Transmembrane protein structure prediction using machine learning

Timothy Nugent

Bioinformatics Group

Department of Computer Science

University College London

A thesis submitted to University College London for the degree of

Doctor of Philosophy

August 2010

Abstract

This thesis describes the development and application of machine learning-based methods for the prediction of alpha-helical transmembrane protein structure from sequence alone. It is divided into six chapters.

Chapter 1 provides an introduction to membrane structure and dynamics, membrane protein classes and families, and membrane protein structure prediction.

Chapter 2 describes a topological study of the transmembrane protein CLN3 using a consensus of bioinformatic approaches constrained by experimental data. Mutations in CLN3 can cause juvenile neuronal ceroid lipofuscinosis, or Batten disease, an inherited neurodegenerative lysosomal storage disease affecting children, therefore such studies are important for directing further experimental work into this incurable illness.

Chapter 3 explores the possibility of using biologically meaningful signatures described as regular expressions to influence the assignment of inside and outside loop locations during transmembrane topology prediction. Using this approach, it was possilbe to modify a recent topology prediction method leading to an improvement of 6% prediction accuracy using a standard data set.

Chapter 4 describes the development of a novel support vector machine-based topology predictor that integrates both signal peptide and re-entrant helix prediction, benchmarked with full cross-validation on a novel data set of sequences with known crystal structures. The method achieves state-of-the-art performance in predicting topology and discriminating between globular and transmembrane proteins. We also present the results of applying these tools to a number of complete genomes.

Chapter 5 describes a novel approach to predict lipid exposure, residue contacts, helix-helix interactions and finally the optimal helical packing ar-

rangement of transmembrane proteins. It is based on two support vector machine classifiers that predict per residue lipid exposure and residue contacts, which are used to determine helix-helix interaction with up to 65% accuracy. The method is also able to discriminate native from decoy helical packing arrangements with up to 70% accuracy. Finally, a force-directed algorithm is employed to construct the optimal helical packing arrangement which demonstrates success for proteins containing up to 13 transmembrane helices.

The final chapter summarises the major contributions of this thesis to biology, before future perspectives for TM protein structure prediction are discussed.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Membrane structure and dynamics

1.1.1 Functions

The cell membrane, also referred to as the plasma membrane or phospholipid bilayer, is an organised, sheet-like assembly composed primarily of lipids and proteins that provides cells with individuality by separating them from their environment. Rather than acting as impervious walls, membranes are highly selective permeability barriers containing specific channels and pumps that allow the ionic and molecular composition of the intracellular medium to be closely regulated. The movement of essential substances across the membrane can be either passive, occurring without the input of cellular energy, or active, requiring the cell to expend energy. The size, charge and other chemical properties of the atoms and molecules attempting to cross the membrane will determine their route and whether they succeed. Biological membranes also have certain mechanical properties. The cell membrane plays a role in anchoring the cytoskeleton to provide shape to the cell, and in attaching to the extracellular matrix to help group cells together in the formation of tissues. Eukaryotic cells contain numerous internal membranes that allow the compartmentalisation of specific organelles such as a nucleus, mitochondria, chloroplasts, lysosomes and endoplasmic reticulum. Such membrane-bound organelles allow chemical or biochemical environments that differ from the rest of the cell to be maintained. Formation of these compartments has been closely linked to functional specialisation over the course of evolution. In plants, fungi and bacteria, an additional membrane forms the outermost boundary. The cell wall primarily provides structural support, but also acts as a selective barrier as porins render it largely permeable to molecules less than about 1500 daltons.

Membranes are involved in a vast array of cellular processes that are indispensable for life. Protein receptors embedded in the cell membrane can act as molecular signals allowing cells to respond to their environment and communicate

with each other. The movement of bacteria towards food and the response of target cells towards hormones are processes in which the primary event is the detection of external stimuli by membrane-bound receptors. Membranes are also able to generate chemical or electrical signals, as in the transmission of nerve impulses. Membranes thus control the flow of information between cells and their environment and play a central role in biological communication. Other proteins embedded in the cell membrane serve as markers which identify a cell to other cells. The interaction of these markers with their respective receptors forms the basis of cell-cell interaction in the immune system.

Two important energy conversion processes occur in membrane systems containing ordered arrays of enzymes and other proteins arranged in an electron transport chain. Oxidative phosphorylation, in which adenosine triphosphate (ATP) is produceded via the oxidation of nutrients, occurs in the inner membranes of mitochondria. In plants, algae and some bacteria, light energy is converted into chemical energy in the thylakoid membranes of chloroplasts during photosynthesis. Membranes therefore play an essential role in the cellular energy cycle.

1.1.2 Common features

While membranes are diverse in both structure and function, they share many common attributes. Membranes consist mainly of lipids held together by non-covalent interactions, and also proteins and carbohydrate molecules. Membrane lipids are relatively small molecules containing both a hydrophobic and hydrophilic moiety, resulting in the spontaneous formation of a closed bimolecular sheet in aqueous media composed of two asymmetric monolayers often called a lipid bilayer. The thickness of the bilayer, which is is believed to vary considerably, is thought to be between 25 Å (2.5 nm) and 100 Å (10 nm) (Lewis & Engelman, 1983; Rawicz *et al.*, 2000). It is also known to be electrically polarised with the cell facing side negatively charged (typically -60 millivolts).

Embedded proteins mediate the distinctive functions of membranes, serving as pumps, channels, receptors, energy transducers and enzymes. Both lipid and protein molecules are able to diffuse rapidly in the plane of the membrane (Calvert *et al.*, 2001), unless anchored by specific interactions, though are unable to rotate across the membrane. Membranes can thus be thought of as fluid structures, effectively composed of two-dimensional solutions of proteins and lipids (Singer & Nicolson, 1972).

1.1.3 Phospholipids

Lipids are water-insoluble molecules that are highly soluble in organic solvents. They have a variety of roles in biological systems: they serve as fuels, concentrated energy cores, signal molecules and components of membranes. There are three major types of membrane lipid: phospholipids, glycolipids and cholesterol (Figure 1.1).

Phospholipids are abundant in all types of biological membrane. They are derived from either glycerol, a three-carbon alcohol, or sphingosine, a more complex alcohol. Phospholipids derived from glycerol are known as phosphoglycerides and consist of a glycerol backbone, two fatty acid chains and a phosphorylated alcohol. Typically, the fatty acid chains in phospholipids and glycolipids, which are always unbranched in animals, contain between 14 and 24 carbon atoms (16 and 18 are most common) and may be saturated or unsaturated, with double bonds arranged in cis conformation (Figure 1.1A). Both the length and degree of unsaturation of the fatty acid chains is known to have a profound effect on membrane fluidity. In phosphoglycerides, the C-1 and C-2 of glycerol are esterified to the carboxyl groups of the two fatty acid chains. The C-3 hydroxyl group of the glycerol is esterified to phosphoric acid, producing the compound phosphatidate, the simplest phosphoglyceride, from which almost all other phosphoglycerides are biosynthesised. The exception is sphingomyelin (Figure 1.1B), which is derived from an amino alcohol containing a long, unsaturated hydrocarbon chain called sphingosine, rather than

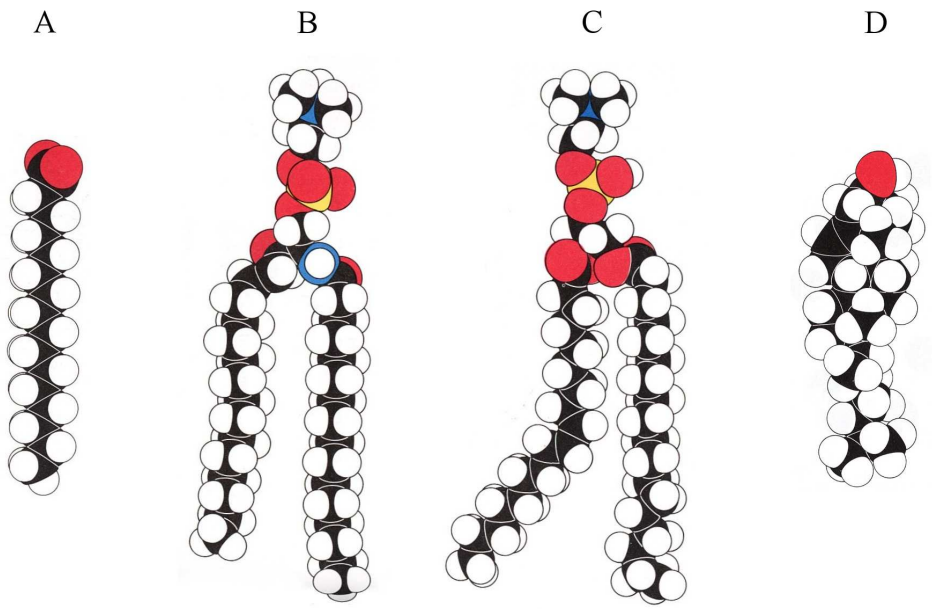

Figure 1.1: Common fatty acids. (A) Palmitate, a 16-carbon phosphoglyceride (B) Sphingomyelin, (C) Phospatidyl choline and (D) Cholesterol (Stryer, 1995).

glycerol. In sphingomyelin, the amino group of the sphingosine backbone is linked to a fatty acid by an amide bond, and the primary hydroxyl group is esterified to phosphoryl choline.

1.1.4 Glycolipids and cholesterol

Glycolipids are sugar-containing lipids which in animal cells are derived from sphingosine. The amino group of the sphingosine backbone is acylated by a fatty acid as in sphingomyelin. However the primary hydroxyl group has one or more sugars attached, rather than phophoryl choline. The simplest glycolipid cerebroside contains only one sugar residue, either glucose or galactose. More complex glycolipids may contain a branched chain with up to seven sugar residues.

In eukaryotes, cholesterol is also present in membranes (Figure 1.1D). Cholesterol contains an oxygen atom in its 3-OH group that comes from O_2. Typically, plasma membranes are rich in cholesterol whereas the membranes of organelles contain much smaller amounts of this lipid.

All membrane lipids share a critical common structural theme: they are all amphipathic molecules that contain both a hydrophobic and hydrophilic moiety (Table 1.1).

Membrane Lipid	**Hydrophobic Unit**	**Hydrophilic Unit**
Phosphoglycerides	**Fatty acid chains**	**Phosphorylated alcohol**
Sphingomyelin	**Fatty acid chain and hydrocarbon chain of sphingosine**	**Phophoryl choline**
Glycolipid	**Fatty acid chain and hydrocarbon chain of sphingosine**	**Sugar residues**
Cholesterol	**Entire molecule except C-3 OH group**	**C-3 OH group**

Table 1.1: Hydrophobic and hydrophilic components of membrane lipids (Stryer, 1995).

Within an aqueous medium, there are two arrangements of phospholipids and glycolipids that satisfy both the water-loving and water-hating passions of the amphipathic molecules. One way is to form a micelle, a globular structure in which the polar head groups are surrounded by water and the hydrocarbon tails are sequestered inside. However, the formation of this structure is unfavourable as phospholipids and glycolipids have two fatty acyl chains that are too bulky to fit into the interior of the micelle. In contrast, salts of fatty acids containing only one chain readily form micelles. The alternative and thus favoured arrangement is the bimolecular sheet, composed of two asymmetric monolayers with the polar head groups on the outside facing the water while the hydrocarbon tails line up against one another on the inside (Figure 1.2).

The formation of lipid bilayers in water is a rapid and spontaneous process driven by hydrophobic forces. Due to the inherent amphipathic character of the lipid molecules, the formation is a self-assembly process. As water molecules are released by the hydrocarbon tails, these tails are sequestered in the non-polar interior of the bilayer where they are then stabilised by van der Waals attractive forces which favour close packing. Electrostatic and hydrogen-bonding then occurs

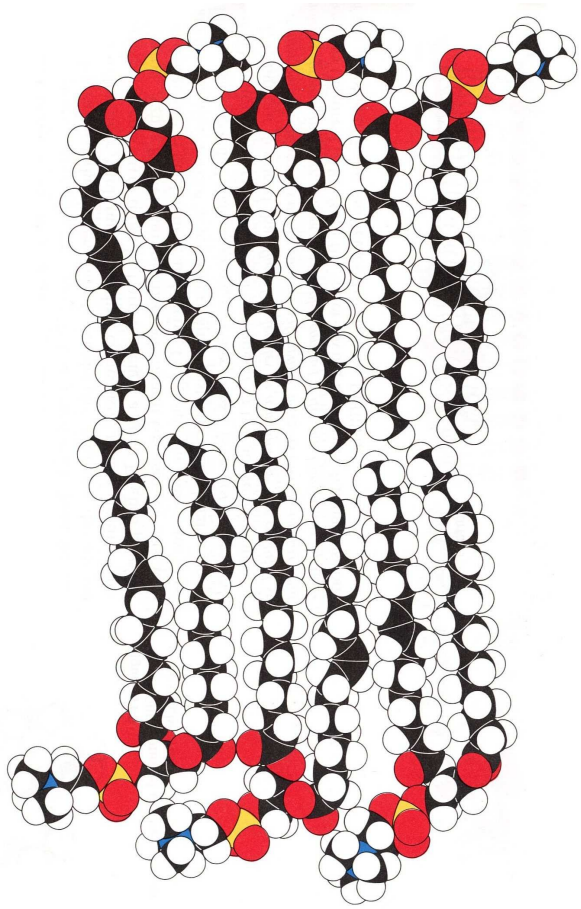

Figure 1.2: A section of the assembled lipid bilayer (Stryer, 1995).

between the polar head groups and water molecules, ensuring that the lipid bilayer is stabilised by the full repertoire of forces that mediate molecular interaction in biological systems.

Interactions between chains are said to be reinforcing. In order to minimise the total number of exposed hydrocarbon chains, the individual components of the lipid bilayer behave in a cooperative fashion, analogous to the huddling of sheep in the cold to minimise the are of exposed body surface. These energetic factors have significant biological consequences: (1) lipid bilayers have an inherent tendency to be extensive, (2) lipid bilayers tend to close on themselves so that no edges are exposed, resulting in compartmentalisation, and (3) lipid bilayers are self-sealing, as a hole in the bilayer which exposes the hydrocarbon tails to water is energetically unfavourable.

1.2 Membrane proteins

Membrane proteins are responsible for most of the dynamic processes carried out by membranes. While membrane lipids form a permeability barrier and thereby establish compartments, it is the role of specific proteins to mediate nearly all the other membrane functions. Membrane proteins can be classified as either peripheral or integral.

Peripheral membrane proteins do not span the membrane and are bound primarily by electrostatic and hydrogen-bond interactions to integral membrane proteins or peripheral regions of the membrane. They have relatively little interaction with the hydrocarbon tails of membrane lipids. Such polar interactions can be disrupted by changes in salt content or pH, unless anchored to the bilayer by a covalently attached chain such as a fatty acid. The regulatory protein subunits of many ion channels and transmembrane receptors are defined as peripheral membrane proteins and have been shown to regulate cell signalling and many other cellular events through a variety of mechanisms. For example, membrane binding may promote rearrangement, dissociation, or conformational changes within many protein structural domains, resulting in activation of their biological activity (Johnson & Cornell, 1999; Thuduppathy *et al.*, 2006). Close association between an enzyme and a biological membranes may also increase proximity with its lipid substrate (Ghosh *et al.*, 2006).

In contrast, integral membrane proteins, or transmembrane (TM) proteins, span the bilayer and interact extensively with the hydrocarbon tails of membrane lipids via hydrophobic interactions. Such proteins can only be studied by disrupting the membrane using an organic solvent or detergent. The TM regions of the proteins are composed of either alpha-helical or beta-barrel structures.

1.2.1 Alpha-helical transmembrane proteins

Alpha-helical membrane proteins, the major category of TM proteins, are present in all type of biological membranes including outer membranes and fulfill a wide

range of functions (Table 1.2). They consist of one of more alpha-helices, containing a stretch of hydrophobic amino acids, embedded in the membrane and linked to subsequent TM helices by extramembranous loop regions. It is thought such proteins may have up to 20 TM helices allowing a wide range of differing topologies. Loop regions are known to contain various substructures including amphipathic helices that lie parallel to the membrane plane, globular domains, and re-entrant helices - short alpha helices that enter and exit the membrane on the same side.

Alpha-helical TM proteins can be further divided into a number of subtypes. Type I proteins have a single TM alpha helix, with the amino terminus exposed to the exterior side of the membrane and the carboxy terminus exposed to the cytoplasmic side. These proteins are subdivided into two types. Type Ia - which constitute most eukaryotic membrane proteins - contain cleavable signal sequences, while type Ib do not. Type II membrane proteins are similar to type I in that they span the membrane only once but their orientation is reversed; they have their amino terminus on the cytoplasmic side of the cell and the carboxy terminus on the exterior.

Type III membrane proteins have multiple TM helices in a single polypeptide chain and are also subdivided into types a and b: type IIIa have cleavable signal sequences while type IIIb have their amino termini exposed on the exterior surface of the membrane, but do not have cleavable signal sequences. Type III membrane proteins include the G-protein-coupled receptors (GPCR) family, members of which consist of seven transmembrane helices (Figure 1.3). GPCRs comprise a large protein family of receptors that sense molecules outside the cell, activate signal transduction pathways and ultimately invoke cellular responses.

Type IV membrane proteins have multiple domains which form an assembly that spans the membrane multiple times. Domains may reside on a single polypeptide chain but are often composed of more than one. Examples include Photosystem I

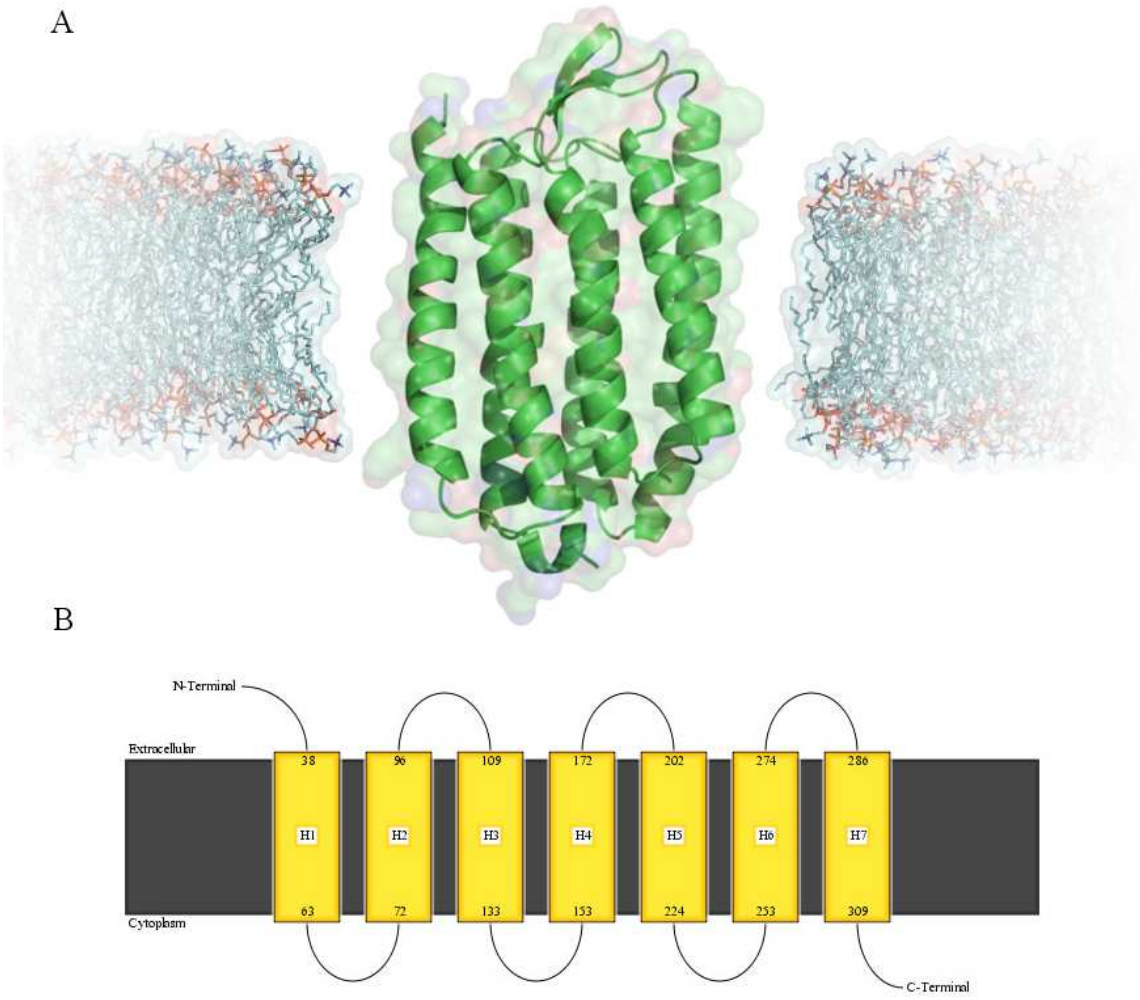

Figure 1.3: A) Bacteriorhodopsin from *Halobacterium salinarium*, a seven transmembrane helix G-protein coupled receptor (GPCR). It acts as a proton pump, using captured light energy to move protons across the membrane out of the cell. PDB code 1PY6. Other GPCRs include halorhodopsin, a light-driven chloride pump, PDB code 1E12. B) Cartoon representation of bacteriorhodopsin topology.

which is comprised of nine unique chains (PDB: 1JB0).

1.2.2 Beta-barrel transmembrane proteins

Beta-barrel TM proteins have been found in the outer membranes of Gram-negative bacteria, cell walls of Gram-positive bacteria, and the outer membranes of mitochondria and chloroplasts (Table 1.3). They consist of a series of anti-parallel beta strands embedded in the membrane, each of which is hydrogen-bonded to the strands immediately before and after it in the primary sequence, connected by extramembranous loops. The beta strands contain alternating polar and hydrophobic amino acids

Function	**Superfamily**
Light-driven transporters	Rhodopsin-like proteins Photosystems Light-harvesting complexes
Oxidoreduction-driven transporters	Transmembrane cytochrome b like Cytochrome c oxidases Multi-heme cytochromes
Electrochemical potential-driven transporters	Proton or Sodium translocating F/V/A-type ATPases
Hydrolysis-driven transporters	P-type ATPase (P-ATPase) Vitamine B12transporter-like ABC transporters Single-helix ATPase regulators Lipid flippase-like ABC transporters Molybdate uptake ABC transporter General secretory pathway (Sec)
Porters	Mitochondrial Carrier (MC) Major Facilitator Superfamily (MFS) Resistance-nodulation-cell division Monovalent cation/proton antiporter (CPA) Neurotransmitter sodium symporter Ammonia transporter (Amt) Drug/Metabolite Transporter (DMT)
Channels including ion channels	Voltage-gated channel like Large conductance mechanosensitive ion channel (MscL) Small conductance mechanosensitive ion channel (MscS) CorA Metal Ion Transporters (MIT) Ligand-gated ion channel (LIC) of neurotransmitter receptors Chloride Channel (ClC) Epithelial sodium channel (EnaC) Magnesium ion transporter-E (MgtE) Major Intrinsic Protein (MIP)
Enzymes	Methane monooxygenase Rhomboid proteins Disulfide bond oxidoreductase-B (DsbB) MAPEG (Eicosanoid and Glutathione metabolism proteins)
Proteins with transmembrane anchors	T cell receptor transmembrane dimerisation domain Steryl-sulfate sulfohydrolase Glycophorin A Inovirus (filamentous phage) major coat protein Pulmonary surfactant-associated protein

Table 1.2: Alpha-helical transmembrane protein superfamilies (Lomize *et al.*, 2006b).

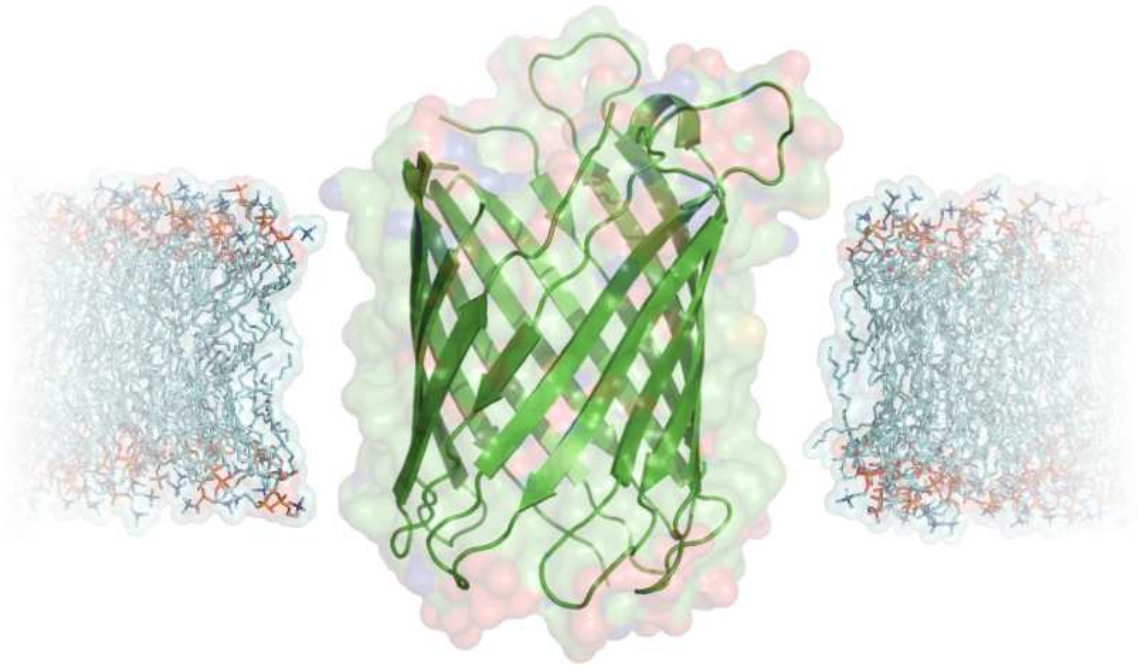

Figure 1.4: A canonical beta-barrel protein, the monomeric porin OmpG from *Escherichia coli*, viewed from the side. Porins are transmembrane proteins with hollow centres through which small molecules can diffuse. PDB code 2F1C.

so that the hydrophobic residues are orientated toward the exterior where they contact the surrounding lipids, and hydrophilic residues are oriented toward the interior pore. All beta-barrel transmembrane proteins have simple up-and-down topology, which may reflect their common evolutionary origin and similar folding mechanism. Beta-barrel TM proteins commonly form porins, sixteen or eighteen-stranded beta-barrels, which assemble into water-filled channels that allow the passive diffusion of nutrients and waste products across the outer membrane (Figure 1.4). Larger, potentially toxic compounds are prevented from entering the cell by the restrictive size of the channel. Porin-like barrel structures are encoded by as many as 2-3% of genes in Gram-negative bacteria (Wimley 2003).

1.2.3 The fluid mosaic model

In 1972, Singer and Nicolson proposed the 'fluid mosaic model' for the organisation and structure of the proteins and lipids of biological membranes (Singer & Nicolson, 1972). The major features of this model are (1) the lipid bilayer has a dual role: it is a solvent for TM proteins and it forms a permeability barrier, (2) a proportion of

Function	**Superfamily**
Outer membranes of Gram-negative bacteria	Nucleoside-specific channel-forming membrane porin OMPT-like Autotransporter (AT) Trimeric autotransporter OM phospholipase FadL outer membrane protein (FadL) OmpG porin Trimeric porins OMPA-like Sugar porins Omp85-TpsB transporters Ligand-gated protein channels Outer Membrane Factor (OMF)
Oligomeric beta-barrels of Gram-positive bacteria	Leukocidin-like

Table 1.3: Beta-barrel transmembrane protein superfamilies (Lomize *et al.*, 2006b).

membrane lipids interact with TM proteins and are likely to be essential for their function, and (3) TM proteins are free to diffuse laterally in the lipid matrix unless restricted by special interactions. The essence of the fluid mosaic model is that membranes are two-dimensional solutions of orientated lipids and globular proteins (Figure 1.5).

The model suggests that proteins in a membrane are dispersed, are at low concentrations and match the dimensions of the unperturbed bilayer. The lipid is seen as a sea in which proteins float, and the bilayer is exposed to the aqueous environment. However, the findings during the last 35 years have weakened this rather generalised view. In the Singer-Nicolson model, molecules are distributed randomly in two dimensions. It is now believed that membranes are patchy, with segregated regions of structure and function. Given the thousands of TM proteins in a proteome and thus vast number of pairwise combinations, a wide range of interaction energies is highly probable. It should therefore be expected that regions of biased composition exist and that the environment in which TM proteins exist must vary, as it is highly improbable that interaction energies will match each other across all protein and lipid species in a membrane (Engelman, 2005). Inspection of known TM

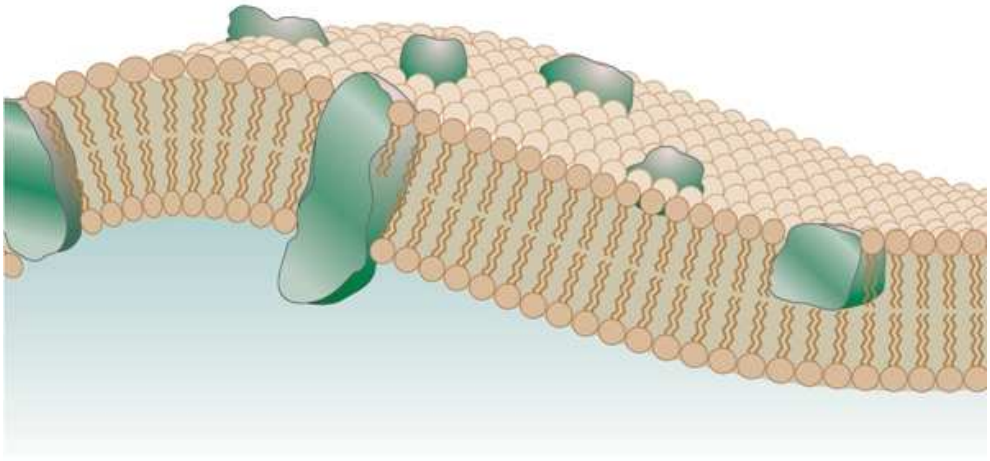

Figure 1.5: Singer-Nicolson proposed the 'fluid mosaic model' (Engelman, 2005).

protein crystal structures leads us to believe that membrane thickness is not uniform as is suggested by the Singer-Nicolson model, but varies from place to place. The lengths of TM helices vary in their hydrophobic dimensions suggesting that either the protein distorts to match the dimensions of the bilayer, or the lipid distorts to match the protein, or both. The fluidity of the lipid and the relative rigidity of proteins suggests it is the lipid that distorts to match the protein, a view supported by experimental and modelling data (Mitra *et al.*, 2004). Crystal structures also indicate that, while the TM component of a TM protein may be fairly compact, extramembranous domains often occupy much larger areas in projection on to the membrane. Proteins anchored by single helices often have ectodomains that cover large areas of the membrane with protein. In many proteins, such as tyrosine kinase receptors, this interaction is functionally important (Binda *et al.*, 2002; Bracey *et al.*, 2002; Ferguson *et al.*, 2003). The combination of crowding and the presence of large ectodomains is likely to limit the exposure of lipid to adjacent aqueous regions. These new themes are illustrated in an amended and updated version of the 'fluid mosaic model' (Figure 1.6).

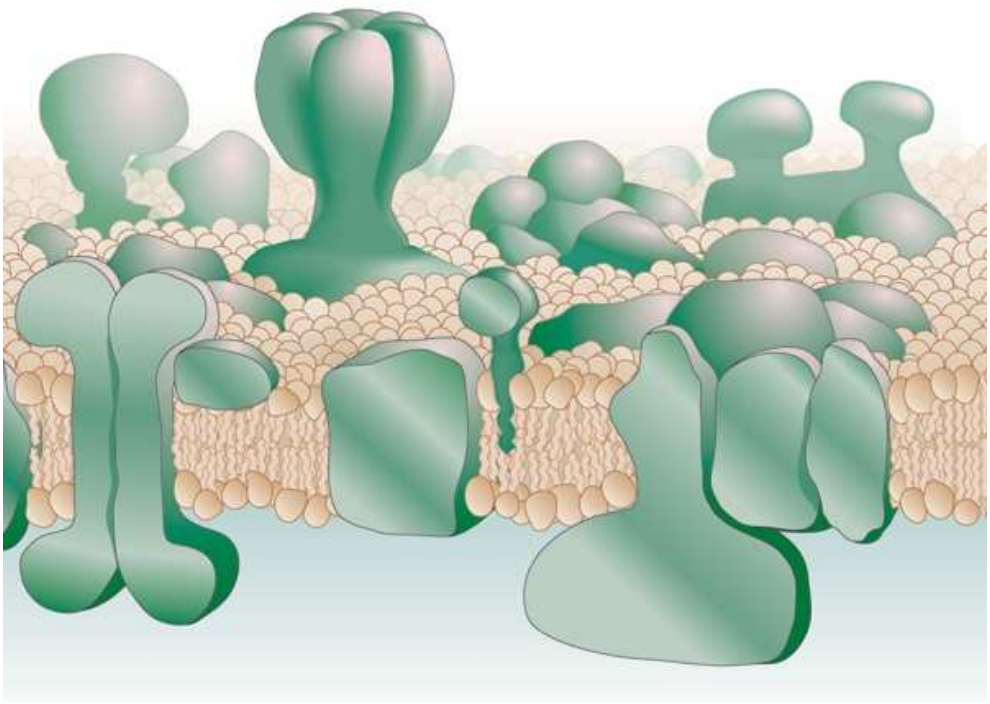

Figure 1.6: An amended and updated 'fluid mosaic model' (Engelman, 2005).

1.2.4 Membrane targeting and insertion

Like all other proteins, a TM protein begins its journey on a ribosome. From this point on, alpha-helical and beta-barrels TM proteins are handled differently. Ribosomes upon which alpha-helical TM proteins are being assembled typically bind cotranslationally to translocons in the target membrane (the inner membrane in bacteria or the endoplasmic reticulum (ER) in eukaryotes). They proceed to move laterally from the translocon channel into the surrounding lipid bilayer. Depending on the local hydrophobicity and the ability of adjacent helices to form stable interactions with each other, this may occur one helix at a time or in pairs. Evidence suggests that the molecular features of the TM protein that enable the translocon to identify a region as TM or non-TM are the same as those seen to mediate protein-lipid interaction in known TM protein structures. This indicates that the translocon allows a translocating nascent chain to sample the surrounding bilayer (Elofsson & von Heijne, 2007; White & von Heijne, 2004).

Beta-barrel proteins are initially transferred from the ribosome to a soluble cyto-

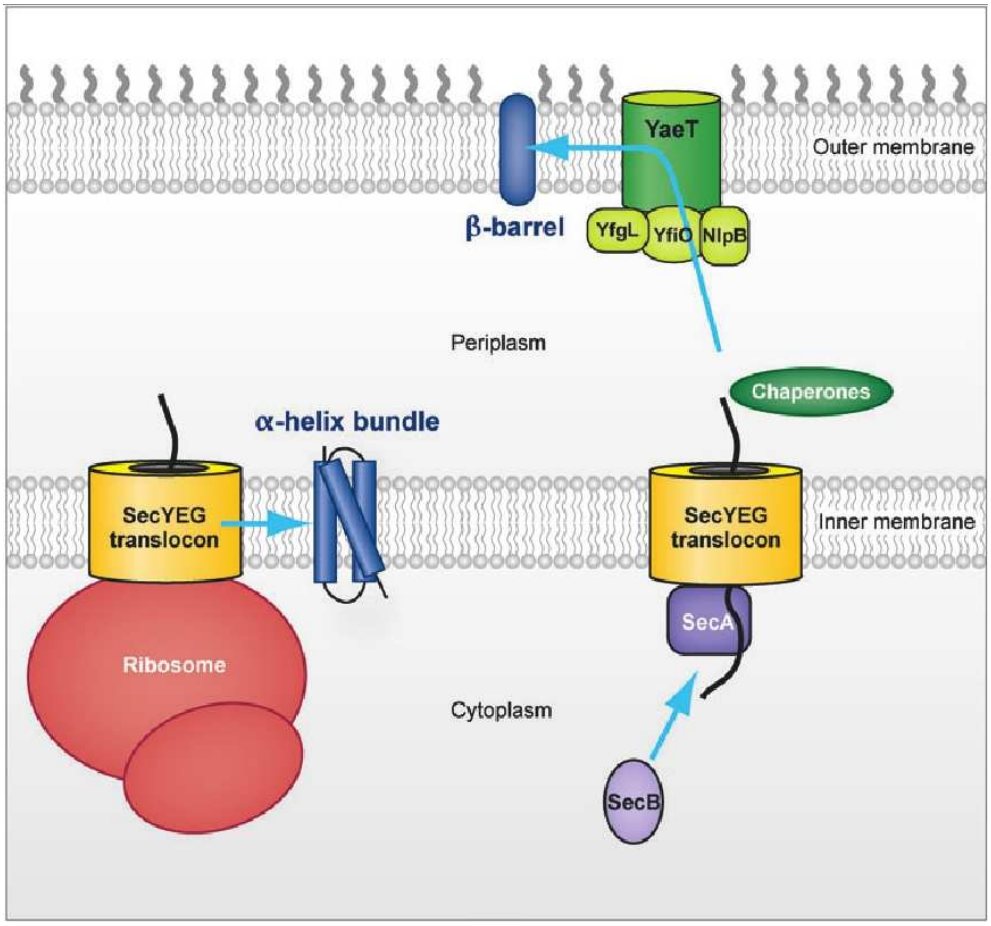

Figure 1.7: Biogenesis of alpha-helical (left) and beta-barrel (right) transmembrane proteins in *Escherichia coli* (Elofsson & von Heijne, 2007).

plasmic chaperone, SecB (Elofsson & von Heijne, 2007; White & von Heijne, 2004; Luirink *et al.*, 2005). They are then translocated through the inner membrane translocon with the aid of SecA ATPase, but do not become embedded in the inner membrane as their short beta-strands are not sufficiently hydrophobic. They are then chaperoned through the periplasmic space and finally insert into the outer membrane with the aid of YaeT hetero-oligomeric outer membrane integrating complex (Figure 1.7) (Luirink *et al.*, 2005; MacIntyre *et al.*, 1988; Ruiz *et al.*, 2006).

1.2.5 Signal peptides and anchors

Signal peptides are short sequences that govern the transport and localisation of a protein in a cell. They targets a protein for translocation across the plasma membrane in prokaryotes and across the ER membrane in eukaryotes (van Vliet *et al.*, 2003). They are typically N-terminal peptides 15-30 amino acids long,

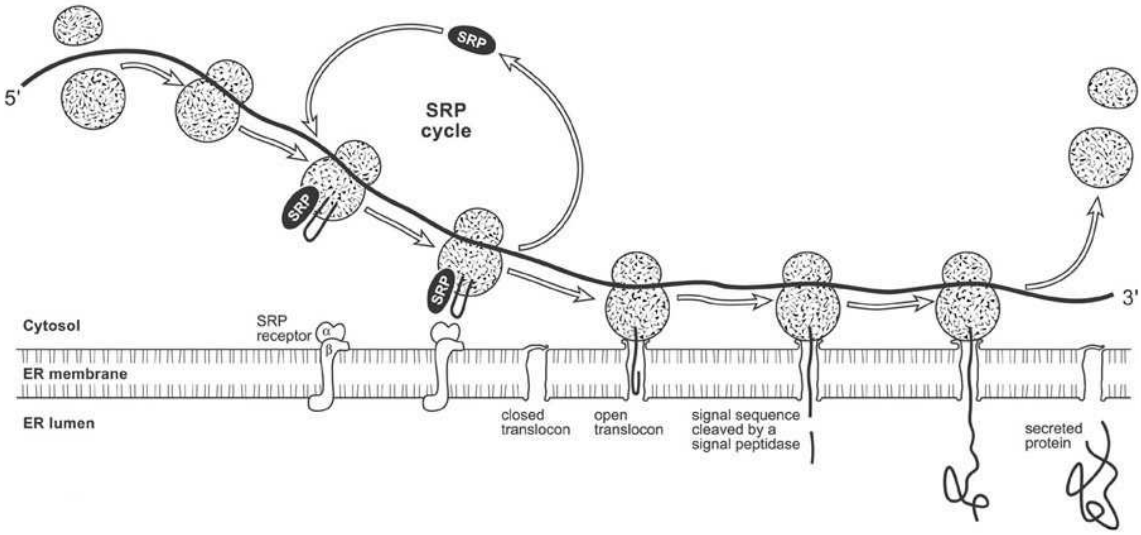

Figure 1.8: Mechanism of synthesis of membrane bound or secreted proteins (King, 2009).

and are cleaved off during translocation by signal peptidase I (SPase). While there is no consensus sequence for a signal peptide, there are three distinct compositional zones. Firstly, an N-terminal region that usually contains charged residues. Next follows a hydrophobic region of as least 6 residues, and finally a C-terminal region of uncharged polar residues that directly proceeds a cleavage site, around which there is conservation at positions -3 and -1 (Emanuelsson *et al.*, 2007). Signal anchors, features of type II TM proteins, are effectively uncleaved signal peptides which share similar composition to signal peptides but have no SPase recognition site. Signal anchors are also known to occur at the C-terminus.

In eukaryotes, signal peptides are recognised by a recognition particle (SRP) during synthesis on a ribosome. The SRP then binds to an SRP receptor embedded in the ER membrane. After sufficient synthesis the signal peptide is removed by SPase. Synthesis will continue and if the protein is secreted it will end up completely in the lumen of the ER. TM proteins possess a stop transfer motif that prevents the transfer of the protein through the ER membrane. The TM protein will then become embedded in the ER membrane (Figure 1.8) (King, 2009).

1.3 Transmembrane protein topology prediction

1.3.1 Membrane proteins are difficult to crystallise

TM proteins, which have both hydrophobic and hydrophilic regions on their surfaces, are much more difficult to isolate than water-soluble proteins, as the native membrane surrounding the protein must be disrupted and replaced with detergent molecules without causing any denaturation. Despite considerable efforts, relatively few TM proteins have yielded crystals that diffract to high resolution. While it is thought that TM proteins comprise approximately 30% of a proteome, they are significantly under-represented in structural databases such as the Protein Data Bank (Bernstein *et al.*, 1978) where they comprise only about 1% of total deposited structures (White, 2004). Tables 1.2 and 1.3 summarise the alpha-helical and beta-barrel crystal structures currently available (Lomize *et al.*, 2006b). However, with advanced technologies such as synchrotron light sources becoming available, it is now possible to determine X-ray structures from ever-smaller protein crystals. Combined with novel crystallisation methods such as the use of antibodies to solubilise proteins, the rate at which TM protein structures are being elucidated should increase over the coming years.

