Sequence-based predictions of membrane-protein topology, homology and insertion

Andreas Bernsel

Stockholm University
"Trying is the first step towards failure"

-- Homer
List of publications

Publications included in this thesis


* These authors contributed equally

Reprints were made with permission from the publishers.
Additional publications


Prediction servers

Two web servers have been developed that implement methods described in the papers:

**ΔG-pred**  Prediction of ΔG for TM-helix insertion  
http://www.cbr.su.se/DGpred/

**TOPCONS**  Consensus prediction of membrane protein topology  
http://topcons.cbr.su.se/
Membrane proteins comprise around 20-30% of a typical proteome and play crucial roles in a wide variety of biochemical pathways. Apart from their general biological significance, membrane proteins are of particular interest to the pharmaceutical industry, being targets for more than half of all available drugs. This thesis focuses on prediction methods for membrane proteins that ultimately rely on their amino acid sequence only.

By identifying soluble protein domains in membrane protein sequences, we were able to constrain and improve prediction of membrane protein topology, i.e. what parts of the sequence span the membrane and what parts are located on the cytoplasmic and extra-cytoplasmic sides. Using predicted topology as input to a profile-profile based alignment protocol, we managed to increase sensitivity to detect distant membrane protein homologs.

Finally, experimental measurements of the level of membrane integration of systematically designed transmembrane helices in vitro were used to derive a scale of position-specific contributions to helix insertion efficiency for all 20 naturally occurring amino acids. Notably, position within the helix was found to be an important factor for the contribution to helix insertion efficiency for polar and charged amino acids, reflecting the highly anisotropic environment of the membrane. Using the scale to predict natural transmembrane helices in protein sequences revealed that, whereas helices in single-spanning proteins are typically hydrophobic enough to insert by themselves, a large part of the helices in multi-spanning proteins seem to require stabilizing helix-helix interactions for proper membrane integration. Implementing the scale to predict full transmembrane topologies yielded results comparable to the best statistics-based topology prediction methods.
Contents

1. Introduction ........................................................................................................... 12

2. Biological membranes ....................................................................................... 15
   2.1 Membrane lipids ............................................................................................ 15
   2.2 Lipid bilayers .................................................................................................. 16

3. Membrane Proteins ............................................................................................... 18
   3.1 Types of membrane proteins ......................................................................... 18
      3.1.1 α-helical transmembrane proteins ......................................................... 19
      3.1.2 β-barrel transmembrane proteins ......................................................... 19
   3.2 Membrane protein targeting and translocation in the ER ................................ 20
   3.3 Amino acid statistics of TM-helices ............................................................... 21
      3.3.1 Helix-breaking residues ......................................................................... 21
      3.3.2 The aromatic belt .................................................................................... 22
      3.3.3 The positive-inside rule ......................................................................... 23
   3.4 Membrane protein topology ............................................................................ 23
      3.4.1 Topology mapping techniques ................................................................ 24
      3.4.2 Dual topology proteins ......................................................................... 24
   3.5 Membrane protein structure ............................................................................ 25
      3.5.1 Structure determination techniques ....................................................... 25
      3.5.2 Structural features of membrane proteins .............................................. 26
      3.5.3 Helix-helix interactions ......................................................................... 27

4. Computational approaches ................................................................................... 28
   4.1 Substitution matrices, profiles and HMMs ...................................................... 28
      4.1.1 Sequence alignment and substitution matrices ........................................ 28
      4.1.2 Sequence profiles ................................................................................... 29
      4.1.3 Hidden Markov Models ......................................................................... 30
   4.2 Artificial neural networks ............................................................................... 31
   4.3 Hydrophobicity scales .................................................................................... 32
   4.4 Transmembrane topology predictors ............................................................. 33
      4.4.1 TopPred .................................................................................................. 33
      4.4.2 TMHMM / PRO-TMHMM / PRODIV-TMHMM ...................................... 33
      4.4.3 HMMTOP ............................................................................................ 34
      4.4.4 MEMSAT ............................................................................................ 34
      4.4.5 PHDhtm ............................................................................................... 34
5. Present investigation

5.1 Improving transmembrane topology prediction by domain assignments (Paper I) 36
5.2 Using topology to predict homology (Paper II) 38
5.3 Deciphering the molecular code for transmembrane helix recognition (Paper III) 39
5.4 Prediction of membrane-protein topology from first principles (Paper IV) 44

Acknowledgements 46

References 48
Abbreviations

ANN  Artificial neural network
BLAST Basic local alignment search tool
DOPC Dioleoyl-phosphatidylcholine
ER  Endoplasmic reticulum
GFP  Green fluorescent protein
GPCR G-protein coupled receptor
GPI Glycosyl-phosphatidylinositol
HMM Hidden Markov model
MD Molecular dynamics
NMR Nuclear magnetic resonance
ORF Open reading frame
PDB Protein data bank
PhoA Alkaline phosphatase
PSI-BLAST Position specific iterative BLAST
PSSM Position specific scoring matrix
SP Signal peptide
SRP Signal recognition particle
TM Transmembrane

Amino acids

<table>
<thead>
<tr>
<th>Amino Acid</th>
<th>One Letter Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ala</td>
<td>A</td>
</tr>
<tr>
<td>Cys</td>
<td>C</td>
</tr>
<tr>
<td>Asp</td>
<td>D</td>
</tr>
<tr>
<td>Glu</td>
<td>E</td>
</tr>
<tr>
<td>Phe</td>
<td>F</td>
</tr>
<tr>
<td>Gly</td>
<td>G</td>
</tr>
<tr>
<td>His</td>
<td>H</td>
</tr>
<tr>
<td>Ile</td>
<td>I</td>
</tr>
<tr>
<td>Lys</td>
<td>K</td>
</tr>
<tr>
<td>Leu</td>
<td>L</td>
</tr>
<tr>
<td>Met</td>
<td>M</td>
</tr>
<tr>
<td>Asn</td>
<td>N</td>
</tr>
<tr>
<td>Asp</td>
<td>P</td>
</tr>
<tr>
<td>Glu</td>
<td>Q</td>
</tr>
<tr>
<td>Phe</td>
<td>R</td>
</tr>
<tr>
<td>Gly</td>
<td>S</td>
</tr>
<tr>
<td>His</td>
<td>T</td>
</tr>
<tr>
<td>Ile</td>
<td>V</td>
</tr>
<tr>
<td>Lys</td>
<td>W</td>
</tr>
<tr>
<td>Leu</td>
<td>Y</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Amino Acid</th>
<th>One Letter Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Methionine</td>
<td>M</td>
</tr>
<tr>
<td>Asparagine</td>
<td>N</td>
</tr>
<tr>
<td>Proline</td>
<td>P</td>
</tr>
<tr>
<td>Glutamine</td>
<td>Q</td>
</tr>
<tr>
<td>Arginine</td>
<td>R</td>
</tr>
<tr>
<td>Serine</td>
<td>S</td>
</tr>
<tr>
<td>Threonine</td>
<td>T</td>
</tr>
<tr>
<td>Valine</td>
<td>V</td>
</tr>
<tr>
<td>Tryptophan</td>
<td>W</td>
</tr>
<tr>
<td>Tyrosine</td>
<td>Y</td>
</tr>
</tbody>
</table>
1. Introduction

Life is a complex thing (Keyes 1966). During the short time a biological cell spends on earth before dividing itself into two new cells, or worse, committing suicide (Hengartner 2000), a myriad of biochemical reactions will have taken place, each of which is required for the development, survival and propagation of the cell and the genetic material it carries. Key players in this chemical factory of life are the proteins, catalyzing the specific conversion of substrates to products, regulating each others’ concentrations, transporting molecules across barriers, and converting light into sugar, among many other things. The specific function of a protein molecule is determined from its three-dimensional shape, which in turn is a function only of its amino acid sequence (Anfinsen 1973), and ultimately of its corresponding DNA sequence. In a more general sense, not focusing on a specific molecule, however, the function of a protein also depends on its concentration and localization in the cell.

In order to separate itself from the rest of the world, all living cells are equipped with a plasma membrane, the advantage being that the cell can modify its own internal conditions independently of, or sometimes in response to, what is going on in the external environment. Within eukaryotic cells, organelles are also delimited by at least one membrane, resulting in different compartments with disparate biochemical conditions. However, the isolation between a cell and its surroundings is not complete. The interior of the cell is only partially shielded from the exterior, and this balance is actively regulated by membrane spanning channels, transporters and receptors. Thus, in response to external changes, channels open and close, energy-driven pumps are activated and de-activated, and ligand-binding receptors give rise to signaling cascades that propagate through the cell. Indeed, the concept of regulated partial isolation is general, and applies not only to cells with respect to their environments, but also to organelles, organs, individuals and even species (Weiss 2005).

