
0.00	6.25	12.50	6.25	37.50	12.50	18.75	6.25	Lysosome
0.00	14.29	16.67	11.90	4.76	33.33	14.29	4.76	Golgi
0.50	8.96	9.45	8.46	3.48	22.89	36.32	9.95	ER
0.00	17.50	5.00	5.00	10.00	20.00	20.00	22.50	Nucleus

Table 7.27. Percentage confusion matrix for the second variant of the tree based set of classifiers. Cells coloured in grey correspond to the percentage of true positives, cells coloured in yellow correspond to misclassifications where the percentage of erroneously predicted membrane proteins is ≥ 15% but < 30%. Cells coloured in red correspond to misclassifications where the percentage of erroneously predicted membrane proteins is ≥ 30%.

The different organellar distances were plotted using biolayout visualization software (Enright and Ouzounis, 2001). **Figure 7.4** and **figure 7.5** show the distance relationships between different organelles where **figure 7.4** shows all distance relationships and **figure 7.5** shows all distance relationships whose weight is ≥ 15 percentage points. It is evident that these graphs can be used to imply evolutionary relationships between the different organelles. The organelles involved in the secretory pathway (including the plasma membrane) form a cluster that can be easily observed when only distance relationships with weight ≥ 15 percentage points are visualized **(figure 7.5)**. An interesting link observed in these graphs is that between the mitochondria and the peroxisome. According to the percentage confusion matrices (**table 7.26** and **table 7.27**), the most abundant error corresponds to proteins located in the peroxisome but predicted as mitochondrial proteins (37%). However, this relationship seems to be unidirectional as it was not observed that a significant proportion of mitochondrial proteins were predicted as peroxisomal proteins. Recent work carried out by Gabaldón and colleagues described the evolutionary relationship between these two organelles (Gabaldon et al., 2006). According to this work, throughout evolution of eukaryotic cells, proteins from different cellular compartments have been retargeted to the peroxisome, and three different retargeting origins were described where two of them include retargeting from the mitochondria to the peroxisome. Some of the peroxisomal proteins were found to be derived from the alpha-proteobacterial ancestor of the mitochondrion where retargeting involved the transfer of the

corresponding genes to the nucleus. Other proteins, without an observable alpha-proteobacterial origin, have also been retargeted from the mitochondria whereas a further class of proteins has been retargeted from other cellular compartments (e.g the endoplasmic reticulum). Therefore, the results obtained from the confusion matrices can be used not only to evaluate the predictive performance of the method but also to imply evolutionary relationships between organelles.

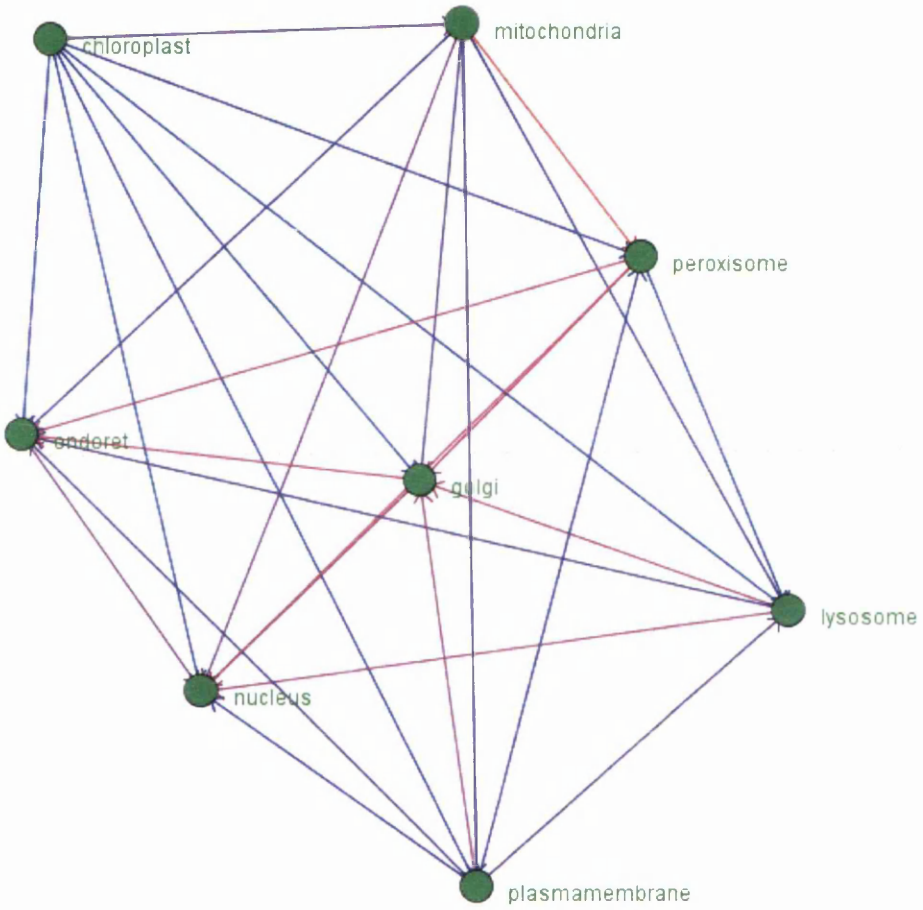

Figure 7.4 Organelle distance relationships extracted from the percentage confusion matrix obtained from the first variant of the predictive tree. The nodes represent the different classes whereas the edges join two related nodes. Edges interconnecting the nodes have been coloured in order to represent the weight of the relationship between two interconnected nodes. Image generated with biolayout (Enright and Ouzounis, 2001).

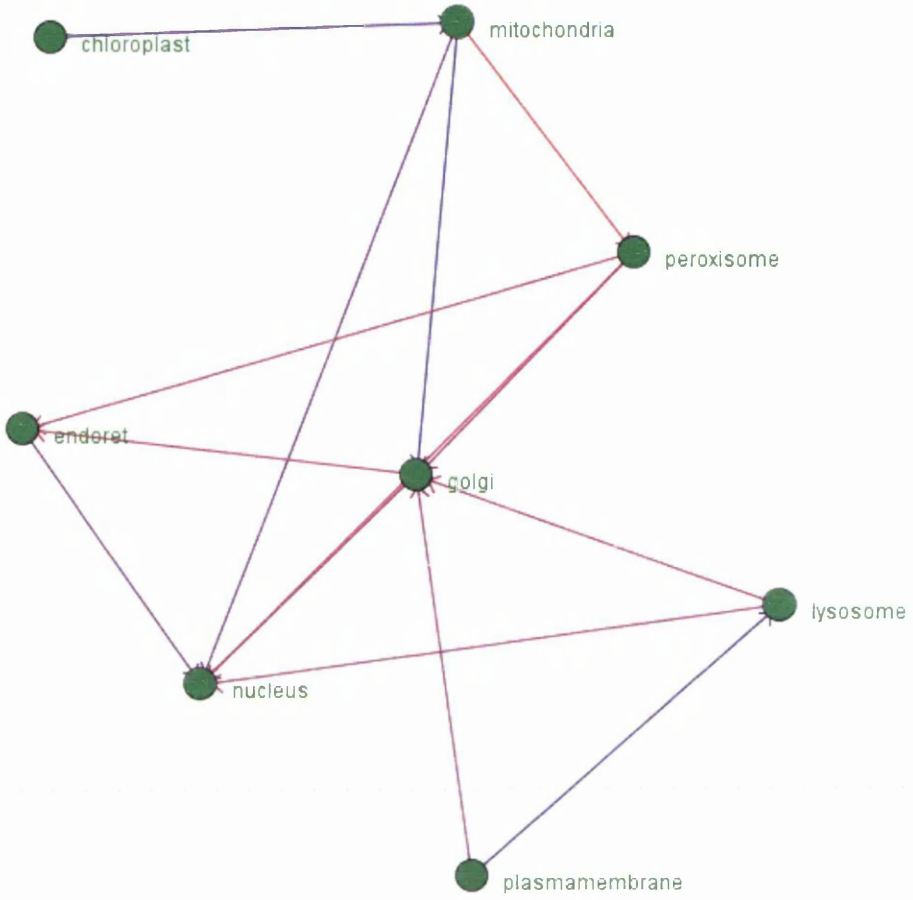

Figure 7.5. Organelle distance relationships extracted from the percentage confusion matrix obtained from the first variant of the predictive tree. The nodes represent the different classes whereas the edges join two related nodes. Edges interconnecting the nodes have been coloured in order to represent the weight of the relationship between two interconnected nodes.. The edges have been filtered using a minimum weight of 15% in order to remove background noise. Image generated with biolayout (Enright and Ouzounis, 2001).

A separate analysis was performed considering only non-plant polytopic membrane proteins as it is also possible that the subcellular signature contained in transmembrane regions may have diverged in different kingdoms. Comparison of the assembled data set **(table 7.3)** with the data set corresponding to eukaryotic cells (including animalia, fungi and plantae kingdoms) **(table 7.2)** showed that the majority of proteins contained in the data set corresponding to eukaryotic cells belongs to non-plant organisms and therefore the development of an algorithm to predict the subcellular location of membrane proteins of plant cells it is not feasible. In the testing of the newly generated data set, prediction was

only performed at the first level (discriminating between mitonchondria, plasma membrane and secretory) as further levels were unlikely to report different accuracy values to the values already obtained due to the taxonomic distribution of proteins within the secretory class. As with previous analyses, different architectures were tested but the single-step mode proved to be the more accurate architecture to distinguish between these three classes. The SVM(7) and SVM(8) reported the highest accuracy values for the given data set **(table 7.28)**. These results did not show an improvement compared to the results obtained from the ten fold cross-validation of level 1 using plant and non-plant polytopic membrane proteins, which possibly indicates that the subcellular signature located in transmembrane regions has not diverged significantly between different eukaryotic kingdoms. Further research on the evolution of subcellular location signals in polytopic membrane proteins will involve the analysis of larger sets for the *animalia*, *plantae* and *fungi* kingdom that are not available yet.

	Sensitivity	Specificity	GAv	Q	nQ	GC
Secretory	76.57	72.28	74.39	74.2	72.5	0.61
Mitochondria	82.59	77.93	80.23			
Plasma membrane	58.39	72.50	65.06			

Table 7.28. Ten fold cross-validation of the first level to distinguish between secretory, mitochondrial and plasma membrane proteins. The data set was composed by proteins belonging to non-plant organisms. The results showed relate to the SVM(7) and SVM(8) without prior attribute selection. The accuracy values do not show an improvement over level 1, which was based on plant and non-plant membrane proteins.

In order to appropriately estimate the obtained results, it is necessary to consider current topology prediction methods and organelle-specific membrane proteomes. Current topology prediction methods have an accuracy of 70-80%, which might have a significant effect on the obtained results. The reason for this is that even if a single transmembrane region is either missed or incorrectly predicted, the extracted features obtained by TMDEPTH **(Chapter 6)** might change dramatically **(figure 6.4)**. Likewise, some of the organelle-specific membrane proteins considered in this approach were found to be under-represented, which obviously has a negative effect during the data mining analysis. Considering these facts the results shown by the predictive method are quite promising.

Future improvements in topology prediction and the identification of new membrane proteins, the predictive accuracy of the developed method will surely increase. However, there is some contention regarding how the data set has been treated in terms of the minimization of sequence redundancy. Similar studies based on globular proteins have reduced the sequence similarity to levels as low as 40%. Such levels of sequence similarity filtering could not be applied to the set of polytopic membrane proteins because some of the organelles described in the developed data set (peroxisome, lysosome and nucleus) could not be feasibly considered as their size would have been too small to be statistically significant.

In closing, the scope of this aspect of the project research was not to develop a "YAPA" (Yet Another Paper About...) prediction of subcellular location from sequence, but to demonstrate for the first time that combining sequence and topology information can be used, along with the detection of other protein features, to predict the subcellular location of polytopic membrane proteins. It is believed that proteins localize to their appropriate organelle using a variety of mechanisms and this is probably the reason why methods based solely on sorting signals or amino acid compositions seem to have reached a plateau in their prediction accuracy. This method used only information contained in the transmembrane region and the predictive accuracy of the method could be increased further by incorporating other features located outside the membrane (e.g N-terminus,C-terminus, extramembraneous loops and signal peptides) that can be used to guide cellular sorting. It is also possible that extramembraneous targeting motifs may be responsible for initial localization of membrane proteins to the appropriate organelle, but that transmembrane regions provide the means by which a particular membrane protein may be retained in the appropriate organellar membrane. Further investigation of these properties of transmembrane regions will also cast light on the relationships between the variations observed in membrane lipid composition and thickness between different organelles, and the specific compositional signatures of transmembrane region associated with the subcellular location of membrane proteins.

7.4 Conclusion

Prediction of subcellular location is an important step while functionally characterizing unknown proteins. Although extensive studies have been carried out to predict the subcellular location of soluble proteins, little effort has been done to predict the subcellular location of membrane proteins. The developed method combines different data mining techniques to predict the subcellular location of polytopic α-helical membrane proteins based on the TMDEPTH feature extraction method **(Chapter 6)**. This method computes a vector, based solely on the amino acid sequence and a refined topological model **(Chapter 5)**, that represents all pairs of residues located at a similar depth in the membrane. The developed method showed a normalized accuracy of 75% to discriminate between proteins belonging to the chloroplast, mitochondria, plasma membrane and secretory organelles. Thus, reflecting the importance of the transmembrane domain to assign the location of polytopic membrane proteins. Additionally, the obtained results were used to infer evolutionary relationships between polytopic membrane proteins belonging to different organelles. The developed method is bound to increase its accuracy as more accurate topology prediction methods are implemented and more proteins belonging to under-represented subsets are found.

7.5 References

ANDRADE, M. A., O'DONOGHUE, S. I. & ROST, B. (1998) Adaptation of protein surfaces to subcellular location. *J Mol Biol,* 276, 517-25.

AOKI, D., LEE, N., YAMAGUCHI, N., DUBOIS, C. & FUKUDA, M. N. (1992) Golgi retention of a trans-Golgi membrane protein, galactosyltransferase, requires cysteine and histidine residues within the membrane-anchoring domain. *Proc Natl Acad Sci U S A,* 89, 4319-23.

ATURALIYA, R. N., FINK, J. L., DAVIS, M. J., TEASDALE, M. S., HANSON, K. A., MIRANDA, K. C., FORREST, A. R., GRIMMOND, S. M., SUZUKI, H., KANAMORI, M., KAI, C., KAWAI, J., CARNINCI, P., HAYASHIZAKI, Y. & TEASDALE, R. D. (2006) Subcellular localization of mammalian type II membrane proteins. *Traffic,* 7, 613-25.

BALDI, P., BRUNAK, S., CHAUVIN, Y., ANDERSEN, C. A. & NIELSEN, H. (2000) Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics,* 16, 412-24.

BANNAI, H., TAMADA, Y., MARUYAMA, O., NAKAI, K. & MIYANO, S. (2002) Extensive feature detection of N-terminal protein sorting signals. *Bioinformatics,* 18, 298-305.

BHASIN, M. & RAGHAVA, G. P. (2004) ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. *Nucleic Acids Res,* 32, W414-9.

BICKMORE, W. A. & SUTHERLAND, H. G. (2002) Addressing protein localization within the nucleus. *Embo J,* 21, 1248-54.

BIERMANNS, M., VON LAAR, J., BROSIUS, U. & GARTNER, J. (2003) The peroxisomal membrane targeting elements of human peroxin 2 (PEX2). *Eur J Cell Biol,* 82, 155-62.

BODEN, M. & HAWKINS, J. (2005) Prediction of subcellular localization using sequence-biased recurrent networks. *Bioinformatics,* 21, 2279-86.

BONIFACINO, J. S., COSSON, P., SHAH, N. & KLAUSNER, R. D. (1991) Role of potentially charged transmembrane residues in targeting proteins for retention and degradation within the endoplasmic reticulum. *Embo J,* 10, 2783-93.

BULASHEVSKA, A. & EILS, R. (2006) Predicting protein subcellular locations using hierarchical ensemble of Bayesian classifiers based on Markov chains. *BMC Bioinformatics,* 7, 298.

CAI, Y. D. & CHOU, K. C. (2003) Nearest neighbour algorithm for predicting protein subcellular location by combining functional domain composition and pseudo-amino acid composition. *Biochem Biophys Res Commun,* 305, 407-11.

CAI, Y. D. & CHOU, K. C. (2004) Predicting subcellular localization of proteins in a hybridization space. *Bioinformatics,* 20, 1151-6.

CAI, Y. D., LIU, X. J. & CHOU, K. C. (2002a) Artificial neural network model for predicting protein subcellular location. *Comput Chem,* 26, 179-82.

CAI, Y. D., LIU, X. J., XU, X. B. & CHOU, K. C. (2002b) Support vector machines for prediction of protein subcellular location by incorporating quasi-sequence-order effect. *J Cell Biochem,* 84, 343-8.

CEDANO, J., ALOY, P., PEREZ-PONS, J. A. & QUEROL, E. (1997) Relation between amino acid composition and cellular location of proteins. *J Mol Biol,* 266, 594-600.

CHOU, K. C. (2000) Prediction of protein subcellular locations by incorporating quasi-sequence-order effect. *Biochem Biophys Res Commun,* 278, 477-83.

CHOU, K. C. (2001) Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins,* 43, 246-55.

CHOU, K. C. & CAI, Y. D. (2002) Using functional domain composition and support vector machines for prediction of protein subcellular location. *J Biol Chem,* 277, 45765-9.

CHOU, K. C. & CAI, Y. D. (2003) Prediction and classification of protein subcellular location-sequence-order effect and pseudo amino acid composition. *J Cell Biochem,* 90, 1250-60.

CHOU, K. C. & CAI, Y. D. (2004) Predicting subcellular localization of proteins by hybridizing functional domain composition and pseudo-amino acid composition. *J Cell Biochem,* 91, 1197-203.

CHOU, K. C. & ELROD, D. W. (1999a) Prediction of membrane protein types and subcellular locations. *Proteins,* 34, 137-53.

CHOU, K. C. & ELROD, D. W. (1999b) Protein subcellular location prediction. *Protein Eng,* 12, 107-18.

CLAROS, M. G. & VINCENS, P. (1996) Computational method to predict mitochondrially imported proteins and their targeting sequences. *Eur J Biochem,* 241, 779-86.

COCQUEREL, L., DUVET, S., MEUNIER, J. C., PILLEZ, A., CACAN, R., WYCHOWSKI, C. & DUBUISSON, J. (1999) The transmembrane domain of hepatitis C virus glycoprotein E1 is a signal for static retention in the endoplasmic reticulum. *J Virol,* 73, 2641-9.

COCQUEREL, L., WYCHOWSKI, C., MINNER, F., PENIN, F. & DUBUISSON, J. (2000) Charged residues in the transmembrane domains of hepatitis C virus glycoproteins play a major role in the processing, subcellular localization, and assembly of these envelope proteins. *J Virol,* 74, 3623-33.

COKOL, M., NAIR, R. & ROST, B. (2000) Finding nuclear localization signals. *EMBO Rep,* 1, 411-5.

COLLEY, K. J. (1997) Golgi localization of glycosyltransferases: more questions than answers. *Glycobiology,* 7, 1-13.

CUI, Q., JIANG, T., LIU, B. & MA, S. (2004) Esub8: a novel tool to predict protein subcellular localizations in eukaryotic organisms. *BMC Bioinformatics,* 5, 66.

DRAWID, A. & GERSTEIN, M. (2000) A Bayesian system integrating expression data with sequence patterns for localizing proteins: comprehensive application to the yeast genome. *J Mol Biol,* 301, 1059-75.

EMANUELSSON, O., NIELSEN, H., BRUNAK, S. & VON HEIJNE, G. (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J Mol Biol,* 300, 1005-16.

EMANUELSSON, O., NIELSEN, H. & VON HEIJNE, G. (1999) ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. *Protein Sci,* 8, 978-84.

ENRIGHT, A. J. & OUZOUNIS, C. A. (2001) BioLayout--an automatic graph layout algorithm for similarity visualization. *Bioinformatics,* 17, 853-4.

FENG, Z. P. & ZHANG, C. T. (2001) Prediction of the subcellular location of prokaryotic proteins based on the hydrophobicity index of amino acids. *Int J Biol Macromol,* 28, 255-61.

FUJIWARA, Y. & ASOGAWA, M. (2001) Prediction of subcellular localizations using amino acid composition and order. *Genome Inform,* 12, 103-12.

FUJIWARA, Y., ASOGAWA, M. & NAKAI, K. (1997) Prediction of Mitochondrial Targeting Signals Using Hidden Markov Model. *Genome Inform Ser Workshop Genome Inform,* 8, 53-60.

GABALDON, T., SNEL, B., VAN ZIMMEREN, F., HEMRIKA, W., TABAK, H. & HUYNEN, M. A. (2006) Origin and evolution of the peroxisomal proteome. *Biol Direct,* 1, 8.

GARDY, J. L., SPENCER, C., WANG, K., ESTER, M., TUSNADY, G. E., SIMON, I., HUA, S., DEFAYS, K., LAMBERT, C., NAKAI, K. & BRINKMAN, F. S. (2003) PSORT-B: Improving protein subcellular localization prediction for Gram-negative bacteria. *Nucleic Acids Res,* 31, 3613-7.

GUDA, C., GUDA, P., FAHY, E. & SUBRAMANIAM, S. (2004) MITOPRED: a web server for the prediction of mitochondrial proteins. *Nucleic Acids Res,* 32, W372-4.

GUDA, C. & SUBRAMANIAM, S. (2005) pTARGET [corrected] a new method for predicting protein subcellular localization in eukaryotes. *Bioinformatics,* 21, 3963-9.

HAWKINS, J. & BODEN, M. (2006) Detecting and sorting targeting peptides with neural networks and support vector machines. *J Bioinform Comput Biol,* 4, 1-18.

HOBMAN, T. C., LEMON, H. F. & JEWELL, K. (1997) Characterization of an endoplasmic reticulum retention signal in the rubella virus E1 glycoprotein. *J Virol,* 71, 7670-80.

HOBMAN, T. C., WOODWARD, L. & FARQUHAR, M. G. (1995) Targeting of a heterodimeric membrane protein complex to the Golgi: rubella virus E2 glycoprotein contains a transmembrane Golgi retention signal. *Mol Biol Cell,* 6, 7-20.

HOLM, L. & SANDER, C. (1998) Removing near-neighbour redundancy from large protein sequence collections. *Bioinformatics,* 14, 423-9.

HONSHO, M., HIROSHIGE, T. & FUJIKI, Y. (2002) The membrane biogenesis peroxin Pex16p. Topogenesis and functional roles in peroxisomal membrane assembly. *J Biol Chem,* 277, 44513-24.

HUANG, Y. & LI, Y. (2004) Prediction of protein subcellular locations using fuzzy k-NN method. *Bioinformatics,* 20, 21-8.

JONES, J. M., MORRELL, J. C. & GOULD, S. J. (2004) PEX19 is a predominantly cytosolic chaperone and import receptor for class 1 peroxisomal membrane proteins. *J Cell Biol,* 164, 57-67.

KAWASHIMA, S. & KANEHISA, M. (2000) AAindex: amino acid index database. *Nucleic Acids Res,* 28, 374.

KROGH, A., LARSSON, B., VON HEIJNE, G. & SONNHAMMER, E. L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol,* 305, 567-80.

LETOURNEUR, F. & COSSON, P. (1998) Targeting to the endoplasmic reticulum in yeast cells by determinants present in transmembrane domains. *J Biol Chem,* 273, 33273-8.

LETUNIC, I., COPLEY, R. R., PILS, B., PINKERT, S., SCHULTZ, J. & BORK, P. (2006) SMART 5: domains in the context of genomes and networks. *Nucleic Acids Res,* 34, D257-60.

LI, W., JAROSZEWSKI, L. & GODZIK, A. (2001) Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics,* 17, 282-3.

MA, J., HAYEK, S. M. & BHAT, M. B. (2004) Membrane topology and membrane retention of the ryanodine receptor calcium release channel. *Cell Biochem Biophys,* 40, 207-24.

MARCOTTE, E. M., XENARIOS, I., VAN DER BLIEK, A. M. & EISENBERG, D. (2000) Localizing proteins in the cell from their phylogenetic profiles. *Proc Natl Acad Sci U S A,* 97, 12115-20.

MATSUDA, S., VERT, J. P., SAIGO, H., UEDA, N., TOH, H. & AKUTSU, T. (2005) A novel representation of protein sequences for prediction of subcellular location using support vector machines. *Protein Sci,* 14, 2804-13.

MATTHEWS, B. W. (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta,* 405, 442-51.

MOTT, R., SCHULTZ, J., BORK, P. & PONTING, C. P. (2002) Predicting protein cellular localization using a domain projection method. *Genome Res,* 12, 1168-74.

MUNRO, S. (1995) An investigation of the role of transmembrane domains in Golgi protein retention. *Embo J,* 14, 4695-704.

NAIR, R., CARTER, P. & ROST, B. (2003) NLSdb: database of nuclear localization signals. *Nucleic Acids Res,* 31, 397-9.

NAIR, R. & ROST, B. (2002a) Inferring sub-cellular localization through automated lexical analysis. *Bioinformatics,* 18 Suppl 1, S78-86.

NAIR, R. & ROST, B. (2002b) Sequence conserved for subcellular localization. *Protein Sci,* 11, 2836-47.

NAIR, R. & ROST, B. (2003a) Better prediction of sub-cellular localization by combining evolutionary and structural information. *Proteins,* 53, 917-30.

NAIR, R. & ROST, B. (2003b) LOC3D: annotate sub-cellular localization for protein structures. *Nucleic Acids Res,* 31, 3337-40.

NAIR, R. & ROST, B. (2005) Mimicking cellular sorting improves prediction of subcellular localization. *J Mol Biol,* 348, 85-100.

NAKAI, K. & HORTON, P. (1999) PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem Sci,* 24, 34-6.

NAKAI, K. & KANEHISA, M. (1991) Expert system for predicting protein localization sites in gram-negative bacteria. *Proteins,* 11, 95-110.

NAKAI, K. & KANEHISA, M. (1992) A knowledge base for predicting protein localization sites in eukaryotic cells. *Genomics,* 14, 897-911.

NEUBERGER, G., KUNZE, M., EISENHABER, F., BERGER, J., HARTIG, A. & BROCARD, C. (2004) Hidden localization motifs: naturally occurring peroxisomal targeting signals in non-peroxisomal proteins. *Genome Biol,* 5, R97.

OP DE BEECK, A., ROUILLE, Y., CARON, M., DUVET, S. & DUBUISSON, J. (2004) The transmembrane domains of the prM and E proteins of yellow fever virus are endoplasmic reticulum localization signals. *J Virol,* 78, 12591-602.

PAN, Y. X., LI, D. W., DUAN, Y., ZHANG, Z. Z., XU, M. Q., FENG, G. Y. & HE, L. (2005) Predicting protein subcellular location using digital signal processing. *Acta Biochim Biophys Sin (Shanghai),* 37, 88-96.

PAN, Y. X., ZHANG, Z. Z., GUO, Z. M., FENG, G. Y., HUANG, Z. D. & HE, L. (2003) Application of pseudo amino acid composition for predicting protein subcellular location: stochastic signal processing approach. *J Protein Chem,* 22, 395-402.

PARK, K. J. & KANEHISA, M. (2003) Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs. *Bioinformatics,* 19, 1656-63.

PEDRAZZINI, E., VILLA, A. & BORGESE, N. (1996) A mutant cytochrome b5 with a lengthened membrane anchor escapes from the endoplasmic reticulum and reaches the plasma membrane. *Proc Natl Acad Sci U S A,* 93, 4207-12.

PETSALAKI, E. I., BAGOS, P. G., LITOU, Z. I. & HAMODRAKAS, S. J. (2006) PredSL: a tool for the N-terminal sequence-based prediction of protein subcellular localization. *Genomics Proteomics Bioinformatics,* 4, 48-55.

PIERLEONI, A., MARTELLI, P. L., FARISELLI, P. & CASADIO, R. (2006) BaCelLo: a balanced subcellular localization predictor. *Bioinformatics,* 22, e408-16.

QUINLAN, R. (1993) *C4.5: Programs for Machine Learning,* San Mateo, CA, Morgan Kaufmann Publishers.

RECZKO, M. & HATZIGERRORGIOU, A. (2004) Prediction of the subcellular localization of eukaryotic proteins using sequence signals and composition. *Proteomics,* 4, 1591-6.

REINHARDT, A. & HUBBARD, T. (1998) Using neural networks for prediction of the subcellular location of proteins. *Nucleic Acids Res,* 26, 2230-6.

ROLLS, M. M., STEIN, P. A., TAYLOR, S. S., HA, E., MCKEON, F. & RAPOPORT, T. A. (1999) A visual screen of a GFP-fusion library identifies a new type of nuclear envelope membrane protein. *J Cell Biol,* 146, 29-44.

SARDA, D., CHUA, G. H., LI, K. B. & KRISHNAN, A. (2005) pSLIP: SVM based protein subcellular localization prediction using multiple physicochemical properties. *BMC Bioinformatics,* 6, 152.

SCOTT, M. S., CALAFELL, S. J., THOMAS, D. Y. & HALLETT, M. T. (2005) Refining protein subcellular localization. *PLoS Comput Biol,* 1, e66.

SCOTT, M. S., THOMAS, D. Y. & HALLETT, M. T. (2004) Predicting subcellular localization via protein motif co-occurrence. *Genome Res,* 14, 1957-66.

SONNHAMMER, E. L., VON HEIJNE, G. & KROGH, A. (1998) A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc Int Conf Intell Syst Mol Biol,* 6, 175-82.

SZCZESNA-SKORUPA, E. & KEMPER, B. (2000) Endoplasmic reticulum retention determinants in the transmembrane and linker domains of cytochrome P450 2C1. *J Biol Chem,* 275, 19409-15.

WITTEN, I. H. & FRANK, E. (2005) *Data Mining: Practical machine learning tools and techniques,* San Francisco, Morgan Kaufmann.

XIE, D., LI, A., WANG, M., FAN, Z. & FENG, H. (2005) LOCSVMPSI: a web server for subcellular localization of eukaryotic proteins using SVM and profile of PSI-BLAST. *Nucleic Acids Res,* 33, W105-10.

YANG, M., ELLENBERG, J., BONIFACINO, J. S. & WEISSMAN, A. M. (1997) The transmembrane domain of a carboxyl-terminal anchored protein determines localization to the endoplasmic reticulum. *J Biol Chem,* 272, 1970-5.

YU, C. S., LIN, C. J. & HWANG, J. K. (2004) Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions. *Protein Sci,* 13, 1402-6.

CHAPTER 8

TMFUN: Prediction of molecular function of eukaryotic membrane proteins based on sequence and topological information

8.1 Introduction

The exponential growth experienced over the last two decades in biological sequence databases **(Chapter 3)** is constantly increasing the gap between the rate at which new sequences are obtained and the rate at which these new sequences can be experimentally characterized. Functional characterization of a single gene takes approximately one year using the current experimental approaches. Therefore, computational approaches are an essential complementary tool in predicting structural and functional properties of newly sequenced genes including functionally important residues, active and binding sites, molecular function, subcellular localization, molecular pathways, interacting partners and post-translational modifications. Ultimately, these predictions should be experimentally confirmed.

The most popular methods used to annotate new gene products are based on the sequence similarity concept. These methods have proven to be a useful approach but they have limitations, which have led to the development of complementary methods for the annotation of genes and gene products based on a wide range of different techniques. Below is a description of the main methods for functional prediction:

8.1.1 Sequence similarity based methods

The basis underlying these methods is derived from the concept of molecular evolution where homologues derived from a common ancestor and sharing a significant sequence similarity are deemed to possess similar structure and function. Following this principle, two sequences obtained from different species but sharing significant sequence similarity are believed to possibly share structural and functional properties. Pair-wise sequence similarity methods, such us the Smith-Waterman algorithm (Smith and Waterman, 1981), FASTA (Pearson and Lipman, 1988) and BLAST (Altschul et al., 1990), are commonly used to detect homologues in large protein databases. The queried sequence is matched against all sequences in a given database and significant matches are reported in order to infer structural and functional properties. While homologues sharing at least 30% of their residues can be used to infer similar protein structure, it is necessary to find homologues with a sequence similarity of at least 60% in order to infer similar function. Pair-wise sequence similarity methods often fail to detect distant homologues with sequence similarity lower than 30%. Likewise, identification of conserved residues with a structural or functional role based only on a pair-wise comparison is more difficult than identifying important residues using a larger set of sequences. In order to overcome these limitations other methods have been implemented (Park et al., 1998). The intermediate sequence search method (ISS) (Park et al., 1997, Abascal and Valencia, 2002, Gerstein, 1998) is used to detect distantly related sequences, which have diverged beyond the point where their evolutionary relationships can be recognized, by relating both sequences to a third intermediate sequence that is a homologue of both sequences. This method is only applicable when the matching region between the 1st sequence and the intermediate sequence and the 2nd sequence and intermediate sequence corresponds to the same domain. Multi-domain proteins matching different domains of the same intermediate sequence will lead to a false distant relationship. Other methods use shared properties of sets of related sequences extracted from multiple sequence alignments. Based on these alignments, different motif and sequence representation techniques and searching algorithms have been developed such as single (Falquet et al., 2002) or multiple motif methods (Attwood et al., 1999, Henikoff et al., 1999a, Wu and Brutlag, 1995), templates (Bashford et al., 1987, Tatusov et al., 1994, Taylor, 1986, Yi and Lander, 1994), profiles (Bucher et al., 1996,

Gribskov et al., 1987, Luthy et al., 1994), hidden Markov models (Baldi et al., 1994, Eddy, 1996, Krogh et al., 1994) and the position-specific-iterated (PSI) BLAST (Altschul et al., 1997).

Although sequence similarity based methods are the most common computational approach used to characterize new sequences, these methods are problematic and should be used with caution. The "similar sequence-similar structure-similar function" paradigm does not always hold true as sequence similarity does not guarantee identical function. Through evolution proteins are subject to mutations. Function directly involves fewer residues than structure. Therefore, random mutations are more likely to have an effect on function rather than structure (Rost et al., 2003). That is, homologues derived from a common ancestor can diverge, and lose, modify or acquire functional properties before they lose their folding. Similarly, proteins with different evolutionary backgrounds can converge into similar functions despite not having a common ancestor. Consequently, sequence similarity based methods might erroneously assign functions to homologues that have diverged and have different functional properties and non-homologues that have evolutionary converged to achieve similar biological processes. Sequence similarity based methods directly rely on the quality of databases and automatic methods based on sequence similarity iteratively facilitate the error propagation in such databases (Gilks et al., 2002). Prediction reproducibility is often poor resulting in different prediction (either using different methods or different databases), which can not be merged in a consensus prediction. Error estimation analyses concluded that 30% of the functional annotations based on homology detection might contain significant errors (Devos and Valencia, 2001). Other studies estimated that in order to assign full enzymatic activity with less than 30% error, it is necessary to have levels of pair-wise sequence similarity higher than 60%, whereas to obtain predictions with an accuracy higher than 90% it is required levels of pair-wise sequence similarity higher than 75% (Rost, 2002). Further research showed that only 35% of the proteins could be predicted with an error rate lower than 5% whereas 70% of the proteins can be predicted when the error rate is set up to 70% (Rost et al., 2003). Likewise, 25-40% of the newly sequenced proteins can not be predicted using sequence similarity based methods either because no homologues have been found yet or because their

corresponding homologues have not been characterized (Rost et al., 2003, Iliopoulos et al., 2001).

8.1.2 Gene orthology detection methods

Orthologues are homologue proteins from different species that evolved from a common ancestor via a speciation event (Fitch, 1970). By contrast, paralogues are homologue proteins that have evolved by a gene duplication event. When such gene duplication events occur within the same species (lineage-specific duplicates) after the speciation event, duplicated proteins are in-paralogues whereas if the gene duplication event occurred before the speciation event (duplicated genes are present in the common ancestor of two species) the duplicated proteins are out-paralogues. Orthology prediction has a relevant role in annotation of newly sequenced genomes as orthologues are more likely to conserve their ancestral functional activity whereas the in-paralogues often evolve to develop new functions that can be line-specific. The classical approach to predicting and distinguishing between orthologues and paralogues are based on phylogenetic trees, which reconstruct the evolutionary relationships between proteins based on alignment methods (Storm and Sonnhammer, 2002, Eisen et al., 1995, Saier et al., 1999, Page, 1998). These methods usually require manual correction and are not suitable for the automated prediction of orthologues, however recent methods have been implemented to perform automated phylogenetic based orthology prediction (Chiu et al., 2006, Goodstadt and Ponting, 2006). Alternatively, other methods have been implemented based on sequence similarity under the assumption that orthologues often have a higher level of sequence conservation than paralogues. The genome-specific best hit (BeT) (Tatusov et al., 1997) identifies orthologues when a pair of proteins from different species are found to mutually have the highest level of sequence similarity if genes from both genomes are compared, and both proteins also have the highest level of sequence similarity with a third protein, which belongs to a different genome. Therefore, the identified Bet includes three orthologues from different species creating a triangular cluster. BeTs are clustered if two triangular clusters (BeTs) have a common side, which has led to the development of the COG database (Tatusov et al., 1994, Tatusov et al., 2003). The sequence similarity idea underlying this method has

been derived to implement further algorithms such as the best bidirectional method (Huynen and Bork, 1998) and the INPARANOID algorithm (Remm et al., 2001).

8.1.3 Genomic context methods

Genomic context methods are used to explore the relations between genes in multiple genomes (Gabaldon and Huynen, 2004). These methods have been used as a complimentary approach to sequence similarity based methods to predict higher order functions such as the interacting partners and the pathway or process in which the protein is involved (Huynen et al., 2000). The type of protein-protein interactions predicted using these methods do not necessary imply a physical interaction but it could imply instead a functional interaction between two proteins, which are indirectly linked in the same process or pathway but do not physically interact.

8.1.3.1 Gene fusion

This method is the most direct method. It is based on the observation that two or more proteins encoded by different genes in a particular species are also encoded by the corresponding orthologues in a single gene inferring a gene fusion event (Marcotte et al., 1999, Enright et al., 1999). Gene fusion events enhance the affinity between different proteins in order to facilitate a particular biological process where both proteins are needed. However, gene fusion is not the only mechanism developed throughout evolution to facilitate the interaction between proteins as pairs of proteins that develop binding sites with higher affinity would not be detected using this method (Marcotte et al., 1999). Likewise, predicted gene fusion events involving promiscuous domains might erroneously indicate functional interaction between two unrelated proteins (Gabaldon and Huynen, 2004).

8.1.3.2 Gene order conservation or co-*occurrence* of genes in potential operons

The principle underlying this method is based on the fact that conserved gene clusters through evolution usually encode proteins that functionally interact within the same pathway (Dandekar et al., 1998, Overbeek et al., 1999). Co-occurrence of genes have been detected by measuring the conservation of neighbouring genes (Dandekar et al., 1998) and measuring the conservation of genes located in the same DNA strand and with an intergenic distance fewer than 300 bases (Overbeek et al., 1999). This method is applicable to prokaryotic and eukaryotic genomes.

