Sequence based methods for the prediction and analysis of the structural topology of transmembrane beta barrel proteins

Dissertation

zur Erlangung des Grades

des Doktors der Naturwissenschaften

der Naturwissenschaftlich-Technischen Fakultät III

Chemie, Pharmazie, Bio- und Werkstoffwissenschaften

der Universität des Saarlandes

von
Sikander Hayat

Saarbrücken
31. März 2010

Tag des Kolloquiums:	.....
Dekan:	.....
Berichterstatter:	.....
	.....
	.....
Vorsitz:	.....
Akad. Mitarbeiter:	.....

Acknowledgements

First and foremost, I am grateful to my supervisor Prof. Dr. Volkhard Helms for providing me with the opportunity to work in his research group. My work would not have been possible without his constant guidance, suggestions and support. I am also thankful to the Graduiertenkolleg 1276/1 of the DFG for providing me with an active scientific environment and sustaining my work with a generous fellowship. I would also like to thank the center for bioinformatics at the Saarland University.

I want to thank the "Coffee-club" for fun times in the afternoon breaks and my colleagues at the research group for a pleasant working atmosphere. I am especially thankful to Mazen Ahmad for lively and fruitful discussions. I am also grateful to Tihamér Geyer for his invaluable technical help with the computer systems. Furthermore, I am thankful to Dr. Yungki Park for helping me with starting my research work. I am also thankful to Nitesh Kumar Singh, and Aaron Goodman for their contributions to the HMM project.

I am especially thankful to Katharina Rudelt for her understanding, patience and support, and to Imran Rauf and Muhammad Zeshan Afzal for their help.

Last but not least, I am grateful to my parents Hayat Mohammad Khan and Badar Hayat and my sisters Salma and Amber Hayat, who always believed in me and provided me with endless moral support and encouragement.

Abstract

Transmembrane proteins play a major role in the normal functioning of the cell. Many transmembrane proteins act as a drug target and hence are of utmost importance to the pharmaceutical industry. In spite of the significance of transmembrane proteins, relatively few transmembrane 3D structures are available due to experimental bottlenecks. Due to this, it is imperative to develop novel computational methods to elucidate the structure and function of these proteins. The two major classes of transmembrane proteins are helical membrane proteins and transmembrane beta barrel proteins. Relatively more 3D structures of helical membrane proteins have been experimentally determined and in general, the majority of computational methods in the realm of transmembrane proteins deal with helical membrane proteins. However, in the recent years there has been an increased interest in the development of computational methods for the transmembrane beta barrel proteins. In this study, I focus on the transmembrane beta barrel proteins. More specifically, I present here computational methods for the prediction of the exposure status of the residues in the membrane spanning region of the transmembrane beta barrel proteins. To the best of our knowledge, the exposure status prediction is a novel problem in the realm of transmembrane beta barrel proteins. The knowledge about the exposure status of the membrane spanning residues is then used to analyse the structural properties of transmembrane beta strands. The exposure status information is also employed to identify relevant physico-chemical properties that are statistically significantly different in the transmembrane beta strands at the oligomeric interfaces and the rest of the protein surface. A method for the prediction of the beta strands in the membrane spanning regions of putative transmembrane beta barrel proteins from protein sequence has also been developed. The computational method for strand prediction is novel in the respect that it also gives the exposure status information of the residues predicted to be in the predicted transmembrane beta strands. The two computational methods developed in this study have been made available as web services. In the future, the information about the exposure status of the residues in the transmembrane beta strands can be used to identify putative transmembrane beta barrels from proteomic data. The exposure status prediction can also be extended to predict the pore region of transmembrane beta barrel proteins from sequence, which could in turn be used in the function prediction of putative transmembrane beta barrels.

