Development of Computational Methods for Predicting Structural Characteristics of Helical Membrane Proteins

Dissertation

zur Erlangung des Grades
des Doktors der Naturwissenschaften
der Naturwissenschaftlich-Technischen Fakultät III
Chemie, Pharmazie, Bio- und Werkstoffwissenschaften
der Universität des Saarlandes

vorgelegt von

Yungki Park

Saarbrücken, June 2007

Tag des Kolloquiums: 2007 November 15

Dekan: Prof. Dr. Uli Müller

Berichterstatter: Prof. Dr. Volkhard Helms
Prof. Dr. Thomas Lengauer

Eidesstattliche Erklärung

Hiermit versichere ich an Eides statt, dass ich die vorliegende Arbeit selbständig und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus anderen Quellen oder indirekt übernommenen Daten und Konzepte sind unter Angabe der Quelle gekennzeichnet.

Die Arbeit wurde bisher weder im In- noch im Ausland in gleicher oder ähnlicher Form in einem Verfahren zur Erlangung eines akademischen Grades vorgelegt.

Saarbrücken, June 2007

Yungki Park

Contents

1. Zusammenfassung

Membranproteine (MPs) spielen in diversen zellulären Prozessen eine wichtige Rolle. Die Mehrheit der MPs bestehen aus einem Bündel von Transmembran (TM) Helices, die die Bezeichnung helikale Membranproteine (HMPs) tragen. Derzeit schätzt man, dass etwa 20 - 30% der Leseraster im sequenzierten Genom HMPs kodieren. Ungeachtet ihrer allgemeinen Bedeutung in der Biologie und ihrer großen Ansammlung im Genom sind bis jetzt lediglich etwa 50 dreidimensionale HMP Strukturen bekannt. Dem gegenüber stehen mehr als 1000 bekannte Strukturen von verschiedenen Faltungsmuster von wasserlöslichen Proteinen. Erkenntnisse zu den molekularen Mechanismen der HMPs können dadurch schwer gesammelt werden. Diese ungünstige Situation hängt zum einen stark mit der hohen Schwierigkeit zusammen, qualitativ hochwertige Kristalle dieser Proteine für Röntgenstrahl-Analysen zu generieren, und zum anderen mit der Tatsache, dass HMPs durch ihre meist enorme Größe für die Strukturaufklärung mittels NMR Spektroskopie ungeeignet sind. Computerunterstützte Methoden zur Vorhersage struktureller Eigenschaften der HMPs basierend auf der Aminosäure-Sequenz könnten in diesem Sachverhalt von großem Interesse sein. Zusätzlich könnten computerunterstützte Analysen interessante Aspekte ihrer Struktur und Funktion aufzeigen, die experimentelle Studien nicht aufzeigen können.

Diese Arbeit fasst jahrelange Anstrengungen zusammen, die sich auf 4 publizierte Artikel (Artikel I - IV) aufteilen. In Artikel I wurde der Versuch unternommen, mittelmäßig aufgelöste HMP-Strukturen mit einer geringen Zahl von TM Helices durch Packungsregeln und konservierte Sequenzmotive zu modellieren. In Artikel II wurde eine grundlegende Untersuchung durchgeführt, um den Grad der Korrelation zwischen exponierten Motiven in den TM Helices zur Doppelmembranschicht und ihren Eigenschaften wie Hydrophobizität und konservierten Motiven zu analysieren. Darauf aufbauend wurde in Artikel III ein optimaler Weg zur Generierung von Skalen vorgestellt, die Paarungspräferenzen der 20 Aminosäuren mit der Doppelmembranschicht auf der Basis von bekannten HMP-Strukturen zu bewerten. Die Ergebnisse zeigen überraschenderweise, dass das architektonische Prinzip von HMPs am besten durch die partiellen spezifischen Volumina der Aminosäuren beschrieben werden kann. Artikel IV präsentiert TMX (TransMembrane eXposure), eine neue computerunterstützte Methode zur Vorhersage der Lipidzugänglichkeit der TM-Aminosäuren in HMPs, die bisherige Methoden deutlich an Genauigkeit übertrifft. Unter http://service.bioinformatik.uni-saarland.de/tmx ist eine Web-Schnittstelle für TMX aufrufbar.

Die Ergebnisse dieser Arbeit dienen als solides Sprungbrett für weitere interessante Studien, denen die Gruppe von Prof. Volkhard Helms nachgehen wird. Schließlich wird experimentell arbeitenden Biologen ein umfangreicher Satz an Methoden für Rechnungen und Vorhersagen zur Verfügung gestellt, der ihnen helfen wird, biochemische und biophysikalische experimentelle Daten zu rationalisieren, und anschliessende Experimente zu entwerfen. Zusätzlich zu diesen ergänzenden Aufgaben können komplexere computerunterstützte Analysen unser Verständnis über Struktur und Funktion von HMPs ausweiten und vertiefen, so wie es in den Artikeln II und III bereits diskutiert wurde.

2. Abstract in English

Helical membrane proteins (HMPs) play a crucial role in diverse cellular processes. Given the difficulty in determining their structures by experimental techniques, it is desired to develop computational methods for predicting their structural characteristics. In addition, computational analysis can provide interesting insights into their structure and function that experimental work can not provide. This thesis summarizes years of such computational endeavours, comprising 4 published papers (Paper I ~ IV). In Paper I, it was attempted to model low-resolution tertiary structures of HMPs with a modest number of transmembrane (TM) helices from packing constraints and sequence conservation patterns. In Paper II, a fundamental investigation was undertaken to analyze the degree of correlation between exposure patterns of TM helices to the membrane and their properties such as their hydrophobicities and conservation patterns. In Paper III, on the basis of the work presented in Paper II, an optimal way of deriving the propensity scales of the 20 amino acids to preferentially interact with the membrane as reflected in known HMP structures was presented, which revealed a surprising fact that the architectural principle of HMPs is best captured by the partial specific volumes of the amino acids. In Paper IV, the development of TMX (TransMembrane eXposure), a novel computational method for predicting the lipid accessibility of TM residues of HMPs, was described, which significantly outperforms other existing methods. A web interface for TMX is available at http://service.bioinformatik.uni-saarland.de/tmx.

3. Abstract in German

Helikale Membranproteine (HMPs) spielen in diversen zellulären Prozessen eine bedeutende Rolle. In Anbetracht der Schwierigkeit, die Struktur dieser Proteine mittels experimenteller Techniken aufzuklären, ist es erstrebenswert, computerunterstützte Methoden für ihre Strukturaufklärung zu entwickeln. Zusätzlich könnten computerunterstützte Analysen interessante Aspekte ihrer Struktur und Funktion aufzeigen, die experimentelle Studien nicht aufzeigen können. Meine Doktorarbeit fasst jahrelange Anstrengungen zusammen, die sich auf 4 publizierte Artikel (Artikel I ~ IV) aufteilen. In Artikel I wurde der Versuch unternommen, mittelmäßig aufgelöste HMP-Strukturen mit einer geringen Zahl von Transmembranprotein (TM) Helices mit Hilfe von Packungsregeln und konservierten Sequenzmotiven zu modellieren. In Artikel II wurde eine grundlegende Untersuchung durchgeführt, um den Grad der Korrelation zwischen exponierten Motiven in den TM Helices zur Doppelmembranschicht und ihren Eigenschaften wie Hydrophobizität und konservierten Motiven zu analysieren. Darauf aufbauend wurde in Artikel III ein optimaler Weg zur Generierung von Skalen vorgestellt, die Paarungspräferenzen der 20 Aminosäuren mit der Doppelmembranschicht auf Basis von bekannten HMP Strukturen zu bewerten. Die Ergebnisse zeigen überraschenderweise, dass das architektonische Prinzip von HMPs am besten durch die partiellen spezifischen Volumina der Aminosäuren beschrieben werden kann. Artikel IV präsentiert TMX (TransMembrane eXposure), eine neue computerunterstützte Methode zur Vorhersage der Lipidzugänglichkeit der TM-Aminosäuren in HMPs, die bisherige Methoden deutlich an Genauigkeit übertrifft. Unter http://service.bioinformatik.uni-saarland.de/tmx ist eine Web-Schnittstelle für TMX aufrufbar.

4. Membrane proteins

4.1 Introduction to membrane proteins

Integral membrane proteins (denoted hereafter as MPs) permanently reside in biological membranes for their functions. Due to the peculiar chemical character of biological membranes (compared to aqueous media where water-soluble proteins reside), MPs display structural characteristics distinct from those of water-soluble proteins [1]. The most notable is that only two structural types are observed in known MP structures. The first type is of a helix-bundle type whereas the other is of a β-barrel type. It is currently estimated that helix-bundle type MPs account for 20 ~ 30% of open reading frames of sequenced genomes [2]. Furthermore, all the MPs that are drug targets are of a helix-bundle type. In contrast, β-barrel MPs have been restricted to the outer membrane of Gram-negative bacteria, mitochondria and chloroplasts [3]. Thus, the helix-bundle type is overwhelmingly predominant and more important from a biological viewpoint. For these reasons, we concentrate on helix-bundle type MPs (denoted hereafter as HMPs) in this work.

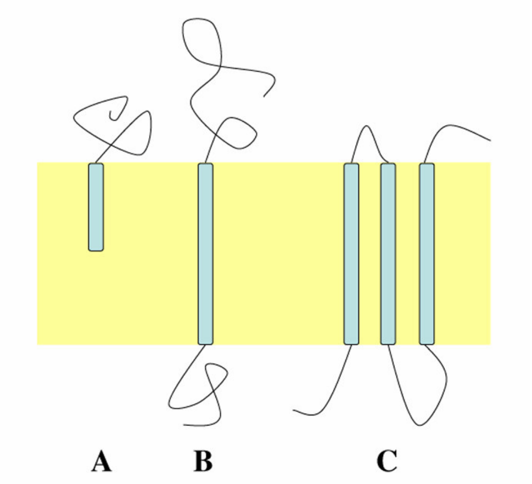


Figure 1. Schematic representation of three types of HMPs. The yellow slab represents a biological membrane and the light blue sticks helical segments that penetrate into the membrane. A: HMPs which span the membrane only halfway (monotopic HMPs). B: HMPs which span the entire membrane only once (bitopic HMPs). C: HMPs which span the entire membrane more than once (polytopic HMPs).

Depending on the way HMPs are embedded in the membrane, HMPs can be classified into three groups as shown in Fig. 1. Few cases are known for monotopic HMPs, and most studies on HMPs have been on bitopic and polytopic HMPs.

HMPs are involved in diverse cellular processes, which can be briefly summarized as follows:

- Active transport of solutes across membranes: Active transport means transport against concentration gradients and thus requires an energy source. Membrane transporters are classified into two types depending on which energy source they use for active transports. The first type is those that utilize the energy released upon ATP hydrolysis, including P-type adenosine triphosphatases (ATPase) [4] and ABC transporters [5, 6]. The second type is those that couple active transport of one type of solutes with downhill transport of another type of solutes, including the well-known major facilitator superfamily (MFS) transporters [7, 8]. Several crystal structures of membrane transporters have become available over the last few years, revealing the molecular basis for active transport across the membrane (Fig. 2).
- Channels across membranes: Unlike transporters, channels usually mediate passive transport. However, like transporters, channels perform highly selective transport (e.g. water channels transport only H_2O, not H_3O^+, in spite of their similarity). Ion channels are of significant

importance in nervous systems because they propagate action potentials. Moreover, ion channels play a crucial role in the homeostasis of most cells. In 2003, the Nobel Prize in Chemistry was awarded to Roderick MacKinnon for his work on potassium channel [9] and Peter Agre for his work on water channel [10] (Fig. 3).

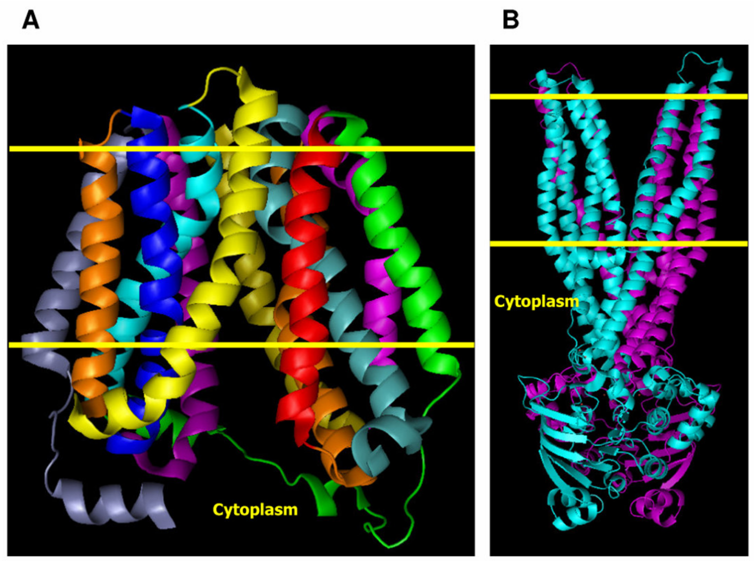


Figure 2. Crystal structures of membrane transporters. Yellow lines indicate putative membrane boundaries. A: Lactose permease of *E. coli* (PDB ID: 1PV7) [7], a member of the MFS transporters. It consists of 12 TM helices with a long loop on the cytoplasmic side connecting the 6 TM helices in the N terminus and the 6 TM helices in the C terminus. It couples the downhill transport of proton into the cell with the uphill uptake of lactose into the cell, i.e. a symporter of proton and lactose. B: Multidrug transport of the ABC family from *S. aureus* (PDB ID: 2HYD) [5]. The ATP binding domain is found in the cytoplasmic side. It is a dimeric structure, each consisting of 6 TM helices, and is responsible for the efflux of diverse cytotoxic substances from the inside, thus being of clinical importance.

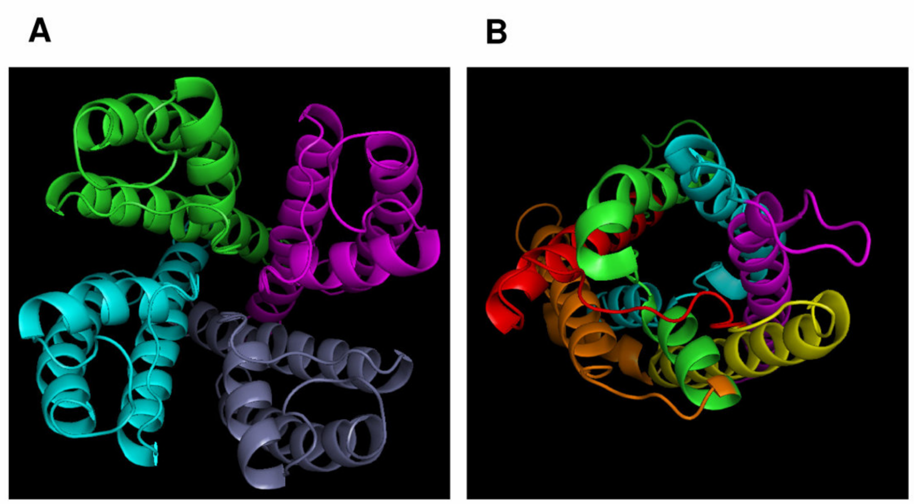


Figure 3. Crystal structures of membrane channels. A: Potassium channel from *S. lividans* (PDB ID: 1BL8) [9], as viewed from the exoplasmic side. It is a tetrameric protein, and the junction of the 4 monomers forms the pore through which K^+ is transported. It is remarkable

to note that this channel nearly completely blocks the transport of Na^+, even though the van der Waals radii of K^+ and Na^+ are quite similar (K^+: 275 pm and Na^+: 227 pm) and there exist channels permitting the transport of both K^+ and Na^+. B: Bovine water channel (PDB ID: 1J4N) [11], as viewed from the exoplasmic side. It is a monomeric protein with 6 TM helices and 2 helices penetrating into the membrane only halfway, one from the cytoplasmic side and the other from the exoplasmic side.

- Receptors: Cells must be able to respond to external signals properly for survival. The primary loci of signal reception are receptor proteins in membranes. Of many receptors, the most prominent example would be G-protein coupled receptors (GPCRs) [12]. Genes encoding this huge family are estimated to occupy ~ 5% of the worm genome and ~ 3% of the human genome. GPCRs recognize and relay external signals to G-proteins in the cytoplasm, which then activate downstream signal transduction pathways. It is remarkable that they can process external signals in a vast array of diverse forms such as electromagnetic waves, small molecules and macromolecules (Fig. 4).

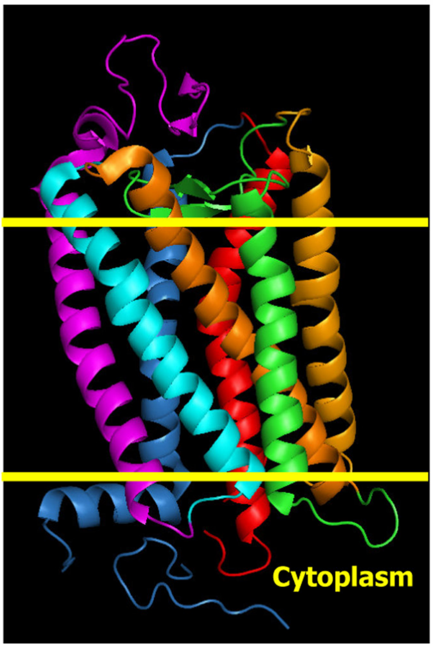


Figure 4. Crystal structure of bovine rhodopsin (PDB ID: 1F88) [12], a member of GPCRs. Yellow lines demarcate putative membrane boundaries. This long-awaited crystal structure confirms the existence of 7 TM helices. Upon light activation, the bound retinal (not shown here) undergoes a conformational change, which is then thought to induce a conformational change of the protein. The protein's conformational change is then further propagated to G proteins docked onto the cytoplasmic side of the protein.

- Energy generation: All key proteins in the energy generation process, e.g. oxidative phosphorylation and photosynthesis, are HMPs. They usually incorporate cofactors and mediate oxidation/reduction of substrates. In fact, the first high-resolution HMP structure was that of bacterial photosynthetic reaction center in 1984 by Johann Deisenhofer, Robert Huber and Hartmut Michel [13]. It was not only the first HMP crystal structure, but also the most complex molecular structure that had been solved by X-ray crystallography up to that point in time. For that achievement, they shared the Nobel Prize in Chemistry in 1988 (Fig. 5).
- Other functions: HMPs play an important role in signal processing by metabolizing lipid molecules, as in the case of Membrane-Associated Proteins in Eicosanoid and Gluthathione Metabolism (MAPEG) family members [14]. Also, it has become increasingly clear that intramembrane proteolytic processes mediated by HMPs represent critical steps in intracellular signal transductions [15].

In addition, several pathological roles have been ascribed to HMPs. Most notable would be the TM glycine zipper motif of HMPs implicated in Alzheimer's disease and prion disease [16]. In this regard, it is worth noting that two of the most widely prescribed drugs in the world – fluoxetine (Prozac) and omeprazole (Prilosec) – target membrane transporters.

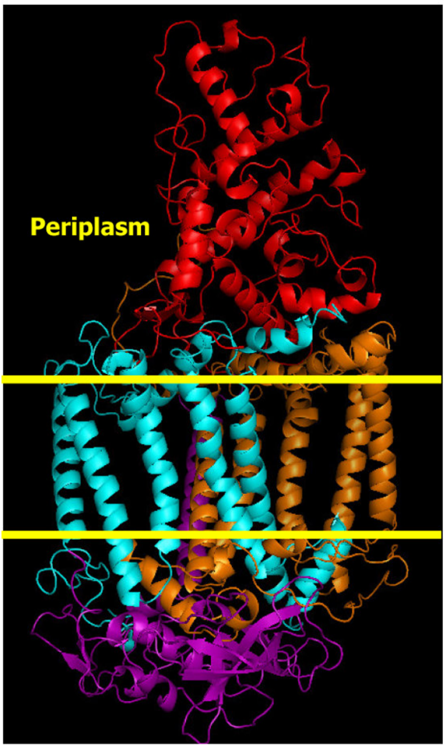


Figure 5. Crystal structure of photosynthetic reaction center from *R. viridis* (PDB ID: 1PRC) [13]. As shown, it consists of 4 subunits. Each of the two TM subunits contains 5 TM helices. Cytochrome *c* (shown in red) and the two TM subunits contain all the necessary pigments (not shown here) for photosynthesis. In summary, light energy absorbed by chlorophylls is used to eject electrons, which flow through cytochrome bc_1 complex to return to the reaction center. Cytochrome bc_1 complex generates a proton gradient during this flow of electrons, which is then used to drive the synthesis of ATP.

4.2 Biogenesis of HMPs

Water-soluble proteins usually fold on their own upon being released from the ribosome. The biogenesis of HMPs is a bit more complicated due to the additional need for them to be embedded in biological membranes. The endoplasmic reticulum (ER) membrane is the main place where biogenesis of HMPs takes place.

Upon emerging from the ribosome, the signal sequences of HMPs are recognized by the signal recognition particle (SRP), a ribonucleo-protein complex, forming a ribosome nascent chain (RNC)-SRP complex [17, 18]. In eukaryotes, this complex formation causes a temporary arrest of the nascent chain elongation in the ribosome. The signal sequences are highly degenerate: they are stretches of 7 ~ 25 amino acids with hydrophobic ones overrepresented [19]. Signal sequences can be classified into 3 groups according to the mode of translocation and the cleavability [20] (see below). The degenerate nature of signal sequences and yet their exquisite recognition by the SRP are remarkably contrasting. The RNC-SRP complex is then targeted to the SRP receptor (SR) embedded in the ER membrane. Both SRP and SR contain GTPase domains. Upon the RNC-SRP complex being targeted to the SR, the GTPase domains of SRP and SR form a dual active site where the O3´ of one GTP becomes a key component of the active site of the other GTPase. This duality allows a reciprocal activation of GTP hydrolysis by both GTPase domains, which is thought to be coupled to and thus to govern docking of the RNC-SRP complex onto the translocon [18].

The translocon is a heterotrimeric complex of HMPs, called the Sec61 complex in eukaryotes and the SecY complex in eubacteria and archaea [21]. The α subunits (Sec61α in mammals, Sec61p in *S. cerevisiae*, SecY in bacteria and archaea) contain 10 TM helices, forming a pore through which secretory proteins cross the membrane and TM segments get shunted laterally to the surrounding membrane milieu (Fig. 6). The 10 helices can be related to each other by a pseudo-symmetry through a two-fold rotation axis in the membrane plane. However, the symmetric nature of the two halves is not obvious from the primary structure. The γ subunits contain 1 TM helix and 1 helix lying on the cytoplasmic surface of the membrane, clamping the two halves of the α subunit. The β subunits contain 1 helix and are not essential for the function of the translocon.

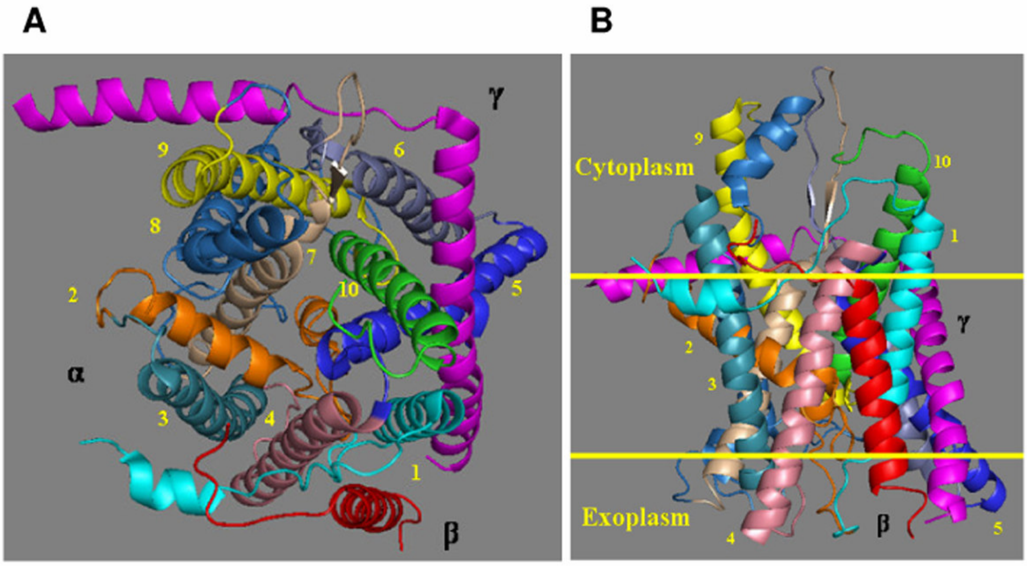


Figure 6. Crystal structure of the archaeal translocon (PDB ID: 1RHZ). The panel A shows the view from the cytosol while the panel B shows the view from the membrane phase. The three subunits and the 10 TM helices of the alpha subunit are labelled. Since the 5 TM helices in the N terminus can be related to the 5 TM helices in the C terminus by a pseudo symmetry through the membrane plane, the overall structure can be viewed as a clamshell. In contrast to the relatively smooth exoplasmic side (panel B), the cytoplasmic side of the translocon is featured by many protrusions, which might be related to the docking of the RNC-SRP complex on this side.

The translocon essentially works as a switching station. It lets secretory proteins cross the ER membrane while it shunts TM helices of HMPs laterally into the ER membrane via its lateral exit (Fig. 7). This dual mode of channelling, along with the diversity of substrates it handles, makes the translocon distinct from other channels. It has been controversial whether the translocon works as a monomer or as a homo-oligomer [21-23]. Recent crystal structures of archaeal translocon (depicted in Fig. 6) suggested that single translocon molecules form a functional channel [21], and this suggestion was recently corroborated by biochemical experiments [23].

Upon the RNC-SRP complex being docked onto the translocon complex, the ribosome resumes the elongation of the nascent chain, and the SRP gets dissociated from the complex, returning to the cytosolic pool of SRPs. Since the translocon complex provides a passive conduit, the driving force for the translocation process is provided by the elongation reaction taking place in the ribosome [17, 18].

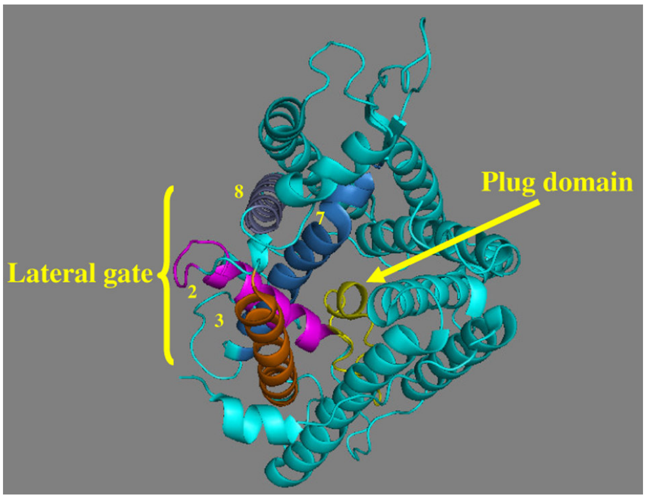


Figure 7. The view of the translocon from the cytosol. In the closed state, the pore of the translocon is blocked by the plug domain. Once a translocation substrate or a TM segment enters the pore region, the plug domain is thought to move away by a rigid-body swing motion, opening the pore. The cleft between TM2-TM3 and TM7-TM8, which can be likened to the mouth of the clamshell, is believed to function as a lateral gate for the membrane insertion of TM segments.

4.2.1 Topogenesis of HMPs

Once the ribosome docked onto the translocon resumes the elongation process of the nascent chain, a series of events unique to the biogenesis of HMPs and quite complicated begin to take place. It is well worthwhile to go over the topology of HMPs [20, 24] before we look into these poorly-understood complicated events. This concept is unique to HMPs since they are essentially constrained in the 2D space of the membrane plane. What is meant by "topology" here is the boundaries of TM segments in the primary structure and the location of intervening loops as well as N and C termini with respect to the membrane plane. For example, if a protein with three TM segments places its N terminus in the cytoplasmic (in) side and its C terminus in the exoplasmic (out) side, its overall topology is referred to as N_{in}–C_{out}. A simple example would make clear why this concept is so important for HMPs. Lactose permease (LacY, Fig. 2A) is an HMP in the inner membrane of *E. coli* responsible for importing lactose from the outside to inside of the cell. As with other membrane transporters, the transport of lactose by LacY happens only in one direction, i.e. in the direction from the outside to inside of the cell. LacY displays an N_{in}–C_{in} topology. Hence, LacY of an N_{out}–C_{out} topology would, if ever possible, transport lactose from the inside to outside of the cell and thus be useless or even detrimental to *E. coli* (due to constraints, other possible topologies are not feasible). As expected, *E. coli* generates only LacY of an N_{in}–C_{in} topology.

As mentioned briefly above, the first major events in the topogenesis of HMPs are how a signal sequence inserts into the membrane and whether it remains intact or is cleaved off upon insertion. As shown in Fig. 8, cleavable signals and signal anchors induce the translocation of the C terminal sequence while reverse signal anchors induce the translocation of the N terminal sequence. As the name implies, cleavable signals get cleaved off upon being inserted. Thus, secreted proteins usually possess only a cleavable signal (case a in Fig. 8). If proteins have a cleavable signal and a TM segment, the final topology becomes N_{out}–C_{in}, and this type of proteins are collectively referred to as type I MPs (case b). If the signal remains intact, the topology gets reversed, resulting in an N_{in}–C_{out} topology (case c, type II MPs). If proteins possess only a reverse signal anchor, its final topology is the same as that of type I MPs (case d). Yet, type I MPs need the assistance of signal peptidases for maturation while proteins with a reverse signal anchor do not. Thus, the latter class of proteins is referred to as type III MPs. The story gets more complicated when it comes to the topogenesis of polytopic HMPs (cases e, f and g).

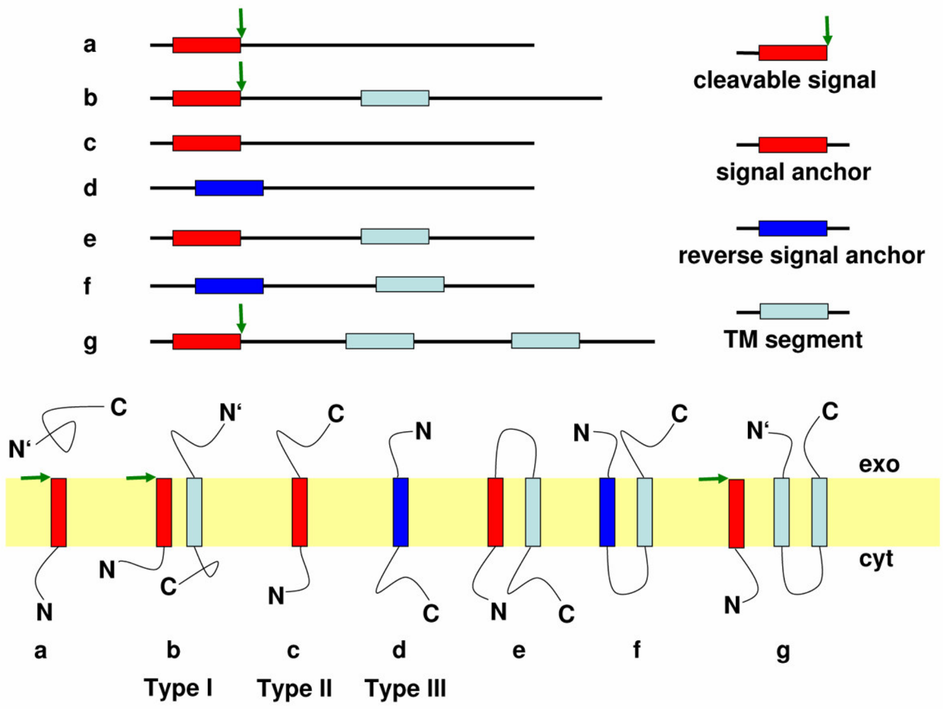


Figure 8. Different modes of HMP topogenesis. Green arrows indicate the proteolytic cleavage by signal peptidases. See the main text for details.

Having overviewed the different modes of HMP topogenesis, now the question is how the topogenesis of HMPs is controlled in the cell. Referring back to the LacY example above, how does *E. coli* make sure that only LacY of the N_{in}–C_{in} topology is synthesized? Remarkably, the topogenic signal is encoded in the amino acid sequence of HMPs, and the machinery for biogenesis of HMPs exquisitely decodes the topogenic code of HMPs and produces only HMPs of correct topology. In 1986, von Heijne made a seminal discovery that the loops enriched with Arg and Lys tend to be located in the cytoplasmic side [25, 26] and subsequently demonstrated that it is possible to predictably manipulate the topology of HMPs by changing charge distributions [27, 28]. This tendency has been dubbed the positive-inside rule. Since there is no general electric potential difference between the two sides of the ER membrane, local charges from the machinery for biogenesis of HMPs should be responsible for the positive-inside rule [20]. Even though this may sound deceptively simple, the exact molecular mechanism for the positive-inside rule still remains unknown. Since the discovery by von Heijne, a number of other factors have also been shown to be involved in the topogenesis of HMPs. As intuitively expected, the folding of sequence segments N terminal to a reverse signal anchor sterically hinders N-terminal translocation, irrespective of the charge distribution, which suggests that polypeptides should unfold in order to be translocated through the translocon pore [29]. The hydrophobicity of signal sequences themselves also contributes to the topogenesis of HMPs. Strongly hydrophobic sequences insert with an N_{out}–C_{in} topology, even when the sequence N terminal to the signal contains more positive charges than that C terminal [30, 31].

4.2.2 Recognition of TM segments by the translocon

Another important element in the biogenesis of HMPs is how TM segments are actually recognized and inserted by the translocon. Overrepresentation of hydrophobic amino acids in TM segments of HMPs clearly suggests that hydrophobic amino acids are important for the recognition of TM segments. Yet, exactly how? This important question was answered by experimental studies from von Heijne's group [32].

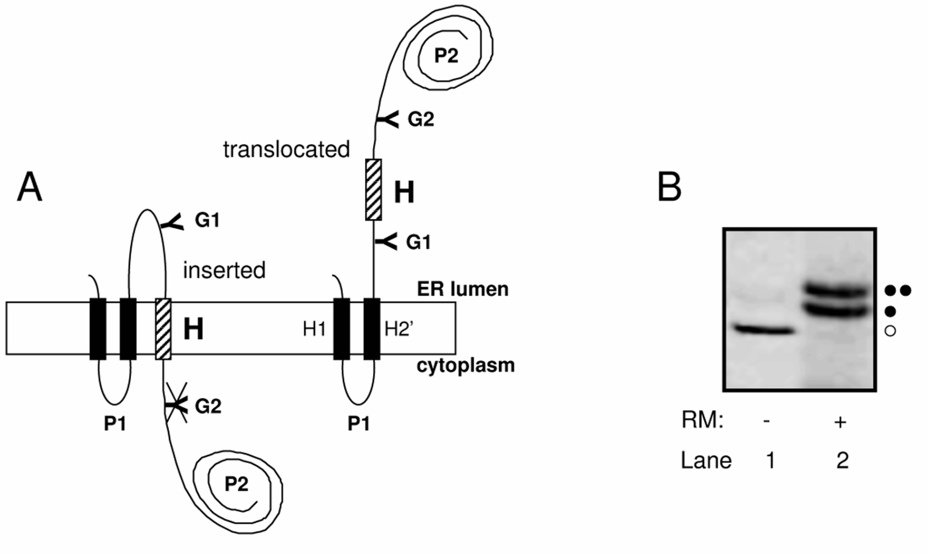


Figure 9. The experimental system for investigating the insertion of TM segments into the membrane. A: Wild-type leader peptidase (Lep) has two TM helices (H1 and H2′) in the N terminus and a large luminal domain (P2) in the C terminus, displaying an N_{out}–C_{out} topology. Synthetic H segments were inserted between residues 226 and 253 in the P2 domain. Glycosylation acceptor sites (G1 and G2) were placed in positions 96–98 and 258–260, flanking the H segment. If the H segment gets inserted, only G1 becomes glycosylated. In contrast, if the H segment gets translocated, both G1 and G2 become glycosylated. These two different cases can be differentiated by SDS-PAGE as shown in panel B. Doubly glycosylated species (denoted by two black dots) migrate more slowly than singly glycosylated species (denoted by one black dot). By quantifying the intensities of the two bands, one can determine the fraction of inserted or translocated H segments. In this experiment, plasmids encoding the Lep/H construct were expressed *in vitro* in the presence of dog rough microsomes (RM). If RM is omitted in the reaction mixture, only unglycosylated species are observed as expected, which migrate faster than both glycosylated species (Lane 1 of panel B).

The experimental system explained in Fig. 9 opens a fruitful avenue to interrogating the translocon to find out how it actually recognizes TM segments. By measuring the insertion efficiency of synthetic H segments, von Heijne and his coworkers were able to derive a scale that the translocon uses to distinguish between TM and non-TM segments [32] (Fig. 10).

In addition to uncovering the biological scale shown in Fig. 10, a couple of important observations were also made, revealing interesting insights into the molecular mechanism of the recognition of TM segments by the translocon and their subsequent insertion into the membrane. First, a strong correlation was observed between the amphiphilicity of H segments and the degree of integration into the membrane, suggesting that, during the recognition by the translocon and subsequent insertion into the membrane, H segments form a helix where the hydrophobic part contacts the membrane phase while the hydrophilic part contacts the protein components of the translocation machinery. This is in agreement with cross-linking data indicating that emerging signal sequences contact not only protein molecules but also lipid molecules. The second noteworthy point was that the effect of a Pro residue on the membrane insertion of the H segment was strongly dependent on its position in the H segment. When placed in the N terminal positions, it displayed only mild effects. Yet, when placed in the C terminal positions, it strongly hindered the membrane insertion of the H segment. In water-soluble proteins, Pro is frequently found in the N terminal positions of helices, but is almost completely absent in the C terminal positions. In light of these patterns of Pro occurrence in helices, the position-dependent effect of Pro suggests that helix formation is somehow necessary for the translocon to properly recognize TM segments and for their subsequent insertion into the membrane, as also suggested above.

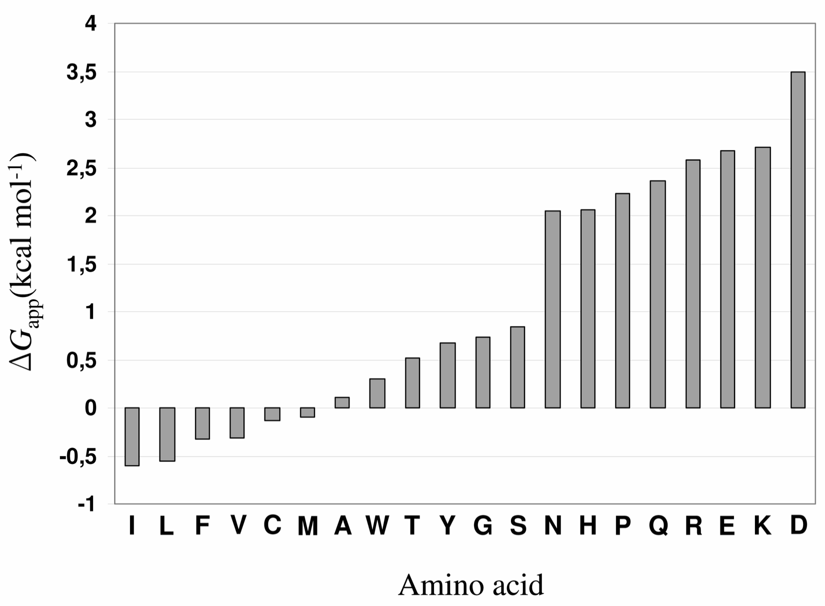


Figure 10. The biological scale that the translocon uses for recognizing TM segments. Using the experimental system shown in Fig. 9, the apparent equilibrium constant (K_{app}) for the membrane integration of an H segment was computed to be $f_{\mathrm{ig}}/f_{\mathrm{2g}}$, where f_{ig} and f_{2g} indicate the fractions of singly and doubly glycosylated H segments, respectively. ("apparent" because we are not certain about whether the balance between being inserted and being translocated truly represents an equilibrium process) From this, apparent free energy (ΔG_{app}) was obtained via $\Delta G_{\mathrm{app}} = - RT \ln K_{\mathrm{app}}$. The results are largely consistent with the intuitive expectation that hydrophobic amino acids such as I and L would be easily integrated into the membrane while hydrophilic amino acids such as K and D would not be easily integrated into the membrane.

4.2.3 Interaction among TM segments during their membrane insertion

Polytopic HMPs contain more than one TM segment, each of which should be inserted into the membrane. Then, the following question arises naturally of whether TM segments of polytopic HMPs get inserted independently on their own or there is any sort of systematic interactions among TM segments so that the whole insertion process gets facilitated. Sketchy survey of crystal structures of HMPs would make even a novice to the membrane protein research ask this question because there exist so many TM segments that are not hydrophobic enough for an independent insertion into the membrane. Interactions among TM segments during the insertion process should be responsible for the assisted insertion of weakly hydrophobic TM segments. Even though there had been circumstantial evidence supporting this idea, direct evidence was obtained rather recently.

In 2000, two research groups independently demonstrated that hydrogen bonding interactions can drive a strong association of model TM helices in detergents and the inner membrane of *E. coli* [33, 34]. Given a low dielectric constant of the membrane milieu and thus strongly promoted hydrogen bonding interactions therein, this may sound trivial today. However, it had a broad and deep impact on our understanding of the association of TM helices, a central component of the folding of HMPs (see below). Upon these observations, an immediately emerging question was then whether we could observe similar effects of hydrogen bonding interactions in the context of the translocon-mediated membrane insertion of TM helices of polytopic HMPs. This is exactly what the experimental studies to be discussed dealt with.

von Heijne and his coworkers devised an interesting experimental system shown in Fig. 11. The original motivation for this experimental system was the observation that one of the most difficult cases for computational methods of predicting the topology of HMPs is when a very long, apolar stretch of amino acids is identified in the sequence. In such cases, it is not clear whether to predict it as a single long TM segment or as a pair of TM helices connected by a tight turn. By mutating residues occurring in the middle portion of H2 of Fig. 11, they attempted to find out what determines the topologic outcome of very long, apolar stretches of amino acids [35].

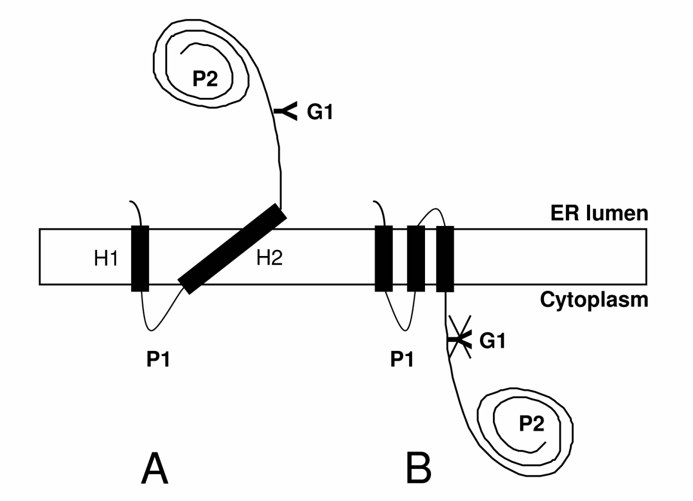


Figure 11. The experimental system for studying the interaction of TM segments during their membrane insertion. A: The H2 segment of wild-type Lep (see also Fig. 9) was replaced by a poly-Leu-based sequence ($LIK_4L_{29}VL_{10}Q_3P$) into which one or two Asn or Asp residues had been inserted. Given its length, H2 can be inserted into the membrane either as a single long TM segment or as a helical hairpin (a pair of closely spaced TM segments). Constructs where

H2 spans the membrane once as a long single TM segment will be glycosylated on a unique glycosylation acceptor site (Y) by the lumenally disposed oligosaccharyl transferase enzyme. B: If H2 spans the membrane twice as a helical hairpin, there will be no glycosylation at all. These two cases can be differentiated by SDS-PAGE as shown in Fig. 9.

