Developing a Novel Method for Homology Detection of Transmembrane Proteins

Naama Hurwitz

Bioinformatics Group
Department of Computer Science
University College London

A thesis submitted to University College London for the degree of
Doctor of Philosophy

July 2012

Abstract

Analysis of the complete genomic sequences for several organisms indicates that 20-25% of all genes code for transmembrane proteins (Jones, 1998, Wallin and von Heijne, 1998), yet only a very small number of transmembrane 3D structures are known. Hence, it is of great importance to develop theoretical methods capable of predicting transmembrane protein structure and function based on protein sequence alone. To address this, we sought to devise a systematic and high throughput method for identifying homologous transmembrane proteins. Since protein structure is more evolutionarily conserved than amino acid sequence, we predicted that adding structural information to simple sequence alignment would improve homology detection of transmembrane proteins. In the present work, we describe development of a search method that combines sequence alignment with structural information.

In our method the initial sequence alignment searches are performed using PSI-BLAST. Then profiles derived from the multiple sequence alignments are input into a neural network, developed in this work to predict which transmembrane residues are buried (core of the helix-bundle) or exposed (to the lipid environment). A maximum accuracy of 86% was achieved. Moreover, for almost half of the query set, the predicted residue orientation was more than 70% accurate. In the last step of the work presented here, the predicted helix locations, residue orientations and loop length scores are added to the PSI-BLAST E-value, to create a ‘combined’ classifier. A linear equation was built for calculating the 'combined’ classifier score.

Our method was evaluated using two databases of proteins: Pfam and GPCRDB. The Pfam database was chosen, as transmembrane proteins in this database have been

classified into various families. GPCRDB was employed as this database, though narrow, is well-studied and maintained. Before building the ‘combined’ classifier, PSI-BLAST sequence alignment was benchmarked using the Pfam database.

We found that our 'combined’ classifier, as compared to a classifier based solely on PSI-BLAST, resulted in more true positives with less false positives when tested using GPCRDB and could differentiate between GPCRDB families. However, our ‘combined’ classifier did not improve homology detection when searching transmembrane proteins from the Pfam database.

A comparison of our ‘combined’ classifier method with two other published methods suggested that profile-profile based searches could be more powerful than profile-sequence based searches, even after the addition of structural information as described here. In light of our study, we propose that combining structural information with profile-profile sequence alignment into a 'combined’ classifier could result in a search method superior to any existing ones for detecting homologous transmembrane proteins.

Declaration Page

I, Naama Hurwitz, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Acknowledgments

I would like to thank my supervisors David Jones for his patience and guidance and Kevin Bryson for always being willing to help.

I would also like to thank the members of the Bioinformatics Group with which I had a chance to study, and more specially, Tim Nugent for assisting me without even knowing me.

Most of all, I would like to thank my family – my children for the hours they let me study without any disturbance, my husband for supporting me over the years and my parents for encouraging me to continue when I was ready to give it up.

This work was generously supported by Engineering and Physical Sciences Research Council (EPSRC).

Table of Content

List of abbreviations

Abbreviation	Detailes
BLAST	Basic Local Alignment Search Tool
CATH	Class, Architecture, Topology, Homologous Superfamily
Fp	False positives
Fn	False negetives
GPCR	G Protein coupled Receptor
HMM	Hidden Markov Model
MCC	Matthews Correlation Coefficient
NMR	nuclear magnetic resonance
NN	Neural Network
PDB	Protein Data Bank
PSI-BLAST	Position-Specific Iterated BLAST
PSSM	Position-Specific Scoring Matrix
SCOP	Structural Classification of Proteins
SVM	Support Vector Machine
Tp	True positives
Tn	True negatives
3D	three–dimensional

Figures list

Tables List

Chapter 1

Introduction

1.1 Membrane proteins

A wide range of fundamental biological processes such as cell signaling, transport of membrane-impermeable molecules, cell-cell communication, cell recognition and cell adhesion are mediated by membrane proteins. Therefore, understanding the structure and function of membrane proteins is of high biological and pharmacological importance.

Analysis of the complete genomic sequences for several organisms indicates that 20-25% of all genes code for transmembrane proteins (Jones , 1998, Wallin and von Heijne, 1998). Despite their large number and importance, less than 1% of all 3D protein structures deposited in the Protein Data Bank (PDB) are of membrane proteins (Berman *et al*., 2000), likely due to the challenges of crystallizing such proteins or performing nuclear magnetic resonance (NMR) analyses. In light of this deficit of empirical information, it is particularly important to develop efficient theoretical methods for predicting the structure of transmembrane proteins.

1.1.1 The biology of membranes

Biological membranes are composed of a lipid bilayer and serve to separate different cellular compartments or the cell from its environment. The lipid bilayer is impermeable to polar (soluble in water) molecules and ions.

The membrane can be represented three-dimensionally as shown in Figure 1. Each phospholipid is composed of a negatively charged phosphate group and two tails, which are two highly hydrophobic hydrocarbon chains. The hydrophobic effect ensures that the tails of the phospholipids in each layer orient towards each other creating a highly hydrophobic environment within the membrane. The charged phosphate groups face out into the hydrophilic environment.

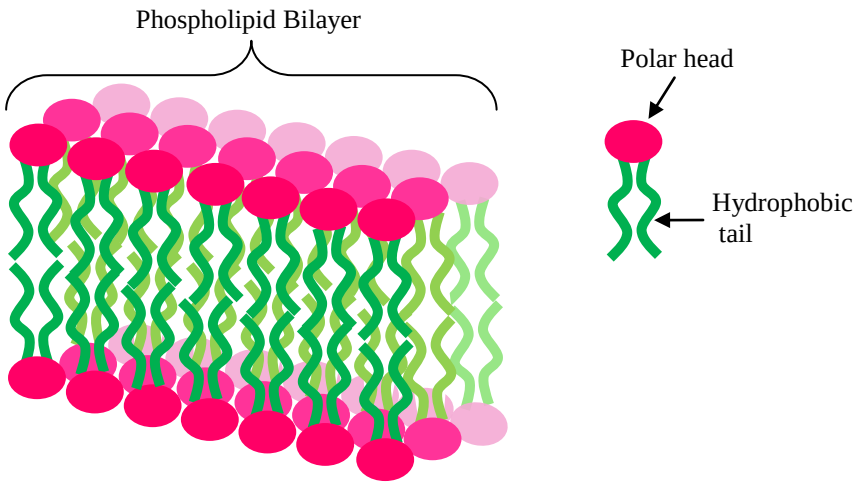


Figure 1: Biological membranes.

Membrane proteins carry out most of the dynamic processes of the membrane. Membrane lipids create the appropriate environment for the action of such proteins.

1.1.2 Types of membrane protein

Membrane proteins can be classified as either peripheral (membrane associated

proteins) or integral, on the basis of how readily they dissociate from the membrane. Peripheral membrane proteins are loosely associated with the membrane and usually interact with the polar head groups of the membrane phospholipids. These proteins can therefore be solubilized under relatively mild conditions, such as exposure to high ionic strength. In contrast, integral membrane proteins, also termed transmembrane proteins, are found to interact extensively with the hydrocarbon chains of the membrane lipids (Figure 2) and can only be solubilized using detergents or an organic solvent.

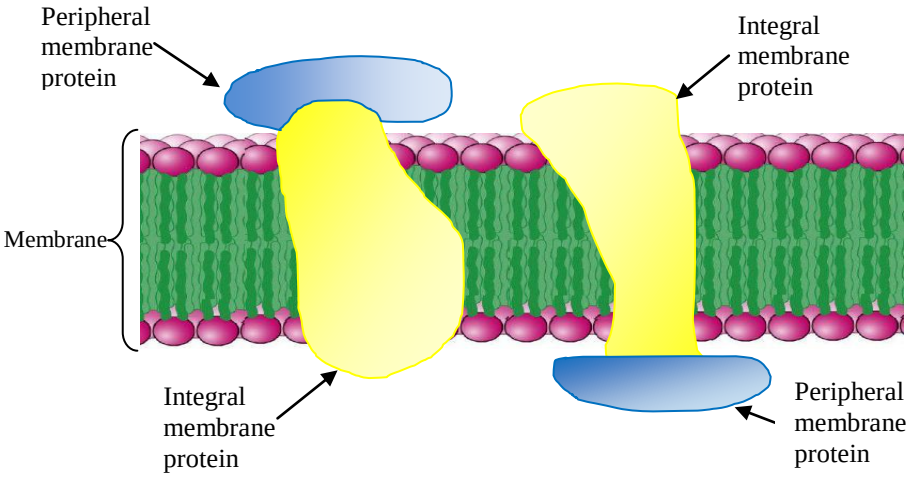


Figure 2: Membrane proteins main types: peripheral membrane proteins and integral membrane proteins.

1.1.2.1 Integral membrane proteins

Integral membrane proteins display particular structures that are remarkably stable despite the high energetic cost of dehydrating the peptide bond during transfer into a non-polar phase (White *et al.*, 2001). This is enabled by two features. Firstly, and perhaps most obviously, most of the amino acid side chains found within integral membrane segments are non-polar. Secondly, the polar groups of the polypeptide

backbone of the transmembrane segments participate in hydrogen bonds to lower the energetic cost of membrane insertion. This second constraining feature of integral membrane proteins is typically accomplished through two structural motifs: the membrane-spanning alpha-helix bundle and the beta-barrel (White and Wimley, 1999).

1.1.2.1.1 Beta -barrel integral membrane proteins

The beta-barrel proteins, consist of beta-strands spanning the membrane connected by short loops facing the periplasm and larger loops protruding outside the outer membrane (von Heijne, 1996). The beta-strands are amphiphilic, i.e., the side chains of the strand residues are alternately polar and hydrophobic with polar residues toward the central pore. Thus the structure forms a pore with a polar environment (see Figure 3 for example).

The beta-barrel proteins are found in the outer membrane of Gram-negative bacteria and in the outer membrane of chloroplasts and mitochondria. Their function is to facilitate diffusion of salts and polar compounds.

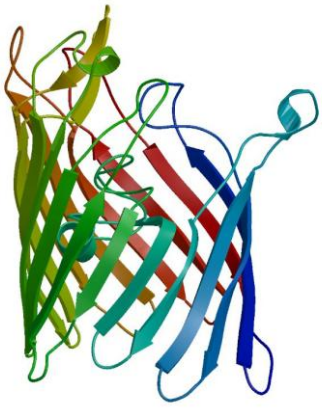


Figure 3: Example of beta-barrel protein, Porin (1OPF).

1.1.2.1.2 Alpha-helical integral membrane proteins

The alpha-helical integral membrane proteins consist of alpha-helices, 17-25 residues in length, which cross the membrane once or several times.

There are two types of alpha-helical integral membrane proteins:

- Bitopic proteins (or membrane-anchored proteins), which cross the membrane once (or sometimes twice), exposing water-soluble domains on the extracellular and cytoplasmic sides. Such proteins typically act as cell surface markers, adhesion factors or receptors. The cytoplasmic domains often play a role in cell signaling (e.g., tyrosine kinases) or connect to the cellular cytoskeleton.

- Polytopic (multi-spanning) alpha-helical membrane proteins have more than one alpha-helical transmembrane segment and the helices are arranged into a bundle (Figure 4).

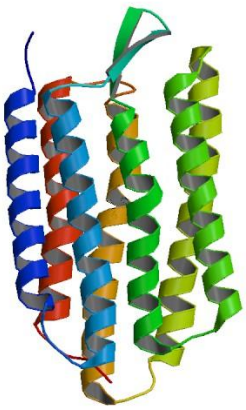


Figure 4: Example of an alpha helical bundle integral membrane protein – Bacteriorhodopsin.

In the current work we study only polytopic alpha-helical integral membrane proteins, therefore these proteins hereafter are referred to simply as ‘transmembrane proteins’.

1.2 Transmembrane proteins

1.2.1 Transmembrane protein functions

Transmembrane proteins are involved mainly in the following cellular processes:

- **Channels:** Channel proteins mediate passive transport through the membrane, but are typically highly selective. For example, ion channels play a key role in the nervous system and in homeostasis of most cells.

- **Transporters:** Transporter proteins mediate active transport of solutes across the membrane. An example of active transport is the transportation of sodium out of the cell and potassium into the cell by the sodium-potassium pump; this process is mediated by ATP energy. There are also transporters, such as the sodium-calcium exchanger, which transport one of the two substances in the direction of its concentration gradient and yields the energy derived from this transport to transport the other substance against its concentration gradient.

- **Receptors:** Are transmembrane proteins that take part in communication between the cell and the outside world. Extracellular signaling molecules attach to the receptor, triggering signaling pathways within the cell. The process is called signal transduction.

  Polytopic alpha–helical membrane receptors can be sub-categorized into two classes: G-protein-coupled receptors and ion channel-linked receptors.

  G-Protein Coupled receptors (GPCRs), a pharmacologically important class, which includes receptors for hormones, neurotransmitters, growth factors,

light and other diverse ligands (Dewji and Singer, 1997). GPCRs possess seven transmembrane helices. After a ligand binds the GPCR, it causes a conformational change in the GPCR, which then activates G-protein by exchanging its bound GDP for a GTP.

Ion channel-linked receptors, also called ligand-gated ion channels, are involved in rapid signaling events mostly found in electrically excitable cells such as neurons.

- **Oxidative phosphorylation and photosynthesis transmembrane processes**: Helical transmembrane proteins are involved in energy generation processes, typically incorporating cofactors and mediating oxidation of substrates.

1.2.2 Transmembrane protein folding process

While water-soluble proteins exist in only one kind of environment, transmembrane proteins are present in three different environments: the hydrophilic environment, the water-membrane interface and the inner-membrane phase. Accordingly, the transmembrane protein folding process differs from that for soluble proteins (White and Wimley, 1999). The interactions of transmembrane proteins with the lipid are important for folding and stability (Lee, 2004). Possible driving forces for helix-helix association in the lipid bilayer are van der Waals interactions and interhelical polar interactions, including hydrogen bond and electrostatic interactions (Popot and Engelman, 2000).

The folding process of transmembrane proteins comprises two stages (Engelman *et al*., 2003). The first stage involves formation of stable helices across the hydrophobic

region of the membrane lipid bilayer. In the second stage, the helices interact to generate a functional membrane protein (Popot and Engelman, 1990). Assembly is carried out by a translocon apparatus and involves the transient attachment of an active ribosome to a translocon embedded in the membrane. As soon as the protein is synthesized into the translocon and transferred into the membrane, the apparatus disassembles leaving the folded protein within the membrane (White and Wimley,1999).

1.2.3 Transmembrane protein structure

The secondary structure, i.e. the topology of transmembrane proteins, describes which segments of the amino acid sequence span the membrane, the number of spanning segments, and which ones protrude into the respective compartments on opposite sides of the membrane (i.e., in-out location of the N and C termini relative to the membrane). Knowing a protein’s topology is a significant step toward understanding its structure and function. A topological description has also been referred to as 'low resolution structure' (Kernytsky and Rost, 2003).

When alpha-helical transmembrane proteins are grouped according to topology, differences between various species can be observed. In general eubacteria, archaea, fungi and plants have a large collection of membrane proteins with 6 or 12 transmembrane segments, whereas in C. elegans and Homo sapiens the predominant topology is membrane proteins with 7 segments (Wallin and von Heijne ,1998).

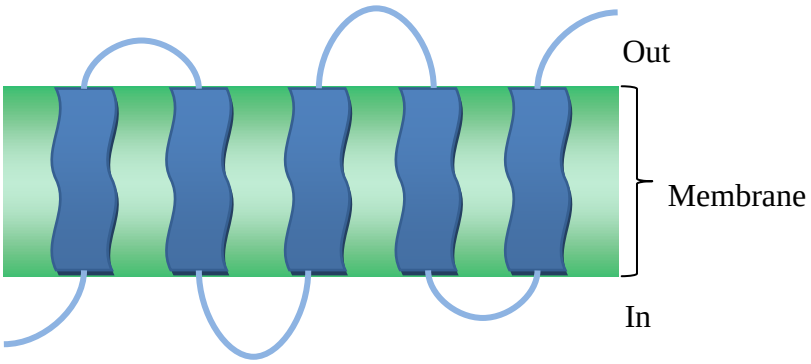


Figure 5: Schematic presentation of a transmembrane protein with five membrane spanning helical segments (blue boxes). The outer membrane regions of the protein are drawn in light blue (the loops). The membrane is drawn in green. "In" designates the inner side of the membrane and "Out" designates the outer side.

The transmembrane protein topology can be represented as boxes spanning the membrane connected by protein loops, with each box representing a transmembrane helix structure. The N-terminus and the C-terminus can be on either side of the membrane (Figure 5).

Alpha-helical transmembrane proteins are comprised of a number of transmembrane helices. Classically, the helices were considered to assemble mostly in parallel or anti-parallel to one another and perpendicular to the membrane. However, recent studies have revealed deviation from this structure. It was found that about 50% of transmembrane proteins contain non-canonical elements (e.g., wide turns, tight turns, and kinks) (Riek *et al*., 2008) and 10% contain reentrant loops, which go half way through the membrane (Viklund et. al, 2006).

Yohannan et al. (2004) showed that 60% of transmembrane helix deformations occur at proline residues. Yohannan *et al*. proposed an evolutionary hypothesis whereby a mutation to proline initially induces a kink, and then further mutations occur locking the kink in the structure. In an extension of this hypothesis, the premise is that

nonproline kinks are places where prolines have been removed during evolution.

Reentrant loops are mostly found in ion and water channel proteins and less in cell surface receptors. It was shown that the difference in residue composition makes the reentrant loops less hydrophobic than the transmembrane helices. This reduced hydrophobicity makes the reentrant loops less stable inside the hydrophobic environment of the lipid membrane (Changhui and Jingru, 2010). An independent study found that reentrant loops have very low hydrophobicity around the deepest point buried in the membrane but relatively high hydrophobicity close to the membrane surfaces (Yan and Luo, 2010). Moreover, the residues situated in reentrant regions are significantly smaller on average as compared to those in other parts of the protein. These unique features allow reentrant loops to be detected based on amino acid composition (Viklund *et al*., 2006). Additionally, reentrant loops often contain functional motifs that differentiate them from regular helices (Lasso *et al*., 2006).

1.3 Empirical approaches to solving transmembrane protein structure

The first three-dimensional structure of a transmembrane protein, *Rhodopseudomonas viridis* photosynthetic reaction centre, was solved in 1985 using X-ray crystallographic analysis by Deisenhofer, Michel and Huber, who won a Nobel prize for their work (Deisenhofer *et al*., 1985, Deisenhofer and Michel, 1989). Since then the three-dimensional structures of only 263 transmembrane proteins have been solved (von Heijne , 2011). Oberai *et al*. have estimated that ~1700 transmembrane proteins structures are needed to cover each structural family (Oberai *et al*., 2006). At

the current pace, as noted by White (2009), it will take approximately 30 years to obtain these 1700 membrane protein structures.

The number of empirically determined transmembrane protein structures is small because of the difficulties involved in expressing and crystallizing these proteins (Grisshammer and Tate, 1995). As discussed above, transmembrane proteins are hydrophobic in the transmembrane regions and consequently are difficult to unfold and refold in vitro. In addition, transmembrane proteins are typically only expressed at low concentrations and therefore it is necessary to over express them in a membrane system, which has proved very difficult. There are various expression systems but all are technically problematic. The technical problems include low yield, post-translational modification, low stability and partial proteolysis (Grisshammer and Tate, 1995).

The difficulty in determining high-resolution structures of membrane proteins has prompted development of alternative methods. The idea is to obtain structural hints concerning the packing of transmembrane helices, which can be used to build and model the whole structure of the protein. Several experimental approaches are used to obtain such structural hints and are summarized in the next sections.

1.3.1 Fusion with a reporter protein

The most common procedure for determining transmembrane protein topology is to fuse the C-terminal part of the protein with a reporter protein. The reporter proteins are chosen according to properties, such as subcellular location or enzymatic activity, and typically are active only in a specific compartment of the cell (van Geest and

Lolkema , 2000). Constructs are created in which the gene encoding the reporter protein is fused at different points in the gene encoding the membrane protein; the resulting set of fusion proteins can be exploited to determine at which side of the membrane the fusion sites reside and gain insight into the topology of the membrane protein.

1.3.2 Proteolytic digestion in situ

In a typical approach, proteolytic enzymes are used to cut the loops outside the membrane. It is then possible to analyze the segments protected by the membrane using SDS-PAGE (Kuroiwa *et al*., 1996) .

Alternatively, the rhomboid family of intramembrane proteases can be used. These enzymes cleave specifically transmembrane regions in a specific sequence, enabling identification of membrane spanning segments that contain the target site (Strisovsky *et al*., 2009).

1.3.3 Site directed mutagenesis

In this approach, residues hypothesized to be important for structure or function, such as N-glycosylation sites, Cys residues, iodinatable sites and antibody epitopes, are changed using site directed mutagenesis and the resulting mutant protein analyzed (van Geest and Lolkema , 2000). In addition, tags added at different positions in the protein can help predict the overall topology. Furthermore, antibodies directed against the loop regions can be exploited to determine if a loop is positioned on the other side of the membrane in the mutant protein.

Despite the advances in empirical methods for determining structure, in silico methods for predicting structure are required to better our understanding of diverse transmembrane proteins. Currently available computational tools for predicting structure of transmembrane proteins will be discussed in the next section.

1.4 Predicting structural features of transmembrane proteins

At present there is no general-purpose method for predicting three–dimensional (3D) structures for transmembrane proteins. For water-soluble proteins the most reliable methods for predicting 3D structures use comparative or homology modeling. Homology modeling is based on the identification of known protein structures that resemble the structure of the query sequence, and on the production of an alignment that maps residues in the query sequence to residues in the template sequence. The sequence alignment and template structure are then used to produce a structural model of the query protein. As protein structures are more conserved than protein sequences, detectable levels of sequence similarity are typically associated with significant structural similarity (Marti-Renom *et al*., 2000). The quality of the homology modeling depends on the accuracy of the sequence alignment as well as the quality of the template structure.

In the early 90s, efforts were made to comparatively model the rhodopsin protein from the GPCR family, based on templates derived from bacteriorhodopsin. However, it became clear after the crystal structure of rhodopsin was available that despite similarities in the overall topology and approximate positioning of the helices, the

structure of bacteriorhodopsin is substantially different in terms of helix packing arrangements. The limited sequence similarity observed between the rhodopsin and bacteriorhodopsin sequences also contributed to the inaccuracy of the predicted structural models. At the time, it was concluded that homology modeling alone could not provide accurate structures for GPCRs (Beeley and Sage, 2003). More generally, it was pointed out that a lack of experimentally determined transmembrane structures makes it difficult to find suitable template structures when performing homology modeling.

Our appreciation of the complexity of transmembrane protein 3D structures is growing as more structures are solved. For example, constraints on the length of the transmembrane helices and the packing angles are not as strict as previously thought (Gimpelev *et al*., 2004). Transmembrane proteins containing non-canonical elements (e.g., wide turns, tight turns, kinks and reentrant loops) have been characterized (Riek *et al*., 2001, Riek *et al*., 2008, Viklund *et al*., 2006). Unexpectedly, it was also shown that the 3D structure of transmembrane proteins is not determined purely by protein sequence but is influenced by insertion into the membrane, which is implemented by the translocon complex (Goder *et al*., 2004). Furthermore, functional transmembrane proteins could be part of larger complexes, such as photosynthetic reaction center of the bacterium Rhodopseudomonas viridis, which consists of four subunits L, M, H, each containing 5 alpha-helices and a cytochrome (Deisenhofer *et al*., 1985). Finally, there have been reports of transmembrane proteins that accommodate water molecules (Renthal, 2008) or other ligands and form more diverse structures.

These complexities notwithstanding, methods have been developed for predicting

structural features of transmembrane proteins that can contribute toward predicting 3D structure. Topology predictions provide initial structural information, but the next step towards full 3D modeling requires delineation of the orientations of each transmembrane helix, i.e., the identification of which residues are exposed to the lipid phase and which are packed against the interior of the transmembrane bundle. In addition, predicting structural features such as kinks and reentrant loops is also crucial for full 3D modeling.

1.4.1 Transmembrane protein topology prediction

Consideration of the strong physiological constraints on transmembrane proteins facilitates prediction of which regions are helical and membrane spanning. As described above, the membrane spanning segments have to possess hydrophobic side chains interfacing with the lipids because the lipid bilayer is highly hydrophobic. In summary, constraints imposed by the membrane reduce the number of possible conformations of the protein.

Methods for predicting the topology of transmembrane proteins rely on two key topological features. The first is that transmembrane helices are generally formed by hydrophobic stretches of residues. The second is that regions flanking the hydrophobic stretches contain predominantly positively charged residues, especially on the intracellular side of the membrane: “the positive-inside” rule, whereby short loops enriched with Lys and Arg residues are typically on the intracellular side and vice versa (von Heijne, 1999, Wallin and von Heijne, 1998). Additional features have been identified as characteristic of helix-bundle membrane proteins and are exploited when predicting topology. For instance, connecting loops between membrane helices

are typically shorter than 60 residues (Wallin and von Heijne, 1998).

More than 30 methods have been developed for predicting the topology of helix-bundle membrane proteins (Kernytsky and Rost, 2003). Below is a brief, chronological summary of the main methods, with particular emphasis given to advances made over the last two decades.

Initially, hydrophobicity scales were developed (Kyte and Doolittle, 1982, Engelman *et al*., 1986). These scales classify amino acids according to propensity to contact polar versus non-polar environments, with a high hydrophobicity score indicating tendency to interact with non-polar environments i.e., the membrane. One way of assigning a hydrophobicity score to a given amino acid is to evaluate its hydrophilic and hydrophobic tendencies (Nozaki and Tanford, 1971, Kyte and Doolittle, 1982, Fauchere and Pliska, 1983, Engelman *et al*., 1986, Radzicka and Wolfenden, 1988, Karplus, 1997). Another approach involves analyzing existing protein structures and calculating the probability of a given amino acid to be exposed to the lipid (Wallin *et al*., 1997).

Taking advantage of the hydrophobicity scales they devised, Kyte and Doolittle developed a "moving-window" approach to identify membrane segments. A window of 19 residues is moved along the protein sequence and the sum total of the 19 hydrophobicity scores is calculated for each window. Based on analysis of known structures, Kyte and Doolittle designated a threshold value, above which a window is considered as containing a membrane helix. This approach was designed only to identify transmembrane segments and did not address the inside-outside location of segments relative to the membrane.

The first major advance in transmembrane topology prediction was the TopPred method described by von Heijne (1992). Like previous approaches, TopPred exploited hydrophobicity scales to predict transmembrane segments, but for the first time, these predictions were combined with a simple topological rule: the positive-inside rule (von Heijne, 1992). The observation that there is a strong bias for positively charged residues on the inside facing segments of a transmembrane protein provided a means to identify which predicted topology is most likely correct from a small number of alternatives. Even though the starting point for TopPred was a basic hydrophobicity plot, this method stands out as the first transmembrane topology prediction method.

The MEMSAT method (Jones *et al.*, 1994) generates statistical tables (log likelihoods) from membrane protein data and utilizes a dynamic programming algorithm to evaluate membrane topology models by expectation maximization. The propensity of each amino acid to be in one of five states (inside loop, outside loop, inside helix end, helix middle and outside helix end) is derived from experimentally well-described membrane proteins. Using these propensities, MEMSAT calculates the most probable length, location and topological orientation for each transmembrane segment.

Similarly, TMAP (Persson and Argos, 1994) uses multiple sequence alignments to produce a preference scale. The scores are calculated by statistically analyzing known membrane proteins and serve to locate transmembrane segments. A notable advantage of this method is incorporation of an algorithm for splitting long hydrophobic regions into pairs of transmembrane helices, such regions are a common problem for other methods.

PHDhtm (Rost *et al*., 1996) was the first method to use neural networks (explained in Appendix A) for predicting transmembrane helices. The method initially employs information derived from multiple sequence alignments as input for a system of neural networks. The neural network serves to calculate the likelihood of each residue residing in a transmembrane helix or a loop. Then protein regions of 18 residues are searched for having the highest propensity to be in a transmembrane helix. The preferences are input to a dynamic programming algorithm that identifies the segments most likely to span the membrane.

TMHMM (Sonnhammer *et al.*, 1998) and HMMTOP (Tusnady and Simon, 1998) were the first methods based on Hidden Markov Models (see section 1.5.1.6). TMHMM implements a cyclic model with seven states for transmembrane helix. HMMTOP uses a Hidden Markov Model to distinguish between five structural states (helix core, inside loop, outside loop, helix caps (C and N) and water-soluble domains). The states are connected by transfer probabilities. Dynamic programming is used to match a sequence against the model in order to find the most probable match. Prodiv – TMHMM was developed by Viklund and Elofsson (2004) and incorporates the best features of the earlier TMHMM method.

The first consensus approach was developed by Nilsson *et al*., (2002). Consensus approaches derive from the consensus of topology prediction methods, in this case the methods were: TMHMM, HMMTOP, MEMSAT, PHDhtm and TopPred. Nilsson *et al*. reported that their approach correctly predicts topology for approximately 90% of the structurally determined membrane proteins from both prokaryotic and eukaryotic organisms, a higher accuracy than achieved by any previous method. Furthermore,

they demonstrated that a consensus topology can be predicted for 70% of all membrane proteins in a bacterial genome and for ~55% of all membrane proteins in the eukaryotic genome (Nilsson *et al*., 2002).

Three other consensus methods have been developed. The first by Fariselli *et al*. (2003), who combined a neural-network method with two different HMM methods for predicting topology. The second by Taylor *et al*. (2003), who combined five methods for predicting topology. The third called MetaTM, was developed recently by Klammer *et al*. (2009), who combined six transmembrane helix prediction methods: TopPred, PHDhtm, HMMTOP, TMHMM, PolyPhobius and MEMSAT. Klammer *et al*. claim MetaTM achieves the greatest accuracy yet, with an average prediction accuracy of 86.3%.

Kall *et al*. (2004) developed a method called Phobius, a HMM-based method that simultaneously predicts transmembrane regions and signal peptides. This advance solved the problem of discriminating between signal peptides and transmembrane helices. PolyPhobius, a method developed by the same group (Kall *et al*., 2005), incorporates homology information and further increases the accuracy of predictions.

Support Vector Machines (SVM) have also been used to predict transmembrane protein topology. For example, Yuan *et al*. (2004) used an SVM for per-residue prediction of helices, with a sliding window.

The MINNOU method (Cao *et al*., 2006) is considered an alternative strategy to predicting topology for membrane proteins. Instead of using evolutionary sequence profiles (see section 1.5.1.4), this method uses prediction-based ‘structural profiles’

comprising predictions of relative solvent accessibility and secondary structure for each residue. Though evolutionary profiles in the form of a multiple alignment are indeed used to derive these simple 'structural profiles', the alignments are not used explicitly for the membrane domain prediction and the overall number of parameters in the model is significantly reduced.

MEMSAT3 was described by Jones (2007) and employs a neural network in addition to the dynamic programming algorithm, the latter devised for MEMSAT (1994). The advanced MEMSAT3 uses sequence profiles to train the neural network, in order to produce a consensus topology score across an aligned family of sequences.

Recent prediction methods also consider reentrant loops, which as mentioned above were only appreciated recently. Although the hydrophobic profiles of reentrant loops and transmembrane helices are similar, predicting both structures simultaneously can corrupt topology prediction. Therefore, it was important to develop methods that can differentiate between the structures. Two such methods exist:

OCTOPUS developed recently by Viklund and Elofsson (2008) uses a combination of hidden Markov models and artificial neural networks. OCTOPUS predicts the correct topology for 94% of the sequences.

MEMSAT-SVM was developed recently by Nugent and Jones (2009). The method is a support vector machine-based (SVM) TM protein topology predictor with reported topology prediction accuracy of 89%. The method discriminates between water-soluble and TM proteins with zero false positives. MEMSAT-SVM also attempts to differentiate between signal peptides and reentrant helices, and predicts these

structures with an accuracy of 93% and 44%, respectively.

1.4.2 Predicting helix orientation

The orientation of helices in the lipid membrane (Figure 6) is defined by the helix tilt (τ) and rotation (ρ). The value ρ is defined as the angle between the perpendicular vector (r) from the helical axis (H) to the selected Cα reference residue (blue circle). The value τ is the angle formed between helical axis (H) and the membrane normal (*N*).

Lipid exposure prediction provides information about the probable orientation of the helices. Early attempts to predict helix orientation employed the hydrophobic moment concept (Eisenberg, 1984, Rees *et al*., 1989). The hydrophobic moment is essentially a vector pointing from the helix axis to the most hydrophobic surface of the helix. In these methods, the orientations of transmembrane helices were predicted on the assumption that the helical hydrophobic moments should point out into the lipid phase. Later, however, it was found that hydrophobic moments are poor indicators of the angular orientation of transmembrane helices due to the fact that hydrophobic residues often face both the core of the protein and the lipid (Stevens and Arkin, 1999, Rees and Eisenberg, 2000).

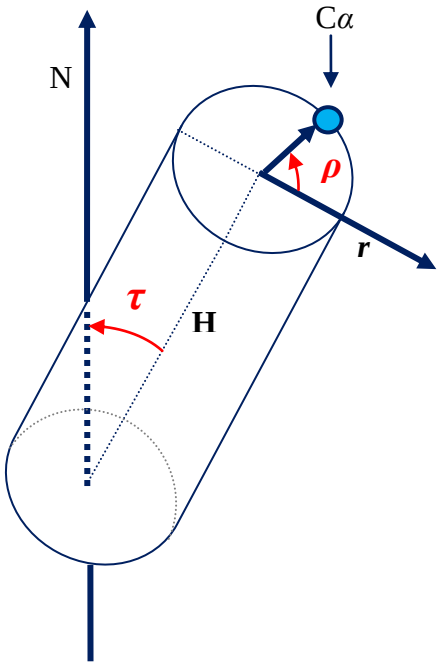


Figure 6: The definition of helix tilt (τ) and rotation (ρ). The value ρ is defined as the angle between the perpendicular vector (r) from the helical axis (H) to the selected Cα reference residue (blue circle). The value τ is the angle formed between helical axis (H) and the membrane normal (N).