8.1.3.3 Phylogenetic profile

The concept of phylogenetic profiling was first introduced by Huynen and colleagues (Huynen and Bork, 1998) and Pellegrini and colleagues (Pellegrini et al., 1999). This method extracts the functional interactions between genes by comparing the presence and absence of a set of genes in different genomes. This method was first applied to detect sub-sets of genes with similar phylogenetic distribution inferring a functional interaction between those genes (Huynen and Bork, 1998, Pellegrini et al., 1999). This method was also used to detect non-orthologous gene displacements where a certain protein is found to be missing and a different protein has evolved to catalyze the missing reaction (Koonin et al., 1996). Such events were identified when different proteins showed complementary or anti-correlated profiles (Galperin and Koonin, 2000). A variant of this method (Pazos and Valencia, 2001), is based on the fact that protein families known to interact have more similar phylogenetic trees than expected (Fryxell, 1996, Goh et al., 2000, Hughes and Yeager, 1999).

8.1.3.4 Conservation of co-expression

This method is the only experimental method that is considered to be a genomic context method. Functional interactions are extracted from gene expression data, which can be obtained under different conditions thus relating proteins with a similar expression pattern. The classical experimental approach for this kind of data is microarray analysis (Schena et al., 1995).

8.1.4 Methods to predict functionally important residues

These methods are essential to understanding the mechanism used by different proteins to carry out their molecular function and classify uncharacterized proteins. Moreover, prediction of functionally important residues is often used to identify residues that can be used to distinguish between different sub-types of proteins. Proteins belonging to particular families can also be sub-classified into sub-families that might have diverged and developed different ligand binding specificity and even very different functions while conserving a similar fold. Homology based functional prediction methods often fail to distinguish between different subfamilies as these tree determinant positions (conserved residues specific to a particular subfamily) might only involve few residues. The main computational approaches used to predict functionally important residues are based on the identification of conserved residues from multiple sequence alignments, pattern discovery or identification of functionally important residues from protein structures.

The method developed by Livingston and colleagues characterized the physico-chemical properties of each position in the multiple sequence alignment in order to identify conserved residues sharing similar physicochemical properties (Livingstone and Barton, 1993). The SequenceSpace method (Casari et al., 1995) was implemented to translate aligned sequences into protein vectors, the sequence space, which are projected by principal component analysis in order to sub-classify a given protein family into the corresponding sub-families and identify the corresponding TDP (tree determinant positions) and

conserved residues across the family. The Evolutionary Trace Method (Lichtarge et al., 1996) iteratively divides a given gene tree at different levels, based on sequence similarity, into an increasing number of subgroups, which include different branches in the tree. At each level, the method identifies TDP conserved uniquely within a particular subgroup. The identified residues are then mapped onto a known structure and active sites and functional interfaces are predicted when spatial clusters of residues specific to a particular subgroup are found. A variation of this method, the weighted evolutionary trace method, weights each amino acid sequence according to its uniqueness and the variability in each position is assigned by an amino acid substitution matrix in order to decrease the influence of highly homologous sequences (Landgraf et al., 1999). Similarly other methods, which included more accurate algorithms to build phylogenetic trees and calculate the residue conservation of each position (Armon et al., 2001, Landau et al., 2005, Pupko et al., 2002), have been developed based on the same principle as the evolutionary trace method. The method developed by Johnson and Church (Johnson and Church, 2000) used a multiple structural alignment to integrate ligand binding information with multiple sequence alignments. Following this, a phylogenetic tree was generated to correlate the sub-branches of the tree with ligand-binding specificity. The phylogenetic analysis performed along with sub-alignments of binding site residues were used to identify sequences with similar binding pockets and predict uncharacterized protein sequences. The method developed by Hannenhalli and Russell (Hannenhalli and Russell, 2000) uses a multiple sequence alignment and a set of proteins sub-classified according to a particular definition of function. Subsequently, TDP were predicted by comparison of different hidden Markov model profiles. The authors argued that protein classification based on phylogenetic trees was not suitable for identifying protein families of highly diverged sequences or groups of proteins with similar function but different evolutionary background. The three-dimensional cluster analysis method (Landgraf et al., 2001) uses sequence and structural information to predict functional sites located at the protein surface. This method compares global and regional similarity matrices obtained from a global alignment and different regional alignments that reflect the local structural environment and evolutionary variation. Functional residue clusters could be identified using the regional conservation score, which defines the conservation of each residue and its three-dimensional neighbours. Following a

similar principle, del Sol Mesa and colleagues (del Sol Mesa et al., 2003) developed the mutational behaviour method under the assumption that the mutational behaviour of tree determinant positions is similar to the mutational behaviour of the whole family. Mirny and Gelfand (Mirny and Gelfand, 2002) developed a method to predict residues that determine the protein specificity based on the assumption that the functional specificity of orthologues remains conserved whereas it evolves among paralogues. Oliveira and colleagues (Oliveira et al., 2003) introduced a new method, which computed the Shannon entropy and the residue variability at each position in a multiple sequence alignment (sequences were weighted to reduce the influence of highly identical sequences). Clustering of the residue positions according to the different residue conservation evaluation scores identified a group of residue positions (with lower Shannon entropy and residue variability) correlated with the main functional sites of a given protein.

Pattern discovery methods have also been applied to the problem of detecting conserved residues within a set of proteins. These methods aim to detect sequence domains responsible for specific structural roles or biochemical functions. Proteins might allocate more than one domain in its sequence and it is necessary to analyze the domain space of proteins (including possible interactions between these domains) to understand how a protein works. Protein evolution also involves gain, loss and re-shuffling of existing domains in order to create new functions. These methods have proven to be successful and as a consequence of this several databases containing different motifs have been created. The Prosite database (Hulo et al., 2006) is the most pre-eminent example of the single motif representation where a protein family is represented by a single motif. The Blocks (Henikoff et al., 1999b) and Prints (Attwood et al., 1999) databases include multiple motifs in order to model the mutual context provided by motif neighbours and improve the sensitivity and specificity of the method (Orengo et al., 2003). The classical approach to detect these motifs is based on multiple sequence alignment. However, other approaches such as the TEIRESIAS algorithm (Rigoutsos and Floratos, 1998) have been developed to detect functionally important motifs (Lasso et al., 2006, Darzentas et al., 2005) without the need for prior multiple sequence alignment.

Usually, protein binding and active sites involve residues located far apart from each other in the sequence but spatially clustered in the protein structure. Predicting these spatial clusters of residues is still a challenge for methods based solely on sequence. Alternatively, analysis of crystallized structures can be used to predict spatial clusters of functionally important residues. The classical approach is to compare different structures of functionally related proteins and identify those conserved residues and three-dimensional folds. Other methods explore other properties such as physico-chemical properties, side-chain patterns or surface-solvent accessibility of active sites in the protein molecular surface. Some of these methods can only be applied to crystallized structures with unknown function as the spatial clusters identified depend on the three-dimensional coordinates of the atoms. Therefore, predicted structures obtained by homology modeling, threading or *ab initio* methods are not suitable as the predicted structure does not attain sufficient resolution. The fuzzy functional forms (FFF) method (Fetrow and Skolnick, 1998) overcome this limitation. Using information from multiple sequence alignments, scientific literature and the structure of the protein, this method identified, spatial clusters in crystallized structures that were found in non-local residues in the sequences. Rather than making the descriptors of the spatial clusters highly specific by including the atomic coordinates, the descriptors were made "fuzzy" while still being able to be used to identify functional sites. By relaxing the constraints imposed on these descriptors, the developed method could be used to identify active sites in low-to-moderate resolution models obtained by threading or *ab-initio* methods.

8.1.5 Data mining prediction methods

During the last decade, the use of different data mining techniques to predict protein functional classes has been continuously increasing. Generally, these methods use a set of extracted features to classify a given training set and classify uncharacterized proteins. The extracted features can be computed solely from sequence, structure or include a combination of both. Sequential features such as the amino acid composition or the pseudo-amino acid composition of the whole sequence, physicochemical features such the isoelectric point, molecular weight, hydrophobicity, and structural features such as the

secondary structure or solvent accessibility have been successfully applied. Among the different data mining techniques used, the Support Vector Machines (SVM) are undoubtedly the most popular technique. However, other techniques such as Bayesian methods, neural networks, decision trees and K-nearest neighbour have been applied. Des Jardins and colleagues (des Jardins et al., 1997) applied data mining methods for the first time to classify enzymes (according to the EC classification) based on the isoelectric point, molecular weight and amino-acid composition. King and colleagues developed a method, which combined inductive logic programming clustering and rule learning to predict the function of uncharacterized open reading frames (ORF) from *M. tuberculosis* (King et al., 2000a) and *E. coli* (King et al., 2000b). The authors explored the feature space of proteins in order to maximize the predictive power of the method by individually using features extracted from sequence, phylogeny and predicted secondary structure (King et al., 2001). By combining different classes of features, 40% of the unassigned ORFs were predicted at an estimated accuracy of 60%. The pseudo-amino acid composition was introduced in **chapter 7** with regard to the prediction of the subcellular location of proteins (Chou, 2001, Chou and Cai, 2003). The first 20 attributes computed by the developed method reflect the effect of the amino acid composition whereas the following attributes correspond to the physicochemical differences (hydrophobicity, hydrophilicity and side-chain mass) between sequential pairs of residues (i, $i+j$ where j = 1,2,3...). Cai and Lin (Cai and Lin, 2003) combined this method with a SVM to predict rRNA-, RNA- and DNA-binding proteins obtaining an accuracy of 76%-97%. SVM-Prot (Cai et al., 2003), a SVM based prediction tool, was implemented to classify functional families obtained from the Pfam database (Bateman et al., 2002). SecretomeP (Bendtsen et al., 2004) was implemented to predict mammalian secretory proteins targeted to the non-classical secretory pathway. This algorithm used a trained neural network with a single layer of hidden neurons that were trained by features such as predicted post-translational modifications, predicted structure, degradation signals, composition, size and charge. This method achieved a sensitivity of 40% with a false positive rate lower than 5% by five fold cross-validation. Chou and Cai (Cai and Chou, 2005, Chou and Cai, 2004) developed various algorithms to predict the enzyme family class and subclass under low sequence similarity (>40%). The GO-PseAA predictor (Chou and Cai, 2004) combined the gene ontology and pseudo amino acid

composition to predict the different enzymatic classes. The Fund-PseAA predictor (Cai and Chou, 2005) combined the composition of functional domains with the pseudo amino acid composition to predict different enzymatic subclasses. Similarly, the Fund-PseAA predictor was also applied to predict different types of proteases with low sequence similarity (<25%), obtaining an overall accuracy over 90%. Lin and colleagues (Lin et al., 2006) trained a SVM to predict different classes of lipid binding proteins. The extracted features included amino acid composition, hydrophobicity, normalized Van der Waals volume, polarity, polarizability, charge, surface tension, secondary structure and solvent accessibility. Additionally other features were also included: i) the composition of equivalent residues (e.g. hydrophobic residues), ii) the transition frequency by which a particular group of equivalent residues is followed by a different group of equivalent residues and iii) the sequence length at which the first 25, 50, 75 and 100% of the residues belonging to a equivalent group of residues are found. The developed method showed sensitivity values from 76% to 91% and specificity values from 97% to 100%.

8.1.6 Functional prediction of membrane proteins

Membrane proteins are by nature different to globular membrane proteins. The environment in which membrane proteins are embedded imposes a compositional constraint in the transmembrane regions of the protein. Following this principle, functional prediction methods based on features that can be extracted from sequence, such as the amino acid composition and different physicochemical properties, might also reflect the differences between globular and membrane proteins. Sugiyama and colleagues (Sugiyama et al., 2003) stated that sequence similarity based methods are less accurate in predicting the molecular function of membrane proteins probably due to the hydrophobic nature of the transmembrane regions. This compositional bias of membrane proteins might affect different methods of predicting molecular function to varying extents. While methods based on features extracted from sequence can be severely affected, pattern discovery methods can be variably affected depending on the location of the functional domain. If the functional domain is located in an extra-membranous loop, it is likely that it resembles a homologue domain located in a globular protein. However, if the functional domain is

located in the transmembrane domain, it is likely that this domain evolves in order to adapt itself to the new environment and the interacting hydrophobic secondary structures of the protein. Therefore, functional prediction methods should be independently trained by two different sets of proteins, one corresponding to globular proteins and a different set composed of membrane proteins, in order to overcome the compositional differences between these two classes.

Despite the importance of membrane proteins in many physiological processes, there are only a few methods to specifically predict the function of membrane proteins. Shimizu and colleagues developed a method to predict the function of transmembrane proteins based on a binary topology pattern (Polulyakh et al., 2000, Sugiyama et al., 2001, Sugiyama et al., 2003, Inoue et al., 2001). The authors combined different topological properties to generate binary vectors: the number of transmembrane regions, loop length (either short or long) and the location of the N-terminus. The developed method showed accurate self-consistency values but it was not appropriately evaluated. The authors also developed a single-linkage clustering method based on a binary topology pattern similarity (Arai et al., 2004). The developed binary topology pattern was simply based on the length of the different loops compared with a particular threshold ("0" and "1" corresponded to short loops and long loops respectively). The clustering method was applied to 87 prokaryotic transmembrane proteomes and was capable of predicting 61% of the proteome whereas homology based methods could only annotate 24% of the proteome. Apart from the binary topology approach introduced by Shimizu and collaborators, remaining research has focussed on the functional prediction of membrane proteins based on the classification and sub-classification of G-protein coupled receptors (GPCR). The GPCR superfamily stands out from other functional types of membrane proteins due to its great pharmacological importance. GPCRs have an essential role in cellular signalling of diverse physiological processes such as neurotransmission, cellular metabolism, secretion, cellular differentiation, growth, inflammatory and immune responses, smell, taste and vision (Hebert and Bouvier, 1998). These receptors bind a wide range of different ligands (e.g. protons, amines, peptides, ions and pheromones), which activates a conformational change of the receptor that ultimately transduces the signal across the membrane. The signal is then

coupled to the cytoplasm by association of the inactive heterotrimeric G-protein (subunits α, β and γ) with the receptor. Such association triggers the substitution of the GDP bound to the α-subunit with GTP. Subsequently the α-subunit dissociates from the receptor and the βγ complex. The dissociated subunits can then act as activators or inhibitors of a variety of effectors (e.g. adenylate cyclase or ion channels). The GTPase activity of the α-subunit promotes the re-association of the subunits and the inactivation of the G-protein by hydrolysis of the bound GTP to GDP. Despite the divergence in ligands bound and processes involved, the GPCR superfamily has a marked structural homology. These proteins show a conserved structural arrangement in the transmembrane domain composed by seven α-helices that completely traverse the membrane. The importance of these proteins is evidenced by the fact that approximately 40-50% of the current drugs act through GPCRs. Karchin and collaborators (Karchin et al., 2002) first introduced SVM for the classification and sub-classification of the GPCR superfamily. The authors compared the predictive accuracy of the SVM with that of BLAST (Altschul et al., 1997) and profile HMM and concluded that the SVM was a more accurate predictive technique to sub-classify the GPCR superfamily. However, no sequence similarity threshold was applied and the negative class that was used to train GPCR superfamily classifier under-represented polytopic membrane proteins. PRED-GPCR was implemented based on a probabilistic approach that used profile HMM from multiple sequence alignments to derive family signatures but no sequence similarity threshold was applied. The program was tested against a positive and negative test set obtaining a sensitivity of 96% and a specificity of 99.6%, however the sequential relationship between the training and test set is not described. Inoue and collaborators (Inoue et al., 2004) applied the topology binary approach to the GPCR superfamily. The results achieved corresponded to the self-dependency of the method but no appropriate evaluation was carried out. GPCRpred (Bhasin and Raghava, 2004) was implemented based on a SVM where the extracted features corresponded to the dipeptide composition of the polypeptide sequences (sequence similarity threshold = 90%). The program achieved 99.5% accuracy for predicting GPCR proteins (5 fold cross-validation). However, the negative class of the corresponding training set was found to under-represent polytopic membrane proteins. GPCRpred was also designed to sub-classify the predicted GPCRs into sub-classes at various levels showing

accurate predictive scores. Guo and collaborators (Guo et al., 2005) designed a method based on a fast fourier transform-based SVM using hydrophobicity to discriminate between different GPCR subfamilies. The method achieved an overall accuracy of 93% to sub-classify GPCR classes B, C, D and F using a jackknife test although no sequence similarity threshold was applied. GPCRsclass (Bhasin and Raghava, 2005) was implemented to predict amine type GPCRs (it belongs to the class A GPCR) using a SVM that examines the dipeptide composition. Interestingly, rather than using a negative class composed of non-Amine GPCRs, the authors used a set of globular proteins as the negative class used to train the Amine GPCR classifier. The method sub-classified the Amine GPCR sub-class with an overall accuracy of 96%. However, no sequence similarity threshold was applied to the training set prior to the data mining process. The methods described above have been implemented to predict GPCR proteins according to the ligand-based GPCRDB classification scheme (Horn et al., 1998). However, other methods have been implemented to predict GPCR coupling specificity to G-proteins based on techniques such as HMM (Sgourakis et al., 2005, Sreekumar et al., 2004), Naïve Bayesian methods (Cao et al., 2003) and SVM (Guan et al., 2005).

To our understanding, few of the data mining approaches described above included permissive conditions that might have resulted in overestimation of the predictive accuracy of the developed methods. The first permissive condition is the development of negative training sets that do not include or under-represent polytopic membrane proteins. In order to predict the GPCR superfamily, SVMs need to be trained with a positive (GPCRs) and a negative class (non-GPCRs). The negative class should only include non-GPCR polytopic membrane proteins in order to avoid that classifiers base their prediction upon features describing environmental differences rather than structural or functional properties. If the negative class only includes, or is mainly composed of, globular proteins the predictive algorithm will be based more upon properties that reflect the constraints imposed by the membrane in the transmembrane regions. On the other hand, if the negative class is composed of other polytopic membrane proteins the features used to predict GPCRs will reflect structural or functional properties more specific to the GPCR superfamily. The second permissive condition is the absence of a sequence similarity threshold. Data sets

containing large subsets with highly identical sequences tend to bias the classifier towards these subsets in order to maximize accuracy of the overall prediction. This would imply that sequences not belonging to any of these subsets are underestimated by the trained classifier. Therefore, sequence similarity thresholds need to be applied in order to minimize such bias.

8.1.7 The TMFUN approach

This chapter describes the implementation of an algorithm to predict the molecular function of polytopic α-helical membrane proteins by combining different data mining techniques. The data set was extracted from the Swiss-Prot database (release 50.2 of 27-06-2006) using the PROCLASS software **(Chapter 3).** Subsequently, the assembled data set was filtered at a sequence similarity threshold of 40% and 90% using CD-HIT (Li et al., 2001). Prior to the feature extraction stage, TMLOOP and TMLOOP writer **(Chapter 5)** were applied to the filtered data sets (two data sets were created after the sequence redundancy stage) in order to refine the transmembrane statement by including the description of membrane dipping loops. The feature extraction was performed by TMDEPTH **(Chapter 6)**, which combines sequence and topology and converts each sequence into a feature vector. The feature vector describes the associations of pairs of residues located at a similar depth in the membrane. Using the Weka platform (Witten and Frank, 2005) different data mining techniques (e.g. Bayesian methods and support vector machines) and architectures were evaluated by ten fold cross-validation in order to maximize the predictive accuracy for each functional class. The different classifiers selected during the data mining stage were combined into a predictive algorithm named TMFUN. This program performs functional prediction of membrane proteins at increasing levels of molecular complexity. At the first level proteins can be classified as enzymes, GPCRs, ion channels and molecular transporters. Further levels of prediction distinguished between ion channels specifically transporting cations or anions, class A GPCRs, non-class A GPCRs, amine GPCRs, peptide GPCRs, olfactory GPCRs and rhodopsin GPCRs. At the sequence similarity threshold of 40%, TMFUN predicted enzymes, GPCRs, and transporters with a sensitivity of 64.1%, 87.5% and 71.4% respectively **(table 8.53)**. The

ion channel class could not be accurately predicted as the algorithm tended to misclassify proteins with transport activities such as ion channels, molecular transporters and particular enzymes. At the most informative level, TMFUN predicted 70% of the olfactory GPCRs under conditions of low sequence similarity using a sequence similarity threshold of 40%. At the 90% sequence similarity threshold, TMFUN achieved higher predictive accuracies. Enzymes, GPCRs, ion channels and molecular transporters were predicted with a sensitivity of 87.8%, 92.8%, 58.3% and 75.6% respectively **(table 8.57)**. At the most informative level the different subclasses of class A GPCRs were predicted at sensitivity values of 84.5%-92.9%. Unlike previously described methods, no globular proteins have been included in the training sets in order to avoid the effect of inherent compositional differences between polytopic α-helical membrane proteins and soluble proteins. In order to fully evaluate the results described above it is necessary to consider the current context in which such predictions are carried out. The applied feature extraction method depends directly on the accuracy of the topological models contained in the Swiss-Prot database. The incorrect prediction of even a single transmembrane region can completely alter the protein vector used to classify a given membrane protein **(figure 6.4)**. Currently, topology prediction methods achieve an accuracy of only 70%-80% in correctly predicting the topology of polytopic α-helical membrane proteins. Additionally, the predictions carried out by TMFUN are solely based on the information contained in the transmembrane regions. Therefore, these results confirm the essential role that transmembrane regions play in the assignment of molecular function for polytopic α-helical membrane proteins.

8.2 Methods

8.2.1 Data set development

The data set was retrieved from the Swiss-Prot database (release 50.2 of 27-06-2006) using PROCLASS **(Chapter 3)** regardless of their taxonomic classification. The data set assembled with PROCLASS contained 10,022 α-helical membrane proteins (9,907 proteins were clustered in 808 clusters whereas 115 non-clustered proteins were manually retrieved) classified according to their corresponding functional annotation space **(table**

3.4). The obtained data set was filtered at the protein and functional level, as described below, in order to develop a data set suitable for analysis with TMDEPTH and the data mining process. At the protein level, both structural and sequence redundancy were analyzed. Membrane proteins whose structure was found to be composed of β-sheet forming a β-barrel structure were not to be included in the data set as TMDEPTH was implemented to detect associations of residues belonging to different α-helical structures located at a similar depth in the membrane. The estimation of the membrane thickness and depth values for the different amino acids is not applicable to β-barrel membrane proteins due to their different structural properties. In order to remove β-barrel membrane proteins, Swiss-Prot like text files containing the term "porin" (used to describe a β-barrel membrane proteins) were excluded from the data set. Predicted membrane dipping loops (only those found to be true positives) **(Chapter 5)** were included in the transmembrane statement of corresponding Swiss-Prot like text files using the TMLOOP writer **(Chapter 5)**. As explained in **Chapter 7**, TMDEPTH was implemented to consider transmembrane regions whose length is lower than 14 residues long as false positives, ignoring such segments while retrieving transmembrane information from the Swiss-Prot database. Although it is possible that those predicted segments are false positive, it is also possible that the topology prediction underestimated the length of the given segment or that the predicted segment corresponds not to a helical structure but to an unstructured loop or loop-helical structure that completely traverses the membrane. In order to minimize potential errors when calculating the matrices using TMDEPTH, those proteins containing at least one transmembrane region with less than 14 residues were excluded from the data set. Sequence similarity was also analyzed within each cluster using CD-HIT (Li et al., 2001). As the size of the assembled data set was large enough to apply a stringent filter using CD-HIT two different sets were designed, the first set was composed of sequences not sharing more than 40% of its residues with another sequence in the same cluster (CD-HIT parameters: n = 2, c = 0.4) whereas a more flexible filter was applied to the second set (CD-HIT parameters: n = 5, c = 0.9) where no clusters contained 2 or more proteins sharing at least 90% of their residues.

At the functional level, protein complexes and multifunctional proteins were not considered. TMDEPTH has also been implemented to calculate the percentage of interhelical associations between residues located at a similar depth between different subunits within protein complexes. However, protein complexes were discarded because there is not yet full description of the different subunits forming such complexes (unless they have been experimentally analyzed). Theoretically, the designed data mining workflow and the TMFUN algorithm is compatible with the multifunctional behaviour of proteins but in order to capture the profile of particular molecular activities (including molecular function and ligand-binding specificity) it is more appropriate to use membrane proteins with a single functional behaviour during the data mining process. From the data set assembled with PROCLASS, some functional clusters whose molecular activity needed further exploration were also analyzed, namely adhesion proteins, chloroplast envelope proteins, tetraspanin proteins and photosynthetic proteins. Proteins belonging to the chloroplast envelope cluster and photosynthetic cluster were not considered because no defined molecular activity has been defined yet or because a large fraction of these proteins belong to protein complexes (e.g. photosystems I and II). Adhesion proteins were not considered due to the small size of this cluster (38 sequences where more than 75% of those share more than 90% of their residues). Tetraspanin proteins carry out a particular function acting as linkers in the membrane organizing other proteins into a network of multimolecular membrane microdomains (also known as tetraspanin web). Unfortunately, as with adhesion proteins, the size of this unique protein cluster was not large enough for the analysis during the data mining process. After filtering at protein and functional level using a sequence redundancy threshold of 40% the assembled data set contained 1,663 α-helical membrane proteins **(table 8.1)** whereas if a sequence redundancy threshold of 90% was applied using CD-HIT the assembled data set contained 4,470 α-helical membrane proteins **(table 8.2)**.

Level 1	Level 2	Level 3
Enzyme: 806	EC 1: 70	EC 1.1: 16
		EC 1.14: 24
		Other: 30
	EC 2: 396	EC 2.3: 107
		EC 2.4: 92
		EC 2.7: 177
		Other: 20
	EC 3: 309	EC 3.1: 28
		EC 3.4: 74
		EC 3.6: 208
	EC 4: 18	-
	EC 6: 13	-
GPCR: 241	Frizzled GPCR: 8	-
	GPCR class A: 192	Amine: 45
		Olfactory: 37
		Peptide: 61
		Rhodopsin: 19
		Other: 30
	GPCR class B : 14	-
	GPCR class C : 26	-
Receptor: 69	Acetylcholine receptor: 2	-
	Serpentine: 67	-
Ion channel: 269	Anionic channel: 49	-
	Cationic channel: 203	-
	Other: 17	-
Transporters: 278	Amino acid transporter: 84	-
	Sugar transporter: 80	-
	Other: 114	-

Table 8.1. Assembled data set using PROCLASS and filtered at a sequence similarity threshold of 40%. Molecular function is defined at various levels of molecular complexity. Level 1 corresponds to the less informative definition of molecular function whereas higher levels describe molecular function in more detail. The numbers within each cell ***are*** the number of proteins with the corresponding function.

Level 1	Level 2	Level 3	Level 4
Enzyme: 1904	EC 1: 165	EC 1.1: 34	-
		EC 1.3: 27	-
		EC 1.8: 22	-
		EC 1.14: 62	Fatty acid desaturase: 23
			Other: 39
		Other: 20	-
	EC 2: 932	EC 2.3: 205	Apolipoprotein N-acyltransferase: 60
			Palmitoyl transferase: 95
			Other: 50
		EC 2.4: 264	Alpha-1,2-glucosyltransferase: 20
			Chitinsynthase: 32
			Prolipoprotein diacylglyceryl transferase: 139
			Other: 73
		EC 2.7: 413	Phosphatidate cytidyltransferase: 25
			Phospho-N-acetylmuramoyl-pentapeptide-transferase: 149
			Cobalamin synthase: 50
			Histidine kinase: 109
			Cardiolipin synthetase: 23
			Other: 57
		Other: 50	
	EC 3: 757	EC 3.1: 63	
		EC 3.4: 168	
		EC 3.6: 526	Pyrophosphate-energized proton pump: 27
			Undecaprenyl-diphosphatase: 157

			Copper ATPase: 17
			Proton ATPase: 21
			Calcium ATPase: 29
			Other ATPase: 20
			Multidrug resistance ABC transporter 58
			Lipid export ABC transporter: 30
			Other ABC transporter: 167
	EC 4: 26	-	-
	EC 6: 24	-	-
GPCR: 1156	Frizzled GPCR: 26	-	-
	GCPR class A: 1007	Amine: 144	Adrenergic receptor: 36
			Dopamine: 22
			Serotonin: 37
			Trace amine: 20
			Other: 29
		Hormone: 32	-
		Nucleotide like: 40	-
		Olfactory: 432	-
		Peptide: 221	CC Chemokine: 30
			Melanocortin: 21
			Other chemokine: 43
			Other: 127
		Rhodopsin: 95	Chromophore: 61
			Opsin: 34
		Other: 73	-
	GPCR class B: 43	-	-
	GPCR class C: 82	Taste: 63	-
		Other: 19	-
Receptor: 116	Acetylcoline: 5	-	-
	Serpentine: 111	-	-
Ion channel: 576	Anionic channel: 134	Chloride: 31	-
		Phosphate: 35	-
		Sulfate: 20	-

		Other: 48	-
	Cationic channel: 387	Iron: 94	-
		Magnessium: 22	-
		Potassium: 67	-
		Sodium/Proton antiporter: 29	-
		Zinc: 55	-
		Other: 120	-
	Other: 58	Cation/Anion antiporter: 33	-
	Mechanosensitive channel	25	-
Molecular transporter: 718	ADP/ATP translocase: 27	-	-
	Amino acid: 163	-	-
	Amonium: 24	-	-
	Aquaglyceroporin: 74	-	-
	DNA translocase: 61	-	-
	Multidrug: 43	-	-
	Oligopeptide: 31	-	-
	Sodium dicarboxylate: 30	-	-
	Sugar transporter: 198	-	-
	Other: 67	-	-

Table 8.2. Assembled data set using PROCLASS and filtered at a sequence similarity threshold of 90%. Level 1 corresponds to the less informative definition of molecular function whereas higher levels describe molecular function in more detail. The numbers within each cell ***are*** the number of proteins with the corresponding function.

Both sets (those with a sequence similarity threshold of 40% and 90%) are composed of 5 different large functional classes: Enzymes, GPCR, receptors, ion channels and transporters. GPCR proteins are a distinct type of receptor, this large protein superfamily is composed by integral α-helical membrane proteins with seven transmembrane helices and was regarded as a functional class by itself due to the conserved topology of the receptor among subfamilies and its pharmacological importance (more than 50% of current drugs target membrane proteins). Likewise, the fact that the smaller size of

the wider receptor class compared with that of the GPCR class can bias the profiling of a combined group towards GPCR activity did not encourage the analysis of a receptor class, which would encapsulate GPCR and non-GPCR proteins. However, it was found that the receptor family was composed mainly by serpentine receptors (97.1% and 95.7 for the set filtered at a threshold of 40% and 90% respectively). Therefore, the receptor class was discarded because rather than detecting profiles for general receptor activity only the signature corresponding to serpentine receptors would have been detected.

As with the data set development for the prediction of subcellular location of eukaryotic α-helical membrane proteins, TMDEPTH was used to calculate the corresponding matrices for each of the proteins contained in the different classes using the sequence and topological information described in the local Swiss-Prot like text files. TMDEPTH was required to save the computed interhelical associations of residues located at a similar depth (percentage of amino acid participation in interhelical associations, the normalized 20x20 triangular matrix of interhelical associations and the normalized 3x3 triangular matrix of interhelical associations of clusters of biochemically equivalent residues) in C4.5 format, which can be processed not only by C4.5 and its latest Windows version (See5) but also by the Weka platform. The data sets filtered using a sequence similarity threshold of 40% and 90% were both composed of four different classes (Enzymes, GPCRs, ion channels and molecular transporters) but the set filtered with a sequence similarity threshold of 40% was composed of 1,594 data points whereas the more flexible set filtered at a sequence similarity threshold of 90% was composed of 4,354 data points. Each data point corresponds to a single protein and is composed of 236 attributes where the first 20 attributes correspond to the percentage of interhelical association participation for each residue, the following 210 attributes correspond to the normalized 20x20 triangular matrix of interhelical associations and the last six attributes correspond to the normalized 3x3 triangular matrix of interhelical associations of clusters of biochemically equivalent residues.

Both data sets were also filtered using the attribute selection filtering toll built within the Weka platform, as described in **chapter 7**. The different data mining methods and architectures (see below) were applied to both the non-filtered data sets and the filtered data sets to maximize the accuracy prediction of a given data mining tool.

8.2.2 Development of the data mining workflow

The Weka platform (Witten and Frank, 2005) was used to design and evaluate different data mining analyses carried out in a similar fashion, as described in **chapter 7**. Both single-step architectures and multilayer architectures (also known as multi-step or tree-based architectures) **(figure 7.1)** were designed and evaluated using different data mining methods **(table 7.4)** in order to maximize the predictive accuracy for a given data set. The single-step architectures involved two different variations: the all-against-all method and one-against-all method. The all-against-all applies a particular data mining technique for the prediction of various classes in a single-step (also known as multi-class classifier) fashion. Data mining techniques based on the one-against-all architecture can only be applied to predict a particular class, so if a given data set is composed of 3 different classes it is necessary to design 3 different classifiers (e.g. for the classification of enzymes, the sets filtered with a sequence similarity threshold of 40% and 90% similarity were re-classified into two classes: the enzyme class and the non-enzyme class, which comprises GPCR proteins, ion channels and molecular transporters). This second variation is considered as a special type of single-step architecture as it depends on the final arrangement of the classifiers in a predictive software. If the different classifiers are arranged in a single layer and the final prediction is based upon a consensus prediction the one-against-all variation remains a variation of single-step architecture **(figure 8.1a)**. However, if the classifiers are arranged sequentially giving preference to the earlier classifiers, the given variation no longer belongs to the single-step architecture but becomes a classical example of multi-step architecture **(figure 8.1b)**. Due to the functional relationship between ion channels and molecular transporters, both classes were also classified according to the multi-step architecture where first proteins with any kind of transport activity were predicted and in a subsequent node α-helical membrane proteins

with transport activity were more specifically classified as ion channels or molecular transporters **(figure 8.2)**.

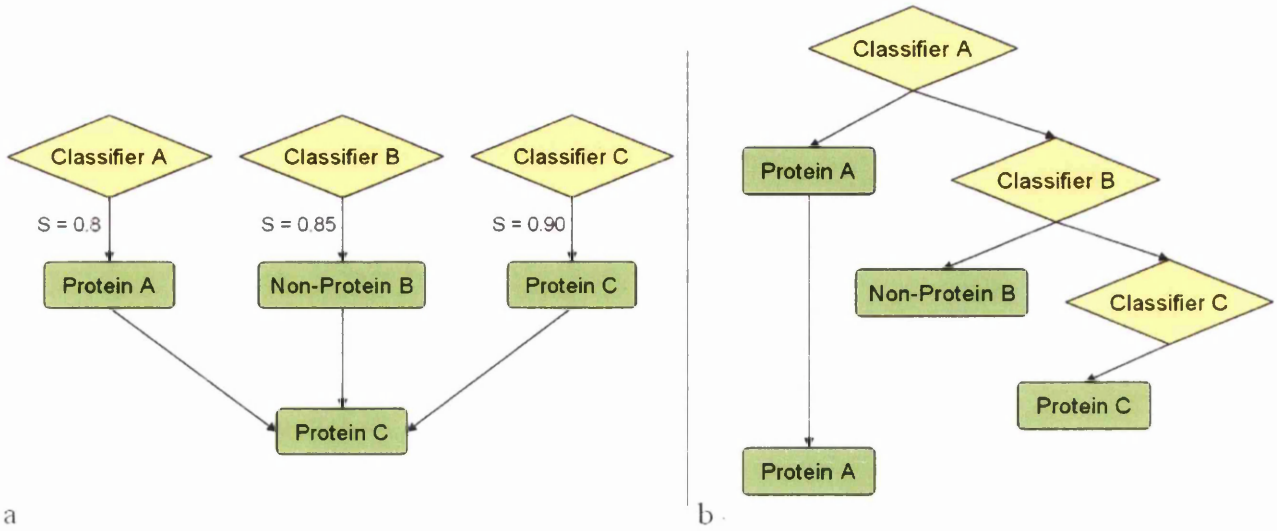

Figure 8.1. Example of different predictive architectures based on the re-arrangement of the classifiers: a) when the classifiers are arranged in a single layer a consensus prediction (either based on the corresponding support for each classifier or using equally weighted classifiers); b) if the same classifiers are arranged following a tree-based model, earlier classifiers are given preference over later classifiers.

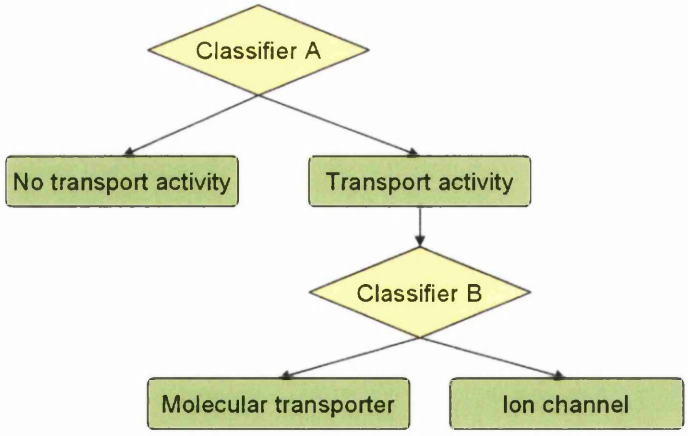

Figure 8.2. Multilayer predictive architecture for the prediction of molecular transporters and ion channels. The earlier classifier is designed to distinguish between proteins with and without any kind of transport activity. The later classifier sub-classifies membrane proteins with transport activity as molecular transporters or ion channels.

The main objective of the developed research in this chapter is to prove that the TMDEPTH method **(Chapter 6)** can be used to demonstrate that signatures derived from the predicted topology of membrane proteins can be associated with specific molecular functions of membrane proteins. As described in **tables 8.1** and **8.2** the molecular function of a protein can be defined at differing levels of complexity. Though the data mining workflow can be applied at the various levels described, the ending of the data mining workflow was dictated by the predictive accuracy of a given class using the data set filtered at a sequence identity threshold of 40%. When low accuracy was obtained for predicting a particular class at a particular level, no further data mining techniques were applied at higher levels of complexity of molecular activity. Although data mining analysis using the more flexible data set filtered at a sequence identity threshold of 90% is likely to give equal or higher accuracy values for a given class at a particular level, no data mining of further levels was applied, in order to focus on the comparison of the predictive method based on a heterogeneous set (where the sequence similarity levels for each class are close to the twilight zone) and a more representative set filtered at a sequence identity threshold of 40 and 90% respectively.