Incidentally, this experimental system turned out to be as fruitful for studying stabilizing interactions of TM helices during their membrane insertion as for the original purpose. Hermansson and von Heijne introduced two Asn or Asp residues to the H2 segment of Fig. 11, one to each half of the H2 segment [36]. If the polar residues were introduced to proper positions in the sense that they could form a hydrogen bond in case the H2 segment was inserted into the membrane as a helical hairpin (B of Fig. 11), the introduced polar residues would be expected to promote the formation of a helical hairpin. The results were interesting in several folds [36]. First, it confirmed the expectation that the formation of helical hairpin is promoted by hydrogen bonding interactions between a pair of polar residues. Second, the promoting polar residues displayed position-specific effects. Namely, the pairs of positions, 8-33 and 8-37, were stabilizing whereas 8-35 and 10-33 were not. It was not immediately obvious why these faces are special. On the other hand, it suggests that TM helices while being inserted can not rotate freely, the reason of which is not yet known. Overall, these observations are strongly indicative of stabilizing interactions among TM helices during their membrane insertion.
A second piece of evidence supporting stabilizing interactions among TM segments during their membrane insertion came also from von Heijne and his coworkers, this time using a slightly different experimental scheme (Fig. 12).

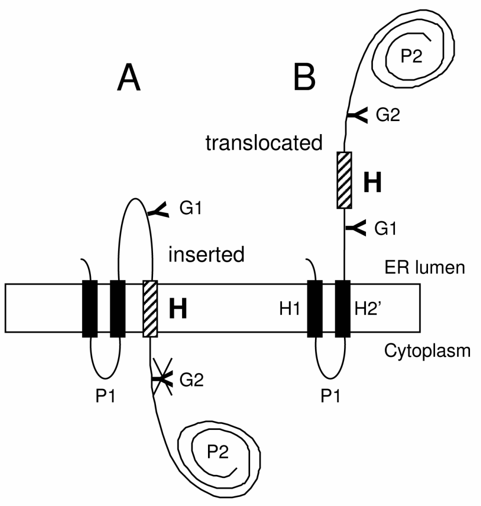


Figure 12. The experimental system for studying the interaction of TM segments during their membrane insertion.

The motivation for the experimental setup in Fig. 12 is to directly investigate whether hydrogen bonding interactions can help marginally hydrophobic TM helices to be inserted into the membrane. Some strategically chosen positions in the H2' and H segments were mutated to either Asn or Asp, and the expectation was that if the introduced polar residues formed a stabilizing hydrogen bond, it would help the marginally hydrophobic TM helix (the H segment in Fig. 12) to be pulled into the membrane, the degree of which could be measured by SDS-PAGE as for the above cases. The experimental results confirmed the expectation [37]. This confirmation has a broad relevance to the biogenesis of polytopic HMPs. The insertion efficiencies of single TM helices taken from natural polytopic HMPs are estimated to be much lower than 1 (unpublished data). For the case of bacteriorhodopsin, a prototype HMP consisting of 7 TM helices, even if we assume that each TM helix gets inserted on its own (i.e. no stabilizing interactions at all among TM helices during the membrane insertion) with an efficiency of 0.9, the overall biogenesis efficiency for this protein would be $7^{0.9} \approx 0.48$, which is unacceptably low. This low number strongly suggests the existence of some sort of biological safeguard mechanisms for the inherently complex biogenesis of polytopic HMPs. Stabilizing interactions among TM helices during their membrane insertion seem a favourable candidate. Clearly, further experimental studies would be needed to resolve this issue completely.

4.3 Structure prediction of HMPs

It still remains at its infancy to predict tertiary or quaternary structures of HMPs, in contrast to the progress being made in water-soluble proteins as reported by the biannial CASP contest [38]. Hence, most efforts have so far been focused on predicting either simple structural characteristics of HMPs such as their topology and solvent accessibility or tertiary structures of homo-oligomeric complexes of bitopic HMPs for which symmetry-based constraints can be easily implemented to drastically reduce the conformational search space.

4.3.1 Topology prediction of HMPs

The earliest efforts in predicting structural characteristics of HMPs focused on identifying TM segments from the amino acid sequence based on experimentally determined hydropathy indices of the 20 amino acids [39, 40]. For each target residue, the average hydropathy index was computed by considering a window of residues centering on the target one, which was then compared with a heuristically determined cutoff value for classification into either TM segment or non-TM segment. An important improvement in this field was made upon the observation (the positive-inside rule), as mentioned above, that intervening loops enriched with positive charges tend to be disposed on the cytoplasmic side [25, 26]. In addition to making it possible to predict the location of intervening loops, this observation also led to improvements in identifying TM segments from the sequence. Consider the case depicted in Fig. 13 where, upon scanning the query sequence, you identify a sequence of segments L1 – TM1 – L2 – TM2 – L3 where TM1 is a marginally hydrophobic TM segment and TM2 a strongly hydrophobic TM segment. Without the positive-inside rule, it might not be so clear how much one can trust the prediction result for TM1 – the marginally hydrophobic TM segment, i.e. it is quite likely to be a false positive. However, with the positive-inside rule, one can classify it as a TM segment with certainty because L1 and L3 are constrained to be on the cytoplasmic side and TM2 must cross the membrane and the only way to satisfy these constraints is for TM1 to cross the membrane as well. The discovery of the positive-inside rule led to the development of the first fully automated topology prediction program, TOPPred [41]. TOPPred first scans the sequence for certain and putative TM segments and then sorts out the most likely topology, including none, some or all of the putative TM segments, based on the charges of loop segments. In this framework, other techniques might be used to detect potential TM segments. For example, PHDhtm utilizes an artificial neural network [42], and DAS is based on a sequence profile [43].

Instead of scanning the sequence for potential TM segments and then sorting out possible topologies, the search for potential TM segments can be integrated with the evaluation of possible topologies in a single step. This idea was embodied in Memsat [44]. In Memsat, the most likely topology is predicted on the basis of correlations between the amino acid distribution of the query sequence and known distributions for each type of topologically distinct regions (cytoplasmic and exoplasmic segments and TM segments). Probabilistic approaches based on hidden Markov model (HMM) have also been exploited for the topology prediction of HMPs. Some popular examples based on HMM are HMMTOP [45] and TMHMM [46]. In fact, so many different approaches have been proposed so far that it is beyond the scope of this thesis to discuss all of them. Recently, Rost and his coworkers have set up a standard benchmark webserver so that any newly developed hydrophobicity scale or topology prediction method can be objectively assessed using a battery of measures that have been found applied in the literature [47]. Assessment on the basis of this benchmark set showed that PHDhtm and TMHMM took the lead in terms of both sensitivity and specificity, achieving the per-residue accuracy of up to 80% and the per-segment accuracy of up to 84%. The confusion rate for distinguishing between TM proteins and non-TM proteins was estimated to be 1% whereas that for distinguishing between TM helices and cleavable signal sequences was estimated to be 20 ~ 30%. Other studies also evaluated the performance of available prediction methods, revealing the overall limited accuracy of predicting TM segments [48, 49].

Predicting the topology of HMPs (or put simply, identifying TM segments) has been one of the most profitable tasks in structural bioinformatics of HMPs. Arguably, it is one of the fields in structural bioinformatics for which the highest prediction accuracy has been achieved. Nevertheless, the recent evaluation studies point to the need for new ideas/algorithms for this classical problem. In this regard, it is interesting to note that MINNOU, a novel method based on neural networks along with a novel input encoding scheme, seems to outperform PHDhtm and TMHMM [50].

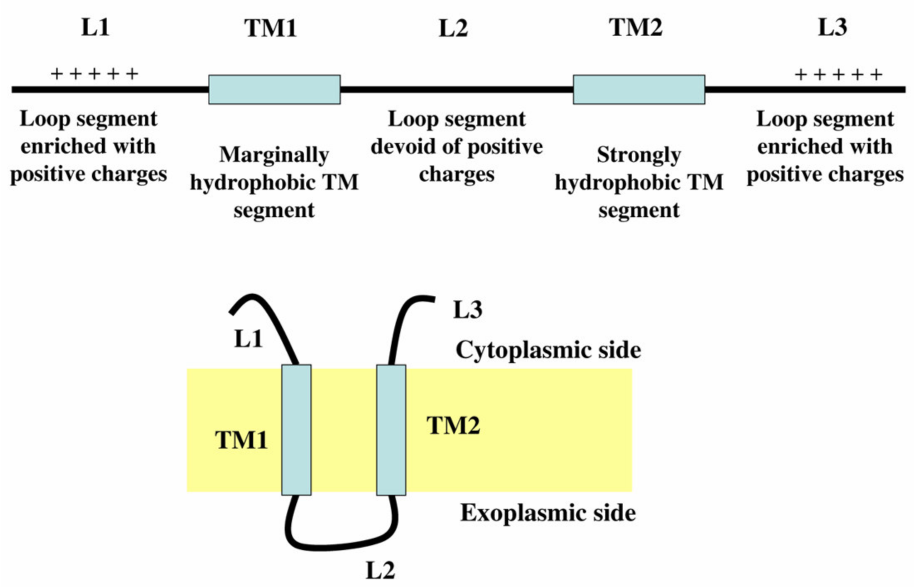

L1
TM1
L2
TM2
L3
+++++
+++++
Loop segment enriched with positive charges
Marginally hydrophobic TM segment
Loop segment devoid of positive charges
Strongly hydrophobic TM segment
Loop segment enriched with positive charges
L1
L3
Cytoplasmic side
TM1
TM2
Exoplasmic side
L2

Figure 13. The case for which the positive-inside rule helps improve the identification of TM segments from the sequence.

4.3.2 Predicting rotational conformations of TM helices

Prediction methods usually need a training set to be fine-tuned, and the developmental course of prediction methods seems strongly dependent on the course of data availability. For example, the best-performing protein structure prediction methods for water-soluble proteins such as ROSETTA [51] and TASSER [52] would not have been possible without the immense number of known structures. Unlike water-soluble proteins, it still remains extremely difficult to determine high-resolution structures of HMPs via conventional experimental techniques such as X-ray crystallography and NMR spectroscopy. Presumably for this reason, each time a new structure is determined, it is published in either Science or Nature.

As shown in Fig. 14, the first HMP crystal structure came out in 1985, which was the photosynthetic reaction center from *R. virdis* as mentioned in Section 4.1. Prior to that, the only structural information available about HMPs was a modest-resolution view of bacteriorhodopsin, which was established in 1975 by electron microscopy (EM) by Henderson and his coworker [53]. It revealed a bundle of long helical rods perpendicular to the membrane plane, suggesting that HMPs might well be thought of as an assembly of TM helices. This view well matched up with the stretches of ~ 20 hydrophobic residues found in MPs that were assumed to traverse the membrane and gained support from subsequent structure determinations for bacterial photosynthetic reaction center [13, 54], bacteriorhodopsin [55], halorhodopsin [56], cytochrome *c* oxidase [57], and glycophorin A [58].

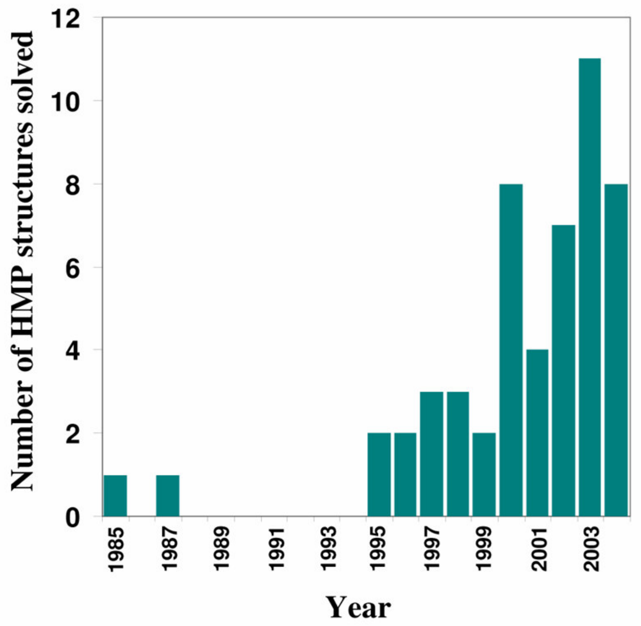


Figure 14. Pace of HMP structure determination by experimental means. Only unique structures (including same proteins from different species) are considered. The data were obtained from the list of known HMP structures compiled by White (https://blanco.biomol.uci.edu) [59]. However, presumably due to inherently poor crystal packing, many structures are of a low-resolution, and the number of high-resolution structures (better than 3.0 Å) is only 21 as of Feb. 2007.

Since HMPs are essentially constrained in the 2D space of the membrane plane, they form relatively easily 2D crystals, which are well suited to structural investigation by EM. The outcome is usually a low- or medium-resolution projection view of the arrangement of TM helices onto the membrane plane. Although quite useful, the main limitation of this medium-resolution structural information is that 1) it is nearly impossible to identify which TM segments in the sequence correspond to which density rods in the projection view and 2) even if one manages to identify the identity of density rods, it is never trivial to model a TM helix structure onto a density rod (i.e. put simply, one has to predict the rotation angle of a TM helix about its helix axis if it is reasonably straight). No computational methods have been developed to date for the first issue, and it has been usual to rely on indirect biochemical or biophysical data to sort out possible arrangements. Assuming the first issue is settled, most computational methods have been directed to the second issue of predicting the rotational angle of TM helices about their helix axes [60-63].

A natural approach to predicting the face of a TM helix exposed to the membrane would be to derive a propensity scale of the 20 amino acids to interact with the membrane and then to score different rotational conformations of the TM helix according to the propensity scale to find an optimal one. In this framework, the most important element is the propensity scale, which explains the quest for an optimal propensity scale from the early 80's [61, 62, 64-67]. Hydrophobicity scales were the first candidate for it, based on the argument that the core part of the membrane is very hydrophobic and thus the face of a TM helix exposed to it should also be hydrophobic. This argument was later found to be partially correct. Namely, the face of a TM helix exposed to the membrane is usually hydrophobic, perhaps due to the reason mentioned, yet the inside of HMPs is found to be as hydrophobic as their outside in the TM region [68]. The lack of difference in hydrophobicity between the inside and outside of HMPs in the TM region rendered hydrophobicity scales incompetent for predicting the face of TM helices exposed to the membrane. Early efforts also included the derivation of new scales from sequence data different from conventional hydrophobicity scales. For example, Taylor and colleagues derived a scale based on the assumption that the tendency of amino acids to be exposed to the membrane versus to be buried in the protein structure would be reflected in their relative occurrences in TM and non-TM regions of bitopic MPs [69]. Samatey and colleges devised a scale based on the assumption that amino acids with a similar propensity to be exposed to the membrane would be positioned on the same face of TM helices [62]. Finally, Pilpel and colleagues derived a scale (kPROT) based on the assumption that the tendency of amino acids to be exposed to the membrane versus to be buried in the protein structure would be reflected in their relative occurrences in bitopic and polytopic HMPs [61]. These early approaches, however, turned out to be rather ineffective.

A breakthrough in this field was made by exploiting the observation that the inside of HMPs tends to be more conserved than the outside in the TM region. In fact, the observation was already made in 1987 [54], and there were even active suggestions in the early 90's to incorporate it as constraints for structural modelling of HMPs [69, 70]. However, no systematic efforts were made until 2001 to tap this observation for structural modelling of HMPs, even though there were sporadic cases where conservation properties of TM residues were fruitfully exploited as constraints for predicting rotational conformations of TM helices [60]. In 2001, Arkin and his coworkers showed that TM residues buried in the interface of different subunits tend to be less conserved than those buried in single subunits [71], reigniting the interest in the application of the conservational behaviour of TM residues. Subsequently, Ben-Tal and his coworkers [63] and Weinstein and Beuming [72] exploited it in predicting rotational conformations of TM helices. Quite recently, Liang and Adamian developed a highly effective scheme for predicting exposed faces of TM helices from the sequence [73], taking advantage of conservation properties of TM residues and the canonical structure of TM helices. This new method achieved an impressive prediction accuracy of ~ 88% for known HMP structures.

The problem of predicting the face of TM helices exposed to the membrane can be cast in several different forms. One can embark on it as it is – devise a scoring scheme and find optimal rotational conformations, as mentioned above. A main limitation here is that it assumes the availability of helix parameters, most commonly from EM studies. Even though it was stated above that HMPs form relatively easily 2D crystals for structural analysis by EM, it should be stressed that it is never easy to obtain 2D crystals of HMPs. Thus, from a computational biologist's viewpoint, a better treatment of the problem would be to predict the exposed face of TM helices from the sequence, without resorting to any experimental data. This treatment becomes possible upon recognizing the predominant canonical structure of TM helices, as was first noted by Liang and Adamian [73]. At the heart of this treatment is to predict which TM residues get exposed to the membrane. This task – predicting TM residues exposed to the membrane and those buried inside – is essentially a binary classification problem, and there are tons of well-established machine-learning techniques available for it. Nevertheless, this point has never been appreciated, and only one approach – rather empirical – has been proposed so far [66]. In Paper IV, we describe the development of TMX (TransMembrane eXposure), a novel computational method for predicting the binary burial status of TM residues from the amino acid sequence, which greatly outperforms the empirical approach.

4.3.3 Structural modelling of homo-oligomeric complexes of bitopic HMPs

We now turn to modelling of the structures of homo-oligomeric complexes of bitopic HMPs, another classical structural bioinformatics problem of HMPs. There have been great interests in developing computational methods for modelling such complexes for two reasons. First, polytopic HMPs are way too complicated, and there have been few known structures from which to learn how they fold and maintain their stability. Second, unlike polytopic HMPs and hetero-oligomeric complexes of bitopic HMPs, homo-oligomeric complexes of bitopic HMPs are relatively easy to model due to inherent symmetry constraints that can be easily implemented to drastically reduce the conformational search space. Moreover, the two-stage model for the folding of polytopic HMPs [74] has justified the strategy of focusing first on homo-oligomeric complexes of bitopic HMPs and then transferring knowledge garnered there to polytopic HMPs.

It may be in order to look first into the two-stage model for the folding of polytopic HMPs. In 1990, Engelman and Popot proposed that the folding of polytopic HMPs can be separated into two energetically distinct steps [74]. In the first step, TM helices are established to satisfy the hydrogen bonding potentials of backbone polar atoms in the hydrophobic environment of the membrane. Once established, these TM helices are usually independently stable and often regarded as autonomous folding units. Unfolding of TM helices and the resulting exposure of polar backbone atoms to the hydrophobic environment would be energetically unfavourable. In the second step, independently stable TM helices associate laterally to form a functional tertiary structure. The two-stage model was based on a number of experimental observations. First, the lateral association of different subunits is frequently observed in HMP structures. Second, experimental studies on the refolding of proteolytically cleaved and denatured bacteriorhodopsin (bR) fragments support the notion behind the two-stage model. bR consists of 7 TM helices and a retinal prosthetic group. It functions as a light-activated proton pump. One big advantage of using bR for studying the folding and assembly of polytopic HMPs was that the retinal group functions as a very sensitive probe for protein conformations and thus it is possible to infer protein conformations by spectroscopic techniques. Chymotrypsin cleaves the loop segment connecting the second and third TM helices of bR, generating two peptide fragments. It was experimentally demonstrated already in the early 80's that the two initially denatured peptide fragments, upon renaturation and addition of the retinal group, can form a complex indistinguishable from native bR in terms of spectroscopic properties and light-activated proton-pumping activities [75, 76]. Later, crystallographic analysis of the complex showed that its structure is nearly the same as that of native bR [77], suggesting that native bR structure sits in a free energy minimum. It had an important implication given that the anisotropic environment of the membrane and the complexity of the translocon-mediated biogenesis of HMPs make it possible for native HMP structures to get biosynthetically trapped at a state of high energy that could not be reached during the refolding event *in vitro*. Moreover, these results are such as would be predicted from the two-stage model. In light of the two-stage model, understanding of the factors driving the homo-oligomeric association of biotopic HMPs in the membrane would also be useful for elucidating the folding and thermodynamic stability of polytopic HMPs. As a model system for the homo-oligomeric association of biotopic HMPs in the membrane, the dimerization of the TM helix of glycophorin A (GpA) has been intensively studied using a variety of tools.

GpA is a bitopic HMP found in erythrocytes with unknown function. Earlier studies showed that the dimerization of GpA is mediated by its single TM helix and sequence-specific [78]. Its dimerization is so strong that it is stable even in sodium dodecyl sulphate (SDS), a very strong detergent that denatures almost all non-covalent interactions. This persisting stability in SDS was exploited in a pioneering study by Engelman and his coworkers that identified the residues mediating the dimerization [78]. At that time, the prevailing view about the oligomerization of HMPs was that charged residues generally play a more important role than tight packing interactions via van der Waals interactions, which was primed by several observations. The interhelical salt bridge between the single TM domains of the T-cell receptor complex (TCR) α subunit and CD3δ was shown to be central to the assembly of TCR [79]. In the assembly of the Fcγ receptor, interactions between the TM domain of the α subunit of the Fcγ receptor and the equivalent domain of the ζ subunit of TCR/CD3 were shown to play a critical role, which were suggested to be mediated by single aspartic acid residues of the two TM domains [80]. In contrast to these examples, the TM segment of GpA does not possess any charged residues. To elucidate the forces driving the dimerization of the GpA TM helix, Engelman and his cowokers carried out systematic mutagenesis of each of the 23 residues in the TM segment. The results were interesting in several respects. First, the dimerization of GpA was sequence specific to a surprising extent. In some sequence positions, even the most conservative mutations, e.g the replacement of Gly83 with Ala, were not tolerated, completely disrupting the dimerization capacity. Second, interestingly, these hypersensitive sequence positions occurred with a period of 3.9 residues. These observations pointed to the existence of a special face of the GpA TM helix for dimerization, and the face was featured by small amino acids flanked by β-branched amino acids that would be well suited to the "knobs into holes" type of tight packing interactions. Molecular modelling based on these experimental constrains and the symmetry constraint for a homo-dimeric complex yielded a right-handed dimer with a helix-helix crossing angle of ~ 40° [81, 82]. In this molecular model, the two most hypersensitive glycine residues (G79 and G83) were located in the helix-helix interface, facilitating a close contact between the two helices. Later, the structure of the dimeric TM domain of GpA was determined by NMR (Fig. 15) [58], and it was found that the root mean squared deviation (RMSD) for C alpha atoms between the molecular model and the NMR models is just ~ 1Å.

Before we move on, it would be rewarding to look more into the broad impact that the studies on the dimerization behaviour of GpA have had on our current understanding of what drives the association of TM helices. Visual inspection of the NMR models made it clear that, as predicted by the molecular model above, the exquisite degree of packing in the dimeric TM domain of GpA is largely driven by the $G_{79}XXXG_{83}$ motif, explaining the intolerance of Gly79 and Gly83 even to conservative mutations. The groove formed by Gly is filled by the ridge formed by a neighboring β-branched amino acid (Val for GpA) on the opposite chain. In addition to filling the groove, β-branched amino acids further promote dimerization by reducing entropic penalties incurred upon dimerization: β-branched amino acids are allowed to populate only one rotameric state in a helix conformation and thus there is no reduction in rotameric freedom upon dimerization. There can still be one more way in which the GXXXG motif can positively contribute to a tight packing of TM helices: the Cα–H···O hydrogen bond [83], i.e. the hydrogen bond between the hydrogen atom bonded to a Cα atom and an oxygen atom, which is a bit different from conventional hydrogen bonds in that the hydrogen atom is activated not by N or O but by Cα. There were two synergistic observations behind this controversial proposal. One was a careful survey of known HMP structures by Senes *et al.* that revealed the presence of a multitude of candidate Cα–H···O hydrogen bonds [83]. The others were quantum chemical calculations estimating the energy of the Cα–H···O hydrogen bond to be as much as ~2.5 kcal/mol in vacuum, which is approximately half the energy of conventional hydrogen bonds [84, 85]. However, experimental studies raised doubts on this proposal. Bowie and his coworkers found that one of the candidate Cα–H···O hydrogen bonds highlighted by Senes *et al.* is not stabilizing at all [86]. The energy of candidate Cα–H···O hydrogen bonds in the GpA NMR models as measured by Arbely and Arkin was only ~0.88 kcal/mol [87]. Thus, further studies are warranted before definite conclusions can be made as to whether stabilizing Cα–H···O hydrogen bonds are really there in HMP structures [88].

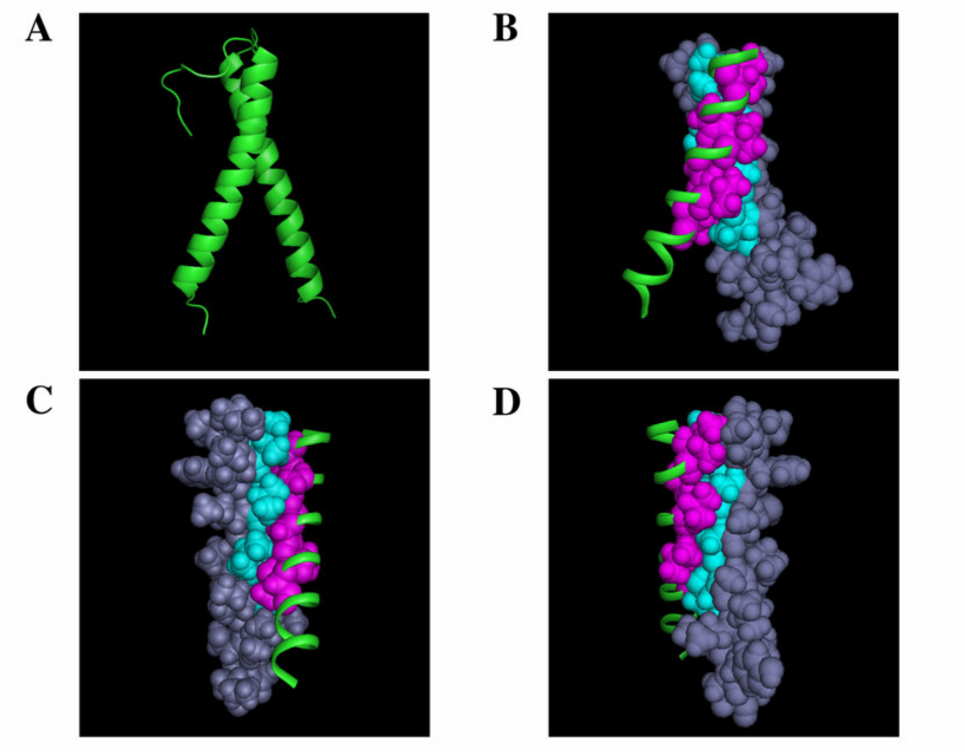


Figure 15. An NMR model of the dimeric TM domain of GpA (PDB ID: 1AFO). A: Overall view from the membrane plane including domains out of the TM domain. B, C, D: Views highlighting the residues (the interface residues, including Ile76, Gly79, Gly83 and Thr87) important for the dimerization capacity as identified by Engelman and his coworkers. Chains A (in grey) and B (in green) are depicted as spheres and ribbon, respectively. The interface residues of chains A and B are in cyan and magenta, respectively.

Now, it seemed that the GXXXG motif is crucial to the association of TM helices. But how general is its importance? If nature has adopted the GXXXG motif as a general mechanism for helix-helix interactions in the membrane, three easily verifiable predictions are to be envisaged. First, the GXXXG motif should be overrepresented in TM sequences. To see whether this is indeed the case, Senes and his colleagues devised a novel formalism for calculating the exact expectancies of pairs of amino acids in individual TM sequences [89]. The main themes observed were pairs of small amino acids (Gly, Ala and Ser) separated by three residues and flanked by large aliphatic amino acids (Ile, Val and Leu) as seen in the GpA TM sequence. The most overrepresented motif was the GXXXG motif with a p value of less than 10^{-33}, and it was in most cases flanked by β-branched amino acids (Ile and Val). Second, one should be able to retrieve the GXXXG motif from a randomized sequence library in quest for motifs mediating helix-helix interactions in the membrane. Russ and Engelman designed a library of random sequences based on the right-handed dimerization motif of GpA that covered ~10^7 sequences [90]. Using the newly developed TOXCAT system [91] that selects TM sequences capable of helix-helix association in the inner membrane of *E. coli*, they were able to identify sequence patterns with a high-affinity helix-helix interaction in the membrane. The most common motif isolated was GXXXG, occurring in more than 80% of the isolates. Third, the GXXXG motif is to be found to play an important role in other HMPs as well. In fact, since 2000, the GXXXG motif has been steadily demonstrated to be important in a diverse set of bi- and poly-topic HMPs. For it is simply out of reach to discuss all of them here, just a few examples are mentioned. The GXXXG-like motifs were shown to mediate homo- and hetero-dimerization of the single TM domains of ErbB family members [92, 93]. APH-1 is a polytopic HMP involved in the formation of the γ-secretase complex and essential for the notch/glp-1 signal transduction pathway important for embryogenesis. The conserved GXXXG tandem repeat (GXXXGXXXG) in its fourth TM domain was shown to be critical to its stable association with the γ-secretase complex [94]. The GXXXG motif was also shown to be involved in the gating transition of the MscS mechanosensitive channel [95].

The last question on the GXXXG motif for helix-helix interactions in the membrane is whether the GXXXG motif alone is sufficient for inducing a strong association of TM helices. Put another way, the dimerization (or generally association) potential – is it encoded throughout TM sequences or in just a few hot spots such as the two glycines of the GXXXG motif? This question was in part motivated by the observation that the TM helix of the major coat protein of the M13 bacteriophage (MCP-TM), in spite of possessing the GXXXG motif, forms a relatively weak homo-dimer [96]. Interestingly, such low-affinity homo-dimerization appeared to be a requirement for phage viability [97], raising the intriguing possibility that the dimerization potential of MCP-TM is tailored by the neighboring residues of the GXXXG motif such that it does not get trapped to the dimeric state, which would be detrimental to phage viability. As expected, when the neighboring residues were mutated to those found in the GpA TM helix, the dimerization capacity was significantly enhanced, nearly to two thirds the level of the dimerization of the GpA TM helix [98]. Remarkably, mutations in just two sequence positions were already sufficient for a significant boosting of the dimerization potential of MCP-TM (mutation of either position alone did not enhance the dimerization potential, though). The two residues are as long as ~ 18Å (12 residues) apart in space, revealing that neighboring residues that may be far apart from one another act together to modulate the dimerization potential of a TM helix. The emerging picture was then the one in which the GXXXG motif act in concert with neighboring residues to fine-tune the dimerization potential of a TM helix such that it can fulfil its biological requirements. Later, it was demonstrated that the GpA TM helix can dimerize even when the GXXXG motif is abolished by mutations [99], reinforcing the notion that sequence context can considerably modulate the inherent dimerization potential of the GXXXG motif.

At the heart of the GXXXG motif-mediated association of TM helices is van der Waals interaction. Then, what about electrostatic interactions in the association of TM helices (more generally in the folding of HMPs)? Since the membrane environment displays a low dielectric constant, electrostatic interactions can be significantly promoted in it compared to in an aqueous environment where solvation by polar media screens off much of electric charges. Hence, electrostatic interactions could be considered a determinant for the association of TM helices, with the possibility of even inducing promiscuous association of TM helices. Nevertheless, electrostatic interactions had not been a favourable candidate until 2000, partly due to the overwhelming importance of van der Waals interactions in the dimerization of the GpA TM helix, as discussed.

In 2000, two research groups, independently, came up with similar ideas to find out whether complementary packing interface alone is generally enough for driving the association of TM helices, as suggested by the case of GpA (the strongest known TM helix dimer) [33, 34]. Their experimental systems were designed on the basis of a dimeric leucine zipper structure found in water-soluble proteins. Leucine zipper sequences exhibit a heptad repeat pattern (*abcdefg*)$_n$, where the residues at *a* and *d* are usually hydrophobic and create a tight 'knobs into holes' packing interface in the dimeric structure, as seen in the yeast transcription factor GCN4 [100]. In the case of the GCN4 leucine zipper sequence, the residues at *a* and *d* are Val and Leu, respectively, except for one Asn at position *a*, which forms an asymmetric hydrogen bond across the dimer interface. What kind of role does this pair of buried Asn residues play in the thermodynamic stability of the GCN4 structure? A study by Harbury *et al.* revealed that an Asn to Val mutation increases the stability of the dimeric structure [101]. However, upon the mutation, a mixture of dimeric and trimeric forms began to appear. Hence, the Asn residue provides specificity for the dimeric state at the expense of stability. Then, what would happen if we convert the leucine zipper helix into a membrane-soluble analogue? Would the Asn residue provide specificity for the dimeric form at the expense of stability also in a membrane environment? This is exactly what the two research groups pursued. Interestingly, the Asn residue turned out to be also critical to the dimerization of TM helices bearing the GCN4 leucine zipper sequence, yet in a wholly different way from what was observed in water-soluble counterparts.

The two groups converted the GCN4 helix into a membrane-soluble analogue by mutating surface residues to hydrophobic amino acids while conserving buried ones. Except for the buried Asn, the internal packing of this membrane-soluble leucine zipper appears perfect for dimerization in a membrane since it is long enough to span the membrane, adopts a common helix-helix crossing angle and, most importantly, its complementary internal packing is reminiscent of that found in GpA. Then, if the Asn residue is mutated to a hydrophobic amino acid to bring the internal packing state to completion, would it display a similar degree of dimerization to GpA? Surprisingly, no dimerization was observed [33, 34]. Then, what if we conserve the Asn residue? Again surprisingly, a strong level of oligomerization (dimerization and trimerization) was observed, although not at the level of GpA. NMR experiments indicated that the Asn side chains were involved in hydrogen bonding interactions [34]. Thus, Asn-mediated hydrogen bonding interactions seemed to drive a strong association of model TM helices. Furthermore, it was demonstrated that, when a single Asn residue was grafted onto a poly-Leu background, it induced a strong level of oligomerization while the poly-Leu background sequence itself did not display any level of oligomerization [34]. Two Asn residues, when grafted onto the poly-Leu background, provided additive effects in driving the oligomerization of TM helices. Follow-up studies further demonstrated that other polar amino acids such as Asp and Glu can also induce similar degrees of oligomerization [102, 103]. These results led to the conclusion that 1) tight packing interaction alone may not be sufficient for the association of TM helices and 2) hydrogen bonding interactions mediated by polar residues could play a crucial role in the association of TM helices.

We now go back to the original subject of this subsection: structural modelling of homo-oligomeric complexes of bitopic HMPs. Perhaps, it would be fair to say that the foundation of this field was laid down by the Engelman group when they generated structural models for the dimeric TM domain of GpA [81, 82], which, as mentioned above, turned out to be very similar to NMR models. This was a tour de force at that time, given that structure determination of HMPs was far from a feasible task as it is still the case and this was the first concrete case realizing the proposition of aiding structural studies of HMPs by molecular modelling supplemented with experimental constraints. The molecular modelling protocol used by the Engelman group was very simple. Starting conformations were generated by systematically varying the helix-helix crossing angle and the rotational angle of each helix about its helix axis. Then, for each of the starting conformations, they carried out simulated annealing Monte Carlo (MC) calculations. Resulting conformations were then clustered, and the clusters meeting 1) symmetry-based constraints and 2) experimental constraints, usually in the form that mutation-sensitive residues should be located in a helix-helix interface, were selected and their average structures became final models. Since then, a number of modifications have been proposed such that one can predict structures even without experimental constraints. For example, it was proposed that one can rely on conservation properties instead of mutational data [104] because conserved residues tend to be located in helix-helix interfaces while variable ones tend to get exposed to the membrane, as mentioned above. One of the most successful approaches – the one proposed by Kim and Bowie [105] – does not even use conservation properties. The core idea of their method was that 1) walk through the search space exhaustively, identifying low-energy conformations (as evaluated by their custom-made scoring function heavily dependent on van der Waals interaction) and 2) cluster low-energy conformations and the average structure of the most populated cluster becomes a final model. In spite of its simple nature, it generated structural models that are consistent with available experimental data for a number of biotopic HMPs.

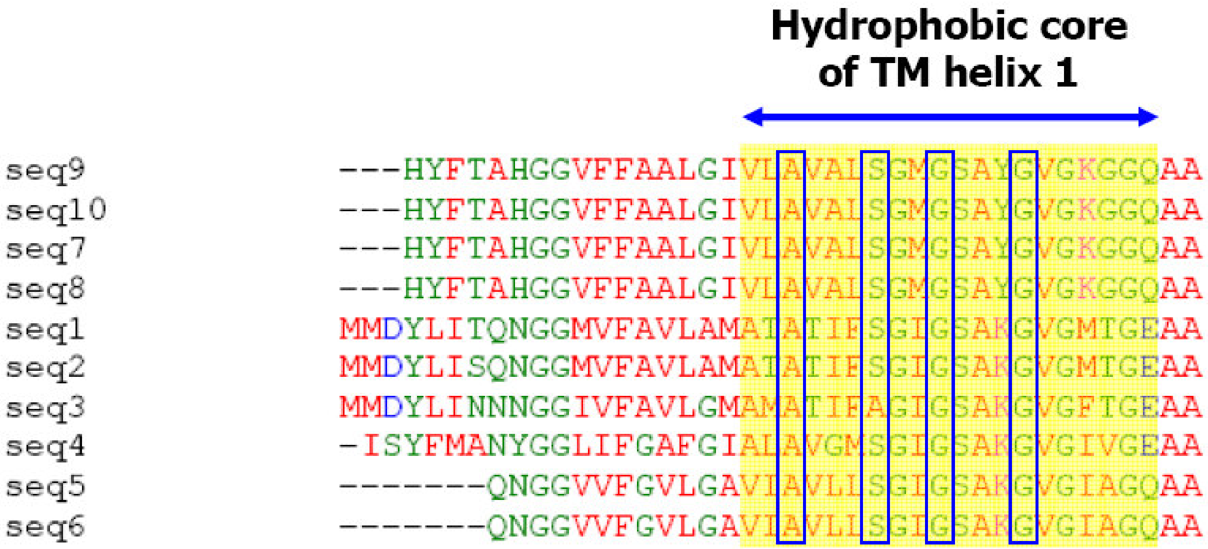
Almost all approaches proposed thus far use symmetry-based constraints and thus are not applicable for modelling hetero-oligomeric complexes, e.g. a hetero-dimer of TM helices. Given the simplicity of the system, i.e. the association of two (rather) rigid TM helices in a membrane environment, it seems feasible that one develops computational methods based on sound physical theories that properly deal with the fine balance between enthalpic and entropic components so that hetero- as well as homo-oligomeric complexes can be modelled with a reasonable accuracy without resorting to any constraints. In my Master's thesis, it was attempted to develop a heuristic scoring function that can be applied to both hetero- and homo-oligomeric complexes of bitopic HMPs [106]. Although partially successful, it was significant in that it opened an avenue for further developments.

5. Computational methodology

5.1 Linear regression

One of the central questions on HMPs is how their thermodynamic stability is maintained in a membrane milieu. It is quite conceivable that one may be able to answer this question, at least partially, by interrogating known HMP structures. In fact, our novel statistical analysis of known HMP structures reported in Paper III revealed a surprising fact that the thermodynamic stability of HMPs can be well approximated by the Cohn & Edsall partial specific volumes [67, 107]. A key to this insightful analysis was to note the direct relationship between the Pearson's correlation coefficient (i.e. covariance normalized by standard deviations) and a sum of squared errors, as shown below.

To make things more concrete, consider a situation shown in Fig. 16. From a known structure, we know which TM residues are buried inside the protein structure and which ones are exposed to the membrane. Moreover, we know the frequencies of occurrence of the 20 amino acid types in each sequence position, based on a multiple sequence alignment (MSA). Now equipped with these pieces of information, how can we answer the question of what known HMP structures tell about their thermodynamic stability in the membrane milieu?



```
seq9                ---HYFTAHGGVFFAALGIVLAVALSGMGSAYGVGKGGQAA
seq10               ---HYFTAHGGVFFAALGIVLAVALSGMGSAYGVGKGGQAA
seq7                ---HYFTAHGGVFFAALGIVLAVALSGMGSAYGVGKGGQAA
seq8                ---HYFTAHGGVFFAALGIVLAVALSGMGSAYGVGKGGQAA
seq1                MMDYLITQNGGMVFAVLAMATATIFSGIGSAKGVGMTGEAA
seq2                MMDYLISQNGGMVFAVLAMATATIFSGIGSAKGVGMTGEAA
seq3                MMDYLINNNGGIVFAVLGMAMATIFAGIGSAKGVGFTGEAA
seq4                -ISYFMANYGGLIFGAFGIALAVGMSGIGSAKGVGIVGEAA
seq5                --------QNGGVVFGVLGAVIAVLISGIGSAKGVGIAGQAA
seq6                --------QNGGVVFGVLGAVIAVLISGIGSAKGVGIAGQAA
```

Figure 16. A snippet of an MSA for the membrane rotor of V-type ATPase. The residues of TM helix 1 that are found to be in the hydrophobic core of the membrane according to the 2bl2 structure [108] are shaded ("seq1" corresponds to the 2bl2 sequence). Buried residues of TM helix 1 (i.e. those that are not touched by lipid molecules) are shown boxed.

It may be easier to answer the question if it is cast in a slightly different form. Namely, what do known HMP structures tell about the propensities of the 20 amino acid types to preferentially interact with the membrane? If this propensity scale turns out to be exactly the same as, for example, the White-Wimley (WW) scale [109], then we can infer that the folding and thermodynamic stability of HMPs are fully governed by the forces modelled in the WW scale. Technically, then, how can one derive such propensity scales from known HMP structures? Our answer was that one derives a propensity scale such that the Pearson's correlation coefficient gets maximized between the mean propensities of TM residues to be exposed to the membrane (Eq. 1) and their degrees of exposure to the membrane.

$$S(i) = \sum_{j=1}^{20} f_i(j) \times \beta(j) \tag{1}$$

In Eq. 1, $S(i)$ is the mean propensity of TM residue i to be exposed to the membrane, the index j runs over the 20 amino acid types, $f_i(j)$ the frequency of amino acid type j in TM residue i (obtained from an MSA), and $\beta(j)$ is the propensity value of amino acid type j. A nice point of this insightful formulation is that the correlation coefficient is directly related to the sum of squared errors (Eq. 2).

$$\mathrm{SSE}(\beta) = k(1 - r(\beta)^2) \tag{2}$$

In Eq. 2, SSE(β) is the sum of squared errors between the mean propensities to be exposed to the membrane and the observed degrees of exposure (Eq. 3), $r(\beta)$ the correlation coefficient between them and k a constant ≥ 0. SSE(β) and $r(\beta)$ are functions of β, as indicated.

$$\mathrm{SSE}(\beta) = (\mathrm{Y} - \mathrm{X}\beta)^{\mathrm{T}}(\mathrm{Y} - \mathrm{X}\beta) \tag{3}$$

In Eq. 3, Y is a column vector of size N (the degrees of exposure to the membrane of TM residues constituting the data set), X a matrix of N by 21 (the frequencies of occurrence of the 20 amino acid types and 1) and β a column vector of size 21 (the propensity values of the 20 amino acid types and an intercept value).

Eq. 2 reveals that maximization of the correlation coefficient with respect to β is equivalent to minimization of the sum of squared errors with respect to β. Minimization of the sum of squared errors in the realm of Eq. 1 is the task of linear regression. Hence, Eq. 2 enables one to derive a propensity scale from known HMP structures in an analytically exact manner. The propensity scale encoded in HMP structures is given by the first 20 elements of β in Eq. 4.

$$\beta = (\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}}\mathrm{Y} \tag{4}$$

Is Eq. 4 the end of a story? A careful look would reveal that there might be a problem of overfitting in the answer as given by Eq. 4. When counting only non-redundant high-resolution structures, one is left with only ~25 HMP structures. This is way too low to fairly represent the whole population of HMPs found in a typical genome. Nonetheless, we desire to extract a generalizable picture from such small data set. One way of doing so is to decay weights, known as ridge regression in linear regression. The modified answer is then given by Eq. 5.

$$\beta = (\mathrm{X}^{\mathrm{T}}\mathrm{X} + \lambda \mathrm{I})^{-1}\mathrm{X}^{\mathrm{T}}\mathrm{Y} \tag{5}$$

In Eq. 5, λ is a complexity parameter that determines the extent of weight decay. If it goes to zero, one is faced with overfitting problems. In contrast, if it is assigned very large numbers, all the propensity values converge to 0, destroying distinct features of the 20 amino acid types. In our case, values between 0.00001 and 10 seemed to be fine.

We computed correlation coefficients between the derived propensity scale (the first 20 elements of β in Eq. 5, termed the MO scale) and various empirical scales. The Cohn & Edsall partial specific volumes turned out to be most strongly correlated with the MO scale. Thus, one may conclude that the thermodynamic stability of HMPs in the membrane milieu can be best captured by the Cohn & Edsall partial specific volumes. Notably, the Cohn & Edsall partial specific volumes were derived in 1943 when there were no known protein structures at all.