In later studies, statistical analyses were conducted using known high-resolution structures of transmembrane proteins with the goal of defining the lipid exposure propensities for each residue in a given transmembrane helix (Donnelly *et al*., 1993, Donnelly 1994). The work of Donnelly is notable, in that it described very clearly the utility of sequence conservation in discriminating between lipid exposed and buried residues. Lipid exposed residues, though required to be highly hydrophobic, are not under any significant steric constraints and so can be evolutionarily quite variable. In contrast, buried residues though also typically hydrophobic, are indeed subject to steric constraints and so are commonly highly conserved evolutionarily in sequence alignments. Unfortunately, these studies were based on a dataset of insufficient size to generate good statistics.

Taking an alternative approach to defining the lipid exposure propensities, Pilpel *et al*.

(1999) proposed a knowledge-based scale. The authors made the assumption that residues which tend to be exposed to the membrane (as opposed to buried within) will be more frequent in the transmembrane segments of single spanning transmembrane proteins than in multi-spanning proteins, whereas residues that prefer to be buried in the transmembrane bundle interior would exhibit the opposite trend.

Other, more advanced methods developed for predicting residue orientation in transmembrane proteins (Beuming and Weinstein, 2004, Adamian and Liang, 2006, Hildebrand *et al*., 2006, Yuan *et al.*, 2006, Park *et al*., 2007, Illergard *et. al*., 2010) will be discussed in Chapter 2.

1.4.3 Predicting kinks and reentrant loops

A significant proportion of transmembrane proteins contain kinks (Riek *et al*., 2001). Yohannan *et al*. (2004) developed an algorithm that predicts kinks with an accuracy of > 90% by identifying peaks of proline in sequence alignments, as they had surmised that 50% of the kinks are due to proline. Another method, based on sequence pattern descriptors, predicts kinks and also other non-canonical helical conformations (Rigoutsos *et al*., 2003). More recently, Hall *et al*. (2009) employed molecular dynamic simulation using isolated helices with the goal of identifying the position of helical kinks in transmembrane helices. The authors reported a capacity to identify about 79% of the proline kinks. Furthermore, recently, Langelaan *et al*. (2010) developed a method for predicting kinks using machine learning and concluded that although kinks are somewhat predicted by sequence, kink formation appears to be driven predominantly by other factors. Langelann *et al*. showed that although the proline amino acid has been advanced as being essential for kinks

formation (Yohannan *et al*., 2004) there are proline residues that do not induce a kink and there are kinks in the absence of proline. Langelann *et al*. remarked that Yohannan *et al*. tested their algorithm on a relatively small set of transmembrane proteins.

More than 10% of transmembrane proteins contain reentrant loops (Viklund *et al*., 2006). A reentrant loop goes part way through the membrane and turns and exits the membrane in the same side it has entered.

Very few reentrant loop predicting methods exist. Viklund *et al*. (2006) developed the method TOP-MOD for predicting reentrant regions with an accuracy of ~70% based on their amino acid composition. TMLOOP also identifies re-entrant loops (Lasso *et al*., 2006). As mentioned above, the method OCTOPUS developed by Viklund and Elofsson (2008) and the method MEMSAT-SVM developed by Nugent and Jones (2009) can predict the existence of reentrant loops in transmembrane proteins as well as their topology.

Other methods that predict the existence of motifs such as signal peptides and signal anchors are SignalP (Bendtsen *et al*., 2004) and TargetP (Emanuelsson *et al*., 2007).

1.4.4 Transmembrane protein 3D structure prediction

Various attempts have been made to develop prediction methods for transmembrane protein 3D structure. Taylor *et al*. (1994) adapted some programs originally developed for predicting water-soluble protein structures to derive a method for predicting 3D structures of integral membrane proteins. The method uses the "variphobicity" (evolutionarily variable and hydrophobic) faces of transmembrane

helices to predict the structure. The method was successfully applied to two protein family sequence alignments (bacteriorhodopsin and rhodopsin).

Helix-helix associations were modeled by Adams *et al*. (1996), on Glycophorin A, using an energy function which searches for the best possible packing interactions between helices. In 2001, a modeling approach was developed by Nikiforovich *et al*. (2001). Their approach combined helical packing, based on the bacteriorhodopsin template, and selection of low-energy conformers for loops that are closest to the bacteriorhodopsin X-ray structure. Using this method the authors were able to reproduce the bacteriorhodopsin structure.

Fleishman and Ben-Tal (2002) used knowledge of residue environment preferences to predict the likely arrangement of transmembrane helices, on the basis of a rule: "small residues go inside". This method predicted successfully the native structure of transmembrane protein glycophorin A. In the same year Ledesma *et al*. (2002) produced a model for Uncoupling protein 1 (UCP1), using a computational docking method. Later Chen and Chen (2003) used a Monte Carlo method for protein folding and successfully predicted the seven helix bundle structure of rhodopsin I.

Pellegrini-Calace *et al*. (2003) developed a method (FILM) for predicting small membrane protein structure based on a method previously developed for predicting tertiary structure of water-soluble proteins (FRAGFOLD). The method is based on the assembly of super secondary structural fragments taken from a library of proteins with known structure. A standard simulated annealing algorithm is used to narrow the search of conformational space, which pre-selects fragments from a library of highly resolved protein structures. The method was applied to small membrane proteins of

known structure and was able to predict with a reasonable accuracy level the helix topology and protein conformation.

Another modification of a water-soluble protein modeling program (ROSETTA) was developed by Barth *et al*. (2007), which attained near atomic accuracy for several small membrane proteins. More recently, the same group developed a method for predicting the structure of large transmembrane proteins. The newer method constrains helix-helix packing arrangements at particular positions according to predictions from sequence analysis or in line with empirical data and produced near-native models for 9 out of 12 tested proteins (Barth *et al*., 2009)

Fuchs *et al*. (2009) showed that applying water-soluble methods that predict helix-helix interaction (contact map) to membrane proteins was not very effective. To address this issue, they developed a method (TMHcon) based on neural networks, which predicts helix-helix contacts in transmembrane proteins. In addition to the input features commonly used for contact prediction of soluble proteins, such as windowed residue profiles and residue distance in the sequence, the network also incorporates features that apply to membrane proteins only, such as residue position within the predicted transmembrane segment and orientation toward the lipid environment. The obtained neural network can predict contacts between residues in transmembrane segments with nearly 26% accuracy.

TMhit is another method that predicts helix-helix interaction in transmembrane proteins (Lo *et al*., 2009). The method incorporates contact propensities, various sequence, physico-chemical, and structural information in a two-level architecture using support vector machines (SVMs). In the first level, contact residues are

predicted and their pairing relationship or connectivity is further predicted in the second level.

Recently Nugent and Jones (2010) developed a novel approach to predicting lipid exposure, residue contacts, helix-helix interactions and the optimal helical packing arrangement of transmembrane proteins. They employed molecular dynamics data to label residues potentially exposed to lipid, trained and cross-validated a support vector machine (SVM) classifier to predict for each residue the probability of lipid exposure, reporting an accuracy rate of 69%. The resulting information is combined with additional features to train a second SVM to predict residue contacts, which in turn are used to determine helix-helix interactions. An accuracy rate of up to 65% was reported when using stringent cross-validation conditions for a non-redundant test set.

Despite this progress in predicting the 3D structure of membrane proteins, more advanced methods are needed that are reliable and fast enough to apply on a genomic scale.

1.5 Computational approaches to characterizing proteins

1.5.1 Sequence similarity methods

If two proteins have diverged from a common ancestor they are defined as homologous proteins and are likely to have similar sequences. Computational sequence comparison detects homologous proteins; it takes as input two sequences and outputs the similarity between them. Sequence comparison can also serve to

delineate the most probable set of point mutations, deletions and insertions that define the evolutionary relationship between the two proteins.

Computational approaches to sequence alignment generally fall into two categories: global alignment and local alignment. Global alignments span the entire length of the query sequence. Conversely, local alignments identify regions of similarity within long sequences that are often widely divergent elsewhere.

The algorithm for calculating either local or global sequence similarity does not give equal weight to each amino-acid aligned. Instead, scoring matrices are used, which give dissimilar weights to replacement of different amino acids. The most commonly used scoring matrices are PAM (Dayhoff *et al.*, 1978) and BLOSUM (Henikoff and Henikoff, 1992), described in detail in Appendix B.

In addition, sequence alignment methods can be divided into three classes based on the information used for the alignment:

1. Sequence-sequence alignments are pairwise methods that compare sequences one against one.

2. Profile-sequence methods compare one sequence to an aligned family of sequences.

3. Profile-Profile methods compare two aligned families of sequences.

Pairwise methods align the sequences assuming that all amino acids are equally important. However, in reality, this is not the case; at some positions the amino acids are conserved while at others they are not. The conserved amino acids are likely to be more important for the protein structure and function. Profile based methods exploit

this information for the alignment and therefore, are more sensitive than pairwise methods. The next sections summarize the main algorithms for sequence alignment.

1.5.1.1 The Needleman -Wunsch algorithm

The Needleman–Wunsch algorithm performs global alignments of pairwise sequences. The Needleman-Wunsch algorithm applied dynamic programming for the first time to sequence comparison (Needleman and Wunsch, 1970). It maximizes the number of matches between the sequences along the entire length of the two sequences, thus the algorithm aligns the two sequences from the first residue to the last even if only the middle of the sequences is similar. Insertions and deletions are considered by conferring appropriate costs to gap opening and gap extensions.

This method is applied in the current thesis and therefore is explained in more detail.

The Needleman-Wunsch algorithm starts with initialization of the score matrix: a matrix with M+1 columns and N+1 rows is created where M and N correspond to the length of the sequences to be aligned. Then the matrix is filled: scores for aligned residues are specified by the designated substitution matrix. Substitution matrices describe the evolutionary rate at which one character in a sequence changes to another character over time, where S(i,j) is the similarity score for residues i and j.

In the next step, for each position, $M_{i,j}$ the maximum score at position i,j is calculated. In the original publication from 1970, gap is not penalized and the maximum score is:

$$M_{i,j} = MAX_{h<i,k<j} \{ M_{h,j-1} + S_{i,j}(Ai,Bj), M_{i-1,k} + S_{i,j}(Ai,Bj)\} \quad (1)$$

When adding gap penalty (d) to the algorithm, which is a negative score, the

maximum score is:

$$M_{i,j} = MAX\{ M_{i-1, j-1} + S_{i,j}(Ai,Bj), M_{i,j-1} + d , M_{i-1,j} + d\} \qquad (2)$$

In the last step traceback is performed. The traceback begins with the last cell to be filled with the score, i.e., the bottom right cell. Traceback takes the current cell and looks to the neighbor cells that could be direct predecessors. There are three possible moves: diagonally (toward the top-left corner of the matrix), up or left. The algorithm for traceback chooses as the next cell in the sequence one of the possible predecessors. Continuing with the traceback step, the algorithm gets to a position in column 0, row 0 which tells us that traceback has completed with the best scored global alignment. The alignment is deduced from the values of cells along the traceback path, taking into account the values of the cell in the traceback matrix.

A similar algorithm to Needleman-Wunch is the Smith-Waterman algorithm, which applied dynamic programming to local alignment of sequences.

1.5.1.2 The Smith-Waterman and FASTA algorithms

The Smith-Waterman algorithm (Smith and Waterman, 1981) performs local alignments of pairs of sequences, i.e. it identifies the most similar region shared between two sequences. The method employs a dynamic programming algorithm in a similar way to the Needleman-Wunsch algorithm except that negative scoring matrix cells are set to zero. Backtracking starts at the highest scoring matrix cell and proceeds until a cell with score zero is encountered, producing the highest scoring local alignment.

The Smith-waterman algorithm is time demanding. A more efficient alignment method is FASTA. The FASTA algorithm (Lipman and Pearson, 1985, Pearson and Lipman, 1988) is a heuristic approximation to the Smith-Waterman algorithm, which reduces the time required by matching words of a given length. The length chosen for the word impacts the speed and sensitivity of the algorithm. The method identifies regions of similar sequences before performing an optimized search using a Smith-Waterman type of algorithm.

The FASTA algorithm can be used to search databases for homologous proteins, but is still not fast enough. A more advanced and faster algorithm is BLAST.

1.5.1.3 The BLAST algorithm

The BLAST (Altschul *et al*., 1990) algorithm searches a corresponding sequence database by using a heuristic algorithm to find similar database sequences. First BLAST locates words (with k letters) in the query sequence with match score above a defined threshold, T, when compared to sequences in the database, using a scoring matrix. Then BLAST begins to make local alignments from these initial matches, by locating neighborhood words that again must have a match score of at least the threshold. However, if the score is lower than this pre-determined T, the alignment will cease to extend, preventing areas of poor alignment from being included in the BLAST results. The algorithm extends the alignment in both directions.

By aligning only to sequences that satisfy a requirement of having a score of at least the threshold, BLAST performs far fewer local alignments than FASTA which performs local alignments on the full sequences. BLAST is therefore much faster than

FASTA.

A more advanced method than BLAST is PSI-BLAST (Altschul *et al*., 1997). The PSI-BLAST method is applied in the current work and is therefore explained in more detail in the section below.

1.5.1.4 The PSI-BLAST algorithm

PSI-BLAST (Altschul *et al*., 1997), Position-Specific Iterated BLAST, identifies homologous proteins iteratively. PSI-BLAST is one of the most commonly used and powerful methods for detecting sequence similarity (Jones and Swindells, 2002).

PSI-BLAST, a profile–sequence alignment method, introduces evolutionary information by constructing protein sequence profiles. Multiple sequence alignments and corresponding sequence profiles represent one of the most significant methodological improvements with impact on alignment accuracy. This methodological approach was not new at the time PSI-BLAST was published. Already in 1987, Gribskov *et al.* used profiles for homology searches, but PSI-BLAST appeared to work better than any other profile-based search tool that had existed previously (Jones and Swindells, 2002). The profiles are obtained by computing the frequency of different residues in each alignment position. A sequence profile lists a preference for the 20 standard amino acid residue types at each position in a given multiple sequence alignment. Using sequence profiles adds more information to the alignment regarding importance and conservation of specific regions. The profile contains more information about the sequence family than a single sequence.

The PSI-BLAST algorithm (Figure 7) automatically generates a multiple alignment from the output of an initial BLAST similarity search. This alignment is then used to create a position-specific score matrix (PSSM), or profile, with dimensions n x 20, where n is the length of the sequence. For each row, a substitution score for each of the 20 amino acids is given. The main difference between PSSMs and standard substitution matrices is that the score for the same amino acid type can differ depending on its position within the sequence. The PSSM is used to search the database. While searching for additional similar protein sequences the PSSM matrix is updated after each iteration.

The search may be iterated many times, as new significant similarities are found. The result of such a search is a list of possible homologues, sorted by E-value. The E-value is a statistical score which represents the number of times one would expect to get a hit with the same or better score by chance. The E-value for a given alignment depend on the length m and n of the sequences and on the alignment score S. The parameters K and λ are constants that depend on the search space size and the scoring system used. The E-value is calculated as:

$$E = K * n * m * e^{-\lambda S} \quad (3)$$

The lower the E-value is, the higher the probability that the query and match are homologous. For example, the meaning of an E-value equal to 1 is that in a database of the current size one might expect to see one match with this score or better, simply by chance.

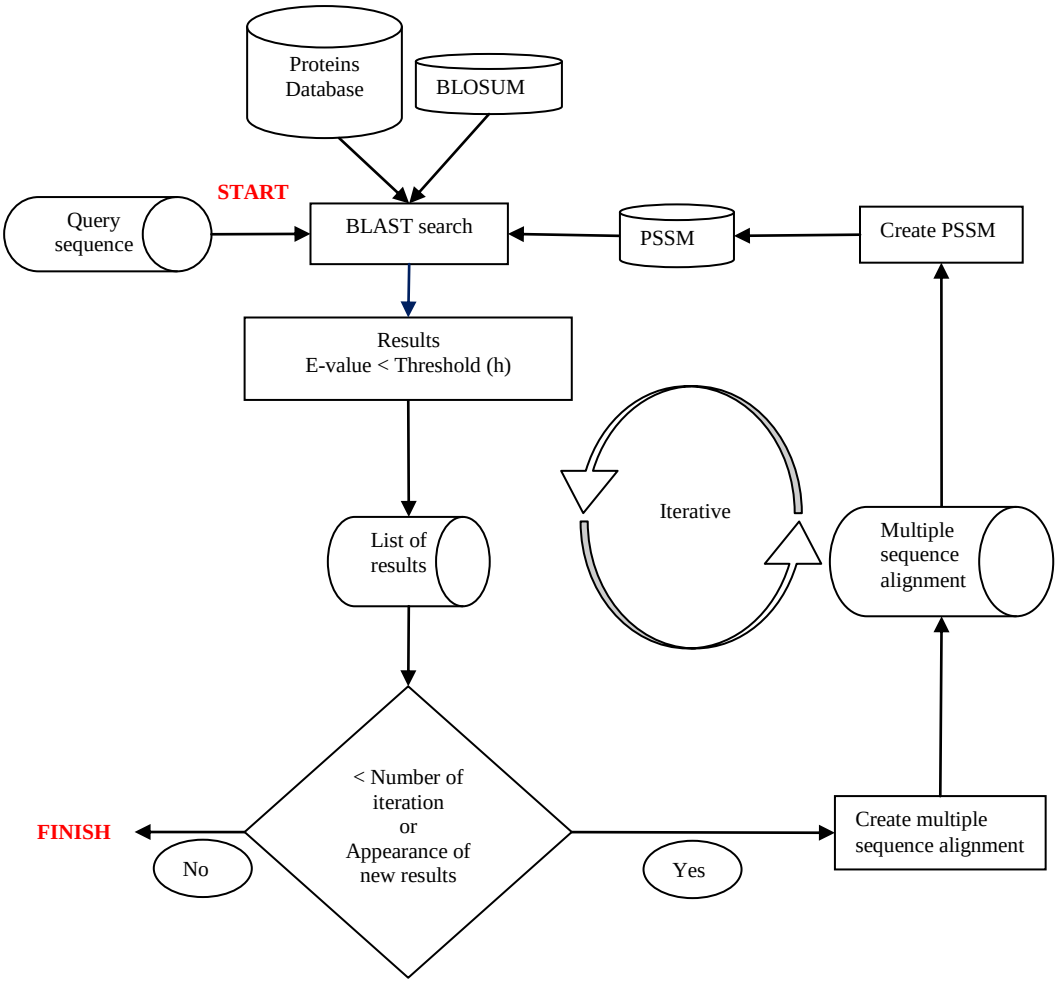


Figure 7: Schematic overview of PSI-BLAST: In the *first step* a BLAST search is performed using a substitution matrix (BLOSUM). Sequences below a given E-value threshold are listed and used for multiple sequence alignment and converted into a PSSM. In the *second step* the PSSM constructed in the first step is used to search the sequence database. *Following steps*: Second step is repeated iteratively, each time a new PSSM is constructed, until no more sequences under a threshold E-value are added or until a given maximum number of rounds have been accomplished. The result is a list of sequence alignments from the final round.

The development of profile–sequence alignment methods such as PSI-BLAST has led to a great improvement in sensitivity over sequence–sequence alignment methods.

1.5.1.5 Profile – profile algorithms

Another significant improvement in sequence alignment algorithms was achieved by

developing profile-profile algorithms. Profile-profile algorithms align two sequence profiles against each other; evolutionary information is included for both query and database sequences.

Several groups have developed profile–profile alignment methods (Pietrokovski,1996, Rychlewski *et al*., 1998, Yona and Levitt , 2002, Sadreyev and Grishin, 2003). The idea behind all the methods is identical; a pair of sequence profiles is used instead of a pair of sequences for the alignment. However, the alignment calculation differs: Rychlewski *et al*. (1998) calculate the similarity score between positions in two profiles by calculating the average of scores between all amino acid pairs according to the probability distributions in each profile; Yona and Levitt (2002) proposed a scoring formula based on a theoretical measure of differences between the two probability distributions represented by the profiles; Sadreyev and Grishin (2003) generates scores for matching positions of the two profiles by using a scheme of log-odds ratios and Pietrokovski (1996) used Euclidean distances between the profiles.

Profile-profile methods have been shown to improve homology detection among proteins to a greater extent than profile–sequence methods (Rychlewski *et al*., 1998, Sadreyev and Grishin, 2003).

1.5.1.6 Hidden Markov Model based methods

Hidden Markov Models (HMMs) are probabilistic models that were originally applied to the problem of speech recognition (Jelinek *et al*., 1975), and were later applied to biological sequence analysis (Churchill, 1989). HMMs have been applied to many problems in computational biology.

Krogh *et al*. (1994) were the first to delineate an HMM architecture for protein sequence alignment, termed the profile HMM, which is another representation of multiple sequence alignment profiles. Profile HMMs are thus similar to simple sequence profiles, but in addition to the amino acid frequencies in the columns of a multiple sequence alignment, the columns also contain information about the frequency of inserts and deletions and can also incorporate other types of data (such as secondary structure propensities). In building a profile HMM, an existing multiple alignment is given as input. For each column of the multiple alignment, a 'Match' state models the frequencies of the residues in the column. An 'Insert' state for each column enables insertion of residues between that column and the next one, and 'Delete' state enables deleting of the residue between that column and the next one. The states in the profile HMMs are sequentially connected so that each position in the multiple sequence alignment is represented by a 'Match' state, an 'Insert' state and a 'Delete' state. The model starts in 'Begin' state and ends with 'End' state. The probabilities of the profile HMM are converted to log-odds scores, which can then be summed.

One of the most well-known software packages used for generating profile HMMs automatically from multiple sequence alignments is HMMER (Eddy, 1998). Figure 8 shows the architecture of HMMER model. The architecture is linear and corresponds to a multiple sequence alignment, i.e., match states correspond to the conserved columns of the alignment, insert states to the insertions and delete states to the deletions. In addition, transitions between the states represent the deletions and the insertions.

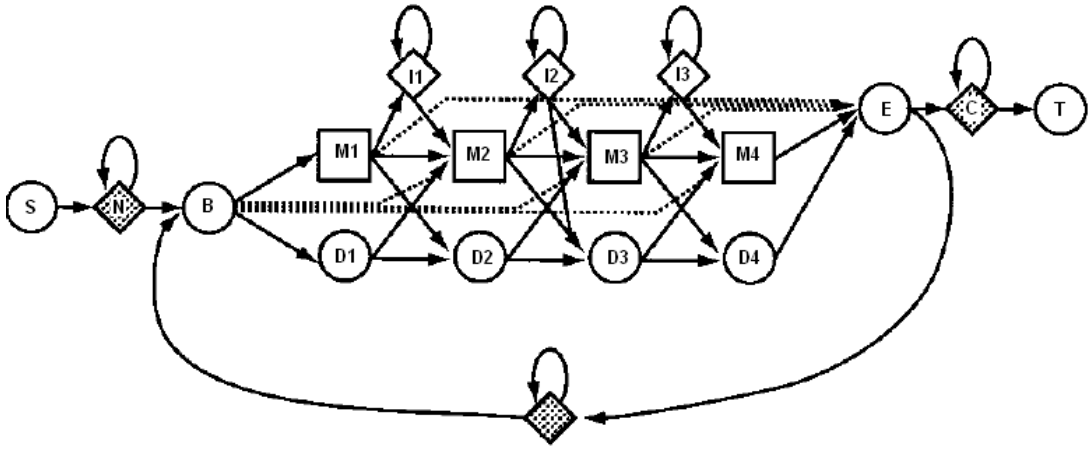


Figure 8: The HMMER model architecture (Eddy, 1998). It starts from Begin (B) state and finishes at the End (E) state. The N and C states are the N and C terminals. Match states (M) correspond to the conserved amino acids, insert state (I) to the insertion and delete state (D) to the deletions.

HMMs can also be used in profile HMM-profile HMM, methods which are similar to profile-profile methods. HHpred (Soding *et al*., 2005) was the first server to employ profile HMM-profile HMM comparison (uses the program HHsearch), based on a novel statistical method. Using HMMs both for the query and the database greatly enhances the sensitivity and selectivity of the method (Soding *et al*., 2005).

1.5.2 Computational approaches for classifying proteins

Only a small fraction of annotated proteins have been characterized functionally (Ursing *et al*., 2002). The most powerful method for characterizing the biological function of a protein is to search for other proteins in databases with sequence or structure similarity.

Databases that classify proteins into families are based on protein resemblance in sequence, structure and/or function. Protein classification is an important task in bioinformatics, as it provides valuable clues to the structure and functions of unknown

proteins and can be employed for evolutionary and statistical studies of protein families. Moreover, classifications can be useful during large scale annotation of proteins, as required by the growing body of sequence data generated by complete genome sequencing.

When classifying according to sequence similarity, the classification can be based on full length sequence or on domains or motifs. A protein domain is defined as a section of protein sequence that encodes for a structure that can function independently of the rest of the protein chain. Typically each domain forms a 3D structure that is independently stable. Wetlaufer was the first to propose the domain concept (1973); he defined domains as stable units of protein structure that fold autonomously. Proteins comprising more than one structural domain are called multidomain proteins and are often multifunctional proteins (Chothia, 1992). A domain can appear more than once and in various configurations with other domains (Apic *et al*., 2001). In a multidomain protein, each domain may function independently or in a concerted manner with its neighbors. Liu *et al*. (2004) examine Pfam classified families and found that most transmembrane proteins (78% for archaea and prokarya and 67% for eukarya) contain only a single classified membrane domain.

The algorithms most commonly used for classifying proteins are based either on sequence similarity, on structural similarities or on combinations of sequence and structural similarities.

1.5.2.1 Classification based on sequence similarity

The underlying assumption of classification based on sequence similarity, is that

proteins with high sequence identity are likely to share the same structure. However, there are many examples of structurally similar proteins that do not display significant sequence similarities. Accordingly, most classification methods that rely solely on sequence data fail to recognize up to 30% of extremely distant homologs (Engelman *et al*., 2003, Gough *et al*., 2001). Classification based on sequence similarity is performed by searching for similarity to a given protein with a chosen specified threshold. Specifying the threshold can be a difficult task. A restrictive threshold can generate few matches and miss sequences that have diverged during evolution; alternatively a less restrictive threshold can result in a list that includes unrelated proteins, i.e., false positives.

There are three types of sequence-based classification:

1. Full length sequence analysis: In which the full length sequence is used for classification.

   ProtoNet (Sasson *et al.,* 2003) and ProtoMap (Yona *et al.,* 2000) are examples for databases which are based on full length classification.

2. Domain/motif analysis**:** This approach was prompted by the observation that some regions have been better conserved than others during evolution. Such conserved regions are generally important for the function of a protein and for the maintenance of its 3D structure. Analysis of the constant and variable properties of sets of similar sequences, enabled derivation of a signature for a protein family or domain, which distinguishes its members from all other unrelated proteins. Thus, the underlying assumption of this type of

classification is that proteins with similar domains are likely related even if overall they display low sequence similarity. The limitation of domain-based classification is that many proteins possess several domains whereas some proteins do not contain recognized domains. In addition, it is not possible to predict domains for some very small families (e.g., that comprise 2 members).

Some of the best known protein-related databases are based on motif or domain classification, for example: Pfam (Bateman *et al.*, 2004), which is broadly used in the current work and therefore will be described in detail (in section 1.5.2.1.1), PROSITE (Falquet *et al.*, 2002), BLOCKS (Henikoff *et al.*, 2000), TIGRFAM (Haft *et al.,* 2003) and PRINTS (Attwood *et al.*, 2002).

3. Phylogenetic analysis: In this analysis, proteins are classified together if they are inferred to be orthologs. Orthologs are genes of common origin that have diverged through evolution. Typically, orthologous proteins have the same domain architecture and the same function, although there are many exceptions and complications to this generalization, particularly among multicellular eukaryotes.

   COGS (Tatusov *et al.*, 2001), is a database which phylogenetically classifies the entire encoded proteins (both predicted and characterized).

1.5.2.1.1 Pfam database

In the Pfam database (Bateman et. al, 2004), the protein sequences from SwissProt and TrEMBL are organized into protein domain families. The classification is semi-automatic and is based on multiple protein alignments that are used to derive profile-

HMMs for the protein families. The HMMs are generated automatically using HMMER (Sonnhammer *et al.*, 1998). 74% of protein sequences have at least one match to Pfam. Given a new sequence, it is possible to evaluate the probability of that sequence belonging to the family modelled by a given HMM. A similarity score is associated with a new sequence based on the most probable path through the HMM which generates the input sequence.

Pfam provides a high quality description of each protein family, including text description about function, cellular location, relevant literature references and links to taxonomic groups in which the family is found.

Pfam families are categorized as A or B. Pfam-A is the partially manually curated portion of the database that contains over 10,000 entries. For each entry a protein sequence alignment and a hidden Markov model is stored. Because the entries in Pfam-A do not cover all known proteins, an automatically generated supplement is provided called Pfam-B. Pfam-B contains a large number of small families automatically generated from clusters produced by the ProDom database in the early releases (Corpet *et al.*, 2000) and by the ADDA database (Heger *et al.*,2005) in recent releases (since release 23.9, 2008) . Although of lower quality, Pfam-B families can be useful when no Pfam-A families are found. Pfam data are freely accessible via the web.

The Pfam-A database is generated in a semi-automated process, starting from a seed based on multiple alignments. After manual inspection, an HMM is built and used to search the database, thus members are added to the seed alignment and the process is repeated.

Pfam proteins are not only classified into families but also, groups of related families are classified into clans (Finn *et al*., 2006). A clan is a collection of Pfam-A entries that are judged likely to be homologous. A clan contains two or more Pfam families that have arisen from a single evolutionary origin. Clans are built manually and based on various sources of information: the primary literature, known structures, profile−profile comparisons and other databases such as SCOP. Clan classifications were developed because of the difficulties in classifying proteins into families. It was found that there are many related Pfam families, the members of which effectively overlap. Conversely, it was found that for some large, divergent families it was not possible to build a single HMM that detects all members of the family.

1.5.2.2 Classification based on protein structure

Structural protein classification creates groups according to 3D structure similarity. The most widely used structure classification resources are SCOP (Murzin *et al*., 1995) and CATH (Orengo *et al*., 1997).

SCOP provides a detailed and comprehensive description of the structural and evolutionary relationships between proteins with solved 3D structures. SCOP classifies the proteins into a four-level hierarchy: Family (proteins with significant sequence similarity), Superfamily (proteins with low sequence similarity, but with structural and functional features suggesting a common evolutionary origin), Fold (superfamilies with major structural similarity) and Class (high level classification). SCOP classification is manual.

In CATH, the classification of protein domain structures is created using a

combination of manual and automatic methods. There are four hierarchical levels in CATH database: Class, Architecture, Topology (fold family) and Homologous superfamily (Orengo *et al*., 1997). When classifying a new domain, if it has sufficiently high sequence and structural similarity with a domain that has been previously classified in CATH, the classification is automatically assigned. Otherwise, the domain is classified manually.

1.5.2.3 Classification based on sequence and structure

InterPro (Apweiler *et al.*, 2001) is a database that attempts to integrate the advantages of each approach to classification.

InterPro is an integrated documentation resource for protein families, domains, regions and sites. InterPro combines a number of databases (Pfam, PRINTS, PROSITE, ProDom and TIGRFAMs) that use different methodologies and a varying degree of biological information on well-characterized proteins to derive protein signatures. By collating databases, InterPro capitalizes on individual strengths, producing a powerful integrated database and diagnostic tool (Apweiler *et al.*, 2001).

1.5.2.4 Classification of transmembrane proteins

As mentioned above, too few transmembrane protein structures have been solved to allow classification of transmembrane proteins based on structure. In a similar approach to that underlying the Pfam database, Liu *et al*. (2002, 2004) classified transmembrane proteins from 26 genomes into 637 families according to number of transmembrane helices and sequence similarity. They report that the majority of integral transmembrane proteins have single domains unlike soluble proteins, which

typically encompass several domains.

Other attempts have been made to classify efficiently transmembrane proteins. Aria *et al*. (2004) focused on 87 complete prokaryotic genome sequences to develop a method based on 'topology similarity', in which a score was calculated by comparing the length of loop regions. Suwa *et al*. (2000) developed a classification method based on computing the polar energy surface, which can reveal characteristic interaction patterns for individual helices. The transmembrane proteins families in C. elegans (and human orthologs) were classified by Remm and Sonnhammer (2000) on the basis of sequence similarity using Hidden Markov Model techniques.

There have been many studies focused on classifying G protein coupled receptors (GPCRs), a large membrane protein family important physiologically and pharmacologically due to key roles in regulating cellular growth, death and metabolism. This family is difficult to classify using only sequence homology as members are highly divergent at the sequence level. Hedman *et al*. (2002) developed a method to classify GPCRs that combines topological information with sequence alignment (discussed in more detail in chapter 4). Recently, Huang *et al*. (2004) attempted to classify GPCRs using a bagging classification tree algorithm based on amino acid composition. Inoue *et al*. (2004) developed a binary topology pattern method for GPCR classification, in which a binary pattern was obtained for each functional class by assigning binary loop threshold lengths (short loop/long loop).

More recently, Marsico *et al*. (2010) developed a technique called structural fragment clustering, which learns sequential motifs from 3D structural fragments in transmembrane proteins. They concluded that structural fragment clustering enables

sequence motifs to be linked to function. Once characterized, sequence motifs can be used to identify and characterize membrane proteins in novel genomes.

Helical transmembrane proteins from the SCOP and CATH databases were analyzed by Neumann *et al*. (2010). They concluded that effective classification of transmembrane proteins with only a few membrane-spanning helices requires integration of more fine-grained structural features such as helix-helix interactions and reentrant regions.

Several databases of transmembrane proteins have been constructed and are accessible through the Web. These databases are summarized in Table 1.

Table 1: Existing transmembrane protein databases.