The amount of biological sequence data is currently increasing at an exponential rate (Overbeek 2000). To date, the sequences of more than 700 complete genomes are publicly available (http://www.ebi.ac.uk/GenomeReviews/stats/; August 2008), together comprising the result of one of the largest scientific endeavors in history. This rapid increase in the amount of biological information available has played a
major role in the origin of the field of bioinformatics, which deals with annotation, storage and analysis of such biological data.

Accordingly, one of the first disciplines within bioinformatics, and still one of the largest, has been that of sequence analysis. Methods for automatic annotation of new DNA and protein sequences are often based on the assumption that only the sequence is known, and try to predict various features such as DNA exons and introns, protein domains or subcellular location using only sequence as input. Thus, there is a certain rationale for developing prediction methods that are fully sequence based, not just because they do not require data from expensive biochemical experiments, but also for the prospect that the wealth of training data can only be expected to increase. Development of such sequence based prediction methods is only becoming increasingly important as the sequence collection continues to grow.

In particular, sequence based predictions on membrane spanning proteins are of great importance. Membrane proteins comprise about 25% of all proteins encoded by most genomes (Wallin and von Heijne 1998; Krogh et al. 2001) and are responsible for numerous vital processes in the cell, but apart from their general biological significance, there are also specific reasons for why additional effort is being put into membrane protein predictions. First, the medical importance of membrane bound receptors, channels and pumps as targets for drugs is well established. It has been estimated that more than 50% of all available drugs target membrane proteins (Drews 1996; Hopkins and Groom 2002). Second, structural information is hard to obtain experimentally, both because membrane proteins are particularly hard to overexpress and purify, but also since the large hydrophobic surfaces complicate the growth of crystals.

This thesis focuses on prediction methods for membrane proteins that ultimately rely on their amino acid sequence only. As a consequence of the peculiar surroundings in which integral membrane proteins find themselves, spanning the apolar hydrocarbon core and projecting into the polar water-phase on either side of the membrane, they are somewhat constrained in terms of what amino acid sequence is compatible with the different environments. This makes it possible to find sequence patterns and predict the topology of the protein, i.e. what parts of the sequence that actually span the membrane, and what parts are located on the cytoplasmic and extracytoplasmic side of the membrane. In paper I, we attempt to improve sequence-based topology prediction of membrane proteins by introducing additional information about the presence of extra-membranous protein domains in the sequence, which are believed to always reside on just one side of the membrane. In paper II, we instead use the sequence constraints placed by the hydrophobic interior of the membrane to improve detection of distant membrane protein homologs. Paper III and IV relate to the actual insertion of membrane proteins into the membrane, and how transmembrane helices are recognized in the cell. In paper III, a scale describing the energetics of
this process is developed based on experimental measurements, and in paper IV, this scale is implemented to predict transmembrane topologies.
2. Biological membranes

All living cells are surrounded by membranes, separating the interior of the cell from the surrounding environment and protecting it from foreign substances. In the case of eukaryotic cells, membranes also enclose subcellular compartments. The fundamental structural component of biological membranes is lipids, although a large fraction of the membrane mass comes from membrane proteins embedded within the lipids (Guidotti 1972).

2.1 Membrane lipids

Membrane lipids are amphiphilic, consisting of one or more hydrophobic hydrocarbon tails, and a hydrophilic sometimes charged headgroup. Depending on the relative sizes of the headgroup and the hydrophobic tails, lipids are more or less prone to form flat bilayer structures rather than micelles or other structures when placed in water, Fig. 2.1. When the headgroup is of approximately the same size as the hydrocarbon tails, the lipid becomes roughly cylindrical, which promotes bilayer formation, whereas if the headgroup is either smaller or larger than the hydrocarbon tails, the conical shape induces curvature in the lipid monolayer (Epand 1998).

Figure 2.1. Different shapes of membrane lipids. Cylindrically shaped lipids (left) stabilize a flat lipid bilayer structure. Conical lipids (middle) have large headgroups or skinny hydrocarbon tails. Inverted conical lipids (right) have small headgroups or bulky unsaturated hydrocarbon tails.
In addition to the variations in lipid shape, there are also differences in charge of the lipid headgroup, being neutral, zwitterionic or negatively charged. The lipids predominantly found in biological membranes are phospholipids and glycolipids, with the overall shape as in Fig. 2.1, and cholesterol, a neutral lipid containing a ring structure that breaks the tight packing of fatty acid chains, making the membrane more fluidic. In general, the headgroup region of a natural mix of lipids in a biological membrane tends to be slightly negative on average.

2.2 Lipid bilayers

As a consequence of their amphiphilic nature, lipids aggregate and spontaneously self-organize in water, generally into either micelles or bilayers. The biological membrane is a bilayer structure, where the hydrophobic fatty acid chains avoid contact with water molecules by arranging themselves into two oppositely oriented fluidic leaflets, Fig. 2.2. An early description of the system was the fluid mosaic model (Singer and Nicolson 1972), in which lipids and proteins diffuse laterally in the plane of the bilayer. The width of the membrane is determined by the lengths of the hydrocarbon tails, and is generally around 60Å, of which about half constitutes the hydrophobic core. However, the actual thickness, as well as fluidity and curvature, depends on...
the lipid composition which varies between different membranes, and local fluctuations may occur due to the presence of lipid rafts (Simons and Ikonen 1997).

The actual distributions of the functional groups in Fig. 2.2 across a DOPC lipid bilayer have been measured experimentally (Wiener and White 1992), and are shown in Fig. 2.3. From these distributions, which are quite consistent with the molecular dynamics simulation in Fig 2.2, the emerging picture is that thermal fluctuations lead to a highly heterogeneous environment, where the interface regions together occupy roughly the same width as the hydrophobic core of the membrane. Thus, the chemical environment across the membrane varies markedly over short distances, which is reflected in the amino acid distributions of TM helices at different depths into the bilayer (see section 3.3).

Lipids with conical shapes (Fig. 2.1; middle and right) might induce curvature when present in different concentrations in the two leaflets of a lipid bilayer. A membrane consisting primarily of phosphatidylcholine (cylindrical) and phosphatidylethanolamine (inverted conical) in both leaflets will experience curvature stress, meaning there will be higher lateral pressure in the lipid tail region than in the headgroup region. Such forces could be important for correct folding and functioning of integral membrane proteins (Dan and Safran 1998; Epand 1998).

**Figure 2.3** Structure of a DOPC bilayer, as determined by X-ray and neutron diffraction measurements. The distributions reflect the probability of finding a particular structural group at a specific location in the bilayer. Colors are as in Fig 2.2. Figure adapted from (Wiener and White 1992).
3. Membrane Proteins

3.1 Types of membrane proteins

For biological membranes, lipids are only half the story. The other half or so of the membrane mass comes from proteins (Guidotti 1972), interspersed among the lipids and contributing to the overall stability of the membrane, although the exact protein-lipid ratio varies between different membranes. Broadly, membrane proteins may be classified as peripheral or integral, and the integral membrane proteins can be further subdivided into monotopic, bitopic and polytopic ones (Blobel 1980). Peripheral membrane proteins are only loosely attached to the membrane through electrostatic or hydrophobic interactions with the lipid headgroups or other membrane proteins, or through a covalently attached GPI anchor (Chatterjee and Mayor 2001). Monotopic integral membrane proteins are largely water soluble, but also contain a hydrophobic part that partly penetrates into the membrane from

![Figure 3.1](image-url) Structural classes of transmembrane proteins. (Left) The α-helical transmembrane protein bacteriorhodopsin, found in the plasma membrane of archaea. (Right) The β-barrel transmembrane protein OmpA, found in the outer membrane of gram-negative bacteria. Meshes indicate the approximate borders of the hydrocarbon core of the respective membranes.
Bitopic integral membrane proteins span the membrane once and polytopic integral membrane proteins, finally, span the lipid bilayer multiple times. Polytopic proteins come in two basic architectures, either as bundles of tightly packed α-helices, or as closed barrels of amphipathic β-strands, Fig. 3.1.