5.2 Linear and non-linear classification

As mentioned above, it still remains extremely difficult to experimentally determine high-resolution structures of HMPs. Thus, it is highly desirable to develop sequence-based computational methods for predicting structural characteristics of HMPs. For water-soluble proteins, two structural characteristics have been the main target of computational prediction methods: secondary structure and solvent accessibility. For HMPs, the prediction of secondary structures does not carry as significant a momentum as for water-soluble proteins because TM segments are known to usually adopt helical conformations to satisfy the hydrogen bonding potentials of polar backbone atoms in a membrane milieu. In contrast, the prediction of solvent accessibility of HMPs has remained to date nearly untouched. The ability to predict which TM residues are buried in the protein structure and which ones are exposed to the membrane would be quite helpful in elucidating not only the structure-function relationship of HMPs but also various cellular processes mediated by this important class of proteins.

In Paper IV, the development of TMX (TransMembrane eXposure), a novel computational method for predicting the burial status of TM residues, is presented. TMX is a two-step prediction method. In the first step, it computes a positional score for a query TM residue. In the second step, the computed positional score is input to a linear support vector classifier (SVC) for predicting the burial status of the query TM residue.

Here, we briefly review the theory of SVCs. The theory of SVCs evolves from a simpler case of optimal separating hyperplanes that, while separating two separable classes, maximize the distance between a separating hyperplane and the closest point from either class.

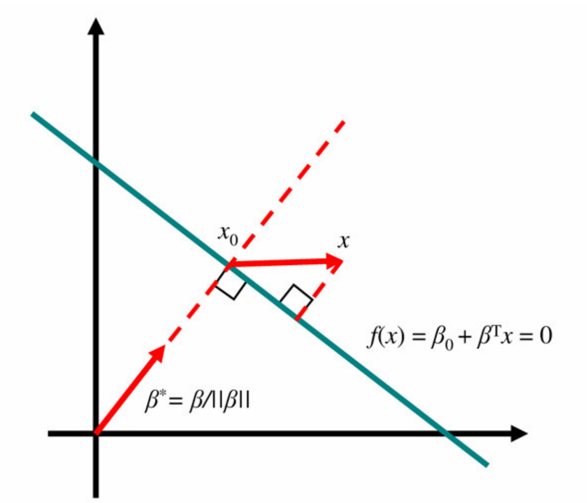


Figure 17. The linear algebra of a hyperplane

Fig. 17 shows a hyperplane L defined by the equation $f(x) = \beta_0 + \beta^T x = 0$. Here, L is a line because we are in the 2D space. The signed distance of any point x to L is computed to be $\beta^{*T}(x - x_0) = (\beta^T x + \beta_0)/\|\beta\|$. The problem to be solved for obtaining optimal separating hyperplanes as defined above becomes then

$$\max_{\beta,\beta_0,\|\beta\|=1} C \qquad (6)$$
$$\text{subject to } y_i(x_i^T\beta + \beta_0) \geq C,\ i = 1,\ldots,N.$$

In Eq. 6, the values of y_i for one class are set to be 1 (the class sitting above the hyperplane in case of Fig. 17) while those for the other class are –1 (the class sitting beneath the hyperplane in case of Fig.

17), as logically expected. The set of constraints in Eq. 6 ensures that all the points are at least a signed distance C from a separating hyperplane defined by β_0 and β. For two separable classes, all the signed distances can be set greater than or equal to zero, and one is simply to maximize C. For a non-separable case, Eq. 6 should be modified (see below).

The norm constraint for β can be eliminated by replacing the constraints with

$$y_i(x_i^{\mathrm{T}}\beta + \beta_0)/\|\beta\| \geq C \tag{7}$$

or equivalently

$$y_i(x_i^{\mathrm{T}}\beta + \beta_0) \geq C\|\beta\| \tag{8}$$

Then, the original optimization problem is recast as

$$\min_{\beta,\beta_0} \frac{\|\beta\|^2}{2} \tag{9}$$

$$\text{subject to } y_i(x_i^{\mathrm{T}}\beta + \beta_0) \geq 1,\ \ i = 1,...,N.$$

This is a typical convex optimization problem, i.e. a quadratic minimization task with a set of linear constraints. The Lagrange primal function to be minimized with respect to β and β_0 is

$$L_P = \frac{1}{2}\|\beta\|^2 - \sum_{i=1}^{N}\alpha_i\left[y_i\left(x_i^{\mathrm{T}}\beta + \beta_0\right) - 1\right] \tag{10}$$

Setting the derivatives to zero, we obtain

$$\beta = \sum_{i=1}^{N}\alpha_i y_i x_i \tag{11}$$

$$0 = \sum_{i=1}^{N}\alpha_i y_i \tag{12}$$

Upon plugging Eqs. 11 and 12 into Eq. 10, we obtain the so-called Wolfe dual

$$L_D = \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{N}\alpha_i\alpha_k y_i y_k x_i^{\mathrm{T}} x_k \tag{13}$$

$$\text{subject to } \alpha_i \geq 0.$$

The solution is obtained by maximizing L_D, a simpler convex optimization problem compared with Eq. 10. The solution must simultaneously satisfy the Karush-Kuhn-Tucker conditions, which include Eqs. 11, 12 and 14.

$$\alpha_i\left[y_i\left(x_i^{\mathrm{T}}\beta + \beta_0\right) - 1\right] = 0\ \forall i \tag{14}$$

From Eq. 14, it is inferred that if α_i is not zero, then $y_i(x_i^{\mathrm{T}}\beta + \beta_0) = 1$, which means that x_i sits on the boundary of the empty slab ("empty" in the sense that there are no training points found within the slab, Fig. 18A) centering on the optimal separating hyperplane. Conversely, if $y_i(x_i^{\mathrm{T}}\beta + \beta_0) > 1$, α_i should be zero and x_i does not contribute to the optimal separating hyperplane, as indicated by Eq. 11. Now, it is clear that the optimal separating hyperplane is defined solely by the set of training points (x_i) with non-zero α_i, i.e. those points sitting on the boundary of the empty slab via $\alpha_i > 0$. Such a set of points is often called "support vectors," explaining the origin of the term "support vector" in "support vector classifiers." Given the estimates of β and β_0, the classifier function is written as

$$\hat{G}(x) = \text{sign of } \hat{f}(x) = \text{sign of } x^{\mathrm{T}}\hat{\beta} + \hat{\beta}_0 \tag{15}$$

In the case shown in Fig. 17, if the sign of a query point is computed to be positive, it is classified to the class sitting above the hyperplane. Otherwise, it is classified to the class sitting beneath the hyperplane.

A notable feature of the formalism behind optimal separating hyperplanes is that it depends only on a relatively small number of data points with $\alpha_i > 0$ in forming a classification boundary while ignoring those points far from the boundary and can thus be more robust against model misspecification. However, if the model specification is 100% right, it might not be perfect because it heavily focuses on potentially noisier data at the boundaries of the classes. To some extent, this is also the case for SVCs (see below).

As mentioned above, the formalism described thus far is applicable only to separable cases (Fig. 18A). For inseparable cases (i.e. the two classes are not separable with a hyperplane as shown in Fig. 18B), Eq. 9 should be modified. One way of doing so is to still maximize C in Eq. 6 while allowing for some points to be on the wrong side. Technically, it is modeled as follows. Each point gets assigned a slack variable ξ as shown in Fig. 18B, and the constraints in Eq. 6 are modified as shown in Eq. 16.

$$\begin{aligned} &y_i(x_i^{\mathrm{T}}\beta + \beta_0) \geq C(1 - \xi_i),\ \ i = 1,\ldots,N. \\ &\xi_i \geq 0,\ \ i = 1,\ldots,N. \end{aligned} \tag{16}$$

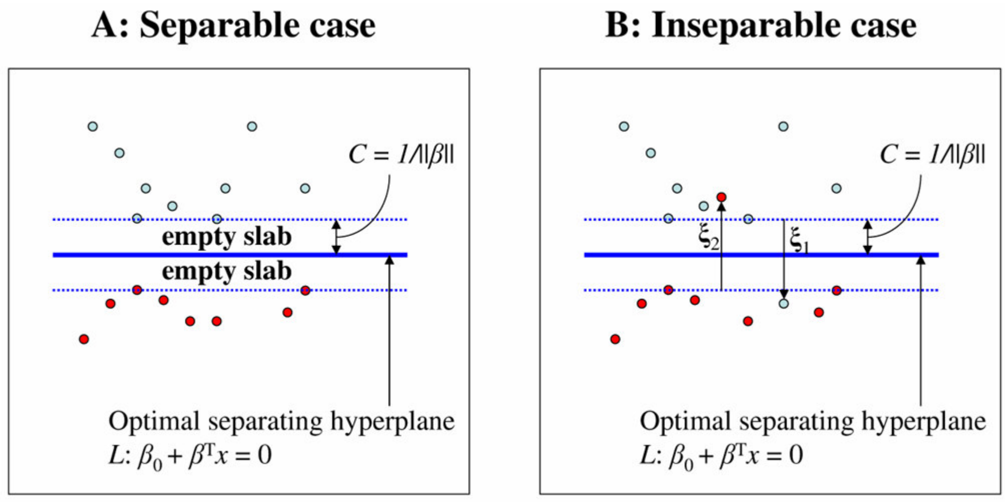

A: Separable case
B: Inseparable case
$C = 1/\|\beta\|$
empty slab
empty slab
Optimal separating hyperplane
$L: \beta_0 + \beta^{\mathrm{T}}x = 0$
$C = 1/\|\beta\|$
ξ_2
ξ_1
Optimal separating hyperplane
$L: \beta_0 + \beta^{\mathrm{T}}x = 0$

Figure 18. A: The two classes can be fully separable by a hyperplane, and the optimal separating hyperplane can be obtained by solving Eq. 9. B: It is not possible to separate the two classes with a hyperplane, and the optimal hyperplane can be obtained by solving Eq. 17.

With these new constraints, there are now two quantities to be minimized. One is $\|\beta\|$ as in Eq. 9, and the other the sum of the values of the slack variable assigned to the points. Thus, one formulates the following for inseparable cases.

$$\min_{\beta,\beta_0} \frac{\|\beta\|^2}{2} + \gamma\sum_{i=1}^{N}\xi_i \tag{17}$$
$$\text{subject to } y_i(x_i^{\mathrm{T}}\beta + \beta_0) \geq 1 - \xi_i,\ \ \xi_i \geq 0,\ \ i = 1,\ldots,N.$$

The variable γ in Eq. 17 is often called a regularization constant and determines how much weight one should put on bounding the sum of the values of the slack variable relative to minimizing the squared length of β in the overall minimization process. The effects of γ will be discussed below in conjunction with the use of non-linear kernels. The way of solving Eq. 17 is the same as that for Eq. 9 for separable cases. Namely, one formulates the Lagrange primal function, from which the Wolfe dual function is derived and solved. The Lagrange primal function is formulated as follows.

$$L_P = \frac{1}{2}\|\beta\|^2 + \gamma\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\alpha_i\left[y_i\left(x_i^{\mathrm{T}}\beta + \beta_0\right) - \left(1-\xi_i\right)\right] - \sum_{i=1}^{N}\mu_i\xi_i \tag{18}$$

Setting the derivatives to zero, we get the conditions in Eq. 19, along with the constraints α_i, μ_i, $\varsigma_i \geq 0$ for i = 1,…, N. Upon plugging the conditions in Eq. 19 into Eq. 18, the Wolfe dual function is derived (Eq. 20).

$$\begin{aligned} \beta &= \sum_{i=1}^{N}\alpha_i y_i x_i \\ 0 &= \sum_{i=1}^{N}\alpha_i y_i \\ \alpha_i &= \gamma - \mu_i \end{aligned} \tag{19}$$

$$L_D = \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{N}\alpha_i\alpha_k y_i y_k x_i^{\mathrm{T}} x_k \tag{20}$$

The values of α_i maximizing L_D and simultaneously satisfying all of the Karush-Kuhn-Tucker conditions become the solution. The only difference from the separable cases addressed above is that there are now two types of support vectors. One type is those sitting on the edge of the slab, for which the slack variable gets assigned a value of 0. The remainders are either inside the slab or on the wrong side of the separating hyperplane.

The formalism described thus far generates linear classification boundaries. Non-linear classifiers are more flexible than linear ones and might achieve better prediction accuracies. The usual tactic to go beyond linearity is to enlarge the feature space using basis expansions and then to apply linear classifiers to the enlarged feature space. Linear classifiers in the enlarged feature space are translated to non-linear classifiers in the original feature space. For the selected basis functions $h_m(x)$, m = 1,…,M, each feature vector x_i is transformed to generate a derived feature vector $h(x_i) = (h_1(x_i), h_2(x_i),\ldots, h_M(x_i))$.

Seen from this perspective, the formalism for separating hyperplanes is particularly well suited for going beyond linearity with the basis expansion technique. The Wolfe dual function (Eq. 20) involves feature vectors only via their inner products. The resulting classifier (Eqs. 15 and 19) also involves feature vectors only via their inner products. With derived feature vectors, the Wolfe dual function is straightforwardly translated as

$$L_D = \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{N}\alpha_i\alpha_k y_i y_k \left\langle h(x_i), h(x_k)\right\rangle \tag{21}$$

Similarly, the solution function $f(x)$ is translated as

$$f(x) = h(x)^{\mathrm{T}}\beta + \beta_0 = \sum_{i=1}^{N}\alpha_i y_i \left\langle h(x), h(x_i)\right\rangle + \beta_0 \tag{22}$$

Usually, it is computationally expensive to enlarge feature vectors using basis expansions. In our case, however, we do not need derived feature vectors themselves, as Eqs. 21 and 22 suggest. All we need is their inner products. Thus, it is not necessary to explicitly specify h. Rather, the knowledge of a kernel that computes the inner products of derived feature vectors is sufficient for all intents and purposes. Theoretical analysis showed that the kernels should be a symmetric positive (semi-) definite function. Three popular kernels found in the literature are (1) polynomial $(\zeta u'v+\mathrm{coef0})^{\mathrm{degree}}$, where ζ is a modulating constant for the inner product of u and v, coef0 a constant in the polynomial expansion and degree the degree of the polynomial, (2) radial $\exp(-\zeta|u-v|^2)$ and (3) sigmoid $\tanh(\zeta u'v+\mathrm{coef0})$.

The role of γ in Eq. 17 is clearer with the use of non-linear kernels. Large values of γ will put more weights on minimizing the sum of the values of the slack variable and thus result in a wiggly boundary, potentially overfitting to the training data. In contrast, small values of γ will lead to smoother boundaries, which might yield smaller generalization errors. Several studies on predicting the solvent accessibility of water-soluble proteins have suggested that the radial kernel seems superior to others, presumably due to its flexibility [110-112]. Yet, in our study in Paper IV, the linear kernel was found to be most effective, suggesting that non-linear kernels suffer from overfitting problems.

5.3 Feature selection

As mentioned above, TMX currently uses an SVC with a linear kernel for a binary classification. It may be a good idea to try other classifiers such as boosted classification trees for boosting the prediction accuracy of TMX. Equally good, though, would be to refine input vectors themselves. Initially, the input vectors for TMX consisted of 441 elements (the frequencies of occurrence of the 20 amino acid types and the conservation index for each in a window of 21 residues centering on the query residue). The reason for considering a window of 21 residues in predicting the burial status of the query residue is that one can better account for long-range effects with enlarged windows. However, the shortcoming of enlarged windows is that the signal-to-noise ratio decreases as a window size increases. For instance, it is intuitively clear that not all of the 441 elements would contribute equally to the prediction; many of them may simply be noises. Thus, one is forced to make an unpleasant tradeoff between long-range effects and signal-to-noise ratio. However, it may be possible to circumvent this unpleasant tradeoff. For instance, by a feature selection, i.e. by filtering out noisy elements. Refining input vectors in this fashion might also be helpful in overcoming the curse of dimensionality, a pathologic phenomenon associated with high dimensionalities of input vectors [113]. The existing feature selection techniques can be classified into two categories depending on whether they generate transformed features (usually via a linear combination of original features) or just cull out original features as they are. We preferred the latter to the former because of the enhanced interpretability (especially considering that the feature vector in our case resides in a 441 dimensional space). Another point of choosing the latter is that the former usually involves inversion of matrices (of a considerable dimension), which is often numerically unstable or costly. Of several alternatives, we adopted the Fisher's index. In our opinion, the Fisher's index is conceptually attractive, well suited to continuous features as in our case and numerically efficient. Put simply, the Fisher's index measures the ability of a feature to simultaneously maximize the distance between the centroids of the two classes and minimize the overlap between the two classes as shown in Figure 19.

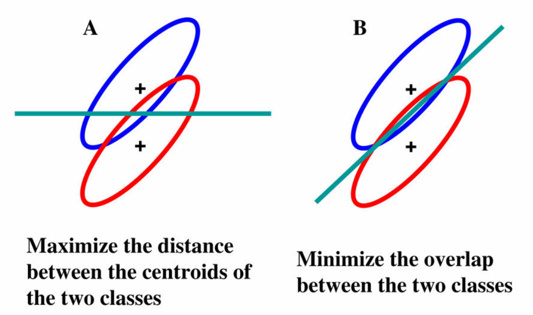

A
B
+
+
+
+
Maximize the distance between the centroids of the two classes
Minimize the overlap between the two classes

Figure 19. Graphic representation of the core idea behind the Fisher's index.

The Fisher's index of the *i*th feature is computed as follows. To turn off the effects of the other features, we set $v_i = (0\ldots0\ 1\ 0\ldots0)^T$, i.e. v_i is a column vector with all the components set to 0 except for the *i*th component set to 1. The distance between the centroids when projected onto v_i is computed as $[v_i^T(c_1\text{-}c_2)]^2$ where c_j is the centroid of the *j*th class. The overlap of the two classes when projected onto v_i is computed as $v_i^T S v_i$ where S is the common covariance matrix for the two classes. Based on these two quantities, the Fisher's index is computed as $\mathrm{FI}(i) = [v_i^T(c_1\text{-}c_2)]^2/v_i^T S v_i$. In our case, a feature selection based on the Fisher's index led to dramatic improvements in the prediction accuracy. Without any feature selection, the prediction accuracy was 76.0% on a benchmark set of 3138 TM residues. Upon feature selections, it was raised to 77.2% (a statistically significant improvement as judged by the Wilcoxon signed rank test). In addition to boosting prediction accuracies, the Fisher's index allowed us to find which elements are discriminating between buried and exposed TM residues, another boon of a feature selection analysis.

6. Conclusions and outlook

The thesis is based on three published papers and one submitted manuscript.

The first paper (Paper I) is
Park, Y. and Helms, V. (2006) *Proteins*, 64, 895-905. Assembly of Transmembrane Helices of Simple Polytopic Membrane Proteins from Sequence Conservation Patterns.
In this paper, it was demonstrated that it is feasible to generate native-like structural models for polytopic HMPs with a modest number of TM helices from packing constraints and sequence conservation patterns. In spite of their low-resolution nature, they should be helpful in rationalizing experimental data and designing further experiments, as exemplified for the case of V-type ATPase in the paper.

The second paper (Paper II) is
Park, Y. and Helms, V. (2006) *Biopolymers*, 83, 389-399. How Strongly do Sequence Conservation Patterns and Empirical Scales Correlate with Exposure Patterns of Transmembrane Helices of Membrane Proteins?
In this paper, sequence conservation patterns and a set of commonly used empirical scales were examined to find how strongly they correlate with exposure patterns of HMPs. In addition, the examination of previously proposed knowledge-based scales suggested that there could be a better way of deriving a knowledge-based scale from known HMP structures.

The third paper (Paper III) is
Park, Y. and Helms, V. (2007) *Bioinformatics*, 23, 701-708. On the Derivation of Propensity Scales for Predicting Exposed Transmembrane Residues of Helical Membrane Proteins.
Encouraged by the suggestion made in Paper II, we developed an analytical formalism for deriving a propensity scale from known HMP structures. The derived scale revealed that the thermodynamic stability of HMPs can be best captured by the Cohn & Edsall partial specific volumes.

The submitted manuscript (Paper IV) is
Park, Y. and Helms, V. Prediction of the Burial Status of Transmembrane Residues of Helical Membrane Proteins
In this manuscript, we describe the development of TMX, a novel computational method for predicting the burial status of TM residues of HMPs. Its prediction accuracy is significantly higher than that of previously proposed methods. Moreover, TMX provides confidence scores for the predictions made, making it well-suited to real application settings. Feature selection incorporated in TMX allowed for interesting insights into the architectural principle of HMPs.

A number of conclusions may be drawn from the study reported in this thesis.
1. For TM residues, it is the conservation property that most strongly correlates with their exposure patterns. The conservation property makes it possible to generate low-resolution structural models of HMPs with a modest number of TM helices.
2. In combination with the frequencies of occurrence of the 20 amino acid types in TM residues, the conservation property enables one to predict the burial status of TM residues with an accuracy of ~80%.
3. Computational analysis of known HMP structures showed that the thermodynamic stability of HMPs in a membrane milieu can be best captured by the Cohn & Edsall partial specific volumes.

The study reported in this thesis lays a firm foundation for several follow-up studies. The formalism for deriving a propensity scale of the 20 amino acids to preferentially interact with the membrane as reflected in known HMP structures (Paper III) can be straightforwardly applied to ß-barrel MPs. Also, it can be applied to the interfacial regions of both HMPs and ß-barrel MPs. A comprehensive cross-analysis of the 4 propensity scales would reveal similarities and dissimilarities in how the folding and thermodynamic stability of the two types of MPs are achieved in two distinct regions (interfacial and hydrophobic core regions) of the membrane.

Surprisingly, Paper III revealed that the thermodynamic stability of HMPs can be best captured by the partial specific volumes of the 20 amino acids. This observation raises several interesting questions, which can be readily answered. One would be to which extent one can predict helix faces exposed to the membrane from partial specific volumes. Here, we are not really interested in prediction accuracies themselves but focus on physical explanations for the structural organization of HMPs. Can we observe periodic variations of partial specific volumes along TM segments using discrete Fourier transform power spectra? Would this also be of help in predicting the boundaries of TM segments? Can we see an enrichment of amino acids with a small partial specific volume in HMPs of thermophiles vs. those of mesophiles?

Regarding TMX reported in Paper IV, it would be desired to incorporate topology prediction methods into TMX so that one can get a global overview, in a single step, of the schematic organization of the polypeptide chain in the membrane. Also, it would be desired to look for other classifiers and novel input encoding schemes for boosting prediction accuracies as much as possible.

7. References

1. White SH, Wimley WC: **Membrane protein folding and stability: physical principles.** *Annu Rev Biophys Biomol Struct* 1999, **28**:319-365.
2. Wallin E, von Heijne G: **Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms.** *Protein Sci* 1998, **7**:1029-1038.
3. Wimley WC: **The versatile beta-barrel membrane proteins.** *Curr Opin Struct Biol* 2003, **13**:404-411.
4. Toyoshima C, Inesi G: **Structural basis of ion pumping by Ca2+-ATPase of the sarcoplasmic reticulum.** *Annu Rev Biochem* 2004, **73**:269-292.
5. Dawson RJ, Locher KP: **Structure of a bacterial multidrug ABC transporter.** *Nature* 2006, **443**:180-185.
6. Locher KP, Lee AT, Rees DC: **The E. coli BtuCD structure: a framework for ABC transporter architecture and mechanism.** *Science* 2002, **296**:1091-1098.
7. Abramson J, Smirnova I, Kasho V, Verner G, Kaback HR, Iwata S: **Structure and mechanism of the lactose permease of Escherichia coli.** *Science* 2003, **301**:610-615.
8. Huang Y, Lemieux MJ, Song J, Auer M, Wang DN: **Structure and mechanism of the glycerol-3-phosphate transporter from Escherichia coli.** *Science* 2003, **301**:616-620.
9. Doyle DA, Morais-Cabral J, Pfuetzner RA, Kuo A, Gulbis JM, Cohen SL, Chait BT, MacKinnon R: **The structure of the potassium channel: molecular basis of K+ conduction and selectivity.** *Science* 1998, **280**:69-77.
10. Murata K, Mitsuoka K, Hirai T, Walz T, Agre P, Heymann JB, Engel A, Fujiyoshi Y: **Structural determinants of water permeation through aquaporin-1.** *Nature* 2000, **407**:599-605.
11. Sui H, Han BG, Lee JK, Walian P, Jap BK: **Structural basis of water-specific transport through the AQP1 water channel.** *Nature* 2001, **414**:872-878.
12. Palczewski K, Kumasaka T, Hori T, Behnke CA, Motoshima H, Fox BA, Le Trong I, Teller DC, Okada T, Stenkamp RE *et al*: **Crystal structure of rhodopsin: A G protein-coupled receptor.** *Science* 2000, **289**:739-745.
13. Deisenhofer J, Epp O, Miki K, Huber R, Michel H: **X-ray structure analysis of a membrane protein complex. Electron density map at 3 A resolution and a model of the chromophores of the photosynthetic reaction center from Rhodopseudomonas viridis.** *J Mol Biol* 1984, **180**:385-398.
14. Jakobsson PJ, Morgenstern R, Mancini J, Ford-Hutchinson A, Persson B: **Common structural features of MAPEG -- a widespread superfamily of membrane associated proteins with highly divergent functions in eicosanoid and glutathione metabolism.** *Protein Sci* 1999, **8**:689-692.

15. Wang Y, Zhang Y, Ha Y: **Crystal structure of a rhomboid family intramembrane protease.** *Nature* 2006, **444**:179-183.

16. Kim S, Jeon TJ, Oberai A, Yang D, Schmidt JJ, Bowie JU: **Transmembrane glycine zippers: physiological and pathological roles in membrane proteins.** *Proc Natl Acad Sci USA* 2005, **102**:14278-14283.

17. Nagai K, Oubridge C, Kuglstatter A, Menichelli E, Isel C, Jovine L: **Structure, function and evolution of the signal recognition particle.** *EMBO J* 2003, **22**:3479-3485.

18. Egea PF, Stroud RM, Walter P: **Targeting proteins to membranes: structure of the signal recognition particle.** *Curr Opin Struct Biol* 2005, **15**:213-220.

19. Hegde RS, Bernstein HD: **The surprising complexity of signal sequences.** *Trends Biochem Sci* 2006, **31**:563-571.

20. Higy M, Junne T, Spiess M: **Topogenesis of Membrane Proteins at the Endoplasmic Reticulum.** *Biochemistry* 2004, **43**:12716-12722.

21. Van den Berg B, Clemons WJ, Collinson I, Modis Y, Hartmann E, Harrison SC, Rapoport TA: **X-ray structure of a protein-conducting channel.** *Nature* 2004, **427**:36-44.

22. Yahr TL, Wickner WT: **Evaluating the oligomeric state of SecYEG in preprotein translocase.** *EMBO J* 2000, **19**:4393-4401.

23. Osborne AR, Rapoport TA: **Protein translocation is mediated by oligomers of the SecY complex with one SecY copy forming the channel.** *Cell* 2007, **129**:97-110.

24. von Heijne G: **Membrane-protein topology**. *Nat Rev Mol Cell Biol* 2006, **7**:909-918.

25. von Heijne G: **The distribution of positively charged residues in bacterial inner membrane proteins correlates with the trans-membrane topology.** *EMBO J* 1986, **5**:3021-3027.

26. von Heijne G, Gavel Y: **Topogenic signals in integral membrane proteins.** *Eur J Biochem* 1988, **174**:671-678.

27. von Heijne G: **Control of topology and mode of assembly of a polytopic membrane protein by positively charged residues.** *Nature* 1989, **341**:456-458.

28. Gafvelin G, von Heijne G: **Topological "frustration" in multispanning E. coli inner membrane proteins.** *Cell* 1994, **77**:401-412.

29. Denzer AJ, Nabholz CE, Spiess M: **Transmembrane orientation of signal-anchor proteins is affected by the folding state but not the size of the N-terminal domain.** *EMBO J* 1995, **14**:6311-6317.

30. Wahlberg JM, Spiess M: **Multiple determinants direct the orientation of signal-anchor proteins: the topogenic role of the hydrophobic signal domain.** *J Cell Biol* 1997, **137**:555-562.

31. Goder V, Spiess M: **Molecular mechanism of signal sequence orientation in the endoplasmic reticulum.** *EMBO J* 2003, **22**:3645-3653.

32. Hessa T, Kim H, Bihlmaier K, Lundin C, Boekel J, Andersson H, Nilsson I, White SH, von Heijne G: **Recognition of transmembrane helices by the endoplasmic reticulum translocon**. *Nature* 2005, **433**:377-381.

33. Choma C, Gratkowski H, Lear JD, DeGrado WF: **Asparagine-mediated self-association of a model transmembrane helix.** *Nat Struct Biol* 2000, **7**:161-166.

34. Zhou FX, Cocco MJ, Russ WP, Brunger AT, Engelman DM: **Interhelical hydrogen bonding drives strong interactions in membrane proteins.** *Nat Struct Biol* 2000, **7**:154-160.

35. Nilsson I, von Heijne G: **Breaking the camel's back: proline-induced turns in a model transmembrane helix.** *J Mol Biol* 1998, **284**:1185-1189.

36. Hermansson M, von Heijne G: **Inter-helical hydrogen bond formation during membrane protein integration into the ER membrane.** *J Mol Biol* 2003, **334**:803-809.

37. Meindl-Beinker NM, Lundin C, Nilsson I, White SH, von Heijne G: **Asn- and Asp-mediated interactions between transmembrane helices during translocon-mediated membrane protein assembly.** *EMBO Rep* 2006, **7**:1111-1116.

38. Kryshtafovych A, Venclovas C, Fidelis K, Moult J: **Progress over the first decade of CASP experiments.** *Proteins* 2005, **61 Suppl. 7**:225-236.

39. Kyte J, Doolittle RF: **A simple method for displaying the hydropathic character of a protein.** *J Mol Biol* 1982, **157**:105-132.

40. Eisenberg D, Schwarz E, Komaromy M, Wall R: **Analysis of membrane and surface protein sequences with the hydrophobic moment plot.** *J Mol Biol* 1984, **179**:125-142.

41. von Heijne G: **Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule.** *J Mol Biol* 1992, **225**:487-494.

42. Rost B, Casadio R, Fariselli P, Sander C: **Transmembrane helices predicted at 95% accuracy.** *Protein Sci* 1995, **4**:521-533.

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

44. Jones DT, Taylor WR, Thornton JM: **A model recognition approach to the prediction of all-helical membrane protein structure and topology.** *Biochemistry* 1994, **33**:3038-3049.

45. Tusnady GE, Simon I: **Principles governing amino acid composition of integral membrane proteins: application to topology prediction.** *J Mol Biol* 1998, **283**:489-506.

46. Kall L, Krogh A, Sonnhammer EL: **A combined transmembrane topology and signal peptide prediction method.** *J Mol Biol* 2004, **338**:1027-1036.

47. Kernytsky A, Rost B: **Static benchmarking of membrane helix predictions.** *Nucleic Acids Res* 2003, **31**:3642-3644.

48. Moller S, Croning MD, Apweiler R: **Evaluation of methods for the prediction of membrane spanning regions.** *Bioinformatics* 2001, **17**:646-653.

49. Chen CP, Rost B: **State-of-the-art in membrane protein prediction.** *Appl Bioinformatics* 2002, **1**:21-35.

50. Cao B, Porollo A, Adamczak R, Jarrell M, Meller J: **Enhanced recognition of protein transmembrane domains with prediction-based structural profiles.** *Bioinformatics* 2006, **22**:303-309.

51. Bradley P, Misura KM, Baker D: **Toward high-resolution de novo structure prediction for small proteins.** *Science* 2005, **309**:1868-1871.

52. Zhang Y, Arakaki AK, Skolnick J: **TASSER: an automated method for the prediction of protein tertiary structures in CASP6.** *Proteins* 2005, **61 Suppl. 7**:91-98.

53. Henderson R, Unwin PN: **Three-dimensional model of purple membrane obtained by electron microscopy.** *Nature* 1975, **257**:28-32.

54. Yeates TO, Komiya H, Rees DC, Allen JP, Feher G: **Structure of the reaction center from Rhodobacter sphaeroides R-26: membrane-protein interactions.** *Proc Natl Acad Sci USA* 1987, **84**:6438-6442.

55. Pebay-Peyroula E, Rummel G, Rosenbusch JP, Landau EM: **X-ray structure of bacteriorhodopsin at 2.5 angstroms from microcrystals grown in lipidic cubic phases.** *Science* 1997, **277**:1676-1681.

56. Kolbe M, Besir H, Essen LO, Oesterhelt D: **Structure of the light-driven chloride pump halorhodopsin at 1.8 A resolution.** *Science* 2000, **288**:1390-1396.

57. Iwata S, Ostermeier C, Ludwig B, Michel H: **Structure at 2.8 A resolution of cytochrome c oxidase from Paracoccus denitrificans.** *Nature* 1995, **376**:660-669.

58. MacKenzie KR, Prestegard JH, Engelman DM: **A transmembrane helix dimer: structure and implications.** *Science* 1997, **276**:131-133.

59. White SH: **The progress of membrane protein structure determination.** *Protein Sci* 2004, **13**:1948-1949.

60. Baldwin JM, Schertler GF, Unger VM: **An alpha-carbon template for the transmembrane helices in the rhodopsin family of G-protein-coupled receptors.** *J Mol Biol* 1997, **272**:144-164.

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

62. Samatey FA, Xu C, Popot JL: **On the distribution of amino acid residues in transmembrane alpha-helix bundles.** *Proc Natl Acad Sci USA* 1995, **92**:4577-4581.

63. Fleishman SJ, Harrington S, Friesner RA, Honig B, Ben-Tal N: **An automatic method for predicting transmembrane protein structures using cryo-EM and evolutionary data.** *Biophys J* 2004, **87**:3448-3459.

64. Cornette JL, Cease KB, Margalit H, Spouge JL, Berzofsky JA, DeLisi C: **Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins.** *J Mol Biol* 1987, **195**:659-685.

65. Adamian L, Nanda V, Degrado WF, Liang J: **Empirical lipid propensities of amino acid residues in multispan alpha helical membrane proteins.** *Proteins* 2005, **59**:496-509.

66. Beuming T, Weinstein H: **A knowledge-based scale for the analysis and prediction of buried and exposed faces of transmembrane domain proteins.** *Bioinformatics* 2004, **20**:1822-1835.

67. Park Y, Helms V: **On the derivation of propensity scales for predicting exposed transmembrane residues of helical membrane proteins.** *Bioinformatics* 2007, **23**:701-708.

68. Stevens TJ, Arkin IT: **Are Membrane Proteins "Inside-Out" Proteins?** *Proteins* 1999, **36**:135-143.

69. Taylor WR, Jones DT, Green NM: **A method for alpha-helical integral membrane protein fold prediction.** *Proteins* 1994, **18**:281-294.

70. Donnelly D, Overington JP, Ruffle SV, Nugent JH, Blundell TL: **Modeling alpha-helical transmembrane domains: the calculation and use of substitution tables for lipid-facing residues.** *Protein Sci* 1993, **2**:55-70.

71. Stevens TJ, Arkin IT: **Substitution rates in alpha-helical transmembrane proteins.** *Protein Sci* 2001, **10**:2507-2517.

72. Beuming T, Weinstein H: **Modeling membrane proteins based on low-resolution electron microscopy maps: a template for the TM domains of the oxalate transporter OxlT.** *Protein Eng Des Sel* 2005, **18**:119-125.

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

74. Popot JL, Engelman DM: **Membrane protein folding and oligomerization: the two-stage model.** *Biochemistry* 1990, **29**:4031-4037.

75. Huang KS, Bayley H, Liao MJ, London E, Khorana HG: **Refolding of an integral membrane protein. Denaturation, renaturation, and reconstitution of intact bacteriorhodopsin and two proteolytic fragments.** *J Biol Chem* 1981, **256**:3802-3809.

76. Liao MJ, London E, Khorana HG: **Regeneration of the native bacteriorhodopsin structure from two chymotryptic fragments.** *J Biol Chem* 1983, **258**:9949-9955.

77. Popot JL, Trewhella J, Engelman DM: **Reformation of crystalline purple membrane from purified bacteriorhodopsin fragments.** *EMBO J* 1986, **5**:3039-3044.

78. Lemmon MA, Flanagan JM, Treutlein HR, Zhang J, Engelman DM: **Sequence specificity in the dimerization of transmembrane alpha-helices.** *Biochemistry* 1992, **31**:12719-12725.

79. Manolios N, Bonifacino JS, Klausner RD: **Transmembrane helical interactions and the assembly of the T cell receptor complex.** *Science* 1990, **249**:274-277.

80. Romeo C, Seed B: **Cellular immunity to HIV activated by CD4 fused to T cell or Fc receptor polypeptides.** *Cell* 1991, **64**:1037-1046.

81. Treutlein HR, Lemmon MA, Engelman DM, Brunger AT: **The glycophorin A transmembrane domain dimer: sequence-specific propensity for a right-handed supercoil of helices.** *Biochemistry* 1992, **31**:12726-12732.

82. Adams PD, Engelman DM, Brunger AT: **Improved prediction for the structure of the dimeric transmembrane domain of glycophorin A obtained through global searching.** *Proteins* 1996, **26**:257-261.

83. Senes A, Ubarretxena-Belandia I, Engelman DM: **The Calpha ---H...O hydrogen bond: a determinant of stability and specificity in transmembrane helix interactions.** *Proc Natl Acad Sci USA* 2001, **98**:9056-9061.

84. Vargas R, Garza J, Dixon DA, Hay BP: **How Strong Is the C-H···O=C Hydrogen Bond?** *J Am Chem Soc* 2000, **122**:4750-4755.

85. Scheiner S, Kar T, Gu Y: **Strength of the Calpha H..O hydrogen bond of amino acid residues.** *J Biol Chem* 2001, **276**:9832-9837.

86. Yohannan S, Faham S, Yang D, Grosfeld D, Chamberlain AK, Bowie JU: **A C alpha-H...O hydrogen bond in a membrane protein is not stabilizing.** *J Am Chem Soc* 2004, **126**:2284-2285.

87. Arbely E, Arkin IT: **Experimental measurement of the strength of a C alpha-H...O bond in a lipid bilayer.** *J Am Chem Soc* 2004, **126**:5362-5363.

88. Mottamal M, Lazaridis T: **The contribution of C alpha-H...O hydrogen bonds to membrane protein stability depends on the position of the amide.** *Biochemistry* 2005, **44**:1607-1613.

89. Senes A, Gerstein M, Engelman DM: **Statistical analysis of amino acid patterns in transmembrane helices: the GxxxG motif occurs frequently and in association with beta-branched residues at neighboring positions.** *J Mol Biol* 2000, **296**:921-936.

90. Russ WP, Engelman DM: **The GxxxG motif: a framework for transmembrane helix-helix association.** *J Mol Biol* 2000, **296**:911-919.

91. Russ WP, Engelman DM: **TOXCAT: a measure of transmembrane helix association in a biological membrane.** *Proc Natl Acad Sci USA* 1999, **96**:863-868.

92. Mendrola JM, Berger MB, King MC, Lemmon MA: **The single transmembrane domains of ErbB receptors self-associate in cell membranes.** *J Biol Chem* 2002, **277**:4704-4712.

93. Gerber D, Sal-Man N, Shai Y: **Two motifs within a transmembrane domain, one for homodimerization and the other for heterodimerization.** *J Biol Chem* 2004, **279**:21177-21182.

94. Lee SF, Shah S, Yu C, Wigley WC, Li H, Lim M, Pedersen K, Han W, Thomas P, Lundkvist J *et al*: **A conserved GXXXG motif in APH-1 is critical for assembly and activity of the gamma-secretase complex.** *J Biol Chem* 2004, **279**:4144-4152.

95. Edwards MD, Li Y, Kim S, Miller S, Bartlett W, Black S, Dennison S, Iscla I, Blount P, Bowie JU *et al*: **Pivotal role of the glycine-rich TM3 helix in gating the MscS mechanosensitive channel.** *Nat Struct Mol Biol* 2005, **12**:113-119.

96. Dawson JP, Melnyk RA, Deber CM, Engelman DM: **Sequence context strongly modulates association of polar residues in transmembrane helices.** *J Mol Biol* 2003, **331**:255-262.

97. Henry GD, Sykes BD: **Detergent-solubilized M13 coat protein exists as an asymmetric dimer. Observation of individual monomers by 15N, 13C and 1H nuclear magnetic resonance spectroscopy.** *J Mol Biol* 1990, **212**:11-14.

98. Melnyk RA, Kim S, Curran AR, Engelman DM, Bowie JU, Deber CM: **The affinity of GXXXG motifs in transmembrane helix-helix interactions is modulated by long-range communication.** *J Biol Chem* 2004, **279**:16591-16597.

99. Doura AK, Kobus FJ, Dubrovsky L, Hibbard E, Fleming KG: **Sequence context modulates the stability of a GxxxG-mediated transmembrane helix-helix dimer.** *J Mol Biol* 2004, **341**:991-998.

100. O'Shea EK, Klemm JD, Kim PS, Alber T: **X-ray structure of the GCN4 leucine zipper, a two-stranded, parallel coiled coil.** *Science* 1991, **254**:539-544.

101. Harbury PB, Zhang T, Kim PS, Alber T: **A switch between two-, three-, and four-stranded coiled coils in GCN4 leucine zipper mutants.** *Science* 1993, **262**:1401-1407.

102. Zhou FX, Merianos HJ, Brunger AT, Engelman DM: **Polar residues drive association of polyleucine transmembrane helices.** *Proc Natl Acad Sci USA* 2001, **98**:2250-2255.

103. Gratkowski H, Lear JD, DeGrado WF: **Polar side chains drive the association of model transmembrane peptides.** *Proc Natl Acad Sci USA* 2001, **98**:880-885.

104. Briggs JA, Torres J, Arkin IT: **A new method to model membrane protein structure based on silent amino acid substitutions.** *Proteins: Struct Funct Bioinformatics* 2001, **44**:370-375.

105. Kim S, Chamberlain AK, Bowie JU: **A simple method for modeling transmembrane helix oligomers.** *J Mol Biol* 2003, **329**:831-840.

106. Park Y, Elsner M, Staritzbichler R, Helms V: **A novel scoring function for modeling structures of oligomers of transmembrane alpha-helices.** *Proteins: Struct Funct Bioinformatics* 2004, **57**:577-585.

107. Cohn EJ, Edsall JT: **Proteins, amino acids and peptides**. New York: Reinhold Publ. Corp.; 1943.

108. Murata T, Yamato I, Kakinuma Y, Leslie AG, Walker JE: **Structure of the rotor of the V-Type Na+-ATPase from Enterococcus hirae.** *Science* 2005, **308**:654-659.

109. Wimley WC, Creamer TP, White SH: **Solvation energies of amino acid side chains and backbone in a family of host-guest pentapeptides.** *Biochemistry* 1996, **35**:5109-5124.

110. Kim H, Park H: **Prediction of protein relative solvent accessibility with support vector machines and long-range interaction 3D local descriptor.** *Proteins* 2004, **54**:557-562.

111. Nguyen MN, Rajapakse JC: **Prediction of protein relative solvent accessibility with a two-stage SVM approach.** *Proteins* 2005, **59**:30-37.

112. Nguyen MN, Rajapakse JC: **Two-stage support vector regression approach for predicting accessible surface areas of amino acids.** *Proteins* 2006, **63**:542-550.

113. Hastie T, Tibshirani R, Friedman J: **The Elements of Statistical Learning**. New York: Springer; 2001.

8. Paper I

Park, Y. and Helms, V. Proteins (2006): 64, 895-905.