Database Name	**Description**	**Reference**
Mptopo	All currently known high-resolution transmembrane protein structures with links to the PDB and PubMed entries. Additionally, the database includes a list of proteins with unknown 3D structure, but with topology that has been experimentally annotated using low-resolution techniques.	Jayasinghe *et al*., 2000
PDBTM	Database of known transmembrane protein structures proteins, listed in the Protein Data Bank (PDB).	Tusnady *et al*., 2004
OPM	Includes all unique experimental structures of transmembrane proteins. In addition it provides spatial arrangements of membrane proteins with respect to the hydrocarbon core of the lipid bilayer.	Lomize *et al*., 2006, Lomize *et al*., 2007
CAMPS	Contain transmembrane proteins with three or more predicted transmembrane helices. Proteins were subjected to single-linkage clustering using only sequence alignments. These clusters were further subdivided into functionally homogeneous subclusters according to the COG. The clusters are thus designed to reflect three main levels of interest for structural genomics: fold, function, and modeling distance.	Martin-Galiano and Frishman , 2006

TMPad	Integrated structural database for helix-packing folds in transmembrane proteins. It integrates experimentally observed helix−helix interactions and related structural information for transmembrane proteins.	Lo *et al.*, 2011
Mplot	Provides a quick and easy way for structural biologists to analyze, visualize and plot tertiary structure contacts of helical transmembrane proteins.	Rose *et al.*, 2010
GPCRDB	Includes all G-protein coupled receptors and provides data about sequences, ligand binding constants and mutation	Horn *et al.*, 1998, Horn *et al.*, 2003
TCDB	Transporter classification (TC) system that classifies all transmembrane transporters.	Saier *et al.* 2006, Saier *et al.*, 2009

1.6 The present work

Currently, there is no efficient and accurate method for classifying all transmembrane proteins in an automated way. Since the number of known 3D structures is low, an effective and reliable way to classify transmembrane protein into families based on their sequence must be developed. Such a classification would require a method that reliably detects distant homology between transmembrane proteins. The aim of the present work was to develop an automated method for detecting homology among transmembrane proteins, which predicts reliable and true relationships for the tested protein based on sequence alone.

In the current work a method was developed that uses sequence similarity, topology, predicted structural features (predicted residue lipid exposure) and loop lengths to find homology between transmembrane proteins. The method relies on the assumption that protein structures are more conserved than protein sequences among

homologs and therefore, combining structural information with a simple sequence alignment will improve homology detection (Chothia and Lesk, 1986, Kaczanowski and Zielenkiewicz, 2010, Marti-Renom *et al*., 2000).

Chapter 2

Predicting the lipid exposure of transmembrane proteins

2.1 Introduction

Predicting the three-dimensional (3D) structure of transmembrane proteins remains a challenging task. A simpler initial task, which can serve as a stepping-stone toward predicting 3D structure, is predicting the relative exposure of each residue to the membrane environment, i.e., predicting whether a residue faces the lipid environment or is buried inside the protein.

For water soluble proteins, calculating solvent accessibility has proved quite informative for identifying protein function and domains (Wodak, 1981). In addition, solvent accessibility can be used as additional information when aligning regions with remote sequence identity (Gaboriaud *et al*., 1987, Lemesle-Varloot *et al*., 1990). The concept of solvent accessibility for water soluble proteins was introduced by Lee and Richards (1971). The driving force during folding is the hydrophobic effect, where folding occurs such that unfavorable interactions between hydrophobic residues and the hydrophilic environment are minimized (Honig *et al*. 1995). Accordingly, folded water soluble proteins consist of a hydrophobic interior and hydrophilic exterior. Therefore, predicted solvent accessibility can indicate whether a given residue is

interior or exterior and typically is defined either numerically, the real-valued solvent accessibility, or as a binary classification into buried versus exposed states. Alternatively, Rost and Sander (1994) classified a relative solvent accessibility into three and ten states when predicting accessibility for water soluble proteins. For water soluble proteins many methods have been developed for predicting accessibility. However, only a few such methods have been developed for transmembrane proteins.

Early studies of the bacteriorhodopsin structure suggested that membrane proteins are "inside-out" relative to water soluble proteins, i.e., that they consist of a hydrophilic interior and a hydrophobic exterior (Engelman *et al*,. 1980, Rees *et al*., 1989). However, later it was found that the "inside-out" rule is not completely accurate (Rees and Eisenberg, 1999, Stevens and Arkin, 1999). Transmembrane proteins typically pass through the membrane multiple times. In order to satisfy the hydrogen-bonding requirements of the polar back-bone atoms, transmembrane proteins adopt the architecture of alpha-helical bundles in the regions situated in the membrane. Accordingly, transmembrane proteins face three distinct environments: a hydrophobic lipid environment inside the membrane, a hydrophilic water environment outside the membrane and an interface region rich in phospholipid head-groups. Therefore, it is energetically favorable for transmembrane proteins to expose different types of residues in the different regions (Illergard *et al*., 2010).

As discussed in detail in Chapter 1, early attempts to predict helix orientation were done using the hydrophobic moment concept (Eisenberg, 1984, Rees *et al*., 1989). However, hydrophobic moments were found out to be a poor indicator of angular rotation for transmembrane helices (Stevens and Arkin, 1999, Rees and Eisenberg,

2000). In later studies, a statistical analysis was conducted on known high-resolution structures of transmembrane proteins to find the lipid exposure propensities of the different residues (Donnelly *et al*., 1993, Donnelly 1994). It was discovered that the buried residues are highly conserved relative to the exposed residues.

Based on this finding, Beuming and Weinstein (2004) developed a method for predicting if transmembrane protein residues are buried in the core of the transmembrane helix bundle or exposed to the lipid environment. The method uses information about residue distribution collected from solved structures and combines it with evolutionary criteria about conservation (Briggs *et al*., 2001). This method performed with at most 80% accuracy when predicting if a residue is lipid exposed or buried.

Later, Adamian and Liang (2006) developed a method for predicting transmembrane helix orientation – LIPS (LIPid-facing Surface). Their method predicts the face of the transmembrane helix exposed to the membrane and not the hydrophobicity status of individual transmembrane residues. Admian and Liang's method is based on a canonical helical face model whereby the surface of each helix is partitioned into seven surface patches (faces) that could interact with lipids or other helices. It allows collective assessment of the evolutionary and physico-chemical properties for each of the seven faces formed by residues centered at one of the seven positions. They identify lipid exposure with an accuracy of about 88% from the sequence information alone. The LIPS server is available online at http://gila.bioengr.uic.edu/lab/larisa/lips.html.

Park and Helms (2006) studied in more detail the correlation between conservation

patterns and empirical scales that score the exposure pattern of transmembrane helices. They carried out a large scale benchmarking of the prediction scales proposed so far. Unsurprisingly, this analysis revealed that scales incorporating structural data show stronger correlation with exposure patterns than hydrophobicity-based scales. This conclusion was expected as structure based scales were parameterized explicitly for the purpose of predicting buried versus lipid-exposed faces of transmembrane helices. The other scales (hydrophobicity-based scales) were developed before high-resolution structural data existed. In light of their analysis, Park and Helms proposed a framework that combines sequence conservation patterns and empirical scales, but found that improvements gained from combining the two sources of information were not dramatic in almost all cases.

Hildebrand *et al*. (2006) described a computational method for predicting whether a given residue is located at a helix-helix interface in the membrane. They show that when the sequence motifs typical for membrane channels and transporters were exploited for predicting helix-helix contacts (i.e., the context of a residue was taken into account), the quality of prediction rose by 16% to an average value of 76%, compared to an equivalent approach when only single amino acid positions were taken into account.

Yuan *et al*. (2006) developed a method to predict the solvent accessible surface areas, with resulting correlation coefficients between predicted and observed accessible surface areas of around 0.65. The method involved finding the best threshold of accessible surface areas to differentiate between residues exposed to the lipid environment or buried inside a protein. The method is based on support vector

regression (SVR).

Park *et al*. (2007) developed TMX (TransMembrane eXposure), a method for predicting the burial status of residues in transmembrane proteins. TMX derives positional scores of transmembrane residues based on profiles and conservation indices. Then, a support vector classifier is used to predict burial status. An accuracy of 78.71% was reported for a benchmark data set.

Rose *et al*. (2009) generated a server for predicting the orientation of transmembrane helices in channels and other membrane proteins (membrane–coils) called RHYTHM **(**http://proteinformatics.de/rhythm). The prediction is based on precalculated packing files and evolutionary information from sequence patterns collected from a representative dataset of transmembrane proteins. The program uses two types of position specific matrices to account for the different geometries of packing in channels and transporters or other membrane proteins. The average AUC-values for the prediction of helix–helix contacts was reported to be 0.72 for channels and 0.68 for membrane–coils, respectively.

Recently, Wang *et al*. (2010) developed an additional method for predicting the burial status of residues in transmembrane proteins. The method incorporates physicochemical scales and conservation indices to produce an efficient prediction model using least squares support vector machine (SVM). In least squares SVM one finds the solution by solving a set of linear equations instead of a convex quadratic programming problem for classical SVMs. Wang *et al*. reported that the prediction accuracy of this method was much better than reported for previous approaches.

Illergard *et al*. (2010) compared the published methods for predicting accessibility in transmembrane regions and concluded that the best one is by Park *et al*. (2007). However, this method performs badly for non-membrane regions. Illergard *et al*. summarized that all existing state-of-the-art predictors for surface area are optimized for one of the environments and therefore perform poorly in the non-optimized environment. To address this, Illergard *et al*. developed a method that predicts the accessibility of transmembrane proteins for regions outside and inside the membrane. The method, termed MPRAP, uses a support vector machine (SVM), which includes the entire protein in the training set. MPRAP was shown to recognize the preferences for exposed sites within and outside the membrane. In parallel, Nugent and Jones (2010) developed another method that predicts lipid exposure, residue contacts, helix-helix interactions and the optimal packing arrangement of transmembrane proteins. Their method is described in chapter 1.

In summary, in the last few years there has been much progress in the ability to predict the buried/exposed state of residues in helical transmembrane proteins. Methods have been developed that combine propensity scales and sequence conservation. In addition, more recently, methods have been generated that include also structural information about contact between helices.

2.1.1 The present work

In the current work, a neural network has been used to predict residue orientation, i.e., to define which faces of transmembrane protein residues are buried and exposed. The approach taken is similar to that employed by Rost and Sander (1994) in their study on water soluble proteins.

Evolutionary information was incorporated using profiles derived from multiple sequence alignments and input into a neural network. The network was trained to determine whether a residue is buried in the core of the helix-bundle or exposed to the lipid environment surrounding the protein. Predicting residue orientation will be a key step in our method aimed at identifying homologous transmembrane proteins.

2.2 Methods

2.2.1 Dataset for the analysis

The dataset used for developing our method comprised transmembrane proteins with known topology and known 3D structure. The list of proteins was prepared from two sources. The first was the MPtopo database, provided by Stephen White's website (http://blanco.biomol.uci.edu/) (Jayasinghe *et al.*, 2001), which is a database of transmembrane proteins with experimentally validated transmembrane segments. The second resource also provided by Stephen White's website, was a list of all transmembrane proteins of known 3D structure. This list does not include information about the transmembrane segments and therefore the locations of the transmembrane helices were predicted using the program MEMSAT (Jones *et al.*, 1994). The transmembrane helices locations was used later to train the neural network (as described below).

Proteins were selected so as to produce a non-redundant list with a 30% sequence identity threshold, i.e., no pair of proteins in the final list had >30% sequence identity. Only helix-bundle proteins were included in the dataset, i.e., the porin-like proteins

were excluded. Furthermore, proteins with only one helix in the membrane were excluded from the dataset in order to improve the prediction, as explained in the results section, and proteins with low structure resolution were excluded as well. The final dataset consisted of 42 protein chains.

A control dataset of 150 water-soluble proteins with known structure (extracted from CATH, S-reps, v1.6, Orengo *et al.*, 1997) were constructed as well. Proteins were removed to produce a non-redundant list with a 30% sequence identity threshold.

2.2.2 Accessibility

The solvent accessible surface area (illustrated in Figure 9), or accessibility, of an atom is the surface area of the van der Waals envelope around each atom that is exposed to solvent; in our case the solvent under consideration is the membrane phase. The residue accessibility is the sum of the accessibilities of the atoms in that residue. The residue accessibility generally serves as an indicator of the residue's location, on the surface or in the core, i.e., exposed to the membrane or buried in the protein.

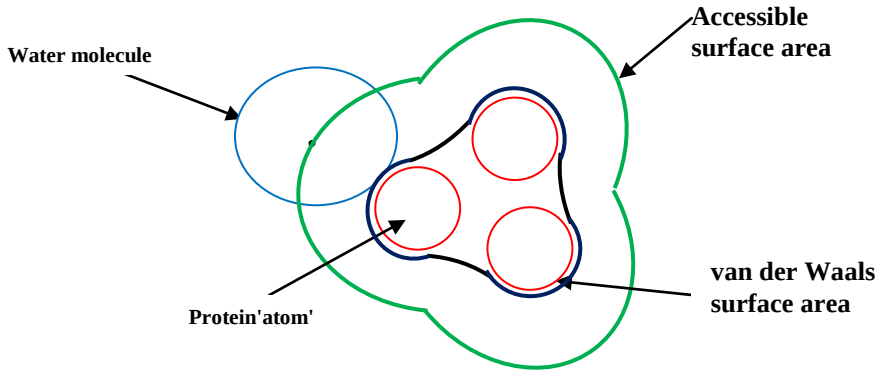


Figure 9: Accessibility estimation. The accessible surface area (green line) is at a water molecule's estimated radius beyond the van der Waals radius (red lines).

Accessibility of each residue was calculated using the DSSP program (Kabsch and Sander, 1983). The program employs the Shrake and Rupley (1973) method, that uniformly distribute a mesh of points equidistant from each atom of the molecule and uses the number of these points that are solvent accessible to determine the surface area. The points are drawn at a water molecule's estimated radius beyond the van der Waals radius. Each point is checked against the surface of neighboring atoms to determine whether they are buried or accessible. For each atom, the number of test points accessible is multiplied by the surface area value corresponding to each test point in order to calculate the accessible surface area.

The DSSP program in the current work considers the whole protein structure taken from the EBI Macromolecular structure database and searches using the Protein Quaternary Structure Form (PQS).

For comparison between amino acids of different sizes, relative accessibility (Accessibility / Maximum Accessibility) was calculated (Rost and Sander, 1994). Henceforth, in this report relative accessibility is referred to simply as accessibility.

The accessibility of each residue was first divided into binary states, i.e., buried or exposed. According to Rost and Sander (1994), when developing a prediction method, the best threshold to use for distinguishing between these two states is 16%. However, since in this study the accessibility investigated is the accessibility to lipid rather than water, it was not clear where to set the thresholds. Several thresholds were tested: 16, 20, 24, 30 and 36 percent. 30 percent was found to be the optimal threshold based on prediction quality. In addition, a three state accessibility was considered, i.e., buried, intermediate and exposed states.

2.2.3 Predicting Accessibility using Neural Networks

A system of neural networks was used in order to predict the lipid accessibility. The architecture of the neural network was based on previous work by Jones (1999).

The inputs to the network were windows of 15 consecutive residues. This window size was found to be the optimal size by Jones (1999). Additional window sizes were tested (Table 2) including a smaller window size of 7, which was found to reduce neural network performance, and windows of 11 and 19 residues, which produced similar results as the 15 residue window. The window was passed only along the sequence of the transmembrane protein helical region.

Profiles of multiple alignments were used as input. The profile was calculated using PSI-BLAST, with the following parameters: NRDB90 database with 2 iterations (-j 2), and an E-value threshold of 10^{-6} used for profile inclusion (-h 10^{-6}).

The profile matrix elements were scaled to the required 0-1 range using the standard logistic function (Jones, 1999): $\frac{1}{1+e^{-x}}$

where x is the raw profile matrix value.

The neural network output was the relative lipid accessibility, buried or exposed, of the central residue. Two different neural networks were compared. First, a network with one output, encoding for buried or exposed. Second, a network with three outputs, encoding for buried, intermediate or exposed.

A standard feed-forward neural network was used with a single hidden layer (see Appendix A) that was trained by backpropagation. The input layer comprised 315 input units, divided into 15 groups of 21 units, one group for each position in the window (overall 6410 residues used as input). 20 units represent each amino acid and the extra unit per amino acid is used to indicate if the window spans either the N or C terminus of the protein chain.

A hidden layer of 75 units was used, based on the neural network architecture described by Jones (1999). Additional sizes of neural network were tested (Table 3), including a smaller hidden layer of 30 nodes, which produced similar results to the 75 units architecture and a bigger hidden layer of 120 nodes that was found to reduce neural network performance.

The neural network system was built (Figure 10) and trained using the Neural Network toolbox for Matlab (MathWorks, version 6.5). The training algorithm used was the batch adaptive steepest descent with Momentum (traingdx), described in detail in Appendix A. Training of the network was halted when the performance of

the network on the test set began to degrade, to prevent over-fitting of the network. For benchmarking leave one out cross-validation was performed.

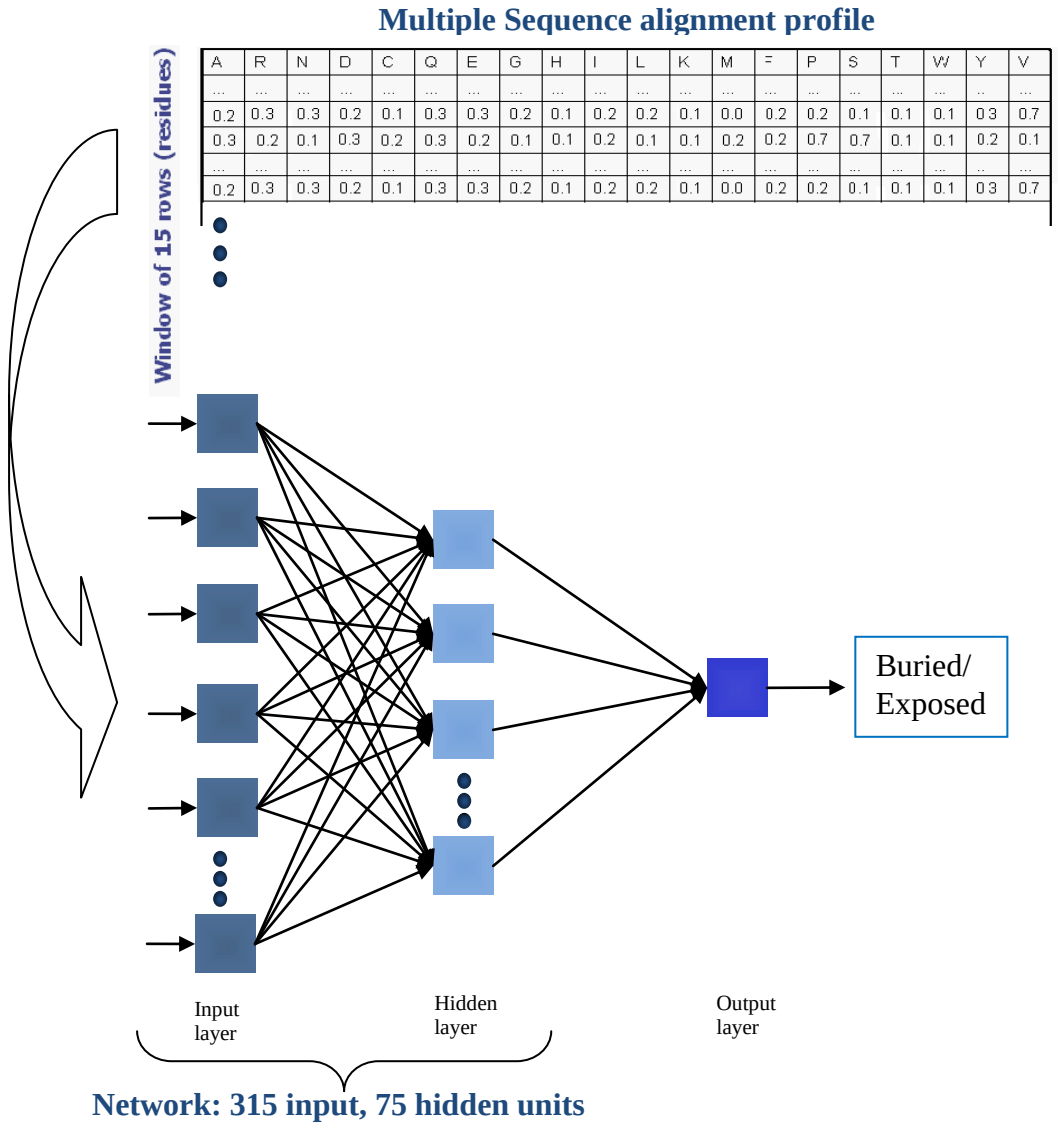


Figure 10: Neural network architecture. A standard feed-forward neural network with a single hidden layer trained by backpropagation. Profiles derived from PSI-BLAST were input into a neural network. The input layer comprised 315 input units, divided into 15 groups of 21 units. The output was the buried state of the central residue. The neural network was trained using data concerning 41 transmembrane proteins with known structure (overall 6410 residues used as input).

2.2.4 Predicting water-soluble protein accessibility

In order to test the neural network system, its ability to reproduce the results of Rost and Sander (1994) was evaluated. Rost and Sander attempted to predict the solvent accessibility of water-soluble proteins. In the present work 150 water-soluble proteins were used as the dataset with the network described above.

The network trained with water-soluble proteins was used also as a control network for transmembrane proteins accessibility prediction.

2.2.5 Assessing the accuracy of predictions

The accuracy of predictions was assessed by four scores (Baldi *et al*., 2000, Rost and Sander, 1994):

1. Percentage of correctly predicted residues.

$$Q2 = \frac{Correctly\_predicted\_residues}{Number\_of\_residues} \times 100 \quad (4)$$

2. Percentage of correctly predicted exposed residues (Sensitivity).

$$Sensitivity = \frac{Tp}{Tp + Fn} \times 100 \quad (5)$$

3. Percentage of correctly predicted buried residues (Specificity).

$$Specificity = \frac{Tn}{Tn + Fp} \times 100 \quad (6)$$

4. Matthews correlation coefficient (MCC).

$$MCC = \frac{Tp \times Tn - Fp \times Fn}{\sqrt{(Tp + Fn) \times (Tp + Fp) \times (Tn + Fp) \times (Tn + Fn)}} \quad (7)$$

Whereas, true positives (Tp) is the number of times the prediction is exposed and the

target is exposed. True negative (Tn) is defined as the number of times the prediction is buried and the target is buried. False positive (Fp) is defined as the number of times the prediction is exposed and the target is buried. False negative (Fn) is defined as the number of times the prediction is buried and the target is exposed.

A Receiver operating characteristic (ROC) curve was generated to further evaluate the accuracy of predictions (Hanley and McNeil, 1982). The ROC curve is a plot of the true positive rate (sensitivity) versus false positive rate (1 – specificity). The area under the ROC curve (AUC) is considered a good measure of the overall accuracy of the prediction method. Hanley and McNeil showed in their paper that there is a correspondence between the area under the ROC curve and Wilcoxon rank-sum statistic with a score of 50% representing random and 100% perfect prediction.

2.3 Results

2.3.1 Prediction with one state output

The architecture of the neural network for predicting the lipid accessibility was based on previous work by Jones (1999). A window is passed along the sequence of the transmembrane protein helix. According to Jones the optimal window size is 15 consecutive residues. A few additional window sizes were tested and the Matthews correlation coefficient was calculated (Table 2). In the current work the 15 residues window size was chosen for building the neural network.

Table 2: Results of predicting the buried/exposed residue state of a transmembrane protein set using different window size as input to the neural network, evaluated using Matthews correlation coefficient (MCC).

Window size	**MCC**
7	0.27
11	0.3
15	0.3
19	0.3

In addition, the hidden layer size was tested. According to Jones (1999) the optimal hidden layer size is 75 nodes. Two additional hidden layer sizes were tested: hidden layer sizes of 30 and 120 nodes. Matthews correlation coefficient for these neural network architectures are shown in Table 3. In the current work, the 75 nodes hidden layer was chosen for building the neural network.

Table 3: Results of predicting the buried/exposed residue state of a transmembrane protein set using different neural network hidden layer sizes, evaluated using Matthews correlation coefficient (MCC).

Hidden layer size (nodes number)	**MCC**
30	0.3
75	0.3
120	0.29

Table 4 shows the accuracy scores for predicting the lipid exposed/buried residues of a set of 41 chains taken from 41 transmembrane proteins. The threshold used for distinguishing between these two states was 30%. The Matthews correlation coefficient (MCC) for these predictions was 0.3. For the control test, the maximum MCC was found to be 0.1. More than 70% of residues were correctly predicted (Q2) in 19 proteins; the highest accuracy was 86% of residues correctly predicted (for 1J4NA). These data indicate that the method is able to predict accessibility of residues with very high accuracy for at least some of the proteins in the test set.

Analysis of the proteins, for which the buried/exposed state of constituent residues was predicted badly, revealed that some of these proteins are channel proteins, such as: 1k4c, 1orq, 1mxm and 1p7b. Notably, the environment in which channel proteins exist in the membrane is different than that experienced by other transmembrane proteins. Residues within channel proteins can exist in three different states in the membrane: exposed to lipid, buried from lipid and exposed to solvent (the channel itself). Therefore, it is not surprising that the buried/exposed state of residues within such more complex protein structures would be harder to predict. Indeed, analysis revealed that the buried/exposed state of residues was predicted with low accuracy for all of the channel proteins.

Table 4: Results of predicting the buried/exposed residue state of a transmembrane protein set using neural network (MCC , Q2, Sensitivity, Specificity: for definition see page 74).

PDB Code	No. of transmembrane Helices	Chain length	MCC	Q2	Sensitivity	Specificity
1j4nA	8	138	0.562	86	86	86
1occC	7	187	0.545	84	60	91
1l7vB	10	221	0.532	80	67	85
1iwgA	11	227	0.420	80	43	92
1ar1A	12	352	0.441	79	69	82
1prcM	5	139	0.432	79	43	93
1fftC	5	158	0.437	77	51	88
1nekD	3	82	0.482	76	50	92
1bgyC	8	204	0.412	76	50	88
1jb0L	2	44	0.567	75	47	100
1jgjA	6	122	0.468	75	63	82
1fx8A	8	165	0.363	75	61	79
1jb0F	2	41	0.576	73	57	100
1otsA	17	422	0.407	73	73	74
1qlaC	5	146	0.369	73	44	88
1ogvL	5	115	0.353	73	40	89
1ar1B	2	65	0.475	72	58	87
1rh5A	10	254	0.396	71	71	71
1pw4A	12	326	0.316	71	46	83
1f88A	7	215	0.281	70	43	82
1nekC	3	88	0.343	69	55	77
1e12A	7	182	0.333	69	44	85
1mslB	2	50	0.391	68	23	100
10ulA	10	255	0.316	68	61	71
1q16C	5	136	0.305	68	53	77
10hkA	13	382	0.276	67	66	67
1okcA	6	169	0.292	66	27	93
1pv7A	12	315	0.124	65	19	89
1p7bA	4	63	0.120	65	18	90
1mxmA	3	58	0.267	63	74	51
1kqfC	4	84	0.162	63	44	71
1orqC	6	120	0.170	61	52	65
1jb0A	11	218	0.135	61	18	90
1l0vC	3	55	0.078	58	21	84
1l0vD	3	64	0.038	57	18	78
1k4cC	2	48	0.068	56	46	60
1oedE	4	116	0.197	55	24	90
1pf4A	6	151	0.123	52	39	72
1rwtA	3	85	0.104	49	26	82

PDB Code	No. of transmembrane Helices	Chain length	MCC	Q2	Sensitivity	Specificity
1fftB	2	60	0.082	45	41	50
1s7bA	4	88	0.020	42	21	80
Total/Average	**6.3**	**156.3**	**0.3**	**67.9**	**46.6**	**81.8**

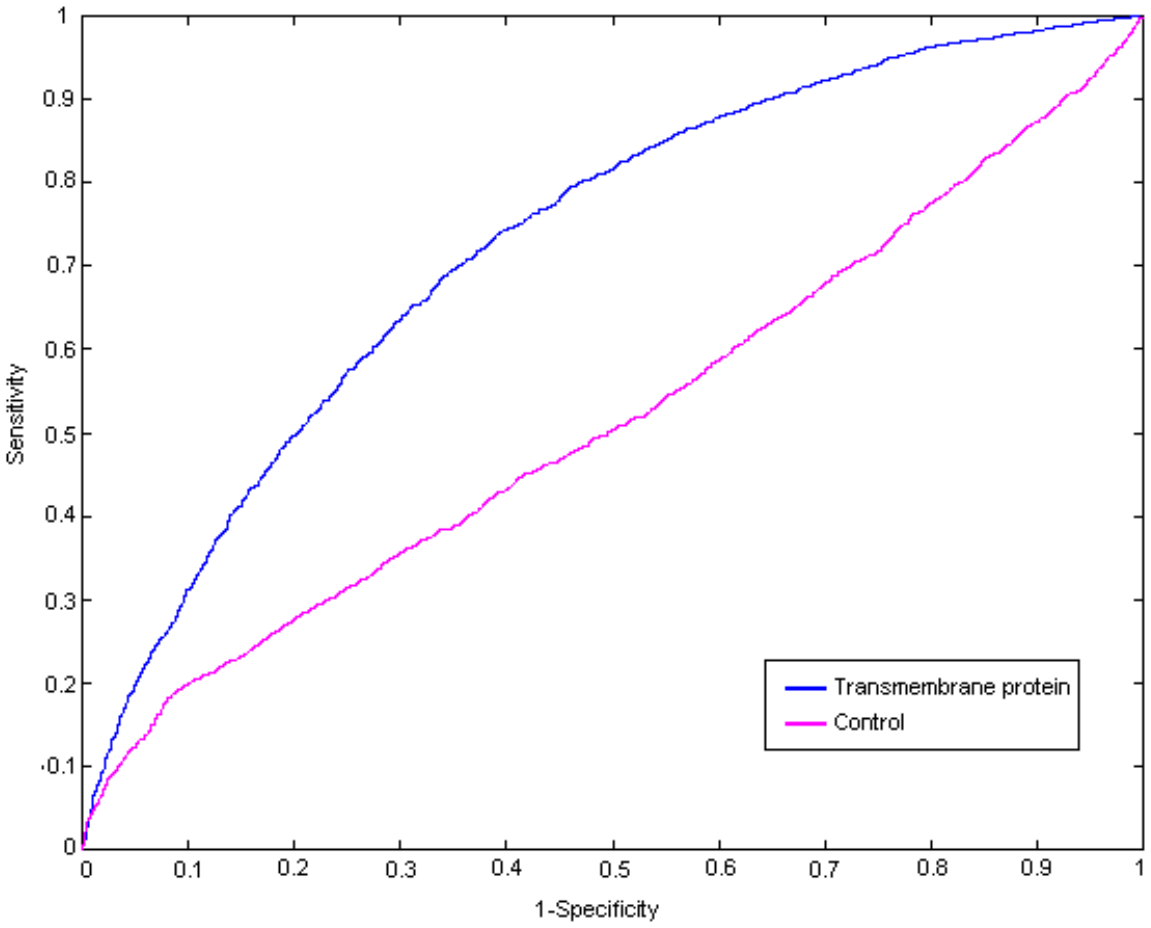


Figure 11: ROC curve for two state output network, predicting the lipid exposure (buried/exposed) for a set of 41 transmembrane proteins with known structure.

The receiver operating characteristic (ROC) curves of predicting exposed/buried residues for both the transmembrane protein and the control test (water-soluble proteins) are shown in Figure 11. The area under the ROC curve (AUC) of the prediction was calculated to be 0.73. As expected the control test shows results close to random.

As mentioned, several other accessibility thresholds were tested for distinguishing between the two states; exposed/buried to lipid. The results are presented in Table 5. Although the differences were not very significant between the thresholds tested, the threshold 30% generated the best results.

Table 5: Results of predicting the buried/exposed residue state of a transmembrane protein set using different accessibility thresholds for two state predictions (buried/exposed), evaluated using AUC and MCC (see page 75 for definition).

Accessibility Thresholds	AUC	MCC
16%	0.67	0.27
20%	0.67	0.26
24%	0.68	0.28
30%	0.73	0.3
36%	0.7	0.29

Training on a preliminary dataset, which included proteins with a single helix in the membrane resulted in a Matthews correlation coefficient score of 0.15, which is close to random. Therefore, these proteins were excluded from the final dataset as described. One explanation for this finding is that helices structured as a bundle possess particular features not exhibited by single helices.

2.3.2 Prediction with three state output

The ROC curve for a three state output network, Exposed (accessibility >30%), Intermediate (10%-30% accessibility) and Buried (accessibility<10%) is shown in Figure 12. The Matthews correlation coefficient (MCC) and area unser ROC curve

(AUC) for these predictions is shown in Table 6.

Table 6: Results of predicting the buried/exposed residue state of a transmembrane protein set for a three state output network (exposed/intermediate/buried), evaluated using MCC and AUC (see page 75 for definition).