8.2.3 Classifier evaluation

Each data mining technique was evaluated by ten fold cross-validation. When evaluating the different classifiers applied in a single-step fashion, the 10 fold cross-validation of the single node is effectively an evaluation of that particular method to predict the functional classes. However, evaluation of a set of classifiers arranged in a tree-based fashion can not be achieved by summing up the independent evaluations of each node contained in a given tree. In order to evaluate a tree-based set of classifiers, it was necessary to train the corresponding classifier at each node with the subset of the training set containing the corresponding classes and to apply the trained tree-based set of classifiers to each data point in the test set until reaching a leaf where no further classification can be achieved.

The predictive accuracy of each data mining technique and tree-based set of classifiers is estimated based on the confusion matrix obtained from the ten fold cross-validation **(Chapter 7)**. The accuracy in predicting a particular class is defined by its sensitivity **(1.16)**, specificity **(1.17)** and geometric average **(1.18)**. The evaluation of the performance of the method is given by the accuracy **(1.19)**, the normalized accuracy **(1.20)** and the Matthews correlation coefficient for data sets with two classes **(1.21)** or the Generalized correlation for data sets with three or more classes **(1.22)**. These predictive values are compared between the different data mining techniques **(table 7.4)** in order to select the most accurate data mining technique for predicting a particular functional class.

8.2.4 TMFUN development

As with previous software implementation, TMFUN has been implemented using the Borland Delphi 7 programming environment and the model-view-controller architecture **(figure 3.3)**. The current version **(figure 8.3)** has been implemented as a web application, which uses a Common Gateway Interface (CGI), with an Apache web server. The code generated by Delphi is a console application, which is placed in the Apache server and called by submission of input at the TMFUN interface by HTML code.
The predictive algorithm has been implemented to combine the different classifiers selected during the classifier evaluation process. The implemented algorithm combines different predictive architectures and performs consensus prediction at each level. Based on the consensus prediction achieved at a particular level of molecular function complexity, the algorithm calls the corresponding classifier to be applied at the subsequent level and further sub-classifies the query sequence. TMFUN has also been designed not to assign a functional molecular class by default. However, the current version has not been implemented to predict multifunctional proteins.

The current web version of TMFUN has been designed to so that the required information (identifier, topology information and amino acid sequence) is input in Swiss-Prot like format. Once the input has been submitted, TMFUN calls TMDEPTH in order to perform the feature extraction **(Chapter 6)** and create a protein vector with 236 attributes,

which summarizes the associations of residues belonging to different transmembrane helices but located at a similar depth. Each of the developed vectors is then tested using the Weka platform using the previously trained classifiers for the lower level of molecular function complexity. TMFUN communicates with the Weka platform with a series of bat files (*.bat) where a specific command is run in order to load the corresponding classifier, load the corresponding test vectors, apply the classifier to the loaded test vectors and save the prediction as a specific text file located in a particular location. Once all classifiers belonging to the 1st level have been run and the corresponding predictions have been analyzed by TMFUN, a consensus prediction is performed to predict the molecular function of the query proteins at its lowest level of complexity. TMFUN performs three different types of consensus prediction. The first consensus prediction method assumes equally weighted classifiers, whereas the second and third consensus prediction methods assume un-equally weighted classifiers where the support for each classifier is obtained from the ten fold cross-validation performed during the evaluation of the classifier, and corresponds to the geometric average (GAv) and the Matthews correlation coefficient respectively (MCC). The former consensus prediction method that assumes equally weighted classifiers has been designed to include up to two different hits (e.g. "the query sequence belongs to an enzyme or ion channel") but when more than two different hits have been made by different classifiers TMFUN does not make any prediction (e.g. "the molecular function of the query sequence can not be assigned"). Similarly, in order to predict the transport activity of the query protein three different classifiers are called (**figure 8.5**). The 1st and 2nd classifier have been design to predict ion channels and molecular transporters respectively, the 3rd classifier is a tree-based set of 2 classifiers where the 1st classifier predicts whether the query protein has transport activity and the 2nd classifier sub-classifies the query as ion channel or molecular transporter (**figure 8.2**). When these 3 classifiers give a hit for the query sequence, assuming equally weighted classifiers, the consensus prediction relies on the prediction with the higher number of hits. By contrast, if two contradictory hits are obtained (e.g. 1st classifier, 2nd classifier and 3rd predict the query sequence as non-ion channel, molecular transporter and ion channel respectively), TMFUN can only indicate the transport function of the query protein without indicating whether it is an ion channel or molecular transporter. Consensus prediction assuming unequally weighted classifiers does

not lead to the situations described above, as there are never two different classifiers (acting at the same level) with the same support.

Depending on the consensus prediction at the first level, TMFUN calls the corresponding classifier to further sub-classify the query sequence at the second level of molecular function complexity, the classification process continues until no further sub-classification can be performed. Based on the results obtained from the classifier evaluation using the data set filtered at a sequence similarity threshold of 40%, TMFUN has been implemented to predict the molecular activity of the query sequence at three different levels of molecular function complexity **(figure 8.5)**. At the first level, TMFUN distinguishes between enzymes, GPCRs, ion channels, molecular transporters and proteins with transport activity. At the second level of complexity, TMFUN sub-classifies GPCR proteins into class A GPCR proteins or other GPCR proteins, and ion channels are also sub-classified into cation channels or anion channels. The 3rd level of molecular function complexity has been designed to further subclassify class A GPCR proteins into amine, olfactory, peptide and rhodopsin GPCR proteins.

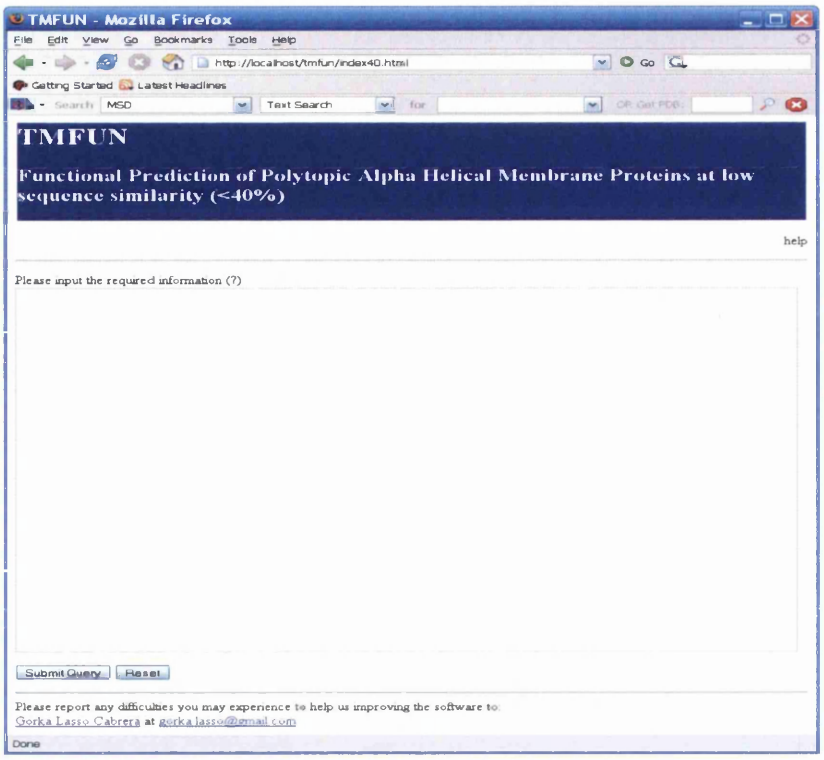

Figure 8.3. Screenshot of TMFUN for prediction of the molecular function of α-helical membrane proteins under low sequence similarity.

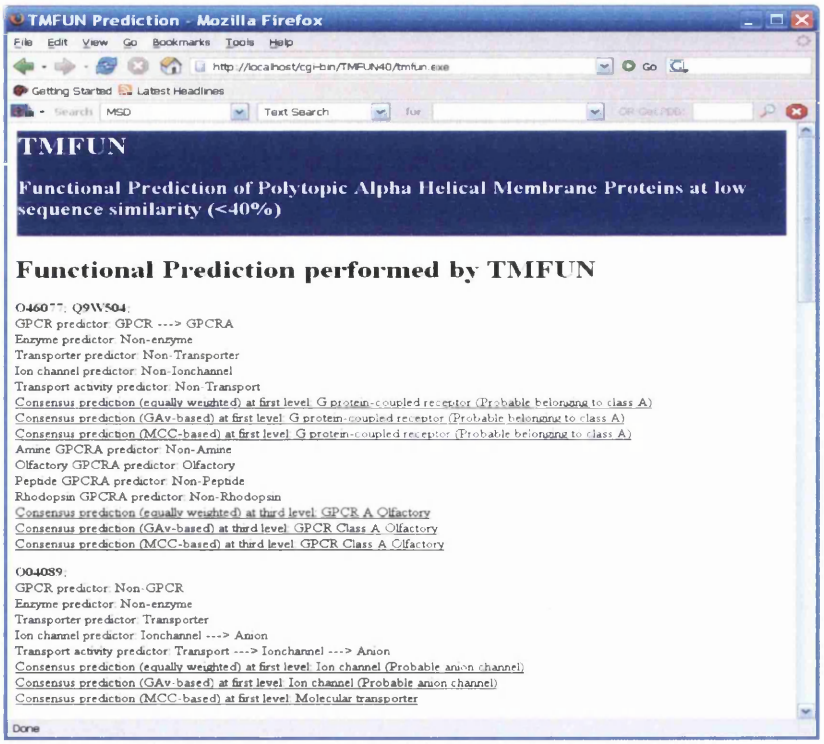

Figure 8.4. Output interface of the current web version of TMFUN.

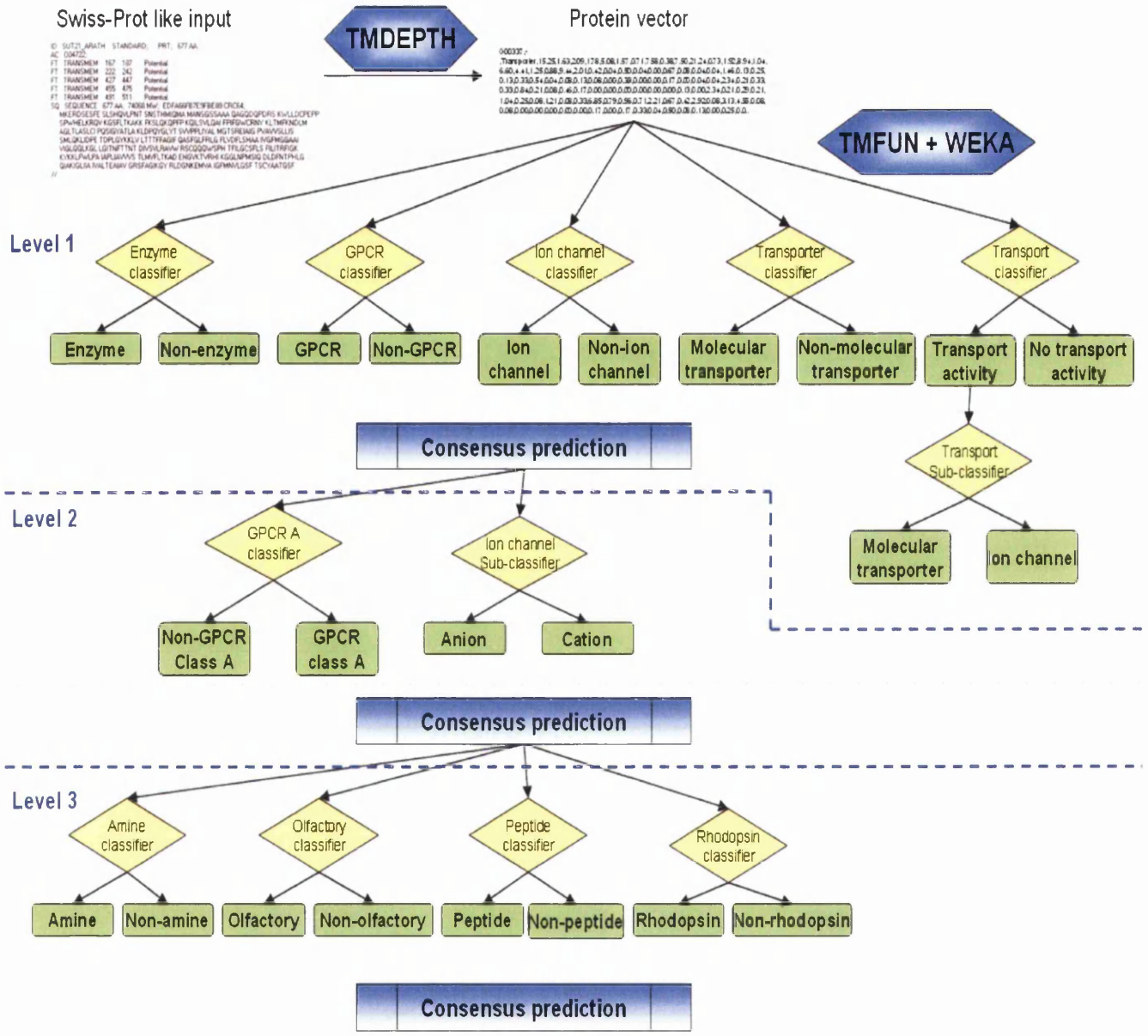

Figure 8.5. TMFUN algorithm. The current version of TMFUN requires an input with Swiss-Prot like format indicating the identifier, the transmembrane regions and the amino acid sequence. TMFUN calls TMDEPTH, which extracts the features corresponding to the percentage of pairs of residues located at a similar depth in the membrane and creates a protein vector to be analyzed by TMFUN and WEKA. TMFUN uses 12 different classifiers to perform the prediction of molecular function at three levels of complexity. Within each level, all classifiers are arranged in a single-step fashion with the exception of the transport classifier in the first level, which is composed of two different classifiers arranged in a tree-based fashion.

8.2.5 TMFUN evaluation

Although each classifier has been individually evaluated, TMFUN combines 12 different classifiers in order to obtain a single prediction of the molecular function. Therefore the program needs to be evaluated in order to estimate the overall accuracy prediction of the algorithm. TMFUN was evaluated by ten fold cross-validation, where iteratively corresponding classifiers are trained with 90% of the corresponding data set and the remaining 10% is used as a test set. The evaluation is repeated ten times so each protein has been included once in the test set.

8.3 Results and Discussion

8.3.1 Data set development

The data set assembled using PROCLASS was composed of 10,022 α-helical membrane proteins clustered into 808 different clusters according to the functional annotation details contained in the Swiss-Prot database. If sharing broad functional properties, the obtained clusters were merged at different levels of complexity of molecular function (**table 8.1** and **table 8.2**). Following this principle, all clusters corresponding to ion channels transporting different anions were clustered into the anion channel class at a less informative level and also merged with other cation channels constructing the ion channel class at a broader level. Classification of these clusters lead to an interesting question while classifying proteins whose overall molecular activity encapsulates two linked functions, where one function is related to the transmembrane section of the protein and the remaining function is related to the extra-membranous section of the protein. This scenario was particularly evident with hydrolases transporting ions through the membrane such as the calcium ATPases. These proteins catalyze the hydrolysis of a molecule at the extra-membranous side after the transport of a specific ion through the membrane triggers the corresponding conformational change.

TMDEPTH has been designed to extract features that are present in the transmembrane section of the protein. Therefore, it is possible to classify proteins based solely on the molecular activity carried out in the membrane. However, the transmembrane section of the proteins with these properties might be indirectly linked to the function carried out at the extra-membranous side either by triggering conformational change, by transduction of the signal that triggers the conformational change or by stabilization of required conformation. Based on this, proteins carrying out two linked functions related to the transmembrane and extra-membranous section were classified according to their overall molecular activity.

The data set was filtered, removing proteins with β-barrel structure or transmembrane regions whose length is lower than 14 residues, and multifunctional proteins and proteins that belong to protein complexes. The size of the filtered data set was large enough to accommodate more stringent filters at the sequence level using CD-HIT. The so-called "twilight zone" ranges from approximately 20% to 30% sequence similarity. Assessment of functional prediction evaluation at sequence similarity levels close to the twilight zone is the ultimate test for any functional prediction method. The lowest possible sequence similarity threshold that may be applied by CD-HIT corresponds to 40% sequence similarity. Additionally a more flexible sequence similarity threshold was also applied (at 90% sequence similarity) where only highly identical pairs of sequences are avoided. The goals of these two data sets is to prove the principle proposed by TMDEPTH (signatures derived from the predicted topology of membrane proteins can be associated with specific molecular functions of membrane proteins) and to implement two versions of TMFUN, one version to be applied under low sequence similarity and a more flexible version of TMFUN whose classifiers have been trained with a more representative set.

8.3.2 Development of predictive architecture

As with the development of TMLOCATE, a range of data mining techniques **(table 7.4)** has been evaluated in order to select the best performing classifiers. As explained earlier, the end of the data mining workflow was dictated by the predictive accuracy of a

given class using the data set filtered using a sequence identity threshold of 40%. When low accuracy was obtained to predict a particular class at a particular level, no further data mining techniques were applied at increasing levels of complexity of molecular function. Although data mining analysis using the more flexible data set filtered at a sequence identity threshold of 90% is certain to give equal or higher accuracy values for a given class at a particular level, no further data mining was applied in order to compare the predictive method based on a heterogeneous set (where the sequence similarity levels for each class are close to the twilight zone) and on a more representative set i.e. the two sets filtered at a sequence identity threshold of 40 and 90% respectively.

8.3.2.1 Data set filtered at a sequence similarity threshold of 40%

Based on this data set, different classification schemes and multilayer classifiers with different combinations of classes were tested. The first classification was based on three different functional classes, which included enzymes, GPCR proteins and proteins with transport activity (including molecular transporters and ion channels). The best performing classifier was found to be the Bayesian network (attribute selection = true) **(table 8.3)**, which achieved a normalized accuracy (nQ) of 74.3. Proteins with transport activity were subsequently classified as ion channels or molecular transporters by a downstream classifier creating a tree-based or multilayer architecture composed of two different classifiers **(figure 8.2)**. The best performing classifier in the discrimination between ion channels and molecular transporters was the SVM(11) (attribute selection = true), which achieved a nQ value of 73.2 **(table 4)**. The combination of both classifiers **(Please see appendix B figure 8.1 on CD)** was found to correctly predict enzymes and GPCR proteins with a geometric average (GAv) of 71 and 79.5 respectively, however the GAv values for ion channels and molecular transporters was found to be 35 and 54.8 respectively **(table 8.5)**. This first classification attempt showed impressive results for the GPCR class where 89.6% of the GPCR proteins were correctly predicted but it was necessary to further explore different combinations in order to maximize the predictive accuracy of each class.

	Attribute selection			No attribute selection		
Data mining method	**Q**	**nQ**	**GC**	**Q**	**nQ**	**GC**
Bayesian networks	**70.2**	**74.3**	**0.61**	67.7	73.3	0.58
Naïve bayesian	54.7	64.1	0.44	50.8	59.8	0.37
Naïve Bayesian simple	55.3	64.4	0.44	-	-	-
Logistic regression	63.9	61.9	0.46	-	-	-
RBF Network	59.1	61.2	0.45	50.8	33.3	-
Kstar	61.9	60.5	0.46	56.0	56.2	0.42
MultiBoostAB	69.6	68.1	0.58	70.3	65.9	0.56
J48	59.5	58.9	0.42	57.8	56.9	0.38
Random forest	68.1	65.9	0.54	67.5	62.2	0.49
Support vector machine (1)	62	58.3	0.44	63.8	62.1	0.47
Support vector machine (2)	64.2	61.5	0.47	61.9	60.8	0.43
Support vector machine (3)	63.8	61.2	0.46	61.6	60.4	0.43
Support vector machine (4)	64.1	59.3	0.46	66.5	66.6	0.5
Support vector machine (5)	60.6	53.3	0.37	64.4	66	0.48
Support vector machine (6)	69.3	67.4	0.55	74.1	70.6	0.61
Support vector machine (7)	65.3	64.1	0.49	71.2	68.9	0.57
Support vector machine (8)	65.3	64.1	0.49	71.2	68.9	0.57
Support vector machine (9)	64.2	64.4	0.47	66.5	66.6	0.5
Support vector machine (10)	71.6	71.6	0.57	72.4	65.9	0.57
Support vector machine (11)	71.1	67.3	0.57	72.5	66.1	0.57

Table 8.3. Ten fold cross-validation of different data mining techniques applied in a single-step fashion to discriminate between enzymes, GPCR proteins and proteins with transport activity (including molecular transporters and ion channels). The highlighted cells correspond to the best performing data mining technique.

	Attribute selection			No attribute selection		
Data mining method	**Q**	**nQ**	**MCC**	**Q**	**nQ**	**MCC**
Bayesian networks	68.4	68.2	0.37	63.1	62.8	0.27
Naïve bayesian	64.2	63.8	0.3	58.5	58	0.18
Naïve Bayesian simple	63.6	63.2	0.29	-	-	-
Logistic regression	65.4	65.3	0.31	61.6	61.6	0.23
RBF Network	66.5	66.2	0.35	51.7	51	0.03
KStar	62.9	62.7	0.26	61.6	61.3	0.24
MultiBoostAB	67.8	67.9	0.36	66.5	66.6	0.33
J48	63.4	63.6	0.27	59.7	59.7	0.19
Random forest	66.4	66.5	0.33	63.6	63.7	0.27
Support vector machine (1)	67.3	67.1	0.35	61	60.9	0.22
Support vector machine (2)	66	65.8	0.32	63.2	63.2	0.26
Support vector machine (3)	65.8	65.7	0.32	62.1	62.1	0.24
Support vector machine (4)	69.1	68.9	0.39	66.7	66.6	0.33
Support vector machine (5)	66.7	66.4	0.35	64.9	64.8	0.3
Support vector machine (6)	61.4	61.2	0.23	64.3	64	0.3
Support vector machine (7)	63.4	63.3	0.27	63.4	63.2	0.27
Support vector machine (8)	63.4	63.3	0.27	63.4	63.2	0.27
Support vector machine (9)	63.4	63.3	0.27	66.7	66.6	0.33
Support vector machine (10)	63.1	62.8	0.27	61.8	61.4	0.25
Support vector machine (11)	**73.2**	**73.2**	**0.46**	68.9	69.1	0.39

Table 8.4. Ten fold cross-validation of different data mining techniques applied in a single-step fashion to discriminate between molecular transporters and ion channels. The highlighted cells correspond to the best performing data mining technique.

Class	Sensitivity	Specificity	GAv
Enzyme	67.25	74.9	71
GPCR	89.6	70.7	79.5
Ion channel	36.7	34.3	35
Molecular transporter	60.1	48.3	54.8

Table 8.5. Predictive accuracy for each class after combining two different classifiers in a tree-based architecture. The first classifier is a Naïve Bayesian method to distinguish between enzymes, GPCR proteins and transport proteins whereas the second classifier, a support vector machine (c = 1, exp = 15, feat. space normalization = true, γ = 0.0001), sub-classifies proteins with transport activities as ion channels and molecular transporters.

The next analysis was based on the evaluation of a multi-class classifier that could distinguish between the four functional classes in a single step **(Please see appendix B figure 8.2 on CD)**. Cross-validation results showed that the best performing classifiers were the Bayesian network (attribute selection = true), the SVM(6), SVM(7) and SVM(8) **(table 8.6)**. Comparison of the GAv values for each class showed that the SVM(6) was a

slightly better classifier than the remaining classifiers **(table 8.7)**. Comparison of this classifier with the multilayer classifier previously designed (**table 8.5** and **table 8.7**) showed that the multi-class SVM(6) was a better classifier. Despite the improvement obtained with this classifier, prediction of proteins with transport activity (ion channels and molecular transporters) could still be further improved and so more data mining analyses were carried out.

	Attribute selection			No attribute selection		
Data mining method	**Q**	**nQ**	**GC**	**Q**	**nQ**	**GC**
Bayesian networks	62.4	60.2	0.51	59.2	59.2	0.49
Naïve bayesian	43.2	51.7	0.39	46.3	51.6	0.4
Naïve Bayesian simple	43.1	51.6	0.39	-	-	-
Logistic regression	58.2	47.6	0.39	-	-	-
RBF Network	57.8	46.1	-	50.8	25	-
KStar	54.9	51.3	0.41	49.7	48.5	0.38
MultiBoostAB	65.3	53.8	0.48	63.5	49.9	0.45
J48	53.6	47.8	0.34	52	47.2	0.33
Random forest	62.1	50.7	0.44	60.1	47.3	0.4
Support vector machine (1)	56.5	37.7	-	57.9	46.2	0.38
Support vector machine (2)	56.9	39.2	-	54.7	47.7	0.36
Support vector machine (3)	57.1	39.3	-	54.2	47.8	0.36
Support vector machine (4)	62.3	50.7	0.44	61	57.9	0.45
Support vector machine (5)	59.4	48.4	0.39	59.1	57.8	0.44
Support vector machine (6)	63.8	56.3	0.47	**69.3**	**59.2**	**0.53**
Support vector machine (7)	60.7	54.7	0.44	66.5	59.3	0.51
Support vector machine (8)	60.7	54.7	0.44	66.5	59.3	0.51
Support vector machine (9)	59	53.6	0.41	61	57.9	0.45
Support vector machine (10)	66.2	56.2	0.49	67.2	54.4	0.49
Support vector machine (11)	66.7	55.6	0.5	66.9	53.8	0.49

Table 8.6. Ten fold cross-validation of different data mining techniques applied in a single-step fashion to discriminate between enzymes, GPCR proteins, molecular transporters and ion channels. The highlighted cells correspond to the best performing data mining technique.

	Bayesian Network (attribute selection)			SMV(6)			SVM(7) & SVM(8)		
Class	**Sensitivity**	**Specificity**	**GAv**	**Sensitivity**	**Specificity**	**GAv**	**Sensitivity**	**Specificity**	**GAv**
Enzyme	68.1	74.7	71.3	89.1	70.7	79.3	81.3	71.0	75.9
GPCR	88.6	68.9	78.1	62.9	81	71.4	67.5	76.2	71.7
Ion channel	18	33.1	24.4	33.3	57.1	43.6	37.8	50.2	43.6
Molecular transporter	66.1	45.5	54.9	51.6	61.9	56.5	50.5	55.3	52.9

Table 8.7. Predictive accuracy for each class using the two data mining techniques that maximize the accuracy of prediction of molecular activity at the lowest complexity level.

The subsequent data mining analyses (**tables 8.8 - 8.12**) were based on the one-against-all principle where a single class is distinguished from the remaining classes one at a time **(Please see appendix B figure 8.3 on CD).**Therefore, in order to predict a set with four classes, four different classifiers need to be evaluated. Prediction of enzymes showed that the SVM(11) (attribute selection = true), SVM(10) and SVM(11) were the best performing classifiers. Further comparison of these classifiers showed that the SVM(11) with attribute selection was the best performing algorithm as it could discriminate the non-enzyme membrane proteins better than the remaining classifiers **(Please see appendix B table 8.1 on CD)**. The evaluation of classifiers used to predict GPCR proteins showed that the Bayesian network was the best performing classifier achieving an nQ value of 92.3 **(table 8.9)**. The best performing classifiers used to predict molecular transporters were the Bayesian network (attribute selection = true) and the SVM(6), SVM(7) and SVM(8) **(table 8.10)**. Comparison of these classifiers **(Please see appendix B table 8.2 on CD)** showed that the Bayesian network was the best classifier as it predicted the molecular transporters with the highest sensitivity but with a lower specificity. At either sequence similarity threshold, prediction of ion channels was found to be the most challenging prediction for any of the data mining methods used. Evaluation of the classifiers showed that the highest level of Matthews correlation coefficient (MCC) found was no higher than 0.36 **(table 8.11)**. With respect to the ion channels, the accuracy (Q), normalized accuracy (nQ) and Matthews correlation (MCC) were not representative of the best performing classifier, as classifiers with higher levels of nQ and MCC consistently underpredicted ion channels **(Please see appendix B table 8.3 on CD)**. The predictive accuracy of the different data

mining techniques evaluated was manually checked and the Naïve Bayesian method was selected to predict membrane proteins transporting ions through the membrane. This method showed the highest value of sensitivity although the specificity of the method was found to be only 24%. Considering the predictive performance of the different data mining techniques to predict ion channels is rather poor, it is obvious that this classifier will have the lowest support among the classifiers to be used at level 1b **(figure 8.5)**. Therefore, if the ion channel classifier and other classifier predict a particular data point as a hit, the ion channel classifier will never be used as the final predictor because only the classifier with the highest support is applied in the consensus prediction (assuming unequally weighted classifiers). Thus, the only chance of predicting ion channels occurs when no other classifiers but the ion channel classifier detects a hit. Following this principle, in order to maximize the prediction of ion channels the classifier should have the highest sensitivity (even if it is achieved at the expense of a lower specificity). Therefore, the Naïve Bayesian method was chosen to predict the ion channel class. The transport activity class (including molecular transporters and ion channels) was also mined using different classifiers, and among the best data mining methods found were the Bayesian network (attribute selection = true), the MultiBoostAB based on Random forest and several support vector machines **(table 8.12)**. Comparison of these data mining methods **(Please see appendix B table 8.4 on CD)** showed that the Bayesian network with attribute selection was found to be the best performing method.

	Attribute selection			No attribute selection		
Data mining method	**Q**	**nQ**	**MCC**	**Q**	**nQ**	**MCC**
Bayesian networks	71.5	71.5	0.43	69.5	69.6	0.40
Naïve bayesian	62.9	63.3	0.30	63.8	64.2	0.31
Naïve Bayesian simple	62.7	63.1	0.30	-	-	-
Logistic regression	67.7	67.7	0.35	66.2	66.3	0.33
RBF Network	65.9	66.2	0.34	50.2	49.7	-0.01
KStar	61.6	61.7	0.23	60.5	60.6	0.22
MultiBoostAB	71.1	71.1	0.42	73.7	73.6	0.48
J48	63.7	63.7	0.27	64.0	63.9	0.28
Random forest	69.2	69.1	0.39	71.2	71.1	0.43
Support vector machine (1)	68.3	68.4	0.37	67.4	67.5	0.35
Support vector machine (2)	68.8	68.9	0.38	67.0	67.1	0.34
Support vector machine (3)	68.7	68.8	0.38	66.5	66.6	0.33
Support vector machine (4)	73.1	73.2	0.46	69.6	69.6	0.39
Support vector machine (5)	69.9	69.9	0.40	69.1	69.2	0.39
Support vector machine (6)	68.9	68.9	0.38	75.0	74.9	0.50
Support vector machine (7)	71.3	71.3	0.43	73.0	73.0	0.46
Support vector machine (8)	71.3	71.3	0.43	73.0	73.0	0.46
Support vector machine (9)	65.3	65.3	0.31	69.6	69.6	0.39
Support vector machine (10)	70.0	69.9	0.40	75.7	75.5	0.53
Support vector machine (11)	**75.2**	**75.1**	**0.50**	75.4	75.3	0.51

Table 8.8. Ten fold cross-validation of different data mining techniques applied in a single-step fashion to discriminate between enzymes and non-enzymes. The highlighted cells correspond to the best performing data mining techniques.

	Attribute selection			No attribute selection		
Data mining method	**Q**	**nQ**	**MCC**	**Q**	**nQ**	**MCC**
Bayesian networks	**93.2**	**92.3**	**0.77**	88.1	90.4	0.67
Naïve bayesian	83.9	85.1	0.56	63.0	75.0	0.36
Naïve Bayesian simple	83.9	85.1	0.56	-	-	-
Logistic regression	89.2	75.4	0.55	84.8	75.7	0.47
RBF Network	90.3	76.0	0.58	85.1	50.0	-
KStar	90.5	74.9	0.58	90.7	75.0	0.59
MultiBoostAB	92.7	82.7	0.70	91.6	76.5	0.63
J48	89.9	80.3	0.60	86.2	72.2	0.45
Random forest	91.7	80.9	0.66	91.2	75.7	0.61
Support vector machine (1)	89.4	71.3	0.52	89.7	74.4	0.55
Support vector machine (2)	89.4	74.8	0.55	87.1	77.3	0.52
Support vector machine (3)	89.2	74.7	0.54	86.9	77.7	0.52
Support vector machine (4)	90.9	76.8	0.61	87.7	79.2	0.55
Support vector machine (5)	88.6	74.0	0.52	86.0	79.3	0.52
Support vector machine (6)	91.9	80.5	0.66	92.8	79.9	0.69
Support vector machine (7)	90.3	79.9	0.61	92.4	80.7	0.68
Support vector machine (8)	90.3	79.9	0.61	92.4	80.7	0.68
Support vector machine (9)	87.7	78.2	0.54	87.7	79.2	0.55
Support vector machine (10)	92.1	78.3	0.66	91.0	73.8	0.60
Support vector machine (11)	92.4	78.0	0.67	91.1	72.7	0.60

Table 8.9. Ten fold cross-validation of different data mining techniques applied in a single-step fashion to discriminate between GPCR proteins and non-GPCR proteins. The highlighted cells correspond to the best performing data mining techniques.

	Attribute selection			No attribute selection		
Data mining method	**Q**	**nQ**	**MCC**	**Q**	**nQ**	**MCC**
Bayesian networks	**79.0**	**76.3**	**0.44**	69.3	72.2	0.34
Naïve bayesian	63.1	72.0	0.33	53.9	65.1	0.23
Naïve Bayesian simple	62.9	72.0	0.33	-	-	-
Logistic regression	81.5	54.3	0.14	77.4	60.4	0.21
RBF Network	82.3	62.9	0.30	82.5	50.0	-
KStar	74.8	64.5	0.26	72.3	64.9	0.25
MultiBoostAB	85.1	65.7	0.40	84.4	62.1	0.34
J48	79.7	63.8	0.28	79.5	62.8	0.27
Random forest	84.1	65.3	0.37	84.3	62.7	0.35
Support vector machine (1)	82.5	50.0	-	82.0	49.7	-0.03
Support vector machine (2)	82.5	50.0	-	78.6	58.7	0.19
Support vector machine (3)	82.5	50.0	-	78.4	59.3	0.20
Support vector machine (4)	82.7	60.8	0.28	78.6	68.2	0.33
Support vector machine (5)	80.3	64.9	0.30	76.9	70.2	0.35
Support vector machine (6)	82.2	68.0	0.37	86.9	71.4	0.50
Support vector machine (7)	81.4	68.5	0.36	84.4	71.5	0.44
Support vector machine (8)	81.4	68.5	0.36	84.4	71.5	0.44
Support vector machine (9)	77.6	64.9	0.28	78.6	68.2	0.33
Support vector machine (10)	84.8	69.2	0.43	86.5	67.6	0.45
Support vector machine (11)	86.1	65.4	0.43	86.1	64.2	0.42

Table 8.10. Ten fold cross-validation of different data mining techniques applied in a single-step fashion to discriminate between molecular transporters and non-molecular transporters. The highlighted cells correspond to the best performing data mining techniques.

	Attribute selection			No attribute selection		
Data mining method	**Q**	**nQ**	**MCC**	**Q**	**nQ**	**MCC**
Bayesian networks	74.6	63.1	0.23	72.6	63.7	0.23
Naïve bayesian	54.4	64.8	0.22	**60.9**	**60.2**	**0.16**
Naïve Bayesian simple	56.2	65.5	0.23	-	-	-
Logistic regression	82.8	52.3	0.11	77.0	54.4	0.10
RBF Network	83.2	50.0	-	83.2	50.0	-
KStar	74.1	61.7	0.21	75.7	59.4	0.18
MultiBoostAB	84.5	59.9	0.30	84.4	56.6	0.26
J48	78.6	56.7	0.15	75.3	56.9	0.14
Random forest	84.1	59.9	0.29	82.5	54.1	0.15
Support vector machine (1)	83.2	50.0	-	82.9	50.0	0.00
Support vector machine (2)	83.2	50.0	-	80.2	52.6	0.08
Support vector machine (3)	83.2	50.0	-	79.3	51.9	0.05
Support vector machine (4)	82.7	50.3	0.02	77.7	63.3	0.25
Support vector machine (5)	82.5	50.8	0.05	73.1	63.2	0.22
Support vector machine (6)	81.8	65.3	0.32	84.4	61.0	0.31
Support vector machine (7)	80.9	64.0	0.29	82.0	62.0	0.28
Support vector machine (8)	80.9	64.0	0.29	82.0	62.0	0.28
Support vector machine (9)	82.7	56.8	0.21	77.7	63.3	0.25
Support vector machine (10)	84.4	64.8	0.36	84.1	58.8	0.28
Support vector machine (11)	84.6	55.6	0.26	83.8	53.7	0.19

Table 8.11. Ten fold cross-validation of different data mining techniques applied in a single-step fashion to discriminate between ion channel and non-ion channel. The cells coloured in yellow correspond to the best performing data mining techniques. The cells highlighted in red correspond to data mining techniques where the predictive scores are not representative of the predictive power of the method (not normalised). This is probably due to a large class being well predicted whereas the smaller class can not be accurately discriminated.

	Attribute selection			No attribute selection		
Data mining method	**Q**	**nQ**	**MCC**	**Q**	**nQ**	**MCC**
Bayesian networks	**75.0**	**74.6**	**0.48**	71.5	70.6	0.40
Naïve bayesian	60.9	66.9	0.33	55.8	61.6	0.23
Naïve Bayesian simple	61.1	66.9	0.33	-	-	-
Logistic regression	72.6	67.3	0.37	68.9	65.2	0.31
RBF Network	72.0	71.3	0.41	65.7	50.0	-
KStar	65.5	65.1	0.29	59.8	60.3	0.20
MultiBoostAB	75.9	72.4	0.46	78.1	73.6	0.50
J48	69.3	65.5	0.31	68.4	65.3	0.30
Random forest	74.3	70.9	0.42	76.1	72.3	0.46
Support vector machine (1)	72.1	66.3	0.35	71.3	66.4	0.34
Support vector machine (2)	72.2	67.5	0.36	69.8	66.1	0.32
Support vector machine (3)	72.3	67.5	0.36	69.8	66.1	0.32
Support vector machine (4)	74.2	66.6	0.39	73.1	71.7	0.42
Support vector machine (5)	71.0	60.8	0.29	72.9	72.1	0.43
Support vector machine (6)	74.1	70.8	0.42	79.6	74.7	0.53
Support vector machine (7)	73.0	70.6	0.41	76.2	72.5	0.46
Support vector machine (8)	73.0	70.6	0.41	76.2	72.5	0.46
Support vector machine (9)	74.8	71.5	0.43	73.2	71.8	0.42
Support vector machine (10)	75.6	71.2	0.44	79.9	73.5	0.54
Support vector machine (11)	78.8	74.6	0.51	79.5	73.3	0.53

Table 8.12. Ten fold cross-validation of different data mining techniques applied in a single-step fashion to discriminate between proteins with transport activity and proteins without transport activity. The highlighted cells correspond to the best performing data mining techniques.