Assembly of Transmembrane Helices of Simple Polytopic Membrane Proteins from Sequence Conservation Patterns

Title: Assembly of Transmembrane Helices of Simple Polytopic Membrane Proteins from Sequence Conservation Patterns

Yungki Park and Volkhard Helms*

Center for Bioinformatics, Saarland University, Germany

*Corresponding author

Abstract

The transmembrane (TM) domains of most membrane proteins consist of helix bundles. The seemingly simple task of TM helix bundle assembly has turned out to be extremely difficult. This is true even for simple TM helix bundle proteins, i.e., those that have the simple form of compact TM helix bundles. Here, we present a computational method that is capable of generating native-like structural models for simple TM helix bundle proteins having modest numbers of TM helices based on sequence conservation patterns. Thus, the only requirement for our method is the presence of more than 30 homologous sequences for an accurate extraction of sequence conservation patterns. The prediction method first computes a number of representative well-packed conformations for each pair of contacting TM helices, and then a library of tertiary folds is generated by overlaying overlapping TM helices of the representative conformations. This library is scored using sequence conservation patterns, and a subsequent clustering analysis yields 5 final models. Assuming that neighboring TM helices in the sequence contact each other (but not that TM helices A and G contact each other), the method produced structural models of CA RMSD of 3 ~ 5 Å from corresponding crystal structures for bacteriorhodopsin, halorhodopsin, sensory rhodopsin II, and rhodopsin. In blind predictions, this type of contact knowledge is not available. Mimicking this, predictions were made for the rotor of the V-type Na$^+$-ATPase without such knowledge. The CA RMSD between the best model and its crystal structure is only 3.4 Å, and its contact accuracy reaches 55%. Furthermore, the model correctly identifies the binding pocket for sodium ion. These results demonstrate that the method can be readily applied to *ab initio* structure prediction of simple TM helix bundle proteins having modest numbers of TM helices.

Introduction

Over the past decade, steady progress has been noted in the structure prediction of soluble proteins.[1] Especially in the new fold category, a couple of methods such as Rosetta[2] and TASSER[3] have shown admirable performance. In contrast, little has been achieved in the structure prediction of membrane proteins. As pointed out by White and von Heijne,[4] two fundamental problems need to be addressed for the development of reliable structure prediction methods: the mechanisms of the biological assembly of membrane proteins and the thermodynamic principles of their structural stability in the lipid bilayer. Even though great strides have been made regarding the two problems over the past years,[5-11] our understanding is not yet sufficient. In fact, as discussed by White,[12] the situation seems to get more complicated than expected, in light of the complex structures recently presented for the ClC chloride channel[13] and the KvAP voltage-gated potassium channel.[14] The end-to-end arrangement of helices within the hydrophobic core of the membrane seen in the aquaporin family[15] also compounds our understanding about membrane protein folding.

Nevertheless, for simple membrane proteins, i.e., those that have the *simple* form of *compact* transmembrane (TM) helix bundles, structural modeling can, in principle, be divided into two steps according to the well-established two-stage model[16]: determination of the portions of the primary sequence that traverse the membrane and assembly of these TM helices. Since TM boundaries can be accurately predicted in many cases,[17-22] the structural modeling boils down to assembly of TM helices. This has been considered an easier problem compared to the structure prediction of soluble proteins. Yet, years of work have attested that this is still too difficult a problem, and successful structural modeling has been mostly confined to homo-oligomeric complexes,[23-26] where symmetry constraints can be easily imposed on to simplify the conformational search problem. The current study tackles structural modeling of simple TM helix bundle proteins (being "simple" as defined above) that consist of modest numbers of TM helices.

So far, a number of studies have been reported about structural modeling of polytopic membrane proteins. In 1997, Baldwin and his coworkers presented a method for assembling the TM helices of the rhodopsin family of G-protein-coupled receptors (GPCRs) based on helix parameters extracted from cryo-electron microscopy (cryo-EM) maps and sequence conservation patterns.[27] The study demonstrated that sequence conservation patterns can be a powerful tool for structural modeling. Yet, the presence of EM maps was a prerequisite for the presented methodology, which is really a heavy requirement given that EM maps are as difficult to get as crystals for the determination of atomic-scale structural models. Subsequently, Goddard and his coworkers developed a computational method of predicting the structures of GPCRs.[28,29] However, their method also assumed the presence of EM maps, since it is based on helix parameters extracted from EM maps. A computational method of different nature has also been proposed, where a set of distance constraints is utilized for the generation of a small number of feasible TM helix bundles.[30,31] This method successfully predicted the structure of bovine rhodopsin using a set of 27 distance constraints.[31] Even though it is relatively easier to get this type of experimental constraints than EM maps, it is still expected to be quite laborious to obtain that many experimental constraints routinely. The method presented in this study does not make any heavy assumptions of this sort. Since it is based on sequence conservation patterns, the only requirement is the presence of more than ~ 30 homologous sequences. Given rapid increases in the size of sequence databases, we regard this requirement as quite light.

Methods

Overview of the prediction protocol

The fundamental assumption of our prediction protocol is that, even though complex tertiary interactions among non-neighboring TM helices are expected to play an important role in the determination of overall structures, we could split, to a large extent, modeling of TM helix bundles into modeling of TM helix pairs and subsequent assembly to TM helix bundles. This assumption might not hold for complex polytopic membrane proteins, yet it is expected to be a reasonably good approximation for simple TM helix bundle proteins, the focus of the current study.

Based on the pairwise separation scheme, we first compute representative well-packed conformations for each pair of contacting TM helices. Then, a library of tertiary folds is generated by overlaying overlapping TM helices of the representative conformations. This library is scored using sequence conservation patterns. As is usually done in the protein structure prediction field, a clustering analysis of the top-scoring folds and subsequent rigid-body refinements produce 8 candidate models. This whole process is repeated 50 times. The generated candidate models are then pooled, and the same clustering analysis yields 5 final models. This overall flow of the prediction protocol is depicted in Fig. 1.

Test proteins

As outlined above, we restricted our attention to simple TM helix bundle proteins as represented by bacteriorhodopsin.[32] We were reluctant to test the prediction protocol against structure fragments, for example, the N- or C- terminal domains of lactose permease,[33] because they are not "clean" domains as understood in soluble proteins. Since we score the library of tertiary folds using sequence conservation patterns, membrane proteins with small numbers of homologous sequences in sequence databases were not suitable, either. With these criteria in mind, we found 5 suitable targets from the list of membrane proteins with known structure summarized by White (http://blanco.biomol.uci.edu): bacteriorhodopsin (bR),[32] halorhodopsin (hR),[34] sensory rhodopsin II (sR),[35] rhodopsin,[36] and the rotor of the V-type Na^+-ATPase (NtpK).[37] The sequence identities of the three bacterial rhodopsins are ~ 30%. Yet it is to be noted that a couple of recent studies considered them to be independent targets.[38,39]

Since the prediction protocol generates a structural model for the TM domains of the test proteins, TM boundaries need to be defined before structural modeling. The current study focuses on the second stage of the two-stage model: assembly of independently stable TM helices to TM helix bundles. Thus we simply took TM boundary information from the PDB-TM database.[40] Once defined, individual TM helices were constructed as ideal right-handed α-helices with backbone dihedral angles of $\Phi = -57°$ and $\Psi = -47°$. Random perturbations in the TM boundaries taken from the PDB-TM database within a variation of ±2 residues did not affect prediction results significantly (data not shown).

Systematic conformational search of the TM helix bundles

As stated above, the way we travel through the conformational space of a given TM helix bundle is by overlaying overlapping helices of the representative conformations. Thus, one first needs to compute representative conformations. This is carried out as follows: For each pair of contacting TM helices, 3888 conformations are to be explored in a systematic way (see below). These are scored by our newly developed scoring function.[41] Then, the 1000 lowest-energy conformations are clustered into a few groups, and the centroid conformations for the groups become the representative well-packed conformations.

It is an open issue how many lowest-energy conformations are to be clustered into how many groups. Empirically, we chose to cluster 1000 lowest-energy conformations into 13 groups. With regards to the number of representative conformations assigned to each pair of contacting TM helices, we investigated the possibilities from as large as 40 to as small as 10. Going below 11 gave significantly poorer results for some test proteins. Going up beyond 15 slightly improved the results, yielding a model of C alpha atom root-mean-square deviation (CA RMSD) of 2.5 Å for some cases. Yet, the total computational time increases rapidly with the number of representative conformations. For example, when using 13 conformations for each pair of contacting TM helices of the 7 TM helix bundle protein, the computational cost for generating a library of tertiary folds is 13^6, taking ~ 10 minutes on a 2.8 GHz processor. However, it takes ~ 150 hours on the same processor when using 40 conformations for each pair (40^6). It is desirable to use sufficient numbers of representative conformations to guarantee an acceptable quality of results, as long as affordable on a typical workstation. Numbers between 11 and 15 seem to be a good choice in this regard for TM helix bundles consisting of 7 TM helices. For TM helix bundles with 4 TM helices, the numbers between 31 and 35 seem to be a good choice. For reasons of limited space, we only present results with 15 representative conformations for bR, hR, sR, and rhodopsin (all have 7 TM helices). For NtpK (having 4 TM helices), the results with 35 representative conformations are reported.

For a systematic and unbiased scanning of the conformational space of a pair of contacting TM helices, we first randomly rotated the two TM helices. Then, four of the six variables describing their relative orientations were manipulated as follows (Fig. 2). Describing the helix-helix distance, ζ, was set to 9.0 Å, which is a typical value observed for contacting TM helices.[42] To allow the two helices to contact each other at different positions, δ was varied in steps of 5.0 Å in the range of –5.0 Å ~ 5.0 Å for both helices. Describing the two rotational angles about the helix axes, α and β were varied in steps of 20°. γ, describing the tilting angle, was allowed among -24°, -12°, 12°, or 24°. In total, 3888 ($3 \cdot 18 \cdot 18 \cdot 4$) conformations were explored. We could have performed a denser scanning of the conformational space, yet we observed that the current degree of scanning is dense enough for simple TM helix bundle proteins. Furthermore, a rather coarse scanning is desirable given that the next step is a clustering calculation. As before, we optimize the side chain conformations of each structure explored using SCWRL[43] and compute interaction energies with a cutoff at 9 Å. Those conformations were removed during the scanning that harbor steric clashes (1.5 Å cutoff for the inter-atomic distance between heavy atoms of the SCWRL-optimized structure). Our earlier report[41] defined the interaction centers only for 11 amino acids. The expanded list of the interaction centers for all 20 amino acids is summarized in Table 1. Clustering was performed using the average linkage clustering algorithm.[44]

Upon generating representative conformations for contacting TM helix pairs, a library of tertiary folds was built by overlaying overlapping TM helices of the representative conformations. To speed up the computation, TM helix bundles were represented only by CAs from this step on. Folds with bad contacts (distance between CAs of different TM helices of 4.0 Å or less) were removed. TM helix bundle proteins are well known to form compact structures.[45] Thus loosely packed folds were removed as well. For this, we calculated the average distance between all pairs of CA-based helix centers. Simple experiments on a 2D grid show that 16.0 Å is a reasonable upper bound for TM helix bundles consisting of 7 TM helices (Fig. 3). This compactness filter was also useful in keeping the sizes of libraries to a manageable level. For TM helix bundles with 4 TM helices, it was not necessary to apply this sort of compactness filter since the sizes of the libraries were inherently small.

Scoring of a library of TM helix bundle folds based on sequence conservation patterns

A number of studies have shown that the more conserved a sequence position is, the less likely it is to be exposed to the lipid bilayer.[27,46,47] We make use of this observation for scoring the libraries of TM helix bundle folds. Specifically, scores were computed using the following equation.

$$Score = \sum_{i} CI_i \cdot rSASA_i \qquad (1)$$

In Eq. 1, the index i runs over the residues making up the TM helix bundle. $rSASA_i$ is the solvent-accessible surface area (SASA) of the sequence position i in the TM helix bundle divided by its SASA when the TM helix containing it is isolated from the TM helix bundle, taking values between 0 (for fully buried residues) and 1 (for fully exposed residues). For computing the SASAs of CA-only models, the probe radius was set to 4.0 Å to mimic all-atom results as best as possible. Calculation of SASAs was performed using the BALL library.[48] CI_i is the conservation index of the sequence position i estimated from a multiple sequence alignment (MSA) using the variance-based method (Eq. 2, see below). The suite of web services at the EBI website (http://srs.ebi.ac.uk/srsbin) was used when obtaining MSAs. A maximum of 500 similar sequences were retrieved from the Uniprot database[49] using BlastP[50,51] with a significance cutoff of 10^{-4} while keeping the other parameters to the default values, and aligned against the query sequence using ClustalW[52] with all the parameters set to the default values. One wishes to obtain MSAs where the diversity of sequences within the significance cutoff is represented as fully as possible for the accurate estimation of conservation indices. This seemed to be satisfied for all test proteins except rhodopsin; the sequence identity of the least similar sequence to rhodopsin was ~ 40%. Thus, we restricted the search for similar sequences to rhodopsin to the Swissprot database.[53] To get rid of sequence fragments in the raw MSAs, a sequence identity filter of 25% was applied, and remaining sequences were realigned using ClustalW with all the parameters set to the default values. A recent analysis has shown that one needs to align at least 20 sequences to accurately estimate conservation indices from MSAs.[54] In all cases, the number of sequences in the refined MSA was greater than 30 (specifically, Bacteriorhodopsin – 45, Halorhodpsin – 30, Sensory rhodopsin II – 30, Rhodopsin – 149, NtpK – 54). In the current implementation, the more conserved a sequence position, the higher its conservation index. Accordingly, the better a given TM helix bundle manifests the sequence conservation pattern, the lower is its score computed by Eq. 1.

Estimation of conservation indices from multiple sequence alignments

Conservation indices were estimated from multiple sequence alignments using the following variance-based method.[54]

$$C(i) = \sqrt{\sum\nolimits_j (f_j(i) - f_j)^2} \qquad (2)$$

In Equation 2, $C(i)$ is the conservation index for the sequence position i in a multiple sequence alignment, f_j is the overall frequency of amino acid j in the alignment, and $f_j(i)$ is the frequency of amino acid j in the sequence position i. Obviously, positions with $f_j(i)$ equal to f_j for all amino acids j are assigned $C(i) = 0$. On the contrary, $C(i)$ takes on its maximum for the position occupied by an invariant amino acid whose overall frequency in the alignment is low. In order to account for the redundancy of aligned sequences, amino acid frequencies were weighted using a modified method of Henikoff and Henikoff as implemented in PSI-BLAST.[51,55] The conservation indices computed by Eq. 2 were then normalized by subtracting the mean from each conservation index and dividing by the standard deviation. Actual calculations were performed using a program written by Pei and Grishin,[54] which is freely available at ftp://iole.swmed.edu/pub/al2co/. As Eq. 2 implies, “conserved” in this study means being identical, disallowing “conserved” substitutions. More sophisticated methods taking “conserved” substitutions into account usually generated similar conservation indices (data not shown). Thus, the seemingly simple Eq. 2 appears sophisticated enough to accurately extract sequence conservation patterns from multiple sequence alignments. The current scoring scheme (Eq. 1 combined with Eq. 2) was found to work remarkably well for predicting the rotational angles of TM helices about the helix axes derived from EM maps (manuscript submitted for publication), which is why we used it for the current study.

Generation of 8 candidate models

Upon scoring of the library based on sequence conservation patterns, 500 top-scoring folds were clustered into 8 groups using the average linkage clustering algorithm, and average structures were generated for the 8 groups, yielding 8 candidate models. Again, it is an open issue how many top-scoring folds are to be clustered into how many groups. The combination 500/8 was chosen empirically. However, any reasonable choice seems fine because the prediction protocol is run 50 times and the candidate models generated are going to be pooled in the end (see below). Average structures were generated by optimal superimposition of the corresponding ideal helices onto the averaged coordinates. As a result, the 8 models do not have deviations from ideal geometry such as distortions in bond length, angle and dihedral angle. Yet, they did contain steric clashes due to their average nature. To correct for deficiencies of this type, a rigid-body refinement step was carried out. There are many possible ways for performing rigid-body refinements. The most natural one complementing the discrete nature of the previous step of overlaying overlapping TM helices of representative conformations seems to relax the angles formed by three neighboring TM helices (Fig. 4). For example, when forming a triple of TM helices ABC from pairs AB and BC, the *position* of TM helix C relative to that of TM helix A is fully determined by the relative *orientation* of A to B and that of B to C. This leaves no room at all for the adjustment of TM helix C’s position relative to that of TM helix A. Relaxing the angle formed by the three projection points onto the membrane plane of the CA-based centers of the three TM helices is a natural way for adjusting the positions of TM helices.

First, TM helices were translated along their helix axes to align their CA-based centers to the midplane of the lipid bilayer. Then to remove any steric clashes present in the average structures, we relaxed all the interhelical distances to 9.6 Å, which is the average distance between contacting TM helices.[42] Then, we relaxed the angles as follows: Starting with the smallest one, we change it in 1° increments over the range of -25° ~ 25° around its current value in search of the angle that minimizes the maximal fractional exposure of individual TM helices to the lipid bilayer while inducing no steric clashes (distance between CAs of different TM helices of 5.0 Å or less). This process was repeated with the next smallest angle until all the angles were relaxed. Figure 4 depicts this rigid-body refinement step. This rigid-body refinement process was iterated two times.

Generation of 5 final models

Since we start with randomly rotated helices for an unbiased scanning, every run of the above procedure does not generate results of the same quality. For this reason, we repeat the procedure 50 times, as depicted in Fig. 1. Upon running the prediction protocol 50 times, the candidate models generated so far are pooled, and a clustering analysis is carried out to yield 5 final models. Each run generates 8 candidate models as described above, and the number of pooled candidate models becomes 400 (8 times 50). Out of these, 50 top-scoring models are clustered into 8 groups. Then, average structures are computed for the 5 most-populated groups, generating 5 final models. Other combinations than 50 top-scoring models and 8 groups generate similar results (data not shown). Thus the choice of 50/8 is not really critical. For the *ab initio* structure predictions of NtpK (having 4 TM helices), the prediction protocol is run for all possible permutations of the connectivity of TM helices (3 in this case, see below). Thus, the number of pooled models becomes 1200 (3 times 8 times 50), and 150 top-scoring models are clustered into 8 groups (in line with the above proportion 400:50 = 1200:150). The reason for generating 5 final models was because predictors are allowed to submit a maximum of 5 models in CASP experiments.[56]

Computational expense

The computation of representative conformations for each pair of contacting TM helices takes ~ 10 minutes on a 2.8 GHz processor. Generation of a library of tertiary folds and concurrent scoring based on sequence conservation patterns for a TM helix bundle with 7 TM helices takes ~ 15 minutes on the same processor. All together, one run of the prediction protocol requires ~ 90 CPU minutes.

Results

Summary of the prediction results

Given that there exist no sophisticated scoring functions for modeling the structures of membrane proteins, sequence conservation patterns might be a viable alternative. Even though this conjecture appears feasible given the previous studies that reported a successful application of sequence conservation patterns for predicting the rotational angles of TM helices about their helix axes.[27,46] it has never been confirmed. In other words, it is not clear whether sequence conservation patterns could be equally powerful for *ab initio* structural modeling where helix parameters extracted from EM maps are not assumed. When we launched the current investigation, we immediately encountered the problem of scarcity of suitable test systems. Since the presented prediction protocol starts with computation of representative conformations of each pair of contacting TM helices, the protocol should be run for all possible permutations of the connectivity of TM helices for an *ab initio* structure prediction. Admittedly, this is nearly impossible for a TM helix bundle consisting of 6 or more helices. Out of the 5 test proteins, only NtpK has less than 6 TM helices. Thus, for the other 4 test systems (bR, hR, sR, and bovine rhodopsin), we were forced to perform a rather limited test. Namely, we assumed that neighboring TM helices in the sequence contact each other: TM helix pairs AB, BC, CD, DE, EF, and FG contact each other. Since we do not assume that TM helices A and G contact each other, the number of possible TM helix bundle folds remains quite high (For example, see Fig. 6). As a result, it is still quite difficult to predict the correct arrangement pattern of TM helices, let alone accurate predictions of tilting and rotational angles of individual TM helices. For NtpK, we do not assume this type of connectivity knowledge, making *ab initio* predictions.

We first present the prediction results for the three bacterial rhodopsins and bovine rhodopsin, for which we assumed the connectivity knowledge as mentioned above. Shown in Fig. 5 are the superimpositions of the best models onto the corresponding crystal structures. Table 2 summarizes quantitative comparisons between the predicted and experimental structures. For the three bacterial rhodopsins (bR, hR and sR), the relative positions of TM helices are correctly modeled. Remarkably, the rotational angles of TM helices about their helix axes are predicted with high precision in most cases, including bovine rhodopsin. This level of accuracy rivals that of the best-performing methods specialized in the prediction of the rotational angles of TM helices about the helix axes extracted from EM maps.[57,58] On the other hand, the coupled prediction of an overall fold pattern and the rotational angles of individual TM helices about their helix axes is, to a large extent, something expected, given the way TM helix bundle folds are generated: overlaying overlapping TM helices of the representative well-packed conformations of contacting TM helix pairs. As expected, the prediction quality for rhodopsin is relatively poor because the pairwise separation scheme is problematic for its complex structure. Specifically, TM helix C is deeply buried in the structural core, and this appears to induce the separation of TM helices D and E. As a result, TM helices D and E do not contact as significantly as other contacting pairs do while, in our approach, we assume they do. Fig. 6 shows the properties of the other 4 final models for bR. It clearly demonstrates that even after assuming that neighboring TM helices in the sequence contact each other, the number of overall fold patterns remains quite high, and thus it is not trivial to predict the correct fold pattern.

When it comes to structural modeling of TM helix bundle proteins, it is of considerable interest to predict helix-helix contacts because they play an important role in maintaining the structural and functional integrity. Since the models generated by the current protocol consist only of CAs, it is not straightforward how to define interhelical contacts. Given that the average interhelical distance of contacting TM helices of polytopic membrane proteins is 9.6 Å,[42] it is reasonable to define a contact for two CAs belonging to different TM helices with the distance between the two smaller than 7.5 Å. We computed two measures of interhelical contact predictions. The first one is accuracy, i.e., TP/N_{pred}, where TP = true positives and N_{pred} = total number of predictions made. The second one is coverage, i.e., TP/N_{obs}, where N_{obs} = total number of observed contacts in experimental structures. Table 2 lists these numbers for each test protein. Complementary to the values of CA RMSD, the values of accuracy and coverage indicate that models of reasonably good quality have been generated for the 4 test proteins. This, in turn, demonstrates that sequence conservation patterns work equally powerfully in a situation where helix parameters extracted from EM maps are not available.

To confirm that sequence conservation patterns played a major role in the positive results of Table 2, we made an experiment with sequence conservation patterns. Namely, we clustered 500 *highest*-score (instead of the 500 *lowest*-score) folds for generating 8 candidate models, and 50 *highest*-score candidate models were clustered to yield 5 final models, performing the other steps in the same way. (Another possible control would have been to select 500 folds and 50 candidate models randomly; we chose the highest-score ones since we believed that this scheme would better reveal the discriminating power of the scoring function between ‘good’ and ‘bad’ folds). Table 3 shows dramatic changes in the results. Considerable increases in CA RMSD are observed for all 4 test proteins. As a result, the values of accuracy and coverage of interhelical contact predictions nearly drop to 0%, confirming that sequence conservation patterns play a critical role in generating native-like structure models.

Ab initio structure prediction

For the current protocol to be a practical method, it should work without assuming the connectivity of TM helices since the connectivity of TM helices is as hard to establish as high- or medium-resolution structures. In other words, the prediction protocol should be run for all possible permutations of the connectivity of TM helices. Due to computational complexity, this is feasible only for TM helix bundles having less than 6 TM helices. Of the 5 test proteins, NtpK is the only one satisfying this criterion. Since it has just 4 TM helices, one has to deal with 3 ((4-1)!/2) permutations only. In our test, we used the permutations, ABCD, ABDC, ACBD. All together 1200 candidate models were generated (400 for each permutation), and 150 top-scoring ones were clustered to generate 5 final models, as described in the Methods section. Table 4 summarizes the results. Again, the rotational angles of TM helices about the helix axes were predicted with reasonable accuracy (Fig. 7-1). Remarkably, the accuracy of interhelical contact predictions reaches 55%. This level of accuracy can be considered exceptionally good, given the standards used in the CAPRI experiments.[59] We made a comparison to the standards of the CAPRI experiments because the current modeling is more like docking of predefined entities (ideal TM helices) rather than structure prediction from scratch. Furthermore, the model correctly predicts the Na^+ binding pocket, revealing a cluster of Thr64, Gln65, Gln110, and Glu139 in the structural core (Fig. 7-2). With the knowledge that NtpK functions as a pump of Na^+,[37] one can readily identify this cluster of highly polar residues in the core of a TM helix bundle as the binding site for a sodium ion. This demonstrates that structural models generated by the current prediction protocol might be useful for designing mutational experiments and rationalizing experimental data for simple TM helix bundle proteins having modest numbers of TM helices.

Discussion

Lessons from the structure prediction of soluble proteins indicate that effective structure prediction techniques can be devised only after we understand the sequence-structure relationship by seeing enough numbers of structures. So far we have seen just ~ 30 distinct folds of membrane proteins, which is clearly not enough. The lack of understanding about the principles of membrane protein folding further complicates the structure prediction problem. Therefore, it is expected to be extremely difficult to develop a generic structure prediction technique for membrane proteins in the near future. Yet, as claimed by many researchers over the years based on the well-known two-stage model, structural modeling of simple membrane proteins, i.e., those that have the *simple* form of *compact* TM helix bundles, can be accomplished in two steps: prediction of the portions of the primary sequence that traverse the membrane and assembly of these TM helices to TM helix bundles. Yet, this claim has never been realized for polytopic membrane proteins, even for simple ones as defined above. To our knowledge, the current study is the first tackling *ab initio* structure prediction of polytopic membrane proteins with concrete positive results. Specifically, it demonstrates that sequence conservation patterns enable us to generate native-like structural models for simple TM helix bundle proteins consisting of less than 6 TM helices. Among 5 final models, there always exists at least one model close to the crystal structure, not only in terms of CA RMSD but also in terms of helix-helix contacts.

One may wonder whether it is feasible to single out the best model out of 5 final ones. Table 5 summarizes the properties of the 5 final models for the test proteins in terms of their cluster size and score as computed by Eq. 1. In some cases, the best model is either the largest cluster or a best-scoring model. Yet, it is not the case for NtpK. Thus, it is not clear whether these indicators actually reflect the genuine difference between correct and wrong models. Furthermore, since we could not carry out a comprehensive benchmark test, we feel that it does not make much sense to look for any discriminating features between correct and wrong models.

As in soluble protein structure prediction,[2,3] the presented method depends on many parameters, i.e., how many top-scoring conformations to cluster into how many representative conformations, how many top-scoring folds to cluster into how many candidate models, how many top-scoring candidate models to cluster into how many final models. Yet, as noted above several times, the results did not critically depend on the particular choices of the parameters: any reasonable choices other than the current ones seem fine. For this reason, we believe that the presented method should be generally applicable to TM helix bundles with less than 6 TM helices, even though, due to the lack of suitable test proteins, we could not carry out large-scale benchmark tests as in soluble protein structure prediction.

The attractive feature of the current approach is that it only requires the availability of ~ 30 homologous sequences. Given rapid increases in the size of sequence databases, we believe that this requirement is relatively light. A couple of recent studies showed that TM helix bundle proteins with less than 6 TM helices are quite prevalent in the sequenced genomes.[20,60] Thus, we expect the presented prediction protocol to be helpful in elucidating the structure and function of simple TM helix bundle proteins. One particular example would be proteins of the tetraspanin family, which play diverse roles in cell adhesion, migration, and cellular activation and signaling.[61,62]

As pointed out by a number of previous studies,[63-65] distortions in the ideal helix geometry are abundant in TM helix bundle proteins, and they often play an important role in maintaining the functional and structural integrity. Thus, a more meaningful structure prediction should consider such aspects. Yet, we feel that such delicate structural aspects are extremely difficult to model with a reasonable accuracy. Clearly, more work is needed in this aspect of structural modeling. On the other hand, structural models obtained by the current approach might be of help for cases where canonical helices are a reasonably good approximation to real structures, as shown by the NtpK test case.

Acknowledgement

This work was supported by grant I/79534 of the Volkswagen Foundation. We acknowledge the extensive use of the BALL library.[48] Figures 4 and 6 were generated using the program Rasmol.[66]

References

1. Venclovas C, Zemla A, Fidelis K, Moult J. Assessment of progress over the CASP experiments. Proteins: Struct Funct Bioinformatics, 2003;53:585-595.
2. Bradley P, Chivian D, Meiler J, Misura KM, Rohl C, Schief W, Wedemeyer WJ, Schueler-Furman O, Murphy P, Schonbrun J, Strauss C, Baker D. Rosetta predictions in CASP5: Successes, failures, and prospects for complete automation. Proteins: Struct Funct Bioinformatics, 2003;53:457-468.
3. Zhang Y, Skolnick J. Automated structure prediction of weakly homologous proteins on a genomic scale. Proc Natl Acad Sci USA 2004;101:7594-7599.
4. White SH, von Heijne G. The machinery of membrane protein assembly. Curr Opin Struct Biol 2004;14:397-404.
5. Goder V, Junne T, Spiess M. Sec61p contributes to signal sequence orientation according to the positive-inside rule. Mol Biol Cell 2004;15:1470-1478.
6. Goder V, Spiess M. Molecular mechanism of signal-sequence orientation in the endoplasmic reticulum. EMBO J 2003;22:3645-3653.
7. Van der Berg B, Clemons WMJ, Collinson I, Modis Y, Hartmann E, Harrison SC, Rapoport TA. X-ray structure of a protein-conducting channel. Nature 2003;427:36-44.
8. White SH, Wimley WC. Membrane protein folding and stability: physical principles. Annu Rev Biophys Biomol Struct 1999;28:319-365.
9. Popot JL, Engelman DM. Helical membrane protein folding, stability, and evolution. Annu Rev Biochem 2000;69:881-922.
10. Hessa T, Kim H, Bihlmaier K, Lundin C, Boekel J, Andersson H, Nilsson I, White SH, von Heijne G. Recognition of transmembrane helices by the endoplasmic reticulum translocon. Nature 2005;433:377-381.
11. Sadlish H, Pitonzo D, Johnson AE, Skach WR. Sequential triage of transmembrane segments by Sec61alpha during biogenesis of a native multispanning membrane protein. Nat Struct Mol Biol 2005;12:870-878.
12. White SH. Translocons, thermodynamics, and the folding of membrane proteins. FEBS Lett 2003;555:116-121.
13. Dutzler R, Campbell EB, Cadene M, Chait BT, MacKinnon R. X-ray structure of a ClC chloride channel at 3.0 A reveals the molecular basis of anion selectivity. Nature 2002;415:287-294.
14. Jiang Y, Lee A, Chen J, Ruta V, Cadene M, Chait BT, MacKinnon R. X-ray structure of a voltage-dependent K+ channel. Nature 2003;423:33-41.
15. Sui H, Han BG, Lee JK, Walian P, Jap BK. Structural basis of water-specific transport through the AQP1 water channel. Nature 2001;414:872-878.
16. Popot JL, Engelman DM. Membrane protein folding and oligomerization: the two-

stage model. Biochemistry 1990;29:4031-4037.

17. von Heijne G. Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule. J Mol Biol 1992;225:487-494.

18. Jones DT, Taylor WR, Thornton JM. A model recognition approach to the prediction of all-helical membrane protein structure and topology. Biochemistry 1994;33:3038-3049.

19. Rost B, Casadio R, Fariselli P, Sander C. Transmembrane helices predicted at 95% accuracy. Protein Sci 1995;4:521-533.

20. Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 2001;305:567-580.

21. Liakopoulos TD, Pasquier C, Hamodrakas SJ. A novel tool for the prediction of transmembrane protein topology based on a statistical analysis of the SwissProt database: the OrienTM algorithm. Protein Eng 2001;14:387-390.

22. Tusnady GE, Simon I. The HMMTOP transmembrane topology prediction server. Bioinformatics 2001;17:849-850.

23. Treutlein HR, Lemmon MA, Engelman DM, Brunger AT. The glycophorin A transmembrane domain dimer: sequence-specific propensity for a right-handed supercoil of helices. Biochemistry 1992;31:12726-12732.

24. Adams PD, Engelman DM, Brunger AT. Improved prediction for the structure of the dimeric transmembrane domain of glycophorin A obtained through global searching. Proteins: Struct Funct Bioinformatics, 1996;26:257-261.

25. Kim S, Chamberlain AK, Bowie JU. A simple method for modeling transmembrane helix oligomers. J Mol Biol 2003;329:831-840.

26. Briggs JA, Torres J, Arkin IT. A new method to model membrane protein structure based on silent amino acid substitutions. Proteins: Struct Funct Bioinformatics, 2001;44:370-375.

27. Baldwin JM, Schertler GF, Unger VM. An alpha-carbon template for the transmembrane helices in the rhodopsin family of G-protein-coupled receptors. J Mol Biol 1997;272:144-164.

28. Trabanino RJ, Hall SE, Vaidehi N, Floriano WB, Kam VW, Goddard WA, III. First principles predictions of the structure and function of g-protein-coupled receptors: validation for bovine rhodopsin. Biophys J 2004;86:1904-1921.

29. Vaidehi N, Floriano WB, Trabanino R, Hall SE, Freddolino P, Choi EJ, Zamanakos G, Goddard WA, III. Prediction of structure and function of G protein-coupled receptors. Proc Natl Acad Sci USA 2002;99:12622-12627.

30. Faulon JL, Sale K, Young M. Exploring the conformational space of membrane protein folds matching distance constraints. Protein Sci 2003;12:1750-1761.

31. Sale K, Faulon JL, Gray GA, Schoeniger JS, Young MM. Optimal bundling of transmembrane helices using sparse distance constraints. Protein Sci 2004;13:2613-2627.

32. Luecke H, Schobert B, Richter HT, Cartailler JP, Lanyi JK. Structure of bacteriorhodopsin at 1.55 A resolution. J Mol Biol 1999;291:899-911.

33. Abramson J, Smirnova I, Kasho V, Verner G, Kaback HR, Iwata S. Structure and

mechanism of the lactose permease of Escherichia coli. Science 2003;301:610-615.

34. Kolbe M, Besir H, Essen LO, Oesterhelt D. Structure of the light-driven chloride pump halorhodopsin at 1.8 A resolution. Science 2000;288:1390-1396.

35. Luecke H, Schobert B, Lanyi JK, Spudich EN, Spudich JL. Crystal structure of sensory rhodopsin II at 2.4 angstroms: insights into color tuning and transducer interaction. Science 2001;293:1499-1503.

36. Palczewski K, Kumasaka T, Hori T, Behnke CA, Motoshima H, Fox BA, Le Trong I, Teller DC, Okada T, Stenkamp RE, Yamamoto M, Miyano M. Crystal structure of rhodopsin: A G protein-coupled receptor. Science 2000;289:739-745.

37. Murata T, Yamato I, Kakinuma Y, Leslie AG, Walker JE. Structure of the rotor of the V-Type Na+-ATPase from Enterococcus hirae. Science 2005;308:654-659.

38. Beuming T, Weinstein H. A knowledge-based scale for the analysis and prediction of buried and exposed faces of transmembrane domain proteins. Bioinformatics, 2004;20:1822-1835.

39. Adamian L, Nanda V, Degrado WF, Liang J. Empirical lipid propensities of amino acid residues in multispan alpha helical membrane proteins. Proteins 2005;59:496-509.

40. Tusnady GE, Dosztanyi Z, Simon I. PDB_TM: selection and membrane localization of transmembrane proteins in the protein data bank. Nucleic Acids Res 2005;33 (Database issue):D275 - D278.

41. Park Y, Elsner M, Staritzbichler R, Helms V. A novel scoring function for modeling structures of oligomers of transmembrane alpha-helices. Proteins: Struct Funct Bioinformatics, 2004;57:577-585.

42. Bowie JU. Helix packing in membrane proteins. J Mol Biol 1997;272:780-789.

43. Canutescu AA, Shelenkov AA, Dunbrack RLJ. A graph-theory algorithm for rapid protein side-chain prediction. Protein Sci 2003;12:2001-2014.

44. Kelley LA, Gardner SP, Sutcliffe MJ. An automated approach for clustering an ensemble of NMR-derived protein structures into conformationally related subfamilies. Protein Eng 1996;9:1063-1965.

45. Eilers M, Shekar SC, Shieh T, Smith SO, Fleming PJ. Internal packing of helical membrane proteins. Proc Natl Acad Sci USA 2000;97(5796-5801).

46. Donnelly D, Overington JP, Ruffle SV, Nugent JH, Blundell TL. Modeling alpha-helical transmembrane domains: the calculation and use of substitution tables for lipid-facing residues. Protein Sci 1993;2:55-70.

47. Stevens TJ, Arkin IT. Substitution rates in alpha-helical transmembrane proteins. Protein Sci 2001;10:2507-2517.

48. Kohlbacher O, Lenhof HP. BALL--rapid software prototyping in computational molecular biology. Biochemicals Algorithms Library. Bioinformatics 2000;16:815-824.

49. Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS. The Universal Protein Resource (UniProt). Nucleic Acids Res 2005;33:154-159.

50. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403-410.

51. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389-3402.

52. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 1994;22:4673-4680.

53. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res 2003;31:365-370.

54. Pei J, Grishin NV. AL2CO: calculation of positional conservation in a protein sequence alignment. Bioinformatics 2001;17:700-712.

55. Henikoff S, Henikoff JG. Position-based sequence weights. J Mol Biol 1994;243:574-578.

56. Moult J, Fidelis K, Zemla A, Hubbard T. Critical assessment of methods of protein structure prediction (CASP)-round V. Proteins: Struct Funct Bioinformatics, 2003;6:334-339.

57. Beuming T, Weinstein H. Modeling membrane proteins based on low-resolution electron microscopy maps: a template for the TM domains of the oxalate transporter OxlT. Protein Eng Des Sel 2005;18:119-125.

58. Fleishman SJ, Harrington S, Friesner RA, Honig B, Ben-Tal N. An automatic method for predicting transmembrane protein structures using cryo-EM and evolutionary data. Biophys J 2004;87:3448-3459.

59. Janin J, Henrick K, Moult J, Eyck LT, Sternberg MJ, Vajda S, Vakser I, Wodak SJ. CAPRI: a Critical Assessment of PRedicted Interactions. Proteins: Struct Funct Bioinformatics 2003;52:2-9.

60. Daley DO, Rapp M, Granseth E, Melen K, Drew D, von Heijne G. Global topology analysis of the Escherichia coli inner membrane proteome. Science 2005;308:1321-1323.

61. Kovalenko OV, Metcalf DG, DeGrado WF, Hemler ME. Structural organization and interactions of transmembrane domains in tetraspanin proteins. BMC Struct Biol 2005;5:11.

62. Hemler ME. Tetraspanin functions and associated microdomains. Nat Rev Mol Cell Biol 2005;6:801-811.

63. Riek RP, Rigoutsos I, Novotny J, Graham RM. Non-alpha-helical elements modulate polytopic membrane protein architecture. J Mol Biol 2001;306:349-362.

64. Yohannan S, Faham S, Yang D, Whitelegge JP, Bowie JU. The evolution of transmembrane helix kinks and the structural diversity of G protein-coupled receptors. Proc Natl Acad Sci USA 2004;101:959-963.

65. Sansom MS, Weinstein H. Hinges, swivels and switches: the role of prolines in signalling via transmembrane alpha-helices. Trends Pharmacol Sci 2000;21:445-451.

66. Sayle RA, Milner-White EJ. RASMOL: Biomolecular graphics for all. Trends

Biochem Sci 1995;20:374.

Table 1. Positions of the interaction centers

Amino acid	Position of the interaction center
Gly	CA atom
Ala, Ser, Cys	0.6*CA atom + 0.4*CB atom
Val, Thr	0.1*CA atom + 0.9*CB atom
Ile	CB atom
Pro	0.8*CA atom + 0.2*CG atom
Leu, Asp, Asn	0.3*CA atom + 0.7*CG atom
His, Phe, Glu, Gln	CG atom
Met	0.5*CA atom + 0.5*SD atom
Tyr	0.3*CG atom + 0.7*CZ atom
Trp	0.5*CG atom + 0.5*CD2 atom
Lys, Arg	CG atom + 0.2* the vector from CA to CG

0.8*CA atom + 0.2*CB atom means a vectorial sum of 0.8 times the vector from an origin to the position of the CA atom and 0.2 times the vector from the same origin to the position of the CB atom.

Table 2. Summary of the prediction results

	Bacterorhodopsin				Halorhodopsin			
CA RMSD	3.0				3.9			
	Acc: 44%		Cov: 46%		Acc: 43%		Cov: 48%	
	Z shift	XY shift	Tilt	Rotation	Z shift	XY shift	Tilt	Rotation
TM helix A	0.4	0.4	20	24	0.9	1.8	16	19
TM helix B	0.7	2.0	7	57	1.8	2.3	1	27
TM helix C	0.6	1.9	13	7	0.9	2.8	7	1
TM helix D	0.1	1.6	11	3	0.1	2.6	11	32
TM helix E	1.0	2.3	2	30	2.1	4.9	12	44
TM helix F	0.8	2.9	5	1	1.8	2.3	7	27
TM helix G	0.2	1.9	2	2	0.4	2.6	8	5
	Sensory rhodopsin II				Rhodopsin			
CA RMSD	3.0				5.2			
	Acc: 45%		Cov: 51%		Acc: 21%		Cov: 28%	
	Z shift	XY shift	Tilt	Rotation	Z shift	XY shift	Tilt	Rotation
TM helix A	0.1	1.5	7	17	0.2	2.3	29	29
TM helix B	0.6	2.4	8	27	0.1	1.0	17	3
TM helix C	0.6	1.6	15	22	0.1	6.1	28	18
TM helix D	0.1	1.8	15	11	0.9	1.6	11	13
TM helix E	0.3	3.0	4	19	0.8	1.7	23	21
TM helix F	0.6	3.3	2	5	1.1	2.1	17	19
TM helix G	0.1	1.2	0	1	0.1	0.6	21	21

CA RMSD: the best (smallest) CA RMSD in angstroms of the 5 final models with respect to the corresponding crystal structure

Z shift: the difference in angstroms between the CA-based centers of predicted and experimental conformations along the membrane normal.

XY shift: the difference in angstroms between the CA-based centers of predicted and experimental conformations along the membrane plane.

Tilt: the angle difference in degrees between the helix axes of predicted and experimental conformations.

Rotation: the angle difference in degrees between the rotational angles of predicted and experimental conformations around the helix axes.

Acc: accuracy of the helix-helix contact predictions, i.e., TP/(TP+FP), where TP and FP stand for true and false predictions, respectively.

Cov: Coverage of the helix-helix contact predictions, i.e., the percentage of the contacts observed in the experimental structures that have been correctly predicted.

Table 3. Effects of inverting the scoring scheme based on sequence conservation patterns

	Bacterorhodopsin	Halorhodopsin
CA RMSD	7.6	7.0
Acc	0%	0%
Cov	0%	0%
	Sensory rhodopsin II	Rhodopsin
CA RMSD	6.6	7.5
Acc	1%	5%
Cov	1%	5%

Definitions for CA RMSD, Acc, and Cov are the same as those given in Table 2.

This is a negative result. As described in the text, it has been obtained by clustering the 500 highest-score folds and 50 highest-score candidate models (instead of the 500 lowest-score folds and 50 lowest-score candidate models as for Table 2). It demonstrates that sequence conservation patterns work as a powerful scoring function, discriminating between ‘good’ and ‘bad’ folds in the library.

Table 4. Summary of the *ab initio* structure prediction for NtpK

	NtpK			
CA RMSD	3.4			
	Acc: 55%		Cov: 23%	
	Z shift	XY shift	Tilt	Rotation
TM helix A	0.5	1.9	14	36
TM helix B	1.2	1.0	14	43
TM helix C	1.3	0.7	4	20
TM helix D	0.8	0.6	17	58

Definitions for CA RMSD, Acc, Cov, Z shift, XY shift, Tilt, and Rotation are the same as those given in Table 2.