1.3.2 Alpha-helical transmembrane protein topology prediction

Due to their severe under-representation in structural databases, the prediction of TM protein structure is extremely difficult. Given the biological and pharmacological importance of TM proteins, an understanding of their topology - the total number of TM helices, their boundaries and in/out orientation relative to the membrane - is therefore an important target for theoretical prediction methods. A number of experimental methods, including glycosylation analysis, insertion tags, antibody studies and fusion protein constructs, allow the topological location of a region to be identified. However, such studies are time consuming, often conflicting (Mao *et al.*, 2003a; Kyttälä *et al.*, 2004), and also risk upsetting the natural

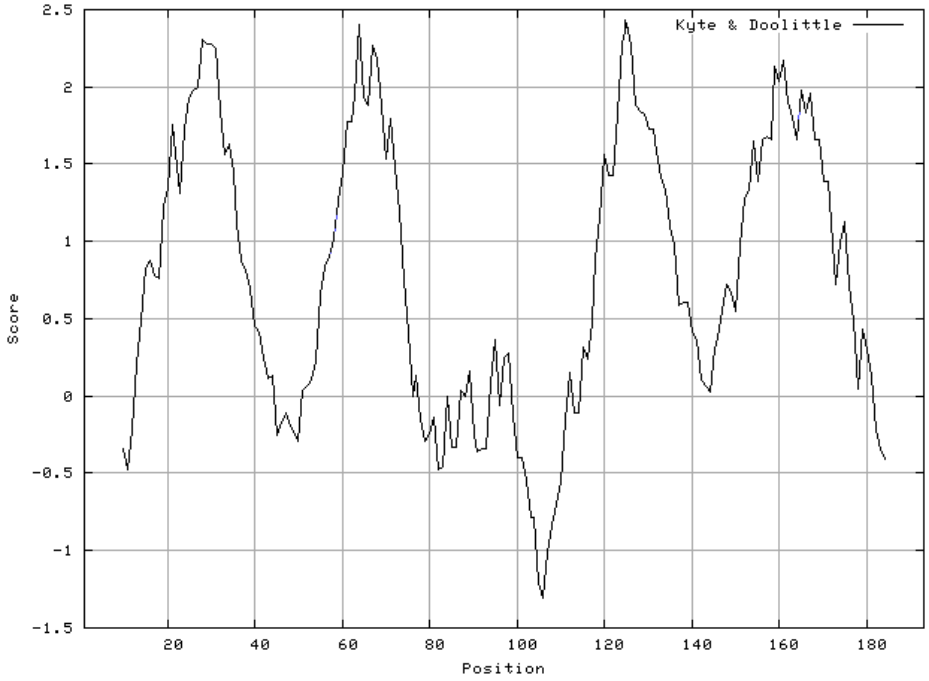

Figure 1.9: A Kyte-Doolittle hydropathy plot. The protein sequence is scanned with a sliding window of size 19-21 residues. At each position, the mean hydrophobic index of the amino acids within the window is calculated and that value plotted as the midpoint of the window. This plot represents a TM protein with 4 TM helices.

topology by altering the protein sequence.

In the absence of structural data, bioinformatic strategies thus turn to sequence-based prediction methods. Long before the arrival of the first crystal structures, stretches of hydrophobic residues long enough to span the lipid bilayer were identified as TM spanning helices. Early prediction methods by Kyte & Doolittle (1982) and Engelman *et al.* (1986), and later by Wimley & White (1996), relied on experimentally determined hydropathy indices to create a hydropathy plot for a protein. This involved taking a sliding window of 19-21 residues and averaging the score with peaks in the plots (regions of high hydrophobicity) corresponding to TM helices (Figure 1.9).

With more structures came the discovery that aromatic Trp and Tyr residues tend to cluster near the ends of the transmembrane segments (Wallin *et al.*, 1997), possibly acting as physical buffers to stabilise TM helices within the lipid bilayer. More recent studies identified the appearance of sequence motifs, such as the GxxxG

motif (Senes *et al.*, 2000), within TM helices and also periodic patterns implicated in helix-helix packing and 3D structure (Samatey *et al.*, 1995). However, perhaps the most important realisation was that positively-charged residues tend to cluster on cytoplasmic loops - the 'positive-inside' rule of von Heijne (von Heijne, 1992). Combined with hydrophobicity-based prediction of TM helices, this led to early topology prediction methods such as TopPred (Claros & von Heijne, 1994).

1.3.3 Machine learning-based approaches

Despite their success, these early methods based on the physicochemical principle of a sliding window of hydrophobicity combined with the 'positive-inside' rule have since been replaced by machine learning approaches which prevail over hydrophobicity methods due to their statistical formulation. A selection of machine learning-based predictors can be found in (Table 1.4).

1.3.3.1 Hidden Markov models

A Hidden Markov model (HMM) is a statistical model in which the system is assumed to be a Markov process - a mathematical model for the random evolution of a system where the likelihood of a given future state, at any given moment, depends only on its present state, and not on any past states. In regular Markov models, the state is directly visible and therefore the state transition probabilities are the only parameters. In HMMs, the states are not directly visible, although the state dependent outputs are visible (Bishop, 2006).

In the context of a biological sequence, we may wish to define a number of states for each label we wish to assign. For TM topology prediction, we might use three states to represent TM helices, inside and outside loop regions. Each state will has its own emission probabilities which models the composition of each state, for example hydrophobic amino acid residues will have higher emission state probabilities within the TM helix state, while positively charged residues will have higher emission states within the inside loop state. Each state also has transition

probabilities, the probabilities of moving from the current state to a different one. The transition edges describe the linear order in which state changes are expected to occur, and create a state path on transition from state to state. This state path is a Markov chain, meaning that the next state only depends on the current state.

In analysing an unknown sequence, we want to infer the hidden state path and therefore identify the topogenic labels that make up the topology. There are potentially many state paths that could generate the same sequence so the task is usually to find the one with the highest probability. The efficient Viterbi algorithm is guaranteed to find the most probable state path given a sequence and an HMM. The Viterbi algorithm is a dynamic programming algorithm similar to those used by various sequence alignment methods. Posterior decoding, which uses forward and backward dynamic programming algorithms that are similar to Viterbi, can then be used to sum over all possible paths in order to calculate the confidence of each state path.

While HMMs are frequently used within bioinformatics, one caveat is that they do not deal with correlations between residues well as they assume that each residue depends on only one underlying state. Long-range pairwise correlations, for example where a salt bridge is formed between two charged residues on non-adjacent TM helices, may be missed by a HMM when attempting to predict secondary structure since an HMM has no way of recalling what was generated by a distant state (Eddy, 2004).

HMMs were first applied to TM topology prediction in HMMTOP (Tusnady & Simon, 1998) and TMHMM (Krogh *et al.*, 2001) and have proved highly successful. TMHMM implements a cyclic model with seven states for a TM helix, while HMMTOP uses HMMs to distinguish between five structural states [helix core, inside loop, outside loop, helix caps and globular domains]. These states are connected by transition probabilities before dynamic programming is used to match a sequence

Method	URL	Algorithm	Features
MEMSAT3	http://bioinf.cs.ucl.ac.uk/psipred/	NN	Signal peptide, MSA, HGA
MINNOU	http://minnou.cchmc.org/	NN	
PHDhtm	http://www.predictprotein.org/	NN	MSA
Phobius	http://phobius.sbc.su.se/	HMM	Signal peptide, MSA, constrained
TMHMM	http://www.cbs.dtu.dk/services/TMHMM/	HMM	HGA
PRODIV-TMHMM	http://www.pdc.kth.se/~hakanv/prodiv-tmhmm/	HMM	Re-entrant region, HGA
HMMTOP	http://www.enzim.hu/hmmtop/	HMM	Constrained
ENSEMBLE	http://pongo.biocomp.unibo.it/pongo/	NN + HMM	MSA
OCTOPUS	http://octopus.cbr.su.se/	NN + HMM	Re-entrant region
SVMtop	http://bio-cluster.iis.sinica.edu.tw/~bioapp/SVMtop/	SVM	
PONGO	http://pongo.biocomp.unibo.it/pongo/	Multiple	Consensus
BPROMPT	http://www.jenner.ac.uk/bprompt/	Multiple	Consensus

Table 1.4: Machine learning-based alpha-helical TM topology predictors. HMM: Hidden Markov model. NN: Neural network. MSA: Topology predictions made using multiple sequence alignments. HGA: Suitable for whole genome analysis.

against a model with the most probable topology. HMMTOP also allows constrained predictions to be made, where specific residues can be fixed to a topological location based on experimental data.

1.3.3.2 Neural networks

Artificial neural networks (NNs) are mathematical models that attempt to simulate the structure and function of biological neural networks. They are non-linear statistical data modelling tools that can be used to model complex relationships between inputs and outputs or to find patterns in data. Originally inspired by the central nervous system and the vastly interconnected neurons which constitute it, NN models are comprised of nodes which are connected together to form a network that processes information. In many cases a NN is an adaptive system that changes its structure depending on the external or internal information that flows through the network during the learning phase.

In supervised learning, where the objective is to deduce a function from a

training set, the NN will attempt to infer the mapping implied by the training data. A cost function is used as a measure of how far away a particular solution is from an optimal solution to the problem to be solved. Learning algorithms search through the solution space to find a function that has the smallest possible cost, which is related to the mismatch between the mapping and the data. Training the NN means selecting one model from a set of allowed models that minimises this cost function. There are various algorithms available for training neural network models; most employ some form of gradient descent. This is achieved by taking the derivative of the cost function with respect to the network parameters and then changing those parameters in a gradient-related direction (Bishop, 2006).

Like HMMs, NNs are commonly used in bioinformatics. The first and most simple type of artificial NN devised was the feed-forward NN. In this network, the information moves in only one direction, forward, from the input nodes, through any hidden nodes and to the output nodes. There are no cycles or loops in the network.

NNs are employed by methods including PHDhtm (Rost *et al.*, 1996) and MEMSAT3 (Jones, 2007). PHDhtm uses multiple sequence alignments to perform a consensus prediction of TM helices by combining two feed-forward NNs. The first creates a 'sequence-to-structure' network which represents the structural propensity of the central residue in a window. A 'structure-to-structure' network then smoothes these propensities to predict TM helices, before the positive-inside rule is applied to produce an overall topology. MEMSAT3 uses a feed-forward neural network and dynamic programming in order to predict not only TM helices, but also to score the topology and to identify possible signal peptides. Additional evolutionary information provided by multiple sequence alignments led to prediction accuracies increasing to as much as 80% using one dataset (Möller *et al.*, 2000).

1.3.3.3 Support Vector Machines

Support vector machines (SVMs) are a group of supervised learning methods that can be applied to classification or regression tasks. Presented with a data set of training examples with each marked as belonging to one of two categories, the SVM training algorithm is able to construct a model that can accurately predict whether novel examples fall into one category or the other. The SVM model is a representation of the examples as points in a high dimensional space, mapped in such a way that that the examples of the two categories can be divided by a clear gap whose width is maximised, allowing positioning of a hyperplane. In searching for the best hyperplane, the SVM finds a set of data points that are the most difficult to classify. These data points are referred to as support vectors. The new examples are then mapped into the same space and are assigned to a category depending on which side of the hyperplane they fall. Good separation is achieved by the hyperplane that has the largest distance to the nearest training data points of any class, known as the functional margin, since in general the larger this margin is, the lower the generalisation error of the classifier. This strategy allows SVM classifiers to provide improved generalisation performance compared with other classification algorithms (Bishop, 2006).

SVMs were first devised by Vapnik (1998) who used a linear separating hyperplane to maximise the distance between two classes in order to create a classifier (Figure 1.10). Data points represented as p-dimensional vectors were separated with a $p-1$-dimensional hyperplane, called a linear classifier - a classifier constructed from a linear combination of the p values contained within the feature vector. While a number of hyperplanes might be used to separate the data classes, the best choice is the one that separates the classes by the largest margin so that the distance between the nearest data points from each class to the hyperplane is maximised. Such a hyperplane is therefore known as the maximum-margin hyperplane and the linear classifier it defines is known as a maximum margin classifier.

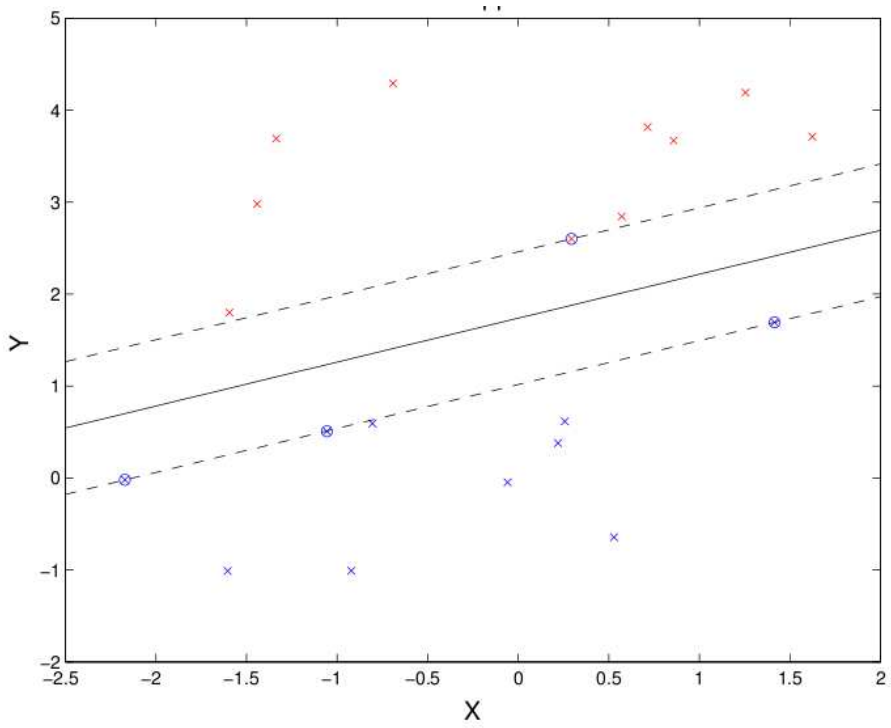

Figure 1.10: Decision surface of an SVM classifier for a linearly separable problem in two dimensions. The decision boundary f(x) = 0 is shown by a solid line. The circled points are the support vectors, which lie on the dashed lines representing the geometric margin (Ward, 2005).

For data classes than cannot be linearly separated in the original Euclidean input space, several adaptations of the maximal margin classifier are required. Soft margin hyperplanes add a penalty function of violation of constraints to the optimisation criterion, allowing input vectors that are corrupted by noise to be separated by a hyperplane as cleanly as possible while still maximising the separation distance. The method introduces slack variables which represent the geometric distance to the margin hyperplanes for examples that fail to have a specified margin. An extra cost term is also included to penalise margin errors by controlling the trade-off between large margin and low empirical risk (Cristianini & Shawe-Taylor, 2000).

Another extension is the use of the kernel trick to solve a non-linear problem by mapping the original non-linear observations into a higher-dimensional space where the linear classifier is subsequently used; this makes a linear classification in the new space equivalent to non-linear classification in the original space. The kernel

trick is based on Mercer's theorem which states that any continuous, symmetric, positive semi-definite kernel function can be expressed as a dot product in a high-dimensional space. The non-linear classifier uses a non-linear kernel function in place of the dot product, allowing the algorithm to fit the maximum-margin hyperplane in a transformed feature space. Common kernels implemented by SVM packages include the linear (Equation 1.1), polynomial (Equation 1.2), sigmoid (Equation 1.3) and radial basis function (Equation 1.4).

$$\psi(x, x_i) = (x \cdot x_i) \tag{1.1}$$

$$\psi(x, x_i) = (\gamma \cdot x \cdot x_i + 1)^d \tag{1.2}$$

$$\psi(x, x_i) = \tanh(\gamma \cdot x \cdot x_i) \tag{1.3}$$

$$\psi(x, x_i) = exp\left(-\gamma \|x - x_i\|^2\right) \tag{1.4}$$

$$y = sign\left\{\sum \alpha_i t_i \psi(x, x_i)\right\} \tag{1.5}$$

Equations 1.1 to 1.5: Equation 1.1: linear kernel function. Equation 1.2: polynomial kernel function. Equation 1.3: sigmoid kernel function. Equation 1.4: radial basis function kernel. Equation 1.5: The decision equation. x is a p-dimensional vector representing the test data point. x_i is a p-dimensional vector representing the ith support vector. $\psi(x, x_i)$ is the kernel function which quantifies the similarity between a test data point and the support vectors. t_i is the class label of the ith support vector. α_i is the positive parameter of the ith support vector determined by the SVM algorithm. d is the degree of the polynomial function. γ is typically set to 1 divided by the number of features.

Perhaps the most difficult aspect of implementing an SVM is the choice or design of an appropriate kernel function. Many kernels, including those described above, are specifically designed for dealing with numerical features. However, when dealing with a data set composed of non-numerical attributes such as protein or DNA sequences, the kernel function must be specially designed or the features must be numerically encoded (Yang, 2004).

Recently, SVMs have been applied to TM protein topology prediction (Yuan *et al.*, 2004; Lo *et al.*, 2008). While NNs and HMMs are capable of producing multiple outputs, SVMs are binary classifiers therefore multiple SVMs must be employed to classify the numerous residue preferences before being combined into a probabilistic framework. Although multiclass ranking SVMs do exist, they are generally considered unreliable since in many cases no single mathematical function exists to separate all classes of data from one another (Abe, 2003). However, SVMs are capable of learning complex relationships among the amino acids within a given window with which they are trained, particularly when provided with evolutionary information, and are also more resilient to the problem of over-training compared to other machine learning methods, although numerous adjustable parameters can result in optimisation becoming extremely time consuming.

1.3.4 Consensus approaches

A number of methods now combine multiple machine learning approaches. ENSEMBLE (Martelli *et al.*, 2002) uses a NN and two HMMs, while OCTOPUS (Viklund & Elofsson, 2008) uses two sets of four NNs and one HMM. Both groups report higher prediction accuracies compared with methods based on only a single classification algorithm. BPROMPT (Taylor *et al.*, 2003), which takes a consensus approach, combines the outputs of five different predictors to produce an overall topology using a Bayesian belief network, while Nilsson *et al.* (2002) used a simple majority-vote approach to return the best topology from their five predictors. The PONGO server (Amico *et al.*, 2006) returns the results of 5 high scoring methods in a graphical format for direct comparison. In most cases, but particularly proteins whose topology is not straightforward, considering a number of predictions by different methods is highly advisable (Figure 1.11).

1.3.5 Signal peptides and re-entrant helices

One problem faced by modern topology predictors is the discrimination between TM helices and other features composed largely of hydrophobic residues. These include

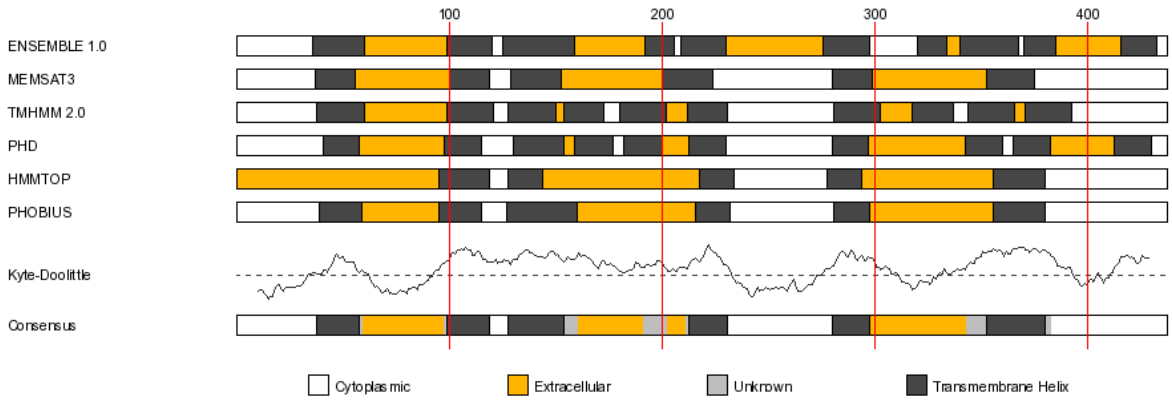

Figure 1.11: Using a number of methods to form a consensus.

targeting motifs such as signal peptides and signal anchors, amphipathic helices, and re-entrant helices which are common in many ion channel families (Figure 1.12). The high similarity between such features and the hydrophobic profile of a TM helix frequently leads to crosstalk between the different types of predictions. Should these elements be predicted as TM helices, the ensuing topology prediction is likely to be severely disrupted. Some prediction methods, such as SignalP (Bendtsen *et al.*, 2004) and TargetP (Emanuelsson *et al.*, 2007) are effective in identifying signal peptides, and may be used as a pre-filter prior to analysis using a TM topology predictor. Phobius (Käll *et al.*, 2004) used a HMM to successfully address the problem of signal peptides in TM protein topology prediction, while PolyPhobius (Käll *et al.*, 2005) further increased accuracy by including homology information. Other methods such as TOP-MOD (Viklund *et al.*, 2006) and OCTOPUS have attempted to incorporate identification of re-entrant regions into a TM topology predictor but there is significant room for improvement. The problem, particularly regarding re-entrant helices, is the lack of reliable data with which to train machine-learning based methods.

1.3.6 Beta-barrel proteins

The relative abundance of alpha-helical TM proteins in both complete proteomes and 3D databases, when compared to beta-barrel TM proteins has resulted in the latter class being somewhat overshadowed in terms of efforts to predict structure

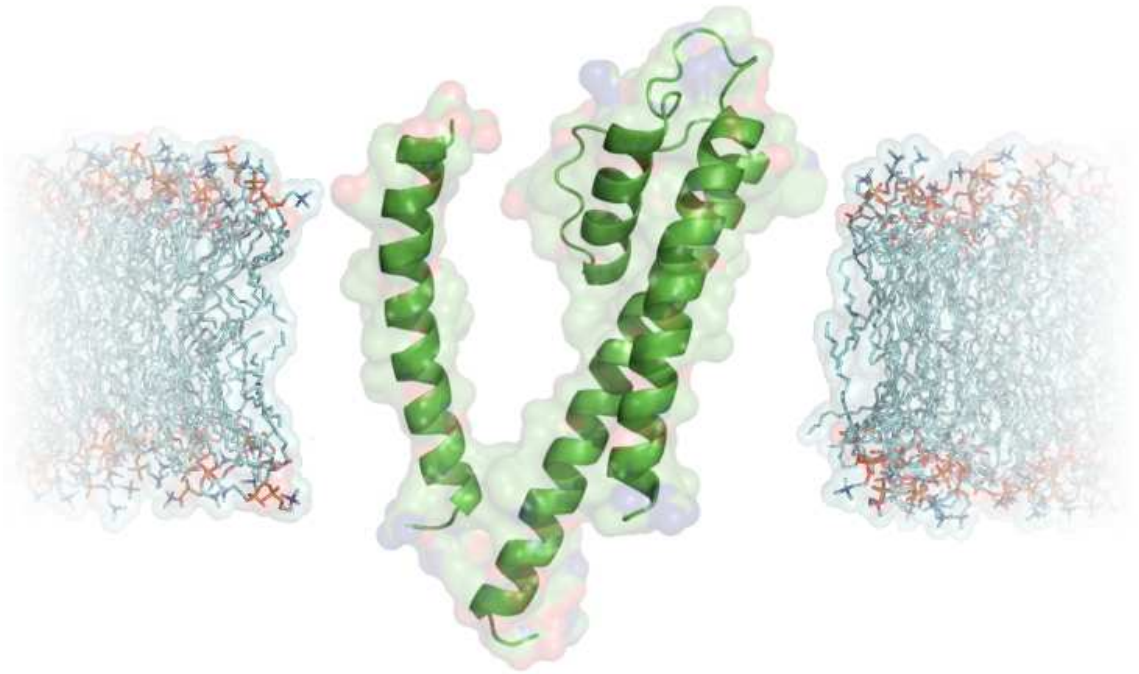

Figure 1.12: Potassium channel subunit from *Streptomyces lividans* showing a short re-entrant helix (PDB: 1R3J).

and topology. Perhaps another reason is the relative ease with which alpha-helical TM helices can be predicted due to their enrichment of hydrophobic residues. The anti-parallel beta-strands of beta-barrel TM proteins contain alternating polar and hydrophobic amino acids, allowing the hydrophobic residues to orientate towards the membrane while the polar residues are oriented toward the solvent-exposed surface. Early methods used to predict such beta-strands relied on sliding window-based hydrophobicity analyses in order to capture the alternating patterns (Schirmer & Cowan, 1993), while other approaches included the construction of special empirical rules using amino acid propensities and prior knowledge of the structural nature of the proteins (Gromiha & Ponnuswamy, 1993). As the number of structures of beta-barrel proteins known at atomic resolution increased, machine learning based methods began to emerge trained on these larger datasets. These include NN, (Jacoboni *et al.*, 2001; Gromiha *et al.*, 2004), HMM (Martelli *et al.*, 2002; Liu *et al.*, 2003b; Bagos *et al.*, 2004a) and SVM-based predictors (Park *et al.*, 2005), using single sequences and multiple sequence alignments. A selection of machine learning-based beta-barrel predictors can be found in Table 1.5.

Method	URL	Algorithm	Features
B2TMR	http://gpcr.biocomp.unibo.it/predictors/	NN	MSA
TMBETA-NET	http://psfs.cbrc.jp/tmbeta-net/	NN	MSA, HGA
HMM-B2TMR	http://gpcr.biocomp.unibo.it/predictors/	HMM	MSA
PROFtmb	http://www.rostlab.org/services/PROFtmb/	HMM	HGA
PRED-TMBB	http://biophysics.biol.uoa.gr/PRED-TMBB/	HMM	HGA
TMBETA-SVM	http://tmbeta-svm.cbrc.jp/	SVM	HGA
TMB-Hunt2	http://bmbpcu36.leeds.ac.uk/	HMM + SVM	HGA

Table 1.5: Machine learning-based beta-barrel TM topology predictors. MSA: Topology predictions made using multiple sequence alignments. HGA: Suitable for whole genome analysis.

1.3.7 Databases

A number of databases now exist that serve as repositories for the sequences and structures of TM proteins. OPM (Lomize *et al.*, 2006b), PDB_TM (Tusnady *et al.*, 2005a), CGDB (Chetwynd *et al.*, 2008), MPDB (Raman *et al.*, 2006) and Stephen White's database (White, 2010) all contain TM proteins of known structure determined using X-ray and electron diffraction, nuclear magnetic resonance and cryoelectron microscopy. OPM, PDB_TM and CGDB additionally contain orientation predictions of the protein relative to the membrane based on water-lipid transfer energy minimisation (Lomize *et al.*, 2006a), hydrophobicity/structural feature analysis (Tusnady *et al.*, 2005b) and coarse grained molecular dynamic simulations (Sansom *et al.*, 2008). For topological studies, OPM provides N-terminus localisation information, while TOPDB (Tusnady *et al.*, 2008) and MPtopo (Jayasinghe *et al.*, 2001) also include TM proteins of unknown 3D structure whose topologies have been experimentally validated using low-resolution techniques such as gene fusion, antibody and mutagenesis studies. A number of TM protein databases collect information on specific families including potassium channels (Li & Gallin, 2004) and GPCRs (Horn *et al.*, 1998), while others such as LGICdb (Donizelli *et al.*, 2006) and TCDB (Saier *et al.*, 2006) focus on particular structural or functional classes.

The Möller dataset (Möller *et al.*, 2000), although in need of modification based

on recent SWISS-PROT annotations (Boeckmann *et al.*, 2003), provides a diverse training and validation set that suffers less from the prokaryotic bias present in 3D structure derived sets. As with all bioinformatics databases, care should be taken to ensure that a given resource is frequently updated. The rate at which new sequences and structures are deposited in Genbank and the PDB (and occasionally retracted e.g. Pornillos *et al.* (2005)) results in significant manual annotation for database administrators, and much evidence suggests that this workload often exceeds the amount of time an administrator is willing to commit.

1.3.8 Multiple sequence alignments

Multiple sequence alignments play an important role in TM protein structure prediction. Homologous sequences identified via database searches can be used to construct sequence profiles which can significantly enhance TM topology prediction accuracy (Käll *et al.*, 2005; Jones, 2007), while template structures can be used for homology modelling.

The most commonly used methods for detecting homologous sequences are the BLAST (Basic Local Alignment Search Tool) and PSI-BLAST (Position-Specific Iterated BLAST) algorithms (Altschul *et al.*, 1997). These methods work on the premise that the greater the similarity between protein or DNA sequences, the more recent the divergence from a common ancestor is likely to be and therefore the more structural and functional characteristics will be shared by the related sequences. BLAST identifies words in the query sequence with a match score above a particular threshold, before searching a sequence database for high-scoring word hits. On detection of a hit, the alignment is extended in both directions producing an alignment score.

PSI-BLAST improves on BLAST by automatically constructing profiles from BLAST alignments. PSI-BLAST first creates a list of all closely related proteins. These proteins are then combined into a general profile sequence, or position-specific

scoring matrix (PSSM), which summarises the significant features present in these sequences. The frequencies of all amino acids at each position in the multiple alignment are used to weight the original scoring matrix to account for the residues that are present in the proteins recovered identified by the search. The sequence database is then searched using this profile, returning a larger number of proteins. This larger group is then used to create another profile before the process is repeated. By including these related proteins in the search, PSI-BLAST is much more sensitive in picking up distant evolutionary relationships than a standard BLAST search.

While conventional pair-wise alignment methods such as BLAST and PSI-BLAST return possible matches based on a scoring function that relies on amino acid substitution matrices such as PAM (Dayhoff, 1979) or BLOSUM (Henikoff & Henikoff, 1992), such matrices are derived from globular protein alignments, and as amino acid composition, hydrophobicity and conservation patterns differ between globular and TM proteins (Jones *et al.*, 1994b) they are in principle unsuitable for TM protein alignment. A number of TM-specific substitution matrices have therefore been developed, which take into account such differences. For example, the JTT TM matrix (Jones *et al.*, 1994b) was based on the observation that polar residues in TM proteins are highly conserved, while hydrophobic residues are more interchangeable. Other matrices such as SLIM (Muller *et al.*, 2001), were reported to have the highest accuracy for detecting remote homologues in a manually curated GPCR dataset, while PHAT (Ng *et al.*, 2000) has been shown to outperform JTT, especially on database searching. However, to date, no independent study has accessed these TM-specific substitution matrices on a common dataset.

Few novel methods have been developed to improve actual TM protein alignment. STAM (Shafrir & Guy, 2004) implemented higher penalties for insertion/deletions in TM segments compared to loop regions, with combinations of different substitution matrices to produce alignments resulting in more accurate homology models. PRALINE^{TM} (Pirovano *et al.*, 2008), which incorporates membrane-specific substi-

tution matrices, was shown to outperform standard multiple alignment techniques such as ClustalW (Thompson *et al.*, 1994) and MUSCLE (Edgar, 2004) when tested on the TM alignment benchmark set within BaliBASE (Bahr *et al.*, 2001). Recent adjustment to BLAST and PSI-BLAST to reflect the composition of the query sequence should theoretically improve results for TM protein searches (Altschul *et al.*, 1997), though again this has not been assessed. An advanced alignment method T-Coffee (Notredame *et al.*, 2000), despite using a single generic scoring matrix, performs well at high sequence identities when tested against a benchmark data set of homologous membrane protein structures, while HMAP (Tang *et al.*, 2003) can improve alignment significantly using a profile-profile based approach incorporating structural information.

1.3.9 Whole genome analysis

Large-scale genomics and proteomics projects are frequently identifying novel proteins, many of which are of unknown localisation and function. While some of the methods outlined above can accurately predict TM topology, fewer are suitable for discriminating between globular and TM proteins. To do so requires the method to be specially trained for this process, and that the program is available as a standalone package as web-based predictors are unsuitable for such large-scale submissions. A number of methods which are suitable for whole genome analysis of alpha-helical and beta-barrel TM proteins are shown in Tables 1.4 and 1.5. In general, error rates are minimised by prior filtering to remove signal and transit peptides using methods such as SignalP and TargetP, since many globular proteins with such signal sequences are frequently predicted as single spanning TM proteins. Currently, the best methods are capable of error rates of less than 1% for alpha-helical TM proteins (Jones, 2007) and less than 6% for beta-barrel TM proteins (Park *et al.*, 2005).

1.3.10 Data sets, homology, accuracy and cross-validation

A key element when constructing any prediction method is the use of a high quality data set for both training and validation purposes. Extracting a training set from

available databases requires a large amount of work and requires a number of critical decisions to be made. As an example in the case of TM proteins, searches of databases such as the PDB using the keyword 'transmembrane' will return both genomically encoded TM proteins as well as TM proteins that are not native, such as PDB entry 1BH1 - a bilayer disrupting peptide found in bee venom - and 1CII, a bacterial colicin used to form pores in the outer membranes of competing bacteria. Furthermore, errors in databases are not infrequent and add an element of noise. While such noise is often well tolerated by machine learning methods, the problem is more significant in smaller data sets.

Another issue that needs to be addressed is homology in the data, with most data sets being reduced at a level of 30-40% sequence identity. Since structural TM protein data is at a premium, this level is perhaps slightly higher than that which would be applied to globular protein data sets. Although there is an increased risk of overfitting, this is necessary to ensure training sets are of sufficient size. All machine learning methods have multiple free parameters and thus have the potential to overfit. That is, rather than identifying a pattern in a sequence, an example may be learned 'by heart', including any noise that the sequence may contain. A method that has been overfitted is typically able to reproduce its training examples accurately, but will perform poorly on examples that it has not seen before. It is important that, when assessing the accuracy of a prediction method, homology in both training and test data sets is reduced in order to avoid overfitting.

In all cases, it is important that stringent cross-validation is performed. Cross-validation is the statistical practice of partitioning a data set into subsets such that a single subset is validated on a model trained using the remaining subsets, and the process is continued until all subsets have been validated. Two types are common in TM topology prediction. In K-fold cross-validation, the data set is partitioned into K subsets. Of the K subsets, a single subset containing a number of sequences

is retained as validation data for testing the model, while the remaining K-1 subsets are used as training data. This process is then repeated K times (folds), with each of the K subsets being used exactly once as the validation data. The K results from the folds can then either be combined or averaged to produce a single estimation. A more stringent, although computationally more intensive form of cross-validation is leave-one-out cross-validation (LOOCV), also referred to as a jack knife test. Jack knifing involves testing a single sequence from the data set against the remaining sequences which make up the training set, then repeating the test such that every sequence is validated once. This is the same as a K-fold cross-validation with K being equal to the number of sequences in the data set.

While some studies have attempted to compare TM topology prediction accuracy between different methods (e.g. Melen *et al.* (2003)), significant progress has been made since then. Currently, the best TM topology predictors claim to predict correct topologies for 80-93% of proteins, though in the absence of independent cross-validation using a common test set it is difficult to accurately compare methods. Those which perform well when tested on a particular data set, e.g. one containing few signal peptides, may perform poorly when tested on a data set which contains many signal peptides. Methods optimised on a data set containing many weakly hydrophobic TM helices may tend to over predict TM helices in other data sets. Current gold-standard TM protein data sets with topologies derived solely from structural data contain no more than 150 sequences when homology reduced (Lomize *et al.*, 2006b), but a lack of consensus amongst these combined with the scarcity of necessary cross-validation data means that differences in accuracy between methods may thus be a result of differences in training and validation data sets rather than significant differences in performance.

1.3.11 3D structure prediction

As with globular proteins, 3D structure prediction of TM proteins can be dealt with via two approaches, homology modelling and *ab initio* modelling.

Homology modelling, also known as comparative modelling, involves the use of a related template structure in order to build a 3D model of a target protein. The method is based on the observation that protein structure is conserved more highly than amino acid sequence, hence even proteins that have diverged significantly in sequence but still share detectable similarity (>30% sequence identity) may also share common structural properties, particularly the overall fold. Due to the difficulties involved in obtaining high-resolution crystal structures, particularly with regard to TM proteins, homology modelling can provide useful structural models for generating hypotheses about a protein's function and directing further experimental work. The process can be subdivided into four steps: template selection, target-template alignment, model construction and model assessment, all of which can be performed iteratively in order to improve the quality of the final model (Sanchez & Sali, 1997; Marti-Renom *et al.*, 2000). A selection of homology modelling programs are shown in Table 1.6.

Aside from SWISS-MODEL (Peitsch, 1996) which has a 7TM/GPCR interface, none of the methods in Table 1.6 are specifically designed to deal with TM proteins. In particular, care must therefore be taken to ensure that models do not contain polar side chains that protrude into the hydrophobic membrane region. Specific side chain modelling tools such as SCWRL (Canutescu *et al.*, 2003) may suffer from this same problem, though the accuracy of extramembranous regions of the model is likely to increase. Despite the lack of TM protein-specific modelling tools, recent research has demonstrated that bioinformatics tools currently applied to soluble proteins, from profile matching to secondary structure prediction and homology modelling, perform at least as well on TM proteins (Forrest *et al.*, 2006b). Indeed, an important application of TM protein modelling lies in the identification and validation of drug targets, as well as the identification and optimisation of lead compounds. Homology model-based drug design has been applied to a number of kinases including epidermal growth factor-receptor tyrosine kinase protein (Ghosh

Method	URL	Description
Modeller	http://www.salilab.org/modeller/	Modelling by satisfying spatial restraints. Includes de novo loop modelling.
SegMod/ENCAD	http://csb.stanford.edu/levitt/segmod/	Modelling by segment matching combined with molecular dynamics refinement.
SWISS-MODEL	http://swissmodel.expasy.org/	Web server modelling by rigid-body assembly.
3D-JIGSAW	http://bmm.cancerresearchuk.org/	Web server modelling server with energy minimisation using CHARMM.
Nest	http://wiki.c2b2.columbia.edu/	Multiple template-based modelling using an artificial evolution method.
Builder	On request	Self Consistent Mean-Field theory (SCMF) (Koehl and Delarue 1996) approach for loop and side chain modelling.
Jackal	http://wiki.c2b2.columbia.edu/	Modelling using a selection of different programs.
SCWRL3	http://dunbrack.fccc.edu/	Backbone-dependent rotamer library-based side chain modelling.

Table 1.6: A selection of commonly used homology modelling programs, adapted from Wallner & Elofsson (2005).

et al., 2001), Bruton's tyrosine kinase (Mahajan *et al.*, 1999) and Janus kinase 3 (Sudbeck *et al.*, 1999).

Ab initio modelling, or *de novo* modelling, involves the construction of a 3D model in the absence of any structural data relating to the target protein or a homolog. Research has focused in three main areas: alternate lower-resolution representations of proteins, accurate energy functions, and efficient sampling methods. While most methods address globular proteins, some efforts have been directed at TM protein structure prediction.

ROSETTA (Rohl *et al.*, 2004) is an *ab initio* modelling program that uses potential functions for computing the lowest energy structure for an amino acid sequence. Feedback from the prediction is used continually to improve potential functions and search algorithms. A modified version of the ROSETTA algorithm (Barth *et al.*, 2007) uses an energy function that describes membrane intraprotein

interactions at atomic level and membrane protein/lipid interactions implicitly, while treating hydrogen bonds explicitly. Results suggest that the model captures the essential physical properties that govern the solvation and stability of membrane proteins, allowing the structures of small TM protein domain (< 150 residues) to be predicted successfully to a resolution of < 2.5 Å. This accuracy compares favourably with predictions obtained on small water-soluble protein domains. The ROSETTA membrane method has also been combined with homology modelling and domain assembly methods to model the structures of the Kv1.2 and KvAP potassium channels, resulting in models with good similarity to their crystal structures. Modelling of the open and closed states of these channels has provided insight into the mechanism of voltage-dependent gating through conformational change, providing testable hypotheses for further experimental work (Yarov-Yarovoy *et al.*, 2006).

FRAGFOLD (Jones, 1997, 2001) is a fragment-based protein tertiary structure prediction method, based on the assembly of supersecondary structural fragments using a simulated annealing algorithm. The strategy attempts to greatly narrow the search of conformational space by preselecting fragments from a library of highly resolved protein structures. FILM (Pellegrini-Calace *et al.*, 2003) adds a membrane potential to the FRAGFOLD energy terms (pairwise, solvation, steric and hydrogen bonding). The membrane potential has been derived by the statistical analysis of a data set made of 640 transmembrane helices with experimentally defined topology and belonging to 133 proteins extracted from the SWISS-PROT database. Results obtained by applying the method to small membrane proteins of known 3D structure show that the method is able to predict, at a reasonable accuracy level, both the helix topology and the conformations of these proteins.

1.3.12 Future developments

Despite the good results obtained using both ROSETTA and FILM, a number of limitations of these approaches need to be addressed in future work. The main

limitation at present is the difficulty in handling large transmembrane structures. The combinatorial complexity of *ab initio* protein folding methods means that it is not feasible to use such method for structures with more than about 150 amino acids. Several approaches might be used to overcome this limitation. The simplest improvement to implement for FILM would be to construct a more restricted supersecondary structure fragment library, perhaps based solely on TM protein structures. This would greatly bias the fragment search to conformations likely to form part of large transmembrane structures. A further improvement could be achieved by using larger structure fragments than just supersecondary motifs. Future challenges to enable ROSETTA to make predictions on larger domains include enhanced conformational sampling strategies and more accurate treatment of electrostatics.

1.4 Structure of thesis

One of the greatest unsolved problems in bioinformatics is understanding how a sequence of amino acids folds into a 3D structure. While most current research in this area focuses on globular proteins, the paucity of structural data means that relatively little effort has been put towards TM protein structure prediction. This thesis will focus on novel approaches to improve alpha-helical TM protein structure prediction.

The next chapter describes a topological study of an uncharacterised TM protein thought to cause a fatal neurodegenerative disease, using a consensus of bioinformatic approaches constrained by experimental data. A number of the tools previously discussed will be applied to a sequence whose topology is not straightforward, in the hope of identifying the correct topology and any interesting structural features. In doing so, it may become possible to direct further experimental work and help understand the disease mechanism. Such an investigation represents the typical use of TM topology prediction for the investigation of an unknown sequence.

The third chapter describes an attempt to increase TM topology prediction

accuracy by modifying an existing NN-based method to enable the presence of biologically meaningful sequence motifs to influence the assignment of topogenic regions during topology prediction. Here we experiment with the idea of assigning topogenic weights to motifs that can be represented as regular expressions, and benchmark the performance of the modified approach on a standard data set. This chapter demonstrates a generic method that could be applied to a range of topology predictors in order to improve prediction performance.

The fourth and fifth chapters describe the development of two novel SVM-based methods. Chapter four describes a novel SVM-based TM topology predictor that integrates both signal peptide and re-entrant helix prediction, benchmarked with full cross-validation on a novel data set of sequences with known crystal structures. Topology prediction performance is compared with a number of recent methods, as is the ability to discriminate between globular and TM proteins. The results of applying these tools to a number of complete genomes are also presented.

The fifth chapter describes a novel approach to predict lipid exposure, residue contacts, helix-helix interactions and finally the optimal helical packing arrangement of TM proteins. It is based on two SVMs that predict per residue lipid exposure and residue contacts, which are then used to determine helix-helix interaction. The method is also able to discriminate native from decoy helical packing arrangements. Finally, a force-directed algorithm is employed to construct the optimal helical packing arrangement. In combination, these tools are likely to assist in reducing the conformational sampling space during *ab initio* modelling.

The final chapter summarises the major contributions of this thesis to biology, before future perspectives for TM protein structure prediction are discussed.

Chapter 2

The transmembrane topology of Batten disease protein CLN3

2.1 Background

2.1.1 Neuronal Ceroid Lipofuscinoses

Neuronal Ceroid Lipofuscinoses (NCLs) are a group of at least eight genetically separate autosomal recessive inherited diseases characterised by progressive blindness, permanent loss of motor ability, neurodegeneration and the severe accumulation of lipopigments, composed of fats and proteins that appear green-yellow when viewed under ultraviolet light, in the lysosomes of neurons and other cell types in organs including the liver, spleen, myocardium, and kidneys. With a frequency of approximately 1 in 10,000 in the United States and northern European populations (Vesa *et al.*, 2002), they are the most common childhood onset neurodegenerative disorders but are currently incurable.