The classifiers designed using the one-against-all principle to predict a particular functional class **(table 8.13)** and the multi-class SVM(6) based on a single-step architecture **(table 8.7)** showed similar values of GAv. The sensitivity scores were found to be larger for the classifiers using the one-against-all principle, but with the cost of lower levels of specificity. However, the results showed in **table 8.13** are not representative of the overall predictive accuracy of the method, as the different classifiers need to be combined in order to apply all classifiers simultaneously and obtain a consensus prediction (level 1 in **figure 8.5**).

Class	Sensitivity	Specificity	GAv
Enzyme	88.5	70.9	79.2
GPCR	91.1	71.3	80.6
Ion channel	59.2	24	37.4
Molecular transporter	72.2	44	56.3
Transport activity	73.2	61	0.48

Table 8.13. Predictive accuracy for each class using the best classifiers to predict a particular functional class. Prediction of enzymes uses a support vector machine (attribute selection = true, c = 1, exp = 15, feat. space normalization = true, γ = 0.0001), prediction of GPCR proteins and ion channels uses the Naïve Bayesian method (attribute selection = true and attribute selection = false respectively), prediction of molecular transporters and proteins with transport activity is based on a Bayesian Network method (attribute selection = true).

Another possible architecture would involve the discrimination of the GPCR class at the first level and the prediction of the remaining functional classes subsequently. GPCR proteins showed the highest predictive accuracy values using the one-against-all principle and so they were deemed to be the best group to be predicted first in the decision architecture. After the GPCR proteins have been predicted, new data mining analysis and architectures needed to be evaluated in order to maximize the prediction of the method **(tables 8.14 – 8.17)**. The best predictive method to distinguish between enzymes, ion channels and molecular transporters based on a single step architecture was found to be the SVM(6) **(table 8.14)**. The results showed in **table 8.15** summarize the prediction accuracy for each of the functional classes predicted by the given classifier. This architecture showed improved levels of accuracy for enzymes but the sensitivity levels for proteins with transport activity was not very high. These results were obtained by training the classifiers with all proteins belonging to the enzyme, ion channel and molecular transporter class and are only an approximation as the predictive architecture where both classifiers were combined **(Please see appendix B figure 8.4 on CD)** has not been evaluated.

	Attribute selection			No attribute selection		
Data mining method	**Q**	**nQ**	**GC**	**Q**	**nQ**	**GC**
Bayesian networks	61.0	53.9	0.35	56.3	51.7	0.3
Naïve bayesian	51.0	50.4	0.3	43.9	45.5	0.24
Naïve Bayesian simple	50.6	50.3	0.29	-	-	-
Logistic regression	61.3	44.2	0.22	-	-	-
RBF Network	64.1	46.3	0.28	59.7	33.3	-
KStar	55.2	49.3	0.24	48.1	46.4	0.18
MultiBoostAB	66.2	50.9	0.33	66.6	49.2	0.32
J48	60.4	50.0	0.29	56.4	47.0	0.23
Random forest	64.2	48.3	0.3	65.6	49.6	0.31
Support vector machine (1)	59.3	33.1	-	60.4	43.3	0.2
Support vector machine (2)	60.8	39.5	-	58.1	47.1	0.23
Support vector machine (3)	60.4	39.1	-	57.9	47.2	0.23
Support vector machine (4)	63.9	49.2	0.29	62.0	56.2	0.34
Support vector machine (5)	61.3	48.5	0.27	59.9	57.0	0.34
Support vector machine (6)	63.6	54.6	0.34	**71.7**	**58.6**	**0.44**
Support vector machine (7)	63.5	55.6	0.35	68.6	57.8	0.4
Support vector machine (8)	63.5	55.6	0.35	68.6	57.8	0.4
Support vector machine (9)	60.6	52.8	0.3	62.0	56.2	0.34
Support vector machine (10)	65.8	54.9	0.36	71.3	56.2	0.43
Support vector machine (11)	68.4	54.7	0.38	71.2	56.1	0.43

Table 8.14. Ten fold cross-validation of different data mining techniques applied in a single-step fashion to discriminate between enzymes, molecular transporters and ion channels. This analysis does not discriminate GPCR proteins because the given classifier was formulated for use after the GPCR proteins had been discriminated by a previous classifier. The highlighted cells correspond to the best performing data mining techniques.

Class	Sensitivity	Specificity	GAv
Enzyme	91.6	75.5	83.1
Ion channel	35.2	59.9	45.9
Molecular transporter	49.1	63.3	55.7

Table 8.15. Predictive accuracy to predict enzymes, ion channels and molecular transporters using a support vector machine (attribute selection = false, c = 50, exp = 9, feat. space normalization = true). GPCR proteins are not distinguished because this classifier was formulated for use subsequent to the classifier used to predict GPCR proteins.

	Attribute selection			No attribute selection		
Data mining method	**Q**	**nQ**	**MCC**	**Q**	**nQ**	**MCC**
Bayesian networks	**72.2**	**72.4**	**0.44**	67.9	68.4	0.36
Naïve bayesian	62.7	66.1	0.33	59.3	63.0	0.27
Naïve Bayesian simple	62.3	65.7	0.32	-	-	-
Logistic regression	68.8	66.7	0.34	67.0	65.8	0.32
RBF Network	67.3	68.5	0.36	59.7	50.0	-
KStar	64.9	65.3	0.30	56.7	57.9	0.16
MultiBoostAB	74.7	72.1	0.46	73.7	70.3	0.44
J48	67.3	65.7	0.32	65.6	63.7	0.28
Random forest	73.2	70.2	0.43	72.4	68.8	0.41
Support vector machine (1)	69.1	67.1	0.35	68.4	67.3	0.35
Support vector machine (2)	69.3	67.8	0.36	67.0	66.5	0.33
Support vector machine (3)	69.0	67.6	0.35	66.7	66.2	0.32
Support vector machine (4)	71.9	69.6	0.40	71.3	71.5	0.42
Support vector machine (5)	69.1	65.4	0.34	68.9	70.1	0.39
Support vector machine (6)	72.1	70.7	0.42	77.8	75.3	0.53
Support vector machine (7)	71.4	70.8	0.41	75.2	73.5	0.48
Support vector machine (8)	71.4	70.8	0.41	75.2	73.5	0.48
Support vector machine (9)	71.9	71.2	0.42	71.3	71.5	0.42
Support vector machine (10)	74.0	72.0	0.45	77.1	73.6	0.52
Support vector machine (11)	75.6	73.2	0.48	77.6	74.5	0.53

Table 8.16. Ten fold cross-validation of different data mining techniques applied in a single-step fashion to discriminate between enzymes, and proteins with transport activity. This analysis does not discriminate GPCR proteins because the given classifier was designed to be applied after GPCR proteins had been discriminated by a previous classifier. The highlighted cells correspond to the best performing data mining techniques.

Class	Sensitivity	Specificity	GAv
Enzyme	71.3	80	75.5
Transport activity	73.5	63.4	68.3

Table 8.17. Predictive accuracy to predict enzymes and proteins with transport activity (including ion channels and molecular transporters) using a Bayesian Network method (attribute selection = true). GPCR proteins are not distinguished because this classifier was formulated for use subsequent to the classifier used to predict GPCR proteins.

Another possible variant would involve the distinction between enzymes and proteins with transport activity at the second level and then sub-classifying the later into the ion channel and molecular transporter classes **(Please see appendix B figure 8.5 on CD)**. Data mining analysis to distinguish solely between enzymes and proteins with transport activity showed that the best performing classifiers were the Bayesian network, the SVM(10) and SVM(11) (all three with prior attribute selection) **(table 8.16)**. Comparison of these data mining techniques showed that the support vector machines tend to

underestimate proteins with transport activity whereas the Bayesian network was a more consistent method for the prediction of both classes. It was necessary to sub-classify the transport activity class into the ion channel and the molecular transporter class by a previously evaluated classifier **(table 8.4)**, which predicts approximately 75% of the ion channels and 71% of the molecular transporters.

All the different architectures analyzed thus far were designed to maximize the prediction of functional classes at their lower complexity level. The accuracy of the architectures combining more than one classifier (with the exception of the first architecture corresponding to **figure 8.1 in appendix B, on CD**) was only estimated, as combining and evaluating such architectures is a tedious and time consuming task. However, the estimation of the accuracy of these architectures can be useful in indicating the best performing classifiers that ultimately need to be evaluated.

Of the five architectures designed (evaluation results corresponding to **tables 8.5, 8.7, 8.13, 8.15, 8.17**), the predictive architecture based on the one-against-all principle showed the highest sensitivity values (true positives) although the specificity values (true negatives) were not as high as for other predictive architectures (especially for proteins with transport activity). However, by introducing an extra classifier that predicts proteins with transport activity and combining such a classifier with those designed to predict ion channels and molecular transporters, the predictive accuracy of the method could be increased. Therefore the selected architecture used to predict the different functional classes at their lowest complexity level was the architecture based on the one-against-all principle (**figure 8.5** level 1).

The subsequent analyses (**tables 8.18 - 8.38**) were carried out to sub-classify the enzyme, GPCR, ion channel and molecular transporter classes into more informative subclasses and identify the end point in the prediction for each functional class. Due to the size of the different classes of GPCR proteins **(table 8.1)**, only the GPCR class A proteins could be distinguished. Evaluation of the different classifiers showed that the best performing classifier was the Bayesian method (attribute selection = true) **(table 8.18)**,

which predicted class A GPCR proteins with a sensitivity of 90.5% and a specificity of 98%. Sub-classification of class A GPCR proteins was also possible due to the accurate prediction values obtained for this protein class. A multi-class classifier was evaluated to distinguish between amine, olfactory, peptide, rhodopsin and other class A GPCR proteins **(table 8.19)**. Comparison of the best performing classifiers **(Please see appendix B table 8.6 on CD)** showed that the different subclasses were predicted with sensitivity values higher than 75% with the exception of other class A GPCR proteins. In order to maximize the prediction of this functional subclass, the SVM(7) and SVM(8) (both with attribute selection = true) were adopted as the equal best performing classifiers. **Table 8.20** shows the sensitivity and specificity scores for each of the class A GPCR subfamilies.

	Attribute selection			No attribute selection		
Data mining method	**Q**	**nQ**	**MCC**	**Q**	**nQ**	**MCC**
Bayesian networks	**90.7**	**91.1**	**0.75**	88.2	85.6	0.67
Naïve bayesian	93.2	89.6	0.79	88.2	81.7	0.63
Naïve Bayesian simple	93.2	89.6	0.79	-	-	-
Logistic regression	86.1	79.6	0.58	78.9	72.8	0.42
RBF Network	91.1	85.1	0.72	79.7	50.0	-
KStar	90.7	82.5	0.70	88.2	73.9	0.60
MultiBoostAB	90.3	81.5	0.68	89.0	74.5	0.63
J48	86.5	79.1	0.58	81.4	71.3	0.43
Random forest	90.7	83.3	0.70	89.0	73.7	0.63
Support vector machine (1)	90.7	84.1	0.70	87.3	75.7	0.58
Support vector machine (2)	87.8	80.7	0.62	87.3	75.7	0.58
Support vector machine (3)	86.9	80.1	0.60	87.3	75.7	0.58
Support vector machine (4)	90.3	86.9	0.71	91.1	82.0	0.71
Support vector machine (5)	86.5	80.7	0.60	86.9	71.6	0.55
Support vector machine (6)	90.3	79.9	0.68	85.7	65.4	0.49
Support vector machine (7)	92.0	84.9	0.74	90.7	80.2	0.69
Support vector machine (8)	92.0	84.9	0.74	90.7	80.2	0.69
Support vector machine (9)	90.3	86.9	0.71	91.1	82.0	0.71
Support vector machine (10)	84.0	62.7	0.41	80.2	51.0	0.13
Support vector machine (11)	83.5	59.4	0.39	79.7	50.0	-

Table 8.18. Ten fold cross-validation of different data mining techniques applied in a single-step fashion to discriminate between GPCR class A proteins and other GPCR proteins. This analysis is specific to GPCR proteins. The cells coloured in yellow correspond to the best performing data mining techniques. The cells highlighted in red correspond to data mining techniques where the predictive scores are not representative of the predictive power of the method (not normalised). This is normally due to a large class being well predicted whereas the smaller class can not be accurately discriminated.

	Attribute selection			No attribute selection		
Data mining method	**Q**	**nQ**	**GC**	**Q**	**nQ**	**GC**
Bayesian networks	77.3	75.8	0.75	73.5	72.5	0.72
Naïve bayesian	76.2	75.3	0.73	64	62.8	0.6
Naïve Bayesian simple	44.7	66.5	0.5	-	-	-
Logistic regression	58.2	56	0.1	66.1	65.8	0.64
RBF Network	66.7	63.9	0.62	32.3	21.6	-
KStar	65.1	61.9	0.58	58.2	57.7	0.51
MultiBoostAB	72	68.1	0.69	64	60	0.59
J48	65.6	65	0.63	58.7	57.7	0.54
Random forest	68.3	64.6	0.65	62.4	58	0.56
Support vector machine (1)	75.1	73.2	0.73	70.4	69.6	0.68
Support vector machine (2)	72	72.3	0.68	70.9	70.3	0.69
Support vector machine (3)	72.5	73	0.69	70.9	70.3	069
Support vector machine (4)	75.7	76.3	0.73	73	68.8	0.69
Support vector machine (5)	72	71.6	0.68	65.6	59.7	0.62
Support vector machine (6)	76.7	75.1	0.73	70.4	63.3	0.68
Support vector machine (7)	**77.8**	**77.3**	**0.75**	80.4	78	0.79
Support vector machine (8)	**77.8**	**77.3**	**0.75**	80.4	78	0.79
Support vector machine (9)	77.7	76.3	0.73	73	68.8	0.69
Support vector machine (10)	71.4	67.1	0.7	47.6	35.9	-
Support vector machine (11)	68.9	63.7	0.68	43.9	31.3	-

Table 8.19. Ten fold cross-validation of different data mining techniques applied in a single-step fashion to discriminate between various GPCR class A subfamilies: amine GPCR, olfactory GPCR, peptide GPCR, rhodopsin GPCR and other GPCR proteins. This analysis is specific to GPCR proteins. The highlighted cells correspond to the best performing data mining techniques.

Class	Sensitivity	Specificity	GAv
Amine	78.6	86.8	82.6
Olfactory	86.5	94.1	90.2
Peptide	73.8	71.4	72.6
Rhodopsin	73.7	87.5	80.3
Other	50.0	39.5	44.4

Table 8.20. Predictive accuracy for subfamilies in the GPCR class A protein family (amine, olfactory, peptide, rhodopsin and other class A GPCR proteins) using a support vector machine (attribute selection = true, c = 50, exp = 9, feat space normalization = true) based on a single step architecture.

Sub-classification of class A GPCR proteins was also evaluated using the one-against-all principle. Prediction of amine GPCR proteins showed that the best performing classifiers were the Bayesian network and the SVM(1) (both with attribute selection = true) **(table 8.21)**. Comparison of these two classifiers **(Please see appendix B table 8.7 on CD)** showed that the SVM(1) classifier was a better classifier with higher values of sensitivity and specificity. Following the same procedure, the best classifiers were identified to predict olfactory and peptide GPCR proteins (**table 8.22** and **table 8.23**). These analyses showed

that the Bayesian network and the Naïve Bayesian method (both with attribute selection = true) were the best performing classifiers to predict olfactory and peptide GPCR proteins respectively. Evaluation of classifiers to predict rhodopsin GPCR proteins **(table 8.24)** showed various classifiers that were consequently compared **(Please see appendix B table 8.8 on CD)** confirming that SVM(2) & SVM(3) predicted rhodopsin GPCR proteins with the highest values of sensitivity. Comparison of the single step classifier using the all-against-all principle **(table 8.20)** with the classifiers designed based on the one-against-all principle **(table 8.25)** showed that using classifiers to predict a single class at a time achieved more accurate values. Therefore, the architecture based on the one-against-all principle was selected to be applied by the TMFUN algorithm (**figure 8.5** level 3). However, it is important to remark that the results showed in **table 8.20** were valid evaluation results whereas the results showed in **table 8.25** are estimations as it would be necessary to combine the prediction from the four classifiers in one analysis in order to properly evaluate the architecture.

	Attribute selection			No attribute selection		
Data mining method	**Q**	**nQ**	**MCC**	**Q**	**nQ**	**MCC**
Bayesian networks	**93.1**	**89.6**	**0.80**	89.9	87.6	0.72
Naïve bayesian	87.3	89.3	0.70	82.0	80.8	0.55
Naïve Bayesian simple	-	-	-	-	-	-
Logistic regression	85.7	79.8	0.59	77.8	72.1	0.41
RBF Network	90.5	87.9	0.73	77.8	50.0	-
KStar	91.5	85.2	0.75	85.2	78.6	0.57
MultiBoostAB	93.1	87.9	0.79	86.2	73.3	0.56
J48	87.8	85.4	0.67	84.7	80.8	0.58
Random forest	93.1	87.9	0.79	86.2	73.3	0.56
Support vector machine (1)	93.7	90.8	0.82	92.1	88.1	0.77
Support vector machine (2)	88.9	84.4	0.68	92.1	88.1	0.77
Support vector machine (3)	88.9	83.5	0.68	92.1	88.1	0.77
Support vector machine (4)	91.0	88.3	0.75	93.1	87.1	0.79
Support vector machine (5)	92.1	88.9	0.77	91.5	83.5	0.74
Support vector machine (6)	89.9	84.2	0.70	89.9	78.2	0.69
Support vector machine (7)	90.5	86.2	0.72	91.5	83.5	0.74
Support vector machine (8)	90.5	86.2	0.72	91.5	83.5	0.74
Support vector machine (9)	91.0	88.3	0.75	93.1	87.1	0.79
Support vector machine (10)	89.4	78.7	0.67	84.7	65.5	0.51
Support vector machine (11)	87.8	75.2	0.62	79.9	54.8	0.28

Table 8.21. Ten fold cross-validation of different data mining techniques applied in a single-step fashion to discriminate between amine GPCR proteins and other GPCR class A proteins. This analysis is specific to GPCR class A proteins. The highlighted cells correspond to the best performing data mining techniques.

	Attribute selection			No attribute selection		
Data mining method	**Q**	**nQ**	**MCC**	**Q**	**nQ**	**MCC**
Bayesian networks	**98.9**	**99.3**	**0.97**	97.4	96.3	0.92
Naïve bayesian	95.8	96.3	0.88	92.6	89.3	0.77
Naïve Bayesian simple	94.7	95.7	0.85	-	-	-
Logistic regression	89.9	86.6	0.70	85.7	73.7	0.52
RBF Network	95.8	94.3	0.87	79.4	51.4	0.06
KStar	88.9	71.6	0.62	88.4	70.3	0.60
MultiBoostAB	94.2	88.2	0.81	92.1	81.8	0.73
J48	91.5	84.5	0.72	84.1	74.8	0.50
Random forest	94.2	88.2	0.81	92.1	81.8	0.73
Support vector machine (1)	96.8	93.9	0.90	95.2	89.9	0.84
Support vector machine (2)	94.7	90.6	0.83	95.2	89.9	0.84
Support vector machine (3)	94.7	90.6	0.83	95.2	89.9	0.84
Support vector machine (4)	94.2	88.2	0.81	93.7	83.8	0.79
Support vector machine (5)	92.1	83.8	0.73	88.9	72.6	0.61
Support vector machine (6)	96.3	92.6	0.88	87.3	67.6	0.55
Support vector machine (7)	96.3	94.6	0.88	97.4	93.2	0.92
Support vector machine (8)	96.3	94.6	0.88	97.4	93.2	0.92
Support vector machine (9)	94.2	88.2	0.81	93.7	83.8	0.79
Support vector machine (10)	84.1	59.5	0.40	80.4	50.0	-
Support vector machine (11)	83.6	58.1	0.37	80.4	50.0	-

Table 8.22. Ten fold cross-validation of different data mining techniques applied in a single-step fashion to discriminate between olfactory GPCR proteins and other GPCR class A proteins. This analysis is specific to GPCR class A proteins. The highlighted cells correspond to the best performing data mining techniques.

	Attribute selection			No attribute selection		
Data mining method	**Q**	**nQ**	**MCC**	**Q**	**nQ**	**MCC**
Bayesian networks	70.9	67.4	0.34	71.4	67.8	0.35
Naïve bayesian	**83.1**	**83.2**	**0.64**	72.5	71.1	0.41
Naïve Bayesian simple	76.2	73.4	0.46	-	-	-
Logistic regression	83.1	81.9	0.62	63.5	58.9	0.18
RBF Network	68.8	70.9	0.39	67.7	50.0	-
KStar	77.8	74.6	0.49	69.8	67.4	0.34
MultiBoostAB	69.8	65.3	0.31	74.1	68.8	0.39
J48	77.8	74.6	0.49	65.1	60.9	0.21
Random forest	82.5	79.0	0.59	74.1	68.8	0.39
Support vector machine (1)	82.5	79.0	0.59	75.1	72.2	0.44
Support vector machine (2)	79.9	76.1	0.53	72.5	69.0	0.38
Support vector machine (3)	79.9	75.7	0.53	72.5	69.0	0.38
Support vector machine (4)	77.2	74.2	0.48	75.1	72.2	0.44
Support vector machine (5)	70.9	68.2	0.35	72.0	70.7	0.40
Support vector machine (6)	79.9	74.9	0.52	78.3	67.7	0.48
Support vector machine (7)	79.4	75.8	0.52	79.9	74.4	0.52
Support vector machine (8)	79.4	75.8	0.52	79.9	74.4	0.52
Support vector machine (9)	77.2	74.2	0.48	75.1	72.2	0.44
Support vector machine (10)	78.3	69.8	0.47	69.3	52.9	0.17
Support vector machine (11)	77.8	69.9	0.46	67.7	50.0	-

Table 8.23. Ten fold cross-validation of different data mining techniques applied in a single-step fashion to discriminate between peptide GPCR proteins and other GPCR class A proteins. This analysis is specific to GPCR class A proteins. The highlighted cells correspond to the best performing data mining techniques.

	Attribute selection			No attribute selection		
Data mining method	**Q**	**nQ**	**MCC**	**Q**	**nQ**	**MCC**
Bayesian networks	98.4	94.4	0.91	97.9	89.5	0.88
Naïve bayesian	97.4	89.2	0.85	88.9	82.1	0.53
Naïve Bayesian simple	97.4	91.5	0.85	-	-	-
Logistic regression	97.9	94.1	0.88	91.5	83.6	0.60
RBF Network	97.9	91.8	0.88	89.9	50.0	-
KStar	95.2	85.7	0.73	93.7	82.4	0.65
MultiBoostAB	97.4	86.8	0.85	94.2	75.7	0.63
J48	94.7	80.7	0.68	92.1	76.9	0.55
Random forest	97.4	86.8	0.85	94.2	75.7	0.63
Support vector machine (1)	98.4	94.4	0.91	96.3	88.6	0.79
Support vector machine (2)	**97.4**	**96.2**	**0.87**	96.3	88.6	0.79
Support vector machine (3)	**97.4**	**96.2**	**0.87**	96.3	88.6	0.79
Support vector machine (4)	96.3	93.3	0.81	95.8	86.0	0.76
Support vector machine (5)	95.2	95.0	0.79	95.8	83.6	0.75
Support vector machine (6)	97.4	86.8	0.85	94.7	73.7	0.67
Support vector machine (7)	97.4	89.2	0.85	97.9	89.5	0.88
Support vector machine (8)	96.3	93.3	0.81	97.9	89.5	0.88
Support vector machine (9)	96.3	81.6	0.78	95.8	86.0	0.76
Support vector machine (10)	96.3	81.6	0.78	90.5	52.6	0.22
Support vector machine (11)	96.3	81.6	0.78	89.9	50.0	-

Table 8.24. Ten fold cross-validation of different data mining techniques applied in a single-step fashion to discriminate between rhodopsin GPCR proteins and other GPCR class A proteins. This analysis is specific to GPCR class A proteins. The cells coloured in yellow correspond to the best performing data mining techniques. The cells highlighted in red correspond to data mining techniques where the predictive scores are not representative of the predictive power of the method (not normalised). This is normally due to a large class being well predicted whereas the smaller class can not be accurately discriminated.

Class	Sensitivity	Specificity	GAv
Amine	83.3	85.4	84.3
Olfactory	100	97.4	97.4
Peptide	83.6	69.9	76.4
Rhodopsin	94.7	81.8	88

Table 8.25. Predictive accuracy for subfamilies in the GPCR class A protein family (amine, olfactory, peptide, rhodopsin and other class A GPCR proteins) using the best classifier to predict a particular subfamily based on the "one-against-all" principle. Prediction of amine and olfactory GPCR proteins is achieved using a Bayesian Network (attribute selection = true), prediction of peptide GPCR proteins is based on a Naïve Bayesian method (attribute selection = true) and prediction of rhodopsin GPCR proteins is achieved using a support vector machine (attribute selection = true, c = 30 or c = 50).

A similar approach was pursued to further sub-classify the enzyme, molecular transporter and ion channel classes. Evaluation of the single-step multi-class classifier and of the one-against-all principle showed that predictions were not accurate enough for enzymes and molecular transporters. The SVM(6) designed to sub-classify enzymes **(table**

8.26) failed to accurately distinguish oxidoreductases and other enzymes **(table 8.27)**. Using the one-against-all principle a Naïve Bayesian classifier **(table 8.28)**, a MultiBoostAB classifier (attribute selection = true) **(table 8.29)** and a Bayesian network (attribute selection = true) **(table 8.30)** were selected to predict oxidoreductases, transferases and hydrolases respectively. Although these classifiers showed fairly good prediction accuracies for transferases and hydrolases **(table 8.31)**, the prediction of oxidoreductases could not be refined. Considering that 88.5% of the enzymes **(table 8.13)** and 70.7% of the transferases **(table 8.31)** are correctly classified, the estimation of correctly predicted transferases is at most 62.6%. The remaining sub-classes were estimated to be predicted at a lower accuracy as the corresponding classifiers reported lower values of sensitivity and specificity **(table 8.31)**. Based on these results it was decided not to further sub-classify the enzyme class, as due to the lower accuracy the obtained predictions would not be meaningful.

	Attribute selection			No attribute selection		
Data mining method	**Q**	**nQ**	**GC**	**Q**	**nQ**	**GC**
Bayesian networks	59.4	38.2	0.27	47.9	32.6	-
Naïve bayesian	45.0	39.8	0.22	48.5	38.0	0.21
Naïve Bayesian simple	45.8	40.9	0.22	-	-	-
Logistic regression	58.6	36.8	0.23	46.4	39.9	0.2
RBF Network	58.4	34.6	0.22	49.1	25.0	-
KStar	51.0	40.7	0.21	49.0	42.1	0.24
MultiBoostAB	61.8	37.8	0.26	62.7	38.7	0.29
J48	51.2	35.2	0.18	49.4	33.9	0.17
Random forest	60.5	37.4	0.25	61.0	37.0	0.25
Support vector machine (1)	59.6	33.4	-	57.1	34.6	0.19
Support vector machine (2)	61.4	36.5	0.25	48.8	39.5	0.23
Support vector machine (3)	61.3	36.5	0.24	48.4	38.7	0.23
Support vector machine (4)	58.4	37.0	0.24	55.8	42.4	0.27
Support vector machine (5)	56.6	37.1	0.22	55.3	43.8	0.26
Support vector machine (6)	54.6	41.1	0.28	61.2	39.7	0.31
Support vector machine (7)	53.5	41.5	0.26	**56.7**	**43.1**	**0.31**
Support vector machine (8)	53.5	41.5	0.26	56.7	43.1	0.31
Support vector machine (9)	51.9	38.9	0.21	55.8	42.4	0.27
Support vector machine (10)	59.2	38.5	0.28	61.8	37.9	0.33
Support vector machine (11)	61.9	38.0	-	60.8	36.3	-

Table 8.26. Ten fold cross-validation of different data mining techniques applied in a single-step fashion to discriminate between oxidoreductases (EC1), transferases (EC2), hydrolases (EC3) and other enzymes. This analysis is specific to enzymes. The highlighted cells correspond to the best performing data mining techniques.

Class	Sensitivity	Specificity	GAv
Oxidoreductase	31.4	15.7	22.2
Transferase	60.6	64.5	62.5
Hydrolase	61.2	67	64
Other	19.4	50	31.1

Table 8.27. Predictive accuracy to distinguish between oxidoreductases, transferases, hydrolases and other enzymes. The given multi-class classifier is a support vector machine (attribute selection = false, c = 50, exp = 9, feat space normalization = true).

	Attribute selection			No attribute selection		
Data mining method	**Q**	**nQ**	**MCC**	**Q**	**nQ**	**MCC**
Bayesian networks	90.4	49.5	-0.03	90.6	49.6	-0.03
Naïve bayesian	89.2	57.2	0.18	**75.7**	**56.3**	**0.09**
Naïve Bayesian simple	89.5	57.6	0.20	-	-	-
Logistic regression	91.2	51.9	0.11	77.7	56.7	0.10
RBF Network	91.8	53.5	0.23	91.3	50.0	-
KStar	90.8	50.4	0.02	80.8	56.5	0.10
MultiBoostAB	90.8	58.1	0.24	91.2	50.6	0.05
J48	91.4	52.0	0.14	87.2	54.9	0.11
Random forest	89.7	57.5	0.20	90.9	51.1	0.07
Support vector machine (1)	91.3	50.0	-	90.2	50.7	0.03
Support vector machine (2)	91.3	50.0	-	84.0	56.3	0.11
Support vector machine (3)	91.3	50.0	-	83.0	55.8	0.10
Support vector machine (4)	91.3	50.0	-	87.7	55.1	0.12
Support vector machine (5)	91.3	50.0	-	88.3	53.5	0.09
Support vector machine (6)	90.3	52.0	0.08	91.4	50.7	0.11
Support vector machine (7)	90.8	53.6	0.15	89.5	52.2	0.07
Support vector machine (8)	90.8	53.6	0.15	89.5	52.2	0.07
Support vector machine (9)	91.6	51.4	0.16	87.7	55.1	0.12
Support vector machine (10)	90.8	53.6	0.15	91.3	50.0	-
Support vector machine (11)	91.3	50.6	0.07	91.3	50.0	-

Table 8.28. Ten fold cross-validation of different data mining techniques applied in a single-step fashion to discriminate between oxidoreductases and other enzymes. This analysis is specific to enzymes. The highlighted cells correspond to the best performing data mining techniques.

	Attribute selection			No attribute selection		
Data mining method	**Q**	**nQ**	**MCC**	**Q**	**nQ**	**MCC**
Bayesian networks	62.3	62.4	0.25	58.1	58.1	0.16
Naïve bayesian	62.8	63.0	0.27	60.7	60.5	0.22
Naïve Bayesian simple	61.9	62.1	0.25	-	-	-
Logistic regression	63.6	63.7	0.27	60.5	60.5	0.21
RBF Network	62.2	62.3	0.25	49.4	48.6	-0.05
KStar	58.6	58.5	0.17	58.3	58.0	0.17
MultiBoostAB	**67.7**	**67.8**	**0.36**	62.8	62.9	0.26
J48	58.9	59.0	0.18	56.8	56.8	0.14
Random forest	65.8	65.8	0.32	61.8	62.0	0.24
Support vector machine (1)	65.5	65.5	0.31	63.0	63.0	0.26
Support vector machine (2)	64.3	64.3	0.29	61.2	61.1	0.22
Support vector machine (3)	64.3	64.3	0.29	61.5	61.5	0.23
Support vector machine (4)	55.5	54.8	0.15	63.4	63.3	0.27
Support vector machine (5)	54.2	53.5	0.13	62.0	61.9	0.24
Support vector machine (6)	63.3	63.1	0.27	62.3	61.9	0.27
Support vector machine (7)	65.5	65.4	0.31	63.0	62.7	0.27
Support vector machine (8)	65.5	65.4	0.31	63.0	62.7	0.27
Support vector machine (9)	61.3	60.9	0.25	63.4	63.3	0.27
Support vector machine (10)	62.4	62.0	0.27	60.7	60.1	0.27
Support vector machine (11)	65.3	65.1	0.31	66.4	66.6	0.35

Table 8.29. Ten fold cross-validation of different data mining techniques applied in a single-step fashion to discriminate between transferases and other enzymes. This analysis is specific to enzymes. The highlighted cells correspond to the best performing data mining techniques.

	Attribute selection			No attribute selection		
Data mining method	**Q**	**nQ**	**MCC**	**Q**	**nQ**	**MCC**
Bayesian networks	**70.6**	**69.9**	**0.39**	64.3	65.3	0.30
Naïve bayesian	65.9	69.2	0.38	59.6	63.7	0.28
Naïve Bayesian simple	65.6	69.1	0.38	-	-	-
Logistic regression	70.3	67.6	0.36	63.0	61.8	0.23
RBF Network	68.7	68.2	0.36	61.7	50.0	-
KStar	62.2	61.7	0.23	64.9	65.8	0.31
MultiBoostAB	74.6	72.3	0.45	74.3	71.9	0.45
J48	66.4	63.8	0.28	61.8	59.2	0.19
Random forest	72.8	71.0	0.42	72.2	70.4	0.41
Support vector machine (1)	70.0	65.4	0.34	67.2	64.9	0.30
Support vector machine (2)	69.7	66.9	0.35	64.5	63.2	0.26
Support vector machine (3)	70.2	67.5	0.36	64.3	62.9	0.26
Support vector machine (4)	71.8	68.2	0.39	67.6	66.8	0.33
Support vector machine (5)	69.2	64.3	0.32	68.4	68.0	0.35
Support vector machine (6)	73.4	69.4	0.42	72.3	67.5	0.39
Support vector machine (7)	72.5	70.1	0.41	71.8	68.7	0.39
Support vector machine (8)	72.5	70.1	0.41	71.8	68.7	0.39
Support vector machine (9)	69.6	68.4	0.36	67.6	66.8	0.33
Support vector machine (10)	70.7	64.5	0.35	70.0	62.9	0.34
Support vector machine (11)	72.7	67.1	0.40	70.5	63.4	0.35

Table 8.30. Ten fold cross-validation of different data mining techniques applied in a single-step fashion to discriminate between hydrolases and other enzymes. This analysis is specific to enzymes. The cells coloured in yellow correspond to the best performing data mining techniques. The cells highlighted in red correspond to data mining techniques where the predictive scores are not representative of the predictive power of the method (not normalised). This is normally due to a large class being well predicted whereas the smaller class can not be accurately discriminated.

Class	Sensitivity	Specificity	GAv
Oxidoreductase	32.9	13.4	21
Transferase	70.7	66	68.3
Hydrolase	67.0	60.5	63.7

Table 8.31. Predictive accuracy to distinguish between oxidoreductases, transferases and hydrolases using the best performing classifier to predict each functional class based on the one-against-all principle. The classifier for predicting oxidoreductases uses the Naïve Bayesian method (attribute selection = false), transferases are distinguished using the MultiBoostAB method (attribute selection = true, classifier = random forest, iterations = 30), the classifier for predicting hydrolases is based on the Naïve Bayesian method (attribute selection = true).

Likewise, sub-classification of molecular transporters into amino acid transporters, sugar transporters and other molecular transporters did not achieve sufficiently accurate predictions. The Bayesian network (attribute selection = true) designed to sub-classify molecular transporters in a single step **(table 8.32)** showed poor prediction accuracy scores for the three subclasses **(table 8.33)**. Based on the one-against-all principle a Naïve

Bayesian classifier (attribute selection = true) **(table 8.34)** and a Naïve Bayesian simple classifier (attribute selection = true) **(table 8.35)** were designed to predict amino acid transporters and sugar transporters respectively. Combination of the classifier to predict molecular transporters whose sensitivity is 72.2% **(table 8.13)** and the classifiers to predict amino acid transporters and sugar transporters **(table 8.36)**, the estimation of correctly predicted amino acid transporters and sugar transporters is at most 45.6% and 46% respectively. As with enzymes, it was decided not to sub-classify the molecular transporter class as it was clear that the accuracy values were not going to be significant.

	Attribute selection			No attribute selection		
Data mining method	**Q**	**nQ**	**GC**	**Q**	**nQ**	**GC**
Bayesian networks	**54.5**	**55.4**	**0.33**	48.0	48.9	0.24
Naïve bayesian	49.5	51.8	0.28	41.9	42.4	0.21
Naïve Bayesian simple	48.7	51.2	0.27	-	-	-
Logistic regression	46.6	46.1	0.21	34.7	35.3	0.08
RBF Network	46.9	47.3	0.23	37.9	31.0	-
KStar	42.2	42.0	0.15	43.3	45.4	0.21
MultiBoostAB	53.4	53.1	0.32	50.5	49.6	0.27
J48	43.7	43.3	0.17	41.5	40.9	0.13
Random forest	53.1	52.7	0.31	44.0	43.8	0.22
Support vector machine (1)	52.0	50.6	0.29	37.2	38.0	0.17
Support vector machine (2)	50.2	49.3	0.26	39.7	40.8	0.15
Support vector machine (3)	48.7	47.8	0.25	39.7	40.8	0.15
Support vector machine (4)	49.5	48.8	0.25	44.4	45.6	0.21
Support vector machine (5)	49.1	49.3	0.25	43.3	44.3	0.18
Support vector machine (6)	47.3	47.5	0.23	51.6	50.1	0.29
Support vector machine (7)	47.7	47.9	0.24	48.7	48.2	0.25
Support vector machine (8)	47.7	47.9	0.24	48.7	48.2	0.25
Support vector machine (9)	45.5	46.0	0.23	44.4	45.6	0.21
Support vector machine (10)	48.7	48.4	0.25	47.7	44.4	0.2
Support vector machine (11)	53.8	52.5	0.33	52.0	48.3	0.29

Table 8.32. Ten fold cross-validation of different data mining techniques applied in a single-step fashion to discriminate between amino acid transporters, sugar transporters and other molecular transporters. This analysis is specific to molecular transporters. The highlighted cells correspond to the best performing data mining techniques.

Class	Sensitivity	Specificity	GAv
Amino acid transporter	56.0	52.8	54.4
Sugar transporter	62.5	51.5	56.8
Other	47.8	59.3	53.3

Table 8.33. Predictive accuracy to distinguish between amino acid transporters, sugar transporters and other molecular transporters. The given multi-class classifier is based on a Naïve Bayesian method (attribute selection = true).