Table 5. Properties of the 5 final models for the test proteins

	Bacteriorhodopsin		Halorhodopsin		Sensory rhodopsin II		Rhodopsin		NtpK	
Model	Cluster size	Score	Cluster size	Score	Cluster size	Score	Cluster size	Score	Cluster size	Score
1	17	-19.1	11	-16.0	12	-19.1	14	-23.4	5	-7.9
2	5	-19.9	5	-17.3	5	-15.9	11	-19.8	11	-8.9
3	6	-17.1	7	-17.8	4	-19.7	4	-20.5	17	-8.6
4	8	-18.1	11	-19.1	7	-18.7	9	-21.1	8	-8.0
5	7	-13.7	10	-11.9	16	-20.4	8	-23.1	12	-7.6

For each test protein, model 1 is the closest to the corresponding crystal structure in terms of CA RMSD. Score means sequence conservation scores computed by Eq. 1

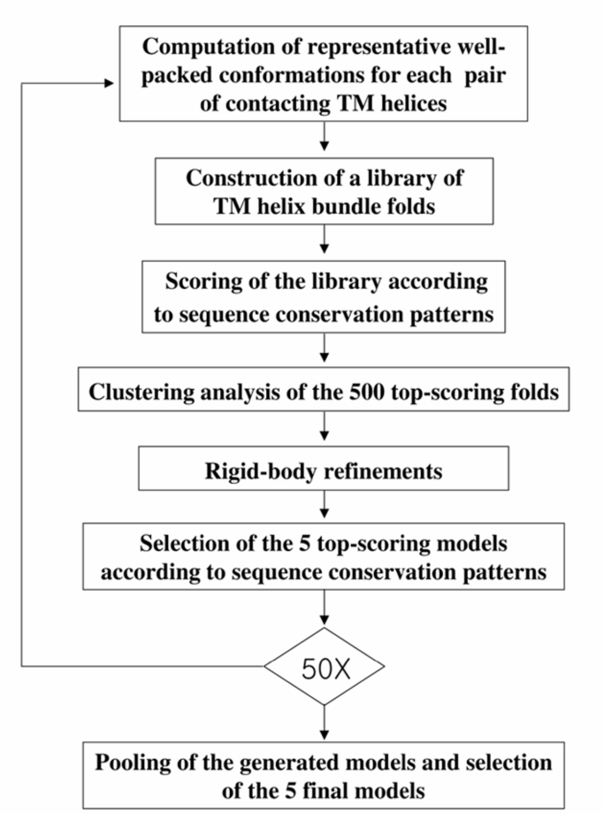

Computation of representative well-packed conformations for each pair of contacting TM helices
Construction of a library of TM helix bundle folds
Scoring of the library according to sequence conservation patterns
Clustering analysis of the 500 top-scoring folds
Rigid-body refinements
Selection of the 5 top-scoring models according to sequence conservation patterns
50X
Pooling of the generated models and selection of the 5 final models

Figure 1. The flowchart of the prediction protocol.

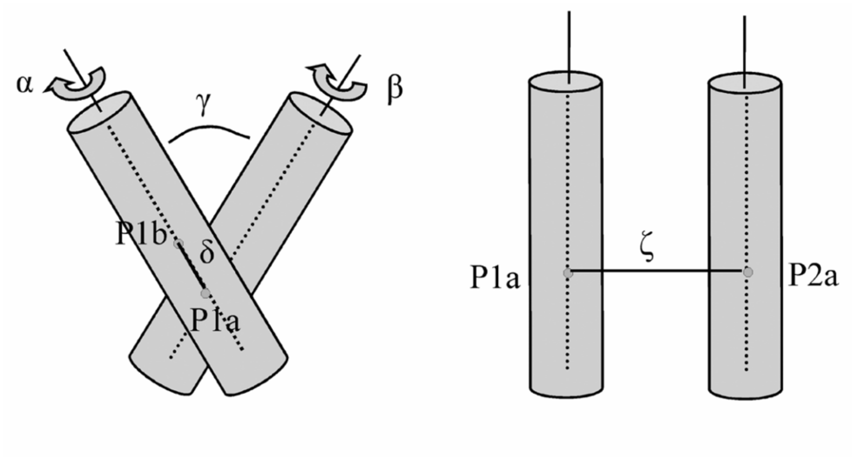


Figure 2. Degrees of freedom for a pair of contacting helices. α and β are the rotational angles about their helix axes. γ indicates the crossing angle between the two helix axes. δ describes the sliding motion along the respective helix axis, measuring the distance between P1a (the crossing point between the helix axis and the axis for γ) and P1b (geometric center of the helix). The remaining degree of freedom describes the motion along the line connecting the two closest points (P1a and P2a) on the helical axes, designated as ζ.

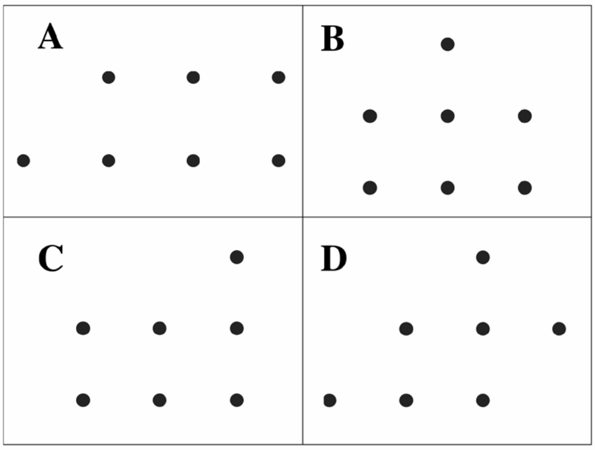


Figure 3. A few examples of compact conformations of a 7 TM helix bundle protein. Helix centers on a 2D grid are shown. The interhelical distance is set to 9.0 Å, and the compactness value is calculated as described in the Methods section. The results are as follows: (A) 14.55 Å, (B) 13.47 Å, (C) 14.08 Å, and (D) 14.65 Å. Thus 16.0 Å used in the present study is a reasonable upper bound.

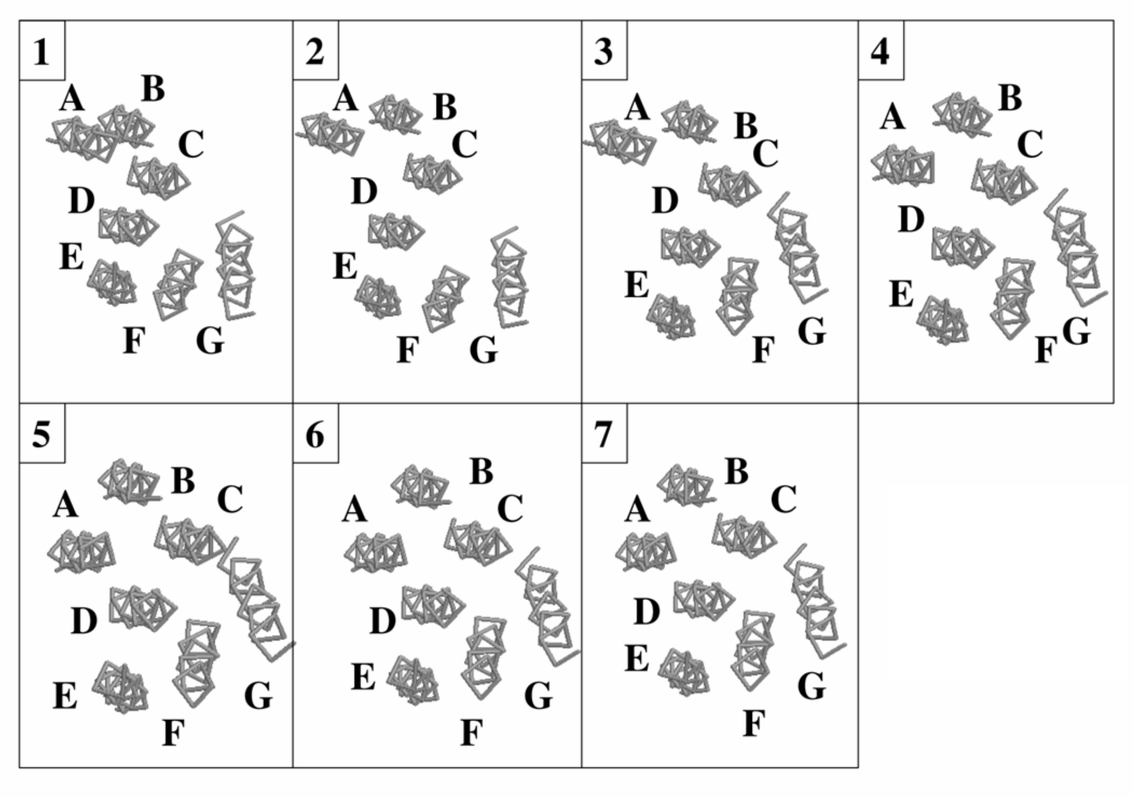


Figure 4. Rigid-body refinements. Panel 1 is the average structure, viewed from the N terminus. The distance between TM helices A and B is too short, resulting in several steric clashes, and TM helix G is overly exposed to the lipid bilayer. Overall, Panel 1 depicts an image untypical of simple well-packed TM helix bundle proteins. In order to correct for the deficiencies, we relax the interhelical distances and the 5 angles formed by the projection points onto the membrane plane of the CA-based centers of each of the 5 triples of neighboring TM helices (from ABC to EFG). All 6 interhelical distances are relaxed to 9.6 Å for the removal of any steric clashes, resulting in Panel 2. Out of the 5 angles, the one formed by TM helices DEF is the smallest. So we relax this angle, leading to Panel 3. Then the angles are relaxed in the order of ABC, BCD, CDE, EFG, resulting in Panels 4 - 7.

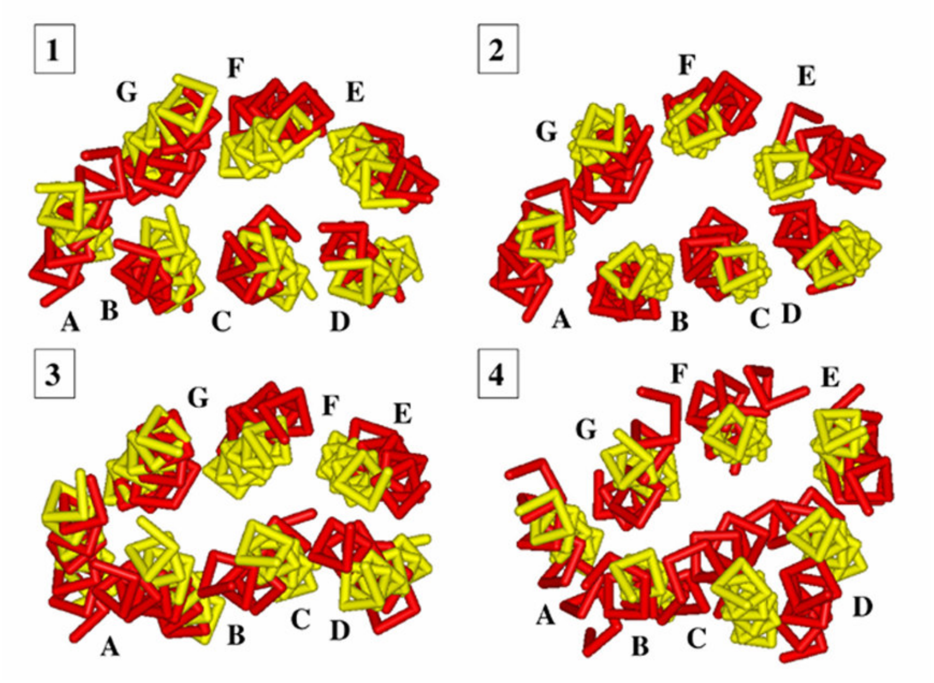


Figure 5. Superimposition of the best models (yellow) onto the corresponding experimental structures (red), viewed from the N-terminus. Panel 1 – bR, panel 2 – hR, panel 3 – sR, panel 4 – rhodopsin. Quantitative comparisons are presented in Table 2. This figure was created using the program Weblab viewer (MSI).

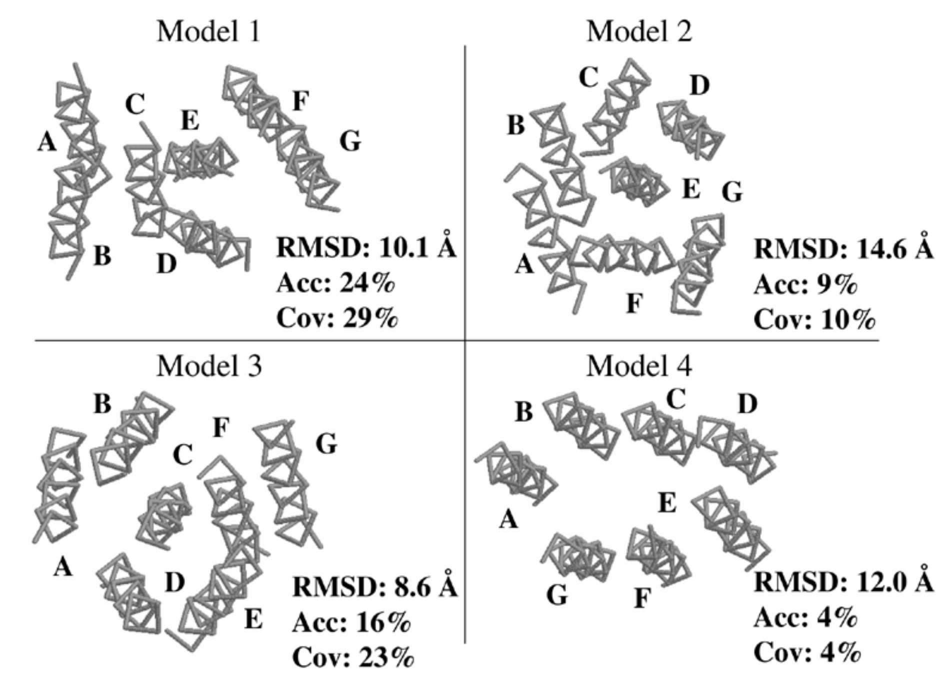


Figure 6. The 4 incorrect models for bacteriorhodopsin.

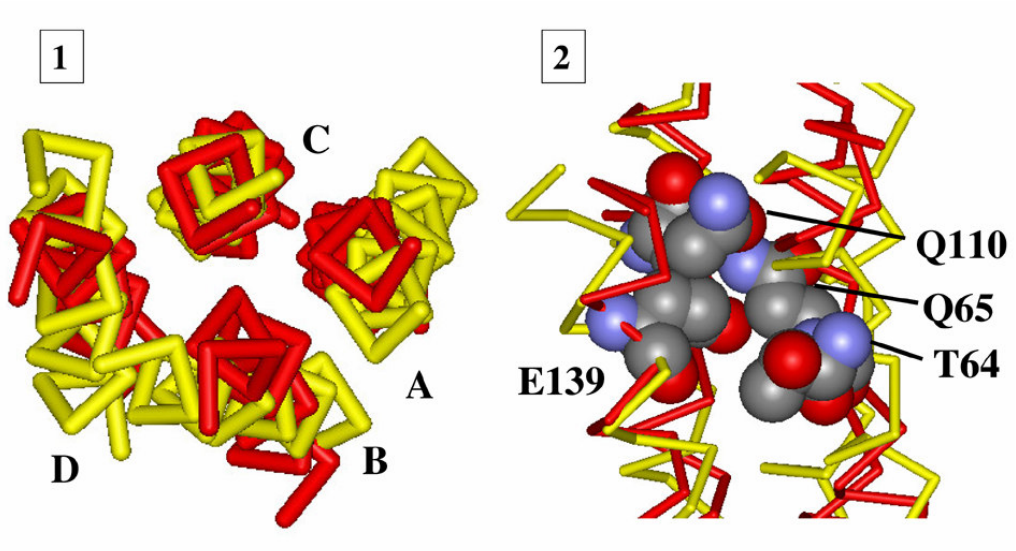


Figure 7. (1) Superimposition of the best model (yellow) onto the 2bl2 structure (red). Quantitative comparisons are presented in Table 4. (2) View of the predicted sodium binding pocket. Q65, Q110, and E139 are clustered in the core of the TM helix bundle, indicating the binding site for sodium ions. The side chain conformations were modeled by using SCWRL.[43]

9. Paper II

Park, Y. and Helms, V. Biopolymers (2006): 83, 389-399.

How Strongly do Sequence Conservation Patterns and Empirical Scales Correlate with Exposure Patterns of Transmembrane Helices of Membrane Proteins?

Title: How Strongly do Sequence Conservation Patterns and Empirical Scales Correlate with Exposure Patterns of Transmembrane Helices of Membrane Proteins?

Yungki Park and Volkhard Helms*

Center for Bioinformatics, Saarland University, Germany

*Corresponding author

Abstract

Given the difficulty in determining high-resolution structures of helical membrane proteins, sequence-based prediction methods can be useful in elucidating diverse physiological processes mediated by this important class of proteins. Predicting the angular orientations of transmembrane (TM) helices about the helix axes, based on the helix parameters from electron microscopy data, is a classical problem in this regard. This problem has triggered the development of a number of different empirical scales. Recently, sequence conservation patterns were also made use of for improved predictions. Empirical scales and sequence conservation patterns (collectively termed as "prediction scales") have also found frequent applications in other research areas of membrane proteins: for example, in structure modeling and in prediction of buried TM helices. This trend is expected to grow in the near future unless there are revolutionary developments in experimental characterization of membrane proteins. Thus, it is timely and imperative to carry out a comprehensive benchmark test over the prediction scales proposed so far to find out their pros and cons. In the current analysis, we use exposure patterns of TM helices as a golden standard, because if one develops a prediction scale that correlates perfectly with exposure patterns of TM helices, it will enable one to predict buried residues (or buried faces) of TM helices with an accuracy of 100%. Our analysis reveals several important points. First, it demonstrates that sequence conservation patterns are much more strongly correlated with exposure patterns of TM helices than empirical scales. Second, scales that were specifically parameterized using structure data (structure-based scales) display stronger correlation than hydrophobicity-based scales, as expected. Third, a non-negligible difference is observed among the structure-based scales in their correlational property, suggesting that not every learning algorithm is equally effective. Fourth, a straightforward framework of optimally combining sequence conservation patterns and empirical scales is proposed, which reveals that improvements gained from combining the two sources of information are not dramatic in almost all cases. In turn, this calls for the development of fundamentally different scales that capture the essentials of membrane protein folding for substantial improvements.

Introduction

Membrane proteins play a crucial role in diverse physiological processes, including signal transduction, the transport of solutes across the membrane, and the maintenance of ionic concentrations in the cell. The transmembrane (TM) domains of many membrane proteins consist of helix bundles. Their formation is believed to be driven, to a large degree, by the high desolvation penalty of exposing polar backbone atoms to the lipid bilayer.[1,2] In spite of recent technical advances,[3-5] it still remains difficult to determine high-resolution structures of helical membrane proteins via X-ray crystallography or NMR spectroscopy. In certain cases, yet, they are amenable to structure determination through cryo-electron microscopy (cryo-EM), which usually results in intermediate- or low-resolution structural information, i.e., the positions of TM helices, but not those of individual amino acids.[6,7] If the correspondence is established between TM segments of the sequence and densities on the EM map assumed to be TM helices with the help of experimental data, then computational tools for predicting the rotational angles of TM helices about the helix axes come into play. A recent study on gap junction channels highlighted how helpful computational methods of predicting the rotational angles of TM helices can be in elucidating important physiological processes.[7]

Early approaches for this classical problem were based on the concept of hydrophobic moments: they predicted the angular orientations of TM helices by aligning the hydrophobic moments to the lipid bilayer.[8,9] Various hydrophobicity scales were utilized for this purpose, (to cite a few, the Eisenberg consensus scale,[9] the Kyte-Doolittle scale,[10] the Goldman-Engelman-Steitz scale[11]) even though the aim of these scales was primarily to identify TM segments from the sequence. However, hydrophobic moments turned out to be a poor indicator of the angular rotations of TM helices.[12,13] This triggered the development of a new generation of empirical scales, including kProt,[14] the tmlip series,[15] and the Beuming-Weinstein scale.[16] The latest approaches rely on sequence conservation patterns as well as empirical scales for improved predictions, following the observation that conserved residues tend to be buried in the interior of TM helix bundles while variable residues tend to be exposed to the lipid bilayer.[17,18]

In addition to the problem of predicting the rotational angles of TM helices, empirical scales and sequence conservation patterns (collectively termed as "prediction scales") have found many applications in other research areas of membrane proteins as well. For example, Park *et al.* fruitfully used sequence conservation patterns as a scoring function in the context of free modeling, i.e., modeling tertiary structures without experimental constraints.[19] Sequence conservation patterns enabled us to model the structure of the rotor of the V-type Na^+-ATPase with an root-mean square deviance (RMSD) of 3.4 Å. A control computation that used scrambled sequence conservation patterns demonstrated that sequence conservation patterns played a crucial role in the positive results. Recently, Adamian and Liang have devised a protocol for predicting buried TM helices based on sequence conservation patterns and their own empirical scale.[15] In a benchmark test, they were able to predict 15 out of 19 buried helices and 39 out of 43 boundary helices, reaching an accuracy of ~ 90%.

Considering the difficulty in determining high-resolution structures of membrane proteins by experimental techniques, the prediction scales, as an inexpensive auxiliary tool, will find increasing applications in diverse areas in the future. Thus, it seems imperative, at the current point, to carry out a large-scale thorough benchmarking of the prediction scales proposed so far to find out their pros and cons. Even for the simplest question of which prediction scale is best for predicting lipid-exposed faces of helical membrane proteins, we do not have a clear answer. The reason is simply that it was not possible to undertake such a comprehensive benchmark test until recently due to lack of enough numbers of high-resolution structures.

One of the most important elements in performing an objective benchmark test is to choose a right golden standard. In comparing different prediction scales in the context of predicting lipid-exposed faces of TM helices, the golden standard should be exposure patterns of TM helices (see the Methods section). This is because if the correlation between the exposure patterns of TM helices and their positional scores calculated from a prediction scale is 1, then it enables one to predict buried/lipid-exposed faces of TM helices with an accuracy of 100%. Therefore, in the current analysis, we adhere to the rule that the more strongly a prediction scale is correlated with exposure patterns of TM helices, the better it is.

Methods

Two different ways of comparing the performance of different prediction scales can be conceived. One strategy considers the average angular errors of predicted rotational angles of TM helices[14] whereas the other one investigates how strongly prediction scales are correlated with exposure patterns of TM helices. In this paper, we adopt the second way because it is defined straightforwardly for all helical membrane proteins.

Benchmark proteins and computation of exposed patterns

A non-redundant set of TM helix bundle proteins (less than 25% sequence identity pairwise and with resolution better than 3.5 Å) was generated based on the lists of membrane proteins with known structure compiled by White (http://blanco.biomol.uci.edu) and by Michel (http://www.mpibp-frankfurt.mpg.de/michel/public/memprotstruct.html). Some proteins, even though satisfying the above criteria, were not included in the analysis for the following reasons: 1rhz_C (chain C of PDB ID 1rhz), 1occ_K, 1vf5_E, 1vf5_F, and 1ehk_C – not enough homologous sequences available; 1fft_C and 1l7v_A – defects in TM helices; 1jb0 and 2axt – due to idiosyncratic extensive protein-cofactor interactions throughout the protein structures. A final list of benchmark proteins is listed in Table 1.

Exposure patterns of TM helices of these benchmark proteins are represented by their relative solvent-accessible surface area (rSASA) values. The rSASA value of a residue was obtained by dividing its SASA value in the crystal structure by its reference value. The reference value of a given residue is its SASA value in the context of a tripeptide helix Gly-X-Gly, where X is the given residue. Thus, the exposure pattern of a TM helix is represented as a string of real numbers between 0 and 1. The probe radius was set to 2.0 Å to approximate the radius of CH_2 group. Calculation of SASAs was performed using the BALL library.[20]

Functionally relevant internal cofactors were included in computation of rSASA values. Since crystal structures of membrane proteins do not include hydrogen atoms, the program REDUCE was used to add hydrogen atoms.[21] Another point to be considered in computing rSASA values is whether to use a monomeric structure or an oligomeric structure. Our guiding principle was to use functionally obligatory oligomeric states because the tendency of conserved residues to be buried is thought to be due to functional and/or structural constraints. We note, however, that there are some ambiguities in a few cases. For example, the H^+/Cl^- exchanger (1KPL), the mitochondrial ADP/ATP carrier (1OKC), and the Na^+/H^+ antiporter (1ZCD) are all known to be a functionally dimer. However, due to different crystallization conditions, only the H^+/Cl^- exchanger exists as a dimer in the crystal structures. So, it appears that even though these proteins primarily function as a dimer *in vivo*, the dimerization is dependent on environmental conditions and not strictly obligatory for keeping their structural (and/or functional) integrity. Unless there is convincing evidence that oligomerization is functionally required (e.g., the dimeric forms of cytochrome bc_1 and b_6f complexes), we always chose the monomeric state. Table 1 lists the oligomeric states adopted for the benchmark proteins.

Since it only makes sense to apply hydrophobicity scales to the hydrophobic core of the lipid bilayer, the most hydrophobic 20 Å thick portion was taken for each of the test proteins, using the Eisenberg hydrophobicity scale[9] and the membrane normal vectors from the PDB_TM database.[22] As in most previous studies, we restrict our attention to boundary TM helices, with "boundary" defined as having an overall fractional exposure to the lipid bilayer of equal or greater than 0.05.

Derivation of positional scores for the prediction scales

Multiple sequence alignments (MSAs) are needed to estimate positional scores from the prediction scales. For each test protein, a maximum of 1000 similar sequences were retrieved from the non-redundant database (nr) in an automatic fashion using BLAST URLAPI (http://www.ncbi.nih.gov/BLAST) with all the parameters set to default values. Initial MSAs were then built by using ClustalW[23] with all the parameters set to default values. Then, sequence fragments (not longer than 80% of the length of the query sequence) were deleted from the MSA. This was crucial for improving the quality of final MSAs. From these refined MSA, 6 different percent identity criteria (from 40% to 15% in steps of 5%) were applied, yielding 6 final MSAs for each test protein. A recent analysis has shown that one needs to align at least 20 sequences to accurately estimate conservation indices from MSAs.[24] Thus, positional scores were computed only for cases of 20 or more sequences in the final MSA.

As described by Pei and Grishin,[24] four different methods were used to estimate conservation indices from MSAs. The first method is an entropy-based one (denoted hereafter as "ENT").

$$C(i) = \sum_{a=1}^{20} f_a(i) \ln f_a(i) \tag{1}$$

In Eq. 1, $C(i)$ is the conservation index for the sequence position i in a multiple sequence alignment, $f_a(i)$ is the frequency of amino acid a in the sequence position i. The second method is a variance-based measure (denoted hereafter as "VAR").

$$C(i) = \sqrt{\sum_{a=1}^{20} (f_a(i) - f_a)^2} \tag{2}$$

In Eq. 2, f_a is the overall frequency of amino acid a in the alignment. A position with $f_a(i)$ equal to f_a for all amino acids a is assigned $C(i) = 0$. On the contrary, $C(i)$ takes on its maximum for the position occupied by an invariant amino acid whose overall frequency in the alignment is low. The third and fourth methods are sum-of-pairs measures.

$$C(i) = \sum_{a=1}^{20} \sum_{b=1}^{20} f_a(i) f_b(i) S_{ab} \tag{3}$$

In Eq. 3, S_{ab} is an amino acid scoring matrix. In the third method (denoted hereafter as "SOP1"), S_{ab} is an identity matrix while, in the fourth method (denoted hereafter as "SOP2"), it is a BLOSUM62 matrix. Thus, unlike other methods, the fourth one takes into account conservative mutations in its computation of conservation indices. In order to account for the redundancy of aligned sequences, amino acid frequencies were weighted in all four methods by using a modified method of Henikoff and Henikoff as implemented in PSI-BLAST.[25,26] Actual calculations were performed using a program written by Pei and Grishin,[24] which is freely available at ftp://iole.swmed.edu/pub/al2co.

For empirical scales, positional scores were computed in an analogous way.

$$score(i) = \sum_{a=1}^{20} f_a(i) \cdot scale(a) \tag{4}$$

In total, 8 empirical scales were considered in the analysis. They are the White-Wimley octanol scale (denoted hereafter as “ww”),[27] the Eisenberg hydrophobicity scale (“eis”),[9] the Kyte-Doolittle hydrophobicity scale (“kd”),[10] the kProt central version (“kP”),[14] the Beuming-Weinstein knowledge-based scale (“bw”),[16] the Goldman-Engelman-Steitz scale (“ges”),[11] and the TMLIP1 and 2 scales.[28]

Analysis of correlation between the prediction scales and exposure patterns of TM helices

Correlation coefficients (CCs) for a set of *n* data points (x_i, y_i) were computed as follows.

$$cc = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{[n\sum x_i^{\,2} - (\sum x_i)^2][n\sum y_i^{\,2} - (\sum y_i)^2]}} \tag{5}$$

To analyze the periodicity of the prediction scales, the discrete Fourier transform power spectra, *P(w)*, was calculated as previously described.[29-31]

$$P(w) = (\sum_{j=1}^{S}(score(j) - \overline{score})\cos(jw))^2 + (\sum_{j=1}^{S}(score(j) - \overline{score})\sin(jw))^2 \tag{6}$$

In Eq. 6, *S* is the number of residues in the TM helix being analyzed, *score*(*j*) is the positional score for the *j*th sequence position (computed by either Eqs. 1 – 4), *w* is the rotational angle between residues around a helix axis, and $\overline{score}$ is the mean score of the TM helix being analyzed. The α-helical character, Ψ, of the *P(w)* curve can be quantified as

$$\Psi = 6\int_{85^\circ}^{115^\circ} P(w)dw \bigg/ \int_{0^\circ}^{180^\circ} P(w)dw \tag{7}$$

An ideal helix with 3.6 residues per turn yields an angle of 360°/3.6 = 100° between successive residues. Thus, the greater the fraction of the P(w) curve that is in the α-helical region (85° - 115°), the greater the resultant Ψ. In other words, the more conforming the prediction scale to the helical periodicity, the greater its α-helical character (Ψ). The normalization factor 6 in Eq. 7 accounts for the difference in integration ranges.

Support vector machine

For predicting buried residues of the test proteins, the support vector machine (SVM) as implemented in the e1071 package[32] of R[33] was used. Unlike the usual usage of SVM, the current study employs SVM as an assessing tool of the prediction scales. Specifically, training errors of a linear SVM are interpreted as directly reflecting the quality of a prediction scale in question.

Results and Discussion

Overall correlation

Table 2 summarizes the overall correlations between exposure patterns of boundary TM helices and the prediction scales. Of the 12 prediction scales considered in the analysis, sequence conservation pattern derived from the variance-based method (Eq. 2) is the most strongly correlated. It is observed that, in general, sequence conservation patterns, derived from any of the four different methods, are always better than empirical scales, demonstrating the superiority of sequence conservation patterns over empirical scales. Significant differences in the correlational property of the 4 sequence conservation scales suggest that not every algorithm estimating conservation indices from a given MSA is equally effective. The poor performance of SOP2 over the simpler ENT or VAR scales is contrary to the naive expectation that it should be better because it takes conservative mutations into account in the derivation of conservation indices. The 8 empirical scales can be divided into 3 groups: a group of classical hydrophobicity scales (kd, eis, ges, ww), kP, and a group of structure-based scales (bw, tmlip1, tmlip2). Among the 8 empirical scales, the structure-based scales are expected to be better than others because they were specifically parameterized using structure data for the purpose of predicting buried/lipid-exposed faces of TM helices. Other scales were developed before high-resolution structural data became available. In addition, the main purpose of classical hydrophobicity scales was to identify TM segments from the primary sequence. Not surprisingly, Table 2 shows that bw is the best among the 8 empirical scales. tmlip1 and tmlip2 are slightly worse than bw. We suspect that this might be due to different assumptions behind their derivation.[28] The discrepancy between bw and tmlip1/tmlip2 again suggests that not every learning algorithm is equally effective. Relating to this point, it is noteworthy that ww, eis, kd are as good as tmlip1 in this comparison. On the other hand, these observations raise the following interesting question: would there then be an even better way of parameterization than that used in the derivation of bw? This point will be addressed in future work. Another noteworthy point in Table 2 is the poor performance of kP (the kProt central version). Unlike the classical hydrophobicity scales, kP was developed, based on sequence data, for the purpose of predicting buried/lipid-exposed faces of TM helices. In its original publication,[14] it was argued to be better than kd, eis, ges, based on the structure data available that time. (Comparison to the ww scale was not made.) Table 2 reveals that even though it performs slightly better than ges, it is not as effective as kd and eis. This points out the importance of a large-scale benchmarking. The overall poor correlation between exposure patterns of TM helices and the 4 hydrophobicity scales corroborates a previous conclusion that membrane proteins are not “inside-out” proteins.[13]

As shown in Fig. 1, the overall correlation of sequence conservation patterns with exposure patterns of TM helices appears rather weak. Due to different quality of MSAs and possibly different functional and structural constraints to which each test protein are subject, the degree of correlation between conservation and exposure to the lipid bilayer may not be the same for all test proteins. Table 3 lists the CCs for individual proteins based on 35% MSAs. Apart from few outliers (those with a CC of lower than 0.15), CCs range over 0.3 ~ 0.7. It is interesting to note that the outliers are all from dimeric structures of analogous proteins. However, inspection of the outliers did not reveal any particular discernable characteristics: the numbers of sequences in the final alignment are much higher than 20 (the cutoff number); some consist of three TM helices while some are single TM helices. Some of these outliers disappear when using different MSAs such as 30% or 25% MSAs (data not shown), which indicates that the quality of MSAs is partly responsible for these poor correlations. Then, it would be natural to further characterize the quality of MSAs by computing confidence intervals of the derived scores using bootstrap techniques.[28] Unfortunately, it is prohibitively expensive in the current context. Instead, we characterize the quality of MSAs in a way that takes advantage of the fact that we are analyzing boundary TM helices. Unlike buried TM helices, the rSASA values of boundary ones display a periodic pattern reflecting the periodic pattern of their environmental heterogeneity. Thus, a prediction scale that exhibits periodic patterns along boundary TM helices is expected to correlate strongly with their exposure patterns. The α-helical characters of the prediction scales (Eq. 7) quantitatively capture how close their periodic patterns are to the α-helical pattern. Table 4 shows that there is a significant correlation between α-helical characters of the prediction scales and their degrees of correlation with exposure patterns of TM helices. It is more pronounced for sequence conservation patterns than for empirical scales. This observation is important because it suggests that α-helical characters of MSAs can be used as a reliability measure. When it is not clear which cutoff value to choose in generating MSAs, as usual, it might be sensible to choose a cutoff value from which the MSAs of largest α-helical characters are generated. Figure 2 shows that this hypothesis holds in most cases.

Combining sequence conservation patterns and empirical scales in an optimal way

Sequence conservation patterns and empirical scales may represent information of different nature (ideally, complementary nature) about the propensity of residues of TM helices to get exposed to the lipid bilayer. Then, combining the two sources of information may lead to improved predictions. In fact, two recent studies exploited this idea. In the study by Fleishman *et al.*,[34] an empirical scoring function combining sequence conservation patterns and their own hydrophobicity scale was parameterized against bacteriorhodopsin crystal structures. In another study, Beuming and Weinstein derived a knowledge-based propensity scale for each type of amino acids to be exposed to the lipid bilayer, to be used in conjunction with sequence conservation patterns for predicting which residues of a TM helix are exposed to the lipid bilayer or buried in the interior of TM helix bundles.[16,35] Due to their empirical nature, it is hard to grasp how optimal their formulations are. However, the two studies concluded that sequence conservation patterns were more crucial to the positive results.

In our context of computing CCs, it is rather straightforward how to combine different sources of information in an optimal manner: we combine different scales such that the combined scales correlate with exposure patterns of TM helices as strongly as possible. Using gradient-descent optimization for this task, we generated every possible binary combination of the 12 prediction scales. Table 5 shows their CCs with exposure patterns of TM helices. Several points are noteworthy in Table 5. First, the best combinations are sequence conservation patterns plus structure-based scales (specifically, the bw scale). Second, for these best combinations, the improvements gained are not dramatic, as previously reported by Beuming and Weinstein.[16] Third, no dramatic improvements are observed in almost all cases, suggesting that sequence conservation patterns and empirical scales are not complementary to each other. We interpret this as implying that fundamentally different propensity scales that capture the essentials of membrane protein folding are needed for substantial improvements. Fourth, the ww scale, which correlates by itself poorly with exposure patterns of TM helices, is as effective as the bw scale as auxiliary information when combined with sequence conservation patterns.

Practical applications

The above analysis indicated that sequence conservation patterns are more strongly correlated with exposure patterns of boundary TM helices than empirical scales. Also, an optimal way of combining the two sources of information was proposed, which revealed that the combination of the VAR and bw scales is most optimal. Its CC is 0.44 whereas that of the VAR scale alone is 0.40 and that of the bw scale alone is 0.26. Practically, what does this difference of 0.04 in the CC values mean? It might be that the VAR scale with a CC of 0.40 and the kd scale with a CC of 0.21 are equally bad in practical prediction settings. To investigate this issue, we made use of a linear support vector machine (LSVM). Due to its linearity, training errors of LSVM are thought to directly reflect the quality of the prediction scales.

So far, we restricted the analysis to boundary TM helices. However, it is usually difficult to reliably distinguish between buried and boundary TM helices from the sequence, even though progress is being made in this problem.[15] Thus, in making predictions for buried residues (those with a rSASA of $\leq$ 0.05), we included all the residues located in the hydrophobic core of the lipid bilayer to mimic real situations. This is also necessary for this analysis to be comparable to a previous study.[16] The 35% MSAs of the test proteins were taken, which resulted in a total of 4058 data points. Of these, 2007 and 2051 data points correspond to exposed and buried residues, respectively. Thus, a prediction accuracy of 51% is the baseline. A point to be clarified in this sort of analysis is with what criterion to classify a given residue as buried. Different criteria such as rSASA of 0 or rSASA of 0.07 were used in previous studies.[16,28] How to choose suitable criteria appears to depend on how to compute rSASA values. In our case, an rSASA value of 0.05 looks appropriate. Figure 3 shows the distribution of the rSASA values of the 4058 data points, which reveals that a proper binary classification should be the one separating the data points at the rSASA value of 0.05.

Table 6 summarizes the results, revealing a number of interesting points. First, in spite of the different natures of the two measures of assessing the quality of the prediction scales (computing CC values and the binary classification of the burial state of residues), a comparison with Table 5 indicates that there is an overall agreement between the two measures. The best performing scales in Table 6 (sequence conservation patterns plus either ww or bw) are also the ones with highest CC values in Table 5. The kP and ges scales, which most poorly correlate with exposure patterns of TM helices, perform most poorly in predicting the binary burial states of residues of TM helices. In fact, the performance of these two scales, when used alone, is no better than the baseline accuracy (51%). Second, as in Table 5, the ww scale is as helpful as the structure-based scales (bw, tmlip1 and tmlip2) as auxiliary information when combined with sequence conservation patterns. A possible explanation for this is that the ww scale captures genuine differences between buried and exposed residues of TM helices, and this gets pronounced when optimally combined with sequence conservation patterns. Relating to this, it is interesting to note the following characteristics of the ww scale: unlike other empirical scales, it is an experimentally determined one, taking into account contributions from backbone atoms. A recent study demonstrated that a biological hydrophobicity scale that was measured *in vitro* strongly correlates with the ww scale.[36] Beuming and Weinstein reported that their empirical formulation combining sequence conservation patterns and the bw scale enabled them to achieve a prediction accuracy of ~ 75% on a test set of 11 proteins.[16] In this work, optimal combinations of either ENT, VAR, or SOP1 with bw yielded a similar prediction accuracy of

74% on the same subset of proteins (data not shown). This implies that the formulation suggested by Beuming and Weinstein is almost optimal.

Conclusion

The current study reports a comprehensive benchmark test of the prediction scales proposed thus far in terms of their correlations with exposure patterns of boundary TM helices. Sequence conservation patterns are shown to display a stronger correlation than empirical scales. Yet, overall correlations are poor for both classes of prediction scales. In addition, the two classes of prediction scales are not complementary to each other, and thus combining the two sources of information yields only marginal improvements. This calls for the development of scales of fundamentally different nature for substantial improvements.

Acknowledgments

This work was supported by grant I/79534 of the Volkswagen Foundation. We gratefully acknowledge the extensive use of the BALL library.[20] We thank Sarel Fleishman and Nir Ben-Tal for correspondence about their data in ref. 34, and Thijs Beuming and Harel Weinstein for sharing their program for computing conservation indices and a preprint of ref. 35 ahead of publication. We also thank crystallographers of membrane proteins because the current study was impossible without their work.

References

1. Popot, J. L.; Engelman, D. M. Biochemistry 1990, 29, 4031-4037.

2. White, S. H.; Wimley, W. C. Annu Rev Biophys Biomol Struct 1999, 28, 319-365.

3. Faham, S.; Bowie, J. U. J Mol Biol 2002, 316, 1-6.

4. Takeda, K.; Sato, H.; Hino, T.; Kono, M.; Fukuda, K.; Sakurai, I.; Okada, T.; Kouyama, T. J Mol Biol 1998, 283, 463-474.

5. Qutub, Y.; Reviakine, I.; Maxwell, C.; Navarro, J.; Landau, E. M.; Vekilov, P. G. J Mol Biol 2004, 343, 1243-1254.

6. Heymann, J. A.; Sarker, R.; Hirai, T.; Shi, D.; Milne, J. L.; Maloney, P. C.; Subramaniam, S. EMBO J 2001, 20, 4408-4413.

7. Fleishman, S. J.; Unger, V. M.; Yeager, M.; Ben-Tal, N. Mol Cell 2004, 15, 879-888.

8. Eisenberg, D.; Weiss, R. M.; Terwilliger, T. C. Nature 1982, 299, 371-374.

9. Eisenberg, D.; Schwarz, E.; Komaromy, M.; Wall, R. J Mol Biol 1984, 179, 125-142.

10. Kyte, J.; Doolittle, R. F. J Mol Biol 1982, 157, 105-132.

11. Engelman, D. M.; Steitz, T. A.; Goldman, A. Annu Rev Biophys Biophys Chem 1986, 15, 321-353.

12. Cronet, P.; Sander, C.; Vriend, G. Protein Eng 1993, 6, 59-64.

13. Stevens, T. J.; Arkin, I. T. Proteins 1999, 36, 135-143.

14. Pilpel, Y.; Ben-Tal, N.; Lancet, D. J Mol Biol 1999, 294, 921-935.

15. Adamian, L.; Liang, J. Proteins 2006, 63, 1-5.

16. Beuming, T.; Weinstein, H. Bioinformatics 2004, 20, 1822-1835.

17. Donnelly, D.; Overington, J. P.; Ruffle, S. V.; Nugent, J. H.; Blundell, T. L. Protein Sci 1993, 2, 55-70.

18. Stevens, T. J.; Arkin, I. T. Protein Sci 2001, 10, 2507-2517.

19. Park, Y.; Helms, V. Proteins 2006, In press.

20. Kohlbacher, O.; Lenhof, H. P. Bioinformatics 2000, 16, 815-824.

21. Word, J. M.; Lovell, S. C.; Richardson, J. S.; Richardson, D. C. J Mol Biol 1999, 285, 1735-1747.

22. Tusnady, G. E.; Dosztanyi, Z.; Simon, I. Nucleic Acids Res 2005, 33, D275-D278.

23. Thompson, J. D.; Higgins, D. G.; Gibson, T. J. Nucleic Acids Res 1994, 22, 4673-4680.

24. Pei, J.; Grishin, N. V. Bioinformatics 2001, 17, 700-712.

25. Henikoff, S.; Henikoff, J. G. J Mol Biol 1994, 243, 574-578.

26. Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. Nucleic Acids Res 1997, 25, 3389-3402.

27. Wimley, W. C.; Creamer, T. P.; White, S. H. Biochemistry 1996, 35, 5109-5124.

28. Adamian, L.; Nanda, V.; Degrado, W. F.; Liang, J. Proteins 2005, 59, 496-509.

29. Cornette, J. L.; Cease, K. B.; Margalit, H.; Spouge, J. L.; Berzofsky, J. A.; DeLisi, C. J Mol Biol 1987, 195, 659-685.

30. Rees, D. C.; Komiya, H.; Yeates, T. O.; Allen, J. P.; Feher, G. Annu Rev Biochem 1989, 58, 607-633.

31. Donnelly, D.; Overington, J. P.; Blundell, T. L. Protein Eng 1994, 7, 645-653.

32. Chih-Chung, C.; Chih-Jen, L. http://wwwcsientuedutw/~cjlin/libsvm.

33. R Development Core Team. http://wwwR-projectorg 2006.

34. Fleishman, S. J.; Harrington, S.; Friesner, R. A.; Honig, B.; Ben-Tal, N. Biophys J 2004, 87, 3448-3459.

35. Beuming, T.; Weinstein, H. Protein Eng Des Sel 2005, 18, 119-125.

36. Hessa, T.; Kim, H.; Bihlmaier, K.; Lundin, C.; Boekel, J.; Andersson, H.; Nilsson, I.; White, S. H.; von Heijne, G. Nature 2005, 433, 377-381.