Kurzfassung

Die Klasse der Transmembranproteine übernimmt eine Reihe wesentlicher Funktionen innerhalb der Zelle. Daher eignen sich viele dieser Proteine als Ziele für medizinische Wirkstoffe und sind daher von außerordentlichem Interesse für die Pharmaindustrie. Trotz ihrer Wichtigkeit wurden bislang nur wenige drei-dimensionale Strukturen von Membranproteinen erfasst, denn deren experimentelle Bestimmung hat sich als ausgesprochen schwierig herausgestellt. Aus diesem Grund erweist sich die Entwicklung von *in silico* Methoden zur *de novo* Vorhersage von Struktur und Funktion dieser Proteine von als notwendige Strategie.

Die beiden wesentlichen Klassen von Transmembranproteinen unterteilt man, basierend auf ihren charakteristischen Sekundärstrukturen, in α-helikale Proteine und β-Barrels. Erstere machen den größeren Anteil an experimentell bestimmten Strukturen aus, und auch die meisten bislang vorgestellten *in silico* Methoden konzentrieren sich auf die Modellierung solch α-helikaler Strukturen. In den vergangenen Jahren stieg daher das Interesse an Methoden zur Modellierung von transmembranen β-Barrels.

Die vorliegende Disseration beschäftigt sich vorrangig mit dieser Klasse von Transmembranproteinen, insbesondere präsentieren wir ein Verfahren zur Vorhersage der Exposition ("*Exposure*") zur Lipidschicht einzelner Residuen innerhalb der Transmembranregion von β-Barrels. Diese Vorhersage der Exposition stellt bislang ein neuartiges Problem im Feld der β-Barrels dar. Die daraus gewonnenen Informationen wurden zur Analyse der strukturellen Eigenschaften von Transmembranketten verwendet. Darüber hinaus können die Exposure-Daten zur Identifikation bedeutender physikochemischer Eigenschaften verwendet werden. Unsere Untersuchungen ergaben, dass zwischen transmembranen β-strands an Oligomer-Interfaces und dem Rest der Proteinoberfläche statistisch signifikante Unterschiede bezüglich dieser Eigenschaften auftreten. Darüber hinaus stellen wir ein Verfahren zur sequenzbasierten Vorhersage von Transmembran-Residuen mutmaßlicher β-Barrels vor, welches in Kombination mit der Vorhersage des Exposure-Status in dieser Form neuartig ist. Die beiden in dieser Studie vorgestellten Methoden sind online als Webdienste verfügbar.

Basierend auf den Exposure-Vorhersagen von β-Faltblättern ist es möglich, in künftigen Studien mutmaßliche transmembrane β-Barrels aus Proteomdaten zu identifizieren.

Zusammenfassung

Die Vorhersage von strukturellen Eigenschaften und der Topologie von Transmembranproteinen sowie deren Identifikation aus Proteomdaten ist bereits seit Jahrzehnten ein ständig behandeltes Forschungsgebiet.

Transmembranproteine unterteilt man im wesentlichen nach ihrer charakteristischen Sekundärstruktur in α-helikale Proteine und β-Barrels. Obwohl jüngst dank wesentlicher Verbesserungen im Bereich experimenteller Methoden zur Strukturbestimmung eine Vielzahl neuer drei-dimensionaler Strukturen bestimmt werden konnte, steht deren Anzahl in einem sehr niedrigen Verhältnis zu den bekannten löslichen Proteinen. Da Transmembranproteine bekanntlich eine Vielzahl wesentlicher Funktionen innerhalb einer Zelle übernehmen, eignen sich viele dieser Proteine als Ziele für medizinische Wirkstoffe und sind daher von äußerster Wichtigkeit für die Pharmaindustrie. Aus diesem Grund erweist sich die Entwicklung von *in silico* Methoden zur *de novo* Vorhersage von Struktur und Funktion dieser Proteine als notwendige Strategie.