	Attribute selection			No attribute selection		
Data mining method	**Q**	**nQ**	**MCC**	**Q**	**nQ**	**MCC**
Bayesian networks	70.8	66.2	0.32	68.2	63.4	0.26
Naïve bayesian	**70.8**	**68.6**	**0.35**	60.6	53.9	0.08
Naïve Bayesian simple	71.1	68.5	0.35	-	-	-
Logistic regression	72.2	61.2	0.27	52.0	48.4	-0.03
RBF Network	66.4	54.1	0.10	70.4	51.5	0.12
KStar	67.5	62.9	0.25	56.7	56.5	0.12
MultiBoostAB	70.4	66.3	0.32	63.9	52.9	0.07
J48	70.4	55.6	0.17	63.9	58.0	0.16
Random forest	69.3	64.9	0.29	62.1	51.3	0.03
Support vector machine (1)	69.0	49.8	-0.01	61.0	52.2	0.05
Support vector machine (2)	69.3	53.8	0.12	57.8	52.5	0.05
Support vector machine (3)	70.4	55.9	0.17	57.8	52.2	0.04
Support vector machine (4)	70.0	55.3	0.16	62.8	54.5	0.09
Support vector machine (5)	70.4	58.3	0.21	63.2	58.1	0.16
Support vector machine (6)	65.7	57.2	0.15	68.2	55.7	0.14
Support vector machine (7)	68.2	56.4	0.15	66.1	55.1	0.12
Support vector machine (8)	68.2	56.4	0.15	66.1	55.1	0.12
Support vector machine (9)	68.2	56.4	0.15	62.8	54.5	0.09
Support vector machine (10)	63.2	55.4	0.11	69.0	52.8	0.10
Support vector machine (11)	69.3	56.1	0.16	68.6	50.2	0.01

Table 8.34. Ten fold cross-validation of different data mining techniques applied in a single-step fashion to discriminate between amino acid transporters and other molecular transporters. This analysis is specific to enzymes. The cells coloured in yellow correspond to the best performing data mining techniques. The cells highlighted in red correspond to data mining techniques where the predictive scores are not representative of the predictive power of the method (not normalised). This is normally due to a large class being well predicted whereas the smaller class can not be accurately discriminated.

	Attribute selection			No attribute selection		
Data mining method	**Q**	**nQ**	**MCC**	**Q**	**nQ**	**MCC**
Bayesian networks	74.7	70.0	0.39	71.1	67.8	0.34
Naïve bayesian	71.8	68.7	0.36	55.6	61.7	0.22
Naïve Bayesian simple	**72.6**	**69.9**	**0.38**	-	-	-
Logistic regression	74.0	63.9	0.31	58.5	53.7	0.07
RBF Network	73.6	62.5	0.29	71.1	50.0	-
KStar	66.4	60.1	0.20	68.6	64.2	0.27
MultiBoostAB	73.3	67.1	0.34	74.7	64.8	0.33
J48	69.0	60.0	0.21	64.6	55.1	0.11
Random forest	72.6	65.9	0.32	73.6	64.4	0.31
Support vector machine (1)	72.9	59.1	0.24	66.8	59.2	0.19
Support vector machine (2)	76.2	65.8	0.37	58.8	54.0	0.08
Support vector machine (3)	75.8	65.2	0.35	58.8	54.0	0.08
Support vector machine (4)	72.9	65.4	0.32	66.8	60.7	0.21
Support vector machine (5)	69.3	64.3	0.28	69.3	60.2	0.22
Support vector machine (6)	70.8	64.6	0.29	74.4	61.9	0.30
Support vector machine (7)	70.4	65.8	0.31	74.7	65.5	0.34
Support vector machine (8)	70.4	65.8	0.31	74.7	65.5	0.34
Support vector machine (9)	66.1	59.8	0.19	66.8	60.7	0.21
Support vector machine (10)	71.5	62.9	0.27	76.9	62.6	0.37
Support vector machine (11)	78.0	65.6	0.41	75.1	58.0	0.30

Table 8.35. Ten fold cross-validation of different data mining techniques applied in a single-step fashion to discriminate between sugar transporters and other molecular transporters. This analysis is specific to enzymes. The highlighted cells correspond to the best performing data mining techniques.

Class	Sensitivity	Specificity	GAv
Amino acid transporter	63.1	51.5	57
Sugar transporter	63.8	52	57.6

Table 8.36. Predictive accuracy of the distinction between amino acid transporters and sugar transporters using the best performing classifier to predict each functional class based on the one-against-all principle. Amino acid transporters are predicted using a Naïve Bayesian classifier (attribute selection = true) and sugar transporters are predicted using a Naïve Bayesian simple method (attribute selection = true).

In the case of ion channels, the first sub-classification was designed to distinguish between cation channels and anion channels. Although the predictive scores for the ion channel class were already discouraging, the combination of the classifier to predict ion channels with the classifier to predict proteins with transport activity and it subsequent sub-classification into molecular transporters and ion channels was yet to be evaluated. Therefore, it was decided to further sub-classify the ion channels. Evaluation of the different classifiers showed various classifiers **(table 8.37)** that were subsequently compared **(Please see appendix B table 8.9 on CD)**. The Naïve Bayesian classifier was

found to be the more consistent classifier and was selected as the predictive classifier to be used in the predictive algorithm. This classifier predicted cation and anion channels with sensitivity values of 72.6% and 75.5% **(table 8.38)**. However, these results do not take into account the previous percentage of correctly predicted ion channels. At most, prediction of cation and anion channels would reach sensitivity values of approximately 50% after combining the different classifiers. Therefore, no further sub-classification was pursued as the predictive accuracies were not believed to be significant.

	Attribute selection			No attribute selection		
Data mining method	**Q**	**nQ**	**MCC**	**Q**	**nQ**	**MCC**
Bayesian networks	75.2	71.5	0.37	70.4	66.9	0.28
Naïve bayesian	**73.2**	**74.1**	**0.40**	53.6	62.7	0.20
Naïve Bayesian simple	70.8	71.0	0.34			
Logistic regression	78.8	59.8	0.23	61.2	51.2	0.02
RBF Network	82.8	60.8	0.33	80.4	50.0	-
KStar	83.2	72.6	0.46	66.8	60.8	0.18
MultiBoostAB	80.8	62.6	0.30	82.8	59.2	0.32
J48	80.0	61.3	0.27	78.0	63.9	0.29
Random forest	82.0	63.3	0.34	80.0	54.4	0.16
Support vector machine (1)	80.0	51.3	0.07	76.0	57.3	0.16
Support vector machine (2)	82.0	60.3	0.30	66.8	57.0	0.12
Support vector machine (3)	82.4	61.3	0.32	66.4	56.7	0.12
Support vector machine (4)	82.0	54.9	0.25	72.0	57.9	0.15
Support vector machine (5)	81.2	54.4	0.20	71.6	61.5	0.21
Support vector machine (6)	81.2	70.6	0.41	83.2	60.2	0.34
Support vector machine (7)	79.6	67.3	0.35	81.2	62.1	0.30
Support vector machine (8)	79.6	67.3	0.35	81.2	62.1	0.30
Support vector machine (9)	80.0	59.8	0.25	72.0	57.9	0.15
Support vector machine (10)	81.6	69.3	0.40	82.0	54.9	0.25
Support vector machine (11)	84.4	61.0	0.41	80.8	51.0	0.13

Table 8.37. Ten fold cross-validation of different data mining techniques applied in a single-step fashion to discriminate between cation channels and anion channels. This analysis is specific to ion channels. The highlighted cells correspond to the best performing data mining techniques.

Class	Sensitivity	Specificity	GAv
Cation channels	72.6	92.4	81.9
Anion channels	75.5	40.2	55.1

Table 8.38. Predictive accuracy for distimction between cation channels and anion channels. The given multi-class classifier is based on a Naïve Bayesian method (attribute selection = true).

Combination of the different classifiers and architectures designed led to the development of a predictive algorithm that predicted the molecular function of α-helical membrane proteins at various levels of complexity **(figure 8.5)**. Sub-classification of

molecular transporters and enzymes was not performed as the corresponding predictive accuracy scores were not high enough. However, the algorithm included classifiers that were designed to further sub-classify class A GPCR proteins. To our knowledge, this is the first algorithm designed to predict the molecular function of α-helical membrane proteins at such low levels of sequence similarity.

8.3.2.2 Data set filtered at a sequence similarity threshold of 90%

The data set used for this analysis is different to the previous set analyzed. This data set has been filtered at a sequence similarity threshold of 90% and it could be considered as a more representative data set where only highly identical sequences have been removed. In order to design a predictive architecture to maximize the prediction of the different functional classes contained in this data set, it is necessary to evaluate different data mining techniques and architectures as was done with the previous data set **(tables 8.3 - 8.38)**. However, the different number of possible combinations and architectures to be evaluated increases exponentially with the number of different functional classes to be predicted. Instead, the data mining analysis performed was based on the architecture of the algorithm designed to predict the molecular activity of proteins under low sequence similarity **(figure 8.5)**. The main disadvantage of using more flexible sequence similarity thresholds is that the trained classifiers can be biased towards large subsets of proteins with a significant sequence similarity. This problem would lead to an under-prediction of less common subsets of proteins that belong the same functional class. However, the implementation of the previous algorithm, trained with highly heterogeneous sets, can be used instead to predict the functional class of the corresponding proteins. On the other hand, the advantage of applying higher sequence similarity thresholds is that for general use in predicting the molecular function of membrane proteins, the resulting predictions would be likely to be more accurate as they are based on a larger training set.

Following the designed architecture, enzymes, GPCR proteins, ion channels, molecular transporters and proteins with transport activity (including ion channels and

molecular transporters) were predicted using separate classifiers that were trained following the one-against-all principle.

Evaluation of the different classifiers showed various classifiers that maximized the accuracy of the prediction of enzymes **(table 8.39)**. Comparison of these classifiers showed that the SVM(10) **(Please see appendix B table 8.10 on CD)** predicted enzymes with the highest sensitivity (90.5%) and was selected to be used in the TMFUN algorithm. Evaluation of the different classifiers to predict GPCR proteins showed various methods that reported Q, nQ and MCC values higher than 90 **(table 8.40)**. Among the different classifiers SVM(6) showed the highest values of sensitivity and **(Please see appendix B table 8.11 on CD)**. Prediction of molecular transporters using the one-against-all principle also showed that various methods could be successfully used for the prediction of this functional class **(table 8.41)**. Further comparison of these methods **(Please see appendix B table 8.12 on CD)** showed that the Bayesian network (attribute selection = true) and the SVM(5) predicted this functional class with the highest values of sensitivity. Although the SVM(5) reported a slightly lower sensitivity, the specificity score was more than ten percentage points higher than the sensitivity score for the Bayesian network. As a result, the SVM(5) was selected as the best performing classifier.

	Attribute selection			No attribute selection		
Data mining method	**Q**	**nQ**	**MCC**	**Q**	**nQ**	**MCC**
Bayesian networks	77.4	77.3	0.54	76.6	75.9	0.52
Naïve bayesian	67.6	64.1	0.35	69.7	66.9	0.38
Naïve Bayesian simple	68.4	65.2	0.36	-	-	-
Logistic regression	73.5	72.6	0.46	-	-	-
RBF Network	69.1	70.0	0.40	56.3	50.0	-
KStar	78.9	78.0	0.57	77.5	76.0	0.54
MultiBoostAB	81.6	81.5	0.63	85.4	85.4	0.71
J48	74.0	73.5	0.47	73.1	72.4	0.45
Random forest	79.6	79.6	0.59	81.0	81.2	0.62
Support vector machine (1)	73.7	72.8	0.46	75.1	74.1	0.49
Support vector machine (2)	73.7	72.8	0.46	75.1	74.1	0.49
Support vector machine (3)	73.6	72.7	0.46	75.0	74.0	0.49
Support vector machine (4)	76.2	76.8	0.53	79.9	79.1	0.59
Support vector machine (5)	71.9	73.5	0.48	77.8	76.0	0.55
Support vector machine (6)	78.8	78.4	0.57	86.1	86.2	0.72
Support vector machine (7)	79.6	79.1	0.58	84.4	84.1	0.68
Support vector machine (8)	79.6	79.1	0.58	84.4	84.1	0.68
Support vector machine (9)	78.7	78.5	0.57	79.9	79.1	0.59
Support vector machine (10)	80.2	79.9	0.60	**87.2**	**87.6**	**0.75**
Support vector machine (11)	81.9	81.6	0.63	86.6	86.8	0.73

Table 8.39. Ten fold cross-validation of different data mining techniques applied in a single-step fashion to discriminate between enzymes and non-enzymes. The highlighted cells correspond to the best performing data mining techniques.

	Attribute selection			No attribute selection		
Data mining method	**Q**	**nQ**	**MCC**	**Q**	**nQ**	**MCC**
Bayesian networks	96.0	95.9	0.90	94.3	94.7	0.86
Naïve bayesian	87.1	88.5	0.71	72.8	79.6	0.52
Naïve Bayesian simple	87.1	88.6	0.71	-	-	-
Logistic regression	88.3	84.6	0.70	90.5	88.3	0.76
RBF Network	88.5	88.5	0.73	77.9	65.2	0.37
KStar	94.4	93.2	0.86	96.0	94.9	0.90
MultiBoostAB	95.5	94.6	0.88	97.0	95.8	0.92
J48	92.0	89.5	0.79	91.4	89.2	0.78
Random forest	94.7	93.7	0.87	94.5	93.2	0.86
Support vector machine (1)	88.5	85.0	0.70	90.3	87.4	0.75
Support vector machine (2)	88.4	85.0	0.70	90.0	87.7	0.75
Support vector machine (3)	88.4	85.1	0.70	90.0	87.6	0.75
Support vector machine (4)	90.0	84.0	0.73	93.2	92.9	0.83
Support vector machine (5)	84.1	71.8	0.56	93.7	93.5	0.84
Support vector machine (6)	92.9	91.5	0.82	**97.3**	**96.0**	**0.93**
Support vector machine (7)	92.7	91.3	0.81	96.5	95.2	0.91
Support vector machine (8)	92.7	91.3	0.81	96.5	95.2	0.91
Support vector machine (9)	91.3	87.6	0.77	93.2	92.9	0.83
Support vector machine (10)	93.9	92.4	0.85	96.9	95.1	0.92
Support vector machine (11)	93.5	91.0	0.83	96.5	94.5	0.91

Table 8.40. Ten fold cross-validation of different data mining techniques applied in a single-step fashion to discriminate between GPCR proteins and non-GPCR proteins. The highlighted cells correspond to the best performing data mining techniques.

	Attribute selection			No attribute selection		
Data mining method	**Q**	**nQ**	**MCC**	**Q**	**nQ**	**MCC**
Bayesian networks	81.5	78.8	0.48	77.4	76.5	0.43
Naïve bayesian	65.8	73.8	0.35	56.4	68.2	0.27
Naïve Bayesian simple	65.3	73.4	0.35	-	-	-
Logistic regression	84.1	58.0	0.25	-	-	-
RBF Network	83.7	50.5	0.09	83.6	50.4	0.07
KStar	86.5	80.0	0.56	85.1	81.2	0.55
MultiBoostAB	89.9	74.9	0.59	90.6	74.5	0.62
J48	85.0	70.5	0.43	84.3	71.6	0.43
Random forest	89.0	74.8	0.56	88.7	71.4	0.53
Support vector machine (1)	83.5	50.0	-	83.6	51.4	0.11
Support vector machine (2)	83.5	50.0	-	84.3	60.2	0.29
Support vector machine (3)	83.5	50.0	-	84.3	60.6	0.30
Support vector machine (4)	86.3	61.2	0.38	87.1	78.9	0.55
Support vector machine (5)	85.3	60.4	0.33	**86.8**	**81.2**	**0.57**
Support vector machine (6)	88.7	79.4	0.59	92.1	81.6	0.69
Support vector machine (7)	87.5	77.3	0.55	90.5	81.6	0.65
Support vector machine (8)	87.5	77.3	0.55	90.5	81.6	0.65
Support vector machine (9)	87.0	71.9	0.49	87.1	78.9	0.55
Support vector machine (10)	89.6	79.7	0.61	92.4	80.5	0.70
Support vector machine (11)	89.4	72.3	0.56	91.5	77.5	0.66

Table 8.41. Ten fold cross-validation of different data mining techniques applied in a single-step fashion to discriminate between molecular transporters and non-molecular transporters. The cells coloured in yellow correspond to the best performing data mining techniques. The cells highlighted in red correspond to data mining techniques where the predictive scores are not representative of the predictive power of the method (not normalised). This is normally due to a large class being well predicted whereas the smaller class can not be accurately discriminated.

Evaluation of the different classifiers to predict ion channels **(table 8.42)** revealed several classifiers that needed further comparison **(Please see appendix B table 8.13 on CD)**. As with the previous data mining analysis, ion channels were the most difficult functional class to predict. In order to maximize the sensitivity of the prediction, the KStar method was selected. Evaluation of classifiers to predict proteins with transport activity showed that the SVM(7) and SVM(8) were the best performing classifiers **(table 8.43)**. Both classifiers performed equally well so any of these two classifiers could be used by the predictive architecture. This classifier was to be linked with another classifier to distinguish between molecular transporters and ion channels. The evaluation of the different data mining methods **(table 8.44)** showed that the SVM(11) (attribute selection = true) was the best performing classifier, which correctly predicted 84.2% and 80.5% of the ion channels and molecular transporters respectively. Evaluation of classifiers to distinguish between

class A GPCR proteins and other GPCR proteins showed a wide range of techniques with accurate values of Q, nQ and MCC **(table 8.45)**. Comparison of these methods showed that **(Please see appendix B table 8.14 on CD)** the Bayesian network classifier (attribute selection = true) predicted both class A GPCR and other GPCR proteins with sensitivity scores higher than 90% and was selected as the predictive classifier to be used in TMFUN.

	Attribute selection			No attribute selection		
Data mining method	**Q**	**nQ**	**MCC**	**Q**	**nQ**	**MCC**
Bayesian networks	75.0	69.3	0.29	68.1	66.3	0.23
Naïve bayesian	81.1	53.9	0.09	79.0	54.4	0.09
Naïve Bayesian simple	81.1	54.3	0.10	-	-	-
Logistic regression	86.5	52.1	0.11	85.6	56.8	0.20
RBF Network	86.7	50.0	-0.01	86.7	50.0	-0.01
KStar	87.8	75.0	0.48	**87.4**	**75.1**	**0.48**
MultiBoostAB	89.8	63.5	0.45	90.4	64.4	0.49
J48	85.1	65.7	0.33	85.3	66.2	0.34
Random forest	89.2	64.7	0.43	89.2	62.0	0.40
Support vector machine (1)	86.8	50.0	-	86.8	50.2	0.05
Support vector machine (2)	86.8	50.0	-	86.9	51.6	0.12
Support vector machine (3)	86.8	50.0	-	86.9	51.6	0.12
Support vector machine (4)	87.4	54.4	0.22	88.1	74.5	0.49
Support vector machine (5)	87.2	55.1	0.22	88.2	75.3	0.50
Support vector machine (6)	89.8	74.4	0.53	91.7	74.8	0.59
Support vector machine (7)	88.4	73.0	0.48	90.6	75.0	0.56
Support vector machine (8)	88.4	73.0	0.48	90.6	75.0	0.56
Support vector machine (9)	87.8	65.9	0.38	88.1	74.5	0.49
Support vector machine (10)	90.9	72.7	0.55	92.1	72.9	0.61
Support vector machine (11)	89.7	63.8	0.44	90.7	66.2	0.51

Table 8.42. Ten fold cross-validation of different data mining techniques applied in a single-step fashion to discriminate between ion channels and non-ion channels. The highlighted cells correspond to the best performing data mining techniques.

	Attribute selection			No attribute selection		
Data mining method	**Q**	**nQ**	**MCC**	**Q**	**nQ**	**MCC**
Bayesian networks	78.5	76.3	0.51	77.2	73.4	0.46
Naïve bayesian	68.1	72.0	0.40	61.9	68.0	0.33
Naïve Bayesian simple	68.0	71.7	0.40	-	-	-
Logistic regression	74.7	65.5	0.35	-	-	-
RBF Network	74.4	66.5	0.35	70.3	50.0	-
KStar	81.6	79.8	0.58	79.1	78.6	0.54
MultiBoostAB	84.8	80.5	0.63	86.7	81.4	0.67
J48	77.1	72.0	0.44	77.2	72.6	0.45
Random forest	82.9	78.6	0.58	83.0	77.5	0.58
Support vector machine (1)	76.0	66.1	0.38	77.4	68.9	0.42
Support vector machine (2)	76.1	67.1	0.38	77.0	69.8	0.42
Support vector machine (3)	76.1	67.1	0.38	77.1	70.0	0.42
Support vector machine (4)	77.1	64.3	0.39	83.9	81.5	0.62
Support vector machine (5)	77.1	64.3	0.39	80.9	79.8	0.57
Support vector machine (6)	85.1	81.2	0.64	87.4	82.9	0.69
Support vector machine (7)	83.9	81.0	0.62	**86.7**	**83.3**	**0.68**
Support vector machine (8)	83.9	81.0	0.62	**86.7**	**83.3**	**0.68**
Support vector machine (9)	81.5	75.3	0.54	83.9	81.5	0.62
Support vector machine (10)	81.5	75.3	0.54	88.0	82.3	0.70
Support vector machine (11)	85.7	79.7	0.64	87.1	81.0	0.68

Table 8.43. Ten fold cross-validation of different data mining techniques applied in a single-step fashion to discriminate between proteins with transport activity (including molecular transporters and ion channels) and proteins without transport activity. The highlighted cells correspond to the best performing data mining techniques.

	Attribute selection			No attribute selection		
Data mining method	**Q**	**nQ**	**MCC**	**Q**	**nQ**	**MCC**
Bayesian networks	70.9	70.0	0.41	66.9	65.3	0.32
Naïve bayesian	65.1	62.3	0.29	63.8	60.8	0.26
Naïve Bayesian simple	65.1	62.3	0.29	-	-	-
Logistic regression	68.3	67.3	0.35	69.3	68.6	0.38
RBF Network	65.8	63.7	0.30	56.4	51.2	0.08
KStar	80.4	79.7	0.60	80.2	79.1	0.60
MultiBoostAB	78.7	78.6	0.57	82.1	82.1	0.64
J48	70.2	69.9	0.40	68.0	67.7	0.35
Random forest	76.3	76.5	0.53	76.4	76.6	0.53
Support vector machine (1)	67.6	66.1	0.34	70.9	69.7	0.41
Support vector machine (2)	67.4	66.2	0.33	70.5	69.7	0.40
Support vector machine (3)	67.6	66.4	0.34	70.5	69.7	0.40
Support vector machine (4)	78.0	76.9	0.55	79.4	78.7	0.58
Support vector machine (5)	76.8	76.2	0.53	79.8	79.1	0.59
Support vector machine (6)	76.0	74.9	0.51	78.2	76.4	0.57
Support vector machine (7)	78.9	78.4	0.57	78.7	77.4	0.57
Support vector machine (8)	78.9	78.4	0.57	78.7	77.4	0.57
Support vector machine (9)	75.3	74.7	0.50	79.4	78.7	0.58
Support vector machine (10)	75.8	74.3	0.51	76.9	74.5	0.56
Support vector machine (11)	80.6	80.2	0.61	**82.1**	**82.4**	**0.64**

Table 8.44. Ten fold cross-validation of different data mining techniques applied in a single-step fashion to discriminate between molecular transporters and ion channels. The highlighted cells correspond to the best performing data mining technique.

	Attribute selection			No attribute selection		
Data mining method	**Q**	**nQ**	**MCC**	**Q**	**nQ**	**MCC**
Bayesian networks	**94.5**	**93.7**	**0.79**	90.3	89.4	0.67
Naïve bayesian	91.5	91.5	0.71	93.5	92.3	0.76
Naïve Bayesian simple	91.6	91.2	0.71			
Logistic regression	94.7	87.4	0.76	91.9	80.1	0.63
RBF Network	93.3	89.6	0.73	86.9	50.0	-
KStar	97.1	89.9	0.87	97.7	91.6	0.89
MultiBoostAB	95.8	86.6	0.81	97.1	90.2	0.87
J48	94.4	88.6	0.76	94.0	86.4	0.74
Random forest	95.9	87.0	0.81	95.8	85.2	0.80
Support vector machine (1)	96.0	89.6	0.82	96.5	91.5	0.85
Support vector machine (2)	95.3	88.6	0.79	96.6	91.9	0.85
Support vector machine (3)	94.8	88.3	0.77	96.6	91.9	0.85
Support vector machine (4)	96.5	92.3	0.84	98.0	93.8	0.91
Support vector machine (5)	94.6	89.5	0.77	97.8	92.5	0.90
Support vector machine (6)	97.1	89.9	0.87	96.8	88.0	0.85
Support vector machine (7)	97.4	92.3	0.88	97.8	92.2	0.90
Support vector machine (8)	97.4	92.3	0.88	97.8	92.2	0.90
Support vector machine (9)	96.5	92.3	0.84	98.0	93.8	0.91
Support vector machine (10)	95.8	84.1	0.80	94.2	77.8	0.72
Support vector machine (11)	96.1	85.7	0.82	94.3	78.1	0.73

Table 8.45. Ten fold cross-validation of different data mining techniques applied in a single-step fashion to discriminate between GPCR class A proteins and other GPCR proteins. This analysis is specific to GPCR proteins. The highlighted cells correspond to the best performing data mining techniques.

Sub-classification of class A GPCR proteins was also pursued following the one-against-all principle. Evaluation of the classifiers to predict amine GPCR proteins and other class A GPCR proteins showed that the SVM(5) was the best performing classifier **(table 8.46)**. Olfactory GPCR proteins were best predicted with SVM(7) and SVM(8) **(table 8.47)**. Both methods performed equally well so either of these two classifiers can be used by TMFUN to predict olfactory GPCR proteins. Evaluation of classifiers to predict peptide GPCR proteins showed various techniques **(table 8.48)** that required further comparison **(Please see appendix B table 8.15 on CD)**. The KStar method showed the highest sensitivity score for predicting peptide GPCR proteins and was considered the most consistent method. Prediction of rhodopsin GPCR proteins **(table 8.49)** showed several support vector machines that performed equally well, reporting impressive results. Consequently any of those support vector machines could be used in the TMFUN algorithm.

	Attribute selection			No attribute selection		
Data mining method	**Q**	**nQ**	**MCC**	**Q**	**nQ**	**MCC**
Bayesian networks	94.3	93.5	0.79	92.1	91.9	0.74
Naïve bayesian	91.5	91.8	0.72	84.5	86.3	0.57
Naïve Bayesian simple	91.6	91.6	0.72	-	-	-
Logistic regression	95.1	90.7	0.80	93.8	89.3	0.76
RBF Network	94.7	91.3	0.79	85.9	50.0	-
KStar	98.3	95.5	0.93	97.6	94.8	0.90
MultiBoostAB	98.2	95.7	0.93	97.5	93.3	0.90
J48	96.6	92.4	0.86	96.4	91.7	0.85
Random forest	98.2	95.7	0.93	97.2	92.2	0.88
Support vector machine (1)	95.8	88.7	0.82	98.3	96.7	0.93
Support vector machine (2)	96.7	93.1	0.86	98.3	96.4	0.93
Support vector machine (3)	96.5	93.3	0.86	98.3	96.4	0.93
Support vector machine (4)	98.3	96.1	0.93	98.8	97.0	0.95
Support vector machine (5)	98.4	96.7	0.93	**99.2**	**97.8**	**0.97**
Support vector machine (6)	97.6	93.3	0.90	98.2	94.3	0.92
Support vector machine (7)	97.8	94.3	0.91	98.5	95.6	0.94
Support vector machine (8)	97.8	94.3	0.91	98.5	95.6	0.94
Support vector machine (9)	98.3	96.1	0.93	98.8	97.0	0.95
Support vector machine (10)	96.8	90.2	0.86	96.8	88.7	0.86
Support vector machine (11)	96.7	89.6	0.86	96.8	88.7	0.86

Table 8.46. Ten fold cross-validation of different data mining techniques applied in a single-step fashion to discriminate between amine GPCR proteins and other GPCR class A proteins. This analysis is specific to GPCR class A proteins. The highlighted cells correspond to the best performing data mining techniques.

	Attribute selection			No attribute selection		
Data mining method	**Q**	**nQ**	**MCC**	**Q**	**nQ**	**MCC**
Bayesian networks	98.5	98.4	0.97	97.9	97.9	0.96
Naïve bayesian	96.6	96.5	0.93	96.5	96.5	0.93
Naïve Bayesian simple	96.5	96.4	0.93	-	-	-
Logistic regression	97.4	97.3	0.95	98.4	98.3	0.97
RBF Network	96.8	96.8	0.94	65.1	62.2	0.27
KStar	97.1	97.0	0.94	96.9	96.7	0.94
MultiBoostAB	97.8	97.8	0.96	97.7	97.7	0.95
J48	95.3	95.1	0.90	95.0	94.8	0.90
Random forest	97.4	97.4	0.95	97.5	97.5	0.95
Support vector machine (1)	97.8	97.6	0.96	98.8	98.7	0.98
Support vector machine (2)	97.4	97.2	0.95	98.8	98.7	0.98
Support vector machine (3)	97.5	97.4	0.95	98.8	98.7	0.98
Support vector machine (4)	97.6	97.4	0.95	98.8	98.7	0.98
Support vector machine (5)	97.1	96.9	0.94	98.5	98.3	0.97
Support vector machine (6)	98.3	98.3	0.97	98.7	98.8	0.97
Support vector machine (7)	98.4	98.3	0.97	**99.3**	**99.3**	**0.99**
Support vector machine (8)	98.4	98.3	0.97	**99.3**	**99.3**	**0.99**
Support vector machine (9)	97.6	97.4	0.95	98.8	98.7	0.98
Support vector machine (10)	98.2	98.1	0.96	98.7	98.7	0.97
Support vector machine (11)	97.9	97.6	0.96	97.9	97.6	0.96

Table 8.47. Ten fold cross-validation of different data mining techniques applied in a single-step fashion to discriminate between olfactory GPCR proteins and other GPCR class A proteins. This analysis is specific to GPCR class A proteins. The highlighted cells correspond to the best performing data mining techniques.

	Attribute selection			No attribute selection		
Data mining method	**Q**	**nQ**	**MCC**	**Q**	**nQ**	**MCC**
Bayesian networks	86.6	87.8	0.68	81.7	83.9	0.59
Naïve bayesian	85.0	85.7	0.64	84.5	86.5	0.64
Naïve Bayesian simple	84.4	85.4	0.63	-	-	-
Logistic regression	88.0	80.7	0.64	87.1	83.1	0.64
RBF Network	89.4	86.7	0.70	78.0	50.3	0.04
KStar	95.2	94.0	0.86	**96.5**	**95.8**	**0.90**
MultiBoostAB	93.1	88.5	0.79	93.1	87.2	0.79
J48	88.6	82.3	0.66	88.1	82.3	0.65
Random forest	93.1	88.5	0.79	92.3	85.8	0.77
Support vector machine (1)	88.9	80.8	0.66	92.0	87.9	0.77
Support vector machine (2)	89.5	83.5	0.69	89.4	86.8	0.71
Support vector machine (3)	89.6	83.7	0.69	89.4	86.8	0.71
Support vector machine (4)	94.8	93.8	0.85	96.1	95.4	0.89
Support vector machine (5)	93.9	92.9	0.83	95.9	95.3	0.88
Support vector machine (6)	96.9	94.9	0.91	96.8	93.7	0.91
Support vector machine (7)	96.2	94.5	0.89	96.7	95.1	0.90
Support vector machine (8)	96.2	94.5	0.89	96.7	95.1	0.90
Support vector machine (9)	94.8	93.8	0.85	96.1	95.4	0.89
Support vector machine (10)	96.9	94.4	0.91	95.6	90.2	0.87
Support vector machine (11)	95.6	91.5	0.87	95.5	89.8	0.87

Table 8.48. Ten fold cross-validation of different data mining techniques applied in a single-step fashion to discriminate between peptide GPCR proteins and other GPCR class A proteins. This analysis is specific to GPCR class A proteins. The highlighted cells correspond to the best performing data mining techniques.

	Attribute selection			No attribute selection		
Data mining method	**Q**	**nQ**	**MCC**	**Q**	**nQ**	**MCC**
Bayesian networks	98.6	94.0	0.92	98.0	93.7	0.88
Naïve bayesian	97.9	95.1	0.88	95.4	93.7	0.77
Naïve Bayesian simple	98.0	94.7	0.88	-	-	-
Logistic regression	98.7	97.9	0.93	98.3	97.2	0.91
RBF Network	98.7	95.0	0.92	91.8	57.8	0.35
KStar	99.3	96.8	0.96	99.2	98.1	0.95
MultiBoostAB	99.2	95.8	0.95	98.6	92.6	0.92
J48	97.4	91.0	0.85	97.1	91.8	0.83
Random forest	99.2	95.8	0.95	98.6	92.6	0.92
Support vector machine (1)	99.0	95.2	0.94	**99.9**	**99.9**	**0.99**
Support vector machine (2)	99.4	98.3	0.97	**99.9**	**99.9**	**0.99**
Support vector machine (3)	99.4	98.3	0.97	**99.9**	**99.9**	**0.99**
Support vector machine (4)	99.3	98.2	0.96	**99.9**	**99.9**	**0.99**
Support vector machine (5)	99.0	97.6	0.94	99.8	98.9	0.99
Support vector machine (6)	99.7	98.4	0.98	99.3	96.3	0.96
Support vector machine (7)	99.8	98.9	0.99	99.7	98.9	0.98
Support vector machine (8)	99.8	98.9	0.99	99.7	98.9	0.98
Support vector machine (9)	99.3	98.2	0.96	**99.9**	**99.9**	**0.99**
Support vector machine (10)	99.3	96.3	0.96	98.9	94.2	0.93
Support vector machine (11)	99.2	95.8	0.95	98.4	91.6	0.90

Table 8.49. Ten fold cross-validation of different data mining techniques applied in a single-step fashion to discriminate between rhodopsin GPCR proteins and other GPCR class A proteins. This analysis is specific to GPCR class A proteins. The highlighted cells correspond to the best performing data mining techniques.

Sub-classification of the ion channel class into cation and anion channels showed that the SVM(6) (attribute selection = true) and the KStar method were the best performing classifiers **(table 8.50)**. Further comparison of these methods **(Please see appendix B table 8.17 on CD)** showed that the KStar method was the best performing classifier predicting both classes with sensitivity scores higher than 80%.

	Attribute selection			No attribute selection		
Data mining method	**Q**	**nQ**	**MCC**	**Q**	**nQ**	**MCC**
Bayesian networks	77.0	76.2	0.48	65.6	67.3	0.30
Naïve bayesian	73.4	75.4	0.45	57.5	69.7	0.36
Naïve Bayesian simple	71.6	74.0	0.42	-	-	-
Logistic regression	81.3	72.4	0.48	68.1	64.1	0.26
RBF Network	76.1	64.5	0.32	74.7	51.0	0.10
KStar	86.5	81.1	0.64	**81.1**	**81.1**	**0.57**
MultiBoostAB	85.1	79.4	0.60	88.2	81.0	0.68
J48	78.8	72.7	0.45	78.0	71.2	0.42
Random forest	85.7	80.8	0.62	84.7	77.2	0.58
Support vector machine (1)	79.3	64.0	0.38	79.0	69.4	0.42
Support vector machine (2)	79.2	68.7	0.41	75.9	71.0	0.40
Support vector machine (3)	79.7	69.9	0.43	75.3	71.1	0.40
Support vector machine (4)	81.7	70.2	0.47	82.6	79.2	0.56
Support vector machine (5)	81.9	71.1	0.48	78.6	77.0	0.50
Support vector machine (6)	88.4	84.1	0.69	88.6	81.0	0.69
Support vector machine (7)	86.5	82.5	0.65	88.4	82.9	0.69
Support vector machine (8)	86.5	82.5	0.65	88.4	82.9	0.69
Support vector machine (9)	81.9	77.5	0.54	82.6	79.2	0.56
Support vector machine (10)	87.8	81.0	0.67	88.0	77.4	0.67
Support vector machine (11)	86.9	76.7	0.63	86.1	73.2	0.62

Table 8.50. Ten fold cross-validation of different data mining techniques applied in a single-step fashion to discriminate between cation channels and anion channels. This analysis is specific to ion channels. The cells coloured in yellow correspond to the best performing data mining techniques. The cells highlighted in red correspond to data mining techniques where the predictive scores are not representative of the predictive power of the method (not normalised). This is normally due to a large class being well predicted whereas the smaller class can not be accurately discriminated.

8.3.3 TMFUN evaluation

The different data mining analyses described above were carried out in order to identify the best performing classifier and architecture to predict the molecular activity of membrane proteins at different levels of complexity of molecular function. The prediction of each class has been individually evaluated by ten fold cross-validation. However, it is necessary to combine the different classifiers and evaluate such the whole architecture in order to demonstrate the overall predictive accuracy of TMFUN and to determine the importance of the contribution of the transmembrane region to determining and undertaking the molecular function of the protein.

TMFUN was evaluated by ten fold cross-validation. In each iteration, the different classifiers were trained using 90% of the assembled data set and the remaining 10% was predicted by TMFUN. Prediction was achieved using three different methods: i) assuming equally weighted classifiers, ii) assuming unequally weighted classifiers where the support for each classifier corresponds to the GAv score and iii) assuming unequally weighted classifiers where the support for each classifier corresponds to the MCC score. The GAv and MCC values for each classifier was obtained from the classifier evaluations previously carried out.