Table 1. Proteins used in the analysis

PDB	Protein name	Chain	Oligomeric State
1. 1BRR	Bacteriorhodopsin	A	Monomer
2. 1F88	Rhodopsin	A	Monomer
3. 1K4C	KcsA potassium channel	C	Tetramer
4. 1MSL	MscL mechanosensitive channel	A	Pentamer
5. 1RHZ	Translocon	A, B	Monomer
6. 1J4N	Aquaporin	A	Monomer
7. 1FX8	Glycerol facilitator channel	A	Monomer
8. 1XQF	Ammonia channel	A	Monomer
9. 1KPL	H^+/Cl^- exchanger	A	Monomer
10. 2A65	Leucine transporter	A	Monomer
11. 1IWG	AcrB multi-drug efflux transporter	A	Monomer
12. 1PV7	Lactose permease	A	Monomer
13. 1PW4	Glycerol-3-phosphate transporter	A	Monomer
14. 1XFH	Glutamate transporter	A	Monomer
15. 1ZCD	Na^+/H^+ antiporter	A	Monomer
16. 1SU4	Calcium ATPase	A	Monomer
17. 2BL2	V-type Na^+-ATPase	A	Decamer
18. 1PRC	Photosynthetic reaction center	L, M, H	Monomer
19. 1L0V	Fumarate reductase (*E. coli*)	C, D	Monomer
20. 1QLA	Fumarate reductase (*W. Succinogenes*)	C	Monomer
21. 1KQF	Formate dehydrogenase N	B, C	Monomer
22. 1Q16	Nitrate reductase A	C	Monomer
23. 1NEK	Succinate dehydrogenase	C, D	Monomer
24. 1ZOY	Complex II	C, D	Monomer
25. 1OKC	Mitochondrial ADP/ATP carrier	A	Monomer
26. 1OCC	Cytochrome C oxidase (aa_3 type)	A, B, C, D, G, I, J, L, M	Monomer
27. 1EHK	Cytochrome C oxidase (ba_3 type)	A, B	Monomer
28. 1FFT	Ubiquinol oxidase	B	Monomer
29. 1BGY	Cytochrome bc_1 complex	C, D, E, G, J	Dimer
30. 1VF5	Cytochrome b_6f complex	A, B, C, D, H	Dimer
31. 2GFP	EmrD multi-drug transporter	A	Monomer

Table 2. Overall correlation between exposure patterns of boundary TM helices and positional scores calculated from the prediction scales

Identity	10	11	12	13	ww	eis	kd	kP	bw	ges	tmlip1	tmlip2
15	0.30	0.33	0.30	0.26	0.21	0.20	0.23	0.15	0.27	0.13	0.22	0.25
20	0.33	0.35	0.32	0.28	0.21	0.21	0.23	0.15	0.26	0.13	0.22	0.25
25	0.35	0.37	0.34	0.29	0.21	0.20	0.22	0.14	0.27	0.13	0.22	0.25
30	0.39	0.39	0.36	0.31	0.20	0.19	0.21	0.13	0.26	0.12	0.20	0.23
35	0.40	0.40	0.38	0.32	0.21	0.19	0.21	0.13	0.26	0.12	0.21	0.24
40	0.39	0.39	0.37	0.31	0.21	0.19	0.21	0.13	0.26	0.12	0.20	0.24

Table 3. Correlation between exposure patterns of boundary TM helices and the 11 scale based on 35% MSAs

ID	chain	CC	ID	chain	CC	ID	chain	CC	ID	chain	CC
1brr	A	0.47	1su4	A	0.62	1occ	D	0.36	1vf5	C	0.12
1f88	A	0.65	1prc	L	0.55	1occ	G	0.47	1vf5	D	0.06
1k4c	C	0.56	1prc	M	0.37	1occ	I	0.54	1vf5	H	0.02
1rhz	A	0.49	1prc	H	0.55	1occ	J	0.47	2gfp	A	0.14
1j4n	A	0.40	1kqf	B	0.30	1occ	L	0.52	1ehk	A	0.46
1fx8	A	0.55	1kqf	C	0.45	1occ	M	0.33	1fft	B	0.17
1xqf	A	0.44	1nek	C	0.15	1bgy	C	0.39	1l0v	C	0.13
1kpl	A	0.49	1nek	D	0.44	1bgy	D	0.53	1l0v	D	0.34
1iwg	A	0.37	1okc	A	0.46	1bgy	E	0.03	1zoy	C	0.47
1pv7	A	0.45	1q16	C	0.53	1bgy	G	0.01	1zoy	D	0.44
1pw4	A	0.38	1occ	A	0.29	1bgy	J	0.17			
1xfh	A	0.32	1occ	B	0.44	1vf5	A	0.53			
1zcd	A	0.65	1occ	C	0.56	1vf5	B	0.50			

Table 4. Overall correlation between the α-helical characters of the prediction scales and their degrees of correlation with exposure patterns of boundary TM helices

Identity	10	11	12	13	ww	eis	kd	KP	bw	Ges	tmlip1	tmlip2
15	0.60	0.60	0.59	0.55	0.43	0.48	0.48	0.37	0.46	0.38	0.40	0.42
20	0.59	0.57	0.55	0.55	0.40	0.45	0.43	0.37	0.45	0.37	0.36	0.40
25	0.58	0.57	0.56	0.56	0.43	0.44	0.40	0.30	0.47	0.33	0.37	0.40
30	0.56	0.56	0.57	0.56	0.41	0.44	0.41	0.31	0.43	0.33	0.36	0.38
35	0.53	0.53	0.52	0.49	0.39	0.48	0.45	0.35	0.42	0.36	0.37	0.38
40	0.53	0.53	0.53	0.50	0.37	0.48	0.46	0.38	0.40	0.39	0.39	0.40

Table 5. Overall correlation between exposure patterns of boundary TM helices and optimally combined prediction scales based on 35% MSAs.

Identity	10	11	12	13	ww	eis	kd	kP	bw	ges	tmlip1	tmlip2
10	0.40	0.40	0.40	0.40	0.43	0.42	0.41	0.40	0.44	0.40	0.42	0.42
11		0.40	0.40	0.41	0.42	0.41	0.41	0.40	0.44	0.40	0.41	0.42
12			0.38	0.38	0.41	0.40	0.39	0.38	0.42	0.38	0.39	0.40
13				0.32	0.37	0.34	0.33	0.32	0.39	0.32	0.34	0.36
ww					0.21	0.21	0.23	0.22	0.27	0.22	0.22	0.24
eis						0.19	0.21	0.19	0.26	0.23	0.21	0.24
kd							0.21	0.23	0.27	0.24	0.22	0.24
kP								0.13	0.27	0.14	0.21	0.24
bw									0.26	0.26	0.26	0.26
ges										0.12	0.23	0.26
tmlip1											0.21	0.25
tmlip2												0.24

Table 6. Accuracies of predicting buried residues of the test proteins by a linear SVM using the prediction scales from 35% MSAs.

Identity	10	11	12	13	ww	eis	kd	kP	bw	ges	tmlip1	tmlip2
10	68	68	69	69	70	69	68	68	70	68	69	69
11		67	68	68	69	68	68	68	70	67	68	68
12			66	66	69	68	67	66	70	66	67	68
13				66	68	66	64	66	67	64	66	65
ww					64	64	64	64	63	64	65	62
eis						62	60	62	64	63	62	62
kd							60	64	64	60	61	64
kP								52	64	51	60	64
bw									64	64	64	64
ges										52	58	64
tmlip1											61	64
tmlip2												64

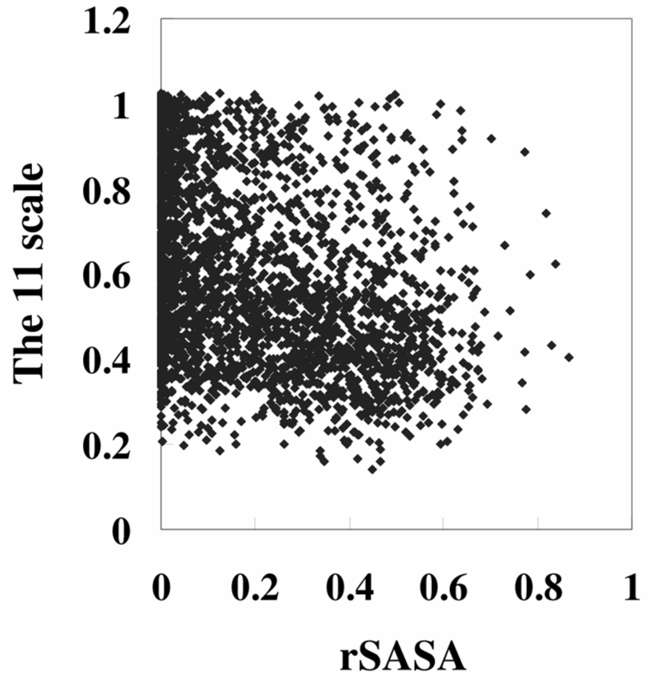


Figure 1. The correlation between the VAR scale and the rSASA values for the 3191 residues of the boundary TM helices of the test proteins located in the hydrophobic core of the membrane bilayer, based on 35% MSAs. A weak tendency is observed that the lower the rSASA value, the higher the VAR scale value. The CC is –0.40.

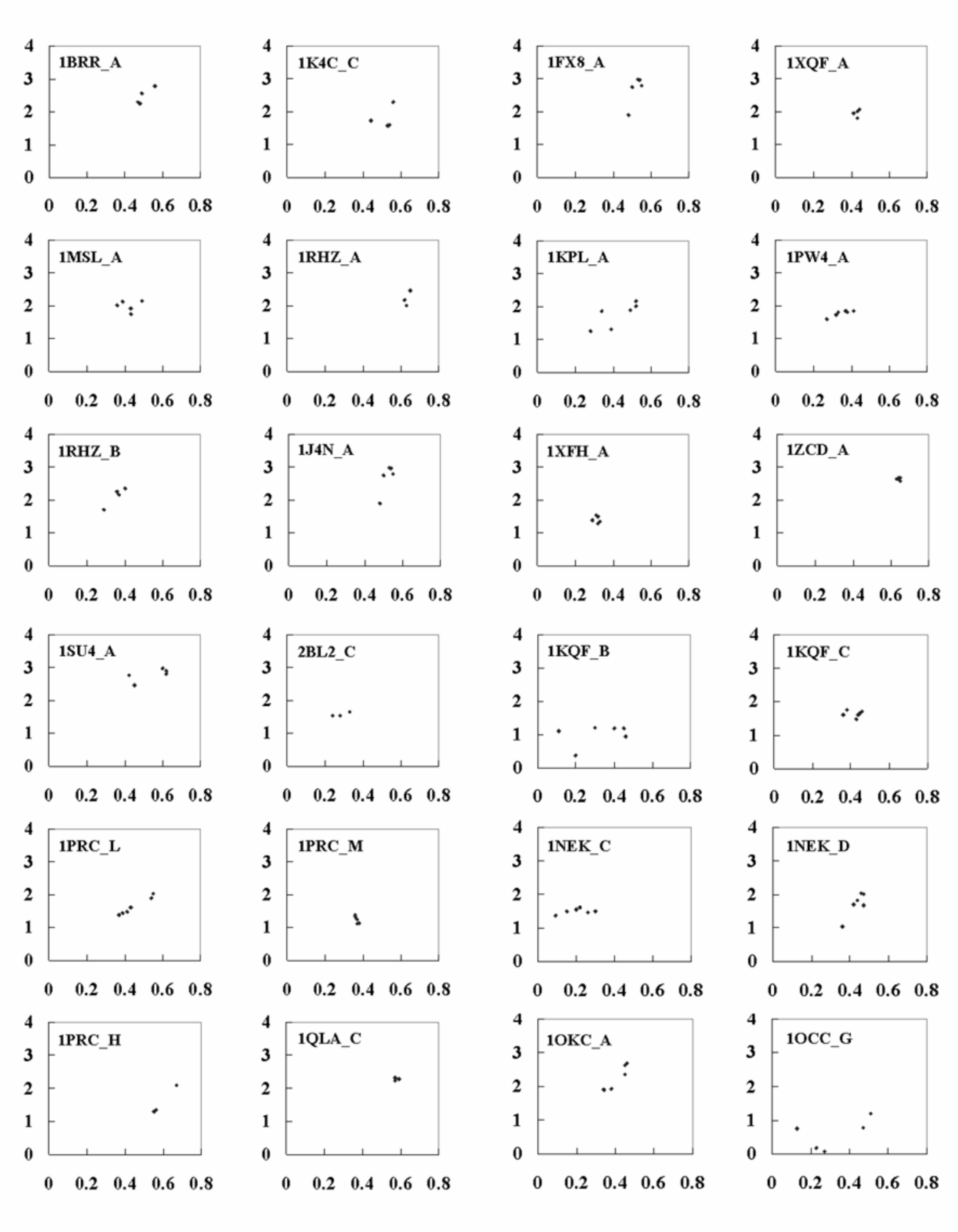


Figure 2. A plot displaying the α-helical character, Ψ, of the VAR scale (the vertical axis) and its correlation with exposure patterns of TM helices (the horizontal axis), based on 35% MSAs for a subset of the test proteins. Overall, the higher the α-helical character of the VAR scale, the more strongly correlated it is with exposure patterns of TM helices.

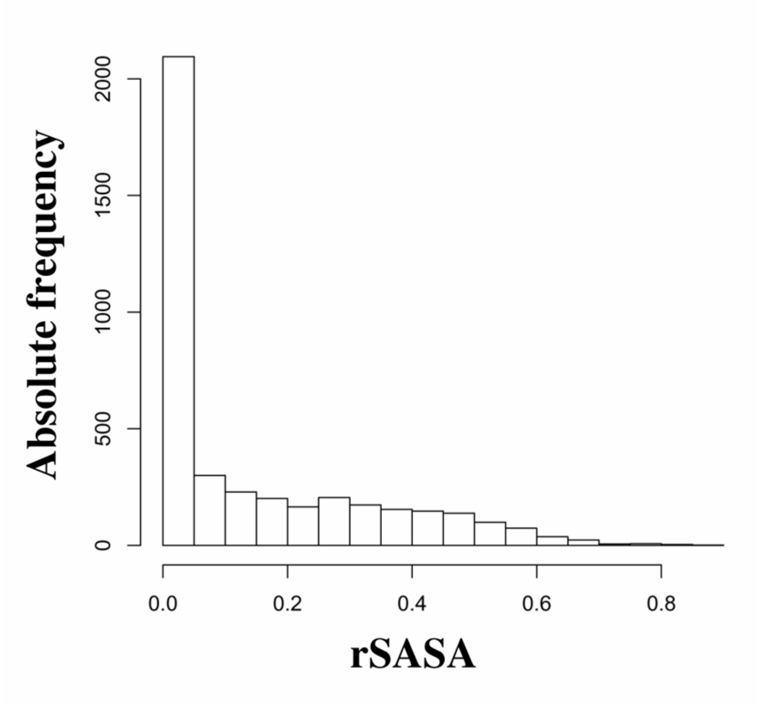


Figure 3. The distribution of rSASA values of the 4058 residues of the test proteins.

10. Paper III

Park, Y. and Helms, V. Bioinformatics (2007): 23, 701-708.

On the Derivation of Propensity Scales for Predicting Exposed Transmembrane Residues of Helical Membrane Proteins

Title: On the Derivation of Propensity Scales for Predicting Exposed Transmembrane Residues of Helical Membrane Proteins

Yungki Park and Volkhard Helms*

Center for Bioinformatics, Saarland University, Germany

*Corresponding author

Abstract

Helical membrane proteins (HMPs) play a crucial role in diverse physiological processes. Given the difficulty in determining their structures by experimental techniques, it is desired to develop computational methods for predicting the burial status of transmembrane residues. Deriving a propensity scale for the 20 amino acids to be exposed to the lipid bilayer from known structures is central to developing such methods. A fundamental problem in this regard is what would be the optimal way of deriving propensity scales. Here, we show that this problem can be reformulated such that an optimal scale is straightforwardly obtained in an analytical fashion. The derived scale favorably compares with others in terms of both algorithmic optimality and practical prediction accuracy. It also allows interesting insights into the structural organization of HMPs. Furthermore, the presented approach can be applied to other bioinformatics problems of HMPs, too.

All the data sets and programs used in the study and detailed primary results are available upon request.

Introduction

Helical membrane proteins (HMPs) play a crucial role in diverse physiological processes, including energy generation, signal transduction, the transport of solutes across the membrane, and the maintenance of ionic and proton gradients. Several studies have suggested that HMPs account for 20 – 30% of open reading frames of sequenced genomes (Liu*, et al.*, 2002; Wallin and von Heijne, 1998). In spite of their physiological importance and genomic abundance, less than 1% of the proteins with known structure are HMPs (Chen and Rost, 2002).

Given this circumstance, it is desirable to develop computational methods for predicting structural aspects of HMPs. At the heart of such efforts lies the development of a propensity scale for the 20 amino acids to be exposed to the lipid bilayer (Adamian*, et al.*, 2005; Beuming and Weinstein, 2004; Pilpel*, et al.*, 1999). Based on the recently increased number of experimentally determined 3D structures, Beuming and Weinstein derived a knowledge-based scale (the BW scale), which in combination with sequence conservation patterns enabled them to predict the burial status of TM residues with an accuracy of 80% (Beuming and Weinstein, 2004). Remarkably, Adamian and Liang (the TMLIP1/TMLIP2 scales) achieved a prediction accuracy of 88% in a similar study by taking advantage of the fact that most helix-helix interactions in the TM region can be recapitulated as occurring between two heptad repeat frames originally developed for coiled coils (Adamian and Liang, 2006).

The ways the BW and TMLIP1/TMLIP2 scales were derived represent three different learning algorithms for deriving a propensity scale from known structures. Our previous study revealed that the three algorithms are not equally effective (Park and Helms, 2006). A natural question that arises is which algorithm works better and why. Or more fundamentally, what would be the optimal way of deriving a propensity scale? This is an important problem not only from an algorithmic viewpoint but also from a practical viewpoint. An optimal algorithm would yield a propensity scale that faithfully captures the affinities of the 20 amino acids to preferentially interact with the lipid bilayer as reflected in experimental HMP structures. The knowledge of such a scale might allow interesting insights into the folding of HMPs. In this report, we show that one can reformulate this problem by selecting a sensible objective function such that an optimal scale (the MO scale) is straightforwardly obtained in an analytical fashion. A comparative analysis reveals that the MO scale favorably compares with others not only in terms of algorithmic optimality but also in terms of practical prediction accuracy. The MO scale also suggests interesting insights into the structural organization of HMPs compared to that of soluble proteins. Moreover, we show that the algorithm used for deriving the MO scale can be also applied to other bioinformatics problems of HMPs.

Methods

Generation of the data set

A non-redundant high-quality data set (less than 25% pairwise sequence identity and resolution better than 3.0 Å) was generated based on the lists of HMPs with known structure compiled by White (http://blanco.biomol.uci.edu) and by Michel (http://www.mpibp-frankfurt.mpg.de/michel/public/memprotstruct.html) as of September 2006. Some protein chains were omitted in spite of satisfying the above criteria either because we could not retrieve enough numbers of homologous sequences from sequence databases or the average pairwise identity of aligned sequences is greater than 80% (i.e., a diverse set of homologous sequences is not available). The final data set comprises 41 protein chains of 2901 TM residues (Table 1). To ensure the high homogeneity of the data set, residues located outside of the hydrophobic core of the lipid bilayer were excluded from the data set. The hydrophobic core of each protein chain, defined to be the region for which the probability of occurrence of the hydration waters of the lipid head groups is zero (White and Wimley, 1999), was derived from the location of the carbonyl groups of the lipid molecules along the membrane normal and the effective hydration profile obtained from the OPM database (Lomize*, et al.*, 2006a; Lomize*, et al.*, 2006b).

Computation of exposure patterns

The classification of a residue as being exposed or buried was based on its relative solvent-accessible surface area (rSASA) value. Several choices need to be made for the accurate computation of rSASA values. First, the probe radius should be properly chosen. Previous studies used the probe radii of 1.4 Å (the approximate radius of a free water molecule) or 1.9 Å (the approximate radius of a $–CH_2–$ group) (Adamian*, et al.*, 2005; Beuming and Weinstein, 2004). Given that the solvents surrounding the hydrophobic core parts of HMPs are hydrocarbon chains of phospholipids, we believe that 1.4 Å is not a proper choice. 1.9 Å is not suitable, either, because the CH_2 group of phospholipids is part of a long hydrocarbon chain and would not have a full mobility like a free $–CH_2–$ group. Thus, a value larger than 1.9 Å that well approximates the effective radius of the CH_2 group of hydrocarbon chains should be chosen. In this study, we empirically set the probe radius to 2.2 Å. Second, when necessary, the two faces of the TM region (the cytoplasmic and exoplasmic faces) were capped with dummy atoms before computing SASA values. Many HMPs contain large interval cavities, and, without capping, large SASA values were assigned to residues lining internal cavities, making these residues look as if they were facing the lipid bilayer. Upon capping, internal

cavities that are inaccessible to the probe were identified and excluded in computing SASA values. Actual computations were carried out by using the program suite VOLBL (Edelsbrunner, 1995; Edelsbrunner*, et al.*). SASA values were normalized by dividing them by reference values to yield rSASA values. The reference value for an amino acid, X, is its SASA in the context of a nonapeptide helix GGGG-X-GGGG. It is an open issue which reference state to use. GGGG-X-GGGG and G-X-G have been used in similar work (Adamian*, et al.*, 2005; Beuming and Weinstein, 2004). In our case, essentially the same results were obtained using G-X-G of a helical conformation as a reference state (see Supplementary Information).
Another point to be clarified in computing rSASA values is whether to use a monomeric or oligomeric form. Since there are few experimental data for the oligomerization of HMPs, it is not clear in most cases which form to choose. Presumably for this reason, different studies from different groups as well as different studies from the same group adopted HMPs of different oligomeric status in deriving propensity scales (Adamian and Liang, 2006; Adamian*, et al.*, 2005; Beuming and Weinstein, 2004). Our guiding principle was the degrees of conservation for the residues involved in the oligomerization. The very reason that buried residues tend to be more conserved than exposed ones (Baldwin*, et al.*, 1997; Donnelly*, et al.*, 1993; Stevens and Arkin, 2001; Yeates*, et al.*, 1987) is that they are central to maintaining structural and/or functional integrity. Thus, we reasoned that if oligomeric forms are absolutely necessary for whatever reasons, this obligatory nature should be reflected in the degrees of conservation for the residues involved in the oligomerization. The analysis revealed that the potassium channel (1R3J in Table 1) is the only one for which the use of oligomeric form is justified (see Supplementary Information). Cytochrome bc$_1$ complex is known to function as a dimer, and the 1PP9 structure in fact reveals a dimeric form (Huang*, et al.*, 2005). However, the dimerization is mediated by residues located outside of the hydrophobic core, which is why a monomeric form is also adopted for this protein.

Computation of profiles, positional scores and conservation indices

In general, the use of a profile (the frequencies of the 20 amino acids for a sequence position) improves the performance of sequence-based prediction methods. Thus, we derived the profiles of the protein chains in Table 1 as described before (Park and Helms, 2006). Briefly, for each protein chain, a maximum of 1000 homologous sequences were retrieved from the non-redundant database using BLAST (Altschul*, et al.*, 1997; Henikoff and Henikoff, 1994). Initial MSAs were then built by using ClustalW (Thompson*, et al.*, 1994). Then, sequence fragments were deleted from the MSA. Sequences that are less than 25% identical to the query sequence were also removed. The remaining sequences were realigned using ClustalW to yield a final MSA, which was used to obtain the profiles. When deriving profiles from an MSA, amino acid frequencies were weighted using a modified method of Henikoff and Henikoff as implemented in PSI-BLAST (Altschul*, et al.*, 1997; Henikoff and Henikoff, 1994). Actual computations were performed using the program AL2CO (Pei and Grishin, 2001), which is freely available at ftp://iole.swmed.edu/pub/al2co.
For a given propensity scale P, the positional score of sequence position i, S$_P$(i), is computed to be

$$S_P(i) = \sum_{j=1}^{20} f_i(j) \times P(j) \qquad (1)$$

where the index *j* runs over the 20 amino acids, $f_i(j)$ is the frequency of amino acid *j* in the sequence position *i*, and *P*(*j*) is the propensity value of amino acid *j*.

Conservation indices were estimated by using the variance-based method (Pei and Grishin, 2001). Our previous study showed that this method performs slightly better than other alternatives.

$$C(i) = \sqrt{\sum_j (f_i(j) - f(j))^2} \tag{2}$$

In Eq. 2, the index *j* runs over the 20 amino acids, *C*(*i*) is the conservation index for the sequence position *i*, *f*(*j*) is the overall frequency of amino acid *j* in the alignment. The conservation indices computed by Eq. 2 were normalized by subtracting the mean from each conservation index and dividing by the standard deviation. Actual calculations were performed again by using the AL2CO program.

Performance measures

The correlation coefficient (*cc*) for a set of *n* data points (x_i, y_i) was computed as follows:

$$cc = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{[n\sum x_i^2 - (\sum x_i)^2][n\sum y_i^2 - (\sum y_i)^2]}} \tag{3}$$

Prediction accuracy was calculated as

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (4),$$

where TP is the number of correctly predicted buried residues, TN is the number of correctly predicted exposed residues, FP is the number of falsely predicted buried residues and FN is the number of falsely predicted exposed residues.

Results and Discussion

Problem statement

Given a set of known HMP structures, the task is to derive an optimal propensity scale of the 20 amino acids to be exposed to the lipid bilayer. The derived scale would faithfully capture the affinities of the 20 amino acids to preferentially interact with the lipid bilayer. Also, it would allow one to predict exposed residues from the sequence with a highest possible accuracy under a linear regime as represented by Eq. 1.

Overview of the previous algorithms

Before introducing our novel learning algorithm, it is helpful to review the previous algorithms that were used to derive the BW and TMLIP1/TMLIP2 scales. It is often difficult to directly compare performance values of different learning algorithms reported in different studies because of the variety of data sets used and the discrepancy in state definitions. To facilitate a transparent performance comparison, we implemented the algorithms for the BW and TMLIP1/TMLIP2 scales and carried out comparisons on the common data set (Table 1).

The BW scale was derived in the following way (Beuming and Weinstein, 2004).

1. For each amino acid type, compute its SF value, which is a sum of surface fraction values of exposed TM residues (defined as those with an rSASA > 0.10 when computed by using a probe with radius of 1.4 Å and the reference value from a tripeptide G-X-G in extended conformation).
2. Identify the highest and lowest SF values (SF_{high} and SF_{low}).
3. Compute a propensity value for amino acid type *j* as $(SF_j\text{-}SF_{low})/(SF_{high} - SF_{low})$

As the 20 amino acids are not equally abundant in the TM region, it might be necessary to correct for compositional bias in deriving a propensity scale. This type of correction was not done in the derivation of the BW scale.

The TMLIP1 scale was derived as follows (Adamian*, et al.*, 2005).

1. Compute $N_{j,s}$ (the number of exposed TM residues of type *j*, with "exposed" being defined as rSASA > 0.0 when computed by using a probe with radius of 1.9 Å), N_s (the number of exposed TM residues of all types), N_j (the number of TM residues of type *j*), *N* (the number of all TM residues).
2. $P_{j,s} = N_{j,s}/N_s$, and $P_j = N_j/N$.
3. Compute a propensity value for amino acid type *j* as $\log(P_{j,s}/P_j)$.

The TMLIP2 scale was derived in a similar way (Adamian*, et al.*, 2005).

1. Compute $N_{j,s}$, N_s as for the TMLIP1 scale.
2. Compute N_j (the number of buried TM residues of type *j*), *N* (the number of buried TM residues of all types).
3. $P_{j,s} = N_{j,s}/N_s$, and $P_j = N_j/N$.
4. Compute a propensity value for amino acid type *j* as $\log(P_{j,s}/P_j)$.

The only difference between the TMLIP1 and TMLIP2 scales is how normalization is performed. In the TMLIP1 scale, normalization is based on compositional bias in the TM region. In contrast, the TMLIP2 scale is more like an odds-ratio type, explicitly taking into account the random probabilities of amino acid type *j* to be exposed and buried. In addition, the TMLIP1/TMLIP2 scales are based on count statistics, while the BW scale is not. Another point to be noted about the derivation of the BW and TMLIP1/TMLIP2 scales is that they require residues in the training set to be labeled as either being buried or exposed, which in turn requires a threshold rSASA value. Since there is no consensus on which rSASA value to use as a threshold, we explored all reasonable values (from 0.00 to 0.05 in steps of 0.01) in our implementation.

Derivation of an optimal propensity scale

Our starting point is fundamentally different from the approaches for the above three scales. We first ask what is meant by a propensity scale being "optimal". In other words, how would one compare different propensity scales? Our answer is how strongly correlated the positional scores derived from a scale for given profiles (Eq. 1) are with the corresponding exposure patterns (rSASA values in our context). In fact, our answer is not novel. This measure (the Pearson's correlation coefficient between the positional scores and rSASA values) has been, for a long time, known to be a key property measuring the fundamental goodness of a prediction method and extensively used in the realm of bioinformatics of soluble proteins (Adamczak, *et al.*, 2004; Ahmad, *et al.*, 2003; Chen and Zhou, 2005; Li and Pan, 2001; Nguyen and Rajapakse, 2006; Pollastri, *et al.*, 2002; Rost and Sander, 1994; Sim, *et al.*, 2005; Thompson and Goldstein, 1996).

Now that the meaning of a scale being "optimal" is clear, our task is to derive a propensity scale in such a way that the positional scores derived from it for given profiles are maximally correlated with the corresponding exposure patterns. Optimization techniques such as gradient-based optimization and Monte Carlo techniques may be used for this purpose. However, they usually do not guarantee the optimality of obtained solutions. For this reason, they have to be run several times with different starting points.

We find out, however, that an exact solution for this problem can be straightforwardly obtained in an analytical fashion. Eq. 5 provides the essential hint for our finding.

$$\mathrm{SSE}(\beta) = k(1 - r(\beta)^2) \qquad (5),$$

where SSE(β) is a sum of squared errors between the positional scores and rSASA values (formally defined in Eq. 6), k a constant ≥ 0, and $r(\beta)$ the Pearson's correlation coefficient between them.

$$\mathrm{SSE}(\beta) = (\mathrm{Y} - \mathrm{X}\beta)^{\mathrm{T}}(\mathrm{Y} - \mathrm{X}\beta) \qquad (6)$$

In Eq. 6, Y is a column vector of size N (the rSASA values of the training data set), X a matrix of N by 21 (a profile and 1) and β a column vector of size 21 (the propensity values of the 20 amino acids and an intercept value). Eq. 5 reveals that maximizing the Pearson's correlation coefficient with respect to β is equivalent to minimizing SSE with respect to β. Minimizing SSE with respect to β in a linear regime is the task of a linear regression analysis. The analytical solution is given as follows (Hastie, *et al.*, 2001):

$$\beta = (\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}}\mathrm{Y} \qquad (7)$$

Remarkably, this means that an optimal propensity scale (the first 20 elements of β in Eq. 7, the MO scale) has been obtained in an exact, analytical manner, without involving any explicit summing or counting steps as in the derivation of the BW and TMLIP1/TMLIP2 scales. Another advantage of this formulation is that, unlike those for the BW and TMLIP1/TMLIP2 scales, it does not require residues in the training set to be labeled beforehand as either being buried or exposed and thus is free from a priori assumptions.

We would like to extract a general picture on the structural characteristics of HMPs from the limited data set of Table 1. The MO scale obtained by Eq. 7 might represent an overfitting to the data set of Table 1. In order to extract a generalizable picture, we used the ridge regression analysis (equivalent to weight decay methods) with the complexity parameter empirically set to 0.00001 (Hastie, *et al.*, 2001). Regarding the choice of complexity parameters, it is to be noted that too small complexity parameters, e.g. 10^{-10}, are likely to generate an MO scale overfitting to the used data set while too large complexity parameters might induce an unreasonably high degree of compression, assigning propensity values close to 0 to all amino acids.

The justification for using the correlation measure as an objective function is now clear. It is naturally connected to the sum of squared errors loss function, which enables one to treat the whole problem in an analytical fashion. Accordingly, the MO scale is guaranteed to be optimal in the linear regime, unlike others. Most importantly, this guaranteed optimality allows one to perform novel analyses with it (see Section 3.5).

Comparative analysis

A jack-nife test was used for measuring the performances of the propensity scales. For each protein chain in Table 1, four different positional scores were derived from its profile and temporary BW, TMLIP1, TMLIP2, MO scales that were derived from the data set of Table 1 excluding the protein chain in question. Then, the performance of each scale was assessed in two complementary ways. First, by the Pearson's correlation coefficient between the computed positional scores and rSASA values, corresponding to algorithmic optimality since we define being "good" as being strongly correlated. Second, by the accuracy of predicting the burial status of TM residues (Eq. 4). This corresponds to what we mean by "practical prediction accuracy." Upon deriving positional scores, residues whose scores are higher than a cutoff value are classified as being exposed while those with a lower score as being buried. The cutoff value is objectively defined by a linear support vector machine on the basis of a training data set excluding the protein chain in question. We made use of the SVM implemented in R for this task with all parameters set to default values (Hsu and Lin, 2002; Karatzoglou*, et al.*, 2006; R Development Core Team, 2004).

The results of the comparative analysis are shown in Table 2. (It is to be noted that figures in Table 2 are only for the purpose of comparing the intrinsic goodness of the 4 propensity scales. For predicting the burial status in real applications, one would rely on more advanced methods along with other information available, e.g. conservation indices.) Table 2 reveals that the MO scale outperforms the others in terms of algorithmic optimality. In terms of practical prediction accuracy, the MO scale is better than the BW and TMLIP1 scales and compares favorably with the TMLIP2 scale. Thus, Table 2 "experimentally" validates the practical virtues of the logic behind the derivation of the MO scale.

A detailed analysis revealed two trends in the prediction results (see Supplementary Information). One is that the MO scale achieved more balanced predictions than others. The other is that, as the proportion of buried residues in the data set increased, the accuracy of predicting buried residues as being buried improved. This is possibly due to the weakening of bias introduced in the partitioning of the data set. The better performance of the TMLIP1/TMLIP2 scales over the BW scale suggests that it pays off to include normalization steps in deriving a propensity scale. The better performance of the TMLIP2 scale over the TMLIP1 scale indicates that not every null model is equally effective for normalization. On the other hand, the overall weak correlation between the positional scores computed from the MO scale and rSASA values of TM residues supports the suggestion that the lipophobic effect does not play a dominant role in folding of HMPs (Faham*, et al.*, 2004).

The MO scale

In addition to checking how the performance of the MO scale compares with that of others, it is of interest to find out how the MO scale itself compares with other scales because the MO scale accurately captures the affinities of the 20 amino acids to preferentially interact with the hydrophobic core of the lipid bilayer as reflected in experimental HMP structures.

Table 3 lists the MO scale, and Table 4 shows its correlations with other scales. As shown in Table 4, there is a strong correlation between the MO scale and other structure-based propensity scales. In contrast, the MO scale correlates poorly with hydrophobicity scales such as KD, EIS, GES, WW and Hessa. This observation supports the suggestion that the scale used by the translocon for recognizing TM segments is not the same as that for constrained partitioning of TM residues between being buried and exposed to the lipid bilayer (Pilpel*, et al.*, 1999).

Perhaps most striking is the observation that the MO scale for HMPs exhibits the strongest correlation with PSV. Since the propensity values captured by the MO scale should reflect a net result of numerous complex interactions involved in the folding of HMPs, they are not expected to display a strong correlation with a single scale. The correlation with PSV is even stronger than that with the structure-based ones. However, the correlation with a related scale, the bulkiness scale, does not stand out as strongly. As a control experiment, we derived an analogous MO scale for a representative set of soluble protein structures (see Supplementary Information). As expected, the MO scale for soluble proteins strongly correlates with hydrophobicity scales. Yet, it displays only a weak correlation with PSV.

In an effort to facilitate the interpretation of the MO scales for HMPs and soluble proteins in terms of more intuitive scales such as the hydrophobicity scales and the Cohn & Edsall partial specific volumes, they were decomposed into binary combinations of these scales using a linear regression analysis (For a meaningful analysis, each scale was standardized to have mean of 0 and standard deviation of 1). As shown in Table 5, the MO scale for HMPs can be interpreted as a hybrid of ~ 0.7 of PSV with ~ 0.3 of a hydrophobicity scale. In contrast, the MO scale for soluble proteins appears almost the same as hydrophobicity scales.

These analyses reveal the organizing principles of each type of protein structures. The strong correlation of the MO scale for soluble proteins with hydrophobicity scales and the decomposition analysis indicate that the hydrophobic effect is a major force behind the folding of soluble proteins, as has been long known (Pace*, et al.*, 1996). Regarding HMPs, two points are to be noted. First, the moderate correlation of the MO scale for HMPs with hydrophobicity scales and the decomposition analysis indicate that hydrophobicity still plays a role, albeit much weaker compared to the case of soluble proteins, in the thermodynamic stability of HMPs. Second, the strong correlation of the MO scale for HMPs with PSV and the decomposition analysis suggest, unexpectedly, that the structural organization of HMPs is better captured by PSV than by hydrophobicity. Our interpretation of this observation is as follows. In soluble proteins, functionally important residues are usually on the surface, and structural scaffolds are maintained by other residues buried inside. In this sense, functional and structural integrities are served by separate groups of residues. If this is not the case, a tradeoff between function and structural stability is to be made (Meiering*, et al.*, 1992; Schreiber*, et al.*, 1994; Shoichet*, et al.*, 1995; Zhang*, et al.*, 1992). In HMPs, functionally important residues are usually found buried inside. Also, HMPs are usually well-packed, presumably to compensate for the lack of the hydrophobic effect as a driving force for folding. Thus, functional and structural integrities are served by similar groups of residues, and one has to compromise between function and structural stability. One suitable way of doing so would be, whenever possible, to select amino acids with smaller partial specific volumes over those with larger partial specific volumes in the buried positions. As specific examples, we show the values from the MO and PSV scales for similarly charged/polar amino acids: S (MO: -0.19, PSV: 0.63) versus T (MO: -0.18, PSV: 0.70), R (MO: -0.21, PSV: 0.70) versus K (MO: -0.10, PSV: 0.82), D (MO: -0.27, PSV: 0.60) versus E (MO: -0.20, PSV: 0.66) and N (MO: -0.23, PSV: 0.62) versus Q (MO: -0.22, PSV: 0.67), suggesting that amino acids with a stronger tendency to be buried tend to display smaller partial specific volumes. Understandably, the intriguing analyses presented in this section are with due caveats owing to the small size of the current data set.

Applications

The approach for the derivation of the MO scale can be also applied to other bioinformatics problems of HMPs. We use the ProperTM method as an example (Beuming and Weinstein, 2004) for demonstration. ProperTM combines positional scores from the BW scale and sequence conservation patterns for improved predictions of the burial status of TM residues. Technically, it computes the overall score for sequence position *i*, *OS*(*i*), as 0.5×(*C*(*i*) – $S_{BW}(i)$), where *C*(*i*) is the conservation index for sequence position *i* and $S_{BW}(i)$ the positional score for sequence position *i* computed from the BW scale via Eq. 1. Since $S_{BW}(i)$ is a linear combination of the profile elements, the overall approach of ProperTM to derive an overall score for sequence position *i* can be cast as follows:

$$OS(i) = C_c C(i) + \sum_{j=1}^{20} C_j \times f_i(j) \qquad (8),$$

where C_c is the coefficient for the conservation index (set to 0.5 in ProperTM) and C_j the coefficient for the *j*th element of the profile (set to -0.5×*BW*(*j*) in ProperTM). The natural question is, then, whether the set of coefficients adopted by ProperTM are optimal. They can be optimized via Eq. 7. In this case, X becomes a matrix of *N* by 22 (*C*(*i*), a profile and 1), and *ß* is a vector of size 22 (the coefficients for *C*(*i*) and the 20 profile elements, and an intercept). Table 6 shows the results. “P_BW” denotes the combination of the conservation index and the positional score derived from the BW scale according to the way proposed in ProperTM. “P_TMLIP1”, “P_TMLIP2”, and “P_MO” are analogously defined. “SO” denotes the combination based on the simultaneous optimization of the coefficients for *C*(*i*) and the 20 profile elements. As expected, “SO” excels the others in terms of both algorithmic optimality and practical prediction accuracy.

Conclusion

The current study introduces a novel way of deriving a propensity scale for the 20 amino acids to be exposed to the lipid bilayer from known structures. The derived scale (the MO scale) favorably compares with others in terms of both algorithmic optimality and practical prediction accuracy. The MO scale also suggests interesting insights into the structural organization of HMPs. In addition, the same approach can be applied to the problem of optimally combining a propensity scale and sequence conservation patterns under a linear regime, as well.

Acknowledgement

This work was supported by Grant I/80469 of the Volkswagen Foundation. We thank crystallographers of HMPs because the current work was impossible without their work.

References

Adamczak, R., Porollo, A. and Meller, J. (2004) Accurate prediction of solvent accessibility using neural networks-based regression. *Proteins*, **56**, 753-767.

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

Adamian, L., Nanda, V., Degrado, W.F. and Liang, J. (2005) Empirical lipid propensities of amino acid residues in multispan alpha helical membrane proteins. *Proteins*, **59**, 496-509.

Ahmad, S., Gromiha, M.M. and Sarai, A. (2003) Real value prediction of solvent accessibility from amino acid sequence. *Proteins*, **50**, 629-635.

Altschul, S.F., Madden, T.L., Schaffer, 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. *Nucleic Acids Res.*, **25**, 3389-3402.

Baldwin, J.M., Schertler, G.F. and Unger, V.M. (1997) An alpha-carbon template for the transmembrane helices in the rhodopsin family of G-protein-coupled receptors. *J. Mol. Biol.*, **272**, 144-164.

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**, 1822-1835.

Chen, C.P. and Rost, B. (2002) State-of-the-art in membrane protein prediction. *Appl. Bioinformatics*, **1**, 21-35.

Chen, H. and Zhou, H.X. (2005) Prediction of solvent accessibility and sites of deleterious mutations from protein sequence. *Nucleic Acids Res.*, **33**, 3193-3199.

Cohn, E.J. and Edsall, J.T. (1943) *Proteins, amino acids and peptides*. Reinhold Publ. Corp., New York.

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

Edelsbrunner, H. (1995) The union of balls and its dual shape. *Discrete Comput. Geom.*, **13**, 415-440.

Edelsbrunner, H., Facello, M., Fu, P. and Liang, J. Measuring proteins and voids in proteins. In "Proc. 28th Ann. Hawaii Internat. Conf. System Sciences, 1995", vol. V: Biotechnology Computing, 256-264.

Eisenberg, D., Schwarz, E., Komaromy, M. and Wall, R. (1984) Analysis of membrane and surface protein sequences with the hydrophobic moment plot. *J. Mol. Biol.*, **179**, 125-142.

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

Faham, S., Yang, D., Bare, E., Yohannan, S., Whitelegge, J.P. and Bowie, J.U. (2004) Side-chain contributions to membrane protein structure and stability. *J. Mol. Biol.*, **335**, 297-305.

Hastie, T., Tibshirani, R. and Friedman, J. (2001) *The Elements of Statistical Learning*. Springer.

Henikoff, S. and Henikoff, J.G. (1994) Position-based sequence weights. *J. Mol. Biol.*, **243**, 574-578.

Hessa, T., Kim, H., Bihlmaier, K., Lundin, C., Boekel, J., Andersson, H., Nilsson, I., White, S.H. and von Heijne, G. (2005) Recognition of transmembrane helices by the endoplasmic reticulum translocon *Nature*, **433**, 377-381.

Hsu, C.W. and Lin, C.J. (2002) A comparison on methods for multi-class support vector machines. *IEEE Trans. Neural Networks*, **13**, 415-425.

Huang, L.S., Cobessi, D., Tung, E.Y. and Berry, E.A. (2005) Binding of the respiratory chain inhibitor antimycin to the mitochondrial bc1 complex: a new crystal structure reveals an altered intramolecular hydrogen-bonding pattern. *J. Mol. Biol.*, **351**, 573-597.