Die meisten bislang vorgestellten *in silico* Methoden konzentrieren sich auf die Analyse von Proteinen mit α-helikalen Strukturen. Basierend auf strukturellen und physikochemischen Gesetzmäßigkeiten können die Topologien dieser Proteine heutzutage mit hoher Genauigkeit vorhergesagt werden. Andererseits existieren zu diesem Zeitpunkt nur wenige Methoden zur Identifikation und Vorhersage von Struktur sowie Topologie transmembraner β-Barrels. Prinzipiell kann deren Struktur innerhalb einer einfachen Grammatik gefasst werden: sie bestehen aus anti-parallelen β-Faltblättern, langen extrazellulären Loops, und kurzen periplasmischen Loops. Trotz der im Vergleich zu α-helikalen Proteinen simplen Struktur hat sich herausgestellt, dass die Identifikation von transmembranen β-Barrels schwieriger ist. Man führt dies auf das Fehlen von charakeristischen langen hydrophoben Aminosäurenketten innerhalb der Membranregion zurück. Die geringere Hydrophobizität, zusammen mit der strukturellen Ähnlichkeit zu den löslichen β-Barrels, macht die Identifikation von transmembranen β-Barrels zur Herausforderung. In den letzten Jahren wurden einige wenige Verfahren mit annehmbarer Genauigkeit publiziert, jedoch fehlen bislang Methoden zur umfassenden Charakterisierung der strukturellen Eigenschaften, so wie sie bereits für α-helikale Proteine entwickelt wurden.

Es existieren verschiedene Regeln, die zur Definition von Transmembranregionen herangezogen werden können und auf topologischen und geometrischen Gesetzmässigkeiten beruhen. Ein Beispiel für eine solche Regel ist das "Doppel-Repeat-Pattern". Für Ketten mit zwei alternierenden Residuen hat man demnach festgestellt, dass diese entweder zur Lipiddoppelschicht oder aber zum Proteinkern hin zeigen. Das Ziel der vorliegenden Studie besteht in der Entwicklung neuartiger Verfahren zur Vorhersage der strukturellen Eigenschaften der Residuen innerhalb von Transmembran-β-Barrels mithilfe ebensolcher Regeln. Eine dieser Eigenschaften ist der "*Exposure*"-Status einzelner Aminosäuren. Dieser gibt Aufschluss darüber, ob eine bestimmte Residue zur Membran hin ("*exposed*") zeigt oder eher im Proteininneren verborgen ist ("*buried*"). Wir zeigen, dass Residuen zwar nach außen gerichtet sein können, aber durch weitere Aminosäuren zur Membranrichtung hin abgeschirmt werden und demnach nicht

direkt mit der Membran interagieren können. Umkehrt haben wir festgestellt, dass innenliegende Residuen teilweise zur Lipiddoppelschicht hin exponiert sind.

Zunächst bestimmen wir für alle Residuen innerhalb eines Proteins, ob sie tendenziell eher exponiert oder vergraben sind. Diese Tendenzen leiten wir korrelierend zu den Frequenzprofilen der jeweiligen Aminosäuren ab, was unseres Wissens neuartig ist und daher bislang noch nicht vorgenommen wurde. Darüberhinaus unterscheiden wir zwischen Residuen innerhalb des Proteinkerns und den Residuen an der Lipid-Wasser Grenzschicht, indem wir separate Tendenzwerte für diese berechnen. Anschließend betrachten wir die für β-Barrels neu hergeleiteten Werte zusammen mit den vergleichbaren Maßstäben, die bereits für α-helikale Membranproteine bestimmt wurden, und stellen Unterschiede und Gemeinsamkeiten zwischen den Tendenzen der Residuen beider Proteinklassen heraus. Wir vergleichen die hergeleiteten Präferenzen mit bereits bekannten physikochemischen Werten aus der Literatur. Darüber hinaus vergleichen wir die "*Exposure*"-Tendenz von Residuen innerhalb des Proteinkerns für einen nicht-redundanten Datensatz von oligomeren β-Barrels. Um die Praxistauglichkeit unserer Skala nachzuweisen, entwickelten wir eine einfache Methode zur Exposure-Vorhersage transmembraner β-Barrels basierend auf der Tichonow-Regularisierung und unserer hergeleiteten Tendenzen.