8.3.3.1 Data set filtered at a sequence similarity threshold of 40%

Evaluation of TMFUN using training sets of low sequence similarity showed that the overall prediction of the method assuming equally weighted classifiers **(table 8.51)** was less accurate than that assuming unequally weighted classifiers (**table 8.52** and **table 8.53**). The overall prediction scores of the method assuming unequally weighted classifiers based on the GAv **(table 8.52)** and the MCC **(table 8.53)** score showed similar values. However, at the most informative level (level 3) the predictive architecture based on the MCC scores were slightly better. Prediction of molecular function assuming weighted classifiers based on the corresponding MCC values showed identical or higher values of sensitivity for each functional class than assuming weighted classifiers based on the corresponding GAv values with the exception of ion channel. At the most informative level, with such low levels of sequence similarity, the best result was the 70% sensitivity to predict olfactory GPCR proteins. Other encouraging results were the sensitivity values to predict amine GPCR proteins and rhodopsin GPCR proteins (65%). At less informative levels, class A GPCR proteins were predicted with a sensitivity of 83.2% (level 2) and the GPCR superfamily was predicted with a sensitivity of 87.5%. 71.4% of the molecular transporters were correctly predicted and enzymes were predicted with a sensitivity of 64.1% (level 1b). Finally, proteins with transport activity were distinguished from enzymes and GPCR proteins with 70% accuracy (level 1a). The functional classes predicted with the lowest sensitivity scores include class A GPCR proteins other than amine, olfactory, peptide and rhodopsin GPCR proteins and ion channels. The confusion matrix derived from the ten fold

cross-validation of the TMFUN algorithm assuming weighted classifiers based on the corresponding MCC values **(table 8.54)** has been used to further investigate the incorrect predictions. Class A GPCR proteins other than amine, olfactory, peptide and rhodopsin GPCR proteins were found to be mainly predicted as other types of GPCR proteins (26.7% was predicted as peptide GPCR proteins and 13.3% was predicted as non-class A GPCR proteins). No classifier was actually designed to predict "other class A GPCR" – instead, when a protein predicted as class A was not further sub-classified as amine, olfactory, peptide or rhodopsin GPCR, TMFUN classified the given protein as "other class A GPCR", i.e. an unknown class A GPCR. The introduction of a classifier to predict this functional class might increase the sensitivity of the method. Interestingly, 33.3% of the peptide GPCR proteins were found to be predicted as "other class A GPCR". Therefore, there seems to be a reciprocal relationship between both classes that might reflect an evolutionary relationship between peptide GPCR proteins and a subset of proteins belonging to the "other class A GPCR" class. The ion channel class was found to be predicted at a sensitivity value of 12% **(table 8.53)**, the prediction of this functional class increased up to 24.4% and 30% assuming weighted classifiers based on the GAv values and equally weighted classifiers respectively. This functional class was found to be the most difficult class to predict and for considering the different classifiers to be used by TMFUN **(tables 8.5, 8.7, 8.13, 8.15)**. The confusion matrix **(table 8.54)** showed that the majority of ion channels tend to be predicted either as enzymes or molecular transporters. Similarly, the most common misclassification of molecular transporters involves enzymes and the most common misclassification of enzymes involved molecular transporters. Further analysis showed that approximately 50% of the enzymes predicted either as ion channels or molecular transporters do indeed have transport activity (e.g. ABC transporter, proton pump pyrophosphatase, and calcium ATPase). Using the information contained in the confusion matrix, a distance between classes was computed and plotted using the Biolayout software (Enright and Ouzounis, 2001). The obtained plot **(figure 8.6)** showed how enzymes, anion channels, cation channels and molecular transporters (edges in red) tend to form a cluster whereas GPCR proteins tend to from a separate cluster. These clusters are more obvious after filtering the distances between classes, reducing the background noise **(figure 8.7)**. With the exception of rhodopsin GPCR proteins, enzymes, ion channels and

molecular transporters form a cluster while the GPCR proteins form a separate cluster. The observed clusters and the fact that approximately 50% of the enzymes predicted as proteins with transport activity do indeed transport substances through the membrane might indicate that the current classifiers detect the corresponding signature of the function carried out in the membrane by the transmembrane regions rather than the overall protein function. Many enzymes (specially hydrolases) do transport substances through the membrane and also catalyze a chemical reaction. Based on the IUMB, these proteins were classified according to their overall enzymatic activity rather than the function carried out in the membrane. However, the features extracted by TMFUN only involve residues located in the membrane, which probably explains the observed misclassifications between enzymes, ion channels and molecular transporters. In order to further investigate this matter, it might be necessary to re-classify the assembled data set based solely on the function carried out in the membrane and restrict only the developed method to predicting the specific function local to the membrane. On the other hand, by combining TMDEPTH with data mining and pattern recognition methods applied to the extra-membranous region of the protein, the limitation of predicting local functions rather than the overall function of a membrane protein could be mitigated.

		Sensitivity	Specificity	GAv	Q	nQ	GC
1a	enzyme	38.3	80.9	55.7	51.0	54.4	0.56
	gpcr	57.5	85.2	70.0			
	transport	67.5	62.9	65.2			
1b	enzyme	38.3	80.7	55.6	43.0	45.4	0.48
	gpcr	57.5	85.2	70.0			
	ionchannel	30.4	34.4	32.3			
	transporter	55.4	54.0	54.7			
2	enzyme	38.3	80.7	55.6	40.2	35.4	0.41
	gpcra	60.5	85.8	72.1			
	non-gpcra	22.0	39.3	29.4			
	anion	18.0	9.6	13.1			
	cation	21.0	33.1	26.4			
	transporter	52.9	51.6	52.2			
3	enzyme	38.3	80.7	55.6	38	35	0.45
	amine	57.5	76.7	66.4			
	olfactory	52.5	87.5	67.8			
	peptide	43.3	51.0	47.0			
	rhodopsin	25.0	55.6	37.3			
	other gpcra	20.0	18.2	19.1			
	non-gpcra	22.0	44.0	31.1			
	anion	18.0	9.6	13.1			
	cation	21.0	24.5	22.7			
	transporter	52.9	51.6	52.2			

Table 8.51. Ten fold cross-validation of TMFUN. Where the consensus prediction was achieved assuming equally weighted classifiers.

		Sensitivity	Specificity	GAv	Q	nQ	GC
1a	enzyme	64.1	77.7	70.6	69.8	73.8	0.61
	gpcr	86.7	70.5	78.2			
	transport	70.8	64.1	67.3			
1b	enzyme	64.1	77.7	70.6	60.6	59.0	0.52
	gpcr	86.7	70.7	78.3			
	ionchannel	24.4	37.7	30.3			
	transporter	60.7	48.0	54.0			
2	enzyme	64.1	77.7	70.6	57.8	49.7	0.46
	gpcra	82.1	70.9	76.3			
	non-gpcra	60.0	41.1	49.7			
	anion	16.0	10.4	12.9			
	cation	15.5	35.2	23.4			
	transporter	60.7	48.0	54.0			
3	enzyme	64.1	77.7	70.6	54.7	49.5	0.5
	amine	65.0	74.3	69.5			
	olfactory	70.0	73.7	71.8			
	peptide	51.7	55.4	53.5			
	rhodopsin	55.0	42.3	48.2			
	other gpcra	36.7	16.9	24.9			
	non-gpcra	60.0	41.1	49.7			
	anion	16.0	10.4	12.9			
	cation	15.5	26.0	20.1			
	transporter	60.7	48.0	54.0			

Table 8.52. Ten fold cross-validation of TMFUN, where the consensus prediction was achieved assuming unequally weighted classifiers. The support for each classifier was obtained from the ten fold cross-validation of the best performing classifier and corresponded to the geometric average of the predicted functional class.

		Sensitivity	Specificity	GAv	Q	nQ	GC
1a	enzyme	64.1	77.7	70.6	69.6	73.9	0.61
	gpcr	87.5	70.5	78.5			
	transport	70.0	64.1	67.0			
1b	enzyme	64.1	77.7	70.6	60.7	58.7	0.52
	gpcr	87.5	70.5	78.5			
	ionchannel	12.0	36.6	21.0			
	transporter	71.4	45.8	57.2			
2	enzyme	64.1	77.7	70.6	58.8	48.8	0.46
	gpcra	83.2	70.9	76.8			
	non-gpcra	62.0	40.8	50.3			
	anion	2.0	3.0	2.5			
	cation	10.0	40.8	20.2			
	transporter	71.4	45.8	57.2			
3	enzyme	64.1	77.7	70.6	55.7	49.8	0.51
	amine	65.0	74.3	69.5			
	olfactory	70.0	71.8	70.9			
	peptide	51.7	55.4	53.5			
	rhodopsin	65.0	46.4	54.9			
	other gpcra	36.7	16.9	24.9			
	non-gpcra	62.0	41.3	50.6			
	anion	2.0	3.4	2.6			
	cation	10.0	24.1	15.5			
	transporter	71.4	45.8	57.2			

Table 8.53. Ten fold cross-validation of TMFUN, where the consensus prediction was achieved assuming unequally weighted classifiers. The support for each classifier was obtained from the ten fold cross-validation of the best performing classifier and corresponded to the Matthews correlation coefficient.

enzyme	amine	olfactory	peptide	rhodopsin	other gpcra	non-gpcra	anion	cation	transporter	
64.1	0.2	0.9	1.0	1.5	2.5	3.3	1.9	2.7	15.2	enzyme
0	65.0	0	12.5	2.5	15	0	0	0	5.0	amine
0	0	70.0	0	0	2.5	20	0	0	7.5	olfactory
1.7	6.7	1.7	51.7	0	33.3	3.3	0	1.7	0.0	peptide
0	0.0	0	0	65.0	5	0	5	0	25	rhodopsin
0	3.3	0	26.7	3.3	36.7	13.3	0	3.3	13.3	other gpcra
14	0	4	4	2	6	62	0	2	4.0	non-gpcra
12	4	0	0	0	0	0	2	4	64	anion
44.5	0	0	0	0	0.5	0.5	3.5	10	33	cation
16.1	0	0.4	0.7	0	0.7	0.7	1.8	2.1	71.4	transporter

Table 8.54. Confusion matrix corresponding to the ten fold cross-validation of TMFUN using weighted classifiers based on their corresponding Matthews correlation coefficient. Cells coloured in grey correspond to true positives, cells coloured in yellow correspond to misclassifications with a percentage error between 10% and 20%, cells coloured in red correspond to misclassifications with a percentage error equal or higher than 20%.

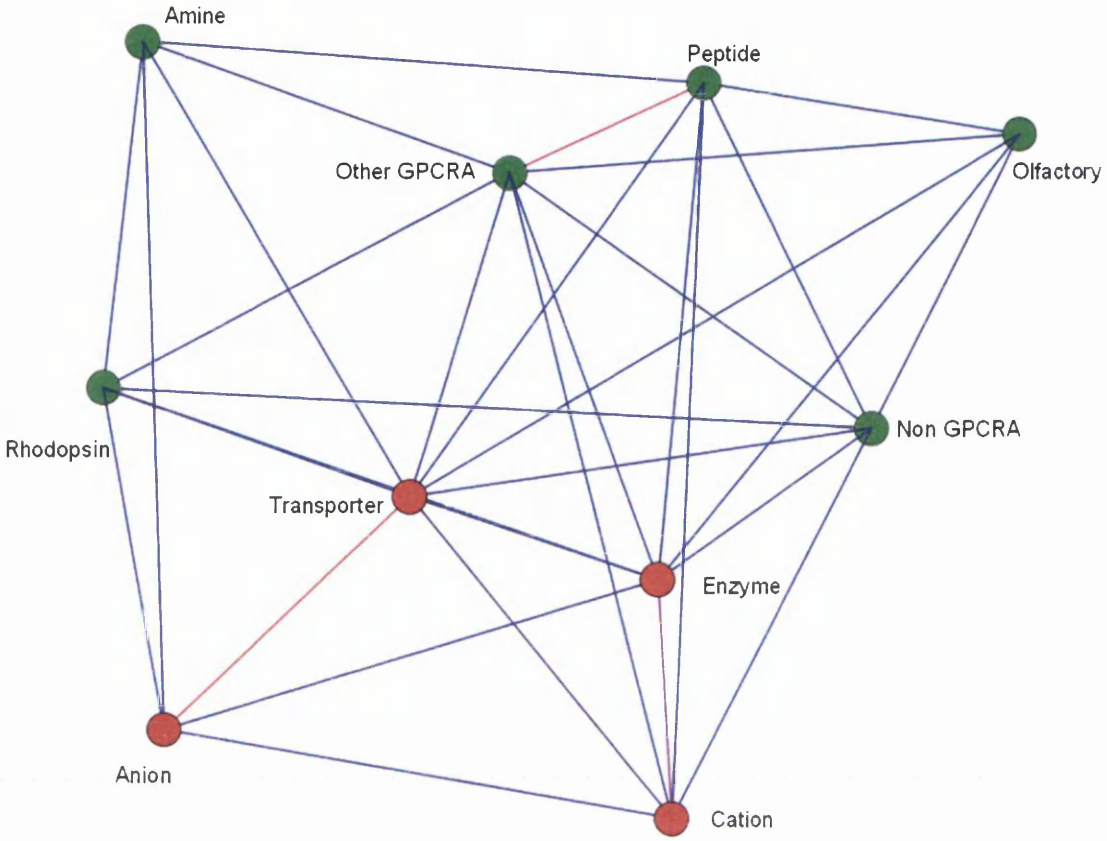

Figure 8.6. Distance relationships between the different functional classes extracted from the percentage confusion matrix obtained from the cross-validation of TMFUN assuming unequally weighted classifiers whose support corresponded to the Matthews correlation coefficient. The red edges correspond to enzymes and proteins with transport activities (molecular transporters, cation channels and anion channels).

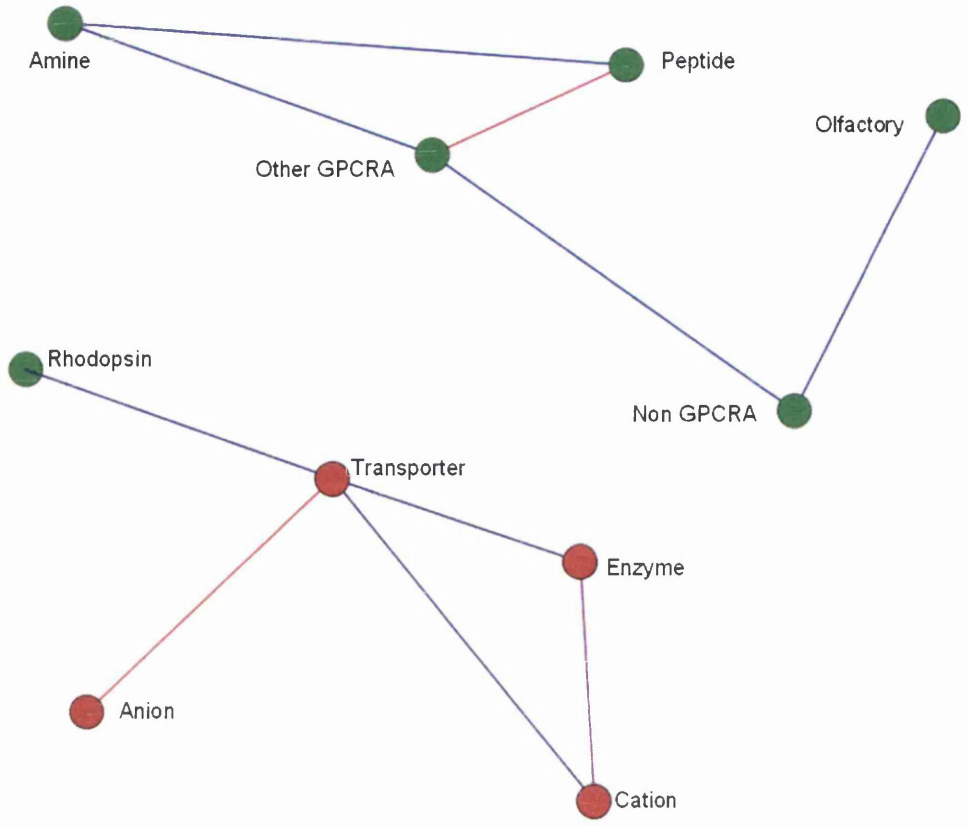

Figure 8.7. Distance relationships between the different functional classes extracted from the percentage confusion matrix obtained from the cross-validation of TMFUN assuming unequally weighted classifiers whose support corresponded to the Matthews correlation coefficient. The red edges correspond to enzymes and proteins with transport activities (molecular transporters, cation channels and anion channels). Here, the edges have been filtered to reduce the background noise using a threshold of 76%.

Although some of the predicted functional classes are broad families that could also be predicted using other methods, it is important to emphasize the importance of these results considering the limitations of our method. Firstly, topology prediction methods have a current success of 70% to 80% in correctly identifying all transmembrane regions of α-helical membrane proteins. Incorrect prediction or over-prediction of a single transmembrane region can completely alter the protein vector used to classify a given membrane protein. Therefore, these result are all the more impressive considering that the topological model of a significant proportion of proteins contained in the given data sets have some kind of error, which lead to erroneous protein vectors. Secondly, TMFUN exploits solely the information contained in the transmembrane regions. Therefore, these

results reflect the importance of this region to assigning the molecular function of α-helical membrane proteins. The functional work of the transmembrane regions has often been thought to be restricted to the less protein-specific roles of maintaining structure and facilitating conformational changes (with the exception, of course, of transport proteins and channels), while the extra-membranous loops of integral membrane proteins have been considered to play the major protein-specific functional roles such as ligand binding, chemical catalysis and signal transduction. There is no doubt about the importance of extra-membranous loops in accomodating the different functions carried out by the protein. However, these results indicate that the transmembrane regions of multi-spanning membrane proteins can also play a major part in refining molecular function. The most representative example supporting this idea is that by exploiting only the information contained in the transmembrane regions of GPCR proteins, the nature of its ligand can be predicted, even employing a training protocol that uses sequences of low sequence similarity.

TMFUN has been trained using severe conditions, where no sequence in the training set shares a sequence identity equal or higher than 40%. To our knowledge, no previous work has been carried out under such conditions. Despite the severe level of sequence similarity, TMFUN has proven the principle underlying the TMDEPTH approach, which states that signatures derived from the predicted topology of membrane proteins can be associated with specific molecular functions of membrane proteins.

8.3.3.2 Data set filtered at a sequence similarity threshold of 90%

As expected, evaluation of TMFUN using the data set filtered at a sequence similarity threshold of 90% reported higher predictive scores than using a more constrained data set. As explained above, using more flexible sequence similarity thresholds might result in a given classifier being biased towards the subset of proteins sharing a higher sequence identity. However, TMFUN was also trained under low sequence similarity not only to prove that the principle underlying TMDEPTH can be used to predict the molecular function of membrane proteins but also to facilitate the prediction of those less represented

proteins, which do not share significant sequence similarity with other proteins in the data set. On the other hand, the data set filtered at a sequence similarity threshold of 90% is a larger data set, which better represents the population of α-helical membrane proteins found in nature.

The differences between TMFUN assuming equally weighted classifiers **(table 8.55)** and using weighted classifiers based on the GAv scores **(table 8.56)** and MCC scores **(table 8.57)** were found to be smaller using a larger and more flexible data set. However, the former method was still the less accurate method for prediction of the molecular function of a membrane protein. Consensus prediction based on weighted classifiers (either using the GAv or MCC scores) predicted the molecular function of membrane proteins at its most informative level (level 3) with a generalized correlation (GC) of 0.81. Amine, olfactory and rhodopsin GPCR proteins were predicted with sensitivity values higher than 90%, peptide GPCR proteins were predicted with a sensitivity of 84.5% and 77.5% of class A GPCR proteins other than amine, olfactory, peptide and rhodopsin were correctly predicted. At level 2, class A GPCR proteins were predicted with 91.7% sensitivity and other classes of GPCR proteins were predicted with a sensitivity of 70%.Interestingly, 70% of the anion channels were correctly predicted whereas only 49% of the cation channels were correctly identified. At level 1b, enzymes, GPCR proteins, ion channels and molecular transporters were identified with a sensitivity of 87.8%, 92.8%, 58.3% and 75.6% respectively. Finally, 74.9% of proteins with transport activity (ion channels and molecular transporters) were correctly predicted at the less informative level (level 1a). The observed results re-emphasize the importance of the transmembrane regions in being able to define molecular function. This is particularly obvious for amine (92.9%), olfactory (92.3%) and rhodopsin (93%) GPCR proteins. Some of the misclassifications reported under low sequence similarity have been hidden by a larger fraction of correct predictions probably corresponding to subsets of proteins sharing a significant sequence similarity. However, the relationship between enzymes, molecular transporters and ion channels is still obvious **(table 8.58)**. The majority of misclassifications corresponded to erroneous predictions between two of these functional classes, 33.6% and 11.5% of the cation channels were predicted as enzymes and molecular transporters respectively, 12.3% of the anion channels

were predicted as molecular transporters and 17.4% of the molecular transporters were predicted as enzymes. The relationship between these functional classes is evident in the plot made based on the distance relationships between the different functional classes **(figure 8.8)**. Enzymes, anion channels, cation channels and molecular transporters (red edges) formed a cluster, whereas GPCR subclasses (green edges) formed a separated cluster. These clusters are more obvious after filtering the distances between classes, reducing the background noise **(figure 8.9)**. As explained earlier, these results probably reflect the transport activity of some enzymes such as ABC transporters and copper ATPases.

		Sensitivity	Specificity	GAv	Q	nQ	GC
1a	enzyme	74.1	87.0	80.3	78.7	79.8	0.81
	gpcr	89.7	98.4	94.0			
	transport	75.6	84.6	80.0			
1b	enzyme	74.1	87.0	80.2	76.3	74.0	0.76
	gpcr	89.7	98.4	94.0			
	ionchannel	59.0	75.6	66.8			
	transporter	73.1	81.0	76.9			
2	enzyme	74.1	87.0	80.2	74.8	69.6	0.71
	gpcra	89.7	97.9	93.7			
	non-gpcra	62.0	69.9	65.8			
	anion	69.2	64.3	66.7			
	cation	49.5	72.6	59.9			
	transporter	73.1	81.0	76.9			
3	enzyme	74.1	87.0	80.2	74.3	76.1	0.8
	amine	90.7	99.2	94.9			
	olfactory	91.6	98.3	94.9			
	peptide	83.6	92.5	87.9			
	rhodopsin	91.0	98.9	94.9			
	other gpcra	75.8	77.8	76.8			
	non-gpcra	62.0	69.9	65.8			
	anion	69.2	64.3	66.7			
	cation	49.5	9.3	21.4			
	transporter	73.1	81.0	76.9			

Table 8.55. Ten fold cross-validation of TMFUN, where the consensus prediction was achieved assuming equally weighted classifiers.

		Sensitivity	Specificity	GAv	Q	nQ	GC
1a	enzyme	87.8	84.8	86.3	85.4	85.2	0.81
	gpcr	92.8	94.8	93.8			
	transport	74.8	84.7	79.6			
1b	enzyme	87.8	84.8	86.3	83.5	78.6	0.77
	gpcr	92.8	95.9	94.4			
	ionchannel	58.3	78.5	67.6			
	transporter	75.6	77.7	76.6			
2	enzyme	87.8	84.7	86.2	82	74	0.72
	gpcra	91.7	96.4	94.0			
	non-gpcra	70.0	63.3	66.5			
	anion	70.0	68.4	69.2			
	cation	49.0	74.3	60.3			
	transporter	75.6	77.7	76.6			
3	enzyme	87.8	84.7	86.2	81.4	79.3	0.81
	amine	92.9	100.0	96.4			
	olfactory	92.3	96.1	94.2			
	peptide	84.5	92.5	88.5			
	rhodopsin	93.0	98.9	95.9			
	other gpcra	77.5	75.6	76.5			
	non-gpcra	70.0	63.3	66.5			
	anion	70.0	68.4	69.2			
	cation	49.0	9.8	21.9			
	transporter	75.6	77.7	76.6			

Table 8.56. Ten fold cross-validation of TMFUN, where the consensus prediction was achieved assuming unequally weighted classifiers. The support for each classifier was obtained from the ten fold cross-validation of the best performing classifier and corresponded to the geometric average of the predicted functional class.

		Sensitivity	Specificity	GAv	Q	nQ	GC
1a	enzyme	87.8	84.7	86.2	85.4	85.2	0.82
	gpcr	92.8	95.6	94.2			
	transport	74.9	84.8	79.7			
1b	enzyme	87.8	84.7	86.2	83.5	78.6	0.76
	gpcr	92.8	95.7	94.3			
	ionchannel	58.3	77.7	67.3			
	transporter	75.6	77.7	76.6			
2	enzyme	87.8	84.7	86.2	82	74	0.72
	gpcra	91.7	96.4	94.0			
	non-gpcra	70.0	63.3	66.5			
	anion	70.0	68.4	69.2			
	cation	49.0	74.3	60.3			
	transporter	75.6	77.7	76.6			
3	enzyme	87.8	84.7	86.2	81.4	79.3	0.81
	amine	92.9	100.0	96.4			
	olfactory	92.3	96.1	94.2			
	peptide	84.5	92.5	88.5			
	rhodopsin	93.0	98.9	95.9			
	other gpcra	77.5	75.0	76.2			
	non-gpcra	70.0	63.6	66.7			
	anion	70.0	68.4	69.2			
	cation	49.0	9.8	21.9			
	transporter	75.6	77.7	76.6			

Table 8.57. Ten fold cross-validation of TMFUN, where the consensus prediction was achieved assuming unequally weighted classifiers. The support for each classifier was obtained from the ten fold cross-validation of the best performing classifier and corresponded to the Matthews correlation coefficient.

enzyme	amine	olfactory	peptide	rhodopsin	other gpcra	non-gpcra	anion	cation	transporter	
87.8	0	0.6	0.3	0	0.4	0.8	1.2	1.9	4.6	enzyme
0	92.9	0	1.4	0	2.9	0.7	0	0	0	amine
3.3	0	92.3	0.2	0	0.2	1.4	0	0.5	0.5	olfactory
0	0	0	84.5	0.5	5.5	7.7	0	0	0	peptide
2	0	0	0	93.0	1	2	0	0	1	rhodopsin
0.8	0	1.7	3.3	0	77.5	11.7	0	0.8	0	other gpcra
12	0	0	2.0	0	2.7	70	0	1.3	3.3	non-gpcra
8.5	0	0	0	0	0	0	70	6.2	12.3	anion
33.6	0	0	0	0	0	0.3	3.3	49	11.5	cation
17.4	0	0.3	0	0	0.3	0.4	0.8	2.2	75.6	transporter

Table 8.58. Confusion matrix corresponding to the ten fold cross-validation of TMFUN using weighted classifiers based on their corresponding Matthews correlation coefficient. Cells coloured in grey belong to true positives, cells coloured in yellow correspond to misclassifications with a percentage error between 10% and 20%, and cells coloured in red correspond to misclassifications with a percentage error equal or higher than 20%.

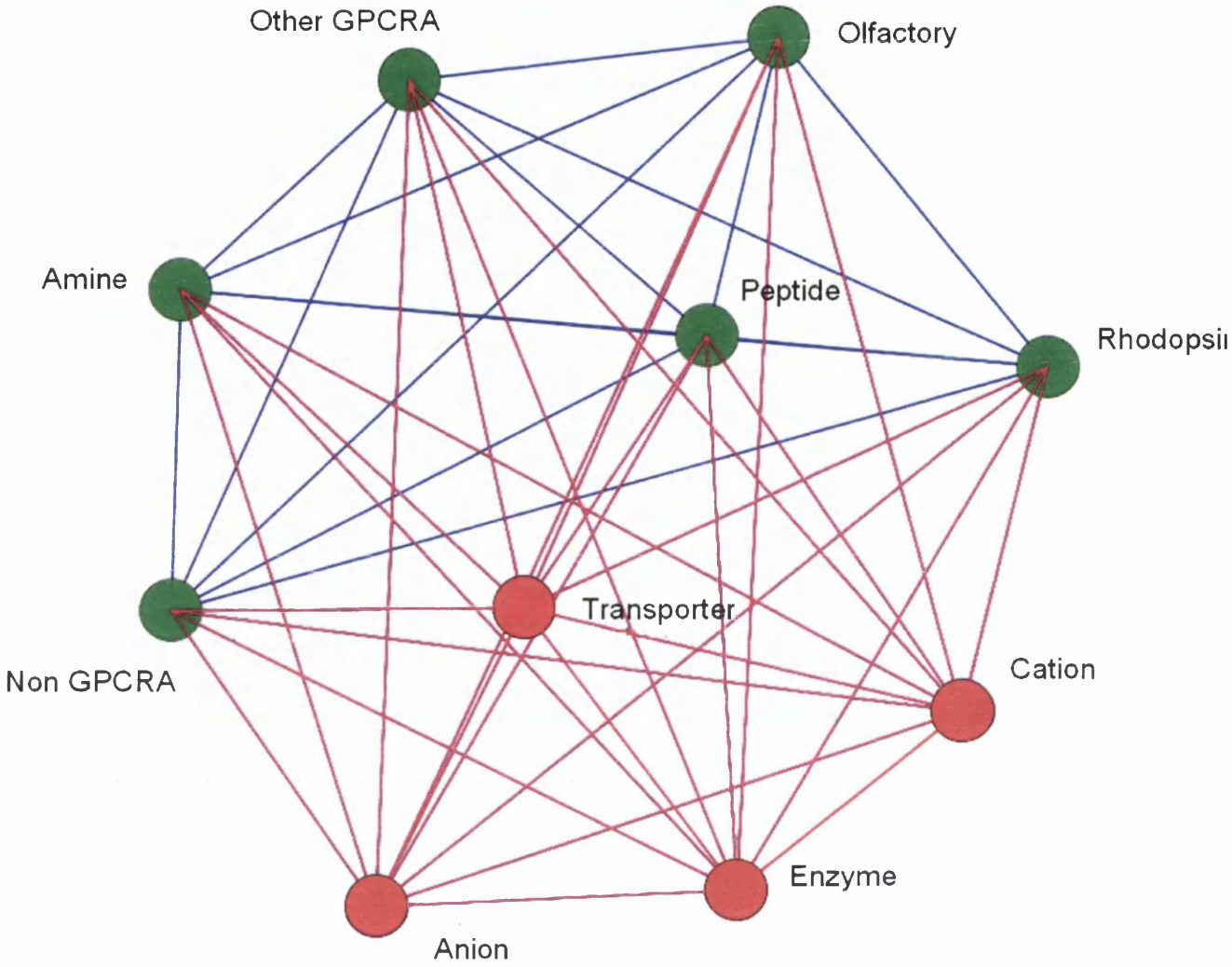

Figure 8.8. Distance relationships between the different functional classes extracted from the percentage confusion matrix obtained from the cross-validation of TMFUN assuming unequally weighted classifiers whose support corresponded to the Matthews correlation coefficient. The red edges correspond to enzymes and proteins with transport activities (molecular transporters, cation channels and anion channels).

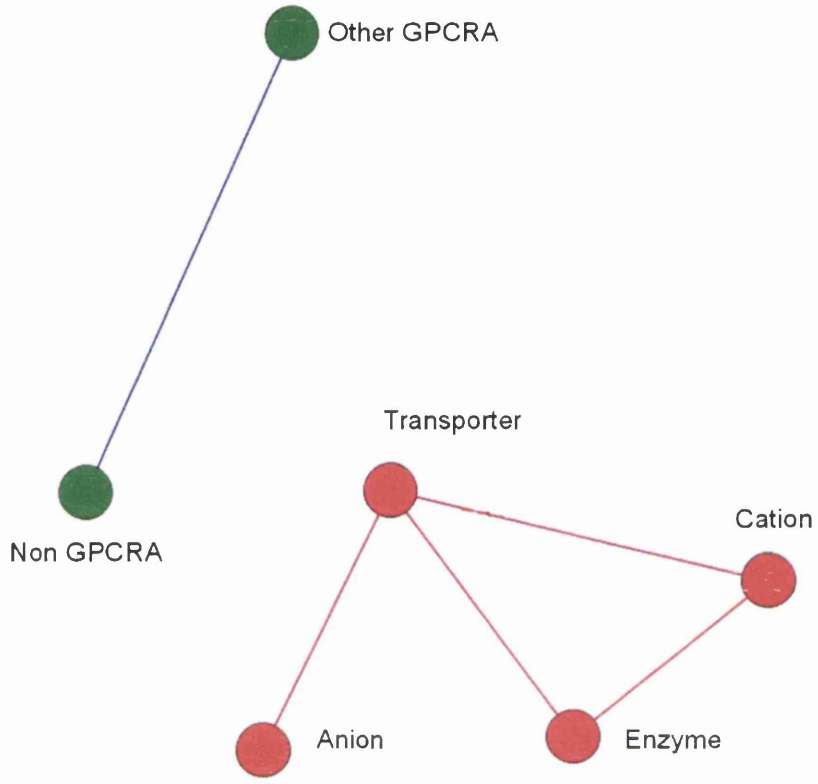

Figure 8.9. Distance relationships between the different functional classes extracted from the percentage confusion matrix obtained from the cross-validation of TMFUN assuming unequally weighted classifiers whose support corresponded to the Matthews correlation coefficient. The red edges correspond to enzymes and proteins with transport activities (molecular transporters, cation channels and anion channels). Here, the edges have been filtered to reduce the background noise using a threshold of 42%.

In order to compare the predictive accuracy of TMFUN based on sequence similarity thresholds of 40% and 90%, the Euclidean distance **(1.11)** between both versions was computed **(table 8.59)**. This distance was computed at each level using the three predictive scores obtained from the evaluation of the method, namely the overall predictive accuracy (Q), the normalized accuracy (nQ) and the generalized correlation (GC). The obtained results showed that the Euclidean distance between the TMFUN method trained using a sequence similarity threshold of 40% and 90% increases with the level of functional complexity. Therefore, as the prediction given by TMFUN further specifies the molecular role of the unknown protein, the TMFUN version trained using a data set filtered with a sequence similarity threshold of 90% increases its accuracy with respect to the TMFUN version trained using a sequence similarity threshold of 40%. The observed relationship

between the predictive power of both versions seems to be logarithmic **(figure 8.10)**. Extrapolating these results the differential magnitude between the predictive accuracy using TMFUN trained at a sequence similarity threshold of 90% and TMFUN trained at a sequence similarity threshold of 40% could increase until no less than 35 different classes can be discriminated (at an expected Euclidean distance value of approximately 60-65). There are two major factors that might explain why using a training set filtered at increasing sequence similarity filter reports more accurate values. One factor is the number of sequences contained in the training set, in order to further sub-classify newly predicted classes a significant number of instances is required. At more severe levels of sequence similarity it is likely to reach earlier down the tree a node where there are not enough instances to appropriately train the classifier. The second reason is that by using a more flexible training set, where the sequence similarity filter appropriately removes highly identical instances but still maintains other instances with marked similarity, better trained classifiers can be obtained. By using a sequence similarity filter of 90%, highly redundant proteins sequences were removed. However, protein families having a marked sequence similarity among their corresponding members are not constraint by severe sequence similarity thresholds and therefore a better classifier can be implemented. Following this principle, the TMFUN version trained at a sequence similarity of 90% implemented better classifiers for those protein classes with higher number of sequences (compared to the data set filtered at a sequence similarity threshold of 40%) while still not falling into bias due to the presence of highly identical sequence. Comparison of these two approaches showed that on average 83 additional instances were misclassified using the TMFUN version trained at a sequence similarity threshold of 90%. This difference is not considered important taking into account that the data set filtered at a sequence similarity threshold of 90% contains 2,807 more instances that the data set filtered at a sequence similarity threshold of 40%.

	Seq sim filter <40%			Seq sim filter <90%			
	Q	nQ	GC	Q	nQ	GC	Euclidean distance
Level 1a	69.6	73.9	0.61	85.4	85.2	0.82	19.4
Level 1b	60.7	58.7	0.52	83.5	78.6	0.76	30.3
Level 2	58.8	48.8	0.46	82	74	0.72	34.3
Level 3	55.7	49.8	0.51	81.4	79.3	0.81	39.1

Table 8.59. Comparison of the predictive accuracy of TMFUN using different thresholds of sequence similarity. The columns coloured in orange correspond to the ten fold cross-validation of TMFUN based sequence similarity threshold of 40%. The columns coloured in blue correspond to the ten fold cross-validation of TMFUN based sequence similarity threshold of 90%. In order to compare the predictive accuracy between both versions of TMFUN at a particular level, the Euclidean distance was calculated using the overall accuracy score (Q), the normalized accuracy score (nQ) and the generalized correlation (GC).

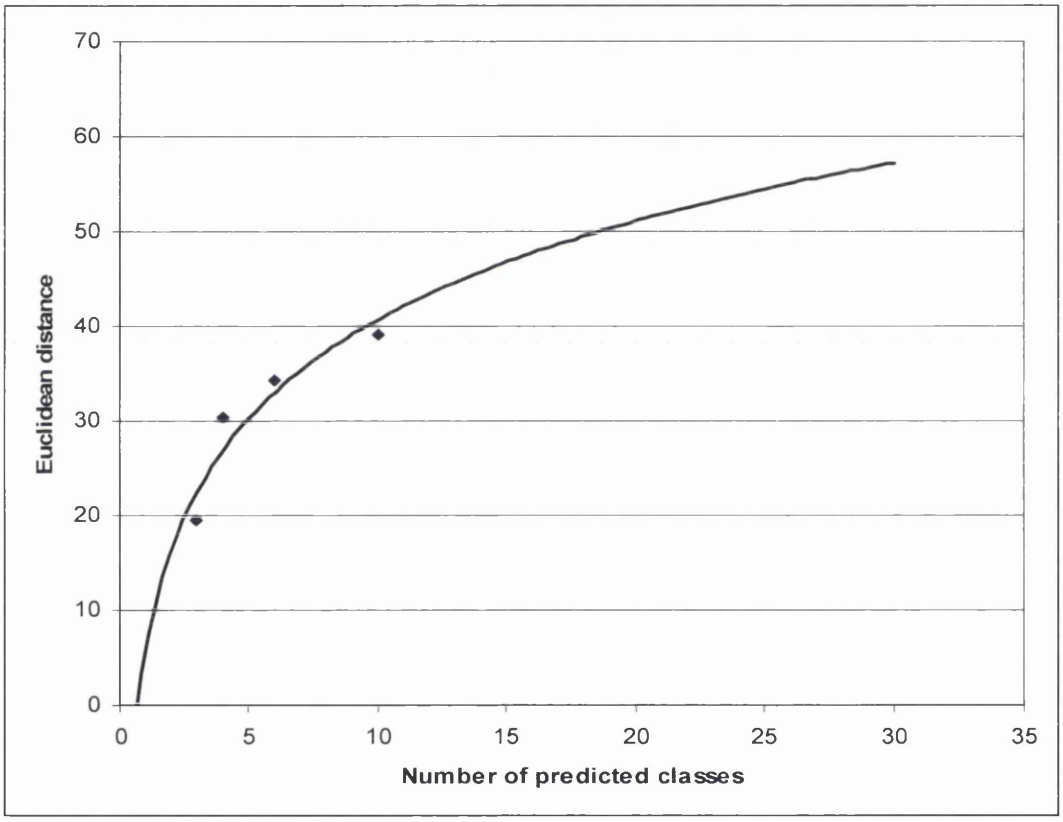

Figure 8.10. Comparison of the predictive accuracy of TMFUN using different thresholds of sequence similarity. The graph shows the Euclidean distance between evaluation results obtained at different levels between TMFUN trained with a sequence similarity threshold of 40% and 90%. Increasing values of Euclidean distance shows the increase of the predictive power of the TMFUN version trained at a sequence similarity threshold of 90% with respect to that found for the TMFUN version trained at a sequence similarity threshold of 40%.

Although the TMFUN version trained at a filtered sequence similarity threshold of 40% has proven that signatures derived from the predicted topology of membrane proteins can be associated with specific molecular functions of membrane proteins, the obtained results showed that in order to provide a tool for the automated prediction of membrane proteins, the TMFUN version trained at a sequence similarity threshold of 90% might be more appropriate. However, the TMFUN version trained at a sequence similarity threshold of 40% might still be useful to predict the function of orphan proteins despite the fact that less highly informative predictions might be obtained.

An alternative option would involve the development of a hypothetical TMFUN version trained using sequence similarity thresholds that range between 40%-90% in order to incorporate the advantages of the approaches trained at a sequence similarity threshold of 40% and 90%. However, to our understanding this would not show any additional improvement for various reasons: i) it will not prove the method as well as if a sequence similarity threshold of 40% was applied and ii) it might constrain protein classes with marked sequence similarity so that the obtained classifier does not maximize its prediction.