Karatzoglou, A., Meyer, D. and Hornik, K. (2006) Support Vector Machines in R. *Journal of Statistical Software*, **15**.

Kyte, J. and Doolittle, R.F. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, **157**, 105-132.

Li, X. and Pan, X.M. (2001) New method for accurate prediction of solvent accessibility from protein sequence. *Proteins*, **42**, 1-5.

Liu, Y., Engelman, D.M. and Gerstein, M. (2002) Genomic analysis of membrane protein families: abundance and conserved motifs. *Genome Biol.*, **3**, research0054.0051–research0054.0012.

Lomize, A.L., Pogozheva, I.D., Lomize, M.A. and Mosberg, H.I. (2006a) Positioning of proteins in membranes: a computational approach. *Protein Sci.*, **15**, 1318-1333.

Lomize, M.A., Lomize, A.L., Pogozheva, I.D. and Mosberg, H.I. (2006b) OPM: orientations of proteins in membranes database. *Bioinformatics*, **22**, 623-625.

Meiering, E.M., Serrano, L. and Fersht, A.R. (1992) Effect of active-site residues in barnase on activity and stability. *J. Mol. Biol.*, **225**, 585-589.

Nguyen, M.N. and Rajapakse, J.C. (2006) Two-stage support vector regression approach for predicting accessible surface areas of amino acids. *Proteins*, **63**, 542-550.

Pace, C.N., Shirley, B.A., McNutt, M. and Gajiwala, K. (1996) Forces contributing to the conformational stability of proteins. *FASEB J.*, **10**, 75-83.

Park, Y. and Helms, V. (2006) How strongly do sequence conservation patterns and empirical scales correlate with exposure patterns of transmembrane helices of membrane proteins? *Biopolymers*, **83**, 389-399.

Pei, J. and Grishin, N.V. (2001) AL2CO: calculation of positional conservation in a protein sequence alignment *Bioinformatics*, **17**, 700-712.

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. Mol. Biol.*, **294**, 921-935.

Pollastri, G., Baldi, P., Fariselli, P. and Casadio, R. (2002) Prediction of coordination number and relative solvent accessibility in proteins. *Proteins*, **47**, 142-153.

R Development Core Team (2004) R: A Language and Environment for Statistical Computing *R Foundation for Statistical Computing, Vienna, Austria.*

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

Schreiber, G., Buckle, A.M. and Fersht, A.R. (1994) Stability and function: two constraints in the evolution of barstar and other proteins. *Structure*, **2**, 945-951.

Shoichet, B.K., Baase, W.A., Kuroki, R. and Matthews, B.W. (1995) A relationship between protein stability and protein function. *Proc. Natl. Acad. Sci. USA*, **92**, 452-456.

Sim, J., Kim, S.-Y. and Lee, J. (2005) Prediction of protein solvent accessibility using fuzzy k-nearest neighbor method. *Bioinformatics*, **21**, 2844-2849.

Stevens, T.J. and Arkin, I.T. (2001) Substitution rates in alpha-helical transmembrane proteins. *Protein Sci.*, **10**, 2507-2517.

Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673-4680.

Thompson, M.J. and Goldstein, R.A. (1996) Predicting solvent accessibility: higher accuracy using Bayesian statistics and optimized residue substitution classes. *Proteins*, **25**, 38-47.

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

White, S.H. and Wimley, W.C. (1999) Membrane protein folding and stability: physical principles. *Annu. Rev. Biophys. Biomol. Struct.*, **28**, 319-365.

Wimley, W.C., Creamer, T.P. and White, S.H. (1996) Solvation energies of amino acid side chains and backbone in a family of host-guest pentapeptides. *Biochemistry*, **35**, 5109-5124.

Yeates, T.O., Komiya, H., Rees, D.C., Allen, J.P. and Feher, G. (1987) Structure of the reaction center from Rhodobacter sphaeroides R-26: membrane-protein interactions. *Proc. Natl. Acad. Sci. USA*, **84**, 6438-6442.

Zhang, J.H., Liu, Z.-P., Jones, T.A., Gierasch, L.M. and Sambrook, J.F. (1992) Mutating the charged residues in the binding pocket of cellular retinoic acid-binding protein simultaneously reduces its binding affinity to retinoic acid and increases its thermostability. *Proteins*, **13**, 87-99.

Zimmerman, J.M., Eliezer, N. and Simha, R. (1968) The characterization of amino acid sequences in proteins by statistical methods. *J. Theor. Biol.*, **21**, 170-201.

Table 1. Protein chains used in the analysis

PDB ID	Protein	Chains
1. 1M0L	Bacteriorhodopsin	A
2. 1GZM	Rhodopsin	A
3. 1R3J	KcsA potassium channel	C
4. 1J4N	Aquaporin	A
5. 1LDF	Glycerol facilitator channel	A
6. 1XQF	Ammonia channel	A
7. 1OTS	H^+/Cl^- exchanger	A
8. 2A65	Leucine transporter	A
9. 2CFQ	Lactose permease	A
10. 1YEW	Methane monooxygenase	B, C
11. 1SU4	Calcium ATPase	A
12. 2BL2	Rotor of V-type Na^+-ATPase	A
13. 1DXR	Photosynthetic reaction center	L, M, H
14. 1KF6	Fumarate reductase (*E. coli*)	C, D
15. 1QLA	Fumarate reductase (*W. succinogenes*)	C
16. 1KQF	Formate dehydrogenase N	B, C
17. 1Q16	Nitrate reductase A	C
18. 1NEK	Succinate dehydrogenase	C, D
19. 1ZOY	Complex II	C, D
20. 1OKC	Mitochondrial ADP/ATP carrier	A
21. 1V55	Cytochrome C oxidase (aa_3 type)	B, D, G, I, J, L, M
22. 1EHK	Cytochrome C oxidase (ba_3 type)	A, B
23. 1PP9	Cytochrome bc_1 complex	D, E, G, J
24. 2GIF	AcrB multidrug efflux transporter	A

Table 2. Performance comparison of the four scales

	0.00[a](0.58)[b]		0.01 (0.55)		0.02 (0.53)		0.03 (0.51)		0.04 (0.50)		0.05 (0.48)	
Scale	CC[c]	Acc[d]	CC	Acc	CC	Acc	CC	Acc	CC	Acc	CC	Acc
BW	0.27	67.1	0.27	66.6	0.27	66.0	0.27	65.7	0.27	65.4	0.27	64.5
TMLIP1	0.27	67.5	0.27	65.8	0.27	66.6	0.27	66.8	0.27	65.8	0.27	65.9
TMLIP2	0.28	69.6	0.28	68.9	0.29	67.9	0.29	67.3	0.29	66.8	0.28	66.4
MO	0.30	69.3	0.30	68.5	0.30	68.0	0.30	67.8	0.30	67.5	0.30	67.5

[a] The threshold rSASA value for specifying the residues in the data set as either being buried or exposed.

[b] The fraction of exposed residues in the data set.

[c] Absolute magnitude of correlation coefficient (Eq. 3)

[d] Accuracy of predicting the burial status in percentage (Eq. 4)

Table 3. The MO scale for HMPs

A	-0.09	G	-0.18	M	-0.23	S	-0.19
C	-0.16	H	-0.24	N	-0.23	T	-0.18
D	-0.27	I	0.05	P	-0.10	V	0.02
E	-0.20	K	-0.10	Q	-0.22	W	-0.03
F	-0.01	L	0.02	R	-0.21	Y	-0.15

Table 4. Correlation coefficients between the MO scale and others

Character of the scale	Scale (reference)	MO scale for HMPs	MO scale for soluble proteins
Structure-based	BW	0.82	ND[a]
	TMLIP1	0.75	ND[a]
	TMLIP2	0.84	ND[a]
Hydrophobicity	KD (Kyte and Doolittle, 1982)	0.73	-0.81
	EIS (Eisenberg, *et al.*, 1984)	0.67	-0.77
	GES (Engelman, *et al.*, 1986)	0.55	-0.80
	WW (Wimley, *et al.*, 1996)	0.65	-0.81
	Hessa (Hessa, *et al.*, 2005)	-0.66	0.87
Size	Bulkiness[b] (Zimmerman, *et al.*, 1968)	0.70	-0.42
Packing	PSV (Partial specific volume)[b] (Cohn and Edsall, 1943)	0.85	-0.34
Others	KPROT (Pilpel, *et al.*, 1999)	0.64	-0.65

[a] Not Defined: these three scales are not defined for soluble proteins

[b] See also Supplementary Information.

Table 5. 3 Best binary decomposition of the MO scales for HMPs and soluble proteins

The MO scale for HMPs			
	Scale 1 (coefficient for the decomposition)	Scale 2 (coefficient for the decomposition)	Correlation with the original MO scale
Decomposition 1	PSV (0.69)	EIS (0.34)	0.90
Decomposition 2	PSV (0.65)	KD (0.34)	0.89
Decomposition 3	PSV (0.70)	Hessa (-0.31)	0.89

The MO scale for soluble proteins			
	Scale 1 (coefficient for the decomposition)	Scale 2 (coefficient for the decomposition)	Correlation with the original MO scale
Decomposition 1	Hessa (0.67)	WW (-0.24)	0.88
Decomposition 2	Hessa (0.94)	PSV (0.13)	0.88
Decomposition 3	Hessa (0.84)	Bulkiness (-0.08)	0.88

Table 6. Performance comparison of the four scales combined with sequence conservation patterns.

	0.00[a](0.58)[b]		0.01 (0.55)		0.02 (0.53)		0.03 (0.51)		0.04 (0.50)		0.05 (0.48)	
	CC[c]	Acc[d]	CC	Acc	CC	Acc	CC	Acc	CC	Acc	CC	Acc
P_BW	0.43	71.3	0.43	70.9	0.43	70.7	0.43	71.2	0.43	70.8	0.43	70.5
P_TMLIP1	0.44	72.2	0.43	71.9	0.44	71.3	0.44	71.6	0.44	71.3	0.44	71.1
P_TMLIP2	0.44	73.0	0.44	72.5	0.44	72.2	0.44	72.3	0.44	72.4	0.44	72.3
P_MO	0.43	70.5	0.43	70.5	0.43	69.8	0.43	69.7	0.43	69.9	0.43	69.4
SO	0.46	74.1	0.46	72.9	0.46	72.6	0.46	72.9	0.46	72.9	0.46	72.9

[a,b,c,d] Defined in the same way as in Table 2.

Supplementary Information

1. Soluble proteins

A non-redundant set of 572 soluble protein chains (less than 25% pairwise sequence identity and resolution better than 2.5Å) were extracted from the PDB (Berman, H. M., et al. Nucleic Acids Res. 28, 235-242) by using the web server of PDB-REPRDB (Noguchi, T. & Akiyama, Y. Nucleic Acids Res. 31, 492-493). Then, for each of them, a multiple sequence alignment was generated as described in the Methods section. rSASA values for soluble protein chains were obtained by dividing the DSSP accessibility (Kabsch, W. & Sander, C. Biopolymers 22, 2577-2637) by the reference values reported by Samanta, U., et al. (Protein Eng. 15, 659-667)

The list of 572 soluble protein chains (PDB ID+Chain ID)

2B97A
1NWZA
1MUWA
1V6PA
1IQZA
1GVKB
4LZT_
1NKIA
1ZLBA
1GQVA
1UG6A
1P1XA
1EB6A
1LNIB
1A6M_
1K6UA
1EXRA
1QTWA
1WUIS
1PSRB
2C9VA
1W66A
2CARA
1N62B
1C5EA
2AB0A
1SU8A
2FBAA
1RG8A
1AXN_
2HVM_
1PT7A
1MH9A
1HPI_
2FJEA
2AQJA
1Y4TA
1ML4A
1J1QA
1GBG_
1H0HB
1LENC
1UBI_
1X1NA
1RYIA
1LD8A
1JUEA
1DOZA
1TML_
1GBS_
1XXOA
1LY2A
1ULKA
1UNKA
1EPTB
1OWLA
1AFWB
1G5TA
2GRNA
1TIF_
1LJ5A
1MML_
2PII_
1ODSA
1GYCA
1H7WD
1TG7A
1ON3E
1BGVA
1B8AA
1B8PA
1QGJA
1BUDA
1KEKA
2AKAA
1R8CA
1UYPA
1CHMA
1M0SA
2FZSB
1PNKB
1F20A
1SFTA
1IDK_
1TM2A
1PNKA
1U7PA
1K12A
2CHSA
1Q16A
1IVUA
1Q16B
2CK3C
1EU8A
1JK7A
2ESBA
1ESL_
1CGT_
1ECFB
2HPDA
1FP3A
1NSYA
1Z5GB
1APYA
2CQSA
1PMMA
1VPNB
1IGS_
1UGQA
1UOK_
2ACVA
2PIA_
1LKI_
1WPBA
1R5TA
1CB6A
1MW3A
1OGSA
2PGD_
1XCLA
2H29A
1BV1_
1N3FA
1COZA
1JS1X
1BPLB
2D5IA
1BPLA
1HCGB
1LNSA
1KKTA
1TARA
1VRNM
2FW2C
1WDTA
1RYDA
1M50A
1U19A
1JJIA
1DYNA
1AW8B
1GCB_
1QHWA
1VQOM
1FROA
1VQO3
1VQO1
1VQOU
1YXWA
1Z9UB
2BJ3A
1ZJHA
1FFYA
1QVCB
1MHLC
1KT6A
1I1XA
1SAUA
1TUKA
1Z53A
1P6OB
1OE3A
1J0OA
1HEUA
1HG7A
2D8DB
1M1NB
1M1NA
1C9OA
1CZPA
1WKQA
1T1EA
2DFCA
1IFC_
2FJ8A
1I6TA
1JBC_
2BMOA
2C1VA
1X6IB
1AMM_
1WN2A
2SN3_
1HXHB
2DEAA
1GNLA
1MEXL
1LXZA
1WDPA
1MQKH
1QKSA
1RTTA
2AEBB
1G61A
1OX0A
1K3YA
1WVFA
2A50B
2A50A
1SG4A
7FD1A
1BXAA
1RRO_
1ISPA
1F41A
1ULRA
1XEOA
1VF8A
2BOQA
1YPHC
1YPHE
1VH5A
1UTG_
1U1WB
2LISA
2A8FB
1HZTA
1QH5A
1V4PA
1C8CA
2AW1A
1JU2A
1SMOB
1GQIA
2RN2_
1HBZA
1HT6A
1Y0PA
1G6SA
1OFNB
2AD6A
1QWOA
1F1UA
1DJ0A
1V5DA
1KQ3A
1OC2B
1UV4A
1DQZA
2BKVB
1IQQA
2DQ6A
1I1NA
1UI0A
2B5HA
1NC5A
2GASA
1EG9B
1CG5B
1WS8B
1YT3A
1O26B
1YW5A
1MD6A
1Y0HB
1NM8A
1U0EA
2FUKA
1IT2A
1GY6A
1V5EA
1FWXA
1IWDA
1XWWA

1I9DA
2FMPA
2A7BA
1YO3A
1EZWA
1DHN_
1WUBA
1E3UB
1DOSA
1B4KA
1CZFA
2F2BA
1TH7A
1T0BH
1YKIA
1FZQA
1ZPSB
1ERT_
1TXLA
2EV6B
1QSTA
2FUJA
1GXYA
1X2TA
1B5FA
1LTUA
1K6WA
1PZ3A
1UMKA
1CQXA
1NPYB
1CHD_
1WD3A
1U8FO
1XGSA
1XX1A
7YASA
1UDH_
1UJ8A
2B0TA
1GNUA
1H4PA
1R8NA
1K66A
1XKRA
1ZAIA
1C8KA
1R4PB
1Y1TA
2A6SB
1CMBA
1V54H

1SHG_
1R9DA
1H8UA
1X8DA
2IW0A
1GND_
1DMR_
2IW2A
1V0ZA
1A44_
1QBA_
2GB0B
1IWBA
2GAGA
2GAGD
1EX2A
1VLS_
1PK6A
2B3YA
1NZOA
1J7DA
1DYR_
1YB0B
1ONRA
1XX2A
1WF3A
1VIYA
1QGDA
1J9JA
1T06A
1AGQD
1EJJA
1C7NA
1UJ1B
1QNXA
1I5ZB
1IAKB
1F5MB
1F4QB
1C8UA
1Z6OM
1K7HA
1AK2_
1KKJA
1EKJG
1KBLA
1KX5D
1QHDA
1FO6A
1KLXA
1PTQ_
1NFVI

1YNHB
1I39A
1K04A
1HCZ_
2C59A
1BF2_
1RQHA
2FCAB
1ENH_
1EZ0A
1UHVA
1YZYA
2CUNA
1FCUA
1UN1A
1PNOA
1G5CA
1A32_
1NU6A
1JL5A
1J2ZA
1SHXA
2CW6A
2FBIA
1BY4B
1XQZA
1ZDZA
1UMPA
1DTYA
1UC2A
1K3BB
1K3BA
1WU3I
2BS3A
2BS3C
1DJNA
1BOUB
1IG8A
1JI3A
1SNZA
1EI9A
1R44A
2F0RA
2BC4A
1V4EA
1S13A
1HPLA
3OVWA
1ITQA
3SSI_
1CAUA
1V6AA

1M73E
1YGYA
1WPGA
1I1IP
1XMEA
1MWKA
2GNNB
1F2KA
1MBB_
1REOA
1AYYD
1V7UA
1NHWA
1NHWC
1A6JA
2GP3A
1L5JA
1KYVC
1RKM_
1UB2A
2CMZB
2A0ZA
1V8BA
1TLBQ
1QGOA
1JDIA
1H3QA
1E5LA
1NBWA
1J36A
1KFQA
2A6MB
2B3ZD
1TJ7A
1NMTA
1F52A
1G99A
1FUIA
1HJRA
2A96B
3PBH_
2B5OA
1YEPA
1CLIB
1POIA
1POIB
1D0NA
2PVAA
1ND2B
1ND2C
1BM0A
1QGKA

1LCSA
1B4AA
1K5HA
2G5HA
1HR6A
2G5HB
1SZBA
1J8QA
1BSMA
1TZVA
1MQOA
1JO0A
1THM_
1H2WA
1LLFA
1GK8I
2BJKA
2FI1A
1EZGA
1RKQA
1WBIA
1QK8A
3VUB_
1YGE_
1YRCA
1G8AA
1Y43B
2DDRC
1V37B
1H2RL
1FP2A
1QH4A
1MXRA
1E6UA
1X7QB
2GDGA
1HKA_
1LFMA
1M2AA
1WHI_
1P1MA
1V5VA
1WRVB
1AH7_
1F46B
1AGI_
1S67L
2CVIA
1KPF_
1IO7A
2C3NA
1LLMC

1KRHA
1OMRA
1Y1NA
1YPYA
1YME_
1T6CA
1USGA
1RV9A
2AVKA
1SZNA
2B2HA
1YQZA
3GRS_
2BEMA
1QQJA
1EDQA
1WTJB
1ZG4A
1TXGA
1Q8FA
1BKPB
1DP0A
2A14A
1KNB_
1CSS_
1JHDA
2D80A
1ZXIC
1YOCB
1MTYD
1PMI_
1PB1A
1UQRD
2C65B
1PXZA
1NVMG
2CI3A
1F5VA
1V58A
2FA1B
1IS6A
1B2PA
1YGTA
1FTRA
2FZVB
1WAB_
1MXIA
1W4XA
1R12A
1FVAB

2. Bulkiness and partial specific volumes

The accession code for the Bulkiness scale from the AAindex database (Kawashima, S. and Kanehisa, M. Nucleic Acids Res. 28, 374) is ZIMJ680102. The accession code for the partial specific volumes of the 20 amino acids of Cohn and Edsall from the AAindex database is COHE430101. However, the values deposited in COHE430101 are not the same as those reported in the original publication. In the current study, we use the values reported in the original 1943 publication (*Protein, Amino Acid, and Peptides*. Reinhold, New York, pages 155 – 176 &370-381), which are listed below (Table S1).

Table S1. The partial specific volumes from Cohn and Edsall

A	0.74	G	0.64	M	0.75	S	0.63
C	0.61	H	0.67	N	0.62	T	0.70
D	0.60	I	0.90	P	0.76	V	0.86
E	0.66	K	0.82	Q	0.67	W	0.74
F	0.77	L	0.90	R	0.70	Y	0.71

3. Results with G-X-G as a reference state

Table S2. Performance comparison of the four scales (equivalent to Table 2 in the main text)

	0.00[a](0.58)[b]		0.01 (0.54)		0.02 (0.51)		0.03 (0.49)		0.04 (0.47)		0.05 (0.45)	
Scale	CC[c]	Acc[d]	CC	Acc	CC	Acc	CC	Acc	CC	Acc	CC	Acc
BW	0.30	68.1	0.30	67.1	0.30	66.5	0.30	65.5	0.30	66.2	0.30	66.4
TMLIP1	0.29	66.8	0.29	66.9	0.30	66.3	0.29	65.7	0.29	65.4	0.29	65.9
TMLIP2	0.31	69.2	0.31	68.3	0.31	67.7	0.31	66.5	0.31	66.5	0.31	66.5
MO	0.32	70.7	0.32	69.4	0.32	68.5	0.32	67.5	0.32	67.5	0.32	67.3

[a,b,c,d] Defined in the same way as in Table 2 in the main text

Table S3. The MO scale for HMPs (equivalent to Table 3 in the main text)

A	-0.07	G	-0.13	M	-0.16	S	-0.13
C	-0.12	H	-0.16	N	-0.15	T	-0.13
D	-0.18	I	0.04	P	-0.09	V	0.01
E	-0.14	K	-0.06	Q	-0.15	W	0.00
F	0.01	L	0.02	R	-0.14	Y	-0.10

Table S4. Correlation coefficients between the MO scale for HMPs and others (equivalent to Table 4 in the main text)

Character of the scale	Scale	Correlation coefficient
Structure-based	BW	0.81
	TMLIP1	0.74
	TMLIP2	0.84
Hydrophobicity	KD	0.70
	EIS	0.65
	GES	0.52
	WW	0.66
	Hessa	-0.65
Size	Bulkiness	0.74
Packing	PSV	0.85
Others	KPROT	0.60

Table S5. Best binary decomposition of the MO scale for HMPs (equivalent to Table 5 in the main text)

	The MO scale for HMPs		
	Scale 1 (coefficient for the decomposition)	Scale 2 (coefficient for the decomposition)	Correlation with the original MO scale
Decomposition 1	PSV (0.69)	WW (0.32)	0.89
Decomposition 2	PSV (0.70)	EIS (0.31)	0.89
Decomposition 3	PSV (0.70)	Hessa (-0.28)	0.88

Table S6. Performance comparison of the four scales combined with sequence conservation patterns (equivalent to Table 6 in the main text)

	0.00[a](0.58)[b]		0.01 (0.54)		0.02 (0.51)		0.03 (0.49)		0.04 (0.47)		0.05 (0.45)	
Scale	CC[c]	Acc[d]	CC	Acc	CC	Acc	CC	Acc	CC	Acc	CC	Acc
P_BW	0.44	71.6	0.44	71.1	0.44	71.2	0.44	70.6	0.44	70.0	0.44	69.8
P_TMLIP1	0.44	72.3	0.44	71.6	0.44	71.6	0.44	71.2	0.44	70.5	0.44	70.4
P_TMLIP2	0.45	73.0	0.45	72.5	0.45	72.2	0.45	72.1	0.45	71.7	0.45	72.0
P_MO	0.42	70.7	0.42	70.3	0.42	69.5	0.42	69.2	0.42	68.7	0.42	68.4
SO	0.48	74.2	0.48	73.5	0.48	73.6	0.48	72.5	0.48	72.6	0.48	72.8

[a,b,c,d] Defined in the same way as in Table 2 in the main text

4. Average conservation of interface residues of protein chains in Table 1 in the main text

Average conservation indices for buried, exposed and interface residues of the protein chains in Table 1 of the main text. Exposed residues are those with an rSASA > 0.00. Interface residues are those that are exposed in the monomeric state and buried in the oligomeric state. Protein chains lacking interface residues are excluded in the analysis.

Table S7. Average conservation indices for buried, exposed and interface residues

	1m0l_A	1r3j_C	1j4n_A	1ldf_A	1xqf_A	1ots_A
Buried	31 (1.41)[a]	6 (0.18)	54 (1.07)	50 (0.98)	88 (0.73)	111 (0.80)
Exposed	65 (-0.24)	36 (-0.21)	57 (-0.09)	54 (-0.17)	72 (-0.38)	73 (-0.41)
Interface	17 (-0.13)	8 (0.89)	24 (-0.05)	18 (-0.29)	29 (-0.22)	16 (0.45)

[a]Number of residues (their average conservation index)

	2a65_A	1yew_B	1yew_C	2bl2_A	1qla_C	1kqf_C
Buried	97 (0.31)	39 (0.16)	14 (0.22)	18 (0.82)	25 (0.56)	24 (0.63)
Exposed	74 (-0.40)	37 (-0.36)	35 (-0.66)	47 (0.16)	52 (-0.48)	54 (-0.45)
Interface	1 (-0.80)	20 (-0.06)	2 (0.02)	18 (-0.11)	5 (-0.54)	1 (1.06)

	1nek_D	1q16_C	1v55_B	1v55_G	2gif_A
Buried	13 (0.41)	20 (0.34)	2 (0.77)	1 (0.20)	75 (0.85)
Exposed	37 (-0.61)	69 (-0.27)	22 (-0.18)	15 (-0.30)	90 (-0.18)
Interface	13 (-0.50)	1 (0.00)	1 (0.15)	5 (-0.36)	25 (-0.21)

Only in the case of 1r3j_C, a significant degree of conservation is found for the interface residues. For 1kqf_C, there is only one residue at the interface, which is not deemed to be significant.

5. Analysis of the nature of prediction errors

Table S8. Analysis of the nature of the prediction errors reported in Table 2 of the main text

0.00[a] (0.58)[b]	BW			TMLIP1			TMLIP2			MO		
		Observed			Observed			Observed			Observed	
	Predicted	Buried	Exposed	Predicted	Buried	Exposed	Predicted	Buried	Exposed	Predicted	Buried	Exposed
	Buried	606 (50.2)	353 (20.9)	Buried	424 (35.1)	160 (9.5)	Buried	543 (45.0)	218 (12.9)	Buried	655 (54.2)	339 (20.0)
	Exposed	602 (49.8)	1340 (79.2)	Exposed	784 (64.9)	1533 (90.6)	Exposed	665 (55.1)	1475 (87.1)	Exposed	553 (45.8)	1354 (80.0)
0.01 (0.55)	BW			TMLIP1			TMLIP2			MO		
		Observed			Observed			Observed			Observed	
	Predicted	Buried	Exposed	Predicted	Buried	Exposed	Predicted	Buried	Exposed	Predicted	Buried	Exposed
	Buried	688 (52.9)	357 (22.3)	Buried	484 (37.2)	176 (11.0)	Buried	655 (50.4)	258 (16.1)	Buried	722 (55.5)	335 (20.9)
	Exposed	612 (47.1)	1244 (77.7)	Exposed	816 (62.8)	1425 (89.0)	Exposed	645 (49.6)	1343 (83.9)	Exposed	578 (44.5)	1266 (79.1)
0.02 (0.53)	BW			TMLIP1			TMLIP2			MO		
		Observed			Observed			Observed			Observed	
	Predicted	Buried	Exposed	Predicted	Buried	Exposed	Predicted	Buried	Exposed	Predicted	Buried	Exposed
	Buried	747 (54.7)	368 (24.0)	Buried	642 (47.0)	245 (16.0)	Buried	750 (55.0)	316 (20.6)	Buried	774 (56.7)	337 (21.9)
	Exposed	618 (45.3)	1168 (76.0)	Exposed	723 (53.0)	1291 (84.1)	Exposed	615 (45.1)	1220 (79.4)	Exposed	591 (43.3)	1199 (78.1)
0.03 (0.51)	BW			TMLIP1			TMLIP2			MO		
		Observed			Observed			Observed			Observed	
	Predicted	Buried	Exposed	Predicted	Buried	Exposed	Predicted	Buried	Exposed	Predicted	Buried	Exposed
	Buried	796 (56.0)	370 (25.0)	Buried	745 (52.4)	288 (19.5)	Buried	820 (57.7)	349 (23.6)	Buried	825 (58.1)	337 (22.8)
	Exposed	625 (44.0)	1110 (75.0)	Exposed	676 (47.6)	1192 (80.5)	Exposed	601 (42.3)	1131 (76.4)	Exposed	596 (41.9)	1143 (77.2)
0.04 (0.50)	BW			TMLIP1			TMLIP2			MO		
		Observed			Observed			Observed			Observed	
	Predicted	Buried	Exposed	Predicted	Buried	Exposed	Predicted	Buried	Exposed	Predicted	Buried	Exposed
	Buried	839 (57.5)	386 (26.8)	Buried	801 (54.9)	335 (23.2)	Buried	857 (58.8)	362 (25.1)	Buried	860 (59.0)	345 (23.9)
	Exposed	619 (42.5)	1057 (73.3)	Exposed	657 (45.1)	1108 (76.8)	Exposed	601 (41.2)	1081 (74.9)	Exposed	598 (41.0)	1098 (76.1)
0.05 (0.48)	BW			TMLIP1			TMLIP2			MO		
		Observed			Observed			Observed			Observed	
	Predicted	Buried	Exposed	Predicted	Buried	Exposed	Predicted	Buried	Exposed	Predicted	Buried	Exposed
	Buried	884 (59.1)	417 (29.7)	Buried	838 (56.0)	331 (23.6)	Buried	886 (59.2)	365 (26.0)	Buried	904 (60.4)	351 (25.0)
	Exposed	613 (41.0)	987 (70.3)	Exposed	659 (44.0)	1073 (76.4)	Exposed	611 (40.8)	1039 (74.0)	Exposed	593 (39.6)	1053 (75.0)

[a,b]Defined in the same way as in Table 2 in the main text

Table S9. Analysis of the nature of the prediction errors reported in Table 6 of the main text

0.00[a] (0.58)[b]	SO			BW			TMLIP1			TMLIP2			MO		
	Observed			Observed			Observed			Observed			Observed		
	Predicted	Buried	Exposed	Predicted	Buried	Exposed	Predicted	Buried	Exposed	Predicted	Buried	Exposed	Predicted	Buried	Exposed
	Buried	730 (60.4)	273 (16.1)	Buried	634 (52.5)	259 (15.3)	Buried	650 (53.8)	248 (14.7)	Buried	675 (55.9)	249 (14.7)	Buried	615 (50.9)	262 (15.5)
	Exposed	478 (39.6)	1420 (83.9)	Exposed	574 (47.5)	1434 (84.7)	Exposed	558 (46.2)	1445 (85.4)	Exposed	533 (44.1)	1444 (85.3)	Exposed	593 (49.1)	1431 (84.5)
0.01 (0.55)	SO			BW			TMLIP1			TMLIP2			MO		
	Observed			Observed			Observed			Observed			Observed		
	Predicted	Buried	Exposed	Predicted	Buried	Exposed	Predicted	Buried	Exposed	Predicted	Buried	Exposed	Predicted	Buried	Exposed
	Buried	808 (62.2)	293 (18.3)	Buried	733 (56.4)	278 (17.4)	Buried	747 (57.5)	262 (16.4)	Buried	773 (59.5)	270 (16.9)	Buried	717 (55.2)	272 (17.0)
	Exposed	492 (37.9)	1308 (81.7)	Exposed	567 (43.6)	1323 (82.6)	Exposed	553 (42.5)	1339 (83.6)	Exposed	527 (40.5)	1331 (83.1)	Exposed	583 (44.9)	1329 (83.0)
0.02 (0.53)	SO			BW			TMLIP1			TMLIP2			MO		
	Observed			Observed			Observed			Observed			Observed		
	Predicted	Buried	Exposed	Predicted	Buried	Exposed	Predicted	Buried	Exposed	Predicted	Buried	Exposed	Predicted	Buried	Exposed
	Buried	876 (64.2)	307 (20.0)	Buried	797 (58.4)	282 (18.4)	Buried	808 (59.2)	275 (17.9)	Buried	841 (61.6)	283 (18.4)	Buried	781 (57.2)	293 (19.1)
	Exposed	489 (35.8)	1229 (80.0)	Exposed	568 (41.6)	1254 (81.6)	Exposed	557 (40.8)	1261 (82.1)	Exposed	524 (38.4)	1253 (81.6)	Exposed	584 (42.8)	1243 (80.9)
0.03 (0.51)	SO			BW			TMLIP1			TMLIP2			MO		
	Observed			Observed			Observed			Observed			Observed		
	Predicted	Buried	Exposed	Predicted	Buried	Exposed	Predicted	Buried	Exposed	Predicted	Buried	Exposed	Predicted	Buried	Exposed
	Buried	935 (65.8)	300 (20.3)	Buried	876 (61.7)	291 (19.7)	Buried	881 (62.0)	283 (19.1)	Buried	903 (63.6)	287 (19.4)	Buried	849 (59.8)	306 (20.7)
	Exposed	486 (34.2)	1180 (79.7)	Exposed	545 (38.4)	1189 (80.3)	Exposed	540 (38.0)	1197 (80.9)	Exposed	518 (36.5)	1193 (80.6)	Exposed	572 (40.3)	1174 (79.3)
0.04 (0.50)	SO			BW			TMLIP1			TMLIP2			MO		
	Observed			Observed			Observed			Observed			Observed		
	Predicted	Buried	Exposed	Predicted	Buried	Exposed	Predicted	Buried	Exposed	Predicted	Buried	Exposed	Predicted	Buried	Exposed
	Buried	972 (66.7)	299 (20.7)	Buried	913 (62.6)	302 (20.9)	Buried	917 (62.9)	293 (20.3)	Buried	938 (64.3)	282 (19.5)	Buried	889 (61.0)	305 (21.1)
	Exposed	486 (33.3)	1144 (79.3)	Exposed	545 (37.4)	1141 (79.1)	Exposed	541 (37.1)	1150 (79.7)	Exposed	520 (35.7)	1161 (80.5)	Exposed	569 (39.0)	1138 (78.9)
0.05 (0.48)	SO			BW			TMLIP1			TMLIP2			MO		
	Observed			Observed			Observed			Observed			Observed		
	Predicted	Buried	Exposed	Predicted	Buried	Exposed	Predicted	Buried	Exposed	Predicted	Buried	Exposed	Predicted	Buried	Exposed
	Buried	1015 (67.8)	305 (21.7)	Buried	947 (63.3)	306 (21.8)	Buried	960 (64.1)	302 (21.5)	Buried	980 (65.5)	286 (20.4)	Buried	927 (61.9)	318 (22.7)
	Exposed	482 (32.2)	1099 (78.3)	Exposed	550 (36.7)	1098 (78.2)	Exposed	537 (35.9)	1102 (78.5)	Exposed	517 (34.5)	1118 (79.6)	Exposed	570 (38.1)	1086 (77.4)

[a,b]Defined in the same way as in Table 2 in the main text

11. Paper IV

Park, Y., Hayat, S. and Helms, V. BMC Bioinformatics (2007): 8, 302.

Prediction of the Burial Status of Transmembrane Residues of Helical Membrane Proteins

Title: Prediction of the Burial Status of Transmembrane Residues of Helical Membrane Proteins

Yungki Park, Sikander Hayat and Volkhard Helms*

Center for Bioinformatics, Saarland University, Germany

*Corresponding author

Abstract

Background: Helical membrane proteins (HMPs) play a crucial role in diverse cellular processes, yet it still remains extremely difficult to determine their structures by experimental techniques. Given this situation, it is highly desirable to develop sequence-based computational methods for predicting structural characteristics of HMPs.

Results: We have developed TMX (TM eXposure), a novel method for predicting the burial status (i.e. buried in the protein core vs. exposed to the membrane) of transmembrane (TM) residues of HMPs. TMX derives positional scores of TM residues based on their profiles and conservation indices. Then, a support vector classifier is used for predicting their burial status. Its prediction accuracy is 78.71% on a benchmark data set, representing considerable improvements over 68.67% and 71.06% of previously proposed methods. Importantly, unlike the previous methods, TMX automatically yields confidence scores for the predictions made. In addition, a feature selection incorporated in TMX reveals interesting insights into the structural organization of HMPs.

Conclusions: A novel computational method, TMX, has been developed for predicting the burial status of TM residues of HMPs. Its prediction accuracy is much higher than that of previously proposed methods. It will be useful in elucidating structural characteristics of HMPs as an inexpensive, auxiliary tool. TMX will be made freely available to academic users through a web server at http://service.bioinformatik.uni-saarland.de/tmx.

Background

Helical membrane proteins (HMPs) play a crucial role in diverse cellular processes, including energy generation, signal transduction, the transport of solutes across the membrane, and the maintenance of ionic and proton concentrations. Several studies have suggested that HMPs account for 20 – 30% of the open reading frames of sequenced genomes [1, 2]. In spite of their biological importance and genomic abundance, less than 1% of the proteins with known structure are HMPs [3], and this situation is not expected to improve dramatically in the near future. Hence, it is desirable to develop sequence-based computational methods for predicting structural characteristics of HMPs. In the realm of soluble proteins, two particular structural characteristics have been the main target of computational prediction methods: secondary structure [4-10] and solvent accessibility [11-26] (often in a form of binary burial status; buried inside vs. exposed to the environment). For HMPs, the prediction of secondary structures does not carry as significant a momentum as for soluble proteins because transmembrane (TM) segments, which can be relatively reliably identified from the sequence by several techniques [27-37], are known to usually adopt helical conformations to satisfy the hydrogen bonding capacity of the backbone polar atoms. On the other hand, the problem of predicting the burial status (i.e. buried in the protein core vs. exposed to the membrane) of TM residues of HMPs has remained nearly untouched until now, in contrast to the situation for soluble proteins, which have been extensively studied (see the references listed above) following the pioneering work of Rost and Sander [12]. This is quite “remarkable” given that it is much more difficult to determine the structures of HMPs than those of soluble proteins by experimental techniques. The ability to predict the burial status of TM residues of HMPs from the sequence should be useful in several tasks. One simple example would be to help design mutational experiments aimed at identifying catalytically important TM residues of transporters [38, 39] by providing a list of TM residues highly likely to be buried in the protein core because TM residues important for the transport function would not be expected to be exposed to the membrane. Another simple example would be to help design mutational experiments aimed at identifying TM residues important for protein-protein interactions in the membrane by providing a list of TM residues highly likely to be exposed to the membrane.

In 2004, Beuming and Weinstein pioneered the first sequence-based computational method for predicting the burial status of TM residues of HMPs (denoted hereafter as the BW method), which was based on sequence conservation patterns and a newly derived knowledge-based propensity scale of the 20 amino acids to be exposed to the membrane [40]. For a rather small benchmark set, the BW method achieved an impressive prediction accuracy of 80%. Recently, Adamian and Liang reported the development of a similar method [41], but it predicts the face of a TM helix exposed to the membrane, not the burial status of individual TM residues. Hildebrand and his coworkers described a computational method for predicting whether a given residue is located at a helix-helix interface in the membrane [42]. Yet, this is a distinct prediction problem from the one the current study deals with: a residue located outside of a helix-helix interface can still be buried. Quite recently, Yuan and his coworkers developed a method for predicting the relative solvent-accessible surface area (rSASA) of TM residues based on support vector regression (SVR, denoted hereafter as the YU method) [43]. Even though the YU method does not explicitly predict the burial status of TM residues, it is possible to do so using the predicted rSASA values. To our best knowledge, the BW and YU methods are the only ones currently available for predicting the burial status of TM residues of HMPs.

We have developed TMX (TM eXposure), a novel sequence-based computational method for predicting the burial status of TM residues of HMPs. Its accuracy is 78.71% over a much larger data set of 3138 TM residues, representing a considerable improvement over 68.67% of the BW method when evaluated on the same data set. This prediction accuracy is also higher than 71.06% of the YU method. Importantly, unlike the BW and YU methods, TMX automatically yields confidence scores for the predictions made, a highly desirable component for any computational prediction method, which allows the user to selectively utilize prediction results depending on confidence scores in real application situations. In addition, a feature selection incorporated in TMX reveals interesting insights into the structural organization of HMPs.

Results and Discussion

Analysis of the BW method

TMX is novel in several aspects compared to the BW and YU methods and can be described without any reference to these previous methods. However, we prefer to describe the logic behind its development in reference to the BW method in order to contrast it with the BW method and highlight its novelties.

For predicting the burial status of a TM residue, the BW method computes its positional score and compares the score with a threshold [40]. If the score is higher than the threshold, it is predicted to be buried. Otherwise, it is predicted to be exposed to the membrane. Formally, the BW method computes a positional score for sequence position i, $S(i)$, as $0.5\times(C(i) - P_{BW}(i))$, where $C(i)$ is the conservation index for sequence position i, and $P_{BW}(i)$ the propensity of sequence position i for being exposed to the membrane, which is in turn derived from the BW scale as shown in Eq. 1. The BW scale is derived from a set of HMPs with known structure.

$$P_{BW}(i) = \sum_{j=1}^{20} BW(j) \times f_i(j) \tag{1}$$

In Eq. 1, the index j runs over the 20 naturally occurring amino acids, $BW(j)$ is the propensity value of amino acid j in the BW scale, and $f_i(j)$ the frequency of amino acid j in sequence position i. Plugging Eq. 1 into $0.5\times(C(i) - P_{BW}(i))$, the overall approach of the BW method for deriving a positional score can be cast as follows.

$$S(i) = 0.5 \times C(i) - \sum_{j=1}^{20} 0.5 \times BW(j) \times f_i(j) \tag{2}$$

Eq. 2 indicates that $S(i)$ is a linear combination of the conservation index and the 20 elements of the profile. Thus, it can be written more generally as follows.

$$S(i) = C_c \times C(i) + \sum_{j=1}^{20} C_j \times f_i(j) \qquad (3),$$

where C_c is the coefficient for the conservation index (set to 0.5 in the BW method) and C_j the coefficient for the jth element of the profile (set to $-0.5\times BW(j)$ in the BW method). With achieving highest possible prediction accuracies in mind, we raise the question of whether setting the coefficients in Eq. 3 empirically as in the BW method is optimal or not. Our answer is no. Optimizing the coefficients would be a better idea. Confirming this expectation, the coefficients optimized by linear regression led to a prediction accuracy of 71.13%, compared to 68.67% of the BW method as shown in Table 1. Specifically, ridge linear regression with the complexity parameter set to 0.001 was used throughout this study in an effort to minimize generalization errors [44]. It is noteworthy that we use the same formula as the BW method – Eq. 3 – but with an entirely different philosophy. In the BW method, one first derives a propensity scale of the 20 amino acids to be exposed to the membrane from known HMP structures and then uses it for computing the propensity of a target residue to be exposed (Eq. 1). This propensity of the target residue is combined with its degree of conservation to yield its positional score. Our analysis reveals that this overall idea of the BW method can be concisely summarized by Eq. 3, which immediately suggests that there is a better way of doing the job.

There is an issue to be clarified before we move on. We implemented the BW method, and its performance was evaluated on the same data set as for TMX. This was necessary since it is often difficult to directly compare performance values of different prediction methods reported in different studies because of the variety of data sets used and the discrepancy in state definitions. A serious difficulty arose in implementing the BW method, namely setting thresholds manually. As mentioned above, upon computing the positional score of a target residue, the BW method compares it with a threshold that has been manually set. If the positional score is greater than the threshold, it is predicted to be buried. Otherwise, it is predicted to be exposed. In a leave-one-out (jack-knife) testing scheme, thresholds need to be manually set separately for each of 43 protein chains in the benchmark data set (see Methods). Admittedly, it is impossible for us to exactly reproduce this step in the way it was performed in the original publication for the BW method [40]. In addition, we feel that it might be subjective to set thresholds manually. Then, is there any mathematical formalism that allows thresholds to be set in such a manner that (1) we exactly mimic the manual setting of thresholds as was done in the BW method and (2) yet, thresholds are set objectively and reproducibly? Our answer is a linear support vector classifier (lSVC, i.e. an SVC with a linear kernel). Since the hyperplane – $f(x) = \beta_0 + \beta^T x = 0$, where β^T is the transpose of a column vector β – obtained by an lSVC in a one-dimensional space represents a scalar value of $-\beta/\beta_0$ [44], setting a threshold via an lSVC is an exact computational analogue to setting it manually, yet in an objective, reproducible way. It is to be noted that the introduction of an lSVC to the prediction scheme transforms it to a two-step scheme because an lSVC also needs training and, as a result, the jack-knife scheme should be applied to both steps. We want to stress that the sole purpose of using an lSVC here is to mimic the manual assignment of thresholds as exactly as possible yet in an objective, reproducible fashion. Thus, we intentionally did not seek SVCs with a non-linear kernel or other sophisticated classifiers at this stage (but see below).