State	AUC	MCC
Exposed	0.7	0.29
Intermediate	0.53	0.05
Buried	0.67	0.25

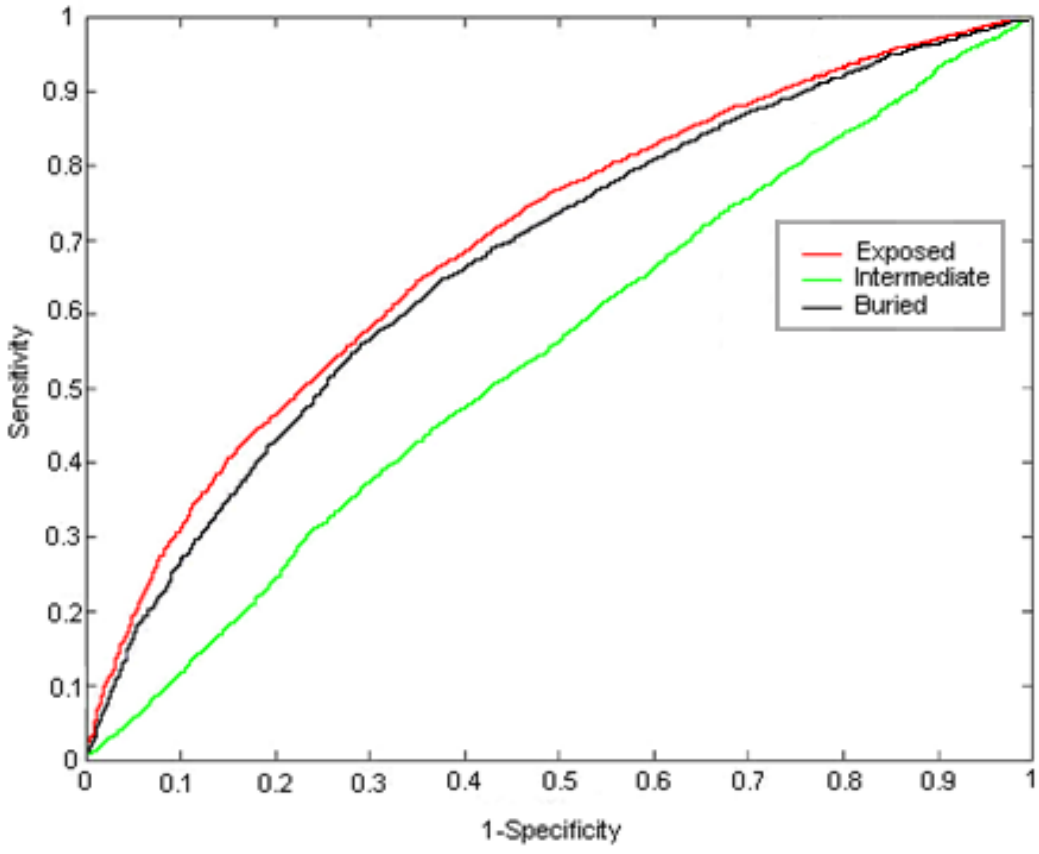


Figure 12: ROC curve for three state output network predicting the lipid exposure (buried/intermediate/exposed) of a set of 41 transmembrane proteins with known structure.

Another setting of accessibility thresholds was tested (<7%, 7%-36% , >36%) which resulted in even lower Matthews correlation coefficients. In summary, a neural

network with three state output was not able to predict the Intermediate state. We did not study any further neural networks with a three state output.

2.3.3 Comparing accessibility predictions for transmembrane versus water-soluble proteins

It was interesting to compare the accuracy of predicting lipid exposure for transmembrane proteins to the accuracy of predicting solvent accessibility for water-soluble proteins. The Matthews correlation coefficient score was 0.54 when the method described here was used to predict the solvent accessibility of water-soluble proteins, which is consistent with that reported by Rost and Sander, 1994. This score is higher than the one obtained when our method is used to predict lipid accessibility of transmembrane proteins (0.3).

2.3.4 Visualization of accessibility

The accessibility prediction can be visualized using the program RasMol (Sayle and Milner-White, 1995). Figure 13 shows visualization of three predicted chains. Only the helices are shown for each chain (i.e. without the loops). Lipid exposed residues are colored red whereas buried residues are colored blue. As seen in the table, the predicted and the observed accessibility patterns are similar. The quaternary protein structure with highlighted chain (including loops) is represented in the table as well, for better understanding of the accessibility patterns of the single chain. For example, 1j4n protein (AQP1 water channel) is built from four identical chains; the buried helices in blue, can be easily identified inside the quaternary protein structure, whereas the exposed to the lipid residues in red can be identified around the structure.

The 1l7v protein (ABC transporter) is built from two chains; therefore in the predicted chain one can see the buried area, between the chains. Similarly, the 1qla protein (Fumarate reductase flavor protein) is built from two chains in the same way as in the 1j4n protein, and it is possible to see the buried contact area between the chains although the prediction in this case was not as accurate.

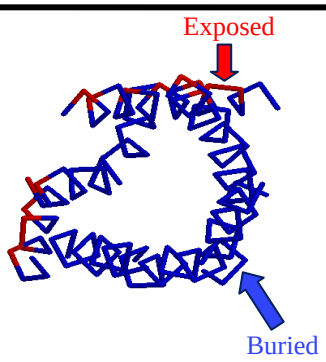
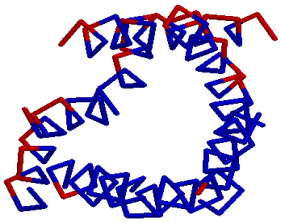
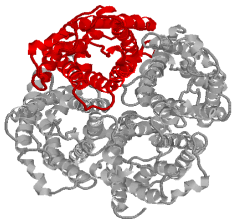
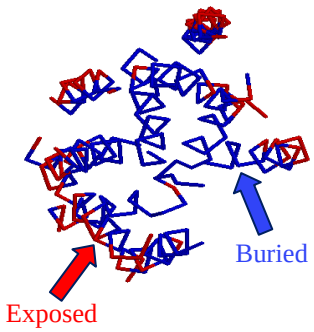
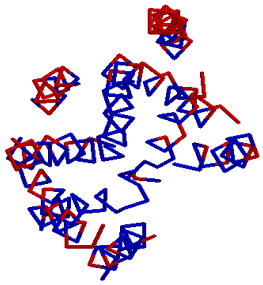
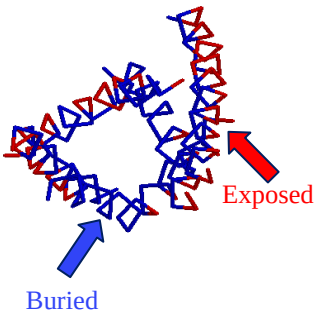
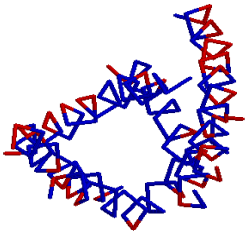
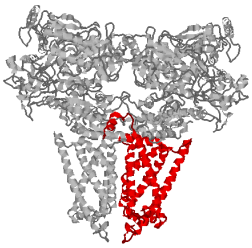
PDB code	Observed	Predicted	Protein structure
1j4nA			
1l7vB			
1qla			

Figure 13: Visual representation using RasMol of three of the predicted transmembrane chains (helices only). Lipid - exposed residues are colored red, buried residues are colored blue. The quaternary protein structure is represented, with the chosen chain colored in red.

2.4 Discussion

The method presented here uses a neural network for predicting which residues of the transmembrane helices are lipid-exposed versus buried inside the protein. When we started the current work there had been very few attempts to predict transmembrane residue orientation. None of these previous attempts achieved the high accuracy that we managed to achieve here using a neural network system. While performing the present study, others reported methods to predict residue orientation. Our results exhibit accuracy comparable to such published studies. For example, the Matthews correlation coefficient for the method developed by Nugent and Jones (2010) is 0.38, which is comparable to the 0.3 of our method.

When applying our method to water soluble proteins we obtained a higher Matthews correlation coefficient than when it was applied to transmembrane proteins. One possible explanation could be the larger size of the water-soluble protein set. As more transmembrane structures become available, it is likely that prediction efficiency will improve. Another explanation is the intrinsically more complex nature of transmembrane proteins. Water-soluble proteins have hydrophobic buried residues versus hydrophilic exposed residues, whereas, transmembrane proteins hydrophobic residues face both the core of the protein and the lipid (Stevens and Arkin, 1999, Rees and Eisenberg, 2000). Another level of complexity is that the hydrophobic residues in transmembrane proteins face two distinct environments, internally versus lipid. In this regard, it is noteworthy that the measurement of accessibility uses a water molecule as a probe (Shrake and Rupley, 1973). Perhaps water is not the correct size probe when considering accessibility to lipids.

Prediction accuracy could also be influenced by the use of only helical regions for the prediction; it could be that using the entire protein would result in better predictions. Moreover, the prediction is calculated based on single protein chains and not on the multimeric protein complex. Since transmembrane proteins could constitute multimeric complexes, predicting the accessibility of only a single chain could compromise the prediction accuracy.

Finally, the multi-helical nature of many transmembrane proteins could affect prediction accuracy. It is expected that it would be harder to predict the lipid exposure for chains containing large numbers of helices as the structure is more complex. This said, a survey of our data did not support such a premise (Table 4), as among the most accurate predictions were chains, comprising 8-12 helices, as well as a small number of helices.

In summary, the prediction method presented here was highly accurate in many cases and comparable to other prediction methods. Therefore, we incorporated this method into our overall strategy for improving homology detection, as described in the following chapters.

Chapter 3

Evaluating the performance of PSI-BLAST for transmembrane proteins

3.1 Introduction

3.1.1 Benchmarking homology detection methods

Homology detection methods aim to identify all, and only, the proteins in the database that are homologous to a query protein. In practice, the methods often designate non-homologous proteins as homologous and miss genuinely homologous proteins. The challenge of designing a detection method that identifies only true positives (TP) is illustrated in Figure 14 (Karwath and King, 2002). In the figure two distributions are shown, for homologous (true positives, TP) and non-homologous (true negatives, TN) proteins. False negatives (FN) are genuinely homologous proteins that are mistakenly predicted to be non-homologous proteins. Conversely, false positives (FP) are the non-homologous proteins that are mistakenly predicted to be homologous. Depending on the threshold (E-value) used, the method detects different proportions of TP, TN, FP and FN.

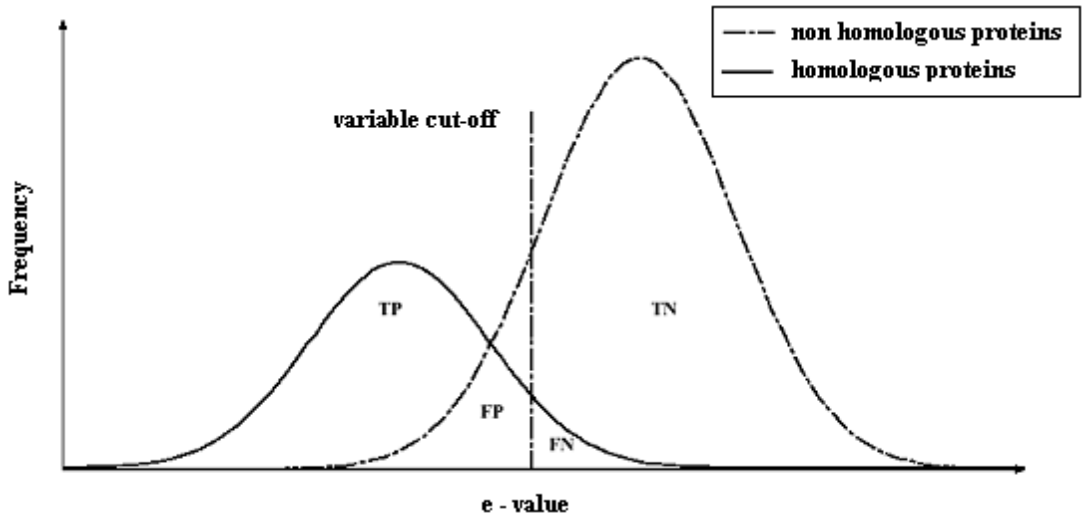


Figure 14: A graphical representation of two different distributions of a homology search (Karwath and King, 2002).

The ability to evaluate the performance of a homology detection method depends mainly on the quality of the database used. Specifically, the database should be annotated such that the true relationship between query and database proteins is known. There are a number of structural classification databases for proteins, based on analysis of protein 3D structure, which serve as a benchmark when evaluating the performance of homology detection methods. For example, the true relationships between proteins in the SCOP (Murzin *et al*., 1995) and CATH (Orengo *et al*., 1997) databases are known. Unfortunately, the number of transmembrane proteins in these databases is still low; only 381 non-redundant chains are described by Neumann *et al*. (2010) in the PDB database, and therefore such datasets cannot be employed to evaluate the performance of transmembrane protein homology detection methods. There is a database called OPM (Lomize *et al*., 2006), that includes all unique experimental structures of transmembrane proteins. When OPM was first published it contained only 126 unique 3D structures that represented 506 PDB entries (Lomize *et*

al., 2006), due to the low number of known structures. In light of the paucity of known transmembrane protein structures, a non-structural classification database, Pfam, is employed, in the current work, when evaluating the performance of homology detection method for transmembrane proteins. The proteins in the Pfam database are classified according to sequence, based on an optimized set of Hidden Markov Model (HMM) profiles for protein domain families. The proteins are classified into families, which are in turn, grouped into clans.

3.1.2 Benchmarking PSI-BLAST

PSI-BLAST (Altschul *et al*., 1997) is a widely used sequence–based homology detection method. In Chapter 1 we discussed the method in detail. Briefly, the PSI-BLAST algorithm first searches the sequence database to collect obviously homologous sequences, decided by considering E-values smaller than a chosen parameter h. These sequences are collected and aligned to generate a position specific scoring matrix (a PSSM). The PSSM is used in the next iterations to identify more homologous sequences, which are added to the PSSM if their E-value is below the cut-off. PSI-BLAST is usually run for a defined number of iterations or until no new homologous proteins are found.

A key parameter of the detection method, which can be set by the user, is the E-value cut-off, which effectively determines the level of confidence in the conclusion that the proteins under consideration are indeed homologs (h-value). Setting this parameter to a low value can lower the number of false positives but concomitantly also lower the total number of true positives. In other words, if the h-value is set too low, only closely homologous proteins are used to make the PSSM and the sequence variation is

too limited to efficiently find homologues. Conversely, if the h-value is set too high, non-homologous proteins will be incorporated into the PSSM, and the next iteration is likely to mistakenly select more non-homologous proteins. Accordingly, it is essential to fine-tune the h-parameter in order to get the optimal output.

In the following sections, previous studies of PSI-BLAST benchmarking are described, where the PSI-BLAST method h-value parameter was fine-tuned to optimize performance.

3.1.2.1 Benchmarking PSI-BLAST for water-soluble proteins

PSI-BLAST was shown to be very effective for detecting homology among water-soluble proteins (Schaffer *et al*., 2001, Lindahl and Elofsson 2000). Schaffer *et al*. (2001) used 103 queries, for which human experts had annotated the true positives in yeast. Sensitivity curves were created that plotted the number of true positive PSI-BLAST search results against the number of false positives hits when using increasing E-values for inclusion in the multiple alignment profile (h-parameter). Based on this analysis, the threshold 10^{-6} was determined to be the best threshold for attaining the highest accuracy, i.e., low number of false positives and high number of true positives. In earlier work, Park *et al*. (1998) concluded that an h-value of $5x10^{-4}$ is optimal and results in a low rate of false positives.

In the study of Shaffer *et al*. (2001), a drawback of the PSI-BLAST search method, termed ‘PSSM corruption’ was delineated. After each iteration, PSI-BLAST constructs a profile, from which a PSSM is generated. In situations when a sequence unrelated to the query sequence is included, then the next PSSM contains more

unrelated sequences and such a PSSM is termed corrupted. Schaffer *et al*. defined the corruption threshold as a PSSM containing a false positive alignment with E-value < 10^{-4} compared to the database. They suggested that, since a single corrupted sequence can affect greatly the plot and reliability of the sensitivity curve, one should consider ignoring such sequences. Schaffer *et al*. showed that it was possible to avoid corruption during PSI-BLAST searches of water-soluble proteins by setting the PSI-BLAST h-parameter to a low value: a threshold of h = 10^{-3} avoided most corrupted sequences and a threshold of h = 10^{-9} had none. However the obvious consequence of lowering the h-parameter to avoid corruption is that the number of true positives detected is smaller.

3.1.2.2 Benchmarking PSI-BLAST for transmembrane proteins

The lipid environment constrains the structural and sequence diversity of transmembrane proteins and therefore increases the likelihood of false resemblance to unrelated transmembrane proteins. Therefore homology searches for transmembrane proteins using sequence alignment alone are more prone to false positives. Indeed, Jones and Swindells (2002) remarked in their study that homology searches are most powerful for proteins with high complexity, as in these cases all 20 amino acids are exploited.

PSI-BLAST was found to be less effective for detecting homology among transmembrane proteins (Hedman *et al*., 2002). Hedman *et al*. benchmarked PSI-BLAST using only G-protein-coupled receptors (GPCRs) and concluded that there is a difference in the performance of PSI-BLAST when dealing with closely related versus distantly related GPCRs. They showed that for closely related proteins the best

performance is obtained using a very restrictive E-value (10^{-15}) whereas for distantly related proteins PSI-BLAST performs better when the E-value is less restrictive.

Forrest *et al*. (2006) also benchmarked PSI-BLAST. However, instead of using GPCRs, they built a database of transmembrane protein structures called HOMEP, which included all available transmembrane protein structures with more than four helices, and as such, comprised 36 structures. Homology searches were performed using PSI-BLAST combined with multiple sequence alignments (ClustalW) on HOMEP, and based on these data they concluded that PSI-BLAST based methods can be effective for transmembrane proteins.

Of note, the studies of Hedman *et al.* and Forrest *et al*. examining the utility of PSI-BLAST based methods for transmembrane homology searches were conducted using small datasets. An open question addressed by the present study is the ability of PSI-BLAST methods to detect transmembrane protein homology when considering larger datasets.

3.1.4 The present work

The ability of PSI-BLAST to detect homologous transmembrane proteins was investigated. The query sequences included representatives of various transmembrane protein families classified in the Pfam database and searched against a database comprised of all non-redundant Pfam protein domains recognized to be transmembrane.

As a control, a water-soluble domain set from Pfam was employed as query set against a database of the entire non-redundant Pfam protein domains.

Our benchmarking of PSI-BLAST for transmembrane proteins considered two homology levels. First, we tested the ability to detect sequences within a Pfam family. In this case, only proteins inside the query protein family were considered true positives. Next, we tested the ability to detect sequences within a Pfam clan. In this case, only proteins inside the query protein clan were considered true positives.

Our goal was to improve the capacity of the sequence alignment method, PSI-BLAST, to detect homologous transmembrane proteins. The information retrieved from this step, of benchmarking PSI-BLAST, was necessary for the development of our more complex search method described in Chapter 4, in particular, for choosing the h- parameter.

3.2 Methods

3.2.1 Databases

Protein domains were extracted from the Pfam database (Bateman *et al*., 2004, Finn *et al*., 2006). The Pfam database, as described in detail in chapter 1, contains protein domains classified using multiple alignments and profile-HMMs into families, and the families grouped into clans. Pfam consists of two parts, Pfam-A, which is curated manually and Pfam-B, an automatically generated supplement. Only Pfam-A is used in the current study. The following files describe Pfam data and are available for downloading: the "pfamseq" file contains all the protein sequences and corresponding descriptions (from SWISS-PROT and SP-TrEMBL); "Pfam-A.fasta" contains the domain sequences; "Pfam-A.full" contains the families, their description and

domains; and "Pfam-C" contains each clan in the database and its constituent families.

The sequences considered when benchmarking PSI-BLAST were the parts of the protein sequences aligned in the Pfam database. A non-redundant query set was generated that had a 50% sequence identity threshold (i.e., no pair of proteins in the final list had >50% sequence identity). In addition, a 90% sequence identity threshold database used for running the PSI-BLAST searches was generated. Redundant sequences were found using CD-HIT (Li *et al*., 2001, Li *et al*., 2002). In addition, domains were excluded from the database if their description (from "pfamseq" file) encompassed any of the following terms: uncharacterized, unidentified, unknown, predicted, hypothetical, undetermined or probable.

Two query sets and corresponding databases were created, for transmembrane proteins and water-soluble proteins, described in the next sections.

3.2.1.1 Transmembrane protein query set and database

The database of Pfam transmembrane domains, for sequence alignment, was built by selecting domains in the Pfam database version 19.0 (Pfam-A file) that had at least one of the following transmembrane protein terms in their description: transmembrane, membrane, membranous, intramembrane, transporter, pump, channel and receptor. The final transmembrane proteins database used for PSI-BLAST searches contained 909,822 protein domains.

The query set (targets to be tested) from the Pfam database was selected as follows. Initially, all clans with transmembrane terms in their description (as described above) were listed (using the Pfam-C file). Then the families within each clan and the

domains they contain were listed. Domain sequences were extracted from the "Pfam-A.fasta" file, which contains each domain name and the family with which it is associated. Furthermore, only domains with more than one transmembrane helix according to TMHMM (Krogh *et al*., 2001) were retained. Finally, the domain query set was chosen randomly from this list of domain sequences. The final query set included 112 randomly chosen proteins, from 29 different clans.

3.2.1.2 Water-soluble protein query set and database

To create the water-soluble protein database, all proteins with the transmembrane proteins terms (listed above) in their Swiss-Prot description were removed. The final water-soluble protein database used for PSI-BLAST searches contained 3,912,930 protein domains.

The set of queries were chosen randomly. Domain sequences were extracted from the "Pfam-A.fasta" file as described. The final water-soluble query set included 71 domains.

3.2.2 Sequence alignment using PSI-BLAST

Sequence alignment searches were performed using PSI-BLAST (Altschul *et al*.,1997) to identify all the protein domains homologous to a given query in the corresponding test database.

PSI-BLAST was performed to detect all possible homologous protein domains using various E-values (-h parameter: 10^{-3}, 10^{-6}, 10^{-8}, 10^{-15}), with the parameter that determines the maximum number of aligned proteins (-v 3000) set to 3,000. Up to 5

iterations were allowed (-j 5). Remaining PSI-BLAST parameters were left at default values. PSI-BLAST results with an E-value smaller than 1 were listed and analyzed (-e 1).

3.2.2.1 Running PSI-BLAST with NRDB90 database before Pfam database

Schaffer *et al*. (2001) claimed that PSI-BLAST is more sensitive to distant relationships when score matrices are created from larger and diverse sets of related sequences. In other words, they recommend searching a comprehensive sequence database for a few iterations, saving the resulting position-specific matrix (PSSM) as a checkpoint, and then restarting PSI-BLAST using that checkpoint matrix to search the narrower database of interest. In the present work, we compared this protocol, whereby PSI-BLAST is run for 4 iterations using the NRDB90 database and then restarted for one iteration using the Pfam database, with running PSI-BLAST for 5 iterations with our constructed Pfam database.

It was not found to improve detection of homologous proteins when PSI-BLAST was run using the NRDB90 database before the Pfam A-derived database.

3.2.3 Assessment of homology detection

A PSI-BLAST result file (list of hits) was generated for each query domain and each result was checked and defined as true positive or false positive. The performance was evaluated by generating sensitivity curves in which true positives are plotted against false positives for each h-parameter: 10^{-3}, 10^{-6}, 10^{-8} and 10^{-15}. In addition, all PSI-

BLAST results were collected in a list that was ordered according to E-value. Generally, it is desirable for more true positives to appear before a given number of false positives, with the number of false positives as low as possible.

3.3. Results

3.3.1 Evaluating the performance of PSI-BLAST on water-soluble proteins

The results presented in the next sections will be divided according to the homology level tested.

3.3.1.1 Evaluating PSI-BLAST at the Pfam family level

Initially, PSI-BLAST was benchmarked using Pfam A-derived water-soluble proteins as a control for benchmarking PSI-BLAST using Pfam A-derived transmembrane proteins.

In this experiment the family level was considered, i.e, true positives are proteins in the same Pfam family as the query. Sensitivity curves resulting from PSI-BLAST run using the water-soluble protein query set and corresponding Pfam database for 5 iterations, at four different settings of threshold parameter (h-parameter) are shown in Figure 15.

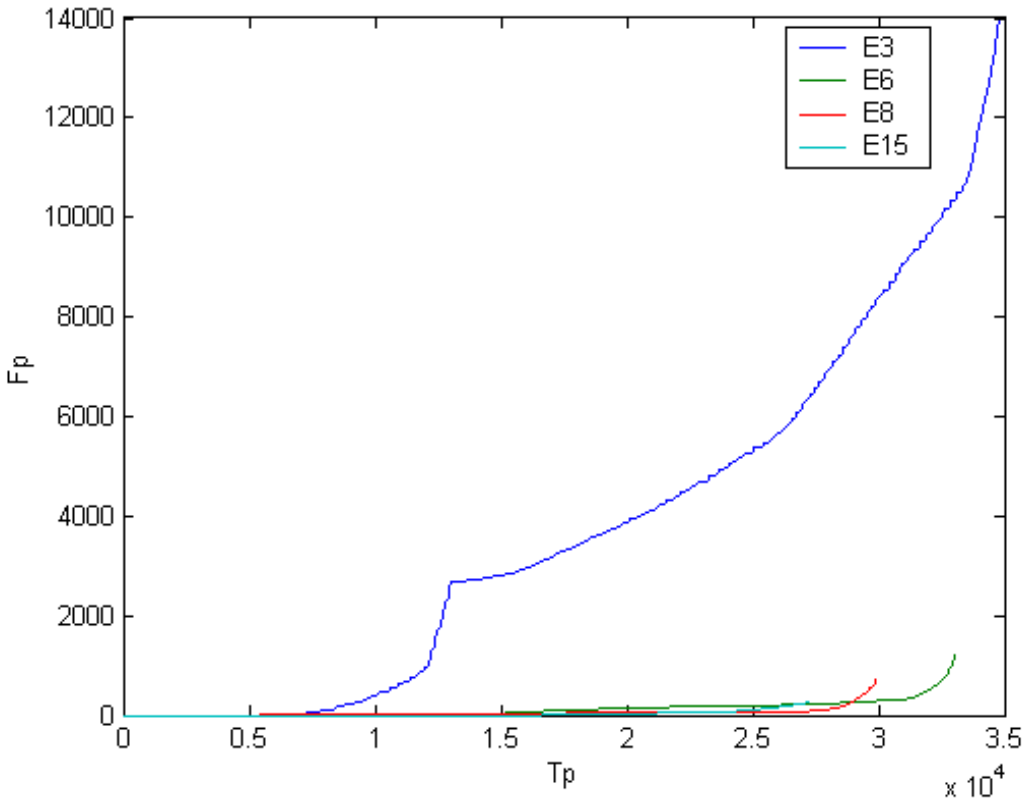


Figure 15: Sensitivity curves for homology searches performed using the Pfam water-soluble test database and query set with four settings of the threshold parameter (h-parameter): 10^{-3} (blue), 10^{-6} (green), 10^{-8} (red), 10^{-15} (light blue). Pfam family homology level.

h-parameters of 10^{-6}, 10^{-8}, 10^{-15} resulted in similar curves, although the overall number of true positives was smaller the smaller the h-value. The higher h-parameter 10^{-3} resulted in a dramatically increased number of false positives and fewer true positives.

One explanation for this behavior is the phenomenon of PSSM 'corruption', described above and characterized by Schaffer *et al*. (2001). To examine this premise, we scored the number of corrupted queries at each h-value (Table 7), using Schaffer *et al*. definition of corruption (PSSM containing a false positive alignment with E-value < 10^{-4}). However this approach to scoring PSSM corruption did not reveal significantly increased corruption at h = 10^{-3}.

Table 7: The number of corrupted queries for each h-value for the water-soluble test database, taken from Pfam (family homology level).

h-value used for running PSI-BLAST	Corrupted queries
$h = 10^{-3}$	2
$h = 10^{-6}$	1
$h = 10^{-8}$	1
$h = 10^{-15}$	0

A different way of addressing PSSM corruption is to count the number of false positives that have a smaller E-value than the h–parameter after the first iteration, second iteration and so on. Such false positives would be used to build the PSSM for the next iteration and cause the number of false positives to increase. First iteration PSI-BLAST results were found to have very small numbers of false positives. The number of false positives started to rise only after 2 iterations.

A diagram of the E-value distribution of the PSI-BLAST false positive hits after running two iterations using the h-parameter of 10^{-3} shows that the number of false positives begins to rise above an E-value of 10^{-4} (Figure 16). This could explain the large number of false positives observed after five iterations when using an h-parameter of 10^{-3}.

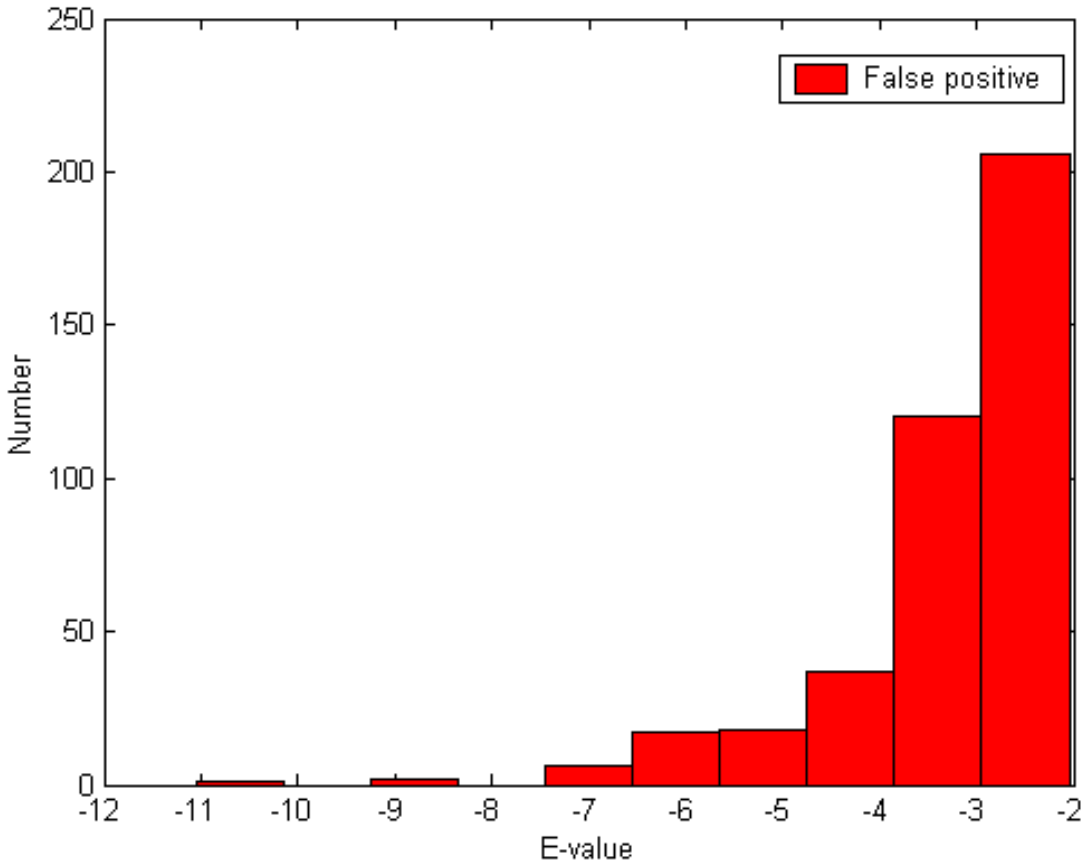


Figure 16: E-value distribution of the PSI-BLAST results found to be a false positive in the second iteration when using h-parameter of 10^{-3} for water-soluble test database, taken from Pfam (family homology level).

Conversely, a corresponding diagram of the E-value distribution of the PSI-BLAST false positive hits after running two iterations using the h-parameter of 10^{-6} shows that the number of false positives is very low under an E-value of 10^{-6} (Figure 17).

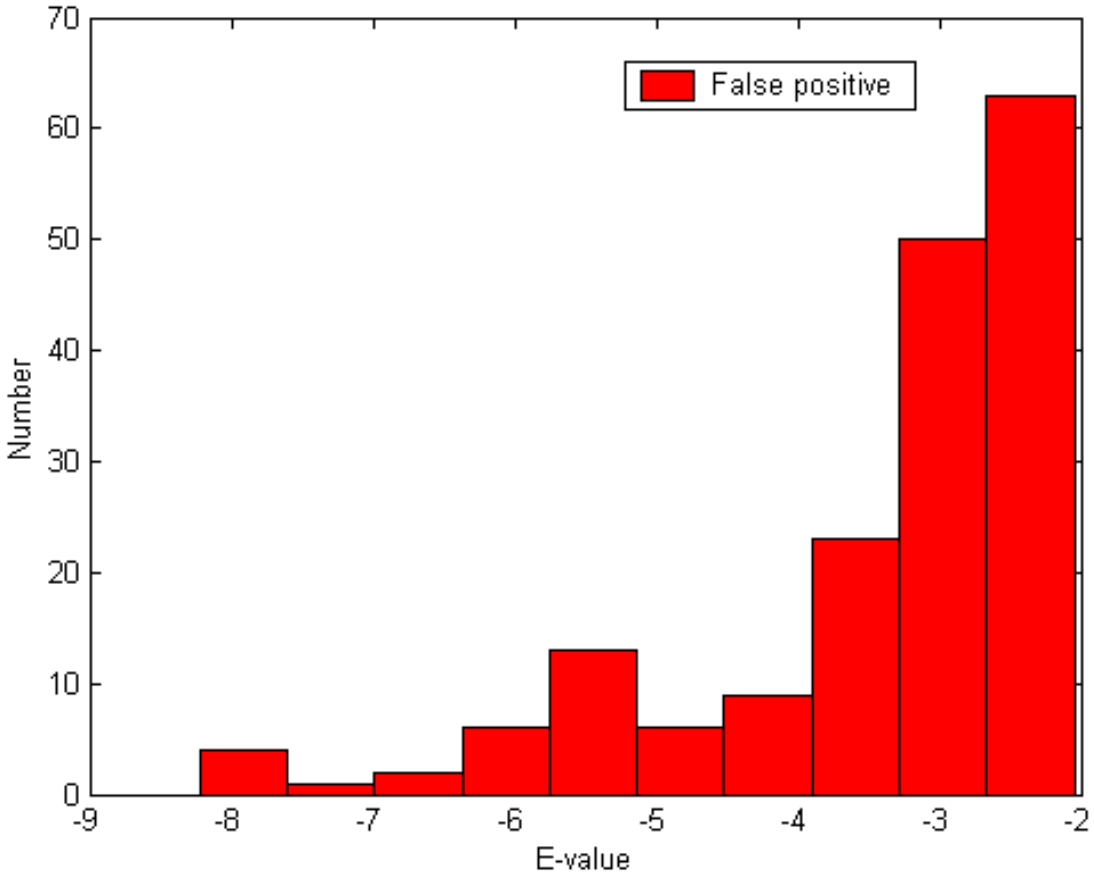


Figure 17: E-value distribution of the PSI-BLAST results found to be false positives in second iteration when using h-parameter of 10^{-6} for water-soluble test database, taken from Pfam (family homology level).