3.1.1 α-helical transmembrane proteins
Because of the low-polarity environment in the hydrophobic core of the membrane, there is a particular need for the protein backbone to form internal hydrogen bonds, and one way of satisfying all backbone hydrogen bonds is to form α-helices in the membrane-spanning parts of the protein. The α-helical bundle type membrane proteins are the most abundant, and are predicted to constitute around 25% of a typical genome (Krogh et al. 2001). Mainly, the TM helices are composed of hydrophobic amino acids that are able to form van der Waals-interactions with fatty acids of the surrounding lipids. However, polar and even charged amino acids are occasionally also found deep into the bilayer, although generally not in direct contact with lipids, but rather buried against a protein surface. The TM-helices are typically roughly perpendicular to the membrane plane, and around 26 residues long (Ulmschneider et al. 2005), although local deformations of the bilayer and tilting of TM-helices, respectively, can accommodate lengths as short as 15 and as long as 43 residues (Granseth et al. 2005). α-helical transmembrane proteins are the most well-studied class of membrane proteins, and the only class investigated in this thesis, and will be referred to simply as ‘membrane proteins' throughout the text.

3.1.2 β-barrel transmembrane proteins
Another way of solving the problem to satisfy internally all backbone hydrogen bonds is seen in the β-barrel type of transmembrane proteins, in which the membrane spanning parts are composed of an even number of antiparallel β-strands, which are tilted around 45° relative to the membrane plane (Galdiero et al. 2007). Each strand is hydrogen bonding to the neighbouring strands, and the first and last strands hydrogen bond with each other to close the barrel (Fig. 3.1). Residues in the β-strands alternately point outwards, facing the lipids, and inwards, facing the inside of the barrel, resulting in a sequence pattern in which the residues are typically alternately polar and hydrophobic. β-barrel type membrane proteins have so far only been found in the outer membrane of gram-negative bacteria (Fischer et al. 1994), although predictions suggest they are also present in the outer membranes of chloroplasts and mitochondria (Schleiff et al. 2003). Although difficult to predict, they have been estimated to account for less than 3% of the ORFs in the genomes in which they are present (Garrow et al. 2005).
3.2 Membrane protein targeting and translocation in the ER

Most α-helical transmembrane proteins, and water-soluble proteins destined for the secretory pathway, are targeted by an N-terminal signal sequence to the ER membrane, where protein translation and translocation or membrane insertion occur simultaneously (Simon and Blobel 1991; White and von Heijne 2005; White 2007). Briefly, as the nascent polypeptide chain emerges from the ribosomal exit tunnel, an N-terminal hydrophobic segment of ~20 amino acids is recognized by a ribonucleoprotein complex known as the Signal Recognition Particle (SRP). The binding of SRP to the signal sequence temporarily halts elongation, and the SRP-ribosome-polypeptide complex, with mRNA also bound to it, is targeted to the SRP receptor protein at the ER membrane. SRP is released as the ribosome binds to a heterotrimeric channel complex known as the Sec translocon, whereupon translation is resumed and the nascent polypeptide chain is threaded directly through the channel as it is being synthesized, Fig. 3.2. The Sec translocon is responsible both for the membrane integration of membrane proteins and for the export of secreted proteins, first into the ER lumen and then, by means of vesicular transport, through the Golgi and out of the cell. Membrane spanning α-helices, as well as the hydrophobic signal peptide, are recognized by the translocon and transferred laterally into the lipid phase, whereas the hydrophilic luminal loops are translocated through the channel and the cytoplasmic loops are retained on the cytoplasmic side. For secreted proteins, the

Figure 3.2 Co-translational translocation and membrane insertion in the ER membrane. As the newly synthesized polypeptide chain emerges from the ribosome, the translocon channel must decide whether to insert it into the membrane or translocate it. Figure adapted from (White and von Heijne 2004).
signal peptide is cleaved off after translocation, releasing the folded protein on the lumenal side of the membrane. Membrane proteins either contain a cleavable signal peptide or an uncleaved signal sequence which then effectively becomes the first TM-helix of the mature protein.

How is the recognition of TM segments accomplished by the translocon channel? It has been suggested that the channel opens and closes rapidly as translocation occurs, thereby allowing the TM segments to partition into the lipid environment (Rapoport et al. 2004). Favourable interactions between lipids and amino acid side chains should then promote membrane integration rather than translocation. With the view of the membrane as a highly heterogeneous environment (Fig. 2.2) where chemical conditions change markedly over short distances, it follows that amino acid side chains will have different preferences for different slabs of the lipid bilayer. On the sequence level, the efficiency with which a TM segment is integrated by the translocon should then be sensitive to the relative positions of amino acids within the segment. This trend is evident when looking at amino acid distributions in known membrane protein structures.

3.3 Amino acid statistics of TM-helices

Although the most prominent feature of TM helices is their overall hydrophobicity, there is also a considerable variation in the preferences of amino acids along the membrane normal, as revealed in the known membrane protein structures. A recent statistical analysis of the frequencies of amino acids in TM helices shows a notable positional dependence, where hydrophobic residues (Ala, Ile, Val, Leu) mostly populate the hydrophobic core of the membrane, while polar (Asn, Gln) and polar aromatic (Trp, Tyr, His) residues avoid the core but are more abundant in the interface regions (Ulmschneider et al. 2005). The small polar residues Ser and Thr were found to be equally well represented in core as interface, which could possibly be explained by their ability to sometimes form hydrogen bonds with the backbone (Gray and Matthews 1984). Charged residues (Arg, Asp, Glu, Lys) show a sharp peak indicating that they are almost absent from the very center of the hydrophobic core, but seem to be well tolerated in the interface regions.

3.3.1 Helix-breaking residues

Pro plays a special role in membrane helices, in that it breaks the helix and induces a backbone kink that can be important for function or stabilization of the structure (von Heijne 1991; Sansom and Weinstein 2000; Cordes et al. 2002). It even seems that the kink-inducing Pro residue can subsequently be substituted, as long as the kink is stabilized by specific interactions with
surrounding residues (Yohannan et al. 2004). Pro is also overrepresented at TM-helix ends in the interface regions (Granseth et al. 2005).

In addition to Pro, Gly is also known to induce breaks in helices and promote turns, and also occurs with high frequency near TM-helix ends in the interface regions (Granseth et al. 2005). Despite its helix breaking nature, it is also quite abundant within the membrane. One reason for this is that its small side chain allows for two TM-helices to come close together, and it thus plays an important role in helix-helix packing (see section 3.5.3).

3.3.2 The aromatic belt

The polar aromatic residues Tyr and Trp are often found near the ends of TM-helices, sometimes referred to as the "aromatic belt", where they have been suggested to anchor and position the helix in the membrane (de Planque et al. 2003). These residues have a hydrophilic part as well as an apolar aromatic ring, and being positioned near the borders of the hydrocarbon core region, they can both interact with the lipid headgroups and at the same time bury the aromatic ring among the apolar lipid tails (Yau et al. 1998).

![Figure 3.3](image)

**Figure 3.3** Three-dimensional structure of the AQPM aquaporin water channel from *Methanothermobacter marburgensis*. The structure is colored according to hydrophobicity, where red is hydrophobic and white hydrophilic. Tyr and Trp residues are colored green and Lys and Arg residues are colored blue. Cytoplasm is downwards in the picture. The aromatic belts are evident, as is the overrepresentation of Lys and Arg residues on the cytoplasmic side of the membrane. Two re-entrant loops (section 3.5.2) are visible in the middle of the structure.
3.3.3 The positive-inside rule

A strong determinant for the topology and overall orientation of membrane proteins is the asymmetric distribution of positively charged amino acids (Lys and Arg) between loops on opposite sides of the membrane, being over-represented in short cytoplasmic loops (von Heijne 1992), known as the positive-inside rule. This charge imbalance was first discovered for signal peptides (von Heijne 1984), and was subsequently found to apply also to membrane helices (von Heijne and Gavel 1988). Site-directed mutagenesis studies have shown that the positive charges are sufficient to determine overall topology (von Heijne 1989; Nilsson and von Heijne 1990), and the rule seems to be universal across all types of membranes in a large number of species (Nilsson et al. 2005).