Improved use of conservation indices

Another point well worth considering in Eq. 3 is how conservation indices are incorporated. The average identities of sequences retrieved from sequence databases for different query sequences can be appreciably varying. Thus, without normalization, one may assign overall high conservation indices to one protein chain while assigning overall low conservation indices to another. Normalization of conservation indices effectively solves this bias problem, just as in microarray data processing. In the BW method, conservation indices are not normalized. We found that normalizing conservation indices leads to a significant improvement in the prediction accuracy, raising it from 71.13% to 73.84%.

A second, minor aspect to be considered is how conservation indices are actually computed in the first place. The BW method computes conservation indices as follows.

$$C(i) = 0.5 \times V(i) + 0.5 \times IC(i) \qquad (4)$$

In Eq. 4, $V(i)$ is the volume of the polytope for sequence position i derived from a multiple sequence alignment (MSA), estimating the probability for the presence of a set of different amino acids from a set of pairwise distribution probabilities, and $IC(i)$ is the information content of sequence position i [40]. Eq. 4 relies on many assumptions that are yet to be validated. The first are ad hoc measures taken to enforce the Euclidean space to the distances between aligned sequences for computing $V(i)$ [45]. The second is the assumption used in computing $IC(i)$ that the 20 naturally occurring amino acids are equally likely to occur in the TM region. The third is that even though it seems reasonable to assign equal weights to both terms in Eq. 4, it is not clear whether that choice is optimal.

As in our previous studies [46, 47], we derived conservation indices using Eq. 5, which is mathematically well-defined and relatively free from potentially problematic assumptions.

$$C(i) = \sqrt{\sum_j (f_i(j) - f(j))^2} \tag{5}$$

In Eq. 5, the index *j* runs over the 20 naturally occurring amino acids, *C*(*i*) is the conservation index for sequence position *i*, $f_i(j)$ is the frequency of amino acid *j* in sequence position *i*, and *f*(*j*) is the overall frequency of amino acid *j* in the alignment. As expected, the use of Eq. 5 instead of Eq. 4 improved the prediction accuracy from 73.84% to 74.51%. It is to be noted that conservation indices obtained by Eqs. 4 and 5 were from the same MSAs.

Extending the window size for the input vector

At this stage, the input vector for the prediction method consists of 21 elements (20 profile elements and a conservation index for the target residue). Another measure that we can take to further improve the prediction accuracy is to additionally consider the neighboring residues of the target residue (i.e. increasing a window size for the input vector from 1 to any larger number). In fact, nearly all techniques developed for water-soluble proteins exploit this possibility [12-26]. We explored all symmetric window sizes (Table 2). There are a couple of points to be noted in Table 2. When increasing the window size from 1 to 3 or 5, the prediction accuracy is decreased, suggesting that the signal-to-noise ratio deteriorated (see also below). The first peak in the prediction accuracy is observed at a window size of 9. It is interesting to note that, assuming the canonical helix conformation, when the length of a helix gets to 9, the first and last residues (residues at positions *i*-4 and *i*+4) face in the same direction as the central residue (residue at position *i*, corresponding to the target residue in our context). Thus, our results suggest that the identities of the residues lying just above and below the target residue on the same helix face are most indicative of the burial status of the target residue, as expected from the canonical helix conformation. As it is actually 3.6 residues per turn in the canonical helix conformation, a certain improvement is already found by including the positions *i* ± 3. Consistent with this line of reasoning, the best prediction accuracy, 75.97%, is observed at a window size of 15. Based on a similar observation, Adamian and Liang recently developed a highly effective method for predicting membrane-exposed faces of TM helices [41].

Feature selection

The logic behind increasing window sizes for better predictions is that one can better account for long-range effects with enlarged windows. However, the shortcoming of enlarged windows is that the signal-to-noise ratio deteriorates as the window size is increased, as demonstrated in Table 2. For example, compare the prediction accuracies for window sizes of 15 and 21. The tradeoff between long-range effects and signal-to-noise ratios would suggest a window size of 15 instead of 21. Is there any way of circumventing this unpleasant tradeoff? Feature selection might be an answer. A simple illustration will make this point clearer. An input vector for a window size of 21 consists of 441 elements (21 elements for each of the 21 residues). It is intuitively clear that not all 441 elements will contribute equally to the prediction. Many of them might simply be noise. Thus, it might be possible to use enlarged windows for a better consideration of long-range effects and still maintain a high signal-to-noise ratio by filtering out noisy elements.

Of many techniques available for feature selection, we chose the Fisher's index for the following reasons. First, the Fisher's index is conceptually attractive, having a clear meaning easy to understand [44]. Put simply, the Fisher's index represents the ability of a given element to maximize the distance between the centroids of the two given classes and simultaneously minimize the overlap between them. Second, unlike techniques involving linear combinations of feature vectors, the Fisher's index is highly interpretable. This is a big advantage given the high dimensions of our feature spaces. Most importantly, one can gain interesting biological insights into the structural organization of HMPs from the Fisher's index (see below). Third, the Fisher's index can be computed cheaply. Fourth, the Fisher's index is well suited to continuous features (as opposed to discrete ones).

The 441 elements of a window of size 21 were ranked according to their Fisher's indices, and increasing fractions of them (in steps of 0.05) were input to the prediction (see first and second columns of Table 3). The best prediction accuracy, 77.21%, was obtained when using the top 20% elements only. This accuracy is higher than 75.97% obtained by an "unintelligently" increased window of size 15 in the above section. Which elements rank top? As shown in Table 4, the top-ranking elements are mostly conservation indices, in line with previous findings that conservation properties of TM residues correlate strongly with their degree of exposure to the membrane [46, 48-50]. Also, Table 4 shows that the frequencies of occurrence of L, I, V and F at the target residue are highly indicative of its burial status. In this regard, it is interesting to note that our previous study showed that these amino acids possess the highest propensities to preferentially interact with the membrane [47]. The frequency of occurrence of G at the target residue is also strongly correlating with its burial status, ranking at the 9th place, which is consistent with earlier findings that glycine residues play a pivotal role in mediating helix-helix interactions in the membrane [51-54]. Table 4 also shows that the frequencies of occurrence of I, G and L at the 4th residue N terminal to the target one also strongly correlate with the burial status of the target residue, which makes sense considering the canonical helix conformation as mentioned above.

Given dramatic improvements in prediction accuracy and interesting insights into the system under investigation through a feature selection as demonstrated here, it was quite surprising to find that almost all studies on predicting the solvent accessibility of water-soluble proteins [12-26] have not considered it. Hence, it would be worthwhile to investigate whether feature selection can similarly pay off in predicting the solvent-accessibility of water-soluble proteins.

Non-linear regression

All approaches for computing positional scores thus far can be understood as an extension of Eq. 3. Namely, they are all linear methods. Additional improvements might be achieved by relying on non-linear methods. The power of non-linearity is illustrated by conservation indices. Conservation indices are non-linear combinations of profile elements (Eq. 5), which was motivated by the prior knowledge that conserved TM residues tend to be buried while variable ones tend to be exposed to the membrane [46, 48-50]. In fact, Table 4 showed that conservation indices were the features most strongly correlated with the burial status of TM residues. Also, it is shown below that conservation indices play a much greater role than profile elements in boosting prediction accuracies. In theory, a perfect non-linear method should be able to find such non-linear combinations of profile elements when fed only profile elements. However, this is usually not the case. Whenever prior knowledge on the system under investigation permits sensible non-linear combinations of raw features (e.g., conservation indices out of profile elements), it is always good to do so explicitly.

If there still remain untapped non-linear combinations of profile elements or profile elements and conservation indices that correlate with the burial status of TM residues, the use of non-linear methods might be profitable. Of the vast array of available non-linear regression techniques, we made use of SVR with a radial kernel because a nice interface with R is already available (see Methods) and it has performed respectably in studies of water-soluble proteins [23, 25]. Our preliminary analysis showed that SVR with a radial kernel tends to rival SVR with other kernels. Once a kernel type is chosen, another important parameter to be fine-tuned is the regularization constant C, i.e. how much weight one should put on minimizing the costs of violating a decision boundary relative to maximizing the closest distance of a data point to the boundary [44, 55]. The general expectation from the theory is that as the regularization constant gets higher, a heavier weight is put on minimizing the violation costs and, as a result, a more wiggly decision boundary is obtained with a possibly larger generalization error. The default C value is 1, and we tried 4 different C values, 10, 1, 0.1 and 0.01. As above, the 441 elements of a window of size 21 were ranked according to their Fisher's indices, and increasing fractions of them (in steps of 0.05) were input to the prediction via SVR with a radial kernel. Table 3 shows the results (columns 3 – 6). It is immediately clear that, in almost all cases, linear regression outperforms SVR, indicating that the generalization errors of SVR are larger than those of linear regression, presumably due to its over-flexibility in fitting a separating boundary to a given data set. Thus, SVR does not seem advantageous over linear regression on this data set. Admittedly, we can not rule out the possibility that highly fine-tuned SVR can outperform linear regression. Given limited computational resources and considerable amounts of computation required for a leave-one-out validation of a two-step prediction method (~ 40 CPU hours on a 2.4 GHz processor), it is beyond our capability to exhaustively scan all possible combinations of SVR parameters. However, it is our experience that SVR with all parameters set to default values generally performs nearly optimally. Thus, we are quite certain that, at least for the current purpose of predicting the burial status of TM residues of HMPs, linear regression is at least as effective as SVR. Supporting this conclusion, previous studies on water-soluble proteins demonstrated that sophisticated linear methods can rival non-linear ones in performance [14, 21, 26].

Optimizing classifiers

Upon computing a positional score for the target residue, a classifier is invoked to classify it as either buried in the protein core or exposed to the membrane. Although any machine-learning technique can be used as a classifier, we have only considered lSVCs so far. The original reason for choosing lSVCs was, as mentioned earlier, to implement the BW method as exactly as possible, yet in an objective, reproducible manner, so that the BW method can be justly compared with ours. However, we may choose other classifiers for our prediction method. Although there are tons of available classifiers, we primarily focused on SVCs for practical reasons as mentioned above. Preliminary analysis showed that SVCs with a linear or radial kernel tend to outperform others. Thus, SVCs with a linear or radial kernel were pursued further in combination with 5 different regularization constants, 1, 0.5, 0.1, 0.05 and 0.01, chosen on the basis of the results shown in Table 3. In addition to searching for a better classifier, it might also be helpful in boosting prediction accuracies to refine input vectors themselves. So far, the input vectors for a classifier have been one-dimensional, i.e. consisting of a positional score for a given target residue. The input vectors for a classifier can be straightforwardly refined exactly in the same way as the input vectors for computing positional scores were refined in Table 3.

Table 5 shows the best prediction accuracies for each combination of an SVC kernel and a regularization constant. An SVC with a linear kernel outperforms that with a radial one, and a regularization constant of 0.5 is optimal among those investigated. The best prediction accuracy, 78.71%, was obtained by an SVC with a linear kernel that considers the top 16 positional scores out of the 21 positional scores (i.e. the positional scores of the target residue and its 10 neighbors on the N terminus and its 10 neighbors on the C terminus) derived from considering the top 10% of the 441 elements of a window of size 21 (Table 3). An SVC with a radial kernel also achieved this prediction accuracy at a regularization constant of 0.5. Due to its sustained performance over the examined range of regularization constants, however, an SVC with a linear kernel is preferred. The method that gives rise to the best performance becomes the method of choice and is named "TMX (TM eXposure)."

The performance of TMX – is it "significantly" higher than that of the BW method? As mentioned earlier, the prediction accuracy of the BW method is 68.67% when tested on the same data set. The *p* value estimating the statistical significance of the 10.04% increase in the prediction accuracy achieved by TMX relative to the BW method is $< 10^{-5}$ according to the Wilcoxon signed rank test. Accordingly, TMX is judged to be a truly better method for predicting the burial status of TM residues of HMPs. A final point worthy of noting is the architecture of TMX. TMX is a two-step prediction method, where binary classifications are made in the second step on the basis of positional scores computed in the first step. The two-step architecture – is it really worthwhile? Obviously, one can directly apply SVCs to the profiles and conservation indices of the target residue and its neighbors for predicting its burial status, without computing positional scores in the first place. Several studies on water-soluble proteins noted that a two-step prediction scheme can better account for correlated patterns of properties to be predicted, leading to higher prediction accuracies [7-10, 21, 23, 24]. To test whether this is also the case for us, we investigated the performance of SVCs that were directly fed profiles and conservation indices for the prediction. Specifically, as shown in Table 3, the 441 elements of a window of size 21 were sorted according to the Fisher's index, and increasing fractions of them were fed to SVCs. The best prediction accuracy for an SVC with a linear kernel was 77.53%, and that for an SVC with a radial kernel 77.21%. Therefore, a two-step prediction scheme appears to pay off in our case, too.

Comparison with the YU method

The YU method computes the positional score of a target residue via SVR using position-specific scoring matrices (PSSMs) obtained by PSI-BLAST [56]. In studies of water-soluble proteins, it has been very popular to use PSSMs as input vectors in order to boost the accuracy of predicting solvent accessibility [8-10, 17, 19-24]. The popularity of PSSMs has partially stemmed from the fact that one does not have to explicitly generate an MSA for obtaining PSSMs. As with the BW method, we implemented the YU method for a transparent performance comparison using the R interface [57, 58] of the LIBSVM library [59]. In implementing the YU method, we set all the parameters of SVR as optimized by Yuan *et al.* and did not intentionally seek any further optimizations.

The best prediction accuracy of the YU method on the benchmark data set is 71.06% (fourth column of Table 6), much lower than 78.71% achieved by TMX (p value of < 0.001 from the Wilcoxon signed rank test). It is of interest to find out where the performance difference between TMX and the YU method comes from, except for the novelties introduced to TMX such as feature selection and a sophisticated classification in the second step. To this end, we replaced PSSMs by profiles or conservation indices to find out how different input vectors affect prediction accuracies. Table 6 shows that profiles alone perform similarly to (or only slightly better than) PSSMs. Compared with the performance of profiles or PSSMs, the performance of normalized conservation indices is really standing out. Moreover, a comparison of the performance of profiles plus normalized conservation indices shown in Table 2 with that of normalized conservation indices alone (C set to 1, a default value, in Table 6) also indicates that conservation indices play a crucial role in boosting prediction accuracies. Thus, it may be concluded that the poor performance of the YU method is partly due to the fact that its input vectors – PSSMs – do not contain the information captured by conservation indices. In this regard, it is interesting to note that the most effective method for predicting the solvent accessibility of water-soluble proteins uses PSSMs as its sole input [24]. Thus, it would be worthwhile to check out whether replacing PSSMs by profiles plus normalized conservation indices would be similarly successful for water-soluble proteins.

Analysis of the TMX predictions

In addition to prediction accuracies, there are other interesting aspects worthy of analyzing. For example, are there any amino acids for which it is easier to predict the burial status? Is it easier to correctly predict buried residues as being buried than exposed residues as being exposed?

Table 7 shows the results for each amino acid. The highest prediction accuracies were achieved for R, H, D and K, all of which are charged or strongly polar. Their average conservation indices are among the highest (data not shown). Thus, it appears that these amino acids are well conserved for functional (and/or structural) reasons and that their high conservation indices make it easier for TMX to correctly predict their burial status. In this regard, the case of proline is a contrasting example. Its average conservation index is among the highest, yet the prediction accuracy for it is among the lowest. The data set contains 43 buried and 46 exposed proline residues, and the average conservation indices for buried and exposed proline residues are 1.22 and 1.07, respectively. Thus, proline residues exposed to the membrane appear as strongly conserved for their structural (and/or functional) role as those buried inside. The lack of correlation between conservation and the burial status for proline seems to make it difficult for TMX to correctly predict its burial status. Surprisingly, E is the amino acid with the lowest prediction accuracy. Inspection of the individual incorrect predictions for E suggests a couple of plausible explanations for this unexpected result. First, conserved E residues are sometimes exposed to the membrane (2GIF_A_346: 1.54 [residue 346 of chain A in the PDB file 2GIF: its conservation index is 1.54], 1OTS_A_414: 2.46, 1YEW_B_201: 1.43 and 2BL2_A_139: 3.47), and TMX predicted them to be buried. Second, there are several buried E residues that are not conserved (2A65_A_112: -0.46, 2A65_A_419: -0.94, 1SU4_A_908: -0.30 and 1QLA_C_180: 0.14), and presumably their low conservation indices hinder the accurate prediction of their burial status. The prediction accuracies for abundant amino acids (L, A, V, I, G and F) are all higher than the overall accuracy of 78.71%.

Table 8 shows the specificity and sensitivity of TMX. The lower sensitivity (70.61%) compared to the specificity (84.77%) seems to reflect the biased composition of the benchmark data set comprising 55.86% of exposed residues and 44.14% of buried residues.

Confidence scores for the predictions made

It is highly desirable to have confidence scores available for the predictions made. Confidence scores allow the user to selectively utilize prediction results in real application settings. In TMX, the absolute magnitude of a decision value generated by the SVC is taken to be a confidence score for the prediction [44, 55]. As shown in Fig. 1, predictions with a high confidence score tend to be more accurate than those with a low one. The prediction accuracy rises to 90.21% when considering the 1440 predictions with a confidence score ≥ 1.2. 1440 out of 3138 means a coverage of 45.89%. Thus, a fairly high coverage is maintained for prediction accuracies of ~ 90%, which makes TMX well suited to real application settings.

Conclusions

We have presented TMX, a novel sequence-based computational method for predicting the burial status of TM residues of HMPs. It significantly outperforms previously proposed methods. In addition, feature selection incorporated in TMX revealed interesting insights into the structural organization of HMPs. Importantly, unlike the previous methods, TMX automatically generates confidence scores for the predictions made, and it was shown that predictions with a high confidence score tend to be more accurate than those with a low one. Thus, in a real application setting, the user of TMX can selectively utilize prediction results on the basis of their confidence scores. The developmental course of TMX clearly highlighted the importance of conservation indices and feature selection in boosting prediction accuracies. In this regard, it was rather surprising to find that the most effective method for predicting the solvent accessibility of water-soluble proteins considers neither conservation indices nor feature selection. It would be interesting to investigate whether these two “new” findings can be favorably transferred to one of the classical bioinformatics problems of predicting the solvent accessibility of water-soluble proteins.

Methods

Generation of the benchmark data set

As is always the case in machine-learning studies, constructing a well-curated data set was the starting point of the current study. Special care was taken in selecting protein chains, delineating their TM boundaries and computing the rSASA values of TM residues.

The details of generating a non-redundant high-quality data set have been described elsewhere [47]. Briefly, based on the lists of HMPs with known structure compiled by White (http://blanco.biomol.uci.edu) and by Michel (http://www.mpibp-frankfurt.mpg.de/michel/public/memprotstruct.html) as of February 2007, protein chains with less than 25% pairwise identity and a resolution better than 3.0 Å were gathered, resulting in 43 protein chains of 3138 TM residues (Table 9). To keep the data set homogeneous, residues located outside of the hydrophobic core of the membrane were excluded from the data set. The classification of a residue as being exposed vs. buried was based on its rSASA value. To approximate the effective radius of the CH_2 group of hydrocarbon chains of phospholipids, the probe radius was set to 2.2 Å. When necessary, the two faces of the TM region (the cytoplasmic and exoplasmic faces) were capped with dummy atoms before computing SASA values. Many HMPs contain large internal cavities, and, without capping, large SASA values were assigned to residues lining internal cavities, making these residues look as if they were facing the membrane. Upon capping, internal cavities that are inaccessible to the probe were identified and excluded in computing SASA values. Actual computations were carried out using the program suite VOLBL [60, 61]. SASA values were normalized by dividing them by reference values to yield rSASA values. The reference value for an amino acid, X, is its SASA in the context of a nonapeptide helix GGGG-X-GGGG computed with a probe radius of 2.2 Å as above.

Exposed residues were defined as those with an rSASA greater than 0.00, as in a previous study [62]. This threshold rSASA value is justified for HMPs given the large probe radius chosen in this study. As discussed before [25], this threshold is also free from artifacts arising from normalization and subsequent binary classification. Nevertheless, it was argued that the threshold for a binary classification should be set such that the data set is equally partitioned into the two classes to avoid statistical artifacts. An rSASA of 0.00 induces a slightly skewed partitioning of 44.14% of buried residues and 55.86% of exposed ones. Equipartitioning of the data set was achieved with an rSASA of 0.04. Additional analysis showed that the conclusions drawn in this study remain fully valid for this new threshold (data not shown).

Computation of profiles and conservation indices

In general, the use of a profile (the frequencies of the 20 amino acids for a sequence position) improves the performance of sequence-based prediction methods. For extracting profiles, one needs to generate MSAs. As with any sequence-based prediction methods, the careful choice of sequences in MSAs is very important for the performance of the prediction method. MSAs generated using different criteria would yield results of differing quality. Thus, it would be desirable to generate "optimal" MSAs for different query sequences. Unfortunately, it is currently impossible to do so in an objective, consistent manner without any prior knowledge about the three-dimensional structures of the query sequences. Thus, a reasonable approach that is also objective, consistent and easily reproducible by others, was taken for generating

MSAs, even though it might produce suboptimal MSAs for some query sequences. Its detail has been described elsewhere [46, 47]. Briefly, for a given query sequence, a maximum of 1000 homologous sequences were retrieved from the non-redundant database using BLAST [56]. Initial MSAs were built using ClustalW [63]. Then, sequence fragments were deleted from the MSA. Sequences that are less than 25% identical to the query sequence were also removed. The remaining sequences were realigned using ClustalW to yield a final MSA, which was used to obtain profiles. When deriving profiles from an MSA, amino acid frequencies were weighted using a modified method of Henikoff and Henikoff as implemented in PSI-BLAST [56, 64]. Actual computations were performed using the program AL2CO [65], which is freely available at ftp://iole.swmed.edu/pub/al2co. Conservation indices (Eq. 5) were also derived using AL2CO.

Support vector machines

The support vector classifier (SVC)/ support vector regression (SVR) [44, 55] implementation in R [57-59] was used for the current work. The parameters for SVC/SVR were set to default values unless otherwise noted. The YU method was implemented using the SVR implementation in R as described in [43].

Performance evaluation

A leave-one-out ('jack-knife') test was carried out to measure the performance of different prediction methods examined in this study. For two-step prediction methods, the jack-knife scheme was applied to both steps as it should be. Prediction accuracies mean the fractions of the benchmark data set for which the burial status was correctly predicted.

Acknowledgements

This work was supported by Grant I/80469 of the Volkswagen Foundation. We thank Thijs Beuming and Harel Weinstein for sharing their program for computing conservation indices. We thank crystallographers of HMPs because the current work was impossible without their work.

References

1. Wallin E, von Heijne G: **Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms.** *Protein Sci* 1998, **7**:1029-1038.
2. Liu Y, Engelman DM, Gerstein M: **Genomic analysis of membrane protein families: abundance and conserved motifs.** *Genome Biol* 2002, **3**:research0054.0051–research0054.0012.
3. Chen CP, Rost B: **State-of-the-art in membrane protein prediction.** *Appl Bioinformatics* 2002, **1**:21-35.
4. Lim VI: **Algorithms for prediction of alpha-helical and beta-structural regions in globular proteins.** *J Mol Biol* 1974, **88**:873-894.

5. Chou PY, Fasman GD: **Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins.** *Biochemistry* 1974, **13**:211-222.
6. Garnier J, Osguthorpe DJ, Robson B: **Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins.** *J Mol Biol* 1978, **120**:97-120.
7. Rost B, Sander C: **Prediction of protein secondary structure at better than 70% accuracy.** *J Mol Biol* 1993, **232**:584-599.
8. Jones DT: **Protein secondary structure prediction based on position-specific scoring matrices.** *J Mol Biol* 1999, **292**:195-202.
9. Cuff JA, Barton GJ: **Application of multiple sequence alignment profiles to improve protein secondary structure prediction.** *Proteins* 2000, **40**:502-511.
10. Guo J, Chen H, Sun Z, Lin Y: **A novel method for protein secondary structure prediction using dual-layer SVM and profiles.** *Proteins* 2004, **54**:738-743.
11. Lee BK, Richards FM: **The interpretation of protein structures: Estimation of static accessibility.** *J Mol Biol* 1971, **55**:379-400.
12. Rost B, Sander C: **Conservation and prediction of solvent accessibility in protein families**. *Proteins* 1994, **20**:216-226.
13. Thompson MJ, Goldstein RA: **Predicting solvent accessibility: higher accuracy using Bayesian statistics and optimized residue substitution classes.** *Proteins* 1996, **25**:38-47.
14. Li X, Pan XM: **New method for accurate prediction of solvent accessibility from protein sequence.** *Proteins* 2001, **42**:1-5.
15. Pascarella S, De Persio R, Bossa F, Argos P: **Easy method to predict solvent accessibility from multiple protein sequence alignments.** *Proteins* 1998, **32**:190-199.
16. Yuan Z, Burrage K, Mattick JS: **Prediction of protein solvent accessibility using support vector machines.** *Proteins* 2002, **48**:566-570.
17. Pollastri G, Baldi P, Fariselli P, Casadio R: **Prediction of coordination number and relative solvent accessibility in proteins.** *Proteins* 2002, **47**:142-153.
18. Ahmad S, Gromiha MM, Sarai A: **Real value prediction of solvent accessibility from amino acid sequence.** *Proteins* 2003, **50**:629-635.
19. Kim H, Park H: **Prediction of protein relative solvent accessibility with support vector machines and long-range interaction 3D local descriptor.** *Proteins* 2004, **54**:557-562.
20. Adamczak R, Porollo A, Meller J: **Accurate prediction of solvent accessibility using neural networks-based regression.** *Proteins* 2004, **56**:753-767.
21. Chen H, Zhou HX: **Prediction of solvent accessibility and sites of deleterious mutations from protein sequence.** *Nucleic Acids Res* 2005, **33**:3193-3199.
22. Sim J, Kim S-Y, Lee J: **Prediction of protein solvent accessibility using fuzzy k-nearest neighbor method.** *Bioinformatics* 2005, **21**:2844-2849.
23. Nguyen MN, Rajapakse JC: **Two-stage support vector regression approach for predicting accessible surface areas of amino acids.** *Proteins* 2006, **63**:542-550.
24. Nguyen MN, Rajapakse JC: **Prediction of protein relative solvent accessibility with**

a two-stage SVM approach. *Proteins* 2005, **59**:30-37.

25. Yuan Z, Huang B: **Prediction of protein accessible surface areas by support vector regression.** *Proteins* 2004, **57**:558-564.
26. Wang JY, Lee HM, Ahmad S: **Prediction and evolutionary information analysis of protein solvent accessibility using multiple linear regression.** *Proteins* 2005, **61**:481-491.
27. von Heijne G: **Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule.** *J Mol Biol* 1992, **225**:487-494.
28. Jones DT, Taylor WR, Thornton JM: **A model recognition approach to the prediction of all-helical membrane protein structure and topology.** *Biochemistry* 1994, **33**:3038-3049.
29. Rost B, Casadio R, Fariselli P, Sander C: **Transmembrane helices predicted at 95% accuracy.** *Protein Sci* 1995, **4**:521-533.
30. Krogh A, Larsson B, von Heijne G, Sonnhammer EL: **Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes.** *J Mol Biol* 2001, **305**:567-580.
31. Tusnady GE, Simon I: **Principles governing amino acid composition of integral membrane proteins: application to topology prediction.** *J Mol Biol* 1998, **283**:489-506.
32. Liakopoulos TD, Pasquier C, Hamodrakas SJ: **A novel tool for the prediction of transmembrane protein topology based on a statistical analysis of the SwissProt database: the OrienTM algorithm.** *Protein Eng* 2001, **14**:387-390.
33. Yuan Z, Mattick JS, Teasdale RD: **SVMtm: support vector machines to predict transmembrane segments.** *J Comput Chem* 2004, **25**:632-636.
34. Cao B, Porollo A, Adamczak R, Jarrell M, Meller J: **Enhanced recognition of protein transmembrane domains with prediction-based structural profiles.** *Bioinformatics* 2006, **22**:303-309.
35. Cserzo M, Wallin E, Simon I, von Heijne G, Elofsson A: **Prediction of transmembrane alpha-helices in prokaryotic membrane proteins: the dense alignment surface method.** *Protein Eng* 1997, **10**:673-676.
36. Granseth E, Viklund H, Elofsson A: **ZPRED: predicting the distance to the membrane center for residues in alpha-helical membrane proteins.** *Bioinformatics* 2006, **22**:e191-e196.
37. Viklund H, Elofsson A: **Best alpha-helical transmembrane protein topology predictions are achieved using hidden Markov models and evolutionary information.** *Protein Sci* 2004, **13**:1908-1917.
38. Abramson J, Smirnova I, Kasho V, Verner G, Kaback HR, Iwata S: **Structure and mechanism of the lactose permease of Escherichia coli.** *Science* 2003, **301**:610-615.
39. Huang Y, Lemieux MJ, Song J, Auer M, Wang DN: **Structure and mechanism of the glycerol-3-phosphate transporter from Escherichia coli.** *Science* 2003, **301**:616-620.
40. Beuming T, Weinstein H: **A knowledge-based scale for the analysis and prediction of buried and exposed faces of transmembrane domain proteins.** *Bioinformatics* 2004, **20**:1822-1835.

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

42. Hildebrand PW, Lorenzen S, Goede A, Preissner R: **Analysis and prediction of helix-helix interactions in membrane channels and transporters.** *Proteins* 2006, **64**:253-262.

43. Yuan Z, Zhang F, Davis MJ, Boden M, Teasdale RD: **Predicting the solvent accessibility of transmembrane residues from protein sequence.** *J Proteome Res* 2006, **5**:1063-1070.

44. Hastie T, Tibshirani R, Friedman J: **The Elements of Statistical Learning**. New York: Springer; 2001.

45. Shi L, Simpson MM, Ballesteros JA, Javitch JA: **The first transmembrane segment of the dopamine D2 receptor: accessibility in the binding-site crevice and position in the transmembrane bundle.** *Biochemistry* 2001, **40**:12339-12348.

46. Park Y, Helms V: **How strongly do sequence conservation patterns and empirical scales correlate with exposure patterns of transmembrane helices of membrane proteins?** *Biopolymers* 2006, **83**:389-399.

47. Park Y, Helms V: **On the derivation of propensity scales for predicting exposed transmembrane residues of helical membrane proteins.** *Bioinformatics* 2007, **23**:701-708.

48. Yeates TO, Komiya H, Rees DC, Allen JP, Feher G: **Structure of the reaction center from Rhodobacter sphaeroides R-26: membrane-protein interactions.** *Proc Natl Acad Sci USA* 1987, **84**:6438-6442.

49. Baldwin JM, Schertler GF, Unger VM: **An alpha-carbon template for the transmembrane helices in the rhodopsin family of G-protein-coupled receptors.** *J Mol Biol* 1997, **272**:144-164.

50. Stevens TJ, Arkin IT: **Substitution rates in alpha-helical transmembrane proteins.** *Protein Sci* 2001, **10**:2507-2517.

51. Lemmon MA, Flanagan JM, Treutlein HR, Zhang J, Engelman DM: **Sequence specificity in the dimerization of transmembrane alpha-helices.** *Biochemistry* 1992, **31**:12719-12725.

52. Javadpour MM, Eilers M, Groesbeek M, Smith SO: **Helix packing in polytopic membrane proteins: role of glycine in transmembrane helix association.** *Biophys J* 1999, **77**:1609-1618.

53. Russ WP, Engelman DM: **The GxxxG motif: a framework for transmembrane helix-helix association.** *J Mol Biol* 2000, **296**:911-919.

54. Senes A, Gerstein M, Engelman DM: **Statistical analysis of amino acid patterns in transmembrane helices: the GxxxG motif occurs frequently and in association with beta-branched residues at neighboring positions.** *J Mol Biol* 2000, **296**:921-936.

55. Vapnik V: **The Nature of Statistical Learning Theory**. New York: Springer; 2000.

56. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25**:3389-3402.

57. R Development Core Team: **R: A Language and Environment for Statistical Computing**; 2004.

58. Karatzoglou A, Meyer D, Hornik K: **Support Vector Machines in R.** *Journal of Statistical Software* 2006, **15**(9).

59. Chang CC, Lin CJ: **LIBSVM : a library for support vector machines**. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm; 2001.

60. Edelsbrunner H, Facello M, Fu P, Liang J: **Measuring proteins and voids in proteins.** In: *"Proc 28th Ann Hawaii Internat Conf System Sciences, 1995"*. vol. V: Biotechnology Computing; 1995: 256-264.

61. Edelsbrunner H: **The union of balls and its dual shape.** *Discrete Comput Geom* 1995, **13**:415-440.

62. Adamian L, Nanda V, Degrado WF, Liang J: **Empirical lipid propensities of amino acid residues in multispan alpha helical membrane proteins.** *Proteins* 2005, **59**:496-509.

63. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** *Nucleic Acids Res* 1994, **22**:4673-4680.

64. Henikoff S, Henikoff JG: **Position-based sequence weights.** *J Mol Biol* 1994, **243**:574-578.

65. Pei J, Grishin NV: **AL2CO: calculation of positional conservation in a protein sequence alignment**. *Bioinformatics* 2001, **17**:700-712.

Table 1. Prediction accuracies of different methods examined in the study

Prediction method	Prediction accuracy [%][1]
The BW method	68.67
TMX	78.71
The YU method	71.06[2]

[1] Defined as the fraction of the TM residues in the data set whose burial status was correctly predicted.

[2] Best prediction accuracy among 16 ones shown in Table 6.

Table 2. Prediction accuracies obtained by linear regression with different window sizes

Window size	Prediction accuracy
1	74.51
3	73.55
5	73.96
7	74.82
9	75.69
11	75.37
13	75.81
15	75.97
17	75.46
19	75.75
21	75.59

Table 3. Prediction accuracies obtained by increasing fractions of the 441 elements of a window of size 21

Fraction used in the prediction	Linear regression	SVR C – 10[1]	SVR C – 1	SVR C – 0.1	SVR C – 0.01
0.05[2]	75.65	70.36	73.45	75.21	74.89
0.1	76.61	71.06	75.88	75.97	74.57
0.15	76.90	70.78	75.11	76.20	74.09
0.2	77.21	70.65	75.14	75.78	73.58
0.25	76.45	70.01	75.59	75.78	73.04
0.3	76.04	71.13	75.11	75.24	72.56
0.35	75.72	71.54	74.79	75.27	71.54
0.4	75.91	72.69	74.76	75.11	71.80
0.45	75.91	72.72	75.11	74.95	71.86
0.5	76.13	72.82	75.33	75.11	71.67
0.55	76.39	72.63	75.43	75.43	72.08
0.6	76.04	72.69	75.43	75.14	71.61
0.65	75.33	72.94	75.24	75.30	70.40
0.7	74.86	73.01	74.98	74.92	70.24
0.75	75.33	73.20	75.43	74.86	69.31
0.8	75.75	72.75	75.75	74.44	68.48
0.85	75.62	72.59	75.75	74.22	67.97
0.9	75.24	72.34	75.14	74.00	67.65
0.95	75.21	72.79	75.46	74.22	67.53
1.0	75.59	73.26	75.97	74.16	67.85

[1] Regularization constant C was set to 10.

[2] Meaning that the top 5% of the 441 elements when ranked by the Fisher's index were input for the prediction.

Table 4. Top 20 elements of the 441 ones of a window of size 21 according to the Fisher's index

Rank	Position[1]	Type	Fisher's index
1	T	conservation index	0.987
2	C4	conservation index	0.534
3	N4	conservation index	0.469
4	N3	conservation index	0.307
5	N7	conservation index	0.306
6	C3	conservation index	0.248
7	T	L	0.243
8	C7	conservation index	0.240
9	T	G	0.203
10	T	I	0 .143
11	C8	conservation index	0.132
12	C1	conservation index	0.092
13	T	V	0.059
14	N1	conservation index	0.057
15	N4	I	0.057
16	N4	G	0.056
17	T	F	0.053
18	T	S	0.052
19	N8	conservation index	0.045
20	N4	L	0.041

[1] T: the target residue, C4: the 4th residue C terminal to the target residue, N4: the 4th residue N terminal to the target residue. Thus, the conservation index of the target residue is most indicative of its burial status, and the conservation index of the 4th residue C terminal to the target residue is second most indicative of the burial status of the target residue.

Table 5. Best prediction accuracies for each combination of an SVC kernel and a regularization constant C

	Regularization constant C				
Kernel	1	0.5	0.1	0.05	0.01
Linear	78.62	78.71	78.65	78.62	78.01
Radial	78.55	78.71	78.23	78.30	77.25

Table 6. Prediction accuracies obtained by SVR with different input vectors

Window size	Profile	Conservation index	PSSM (the YU method)
		C - 1[1]	
11	70.87	73.68	70.08
13	71.16	73.65	69.47
15	71.51	74.16	71.06
17	71.19	74.16	68.23
19	70.91	74.41	61.85
21	70.84	74.19	59.46
		C - 2	
11	70.55	73.17	69.85
13	70.94	73.26	69.31
15	71.16	74.31	70.52
17	71.64	73.61	67.11
19	70.68	73.90	61.54
21	70.49	74.09	59.31
		C - 5	
11	70.36	72.37	69.79
13	71.19	72.53	69.12
15	70.91	73.36	70.43
17	71.32	72.72	66.92
19	70.46	73.07	61.60
21	70.59	72.85	59.18
		C - 7	
11	70.81	72.05	70.17
13	71.03	72.15	69.15
15	71.13	72.94	70.43
17	71.19	72.37	66.89
19	70.43	72.50	61.63
21	70.52	72.08	59.18

[1] Regularization constant C was set to 1.

Table 7. Prediction accuracies for each amino acid

Amino acid	Number of occurrence	Prediction accuracy [%]	Fraction of exposed residues in the data set [%]
A	381	80.31	45.93
C	50	74.00	46.00
D	19	89.47	0.00
E	30	56.67	40.00
F	294	80.95	73.13
G	316	80.38	27.22
H	42	90.48	16.67
I	328	80.49	72.26
K	20	85.00	55.00
L	521	80.42	73.70
M	128	75.78	57.03
N	41	75.61	17.07
P	89	61.80	48.31
Q	26	76.92	26.92
R	15	100.00	6.67
S	153	71.24	39.22
T	169	73.37	48.52
V	365	79.45	65.48
W	74	81.08	70.27
Y	77	72.73	50.65

Table 8. Specificity and sensitivity of TMX

	Observed	
Predicted	Buried	Exposed
Buried	978 (70.61%)	267 (15.23%)
Exposed	407 (29.39%)	1486 (84.77%)
Sum	1385	1753

Table 9. 43 Protein chains used in the study

PDB ID	Protein	Chains
1. 1M0L	Bacteriorhodopsin	A
2. 1GZM	Rhodopsin	A
3. 1R3J	KcsA potassium channel	C
4. 1J4N	Aquaporin	A
5. 1LDF	Glycerol facilitator channel	A
6. 1XQF	Ammonia channel	A
7. 1OTS	H^+/Cl^- exchanger	A
8. 2A65	Leucine transporter	A
9. 2CFQ	Lactose permease	A
10. 1YEW	Methane monooxygenase	B, C
11. 1SU4	Calcium ATPase	A
12. 2BL2	Rotor of V-type Na^+-ATPase	A
13. 1DXR	Photosynthetic reaction center	L, M, H
14. 1KF6	Fumarate reductase (*E. coli*)	C, D
15. 1QLA	Fumarate reductase (*W. succinogenes*)	C
16. 1KQF	Formate dehydrogenase N	B, C
17. 1Q16	Nitrate reductase A	C
18. 1NEK	Succinate dehydrogenase	C, D
19. 1ZOY	Complex II	C, D
20. 1OKC	Mitochondrial ADP/ATP carrier	A
21. 1V55	Cytochrome C oxidase (aa_3 type)	B, D, G, I, J, L, M
22. 1EHK	Cytochrome C oxidase (ba_3 type)	A, B
23. 1PP9	Cytochrome bc_1 complex	D, E, G, J
24. 2GIF	AcrB multidrug efflux transporter	A
25. 2IC8	GlpG rhomboid-family intramembrane protease	A
26. 2NQ2	Putative metal-chelate-type ABC transporter	A

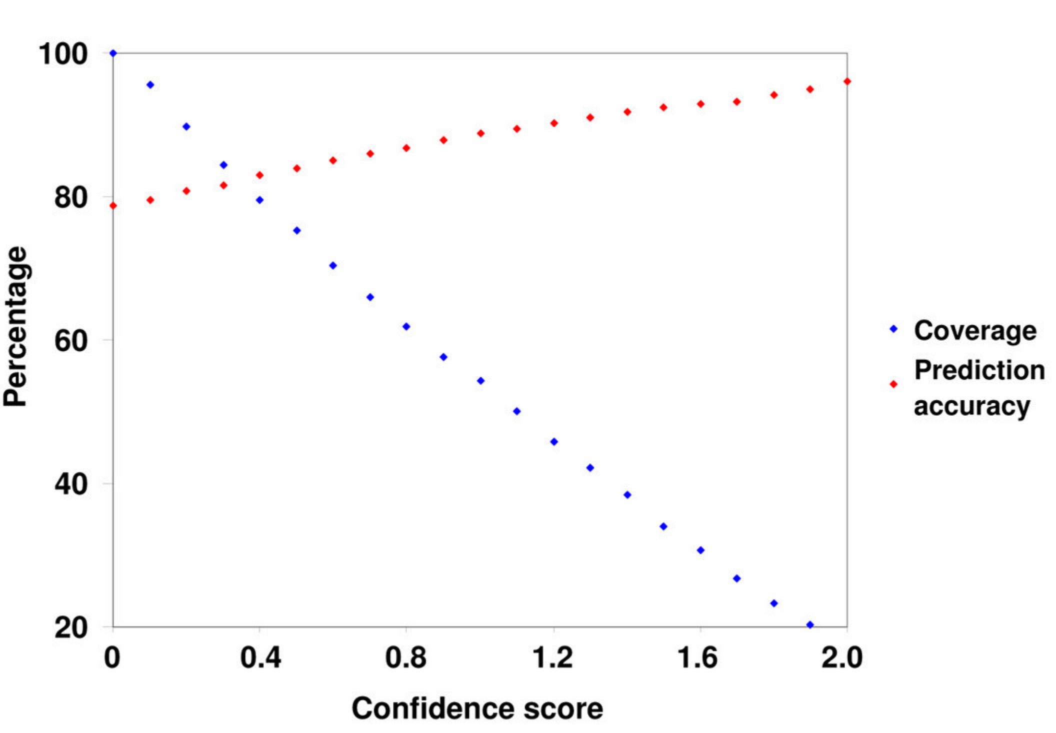


Figure 1. Prediction accuracy and coverage depending on confidence scores. When considering all predictions (i.e. predictions with a confidence score ≥ 0.00), the prediction accuracy is 78.71% and the coverage is 100%. When considering only the 1440 predictions with a confidence score ≥ 1.20, the prediction accuracy rises to 90.21% and the coverage falls down to 45.89%.

12. Acknowledgements

Verlaß dich auf den HERRN von ganzem Herzen, und verlaß dich nicht auf deinen Verstand, sondern gedenke an ihn in allen deinen Wegen, so wird er dich recht führen. Dünke dich nicht weise zu sein, sondern fürchte den HERRN und weiche vom Bösen. Das wird deinem Leibe heilsam sein und deine Gebeine erquicken.

Sprichwörter 3.5-8

I thank God of Jesus Christ for his unlimited grace towards me. About 2000 years ago, He sent Jesus Christ to save me. Furthermore, He let me have a wonderful time in Saarbruecken, Germany for the past 4 years. It has been my privilege to learn little by little how wonderfully God has created different forms of life and the whole universe. Let His name be praised forever…

My special thanks go to my advisor, Prof. Volkhard Helms. Without his unlimited patience and support for my work, this thesis would not be possible. I am also indebted to all the group members of Prof. Volkhard Helms. I would like to apologize for occupying almost all nodes on fauna occasionally.

Finally, my loving thanks go to my wife for her warm supports in all occasions. Also my lovely three sons have been a major driving force for me.