8.4 Conclusions

Experimental approaches to annotate genes can not cope with the rate at which new genomes are being sequenced. Alternatively, different computational methods are being developed to predict the function of these genes, thus facilitating experimental validation. However, the majority of these computational methods have been designed for the functional prediction of soluble protein. Furthermore, many of the few specific computational approaches to predict particular functional classes specifically found in the membrane have not been appropriately evaluated. The developed approach is probably the first rigorous data mining method applied for the functional prediction of polytopic membrane proteins. The data set was filtered at sequence similarity thresholds of 40% and 90% and the negative control sets used during the data mining process corresponded to other polytopic membrane proteins. The underlying principle for the feature extraction method is that signatures derived from the predicted topology of membrane proteins can be

associated with specific molecular functions of membrane proteins. Obtained results with the data set filtered at a sequence similarity threshold of 40% proved this principle. Thus, reflecting the importance of this region to assigning the molecular function of α-helical membrane proteins and proving wrong the assumption of transmembrane regions being mainly restricted to structural and conformational roles (with few exceptions). Classification analysis using the data set trained at a sequence similarity threshold of 90% showed sensitivity values of 70%-93% for predicting the different functional classes considered. More flexible data sets that still avoid pairs of highly identical proteins might be more appropriate as larger sets permit more informative predictions and protein families with marked sequence similarity are not constrained by severe sequence similarity thresholds. This approach showed an MCC value of 0.81 at the most informative level. To our knowledge, TMFUN is the most thorough single computational approach for prediction of the molecular function of polytopic membrane proteins developed to date, and should serve as a good baseline method for future computational developments.

8.5 References

ABASCAL, F. & VALENCIA, A. (2002) Clustering of proximal sequence space for the identification of protein families. *Bioinformatics,* 18, 908-21.

ALTSCHUL, S. F., GISH, W., MILLER, W., MYERS, E. W. & LIPMAN, D. J. (1990) Basic local alignment search tool. *J Mol Biol,* 215, 403-10.

ALTSCHUL, S. F., MADDEN, T. L., SCHAFFER, A. A., ZHANG, J., ZHANG, Z., MILLER, W. & LIPMAN, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res,* 25, 3389-402.

ARAI, M., OKUMURA, K., SATAKE, M. & SHIMIZU, T. (2004) Proteome-wide functional classification and identification of prokaryotic transmembrane proteins by transmembrane topology similarity comparison. *Protein Sci,* 13, 2170-83.

ARMON, A., GRAUR, D. & BEN-TAL, N. (2001) ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *J Mol Biol,* 307, 447-63.

ATTWOOD, T. K., FLOWER, D. R., LEWIS, A. P., MABEY, J. E., MORGAN, S. R., SCORDIS, P., SELLEY, J. N. & WRIGHT, W. (1999) PRINTS prepares for the new millennium. *Nucleic Acids Res,* 27, 220-5.

BALDI, P., CHAUVIN, Y., HUNKAPILLER, T. & MCCLURE, M. A. (1994) Hidden Markov models of biological primary sequence information. *Proc Natl Acad Sci U S A,* 91, 1059-63.

BASHFORD, D., CHOTHIA, C. & LESK, A. M. (1987) Determinants of a protein fold. Unique features of the globin amino acid sequences. *J Mol Biol,* 196, 199-216.

BATEMAN, A., BIRNEY, E., CERRUTI, L., DURBIN, R., ETWILLER, L., EDDY, S. R., GRIFFITHS-JONES, S., HOWE, K. L., MARSHALL, M. & SONNHAMMER, E. L. (2002) The Pfam protein families database. *Nucleic Acids Res,* 30, 276-80.

BENDTSEN, J. D., JENSEN, L. J., BLOM, N., VON HEIJNE, G. & BRUNAK, S. (2004) Feature-based prediction of non-classical and leaderless protein secretion. *Protein Eng Des Sel,* 17, 349-56.

BHASIN, M. & RAGHAVA, G. P. (2004) GPCRpred: an SVM-based method for prediction of families and subfamilies of G-protein coupled receptors. *Nucleic Acids Res,* 32, W383-9.

BHASIN, M. & RAGHAVA, G. P. (2005) GPCRsclass: a web tool for the classification of amine type of G-protein-coupled receptors. *Nucleic Acids Res,* 33, W143-7.

BUCHER, P., KARPLUS, K., MOERI, N. & HOFMANN, K. (1996) A flexible motif search technique based on generalized profiles. *Comput Chem,* 20, 3-23.

CAI, C. Z., HAN, L. Y., JI, Z. L., CHEN, X. & CHEN, Y. Z. (2003) SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence. *Nucleic Acids Res,* 31, 3692-7.

CAI, Y. D. & CHOU, K. C. (2005) Predicting enzyme subclass by functional domain composition and pseudo amino acid composition. *J Proteome Res,* 4, 967-71.

CAI, Y. D. & LIN, S. L. (2003) Support vector machines for predicting rRNA-, RNA-, and DNA-binding proteins from amino acid sequence. *Biochim Biophys Acta,* 1648, 127-33.

CAO, J., PANETTA, R., YUE, S., STEYAERT, A., YOUNG-BELLIDO, M. & AHMAD, S. (2003) A naive Bayes model to predict coupling between seven transmembrane domain receptors and G-proteins. *Bioinformatics,* 19, 234-40.

CASARI, G., SANDER, C. & VALENCIA, A. (1995) A method to predict functional residues in proteins. *Nat Struct Biol,* 2, 171-8.

CHIU, J. C., LEE, E. K., EGAN, M. G., SARKAR, I. N., CORUZZI, G. M. & DESALLE, R. (2006) OrthologID: automation of genome-scale ortholog identification within a parsimony framework. *Bioinformatics,* 22, 699-707.

CHOU, K. C. (2001) Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins,* 43, 246-55.

CHOU, K. C. & CAI, Y. D. (2003) Prediction and classification of protein subcellular location-sequence-order effect and pseudo amino acid composition. *J Cell Biochem,* 90, 1250-60.

CHOU, K. C. & CAI, Y. D. (2004) Predicting enzyme family class in a hybridization space. *Protein Sci,* 13, 2857-63.

DANDEKAR, T., SNEL, B., HUYNEN, M. & BORK, P. (1998) Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem Sci,* 23, 324-8.

DARZENTAS, N., RIGOUTSOS, I. & OUZOUNIS, C. A. (2005) Sensitive detection of sequence similarity using combinatorial pattern discovery: a challenging study of two distantly related protein families. *Proteins,* 61, 926-37.

DEL SOL MESA, A., PAZOS, F. & VALENCIA, A. (2003) Automatic methods for predicting functionally important residues. *J Mol Biol,* 326, 1289-302.

DES JARDINS, M., KARP, P. D., KRUMMENACKER, M., LEE, T. J. & OUZOUNIS, C. A. (1997) Prediction of enzyme classification from protein sequence without the use of sequence similarity. *Proc Int Conf Intell Syst Mol Biol,* 5, 92-9.

DEVOS, D. & VALENCIA, A. (2001) Intrinsic errors in genome annotation. *Trends Genet,* 17, 429-31.

EDDY, S. R. (1996) Hidden Markov models. *Curr Opin Struct Biol,* 6, 361-5.

EISEN, J. A., SWEDER, K. S. & HANAWALT, P. C. (1995) Evolution of the SNF2 family of proteins: subfamilies with distinct sequences and functions. *Nucleic Acids Res,* 23, 2715-23.

ENRIGHT, A. J., ILIOPOULOS, I., KYRPIDES, N. C. & OUZOUNIS, C. A. (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature,* 402, 86-90.

ENRIGHT, A. J. & OUZOUNIS, C. A. (2001) BioLayout--an automatic graph layout algorithm for similarity visualization. *Bioinformatics,* 17, 853-4.

FALQUET, L., PAGNI, M., BUCHER, P., HULO, N., SIGRIST, C. J., HOFMANN, K. & BAIROCH, A. (2002) The PROSITE database, its status in 2002. *Nucleic Acids Res,* 30, 235-8.

FETROW, J. S. & SKOLNICK, J. (1998) Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T1 ribonucleases. *J Mol Biol,* 281, 949-68.

FITCH, W. M. (1970) Distinguishing homologous from analogous proteins. *Syst Zool,* 19, 99-113.

FRYXELL, K. J. (1996) The coevolution of gene family trees. *Trends Genet,* 12, 364-9.

GABALDON, T. & HUYNEN, M. A. (2004) Prediction of protein function and pathways in the genome era. *Cell Mol Life Sci,* 61, 930-44.

GALPERIN, M. Y. & KOONIN, E. V. (2000) Who's your neighbor? New computational approaches for functional genomics. *Nat Biotechnol,* 18, 609-13.

GERSTEIN, M. (1998) Measurement of the effectiveness of transitive sequence comparison, through a third 'intermediate' sequence. *Bioinformatics,* 14, 707-14.

GILKS, W. R., AUDIT, B., DE ANGELIS, D., TSOKA, S. & OUZOUNIS, C. A. (2002) Modeling the percolation of annotation errors in a database of protein sequences. *Bioinformatics,* 18, 1641-9.

GOH, C. S., BOGAN, A. A., JOACHIMIAK, M., WALTHER, D. & COHEN, F. E. (2000) Co-evolution of proteins with their interaction partners. *J Mol Biol,* 299, 283-93.

GOODSTADT, L. & PONTING, C. P. (2006) Phylogenetic reconstruction of orthology, paralogy, and conserved synteny for dog and human. *PLoS Comput Biol,* 2, e133.

GRIBSKOV, M., MCLACHLAN, A. D. & EISENBERG, D. (1987) Profile analysis: detection of distantly related proteins. *Proc Natl Acad Sci U S A,* 84, 4355-8.

GUAN, C. P., JIANG, Z. R. & ZHOU, Y. H. (2005) Predicting the coupling specificity of GPCRs to G-proteins by support vector machines. *Genomics Proteomics Bioinformatics,* 3, 247-51.

GUO, Y. Z., LI, M. L., WANG, K. L., WEN, Z. N., LU, M. C., LIU, L. X. & JIANG, L. (2005) Fast fourier transform-based support vector machine for prediction of G-protein coupled receptor subfamilies. *Acta Biochim Biophys Sin (Shanghai),* 37, 759-66.

HANNENHALLI, S. S. & RUSSELL, R. B. (2000) Analysis and prediction of functional sub-types from protein sequence alignments. *J Mol Biol,* 303, 61-76.

HEBERT, T. E. & BOUVIER, M. (1998) Structural and functional aspects of G protein-coupled receptor oligomerization. *Biochem Cell Biol,* 76, 1-11.

HENIKOFF, J. G., HENIKOFF, S. & PIETROKOVSKI, S. (1999a) New features of the Blocks Database servers. *Nucleic Acids Res,* 27, 226-8.

HENIKOFF, S., HENIKOFF, J. G. & PIETROKOVSKI, S. (1999b) Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations. *Bioinformatics,* 15, 471-9.

HORN, F., WEARE, J., BEUKERS, M. W., HORSCH, S., BAIROCH, A., CHEN, W., EDVARDSEN, O., CAMPAGNE, F. & VRIEND, G. (1998) GPCRDB: an information system for G protein-coupled receptors. *Nucleic Acids Res,* 26, 275-9.

HUGHES, A. L. & YEAGER, M. (1999) Coevolution of the mammalian chemokines and their receptors. *Immunogenetics,* 49, 115-24.

HULO, N., BAIROCH, A., BULLIARD, V., CERUTTI, L., DE CASTRO, E., LANGENDIJK-GENEVAUX, P. S., PAGNI, M. & SIGRIST, C. J. (2006) The PROSITE database. *Nucleic Acids Res,* 34, D227-30.

HUYNEN, M., SNEL, B., LATHE, W., 3RD & BORK, P. (2000) Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Res,* 10, 1204-10.

HUYNEN, M. A. & BORK, P. (1998) Measuring genome evolution. *Proc Natl Acad Sci U S A,* 95, 5849-56.

ILIOPOULOS, I., TSOKA, S., ANDRADE, M. A., JANSSEN, P., AUDIT, B., TRAMONTANO, A., VALENCIA, A., LEROY, C., SANDER, C. & OUZOUNIS, C. A. (2001) Genome sequences and great expectations. *Genome Biol,* 2, INTERACTIONS0001.

INOUE, Y., IKEDA, M. & SHIMIZU, T. (2004) Proteome-wide classification and identification of mammalian-type GPCRs by binary topology pattern. *Comput Biol Chem,* 28, 39-49.

INOUE, Y., SUGIYAMA, Y., IKEDA, M. & SHIMIZU, T. (2001) Classification of Eukaryotic 7-tms transmembrane proteins by binary topology pattern. *Genome Inform,* 336-337.

JOHNSON, J. M. & CHURCH, G. M. (2000) Predicting ligand-binding function in families of bacterial receptors. *Proc Natl Acad Sci U S A,* 97, 3965-70.

KARCHIN, R., KARPLUS, K. & HAUSSLER, D. (2002) Classifying G-protein coupled receptors with support vector machines. *Bioinformatics,* 18, 147-59.

KING, R. D., KARWATH, A., CLARE, A. & DEHASPE, L. (2000a) Accurate prediction of protein functional class from sequence in the Mycobacterium tuberculosis and Escherichia coli genomes using data mining. *Yeast,* 17, 283-93.

KING, R. D., KARWATH, A., CLARE, A. & DEHASPE, L. (2000b) Genome scale prediction of protein functional class from sequence using data mining. *The Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* New York, The Association for Computing Machinery.

KING, R. D., KARWATH, A., CLARE, A. & DEHASPE, L. (2001) The utility of different representations of protein sequence for predicting functional class. *Bioinformatics,* 17, 445-54.

KOONIN, E. V., MUSHEGIAN, A. R. & BORK, P. (1996) Non-orthologous gene displacement. *Trends Genet,* 12, 334-6.

KROGH, A., BROWN, M., MIAN, I. S., SJOLANDER, K. & HAUSSLER, D. (1994) Hidden Markov models in computational biology. Applications to protein modeling. *J Mol Biol,* 235, 1501-31.

LANDAU, M., MAYROSE, I., ROSENBERG, Y., GLASER, F., MARTZ, E., PUPKO, T. & BEN-TAL, N. (2005) ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. *Nucleic Acids Res,* 33, W299-302.

LANDGRAF, R., FISCHER, D. & EISENBERG, D. (1999) Analysis of heregulin symmetry by weighted evolutionary tracing. *Protein Eng,* 12, 943-51.

LANDGRAF, R., XENARIOS, I. & EISENBERG, D. (2001) Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. *J Mol Biol,* 307, 1487-502.

LASSO, G., ANTONIW, J. F. & MULLINS, J. G. (2006) A combinatorial pattern discovery approach for the prediction of membrane dipping (re-entrant) loops. *Bioinformatics,* 22, e290-7.

LI, W., JAROSZEWSKI, L. & GODZIK, A. (2001) Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics,* 17, 282-3.

LICHTARGE, O., BOURNE, H. R. & COHEN, F. E. (1996) An evolutionary trace method defines binding surfaces common to protein families. *J Mol Biol,* 257, 342-58.

LIN, H. H., HAN, L. Y., ZHANG, H. L., ZHENG, C. J., XIE, B. & CHEN, Y. Z. (2006) Prediction of the functional class of lipid binding proteins from sequence-derived properties irrespective of sequence similarity. *J Lipid Res,* 47, 824-31.

LIVINGSTONE, C. D. & BARTON, G. J. (1993) Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. *Comput Appl Biosci,* 9, 745-56.

LUTHY, R., XENARIOS, I. & BUCHER, P. (1994) Improving the sensitivity of the sequence profile method. *Protein Sci,* 3, 139-46.

MARCOTTE, E. M., PELLEGRINI, M., NG, H. L., RICE, D. W., YEATES, T. O. & EISENBERG, D. (1999) Detecting protein function and protein-protein interactions from genome sequences. *Science,* 285, 751-3.

MIRNY, L. A. & GELFAND, M. S. (2002) Using orthologous and paralogous proteins to identify specificity-determining residues in bacterial transcription factors. *J Mol Biol,* 321, 7-20.

OLIVEIRA, L., PAIVA, P. B., PAIVA, A. C. & VRIEND, G. (2003) Identification of functionally conserved residues with the use of entropy-variability plots. *Proteins,* 52, 544-52.

ORENGO, C., JONES, D. & THORNTON, J. M. (2003) *Bioinformatics*
Genes, Proteins & Computers, BIOS Scientific Publishers.

OVERBEEK, R., FONSTEIN, M., D'SOUZA, M., PUSCH, G. D. & MALTSEV, N. (1999) The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci U S A,* 96, 2896-901.

PAGE, R. D. (1998) GeneTree: comparing gene and species phylogenies using reconciled trees. *Bioinformatics,* 14, 819-20.

PARK, J., KARPLUS, K., BARRETT, C., HUGHEY, R., HAUSSLER, D., HUBBARD, T. & CHOTHIA, C. (1998) Sequence comparisons using multiple sequences detect

three times as many remote homologues as pairwise methods. *J Mol Biol,* 284, 1201-10.

PARK, J., TEICHMANN, S. A., HUBBARD, T. & CHOTHIA, C. (1997) Intermediate sequences increase the detection of homology between sequences. *J Mol Biol,* 273, 349-54.

PAZOS, F. & VALENCIA, A. (2001) Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Eng,* 14, 609-14.

PEARSON, W. R. & LIPMAN, D. J. (1988) Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A,* 85, 2444-8.

PELLEGRINI, M., MARCOTTE, E. M., THOMPSON, M. J., EISENBERG, D. & YEATES, T. O. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci U S A,* 96, 4285-8.

POLULYAKH, N., SAITO, K. & SHIMIZU, T. (2000) Transmembrane Topology Pattern and Detection of Transmembrane Protein Functions. *Genome Inform,* 422-423.

PUPKO, T., BELL, R. E., MAYROSE, I., GLASER, F. & BEN-TAL, N. (2002) Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics,* 18 Suppl 1, S71-7.

REMM, M., STORM, C. E. & SONNHAMMER, E. L. (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol,* 314, 1041-52.

RIGOUTSOS, I. & FLORATOS, A. (1998) Combinatorial pattern discovery in biological sequences: The TEIRESIAS algorithm. *Bioinformatics,* 14, 55-67.

ROST, B. (2002) Enzyme function less conserved than anticipated. *J Mol Biol,* 318, 595-608.

ROST, B., LIU, J., NAIR, R., WRZESZCZYNSKI, K. O. & OFRAN, Y. (2003) Automatic prediction of protein function. *Cell Mol Life Sci,* 60, 2637-50.

SAIER, M. H., JR., ENG, B. H., FARD, S., GARG, J., HAGGERTY, D. A., HUTCHINSON, W. J., JACK, D. L., LAI, E. C., LIU, H. J., NUSINEW, D. P., OMAR, A. M., PAO, S. S., PAULSEN, I. T., QUAN, J. A., SLIWINSKI, M., TSENG, T. T., WACHI, S. & YOUNG, G. B. (1999) Phylogenetic characterization of novel transport protein families revealed by genome analyses. *Biochim Biophys Acta,* 1422, 1-56.

SCHENA, M., SHALON, D., DAVIS, R. W. & BROWN, P. O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science,* 270, 467-70.

SGOURAKIS, N. G., BAGOS, P. G., PAPASAIKAS, P. K. & HAMODRAKAS, S. J. (2005) A method for the prediction of GPCRs coupling specificity to G-proteins using refined profile Hidden Markov Models. *BMC Bioinformatics,* 6, 104.

SMITH, T. F. & WATERMAN, M. S. (1981) Identification of common molecular subsequences. *J Mol Biol,* 147, 195-7.

SREEKUMAR, K. R., HUANG, Y., PAUSCH, M. H. & GULUKOTA, K. (2004) Predicting GPCR-G-protein coupling using hidden Markov models. *Bioinformatics,* 20, 3490-9.

STORM, C. E. & SONNHAMMER, E. L. (2002) Automated ortholog inference from phylogenetic trees and calculation of orthology reliability. *Bioinformatics,* 18, 92-9.

SUGIYAMA, Y., ARAI, M. & SHIMIZU, T. (2001) Comprehensive functional identification of prokaryotic transmembrane proteins by binary topology pattern. *Genome Inform*, 334-335.

SUGIYAMA, Y., POLULYAKH, N. & SHIMIZU, T. (2003) Identification of transmembrane protein functions by binary topology patterns. *Protein Eng,* 16, 479-88.

TATUSOV, R. L., ALTSCHUL, S. F. & KOONIN, E. V. (1994) Detection of conserved segments in proteins: iterative scanning of sequence databases with alignment blocks. *Proc Natl Acad Sci U S A,* 91, 12091-5.

TATUSOV, R. L., FEDOROVA, N. D., JACKSON, J. D., JACOBS, A. R., KIRYUTIN, B., KOONIN, E. V., KRYLOV, D. M., MAZUMDER, R., MEKHEDOV, S. L., NIKOLSKAYA, A. N., RAO, B. S., SMIRNOV, S., SVERDLOV, A. V., VASUDEVAN, S., WOLF, Y. I., YIN, J. J. & NATALE, D. A. (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics,* 4, 41.

TATUSOV, R. L., KOONIN, E. V. & LIPMAN, D. J. (1997) A genomic perspective on protein families. *Science,* 278, 631-7.

TAYLOR, W. R. (1986) Identification of protein sequence homology by consensus template alignment. *J Mol Biol,* 188, 233-58.

WITTEN, I. H. & FRANK, E. (2005) *Data Mining: Practical machine learning tools and techniques,* San Francisco, Morgan Kaufmann.

WU, T. D. & BRUTLAG, D. L. (1995) Identification of protein motifs using conserved amino acid properties and partitioning techniques. *Proc Int Conf Intell Syst Mol Biol,* 3, 402-10.

YI, T. M. & LANDER, E. S. (1994) Recognition of related proteins by iterative template refinement (ITR). *Protein Sci,* 3, 1315-28.

CHAPTER 9

Discussion

9.1 The genome explosion and the functional gap in the membrane proteome

In recent years, we have found that the gap between the number of genes yet to be characterized and those genes already characterized is constantly increasing. The reason for this situation is simply that experimentally based methods cannot yet cope with the rate at which different genomes are being sequenced. This functional gap is even more severe in the membrane proteome, although different analyses have shown that: i) membrane proteins are extremely abundant (Arkin et al., 1997, Wallin and von Heijne, 1998), ii) many crucial cellular and physiological processes involve membrane proteins and iii) membrane proteins are of great pharmacological importance (Wu and Yates, 2003). Such a challenging situation has been reached mainly due to the problems involved when extracting the protein from its lipidic environment. An alternative approach is to computationally analyze uncharacterized genes in order to predict structural and functional properties. Such predictions should eventually be tested in the laboratory in order to ultimately characterize unknown genes. Diverse computational methods have been implemented to characterize newly sequenced genomes: from gene and splice variant predictions to structural and functional prediction (including subcellular location, molecular function, post-translational modifications and protein-protein interactions). Due to the vast quantity of available sequential and structural information concerning soluble proteins, computational methods tend to be biased towards the characterization of soluble proteins. Therefore, the rate at which the functional gap observed in the membrane proteome increases is not being reduced as quickly. Although the computational methods developed to characterize biological features of soluble proteins (e.g. enzymes) could be applied to membrane proteins, it is necessary to specifically train such methods with membrane proteins. Membrane proteins are so structurally and functionally distinct from their soluble

counterparts that they should be regarded as two separate proteins classes and should be independently analyzed.

Membrane proteins are by nature very different to globular membrane proteins. The lipidic environment where membrane proteins are embedded has imposed a compositional constraint upon the transmembrane regions of the protein, such as the requirement for a high proportion of hydrophobic residues. Following this principle, functional prediction methods based on features that can be extracted from sequence, such as the amino acid composition and different physicochemical properties, might reflect instead the differences between globular and membrane proteins. Sugiyama and colleagues (Sugiyama et al., 2003) stated that sequence similarity based methods are less accurate in predicting the molecular function of membrane proteins probably due to the hydrophobic nature of the transmembrane regions. This compositional bias of membrane proteins might disproportionately affect particular methods for predicting molecular function. While methods based on features extracted from sequence can be severely affected, pattern discovery methods can also be differentially affected depending on the location of the functional domain. If the functional domain is located in an extramembranous loop, it is likely that it resembles a homologue domain located in a globular protein. However, if the functional domain is located in a transmembrane region, it is likely that this domain has evolved in order to adapt itself to the new environment and to interact with the hydrophobic secondary structures of the protein.

9.2 Previous research that provided the premise for this work

Considering the current functional gap in the membrane proteome, the proposed research was drafted to provide the scientific community with novel methods for the characterization of unknown membrane proteins. Membrane proteins can be broadly classified into two structural classes: α-helical membrane proteins and β-barrel membrane proteins. Functionally, α-helical membrane proteins are more diverse than β-barrel membrane proteins. The latter structural class has mainly been found to be involved in transport processes of substances through the membrane. On the other hand, α-helical

membrane proteins perform a wide range of molecular activities such as enzymatic activity, photosensitivity, transport of ions and molecules through the membrane, signal transduction, protein networking, receptors and cell adhesion.

Previous research that considered interhelical associations of crystallized membrane proteins using TMDistance (Togawa, PhD Thesis, 2006), showed that functional clusters of α-helical membrane proteins contained specific patterns of inter-helical associations located at a similar depth (Lasso, Honours Thesis, 2001). This was unsurprising as binding and active sites located in the membrane must involve residues from different helices that are located at a similar depth in the membrane. However, two interesting aspects were extracted from these results. The first aspect was the substantial potential (though not exploited before by computational biologists) of the amino acid depth value. The second aspect was the important role of the transmembrane region in shaping the molecular function of polytopic α-helical membrane proteins. Following these emerging observations, an ambitious idea was proposed: the ability to infer two dimensional features solely based on the amino acid sequence and to qualify the role of the transmembrane region in determining molecular function. This task would clearly involve extensive structural prediction and computationally expensive algorithms if soluble proteins were to be considered. However, the spatial constraints imposed in the transmembrane region of membrane proteins can be used to infer two-dimensional properties by looking at pairs of residues located at a similar depth. Current topology prediction methods have an accuracy of 70-80% in correctly predicting the topology of α-helical membrane proteins. Although current topology models can still be improved, there is no reason why these models should not be used in the interim to extract and exploit the information already contained. Therefore, as more accurate topological models are being developed, the computational infrastructure for feature extraction will have already been developed.

9.3 Exploring the membrane proteome space, going up and down transmembrane regions of α-helical membrane proteins

The different computational approaches described in this thesis are combined into a single research project **(figure 9.1)**, the aim of which is to provide computational tools to characterize unknown α-helical membrane proteins and investigate the role played by the transmembrane regions in determining molecular function. This thesis can be reduced to the same basic protocol that is followed by any data mining method where the steps to follow are: 1. Data set assembly (PROCLASS), 2. Feature extraction (TMLOOP+TMDEPTH) and 3. Classification (TMLOCATE & TMFUN).

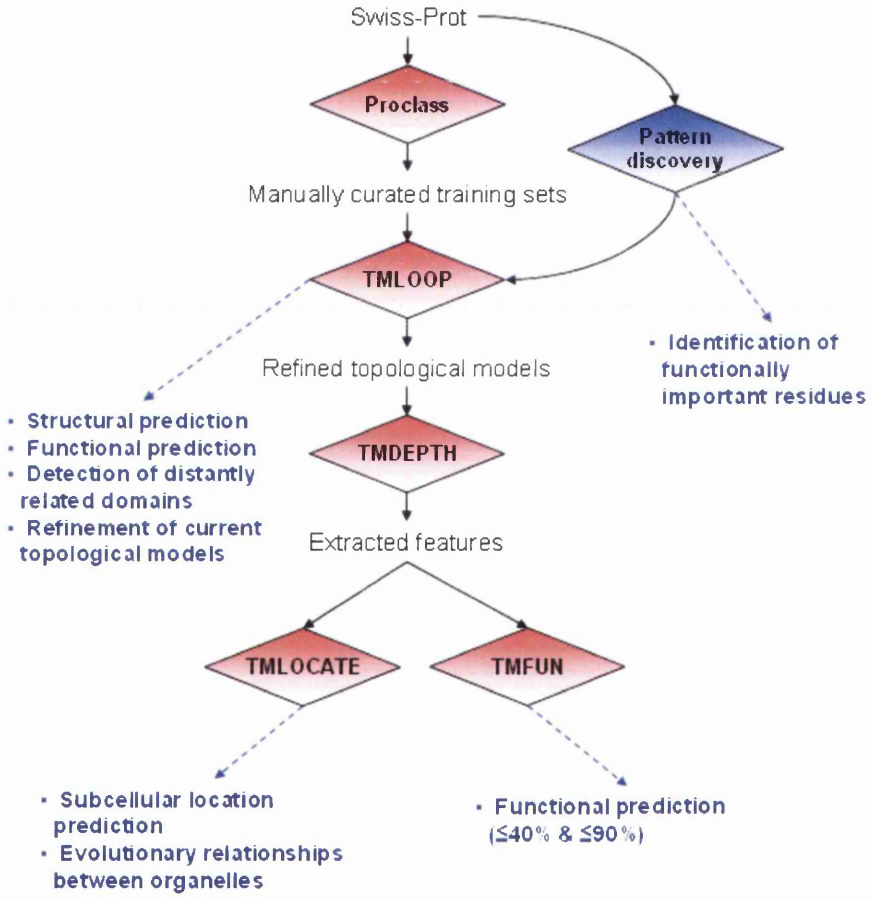

Figure 9.1. Summary of the research carried out. Red rhombus correspond to novel computational tools developed in this study, the blue rhombus corresponds to a computational method, not developed by our group, but used in this research. Statements highlighted in blue correspond to the different functional properties detected or predicted by the different approaches. As this figure illustrates, all implemented methods are interrelated.

9.3.1 Data set assembly

The first obstacle to overcome was the data set assembly from the Swiss-Prot database (Bairoch and Apweiler, 1998, Boeckmann et al., 2003). This is probably one of the most tedious tasks, and yet crucial. In order to assemble large data sets incorporating thousands of proteins, manual assembly is an unthinkable approach. This is not just because it is too time consuming but also because it is error prone. Up until now, the only alternative was to use automatic methods to assemble large data sets based on the functional annotation contained in the Swiss-Prot database. Automatic methods are, however, also error prone and would require an expert to manually analyze each cluster and re-classify proteins that have been erroneously clustered. In between these two extremes, an intermediate approach might be desirable. This approach should facilitate the manual assembly of data sets by means of a computational method that reduces the number of different data points (proteins) down to manually manageable numbers. The principle underlying this idea is straightforward: functional classes can be defined with a unique set of terms, which include ligand binding and molecular activity, therefore all proteins containing this unique set of terms should automatically cluster together (and still be certain that the clustered proteins have the same molecular function). For instance, in order to retrieve sodium channels and glucose transporters the words to be used should be: "sodium", "channel", "transport", "glucose" and "transporter". If the functional annotation space of each protein is then converted into a binary vector composed of five attributes (each binary attribute corresponds to a particular functional term, where "1" indicates that the corresponding term is present and "0" indicates that it is not) it is possible to cluster all proteins contained in the Swiss-Prot data base that share identical binary sequence (exact pattern matching). The cluster defined by the terms "sodium" and "channel" or "sodium", "channel" and "transport" will contain all sodium channels. However, they will also contain other channels such as the sodium / proton antiporter or the chloride / sodium co-transporter because the "proton" and the "chloride" terms were not specified before clustering. Glucose transporters would be clustered in a similar fashion whereas the clusters defined by the terms "glucose", "sodium" and "transport" or "transporter" will probably only contain sodium dependent glucose transporters. If non-sodium channels are required, these proteins

will be grouped together in the cluster defined solely by the "channel" term. Therefore, the method has reduced all non-sodium channels into a single cluster. As this example showed, the key to successfully applying this method to the functional classification of large data sets is to select the appropriate functional terms. PROCLASS was implemented based on the principle described above. In order to ensure that the appropriate terms have been selected prior to clustering the Swiss-Prot database, a training stage was required, where the interaction between PROCLASS and the user leads to the selection of the appropriate terms and descriptions of synonyms and equivalent terms (e.g. Na^+ and sodium; Proton and Proton-associated). Evaluation of PROCLASS showed that 98% of the clusters only contained proteins belonging to the same functional class. Therefore, the protein space containing thousands of proteins has been reduced to hundreds of clusters. Manual evaluation of these clusters should be applied then to merge those clusters corresponding to the same molecular function but different functional annotation space. PROCLASS was applied to develop two different data sets. The first data set was based on the different subcellular location of polytopic membrane proteins and the second data set was based on the different functions carried out by polytopic membrane proteins. Exploration of the obtained clusters showed that specific annotations contained in the Swiss-Prot database needed further refinement (e.g. plasma membrane proteins). To our knowledge, this is the first attempt to classify all functions carried out in the membrane. Furthermore, the different functions carried out in the membrane and contained in the Swiss-Prot database have also been quantified. According to the information contained in the Swiss-Prot database, the enzyme class is the most common functional class found in the membrane (43%), followed by G protein-coupled receptors (20%), molecular transporters (15%), and ion channels (14%). Subsequently, functional subclasses have also being quantified showing interesting results. Lyases (EC4) and ligases (EC6) only accounted for 2% of all membrane proteins with enzymatic activity whereas no isomerases with two or more transmembrane proteins were observed. Sugar transporters were the most common type of molecular transporter (20%), cation channels were found to make up to 69% of all ion channels found in the protein database and the class A of GPCR was found account for 84% of all GPCR. While these results might be used to infer the importance of particular functional classes, these results should not be used to quantify the importance of the

considered functional classes. Analysis with PROCLASS has not been designed to distinguish between subunits, therefore functional groups in the membrane might be overrepresented when a significant number of proteins contained in the cluster are subunits of protein complexes. The Swiss-Prot database might over-represent functional classes where a high percentage of the corresponding protein members do share a marked sequence homology or a signature that make them recognizable and easier to be experimentally tested. Likewise, the Swiss-Prot database might well be biased towards those proteins and organisms of higher research interest (either academic or industrial), although this influence should be mollified in future by the addition of whole genomes. Additionally, trying to generalize the quantification of the functions carried out in the membrane is not realistic. The reason for this is that particular functional classes might play essential roles in particular species whereas in other species the role of such functional classes might not be that important. However, the quantification carried out is still valuable to identify those functional classes that might be under-represented in the database and analyze the current distribution of functional classes contained in the Swiss-Prot database.

PROCLASS is the outcome of a simple and original idea and it has proven to be a very useful tool for the manual curation of large sets of proteins. Future data mining approaches should definitely consider the use of this tool while assembling their datasets.

9.3.2 Feature extraction

9.3.2.1 Prediction of membrane dipping loops and refinement of current topological models

The next step towards functional prediction of polytopic membrane proteins was the feature extraction process. The feature extraction method applied combines sequence and topological models **(Chapter 6)**. However, one of the major shortcomings of current topology prediction methods is the inability to predict membrane dipping loops, also known as re-entrant loops. These structural domains have been shown to play essential functional

roles as selectivity filters and molecular gates. However, due to their residue composition, which differs from that of membrane spanning regions, current topology prediction algorithms often fail to correctly predict these structural domains. It was believed that in order to extract features that could accurately represent the combination of sequence and topology, membrane dipping loops needed to be included and accounted for in such topological models. Therefore, the next obstacle to overcome was to develop a computational tool to predict membrane dipping loops.

The approach was based on the detection of conserved patterns pertaining to the membrane dipping loop in protein families where at least one protein has been crystallized and found to contain a membrane dipping loop **(Chapter 4)**. The main advantage of using a pattern discovery approach is that the information contained in such patterns can be used to identify functionally important residues. Accordingly, the discovered patterns contained some residues that were already identified by experiment as functionally important residues, thus validating the approach. However, additional patterns and residues were identified that have not yet been identified by other methods. Interestingly, different patterns were found in different types of membrane dipping loops (even if they shared a similar arrangement of secondary structures). This observation suggests that the discovered patterns contain residues placed specifically for the purpose of functional mechanisms rather than folding mechanisms. Therefore, this information can not only be used to imply a particular fold (as these patterns are only found in particular membrane dipping loop structures) but also to infer functional similarities (e.g. selectivity filters of channels).
Based on these promising results, a computational tool, named TMLOOP **(Chapter 5)** was implemented. The discovered patterns were used as weighted predictive rules to be matched against queried sequences. Rather than using the pattern with the highest support, a set of partially overlapping patterns with support $\geq 70\%$ was considered in order to detect distantly related membrane dipping loops. During evolution, the constraints imposed by the outside world have been reflected in changes at the molecular level to ensure the adaptability of an organism to a changing environment. Evolution of membrane dipping loops might reflect evolutionary pressures for changes in the gating process of a membrane protein, the re-adjustment of specificity for the corresponding ligand according to new

needs of the cell, or even binding of a different ligand in a similar fashion. All these pressures are ultimately translated into small variations of the residues associated with these motifs and rearrangements with the interacting helices. Divergent evolution would generate two membrane dipping loops from a common ancestral structural motif with small variations in sequence that allow the binding of the same ligand with different specificity or binding different ligands in a similar fashion. On the other hand, phylogenetically unrelated proteins might generate a similar three-dimensional membrane dipping loop to bind the same or a similar ligand, which would involve two different membrane dipping loops containing a similar (but not identical) selective filter or binding site. Evaluation of the approach showed that the consensus motif method maximized the prediction of these particular domains. The predictive tool was applied to the entire membrane proteome contained in the Swiss-Prot database and potential membrane dipping loops not yet discovered by experimental methods were discovered. TMLOOP writer **(Chapter 5)** was implemented to include the predicted membrane dipping loops considered as true positives in the transmembrane statement of the corresponding Swiss-Prot files. Therefore, current topology models of proteins predicted to have membrane dipping loops were refined.