3.3.1.2 Evaluating PSI-BLAST at Pfam clan level

In this experiment the clan level was considered, i.e., true positives are proteins in the same Pfam clan as the query. Sensitivity curves resulting from PSI-BLAST run using the water-soluble protein query set and corresponding Pfam database for 5 iterations, at four different settings of threshold parameter (h-parameter) are shown in Figure 18.

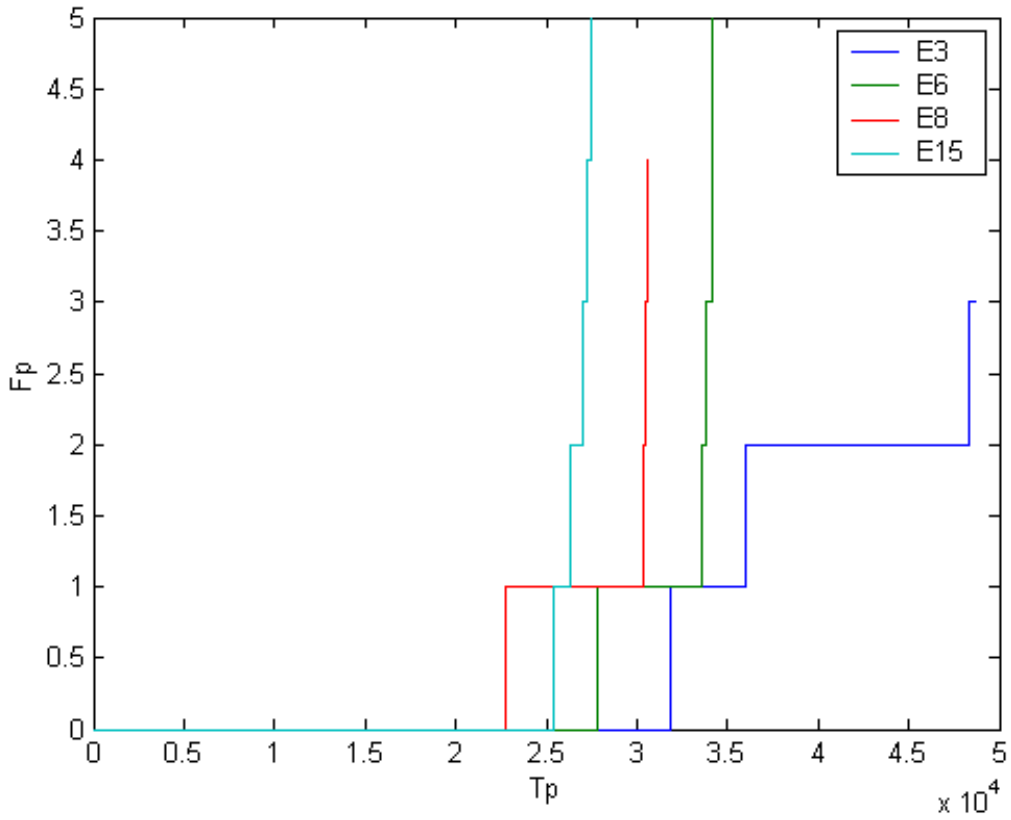


Figure 18: Sensitivity curves for homology searches performed using the Pfam water-soluble test database and query set with four settings of the threshold parameter (h-parameter): 10^{-3} (blue), 10^{-6} (green), 10^{-8} (red), 10^{-15} (light blue). Pfam clan homology level.

The graph shows that PSI-BLAST performed with high accuracy. The number of false positives was low (maximum 5 false positives) while the number of true positives was high for all h-values. Using an h-value of 10^{-3} the number of true positives was higher with fewest false positives. It is hard to compare between the performances of PSI-BLAST with the tested h-parameters due to the fact that there are very few false positives in all cases.

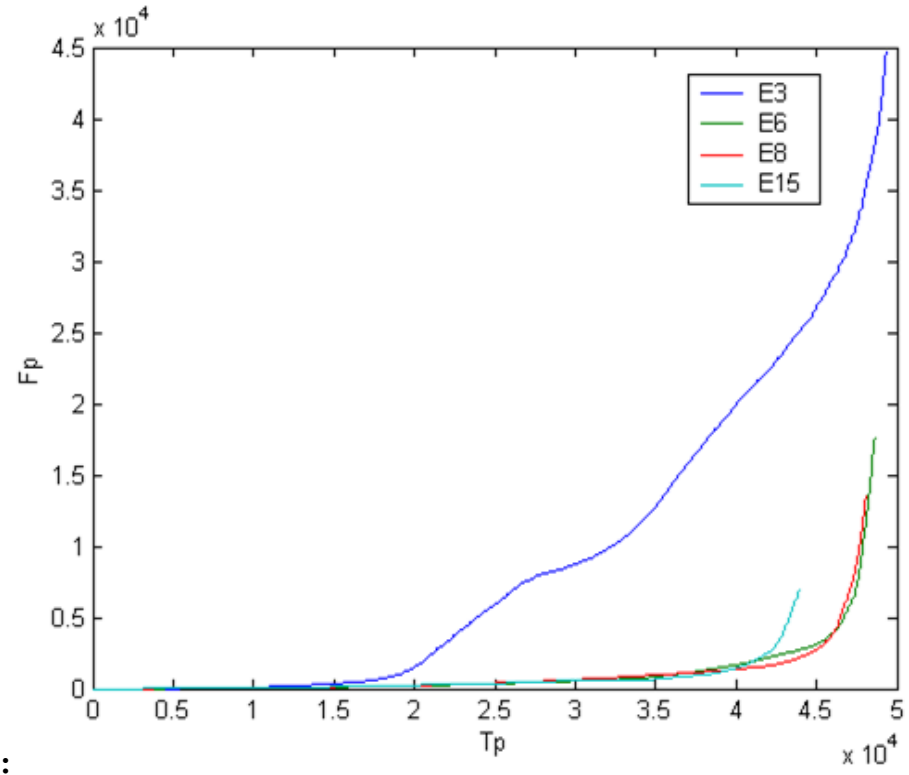
3.3.2 Evaluating the performance of PSI-BLAST on transmembrane proteins

3.3.2.1 Evaluating PSI-BLAST at the Pfam family level

In this experiment, the family level was considered, i.e, true positives are proteins in the same Pfam family as the query. Sensitivity curves resulting from PSI-BLAST run using the transmembrane protein query set and corresponding Pfam database for 5 iterations, at four different settings of threshold parameter (h-parameter: 10^{-3}, 10^{-6}, 10^{-8}, 10^{-15}) are shown in Figure 19 (A and B).

The sensitivity curves for detecting transmembrane proteins behave differently as compared to the sensitivity curves for detecting water-soluble proteins. Notably, for transmembrane proteins, PSI-BLAST run using a bigger h-parameter (10^{-3}) results in the lowest number of false positives. This pattern changes when the number of true positives exceeds 8000, from which point on the h-parameter of 10^{-3} resulted in the highest false positive versus true positive rate. In the case of the sensitivity curves for detecting water-soluble proteins, all 4 h-parameter settings result in almost zero false positives for up to 7000 true positives, and only at this point the number of false positives begin to rise at the h-parameter setting of 10^{-3}.

A:



B:

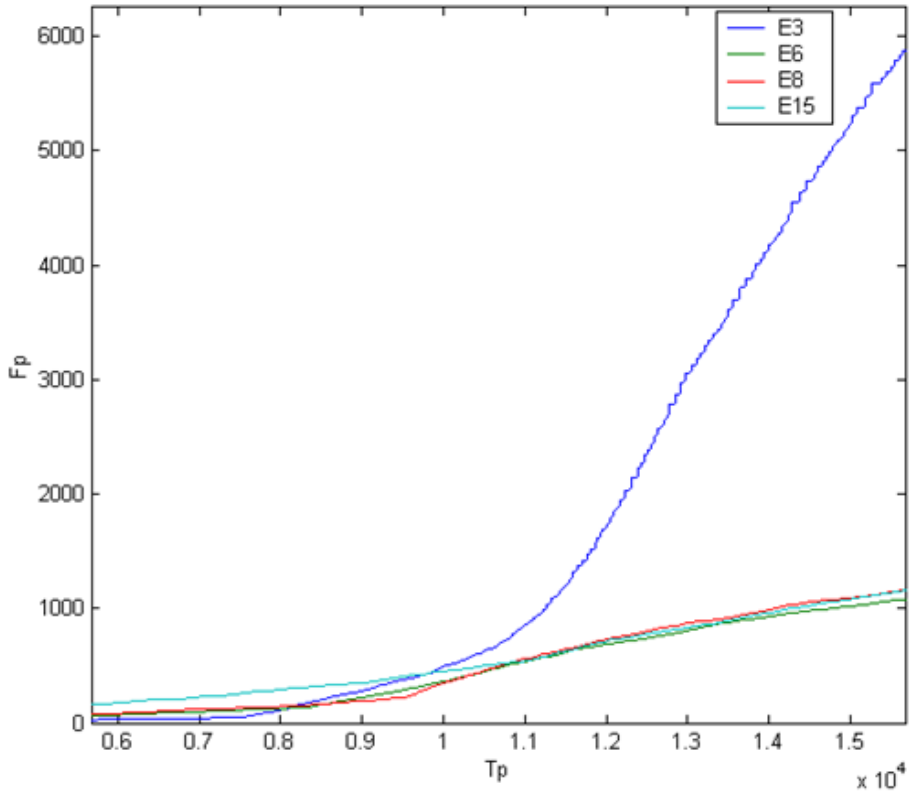


Figure 19 : A: Sensitivity curves for homology searches performed using the Pfam transmembrane test database and query set with four settings of h-parameter: 10^{-3} (blue), 10^{-6} (green), 10^{-8} (red), 10^{-15} (light blue). Pfam family homology level. **B:** Focus on true positives under 1.4×10^4

As observed when detecting water-soluble proteins using the h-parameter setting of 10^{-3}, the sensitivity curve for detecting transmembrane proteins using the h-parameter of 10^{-3} indicates PSSM corruption. Each transmembrane query was scored for corruption at the different h-values (see Table 8). In line with the observed sensitivity curves, the greatest number of corrupted queries were associated with $h = 10^{-3}$.

Table 8: The number of corrupted queries for each h-value for transmembrane test database, taken from Pfam (family homology level).

h-value used for running PSI-BLAST	**Corrupted**
$h = 10^{-3}$	40
$h = 10^{-6}$	27
$h = 10^{-8}$	26
$h = 10^{-15}$	24

3.3.2.2 Evaluating PSI-BLAST at the Pfam clan level

In this experiment the clan level was considered, i.e. true positives are proteins in the same Pfam clan as the query. Sensitivity curves resulting from PSI-BLAST run using the transmembrane protein query set and corresponding Pfam database for 5 iterations, at four different settings of threshold parameter (h-parameter: 10^{-3}, 10^{-6}, 10^{-8}, 10^{-15}) are shown in Figure 20.

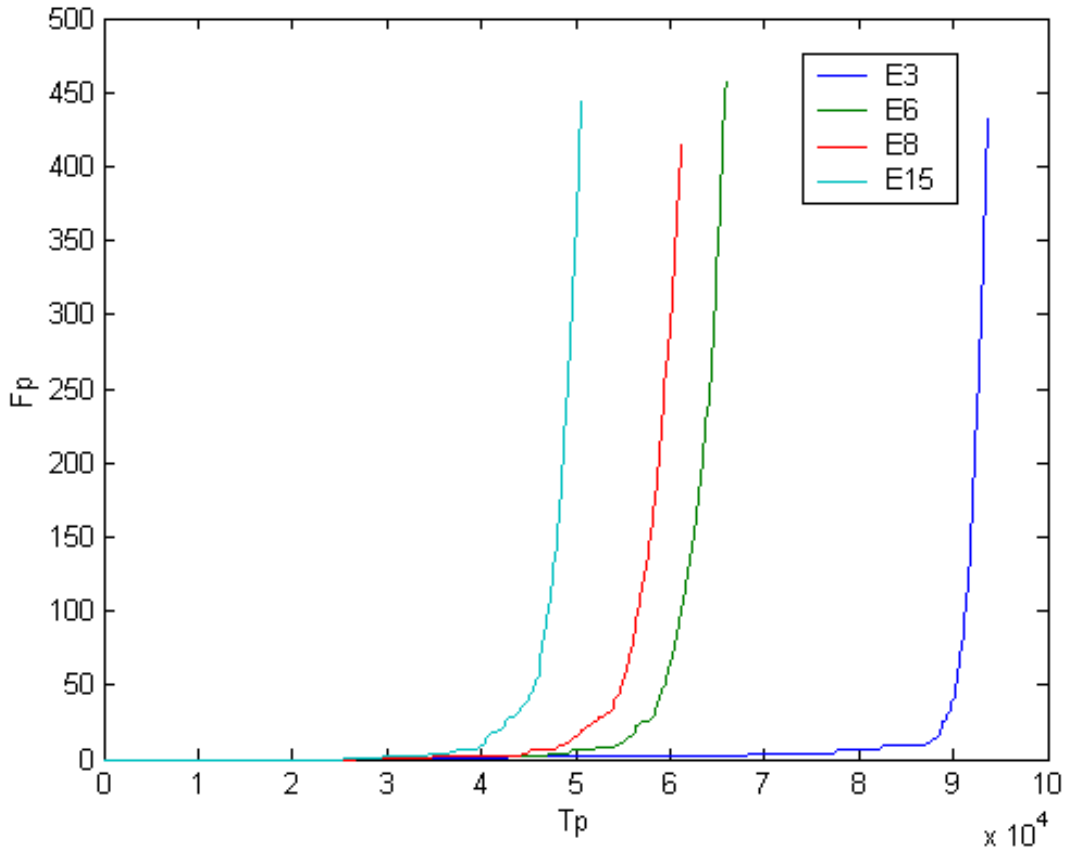


Figure 20: Sensitivity curves for homology searches performed using the Pfam transmembrane test database and query set with four settings of h-parameter: 10^{-3} (blue), 10^{-6} (green), 10^{-8} (red), 10^{-15} (light blue). Pfam clan homology level.

For detecting transmembrane homology at the Pfam clan level, the sensitivity curves show that the higher the h-parameter, the fewer false positives versus true positives.

As for water-soluble proteins, PSI-BLAST performs with greatest accuracy at the Pfam clan homology level. However, unlike as observed for water-soluble proteins, the alignment of transmembrane proteins generates a greater number of false positives.

3.3.3 Comparing the effectiveness of PSI-BLAST for transmembrane versus water-soluble proteins

In order to compare the effectiveness of PSI-BLAST for transmembrane versus water-soluble proteins, a graph of the log of the E-value against the false positive ratio (calculated by $\frac{False\_positive\_number}{Total\_results}$) was drawn for the transmembrane protein set and for the water-soluble proteins set at the Pfam family level (Figure 21).

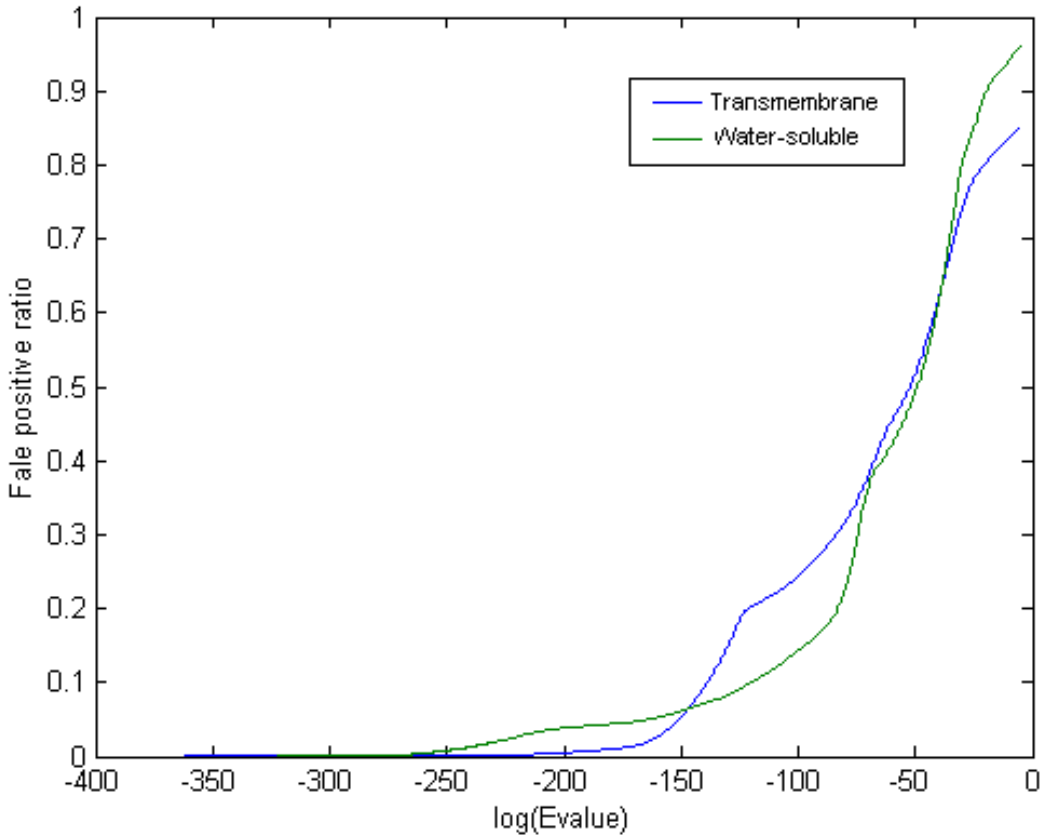


Figure 21: The false positive ratio versus the natural log of E-value of the PSI-BLAST results for transmembrane proteins (blue) and water-soluble proteins (green).

The results shown are for running PSI-BLAST with the h-parameter of 10^{-6}, which appears to result in the best false positive ratio. The false positive rate starts rising above zero at a lower E-value for transmembrane proteins. More generally, the graphs are hard to compare as they cross over in several places.

Since we aimed to use our method (presented in chapter 4) for homology detection at

the level of protein families, which is a more restricted level than the clan homology level, we conducted this comparison at the family homology level.

3.4 Discussion

In this chapter we present an evaluation of the performance of PSI-BLAST both for water-soluble proteins and for transmembrane proteins using Pfam as the source database.

3.4.1 Benchmarking PSI-BLAST for water-soluble proteins

For water-soluble proteins at the Pfam family level**,** we found that h-parameters of 10^{-6}, 10^{-8} and 10^{-15} resulted in similar sensitivity curves; a false positive rate that stays low with a high true positive rate. The lower the h-parameter, the better the sequence alignment performed but the overall number of true positives was smaller. This finding agrees with previous benchmarking studies and supports the idea that a very low h-value is restrictive.

In line with the study of Schaffer *et al.* (2001), we observed a very large number of false positives versus true positives when PSI-BLAST was run using an h-parameter of 10^{-3}, the phenomenon being termed corruption. However, in our study the number of "corrupted" queries, as defined by Schaffer *et al.*, was low. Nevertheless, our result could be explained by our analysis of the distribution of false positives according to E-value when PSI-BLAST was run for two iterations at 10^{-3} versus 10^{-6} h-parameter settings. A great number of false positives were observed at E-values greater than 10^{-4} when running PSI-BLAST with the 10^{-3} h-parameter than are observed at E-values

greater than 10^{-6} when running PSI-BLAST with the 10^{-6} h-parameter.

The reasons why false positives with low E-values were observed in the second iteration of running PSI-BLAST on water-soluble protein is not obvious. It could be that the query proteins chosen (randomly) are relatively similar also to different families and therefore a restrictive h-parameter (smaller than 10^{-3}) is required for accurate homology detection. This explanation is in agreement with a report by Finn *et al*. (2006), the creators of Pfam, describing difficulties in classifying some proteins into Pfam families. Finn *et al*. describe that building new Pfam families and/or revisiting existing families often highlights two confounding issues. (1) Many Pfam families are related and have artificially high thresholds to stop them from overlapping. Thus, two proteins can have evolved from a common ancestor but not be classified in the same family. For example, Globins are haem-containing proteins involved in binding and/or transporting oxygen that share the same folding pattern and are considered to have evolved from a common ancestor. There are two Pfam families containing Globins (PF00042, PF01152) and the separation between these families is not clear. (2) For some large, divergent families a single HMM that detects all family members could not be built.

A closer look at the PSI-BLAST results after the second iteration with h-parameter 10^{-3} supports the premise that some of the false positives were proteins that are homologous to more than one family:

- 50 false positives aligned wrongly to Siderophore-interacting protein (A0QF87) in family PF08021 (Siderophore-interacting FAD-binding domain 9) with low E- value ($<10^{-3}$).

For example:

Both the PSI-BLAST results Q11TL8 and A0JZX0 are in a different family of FAD binding domain 6 (PF00970).

- 23 false positives aligned wrongly to protein A4R870 from the family PF08022 that is FAD-binding domain 8, as well.

  For example:

  The PSI-BLAST result B1I1C3 belongs to the family PF02900 that is a Catalytic LigB subunit of aromatic ring-opening dioxygenase

- 28 false positives aligned wrongly to protein A7UW98 from the family PF08022 mentioned before.

  For example:

  The PSI-BLAST result, A5WDA3 belongs to a different FAD-binding domain (6) (PF00970), Oxidoreductase FAD-binding domain.

All of these false positives belong to the same clan CL0076, raising the possibility that proteins in these families are very similar.

The challenges of classifying proteins based on sequence alone only serve to highlight the need to develop computational approaches to classification that incorporate structural information.

For water-soluble proteins at the Pfam clan level**,** PSI-BLAST run using an h-value of 10^{-3} performed the best; although with all h-parameters tested the number of false positive was very low.

3.4.2 Benchmarking PSI-BLAST for transmembrane proteins

For transmembrane proteins at the Pfam family level, PSI-BLAST performed best for very closely related proteins when using an h-parameter of 10^{-3} and for more distantly related proteins when using an h-parameter of 10^{-6}. These findings are not in agreement with the results of Hedman *et al.* (2002). They concluded that PSI-BLAST performs best with an E-value of 10^{-15} for very closely related proteins but for distantly related proteins it performs better when using a less restrictive E-value (10^{-3}). The key difference between the present study and Hedman's is the query set and corresponding database. Hedman *et al*. were studying GPCR proteins only whereas in our study we benchmarked various transmembrane families.

3.4.3 Comparing the effectiveness of PSI-BLAST for transmembrane versus water-soluble proteins

Since the lipid environment constrains the structural and sequence diversity of transmembrane proteins, it was expected that homology searches, such as PSI-BLAST would be less effective for transmembrane proteins than for water-soluble proteins. Nevertheless, our results did not accord with this expectation. Indeed, PSI-BLAST performed similarly on water-soluble proteins and on transmembrane proteins when tested using Pfam database families. This finding contrasts with that of Hedman *et al*. (2002), who reported that PSI-BLAST performs better on water-soluble proteins. It could be that the way Pfam families are built explains the different findings or the fact that Hedman et al. conducted their research using GPCRDB only. Another possibility

is that the more simple classification of transmembrane proteins into fewer families relative to water-soluble proteins (Oberai *et al*., 2009) compensates for the low complexity in transmembrane protein sequence.

3.4.4 Choosing the best PSI-BLAST h-parameter

Based on our data, 10^{-6} is the best h-parameter when using PSI-BLAST to detect homologous water-soluble proteins at the Pfam family level, as this parameter results in a low number of false positives without limiting severely the number of true positives. This conclusion corroborates earlier findings of Schaffer *et al*. (2001) and Park *et al*. (1998), who each used a different database source for their studies.

Similarly, 10^{-6} is the best h-parameter when using PSI-BLAST to detect homologous transmembrane proteins at the Pfam family level, as this parameter results in a low number of false positives without limiting severely the number of true positives.

When using PSI-BLAST to detect homology at the Pfam clan level, an h-parameter of 10^{-3} is optimal for both water-soluble and transmembrane proteins.

3.4.5 Conclusions

The purpose of benchmarking PSI-BLAST was to determine the best h-parameter when detecting homology among transmembrane proteins. We wanted to benchmark PSI-BLAST using the exact protein set that will be used when developing our new homology search method.

We found that PSI-BLAST run with the h-parameter of 10^{-6} is the best option for

getting a minimal number of false positives with the highest number of true positives when considering the Pfam family homology level. The goal of our more complex detection method, described in the next chapter, was to further decrease this false positive number.

Chapter 4

Integrating sequence similarity and structural information to identify homologous transmembrane proteins

4.1 Introduction

One way of annotating an unknown protein and learning about its function is to search for already characterized homologous proteins. In the past decade this approach has been applied successfully to identify globular proteins but has been less effective for transmembrane proteins. Various transmembrane protein homology detection studies that have attempted to address this problem are summarized below.

4.1.1 Methods based on sequence alignment

The amino acid composition and conservation patterns of transmembrane and water soluble regions differ (Cserzo *et al.*, 1997). This reflects an obvious spatial difference, namely that transmembrane proteins are present simultaneously in two different physicochemical environments whereas soluble proteins are present in only one.

Accordingly, the amino acid composition of the transmembrane regions must be predominantly hydrophobic to be stable within the lipid environment of the membrane. Conversely, surface residues, particularly in soluble proteins are exposed to water and tend to be hydrophilic. More specifically, it has been noted that transmembrane helices display an alternating pattern of conserved and non-conserved amino acids, with the conserved amino acids in the core of the protein structure and the non-conserved hydrophobic amino acids facing the lipids (Donnelly *et al*., 1993). The dissimilar amino acid composition of transmembrane versus soluble proteins has been explored as a signature for sequence alignment methods. Indeed, several groups reasoned that homology detection for transmembrane proteins based on sequence alignment could be improved if amino acid substitution matrices specifically designed for transmembrane proteins were introduced into the search protocol, e.g., the JTT matrix (Jones *et al*., 1994), the PHAT matrix (Ng *et al*., 2000) and the SLIM matrix (Muller *et al*., 2001). Indeed, these groups found that homology searches performed using such matrices proved more effective than searches that employed regular amino acid substitution matrices.

Subsequently, the STAM method (Shafrir and Guy, 2004) was developed that improved alignment accuracy further by combining different subtitution matrics. However, Forrest *et al*. found that using a bipartite scheme (based on BLOSUM62 and PHAT) does not significantly improve transmembrane protein sequence alignment (Forrest *et al*., 2006). In light of this finding, Forrest *et al.* proposed that previous reported improvements in sequence alignment due to acid substitution matrices specifically designed for transmembrane proteins could be attributable to the separation and independent alignment of transmembrane and non-transmembrane

regions and to differences in gap penalties, rather than to the choice of substitution matrix.

Pirovano *et al*. (2008), developed a new method for aligning transmembrane proteins (Praline TM) that not only employs transmembrane specific substitution matrices (PHAT) but also incorporates a higher gap penalty setting, different from the typical one used when searching for homologous globular proteins. Higher gap penalty for the transmembrane regions (15-18) and using PHAT matrix for the transmembrane regions yield better performance. Nevertheless, the effect of the gap penalty on the performance was minor.

In summary, the utility of substitution matrices specifically designed for transmembrane proteins remains debatable. In light of this uncertainty concerning the effectiveness of transmembrane protein specific substitution matrices, we chose in the current work not to use such matrices when conducting sequence alignment for transmembrane proteins but to use regular BLOSUM62.

4.1.2 Methods based on loop lengths

Another feature of transmembrane proteins is the loops between transmembrane helices, which are less conserved than the transmembrane regions (Forrest *et al*., 2006). These loops range in size and exhibit structural flexibility and variability. Thus, the patterns of amino acid insertion/deletion are different in transmembrane versus globular proteins, potentially confounding homology searches for transmembrane proteins based on multiple sequence alignments.

To address this issue, Arai *et al*. (2004) devised a search protocol that incorporates information about loop lengths and performed modified searches using 87 complete prokaryotic genome sequences. Briefly, in this method transmembrane protein function is classified on a proteomic scale by applying a single-linkage clustering method based on sequence similarity and predicted topological similarity, the latter calculated by comparing the lengths of loop regions between helices. Notably, an assumption underlying this approach is that members of a given family possess the same number of transmembrane helices. Proteins are initially divided into groups according to the number of transmembrane helices and only then a "loop score" calculated, which relates to the loop lengths exhibited by each pair of compared proteins. Arai *et al*. reported that this clustering approach raised the rate of transmembrane proteins classified functionally and identified from 24.3% to 60.8%.

Similarly, Sugiyama *et al*. (2003) developed a method for classifying transmembrane proteins based on the number of transmembrane segments, the loop length and the N-terminus location. In this method, the length of each loop is expressed as '1' or '0' depending on whether it is longer or shorter than the threshold length defined for each loop. Next, for each functional group the average of binary loop length is calculated. Using these averages, a binary topology pattern (BTP) is determined for each transmembrane functional group. After testing 37 functional transmembrane protein groups, Sugiyama *et al*. reported that the BTPs are very accurate at identifying the individual functions.

Wistrand *et al*. (2006) designed a method (GPCRHMM) to identify new G Protein Coupled Receptors (GPCRs) based on a Hidden Markov Model. A set of GPCRs were

analyzed to determine distinct loop length patterns and differences in amino acid composition between cytosolic loops, extracellular loops, and membrane regions. The hidden Markov model, GPCRHMM, was designed to fit the observed parameters. When applied to search for novel GPCR superfamily members across five proteomes, GPCRHMM detected 120 sequences that lacked annotation and, as such are novel putative GPCRs.

4.1.3 Methods based on hydropathy profiles

The structure of transmembrane proteins is reflected in the hydropathy profile of the amino acid sequence. Accordingly, the hydropathy profile is often better conserved than the underlying sequence and can be used as an additional tool when searching for homologous transmembrane proteins. Indeed, Lolkema and Slotboom demonstrated that two transmembrane proteins with only marginal sequence identity or two non-related families of membrane proteins can have very similar hydropathy profiles, indicating similar global structures (Lolkema and Slotboom, 1998).

Subsequently, a search method that incorporates patterns of hydropathy profiles was developed by Clements and Martin (2002). A hydropathy profile pattern is the pattern of peaks in the hydropathy profile of a given protein. Searches based on hydropathy profile patterns were shown to identify new members of functional classes of transmembrane proteins not detected by sequence alignment alone.

4.1.4 Methods that combine sequence alignment with secondary structure information

The detection of homologous globular proteins was improved by combining sequence alignment with secondary structure information (Rost *et al*., 1997, Park *et al*., 1998, Rychlewski *et al*., 2000 , Lindahl and Elofsson, 2000). Similarly, a small number of studies have developed search protocols for homologous transmembrane proteins based on sequence comparisons that incorporate topological information. These studies are especially relevant to the current work and therefore are discussed in detail.

Hedman *et al*. (2002) focused on finding homologous members of the G Protein Coupled Receptor (GPCR) family and developed a new approach, called the Pmembr method, which adds information about predicted transmembrane segments to standard Smith-Waterman and profile-sequence (PSI-BLAST) search algorithms. Basically, the alignment score is increased if two residues predicted to belong to transmembrane segments align. A notable advantage of Pmembr, compared to methods using only sequence based algorithms, is that the number of false positives is significantly reduced in searches for closely and distantly related proteins.

The first group to design a homology search that combines transmembrane protein specific sequence constraints with profile-profile based comparisons was Bernsel *et al*. (2007). Termed SHRIMP, the protocol incorporates a Hidden Markov Model (HMM) that integrates sequence information with predicted topology and hydrophobicity data to detect related proteins. The HMM profile is constructed from multiple sequence alignments and expanded using a second alphabet corresponding to predictions of either hydrophobicity or transmembrane topology. Sequence profiles, with the same additional information, are then scored against the model, and paths

through the model corresponding to alignments where transmembrane regions are matched to each other will have a relatively higher probability. This search method was applied initially to the database of G-protein coupled receptors (GPCRDB; Horn *et al*., 2003). To gauge the ability to detect distant homologs, only hits to GPCRs from different classes were considered positives, whereas hits within a GPCR class were ignored and hits to non-GPCRs (from GPCRDB and Swiss-Prot) were considered negative. Evaluation of SHRIMP indicated that introducing structural information to the profile-profile method improves detection of distant homologs. In addition, SHRIMP performed better than the profile-sequence based method Pmembr. The ability of the SHRIMP method to find close homologs within, rather than between, GPCR classes was also assessed. Again, the SHRIMP method performed better than Pmembr method.

Subsequently, the SHRIMP method was applied to the HOMEP database (Forrest *et al*., 2006). The HOMEP database comprises 36 homologous transmembrane proteins with solved crystal structures, which can be classified into 11 SCOP families. Applying SHRIMP to HOMEP corroborated that adding topological information improves homology detection. This notwithstanding, Bernsel *et al*. found that the SHRIMP method does not clearly recognize clan relationships in the Pfam database. Specifically, Bernsel *et al*. reported that although the performance of the SHRIMP classifier was greater across the whole range of false positive rates (relative to a simple classifier based on sequence similarity), the improvement was limited.

4.1.5 Methods based on helix interaction patterns

Recently, Fuchs and Frishman (2010) reported a new search method, which identifies relationships between transmembrane proteins by clustering them according to similarities among transmembrane helix interaction graphs. A helix interaction graph is generated by considering the transmembrane helices as graph nodes and the interactions between helices as the edges of the graph. For each pair of helices, the number of residues in contact is determined from the structure by evaluating a minimal distance between opposing side chains or backbone atoms. All helix pairs with at least one residue in contact were considered as interacting. This method produces a score called HISS, which encapsulates to what extent the architecture of transmembrane helix bundles are conserved. The search method was applied to all the available transmembrane proteins with solved 3D structures and revealed that common helix interaction patterns are indeed conserved among proteins with distinctly different sequences but with the same structure. Moreover, when clustering was performed according to helix interaction similarity on structurally available transmembrane proteins with more than four transmembrane helices, 20 recurrent helix architectures were discovered and 15 singleton proteins. Of note, this classification approach, as it is based on the extent helix interactions are similar, is reminiscent of conventional structural classifications for globular proteins, such as SCOP and CATH, and led to the appreciation that helix interactions are key in determining transmembrane protein structure.