3.4 Membrane protein topology

The topology of a membrane protein describes what segments of the amino acid sequence span the membrane, and what segments protrude into the respective compartments on opposite sides of the membrane. Although this simple specification is clearly not as informative and accurate as the full three-dimensional structure of the protein, there are certain reasons why topology information can still be useful. First, even though full structure prediction is currently almost impossible to do with any reasonable reliability, if such methods are to arise in the future they will most probably have to start by making correct topology assumptions (von Heijne 2006). Thus, development of topology prediction algorithms has the prospect of paying off in the future, since the predicted information is so fundamental. Second, certain characteristics of membrane proteins can be deduced from topology information alone. Frequently, membrane protein superfamilies, e.g. the important 7TM receptors, have highly conserved topologies which can provide some information for detection of new members (Wistrand et al. 2006). Moreover, function can to a certain extent be predicted simply from the number of transmembrane helices in the protein (Sugiyama et al. 2003).

Transmembrane topology can either be mapped experimentally (section 3.4.1), or predicted from amino acid sequence (section 4.4). Attempts have also been made at combining limited experimental information with prediction techniques to provide higher quality topology models for a relatively large number of proteins (Daley et al. 2005; Kim et al. 2006). According to these studies, in both the Escherichia coli and Saccharomyces cerevisiae entire membrane proteomes, proteins with the C-terminus on the cytoplasmic side of the membrane were roughly four times as frequent as those with the C-terminus on the exoplasmic side. In addition, proteins with an even number of TM-helices were clearly over-represented, and consequently, the
N-terminus is also more commonly located on the cytoplasmic side of the membrane. This might indicate that helical hairpins are the basic blocks of membrane insertion by the Sec translocon. This is a debated subject, however, since the X-ray structure of the translocon reveals that the pore seems too small to accommodate more than one helix at a time (Van den Berg et al. 2004).

3.4.1 Topology mapping techniques

A number of techniques exist to experimentally map the locations of a few strategically selected parts of the sequence, in order to produce a topology model of the protein. One approach exploits recognition of sequence segments by agents that have access only to one side of the membrane, like glycosylation mapping (Chang et al. 1994), cysteine labeling (Kimura et al. 1997) or epitope mapping (Canfield and Levenson 1993). Another technique involves fusion of a reporter protein that is active only on one side of the membrane (Manoil and Beckwith 1986). By fusing the reporter protein to a number of differently truncated versions of the membrane protein, the in/out location of several points in the sequence can be derived and used to produce a full topology model. To avoid having to interpret the absence of activity of the reporter protein as a positive result, two reporter proteins with activity on opposite sides of the membrane are often used in combination. Reporter proteins frequently employed include PhoA (active in periplasm only) and GFP (active in cytoplasm only) (Drew et al. 2002; Daley et al. 2005).

3.4.2 Dual topology proteins

Although the large majority of membrane proteins have a well defined topology, recent studies have demonstrated the existence of a few membrane proteins that seem to insert with some probability in oppositely oriented directions, termed dual-topology membrane proteins (Rapp et al. 2006; Rapp et al. 2007). These proteins have in common a small or non-existent bias of Arg and Lys (KR-bias) to either side of the membrane (section 3.3.3). Interestingly, they sometimes have homologues with large KR-biases, and in those cases, two copies of the gene with opposite KR-biases are always present within the same genome, presumably inserting in opposite directions into the membrane. In one family, the two oppositely oriented copies were sometimes fused, such that both protein length and KR-bias were doubled. If both directional variants are required for function, this points towards a possible route of evolution for membrane proteins in which a dual topology protein is duplicated, whereupon the two copies evolve a KR-bias to stably insert in opposite directions, and may subsequently be fused into a single protein. Indeed, internal symmetry is a common phenomenon in known membrane protein structures (Choi et al. 2008).
3.5 Membrane protein structure

Membrane proteins are notoriously difficult to crystallize, which has led to the situation that, although about 25% of a typical genome are predicted to encode membrane proteins (Krogh et al. 2001), they only account for less than 1% of the known protein structures, with about 160 unique structures known (http://blanco.biomol.uci.edu/Membrane_Proteins_xtal.html; August 2008). Encouragingly though, the number of known membrane protein structures is growing exponentially, although at a rate that lags behind the corresponding rate for soluble proteins during the equivalent time period, Fig. 3.4.

![Figure 3.4](image.png)

**Figure 3.4** The number of known membrane protein structures grows exponentially, although at a lower rate than soluble proteins during the equivalent time period. Figure reproduced from (White 2004) with permission.

3.5.1 Structure determination techniques

Structure determination starts with protein overexpression, which is often difficult with membrane proteins, both because large amounts of the protein may be toxic to the cell, and because the proteins sometimes aggregate to form inclusion bodies (Drew et al. 2003). After solubilization with a detergent and purification, there are a number of alternative methods to determine the three-dimensional structure. The majority of currently known membrane protein structures have been solved by X-ray crystallography. The main obstacle using this method is that growth of well-ordered diffracting crystals is difficult due to the amphiphilicity of the protein. To overcome this, membrane proteins are sometimes fused to soluble domains, increasing the polar
Apart from X-ray crystallography, the other main method in use for obtaining structural information of membrane proteins is NMR spectroscopy. Here, no crystals are required, but on the other hand, the method is limited to quite small proteins (Torres et al. 2003). This can be particularly problematic for membrane proteins, since they are typically quite large by themselves (Mitaku and Hirokawa 1999), and need to be solubilized in detergent micelles, adding to the size of the complex (Torres et al. 2003).

Although membrane proteins are reluctant to arrange themselves into three-dimensional crystals, they more easily pack into two-dimensional planar structures, which can be analysed by electron microscopy (EM). A three-dimensional model of the protein can be produced by combining two-dimensional projections at different angles (Henderson et al. 1990). Although the method gives relatively low resolution compared to the 3D methods, higher resolution can be achieved with cryo-EM, where the sample is studied at temperatures approaching -260°C (Torres et al. 2003). Cryo-EM has been used to solve a number of near-atomic membrane protein structures, including aquaporin (Murata et al. 2000) and the acetylcholine receptor (Unwin 2005).

3.5.2 Structural features of membrane proteins

Incidentally, the first membrane protein structures to be solved were relatively simple in their structures (Deisenhofer et al. 1985; Henderson et al. 1990), with straight hydrophobic helices penetrating the membrane at angles roughly perpendicular to the membrane plane and interconnected by short loops. With increasing insight into membrane protein structures, the picture is now somewhat more complex, and certain recurring structural features not evident from just a membrane topology representation have emerged.

One obvious observation is that not all helices are straight rods, but a few contain Pro-induced kinks (section 3.3.1), which might be important for protein function. Whereas Pro is quite rare in α-helices of soluble proteins, they occur with a relatively high frequency in transmembrane α-helices, where they seem to play an important role in the biological activity of the peptide (Cordes et al. 2002).

Another recurring structural element in membrane proteins is the re-entrant loop, which penetrates only halfway through the membrane, and enters and exits on the same side (Lasso et al. 2006; Viklund et al. 2006). Commonly, such loops contain residues important for function of the protein (Lasso et al. 2006), and they are found in, e.g., the translocon channel (Van den Berg et al. 2004), the aquaporin water conducting channel (Fig. 3.3 and Harries et al. 2004), and the chloride ion channel (Dutzler et al. 2003).
In addition to the transmembrane α-helices perpendicular to the membrane plane, some structures also contain shorter interfacial helices, lying roughly parallel to the membrane. The function of these helices is not entirely clear, but they might be involved in positioning of the transmembrane helices (Granseth et al. 2005). They are on average 9 residues long and rich in Trp and Tyr residues (Granseth et al. 2005).

Finally, preferences of amino acid side chains can be studied from the structures. Lys and Arg have long hydrophobic side chains with a positive charge at the very end, and when these residues are situated within the hydrocarbon core of the membrane, they often point their side chain away from the membrane core and out towards the hydrophilic interface region, a phenomenon described as 'snorkelling' (Chamberlain et al. 2004). Similarly, Phe residues tend to direct its hydrophobic aromatic ring towards the membrane core, termed 'anti-snorkelling'.