NCLs were historically classified according to the clinical onset of symptoms as infantile, late-infantile, juvenile, and adult forms, although several variant forms are now recognised. All are believed to progress at different rates. While seven human disease gene loci have been identified, the functions of most of the encoded proteins remain unknown. Mutations in a number of genes are however thought to be responsible; so far CLN1, CLN2, CLN3, CLN5, CLN6, CLN7, CLN8, CTSD and possibly CLCN6 genes are believed to be implicated (Pardo *et al.*, 1994).

Understanding the functions of the proteins encoded by these genes will undoubtedly shed light on the disease mechanism which in turn may lead to the development of novel therapies. CTSD is known to encode the proteolytic enzyme cathepsin D. Mutations in CTSD cause disease evident at or before birth, with very severe brain atrophy and death occurring soon afterwards. CLN1 encodes the enzyme palmitoyl protein thioesterase 1 (PPT1) which removes palmitate residues from proteins. Mutations in this gene cause NCL with a wide clinical spectrum; onset is typical in infancy but can be delayed until adulthood. CLN2 encodes a lysosomal enzyme tripeptidyl peptidase (TPP1), a member of a recently defined

family of serine-carboxyl proteinases involved in the removal of tripeptides from the N-terminus of small proteins; mutations cause classic late infantile NCL, and in some instances a more protracted disease with later onset. Mutations in CLN5, CLN6, CLN7, CLN8 and other as yet unknown genes are known to cause variant late infantile NCL. For disease caused by mutations in CLN5, symptoms may include developmental regression, visual impairment, ataxia, myoclonus and epilepsy. For disease caused by mutations in CLN6 and CLN7, seizures and motor difficulties present before visual failure. Mutations in CLN8 can cause two different disease, one mutation is associated with Northern epilepsy or Progressive epilepsy with mental retardation, only recently recognised as an NCL. Other mutations cause typical NCL with seizures and deteriorating motor skills the leading symptoms followed by myoclonus, visual failure and loss of cognitive skills (Mole, 1998).

2.1.2 CLN3 mutations cause Batten disease

Mutations in the CLN3 gene underlie juvenile onset NCL (JNCL), also known as Batten disease. Batten disease presents with visual failure, typically progressing over 2-3 years to an appreciation of light and dark only. This is followed in most cases by deterioration in cognitive skills, speech and mobility occurs in the early teenage years together with the onset of seizures. Behaviour may also become problematic at this time as aggressiveness, psychosis, mood disturbances and anxiety occur. Speech becomes dysfluent and mobility becomes characteristically slow and shuffling with a slightly stooped posture. As the disease progresses myoclonic jerks and parkinsonian features become prominent. Communication, mobility and self-help skills are lost.

Batten disease sufferers usually carry a 1kb intragenic deletion on at least one disease allele although some mutations causing a mild or more protracted disease in which visual failure occurs but further symptoms can be delayed well into adulthood. A small group of patients assumed to have mutations in an unknown CLN9 gene also cause juvenile onset NCL (Mole, 1998). While the CLN3 gene is known to encode a

438 amino acid TM protein (The International Batten Disease Consortium, 1995), its function remains elusive and currently hinders understanding of the molecular basis of this fatal disease. Elucidating the TM topology of CLN3 may give insight to its function and mechanism of action.

2.1.3 CLN3 topology is controversial

Despite several experimental studies using antibodies, inserted tags and glycosylation mutagenesis, CLN3 topology remains controversial. Previous predicted topological models for CLN3 have proposed between five and eight TM helices and differ with their placement of the amino terminus on either side of the membrane (Janes *et al.*, 1996; Mitchison *et al.*, 1997; Mao *et al.*, 2003a). The bioinformatic methods used in these predictions have relied on early hydrophobicity-based methods to detect TM helices. As discussed in the previous chapter, such methods have now been superseded by machine learning approaches which demonstrate significantly improved performance when benchmarked on standard data sets, particularly when evolutionary information is used to enhance the prediction (Viklund & Elofsson, 2004; Jones, 2007).

In this chapter, we present a topological study of the CLN3 protein using a selection of recent machine learning-based TM protein topology predictors, constrained by experimental data. Our results suggest that CLN3 has a six TM helix topology with cytoplasmic N and C-termini, three large lumenal loops, one of which may contain an amphipathic helix, and one large cytoplasmic loop, a model which is in agreement with almost all published experimental data. While these results support the accuracy of these machine learning based topology prediction methods, surprisingly varied topological predictions made using different subsets of orthologous sequences highlights the challenges still remaining for topology prediction and the importance of using experimental data to confirm such predictions.

2.2 Methods

2.2.1 CLN3 topology prediction using a selection of recent predictors

We analysed CLN3 sequences using a selection of the highest scoring prediction methods via the PONGO server (Amico *et al.*, 2006). The PONGO server provides topological annotation for all alpha-helical TM proteins in the human genome through a web interface as well as via distributed annotation systems (DAS) queries. In order to produce a comprehensive analysis of query sequences, annotations are carried out by four high scoring predictors: TMHMM (Krogh *et al.*, 2001), MEMSAT3 (Jones, 2007), PRODIV (Viklund & Elofsson, 2004) and ENSEMBLE (Martelli *et al.*, 2003). PRODIV is a recent method which uses a HMM similar to TMHMM, but exploits evolutionary information derived from multiple sequence alignments. ENSEMBLE is an combination of two HMMs and one NN. ENSEMBLE also takes advantage of the evolutionary information derived from multiple sequence alignments, both for the NN and HMM systems. Additionally, the signal peptide predictor SPEP, based on combination of NNs, is used (Fariselli *et al.*, 2003). SPEP has performance similar to the most widely used signal peptide predictor SignalP (Bendtsen *et al.*, 2004) and is included since signal peptides are commonly mispredicted as TM helices. Stored and pre-computed predictions for the human proteins can be searched and displayed in a graphical view, while the web service allows the topology prediction of any kind of putative membrane proteins (Figure 2.1).

2.2.2 Using MEMSAT3 with PSI-BLAST profiles derived from custom databases

MEMSAT3 was used in conjunction with a database of 50 diverse multi-species CLN3 sequences that had been manually curated (MEMSAT + CLN3), a database of a subset of 40 microbial CLN3 sequences (MEMSAT + Microbial CLN3) and SWISS-PROT release 54.0 (MEMSAT + SWISS-PROT) in order to produce PSI-BLAST profiles used to enhance the prediction, while the remaining methods were

Figure 2.1: Typical graphical output from the PONGO server for a TM protein containing a single TM helix (shown in red) and a signal peptide (orange).

run using either human or individual microbial CLN3 sequences. By ensuring that all sequences included in the profiles were CLN3 orthologues, we could be certain that false positive hits were excluded.

2.2.3 Additional prediction methods and experimental data

Experimental data using antibody staining on selectively permeabilised cells (Kyttälä *et al.*, 2004) and susceptibility to protease digestion and N-terminal block to Edman degradation (Ezaki *et al.*, 2003) strongly indicated that the amino terminus was located on the cytoplasmic side of the membrane, so for this reason all models with predicted lumenal amino termini were excluded. This information was also used to constrain a prediction using the Phobius server (Käll *et al.*, 2004). Phobius is a combined TM protein topology and signal peptide predictor. The predictor is based on a HMM that models the different sequence regions of a signal peptide and the different regions of a TM protein in a series of interconnected states. Compared to TMHMM and SignalP, errors coming from cross-prediction between TM segments and signal peptides were reduced substantially by Phobius when benchmarked on a manually curated data set, suggesting that Phobius is well suited for whole genome annotation of signal peptides and TM regions (Figure 2.2).

ScanPROSITE (de Castro *et al.*, 2006) was also used to detect potential phosphorylation and N-glycosylation sites, PSIPRED (Jones, 1999) was used to assess

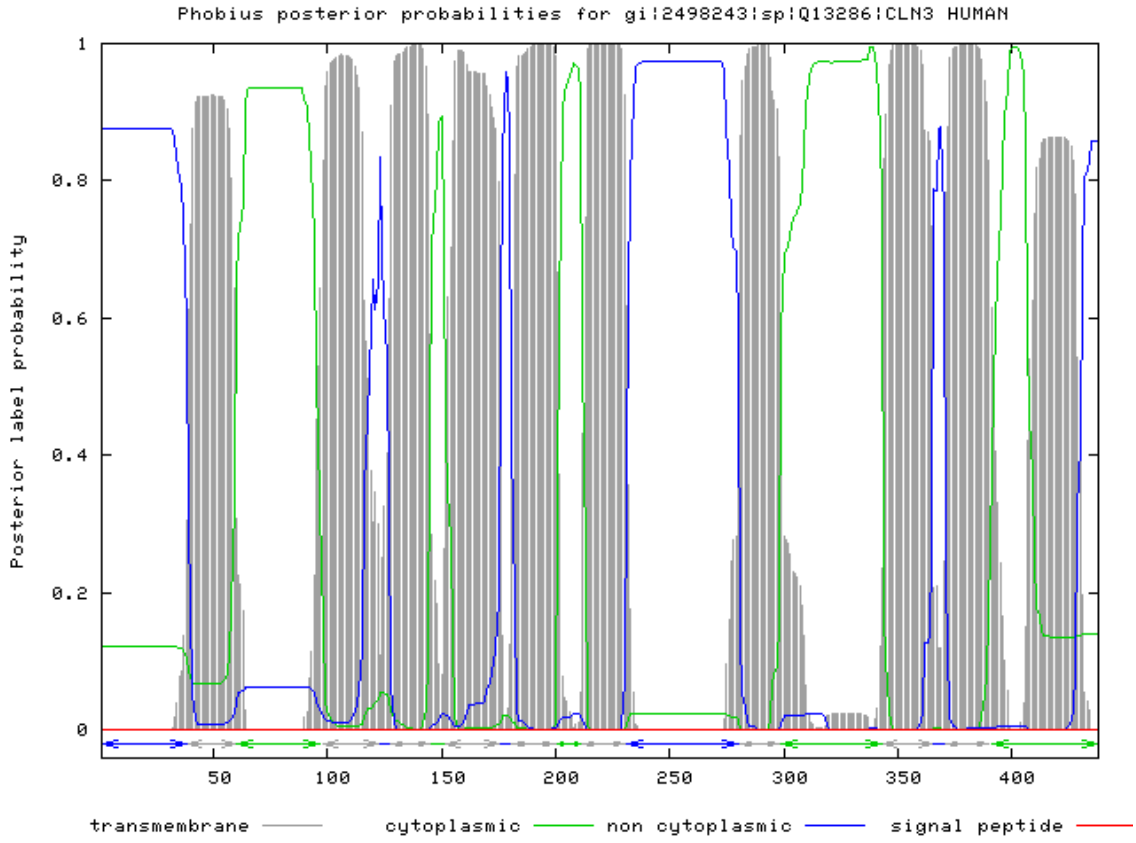

Figure 2.2: Typical graphical output from the Phobius server.

secondary structure and the LIPS (LIPid-facing Surface) server (Adamian & Liang, 2006) was used to predict helix-lipid interfaces. LIPS is based on a canonical model of the heptad repeat originally developed for coiled coils. Is uses an empirical scoring function which combines lipophilicity and conservation of residues in the helix.

2.3 Results

2.3.1 A topology model for human CLN3

A range of different models were produced with between six and eleven TM spanning helices (Figure 2.3). Despite this variation, there are four distinct regions where there is a strong consensus between all prediction methods: amino acid residues 36-60, 97-121, 210-231 and 276-303. A multi-species CLN3 alignment of 50 sequences shows that all four regions are enriched with hydrophobic residues yet show distinct sequence variation - two features which are entirely consistent with

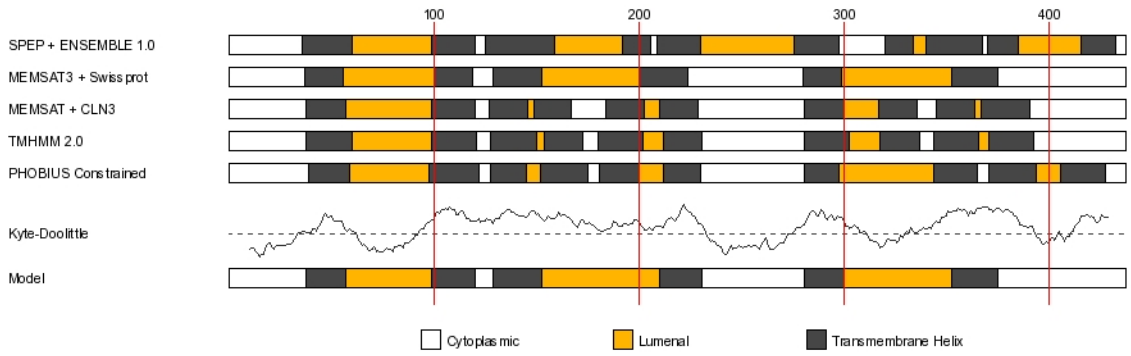

Figure 2.3: Results of topology prediction for CLN3 showing models with cytoplasmic amino terminals and between six and eleven TM spanning helices generated using six different methods, and our consensus prediction that takes into account additional information discussed within the text. The Kyte-Doolittle plot was generated using a 19 residue sliding window.

lipid-exposed membrane-spanning helices.

Of the remaining regions, we ruled out the presence of a TM helix in a number of cases. We believe the region 122-209 - a highly hydrophobic stretch but with no clear peaks of hydrophobicity - contains only one TM helix at about 129-153, roughly in-line with the MEMSAT3 (using SWISS-PROT) prediction. A helix in this position allows for a short loop region after the second helix, and leaves the highly conserved residues at positions 159-195 exposed on a loop. The multi-species alignment of CLN3 orthologues indicates this loop contains insertions in a number of species which casts doubt over the additional TM helices predicted for this region by Phobius and the TMHMM-based methods.

At position 319-336 there is a discrepancy in the consensus prediction, with half the methods predicting a TM helix and half predicting a loop. The Kyte-Doolittle plot indicates this region has relatively low hydrophobicity and it is in fact enriched with polar residues making it an unlikely candidate for a TM helix. A PSIPRED secondary structure prediction does however show a high helix forming propensity for this stretch, and a bias in hydrophobic residue phasing to one side is indicated on construction of a helical wheel (Figure 2.4). This leads us to believe that the region may form an amphipathic helix partially buried in the membrane, a model which is further strengthened by a high lipid exposure score (10.810) for the

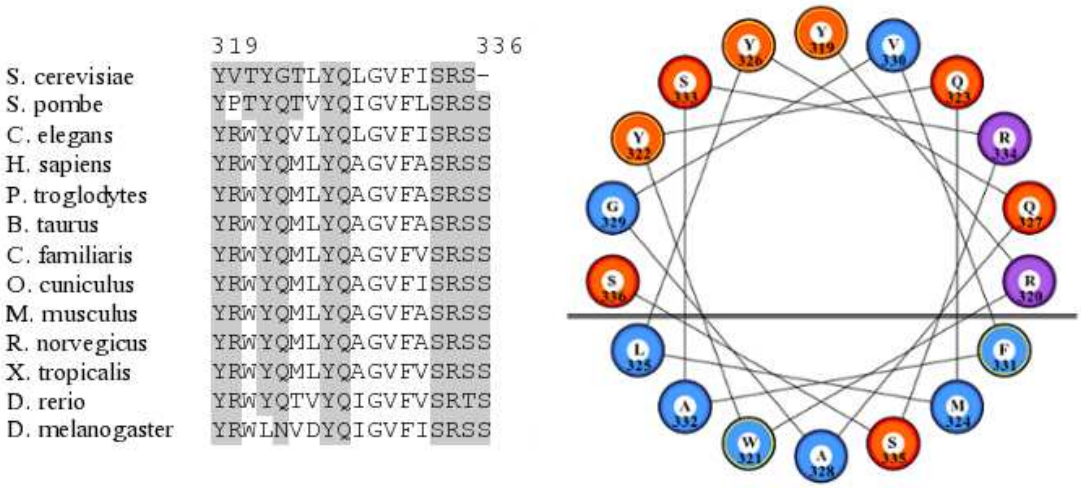

Figure 2.4: Sequence comparison of the potential amphipathic helix from selected species. Numbers refer to amino-acid position in human CLN3. The helix is presented as a helical wheel with hydrophobic residues shaded in light grey and polar residues in dark grey. The area below the line illustrates the position of the membrane.

buried surface (using the LIPS Server). This orientation would result in the highly conserved residues Tyr326, Gln327, Gy329, Val330, Ser333 and Arg334 facing into the lumen away from the membrane and thus free to interact with potential binding partners. The two helix surfaces that contain these residues also score lowest in terms of lipid exposure (2.558 and 2.735).

Between residues 340 and 393 - another highly hydrophobic region with no clear peaks of hydrophobicity - the general consensus is that there are two TM helices connected with a very short loop region. However, we are inclined to accept the MEMSAT3 (using SWISS-PROT) prediction of one helix spanning 353-375 which unifies the two predicted by other methods. A single helix in this position allows the highly conserved flanking residues to be positioned in loop regions, as is usual, whereas the two helix predictions would place both these areas inside the membrane.

Finally, we rule out the last helix, present in four of the six models, spanning 406-433. This region contains two distinctive lysosomal sorting motifs described by Kyttälä *et al.* (2004) that experimentally must reside in the cytosol.

We thus propose a six TM topology model with cytoplasmic N and C-termini,

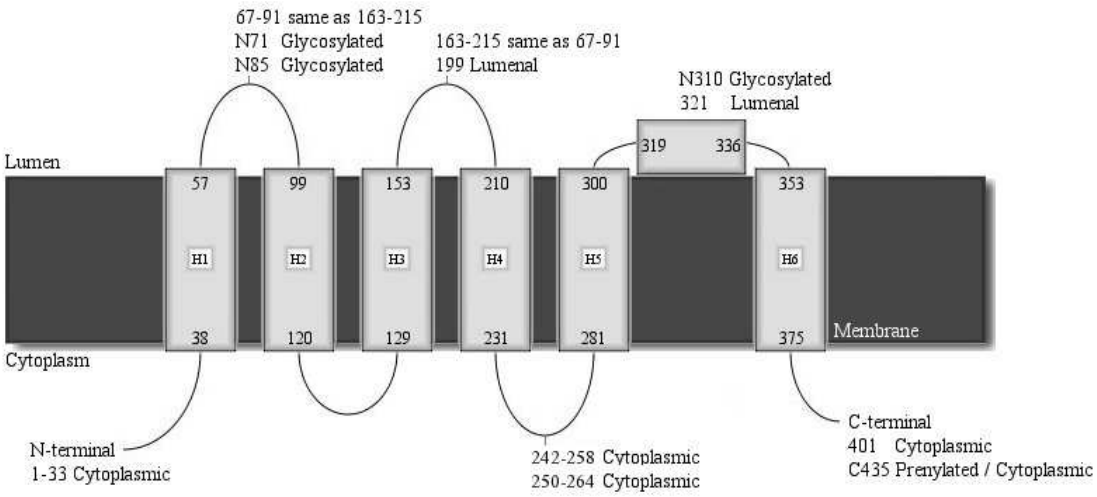

Figure 2.5: Schematic model for human CLN3 showing the six TM helices, proposed amphipathic helix and experimentally determined loop locations (see Table 2.1).

three large lumenal loops, one of which may contain an amphipathic helix, and one large cytoplasmic loop. Our model is derived by applying the latest computational approaches for topological predictions of TM proteins using their primary sequence, and is constrained in only two positions on the basis of reliable experimental data (the cytoplasmic location of the N terminus and the presence of a cytoplasmic trafficking motif at the C terminus). Importantly, this model (Figure 2.5) is supported by almost all experimentally determined loop locations (Table 2.1). The exception is the report in Mao *et al.* (2003a) that the N-terminus might be lumenal, in contrast to data from Kyttälä *et al.* (2004) and (Ezaki *et al.*, 2003). This possible lumenal location for the N-terminus was suggested by (1) inability to immunoprecipitate translated CLN3 from microsomes using antisera that recognised the N-terminus of CLN3 and (2) glycosylation of CLN3 even when Asn310 was mutated (the remaining putative glycosylation sites are Asn49, Asn71 and Asn85). We cannot explain why the immunoprecipitation of CLN3 translated by microsomes did not occur. However we can suggest that, provided Asn71 or Asn85 is glycosylated as later shown by Storch *et al.* (2007), Asn 49 does not have to be lumenal.

2.3.2 Topology models for *Schizosaccharomyces pombe* and *Saccharomyces cerevisiae*

Additionally, we attempted to construct models of two diverse yeast CLN3 sequences (Sipiczki, 1995), *Schizosaccharomyces pombe* and *Saccharomyces cerevisiae*, using 40 microbial CLN3 sequences to create PSI-BLAST profiles for MEMSAT3. In contrast to our human CLN3 model, we found a strong consensus for up to 11 TMH between all the predictors for both species (Figure 2.6). One of these TM helices (helix 8) corresponds to the predicted amphipathic helix (Figure 2.4), suggesting a 10 TMH model that also contains an amphipathic helix for these yeast species (Figure 2.7). Aligning each of the 10 TM helices against the human CLN3 sequence using the Smith-Waterman local sequence alignment algorithm (Smith & Waterman, 1981) resulted in scores greater than 30% sequence identity between only the 1st, 2nd and 5th helices of our human model and the 1st, 2nd and 7th helices of the yeast models. However, helices 4 and 5 of the yeast model contain the most highly conserved residues between all orthologues (including two residues mutated in disease), making it unlikely that they would be present in a membrane in one species and projecting into the lumen in another. These helices also contain the equivalent residues to those experimentally proven to reside in the lumen in human CLN3 (Table 2.1). Similarly helix 10 contains a conserved residue mutated in disease, and the loop between these helices 9 and 10 is highly conserved suggesting that it should have the same orientation in yeast and mammalian species. Our assumption that conserved regions should have the same topological orientation is consistent with the knowledge that human CLN3 protein can functionally substitute for Btn1p in *Schizosaccharomyces pombe* and *Saccharomyces cerevisiae* (Gachet *et al.*, 2005; Pearce & Sherman, 1998). CLN3 and its orthologues, then, have a topology that is not entirely straightforward to predict using currently available methods. In such a situation, our approach that makes use of all available sequences to reach a consensus prediction is even more appropriate. This also explains the discrepancy between MEMSAT3 + SWISS-PROT and MEMSAT3 + CLN3 models which were constructed using PSI-BLAST profiles

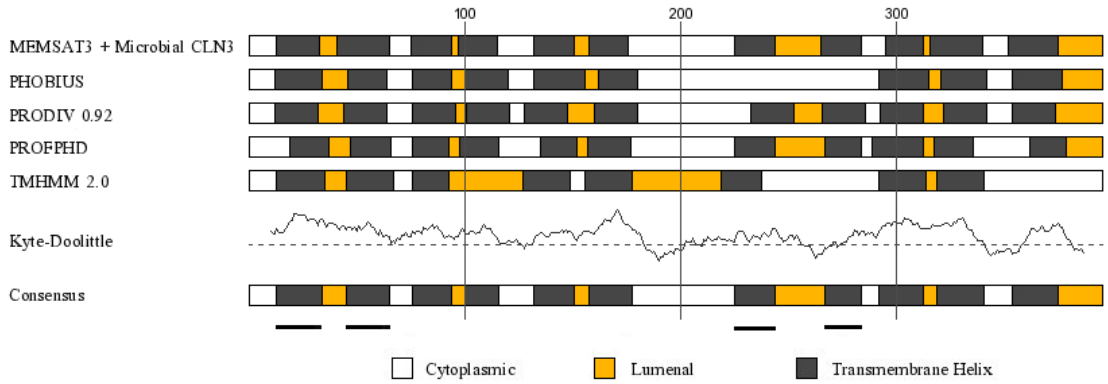

Figure 2.6: Results of topology prediction for *Schizosaccharomyces pombe* Btn1p showing models with cytoplasmic amino terminals and a consensus of eleven TM spanning helices. MEMSAT3 was used in conjunction with a database of 40 microbial CLN3 sequences to construct PSI-BLAST profiles. Predicted helices marked with a dark line are conserved in human and correspond to the 1st, 2nd, 5th and the amphipathic helix of our model.

composed predominantly of mammalian and microbial CLN3 sequences respectively.

The human model contains sixteen positively charged residues in cytoplasmic loops, compared with only eight in lumenal loops, which is consistent with the general observation of a more positively charged cytoplasmic surface - the positive inside rule (von Heijne, 1992). While this is energetically unfavourable, their spatial distribution across four TM helices allows for the formation of ion-pairs between Asp103 (TMH2) and His146 (TMH3), and Lys112 (TMH2) and Glu295 (TMH5) or Asp362 (TMH6). It would also be possible to satisfy these bonding potentials should CLN3 undergo dimerisation. While the formation of either of these salt bridges could help stabilise the structure, they would also impose constraints on the three-dimensional folding of the protein. Further stability could be provided should disulphide bridges form between Cysteine residues which are present on two lumenal loops (although these are not conserved across species), consistent with the expectations for a correct model in which such bridges usually form post-synthesis in the oxidative environment of the endoplasmic reticulum lumen.

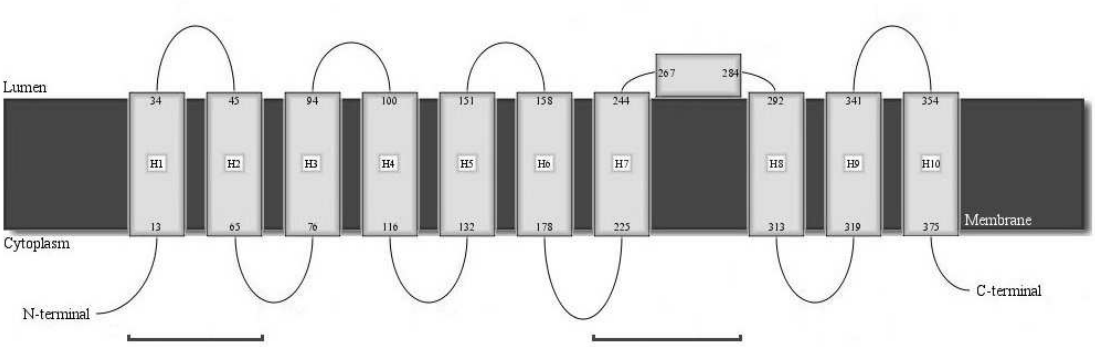

Figure 2.7: Schematic model for *Schizosaccharomyces pombe* Btn1p showing a ten TM spanning model, an amphipathic helix and cytoplasmic N and C-termini. Predicted helices marked with a dark line are conserved in human and correspond to the 1st, 2nd, 5th and the amphipathic helix of our human CLN3 model. Helices 4 and 5 of this model are predicted to project into the lumen in the human CLN3 model, with experimental evidence supporting this, and helices 9 and 10 are predicted to project into the cytoplasm in the human CLN3 model.

2.3.3 Analysis of PROSITE matches

CLN3 has been shown to undergo phosphorylation when incubated with cAMP-dependent protein kinase, cGMP-dependent protein kinase or Casein kinase II (Michalewski *et al.*, 1999), though the specific residues involved are unknown. Using the ScanPROSITE tool, nine potential phosphorylation sites were detected, six on cytoplasmic loops (Ser12, Ser14, Thr19, Thr232, Ser270, Thr400) and three on lumenal loops (Ser69, Ser74, Ser86), with the cytoplasmic signatures showing higher conservation on average than those in the lumen. While the PROSITE phosphorylation signatures offer a large degree of freedom and are thus known to produce high numbers of false positives, the bias in frequency towards cytoplasmic loops can be explained by the observation that the kinases responsible are known to localise exclusively in the cytoplasm (Forrest *et al.*, 2006a).

Of the four potential N-glycosylation sites that were predicted (Asn49, Asn71, Asn85, Asn310), three have had their locations validated experimentally (Asn71, Asn85, Asn310) (Mao *et al.*, 2003a; Storch *et al.*, 2007) and are correctly placed on lumenal loops by our model. This bias in frequency towards lumenal loops may be explained by the ability of glycosylation to prevent proteolysis in the protease-rich lysosomal lumen, thought to be the reason lysosomal membrane proteins are

Region/residue	Location	Reference
N-terminal	Cytoplasmic	Ezaki *et al.* (2003)
1−33	Cytoplasmic	Kyttälä *et al.* (2004)
1−20	Lumenal	Mao *et al.* (2003a), Mao *et al.* (2003b)
71	Lumenal	Storch *et al.* (2007)
67−91	Same as 163-215	Unpublished data cited within Mao *et al.* (2003a)
85	Lumenal	Storch *et al.* (2007)
163−215	Same as 67-91	Unpublished data cited within Mao *et al.* (2003a)
199	Lumenal	Mao *et al.* (2003a)
250−264	Cytoplasmic	Mao *et al.* (2003a), Mao *et al.* (2003b)
242−258	Cytoplasmic	Kyttälä *et al.* (2004)
310	Lumenal	Mao *et al.* (2003a)
321	Lumenal	Mao *et al.* (2003a)
401	Cytoplasmic	Kyttälä *et al.* (2004)
435	Cytoplasmic	Storch *et al.* (2007)
C-terminal	Cytoplasmic	Mao *et al.* (2003a), inferred in Kyttälä *et al.* (2004)

Table 2.1: Locations of experimentally determined regions/positions.

often heavily glycosylated. Both CLN3 and the yeast orthologues traffic to the lysosome/vacuole (Kyttälä *et al.*, 2004; Gachet *et al.*, 2005), although this may not be their only functional location. The remaining potential N-glycosylation motif has been placed within the first TM helix; however, unlike the other three, this site has not been validated experimentally and is most likely a false positive.

2.3.4 Cross-species conservation

CLN3 has orthologues identified in at last 46 diverse eukaryotic species to date. Sequence homology extends from residue 41 of the human CLN3 protein in large stretches to the end of the protein, with certain residues, including most disease-causing missense mutations (Mole, 1998), identical or similar across all species. Our model can be used to examine the topological position of the most conserved residues and regions. Several interesting and striking observations can be made from this. First, we note that the N-terminus, the first lumenal loop and the second cytoplasmic loop of CLN3 are not well conserved across most species, suggesting that these do not contribute directly to the basic function of the protein. Second, in contrast, the second and third lumenal loops and much of the C-terminus are highly conserved

across diverse species, as is the proposed amphipathic helix (contained within the third lumenal loop), suggesting that these regions contribute important structural constraints or domains important for function or, in the case of the C-terminus, trafficking. All known disease-causing missense mutations are located either in or immediately adjacent to a predicted TM helix (Leu101, Glu295, Gln352) or in the conserved lumenal loops (Ala158, Leu170, Gly187) or the amphipathic helix (Val330 and Arg334 which is mutated twice) or the C-terminus (Asp416) consistent with the importance of these regions. Of particular interest are the two conserved lumenal loops/amphipathic helix that may interact with lumenal molecular species such as proteins, carbohydrate moeities or the lipid bilayer, or transduce changes in the lumenal environment (eg pH) to modulate CLN3 activity.

2.3.5 Function prediction

Helices that interact with the lipid bilayer are thought to modulate the activity of many ion channels (Kuo *et al.*, 2003; Enkvetchakul *et al.*, 2007). The proposed amphipathic helix of CLN3 may act similarly and, if so, the local composition of the lipid membrane may influence function. Intriguingly, two methods suggest CLN3 may be involved in transport. FFPred, a protein feature based function prediction method (Lobley *et al.*, 2008), suggests a role in ion transport, while Pfam (Finn *et al.*, 2006) lists CLN3 as a member of the major facilitator superfamily (MFS) clan, again suggesting a possible role as a transporter. Interestingly, the most recently identified NCL gene, MFSD8/CLN7 also encodes a member of the MFS super-family (Siintola *et al.*, 2007). However, careful inspection of global alignments between CLN3 and the 12 TMH members of the MFS family suggest these hits may be false positives, and analysis of CLN3 TM helices fails to identify any which look to be involved in pore formation. Experiments that target residues in these regions and define the phenotypic consequences of mutations may shed light on their role. In addition, the identification of any interacting partners with the lumenal loops will require specialized biochemical approaches, since many commonly used methods (e.g. two-hybrid) are only appropriate for cytoplasmic interactions.

2.4 Discussion

In summary, we propose a six TM helix topology for CLN3, a novel predicted amphipathic helix previously unrecognised, with both termini located in the cytoplasm. While no single topology prediction agrees with our final model, we have shown that a consensus approach combined with careful analysis of evolutionary data can produce a model which agrees with all published experimental data. Previous work had been less confident about the number of possible TM helices, although one model proposed on the basis of experimental findings (Kyttälä *et al.*, 2004) agrees closely with ours. Our unexpected finding that orthologues of CLN3 might produce different topologies may be due either to atypical membrane or hydrophobic structures. Taking into account the location of the conserved residues may help identify the regions critical for structure or function and this can be used to inform topological interpretations. Our approach may have wider applicability in the prediction of the topology of other TM proteins, particularly those containing additional hydrophobic structures that may not be membrane spanning. Determining the correct topology of CLN3 is critical for complete understanding the mechanism of Batten disease. However, until a CLN3 crystal structure can reveal the true TM topology in atomic resolution, likely to be some way beyond current capabilities, a model produced by combining experimental data with topology prediction provides us with one that can be further tested experimentally. Significantly, the presence of a lumenal amphipathic helix and conserved intralumenal domains provides new insight into the possible mechanism of action of CLN3 and its orthologues in model organisms that can also be investigated appropriately.

Chapter 3

Improving topology prediction by topogenic assignment of biologically meaningful sequence motifs

3.1 Background

3.1.1 Topology prediction

3.1.2 Modern topology predictors

The accuracy of modern state-of-the-art predictions methods is currently in the region of 60-80% accuracy (Chen *et al.*, 2002), though this is highly dependent on the assessment dataset and cross-validation strategy. Recently, a novel method MEMSAT3 (Jones, 2007) was described that combined the original MEMSAT (Jones *et al.*, 1994a) approach with an artificial neural network. MEMSAT made use of scores derived from membrane protein data and a dynamic programming algorithm, allowing it to search through all possible topological models by a process of expectation maximisation. Amino acid propensities for each of five states topological states - inside loop, outside loop, inside helix end, helix middle and outside helix end - were calculated from TM proteins with experimentally determined topologies and were expressed as a log-likelihood ratio. This approach allowed MEMSAT to calculate the most probable length, location and topological orientation for each TM helix, therefore returning a list of all possible topologies ranked by overall likelihood and thus guaranteeing a mathematically optimal solution, rather than simply deciding between a limited number of possible topologies. MEMSAT3 replaced the log-likelihood rations with scores from a neural network trained on sequence profiles generated using PSI-BLAST, allowing it to utilise sequence conservation information that has proved powerful in other applications, for example the PSIPRED secondary structure prediction method (Jones, 1999). When benchmarked with full cross-validation on a data set of 184 transmembrane proteins, MEMSAT3 was able to predict the correct topology for 80% of the test set, compared with accuracies of 62-72% for other recent methods on the same benchmark.

3.1.3 Improving topology prediction using experimental constraints

While this level of accuracy represents a significant advance over existing methods, there is still substantial scope for improvement. A number of studies have demonstrated that incorporating additional information into topology prediction can increase accuracy. Tusndy and Simon demonstrated using their HMMTOP topology prediction server (Tusnady & Simon, 2001), an HMM-based method, that a theoretical improvement in performance was possible by incorporating experimental information into the topology prediction. Experimental information provided by the user that showed, for example, that the N-terminus of a sequence was localised in the extracellular space, or that certain sequence motifs were expected to be localised in the cytoplasm, was incorporated into the Baum-Welch algorithm by a conditional probability in order to find the unknown parameters of the HMM. Kim *et al.* (2003) used a similar approach to determine the topology of 37 TM proteins from *S. cerevisiae*. A C-terminal fusion to a dual topology reporter was first used to determine the location of the C-terminus of each protein relative to the endoplasmic reticulum membrane before this information was used in conjunction with a topology prediction method to arrive at a final topology model. A subsequent study used the same method to determine the C-terminal locations of a further 617 *S. cerevisiae* proteins and used this information to present experimentally constrained topology models for 546 of them. By applying this information to homologous sequences, the topologies of 15,000 TM proteins from 38 fully sequenced eukaryotic genomes was reported (Kim *et al.*, 2006). A similar study of the *E. coli* inner membrane proteome which used green fluorescent protein to tag C-terminal locations established the periplasmic or cytoplasmic locations of the C termini for 601 inner membrane proteins, from which high-quality topology models were produced (Daley *et al.*, 2005). While such studies are valuable, the availability of experimental data is frequently limited and results can also be conflicting (Mao *et al.*, 2003a; Kyttälä *et al.*, 2004).

3.1.4 Improving topology prediction using domain assignments

Bernsel & Heijne (2005) adopted a more automated approach in improving topology prediction. They identified a set of 367 domains in the SMART database (Letunic *et al.*, 2004) - a database of well-annotated protein domains represented as profile-HMMs - that had compartment-specific localisation when found in soluble proteins, but which were also relevant to TM proteins. Protein domains are modular, independently evolving, and often structurally similar amino acid regions that exist alone or in combination to form multi-domain proteins. Covalent combinations between soluble domains and TM domains are frequently observed, therefore their localisation in soluble proteins can in some cases be transferred to TM proteins. By using the presence of these domains and their inside/outside locations, which was considered to be entirely correct, topology models were produced using PRO-TMHMM (Viklund & Elofsson, 2004) by fixing the domain containing region to the corresponding side of the membrane. Using these constrained predictions, they were able to provide high-quality topology models for 11% of TM proteins extracted from 38 eukaryotic genomes, although two-thirds of these were single spanning TM proteins.

The use of domains to constrain predictions is a powerful approach, and worthy of incorporation into topology prediction servers, particularly for whole genome studies. However, in most cases, the detection of a domain will indicate that protein already has a well characterised homologue in a database and it may be possible to transfer the topology from the well characterised protein to the unknown sequence should sequence identity be high enough. This approach would however be unable to enhance topology prediction accuracy if presented with a sequence without known homologues, and therefore without matching domains. To constrain prediction of unknown sequences we cannot rely on protein family-specific identifiers, but need more general sequence motifs or features.

3.1.5 The PROSITE database

In this study we explore the possibility that shorter, biologically meaningful sequence motifs that demonstrate a topogenic preference may be used to constrain topology predictions, resulting in an increase in prediction performance. Specifically, we intend to use sequence motifs from the PROSITE database (Falquet *et al.*, 2002) that are not family-specific, therefore allowing the approach to be applicable to unknown sequences without obvious homologues. PROSITE is an annotated collection of motif descriptors, represented as either patterns or profiles, which are derived from multiple alignments of homologous sequences. This gives the descriptors the advantage of identifying distant relationships between sequences that may not have been detected based on a pairwise sequence alignment (Sigrist *et al.*, 2002). Relationships can be revealed by the presence of a cluster of residues, also known as a pattern, motif, signature or fingerprint, typically between 10 and 20 amino acids in length that are involved in an important biological function and are therefore conserved in both structure and sequence during evolution. Such biologically significant regions include enzyme catalytic sites, prosthetic group attachment sites (heme, biotin etc.), amino acids involved in metal ion binding, Cysteine residues involved in the formation of disulphide bonds, and regions involved in the binding of other molecules (ADP/ATP, DNA etc.) or other proteins.

As the sequence of motifs is conserved, it is possible to reduce a multiple alignment of them to a consensus pattern known as a regular expression, where each position in such a pattern can be occupied by any residue from a specified set of acceptable residues and can be repeated a variable number of times within a specified range. Strictly conserved positions may only allow a particular residue, while at other positions, residues with similar physicochemical properties can be acceptable, while specific incompatible residues are excluded. Finally, conserved residues can be separated by gaps of variable lengths. The resulting expression can then be used to scan unknown sequences, a process that can be performed quickly on a modern computer. However, matching a regular expression with a sequence is a qualitative

process in that there is either a match or there is not - there is no threshold or score associated with the presence of the motif above which a match is classed as statistically significant. A potential caveat of this approach is thus the possibility of high levels of false positive matches. However, the accuracy of PROSITE patterns has been evaluated using the number of hits obtained while scanning the manually curated SWISS-PROT database (Boeckmann *et al.*, 2003) and other randomised databases. Whenever a new motif descriptor is added to PROSITE, it is used to scan SWISS-PROT in order to attribute a status match to the SWISS-PROT entry for true positive, false positive, false negative, unknown (proteins that could belong to the set considered by the motif) and partial (proteins belonging to the set being considered but not detected by the motif) cases. These statistics, which are available for each motif, allow the user to assess the sensitivity and specificity of each regular expression, and additionally allow the motifs to be improved with each release of SWISS-PROT.

3.1.6 Using PROSITE to guide topology prediction

In this chapter I will describe a strategy to use PROSITE motifs to guide TM topology prediction. The method utilises PROSITE motifs that display a bias towards a particular topogenic region in TM proteins. To identify this bias, I have assembled a novel high quality data set of TM proteins that have crystal structures available and have scanned the corresponding sequences for PROSITE motifs and assigned matches to topogenic regions. Motifs were identified that occur in a specific topogenic region with significantly different frequencies compared to those expected at random. I will describe the modification of MEMSAT3 so that topology predictions can be constrained or guided in such a way that the resulting topology models are scored more highly if they satisfy the topogenic biases of the matching motifs, and therefore increase overall topology prediction accuracy.

3.2 Methods

3.2.1 Assembling a novel data set of transmembrane proteins

The novel data set was based on crystal structure data. Additional information was collected from MPTOPO (Jayasinghe *et al.*, 2001), OPM (Lomize *et al.*, 2006b), PDB_TM (Tusnady *et al.*, 2005a), SWISS-PROT (Boeckmann *et al.*, 2003) and from the literature. SWISS-PROT files were parsed for entries containing the keyword 'TRANSMEM' in feature table (FT) lines. N-terminal data was also extracted using keyword 'TOPO_DOM' where available. To avoid partial sequences being included, any entries containing keywords such as 'FRAGMENT' were excluded. Sequences were then scanned against the PDB in order to identify entries for which the TM region had complete structural coverage. Alignments occasionally highlighted chain breaks. In these cases, the sequence was excluded unless a visual inspection ensured the topology could not be cast in doubt by the break. This left a redundant data set containing 944 sequences which was then homology reduced at the 40% sequence identity level. A number of sequences were then removed. These included colicins (e.g. PDB: 1COL) and bee venom (2MLT) which are bilayer disrupting and thus are not native integral membrane proteins, sequences labelled as 'secreted protein', and sequences where the N-terminal location or topology could not be verified.

OPM was then used to define TM helix boundaries, or in the absence of an OPM entry, PDB_TM was used. OPM uses a theoretical multi-feature approach to position proteins in a membrane which has been shown to be in good agreement with experimental studies of 24 TM proteins. In some cases where a visual inspection appeared to indicate an incorrect placement of the membrane, PDB_TM helix boundary definitions were used instead. For example, OPM lists Mechanosensitive channel protein MscS (2OAU) as having two TM helices, neither of which fully cross the membrane, whereas the PDB_TM definition of 3 TM helices appears more plausible (Figure 3.1).