4.1.6 The present work

In the current work, a transmembrane protein homology detection method is presented that integrates sequence alignment with structural information. Our method

incorporates into iterative multiple sequence alignments (PSI-BLAST) information regarding the predicted transmembrane segments as well as comparisons between predicted residue orientations and loop lengths. This approach to detecting homologous transmembrane proteins is expected to correctly detect the relationship between transmembrane proteins in cases when simple sequence alignment (such as PSI-BLAST) fails to detect any homology. Of note, the current work is the first method to employ helix orientation predictions when searching for homology among transmembrane proteins. Moreover, our method is the first to combine a number of different structural features of transmembrane proteins with sequence alignment. Another novel feature is that when more structural information is available, it will be fairly easy to incorporate it into our method.

In addition to improving the accuracy of homology detection, we aimed to ensure that this advanced search method could be used on a large scale for automated classification of transmembrane proteins. This feature is critical as transmembrane proteins are challenging to work with experimentally and currently there is no structure-based transmembrane protein database comparable to the water-soluble protein databases, such as SCOP and CATH. In particular, the difficulty in performing structural studies underscores the need for innovative bioinformatics tools that enable automated classification of the functional and evolutionary relationships between transmembrane proteins.

4.2 Methods

4.2.1 Databases

When evaluating a homology detection method, it is crucial to have query set and a test database where both false and true relationships between the queries and database proteins are already known. For globular proteins, structure-based databases such as SCOP (Murzin *et al*., 1995) and CATH (Orengo *et al*., 1995) can serve this purpose. However, a "gold standard" database for transmembrane proteins does not yet exist, as very few transmembrane protein structures have been solved.

This notwithstanding, the G-protein coupled receptor database (GPCRDB; Horn *et al*., 1998, Horn *et al*., 2003) has proved a reliable database when testing searches for homologous transmembrane proteins. It is a well maintained and manually curated database, collating and validating large amounts of heterogeneous data concerning GPCRs. Categorization in GPCRDB is based on classes, which contain proteins with similar function and sequence homology. The GPCRDB database divides the GPCR superfamily into six classes: Class A rhodopsin-like, which account for over 80% of all GPCRs, Class B secretin-like, Class C metabotropic glutamate receptors, Class D pheromone receptors, Class E cAMP receptors and the Class F frizzled/smoothened family. Initially, we used GPCRDB when developing and testing our search method.

To evaluate whether the search method can be applied to other transmembrane proteins, subsequently we tested our method using the Pfam database (Bateman *et al*. 2004; described in detail in Chapter 1). Pfam is a semi-automatically maintained

database that contains a collection of protein families and domains, as well as multiple alignments and profile-HMMs that characterize these families. Of note, Pfam is not as reliable as GPCRDB with regards to categorizing the relationships between proteins, as the former employs a semi-automated classification method unlike GPCRDB that is manually curated. Protein domains are categorized in the Pfam database into families and clans. A clan is a collection of families judged likely to be homologous. Families are classified mostly automatically but clans are built manually, based on various sources of information.

4.2.1.1 GPCRDB

Our search method was tested using the GPCRDB, and we chose to build the test data set in a similar way to previous studies (Hedman *et al*., 2002 and Bernsel *et al*., 2007). The query set was created by downloading six classes of GPCRs from GPCRDB (June 2006 release 10.0) and selecting randomly 127 proteins such that they were proportionately distributed among the GPCRDB classes. Redundant proteins were removed using the CD-HIT program (Li *et al*., 2002) to ensure that the sequence identity between any two proteins was less than 50%. The final query set is shown in Table 9.

Table 9: Query set from GPCRDB.

Class	**No. of proteins in GPCRDB**	**No. of proteins in**
Rhodopsin like	4949	90
Frizzled-Smoothened	113	6
Secretin like	231	12
Fungal pheromone receptors	58	7

Metabotropic glutamate / pheromone	160	12
cAMP receptors	7	1

The test database for sequence alignment included proteins from the six GPCRDB classes with a maximum of 90% redundancy (according to CD-HIT). In addition the test database included ~21,000 proteins from the non-redundant sequence database Uniref90 (Suzek *et al.*, 2007) to serve as potential false hits (negative set). Only proteins containing more than one transmembrane helix according to TMHMM prediction (Krogh *et al.*, 2001) were added to the database. GPCRs were excluded from the ~21,000 protein negative set by screening for GPCR description in Swiss-Prot entries both manually and automatically. In addition, CD-HIT-2D (Li *et al.*, 2002) was utilized, which compares two protein sets and identifies the sequences in a second set that are similar to those in the first set above a preset threshold; the identity threshold was set at 99%. Moreover, proteins were excluded from the test database if their Swiss-Prot description had any of the following terms: uncharacterized, unidentified, unknown, predicted, hypothetical, undetermined or probable. These proteins were excluded from the database because relationships between these proteins and others are known to be confounding.

4.2.1.2 Pfam database

The database of Pfam transmembrane domains, used for testing homology search, was the same database used in chapter 3, for benchmarking PSI-BLAST. It was built by selecting proteins in the Pfam database, version 19.0 (Pfam-A file) that had at least one of the following transmembrane protein terms in their Swiss-Prot description:

transmembrane, membrane, membranous, intramembrane, transporter, pump, channel and receptor. Proteins were excluded from the database if the Swiss-Prot description had any of the following terms: uncharacterized, unidentified, unknown, predicted, hypothetical, undetermined or probable. In addition, highly homologous sequences (greater than 90% identity) were excluded; homology reduction was carried out using the CD-Hit program.

The query set from the Pfam database, was same set used in chapter 3. For building the set initially, all clans with transmembrane terms in their description (as described above) were listed (using the Pfam-C file). Then the families within each clan and the domains they contain were listed. Domain sequences were extracted from the "Pfam-A.fasta" file, which contains each domain name and the family with which it is associated. Homologous sequences (greater than 50% identity) were excluded from the list using the CD-HIT program. Furthermore, only domains with more than one transmembrane helix according to TMHMM (Krogh *et al*., 2001) were retained. Finally, the domain query set was chosen randomly from this list of domain sequences. The final query set included 112 randomly chosen proteins, from 29 different clans.

4.2.2 Sequence alignment searches using PSI-BLAST

Sequence alignment searches were performed using PSI-BLAST (Altschul *et al*., 1997) to identify all the transmembrane proteins homologous to a given query in a corresponding test database.

After inspecting the PSI-BLAST benchmarking results (presented in chapter 3) we

concluded that the best value to use for the h-parameter, which defines the E-value for inclusion when building up the PSI-BLAST profile, is 10^{-6} for Pfam family homology level and 10^{-3} for Pfam clan level. This value resulted in the optimal sensitivity for transmembrane homology detection, namely the number of true positive PSI-BLAST results was high yet the number of false positives was low. Since we aimed to use our method for homology detection at the level of proteins families, which is a more restricted level than clans homology level we chose to run PSI-BLAST with E-value of 10^{-6}.

The parameter that determines the maximum number of aligned proteins printed (-v parameter) was set to 3,000. PSI-BLAST was set to run up to 5 iterations (-j parameter). The remaining PSI-BLAST parameters were left at default values. PSI-BLAST results with an E-value smaller than 1 were listed and analyzed (-e 1).

For each query, PSI-BLAST results were listed and noted for further study. The E-value of each PSI-BLAST result was used as the PSI-BLAST method score. As mentioned previously, the E-value represents the number of times one would expect to get a hit with the same or better score by chance. Thus, the lower the E-value, the greater the sequence similarity between the PSI-BLAST result and the query protein.

4.2.3 Integrating secondary structure information with PSI-BLAST E-values to improve searches for homologous proteins

The following steps were performed to calculate secondary structure scores (Figure 22):

(1) For the GPCRDB query set, all PSI-BLAST results were listed and noted for further analyses. For Pfam database queries, full length sequences were extracted for both queries and PSI-BLAST results. It should be noted that when searching for homologs in the Pfam database the PSI-BLAST search was performed using only domain sequences, nevertheless when considering the secondary structure we chose to look at the full length sequence. According to Liu *et al*. (2004), the majority of transmembrane proteins have only a single transmembrane domain. Therefore, we assumed that working with full length proteins at this point in our method would not impact the results and would enable better, more accurate detection of homologous proteins, especially in cases where the sequence domains are truncated. However, later it was found that this precaution was unnecessary, as the full length sequences and the Pfam domains possessed the same number of helices in all proteins under study.

(2) Each PSI-BLAST result was aligned to the corresponding query protein using sequence to sequence global alignment with the Needleman-Wunsch algorithm, using Blosum62 as the substitution matrix and gap penalty of 8. As mentioned, when considering secondary structure we chose to consider the entire length of all query sequences and therefore could not use the PSI-BLAST output alignments.

(3) The locations of possible transmembrane helices were predicted using TMHMM2.0 (Krogh *et al*., 2001). TMHMM was applied to both the query proteins and the PSI-BLAST result. TMHMM2.0 was chosen because of its speed relative to other methods such as MEMSAT-SVM (Nugent and Jones, 2009), though the latter is

a better predictor of transmembrane helix location and topology.

(4) The following structural scores were calculated: transmembrane helix location score (henceforth referred to as ‘helix score’), residue orientation score, loop score and combined score, the latter representative of the other scores (see below). How each score was calculated is described below.

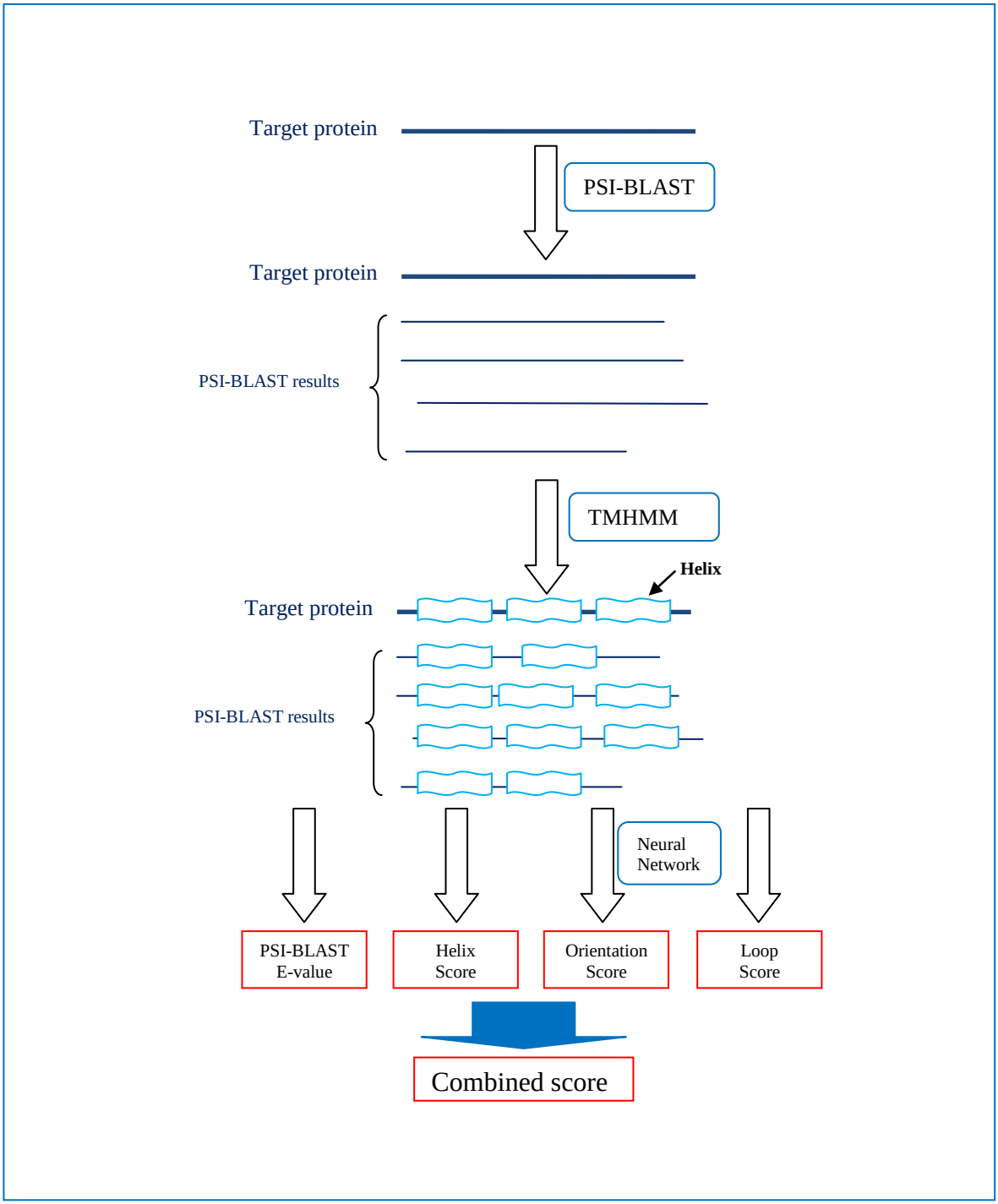


Figure 22: Diagrammatic outline of the developed search method. Homology detection using PSI-BLAST was the fundamental step. Then the locations of helices were predicted using TMHMM. Next, several structural scores (helix score, residues orientation score and loop score) were calculated for each PSI-BLAST result. The last step involved finding the optimal weights to generate a combined score.

4.2.3.1 Helix score

The transmembrane helix location score (helix score) was calculated by counting the number of residues aligned between the query and the PSI-BLAST result that are predicted in both cases to reside in a transmembrane helix. The resulting value was normalized by dividing it by the total number of residues in all the predicted helices in the query protein.

4.2.3.2 Residue orientation score

The orientation of each residue was predicted as described in detail in Chapter 2. Briefly, the sequences of each query and PSI-BLAST results were input to a neural network, which had been trained to determine whether a residue is buried in the core of a helix-bundle or exposed to the lipid environment surrounding the transmembrane protein. The residues orientation score was calculated by counting the number of residues aligned between the query and the PSI-BLAST result that are predicted in both cases to be not only inside a helix but also in the same orientation. The resulting value was normalized by dividing it by the number of residues in all the predicted helices in the query.

Alternative ways of calculating an orientation score were tested and found to be less effective (Table 12). Briefly, we tried scoring the overall orientation of each helix based on a threshold number of residues that are buried or exposed and assigning 1 to the score when the test and query had similar helix orientation. In addition, we tried calculating Euclidian distance between the helixes score of the target and the PSI-BLAST result.

4.2.3.3 Loop score

The loop score was calculated as reported previously by Aria *et al*. (2004):

$$S_{1,2}(\%) = 100 * \sum_{i=1}^{n+1} \min(l_{1,i}, l_{2,i}) / \sum_{i=1}^{n+1} \max(l_{1,i}, l_{2,i}) \qquad (8)$$

Where n is the number of transmembrane helices. $l_{1,i}$ and $l_{2,i}$ are the length of the i-th loop in sequences 1 and 2, namely the sequences of the query and PSI-BLAST result, respectively. When the number of helices in the query protein and PSI-BLAST result were not equal, the best score from aligning any continuous combination of loops was used.

4.2.3.4 Combined score

The combined score for each PSI-BLAST result was defined as the classifier of our search method. The E-value, helix score, residues orientation score and loop score were combined as follows:

$$Combined = w1*(-\log(Eval)) + w2*Helix + w3*orientation + w4*Loop \qquad (9)$$

Additional ways of combining the scores were evaluated but found to be less effective, including:

- Using the scores as they are, in a simple linear equation.
- A linear equation:

$$Combined = w1*Eval + w2*e^{Helix} + w3*e^{Orientation} + w4*e^{Loop} \qquad (10)$$

The steps for calculating the weights used to generate the combined score are

described in Figure 23. In the first step the query set was divided randomly into ten sets of proteins. For each of the ten training sets from GPCRDB or Pfam, the Paramopt program (by Prof. D. Jones, not published) was run separately to determine the optimal weights. The Paramopt program searches for an optimal set of command line parameters using a genetic-style search. Paramopt used the AUC (the area under the ROC curve) for minimization.

This test was repeated five times, resulting in a total of 50 sets of weights for each database. The mode and the average values of the weights were calculated. Then the performance of using the PSI-BLAST E-value score alone was compared to the performance of using the combined score calculated using the mode or average weights.

Overfitting of the weights was avoided by applying a 10-fold cross validation test: the 50 training sets of a particular database were divided into ten sets. Then the mode and the average values of the weights were calculated each time without one of the sets. The resulting weight values were then used to calculate a combined score.

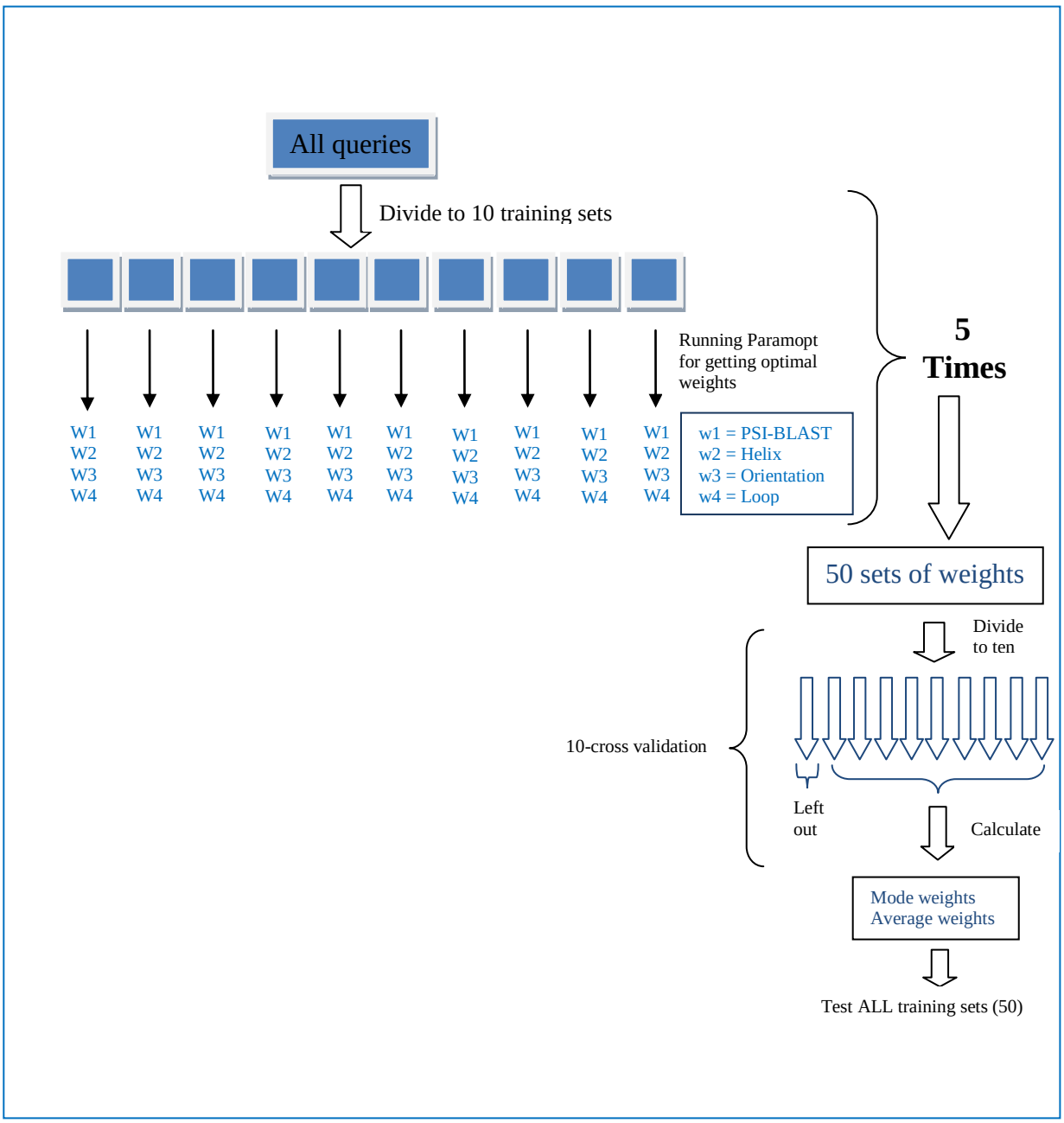


Figure 23: Diagrammatic outline of the steps we took to calculate the combined score weights. The queries were divided to 10 training sets. Paramopt program was run on each set to derive the optimal set of weights. This was repeated five times, resulting in 50 sets of weights. The 10-cross validation test was performed to calculate the mode and average of the weights, which were then tested on all (50) training sets.

For GPCRDB: the mode values of the weights were similar for all of the 10 tests, and were found to work on all 50 training sets individually. Accordingly, the combined score using these weights was found to perform better than PSI-BLAST E-

value score. In contrast, the average values of the weights were comparable across the 50 training sets, but could not be applied to each one of the test sets. Therefore, the mode values of the weights were used to generate the combined score for GPCRDB proteins.

For Pfam database: the mode values of the weights were: E-value weight = 1, Helix weight = 0, combined weight = 0 and Loop weight = 0. Meaning, that the E-value score is the only score that contributes to the combined score in most of the training sets. In addition, the average values could not be applied to each one of the test sets. Similar results were obtained for clan and family homology level.

4.2.4 Evaluating the ability of the search method to identify homologous transmembrane proteins

For each query, PSI-BLAST results were listed and their scores (helix score, orientation score and combined score) were calculated. Then each one of the scores was used as a classifier, i.e., used for discriminating between true positive PSI-BLAST results and false positive PSI-BLAST results, and their performance was evaluated as described in the next sections.

4.2.4.1 Defining a true positive – homologous proteins

In the case of homology searches performed using GPCRDB, a PSI-BLAST result was considered a true positive if it is classified in the GPCRDB database in the same class as the query protein.

For searches performed using the Pfam database, two levels of homology were tested: (i) A PSI-BLAST result was considered true positive if the query and the PSI-BLAST result appear in the same Pfam family. (ii) A PSI-BLAST result was considered true positive if the query and PSI-BLAST result appear in the same Pfam clan.

4.2.4.2 Classifier performance assessment

Receiver operating characteristic (ROC) curve analysis was used in order to assess the performance of using each classifier: PSI-BLAST E-value, helix score, orientation score and the combined score. The ROC curve is a plot of the true positive rate (TPR) against false positive rate (FPR), as the threshold value of the classifier is varied. True positive (Tp) and False negative (Fn) together constitute the total number of true results, in other words the truly homologous proteins, while False positives (Fp) and True negatives (Tn) constitute the total number of false results, namely unrelated proteins.

True positive rate (TPR, equation 5) and False positive rate (FPR) are calculated as follows:

$$FPR = \frac{Fp}{Fp + Tn} \qquad (11)$$

An ROC curve can be interpreted either graphically or numerically. Interpreting the ROC curve numerically involves calculating the AUC (the Area Under the ROC curve, Hanley and McNeil, 1982). An AUC score of 1 indicates that the true positives are perfectly separated from the negatives; i.e., the classifier assigns higher scores to all true positives than to any false positives, so that the true positives are at the top of

the sorted list. An AUC score of 0 indicates that no true positives are found. If one ROC curve is higher than another, it has a greater AUC indicating a better classifier performance. If two ROC curves cross over at any point, then each classifier outperforms the other under some conditions, and comparing AUC values is not very informative.

A ROC curve was plotted for each classifier: PSI-BLAST E-value, helix score, residues orientation score and combined score. Then the corresponding AUC was calculated.

4.2.4.3 Testing the weights used to generate the combined score

A set of weights was tested by comparing the performance of a combined score classifier generated using the weights to the performance of a classifier that is the PSI-BLAST E-value alone. In other words, the area under the ROC curve (AUC) was calculated when the PSI-BLAST E-value was used as a classifier and compared with that calculated when the combined score was used as the classifier.

4.3 Results and discussion

Results for the two different databases tested, GPCRDB and Pfam, are presented separately in the sections below.

4.3.1 Homology detection - GPCRDB

4.3.1.1 Finding the optimal weights for the combined score - GPCRDB

As described above, to integrate structural information with sequence alignment data all four parameters (E-value, helix score, residue orientation score and loop score) were consolidated into one combined score by a linear combination of the log of the E-value and the three other scores (Equation 9). The weights for the equation were found by dividing the query set into ten training sets (repeated five times) and using the Paramopt program for retrieving the optimal weights for these training sets. The mode values of the weights (Table 10) were found to work and are applicable to all test sets.

Table 10: The optimal weights for each parameter used to generate a combined score.

Parameter	**Weight**
-Log(E-value)	10
Helix score	0
Residue orientation score	10
Loop score	0

The PSI-BLAST E-value and the residues orientation score are both high, indicating that they are the key parameters for detecting homology among proteins in this database. However, as the E-value parameter is exponential (in the range used,

smaller than 1), there will be high E-values that reduce its contribution to the combined score such that the other parameter, the orientation score, predominates. Notably, the high number for the orientation score weight reveals that this structural parameter in particular plays an important role in the performance of the classifier and contributes considerably to improving homology detection versus using only the E-value score.

4.3.1.2 Homology detection results - GPCRDB

Homology searches were performed using the GPCRDB test database and query set and each of the following classifiers: PSI-BLAST E-value, helix score, residues orientation score, and the combined score.

Examination of the ROC curve (Figure 24) revealed that when the false positive rate was low, the residues orientation score, helix score and the combined score performed better as classifiers than the PSI-BLAST E-value. Nevertheless, the best classifier overall was the combined score. The AUC values confirmed that the combined score performed as the best classifier (Table 11). In summary, ROC curve analysis shows that integrating the parameters improves homology detection when using the GPCRDB test database.

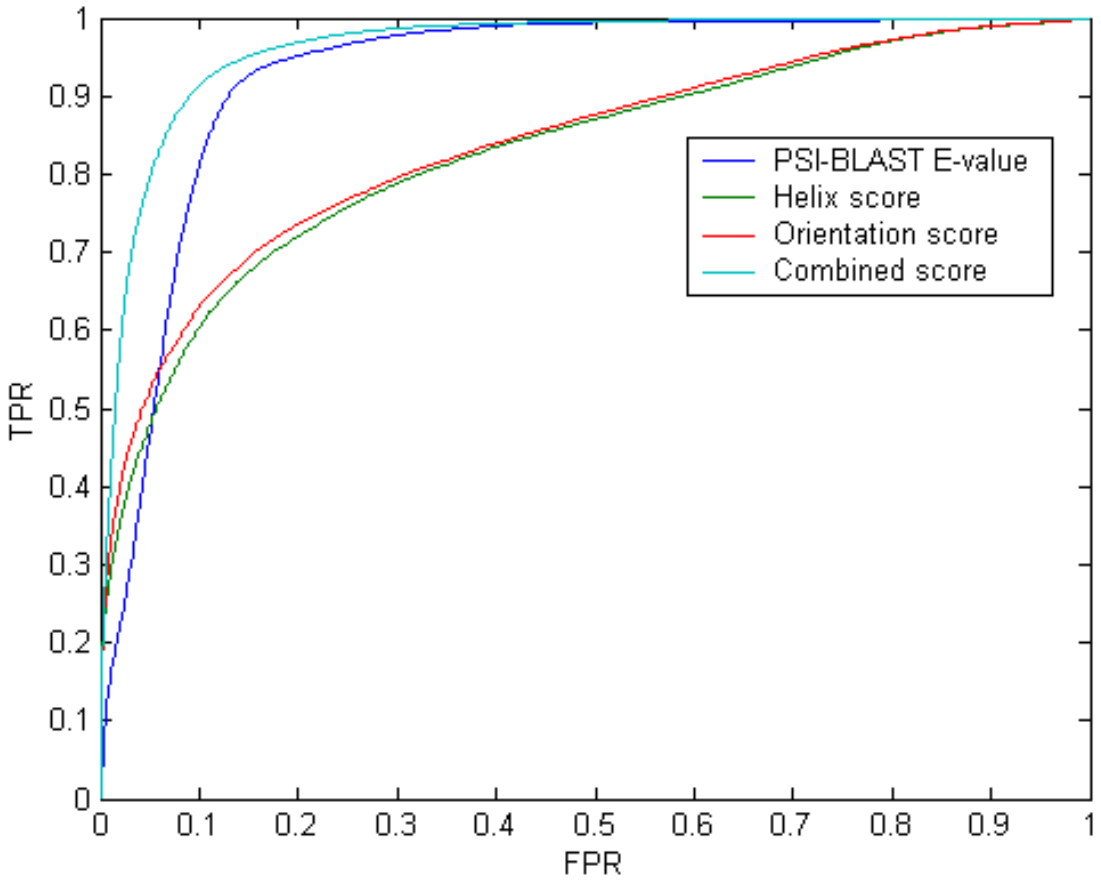


Figure 24: ROC curves for homology searches performed using the GPCRDB test database and query set and each of the following classifiers: E-value (blue), helix score (green), residue orientation score (red), combined score (light blue).

Table 11: AUC values when each classifier is used to search for homologous proteins in the GPCRDB test database.

Classifier	**AUC**
PSI-BLAST E-value	0.93
Helix score	0.83
Residues orientation score	0.84
Combined score	0.96

The results of two alternative ways of calculating an orientation score are shown in Table 12. In one approach a helix orientation score was calculated based on the overall orientation of each helix. In the second method a helix orientation score was calculated based on the Euclidian distance between the helix scores of the target and the PSI-BLAST result. Calculating the orientation score using either of these approaches was not effective.

Table 12: AUC values when each classifier is used to search for homologous proteins in the GPCRDB test database – testing alternative ways of calculating the residue orientation.

Classifier	**AUC**
Residue orientation score	0.84
Helix orientation score	0.54
Helix orientation score – Euclidian distance	0.46

4.3.2 Homology detection - Pfam database

To see if the combined classifier can identify homologous transmembrane proteins when other transmembrane protein families are included in the search, we applied our method to detect homologous proteins in the Pfam database. The ability of the combined classifier to detect two levels of homology was tested; the first level was homologous family membership and the second level was homologous clan membership.

4.3.2.1 Finding the optimal weights for the combined score – Pfam database

Initially, homology searches were performed using the Pfam test database and query set and the combined classifier, the latter based on the weights derived using the GPCRDB test database and query set. However, a combined score based on these GPCRDB weights did not perform well as a classifier. Therefore, we decided to derive independently for each level of homology (family and clan) a set of weights using the Pfam set. We reasoned that each database and Pfam homology level was created in a different way and has distinct features with potential to influence the classifier performance. Specifically, the Pfam families were classified automatically based on domain sequences whereas Pfam clans were built manually by gathering Pfam families together according to structure and function similarity. Thus, domains in a Pfam clan could have distant sequence homology.

Even more dramatically than when the weights were derived using the GPCRDB query set, when weights were derived for the Pfam query set we found that the PSI-BLAST E-value is the key and only parameter for detecting homology. Thus, in the case of the combined score derived using the Pfam query set, it is clear that sequence similarity plays a dominant role. The only weight above 0 in the mode values of the weights was the weight for the E-value score. In addition, when testing the average values of the weights, it was impossible to generalize the weights of the combined score for clan homology level and for family homology level as well. Thus, it was impossible to derive optimal weights that could be used for all training sets of Pfam families and clans. For some training sets, the PSI-BLAST E-value was the only

parameter that contributed to the combined score but for other training sets the residues orientation score was the key parameter and introducing any other reduced the performance of the classifier. The inability to derive an optimal set of weights is likely explained by the fact that though most Pfam families were built automatically and are thus, generally sequence dependent, other Pfam families were manually generated. Alternatively, it is possible that the combined score does not work well when trying to detect homologous proteins from different families. It could be that there are families in which the proteins are similar to each other in sequence and other families in which the sequence similarity is smaller, but the structural similarity is prominent.

4.4 Comparing our search method to other transmembrane homology detection methods

There are only two other methods comparable to the one developed in the current work. The first is the Pmembr method (Hedman *et al*., 2002) and the second one is SHRIMP (Bernsel *et al*., 2007), both were detailed in the introduction of this chapter. Like the present method, these other two methods combine structural and sequence information for transmembrane homology detection.

4.4.1 Comparison with the Pmembr method

Pmembr (Hedman *et al*., 2002), described before in section 4.1.4, incorporates information about predicted transmembrane segments into standard Smith-Waterman

and profile-sequence search algorithm, PSI-BLAST. This method was tested using the GPCRDB on two homology levels. First, individual classes within the GPCRDB were considered. Therefore, hits to GPCR sequences outside a given class were ignored and only hits inside the class defined as true. Second, GPCRDB was considered as a superfamily. Thus, hits to GPCR sequences in all classes were considered true and hits to the same class were ignored. In both tests, non-GPCRs were defined as incorrect hits. Using these tests, the Pmembr search method, which adds topology information to the PSI-BLAST search method, was demonstrated to improve slightly the ability to detect both closely related GPCRs (first level of homology) and distantly related GPCRs (second level of homology), as compared to PSI-BLAST alone. Adding the structural information to standard Smith-Waterman was less effective.

In the present work we tested the ability of our method to classify sequences correctly to the relevant GPCR class. Therefore, hits to GPCRs inside the given class were considered true and all other hits were defined as incorrect, including hits to GPCRs in other classes. Using these strict criteria, we noted that most of the false positive results were GPCRs belonging to other classes that did indeed possess some sequence and structural similarity. It is likely that the performance of our search method would have attained a higher score if we had chosen to ignore such false positives, as in the Hedman *et al*. study. We applied such strict criteria as our goal was to develop a method that is capable of correctly classifying to a specific class, as we consider this capability a requirement for automatic classification.