3.5.3 Helix-helix interactions

Whereas for globular proteins, the hydrophobic effect is the main driving force of folding (Honig and Yang 1995; Lins and Brasseur 1995), the circumstances are somewhat different for membrane proteins, being embedded in the hydrophobic environment of the membrane. For a long time, hydrogen bonding has been considered the major stabilizing force of membrane protein structure through inter-helical interactions (Engelman 1982; Choma et al. 2000; Zhou et al. 2000), but recently, this view has been questioned (Joh et al. 2008). A dominant force of helix packing in the hydrocarbon core might be van der Waals interactions between helices (White and Wimley 1999). Two small residues (Gly, Ala, Ser) separated by three residues will end up on the same side of the α-helix, and thus allow two helices to come close together and pack tightly. The GX₃G-motif has been implicated as the main driving force of homodimerization of the single-spanning Glycophorin A (Lemmon et al. 1992; Treutlein et al. 1992), and both this motif and the GX₃GX₃G- and GX₇G-motifs are overrepresented in TM-helices in general (Liu et al. 2002; Kim et al. 2005).

The assembly of membrane protein chains into complexes can be better understood by considering the thermodynamics of the surrounding membrane lipids. Even though the dimerization of two TM chains into a rigid arrangement seems to reduce the entropy of the proteins, this is counterbalanced and outweighed by the entropy increase resulting from the lipids being released back into the lipid pool of the membrane as the total protein-lipid surface area decreases (Helms 2002).
4. Computational approaches

4.1 Substitution matrices, profiles and HMMs

In principle, all characteristics of a newly synthesized protein, including targeting to different subcellular components, folding of secondary and tertiary structure, complex assembly, function and ultimately degradation, are determined by its amino acid sequence (Anfinsen 1973). One of the most important and challenging areas of bioinformatics is thus the decoding of the biological signals intrinsic in the sequence, and prediction of biochemical properties from this source of information. Among the many different types of signals that can be recognized, evolutionary relationships and classification of proteins into superfamilies and families are among the most important.

4.1.1 Sequence alignment and substitution matrices

Given two protein sequences, the alignment problem can be described as finding the most probable set of point mutations, insertions and deletions defining the evolutionary divergence of the two proteins. In the simplest case, each amino acid substitution is associated with a certain cost, and a dynamic programming algorithm, such as the Smith-Waterman (Smith and Waterman 1981) or Needleman-Wunsch algorithm (Needleman and Wunsch 1970), is employed to search for the set of substitutions that together sum to the lowest total cost for the alignment. Insertions and deletions can be included in the model by also associating appropriate costs with gap opening and gap insertion respectively. Some different versions of the 20x20 amino acid substitution matrix, describing the costs for aligning all amino acids to each other, are regularly used in bioinformatics applications, of which the most popular are the PAM (Dayhoff 1978) and BLOSUM (Henikoff and Henikoff 1992) set of matrices respectively. Incidentally, the BLOSUM matrices were recently shown to contain errors due to a software source code bug, although the 'erroneous' matrices actually perform better than the 'intended' ones (Styczynski et al. 2008).

Dynamic programming algorithms are generally too slow for searching large sequence databases, but heuristic algorithms exist that are significantly
faster at the cost of being slightly less sensitive. BLAST (Altschul et al. 1990) first searches sequences for high-scoring 'words', and the hits can subsequently be extended using dynamic programming, and FASTA (Pearson 1990) uses a similar approach.

### 4.1.2 Sequence profiles

By aligning two sequences using a single substitution matrix, it is impossible to model the very common situation where substitution rates are not constant over the whole sequence, but rather vary according to the relative importance between different parts of the protein. Structurally and functionally important regions in a protein structure should be expected to exhibit higher level of conservation than e.g. flexible loops far from the active site. In order to capture this effect, sequence profiles or Position Specific Scoring Matrices (PSSMs) (Altschul et al. 1997) are commonly used to represent common motifs in protein families. In a sequence profile, each position of the sequence motif or sequence family contains a 20x1-vector of probabilities for observing each of the 20 amino acids in that particular position, as given by a multiple sequence alignment (Fig. 4.1).

**Figure 4.1** (Above) Multiple sequence alignment of a Zinc finger domain. (Below) A sequence profile for the first ten residues of the same alignment. The profile is a compact representation of the multiple sequence alignment. Zeros in the profile are omitted for clarity. A number of conserved residues can be easily detected.
When evaluating the alignment of a sequence to a sequence profile, in principle the probability for the aligned amino acid in the single sequence is given by the corresponding frequency vector in the profile, and these probabilities are multiplied to give the total probability of the full alignment. To avoid the very small accumulated probabilities that can result from long protein sequences, in practice these calculations are always performed in log-space.

It is well established that including evolutionary information in the form of profiles improves homology detection between protein sequences (Lindahl and Elofsson 2000). By aligning two sequence profiles against each other, evolutionary information is included for both query and database sequences, which has been shown to improve remote homology detection even further (Park et al. 1998; Rychlewski et al. 2000; Mittelman et al. 2003; Ohlson et al. 2004).

### 4.1.3 Hidden Markov Models

Hidden Markov Models (HMMs) are probabilistic models that were early applied to the problem of speech recognition (Jelinek et al. 1975; Rabiner 1989), and have later on been used in a number of biological sequence analysis problems (Churchill 1989; Krogh et al. 1994). The model consists of a set of interconnected states, each of which contains a probability distribution, called emission probabilities, over a set of symbols, referred to as the alphabet of the model. In addition, each state also contains a probability distribution over possible transitions to all states in the model (including itself), known as transition probabilities. According to the parameters of the model, the emission and transition probabilities, the model generates a sequence of symbols from the alphabet by first emitting a symbol according to the emission probability distribution of the current state, then transits to a new state according to the transition probability distribution and repeats this until a certain stop state is reached. By tuning the parameters of the model, it can be designed to generate a set of sequences with high probability.

There are basically three different problems associated with HMMs. First, the probability for the model to generate a certain sequence of symbols can be calculated, known as the evaluation problem. The forward and backward algorithms (Rabiner 1989) are used to solve this problem. Second, the most likely state path through the model can be calculated for a certain symbol sequence, known as the decoding problem. This problem is solved by the Viterbi algorithm (Viterbi 1967). Finally, the estimation problem relates to finding the most likely model parameters given a set of symbol sequences, and the Baum-Welch algorithm (Baum 1972) is used for this task.

Profile HMMs (Eddy et al. 1995; Eddy 1998) are a certain kind of HMM, that has achieved widespread use in biological sequence analysis applications. These models resemble sequence profiles (section 4.1.2) in that they
are used to represent common characteristics of a whole protein family. States in profile HMMs are sequentially arranged and connected so that positions in a multiple sequence alignment are represented by one Match state (M), one Insert state (I) and one Delete state (D) (Fig. 4.2). Each Match and Insert state in the model contains both amino acid emission probabilities, in analogy with the probability vector in ordinary sequence profiles, and transition probabilities, whereas the Delete states are 'silent states' representing gaps, and do not emit symbols. A profile HMM can easily be constructed from a multiple sequence alignment by setting emission and transition probabilities according to the observed amino acids and gaps at different positions in the alignment.

4.2 Artificial neural networks

Artificial neural networks (ANNs) (Brian and Hjort 1995) is a standard machine learning technique for pattern recognition and classification, that has been applied to many problems in bioinformatics, including TM topology prediction. The model consists of an interconnected set of nodes (neurons), each taking one or several signals as input. The output from a node is frequently a nonlinear function of a weighted sum of its input values, where hyperbolic tangent functions are particularly common. In the simplest case, the nodes are interconnected in a directed fashion, being arranged into layers, where each layer of nodes is only connected to the neighboring layers and signals through the network are propagated from left to right, called a feed-forward network. The output from the network effectively becomes a 'function of functions' corresponding to a non-linear mapping of a set of input variables to a set of output variables.
ANNs have been successfully applied to e.g. prediction of secondary structure (Qian and Sejnowski 1988), prediction of subcellular location (Bendtsen et al. 2004) and protein homology detection (Frishman and Argos 1992). In TM topology prediction, ANNs are sometimes used to predict a residue preference score, that is subsequently used as input to a dynamic programming algorithm (Rost et al. 1996; Viklund and Elofsson 2008).

### 4.3 Hydrophobicity scales

In order to search for hydrophobicity signals in membrane protein sequences, a prerequisite is a scale assigning hydrophobicity values for all amino acids. Numerous such hydrophobicity scales are available, based on e.g. the partitioning of amino acids between two immiscible liquid phases, chromatographic techniques or accessible surface area calculations. A few of the available scales are outlined below, but many more exist.

One of the most frequently cited hydrophobicity scales, the Kyte-Doolittle scale (Kyte and Doolittle 1982), combines accessible surface area measurements in globular proteins with water-vapor partitioning preferences. By implementing this scale in a sliding-window approach, it was possible both to distinguish exterior from interior in globular proteins, and to identify TM regions in membrane proteins.