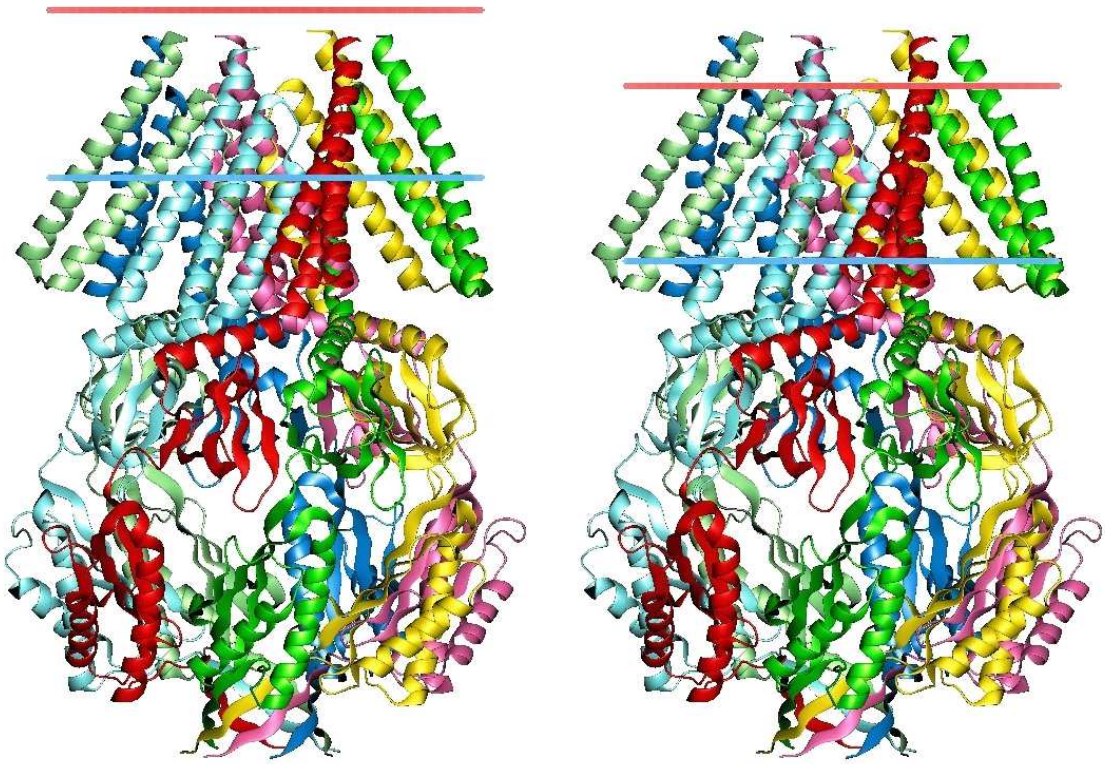

Figure 3.1: Theoretical membrane placement on to the Mechanosensitive channel protein MscS crystal structure (PDB: 2OAU) by OPM (left) and PDB_TM (right). The membrane region is between the red and blue bars. PDB_TM uses hydrophobicity and structural feature analysis to determine the position of the membrane (Tusnady *et al.*, 2005b)

PDB_TM was also used to annotate proteins containing re-entrant helices. A re-entrant helix was defined as a helix-containing region that enters and exits the membrane on the same side, penetrating at least 6 Å but not more than 6 Å from the opposite membrane face. Re-entrant regions that did not contain a helix formed by at least three contiguous amino acid residues were excluded. Sequences containing signal peptides were then labelled according to SWISS-PROT annotations.

The composition of the final data set containing 131 sequences, all with available crystal structures, verifiable topology and N-terminal locations, is show in Table 3.1. The full list of sequences and topologies can be found in Appendix B.

Protein class	Number in set
Prokaryotic	92
Eukaryotic	37
Viral	2
Single-spanning TM segment	57
Multiple-spanning TM segments	74
Contains re-entrant helix	11
Contains signal peptide	14
Total	131

Table 3.1: Crystal structure data set composition.

3.2.2 Identification of PROSITE matches and their respective topogenic biases

All sequences in the novel data set were then scanned using the ScanProsite tool (Gattiker *et al.*, 2002) against release 20.9 of the PROSITE database. Where a motif was detected, the topology definition was used to assign it to inside and outside loops, TM helices and regions containing both loop and TM helices ('multiple'). A χ^2 test for independence was then used to identify motifs which were not evenly distributed between inside and outside loops. With one degree of freedom, a χ^2 value above 0.82 indicates a bias with 95% confidence.

3.2.3 Modification of MEMSAT3 to incorporate PROSITE motif matches

We analysed the sequences from the Möller data set (Möller *et al.*, 2000) whose topologies were incorrectly predicted by MEMSAT3, identifying 24/184 cases where the placement of all TM helices was correct but the location of the N-terminus was incorrect. Initially, we attempted to constrain the predictions in order to satisfy the topological biases of the PROSITE motifs that were matched to each of these sequences, therefore assuming that the biases were correct. This was achieved by filtering the results to remove all topologies that did not satisfy these constraints. However, this proved infeasible as there were a number of instances

where it was impossible to satisfy the biases of all motifs simultaneously, suggesting some matches were false positives or that our assumption that their bias towards a topological region was incorrect. We therefore tried using each motif separately to avoid such scenarios. For the majority of motifs, it was possible to reduce the numbers of incorrect predictions where the helices were correctly predicted but the N-terminus location was incorrect to below 24; however, in all cases, the total number of correctly predicted topologies was reduced as previously correct topologies were rendered incorrect as, presumably due to false positive matches, TM helices were falsely constrained to inside or outside loop regions.

We therefore attempted to modify the topogenic propensities, initially generated using a NN, that the MEMSAT3 dynamic programming algorithm uses to generate topologies. For each residue in the target sequence, a NN score is generated reflecting the likelihood that amino acid resides in an inside loop, outside loop, TM helix or signal peptide region. For every PROSITE motif that expressed a topogenic bias, we used three weights to modify the inside loop, outside loop and TM helix scores. Where the resulting score was below 0 or above 1, the scores were set to 0 and 1 respectively. To optimise these weights, we employed a genetic algorithm (GA, see below). In order to avoid optimising these weights specifically for the Möller data set, we randomly split the set into two halves, then used the GA to determine the set of weights that resulted in the highest topology prediction accuracy for each split before averaging these to determine the final weights.

3.2.4 Genetic algorithms

Genetic algorithms are a class of adaptive heuristic search algorithm often used to find exact or approximate solutions to optimisation problems. GAs are based on techniques inspired by evolutionary biology including inheritance, mutation, selection, and crossover, and therefore represent the intelligent exploitation of a random search within a defined search space in order to solve a problem. GAs are implemented as computer simulations in which an abstract population, typically

chromosomes, evolves towards an optimal solution. The simulation begins with a population of randomly generated individuals which is then evaluated, and the fitness of each individual is assessed. Multiple individuals are stochastically selected from the current population based on their fitness, and are modified by recombination or random mutation to form the subsequent population. The new population is then used in the next iteration of the algorithm, and this process continues until a termination condition is reached (Gondro & Kinghorn, 2007).

Once the search problem has a defined genetic representation and a fitness function - in our case this will be a function to assess topology performance using weights to modify topogenic propensities - the GA is initialised and a population of individual solutions - the list of weights - is generated in order to cover the entire search space. These are usually limited to a defined type and range; our weights were floating point values between -1 and 1. With each successive generation, a proportion of the existing population is then selected to breed a new generation. The fitness of each solution is determined using the fitness function, with fitter solutions more likely to be selected. Different implementations may select only the best solutions, while others may rate only a random sample of the population in order to reduce computation time. Functions are usually stochastic and designed to ensure that a small proportion of less fit solutions are selected; this keep the diversity of the population large, preventing premature convergence towards poor solutions.

A successive population of solutions is generated through crossover (recombination) and mutation. New solutions are produced by a pair of parent solutions from the pool previously selected. Characteristics of each parent are passed on to the child solution through crossover, analogous to the biological crossover of chromosomes, and mutation, analogous to biological mutation. This process continues until a new population of solutions of the appropriate size is generated, and since the fitter individuals are selected for breeding, this successive population

will usually have a higher average fitness than the previous one.

This generational process is then repeated until a termination condition has been reached. This is typically defined by the population reaching a satisfactory fitness level, or by a maximum number of generations, although in the latter case the optimal fitness level may not have been reached. The highest ranking solution's fitness may have reached a plateau, such that successive iterations no longer produce significantly better results, or a computational or time limit may have been reached (Mitchell, 1996).

3.3 Results

3.3.1 PROSITE motifs that express a topogenic bias

Table 3.2 shows the nine PROSITE motifs that were identified as having a topogenic bias with 95% confidence. Myristoylation, the most prevalent motif in our data set, is a post-translational protein modification in which myristic acid is covalently attached to the alpha-amino group of an N-terminal Glycine residue via an amide bond. Myristoylation is known to influence the conformational stability of individual proteins, as well as their ability to interact with various membranes or the hydrophobic domains of other proteins. Myristic acid is able to loosely tether the modified protein to the plasma membrane, endoplasmic reticulum, mitochondrion, or other membrane system, possibly allowing interaction with other proteins localised nearby (Podell & Gribskov, 2004). The PROSITE N-myristoylation site is still widely used, despite the fact that the signature has not been has not been updated since 1989, and is known to produce a large number both false positive and false negative predictions. The main reason for the inaccuracy is that the amino acid choices at each position are fairly broad; as a result, only two of the five positions described are actually restrictive (Table 3.3). More recent data also indicates that residues downstream from the initial five can also influence myristoylation site suitability (Maurer-Stroh et al., 2002). Finally,

only a small number of myristoylated sequences were actually used to construct the signature. Despite these factors, the motif appears to display a significant bias towards outside loops compared to inside loops.

Casein kinase II, protein kinase c and cAMP- and cGMP-dependent protein kinase are all involved in the phosphorylation of a wide range of different proteins. Phosphorylation is a ubiquitous cellular regulatory mechanism that occurs through the reversible addition of phosphate groups from ATP to various amino acid residues. Phosphorylation of proteins is an essential regulatory mechanism that occurs in both prokaryotic and eukaryotic organisms. Many enzymes and receptors are activated and deactivated by phosphorylation and dephosphorylation as a result of the conformational change induced in the structure that occurs upon modification. Phosphorylation usually occurs on Ser, Thr, and Tyr residues in eukaryotic proteins, while often occurring on the basic amino acid residues His or Arg or Lys in prokaryotic proteins (Cortay *et al.*, 1991; Stock *et al.*, 1989).

Again, the PROSITE phosphorylation signatures offer a large degree of freedom. Interestingly though, all three motifs show a significant bias towards inside loops which can be explained by the observation that the kinases responsible for phophorylation are known to localise exclusively in the cytoplasm (Forrest *et al.*, 2006a). In the case of protein kinase c and cAMP- and cGMP-dependent protein kinase phosphorylation sites, the signature contains positively charged Arg and Lys residues so it is possible that this observation is due to the positive-inside rule.

Glycosylation involves the addition of saccharides to proteins and lipids in order to produce glycans, and is a principal post-translational modification in the synthesis of membrane and secreted proteins. N-linked glycosylation is known to be important for the folding of some eukaryotic proteins. The process occurs in eukaryotes and widely in archaea, but very rarely in prokaryotes. It involves the addition of a 14-sugar precursor, containing 3 glucose, 9 mannose, and 2

N-acetylglucosamine molecules, to the Asn in the polypeptide chain of the target protein. Since glycosylated residues are known to be involved in cellular recognition (Rudd *et al.*, 1999), it is interesting that the motif occurs more frequently on the inside loops of TM proteins. Amidation sites generally function as active peptide precursor cleavage sites, though they are also noted to have a high probability of occurrence as all amino acids can be amidated (Kreil, 1984). Our data set indicates a preference for an inside loops bias. The leucine zipper signature is another unspecific pattern that is unlikely to be a true motif unless the protein has been shown to bind DNA, a function in which TM proteins are rarely known to play a role (Landschulz *et al.*, 1988).

The final two motifs identified as significant are both family-specific identifiers, so will not be included in the study as sequences that contain these motifs will easily be identifiable using a homology search. The major intrinsic protein (MIP) signature is used to identify a family of highly related TM channel proteins (Pao *et al.*, 1991). These include mammalian aquaporins, water-specific channels that provide the plasma membranes of red blood cells and kidney proximal and collecting tubules with high permeability to water, thereby permitting water to move in the direction of an osmotic gradient (Chrispeels & Agre, 1994). MIP family proteins seem to contain six transmembrane segments, with the signature pattern mapping to a well conserved region which is located in a probable cytoplasmic loop between the second and third TM regions. Cytochrome c oxidase is an oligomeric enzymatic complex which is a component of the respiratory chain and is involved in the transfer of electrons from cytochrome c to oxygen. In eukaryotes this enzyme complex is located in the mitochondrial inner membrane while is is found in the plasma membrane in aerobic prokaryotes (Capaldi *et al.*, 1980). The signature targets the copper ligands at the centre of the complex.

PROSITE ID	Description	Inside	Outside	Helix	Multiple	n	χ^2
PS00008	N-myristoylation site	20.23%	32.11%	44.82%	2.68%	598	16.11
PS00006	Casein kinase II phosphorylation site	53.97%	38.41%	5.96%	1.66%	302	7.92
PS00005	Protein kinase c phosphorylation site	63.08%	29.23%	7.69%	0.00%	260	32.27
PS00001	N-glycosylation site	49.51%	33.01%	13.59%	3.88%	103	3.40
PS00004	cAMP- and cGMP-dependent protein kinase phosphorylation site	67.86%	28.57%	3.57%	0.00%	28	4.48
PS00009	Amidation site	85.71%	7.14%	0.00%	7.14%	14	9.31
PS00029	Leucine Zipper	27.27%	0.00%	27.27%	45.45%	11	3.00
PS00221	MIP family signature	100.00%	0.00%	0.00%	0.00%	4	4.00
PS50857	Cytochrome c oxidase subunit II signature	0.00%	100.00%	0.00%	0.00%	4	4.00

Table 3.2: PROSITE motifs that were identified as having a topogenic bias. Column 1: PROSITE database identifier. Column 2: Motif description. Column 3: Percentage of matching motifs assigned to inside loop regions. Column 4: Percentage of matching motifs assigned to outside loop regions. Column 5: Percentage of matching motifs assigned to TM helices. Column 6: Percentage of matching motifs that span multiple topogenic regions, e.g. TM helix and loop regions. Column 6: Number of matches identified using crystal structure data set. Column 7: χ^2 value. A threshold of 0.82 indicates significance with 95% confidence.

PROSITE ID	Signature
PS00008	G - {EDRKHPFYW} - x(2) - [STAGCN] - {P} [G is the N - myristoylation site]
PS00006	[ST] - x(2) - [DE] [S or T is the phosphorylation site]
PS00005	[ST] - x - [RK] [S or T is the phosphorylation site]
PS00001	N - {P} - [ST] - {P}[N is the glycosylation site]
PS00004	[RK](2) - x - [ST] [S or T is the phosphorylation site]
PS00009	x - G - [RK] - [RK] [x is the amidation site]
PS00029	L - x(6) - L - x(6) - L - x(6) - L
PS00221	[HNQA] - {D} - N - P - [STA] - [LIVMF] - [ST] - [LIVMF] - [GSTAFY]
PS50857	V-x-H-x(33,40)-C-x(3)-C-x(3)-H-x(2)-M [The C's and H's are copper ligands]

Table 3.3: PROSITE motif signatures. Square brackets [] indicate any of the enclosed amino acids are acceptable in that position. Curly brackets {} indicate any amino acids except those enclosed are acceptable in that position. x indicates any amino acid is acceptable. Parentheses (n) indicates the n copies of the preceding amino acid must be present. As an example, the signature [AC]-x-V-x(4)-{ED} can be translated as [Ala or Cys]-any-Val-any-any-any-any-{any but Glu or Asp}.

3.3.2 Topogenic propensity weights generated using the genetic algorithm

Table 3.4 shows the the weights generated using the GA that, when used to modify the original NN scores for all residues in the sequence covered by the motif, lead to optimum topology prediction performance on the Möller data set. The GA detected convergence and therefore terminated after approximately 1250 iterations. Of the seven PROSITE motifs that were identified as having a topogenic bias and were not family-specific signatures, all of them apart from the N-myristoylation site (PS00008) displayed a bias towards inside loop regions. In fact the N-myristoylation appeared most frequently not on outside loops but on TM helices. The weights generated by the GA are therefore slightly surprising since they do not tend to reflect this inside loop bias. The N-myristoylation site weights increases both loop scores by approximately the same amount, though only increase the TM helix score by 0.175 therefore regions containing this signature are likely to be guided to either inside or outside loops, but not to TM helices. Modifications to the two phosphorylation site motifs that contain positively charged residues, protein kinase c and cAMP- and cGMP-dependent protein kinase (PS00004 and PS00005), are somewhat conflicting. As expected by the positive inside rule, the PS00004 weight increase the inside loop score by approximately 1, and decrease the outside loop by the same amount, inevitably guiding this region to the inside. For PS00005, the inside loop score is increased only very slightly, while the outside loop score is significantly increased by 0.658, unexpectedly guiding this region to the outside. The remaining Casein kinase II phosphorylation site (PS00006) is guided towards a TM helix since both loop region scores are significantly reduced, the inside loop more so, while the TM helix score is reduced only very slightly. The N-glycosylation and leucine zipper motifs (PS00001 and PS00029) weights do favour inside loops as expected, although they do both increase the TM helix score slightly. Finally, the amidation site (PS00009) increases the outside loop score by 1 while reducing the inside loop and TM helix scores, therefore guiding the region to the outside.

In summary, the weights of three motifs corresponded with their topogenic biases (PS00001, PS00004, PS00029) and four do not - two of these guide topology towards the outside rather than the inside as expected (PS00005, PS00009); one guides topology towards a TM helix (PS00006, although a slight preference towards an outside loop compared to an inside loop) while the remaining motif guides topology towards either loop region rather than a TM helix (PS00008).

PROSITE ID	Inside modification	Outside modification	TM helix modification
PS00008	0.518	0.533	0.175
PS00006	-1.000	-0.669	-0.026
PS00005	0.069	0.658	0.031
PS00001	1.000	0.135	0.702
PS00004	0.956	-1.000	-0.172
PS00009	-0.288	1.000	-0.396
PS00029	0.870	0.572	0.298

Table 3.4: Weights generated using the GA used to modify the original NN scores. These weight are added to all residues covered by the matching signature. Where the modified score was below 0 or above 1, the score was set to 0 and 1 respectively.

3.3.3 Topology prediction performance against the Möller data set using PROSITE motif weights

Table 3.5 shows the topology prediction performance against the Möller data set, updated using recent SWISS-PROT annotations, with and without the modification of topogenic propensities using PROSITE motif weights. Results are fully cross-validated, with all proteins homologous to the target being removed from the respective training set. For an overall correct topology prediction (column 7), the correct number of TM helices needed to be predicted (column 2), their locations needed to overlap their observed positions by at least 5 residues (column 3) and the N-terminal had to be localised to the correct side of the membrane (column 4). By modifying the topogenic propensities, overall topology prediction performance was increased from 77.2% to 83.2%, an increase of 6% corresponding to the correct prediction of 11 additional sequences.

In 3 of these 11 cases, the number and locations of TM helices is correct but the original MEMSAT3 prediction places the N-terminal on the wrong side of the membrane. In each case, the PROSITE weights cause the alternate N-terminal location topology to score higher, largely due to the cAMP- and cGMP-dependent protein kinase motif (PS00005) guiding topology towards the outside (CVAA_ECOLI, GEF_ECOLI), while in one case the N-glycosylation site motif (PS00001) guides topology towards the inside (EBR_STAAU) (Figure 3.2).

In a number of cases, the occurrence of motifs on regions where TM helices were incorrectly predicted guides prediction of the matching region to a loop. This accounts for the 4.3% increase in the number of sequences with the correct helix count (88.0% compared to 83.7%) and the reduction in the number of falsely predicted TM helices (0.0% compared to 3.8%). Unfortunately, many of these topologies are still incorrect due to inaccurate locations of TM helices or the N-terminal.

The largest category of sequences whose topologies are corrected using PROSITE motifs is that where MEMSAT3 falsely predicts helices as loop regions, affecting 7 sequences (ADT2_YEAST, FTSH_ECOLI, HLYB_ECOLI, KDPD_ECOLI, PMA1_NEUCR, STE6_YEAST and UPKB_BOVIN). In doing so, the loop region that immediately follows the position of the mis-predicted helix is inevitably correct. In these cases, the modification of topogenic propensities by PROSITE motif weights is sufficient to correct the topology of this loop region, and therefore induce prediction of the missing TM helix/helices (Figure 3.3).

In one case, the N-terminal localisation and TM helix count were correct but the single TM helix was incorrectly positioned (VG1_BPFD). A number of loop inducing motifs that matched the incorrect TM helix reduced this topology's score, resulting in an alternative topology which did have the TM helix correctly positioned being the highest scoring.

Method	Correct TMH count	Correct TMH locations	Correct N-terminus	FP helix	FN helix	Correct topology
MEMSAT3	83.7%	82.1%	87.0%	3.8%	12.5%	142/184
MEMSAT3 + PROSITE	88.0%	87.0%	89.7%	0.0%	12.0%	153/184

Table 3.5: Topology prediction performance against the Möller data set, with and without modification of topogenic propensities using PROSITE motif weights. Column 1: Prediction method. Column 2: Correct TMH count - Fraction of sequences with the correct number of TM helices predicted. Column 3: Correct TMH locations - Fraction of sequences with the correct number and locations of TM helices predicted. Column 4: Correct N-terminus - Fraction of sequences with the correct N-terminal location predicted. Column 5: FP helix - Fraction of sequences with at least one over predicted TM helix. Column 6: FN helix - Fraction of sequences with at least one under predicted TM helix. Column 7: Correct topology: Fraction of sequences that have correct overall topology predicted, requiring the correct number and location of TM helices and correct location of the N-terminal. TM helices must overlap their defined positions by at least 5 residues.

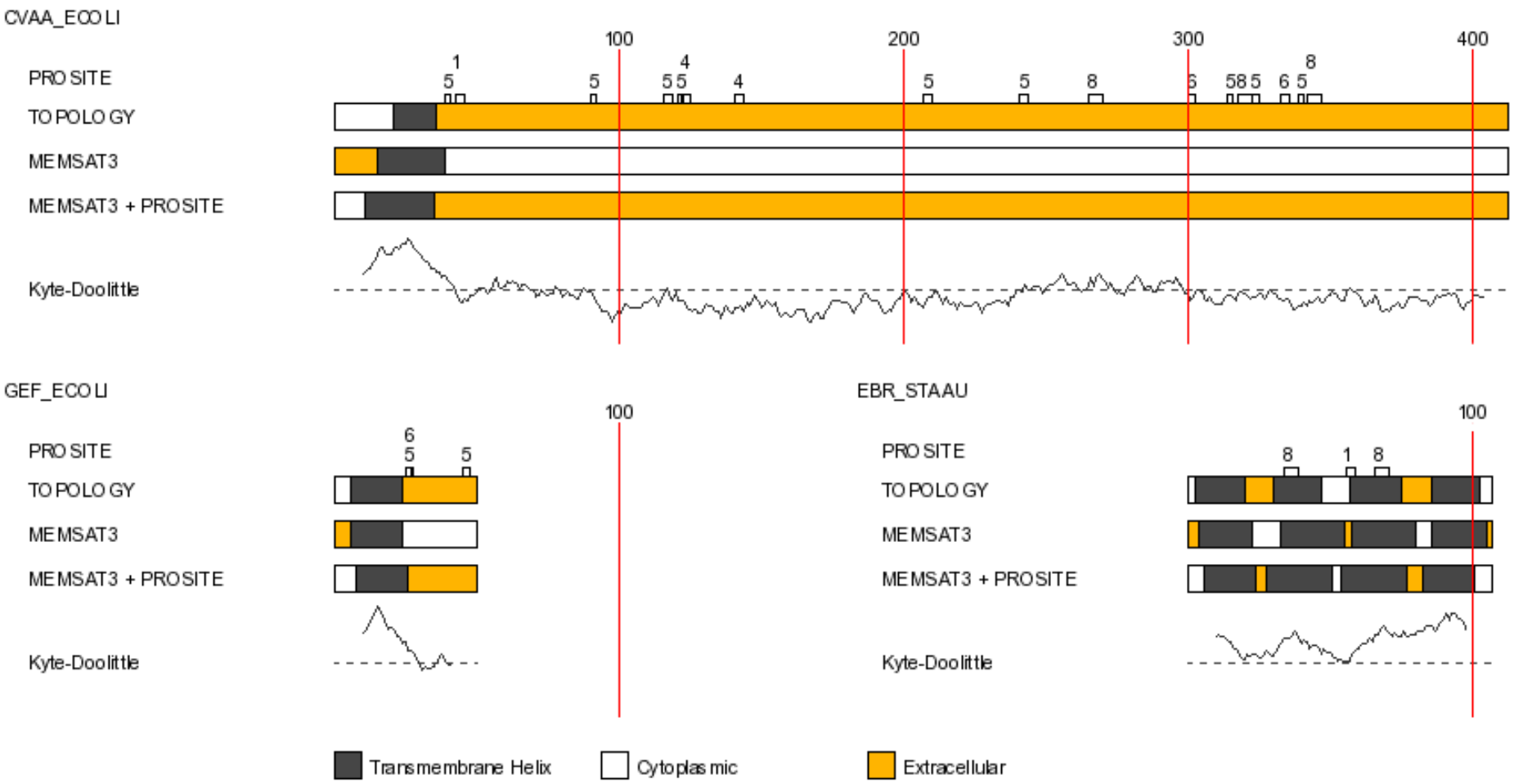

Figure 3.2: Topology predictions corrected by altering N-terminal localisation. cAMP- and cGMP-dependent protein kinase motif (PS00005) guides topology towards the outside (CVAA_ECOLI, GEF_ECOLI), while an N-glycosylation site motif (PS00001) guides topology towards the inside (EBR_STAAU) . PROSITE: PROSITE motif identifier (PS0000X). TOPOLOGY: known topology. MEMSAT3: original MEMSAT3 topology prediction. MEMSAT3+PROSITE: PROSITE motif modified MEMSAT3 topology prediction.

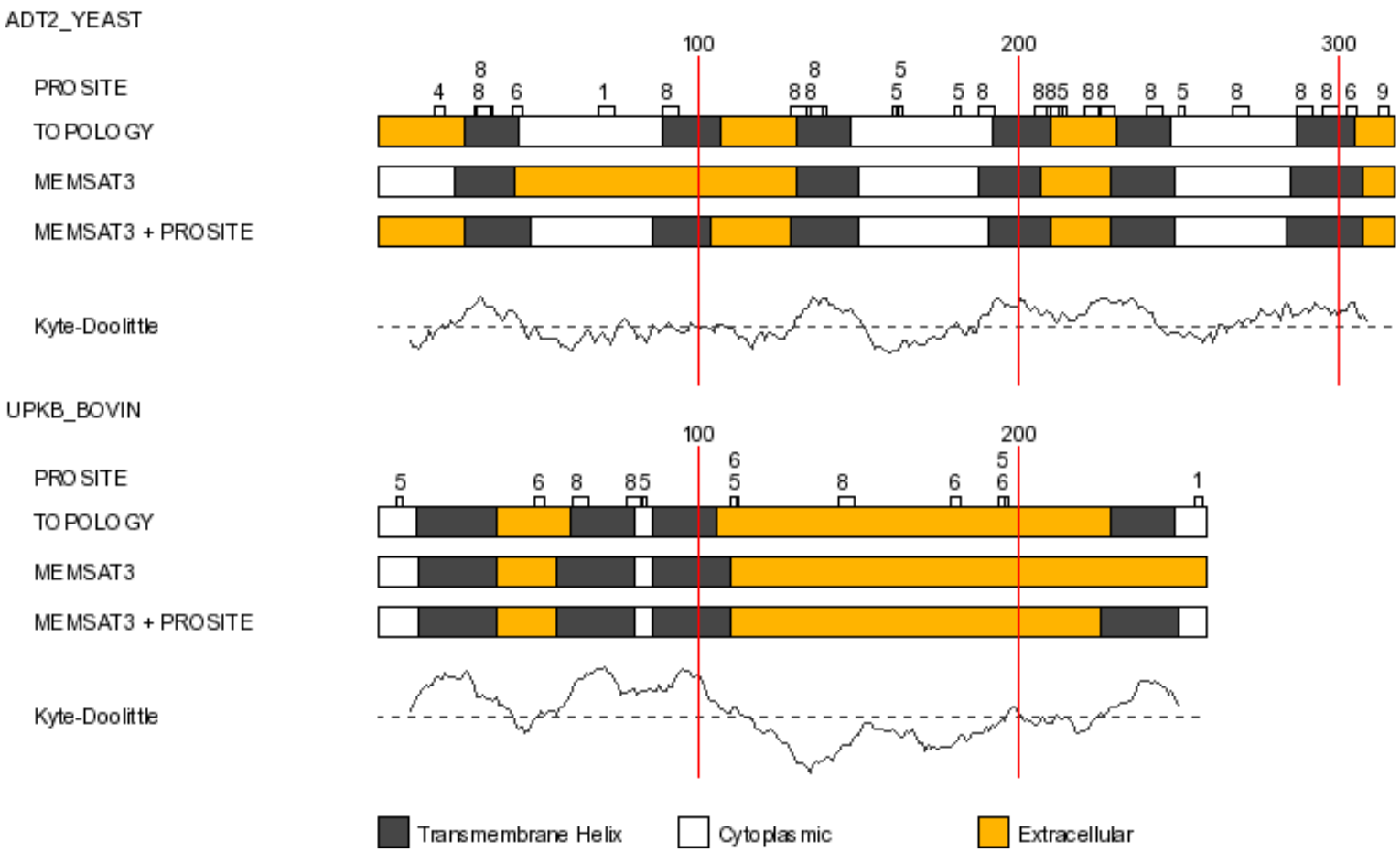

Figure 3.3: Topology predictions corrected by prediction of a TM helix where a loop region was previously predicted. N-glycosylation site motifs (PS00001) guide topology towards the inside therefore inducing prediction of a TM helix at residues 85-103 (ADT2_YEAST) and at residues 226-250 (UPKB_BOVIN). PROSITE: PROSITE motif identifier (PS0000X). TOPOLOGY: known topology. MEMSAT3: original MEMSAT3 topology prediction. MEMSAT3+PROSITE: PROSITE motif modified MEMSAT3 topology prediction.

3.4 Discussion

Previous work has shown that TM protein topology prediction can be improved if constraints are used to fix specific regions to a particular side of the membrane, prior to running a predictor (Kim *et al.*, 2003; Bernsel & Heijne, 2005; Daley *et al.*, 2005). While existing approaches have made use of experimental data or family-specific domain assignments, here we present a method that uses biologically meaningful sequence motifs that are not family-specific in order to guide topology prediction. We have identified seven PROSITE motifs which display a significant topogenic bias, and have used a GA to generate weights for each such that the topogenic propensities for each residue guide the MEMSAT3 prediction towards the correct topology. Using the standard Möller data set we have demonstrated that this approach can lead to an improvement of 6% prediction accuracy compared to standard MEMSAT3 performance, corresponding the correct prediction of an additional 11 sequences.

While this level of improvement is significant, there are a number of peculiarities affecting the method, principally the lack of correlation between the direction of the topogenic biases designated by the χ^2 significance test, and the sign and magnitude of the corresponding weights determined using the GA. The χ^2 test indicated an inside loop bias for six motifs and an outside bias for one. The weightings correspond to this bias for three motifs - the N-glycosylation site motif, the protein kinase c phosphorylation motif and the leucine zipper motif. However, the remaining four motifs weights do not correspond with the bias that was indicated. The cAMP- and cGMP-dependent protein kinase phosphorylation site and amidation site weights strongly favour outside loops as opposed to inside loops, the Casein kinase II phosphorylation site weights favour a TM helix as opposed to an inside loop, and the N-myristoylation site weights favour either loop region rather than a TM helix, when a preference for an outside loop was indicated. The most likely reason for this discrepancy is that these four motifs have matched a large number of false positive hits which have affected the χ^2 test, resulting in

incorrect topogenic biases.

Looking at the weights associated with these four motifs, the N-myristoylation site does not appear to favour either loop region in particular, and also has a very broad signature in which only two of the five positions described are actually restrictive, increasing the likelihood of false positive matches. Interestingly, the Casein kinase II phosphorylation site motif signature contains negatively charged residues which the weights direct topology towards TM helices. While charged residues within the TM region are usually energetically unfavourable unless involved in salt bridge formation, a number of studies have observed that negatively charges in the TM domain are conserved in various proteins including potassium channels, where they may contribute to voltage-sensing during the activation process (Liu *et al.*, 2003a), the neuronal alpha 7 acetylcholine receptor where they increase permeability to divalent cations (Ferrer-Montiel & Montal, 1993) and ubiquinone oxidoreductase where they determine its dependence on YidC for membrane insertion (Price & Driessen, 2010). The relatively broad signature may therefore be matching such conserved residues as false positive phosphorylation sites. Surprisingly, both the cAMP- and cGMP-dependent protein kinase phosphorylation and amidation site motif weights direct topology towards outside loops while the signatures contains positively charged residues, which the positive inside rule dictates should enrich inside loops. While this is unexpected, it is the overall ratio of charged residues on inside loops compared to outside loops that is significant, so as long as there are a greater number of positively charged residues on inside loops the positive inside rule will still hold. Additionally, only a small number of matches were found (28 and 14 respectively) using the crystal structure data set and the χ^2 values is relatively small (4.48) in the phosphorylation site. Again, both signatures are broad indicating a high likelihood of false positive hits.

Despite the discrepancy between the χ^2 test and the generated weightings, and thus the likely high false positive rate of PROSITE matches, a significant improve-

ment in topology prediction is still possible. With hindsight, an alternative approach may have been to include all non family-specific PROSITE motif in the GA optimisation stage without testing for χ^2 significance. While the optimisation time would have been significantly extended, the inconsistency would have been avoided while the sign and magnitude of the weightings would have reflected the likelihood of false positive matches. Another approach would be to use machine learning-based approaches to detect motifs. A number of tools now exist that use HMMs and NNs to identify specific motifs (Julenius *et al.*, 2005; Huang *et al.*, 2005), with improved performance over regular expression based methods. While applying these tools instead of using PROSITE motifs may have resulted in lower false positive rates, they are currently limited to the detection of mammalian mucin-type O-glycosylation sites and protein kinase-specific phosphorylation sites. Should gold standard data sets of a broader range of motifs emerge in future, it may be possible to develop tools that can detect a wider range of PROSITE-like motifs using machine learning approaches rather than regular expressions - the incorporation of such methods into future topology predictors is likely to yield increased performance. In conclusion, the use of PROSITE motifs to guide topology prediction results in improved accuracy by providing additional information not fully captured by the NN employed by MEMSAT3. However, the motifs' broad signatures and high false positive rates suggest that caution should be used when interpreting these PROSITE matches as they may be identifying conserved residues that are accounted for by the positive inside rule or play a role in alternative biological functions.

Chapter 4

Transmembrane protein topology prediction using support vector machines

4.1 Background

4.1.1 Machine learning approaches for topology prediction

Machine learning techniques are increasingly being used to address problems in bioinformatics. Novel computational techniques based on machine learning that have been used to analyse data derived from DNA and protein sequences, microarray experiments, pathways, and images are now vital for understanding diseases and the development of novel therapies. Alogrithms including Hidden Markov models (HMMs) (Krogh *et al.*, 2001), neural networks (NNs) (Jacoboni *et al.*, 2001) and support vector machines (SVMs) (Park *et al.*, 2005) have shown great success in analysing data generated by the life sciences because of their ability to generalise while handling both noise and randomness (Zhang & Rajapakse, 2008). Early topology prediction methods, based on the physicochemical principle of a sliding window of hydrophobicity combined with the 'positive-inside' rule (von Heijne, 1992), have therfore been superseded by machine learning approaches include HMMs, NNs and more recently, SVMs.

Perhaps due to their ability to produce multiple outputs, NN and HMM-based approaches for topology prediction have proved both popular and successful over recent years. SVMs however are predominantly binary classifiers therefore multiple SVMs must be employed to classify the numerous residue preferences before being combined into a probabilistic framework. Like NNs and HMMs, SVMs are capable of learning complex relationships among the amino acids within a given window with which they are trained, particularly when provided with evolutionary information. SVMs also have a number of advantages over other machine learning methods; while NNs can encounter multiple local minima, the solution to an SVM is global and unique. They are also considered more resilient to the problem of overtraining compared to other approaches. These benefits may be in part due to the way in which SVMs were developed. Rather than following a heuristic path, from application and extensive experimentation to theory, as in the case of NNs, SVMs

were developed in the reverse order, having evolved from sound theory through to implementation and experiments (Wang, 2005).

4.1.2 Signal peptides, amphipathic helices, and re-entrant helices

One problem faced by modern topology predictors is the discrimination between TM helices and other features composed largely of hydrophobic residues. These include targeting motifs such as signal peptides and signal anchors, amphipathic helices, and re-entrant helices - membrane penetrating helices that enter and exit the membrane on the same side - common in many ion channel families. The high similarity between such features and the hydrophobic profile of a TM helix frequently leads to crosstalk between the different types of predictions. Should these elements be predicted as TM helices, the ensuing topology prediction is likely to be corrupted. Some prediction methods, such as SignalP (Bendtsen *et al.*, 2004) and TargetP (Emanuelsson *et al.*, 2007), are effective in identifying signal peptides, and may be used as a pre-filter prior to analysis using a TM topology predictor. Phobius (Käll *et al.*, 2004) uses a HMM to successfully address the problem of signal peptides in TM protein topology prediction, while PolyPhobius (Käll *et al.*, 2005) further increases accuracy by including homology information. Other methods such as TMLOOP (Lasso *et al.*, 2006), TOP-MOD (Viklund *et al.*, 2006) and OCTOPUS (Viklund & Elofsson, 2008) have attempted to identify re-entrant regions, the latter two in combination with a TM topology predictor, but there is significant room for improvement.

TM topology predictors also exist that are able to use experimentally derived information in order to guide topology prediction. With reliable experimental data, prediction accuracy is likely to benefit substantially. Methods include HMM-TM (Bagos *et al.*, 2006), HMMTOP (Tusnady & Simon, 2001) and TMHMMfix (Melen *et al.*, 2003). Tools such as SOSUI (Hirokawa *et al.*, 1998) and PRED-CLASS (Pasquier *et al.*, 2001) are designed to discriminate between globular and TM pro-

teins, while others such as PRED-TMBB (Bagos *et al.*, 2004b) specialise in the discrimination and prediction of beta-barrel TM proteins.

4.1.3 The importance of using high quality data

A key element when constructing any prediction method is the use of a high quality data set for both training and validation purposes. Previously described TM data sets such as the Möller set (Möller *et al.*, 2000) have contained relatively few sequences with structures available, but substantially more with TM region annotation based on varying types of biochemical characterisation. A number of experimental methods, including glycosylation analysis, insertion tags, antibody studies and fusion protein constructs, allow the topological location of a region to be identified. However, such studies are often conflicting (Kyttälä *et al.*, 2004; Mao *et al.*, 2003a) and also risk upsetting the natural topology by altering the protein sequence. As a result, orientation and helix boundary errors in databases are not infrequent and add an element of noise. While such noise is often well tolerated by machine learning methods, the problem is more significant in smaller data sets.

This chapter describes the development of a new TM topology predictor trained and benchmarked with full cross-validation on a novel data set of 131 sequences with crystal structures. The method uses evolutionary information and four SVMs, combining the outputs using a dynamic programming algorithm, to return a list of predicted topologies ranked by overall likelihood, and incorporates signal peptide and re-entrant helix prediction. Overall, the method predicted the correct topology and location of TM helices for 89% of the test set, a significant improvement on our previous NN-based method MEMSAT3 (Jones, 2007). An additional SVM has been trained to discriminate between TM and globular proteins with zero false positives and a low false negative rate of 0.4%, making this method highly suitable for whole genome analysis.

4.2 Methods

4.2.1 Support vector machine training

As SVMs are binary classifiers, we chose to combine multiple SVMs to classify each of the residue preferences found in TM proteins. Although multiclass ranking SVMs do exist, they are generally considered unreliable since in many cases no single mathematical function exists to separate all classes of data from one another (Abe, 1998). We therefore trained four SVMs to classify TM helix/¬TM helix, inside loop/outside loop, re-entrant helix/¬re-entrant helix and signal peptide/¬signal peptide. Residue labelling was performed according to our data set definitions.

For SVM training and cross-validation, we used the crystal structure data set described in Chapter 3. PSI-BLAST (Altschul *et al.*, 1997) was used to generate position-specific scoring matrices for each of the proteins in the data set using the UniRef 90 database (Boutet *et al.*, 2007). Two iterations were performed with a profile-inclusion E-value threshold of 0.001 in order to reduce false positive hits, to which TM proteins are more prone than globular proteins (Hedman *et al.*, 2002). The E-value, or expectation value, is a parameter that describes the number of hits one can expect to see by chance when searching a database of a particular size. It decreases exponentially with the S-score, the raw alignment score calculated as the sum of substitution and gap scores using matrices such as PAM and BLOSUM that is assigned to a match between two sequences. Essentially, the E-value describes the random background noise that exists for matches between sequences and is used as a convenient way to create a significance threshold for reporting results. When the E-value is increased, a larger list with more low-scoring hits will be reported, while a lower E-value will result in a shorter list containing more quality hits. For each residue in a sequence, a sliding window approach was used to create a feature vector of length 20 x W, where W is the size of the window centred on the target residue. Where the window extended beyond the protein termini, empty feature values were set to zero. All values for each feature position where then normalised by Z-score

(Equation 4.1) to enable faster SVM convergence. Initial attempts at scaling values between 0 and 1 had resulted in lower overall prediction accuracy.

$$z = \frac{(x - \mu)}{\sigma} \tag{4.1}$$

Equation 4.1: x is the raw score to be normalised. μ and σ are the mean and standard deviation PSI-BLAST scores for each of the 20 amino acid, generated using profiles for all 131 sequences.

In order to accentuate the contribution of re-entrant helices for which data is particularly sparse, the sequences of 64 proteins, all homologous to the 11 re-entrant helix-containing sequences in our initial data set, were also used to train the TM helix/¬TM helix and re-entrant helix/¬re-entrant helix SVMs. Helix, loop and re-entrant helix boundaries were determined by PDB_TM definitions.

We also attempted to train the TM helix/¬TM helix SVM using unlabelled data via transduction. In transduction, the learning task is to assign labels to unlabelled data as accurately as possible (Chen *et al.*, 2003). SVMs can perform transduction by finding the hyperplane that maximises the margin relative to both the labelled and unlabelled data, in order to improve the generalisation performance. We selected sequences from SWISS-PROT identified by the MEMSAT3 TM/globular protein discriminator as TM proteins. Sequences with greater than 40% sequence identity to sequences in the labelled data set were removed, as were those with signal peptides predicted by SignalP. Of those remaining, 135 sequences were used as unlabelled training data.