In summary, in light of the dissimilar test criteria and query set it is impossible to directly compare rigorously our method with the Pmembr method. In addition, the

Pmembr website is not maintained anymore and we have experienced difficulties with running a standalone Pmembr program (e.g., an error message was received saying the profile is too long and this error could not solved. Of note, the authors were contacted but were not able to solve the problem). Nevertheless, Pmembr result files could be downloaded from the Pmembr website from a directory that contained all queries and corresponding PSI-BLAST results files with the Pmembr score, and these files were used to compare the Pmembr method with our method. We chose to download the results of running PSI-BLAST with the h-parameter (threshold for inclusion of new sequences in each iteration of PSI-BLAST) set at 10^{-5}, which is similar to the value employed in our method (10^{-6}) and considered the best h-value to detect both closely and distantly related GPCRs (Hedman *et al*. 2002). For the comparison between Pmembr and our method, the definition of false and true positive hits for Pmembr were changed to meet our test criteria, i.e., true positive hits were defined as GPCRs inside the given class and all other hits were considered false positives, including GPCRs in other classes. In addition, only PSI-BLAST results with an E-value smaller than 1 were listed (Hedman *et al*. listed hits with an E-value smaller than 99).

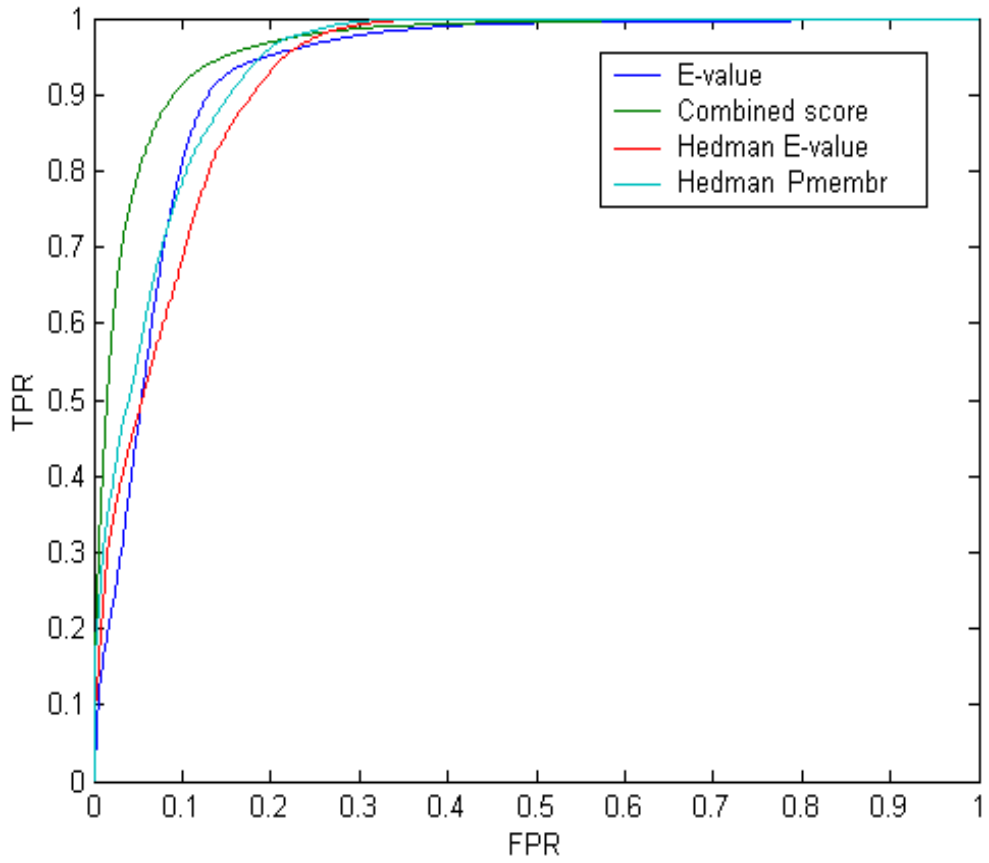


Figure 25: ROC curves for homology searches (a) performed using Hedman *et al*. (2002) test database and query set and each of the following classifiers: PSI-BLAST E-value (red), Hedman score (light blue) (b) performed using the current method database and each of the following classifiers: PSI-BLAST E-value (current work, dark blue) or combined score (current work, green).

ROC curves were plotted for the PSI-BLAST classifier (generated using Hedman *et al*. data) and the Pmembr classifier (Figure 25) and compared to curves generated using the classifiers defined in the current work (PSI-BLAST E-value and combined score, using the data generated in the current work); AUCs were calculated (Table 13).

The AUC values (Table 13) support that the Pmembr score serves as a better classifier than the PSI-BLAST E-value even when our criteria are applied to the analyses. However, examination of the ROC curves (Figure 25) revealed that our classifier (the

combined score) slightly improves homology detection relative to the PSI-BLAST E-value, even more than Pmembr.

Table 13: AUC values for homology searches when each Pmembr classifier is used (Hedman *et al*., 2002).

Classifier	**AUC**
PSI-BLAST E-value –Hedman *et al*.	0.926
Pmembr score – Hedman *et al*.	0.940
PSI-BLAST E-value – current work	0.93
Combined score – current work	0.96

The performance of the classifiers was evaluated also by plotting sensitivity curves (Figure 26), which show the true positive number versus the false positive number per query. The reason for drawing sensitivity curves for the comparison and not only ROC curves was the different number of total results, between the current work set and Hedman *et al*. set, and in addition the E-value classifier of Hedman et. al and the current work result in different plots, making comparison of ROC curves less clear. For sensitivity curves generally, when considering the sorted list it is desirable for more true positives to appear at the top of the list before a given number of false positives, such that the number of false positives is as low as possible.

Examination of the sensitivity curves (Figure 26) revealed that the total number of false positive hits in homology searches performed using the Hedman *et al*. data and Pmembr classifiers is much bigger than in the current study. To understand this finding the list of false positives from the Pmembr data was inspected, with special

attention paid to hits not belonging to any of the GPCRDB classes (from Swiss-Prot). It was found that in some cases the false positives were GPCRs not present in GPCRDB, some were proteins with only one helix and some had one of the following terms: uncharacterized, unidentified, unknown, predicted, hypothetical, undetermined and probable in their Swiss-Prot description (again, these were filtered out of our test database and quey set). Thus, it is likely that some of the false positives detected using the Hedman *et al*. set and Pmembr classifiers are due to the less rigorous filtering of the Hedman *et al*. test database and query set. Nevertheless, because the control plots (PSI-BLAST E-value curves) are dissimilar, it is hard to compare the two methods using the sensitivity curve.

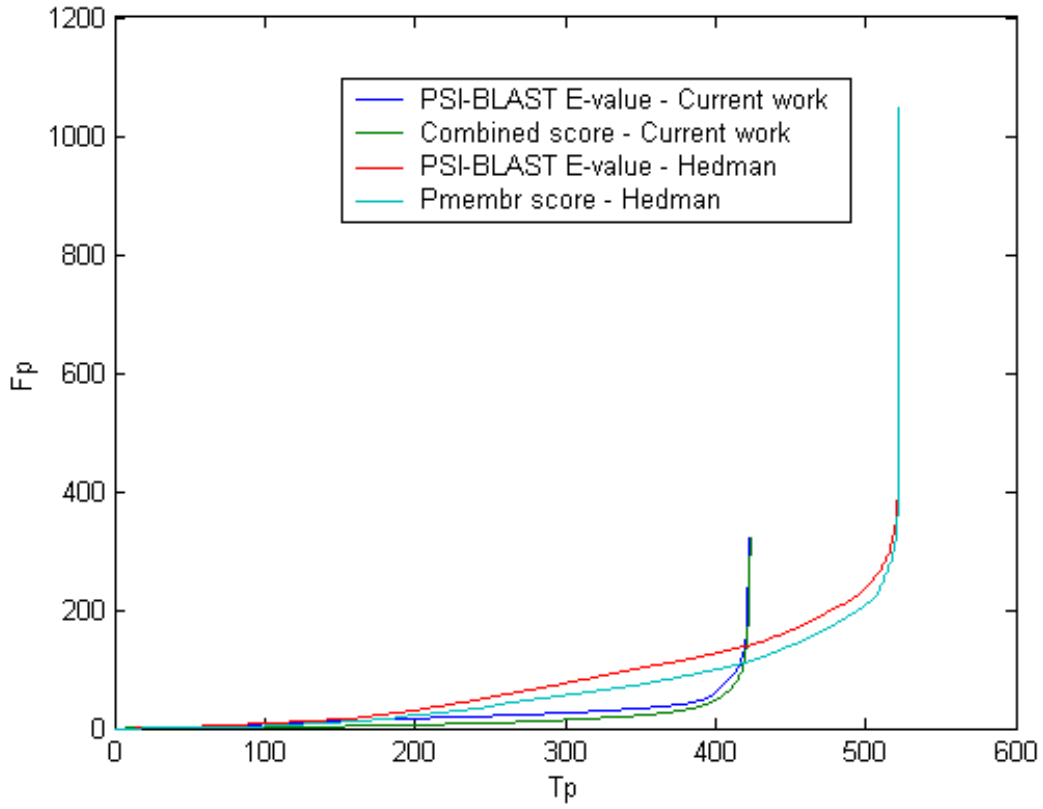


Figure 26: Sensitivity curves for homology searches performed using one of the following classifiers: PSI-BLAST E-value (current work, blue), combined score (current work, green), PSI-BLAST E-value (Hedman *et al*., red) or Pmembr score (Hedman *et al*., light blue). The true and false positive numbers were divided by the number of proteins in the test set (79 in the case of Hedman et al. and 112 in the case of the current work) to show the false positive/true positive per query**.**

4.4.2 Comparison with SHRIMP method

The SHRIMP method (Bernsel *et al*., 2007), described before in section 4.1.4, incorporates a Hidden Markov Model (HMM) that integrates sequence information with predicted topology and hydrophobicity, with each of these structural features added separately to the HMM. The method was tested originally using the GPCRDB on two homology levels, in a similar way to the Pmembr method. It was shown that introducing structural information to the method improves homology detection for distantly related GPCRs but is less helpful for close homologs. Furthermore, for both homology levels, it was demonstrated that the SHRIMP method performs much better than the Pmembr method. When SHRIMP was tested using another database, HOMEP (Forrest *et al*., 2006), similar levels of improvement in detection were reported.

The SHRIMP method was also tested using the Pfam database and found to be unable to clearly recognize clan relationships. Specifically, it was reported that although the performance of the classifier was increased across the whole range of false positive rates, the improvement was limited and the data was not shown. In the supplementary data files of the SHRIMP study there was a list of 126 alignments, which were the false hits that had high scores when detecting Pfam clan homology. Bernsel *et al*. suggested that domains in the list as yet unassigned to a clan were likely genuine homologs. SHRIMP was not tested on Pfam families, or it was not reported.

It appears that the developers of SHRIMP encountered the same problems as we did when testing their search method using Pfam clans. In particular, a difficulty to

differentiate between false and true positives with high scores assigned to alignments between Pfam domains not belonging to the same clan.

Comparing the reported SHRIMP test results directly with our method results is unhelpful as different definitions and query sets were employed in each case. Unfortunately, we were unable to compare our method with the SHRIMP method directly using our database and query set because we encountered problems when running the standalone program. Running a standalone SHRIMP program involves two steps: in the first one a profile database is made containing transmembrane helix predictions for all the sequences in the database (using the create_db.pl script) and in the second step another script is used (search_db.pl) which creates a profile-HMM from the query sequence, then predicts its transmembrane topology and finally uses the HMM to search for homologs in the profile database. Already when running the first step (create_db.pl script) we got an error message and the second step created an empty results file. In line with advice from the developers of SHRIMP (Prof. Arne Elofsson), we tried using older versions of PSI-BLAST (blast-2.2.10 and blast-2.1.3) from the one we typically employ (blast-2.2.21), but still encountered the same problems.

To compare our method indirectly with SHRIMP, we reviewed supplementary data in Bernsel *et al*. (2007), concerning detecting close homologs within, rather than between, GPCR classes. Here we should emphasize that for the SHRIMP evaluation, a result was considered a false positive if it was not classified to any GPCRDB class. Whereas in the current work, false positives are also proteins classified to different GPCR classes (other than the query's class).

In Figure 27 (Bernsel *et al*., 2007), the methods, PSI-BLAST and Pmembr, were compared to the method SHRIMP, which was found to perform better. The SHRIMP results are presented in three ROC curves: sequence + topology information (SHRIMP-tmpred), sequence + hydrophobicity (SHRIMP-hphob) and sequence only (SHRIMP-seq). All three ROC curves reach the true positive rate 1.0 at very low false positive rates. Thus, for closely related homologs, sequence alone is sufficient when using SHRIMP.

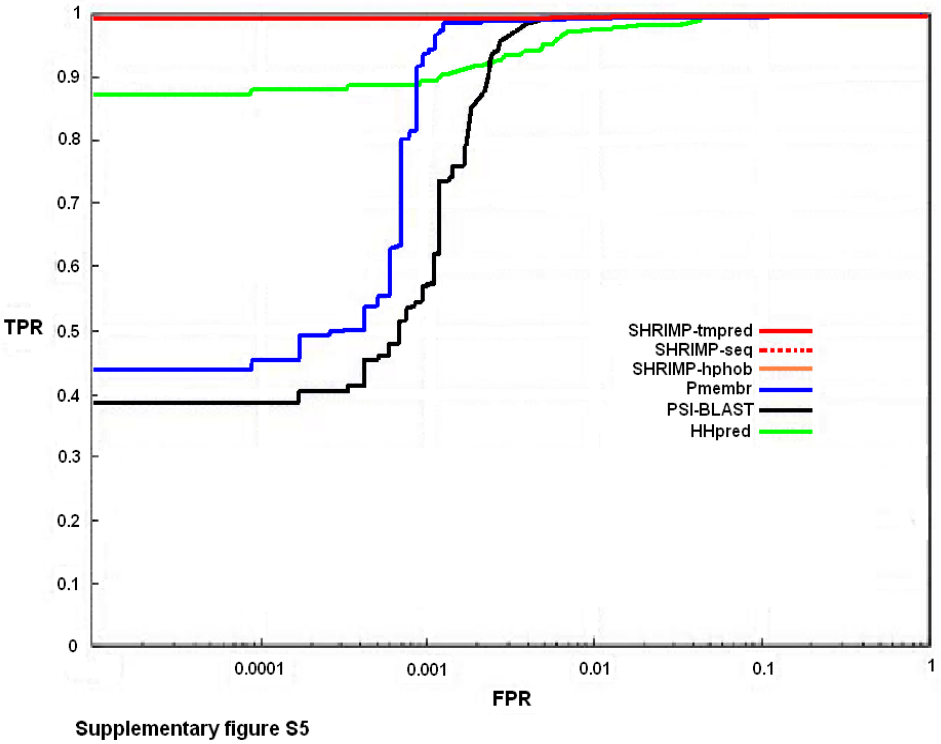


Figure 27: ROC curves presented in supplementary data of Bernsel *et al*., 2007. Comparison between the method PSI-BLAST (black), Pmembr (blue), HHpred (Soding *et al*., 2005, green), and SHRIMP (red) : sequence + topology information (SHRIMP-tmpred), sequence + hydrophobicity (SHRIMP-hphob) and sequence only (SHRIMP-seq).

Given these data, as our method performs only slightly better than Pmembr, we suspect that the SHRIMP method is superior to ours in detecting related transmembrane proteins. If this is indeed the case, then it appears that a method which

is based on a profile HMM-profile HMM algorithm could be more powerful than methods in which structural information has been added to a profile-sequence based method, as presented in the current study. Nevertheless, in light of the present study, we propose that developing a combined score for adding to a system similar to SHRIMP's method could result in an even better performance.

4.4.3 Exploring helices number in Pfam clans and TMHMM performance

In order to check features of the Pfam database which might have contributed to the poor performance of our homology search of this particular database, we checked the number of helices in all the clans in the queries set (Table 14).

By exploring Table 13 we became aware that the Pfam database contains truncated sequence. For example Pfam Clan CL0192 which contains GPCRs with mostly 7 transmembrane helices was predicted to have mean number of 5.7 transmembrane helices.

In addition we compared TMHMM performance with another topology method, MEMSAT-SVM (Nugent and Jones, 2009) and applied to the GPCR database, which contains only proteins with 7 transmembrane helices (Table 15).

Table 14: Statistics of transmembrane helices in Pfam clan protein domains; TMHMM was used when predicting number of helices

Pfam clan name	Mean number of	Standard
CL0015	11.4	1.8
CL0030	8.0	3.4
CL0062	11.1	2.3
CL0111	9.8	2.4
CL0130	7.8	1.6
CL0347	3.5	1.1
CL0192	5.7	1.9
CL0375	3.8	0.6
CL0425	11.1	4.4
CL0404	6.5	2.6

Table 15: Comparing TMHMM and MEMSAT-SVM prediction of number of transmembrane helices in GPCRDB.

Topology method used	Mean number of helices	Standard deviation
TMHMM	6.8	0.8
MEMSAT-SVM	7	0.6

4.5 Conclusions

In this chapter we have presented a new method for detecting homologous transmembrane proteins. On the premise that structure is better conserved than sequence, our method combines multiple sequence alignment (PSI-BLAST) with structural information regarding helical regions, helical residue orientations and loop lengths. We validated that our method has an improved capability to detect true relationships between transmembrane proteins relative to a method based solely on simple multiple sequence alignments.

Specifically, we found that combining the PSI-BLAST E-value with the structural parameter (residues orientation score) generated a combined score that served as a superior classifier, detecting more true positives with less false positives when using the GPCRDB. To combine the parameters, we had to derive optimal weights and thus, we corroborated that the PSI-BLAST E-value, i.e., sequence similarity between the proteins, is a key parameter. This finding is in agreement with data from numerous studies of GPCRs, establishing that helical sequences are strongly conserved among GPCRs. Conserved residues that mediate ligand binding and selectivity of G-protein coupling tend to cluster on the cytoplasmic side of transmembrane helices, while residues unique to each subfamily tend to appear on the extracellular side (Suwa *et al*., 2011). Accordingly, we found that the residues orientation parameter also contributes significantly to the performance of the classifier.

Notably, the loop parameter was found to have no effect on homology detection, indicating that loop lengths are not conserved enough to influence the performance of

a classifier. Accordingly, the length of the third cytoplasmic loop (CL3) and the N- and C-terminal loops have been shown to vary among GPCRs, though the other loops have conserved lengths (EL1, EL2, EL3, CL1, CL2) in nearly all GPCRs (Suwa *et al*., 2011).

The helix parameter also did not improve classifier performance. This finding is likely explained by the fact that the orientation score already encompasses topological information; namely, the orientation score relates specifically to residues predicted to be in helical regions both in the query protein and PSI-BLAST result.

Our newly developed search method proved less effective at detecting homology among Pfam database proteins. Our method was not able to improve the ability to detect homology at the level of Pfam clans or at the level of Pfam families. We suspect that this is due to the way clans are defined, which is based on sequence similarity. In addition, as mentioned above, there are unassigned domains which are possibly genuinely homologous. Moreover, while GPCR is well characterized database, it could be that the Pfam clans are not characterized well enough.

In addition, we suspect that certain features common to GPCRs contribute to the ability of our method to detect GPCR homology. In particular, all or most GPCRs have 7 transmembrane helices and share similar topology. For example, GPCRs do not contain any non-canonical elements such as wide turns, tight turns, kinks and reentrant loops, which makes it easier to accurately predict topology and residue orientation. It is not surprising that familial relationships between transmembrane proteins with dissimilar numbers of helices or more complicated structural features, as present in the Pfam database, are going to be harder to identify.

More generally, our detection method is based on two key predicted parameters: the location of the transmembrane regions, which is predicted using TMHMM, and the orientation of the residues, which is predicted by a neural network developed in the current work. The incorporation of two predicted parameters into the final score raises the possibility of error particularly if the number of helices is not constant within a family.

We also submit that certain features of the Pfam database might have contributed to the poor performance of our homology search of this particular database. First, when checking the Pfam Clan CL0192 (Table 13), which contains GPCRs with mostly 7 transmembrane helices, we became aware that the Pfam database contains truncated sequences. Second, as exemplified by Clan CL0192 that comprises almost all GPCRs, the classification of clans is fairly broad. In contrast, in the GPCRDB GPCRs are categorized into five classes. This second feature likely underscores why we got a smaller number of false positives when testing Pfam clans as opposed to GPCRDB.

We also suspected that the performance of our method was influenced by our choice to use TMHMM when predicting the helical regions. To explore this possibility, TMHMM was compared with another topology method, MEMSAT-SVM (Nugent and Jones, 2009) and applied to the GPCR database, which contains only proteins with 7 transmembrane helices. The results (Table 14) indicate that although MEMSAT-SVM is slightly more accurate at detecting the 7 transmembrane helices of the GPCRs than TMHMM, TMHMM does work well at least for the GPCRDB.

Bernsel et. al. (2007) remark in their study that perhaps Pfam is not an optimal set choice when testing homology detection methods. We share this opinion. A better test

of homology detection among transmembrane protein families, other than GPCR, will require future characterization of a greater number of transmembrane protein structures.

A comparison of our developed method with two other published methods suggested that profile HMM-profile HMM based methods could be more powerful than profile-sequence based methods, even after the addition of structural information as described here. Nevertheless, based on the present study, we propose that combining a profile-profile method with a combined score could improve even further the detection of related transmembrane proteins.

Chapter 5

Discussion and Future work

5.1 Discussion

Transmembrane proteins play crucial roles in a variety of cellular processes and comprise 20-25% of fully sequenced genomes (Jones, 1998, Wallin and von Heijne, 1998). Nevertheless, the tertiary structures of only a small number of transmembrane proteins are known. Hence, it is of great importance to develop theoretical methods capable of predicting transmembrane protein structure and function based on protein sequence alone. To address this, in the current work we aimed to develop a method for identifying homologous transmembrane proteins that could be used for classifying the proteins into structural and functional families based on sequence similarity and predicted structural features.

The method for detecting homology, presented in this thesis, comprises in the first step sequence alignment searches, which are performed using PSI-BLAST. Then profiles derived from the multiple sequence alignments are input into a neural network, developed in this work to predict which transmembrane residues are buried (core of the helix-bundle) or exposed (to the lipid environment). A maximum accuracy of 86% was achieved. Moreover, for almost half of the query set, the

predicted residue orientation was more than 70% accurate. In the last step of the work presented here, the predicted helix locations, residue orientations and loop length scores are combined with the PSI-BLAST E-value, to create a ‘combined’ classifier score. A few approaches to incorporating the information were tested. In the end, a linear equation was chosen for calculating the 'combined score' classifier score. While validating the performance of our ‘combined classifier’, it became clear that the sequence similarity between proteins is a dominant parameter. In addition, however, we found that the residue orientation parameter also contributes significantly to the performance of the classifier. In contrast, the loop parameter had negligible impact on homology detection, suggesting that loop lengths are not conserved enough to influence the performance of a classifier.

Having developed a homology detection method we tested its accuracy using a database of proteins. Ideally the database should be one in which the true relationships between transmembrane proteins are known. The Pfam database was chosen, as transmembrane proteins in this database have been classified into various families, though not entirely reliably. In addition, GPCRDB was employed, as this database, though narrow, is well-studied and maintained.

We found that the 'combined score' classifier, as compared to a classifier based solely on PSI-BLAST, resulted in more true positives with less false positives when it was tested using GPCRDB and could differentiate between GPCRDB families. However, the combined classifier did not improve homology detection when searching transmembrane proteins from the Pfam database. Other attempts to improve homology detection among transmembrane proteins from the Pfam database have failed as well (Bernsel *et al*., 2007), highlighting the challenge of generating

improved approaches to classify transmembrane proteins.

5.2 Future work

A comparison between the homology detection method developed here and two published methods (Hedman *et al*., 2002 and Bernsel *et al*., 2007) leads us to conclude that profile-profile based methods could be more powerful than profile-sequence based methods, even when the latter encompasses structural information as described here. In light of this finding, we propose that a profile-profile method should be developed to incorporate a combined score, as this is likely to improve even further the accuracy of homology searches among transmembrane proteins.

A profile-profile based method for detecting homology among transmembrane proteins that incorporates structural information could be developed in two possible ways:

1. To switch the PSI-BLAST alignment search in the presented ‘combined score’ method with a profile-profile alignment search, such as HHpred (Soding *et al*., 2005). In this way, the effectiveness of profile-profile based searches is exploited to find candidate homologs and then structural information is also considered, to help identify the genuinely homologous proteins.

2. To develop a new method, in which the 'combined score' is encompassed as a second alphabet of a profile HMM model. Bernsel *et al*. (2007) took a similar approach, whereby hydrophobicity was added to a profile HMM and then a profile HMM-profile HMM method was constructed.

In addition to the structural information incorporated in the current work into a 'combined score', there are different types of structural information about transmembrane proteins that should be considered in future studies:

1. Predicted location of re-entrant loops and kinks.
2. Predicted helix-helix interactions and helix tilt angles.
3. Predicted location of special functional motifs in the sequence, such signal peptides. Such motifs strongly influence the folding of transmembrane proteins, as they diminish the length of the final sequence by promoting the cleavage of specific segments at the N-termini (signal peptides) .
4. Predicted disulfide bridges. These are covalent bonds that link closely together two cysteine residues, constraining protein folding (Martelli *et al*., 2004).

The presented method could also be used to improve detection of homologous beta-barrel transmembrane proteins in a very simple way: the training set for the Neural Network used for residues orientation prediction should include beta-barrel proteins. Very few methods exist that detect homology among and classify beta-barrel transmembrane proteins (Remmert *et al*., 2009).

In summary, the method presented here can certainly be improved, but still serves as a useful starting point for developing an effective method for detecting homology. We suspect that as more transmembrane protein structures are characterized and classified, it will become easier to develop better methods for detecting homology among transmembrane proteins.

Appendices

Appendix A

Introduction to backpropagation neural networks

Artificial Neural Network

Artificial Neural Networks are processing devices that are based on the operation of biological neural networks. Neural networks are organized in layers made up of a number of interconnected and interacting components called nodes or neurons which contain an 'activation function'. The activation function of a node defines the output of that node given an input or set of inputs. The first layer of a typical neural network is the 'input layer', which communicates with one or more 'hidden layers' where the actual processing is done. The hidden layers then link to an 'output layer' where the answer is output.

Most neural networks have some type of learning rule which modifies the weights through a learning algorithm, according to the input patterns presented to it. Thus, the network learns by example.

There are numerous kinds of neural network architecture the most commonly used is backpropagation network, which is used in the current work and is presented in the next section.

Backpropagation network architecture

The most commonly used backpropagation network architecture is a feedforward network, as shown in Figure 28.

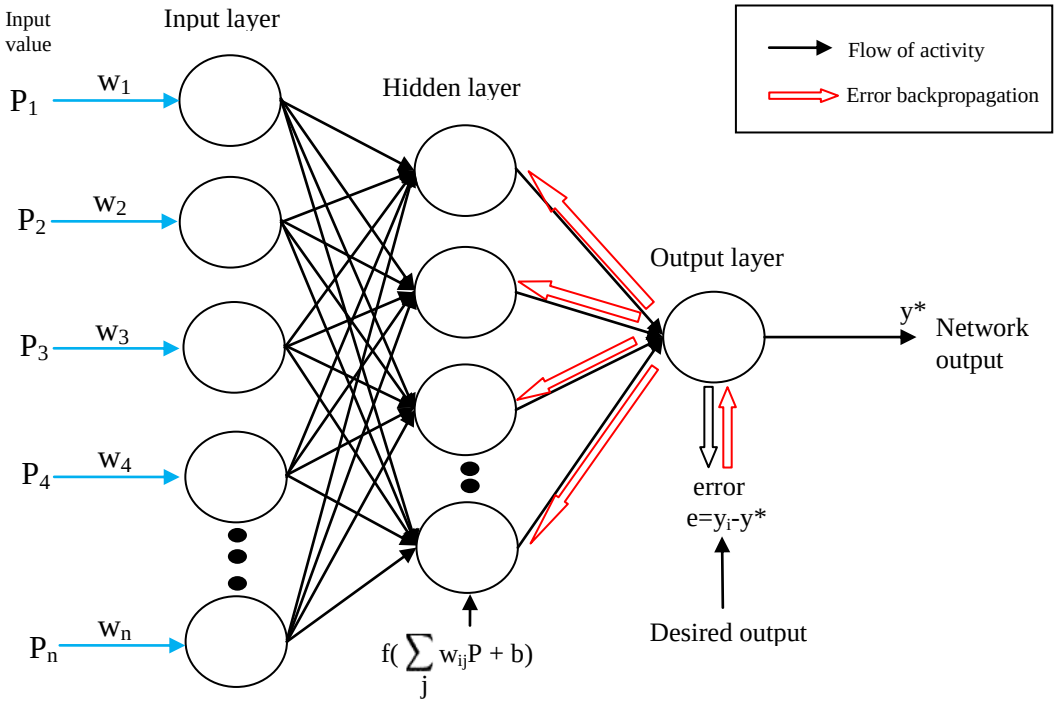


Figure 28: Backpropagation network architecture. Each input is weighted by a real number – w. The sum of the weighted inputs and the bias b forms the input to the transfer function f at each hidden node. Neurons may use any differentiable transfer function to generate their output.

The input vectors are used to train a network until it can approximate a function. The training process requires a set of inputs and outputs as an example of proper network behavior. Through the training process the weights and biases of the network are iteratively adjusted to minimize the mean square error (mse), which is the average squared error between the network outputs and the target outputs.

A standard backpropagation training algorithm is the gradient-descent algorithm, also

known as steepest descent, in which the network updates the weights and biases in the direction of the negative of the gradient.

Adaptive steepest descent with momentum (traingda in Matlab), combines the "Adaptive steepest descent algorithm" with the "Steepest descent with momentum algorithm". "Adaptive steepest descent" is a steepest descent algorithm, which changes the learning rate during the training process. This algorithm trains the network faster than the simple steepest descent algorithm. In "Steepest descent with momentum", a momentum constant regulates the amount of the weight change, which can be a number between 0 and 1. When the momentum constant is 0 a weight change is based solely on the gradient. When the moment constant is 1, the new weight change is set equal to the last weight change and the gradient is ignored. Momentum can prevent the algorithm from getting stuck in a shallow local minimum (Neural Network toolbox for Matlab).

The training can be done in two ways: incremental mode or batch mode. In the incremental mode, the weights are updated after each input is applied to the network. In the batch mode all of the inputs are applied to the network before the weights are updated.

Appendix B

Substitution matrices in sequence similarity methods

Substitution matrices describe the rate at which one character in a sequence changes to another character over time. In the process of evolution, from one generation to the next the amino acid sequences of an organism's proteins are gradually altered through the action of DNA mutations. Each amino acid is more or less likely to mutate into various other amino acids.

A substitution matrix is a 20x20 matrix where the (i, j) *th* entry is equal to the probability of the i th amino acid being transformed into the j th amino acid over a given amount of evolutionary time. There are many different ways to construct such a matrix. The most common substitution matrixes are: PAM and BLOSUM.

PAM

One of the first amino acid substitution matrices, the PAM matrix, was developed by Dayhoff (Dayhoff *et al*., 1978). This matrix is calculated by observing the differences in closely related proteins. The PAM1 matrix estimates what rate of substitution would be expected if 1% of the amino acids had changed. The PAM1 matrix is used as the basis for calculating other matrices by assuming that repeated mutations would

follow the same pattern as those in the PAM1 matrix, and multiple substitutions can occur at the same site (PAM 30 and the PAM70 are most commonly used).

BLOSUM

The BLOSUM was developed by Henikoff and Henikoff (1992). A set of matrices were constructed using multiple alignments of evolutionarily divergent proteins. The probabilities used in the matrix calculation are computed by looking at "blocks" of conserved sequences found in multiple protein alignments. These conserved sequences are assumed to be of functional importance within related proteins. To reduce bias from closely related sequences, segments in a block with a sequence identity above a certain threshold were clustered, giving weight 1 to each such cluster. For the BLOSUM62 matrix, this threshold was set at 62%. Pair frequencies were then counted between clusters, hence pairs were only counted between segments less than 62% identical. One would use a higher numbered BLOSUM matrix for aligning two closely related sequences and a lower number for more divergent sequences. BLOSUM62 matrix works well detecting similarities in distant sequences, and that is the matrix used by default in most recent alignment applications such as BLAST.