Another hydrophobicity scale specifically tailored to TM helices is the Goldman-Engelman-Steitz (GES) scale (Engelman et al. 1986). Here, a semi-theoretical approach is taken, accounting for the attachment of side chains to an α-helical backbone structure. This scale is used to predict TM helices in the TopPred topology prediction algorithm (section 4.4.1).

The Wimley-White scale (White and Wimley 1999) takes into account the (unfavorable) contribution of partitioning the backbone peptide bonds into the bilayer. Partitioning of pentapeptides into POPC bilayer interfaces and n-octanol, respectively, were used to construct this scale.

Finally, the Zhao-London scale (Zhao and London 2006) is based on propensities of amino acids in known membrane protein structures. The scale was refined to separate well between transmembrane and soluble sequences from protein sequence databases, and should in that sense be quite suited to the problem of predicting TM helices.
4.4 Transmembrane topology predictors

Several methods exist that, given an amino acid sequence, predict the full transmembrane topology. These methods are either based on physical properties of the amino acids, such as the hydrophobicity scales described above, or rely on statistical over- or underrepresentation of amino acids in transmembrane helices, as observed from known topologies of membrane proteins. TopPred (section 4.4.1) is the only method of the ones outlined below that is based directly on physical principles (i.e. a hydrophobicity scale) for predicting TM-helices.

4.4.1 TopPred

One of the first topology predictors, TopPred (von Heijne 1992), takes advantage of the GES hydrophobicity scale (section 4.3), and creates a hydrophobicity plot of the sequence, using a sliding window approach. First, all hydrophobicity peaks above a certain cutoff value are identified and marked as 'certain' TM segments, whereas all peaks below this cutoff but above a second lower cutoff value are marked 'putative' TM helices. Second, all possible topologies, including all certain TM segments and either including or excluding each of the putative TM segments are generated, and the topology that best complies with the positive-inside rule (section 3.3.3) is chosen as the final prediction.

4.4.2 TMHMM / PRO-TMHMM / PRODIV-TMHMM

Among the first methods to implement HMMs to predict TM topology, and one of the most frequently used, is TMHMM (Krogh et al. 2001). Common to all HMM based topology predictors is that all states carry a label, and paths through the model correspond to a labeling defining the topology of the protein. The architecture of the model is cyclic, such that two loops with a membrane region in between will always reside on opposite sides of the membrane. The N-best algorithm (Krogh 1997) is used to find the most probable labeling (the most probable topology) of the input sequence, which may correspond to more than one state path through the model.

In PRO-TMHMM (Viklund and Elofsson 2004), a development of TMHMM, sequence profiles rather than single sequences can be scored against the model. This was found to improve prediction performance by a few percentage points. In the same study, another method, PRODIV-TMHMM, was found to improve performance even further. This method uses a re-optimization procedure which was first introduced with HMMTOP (see below).
4.4.3 HMMTOP

Another early method to use HMMs for TM topology prediction is HMMTOP (Tusnady and Simon 2001). The architecture of the model is cyclic, just like TMHMM, but differs in some respects such as the upper and lower boundaries for helix lengths. HMMTOP can take either a single sequence or a sequence profile as input. One notable feature of the method is that the model is re-optimized on every query sequence or sequence profile after the first scoring, and then scored a second time with the new parameters. This has the effect that hydrophobicity relative to the rest of the sequence determines the locations of TM regions, rather than absolute hydrophobicity.

4.4.4 MEMSAT

The very first method to combine all sequence signals and calculate a globally best topology according to the model was MEMSAT (Jones et al. 1994). Five different states (inside loop, outside loop, helix inside, helix outside and helix middle) each contain an amino acid distribution, and dynamic programming is used to align the query sequence to the model. One important difference between this approach and the later HMM-based methods is that the model is not cyclic. Instead, all possible topologies are explored, assuming a linear model with an increasing number of TM-helices, starting from just one. Later versions of MEMSAT (Jones 2007) integrates ANNs with HMMs, and can take sequence profiles as input.

4.4.5 PHDhtm

Another ANN-based approach for TM topology prediction is the PHDhtm method, which is part of a larger package, PHD (Rost et al. 1996), for predicting secondary structure of proteins. The method takes a sequence profile as input and predicts a preference score for each profile column based on a sliding window of neighboring columns. A topology is then generated using a dynamic programming algorithm, and the overall orientation is determined by the 'positive-inside rule'.

4.4.6 Phobius / PolyPhobius

A common problem in TM topology prediction is the confusion of cleaved signal peptides (SPs) and TM regions, since both have an overall hydrophobic character. The first TM topology prediction method to address this problem was the HMM-based method Phobius (Käll et al. 2004), in which TM-regions and SPs are modeled separately. The method is particularly efficient in datasets containing a mixture of proteins with and without TM-regions,
and with and without SPs. A later version of the method, PolyPhobius (Käll et al. 2005), can take a sequence profile as input.
5. Present investigation

5.1 Improving transmembrane topology prediction by domain assignments (Paper I)

Topology prediction of membrane proteins is typically based on sequence statistics, reflecting amino acid preferences for different compartments. In particular, three sequence signals are especially important (Fig. 3.3): the hydrophobicity of membrane spanning segments, the disposition of Trp and Tyr residues to the membrane-interface border (section 3.3.2), and the over-representation of Arg and Lys residues in short cytoplasmic loops (section 3.3.3). In this paper, we explored the possibility that the presence of globular protein domains in membrane protein sequences can be used to constrain topology prediction methods and thereby improve prediction performance.

Protein domains are often compartment specific, and information about domain occurrence has previously been used to predict subcellular location of soluble proteins (Mott et al. 2002). Moreover, although covalent combinations between transmembrane domains (i.e. protein domains with membrane spanning regions) are rare, covalent combinations between one transmembrane domain and one soluble domain are observed frequently (Liu et al. 2004). Taken together, these two observations suggest that soluble domains with compartment specific localization, when found in membrane protein sequences, can be used to constrain that part of the sequence to one side of the membrane, before a sequence-based method is used to predict the topology of the rest of the sequence. This is the basic idea explored in the paper.

Related to this idea are earlier studies, where experimentally determined 'anchor points' were used to constrain topology predictions and provide improved topology models (section 3.4) (Daley et al. 2005; Kim et al. 2006). An important difference to these studies is that, here, no experiments are necessary, but rather one type of prediction (topology prediction) is improved by another type of prediction (domain assignments).

From the SMART domain database (Letunic et al. 2004), 367 domains carrying an annotation about subcellular location (146 cytoplasmic, 221 extracellular) were extracted. Searching for these domains in two test sets, one small set of membrane proteins with known topologies, and one large set of
membrane proteins with predicted topologies, gave the results shown in Fig. 5.1. Briefly, the fraction of sequences containing at least one of the compartment-specific domains was around 10% in both sets, but the fraction domain hits in conflict with the topology was much higher in the large set with predicted topologies, than in the small set with known topologies. These conflicts are the cases where our approach can be used to constrain predictions and thereby improve topology models.

We found that the domains were overrepresented in single-spanning proteins compared to the whole data set. Single-spanning proteins are often mispredicted by topology predictors, mostly due to an inversion such that the TM-segment is correctly located but the overall orientation is wrong. Large extra-membranous domains carry little or no information in other predictors, and our approach thus solves a major weakness in these methods.

After this paper was published, two similar studies have also employed the strategy of using the occurrence of compartment-specific soluble protein domains in membrane protein sequences to improve topology prediction. In LocaloDom (Lee et al. 2006), experimentally verified Swiss-Prot annotations (Boutet et al. 2007) were used to classify Pfam domains (Bateman et al. 2004) as being cytoplasmic or extracellular. In addition, conserved topologies of transmembrane Pfam domains were identified using the Phobius topology prediction algorithm (Käll et al. 2004), and subsequently used to infer TM topologies of sequences in which the Pfam domain was found. TOPDOM (Tusnady et al. 2008) is also based on Swiss-Prot annotations, but also includes predicted information from Swiss-Prot, which makes the domain collection much larger. For the most part, these three studies agree on...
the subcellular location of domains, although there are also a few exceptions (Tusnady et al. 2008).