For training the signal peptide/¬signal peptide SVM, we included data from the Phobius training set which contains 2654 well annotated examples of TM and globular proteins, with and without signal peptides. This was supplemented by a search of SWISS-PROT for sequences labelled with the keyword 'SIGNAL' (but excluding entries labelled 'POTENTIAL' or 'BY SIMILARITY') to add to the signal peptide set, and sequences without keyword 'SIGNAL' to add to the

non-signal peptide set. The combined set was then homology reduced at the 40% sequence identity level, leaving 3205 (1222 with signal peptides, 1983 without signal peptides) sequences for which PSI-BLAST profiles were then generated as outlined above.

Stringent cross validation was performed using a jack knife test (leave-one-out cross validation) for the TM helix/¬TM helix, inside loop/outside loop and re-entrant helix/¬re-entrant helix SVMs. In training, the target sequence, along with any other sequences with greater than 25% sequence identity, were excluded. For the signal peptide/¬signal peptide SVM we used 10-fold cross validation, again excluding sequences from the training set with greater than 25% sequence identity to any sequence in the test set. For training and classification, SVM-Light (Joachims, 1998) was used. The performance of several kernels was investigated in combination with a comprehensive grid search of SVM parameters.

Parameters which had the greatest influence on performance were the C-parameter, the γ parameter of the radial basis function (RBF) kernel and the degree of the polynomial kernel (d-parameter). The C-parameter controls the trade-off between the margin and the size of the slack variables; a low value gives a soft margin while a higher value leads to a hard margin. Adjusting the value of the C-parameter between 0 and 10^6 typically resulted in best performance. In the case of the RBF kernel, γ determines the RBF width; typically a value close to 0.1 was used.

To determine optimal windows sizes, the data set was split randomly into two and the highest scoring window which ranked equally in each split was selected, therefore demonstrating consistency between data sets and reducing the risk of overfitting. We used the MCC to optimise these values which is a more robust measure than using recall or precision alone (Matthews, 1975).

To calculate a list of topologies ranked by overall likelihood, the TM helix/¬TM

helix, inside loop/outside loop and signal peptide/¬signal peptide raw SVM outputs were combined in a modified version of the dynamic programming algorithm used in the original MEMSAT method (Jones *et al.*, 1994a). Dynamic programming is a method of solving complex problems by breaking them down into simpler steps, applicable to cases that consist of overlapping subproblems. MEMSAT3 replaced the log liklihood ratios used by MEMSAT with NN scores for each residue to generate a score for TM helices and the preceding loop segment at each position in the sequence. By defining the minimum and maximum lengths of loops and TM helices, a matrix could be filled with scores for TM helix and loop segments at all possible positions. By traversing the matrix, viable topologies and their corresponding scores could be generated, which were then ranked according to the score generated by summing the individual helix-loop segment scores. The MEMSAT3 algorithm was simplified slightly by treating TM helices as discrete units, rather than separating them into inside, outside and middle components, though a signal peptide state was added. Loop regions between predicted TM helices were scanned for re-entrant helices using the re-entrant helix/¬re-entrant helix raw SVM output and a simple scoring function. For evaluating signal peptide preference, residues with positive signal peptide scores up to position 40 in a target sequence were added to the outside loop score and subtracted from the inside loops score where positive, in order to direct prediction towards a non-cytoplasmic amino-terminus. The value was also scaled by a factor of 10 and subtracted from the TM helix SVM score to prevent TM helix prediction. Residues were therefore predicted to lie in one of five different topological regions: inside loop, outside loop, TM helix, re-entrant helix and signal peptide.

To evaluate performance, four metrics were used. Firstly, correct location of the amino terminus; secondly, correct number of TM helices; thirdly, correct number and location of TM helices (based on an overlap of at least five residues with the helix boundaries in our data set) and fourthly, correct overall topology. For comparison, we also evaluated a number of other leading topology predictors. For

this method and MEMSAT3, the appropriate cross-validated training data was used in assessing performance. Where equivalent data was unavailable for the other methods, performance is likely to be overestimated as it is likely that there is significant overlap between test and training sets. We also assessed performance of the method against proteins containing signal peptides and re-entrant helices.

We also trained an additional SVM to discriminate between TM and globular proteins, to be used as a pre-filter prior to TM topology prediction. For SVM training, we used the data set of 131 TM proteins and 416 globular proteins from non-redundant PDB chains as used by MEMSAT3. To accurately compare with MEMSAT3 we used exactly the same test set consisting of 184 TM proteins from the Möller data set and a separate set of 2269 non-redundant globular protein chains, giving a total of 2453 test cases. PSI-BLAST profiles were generated for all sequences and 10-fold cross validation was used to assess performance, again removing sequences from the training fold with greater than 25% sequence identity to any sequence in the test fold.

For whole genome analysis, ten genomes - nine eukaryotic and one prokaryotic - were downloaded from the Ensembl (Flicek *et al.*, 2008) and NCBI (Benson *et al.*, 2008) websites. Protein sequences were extracted and PSI-BLAST profiles were generated using the SWISS-PROT database. The TM/globular predictor was used to identify TM proteins, which were then subject to full topology prediction.

4.3 Results

4.3.1 Support vector machine performance

Table 4.1 shows the per residue performance of each of the five SVMs used by the method. The TM helix/¬TM helix SVM performs significantly better than the re-entrant helix/¬re-entrant helix and inside loop/outside loop SVMs, and slightly better than the signal peptide/¬signal peptide and TM protein/globular

SVM	Window size	Kernel	MCC
TM Helix/¬TM Helix	33	RBF	0.80
Inside Loop/Outside Loop	35	Polynomial*	0.63
Re-entrant Helix/¬Re-entrant Helix	27	RBF	0.34
Signal Peptide/¬Signal Peptide	27	RBF	0.76
TM Protein/Globular Protein	33	RBF	0.78

Table 4.1: Per residue SVM performance. Column 1: SVM type. Column 2: Window size - the size of the sliding window in residues. Column 3: Kernel - SVM kernel type. RBF = radial basis function. Column 4: MCC - Matthews correlation coefficient. * The Inside Loop/Outside Loop SVM was trained using a third-order polynomial kernel.

protein SVMs, reflecting the relative ease with which the hydrophobic signal of a TM helix is detected compared to sequence features within the other topological regions. The Matthews correlation coefficient (MCC, see Appendix C) value of 0.80 compares favourably with the equivalent value of 0.76 achieved by MEMSAT3 using a NN when cross-validated against the same test set. We found that the inclusion of unlabelled data for transductive learning led to a slightly lower MCC of 0.77, in addition to increasing training time, and thus parameter optimisation time, substantially. As a result we excluded unlabelled data when training the final model.

The inside loop/outside loop SVM was the only SVM to perform optimally using a polynomial kernel, which justifies our use of multiple SVMs to classify each of the residue preferences rather than a single multiclass ranking SVM. The highest MCC value we could achieve using a radial basis function (RBF) kernel for this SVM was 0.35, significantly lower than the value of 0.63 achieved using a third-order polynomial kernel, therefore demonstrating that no single kernel function is capable of optimally separating all the data classes and suggesting the structure of loop data is strongly favoured by this kernel.

Detection of re-entrant helices remains challenging compared to other regions, with lack of training data a significant issue. Despite the addition of 64 proteins to the training set, all were homologous to one of the original 11 re-entrant helix-containing proteins and were therefore removed from the respective training files. Their contribution was therefore reflected by a low false positive rate of 0.008,

but a low true positive rate of 0.478 owing to the lack of positive training examples.

In contrast, the signal peptide/¬signal peptide and TM protein/globular protein SVM performance was close to that of the TM helix/¬TM helix SVM, aided by sufficient quantities of training data. While largely driven by hydrophobicity, the signal peptide/¬signal peptide SVM must accurately discriminate between signal peptides, which contain a 7-15 residue long hydrophobic helix, and an equally hydrophobic but slightly longer TM helix. For all signal peptide-containing proteins, the residues ranked highest by the SVM appear to be close to the C-terminal end of the signal peptide region, suggesting the SVM is efficiently detecting the polar and uncharged 3-8 amino acid residue long C-region and the neutral residues that lie adjacent to the cleavage point (von Heijne, 1983). Similarly, the TM protein/globular protein SVM must discriminate between hydrophobic residues that compose TM helices and those that form the core of globular proteins, a challenge reflected by the difference in MCC compared to the TM helix/¬TM helix SVM.

4.3.2 Overall topology prediction accuracy

Table 4.2 shows the overall topology prediction accuracy when applying the method to the test set of 131 TM proteins, alongside results for a number of other recent topology predictors. MEMSAT-SVM and MEMSAT3 results are fully cross-validated as described above, with all proteins homologous to the target being removed from training sets, while results for the remaining methods were obtained from their respective web servers and consequently are not cross-validated. OCTOPUS was also trained exclusively using proteins with crystal structures available, of which 121 sequences (92%) are present in the test set, therefore results are likely to be significantly overestimated.

Method	Algorithm	Correct HC	Correct locations	Correct N-terminal	FP helix	FN helix	Correct SP	Correct RE	Correct Topology
MEMSAT-SVM	SVM	95%	91%	91%	4%	5%	93%	64%	89%
OCTOPUS	NN + HMM	86%	83%	84%	14%	2%	21%	73%	79%
MEMSAT3	NN	84%	76%	84%	8%	8%	57%	64%	76%
ENSEMBLE	NN + HMM	77%	76%	79%	18%	5%	7%	55%	67%
PHOBIUS	HMM	75%	76%	79%	9%	16%	93%	36%	63%
HMMTOP	HMM	77%	76%	78%	18%	6%	29%	64%	63%
PRODIV	HMM	79%	64%	76%	19%	8%	0%	18%	57%
SVMTOP	SVM	66%	64%	66%	22%	22%	0%	55%	53%
TMHMM	HMM	75%	68%	72%	14%	20%	29%	55%	53%
PHDhtm	NN	75%	54%	55%	23%	30%	29%	18%	45%

Table 4.2: Benchmark results for the SVM-based method ('MEMSAT-SVM') against a selection of leading topology predictors. Column 1: Method - Prediction method. Column 2: Algorithm - Underlying machine-learning algorithm. Column 3: Correct HC - Fraction of sequences with the correct number of TM helices predicted. Column 4: Correct locations - Fraction of sequences with the correct number and locations of TM helices predicted. Column 5: Correct N-terminal - Fraction of sequences with the correct N-terminal location predicted. Column 6: FP helix - Fraction of sequences with at least one over predicted TM helix. Column 7: FN helix - Fraction of sequences with at least one under predicted TM helix. Column 8: Correct SP: Fraction of sequences that contain signal peptides that have correct overall topology predicted. Column 9: Correct RE: Fraction of sequences that contain re-entrant helices that have correct overall topology predicted. Column 10: Correct topology: Fraction of sequences that have correct overall topology predicted, requiring the correct number and location of TM helices and correct location of the N-terminal. TM helices must overlap their defined positions by at least 5 residues.

To assess overall topology prediction accuracy, correct prediction of 3 components were required: the N-terminal location, number of TM helices and TM helix locations, based on an overlap of at least 5 residues with boundary definitions. Correct signal peptide and re-entrant helix predictions were not required for a correct overall topology prediction, though failure to predict these features was likely to result in an incorrect topology. Based on this definition, MEMSAT-SVM correctly predicts topology in 89% (116 out of 131) of cases, a 10% improvement on OCTOPUS which predicted 79% (103) of cases correctly (column 10). Using a more stringent criterion of a 10-residue helix overlap, the margin increases to 11% (MEMSAT-SVM 87%, OCTOPUS 76%), suggesting good segment end point prediction. In terms of the 3 individual components, MEMSAT-SVM is consistently better than all other methods (columns 3-5), and in particular performs well at predicting the correct number of TM helices (95% accuracy). MEMSAT-SVM also had a balanced number of over- and under predictions (columns 6-7) which is favourable to avoid bias towards either type of prediction, and suggests good sensitivity while avoiding over predicting helices. Since this work was completed, an extension to the OCTOPUS method which incorporates signal peptide prediction, SPOCTOPUS (Viklund *et al.*, 2008), has been released. This method achieved 87% accuracy on the test set, largely addressing the poor performance of OCTOPUS on sequences containing signal peptides (column 8). An example of the graphical output for a correct prediction is shown in Figure 4.1.

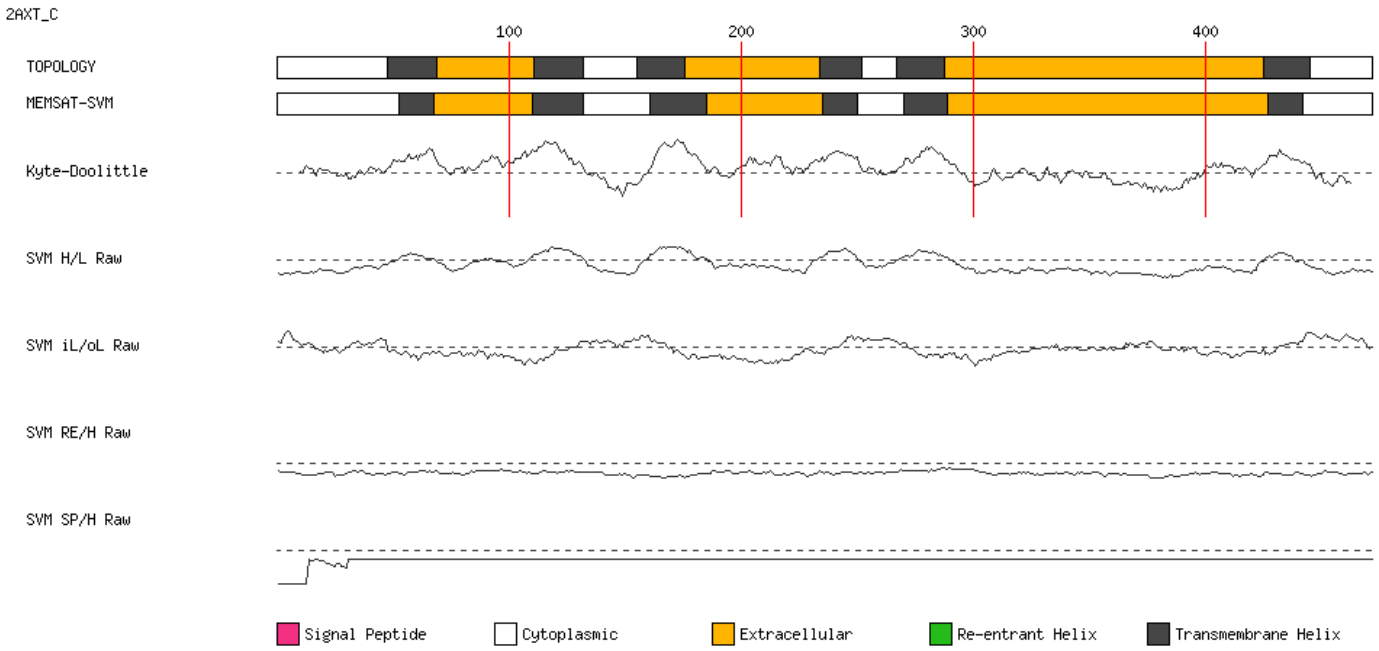

Figure 4.1: Correct topology prediction for Photosystem II chain C from *Thermosynechococcus elongatus* (PDB: 2AXT:C), showing a 6 TM helix prediction with an intracellular N-terminus. The known topology is shown in the track labelled 'Topology' while the prediction is shown in the track labelled 'MEMSAT-SVM'. Beneath this is a Kyte-Doolittle hydropathy plot generated using a window size of 19 residues. The four SVM tracks show the raw SVM score with the dotted line indicating a score of zero. H/L: TM helix/¬TM helix SVM. iL/oL: Inside loop/Outside loop SVM. RE/H: Re-entrant helix/¬Re-entrant helix SVM. SP/H: Signal peptide/¬Signal peptide SVM

4.3.3 Signal peptide and re-entrant helix prediction

MEMSAT-SVM correctly predicts the topology of 93% (13 out of 14) of proteins which contain signal peptides, a substantial improvement on the limited signal peptide prediction capability of our previous method MEMSAT3 (Figure 4.2). In all 13 cases, signal peptides were also predicted. This accuracy is matched by PHOBIUS, the only other method that is specially trained to identify signal peptides in TM proteins. Amongst proteins that did not contain signal peptides, no false positive signal peptides were predicted. Proteins containing re-entrant helices proved much harder to predict, with only 64% (7 out of 11) correctly predicted (Figure 4.3). This is matched by MEMSAT3 and HMMTOP, though is slightly lower than the 73% (8) accuracy achieved by OCTOPUS. However, this additional correct prediction could well be attributed to the overlap between the test and training sets, as, in the absence of cross-validation, MEMSAT-SVM is able to predict 82% (9) topologies correctly. In terms of predicting re-entrant helices, MEMSAT-SVM identifies 44% (8 out of 18) with 2 false positive predictions, which compares favourably with OCTOPUS results of 22% (4) with 4 false positives. Since the numbers of proteins containing re-entrant helices and signal peptides are relatively small (14 and 11 respectively), care should be taken when interpreting these results as a relatively large percentage difference in performance may only reflect the correct prediction of one additional sequence.

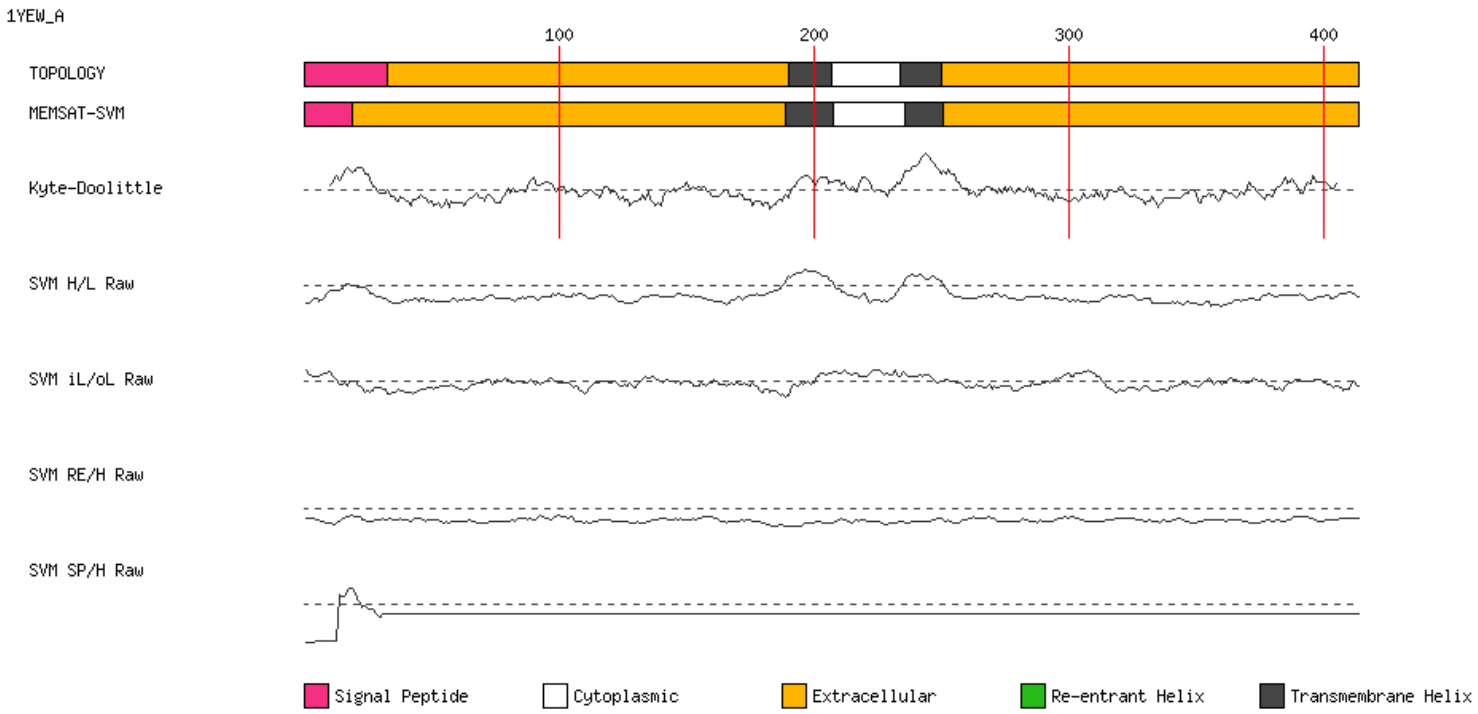

Figure 4.2: Correct topology prediction for Particulate Methane Monooxygenase chain A from *Methylococcus capsulatus* (PDB: 1YEW:A), showing a 2 TM helix prediction with an extracellular N-terminus. In addition to the topology, a signal peptide (shown in pink) was also correctly predicted.

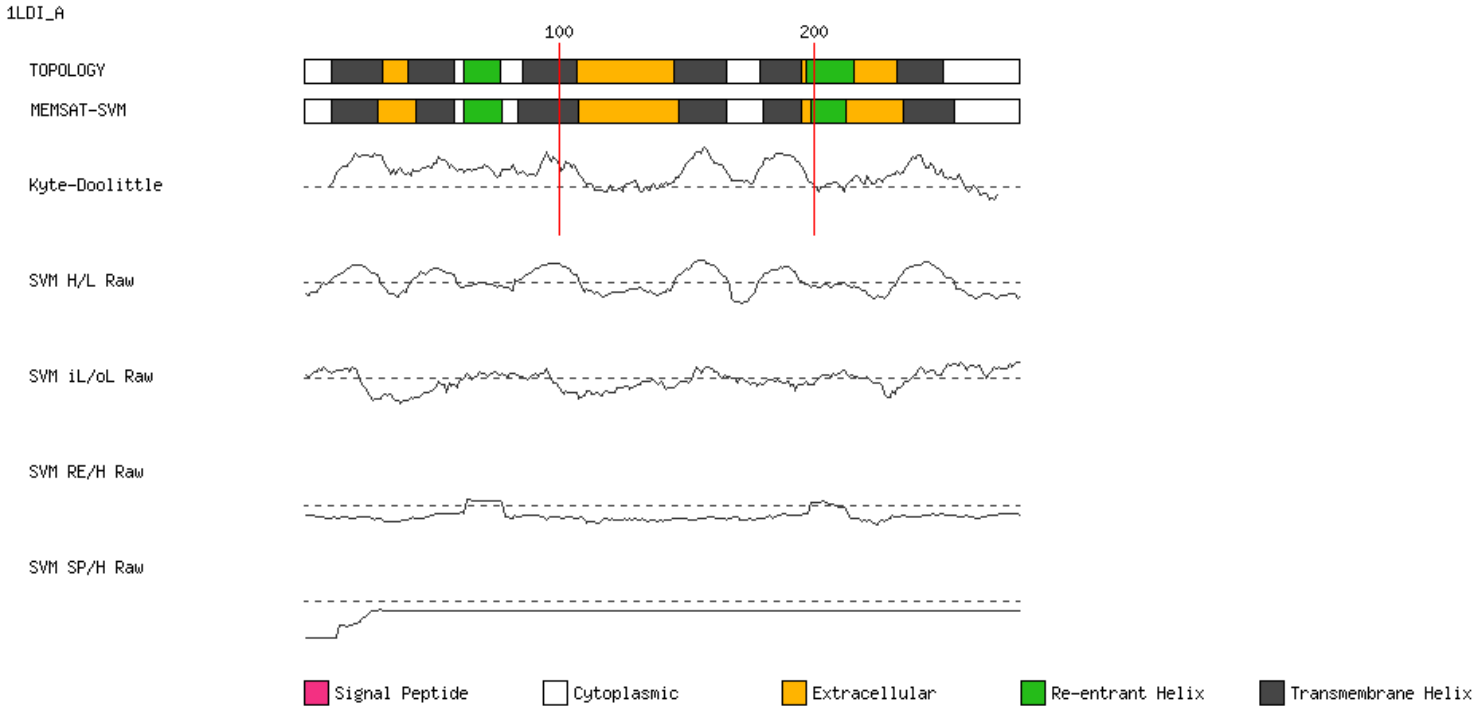

Figure 4.3: Correct topology prediction for Glycerol Uptake Facilitator chain A from *Escherichia coli* (PDB: 1LDI:A), showing a 6 TM helix prediction with an intracellular N-terminus. In addition to the topology, two re-entrant helices (shown in green) were also correctly predicted.

4.3.4 Erroneous predictions

MEMSAT-SVM incorrectly predicts topologies in 15 cases. Four of these correspond to proteins containing re-entrant helices that are erroneously predicted as TM helices - ABC transporter BtuCD (Figure 4.4), Proton Glutamate Symport protein, Aquaporin Z and Clc chloride channel (PDB: 1L7V:B, 1XFH:A, 2ABM:H and 2FEE:B) - accounting for the majority of over predicted TM helices. The remaining over prediction is due to a highly hydrophobic N-terminal region within a chain from Cytochrome bc1 (PDB: 1SQX:D).

In seven cases, incorrect topologies are a result of under predicted TM helices - Photosystem I (chains A, L and K), Steryl-sulfatase, Light-Harvesting Complex II, Particulate Methane Monooxygenase and Sodium/proton antiporter 1 (1JB0:A, 1JB0:L, 1JB0:K, 1P49:A, 1VCR:A, 1YEW:B and 1ZCD:B). These under predictions fall into two categories; weakly predicted helices (1JB0:A, 1JB0:L, 1JB0:K, 1P49:A and 1VCR:A) or prediction of one helix rather than two shorter ones (1YEW:B and 1ZCD:B). Of the weakly predicted helix errors, sequence analysis indicates low hydrophobicity for many of these helices, often due to a large fraction of charged residues. Such helices are therefore extremely difficult to predict and suggest a novel membrane insertion mechanism. Other helices appear sufficiently hydrophobic to be detected; errors are possibly the results of PSI-BLAST alignment which reduce their detectability.

The remaining three incorrect predictions are all single TM helix proteins that are inverted - Photosystem I, Cytochrome bc1 and Cytochrome b6f (1JB0:I, 1P84:I and 1Q90:N). In all three cases, the confidence of the prediction compared to the correct topology (measured by the difference between the two scores) is extremely small. With no clear signal to differentiate between either orientation, interplay with other chains from the same protein may influence the final conformation.

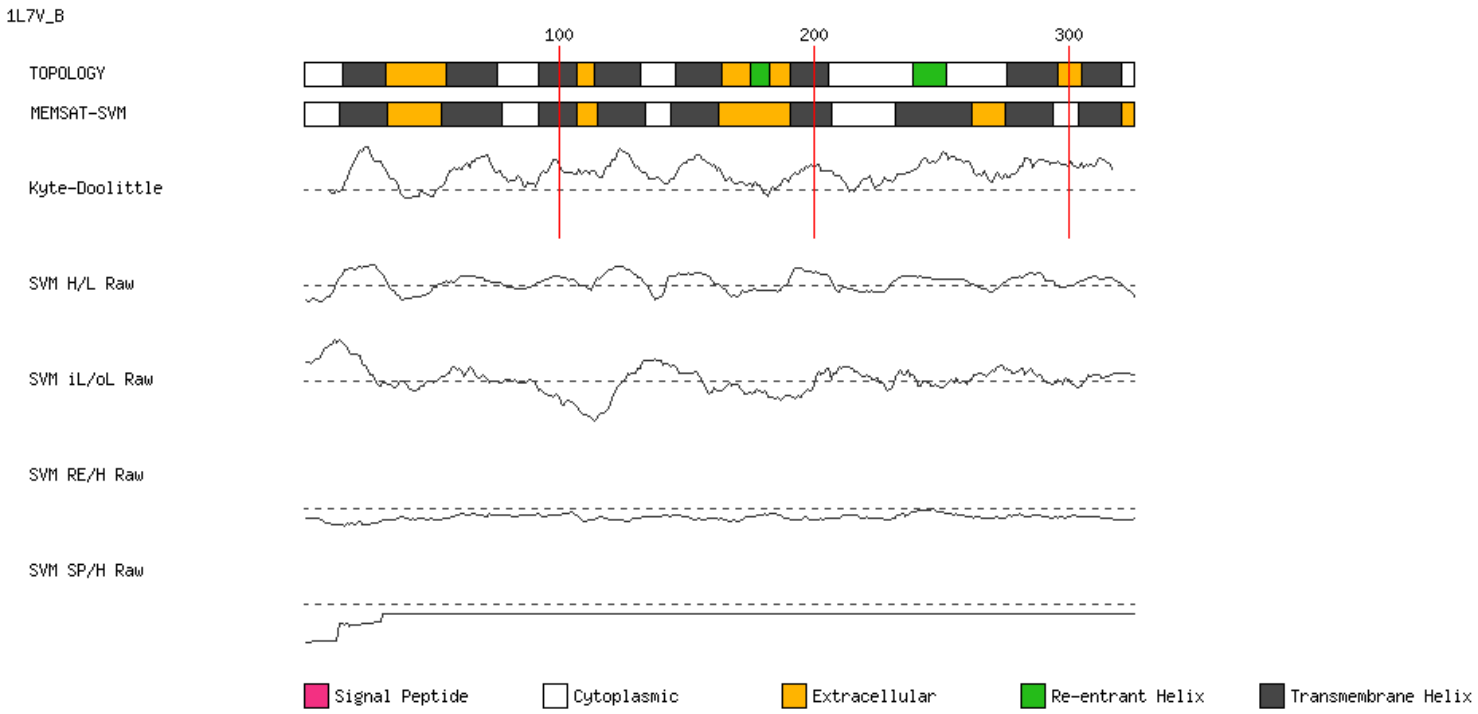

Figure 4.4: Incorrect topology prediction for ABC transporter BtuCD chain B from *Escherichia coli* (PDB: 1L7V:B), showing a 96 TM helix prediction with an intracellular N-terminus. The actual topology consists of 8 TM helices and 2 re-entrant helices; one of the re-entrant helices has been incorrectly predicted as a TM helix (predicted TM helix 7).

Method	**Möller**	**TOPDB**
MEMSAT-SVM	78%	67%
OCTOPUS	69%	64%
MEMSAT3	77%	66%
ENSEMBLE	61%	51%
PHOBIUS	67%	62%
HMMTOP	64%	57%
PRODIV	46%	37%
SVMTOP	70%	42%
TMHMM	60%	56%
PHDhtm	45%	49%

Table 4.3: Prediction performance using the Möller and TOPDB data sets. Column 1: Prediction method. Column 2: Results using the Möller data set. Column 3: Results using the TOPDB data set.

4.3.5 Prediction accuracy using the Möller and TOPDB data sets

We additionally tested prediction performance using a subset of 184 sequences from the Möller set (described in (Bagos *et al.*, 2006; Hirokawa *et al.*, 1998), composed of sequences annotated using both crystal structures and biochemical characterisation (Table 4.3). The Möller set consists of a significantly higher fraction of eukaryotic sequences compared to the data set described above. TM protein crystallisation techniques usually involve over expression hosts, such as *Escherichia coli*, which to date have worked mainly for prokaryotic TM proteins since eukaryotic TM proteins are still very difficult to over express (Granseth *et al.*, 2007). Crystal structure-based sets, while providing more accurate TM helix boundary definitions, thus suffer from this bias towards prokaryotic sequences, so methods trained exclusively using such data sets run the risk of performing poorly when predicting the topologies of eukaryotic sequences. Based on recent updates to SWISS-PROT annotations and under full cross-validation, MEMSAT-SVM achieved 78% accuracy and MEMSAT3 achieved 77%. In the absence of cross-validation, SPOCTOPUS also achieved 77% accuracy, with OCTOPUS the next best method scoring 69%. This performance suggests MEMSAT-SVM offers robust prediction accuracy on proteins from both eukaryotic and prokaryotic domains.

We then tested performance using the TOPDB (Tusnady *et al.*, 2008) data set, a comprehensive collection of TM protein containing experimentally derived topology information (Table 4.3). It currently contains records for 1452 alpha-helical TM proteins. Using this data set, MEMSAT-SVM achieved 67% accuracy, MEMSAT3 66%, OCTOPUS 64% and PHOBIUS 62%. The data set also contains 317 sequences containing signal peptides. Of these, MEMSAT-SVM correctly predicted the topologies for 77% of cases. This value was lower than that of PHOBIUS which achieved 85% accuracy. However, the MEMSAT-SVM false positive rate for signal peptide prediction is 7%, half the PHOBIUS value of 14%. These results show that on this data set, MEMSAT-SVM signal peptide performance is below that of PHOBIUS, though MEMSAT-SVM overall prediction accuracy is 5% higher due to the relatively poor performance of PHOBIUS on sequences that do not contain signal peptides (a substantially larger fraction) - 54% accuracy compared to 63% for MEMSAT-SVM. These results should again be treated with caution as they were not cross-validated.

These results are clearly lower than those attained using the crystal structure-based data set, and we believe this is likely due to errors in TOPDB. We analysed sequences from the original, uncorrected Möller set that at the time did not have crystal structures. 55 of these sequences now have a homologous PDB structure (E-value < 0.001), and of these only 38 (69%) of the original Möller topologies are correct based on current OPM definitions (taking into account only the N-terminal location and TM helix count). There is no reason to believe that the error rate in other data sets such as TOPDB, composed predominantly of sequences whose topologies were determined by biochemical means, should be significantly different. Perfect prediction methods are therefore unlikely to be able to achieve results higher than this, while older methods trained on erroneous topologies have the potential to achieve higher scores but may in reality be poorer predictors, a fact likely to be highlighted when tested against a crystal structure-based set.

4.3.6 Discriminating between globular and transmembrane proteins

Using the combined set of 2453 test cases, we assessed performance in discriminating between globular and TM proteins (Table 4.4). As a discrimination threshold, a number of residues were required to be predicted as part of a TM helix by the SVM in order to classify the protein as TM. This threshold was adjusted in order to minimise the margin between the false positive (FP) and false negative (FN) rates, therefore avoiding bias towards either type of prediction. A 0% FP rate and 0.4% FN rate was achieved using only a single residue as the threshold, an improvement on the MEMSAT3 neural network-based approach (0.5% FP, 0.5% FN) and SOSUI (0.3% FP, 1.1% FN). OCTOPUS matched the FP rate but achieved a higher FN rate, while PHOBIUS matched the FN rate but achieved a higher FP rate. These low error rates suggest that MEMSAT-SVM is extremely well suited to whole genome analysis. An analysis of a selection of known beta-barrel proteins suggested that these are also identified effectively by this method.

Method	Algorithm	False positive rate	False negative rate
MEMSAT-SVM	SVM	0.00%	0.44%
MEMSAT3	NN	0.50%	0.50%
SOSUI	Hydrophobicity analysis	0.33%	1.10%
OCTOPUS	NN + HMM	0.00%	2.51%
PHOBIUS	HMM	2.72%	0.44%

Table 4.4: Results for TM/globular protein discrimination rates.

4.3.7 Application to a number of complete genomes

Table 4.5 shows the results of applying the TM/globular predictor to a number of complete genomes. We estimate that a typical genome contains between 24% and 33% TM proteins, which is slightly higher than previous estimates of between 20% and 30% (Wallin & von Heijne, 1998). Two organisms that have a noticeably higher fraction of TM proteins are *Caenorhabditis elegans* and *Takifugu rubripes*. *Takifugu rubripes* is known to have extensive channel heterogeneity

compared to *Homo sapiens*, with 10 *Homo sapiens* voltage-gated calcium channel α1-subunit genes revealing 21 orthologous genes in *Takifugu rubripes*. Phylogenetic analysis reveals that this is due to fish lineage specific α1-subunit subtype duplication (Wong *et al.*, 2006). Similar increased subtype diversity has also been detected in the appetite receptor neuropeptide Y GPCR family that may have arisen as a result of ray-finned fish tetraploidization (Larsson *et al.*, 2005). *Caenorhabditis elegans* is known to have an exceptionally large number of 7 TM receptors and rhodopsin-like membrane proteins (Liu *et al.*, 2002), thought to have been arisen through duplication events, that possibly imply functional relations between homologous 7 TM domains (Liu *et al.*, 2004). *Escherichia coli* has the lowest fraction of TM proteins of all the species we analysed, which may be a consequence of the lack of internal membrane systems in prokaryotes (Petty, 1993).

Species	Fraction of genome predicted as TM proteins	Fraction of TM proteins predicted to contain re-entrant helices	Fraction of TM proteins predicted to contain signal peptides
Caenorhabditis elegans	33%	2%	33%
Canis familiaris	31%	2%	27%
Danio rerio	29%	2%	26%
Drosophila melanogaster	27%	2%	33%
Escherichia coli	24%	2%	28%
Homo sapiens	26%	2%	35%
Mus musculus	29%	2%	30%
Pan troglodytes	26%	2%	33%
Takifugu rubripes	33%	3%	26%
Xenopus tropicalis	31%	2%	23%

Table 4.5: The fraction of proteins predicted as transmembrane, and to contain re-entrant helices and signal peptides, in a number of complete genomes.

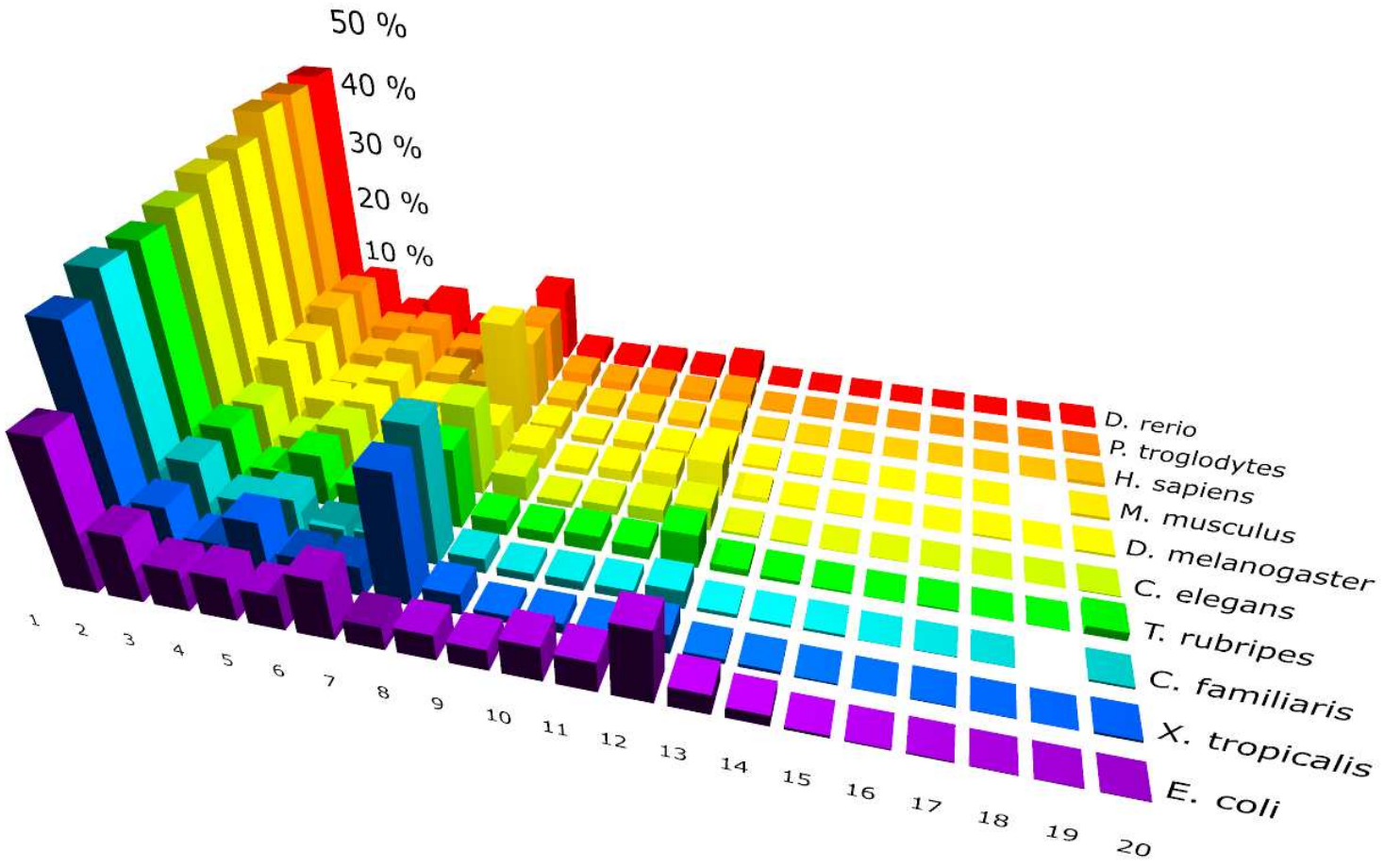

Figure 4.5: Topology prediction results for a number of complete genomes. X-axis: Number of predicted TM helices. Y-axis: Fraction of all predicted TM proteins. Z-axis: Species.

We then carried out full topology prediction on sequences predicted to be TM proteins and analysed these for the presence of re-entrant helices and signal peptides. In most species, re-entrant helices were detected in at least 2% of TM proteins, with more than 3% detected in *Takifugu rubripes* which can be explained by the extensive channel heterogeneity discussed above. However, given the low true positive rate of 44%, this figure is likely to be an underestimate. A positive predictive value (PPV) of 0.8 suggests a value in the range 3-4.5% is more realistic. This range is close to one previous estimate of 5% (Cuthbertson *et al.*, 2005) but below another of 10% (Lasso *et al.*, 2006), although the latter was based on a broader definition of re-entrant regions that did not necessarily contain helical secondary structure.

Topology prediction results illustrate consistent trends across all species, with significant peaks at 7 TM helices representing GPCRs (in eukaryotes) and 12 TM helices representing transporters proteins (Figure 4.5). A slight preference for even-numbered topologies (excluding GPCRs) can be explained by the formation of 2 helix hairpins as independent units during protein assembly, therefore favouring topologies with even numbers of TM helices (Gafvelin *et al.*, 1997). In all species, the most dominant topology is a single TM helix. These results are consistent with previous studies (Krogh *et al.*, 2001).

4.4 Discussion

In this chapter we have implemented a novel SVM-based TM protein topology predictor, an area previously dominated by HMM and NN-based machine learning approaches, and have shown that it outperforms a selection of the best performing prediction methods when fully cross-validated on a novel high resolution data set of 131 protein sequences. This data set includes proteins containing both re-entrant helices and signal peptides, features that this method is also able to predict. The method has also been benchmarked on the Möller data set, which contains a higher fraction of eukaryotic sequences, improving on the best current methods. And we

have achieved extremely low false positive and false negative rates for TM/globular protein discrimination. Using these tools, we have estimated the fraction of TM proteins, re-entrant helices and signal peptides in a number of complete genomes. Overall, our results suggest that MEMSAT-SVM is ideally suited to whole genome annotation of alpha-helical TM proteins.

Chapter 5

Predicting transmembrane helix packing arrangements using residue contacts and a force-directed algorithm

5.1 Background

As discussed in previous chapters, significant effort has been invested in attempting to predict TM protein topology (Jones, 2007; Viklund & Elofsson, 2008; Nugent & Jones, 2009). In contrast, comparatively little attention has been directed toward developing a method to pack the helices together; since the membrane-spanning region is predominantly composed of alpha-helices with a common alignment, this task should in principle be easier than predicting the fold of globular proteins as the longitudinal constraints of helix packing mostly reduces the solution space from three dimensions to two. However, topologies consisting of large numbers of TM helices as well as structural features including re-entrant, tilted and kinked helices render simple approaches that may work for regularly packed proteins unable to predict the diverse packing arrangements now present in structural databases.