A month before our TMLOOP paper was published (Lasso et al., 2006), Viklund and colleagues published similar work (Viklund et al., 2006). The authors analyzed a set of 79 chains obtained from the Protein Data Bank (Berman et al., 2000) and identified 36 membrane dipping loops. As with our approach, detected membrane dipping loops were structurally classified into three different categories: helix-coil-helix motif (corresponding to the helix-in-turn-helix-out domain), helix-coil or coil-helix motif (corresponding to helix-in-turn-loop-out and loop-in-turn-helix-out) and the irregular secondary structure motif. The last structural class differs from our classification scheme. The reason for this is that we believe these domains might well be artefacts caused by crystallization conditions or actual true positives whose secondary structures have been disrupted by crystallization conditions. The authors performed a Principal Component Analysis (PCA) to discriminate between the different structural regions contained in polytopic membrane proteins, namely the transmembrane region, membrane dipping loops, water-interface regions and extramembranous loops. The PCA was based on the amino acid composition of these

regions. The obtained results showed that it was possible to completely describe the differences between these four regions based on the hydrophobicity, small residue content and polar-aromatic residue content (Tyr and Trp). Comparison of these results with the results obtained from TMLOOP revealed an interesting fact: whereas different structural classes of membrane dipping loops do not contain a common pattern, it seems that they have similar overall amino acid composition (the high proportion of small residues, specially Gly and Ala, seem to be a specific property of membrane dipping loops). Furthermore, the authors also implemented a predictive tool based on a Hidden Markov Model (HMM). Evaluation of TOP-MOD showed a sensitivity of 47%-69% and a specificity of 72%. On the other hand, TMLOOP was found to accurately predict membrane dipping loops with a sensitivity of 92% and a specificity of 100%. Although TMLOOP provided a better prediction, it is important to remark that TOP-MOD was trained with only 36 sequences whereas the pattern discovery approach prior to TMLOOP used 580 partial sequences (corresponding to membrane dipping loops). TOP-MOD was applied to predict membrane dipping loops in *E. coli*, *S. cerevisiae* and *H. sapiens*. Interestingly, the obtained results showed that at least 10% of the polytopic membrane proteins contain membrane dipping loops. These results are clearly very different to the results obtained by TMLOOP, where only 2.1% of the polytopic membrane proteins contained in the Swiss-Prot database have membrane dipping loops. There are different possible reasons for this discrepancy. The first possibility is based upon the differential protein content in the Swiss-Prot database and in individual genomes. Particular protein families without membrane dipping loops might be over-represented in the Swiss-Prot database, which underestimates the actual proportion of membrane dipping loops. Another possible explanation could be that the HMM-based predictor has actually found a common structural feature among different membrane dipping loops, whereas TMLOOP has been focussed more on the functional aspects rather than on the structural aspects. Another possibility is that TOP-MOD overestimates the presence of membrane dipping loops.

An interesting statement made by Viklund and colleagues was that the occurrence of membrane dipping loops increases linearly with the number of transmembrane regions, though this relationship may not be so clear cut. Due to the apparent physicochemical

properties of these domains, membrane dipping loops need to interact with neighbouring transmembrane helices in order to minimize the energy penalty imposed when locating polar and charged residues in the membrane. This might indicate that there could be a minimum number of transmembrane regions required to stabilize such domain. Once that minimum requirement has been achieved, the number of transmembrane regions should not be used to imply a higher number of membrane dipping loops contained in the structure. Considering the identified crystallized protein complexes containing membrane dipping loops and the annotated transmembrane regions contained in the PDB_TM database (Tusnady et al., 2004), it was found that the potassium channel complex has six TMs and four membrane dipping loops, the aquaporin structure has six TMs and two membrane dipping loops, the vitamin B12 complex has 20 TMs and two membrane dipping loops, the ClC chloride channel complex has 20 TMs and eight membrane dipping loops, the photosystem I complex is composed of 30 TMs and one membrane dipping loop and the glutamate transporter homologue has 24 TMs and six membrane dipping loops. Considering this data, the relationship described by Viklund and colleagues does not seem to fit the real situation. Interestingly, the minimum number of transmembrane regions for a protein complex containing at least one membrane dipping loop was found to be six. These results seem to be more in accordance with the idea of a requirement for a minimum requirement of number of transmembrane regions in order to accommodate any number of membrane dipping loops.

9.3.2.2 Feature extraction by combining sequence and topology

As described earlier, using a combination of sequence and topology has the objective of being able to associate an approximate depth value for each residue located in the membrane. Subsequently, each possible pair of residues (210 combinations considering the 20 common amino acids) located at a similar depth in the membrane is quantified and stored in a matrix. These matrices are then normalized in order to combine matrices corresponding to proteins belonging to the same cluster (identifying common patterns of inter-helical associations) and to compare matrices corresponding to proteins belonging to different clusters (identifying specific patterns of inter-helical associations). Additionally, some extra features were computed: i) the percentage of participation within the normalized

matrix was obtained for each residue and ii) the normalized associations were clustered according to the physicochemical properties of the 20 residues listed (non-polar, polar and charged). By combining sequence and topology, each protein is converted then into a protein vector containing 236 attributes, which correspond to two dimensional features.

TMDEPTH **(Chapter 6)** is not computationally expensive. The algorithm can process large sets, including hundreds of proteins, within a few minutes. The only limitation of the developed feature extraction method is its direct dependency upon topological models. Considering the implementation of the algorithm, incorrect predictions or over-predictions of a single transmembrane region can completely change the computed protein vector. Although the topological models used have been refined by the inclusion of membrane dipping loops, further improvement is needed in order predict the topology of polytopic membrane proteins, towards an accuracy $\geq 90\%$.

9.3.3 Classification

To some extent, functional prediction of membrane proteins was an integral part of the characterization of membrane dipping loops, as the predictions included reference to ligand specificity. The discovered patterns can be used to identify functionally important residues, where predicted membrane dipping loops are directly related to the ligand-related function of the protein. Therefore, TMLOOP can not only predict structural domains but also infer functional similarities. Different features can be predicted to functionally characterize an unknown gene. The most predominant feature is the prediction of its molecular function. However, other functional properties can be predicted such as the subcellular location, post-translational modifications and protein-protein interactions. In the developed research, the extracted features obtained from TMDEPTH were mined for the purpose of the elucidation of subcellular location and molecular function.

9.3.3.1 TMLOCATE

Different data mining techniques, such as Bayesian methods and support vector machines, were combined into a multilayer predictive algorithm. The final architecture of TMLOCATE **(Chapter 7)** showed that by mimicking the cellular sorting process the predictive accuracy of the method was maximized. These results were in accordance with previously reported algorithms based on data sets of soluble proteins (Nair and Rost, 2005). Rather than just focussing on the positive results of the method, our attention was also focussed on the incorrect predictions performed by TMLOCATE. Evolutionary relationships between the different organelles were then detectable. All organelles involved within the secretory pathway appear to be evolutionarily related. Additionally, the nuclear membrane proteins tended to be predicted as part of the secretory pathway and more specifically, as proteins belonging to the endoplasmic reticulum. These results were in accordance with the arrangement of organelles within the cell as it is well known that the outer nuclear membrane is contiguous with the endoplasmic reticulum (Alberts et al., 1994). Further sub-classification within the secretory class showed that some of the trained classifiers following the one-against-all approach achieved significant values of sensitivity and specificity. However, evaluation of the overall architecture showed significant misclassifications between subcellular compartments involved in the secretory pathway. This might indicate that other features (e.g. length of transmembrane regions or extramembraneous features) not considered here, might be cooperatively involved in protein sorting within the secretory pathway. These results also showed an evolutionary relationship between the peroxisome and mitochondrion, and such a relationship is also in agreement with that published by Gabaldon and colleagues (Gabaldon et al., 2006). The method showed a normalized accuracy of 75% in discriminating between proteins belonging to the chloroplast, mitochondria, plasma membrane and secretory organelles, thus reflecting the importance of the transmembrane region in assigning the organellar location of polytopic membrane proteins. This is the first rigorous attempt to specifically predict the subcellular location of polytopic membrane proteins. Previous work carried out by Chou and colleagues (Chou and Elrod, 1999) was based on a training set where plasma membrane proteins accounted for 80% of the proteins. Furthermore, no sequence similarity filter was applied in order to remove highly identical proteins. Therefore, the two

approaches can not be compared, as the work carried out by Chou and colleagues is likely to be biased towards the prediction of plasma membrane proteins whereas our approach represents a more balanced method.

It is believed that proteins localize to their appropriate organelle using a variety of mechanisms and this is probably the reason why methods based solely on sorting signals or amino acid compositions seem to have reached a plateau in their prediction accuracy. Further improvement of this method should consider other features located outside the membrane (e.g N-terminus, C-terminus, extramembranous loops and signal peptides) that can be used to guide cellular sorting.

These results have shown that signatures derived from the predicted topology of membrane proteins can be associated with specific subcellular locations of membrane proteins. However, it is not clear whether such patterns correspond to retention or targeting signals. Trafficking of proteins to a particular organelle most likely involves a combination of targeting and retention processes in order to appropriately direct and stabilize the given protein. Targeting signals often imply a protein carrier and/or a receptor belonging to the corresponding organelle that recognizes such a signal. Extensive work has been carried out to detect targeting signals in amino acid sequences (Bannai et al., 2002, Bickmore and Sutherland, 2002, Boden and Hawkins, 2005, Claros and Vincens, 1996, Cokol et al., 2000, Emanuelsson et al., 2000, Emanuelsson et al., 1999, Fujiwara et al., 1997, Hawkins and Boden, 2006, Nair et al., 2003, Petsalaki et al., 2006). Previous work has also shown that residues located in particular transmembrane regions are important targeting signals for particular organelles (Biermanns et al., 2003, Honsho et al., 2002, Jones et al., 2004). Such targeting signals are not just composed of transmembrane regions but also portions of the extramembraneous domain. If such signals are to be recognized by other proteins (e.g. carriers and receptors), it is essential that these signals are accessible. Therefore, placing targeting signals in transmembrane regions might pose a difficulty with respect to recognition by proteins carriers and receptors. On the other hand, it is more likely that these signatures contained in the transmembrane domain are involved in the retention processes rather than targeting. Previous work has concluded that the transmembrane regions of

particular membrane proteins are essential retention domains (Aoki et al., 1992, Cocquerel et al., 1999, Colley, 1997, Hobman et al., 1997, Hobman et al., 1995, Ma et al., 2004, Op De Beeck et al., 2004). Retention signals located in the transmembrane domain of polytopic membrane proteins might be explained by the particular lipid composition of lipid bilayers belonging to specific organelles. These compositional differences between organelle membranes might result in different physicochemical properties between these membranes. Particular membrane proteins being transported through the secretory pathway might therefore be retained in specific organelles where the physicochemical properties of the given membrane trigger conformational changes in the transmembrane domain. These conformational changes might involve positional rearrangement of transmembrane regions and new interhelical interactions, which ultimately minimise the potential energy of the molecule. This theory is supported by several experimental analyses that showed that the presence of hydrophilic residues in the middle of transmembrane regions might be important for ER retention (Bonifacino et al., 1991, Cocquerel et al., 2000, Letourneur and Cosson, 1998, Yang et al., 1997). Furthermore, the length of the transmembrane regions has also been pinpointed as a form of retention signal. The plasma membrane has a higher content of cholesterol than the membrane found in the Golgi apparatus and consequently the plasma membrane is thicker than the membrane of the Golgi apparatus. Therefore, membrane proteins with longer transmembrane regions would not be retained in the Golgi apparatus whereas membrane proteins with shorter helices would be retained, as the protein would have reached its minimum potential energy (Munro, 1995). A similar retention property has been found in the ER (Pedrazzini et al., 1996, Yang et al., 1997, Szczesna-Skorupa and Kemper, 2000).

According to this theory, the majority of the transmembrane residues would then probably play important roles for the stabilization of a given membrane protein within a particular lipid bilayer in order to minimize the molecular potential energy.

9.3.3.2 TMFUN

As with prediction of subcellular location, different data mining techniques were combined into a computational tool to predict the molecular function of polytopic membrane proteins. Unlike the data set for subcellular location, the data set assembled using PROCLASS was large enough to apply more stringent filters at the sequence level. Therefore, two versions were implemented, a version trained under low sequence similarity (<40%) and a different version using a filtered set where highly identical proteins were removed (≥90%). At the sequence similarity threshold of 40%, TMFUN predicted enzymes, GPCRs, and transporters with a sensitivity of 64.1%, 87.5% and 71.4% respectively. At the most informative sub-level, TMFUN predicted 70% of the olfactory GPCRs under low sequence similarity. Further analysis showed that proteins belonging to different functional classes but with aspects of similar function carried out in the membrane domain itself might be misclassified. These results were not unexpected as the feature extraction method uses only information contained in the transmembrane regions. Therefore, the extracted features might well explain the local functions carried out in the membrane rather than the overall function of the protein. At the 90% sequence similarity threshold, TMFUN achieved higher predictive accuracies. Enzymes, GPCRs, ion channels and molecular transporters were predicted with a sensitivity of 87.8%, 92.8%, 58.3% and 75.6% respectively. At the most informative level the different subclasses of class A GPCRs were predicted at sensitivity values of 84.5%-92.9%. Although the reported accuracies improved significantly, further analysis also showed that proteins belonging to different functional classes but with a similar function carried out in the membrane were again misclassified.

The results obtained using a sequence similarity threshold of 90% are quite promising. In the short term future, further sub-classification should be carried out in order to obtain more informative predictions. However, the aim proposed at the outset was fully achieved as a computational tool for the prediction of molecular function based solely on the information contained in the transmembrane region was successfully implemented. Although the transmembrane region might have been previously underestimated in terms of

its relevance to molecular function (with the exception, of course of transport proteins and ion channels), the observed results clearly reflect a direct association between the spatial arrangement of residues in the transmembrane regions and the capacity for polytopic membrane proteins to carry out their functions, and therefore the importance of the transmembrane region itself for modulating molecular function that may be primarily associated with extramembranous regions of the protein, such as ligand binding or signalling.

Comparison of these two versions might indicate that misclassifications reported under low sequence similarity have been overshadowed by a larger fraction of correct predictions probably corresponding to subsets of proteins sharing a significant sequence similarity. Therefore, the overall accuracy of the method increases as sequences showing a more significant sequence similarity are included in the set. However, more flexible data sets that still avoid pairs of highly identical proteins might be more appropriate as larger sets permit more informative predictions, and protein families with marked sequence similarity are not constrained by stringent sequence similarity thresholds. Following this idea, the TMFUN version trained at a sequence similarity threshold of 40% has proven that signatures derived from the predicted topology of membrane proteins can be associated with specific molecular functions of membrane proteins. However, in order to obtain a more informative prediction and improve the prediction of those protein families showing a marked sequence similarity, the TMFUN version trained at a 90% sequence similarity might be preferred for most functional annotation purposes.

Comparison of this method to those aiming to predict different subclasses of G-protein coupled receptors showed that our approach was more thorough than previous studies (Guo et al., 2005, Karchin et al., 2002, Inoue et al., 2004, Bhasin and Raghava, 2004, Bhasin and Raghava, 2005). To begin with, all the proteins contained in our data set were polytopic membrane proteins. If soluble proteins were to have been used as the negative training class, the classifier would discriminate proteins based on the broad structural characteristics (membrane protein or soluble protein) rather than on functional characteristics. The reason for this is that the transmembrane region signal is stronger than

the functional signal when soluble and membrane proteins are to be contrasted (as the number of residues involved in transmembrane regions can make up a high proportion of the protein compared to the number of residues directly involved in specific protein function), so like must be compared with like. Additionally, our approach applied filtering at the sequence level in order to remove highly identical sequences that would have biased the classifier. The GPCRpred software (Bhasin and Raghava, 2004) is perhaps the most similar approach to TMFUN. This program was implemented based on different support vector machines where the extracted features corresponded to the dipeptide composition of the polypeptide sequences (sequence similarity threshold = 90%). The program achieved 99.5% accuracy for predicting GPCR proteins (five fold cross-validation). However, the negative class of the corresponding training set was found to under-represent polytopic membrane proteins. GPCRpred was also designed to sub-classify the predicted GPCRs into sub-classes at various levels showing accurate predictive scores (two fold cross-validation). Amine GPCRs, Peptide GPCRs, Rhodopsin GPCRs and Olfactory GPCRs were predicted with an accuracy of 99.1%, 99.7%, 98.9% and 100% respectively. On the other hand, TMFUN (sequence similarity threshold = 90%) evaluation by ten fold cross-validation yielded a sensitivity of 92.9%, 84.5%, 93% and 92.3% respectively. Both methods show significant predictive accuracy and GPCRpred showed higher predictive values for each GPCR sub-family. However, the performance of TMFUN has been evaluated against sets of other polytopic membrane proteins, whereas GPCRpred used, as a non-GPCR class, a set of proteins mostly corresponding to soluble proteins. Ultimately, reliable comparison of these methods can only be achieved when training is performed with the same data set.

Interestingly, these results raise a series of questions related to the biological basis of the signatures found in the membrane. As explained above, these results reflect the importance of the transmembrane domain for indicating the molecular function of the membrane protein. Therefore, even for membrane proteins whose molecular function is believed to take part mostly outside the membrane, there must be some kind of specific functional property contained in the transmembrane domain that is related (either directly or indirectly) to the general molecular function of the protein. These functional properties might relate to a wide range of features such as binding specificities for particular

prosthetic groups found in the membrane or in the lipid-water interface, specific conformational changes triggered by a particular ligand or lipid compositional dependencies. Therefore, when no ligand binding sites are found to be associated with the transmembrane domain of the membrane proteins, other functional features can still be found that might be responsible for those signatures of pairs of residues located at a similar membrane depth. The G protein-coupled receptor superfamily is functional class that is particularly well predicted by the method. Following functional sub-classification, the approach was found to predict specific class A sub-families with a sensitivity of 85%-93%. The functional classification schemes were obtained from the GPCRDB database (Horn et al., 1998), whose classification is based on different ligand specificities. Considering the high predictive accuracies in accordance with the classification scheme, it is feasible to conclude that the extracted signatures related directly to the ligand binding specificity. However, the ligand binding sites might involve extramembraneous regions exclusively, transmembrane regions exclusively, or both regions (Schwartz, 1994). For instance, the ligand binding site in amine GPCRs is composed by different transmembrane regions whereas the ligand binding site in peptide GPCR proteins is located outside the membrane. This suggests that the extracted signatures do not necessarily need to be directly associated with the location of the primary molecular function of the protein. Activation of the G protein on the inside of the membrane must be achieved through different conformational changes depending on the location of the ligand binding site. Furthermore, these different conformational changes might also be translated into different mechanisms of activation of different G-proteins.

TMFUN is the most thorough computational approach for predicting the molecular function of polytopic membrane proteins to date. The approach can be applied not only for the automatic characterization of newly sequenced genomes but also for the functional characterization of orphan genes, thus helping to reduce the current functional gap in the membrane proteome.

9.3.4 Complimentary predictions in the TM project

The different methods developed in this thesis have been implemented to predict different functional features **(figure 9.1)** of polytopic membrane proteins. While TMLOOP accurately predicts a particular structural domain that is directly involved in the molecular function of the protein, TMLOCATE predicts the subcellular location within eukaryotic cells and TMFUN predicts the molecular function of membrane proteins. These different predictions can be combined in order to refine functional characterization by highlighting specific mechanisms of action and functionally important residues, and give insights into possible interacting functional partners, pathways and ultimately physiological processes.

The prediction obtained from TMLOCATE can be used to refine that from TMFUN by excluding those functional classes that are known not to be present in the predicted organelle. By excluding these classes, the trained classifiers might yield higher predictive accuracies as the number of possible classes to be considered decreases. Furthermore, the predicted membrane dipping loops can be mapped onto the sequence in order to characterize a structural domain, give insights about the mechanism of action of the protein and highlight functionally important residues. For example, a hypothetical unknown protein predicted by TMFUN as a molecular transporter might also be found to have a membrane dipping loop similar to that found in the sodium:dicarboxylate symporter family. The information obtained from TMLOOP can be used to preliminarily assume a gating mechanism similar to that found in the glutamate transporter homologue (Yernool et al., 2004) and identify residues of potential functional importance. Additionally, prediction of the subcellular location of the protein might indirectly be used to identify potential interacting partners. Subsequently, further analysis of the predicted protein-protein interaction might be used to identify the pathway involved. Therefore, the three computational methods may be interrelated to characterize the membrane proteins of newly sequenced genomes and thereby significantly reduce the functional gap in the membrane proteome.

The aims and objectives proposed for this research have been fully achieved. The developed methods have for the first time categorically shown that the transmembrane regions of polytopic membrane proteins hold essential information associated with a wide range of functional properties such as filtering and gating processes, subcellular location and molecular function. Future publications derived from this thesis should mark the beginning of a concerted effort in the computational biology field aimed at informed prediction of membrane protein domain structure, function and organellar location that may serve as a genuine complement to the laboratory characterization of membrane proteins.

9.4 Future work

The work carried out has shown very promising results, however further improvements can be made in order to obtain more accurate predictors. TMLOOP has proven to be a highly accurate tool for the prediction of membrane dipping loops. Although TMLOOP writer has been implemented in order to refine current topological models, an interesting option would be to encapsulate the TMLOOP loop algorithm within a reliable topology prediction method. Another interesting line of research would be to apply a data mining technique that can be used to identify common features not yet found. HMM has proven to be a highly accurate method in identifying remote homologues and specific domains. Therefore, it could also be applied to the current set of membrane dipping loops in order to compare different techniques and maximize the prediction of membrane dipping loops.

Although TMLOOP has been implemented to analyze membrane proteins exclusively, it might also be interesting to explore the sequence space of soluble proteins with this algorithm. Possible hits found in soluble proteins should then be located in individual structures and mapped on to the functional properties of the given soluble protein. In theory, these motifs should not be found on the protein surface of soluble proteins unless they have evolved to adapt themselves to a different physicochemical environment. A similar approach would be based on the identification of folds found in specific soluble proteins with similar functional properties to particular membrane dipping

loops (e.g. potassium binding sites in soluble proteins and the potassium selectivity filter found in membrane proteins). Comparison of the corresponding sequences might then yield insights regarding the evolution of the domain.

As explained earlier, both TMLOCATE and TMFUN could benefit from the incorporation of pattern discovery and data mining techniques applied to the extra-membraneous regions of membrane proteins. These extra-membraneous regions contain important signals to define and modulate the subcellular location and function of membrane proteins. Therefore, the predictive accuracy of both methods could be significantly increased by addition of features extracted from these regions (i.e. the N-terminus and C-terminus along with extracellular and intracellular loops). Additionally, other features corresponding to the transmembrane regions might be explored, such as transmembrane region length, as previous experimental research has described the role of this parameter to promote retention along the cellular secretory pathway (Munro, 1995, Pedrazzini et al., 1996, Szczesna-Skorupa and Kemper, 2000, Yang et al., 1997).

TMFUN should also be further extended to report more informative predictions. The developed method has not yet reached it maximum predictive power. Therefore, data mining methods should be applied in order to further sub-classify the current predictions achieved by TMFUN. Likewise, specific (outside) predictors could be implemented to sub-classify a particular protein type. While the other recruited computational methods could be used to imply the general features of an unknown membrane protein (e.g. protease), the developed method could then be used to refine those general predictions (e.g. a protease predictor to sub-classify the different types of proteases). These predictors would benefit from the absence of the accumulative error acquired in previous nodes of the multilayer predictor. Therefore, it is likely to implement highly accurate "sub-predictors" that might be very useful in combination with other computational methods. And last but not least, further work is required to identify the biological relevance of signatures of pairs of residues located at a similar depth in the membrane. GPCR subfamilies such as the peptide GPCR sub-family (where the ligand binding site is located outside the membrane) might represent good target classes with which to start this analysis. It is said that good research

raises more questions than it answers...hopefully, our research to date may be thought of in this light.

9.5 References

ALBERTS, B., BRAY, D., LEWIS, J., RAFF, M., ROBERTS, K. & WATSON, J. (1994) *Molecular biology of the cell,* New York, Garland Publishing, Inc.

AOKI, D., LEE, N., YAMAGUCHI, N., DUBOIS, C. & FUKUDA, M. N. (1992) Golgi retention of a trans-Golgi membrane protein, galactosyltransferase, requires cysteine and histidine residues within the membrane-anchoring domain. *Proc Natl Acad Sci USA,* 89, 4319-23.

ARKIN, I. T., BRUNGER, A. T. & ENGELMAN, D. M. (1997) Are there dominant membrane protein families with a given number of helices? *Proteins,* 28, 465-6.

BAIROCH, A. & APWEILER, R. (1998) The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1998. *Nucleic Acids Res,* 26, 38-42.

BANNAI, H., TAMADA, Y., MARUYAMA, O., NAKAI, K. & MIYANO, S. (2002) Extensive feature detection of N-terminal protein sorting signals. *Bioinformatics,* 18, 298-305.

BERMAN, H. M., WESTBROOK, J., FENG, Z., GILLILAND, G., BHAT, T. N., WEISSIG, H., SHINDYALOV, I. N. & BOURNE, P. E. (2000) The Protein Data Bank. *Nucleic Acids Res,* 28, 235-42.

BHASIN, M. & RAGHAVA, G. P. (2004) GPCRpred: an SVM-based method for prediction of families and subfamilies of G-protein coupled receptors. *Nucleic Acids Res,* 32, W383-9.

BHASIN, M. & RAGHAVA, G. P. (2005) GPCRsclass: a web tool for the classification of amine type of G-protein-coupled receptors. *Nucleic Acids Res,* 33, W143-7.

BICKMORE, W. A. & SUTHERLAND, H. G. (2002) Addressing protein localization within the nucleus. *Embo J,* 21, 1248-54.

BIERMANNS, M., VON LAAR, J., BROSIUS, U. & GARTNER, J. (2003) The peroxisomal membrane targeting elements of human peroxin 2 (PEX2). *Eur J Cell Biol,* 82, 155-62.

BODEN, M. & HAWKINS, J. (2005) Prediction of subcellular localization using sequence-biased recurrent networks. *Bioinformatics,* 21, 2279-86.

BOECKMANN, B., BAIROCH, A., APWEILER, R., BLATTER, M. C., ESTREICHER, A., GASTEIGER, E., MARTIN, M. J., MICHOUD, K., O'DONOVAN, C., PHAN, I., PILBOUT, S. & SCHNEIDER, M. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res,* 31, 365-70.

BONIFACINO, J. S., COSSON, P., SHAH, N. & KLAUSNER, R. D. (1991) Role of potentially charged transmembrane residues in targeting proteins for retention and degradation within the endoplasmic reticulum. *Embo J,* 10, 2783-93.

CHOU, K. C. & ELROD, D. W. (1999) Protein subcellular location prediction. *Protein Eng,* 12, 107-18.

CLAROS, M. G. & VINCENS, P. (1996) Computational method to predict mitochondrially imported proteins and their targeting sequences. *Eur J Biochem,* 241, 779-86.

COCQUEREL, L., DUVET, S., MEUNIER, J. C., PILLEZ, A., CACAN, R., WYCHOWSKI, C. & DUBUISSON, J. (1999) The transmembrane domain of hepatitis C virus glycoprotein E1 is a signal for static retention in the endoplasmic reticulum. *J Virol,* 73, 2641-9.

COCQUEREL, L., WYCHOWSKI, C., MINNER, F., PENIN, F. & DUBUISSON, J. (2000) Charged residues in the transmembrane domains of hepatitis C virus glycoproteins play a major role in the processing, subcellular localization, and assembly of these envelope proteins. *J Virol,* 74, 3623-33.

COKOL, M., NAIR, R. & ROST, B. (2000) Finding nuclear localization signals. *EMBO Rep,* 1, 411-5.

COLLEY, K. J. (1997) Golgi localization of glycosyltransferases: more questions than answers. *Glycobiology,* 7, 1-13.

EMANUELSSON, O., NIELSEN, H., BRUNAK, S. & VON HEIJNE, G. (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J Mol Biol,* 300, 1005-16.

EMANUELSSON, O., NIELSEN, H. & VON HEIJNE, G. (1999) ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. *Protein Sci,* 8, 978-84.

FUJIWARA, Y., ASOGAWA, M. & NAKAI, K. (1997) Prediction of Mitochondrial Targeting Signals Using Hidden Markov Model. *Genome Inform Ser Workshop Genome Inform,* 8, 53-60.

GABALDON, T., SNEL, B., VAN ZIMMEREN, F., HEMRIKA, W., TABAK, H. & HUYNEN, M. A. (2006) Origin and evolution of the peroxisomal proteome. *Biol Direct,* 1, 8.

GUO, Y. Z., LI, M. L., WANG, K. L., WEN, Z. N., LU, M. C., LIU, L. X. & JIANG, L. (2005) Fast fourier transform-based support vector machine for prediction of G-protein coupled receptor subfamilies. *Acta Biochim Biophys Sin (Shanghai),* 37, 759-66.

HAWKINS, J. & BODEN, M. (2006) Detecting and sorting targeting peptides with neural networks and support vector machines. *J Bioinform Comput Biol,* 4, 1-18.

HOBMAN, T. C., LEMON, H. F. & JEWELL, K. (1997) Characterization of an endoplasmic reticulum retention signal in the rubella virus E1 glycoprotein. *J Virol,* 71, 7670-80.

HOBMAN, T. C., WOODWARD, L. & FARQUHAR, M. G. (1995) Targeting of a heterodimeric membrane protein complex to the Golgi: rubella virus E2 glycoprotein contains a transmembrane Golgi retention signal. *Mol Biol Cell,* 6, 7-20.

HONSHO, M., HIROSHIGE, T. & FUJIKI, Y. (2002) The membrane biogenesis peroxin Pex16p. Topogenesis and functional roles in peroxisomal membrane assembly. *J Biol Chem,* 277, 44513-24.

HORN, F., WEARE, J., BEUKERS, M. W., HORSCH, S., BAIROCH, A., CHEN, W., EDVARDSEN, O., CAMPAGNE, F. & VRIEND, G. (1998) GPCRDB: an information system for G protein-coupled receptors. *Nucleic Acids Res,* 26, 275-9.

INOUE, Y., IKEDA, M. & SHIMIZU, T. (2004) Proteome-wide classification and identification of mammalian-type GPCRs by binary topology pattern. *Comput Biol Chem,* 28, 39-49.

JONES, J. M., MORRELL, J. C. & GOULD, S. J. (2004) PEX19 is a predominantly cytosolic chaperone and import receptor for class 1 peroxisomal membrane proteins. *J Cell Biol,* 164, 57-67.

KARCHIN, R., KARPLUS, K. & HAUSSLER, D. (2002) Classifying G-protein coupled receptors with support vector machines. *Bioinformatics,* 18, 147-59.

LASSO, G. (2001) 20x20 matrix classification. *Membrane Protein Bioinformatics Group.* Luton, University of Bedfordshire.

LASSO, G., ANTONIW, J. F. & MULLINS, J. G. (2006) A combinatorial pattern discovery approach for the prediction of membrane dipping (re-entrant) loops. *Bioinformatics,* 22, e290-7.

LETOURNEUR, F. & COSSON, P. (1998) Targeting to the endoplasmic reticulum in yeast cells by determinants present in transmembrane domains. *J Biol Chem,* 273, 33273-8.

MA, J., HAYEK, S. M. & BHAT, M. B. (2004) Membrane topology and membrane retention of the ryanodine receptor calcium release channel. *Cell Biochem Biophys,* 40, 207-24.

MUNRO, S. (1995) An investigation of the role of transmembrane domains in Golgi protein retention. *Embo J,* 14, 4695-704.

NAIR, R., CARTER, P. & ROST, B. (2003) NLSdb: database of nuclear localization signals. *Nucleic Acids Res,* 31, 397-9.

NAIR, R. & ROST, B. (2005) Mimicking cellular sorting improves prediction of subcellular localization. *J Mol Biol,* 348, 85-100.

OP DE BEECK, A., ROUILLE, Y., CARON, M., DUVET, S. & DUBUISSON, J. (2004) The transmembrane domains of the prM and E proteins of yellow fever virus are endoplasmic reticulum localization signals. *J Virol,* 78, 12591-602.

PEDRAZZINI, E., VILLA, A. & BORGESE, N. (1996) A mutant cytochrome b5 with a lengthened membrane anchor escapes from the endoplasmic reticulum and reaches the plasma membrane. *Proc Natl Acad Sci U S A,* 93, 4207-12.

PETSALAKI, E. I., BAGOS, P. G., LITOU, Z. I. & HAMODRAKAS, S. J. (2006) PredSL: a tool for the N-terminal sequence-based prediction of protein subcellular localization. *Genomics Proteomics Bioinformatics,* 4, 48-55.

SCHWARTZ, T. W. (1994) Locating ligand-binding sites in 7TM receptors by protein engineering. *Curr Opin Biotechnol,* 5, 434-44.

SUGIYAMA, Y., POLULYAKH, N. & SHIMIZU, T. (2003) Identification of transmembrane protein functions by binary topology patterns. *Protein Eng,* 16, 479-88.

SZCZESNA-SKORUPA, E. & KEMPER, B. (2000) Endoplasmic reticulum retention determinants in the transmembrane and linker domains of cytochrome P450 2C1. *J Biol Chem,* 275, 19409-15.

TOGAWA, R. (2006) Development of a suite of bioinformatics tools for the analysis and prediction of membrane protein structure. Luton, University of Bedfordshire.

TUSNADY, G. E., DOSZTANYI, Z. & SIMON, I. (2004) Transmembrane proteins in the Protein Data Bank: identification and classification. *Bioinformatics,* 20, 2964-72.

VIKLUND, H., GRANSETH, E. & ELOFSSON, A. (2006) Structural classification and prediction of reentrant regions in alpha-helical transmembrane proteins: application to complete genomes. *J Mol Biol,* 361, 591-603.

WALLIN, E. & VON HEIJNE, G. (1998) Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. *Protein Sci,* 7, 1029-38.

WU, C. C. & YATES, J. R., 3RD (2003) The application of mass spectrometry to membrane proteomics. *Nat Biotechnol,* 21, 262-7.

YANG, M., ELLENBERG, J., BONIFACINO, J. S. & WEISSMAN, A. M. (1997) The transmembrane domain of a carboxyl-terminal anchored protein determines localization to the endoplasmic reticulum. *J Biol Chem,* 272, 1970-5.

YERNOOL, D., BOUDKER, O., JIN, Y. & GOUAUX, E. (2004) Structure of a glutamate transporter homologue from Pyrococcus horikoshii. *Nature,* 431, 811-8.

CHAPTER 10

Conclusions

Despite the important roles of membrane proteins in diverse cellular processes, a severe structural and functional gap has emerged in the membrane proteome. While experimental studies are hampered by the lipid composition of the membrane, very little effort has been made to design computational tools to specifically characterize membrane proteins of known sequence but unknown function. Therefore, the research undertaken has been orientated towards the development of novel computational tools to characterize polytopic membrane proteins at the topological level, subcellular location level and molecular function level.

At the topological level, TMLOOP and TMLOOP writer were designed to predict membrane dipping loops (so called re-entrant loops) and refine existing topological models contained in the Swiss-Prot database, which often does not contain any information regarding these particular domains. A full characterization of all membrane dipping loops known to date was achieved. Sequence patterns were found, with both high sensitivity and specificity **(Chapter 4)**. The corresponding literature highlighted some of the residues contained in these patterns as essential for the function of the protein, thus supporting the pattern discovery approach. Furthermore, a significant group of potential functionally important residues and motifs, not previously characterized, were identified. Based on the discovered patterns, a predictive tool was implemented **(Chapter 5)**. This tool was designed to predict membrane dipping loops using a variation of the single motif approach, named the collective motif approach, which was shown to be capable of detecting distantly related membrane dipping loops. Evaluation of TMLOOP by tenfold cross-validation showed impressive levels of both sensitivity and specificity. TMLOOP was successfully applied to the Swiss-Prot database predicting 75 plausible membrane dipping loops not detected by other methods. The topological model*s* of all true positive hits were subsequently updated by including a description of the predicted membrane dipping loop

(Chapter 5). The added description included the approximated boundaries of the structural domain and the corresponding structural classification.

Subcellular location and molecular function prediction methods relied on the same feature extraction method, named TMDEPTH **(Chapter 6)**. This novel feature extraction method exploits the physicochemical constraints imposed by the lipid bilayer to combine sequence and topology, thus computing two-dimensional features in an original fashion. TMDEPTH uses refined topological models and amino acid sequences to calculate pairs of residues located at a similar depth in the membrane, and stores such information in normalized matrices.

TMLOCATE **(Chapter 7)** was implemented to predict the subcellular location of polytopic membrane proteins by combining a variety of different data mining techniques. The obtained results clearly reflected the importance of the transmembrane region in assigning the organellar location of polytopic membrane proteins. Additionally, the evolutionary relationships between different organelles extracted from the results of the evaluation were in accordance with the literature. The obtained results showed that signatures derived from the predicted topology of membrane proteins can be associated with specific subcellular locations of membrane proteins. Such signatures might in fact correspond to retention signals rather than targeting signals due to the likely poor accessibility of the signals contained in the membrane to be recognized by components of the targeting machinery (e.g. protein carriers).

Similar to the development of TMLOCATE, TMFUN **(Chapter 8)** was implemented to predict the molecular function of polytopic membrane proteins by combining a wide range of different data mining techniques. The method was trained with an assembled data set filtered at two different sequence similarity thresholds: i) 40% and ii) 90%. At the sequence similarity threshold of 40%, the obtained results clearly reflect a direct association between the spatial arrangement of residues in the transmembrane regions and the capacity for polytopic membrane proteins to carry out their functions, and therefore the importance of the transmembrane region itself for modulating molecular

function that may be primarily associated with extramembranous regions of the protein, such as ligand binding or signalling. However, at the sequence similarity threshold of 90%, TMFUN was found to attain maximal predictive accuracy. More flexible data sets, such as this, that still avoid pairs of highly identical proteins might be more appropriate as larger sets permit more informative predictions, and protein families with marked sequence similarity are not constrained by stringent sequence similarity thresholds. Interestingly, the obtained results showed that even for membrane proteins whose molecular function is believed to take part mostly outside the membrane, TMFUN still recognizes transmembrane signatures that can be used to obtain reliable functional predictions. Therefore, there must be a specific functional property contained in the transmembrane domain that is related (either directly or indirectly) to the general molecular function of the protein. These functional properties might relate to a wide range of features such as binding specificities for particular prosthetic groups found in the membrane or in the lipid-water interface, specific conformational changes triggered by a particular ligand or lipid compositional dependencies.

The assembled data set used to train both TMLOCATE and TMFUN were manually curated and yet contained a few thousand proteins. An additional tool was implemented, named PROCLASS **(Chapter 3),** to facilitate the manual curation of large sets of proteins according to their subcellular location and molecular function. Evaluation of this tool showed that if the appropriate terms are selected, the number of data points to be manually curated is extensively reduced and yet the effectiveness of this automated approach that allows user intervention is similar to manually clustering the original data set, but can be undertaken in a fraction of the time. This simple and original tool has shown to be crucial for the manual curation of the assembled data sets and should be considered for many future classification purposes.

The developed research has fully achieved the aims and objectives set up at the early stages of this research. The developed methods have for the first time categorically shown that the transmembrane regions of polytopic membrane proteins hold essential information associated with a wide range of functional properties such as filtering and gating processes,

subcellular location and molecular function. TMLOOP has proven to be the best current predictor of membrane dipping loops, and future developments involving its encapsulation within a reliable topology prediction method will surely improve the accuracy of future topological models. TMLOCATE is, to our knowledge, the first in-silico method that is able to specifically predict the subcellular location polytopic membrane proteins. TMFUN has significantly enhanced the standard of functional prediction of polytopic membrane proteins, and provides a sound basis for future work, involving data mining of the extramembranous domains and integration of the developed tools, which will further enhance our progress towards substantially reducing the current functional gap in the membrane proteome.