5.2 Using topology to predict homology (Paper II)

In this study, we introduced a new profile-profile based alignment method for remote homology detection of membrane proteins. Compared with globular proteins, membrane proteins are surrounded by a more intricate environment and, consequently, amino acid substitution rates in the membrane spanning regions differ from those of globular proteins and hydrophilic loops in TM proteins. Since existing algorithms for homology detection are often developed with globular proteins in mind, they may not be optimal to detect remote membrane protein homologs. In this study, we take advantage of the sequence constraints placed by the hydrophobic core of the membrane to detect distant homology relationships between membrane proteins. The approach is to include information about predicted topology in the alignment, and thereby promote alignments where membrane spanning regions are aligned against each other.

Although for globular proteins, it is well established that homology detection can be improved by including information from secondary structure predictions (Fischer and Eisenberg 1996; Rice and Eisenberg 1997; Hargbo and Elofsson 1999), prior to this study, only a few attempts had been made to utilize the sequence constraints specific to membrane proteins. In one study (Ng et al. 2000), TM regions are clustered to calculate an amino acid substitution matrix specifically designed to TM proteins. In another study (Hedman et al. 2002), topology predictions were used to promote alignments where TM regions are aligned against each other in a profile-sequence based approach.

By including evolutionary information for both query and target sequences, homology detection of globular proteins can be substantially improved (Park et al. 1998; Rychlewski et al. 2000; Mittelman et al. 2003; Ohlson et al. 2004), and this has recently been shown to apply also to membrane proteins (Forrest et al. 2006).

In this paper, we presented the (to our knowledge) first method to combine the membrane protein specific sequence constraints with the strength of profile-profile based methods aiming to detect remote members of membrane protein families. Starting from a query sequence, a two-track profile HMM (section 4.1.3) is constructed, combining sequence information with predicted topology. This HMM is then searched against a database of sequence profiles, also containing predicted topologies, and the HMM emission probabilities are tuned such that alignment between TM regions in the two proteins is favored.
We first applied the method to remote homology detection within the GPCRDB (Horn et al. 2003) database of G-protein coupled receptors. Here, it was apparent both that the profile-profile based methods performed clearly better than the profile-sequence based ones, and that the additional topology information improved prediction results. The advantage of adding topology information was less clear in two other test sets, the OPM database (Lomize et al. 2006) and the HOMEP database (Forrest et al. 2006), although the profile-profile based methods performed consistently better also here. An example of an improved alignment between two GPCRs is shown in Fig. 5.2.

![Figure 5.2](alignment.png)

**Figure 5.2** Alignment between distant homologs BAR1_SCHCO and Q752Q1_ASHGO (both GPCRs), using HMMs with one (only sequence) and two (sequence and topology) alphabets, respectively. Adding topology information shifts the alignment such that the correct TM-regions are aligned against each other. At the same time, the score for the alignment becomes positive and the overlap of TM residues between the two proteins (TM-overlap) is increased.

### 5.3 Deciphering the molecular code for transmembrane helix recognition (Paper III)

Because of the heterogeneous environment in lipid bilayers (section 2.2), amino acid propensities in TM proteins vary with the distance to the center of the hydrocarbon core of the membrane (section 3.3). As secreted or TM
proteins are being synthesized by the ribosome, they are directly threaded through the translocon channel (section 3.2), where membrane spanning segments are recognized and inserted into the bilayer, possibly by the rapid opening and closing of a lateral gate (Rapoport et al. 2004), allowing the segment to partition into the lipid environment if this is energetically favorable. In this paper, we have investigated the 'molecular code' for the recognition of TM helices by the Sec translocon, describing the requirements in terms of amino acid composition for a sequence segment to be recognized as a TM helix and inserted into the membrane, as opposed to being translocated across. An earlier study from our lab (Hessa et al. 2005) laid the basis for the experiments (described below), and derived a 'biological hydrophobicity scale' for the contributions to membrane insertion of each amino acid when placed in the middle of a 19 residue segment.

The experimental data underlying the derivation of the 'code' comes from insertion efficiency measurements of designed putative TM segments in a model membrane protein, Fig. 5.3. Briefly, systematically designed test segments were introduced into the model protein, which was expressed and inserted into rough microsomal membranes in an in vitro translation system. The test segment is flanked by two glycosylation sites and, since glycosylation only occurs on the lumenal side, the inserted and translocated forms of the protein can be separated by size on a polyacrylamide gel. From the relative amounts of the two forms, an apparent free energy of membrane insertion of the segment, \( \Delta G_{\text{app}} \), can be calculated.

Almost 500 designed and natural sequences were expressed and tested for membrane integration in this way. To quantify the contributions, \( \Delta G_{\text{app}}^{\text{nat}} \), from individual amino acids to \( \Delta G_{\text{app}} \), an additive model was assumed in

![Figure 5.3](https://example.com/figure5.png)

**Figure 5.3** Insertion efficiency measurements of systematically designed test segments (red), engineered into a model membrane protein with two TM helices (black). The G2 site is only glycosylated in the translocated form (right) and thus the two forms (left/right) can be size separated on a gel. The relative amounts of the two forms quantify the level of insertion.
which the position, \(i\), of the amino acid within the segment was taken into account:

\[
\Delta G_{\text{app}}^{\text{pred}} = \sum_{i=1}^{l} \Delta G_{\text{app}}^{\text{ag(i)}} + c_0 \mu + c_1 l + c_2 l^2
\]

(1)

where \(l\) is the length of the segment, \(\mu\) is the hydrophobic moment of the helix, \(c_0\) is a weight parameter for the hydrophobic moment, and \(c_1, c_2\) and \(c_3\) are parameters describing the contribution from the helix length. The position-specific amino acid contributions, \(\Delta G_{\text{app}}^{\text{ag(i)}}\), were optimized by iteratively minimizing the squared differences between predicted \(\Delta G_{\text{app}}^{\text{pred}}\) values and measured \(\Delta G_{\text{app}}\) values, using a standard non-linear optimization strategy. To reduce the number of parameters of the model, the contribution from a particular amino acid \(\Delta G_{\text{app}}^{\text{ag(i)}}\) as a function of position, \(i\), was first expressed as a simple gaussian function with two parameters, except for Trp and Tyr, which were both modeled with double gaussian functions containing five parameters each. Using this representation, the model contained in

**Figure 5.4** Position specific amino acid contributions to membrane insertion efficiency of TM helices. Position 0 corresponds to the center of the membrane. (Blue) Curves derived from experimental data as described in Fig. 5.3. (Red) Statistical curves from high-resolution membrane protein structures.
total 50 parameters, optimized using more than 400 constructs. The results from the optimization are shown in Fig. 5.4. Overall, the curves correspond well to statistical curves, obtained from the relative over- or underrepresentation of amino acids at different distances to the membrane center in known membrane protein structures. Charged residues (D, E, K, R) give highly position-specific contributions to membrane insertion, and are much better tolerated towards the interface regions than inside the hydrophobic core. Contributions from polar residues (H, N, Q, S, T) have the same general trend as the charged ones, but are less pronounced. Hydrophobic residues (A, F, I, L, M, V) give negative (favorable) contributions and are less position-specific. Finally, the polar aromatic residues (W, Y) are favorable for membrane insertion when located at a certain distance from the membrane center, but unfavorable in the middle.

To see how well the model would identify TM helices in natural proteins, we collected four test sets of secreted proteins, cytoplasmic proteins, single-spanning membrane proteins, and TM helices from multi-spanning membrane proteins and identified, in each set, the segment with lowest \( \Delta G^{\text{prod}} \).

Figure 5.5 \( \Delta G^{\text{prod}} \) distributions in different sets of proteins. A large fraction of the multi-spanning helices seem to be only inefficiently recognized by themselves, and might rely on stabilizing interactions for membrane integration.

To see how well the model would identify TM helices in natural proteins, we collected four test sets of secreted proteins, cytoplasmic proteins, single-spanning membrane proteins, and TM helices from multi-spanning membrane proteins and identified, in each set, the segment with lowest \( \Delta G^{\text{prod}} \), Fig. 5.5. The overlap between the secreted and single-spanning distribution is small, and they cross close to the zero-point defined by the experimental setup. A relatively large fraction of helices in multi-spanning proteins have
\[ \Delta G_{\text{app}}^{\text{red}} > 0, \] indicating that such segments would not be efficiently integrated by themselves, and might rely on stabilizing interactions from neighboring helices.