5.1.1 Predicting transmembrane protein folds

Early attempts to predict TM protein folds were based on sequence similarity to proteins with a known three-dimensional structure, using statistically derived environmental preference parameters combined with experimentally determined features (Cronet *et al.*, 1993). Another method calculated amino acid substitution tables for residues in membrane proteins where the side chain was accessible to lipid. By comparing observed substitutions obtained from sequence alignments of TM regions, accessibility of residues to the lipid could be predicted. In combination with a Fourier transform method to detect alpha-helices, the buried and exposed faces could then be discriminated and the presence of charged residues used to construct a three-dimensional model (Donnelly *et al.*, 1993). Other methods also made use of exposed surface prediction to allocate helix positions, in combination with an existing framework for globular protein structure prediction involving the combinatorial enumeration of windings over a predefined architecture followed by the selection of preferred folds (Taylor *et al.*, 1994). However, many of these methods were only suitable for 7 TM helix bundles such as rhodopsin and were unsuitable for other topologies.

5.1.2 *Ab initio* methods

More recently, the fragment-based protein tertiary structure prediction method FRAGFOLD (Jones, 2001) was modified to model TM proteins. FRAGFOLD is based on the assembly of super-secondary structural fragments using a simulated annealing algorithm in order to narrow the search of conformational space by pre-selecting fragments from a library of highly resolved protein structures. FILM (Pellegrini-Calace *et al.*, 2003) added a membrane potential to the FRAGFOLD energy terms which was derived from the statistical analysis of a data set of TM proteins with experimentally defined topologies. Results obtained by applying the method to small membrane proteins of known three-dimensional structure showed it could predict both helix topology and conformation at a reasonable accuracy level. Despite these good results, the combinatorial complexity of such *ab initio* protein folding methods means that it is unfeasible to use such approaches for large TM structures, many of which are longer than 150 residues. Modification of another globular protein *ab initio* modelling program, ROSETTA (Rohl *et al.*, 2004), added an energy function that described membrane intra-protein interactions at atomic level and membrane protein/lipid interactions implicitly, while treating hydrogen bonds explicitly (Barth *et al.*, 2007). Results suggest that the model captures the essential physical properties that govern the solvation and stability of TM proteins, allowing the structures of small protein domains, up to 150 residues, to be predicted successfully to a resolution of less than 2.5$\mathring{A}$. A recent enhancement of the algorithm demonstrated that by constraining helix-helix packing arrangements at particular positions based on local sequence-structure correlations for each helix of the interface independently, TM proteins with more complex topologies could be modelled to within 4$\mathring{A}$ of the native structure (Barth *et al.*, 2009).

5.1.3 Helix-Helix interaction motifs

The prediction of helix-helix interactions, derived from residue contacts and topology, has only recently been investigated in TM proteins due to the relative paucity of TM protein crystal structures. In contrast, a number of globular

protein contact predictors exist based on a variety of machine learning algorithms (Punta & Rost, 2005; Cheng & Baldi, 2007), and contact prediction has also been used to assess globular protein models submitted to the Critical Assessment of Structure Prediction (CASP) experiment (Izarzugaza *et al.*, 2007). However, analysis has shown that such globular proteins contact predictors perform poorly when applied to TM proteins, most likely due to differences between TM and globular interaction motifs (Fuchs *et al.*, 2009). A number of studies have identified structural and sequence motifs recurring frequently during helix-helix interaction in TM proteins. One investigation analysed interacting helical pairs according to their three-dimensional similarity, allowing three quarters of pairs to be grouped into one of five tightly clustered motifs (Walters & DeGrado, 2006). The largest of these consisted of an anti-parallel motif with left-handed packing angles, stabilised by the packing of small side chains every seven residues, while right-handed parallel and anti-parallel structures showed a similar tendency though spaced at four-residue intervals. Another study identified a specific aromatic pattern, aromatic-XX-aromatic, which was demonstrated to stabilise helix-helix interactions during assembly (Sal-Man *et al.*, 2007), while others include the GXXXG motif found in glycophorin A (Lemmon *et al.*, 1992), heptad motifs of leucine residues (Gurezka *et al.*, 1999), and polar residues through formation of hydrogen bonds (Zhou *et al.*, 2001).

The discovery of these recurring motifs, and the likelihood that there are more as yet undiscovered, suggests predictability by a generalised pattern search strategy. Recently, two methods have been developed that attempt to predict residue contacts and helix-helix interaction. TMHcon (Fuchs *et al.*, 2009) uses a neural network in combination with profile data, residue co-evolution information, predicted lipid exposure using the LIPS method (Adamian & Liang, 2006), and a number of TM protein specific features, such as residue position within the TM helix, in order to predict helix-helix interaction. TMhit (Lo *et al.*, 2009) uses a two-level hierarchical approach in combination with a support vector machine

(SVM) classifier. The first level discriminates between contacts and non-contacts on a per residue basis, before the second level determines the structure of the contact map from all possible pairs of predicted contact residues therefore avoiding the high computational cost incurred by the quadratic growth of residue pair prediction.

In this chapter, I will describe the development of a novel method to predict lipid exposure, residue contacts, helix-helix interactions and finally the optimal helical packing arrangements of TM proteins. Using molecular dynamics data to label residues potentially exposed to lipid, I have trained and cross-validated a SVM classifier to predict per residue lipid exposure with 69% accuracy. This information is combined with PSI-BLAST profile data and a variety of sequence-based features to train an additional SVM to predict residue contacts. Combining these results with a priori topology information, I was able to predict helix-helix interaction with up to 65% accuracy under stringent cross-validation on a non-redundant test set of 74 protein chains. I then tested the ability of the method to discriminate native from decoy helical packing arrangement using a decoy set of 2811 structures. By comparing our predictions with the test set, I was able to identify the native packing arrangement with up to 70% accuracy. All these performance metrics represents significant improvements over existing methods. In order to visualise the global packing arrangement, I adopted a graph-based approach. By employing a force-directed algorithm, the method attempts to minimise edge crossing while maintaining uniform edge length, attributes common in native structures. Finally, a genetic algorithm is used to rotate helices in order to prevent residue contacts occurring across the longitudinal helix axis.

5.2 Methods

5.2.1 Data sets

For SVM training and cross-validation, we used the crystal structure data set described in Chapter 3 which contained 74 sequences with at least two TM helices.

For 53 of these multi-spanning sequences, and a further 24 single-spanning proteins, we were able to obtain molecular dynamics data from the Coarse Grained Database (CGDB) (Chetwynd *et al.*, 2008) which was used for lipid exposure prediction. We chose not to predict interactions between TM helices and re-entrant helices, found in many channels such as aquaporin, as they are thought to be involved in channel gating and thus move into and out of the membrane region depending on physiological conditions. Including re-entrant helices would therefore be likely to introduce noise into the data set as contacts could be both positive and negative training examples.

5.2.2 Predicting lipid exposure

During TM protein crystallisation, detergents are used extensively for protein solubilisation and then act as mimics of the lipid bilayer due to their self-assembly properties. As a result, crystallographic data rarely contains information regarding the positions of lipid molecules, therefore hindering the study, and prediction, of lipid exposed regions of TM protein. For investigating TM topology, a number of automated methods exist that attempt to position the protein within the membrane (Lomize *et al.*, 2006b; Tusnady *et al.*, 2005a). However, these methods are inappropriate for accurate studies of lipid exposure as they do not take into account the solvent-filled cavities and channels found in many TM proteins. To address this, we used the CGDB, a resource of coarse-grained simulation data, which contains analysis of lipid-protein interactions following 200ns of molecular dynamics using GROMACS (Spoel *et al.*, 2005) to randomly surround TM proteins in dipalmitoylphosphatidylcholine lipids and solvent. A snapshot of each protein in its optimum position within the bilayer and residue statistics throughout the simulation are available. While difficult to validate, the approach has proved successful in reproducing the behaviour of equivalent atomistic simulations of model proteins, as well as allowing the insertion of various test peptides whose final configurations were in agreement with experimental data (Sansom *et al.*, 2008). Additionally, channel-containing proteins such as aquaporin and potassium channels are solvent rather than lipid filled at the end of simulation.

To train the SVM classifier, we used CGDB data to label residues that were lipid exposed. For the 77 proteins within our data set where CGDB data was available, each residue within the membrane was labelled as lipid exposed where the fraction of simulation time exposed to DPPC lipid was greater than 0.5. PSI-BLAST (Altschul *et al.*, 1997) was used to generate position-specific scoring matrices for each of the 77 proteins in the data set using the UniRef 90 database. Two iterations were performed with a profile-inclusion E-value threshold of 0.001. For each residue in a sequence, a sliding window approach was used with a window size of 7, creating a feature vector of length 140 centred on the target residue. To determine this windows size, the data set was split randomly into two and the highest scoring window which ranked equally in each split was selected, therefore demonstrating consistency between data sets and reducing the risk of overfitting. Where the window extended beyond the protein termini, empty feature values were set to zero. All values for each feature position where then normalised by Z-score to enable faster SVM convergence. In training, the target sequence, along with any other sequences with an E-value less than 1e-4, were excluded. We used SVM-Light (Joachims, 1998, chapter 11) and a radial basis function kernel, in combination with a grid search of SVM parameters. Matthews Correlation Coefficient (MCC) was used to optimise these values as it has been shown to be a more robust measure than using recall or precision alone (Matthews, 1975).

5.2.3 Contact definitions

In order to make direct comparisons with other methods, we used three thresholds to consider a pair of residues to be in contact. Firstly, a maximal distance of $8\mathring{A}$ between their C-beta atoms (C-alpha for glycine) (Punta & Rost, 2005; Cheng & Baldi, 2007) (contact definition 1). Secondly, the distance between any two atoms from an interacting pair is less than the sum of their van der Waals radii plus a threshold of $0.6\mathring{A}$ (Lo *et al.*, 2009) (contact definition 2). Thirdly, the minimal distance between side chain or backbone heavy atoms in an interacting pair is less

than 5.5$\mathring{A}$ (Fuchs *et al.*, 2009) (contact definition 3). We defined TM helices as interacting if one residue from each helix was observed to be in contact.

5.2.4 Predicting residue contacts

Using the three contact definitions, all residue pairs from different TM helices were labelled as contacting or non-contacting, resulting in a substantial bias of approximately 1:50. In order to balance training sets and reduce learning time, non-contacting examples were selected randomly in order to achieve approximately equal numbers of positive and negative examples, before fine adjustment of the SVM cost-factor parameter achieved a 1:1 ratio.

SVM input features were based largely on PSI-BLAST profile data, generated as described above. We used a sliding window of 7 residues, centred on each residue in the pair to produce a feature vector of length 280. Again, this window size was determined by randomly splitting the data set. In addition to profile data, the raw SVM scores for predicted lipid exposure were added to the feature vector for each residue. We then added a number of sequence derived statistics. To define the sequence separation between the two residues, a binary vector was used corresponding to distances of 50, 75, 100, 125, 150, 175, 200 and greater than 200 residues. We also added a value which corresponded to the relative position of each residue within the two TM helices, generated by dividing the residue position in the TM helix by the helix length, and subtracting the value from one where the two residues were on adjacent TM helices or are separated by an even number. This value effectively represented a relative Z-coordinate for each residue, the rationale being that residues separated by a large degree on the Z-axis were unlikely to contact. We tried adding a number of additional values including the lengths of each TM helix, average lipid exposure scores for each TM helix, total number of TM helices, sequence length, and a number of residue co-evolution scores (Olmea & Valencia, 1997; Fodor & Aldrich, 2004). However, none of these values increased classification performance so were removed in the final model. Again, each feature position was normalised by Z-score,

before the target sequence and any other sequences with an E-value less than 1e-4 were excluded from training sets. A radial basis function kernel was used and the MCC was used to optimise SVM parameters.

5.2.5 Using helix-helix prediction for discriminating decoy helical packing arrangements

We then tested the ability of the method to discriminate native from decoy helical packing arrangement using the predicted helix-helix interactions. For each of the 74 multi-spanning proteins in our data set, decoys were generated using the REVCAS program (Taylor, 2006). Each chain was expanded into a larger set of structures by making it circular and introducing cyclically permuted breaks. The method involves a triple-point chain reconnection that avoids the restoration of native segments allowing the generation of a set of decoy structures. The method was successfully applied to the pore-forming colicin domain, an all alpha-helical structure that is typical of many TM proteins in that the amino and carboxy termini, which are joined when the structure is circularised, are at opposite ends of the protein, much like TM proteins whose termini are on opposite sides of the membrane (Taylor, 2006). By generating decoys in both forward and reverse directions, 24-48 decoys were generated for each protein resulting in a total set of 2811 structures. Decoys only contained C-alpha atoms, therefore the remaining backbone and side chain atoms were added and the structure was refined and energy minimised using the Jackal package (Petrey *et al.*, 2003). Additionally, homology models of the native structures were constructed using MODELLER (Eswar *et al.*, 2007). Native topologies were then used to define TM helix boundaries allowing observed helix-helix interactions to be extracted which were then compared to the helix-helix interactions predicted from sequence. Decoys and native structures were then scored by the number of interacting/non-interacting helices that matched the predictions and ranked accordingly. We measured the frequency at which the native structure, or a model of the native structure, was ranked first.

5.2.6 Constructing the helical packing arrangement

Once helix-helix interactions had been predicted, the helical packing arrangement was treated as an undirected graph where the helices form vertices and their interactions form edges. A force-directed algorithm is then applied which treats the graph as a virtual physical system. The system is simulated resulting in attractive and repulsive forces being applied to vertices, a process which is repeated iteratively until the system comes to an equilibrium state at which point the final graph layout is constructed.

Using the Boost C++ programming library (http://www.boost.org) we employed a modified version of the Kamada-Kawai force-directed algorithm (Kamada & Kawai, 1989) which generates two-dimensional layouts for connected, undirected graphs. It accomplishes this by treating the graph as a dynamic spring system, where the strength of a spring between two vertices is inversely proportional to the square of the shortest distance between those two vertices, and attempting to minimise the energy within the system. In order to avoid producing a layout with only a local minima, the vertices are first arranged along the vertices of a regular n-sided polygon, where n is the number of TM helices, via a circular layout function.

In their paper, Kamada and Kawai suggested that in many scenarios, the reduction of the number of edge crossings that a graph possesses is not necessarily a good aesthetic criterion for a layout algorithm to implement. They suggested that the total balance of the layout, which is related to the individual characteristics of the graph, can be considered more important than the reduction of edge crossings in the graph. They calculated the total balance of the graph as the square summation of the differences between the ideal distance and the actual distance for all vertices (Equation 5.1). By approximating and minimising the stress in a given system, the Kamada-Kawai method preserves the total balance of a graph, producing layouts

with small numbers of edge crossings.

$$stress(X) = \sum_{i<j} w_{ij}(\parallel X_i - X_j \parallel - d_{ij})^2 \tag{5.1}$$

Equation 5.1: For a pair of nodes i and j, d_{ij} is the ideal distance between vertices corresponding to the shortest path between those vertices. X is the set of 2D or 3D coordinates and w_{ij} is ${d_{ij}}^{-\alpha}$. Kamada and Kawai chose $\alpha = 2$ which seems to produce the best layouts (Kamada & Kawai, 1989). Kamada and Kawai used the Newton-Raphson method to optimise with respect to a single vertex. By iteratively solving for each vertex the overall stress is reduced.

Given that the number of TM helices in a protein is expected to be less than 30, energy minimisation occurs in a number of seconds on a modern computer, avoiding the high running time typically associated with force-directed algorithms and graphs containing a larger number of vertices. Resulting layouts demonstrate uniform edge length, uniform vertex distribution often showing symmetry, and minimisation of edge crossing - attributes that are common to the arrangement of TM helices and their interactions in native TM protein structures.

In a number of cases, multiple helices share the same interactions resulting in numerous possible arrangements. In all cases where this occurs, a recursive function is used to score each arrangement according to the number of observed same-side loop crossovers. The score is determined by drawing a line between a pair of helices adjacent in sequence, before incrementing the helix position by two so that comparisons are between lines on the same side. Each line is compared to every other line on the same side and their intersection is established by determining the cross product. This is repeated for each side, before the total number of intersections per side is compared. Particularly when loops are short, it is unusual for loops to cross each other as this may result in side chain clashes. All arrangements are then returned, with those containing the least number of same-side loop crossovers scored highest.

Finally, the constituent residues are superimposed onto their respective TM helices, before a genetic algorithm is used to rotate all helices around their respective Z-axes such that the sum of all predicted residue-residue contact distances is minimised, therefore preventing residues contacts occurring across the longitudinal helix axis. For each TM helix, a value in the range 0-359 is optimised to an accuracy of one degree.

5.3 Results

5.3.1 Lipid exposure prediction performance

We compared the per residue performance of our lipid exposure predictor to the LIPS method using all TM helix residues from our data set of 77 sequences. The data set contained 336 TM helices composed of 7016 residues, of which 3687 were labelled as lipid exposed and 3329 were not, according to CGDB data. Optimal performance was achieved using a radial basis function kernel, a gamma value of 0.6 and a trade-off value of 1.5. The LIPS method produces a per residue score generated by multiplying lipophilicity by positional entropy. The LIPS score that resulted in the optimal per residue performance was found to be 1.56. Using leave-one-out cross-validation, our method achieved a MCC of 0.38 and accuracy of 69.3%, a significant improvement over the LIPS method which scored 0.23 and 61.7% respectively (Table 5.1). Furthermore, the LIPS method is calculated using sequence profiles from 18 TM protein structures, the majority of which are included in the test set of 77, therefore in the absence of cross-validation these results are likely to be an overestimate. However, as the LIPS method is based on an alternative definition of lipid exposure, we repeated the benchmarking of the two methods using the LIPS definition by labelling residues with a 1.9$\mathring{A}$ probe. Under this definition both methods perform slightly worse although our method still outperforms LIPS, with an MCC value of 0.27 compared to 0.18. This indicates that there is reasonably good correlation between the two definitions although the LIPS definition is slightly harder to predict, most likely because the 1.9$\mathring{A}$ spherical probe is a poor approximation to

Method	Lipid exposure definition	Precision	Recall	FPR	FNR	MCC	Accuracy
MEMPACK	CGDB	0.69	0.56	0.36	0.26	0.38	69.3%
MEMPACK	1.9$\mathring{A}$ probe	0.71	0.61	0.39	0.33	0.27	64.3%
LIPS	CGDB	0.61	0.59	0.48	0.29	0.23	61.7%
LIPS	1.9$\mathring{A}$ probe	0.65	0.65	0.50	0.32	0.18	60.3%

Table 5.1: Per residue lipid exposure prediction performance using a data set of 77 sequences. Lipid exposure definition = test set labelled according to the CGDB definition or using a 1.9 probe. FPR = false positive rate. FNR = false negative rate. MCC = Matthews Correlation Coefficient. Accuracy = (TP + TN)/(TP + TN + FP + FN).

the non-spherical nature of a membrane phospholipid, unlike, for example, a 1.4$\mathring{A}$ spherical probe is to a water molecule.

5.3.2 Residue contact prediction performance

Residue pair contact prediction performance compared with two TM protein contact predictors (TMHcon (Fuchs *et al.*, 2009) and TMhit (Lo *et al.*, 2009)) and two globular protein contact predictors (PROFcon (Punta & Rost, 2005) and SVMcon (Cheng & Baldi, 2007)) using the data set of 74 sequences and three contact definitions is shown in (Table 5.2). Existing methods all had the option of a L5 mode, where only the top L/5 positive results are returned where L is the sequence length, or for TM protein-specific methods, the total length of all TM helices. This generally has the effect of reducing the false positive rate though usually at the expense of increasing the false negative rate; however our method did not benefit from the use of this scoring method, suggesting the SVM hyperplane is already optimally positioned.

Performance at all three contact definitions was consistent, with a MCC value of approximately 0.28 although a slightly lower false positive rate using contact definition 2. All three SVMs achieved optimal performance using radial basis function kernels with gamma and trade-off values of 24 and 1 respectively. Addition of the predicted lipid exposure scores to profile data in the SVM feature vector resulted in an improvement of approximately 0.05 MCC, while the additional

Method	**Contact Definition**	**Precision**	**Recall**	**FPR**	**FNR**	**MCC**
MEMPACK	1	0.69	0.0023	0.0010	0.88	0.28
SVMcon	1	0.06	0.00050	0.0083	0.97	0.03
SVMcon L5	1	0.09	0.00	0.00030	0.99	0.01
PROFcon	1	0.03	0.021	0.46	0.41	0.04
PROFcon L5	1	0.06	0.00010	0.0018	0.99	0.01
MEMPACK	2	0.69	0.0015	0.00070	0.88	0.28
TMhit L5	2	0.57	0.0015	0.0012	0.88	0.26
MEMPACK	3	0.70	0.0022	0.0010	0.89	0.27
TMHcon L5	3	0.09	0.00020	0.0021	0.99	0.02

Table 5.2: Per residue pair contact prediction performance using a data set of 74 sequences. Contact definition 1 = A maximal distance of 8 between their C-beta atoms (C-alpha for glycine). 2 = The distance between any two atoms from an interacting pair is less than the sum of their van der Waals radii plus a threshold of 0.6. 3 = The minimal distance between side chain or backbone heavy atoms in an interacting pair is less than 5.5. Results for contact definition 3 used 58 sequences that had more than 2 TM helices as TMHcon is unable to make predictions for 2 TM helix sequences.

sequence derived statistics contributed approximately 0.03 MCC. Although a combination of residue co-evolution scores did improve performance slightly compared with using profile data alone (0.02 MCC), this increment was lost when scores were added after predicted lipid exposure suggesting the two overlap in feature space.

Compared to existing predictors, our method performed well, with MCC scores substantially higher than both SVMcon and PROFcon (contact definition 1) using either standard or L5 scoring schemes. SVMcon L5 was able to produce a lower false positive rate (FPR) but at the expense of a false negative rate (FNR) of 0.99. Similarly, PROFcon produced a lower FNR of 0.41 but at the expense of a higher FPR of 0.46, compared to 0.001 for our method. On this evidence, globular protein contact predictors appear to perform relatively poorly when applied to TM proteins. In comparison to TMhit, a recent SVM-based TM protein contact predictor, results were more comparable. While our method scores higher on all assessment metrics, the margin of improvement is narrower with a MCC of 0.28 compared to the TMhit value of 0.26. This is not unexpected given that both methods use SVM classifiers, though more significantly there is a considerable overlap of 42 sequences in training sets. Given that we assessed our method using leave-one-out cross-validation

whereas TMhit results were not cross-validated, TMhit results are likely to be overestimated therefore the actual margin of improvement may be larger. Compared to TMHcon, a recent neural network based approach, our method again performed well, with TMHcon results comparable to the globular protein contact predictors.

5.3.3 Helix-helix interaction prediction performance

We assessed performance of helix-helix interaction prediction requiring one residue from each helix to be in contact. Based on observed interactions there were comparable numbers of interacting and non-interacting helices for all contact definitions, with 668 and 733 respectively using contact definition 1. Results using the data set of 74 sequences and three contact definitions is shown in Table 5.3.

Method	**Contact Definition**	**Precision**	**Recall**	**FPR**	**FNR**	**MCC**	**Accuracy**
MEMPACK	1	0.93	0.10	0.0087	0.84	0.29	64.7%
SVMcon	1	0.57	0.11	0.090	0.84	0.11	59.3%
SVMcon L5	1	0.82	0.034	0.0074	0.95	0.13	59.5%
PROFcon	1	0.43	0.16	0.83	0.16	0.02	45.4%
PROFcon L5	1	0.72	0.11	0.043	0.84	0.19	62.0%
MEMPACK	2	0.95	0.11	0.0062	0.84	0.29	63.6%
TMhit L5	2	0.77	0.31	0.12	0.47	0.45	73.2%
MEMPACK	3	0.94	0.11	0.008	0.85	0.27	60.6%
TMHcon L5	3	0.49	0.32	0.37	0.63	0.02	52.3%

Table 5.3: Helix-helix interaction prediction performance using a data set of 74 sequences. Successful prediction of interacting helices requires one residue from each helix to be in contact. Results for contact definition 3 used 58 sequences that had more than 2 TM helices as TMHcon is unable to make predictions for 2 TM helix sequences.

Our method achieved similar scores using contact definitions 1 and 2, with a MCC of 0.29 and accuracies of 64.7% and 63.6%. Using contact definition 3, results were slightly lower with a MCC of 0.27 and accuracy of 60.6%. The FNR was consistent across all definitions at approximately 0.84. Compared to SVMcon and PROFcon, our method performed well with only PROFcon L5 approaching similar performance (MCC 0.19, accuracy 62.0%), suffering only from a higher FPR compared to our method. Other than PROFcon L5 which performed better than expected for a globular protein predictor, results were generally low with MCC values in the range 0.02-0.13. The performance of TMhit surpasses that of our method

with MCC 0.45 and accuracy 72.3%. However, as described above, the TMhit results were not cross-validated and are likely to be substantially overestimated given the overlap of 42 sequences in training sets. To give an estimate of the level of improvement this is likely to have resulted in, we scored our method in the absence of cross-validation for the 42 overlapping sequences and achieved scores of MCC 0.65 and accuracy 82.6%. We additionally compared the two methods using a smaller data set of 14 sequences for which both our method and TMhit results were fully cross-validated (Lo *et al.*, 2009). Requiring a single contacting pair of residues, our method achieved 66.3% accuracy (MCC 0.36) compared to 39.1% for TMhit (standard error 5%). TMHcon achieved MCC 0.02 and accuracy of 52.3%, which reflected the relatively poor performance in residue contact prediction, caused largely by a high FPR of 0.37.

5.3.4 Helical packing arrangement decoy discrimination performance

Using our decoy set, we were able to derive between 1 and 53 (average 18.5) unique helical packing arrangements for 71 sequences in our data set. By combining these with unique helical packing arrangements derived from the native crystal structure and homology models of the native crystal structure, we assessed performance of our and existing methods at discriminating the native or native model arrangements from decoy arrangements. Each arrangement was scored according to the number of interacting/non-interacting helices that matched the prediction from sequence, with interacting/non-interacting helices scored equally. Accuracy was determined by counting the frequency at which the native or native model arrangement achieved the highest score. As discriminating 2 TM helix arrangements, where helices are either interacting or not, is somewhat trivial, Table 5.4 shows results including and excluding 2 TM helix arrangements, where there are a total of 57 sequences with more than one unique packing arrangement.

Consistent with prediction of helix-helix interactions, our method performed

Method	Contact Definition	Accuracy (n=57)	Accuracy (n=71)
MEMPACK	1	68.4%	69.0%
SVMcon L5	1	52.6%	56.3%
PROFcon L5	1	45.6%	52.1%
MEMPACK	2	66.6%	67.6%
TMhit L5	2	59.6%	66.2%
MEMPACK	3	70.2%	70.4%
TMHcon L5	3	40.4%	-

Table 5.4: Helical packing arrangement decoy discrimination using a data set of 71 sequences with 2 or more TM helices (n=71) and a data set of 57 sequences with 3 or more helices (n=57). Accuracy reflects the frequency at which the native or native model helical packing arrangement achieved the highest score compared to the decoy set.

similarly using contact definitions 1 and 2, although unexpectedly performed best using contact definition 3 (70.4% accuracy). Excluding 2 TM helix proteins, using all contact definitions, performance decreased slightly suggesting that, on average, discriminating 2 TM helix arrangements is slightly easier than for other topologies. SVMcon and PROFcon both performed best when evaluated using their L5 modes although both achieved accuracies over 10% lower than our method. TMhit achieved a slightly lower score than our method (66.2%) though again in the absence of cross-validation. Excluding 2 TM helix proteins performance was almost 7% lower. TMHcon was not assessed using the complete set of 71 as it is unable to make predictions on 2 TM helix proteins, and performed below all other methods (40.4% accuracy) on the set of 57.

5.3.5 Assessing the accuracy of helical packing arrangements

Given that the generation of helical packing arrangements is based on the interconnection of vertices within a graph, accuracy is ultimately dependent on the detection of edges via prediction of helix-helix interactions. Out of the data set of 74 sequences, 17 (23%) had all interactions successfully predicted although in 3 of these cases there were no observed interactions between helices. Predicted arrangements were then compared by visual inspection of a two-dimensional slice

taken from the crystal structure approximately normal to the likely plane of the lipid bilayer, and assessed based on the overlap of helices from the predicted arrangement and the slice. Of these 17 cases, 9 arrangements produce overlaps for all TM helices and therefore can be considered as closely resembling the helix packing arrangement observed in the crystal structure.

Among these 9 correct cases, three 7 TM helix proteins (PDB: 1E12:A, 1XIO:A, 2F95:A) produced helical packing arrangements that clearly resembled their respective crystal structures (Figure 5.1). Additionally, for each of these cases the correct arrangement was successfully determined from alternatives by scoring arrangements based on the number of same-side loop crossovers. Overall, this function successfully identified the correct arrangement in 4 out of 6 cases where multiple arrangements were generated when tested using observed helix-helix interaction information; in the remaining 3 cases, 2 had an equal number of crossovers for each of the alternative arrangements (2HYD:A, 1XFH:A) - in these instances, the highest scoring arrangement was the one with the lowest total residue-residue contact distance resulting in one correct and one incorrect prediction, while in the remaining case the correct arrangement contained one more crossover than the incorrect arrangement (1XME:A).

Other cases where all helix-helix interactions were successfully predicted and packing arrangements closely resembled crystal structures included the 5 TM helix ubiquinol oxidase (1FFT:C, Figure 5.2) and 6 TM helix Aquaporin-4 (2D57:A). Below 4 TM helices, arrangements generally resembled crystal structures well although the task becomes more straightforward as the number of TM helices decreases. Where all helix-helix interactions were successfully predicted and packing arrangement resembled the crystal structure, application of a genetic algorithm to rotate helices around their respective Z-axes usually resulted in helix orientations that aligned significantly better with native structures compared to arbitrary degrees of rotation (Figure 5.3).

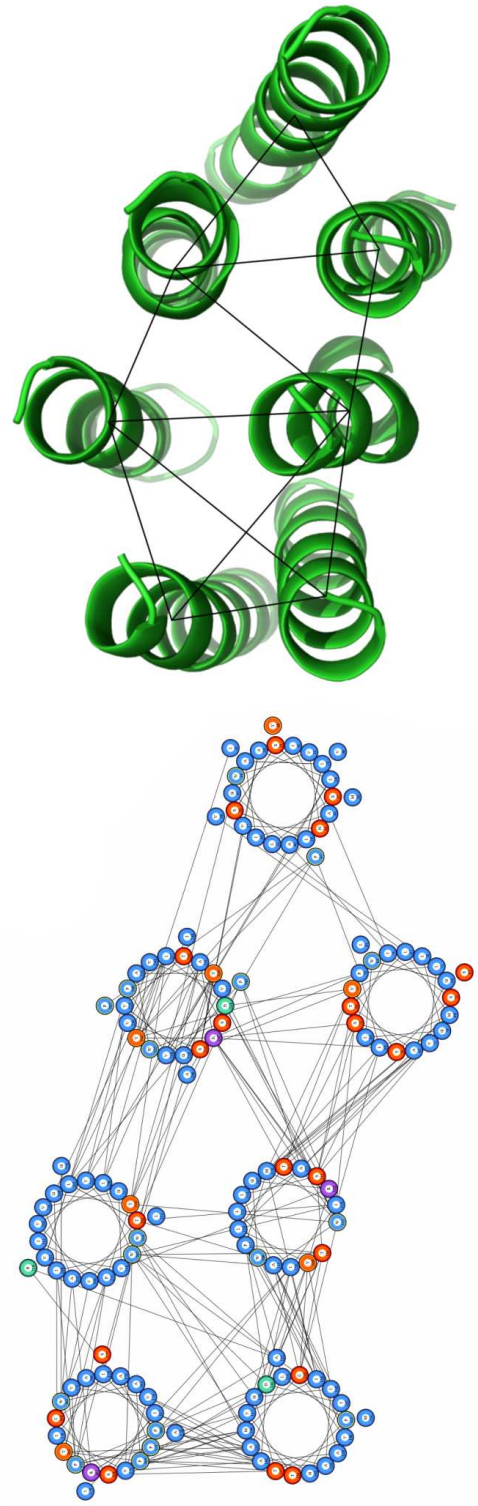

Figure 5.1: Predicted helical packing arrangement and crystal structure of Halorhodopsin (PDB: 1E12:A) from *Halobacterium salinarum*. In this example the two left-most helices share the same interactions. The correct arrangement has been identified as having no same-side loop crossovers, compared to one for the incorrect arrangement. Predicted residue-residue contacts are annotated on the packing arrangement while observed helix-helix interactions are annotated on the crystal structure.

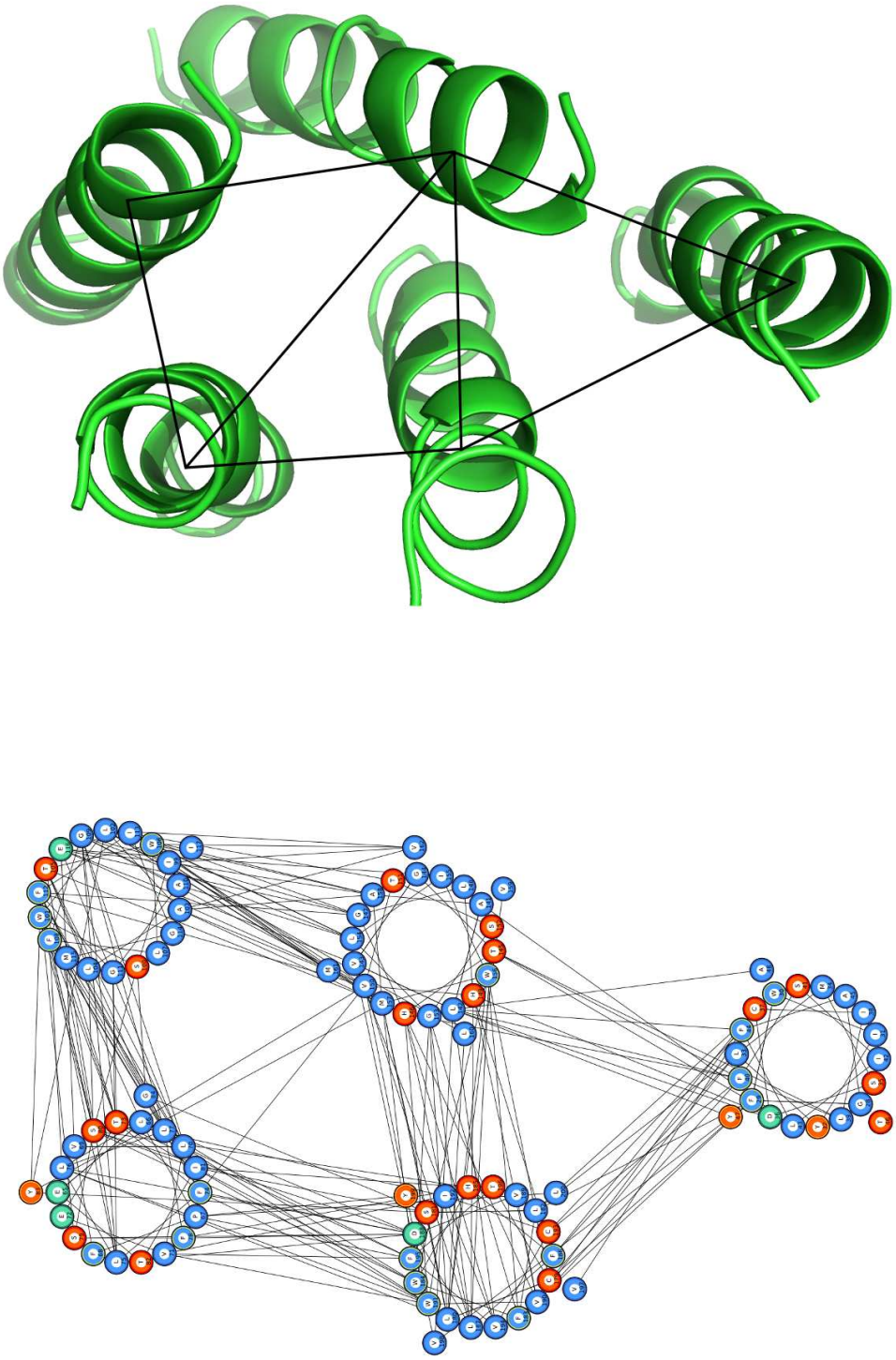

Figure 5.2: Predicted helical packing arrangement and crystal structure of Ubiquinol Oxidase (PDB: 1FFT:C) from *Escherichia coli*. Predicted residue-residue contacts are annotated on the packing arrangement while observed helix-helix interactions are annotated on the crystal structure.

When helices were connected consecutively, for example where a 3 helix protein has interactions between helices 1-2 and 2-3, the program was unable to determine the correct arrangement despite predicting all helix-helix interactions correctly. Under these circumstances, the algorithm defaults to a circular layout, which is frequently closest to the crystal structure as in the case of aquaporin (2D57:A) where helices are arranged around a central pore. In a number of cases though, the correct arrangement is much closer to linear as in the case of Photosystem II (2AXT:A) where there is significant interaction with additional chains in the complex. In such situations, the helix-helix interactions alone do not provide enough information to determine the correct arrangement.

Where prediction of helix-helix interactions falls below 100%, packing arrangements generally fail to accurately resemble crystal structures. In some cases, such as the ammonium transporter (2B2F:A), well connected sub-components of 3-5 TM helices were often correctly formed, but their arrangement in relation to each other was incorrect due to a number of missing helix-helix interactions. In three cases where there was substantial interconnection between TM helices, the arrangement does not succeed, most likely due to the algorithm encountering a local minimum. It is also impossible to generate an arrangement from a disconnected graph, where all helix-helix interactions are incorrectly predicted, which occurs in 12 sequences (16.2%). A summary of results where all interactions were correctly predicted is shown in Table 5.5.

While the successful packing arrangements were achieved with topologies of less than 8 TM helices, we additionally tested the algorithm using observed data to validate its effectiveness at generating arrangements for topologies with large numbers of TM helices using observed helix-helix interaction data rather than predicted contacts. In a number of cases, complex packing arrangements were generated with up to 13 TM helices that clearly resembled the respective crystal structure. Examples include the 8 TM helix proton glutamate symport protein (1XFH:A, Figure 5.4) 10

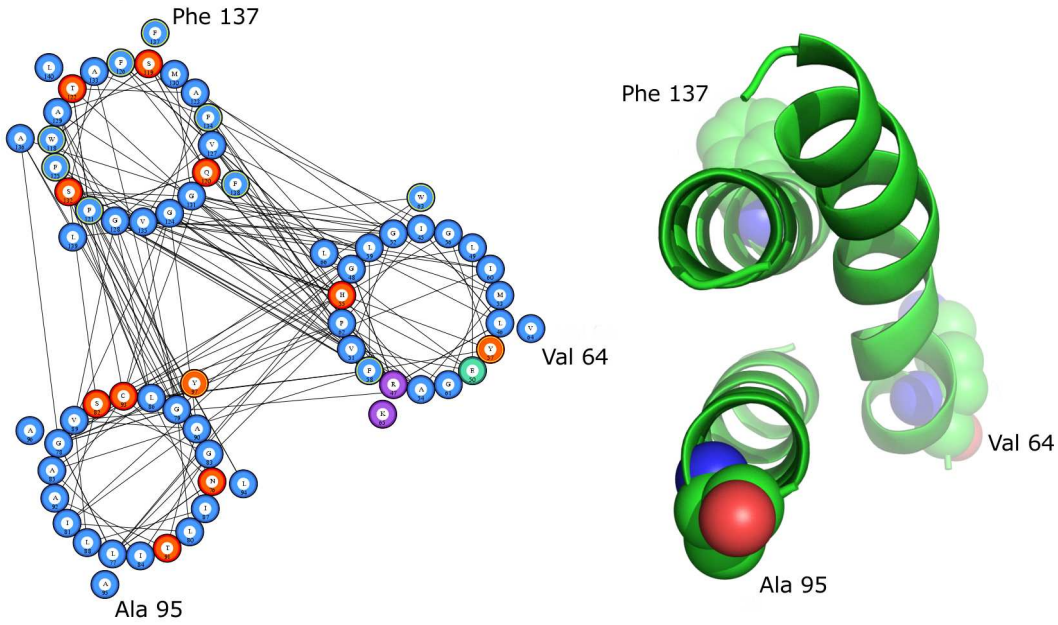

Figure 5.3: Predicted helical packing arrangement and crystal structure of Photosystem I chain D (PDB: 1JB0:L) from *Thermosynechococcus elongatus*. Application of a genetic algorithm to rotate helices about their Z-axes results in the correct positioning of residues Val64, Ala135 and Phe137.

TM helix proton ATPase (1MHS), 12 TM helix multidrug transporter (2GFP:A) and 13 TM helix cytochrome C oxidase (1XME:A, Figure 5.5), although in this case two helices that share the same helix-helix interactions are incorrectly replaced.

Helical packing arrangement prediction	**Count**
Resembles two-dimensional slice from crystal structure	9
No observed helix-helix interactions	3
Incorrect due to linear configuration	3
Incorrect helix placement	2

Table 5.5: Assessment of predicted helical packing arrangements for the 17 sequences where all interactions were successfully predicted. Arrangements were compared to a two-dimensional slice taken from the respective crystal structures and assessed based on the alignment between the helices in the predicted arrangement and in the slice; in 9 cases there was overlap for all helices (PDB: 2F95:A, 1E12:A, 1XIO:A, 2D57:A, 1FFT:C, 1JB0:L, 1C17:A, 1R3J:C, 2AHY:A). In 3 cases, there were no observed helix-helix interactions therefore no arrangement could be predicted (PDB: 1VCR:A, 1YQ3:D, 1ZOY:C). In 3 cases, the arrangement predicted a circular configuration whereas the correct arrangement was approximately linear (1DXR:M, 2AXT:D, 2AXT:A).