Appendix C

Publication arising from the thesis

Hurwitz N., Pellegrini-Calace M. and Jones D.T. (2006)

Towards Genome-scale Structure Prediction for Transmembrane Proteins

Phil. Trans. R. Soc. B 361, 465–475

References

1. Adams, P.D., Engelman, D.M & Brunger, A.T. (1996) Improved prediction for structure of the dimeric transmembrane domain of glycophorin A obtained through global searching, *Proteins*, 26: 256-261
2. Adamian L, Liang J (2006**)** Prediction of transmembrane helix orientation in polytopic membrane proteins, *BMC Struct Biol* , 6**:** 13
3. Altschul S.F, Gish W., Miller W., Myers E.W and Lipman D.J (1990) Basic local alignment search tool, *J Mol Biol*. 215(3): 403-10
4. Altschul S.F, Madden T.L., Schaeffer A.A., Zhang J., Zhang Z., Miller W. and Lipman D.J. (1997), Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res*, 25: 3389-3402.
5. Apic G, Gough G.J and Teichmann S.A.(2001). Domain combinations in archaeal, eubacterial and eukaryotic proteomes, *J Mol Biol*, 310(2):311-325.
6. Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Birney, E., Biswas, M., Bucher, P., Cerutti, L., Corpet,F., Croning, M. D. (2001). The InterPro database, an integrated documentation resource for protein families, domains and functional sites, *Nucleic Acids Res*. *29*: 37-40.
7. Arai M., Okumura K., Satake M., and Shimizu T. (2004) Proteome-wide functional classification and identification of prokaryotic transmembrane proteins by transmembarne topology similarity comparison. *Proteins Science*, 13: 2170-2183
8. Attwood T.K, Blythe M.J, Flower D.R.,Gaulton A., Mabey J.E, Maudling N. , McGregor L., Mitchell A.L, Moulton G., Paine K. and Scordis P. (2002) PRINTS and PRINTS-S shed light on protein ancestry. Nucleic Acids Res, 30(1): 239–241.
9. Baldi P., Brunak S., Chauvin C., Andersen C.A F. and Neilsen H. (2000) Assessing the accuracy of prediction algorithms for classification: an overview *Bioinformatics*, 16 (5): 412-424

10. Bateman A., Coin L., Durbin R., Finn R.D, Hollich V.,Griffiths-Jones S., Khanna A., Marshall M., Moxon S., Sonnhammer E.L.L, Studholme D. J., Yeats C. and Eddy S.R. (2004) The Pfam Protein Families Database. *Nucleic Acids Research*. Database Issue 32: D138-D141
11. Barth P, Schonbrun J, Baker D. (2007) Toward high-resolution prediction and design of transmembrane helical protein structures. *Proc Natl Acad Sci USA* 104: 15682–15687.
12. Barth P, Wallner B, Baker D (2009) Prediction of membrane protein structures with complex topologies using limited constraints. *Proc Natl Acad Sci USA* 106: 1409–1414.
13. Bairoch, A. and R. Apweiler (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000, *Nucleic Acids Res*, 28(1):45-48.
14. Beeley N.R.A and Sage C. (2003) GPCRs: an update on structural approaches to drug Discovery. *Targets*, 2, 1: 19-25
15. Bendtsen JD, Nielsen H, von Heijne G, Brunak S. (2004) Improved prediction of signal peptides: SignalP 3.0, *J Mol Biol*. 340(4):783-95.
16. Berman, H., Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. (2000) The Protein Data Bank. *Nucl. Acids Res* 28: 235 – 242
17. Beuming T and Weinstein H. (2004) A knowledge-based scale for the analysis and prediction of buried and exposed faces of transmembrane domain proteins, *Bioinformatics* 20 (12): 1822-1835.
18. Bernsel A., Viklund H. and Elofsson A (2007) Remote homology detection of integral membrane proteins using conserved sequence features, *Proteins* 71(3):1387-99.
19. Briggs JA, Torres J, Arkin IT (2001) A new method to model membrane protein structure based on silent amino acid substitutions. *Proteins*, 44(3):370-5.
20. Cao B, Porollo A, Adamczak R, Jarrell M, Meller J. (2006) Enhanced recognition of protein transmembrane domains with prediction-based structural profiles, *Bioinformatics* 22 (3):303–9.

21. Chen C.M. and Chen C.C (2003) Computer simulations of membrane protein folding: structure and dynamics, *Biophys. J.* 84:1902–1908

22. Churchill, G.A. (1989) Stochastic models for heterogeneous DNA sequences, *Bull Match Biol* 51(1): 79-94

23. Chothia, C. and Lesk A. M (1986). The relation between the divergence of sequence and structure in proteins, *EMBO J*, 5(4):823-826.

24. Chothia C. (1992). Proteins. One thousand families for the molecular biologist, *Nature* 357 (6379): 543–4.

25. Changhui Y. and Jingru (2010) A Comparison between Transmembrane Helices and Reentrant Loops. *BioInformatics and BioEngineering* (BIBE), 2010 IEEE International Conference, 3: 283 – 284

26. Clements J.D. and Martin M.E. (2002) Identification of novel membrane proteins by searching for patterns in hydropathy profiles, *Eur J Biochem*, 269(8):2101-7.

27. Corpet F, Servant F, Gouzy J, Kahn D (2000) ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons. *Nucleic Acids Res*. 28:267-269

28. Cserzo, M., Wallin, E., Simon, I., von Heijne, G. and Elofsson, A. (1997) Prediction of transmembrane alpha-helices in prokaryotic membrane proteins: the dense alignment surface method, *Protein Eng* 10 (6) : 673-676.

29. Dayhoff, M. O., Schwartz, R. M. and Orcutt, B. C. (1978). A model of evolutionary change in proteins. In Dayhoff, M. O. [ed] Atlas of protein sequece and structure, supplement 3. N*ational Biomedical Research Foundation, Washington DC*, 345-352.

30. Dewji N. and Singer S. J (1997) The seven-transmembrane spanning topography of the Alzheimer disease-related presenilin proteins in the plasma membranes of cultured cells. *Proc Natl Acad Sci U S A.* 94 (25)

31. Deisenhofer J., Epp O., Miki K., Huber R., and Michel H. (1985). Structure of the protein subunits in the photosynthetic reaction centre of Rhodospeudomonas viridis at 3 Å resolution, *Nature* 318**:**618-624.

32. Deisenhofer J. and Michel H. (1989) Nobel lecture. The photosynthetic reaction centre from the purple bacterium Rhodopseudomonas viridis, *EMBO J*. 8(8): 2149–2170.

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

34. Donnelly D., Overington J.P. and Blundel T.L (1994) The prediction and orientation of alpha-helices from sequence alignments: the combined use of environment-dependent substitution tables, Fourier transform methods and helix capping rules, *Protein Eng*., 7(5):645-53.

35. Eddy S.R. (1998) Profile Hidden Markov Models. *Bioinformatics,* 14:755-763

36. Eisenberg D., Weiss R.M. And Terwillger T.C (1984) The hydrophobic moment detects periodicity in the protein hydrophobicity, *Prot Natl Acad Sci USA* 81(1):140-144

37. Emanuelsson O, Brunak S, von Heijne G, Nielsen H (2007) Locating proteins in the cell using TargetP, SignalP and related tools, *Nat Protoc*, 2(4):953-971.

38. Engelman D, Zaccai G (1980) Bacteriorhodopsin is an inside-out protein, *Proc Natl Acad Sci USA*, 77(10)**:**5894-5898

39. Engelman, D., Steitz, T., and Goldman, A. (1986). Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins, *Annu Rev Biophys Biophys Chem*, 15:321-353

40. Engelman DM, Chen Y, Chin CN, Curran AR, Dixon AM, Dupuy AD, Lee AS, Lehnert U, Matthews EE, Reshetnyak YK, Senes A, Popot JL. (2003). Membrane protein folding: beyond the two stage model. *FEBS Lett*. 555, 122-125.

41. Falquet, L., Pagni, M., Bucher, P., Hulo, N., Sigrist, C.J, Hofmann, K. and Bairoch, A.(2002) The PROSITE database, its status in 2002, *Nucleic Acids Res*. 30, 235-238

42. Fauchere, J. L. and Pliska, V. (1983). Hydrophobic parameters pi of amino acid side chains from the partitioning of N-acetyl-amino-acid amides, *Eur J Med Chem*, 18:369-375.

43. Fariselli P, Finelli M, Marchignoli D, Martelli PL, Rossi I, Casadio R. (2003) MaxSubSeq: an algorithm for segment-length optimization. The case study of the transmembrane spanning segments, *Bioinformatics*, 19: 500-505

44. Finn R.D, Mistry J., Schuster-Böckler B., Griffiths-Jones S., Hollich V., Lassmann T, Moxon S, Marshall M., Khanna A., Durbin R., Eddy R.S, Sonnhammer E.L.L., and Bateman A.(2006) Pfam: clans, web tools and services, *Nucleic acids research*, 34 (Database issue), D247-51.

45. Fleishman SJ and Ben-Tal N (2002) A novel scoring function for predicting the alpha-helices, *J Mol Biol*, 321:363-378.

46. Forrest LR, Tang CL, Honig B. (2006) On the accuracy of homology modeling and sequence alignment methods applied to membrane proteins, *Biophysical Journal*, 91:508-517.

47. Fuchs A, Kirschner A, Frishman D (2009) Prediction of helix-helix contacts and interacting helices in polytopic membrane proteins using neural networks. *Proteins*, 74: 857–871.

48. Fuchs A. and Frishman D. (2010) Structural comparison and classification of alpha helical transmembrane domains based on helix interaction patterns, *Proteins*, 78:2587–2599.

49. Gaboriaud, C., Bissery, V., Benchetrit, T., Mornon, J.P. (1987) Hydrophobic cluster analysis: An efficient new way to compare and analyse amino acid sequences, *FEBS Lett*. 224:149-155

50. Gimpelev M, Forrest LR, Murray D, Honig B. Biophys J. *(*2004) Helical Packing patterns in Membrane and Soluble Proteins, *Biophysical Journal*, 87:4075–86

51. Goder V, Junne T, Spiess M. (2004) Sec61p Contributes to Signal Sequence Orientation According to the Positive-Inside Rule, *Mol Biol Cell* 15:1470–8.

52. Gough, J., Karplus K, Hughey R, Chothia C. (2001) Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J. Mol. Biol.* 313: 903–919

53. Grisshammer, R. and Tate, C.G. (1995) Overexpression of integral membrane proteins for structural studies, *Q. Rev. Biophys*. 28: 315–422

54. Gribskov M., McLachlan A.D and Eisenberg D. (1987) Profile analysis: detection of distantly related proteins, *Proc. Natl. Acad. Sci* USA, 84: 4355-4358

55. Hedman M., Deloof H., Von Heijne G., and elofson A. (2002) Improved detection of homologous membrane proteins by inclusion of information from topology predictions, *Proteins Science*, 11:652-6

56. Heger A, Wilton CA, Sivakumar A, Holm L. (2005) ADDA: a domain database with global coverage of the protein universe. *Nucleic Acids Res*. 33(Database issue):188-91.

57. Haft D.H., Selengut J.D. and White O. (2003) The TIGRFAMs database of protein Families, *Nucl. Acids Res*. 31 (1): 371-373.

58. Hall SE, Roberts K and Vaidehi N (2009) Position of helical kinks in membrane protein crystal structures and the accuracy of computational prediction, *Mol Graph Model.* J, 27(8):944-50.

59. Hanley J.A and McNeil B.J (1982), The meaning and use of the area under a receiver operating characteristic (ROC) curve, *Radiology* 143, 29–36.

60. Henikoff S. and Henikoff J. G. (1992). Amino acid substitution matrices from protein blocks, *Proc Natl Acad Sci U S A*, 89(22):10915-10919.

61. Henikoff, J. G., Greene, E. A., Pietrokovski, S., and Henikoff, S. (2000). Increased coverage of protein families with the blocks database servers. *Nucleic Acids Res*. 28: 228-230.

62. Hildebrand PW, Lorenzen S, Goede A, Preissner R (2006) Analysis and prediction of helix helix interactions in membrane channels and transporters, *Proteins* , 4**:** 253

63. Honig B, Yang A (1995) Free energy balance in protein folding**,** *Adv Protein Chem*, 995, 46:27-58.

64. Horn F, Weare J, Beukers MW, Hörsch 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(1):275-9.

65. Horn F., Bettler E, Oliveira L., Campagne F, Cohen F. E. and Vriend G. (2003) GPCRDB information system for G protein-coupled receptors, *Nucleic Acids Res*. 31:294-297

66. Huang Y., Cai J., Ji L. and Li Y. (2004) Classifying G-protein coupled receptors with bagging classification tree, *Computational Biology and Chemistry, 28*, issue 4: 275-280

67. Illergard K., Callergari S. and Elofsson A. (2010) MPRAP: an accessibility predictor for a-helical transmembrane proteins that performs well inside and outside the membrane, BMC *Bioinformatics,* 18;11:333.

68. Inoue Y., Ikeda M. and Shimizu T. (2004) Proteome-wide classification and identification of mammalian-type GPCRs by binary topology pattern, *Computational Biology and Chemistry*, 28: 39-49

69. Jayasinghe S., Hristova K. and White S. H. (2001) MPtopo: A database of membrane protein Topology, *Protein Sci.* 10:455-458

70. Jelinek F., Bahl L., Mercer (1975) R. Design of a linguistic statistical decoder for the recognition of continuous speech, *IEEE Transactions Information Theory* 21(3) : 250-256

71. Jones DT, Taylor WR, Thornton JM. (1994) A model recognition approach to the prediction of all-helical membrane protein structure and topology, *Biochemistry*, 15;33(10):3038-49.

72. Jones DT, Taylor WR, Thornton JM. (1994) A mutation data matrix for transmembrane proteins, *FEBS Lett*. 21;339(3):269-75.

73. Jones D.T (1998) Do transmembrane protein superfolds exist? *FEBS Letters* 27; 423(3):281-5

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

75. Jones DT and Swindells MB. (2002) Getting the most from PSI-BLAST, *TRENDS in Biochemical Sciences*, 27 (3)

76. Jones DT (2007) Improving the accuracy of transmembrane protein topology prediction using evolutionary information, *Bioinformatics*. 1;23(5):538

77. Kaczanowski S and Zielenkiewicz P (2010). Why similar protein sequences encode similar three-dimensional structures?, *Theoretical Chemistry Accounts* 125:543-50

78. Kall L, Krogh A, Sonnhammer ELL (2004) A Combined Transmembrane Topology and Signal Peptide Prediction Method, *J. Mol. Biol.* 338:1027–36

79. Kall L, Krogh A, Sonnhammer EL (2005) An HMM posterior decoder for sequence feature prediction that includes homology information, *Bioinformatics*, 21 Suppl 1:i251-7.

80. Kabsch W. & Sander C. (1983) Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features, *Biopolymers***,** 22:2577-2637.

81. Karwath A. and Hing D.R (2002) Homology Induction: the use of machine learning to improve sequence similarity searches, *BMC Bioinformatics* 3:11

82. Karplus, P. (1997) Hydrophobicity regained, *Protein Sci*., 6(6):1302-1307

83. Kernytsky A. and Rost B. (2003) Static benchmarking of membrane helix predictions, *Nucl. Acids Res*, 31 (13): 3642-2644

84. Klammer M., Messina D.N. , Schmitt T. and Sonnhammer E.L.L (2009) MetaTM - a consensus method for transmembrane protein topology prediction, *BMC Bioinformatics*, 10**:**314

85. Kuroiwa T., Skaquchi M., Omura T. and Lihara K.(1996) Reinitiation of protein translocation across the endoplasmic reticulum membrane for the topogenesis of multispanning membrane proteins, *J. Bio. Chem.*, 271(11): 6423-6428

86. Kyte J. and Doolittle RF (1982) A simple method for displaying the hydrophathic character of protein, *J Mol. Biol.* 157:105-132

87. Krogh A., Brown M., Mian I.S, Sjolander K. and Haussler D. (1994) Hidden Markov models in computational biology. Applicarion to protein modeling, *J Mol Biol*, 305(3): 567-80

88. Krogh A., Larsson B., von Heijne G. and Sonnhammer E.L.L (2001) Predicting transmembrane protein topology with a hidden markov model: application to complete genomes, *J Mol. Biol*. 305:567-580

89. Langelaan DN, Wieczorek M, Blouin C, Rainey JK. (2010) Improved helix and kink characterization in membrane proteins allows evaluation of kink sequence predictors, *J Chem Inf Model*. 50(12):2213-20.

90. Lasso Gorka, Antoniw John F. and Mullins Jonathan G.L. (2006) A combinatorial pattern discovery approach for the prediction of membrane dipping (re-entrant) loops. *Bioinformatics.* 22(14):290 - 297

91. Ledesma A, de Lacoba MG, Arechaga I and Rial E (2002) Modelling the transmembrane arrangement of the considerations of the nucleotide-binding site. *J Bioenerg Biomembr*, 34:473-486.

92. Lee, B.K and Richards, F.M. (1971) The interpretation of protein structure: Estimation of static accessibility, *J. Mol. Bio.* 55:379-400

93. Lee, A. G. (2004) How lipids affect the activities of integral membrane proteins, *Biochimica Et Biophysica Acta-Biomembranes* 1666: 62-87.

94. Lemesle-Varloot, L., Henrisaat, B., Gaboriaud, C., Bissery, V., Morgat, A., Mornon, J.P. (1990) Hydrophobic cluster analysis: Procedure to derive structural and functional information from 2-D representation of protein sequences, *Biochimie* 72:555-574

95. Lindahl E. and Elofsson A. (2000). Identification of related proteins on family, superfamily and fold level, *J.Mol.Bio* 295:613-625

96. Li W. , Jaroszewski L. and Godzik A. (2001) Clustering of highly homologous sequences to reduce the size of large protein database, *Bioinformatics*, 17:282-283

97. Li W. , Jaroszewski L. and Godzik A. (2002) Tolerating some redundancy significantly speeds up clustering of large protein databases, *Bioinformatics*, 18:77-82.

98. Liu Y., Engelman D.M. And Gerstein M (2002) Genomic analysis of membrane protein families: abundance and conserved motifs, *Genome Biology research*, 3(10): 0054

99. Liu Y., Gerstein M. and Engelman, D.M. (2004) Transmembrane protein domains rarely use covalent domain recombination as an evolutionary mechanism, *Proc. Natl. Acad. Sci. U. S. A.* 101: 3495-3497

100. Lipman DJ. and Pearson WR. (1985) Rapid and sensitive protein similarity searches, *Science*, 22;227(4693):1435-41.

101. Lolkema J.S and Slotboom D.J (1998) Estimation of structural similarity of membrane proteins by hydropathy profile alignment, *Mol Membr Biol.*, 15(1):33-42.

102. Lo A, Chiu YY, Rødland EA, Lyu PC, Sung TY, (2009) Predicting helix-helix interactions from residue contacts in membrane proteins, *Bioinformatics* 25: 996–1003.

103. Lo A, Cheng CW, Chiu YY, Sung TY, Hsu WL (2011) TMPad: an integrated structural database for helix-packing folds in transmembrane proteins, *Nucleic Acids Res.* 39: 347-55

104. Lomize, Mikhail A.; Lomize, Andrei L; Pogozheva, Irina D.; Mosberg, Henry I. (2006). OPM:Orientations of Proteins in Membranes database, *Bioinformatics* 22 (5): 623−625.

105. Lomize, Andrei L; Pogozheva, Irina D.; Lomize, Mikhail A.; Mosberg, Henry I. (2006). Positioning of proteins in membranes: A computational approach, *Protein Science* 15 (6): 1318−1333.

106. Lomize, Andrei L; Pogozheva, Irina D.; Lomize, Mikhail A.; Mosberg, Henry I. (2007) The role of hydrophobic interactions in positioning of peripheral proteins in membranes, *BMC Structural Biology* 7 (44): 44.

107. Marsico A., Henschel A., Winter C., Tuukkanen A., Vassilev B., Scheubert K. and Schroeder M. (2010) Structural fragment clustering reveals novel structural and functional motifs in α-helical transmembrane proteins, *BMC Bioinformatics*, 11: 204.

108. Martelli PL, Fariselli P, Casadio R. (2004) Prediction of disulfide-bonded cysteines in proteomes with a hidden neural network. *Proteomics*. Jun;4(6):1665-71.

109. Martin-Galiano AJ, Frishman D.(2006) Defining the fold space of membrane proteins: the CAMPS database, *Proteins*, 1;64(4):906-22.

110. Marti-Renom MA, Stuart AC, 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.

111. Muller T, Rahmann S, Rehmsmeier M. (2001) Non-symmetric score matrices and the detection of homologous transmembrane proteins *Bioinformatics*, 17, Supp l1:S182-9.

112. Murzin, A., Brenner, S. E., Hubbard, T. and Chothia, C. (1995). SCOP: A structural classification of proteins database for the investigation of sequences and structures, *J Mol Biol*, 247: 536-540.

113. Nilsson J., Person B. and von Heijne G. (2002) Prediction of partial membrane protein topologies using a consensus approach, *Protein Science,* 11:2974-2980

114. Nikiforovich G.V, Galaktionov S and Marshal G.R. (2001) Novel approach to computer modeling of seven-helical case of bactriorhodopsin, *Acta Biochim* 48(1): 53-64

115. Neumann S., Fuchs A., Mulkidjanian A., and Frishman D. (2010) Current status of membrane protein structure classification, *Proteins,* 15;78(7):1760-73.

116. Needleman,S.B and Wunsch, CD. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins, *J Mol Biol.* 48(3):443-53

117. Ng PC, Henikoff JG, Henikoff S. (2000) PHAT: a transmembrane-specific substitution matrix. Predicted hydrophobic and transmembrane, *Bioinformatics,* 16(9):760-6.

118. Nozaki, Y. and Tanford, C. (1971). The solubility of amino acids and two glycine peptides in aqueous ethanol and dioxane solutions, *J Biol Chem,* 246:2211-2217

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

120. Nugent T, Jones DT (2010) Predicting Transmembrane Helix Packing Arrangements using Residue Contacts and a Force-Directed Algorithm, PloS *Comput Biol* 6(3): 1000714.

121. Oberai, A., Ihm, Y., Kim, S. & Bowie, J. U. (2006). A limited universe of membrane protein families and folds, *Protein Sci.* 15:1723–1734

122. Orengo, C. A., Michie, A. D., Jones D.T, Swindells, M. B. and Thornton, J. M. (1997). CATH - a hierarchic classification of protein domain structures, *Structure,* 5:1093-1108.

123. Park J., Karplus K., Barrett C., Hughey R., Haussler D., Hubbard T. and Chothia C. (1998) Sequence comparison using multiple sequences detect three times as many remote homologes as pairwise methods, *J.Mol.Bio.* 284: 1201-1210

124. Park y. and Helms (2006) How strongly do sequence conservation patterns and empirical scales correlate with exposure patterns of transmembrane helices of membrane proteins?, *Biopolymers*, 83(4):389–399

125. Park, Y. and Helms, V. (2007).On the derivation of propensity scales for predicting exposed transmembrane residues of helical membrane protein, *Bioinformatics* 23:701-708

126. Park, Y., Hayat, S. and Helms, V. (2007). Prediction of the burial status of transmembrane residues of helical membrane proteins, *BMC Bioinformatics* 8: 302

127. Pearson W.R and Lipman D.J (1988) Improved tools for biological sequence comparison. *Proc Natl Acad Sci USA*,85(8):2444-8

128. Pellegrini-Calace, M., Carotti, A. and Jones, D.T. (2003) Folding in lipid membranes (FILM): a novel method for the prediction of small membrane protein 3-D structures. *Proteins Structure, Function, and Genetics*, 50:537-545

129. Persson, B. and Argos, P. (1994). Prediction of transmembrane segments in proteins utilising multiple sequence alignments. *J Mol Biol*, 237(2):182-192

130. Pietrokovski S (1996) Searching databases of conserved sequence regions by aligning protein multiple-alignments. *Nucleic Acids Res*. 24(19): 3836-3845

131. Pilpel Y., Ben-Tal N. and Lancet D. (1999) kPROT: a knowledge-based scale for the propensity of residue orientation in transmembrane segments. Application to membrane protein structure prediction. *J Mo Bio* 294: 921-935

132. Pirovano W., Feenstra K.A. and Heringa J. (2008) PRALINETM: a strategy for improved multiple alignment of transmembrane proteins, *Bioinformatics*, 24 (4): 492-497.

133. Popot and Engleman (1990) Membrane protein folding and oligomerization: the two- stage model, *Biochemistry*, 29: 17

134. Popot and Engelman (2000) Helical membrane protein folding, stability, and revolution, *Annu. Rev. Biochem*, 69:881-922

135. Radzicka, A. and Wolfenden, R. (1988). Comparing the polarities of the amino acids: side-chain distribution coefficients between the vapor phase, cyclohexane, 1-octanol, and neutral aqueous solution, *Biochemistry*, 27:1664-1670.

136. Rees D.C., DeAntonio L., Eisenberg D, (1989) Hyrophobic organization of membrane proteins, *Science*, 245, 4917: 510-513

137. Rees D.C. And Eisenberg D. (2000) Turning a reference Inside-out: commentary on an article by Stevens and Arkin Entitled: "are membrane proteins 'inside-out' proteins"?, *Proteins,* 38:121-122

138. Remm M. and Sonnhammer E. (2000) Classification of transmembarne proteins families in the Caenorhabditis elegans Genome and identification of human orthologs, *Genome Research*, 10(11):1679-1689

139. Remmert M., Linke D., Lupas A.N. and So ding J. (2009) HHomp—prediction and classification of outer membrane proteins, *Nucleic Acids Research*, 37: 446–451

140. Renthal R. (2008) Buried water molecules in helical transmembrane proteins, *Protein Science* 17:293-298

141. Riek RP, Rigoutsos I, Novotny J, Graham RM.(2001) Non-alpha-helical elements modulate polytopic membrane protein architecture. *J Mol Biol.* 306:349−62.

142. Riek R.P, Finch A. A., Begg G.E., and Graham R.M.(2008) Wide Turn Diversity in Protein Transmembrane Helices Implications for G-Protein-Coupled Receptor and Other Polytopic Membrane Protein Structure and Function, *Mol Pharmacol* 73:1092–1104

143. Rigoutsos I, Riek P, Graham RM, Novotny J.(2003) Structural details (kinks and non- a conformations) in transmembrane helices are intrahelically determined and can be predicted by sequence pattern descriptors, *Nucl. Acids Res*, 31:4625–31.

144. Rose A., Lorenzen S., Goede A, Gruening B. and Hildebrand P.W.(2009), RHYTHM—a server to predict the orientation of transmembrane helices in channels and membrane-coils, *Nucleic Acids Research*, 37, Web Server issue 575–580

145. Rose A, Goede A. and Hildebrand PW (2010) MPlot--a server to analyze and visualize tertiary structure contacts and geometrical features of helical membrane proteins, *Nucleic Acids Res.* ;38

146. Rost B. and Sander C. (1994) Conservation and prediction of solvent accessibility in protein families. *Proteins* 20:216-226

147. Rost B., Casadio R, Fariselli P. (1996) Topology prediction for helical transmembrane proteins at 86% accuracy. *Prot. Sci.* 4:521 -33

148. Rost B., Schineider, R., and Sanedr C.(1997) Protein fold recognition by prediction- based threading. *J. Mol. Bio.* 270:471-480

149. Rychlewski L, Zhang B., Godzik A. (1998) Fold and function prediction for Mycoplasma genitalium proteins, *Folding and Design*, 3 (4): 229-238

150. Rychlewski L, Jaroszewski L., Li W. and Godzik A. (2000) Comparison of sequence profiles. Strategies for structural predictions using sequence information. *Protein Science* , 9:232-241

151. Sadreyev R and Grishin N (2003) COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance. *J.Mol. Bio*. 326(1) 317-336

152. Saier MH Jr, Tran CV, Barabote RD. (2006), TCDB: the Transporter Classification Database for membrane transport protein analyses and information, *Nucl. Acids Res.*, 34: 181-6.

153. Saier MH Jr, Yen MR, Noto K, Tamang DG, Elkan C. (2009), The Transporter Classification Database: recent advances, *Nucl. Acids Res*., 37: D274-8.

154. Sasson O, Vaaknin A, Fleischer H, Portugaly E, Bilu Y, Linial N, Linial M. (2003). ProtoNet: hierarchical classification of the protein space. *Nucleic Acids Res.* 31, 348-352.

155. Sayle R.A and Milner-White E.G (1995) RasMol: Biomolecular graphics for all *Trends in Biochemical Sciences,* 20:374-376

156. Schaffer A. A., Aravind L.,Madden T. L., Shavirin S., Spouge J., Wolf Y. I., Koonin E. V. and Altschul S. F. (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Research,* 29(14): 2994-2005

157. Shafrir Y, Guy HR. (2004) STAM: simple transmembrane alignment method, *Bioinformatics*, 22; 20(5):758-69

158. Shrake A and Rupley JA. (1973) Environment and exposure to solvent of protein atoms. Lysozyme and insulin. *J Mol Biol. 79(2): 351*-71

159. Smith T.F and Waterman M.S (1981) Identification of common molecular subsequences. *J Mol Biol.* 147(1):195-7

160. Soding, J., Biegert, A. and Lupas, A. N. (2005) The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res,* 33, 244 – 248

161. Sonnhammer E.L.L, von Heijne G. and Krogh A. (1998) A hidden Markov model for predicting tarnsmembrane helices in protein sequences. *Intell. Syst. Mol. Biol.* 6, 175-182

162. Stevens T.J and Arkin I.T. (1999) Are membrane proteins "inside-out" proteins? *Proteins: Struct. Funct. Genet.* 36; 135-143

163. Strisovsky K., Sharpe H. A., and Freeman M (2009) Sequence-Specific Intramembrane Proteolysis: Identification of a Recognition Motif in Rhomboid, *Substrates Mol Cell.* 24; 36(6-2): 1048–1059

164. Sugiyama Y., Polulyakh N. and Shimizu T. (2003) IdentifIcation of transmembrane protein functions by binary topology, *Protein Engineering*, 16 (7): 479-488

165. Suwa M., Yudate T. H., Masuho Y. and Mitaku S. (2000) A novel measure characterized by a polar energy surface approxinatiob for recognition and classifcation of transmembrane proteins structures. *Proteins:Struct. Funct. Genet* 41:504-517

166. Suwa M., Sugihara M. and Yukiteru O. (2011) Functional and Structural Overview of G-Protein-Coupled Receptors Comprehensively Obtained from Genome Sequences, *Pharmaceuticals*, 4: 652-664

167. Suzek B.E, Huang H., McGarvey P., Mazumder R. and Wu C.H (2007) UniRef: comprehensive and non-redundant UniProt reference clusters, *Bioinformatics,* 23(10): 1282-1288

168. Taylor W. R., Jones D. T., and Green N.M (1994) A method for alpha-helical integral membrane protein fold prediction, *Proteins.,* 18: 281-294

169. Taylor PD, Attwood TK, Flower DR. (2003) BPROMPT: A consensus server for membrane protein prediction, *Nucleic Acids Res*,31:3698–700.

170. Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND, Koonin EV (2001.) The COG database: new developments in phylogenetic classification of proteins from complete genomes, *Nucleic Acids Res* 29:22–28;

171. Tusnady G.E. and Simon I., (1998) Principles governing amino acid composition of integral membrane proteins: application to topology prediction. *J. Mol.Bio.* 283:489-506

172. Tusnády GE, Dosztányi Z, Simon I. (2004). Transmembrane proteins in the Protein Data Bank: identification and classification, *Bioinformatics,* 20(17):2964-72.

173. Ursing, B. M., F. H. J. van Enckevort, J. A. M. Leunissen and R. J. Siezen (2002). EXProt: a database for proteins with an experimentally verified function, *Nucleic Acids Res*, 30(1):50-51.

174. Van Geest M. and Lolkema J.S (2000) Membrane topology and insertion of membrane proteins: search for topogenic signals Michriniol, *Mol.Biol. Rev.* 64:13-33

175. Viklund H and Elofsson A. (2004) Best Alpha-helical transmembrane protein topology predictions are achieved using hidden Markov models and evolutionary information, *Protein Sci.* 13:1908–17

176. Viklund, H., Granseth, E. and 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

177. Viklund H, Elofsson A. (2008) OCTOPUS: improving topology prediction by two-track ANN-based preference scores and an extended topological grammar. *Bioinformatics.* 1;24(15):1662-8.

178. von Heijne G. (1992) Membrane protein structure prediction, *J Mol. Bio.* 255:487-494

179. von Heijne G. (1996) Principles of membrane protein assembly and structure, *Prog.Biophys. Molec. Biology*, 66 (2): 113-139

180. von Heijne G (1999) Recent advances in the understanding of membranbe prtein assembly and structure, *Quarterly reviews of Biophysics,* 32, 4:285-307

181. von Heijne G. (2011) Introduction to theme "Membrane protein folding and insertion", *Annu. Rev. Biochem*, 80:157-160

182. Wallin E. , Tsukihara T., Yoshikawa S., von Heijne G and Elofsson A. (1997) Arhitecture of helix bundle membrane proteins: An analysis of cytochraome c oxidase from bovine mitochondria, *Protein Science* 6:808-815

183. Wallin E. and von Heijne G (1998) Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms, *Protein Science* 7:1029-1038

184. Wang C, Li S, Xi L, Liu H, Yao X. (2010) Accurate prediction of the burial status of transmembrane residues of α-helix membrane protein by incorporating the structural and physicochemical features, *Amino Acids* . Aug 26

185. Wetlaufer DB. (1973). Nucleation, rapid folding, and globular intrachain regions in proteins, *Proc Natl Acad Sci USA* 70 (3): 697−701.

186. Wodak, S.J., Janin, J (1981) Location of structural domains in proteins, *Biochemistry* 20: 6544-6552

187. White S.H. and Wimley W.C. (1999) Membrane protein folding and stability: physical principles, *Ann. Rev. Biophys. Struct*., 28:319-65

188. White S.H. (2001) How membranes shape protein structure, *J Biol Chem.*,31;276(35):32395-8

189. White, S.H. (2009) Biophysical dissection of membrane proteins. *Nature* 459, 344–346

190. Wistrand M., Kall L., and Sonnhammer L.L. E.(2006) A general model of G protein-coupled receptor sequences and its application to detect remote homologs, *Protein Sci.*, 15: 509-521

191. Yan C. and Luo J. (2010) An analysis of reentrant loops. *Protein J*., 29(5):350-4.

192. Yuan Z, Mattick JS, Teasdale RD. J (2004) SVMtm: support vector machines to predict transmembrane segments, *Comput Chem.* 25:632−6

193. Yuan Z, Zhang F, Davis MJ, Bodén M, Teasdale RD.(2006) Predicting the solvent accessibility of transmembrane residues from protein sequence, *J Proteome Res*., 5(5):1063-70.

194. Yona, G., Linial, N., and Linial, M. (2000). ProtoMap: automatic classification of protein sequences and hierarchy of protein families, *Nucleic Acids Res*. *28*, 49-55.

195. Yona G, Levitt M. (2002) Within the Twilight Zone: A Sensitive Profile-Profile Comparison Tool Based on Information Theory, *J Mol Biol*. 315:1257–1275.

196. Yohannan S, Faham S, Yang D, Whitelegge JP, Bowie JU (2004). The evolution of transmembrane helix kinks and the structural diversity of G protein-coupled receptors, *Proc. Natl. Acad. Sci. USA*, 101:959-63