Molecular dynamics simulations of lipid-embedded TM-helices containing charged residues (Dorairaj and Allen 2007) have resulted in free energies considerably more unfavorable than suggested by the 'biological hydrophobicity scale' first presented in (Hessa et al. 2005) and further developed in this paper. Concerns raised about the experimental setup involve the possibility of interactions between the H-segment and TM1 or TM2 (Fig. 5.3) (Shental-Bechor et al. 2006) and the potential for the H-segment to slide away from its intended position such that the charged residue ends up being closer to the interfacial region (Dorairaj and Allen 2007). However, control experiments have shown that the identity of TM1 and TM2 have little influence on the efficiency with which the H-segment is incorporated (Meindl-Beinker et al. 2006). In addition, series of constructs where two polar or charged residues are scanned symmetrically from the middle towards the ends of the segment agree well with the additive model, arguing against sliding of the H-segment as being a crucial factor (Hessa et al. 2007). When interpreting the experimental results, a few things should be kept in mind. First, the position of an amino acid within the H-segment (Fig. 5.4) refers to the C\(_{\alpha}\)-carbon, meaning that polar and charged residues, especially Arg and Lys, might be expected to point their side chains away from the immediate center of the hydrocarbon core. Second, a discrepancy to the MD-simulations is that biological membranes consist to a large part of proteins (Guidotti 1972), possibly influencing the polarity of the membrane environment compared to pure lipid bilayers. Third, charged residues in the membrane may induce local deformations of the bilayer and pull water into the membrane (Johansson and Lindahl 2006). Potentially, a combination of several explanations will eventually make it possible to reconcile the MD data with the experimental results.

A recent study (Park and Helms 2008) derived a statistical position-specific free-energy scale for the contributions of all amino acids, analogous to the statistical curves in Fig. 5.4. Using this scale to predict TM segments yielded results comparable to the best available TM topology prediction methods. However, scanning globular proteins, this scale did not show any difference between cytoplasmic and secreted proteins, as the one suggested in Fig. 5.5.
5.4 Prediction of membrane-protein topology from first principles (Paper IV)

In an attempt to predict full transmembrane topologies using the model developed in paper III, we implemented the position-specific $\Delta G_{\text{app}}^{\text{out}(i)}$-scale in two prediction algorithms. Topology prediction algorithms (section 4.4) are typically based on machine learning methods that extract statistical signals from known membrane protein structures, thereby optimizing a large number of parameters. An important difference to those machine learning approaches is that our methods are only based on data from experiments as in Fig. 5.3, and should be more directly linked to the physico-chemical principles underlying TM helix recognition in vivo.

One method, TopPred$\Delta G$, basically employs the original TopPred algorithm (section 4.4.1), but the position independent GES hydrophobicity scale is replaced by the position dependent $\Delta G_{\text{app}}^{\text{out}(i)}$-scale. Two parameters, the upper and lower $\Delta G_{\text{app}}^{\text{pred}}$-cutoffs, corresponding to putative and certain TM-helices respectively, needed to be optimized on known structures. A variant of the algorithm that takes a sequence profile rather than a single sequence as input was also developed, which improved performance by a few percentage points.

The $\Delta G_{\text{app}}^{\text{out}(i)}$-scale was also implemented in an HMM-based method, SCAMPI, designed to be as simple as possible, such that only the $\Delta G_{\text{app}}^{\text{out}(i)}$-scale and no other parameters would influence the predicted topology. As such, all transition probabilities throughout the model are set equal to 1, and emission probabilities in the loop regions of the model are flat distributions and emit all amino acids with equal probability. In the membrane region of the model, the $\Delta G_{\text{app}}^{\text{out}(i)}$-scale is used to calculate the free energy of insertion, and this energy is converted to the corresponding expected fraction of insertion in an experiment as in Fig. 5.3, which is used as emission probability in the model. Only two parameters in SCAMPI needed to be optimized on known structures. First, since many TM-helices are predicted not to be hydrophobic enough to insert by themselves (cf. Fig. 5.5), a constant correction factor is subtracted from the predicted $\Delta G_{\text{app}}^{\text{pred}}$ before this value is converted to a corresponding probability. Second, since the $\Delta G_{\text{app}}^{\text{out}(i)}$-scale does not contain any orientational information, the emission probabilities for Lys and Arg are increased by a constant value in the part of inside loops closest to the membrane, in order to comply with the positive-inside rule (section 3.3.3).

For both TopPred$\Delta G$ and SCAMPI, the two parameters were optimized on one dataset, and used for prediction on another independent dataset. An example prediction is shown in Fig. 5.6. Somewhat unexpectedly, our simple methods based on experimental measurements of artificial TM helices were found to match the performance of the best available statistics based methods, predicting 80-85% correct topologies in our benchmark on all known membrane protein structures from PDB. We further analyzed the TM helices
in the dataset with respect to their $\Delta G_{app}^{\text{pred}}$ values, and found that helices that were highly buried in the structure typically had higher $\Delta G_{app}^{\text{pred}}$ values than helices that were exposed towards the lipid phase, and that TM-helices in proteins with many TM-helices tend to have higher $\Delta G_{app}^{\text{pred}}$ values on average. These observations support the hypothesis that helix-helix interactions might be an important factor for incorporation of marginally hydrophobic helices into the membrane.

Figure 5.6 Topology predictions of Cytochrome c oxidase (PDB code 1ehkA) by TopPred$^{\Delta G}$ and SCAMPI. (Top) Crystal structure of the protein chain. (Middle) the $\Delta G_{app}^{\text{pred}}$ profile, with the TopPred$^{\Delta G}$ parameters $\Delta G_{\text{high}}^{\Delta G}$ and $\Delta G_{\text{low}}^{\Delta G}$ shown as dashed lines and the topology predicted by TopPred$^{\Delta G}$ indicated beneath the curve. (Bottom) The path through the SCAMPI model, with the vertical position in the model (see y axis) plotted as a function of sequence position. The structure and predicted topologies are colored according to $\Delta G_{app}^{\text{pred}}$ values. Locations of the TM helices as annotated in the OPM database (Lomize et al. 2006) are shown as gray boxes. Lys and Arg residues are marked with plus signs.
Acknowledgements

Lots of people have contributed, helped and encouraged me during the work of this thesis. Thanks to all of you. I would like to especially thank the following people:

My supervisor, Gunnar, for giving me the opportunity to work in your lab. Thanks for always knowing everything, and for explaining it in a way that the rest of us can understand. Thanks for giving me the freedom to try my own ideas and, when they fail, for having better suggestions. And thanks for always answering emails within 5 minutes, irrespective of time or the timezone you're in.

My co-supervisor, Arne, for creating a great, relaxed atmosphere at work. Thanks for having more ideas than the rest of us taken together, and for enthusiastic explanations of them. And thanks for all the espresso.

Past and present room-mates: Kristoffer, for julmys and for watering the flowers, Håkan, for all the science we learnt in Aspen, Tomas, for the amazing chip-tricks, Diana, for everything I know about the importance of conquering Kas.

Erik Sj., the system guru, for your unique dart-throwing technique and for gathering people for the Thursday pub.

SBC/CBR-people: Erik L., for your positive and encouraging attitude and for the PS3, Åsa, for baking amazing cake-club cakes, Per, for having the only working version of pymol, Erik G., for voting for me in the karaoke-contest, Jenny, for teaching me how to spin the ball in wii-bowling, Björn, for being so reasonable with Zlatans morsa, Joppe, for choosing the sun-exposed part of the table in the "hand-on-hot-table" contest in Brazil, Anna, for never occupying more than necessary of the cluster, Pär, for nerve-wrecking dart rounds, Aron, for taking care of the left-overs, Ali, for never bluffing, Örjan, for all the sparris, Linnea, Rauan, Marcin, Anni, Yana, Christine, Katarina for being great colleagues.

Mia for all the asian lunches.
My Biosapiens work package colleagues: **Rita**, for sightseeing and shopping in Barcelona, **Gert**, for teaching me about GPCRs and for buying me lunch.

My dear family: **Mamma** and **Pappa**, for life, for being so kind, generous and supportive, and for everything else, **Morfar**, for all your encouragement and interest, **Johanna** and **Thomas**, for long phone calls and for the sun deck at Lusten, **Alma** and **Adam**, for never saying no to play with me, **Fredrik** and **Laura**, for fine wines and private cinema shows, **Bianca**, for being cute and for forcing your parents to buy a bigger car, **Flicka**, for always being happy.

My Sweetheart, **Kirsi**, for all your love and support, and for all the good times we've had together. You are my sunshine.
References