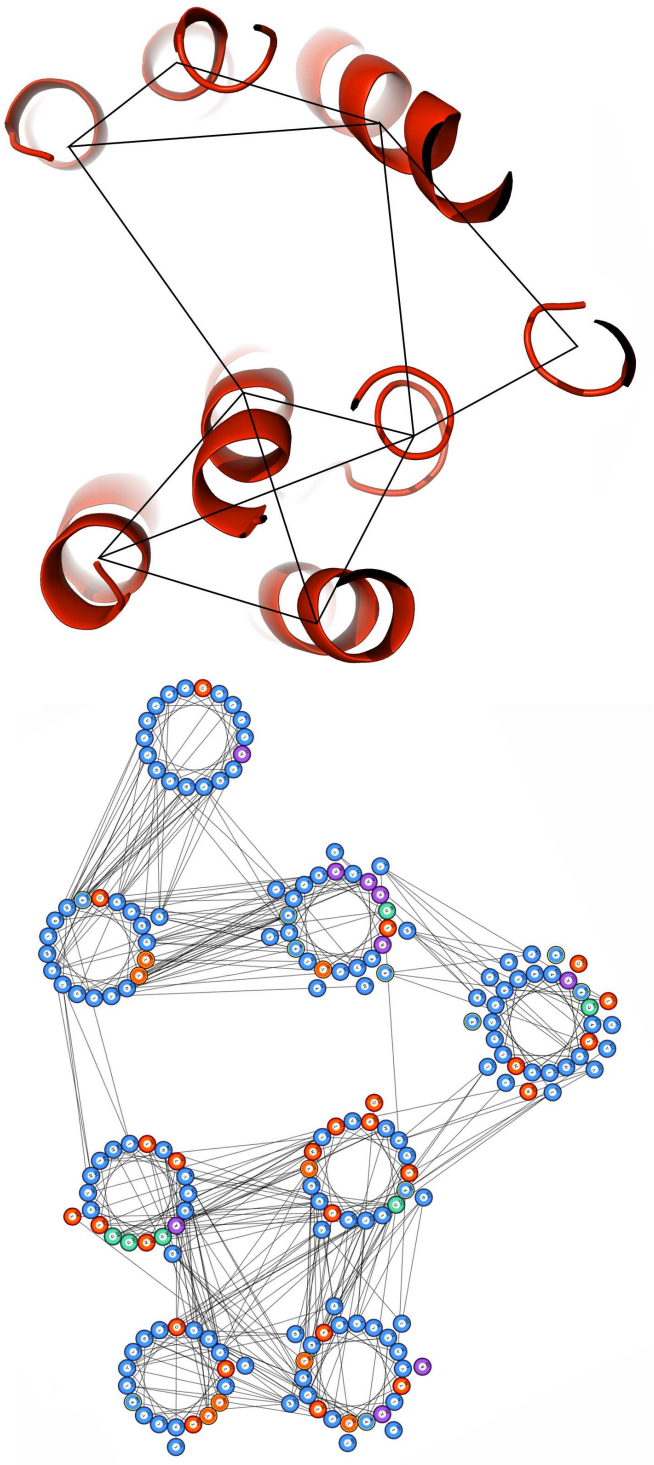

Figure 5.4: Helical packing arrangement and crystal structure of proton glutamate symport protein (PDB: 1XFH:A) from *Pyrococcus horikoshii*, generated using observed rather than predicted helix-helix interactions. Observed residue-residue contacts are annotated on the packing arrangement while observed helix-helix interactions are annotated on the crystal structure.

Figure 5.5: Helical packing arrangement and crystal structure of cytochrome C oxidase (PDB: 1XME:A) from *Thermus thermophilus*, generated using observed rather than predicted helix-helix interactions. Observed residue-residue contacts are annotated on the packing arrangement while observed helix-helix interactions are annotated on the crystal structure. In this example, the two helices at the bottom left of the arrangement are incorrectly placed; they share the same helix-helix interactions but the correct arrangement has one same-side loop crossover whereas the incorrect arrangement has none. The alternative correct arrangement where the placement of these two helices is reversed is returned as the second highest scoring arrangement.

5.4 Discussion

In this chapter we have implemented a novel tool capable of predicting lipid exposure, residue contacts and helix-helix interactions using SVM classifiers. These predictions are then combined to produce the optimal helical packing arrangement using a force-directed algorithm. Firstly, lipid exposure is predicted using evolutionary information labelled by data derived from coarse-grained molecular dynamics simulations. Solvent-exposed residues in both globular and TM proteins are known to be less conserved than buried residues, therefore non-conserved residues are more likely to identify lipid-exposed surfaces of TM helices (Wallin *et al.*, 1997; Stevens & Arkin, 2001). But in contrast to globular proteins, TM proteins do not show large differences in hydrophobicity between lipid-exposed and buried residues, making lipid exposure prediction a harder task (Elofsson & von Heijne, 2007). Using machine learning tools that have been successfully applied to TM protein topology prediction (Nugent & Jones, 2009), we were able to achieve per residue accuracy that compares favourably with a recent existing method suggesting the SVM is efficiently capturing the major distinguishing features of lipid exposure, the periodicity of conserved residues and the polarity of their side chains, from sequence profile data. Predictions may be useful for a number of additional applications including the modification of a TM protein-specific energy functions for *ab initio* modelling (Pellegrini-Calace *et al.*, 2003) where they could be incorporated into the potential, as for example ROSETTA (Rohl *et al.*, 2004) includes the LIPS score in its energy function, or added as an additional term with a separate weighting.

By combining predicted lipid exposure with sequence derived statistics and profile data centred on each residue in a pair, we were able to train an additional SVM to predict residue contacts. Recent methods specifically designed to predict residue contacts in TM proteins have used a variety of features including residue co-evolution scores, contact propensities and a range of global sequence-derived values. By experimenting with different combinations we attained optimal performance using a minimal set of features without the need for a consensus approach,

resulting in significant improvement compared to all existing methods. Our results demonstrate that globular protein contact predictors perform poorly when applied to TM proteins due to extremely high levels of false negative predictions. This is not especially surprising since the amino acid composition of hydrophobic globular protein alpha-helices has recently been shown to contrast from that of TM helices, therefore contact propensities are likely to differ. Generally, hydrophobic globular protein alpha-helices that are long enough to span the bilayer contain three or more charged residues with a relatively even distribution along their lengths, as well as a decreased frequency of occurrence of Ile and Val residues, while charged residues in TM helices tend to be concentrated towards helix termini (Cunningham *et al.*, 2009). Additionally, in the case of PROFcon, all TM proteins were removed from the data set so the neural network had received no training with TM protein data. Compared to the top performing TM protein contact predictor, our method achieves higher performance on all assessment metrics despite the lack of cross-validation of TMhit which was trained on a data set which included 42 sequences that are present in our test set. While our method produces a consistently low FPR, the FNR achieved a maximum score of 0.89. This result may suggest that our SVM is not sampling feature space effectively, although it is reasonable to suggest that many of these contacts are brought together as a consequence of strongly interacting residues that are correctly predicted. Studies of globular proteins have found that folds could be reconstructed using *ab initio* techniques and distance constraints to obtain native-like structures using between N/4 and N/8 restraints, where N is sequence length (Li *et al.*, 2004; Aszodi *et al.*, 1995), which supports the notion that the majority of contacts may be consequential. Ranked by average raw SVM score, the top five predicted contacts include Ala-Ser, Gly-Ile, Ile-Phe, Ala-Trp and Ala-Leu, which is broadly in line with previous observations of a relative enrichment of small and aromatic residues in packing interactions (Walters & DeGrado, 2006; Sal-Man *et al.*, 2007; Gimpelev *et al.*, 2004). Residue contacts involving a pair of charged residues occur in between 16 and 20 of the 74 proteins (depending on the contact definition), with most containing only a single charged

pair. Therefore they are relatively under-represented in the current data set. Out of 53 contacting charged pairs across all contact definitions, only 10 are correct, so compared to uncharged contacts they are poorly predicted by the SVM. Aside from a relative lack of training data, it is difficult to speculate on exactly why this is although most are side-chain to backbone interactions. Additional input features may therefore be required to improve prediction of charged residue pairs. However, contacts between some Arg-Asp and Arg-Glu pairs are predicted relatively strongly and are amongst the top 25 scoring predictions.

Helix-helix interaction results generally mirrored contact prediction performance, though globular protein contact predictors faired slightly better due to the relative ease of only having to predict a single residue contact for a successful helix-helix interaction, particularly when the FPR is reduced using the L5 mode, with PROFcon achieving 62.0% compared with 64.7% compared to our method. While difficult to compare accuracy using the entire test set of 74 sequences, the significant improvement of our method over TMhit when fully cross-validated on a smaller set of 14 sequences suggests state-of-the-art performance. While it is often difficult to successfully predict all helix-helix interactions correctly, the discrimination of decoy helical packing arrangements provides a measure of how well a method predicts enough interactions correctly to identify the native arrangement, a value which is usually below 100%. Results indicate that our method performs well, achieving up to 70.4% accuracy, aided by the fact that 50% of sequences have over 60% of their helix-helix interactions correctly predicted (contact definition 3). PROFcon, achieving only 52.1%, performs much worse than its helix-helix interaction prediction score would suggest, indicating that these successful interaction predictions are limited to a smaller number of sequences, and that prediction generalises poorly across a larger test set, while conversely SVMcon performs better than its interaction prediction score would suggest indicating better generalisation. Again it is difficult to accurately compare TMhit which achieves identical performance.

Using the helix-helix interaction results, helical packing arrangements were constructed using a force-directed algorithm. This task, which was ultimately dependent on the accuracy of predicted interactions, was successful for proteins with up to 7 TM helices, although errors occurred where helices were connected consecutively and even correct interaction data was insufficient to identify the correct arrangement. In these circumstances, interactions with additional chains are likely to play a role. For proteins where helix-helix interactions were not all correctly predicted, testing using observed interaction data showed that the algorithm is capable of constructing packing arrangements for proteins with up to 13 TM helices. These results suggest that where predicted helix-helix interactions can be supplemented with interaction data from experimental sources, for example mutagenesis studies, it may be possible to generate accurate packing arrangements for complex proteins containing large numbers of TM helices. This process would be assisted by the fast run time of the algorithm that will also allow alternative packing arrangements to be explored iteratively. Predictions can be used to generate pseudo three dimensional-structures with which loop regions can be built using programs such as SuperLooper (Hildebrand *et al.*, 2009). Models could then be used to pre-position residues prior to *ab initio* modelling therefore reducing conformational search space and reducing computational requirements.

While our results are encouraging, the paucity of structural data available for training purposes is likely to have limited residue contact and helix-helix interaction prediction performance, particularly as small data sets reduce tolerance to errors and the ability of SVMs to develop large generalisation bounds. Paradoxically, another problem may be the use of crystal structures to derive contact data, which provide only a snapshot of a protein at a given time therefore neglecting the inherent dynamic nature of TM proteins. TM proteins are known to exhibit significant conformational flexibility for a range of functions including modulation of catalytic activity and control of ionic flow, therefore labelling contacts according

to a single crystal structure will inevitably lead to training errors. Should enough data become available, it may be preferable to use ensembles of nuclear magnetic resonance structures in place of crystal structures, though due to the experimental difficulties in obtaining membrane protein structures this is unlikely to be an option in the near future. Another issue is the interaction between chains in multimeric complexes, which the majority of TM proteins in structural databases form. It is reasonable to expect that the interplay between chains in complexes has a degree of influence on the folding of individual chains, therefore satisfying these oligomeric interactions may lead to an improvement in the fold prediction of individual chains. Predicting oligomeric interactions would also allow TM protein quaternary structure to be predicted from sequence for the first time, while revealing the stoichiometry and symmetry of the complex.

Overall, our results demonstrate that residue contacts and helix-helix interactions can be used to accurately predict the helical packing arrangement of TM proteins, and discriminate native from decoy arrangements. This method can be used to gain insights into TM protein folding, while providing testable hypotheses for a variety of studies including protein design, mutagenesis and thermostability experiments, in addition to reducing conformational search space prior to *ab initio* modelling.

Chapter 6

Discussion

This chapter is divided into two parts. Firstly, the major contributions of this thesis to biology will be summarised, before future perspectives for TM protein structure prediction are discussed.

6.1 Biological discoveries

In Chapter 2, a six TM helix topology was proposed for the uncharacterised Batten disease protein, CLN3. It was demonstrated that adopting a consensus approach, in combination with the careful analysis of evolutionary data, allowed a topology model to be produced that agreed with all published experimental data. The model suggested CLN3 may contain a previously unrecognised amphipathic helix, with both termini located in the cytoplasm. Additionally, our findings suggested that orthologues of CLN3 might produce different topologies, possibly due either to atypical membrane or hydrophobic structures. Our strategy demonstrated a generic approach suitable for the analysis of TM proteins whose topologies are controversial - those where experimental data, as well as topology predictions by different algorithms, are conflicting. It may also have wider application to the prediction of topology for other TM proteins which may contain additional hydrophobic structures that do not span the membrane. This study serves to validate previous research that has demonstrated that structure prediction accuracy can be increased by using a consensus of prediction tools (Nilsson *et al.*, 2002; Ward, 2005). While the CLN3 model provides a basis for designing further experiments which may help validate the topology, the true function and mechanism of action may not be revealed until a CLN3 crystal structure becomes available. This highlights the challenges that remain for structural genomics initiatives, particularly with regard to TM proteins.

Chapter 3 described a strategy to use PROSITE motifs to guide TM protein topology prediction by modifying the MEMSAT3 algorithm. The method used PROSITE motifs that displayed a bias towards a particular topogenic region in TM proteins, identified by a chi-squared significance test. Motifs that occurred

in a specific topogenic region with significantly different frequencies compared to those expected at random were used to guide topology prediction in order to satisfy the topogenic biases of the matching motifs using weights determined using a GA, therefore increasing overall topology prediction accuracy. Using this strategy, an improvement of 6% prediction accuracy was possible using the Möller data set, corresponding to the correct prediction of an additional 11 sequences. Despite this improvement in prediction performance, there was a lack of correlation between the direction of the topogenic biases designated by the chi-squared significance test, and the sign and magnitude of the corresponding weights determined using the GA. The most likely reason for this discrepancy was that PROSITE motifs had matched a large number of false positive hits which had affected the chi-squared test, therefore resulting in incorrect topogenic biases. Despite this discrepancy, the use of PROSITE motifs to guide topology prediction did result in improved accuracy by providing additional information not fully captured by the NN employed by MEMSAT3. However, the high false positive rates of PROSITE motifs suggest that caution should be used when interpreting these PROSITE matches. While it is likely they may be identifying conserved residues that are already accounted for by the positive inside rule, they may also play a role in alternative undetermined biological functions.

In Chapter 4 we implemented a novel SVM-based TM protein topology predictor and showed that it could outperform a selection of the best performing prediction methods when fully cross-validated on a novel high resolution data set of 131 protein sequences. The method can also detect both signal peptides and re-entrant helices, while an additional program was able to achieved extremely low false positive and false negative rates for TM/globular protein discrimination. These tools demonstrate for the first time that SVM are well suited to TM protein topology prediction, an area previously dominated by HMM and NN-based machine learning approaches. Despite the strong performance of the method, it is impossible to determine whether SVMs outperform either HMMs or NNs without training and

cross-validating each algorithm using the same protocol on exactly the same data set. The recent SPOCTOPUS method (Viklund *et al.*, 2008) which uses HMMs and NNs achieved comparable prediction performance on a highly similar data set. This indicates that separation of the classes by the layer of hidden units used to solve non-linear problems by NNs achieves comparable performance to the separation by the SVM kernel function achieved by projecting the data into high dimensional feature space. Additionally, the learning bias used by SVMs to select a model of the training data may achieve equal generalisation to the learning bias used to train NNs. However, our method provides another machine learning algorithm which will undoubtedly supplement existing approaches, particularly when used in a consensus as was demonstrated in Chapter 2. Using MEMSAT-SVM, we estimated the fraction of TM proteins, re-entrant helices and signal peptides in a number of complete genomes. We determined that a typical genome contains between 24% and 33% TM proteins with *Caenorhabditis elegans* and *Takifugu rubripes* having a noticeably higher fraction. In all species we analysed, re-entrant helices were detected in between 2% and 3% of TM proteins, although prediction of these features remains a difficult task and will likely remain so until further training examples become available. Signal peptides prediction was more successful, with these features detected in between 23% and 25% of sequences. Topology prediction results illustrated consistent trends across all species, with significant peaks in eukaryotes at 7 TM helices representing GPCRs and 12 TM helices representing transporters proteins. In all species, the most dominant topology is a single TM helix. Overall, our results suggest that this new method, MEMSAT-SVM, is ideally suited to whole genome annotation of alpha-helical TM proteins.

In Chapter 5 we described a novel method capable of predicting lipid exposure, residue contacts and helix-helix interactions, again using SVM classifiers. By combining these predictions, we were able to generate the optimal helical packing arrangement using a force-directed algorithm. The successful prediction of lipid exposure using evolutionary information labelled by data derived from

coarse-grained molecular dynamics simulations provides a tool which may be useful for purposes such as *ab initio* modelling, while validating the effectiveness of SVMs at capturing the major distinguishing features of lipid exposure and demonstrating the potential of molecular dynamics simulations. Our analysis of contact prediction methods demonstrated that globular protein contact predictors perform poorly when applied to TM proteins, most likely due to the differing amino acid composition of hydrophobic globular protein alpha-helices and TM helices. Our tool was able to achieve higher performance on all assessment metrics compared to these methods as well as the top performing TM contact predictors, despite their lack of cross-validation. However, the relatively low sensitivity of our method suggests that our SVM may not be sampling feature space effectively, or that many contacts are brought together as a consequence of strongly interacting residues that are correctly predicted. Raw SVM scores suggested that small and aromatic residues are primarily involved in packing interactions, broadly in line with previous observations, while interactions involving charged residue pairs were relatively rare. Performance of helix-helix interaction generally mirrored contact prediction performance, with prediction accuracy on a small test set demonstrating state-of-the-art performance compared to all existing methods. Using this information and a force-directed algorithm, we were then able to predict the helical packing arrangement for proteins with up to 7 TM helices, or 13 TM helices using observed data. This task remains difficult and the limited success in many cases indicates that interactions with additional chains are likely to play a role in packing, while highlighting the paucity of structural data available for training purpose. The success using observed data suggests that, where predicted helix-helix interactions can be supplemented with interaction data from experimental sources, it may be possible to generate accurate packing arrangements for complex proteins containing large numbers of TM helices.

6.1.1 Future perspectives for transmembrane protein structure prediction

While this thesis has demonstrated that machine learning can be successfully applied to TM protein structure prediction, a number of other predictive tasks remain which may further enhance our ability to predict three-dimensional structure from sequence alone. In combination with the prediction of topology, residue contacts, helix-helix interactions and features such as signal peptides and re-entrant helices, realising these challenges may help to improve TM protein structure prediction to the point where it becomes useful for the application of medicinal chemistry. Among these challenges are the prediction of pore-forming regions and oligomeric interactions in alpha-helical TM proteins.

6.1.2 Prediction of pore-forming regions in alpha-helical transmembrane proteins

Ion channels are TM proteins that regulate the movement of specific ions across the membrane by facilitating ionic flow down electrochemical gradients (Figure 6.1). They play an important role in a number of cell types and occur as large families of related genes with cell and tissue specific expression patterns. Many common diseases including diabetes, hypertension, cardiac arrhythmias, angina pectoris and epilepsy have been related to ion channel dysfunction, therefore ion channels represent one of the most important classes of protein for pharmaceutical intervention. Frequently, pore-lining TM helices are enriched with charged residues, thus facilitating passage of the cognate ion through the channel. However, many TM proteins that are not ion channels contain charged residues within the TM region that are used to stabilise helix-helix interaction, for example via formation of salt bridges, thus the presence of charged residues alone cannot be used to discriminate pore-forming regions.

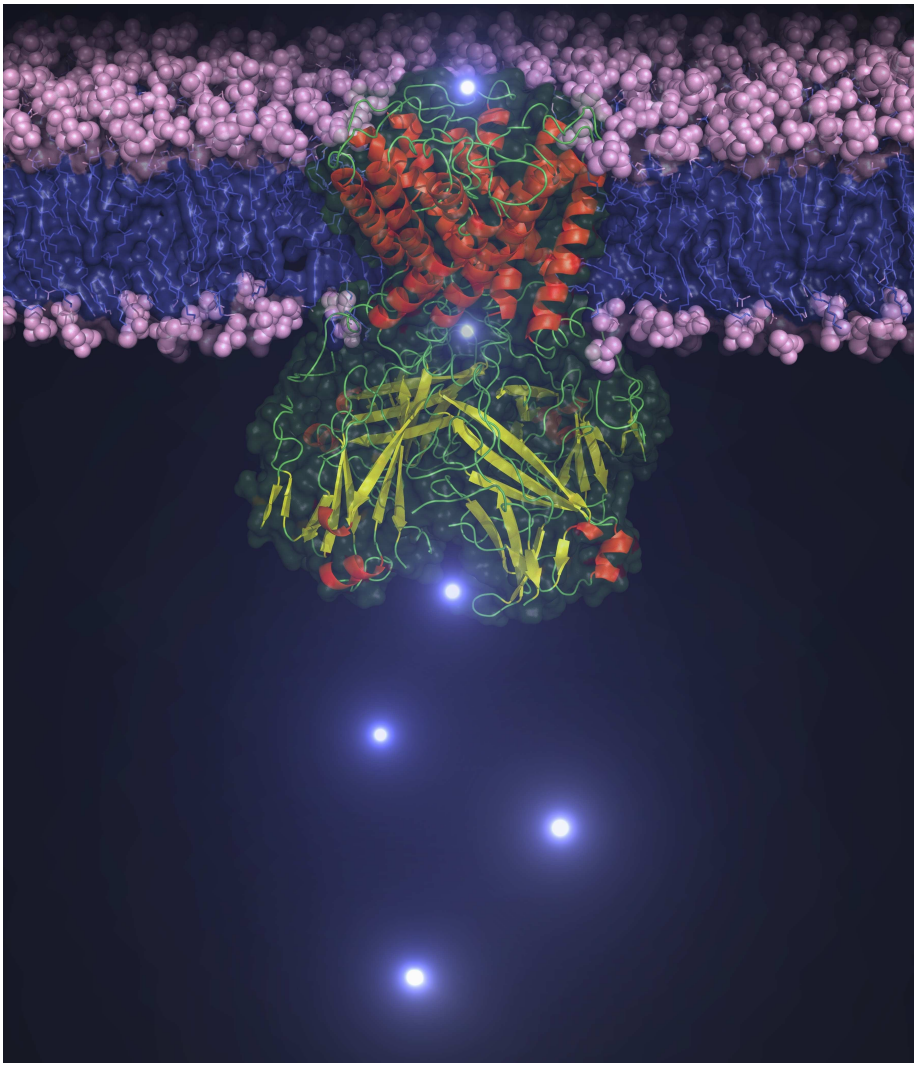

Figure 6.1: Potassium channel KcsA from *Streptomyces lividans* (PDB: 1R3J:A).

By taking advantage of a number of recent methods that allow the identification of pore-lining residues in TM protein crystal structures, it may be possible to use a machine learning approach to predict pore-lining residues within a TM protein. Methods such as HOLLOW (Ho & Gruswitz, 2008) and Pore-Walker (Pellegrini-Calace *et al.*, 2009) allow the identification of the pore centre and pore axis using geometric criteria, allowing the biggest and longest cavity through the channel to be detected. Pore features, including diameter profiles, pore-lining residues, size, shape and regularity can then be calculated.

By labelling pore-lining residues using such methods, training and test sets could be assembled before a supervised learning approach is employed to predict

the likelihood of a TM protein being involved in pore formation (Figure 6.2A). Using the same data set, further classifiers could be developed to predict additional features including the pore size and specific ion type the channel is capable of transporting. Given the success of the SVM-based approaches used in Chapters 4 and 5, this learning algorithm may again prove successful, although HMMs, NNs, and possibly consensus approaches may also perform well. Alternative machine learning algorithms including Adaptive Boosting have recently demonstrated state of the art performance in other areas of computer science, so an assessment of the performance of such methods may also be useful.

When used in conjunction with a whole genome scan for TM proteins and subsequent topology prediction, such a predictive tool has the potential to identify novel ion channels, the discovery of which may be of substantial biochemical and pharmacological significance. From a structural modelling perspective, identification of pore-forming regions may provide insight into quaternary structure geometry and provide information for *ab initio* methods to model such regions so that they are solvent accessible rather than lipid embedded. Furthermore, site-directed mutagenesis of predicted pore-lining residues may allow modification of ionic specificity, providing valuable insight into protein design.

6.1.3 Modelling alpha-helical transmembrane protein quaternary structure from sequence using oligomeric interactions

As discussed in Chapter 5, despite significant efforts to predict TM protein topology, relatively little attention has been directed toward predicting the fold of TM proteins. While methods such as that described now exist to predict interactions within a single protein chain, none are able to model the interaction between chains in multimeric complexes, which the majority of TM proteins in the PDB form. It is reasonable to expect that interplay between chains in complexes has a

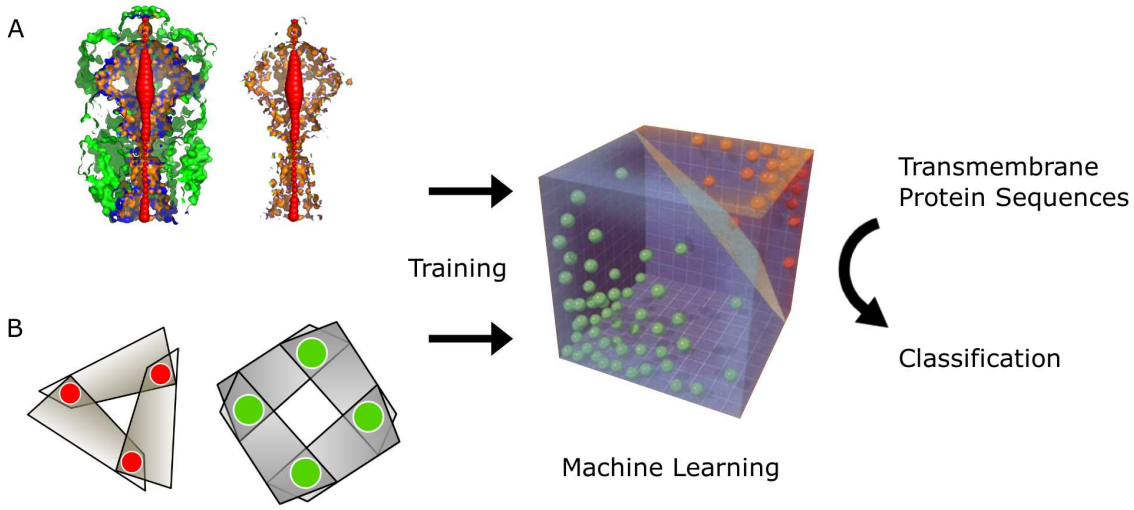

Figure 6.2: (A) Residues defined as pore-lining by programs such as HOLLOW and Pore-Walker (orange) are used to train a machine learning algorithm. Transmembrane protein sequences predicted to contain pore-lining regions can then be detected. (B) Residues within transmembrane protein complexes that form oligomeric interactions are used to train a machine learning algorithm (red and green). Transmembrane protein sequences likely to form complexes can then be predicted, and the interacting protein(s) identified.

degree of influence on the folding of individual chains, therefore satisfying these oligomeric interactions may lead to an improvement in single chain fold prediction by constraining conformational search space. It may therefore be possible to model alpha-helical TM protein quaternary structure from sequence by predicting oligomeric interactions (Figure 6.2B).

The prediction of oligomeric interactions is a natural progression of the work to predict the fold of single TM protein chains using residue-residue contacts. By assembling a data set to include all non-redundant TM protein complexes, a machine learning approach could be used to predict residue-residue contacts involved in oligomeric interactions, before using these predictions to assemble a model of TM protein quaternary structure. It might also be possible include interactions between TM and peripheral membrane proteins.

In combination with topology and single chain fold prediction, this method would allow TM protein quaternary structure to be predicted from sequence for the first time, while revealing the key residues required for oligomeric interaction

and the stoichiometry and symmetry of the complex. Further information used to construct the complex could be provided should a pore-forming region be detected. Should the complex bind a ligand, residues composing a binding site created at the interface of multiple subunits could be revealed possibly identifying sites for pharmaceutical intervention. Such models could also provide testable hypotheses for a variety of studies including protein design, mutagenesis and thermostability experiments.

In summary, while the most successful bioinformatic methods for protein structure prediction will continue to be founded on a solid understanding of the underlying biology, machine learning provides powerful tools with which to supplement experimental techniques. As new algorithms are developed and crystal structure databases expand, advances in this field will help to push TM protein structure prediction to ever increasing resolutions, to the point where such methods begin to have a significant impact on human health and disease.

Appendices

Appendix A

List of abbreviations

Abbreviation	Details
ATP	Adenosine Triphosphate
BLAST	Basic Local Alignment Search Tool
BLOSUM	Blocks Substitution Matrices
CASP	Critical Assessment of Methods of Protein Structure Prediction
CATH	Class, Architecture, Topology, Homologous Superfamily
DNA	Deoxyribonucleic Acid
ER	Endoplasmic Reticulum
FN	False Negative
FP	False Positive
GA	Genetic Algorithm
GPCR	G-Protein-Coupled Receptor
HMM	Hidden Markov Model
JNCL	Juvenile Onset NCL
LOOCV	Leave-One-Out Cross-Validation
MCC	Matthews Correlation Coefficient
MFS	Major Facilitator Superfamily
MIP	Major Intrinsic Protein
NCBI	National Center for Biotechnology Information
NCL	Neuronal Ceroid Lipofuscinose
NN	Neural Network
NMR	Nuclear Magnetic Resonance
PDB	Protein Data Bank
PSI-BLAST	Position Specific Iterated-BLAST
PSSM	Position-Specific Scoring Matrix
RMSD	Root Mean Squared Deviations
SMART	Simple Modular Architecture Research Tool
SRP	Signal Recognition Particle
SVM	Support Vector Machine
TM	Transmembrane

Table A.1: List of Abbreviations.

Appendix B

Data sets

PDB	SWISS-PROT	N-terminus	Helices	Topology
1AFO:A	GLPA_HUMAN	out	1	92,112
2OCC:X	COX7B_BOVIN	in	1	38,56
1RZH:H	RCEH_RHOSH	out	1	12,32
1NKZ:A	LHA4_RHOAC	in	1	18,37
1NYJ:D	M2_IAFOW	out	1	23,43
2OCC:Z	COX81_BOVIN	in	1	40,59
2OCC:Y	COX7C_BOVIN	in	1	38,55
2OCC:Q	COX41_BOVIN	in	1	100,121
2OCC:V	COX6C_BOVIN	in	1	16,33
2OCC:W	CX7A1_BOVIN	in	1	55,75
2OCC:T	CX6A2_BOVIN	in	1	30,48
1QLE:D	COX4_PARDE	in	1	25,46
1Q90:R	UCRIA_CHLRE	in	1	43,66
1P84:I	UCR9_YEAST	in	1	17,31
1SQX:K	UCR11_BOVIN	in	1	18,37
1RHZ:C	SECG_METJA	in	1	32,50
1RHZ:B	SECE_METJA	in	1	37,63
2AXT:H	PSBH_SYNEL	in	1	27,48
2AXT:F	PSBF_SYNEL	in	1	18,41
2AXT:E	PSBE_SYNVU	in	1	18,38
1ZLL:E	PPLA_RAT	in	1	28,51
1WRG:A	LHB_RHORU	in	1	21,40
1XRD:A	LHA_RHORU	in	1	13,33
1LGH:J	LHA_RHOMO	in	1	22,40
1KQG:B	FDNH_SHIFL	out	1	256,277
1Q90:A	CYF_CHLRE	out	1	283,305
1SQX:D	CY1_BOVIN	out	1	291,308

Table B.1: Crystal structure data set. Column 1: PDB chain ID. Column 2: SWISS-PROT ID. Column 3: Location of N-terminus. Column 4: Number of transmembrane helices. Column 5: Transmembrane helix boundaries, in relation to SWISS-PROT sequence.

PDB	SWISS-PROT	N-terminus	Helices	Topology
1XME:B	COX2_THET8	in	1	15,34
1PJE:A	VPU_HV1LW	out	1	9,23
1JB0:F	PSAF_SYNEN	out	1	83,106
1KB9:H	UCRQ_YEAST	in	1	55,72
1L0L:E	UCRI_BOVIN	in	1	112,134
1L0L:G	UCRQ_BOVIN	in	1	47,60
1ZZA:A	SNN_HUMAN	out	1	8,28
2AXT:K	PSBK_SYNEL	out	1	17,41
1IFI:A	COATB_BPFD	out	1	38,59
2E74:G	PETG_MASLA	out	1	5,26
1EHK:C	COXA_THET8	in	1	7,28
2HAC:A	CD3Z_HUMAN	out	1	31,51
2J58:A		out	1	325,353
1Q90:L	PETL_CHLRE	out	1	16,36
1JB0:I	PSAI_SYNEL	out	1	9,32
1JB0:M	PSAM_SYNEL	out	1	7,27
1JB0:X		in	1	5,25
1B9U:A	ATPF_ECOLI	out	1	6,30
1BA4:A	A4_HUMAN	out	1	688,708
2AXT:J	PSBJ_SYNEL	in	1	10,30
2AXT:L	PSBL_SYNEL	in	1	14,35
2AXT:M	PSBM_SYNEL	out	1	7,27
2AXT:T	PSBT_SYNEL	out	1	4,22
2FYN:B		out	1	229,248
2FYN:C	UCRI_RHOSH	in	1	14,35
2AXT:I	PSBI_SYNEL	out	1	3,25
1Q90:N	PETN_CHLRE	out	1	13,34
1RKL:A	OST4_YEAST	out	1	8,29

Table B.2: Crystal structure data set. Column 1: PDB chain ID. Column 2: SWISS-PROT ID. Column 3: Location of N-terminus. Column 4: Number of transmembrane helices. Column 5: Transmembrane helix boundaries, in relation to SWISS-PROT sequence.

PDB	SWISS-PROT	N-terminus	Helices	Topology
1IIJ:A	ERBB2_RAT	out	1	653,677
1S5L:X		out	1	9,23,
2OAR:A	MSCL_MYCTU	in	2	19,39,70,89
1FFT:B	CYOA_ECOLI	out	2	45,66,89,108
1C17:A	ATPL_YERPE	out	2	7,31,53,77
1R3J:C	KCSA_STRLI	in	2	25,47,86,111
2OCC:O	COX2_BOVIN	out	2	29,46,59,76
1P49:A	STS_HUMAN	out	2	182,206,213,236
2AXT:Z	PSBZ_SYNEL	out	2	3,27,38,58
1M57:H	COX2_RHOSH	out	2	59,81,98,118
2F95:B	HTR2_NATPH	in	2	24,40,62,81
2IUB:G	CORA_THEMA	in	2	296,313,326,344
1JB0:K	PSAK_SYNEN	out	2	21,32,57,76
1XL6:A	Q2W6R1_MAGMM	in	2	71,96,133,156
1YCE:L	ATPL_PROMO	out	2	16,39,54,80
1YEW:A	Q49104_METCA	out	2	190,207,234,250
2AHY:A	Q81HW2_BACCR	in	2	24,45,74,95
1LNQ:A	MTHK_METTH	in	2	22,38,72,95
1KF6:D	FRDD_SHIFL	in	3	13,42,62,87,95,117
1KF6:C	FRDC_ECOLI	in	3	27,48,67,88,110,130
1VCR:A	CB22_PEA	in	3	101,122,161,179,216,237
1JB0:L	PSAL_SYNEL	in	3	45,66,76,97,118,140
1Q90:D	PETD_CHLRE	in	3	32,57,95,117,127,146
2OAU:A	MSCS_SHIFL	out	3	32,58,70,89,91,104
1NEK:C	DHSC_ECOLI	in	3	25,51,68,93,109,128
1NEK:D	DHSD_SHIFL	in	3	17,40,55,77,88,113
1YQ3:D	DHSD_HUMAN	in	3	67,85,89,111,122,145
1ZOY:C	C560_HUMAN	in	3	70,91,110,136,151,168

Table B.3: Crystal structure data set. Column 1: PDB chain ID. Column 2: SWISS-PROT ID. Column 3: Location of N-terminus. Column 4: Number of transmembrane helices. Column 5: Transmembrane helix boundaries, in relation to SWISS-PROT sequence.

PDB	SWISS-PROT	N-terminus	Helices	Topology
2CYD:J	NTPK_ENTHR	out	4	16,41,55,78,93,117,129,154
1KQG:C	FDNI_SHIFL	in	4	13,36,52,75,112,134,150,174
2D2C:N	CYB6_MASLA	in	4	33,53,88,109,115,134,185,205
2BG9:A	ACHA_TORMA	out	4	236,258,268,289,300,321,433,457
2BG9:E	ACHG_TORCA	out	4	239,260,270,291,303,326,467,491
2HI7:B	DSBB_ECOLI	in	4	15,35,48,63,72,85,145,161,
1RZH:L	RCEL_RHOSH	in	5	31,53,85,106,116,138,171,192,231,251
1FFT:C	CYOC_SHIFL	in	5	28,48,65,84,99,117,141,162,179,200
1DXR:M	RCEM_RHOVI	in	5	53,74,111,132,144,165,199,219,265,284
2AXT:A	PSBA1_SYNEN	in	5	32,53,114,134,142,160,196,218,271,292
2BS4:F	FRDC_WOLSU	in	5	31,49,77,95,128,149,169,187,211,232
1Q16:C	NARI_ECOLI	out	5	3,26,50,71,85,110,127,147,182,200
2AXT:D	PSBD_PROHO	in	5	31,52,109,130,140,158,195,216,265,287
1LDI:A	GLPF_ECOLI	in	6	11,31,41,59,86,107,145,166,179,195,233,251
2ABM:H	AQPZ_SHIFL	in	6	4,26,34,55,81,102,131,152,161,178,201,223
2F2B:A	AQPM_METTM	in	6	8,25,56,73,100,118,146,162,175,191,222,240
2D57:A	AQP4_RAT	in	6	34,56,70,88,112,136,156,178,189,203,231,252
2C3E:A	ADT1_BOVIN	out	6	14,37,73,87,116,142,169,193,215,238,266,287
2AXT:B	PSBB_ANASP	in	6	19,39,95,115,138,159,197,218,234,255,451,472
2AXT:C	PSBC_SYNY3	in	6	48,69,111,132,155,176,234,252,267,288,425,445
2HYD:A	Q1Y946_STAAU	in	6	13,37,60,85,136,159,161,182,243,266,282,304
2IC8:A	GLPG_ECOLI	in	6	95,114,148,163,171,192,201,213,228,242,251,268
2ONK:C		in	6	3,33,47,77,83,97,128,151,182,198,230,251,
2BRD:A	BACR_HALSA	out	7	22,43,57,75,93,110,121,140,145,166,187,209,214,235
1GZM:A	OPSD_BOVIN	out	7	38,63,72,96,109,133,153,172,202,224,253,274,286,309
1QLE:C	COX3_PARDE	in	7	15,35,49,67,88,107,138,159,172,193,211,232,246,271
2F95:A	BACS2_NATPH	out	7	3,24,38,56,70,87,98,117,122,142,163,180,190,211
1E12:A	BACH_HALSA	out	7	27,50,63,82,106,122,134,153,159,179,201,221,227,248

Table B.4: Crystal structure data set. Column 1: PDB chain ID. Column 2: SWISS-PROT ID. Column 3: Location of N-terminus. Column 4: Number of transmembrane helices. Column 5: Transmembrane helix boundaries, in relation to SWISS-PROT sequence.

PDB	SWISS-PROT	N-terminus	Helices	Topology
1XIO:A	Q8YSC4_ANASP	out	7	3,26,35,56,70,89,99,121,127,148,168,185,195,218
1YEW:B	Q607G3_METCA	in	7	23,45,63,79,88,103,125,137,141,156,200,211,225,239
2FEE:B	CLCA_ECOLI	in	8	36,61,80,97,215,232,252,277,289,306,334,350,357,368,422,438
1L7V:B	BTUC_ECOLI	in	8	15,32,56,76,92,107,114,132,146,164,191,206,276,296,305,321
1XFH:A	O59010_PYRHO	in	8	13,30,38,63,82,106,128,160,200,218,232,251,300,320,390,409
2FYN:A	CYB_RHOSH	in	8	42,63,95,117,126,144,195,215,248,266,329,348,362,380,390,409
2C8L:A	AT2A1_RABIT	in	10	60,77,89,104,259,274,291,306,763,780,789,807,834,854,896,915,932,949,966,986
1RHZ:A	SECY_METJA	in	10	30,43,76,88,110,129,138,158,169,187,210,227,256,276,313,333,382,397,399,411
1MHS:A	PMA1_NEUCR	in	10	113,134,144,164,294,313,320,341,690,713,717,737,760,783,795,813,827,843,854,875,
1JB0:A	PSAA_SYNEN	in	11	72,93,159,179,193,216,298,314,352,374,392,413,439,461,536,558,591,612,674,691,725,745
1XQF:A	AMTB_ECOLI	out	11	34,54,66,90,123,141,147,170,186,200,222,241,249,270,282,300,303,321,335,355,373,399
2B2F:A	O29285_ARCFU	out	11	8,27,39,57,90,106,114,137,153,167,189,208,219,237,246,263,269,287,301,319,338,363
1QLE:A	COX1B_PARDE	in	12	31,55,89,113,129,151,178,201,220,247,269,291,307,326,338,360,371,394,406,429,443,465,487,509
1PW4:A	GLPT_ECOLI	in	12	31,52,65,87,93,112,121,141,159,179,189,207,253,277,290,313,321,341,348,369,385,407,415,437
1ZCD:B	NHAA_ECOLI	in	12	12,30,59,80,98,116,122,140,156,174,181,200,206,219,223,236,254,271,291,310,329,350,358,379
1T9Y:A	ACRB_ECOLI	in	12	10,27,340,358,363,386,395,413,440,457,470,492,539,556,873,892,896,918,927,945,974,991,1003,1022
2A65:A	O67854_AQUAE	in	12	15,32,43,63,92,124,166,184,194,211,243,265,278,298,339,364,378,395,399,422,448,469,483,501
2CFP:A	LACY_ECOLI	in	12	10,34,45,66,75,99,105,127,143,162,167,186,221,244,257,280,288,307,313,334,349,370,381,399
2GFP:A	EMRD_ECOLI	in	12	10,35,43,63,72,92,97,118,133,154,158,176,207,229,236,260,267,284,290,310,326,348,356,378
1XME:A	COX1_THET8	in	13	22,44,67,90,104,125,143,162,186,210,224,248,267,280,293,314,347,367,379,402,420,441,465,490,527,547

Table B.5: Crystal structure data set. Column 1: PDB chain ID. Column 2: SWISS-PROT ID. Column 3: Location of N-terminus. Column 4: Number of transmembrane helices. Column 5: Transmembrane helix boundaries, in relation to SWISS-PROT sequence.

Appendix C

Evaluation metrics

The following table summarises the evaluation metrics used in this thesis.

Metric	**Formula**
Sensitivity or True Positive Rate (TPR)	$TPR = TP/P = TP/(TP + FN)$
False Positive Rate (FPR)	$FPR = FP/N = FP/(FP + TN)$
Accuracy (ACC)	$ACC = (TP + TN)/(P + N)$
Specificity or True Negative Rate (TNR)	$TNR = TN/N = TN/(FP + TN) = 1 - FPR$
Precision or Positive Predictive Value (PPV)	$PPV = TP/(TP + FN)$
Negative Predictive Value (NPV)	$NPV = TN/(TN + FN)$
False Discovery Rate (FDR)	$FDR = FP/(FP + TP)$
Matthews Correlation Coefficient (MCC)	$MCC = \frac{(TP \cdot TN - FP \cdot FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$

Table C.1: Evaluation metrics. T = True. F = False. P = Positive. N= Negative.

Appendix D

Publications

The following chronological list contains peer-reviewed publications that I have authored during the course of my doctoral studies. In the papers where I am listed as first author, the contributions of the other authors are described. All other work was carried out by myself. The papers where I am listed as a secondary author include a summary of my contribution to the work.

Nugent, T., Ward, S. & Jones, D.T. (2010). The MEMSAT alpha-helical transmembrane protein structure prediction server. *Bioinformatics*, Submitted.

Web server implementation by SW. Manuscript was prepared by TN and was read and approved by DTJ.

Buchan, D.W., Ward, S.M., Lobley, A.E., **Nugent, T.**, Bryson, K. & Jones, D.T. (2010). Protein annotation and modelling servers at University College London. *Nucleic Acids Res*, **38**, W563-W568.

Web server implementation by DWB and SW. TN developed MEMSAT-SVM for transmembrane protein topology prediction.

Nugent, T. & Jones, D.T. (2010). Predicting transmembrane helix packing arrangements using residue contacts and a force-directed algorithm. *Plos Comp*

Bio, **10**, 159.

DTJ provided direction for computational aspects of the algorithm and biological/biophysical insight into aspects of membrane protein structure. Manuscript was prepared by TN and was read and approved by DTJ.

Nugent, T. & Jones, D.T. (2009). Transmembrane Protein Topology Prediction using Support Vector Machines. *BMC Bioinformatics*, **6**, e1000714.

Original source code was developed by DTJ. This was re-written and extended by TN. DTJ provided direction for computational aspects of the algorithm and biological/biophysical insight into aspects of membrane protein structure. Manuscript was prepared by TN and was read and approved by DTJ.

Rigden, D., ed. (2009). *From Protein Structure to Function with Bioinformatics.* Springer.

Chapter entitled 'Membrane Protein Structure Prediction' was prepared by TN and was read and approved by DTJ.

Lobley, A.E., **Nugent, T.**, Orengo, C.A. & Jones, D.T. (2008). FFPred: an integrated feature-based function prediction server for vertebrate proteomes. *Nucleic Acids Res*, **36**, W297-W302.

Source code for rendering of transmembrane protein topology predictions written by TN.

Nugent, T., Mole, S.E., Jones, D.T. (2008). The transmembrane topology of Batten disease protein CLN3 determined by consensus computational prediction

constrained by experimental data. *FEBS Lett*, **582**, 1019-24.

DTJ provided direction for computational aspects and biological/biophysical insight into aspects of membrane protein structure. SM provided direction for biological insight. Manuscript was prepared by TN and SM and was read and approved by DTJ.

Appendix E

Acknowledgements

I would like to thank all the members of the Bioinformatics Group, past and present, for their friendship and support over the previous three years. I would specifically like to thank my two supervisors; David Jones for providing a stimulating working environment and for demonstrating the benefits of being both rigorous and pragmatic, and Kevin Bryson for providing the entertainment during journal club, and for his constant help and advice.

Special thanks goes to Anna Lobley for her machine learning and benchmarking expertise, Erik Granseth and Antonis Koussounadis for their membrane protein discussions, Stefano Lise for answering many of my questions, Alexandra Davidson and Anna Thomas for their help with proofreading, and my parents for their constant support throughout my academic career.

This work was generously supported by the Biotechnology and Biological Sciences Research Council (BBSRC).

Bibliography

Abe, S., ed. (1998). *Analysis of Multiclass Support Vector Machines*. Wiley-Interscience.

Abe, S. (2003). Analysis of multiclass support vector machines. *Proc. CIMCA*, 385–396.

Adamian, L. & Liang, J. (2006). Prediction of transmembrane helix orientation in polytopic membrane proteins. *BMC Struct Biol*, **6**, 13.

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.

Amico, M., Finelli, M., Rossi, I., Zauli, A., Elofsson, A., Viklund, H., von Heijne, G., Jones, D., Krogh, A., Fariselli, P., Martelli, P.L. & Casadio, R. (2006). PONGO: a web server for multiple predictions of all-alpha transmembrane proteins. *Nucleic Acids Res*, **34**, W169–72.

Aszodi, A., Gradwell, M.J. & Taylor, W.R. (1995). Global fold determination from a small number of distance restraints. *J Mol Biol*, **251**, 308–26.

Bagos, P.G., Liakopoulos, T.D., Spyropoulos, I.C. & Hamodrakas, S.J. (2004a). A Hidden Markov Model method, capable of predicting and discriminating beta-barrel outer membrane proteins. *BMC Bioinformatics*, **5**, 29.

Bagos, P.G., Liakopoulos, T.D., Spyropoulos, I.C. & Hamodrakas, S.J. (2004b). PRED-TMBB: a web server for predicting the topology of beta-barrel outer membrane proteins. *Nucleic Acids Res*, **32**, W400–4.

Bagos, P.G., Liakopoulos, T.D. & Hamodrakas, S.J. (2006). Algorithms for incorporating prior topological information in HMMs: application to transmembrane proteins. *BMC Bioinformatics*, **7**, 189.

Bahr, A., Thompson, J.D., Thierry, J.C. & Poch, O. (2001). BAliBASE (Benchmark Alignment dataBASE): enhancements for repeats, transmembrane sequences and circular permutations. *Nucleic Acids Res*, **29**, 323–6.

Barth, P., Schonbrun, J. & Baker, D. (2007). Toward high-resolution prediction and design of transmembrane helical protein structures. *Proc Natl Acad Sci U S A*, **104**, 15682–7.

Barth, P., Wallner, B. & Baker, D. (2009). Prediction of membrane protein structures with complex topologies using limited constraints. *Proc Natl Acad Sci U S A*, **106**, 1409–14.

Bendtsen, J.D., Nielsen, H., von Heijne, G. & Brunak, S. (2004). Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol*, **340**, 783–95.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. & Wheeler, D.L. (2008). GenBank. *Nucleic Acids Res*, **36**, D25–30.

Bernsel, A. & Heijne, G.V. (2005). Improved membrane protein topology prediction by domain assignments. *Protein Sci*, **14**, 1723–8.

Bernstein, F.C., Koetzle, T.F., Williams, G.J., Meyer, E.F.J., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1978). The Protein Data Bank: a computer-based archival file for macromolecular structures. *Arch Biochem Biophys*, **185**, 584–91.

Binda, C., Newton-Vinson, P., Hubalek, F., Edmondson, D.E. & Mattevi, A. (2002). Structure of human monoamine oxidase B, a drug target for the treatment of neurological disorders. *Nat Struct Biol*, **9**, 22–6.

Bishop, C.M., ed. (2006). *Pattern recognition and machine learning*. Springer.

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.

Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M. & Bairoch, A. (2007). Uniprotkb/swiss-prot. *Methods Mol Biol*, **406**, 89–112.

Bracey, M.H., Hanson, M.A., Masuda, K.R., Stevens, R.C. & Cravatt, B.F. (2002). Structural adaptations in a membrane enzyme that terminates endocannabinoid signaling. *Science*, **298**, 1793–6.

Buchan, D.W., Ward, S.M., Lobley, A.E., Nugent, T.C., Bryson, K. & Jones, D.T. (2010). Protein annotation and modelling servers at university college london. *Nucleic Acids Res*, **38 Suppl**, W563–W568.

Calvert, P.D., Govardovskii, V.I., Krasnoperova, N., Anderson, R.E., Lem, J. & Makino, C.L. (2001). Membrane protein diffusion sets the speed of rod phototransduction. *Nature*, **411**, 90–4.

Canutescu, A.A., Shelenkov, A.A. & Dunbrack, R.L.J. (2003). A graph-theory algorithm for rapid protein side-chain prediction. *Protein Sci*, **12**, 2001–14.

Capaldi, R.A., Prochaska, L. & Bisson, R. (1980). Structure of cytochrome c oxidase. *Adv Exp Med Biol*, **132**, 197–210.

Chen, C.P., Kernytsky, A. & Rost, B. (2002). Transmembrane helix predictions revisited. *Protein Sci*, **11**, 2774–91.

Chen, Y., Wang, G. & Dong, S. (2003). Learning with progressive transductive support vector machine. *Pattern Recognition Letters*, **24**, 1845–1855.

Cheng, J. & Baldi, P. (2007). Improved residue contact prediction using support vector machines and a large feature set. *BMC Bioinformatics*, **8**, 113.

Chetwynd, A.P., Scott, K.A., Mokrab, Y. & Sansom, M.S.P. (2008). CGDB: a database of membrane protein/lipid interactions by coarse-grained molecular dynamics simulations. *Mol Membr Biol*, **25**, 662–9.

Chrispeels, M.J. & Agre, P. (1994). Aquaporins: water channel proteins of plant and animal cells. *Trends Biochem Sci*, **19**, 421–5.

Claros, M.G. & von Heijne, G. (1994). TopPred II: an improved software for membrane protein structure predictions. *Comput Appl Biosci*, **10**, 685–6.

Cortay, J.C., Negre, D. & Cozzone, A.J. (1991). Analyzing protein phosphorylation in prokaryotes. *Methods Enzymol*, **200**, 214–27.

Cristianini, N. & Shawe-Taylor, J., eds. (2000). *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press.

Cronet, P., Sander, C. & Vriend, G. (1993). Modeling of transmembrane seven helix bundles. *Protein Eng*, **6**, 59–64.

Cunningham, F., Rath, A., Johnson, R.M. & Deber, C.M. (2009). Distinctions between hydrophobic helices in globular proteins and transmembrane segments as factors in protein sorting. *J Biol Chem*, **284**, 5395–402.

Cuthbertson, J.M., Doyle, D.A. & Sansom, M.S.P. (2005). Transmembrane helix prediction: a comparative evaluation and analysis. *Protein Eng Des Sel*, **18**, 295–308.

Daley, D.O., Rapp, M., Granseth, E., Melen, K., Drew, D. & von Heijne, G. (2005). Global topology analysis of the Escherichia coli inner membrane proteome. *Science*, **308**, 1321–3.

Dayhoff, M.O., ed. (1979). *Atlas of Protein Sequence and Structure*. National Biomedical Research Foundation.

de Castro, E., Sigrist, C.J.A., Gattiker, A., Bulliard, V., Langendijk-Genevaux, P.S., Gasteiger, E., Bairoch, A. & Hulo, N. (2006). ScanProsite: detection of PROSITE

signature matches and ProRule-associated functional and structural residues in proteins. *Nucleic Acids Res*, **34**, W362–5.

Donizelli, M., Djite, M.A. & Novere, N.L. (2006). LGICdb: a manually curated sequence database after the genomes. *Nucleic Acids Res*, **34**, D267–9.

Donnelly, D., Overington, J.P., Ruffle, S.V., Nugent, J.H. & Blundell, T.L. (1993). Modeling alpha-helical transmembrane domains: the calculation and use of substitution tables for lipid-facing residues. *Protein Sci*, **2**, 55–70.

Eddy, S.R. (2004). What is a hidden Markov model? *Nat. Biotechnol.*, **22**, 1315–1316.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*, **32**, 1792–7.

Elofsson, A. & von Heijne, G. (2007). Membrane protein structure: prediction versus reality. *Annu Rev Biochem*, **76**, 125–40.

Emanuelsson, O., Brunak, S., von Heijne, G. & Nielsen, H. (2007). Locating proteins in the cell using TargetP, SignalP and related tools. *Nat Protoc*, **2**, 953–71.

Engelman, D.M. (2005). Membranes are more mosaic than fluid. *Nature*, **438**, 578–80.

Engelman, D.M., Steitz, T.A. & Goldman, A. (1986). Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. *Annu Rev Biophys Biophys Chem*, **15**, 321–53.

Enkvetchakul, D., Jeliazkova, I., Bhattacharyya, J. & Nichols, C.G. (2007). Control of inward rectifier K channel activity by lipid tethering of cytoplasmic domains. *J Gen Physiol*, **130**, 329–34.

Eswar, N., Webb, B., Marti-Renom, M.A., Madhusudhan, M.S., Eramian, D., Shen, M.Y., Pieper, U. & Sali, A. (2007). Comparative protein structure modeling using MODELLER. *Curr Protoc Protein Sci*, **Chapter 2**, Unit 2.9.

Ezaki, J., Takeda-Ezaki, M., Koike, M., Ohsawa, Y., Taka, H., Mineki, R., Murayama, K., Uchiyama, Y., Ueno, T. & Kominami, E. (2003). Characterization of Cln3p, the gene product responsible for juvenile neuronal ceroid lipofuscinosis, as a lysosomal integral membrane glycoprotein. *J Neurochem*, **87**, 1296–308.

Falquet, L., Pagni, M., Bucher, P., Hulo, N., Sigrist, C.J.A., Hofmann, K. & Bairoch, A. (2002). The PROSITE database, its status in 2002. *Nucleic Acids Res*, **30**, 235–8.

Fariselli, P., Finocchiaro, G. & Casadio, R. (2003). SPEPlip: the detection of signal peptide and lipoprotein cleavage sites. *Bioinformatics*, **19**, 2498–9.

Ferguson, K.M., Berger, M.B., Mendrola, J.M., Cho, H.S., Leahy, D.J. & Lemmon, M.A. (2003). EGF activates its receptor by removing interactions that autoinhibit ectodomain dimerization. *Mol Cell*, **11**, 507–17.

Ferrer-Montiel, A.V. & Montal, M. (1993). A negative charge in the M2 transmembrane segment of the neuronal alpha 7 acetylcholine receptor increases permeability to divalent cations. *FEBS Lett*, **324**, 185–90.

Finn, R.D., Mistry, J., Schuster-Bockler, B., Griffiths-Jones, S., Hollich, V., Lassmann, T., Moxon, S., Marshall, M., Khanna, A., Durbin, R., Eddy, S.R., Sonnhammer, E.L.L. & Bateman, A. (2006). Pfam: clans, web tools and services. *Nucleic Acids Res*, **34**, D247–51.

Flicek, P., Aken, B.L., Beal, K., Ballester, B., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cunningham, F., Cutts, T., Down, T., Dyer, S.C., Eyre, T., Fitzgerald, S., Fernandez-Banet, J., Graf, S., Haider, S., Hammond, M., Holland, R., Howe, K.L., Howe, K., Johnson, N., Jenkinson, A., Kahari, A., Keefe, D., Kokocinski, F., Kulesha, E., Lawson, D., Longden, I., Megy, K., Meidl, P., Overduin, B., Parker, A., Pritchard, B., Prlic, A., Rice, S., Rios, D., Schuster, M., Sealy, I., Slater, G., Smedley, D., Spudich, G., Trevanion, S., Vilella, A.J., Vogel, J., White, S., Wood, M., Birney, E., Cox, T., Curwen, V., Durbin, R., Fernandez-Suarez, X.M., Herrero, J., Hubbard, T.J.P., Kasprzyk, A., Proctor, G., Smith,

J., Ureta-Vidal, A. & Searle, S. (2008). Ensembl 2008. *Nucleic Acids Res*, **36**, D707–14.

Fodor, A.A. & Aldrich, R.W. (2004). Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. *Proteins*, **56**, 211–21.

Forrest, A.R.R., Taylor, D.F., Fink, J.L., Gongora, M.M., Flegg, C., Teasdale, R.D., Suzuki, H., Kanamori, M., Kai, C., Hayashizaki, Y. & Grimmond, S.M. (2006a). PhosphoregDB: the tissue and sub-cellular distribution of mammalian protein kinases and phosphatases. *BMC Bioinformatics*, **7**, 82.

Forrest, L.R., Tang, C.L. & Honig, B. (2006b). On the accuracy of homology modeling and sequence alignment methods applied to membrane proteins. *Biophys J*, **91**, 508–17.

Fuchs, A., Kirschner, A. & Frishman, D. (2009). Prediction of helix-helix contacts and interacting helices in polytopic membrane proteins using neural networks. *Proteins*, **74**, 857–71.

Gachet, Y., Codlin, S., Hyams, J.S. & Mole, S.E. (2005). btn1, the Schizosaccharomyces pombe homologue of the human Batten disease gene CLN3, regulates vacuole homeostasis. *J Cell Sci*, **118**, 5525–36.

Gafvelin, G., Sakaguchi, M., Andersson, H. & von Heijne, G. (1997). Topological rules for membrane protein assembly in eukaryotic cells. *J Biol Chem*, **272**, 6119–27.

Gattiker, A., Gasteiger, E. & Bairoch, A. (2002). ScanProsite: a reference implementation of a PROSITE scanning tool. *Appl Bioinformatics*, **1**, 107–8.

Ghosh, M., Tucker, D.E., Burchett, S.A. & Leslie, C.C. (2006). Properties of the Group IV phospholipase A2 family. *Prog Lipid Res*, **45**, 487–510.

Ghosh, S., Liu, X.P., Zheng, Y. & Uckun, F.M. (2001). Rational design of potent and selective EGFR tyrosine kinase inhibitors as anticancer agents. *Curr Cancer Drug Targets*, **1**, 129–40.

Gimpelev, M., Forrest, L.R., Murray, D. & Honig, B. (2004). Helical packing patterns in membrane and soluble proteins. *Biophys J*, **87**, 4075–86.

Gondro, C. & Kinghorn, B.P. (2007). A simple genetic algorithm for multiple sequence alignment. *Genet Mol Res*, **6**, 964–82.

Granseth, E., Seppala, S., Rapp, M., Daley, D.O. & Heijne, G.V. (2007). Membrane protein structural biology–how far can the bugs take us? *Mol Membr Biol*, **24**, 329–32.

Gromiha, M.M. & Ponnuswamy, P.K. (1993). Prediction of transmembrane beta-strands from hydrophobic characteristics of proteins. *Int J Pept Protein Res*, **42**, 420–31.

Gromiha, M.M., Ahmad, S. & Suwa, M. (2004). Neural network-based prediction of transmembrane beta-strand segments in outer membrane proteins. *J Comput Chem*, **25**, 762–7.

Gurezka, R., Laage, R., Brosig, B. & Langosch, D. (1999). A heptad motif of leucine residues found in membrane proteins can drive self-assembly of artificial transmembrane segments. *J Biol Chem*, **274**, 9265–70.

Hedman, M., Deloof, H., Heijne, G.V. & Elofsson, A. (2002). Improved detection of homologous membrane proteins by inclusion of information from topology predictions. *Protein Sci*, **11**, 652–8.

Henikoff, S. & Henikoff, J.G. (1992). Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A*, **89**, 10915–9.

Hildebrand, P.W., Goede, A., Bauer, R.A., Gruening, B., Ismer, J., Michalsky, E. & Preissner, R. (2009). SuperLooper–a prediction server for the modeling of loops in globular and membrane proteins. *Nucleic Acids Res*, **37**, W571–4.

Hirokawa, T., Boon-Chieng, S. & Mitaku, S. (1998). SOSUI: classification and secondary structure prediction system for membrane proteins. *Bioinformatics*, **14**, 378–9.

Ho, B.K. & Gruswitz, F. (2008). Hollow: generating accurate representations of channel and interior surfaces in molecular structures. *BMC Struct Biol*, **8**, 49.

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.

Huang, H.D., Lee, T.Y., Tzeng, S.W., Wu, L.C., Horng, J.T., Tsou, A.P. & Huang, K.T. (2005). Incorporating hidden Markov models for identifying protein kinase-specific phosphorylation sites. *J Comput Chem*, **26**, 1032–41.

Izarzugaza, J.M.G., Grana, O., Tress, M.L., Valencia, A. & Clarke, N.D. (2007). Assessment of intramolecular contact predictions for CASP7. *Proteins*, **69 Suppl 8**, 152–8.

Jacoboni, I., Martelli, P.L., Fariselli, P., Pinto, V.D. & Casadio, R. (2001). Prediction of the transmembrane regions of beta-barrel membrane proteins with a neural network-based predictor. *Protein Sci*, **10**, 779–87.

Janes, R.W., Munroe, P.B., Mitchison, H.M., Gardiner, R.M., Mole, S.E. & Wallace, B.A. (1996). A model for Batten disease protein CLN3: functional implications from homology and mutations. *FEBS Lett*, **399**, 75–7.

Jayasinghe, S., Hristova, K. & White, S.H. (2001). MPtopo: A database of membrane protein topology. *Protein Sci*, **10**, 455–8.

Joachims, ed. (1998). *Making large-scale SVM learning practical. In Advances in Kernel Methods-Support Vector Learning.*. Wiley-Interscience.

Johnson, J.E. & Cornell, R.B. (1999). Amphitropic proteins: regulation by reversible membrane interactions (review). *Mol Membr Biol*, **16**, 217–35.

Jones, D.T. (1997). Successful ab initio prediction of the tertiary structure of NK-lysin using multiple sequences and recognized supersecondary structural motifs. *Proteins*, **Suppl 1**, 185–91.

Jones, D.T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol*, **292**, 195–202.

Jones, D.T. (2001). Predicting novel protein folds by using FRAGFOLD. *Proteins*, **Suppl 5**, 127–32.

Jones, D.T. (2007). Improving the accuracy of transmembrane protein topology prediction using evolutionary information. *Bioinformatics*, **23**, 538–44.

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

Jones, D.T., Taylor, W.R. & Thornton, J.M. (1994b). A mutation data matrix for transmembrane proteins. *FEBS Lett*, **339**, 269–75.

Julenius, K., Molgaard, A., Gupta, R. & Brunak, S. (2005). Prediction, conservation analysis, and structural characterization of mammalian mucin-type O-glycosylation sites. *Glycobiology*, **15**, 153–64.

Käll, L., Krogh, A. & Sonnhammer, E.L. (2004). A combined transmembrane topology and signal peptide prediction method. *J Mol Biol*, **338**, 1027–36.

Käll, L., Krogh, A. & Sonnhammer, E.L. (2005). An HMM posterior decoder for sequence feature prediction that includes homology information. *Bioinformatics*, **21 Suppl 1**, i251–7.

Kamada, T. & Kawai, S. (1989). An algorithm for drawing general undirected graphs. *Information Processing Letters*, **31**, 7–15.

Kim, H., Melen, K. & von Heijne, G. (2003). Topology models for 37 Saccharomyces cerevisiae membrane proteins based on C-terminal reporter fusions and predictions. *J Biol Chem*, **278**, 10208–13.

Kim, H., Melen, K., Osterberg, M. & von Heijne, G. (2006). A global topology map of the Saccharomyces cerevisiae membrane proteome. *Proc Natl Acad Sci U S A*, **103**, 11142–7.

King, M.W. (2009). Secreted and membrane-associated proteins, unpublished.

Kreil, G. (1984). Occurrence, detection, and biosynthesis of carboxy-terminal amides. *Methods Enzymol*, **106**, 218–23.

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.

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.

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

Kyttälä, A., Ihrke, G., Vesa, J., Schell, M.J. & Luzio, J.P. (2004). Two motifs target Batten disease protein CLN3 to lysosomes in transfected nonneuronal and neuronal cells. *Mol Biol Cell*, **15**, 1313–23.

Landschulz, W.H., Johnson, P.F. & McKnight, S.L. (1988). The leucine zipper: a hypothetical structure common to a new class of DNA binding proteins. *Science*, **240**, 1759–64.

Larsson, T.A., Olsson, F., Sundstrom, G., Brenner, S., Venkatesh, B. & Larhammar, D. (2005). Pufferfish and zebrafish have five distinct NPY receptor subtypes, but have lost appetite receptors Y1 and Y5. *Ann N Y Acad Sci*, **1040**, 375–7.

Lasso, G., Antoniw, J.F. & Mullins, J.G.L. (2006). A combinatorial pattern discovery approach for the prediction of membrane dipping (re-entrant) loops. *Bioinformatics*, **22**, e290–7.

Lemmon, M.A., Flanagan, J.M., Hunt, J.F., Adair, B.D., Bormann, B.J., Dempsey, C.E. & Engelman, D.M. (1992). Glycophorin A dimerization is driven by specific interactions between transmembrane alpha-helices. *J Biol Chem*, **267**, 7683–9.

Letunic, I., Copley, R.R., Schmidt, S., Ciccarelli, F.D., Doerks, T., Schultz, J., Ponting, C.P. & Bork, P. (2004). SMART 4.0: towards genomic data integration. *Nucleic Acids Res*, **32**, D142–4.

Lewis, B.A. & Engelman, D.M. (1983). Lipid bilayer thickness varies linearly with acyl chain length in fluid phosphatidylcholine vesicles. *J Mol Biol*, **166**, 211–7.

Li, B. & Gallin, W.J. (2004). VKCDB: voltage-gated potassium channel database. *BMC Bioinformatics*, **5**, 3.

Li, W., Zhang, Y. & Skolnick, J. (2004). Application of sparse NMR restraints to large-scale protein structure prediction. *Biophys J*, **87**, 1241–8.

Liu, J., Zhang, M., Jiang, M. & Tseng, G.N. (2003a). Negative charges in the transmembrane domains of the HERG K channel are involved in the activation- and deactivation-gating processes. *J Gen Physiol*, **121**, 599–614.

Liu, Q., Zhu, Y.S., Wang, B.H. & Li, Y.X. (2003b). A HMM-based method to predict the transmembrane regions of beta-barrel membrane proteins. *Comput Biol Chem*, **27**, 69–76.

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

Liu, Y., Gerstein, M. & Engelman, D.M. (2004). Transmembrane protein domains rarely use covalent domain recombination as an evolutionary mechanism. *Proc Natl Acad Sci U S A*, **101**, 3495–7.

Lo, A., Chiu, H.S., Sung, T.Y., Lyu, P.C. & Hsu, W.L. (2008). Enhanced membrane protein topology prediction using a hierarchical classification method and a new scoring function. *J Proteome Res*, **7**, 487–96.

Lo, A., Chiu, Y.Y., Rodland, E.A., Lyu, P.C., Sung, T.Y. & Hsu, W.L. (2009). Predicting helix-helix interactions from residue contacts in membrane proteins. *Bioinformatics*, **25**, 996–1003.

Lobley, A., Nugent, T., Orengo, C.A. & Jones, D.T. (2008). FFPred: an integrated feature-based function prediction server for vertebrate proteomes. *Nucleic Acids Res*, **36**, 297–302.

Lomize, A.L., Pogozheva, I.D., Lomize, M.A. & Mosberg, H.I. (2006a). Positioning of proteins in membranes: a computational approach. *Protein Sci*, **15**, 1318–33.

Lomize, M.A., Lomize, A.L., Pogozheva, I.D. & Mosberg, H.I. (2006b). OPM: orientations of proteins in membranes database. *Bioinformatics*, **22**, 623–5.

Luirink, J., von Heijne, G., Houben, E. & de Gier, J.W. (2005). Biogenesis of inner membrane proteins in Escherichia coli. *Annu Rev Microbiol*, **59**, 329–55.

MacIntyre, S., Freudl, R., Eschbach, M.L. & Henning, U. (1988). An artificial hydrophobic sequence functions as either an anchor or a signal sequence at only one of two positions within the Escherichia coli outer membrane protein OmpA. *J Biol Chem*, **263**, 19053–9.

Mahajan, S., Ghosh, S., Sudbeck, E.A., Zheng, Y., Downs, S., Hupke, M. & Uckun, F.M. (1999). Rational design and synthesis of a novel anti-leukemic agent targeting Bruton's tyrosine kinase (BTK), LFM-A13 [alpha-cyano-beta-hydroxy-beta-methyl-N-(2, 5-dibromophenyl)propenamide]. *J Biol Chem*, **274**, 9587–99.

Mao, Q., Foster, B.J., Xia, H. & Davidson, B.L. (2003a). Membrane topology of CLN3, the protein underlying Batten disease. *FEBS Lett*, **541**, 40–6.

Mao, Q., Xia, H. & Davidson, B.L. (2003b). Intracellular trafficking of CLN3, the protein underlying the childhood neurodegenerative disease, Batten disease. *FEBS Lett*, **555**, 351–7.

Martelli, P.L., Fariselli, P., Krogh, A. & Casadio, R. (2002). A sequence-profile-based HMM for predicting and discriminating beta barrel membrane proteins. *Bioinformatics*, **18 Suppl 1**, S46–53.

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.

Marti-Renom, M.A., Stuart, A.C., Fiser, A., Sanchez, R., Melo, F. & Sali, A. (2000). Comparative protein structure modeling of genes and genomes. *Annu Rev Biophys Biomol Struct*, **29**, 291–325.

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

Melen, K., Krogh, A. & von Heijne, G. (2003). Reliability measures for membrane protein topology prediction algorithms. *J Mol Biol*, **327**, 735–44.

Michalewski, M.P., Kaczmarski, W., Golabek, A.A., Kida, E., Kaczmarski, A. & Wisniewski, K.E. (1999). Posttranslational modification of CLN3 protein and its possible functional implication. *Mol Genet Metab*, **66**, 272–6.

Mitchell, M., ed. (1996). *An Introduction to Genetic Algorithms*. MIT Press.

Mitchison, H.M., Taschner, P.E., Kremmidiotis, G., Callen, D.F., Doggett, N.A., Lerner, T.J., Janes, R.B., Wallace, B.A., Munroe, P.B., O'Rawe, A.M., Gardiner, R.M. & Mole, S.E. (1997). Structure of the CLN3 gene and predicted structure, location and function of CLN3 protein. *Neuropediatrics*, **28**, 12–4.

Mitra, K., Ubarretxena-Belandia, I., Taguchi, T., Warren, G. & Engelman, D.M. (2004). Modulation of the bilayer thickness of exocytic pathway membranes by membrane proteins rather than cholesterol. *Proc Natl Acad Sci U S A*, **101**, 4083–8.

Mole, S.E. (1998). NCL resource - a gateway for batten disease, unpublished.

Möller, S., Kriventseva, E.V. & Apweiler, R. (2000). A collection of well characterised integral membrane proteins. *Bioinformatics*, **16**, 1159–60.

Muller, T., Rahmann, S. & Rehmsmeier, M. (2001). Non-symmetric score matrices and the detection of homologous transmembrane proteins. *Bioinformatics*, **17 Suppl 1**, S182–9.

Ng, P.C., Henikoff, J.G. & Henikoff, S. (2000). PHAT: a transmembrane-specific substitution matrix. Predicted hydrophobic and transmembrane. *Bioinformatics*, **16**, 760–6.

Nilsson, J., Persson, B. & Heijne, G.V. (2002). Prediction of partial membrane protein topologies using a consensus approach. *Protein Sci*, **11**, 2974–80.

Notredame, C., Higgins, D.G. & Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol*, **302**, 205–17.

Nugent, T. & Jones, D.T. (2009). Transmembrane protein topology prediction using support vector machines. *BMC Bioinformatics*, **10**, 159.

Nugent, T. & Jones, D.T. (2010). Predicting transmembrane helix packing arrangements using residue contacts and a force-directed algorithm. *PLoS Comput Biol*, **6**, e1000714.

Nugent, T., Mole, S.E. & Jones, D.T. (2008). The transmembrane topology of batten disease protein cln3 determined by consensus computational prediction constrained by experimental data. *FEBS Lett*, **582**, 1019–1024.

Olmea, O. & Valencia, A. (1997). Improving contact predictions by the combination of correlated mutations and other sources of sequence information. *Fold Des*, **2**, S25–32.

Pao, G.M., Wu, L.F., Johnson, K.D., Hofte, H., Chrispeels, M.J., Sweet, G., Sandal, N.N. & Saier, M.H.J. (1991). Evolution of the MIP family of integral membrane transport proteins. *Mol Microbiol*, **5**, 33–7.

Pardo, C.A., Rabin, B.A., Palmer, D.N. & Price, D.L. (1994). Accumulation of the adenosine triphosphate synthase subunit C in the mnd mutant mouse. A model for neuronal ceroid lipofuscinosis. *Am J Pathol*, **144**, 829–35.

Park, K.J., Gromiha, M.M., Horton, P. & Suwa, M. (2005). Discrimination of outer membrane proteins using support vector machines. *Bioinformatics*, **21**, 4223–9.

Pasquier, C., Promponas, V.J. & Hamodrakas, S.J. (2001). PRED-CLASS: cascading neural networks for generalized protein classification and genome-wide applications. *Proteins*, **44**, 361–9.

Pearce, D.A. & Sherman, F. (1998). A yeast model for the study of Batten disease. *Proc Natl Acad Sci U S A*, **95**, 6915–8.

Peitsch, M.C. (1996). ProMod and Swiss-Model: Internet-based tools for automated comparative protein modelling. *Biochem Soc Trans*, **24**, 274–9.

Pellegrini-Calace, M., Carotti, A. & Jones, D.T. (2003). Folding in lipid membranes (FILM): a novel method for the prediction of small membrane protein 3D structures. *Proteins*, **50**, 537–45.

Pellegrini-Calace, M., Maiwald, T. & Thornton, J.M. (2009). Porewalker: a novel tool for the identification and characterization of channels in transmembrane proteins from their three-dimensional structure. *PLoS Comput Biol*, **5**, e1000440.

Petrey, D., Xiang, Z., Tang, C.L., Xie, L., Gimpelev, M., Mitros, T., Soto, C.S., Goldsmith-Fischman, S., Kernytsky, A., Schlessinger, A., Koh, I.Y.Y., Alexov, E. & Honig, B. (2003). Using multiple structure alignments, fast model building, and energetic analysis in fold recognition and homology modeling. *Proteins*, **53 Suppl 6**, 430–5.

Petty, H.R., ed. (1993). *Molecular Biology of Membranes: Structure and Function*. Springer.

Pirovano, W., Feenstra, K.A. & Heringa, J. (2008). PRALINETM: a strategy for improved multiple alignment of transmembrane proteins. *Bioinformatics*, **24**, 492–7.

Podell, S. & Gribskov, M. (2004). Predicting N-terminal myristoylation sites in plant proteins. *BMC Genomics*, **5**, 37.

Pornillos, O., Chen, Y.J., Chen, A.P. & Chang, G. (2005). X-ray structure of the EmrE multidrug transporter in complex with a substrate. *Science*, **310**, 1950–3.

Price, C.E. & Driessen, A.J.M. (2010). Conserved negative charges in the transmembrane segments of subunit K of the NADH:ubiquinone oxidoreductase determine its dependence on YidC for membrane insertion. *J Biol Chem*, **285**, 3575–81.

Punta, M. & Rost, B. (2005). PROFcon: novel prediction of long-range contacts. *Bioinformatics*, **21**, 2960–8.

Raman, P., Cherezov, V. & Caffrey, M. (2006). The Membrane Protein Data Bank. *Cell Mol Life Sci*, **63**, 36–51.

Rawicz, W., Olbrich, K.C., McIntosh, T., Needham, D. & Evans, E. (2000). Effect of chain length and unsaturation on elasticity of lipid bilayers. *Biophys J*, **79**, 328–39.

Rigden, D., ed. (2009). *From Protein Structure to Function with Bioinformatics*. Springer.

Rohl, C.A., Strauss, C.E.M., Misura, K.M.S. & Baker, D. (2004). Protein structure prediction using Rosetta. *Methods Enzymol*, **383**, 66–93.

Rost, B., Fariselli, P. & Casadio, R. (1996). Topology prediction for helical transmembrane proteins at 861704–18.

Rudd, P.M., Wormald, M.R., Stanfield, R.L., Huang, M., Mattsson, N., Speir, J.A., DiGennaro, J.A., Fetrow, J.S., Dwek, R.A. & Wilson, I.A. (1999). Roles for glycosylation of cell surface receptors involved in cellular immune recognition. *J Mol Biol*, **293**, 351–66.

Ruiz, N., Kahne, D. & Silhavy, T.J. (2006). Advances in understanding bacterial outer-membrane biogenesis. *Nat Rev Microbiol*, **4**, 57–66.

Saier, M.H.J., Tran, C.V. & Barabote, R.D. (2006). TCDB: the Transporter Classification Database for membrane transport protein analyses and information. *Nucleic Acids Res*, **34**, D181–6.

Sal-Man, N., Gerber, D., Bloch, I. & Shai, Y. (2007). Specificity in transmembrane helix-helix interactions mediated by aromatic residues. *J Biol Chem*, **282**, 19753–61.

Samatey, F.A., Xu, C. & Popot, J.L. (1995). On the distribution of amino acid residues in transmembrane alpha-helix bundles. *Proc Natl Acad Sci U S A*, **92**, 4577–81.

Sanchez, R. & Sali, A. (1997). Advances in comparative protein-structure modelling. *Curr Opin Struct Biol*, **7**, 206–14.

Sansom, M.S.P., Scott, K.A. & Bond, P.J. (2008). Coarse-grained simulation: a high-throughput computational approach to membrane proteins. *Biochem Soc Trans*, **36**, 27–32.

Schirmer, T. & Cowan, S.W. (1993). Prediction of membrane-spanning beta-strands and its application to maltoporin. *Protein Sci*, **2**, 1361–3.

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.

Shafrir, Y. & Guy, H.R. (2004). STAM: simple transmembrane alignment method. *Bioinformatics*, **20**, 758–69.

Sigrist, C.J.A., Cerutti, L., Hulo, N., Gattiker, A., Falquet, L., Pagni, M., Bairoch, A. & Bucher, P. (2002). PROSITE: a documented database using patterns and profiles as motif descriptors. *Brief Bioinform*, **3**, 265–74.

Siintola, E., Topcu, M., Aula, N., Lohi, H., Minassian, B.A., Paterson, A.D., Liu, X.Q., Wilson, C., Lahtinen, U., Anttonen, A.K. & Lehesjoki, A.E. (2007). The novel neuronal ceroid lipofuscinosis gene MFSD8 encodes a putative lysosomal transporter. *Am J Hum Genet*, **81**, 136–46.

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

Sipiczki, M. (1995). Phylogenesis of fission yeasts. Contradictions surrounding the origin of a century old genus. *Antonie Van Leeuwenhoek*, **68**, 119–49.

Smith, T.F. & Waterman, M.S. (1981). Identification of common molecular subsequences. *J Mol Biol*, **147**, 195–7.

Spoel, D.V.D., Lindahl, E., Hess, B., Groenhof, G., Mark, A.E. & Berendsen, H.J.C. (2005). GROMACS: fast, flexible, and free. *J Comput Chem*, **26**, 1701–18.

Stevens, T.J. & Arkin, I.T. (2001). Substitution rates in alpha-helical transmembrane proteins. *Protein Sci*, **10**, 2507–17.

Stock, J.B., Ninfa, A.J. & Stock, A.M. (1989). Protein phosphorylation and regulation of adaptive responses in bacteria. *Microbiol Rev*, **53**, 450–90.

Storch, S., Pohl, S., Quitsch, A., Falley, K. & Braulke, T. (2007). C-terminal prenylation of the CLN3 membrane glycoprotein is required for efficient endosomal sorting to lysosomes. *Traffic*, **8**, 431–44.

Stryer, L. (1995). *Biochemistry*. W.H.Freeman & Company.

Sudbeck, E.A., Liu, X.P., Narla, R.K., Mahajan, S., Ghosh, S., Mao, C. & Uckun, F.M. (1999). Structure-based design of specific inhibitors of Janus kinase 3 as apoptosis-inducing antileukemic agents. *Clin Cancer Res*, **5**, 1569–82.

Tang, C.L., Xie, L., Koh, I.Y.Y., Posy, S., Alexov, E. & Honig, B. (2003). On the role of structural information in remote homology detection and sequence alignment: new methods using hybrid sequence profiles. *J Mol Biol*, **334**, 1043–62.

Taylor, P.D., Attwood, T.K. & Flower, D.R. (2003). BPROMPT: A consensus server for membrane protein prediction. *Nucleic Acids Res*, **31**, 3698–700.

Taylor, W.R. (2006). Decoy models for protein structure comparison score normalisation. *J Mol Biol*, **357**, 676–99.

Taylor, W.R., Jones, D.T. & Green, N.M. (1994). A method for alpha-helical integral membrane protein fold prediction. *Proteins*, **18**, 281–94.

The International Batten Disease Consortium (1995). Isolation of a novel gene underlying Batten disease, CLN3. The International Batten Disease Consortium. *Cell*, **82**, 949–57.

Thompson, J.D., Higgins, D.G. & Gibson, T.J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, **22**, 4673–80.

Thuduppathy, G.R., Craig, J.W., Kholodenko, V., Schon, A. & Hill, R.B. (2006). Evidence that membrane insertion of the cytosolic domain of Bcl-xL is governed by an electrostatic mechanism. *J Mol Biol*, **359**, 1045–58.

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.

Tusnady, G.E., Dosztanyi, Z. & Simon, I. (2005a). PDB_TM: Selection and membrane localization of transmembrane proteins in the protein data bank. *Nucleic Acids Res*, **33**, D275–8.

Tusnady, G.E., Dosztanyi, Z. & Simon, I. (2005b). TMDET: web server for detecting transmembrane regions of proteins by using their 3D coordinates. *Bioinformatics*, **21**, 1276–7.

Tusnady, G.E., Kalmar, L. & Simon, I. (2008). TOPDB: topology data bank of transmembrane proteins. *Nucleic Acids Res*, **36**, D234–9.

van Vliet, C., Thomas, E.C., Merino-Trigo, A., Teasdale, R.D. & Gleeson, P.A.

(2003). Intracellular sorting and transport of proteins. *Prog Biophys Mol Biol*, **83**, 1–45.

Vapnik, V.N., ed. (1998). *Statistical Learning Theory*. Wiley-Interscience.

Vesa, J., Chin, M.H., Oelgeschlager, K., Isosomppi, J., DellAngelica, E.C., Jalanko, A. & Peltonen, L. (2002). Neuronal ceroid lipofuscinoses are connected at molecular level: interaction of CLN5 protein with CLN2 and CLN3. *Mol Biol Cell*, **13**, 2410–20.

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.

Viklund, H. & Elofsson, A. (2008). OCTOPUS: improving topology prediction by two-track ANN-based preference scores and an extended topological grammar. *Bioinformatics*, **24**, 1662–8.

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.

Viklund, H., Bernsel, A., Skwark, M. & Elofsson, A. (2008). SPOCTOPUS: a combined predictor of signal peptides and membrane protein topology. *Bioinformatics*, **24**, 2928–9.

von Heijne, G. (1983). Patterns of amino acids near signal-sequence cleavage sites. *Eur J Biochem*, **133**, 17–21.

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

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

Wallin, E., Tsukihara, T., Yoshikawa, S., von Heijne, G. & Elofsson, A. (1997). Architecture of helix bundle membrane proteins: an analysis of cytochrome c oxidase from bovine mitochondria. *Protein Sci*, **6**, 808–15.

Wallner, B. & Elofsson, A. (2005). Pcons5: combining consensus, structural evaluation and fold recognition scores. *Bioinformatics*, **21**, 4248–54.

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

Wang, L., ed. (2005). *Support Vector Machines: Theory and Applications*. Springer.

Ward, J.J. (2005). *Kernel-Based Classification of Protein Structure and Function from Amino Acid Sequences*. Ph.D. thesis, University College London.

White, S. (2010). Membrane proteins of known 3d structure, unpublished.

White, S.H. (2004). The progress of membrane protein structure determination. *Protein Sci*, **13**, 1948–9.

White, S.H. & von Heijne, G. (2004). The machinery of membrane protein assembly. *Curr Opin Struct Biol*, **14**, 397–404.

Wimley, W.C. & White, S.H. (1996). Experimentally determined hydrophobicity scale for proteins at membrane interfaces. *Nat Struct Biol*, **3**, 842–8.

Wong, E., Yu, W.P., Yap, W.H., Venkatesh, B. & Soong, T.W. (2006). Comparative genomics of the human and Fugu voltage-gated calcium channel alpha1-subunit gene family reveals greater diversity in Fugu. *Gene*, **366**, 117–27.

Yang, Z.R. (2004). Biological applications of support vector machines. *Brief Bioinform*, **5**, 328–38.

Yarov-Yarovoy, V., Schonbrun, J. & Baker, D. (2006). Multipass membrane protein structure prediction using Rosetta. *Proteins*, **62**, 1010–25.

Yuan, Z., Mattick, J.S. & Teasdale, R.D. (2004). SVMtm: support vector machines to predict transmembrane segments. *J Comput Chem*, **25**, 632–6.

Zhang, Y. & Rajapakse, J.C., eds. (2008). *Machine Learning in Bioinformatics*. Wiley-Interscience.

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.