Im Anschluss haben wir ein ausgereifteres Verfahren namens BTMX entwickelt, das ein zwei-stufiges "*sliding window*"-Konzept verfolgt. Diese Methode verwendet *Positional Specific Scoring Matrices* (PSSM) aus multiplen Sequenzalignments und berechnet daraus den Exposure-Status für die jeweiligen Eingabesequenzen. Die BTMX-Software ist online als Webservice zugänglich. Das Programm generiert farbig markierte Snakeplots, die mit dem jeweiligen Exposure-Status der einzelnen Residuen annotiert sind. Vergleiche unserer BTMX-Methode mit einer weiteren Methode aus der Literatur zeigen, dass BTMX hinsichtlich der Vorhersagegenauigkeit deutlich bessere Ergebnisse erzielt. Zusätzlich dazu ist BTMX in der Lage, seine Berechnungen mit einer Konfidenzscore zu bewerten, was in der Praxis Aufschluß über die Güte der Vorhersagen gibt.

Darüberhinaus sind Vorhersagen zu oligomeren Interfaces bei transmembranen β-Barrels ein weiteres ausstehendes Problem. Ebensowenig hat man bislang physikochemische Eigenschaften nachweisen können, die zur Unterscheidung solcher oligomeren Interfaces von den anderen β-Faltblättern signifikante Hinweise liefern. Wir haben in dieser Studie relevante physikochemische Eigenschaften für exponierte Residuen analysiert und konnten einige weitere Eigenschaften herausarbeiten, die künftig zur Entwicklung von Methoden zur Identifikation von oligomeren Interfaces nützlich sein können.

Diese Arbeit behandelt weiterhin Methoden zur Vorhersage von transmembranständigen β-Faltblättern basierend auf einer bekannten Proteinsequenz. Wir haben eine *Hidden Markov-Modell* (HMM)-basierte Methode zur Vorhersage der strukturellen Topologie mutmaßlicher β-Barrels entwickelt. Das Verfahren, das wir TMBHMM nennen, liefert bislang die umfassendsten *in silico* Vorhersagen zur Struktur von β-Barrel-Sequenzen. Wie auch BTMX berechnet TMBHMM auch den vermutlichen Exposure-Status der membranständigen Residuen. Die Methode ist ebenfalls als Webserver frei zugänglich.

In künftigen Studien kann das Wissen über den Exposure-Status einzelner Residuen verwendet werden, um mutmaßliche transmembrane β-Barrels aus Proteomdaten zu identifizieren. Zu diesem Zweck lassen sich beispielsweise

neue strukturelle Motive aus den Exposure-Daten herleiten. Schließlich besteht die Möglichkeit, die Exposure-Berechnungen so erweitern, um Vorhersagen der Poren von β-Barrels sequenzbasiert zu treffen. Diese wiederum erlauben es letzten Endes, Rückschlüsse auf die Funktion der betrachteten Proteine zu ziehen.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Overview

Transmembrane (TM) proteins play a crucial role in the functioning of the cell and make for important drug targets as discussed in section 1.3. In spite of their physiological importance and genomic abundance, less than 1% of the known 3D structures are helical membrane proteins (HMPs) [1]. Similarly, only 50 transmembrane beta barrel (TMB) protein crystal structures have been reported since the characterization of the 3D atomic structure of porin from Rhodobacter capsulatus [2]. It is therefore, highly desirable to develop sequence based *in silico* methods for predicting structural properties of TM proteins. In the recent years, the number of 3D structures of membrane proteins at atomic resolution has increased rapidly due to the improvement in the cloning and crystallization techniques [3]. This has led to an increase in the number of computational prediction methods for TM proteins. A close inspection of the literature in the realm of TM proteins reveals that most of the newly determined 3D structures have been HMPs. Moreover, due to the availability of significantly more HMP 3D structures and the notion that HMPs are biologically more important than the TMBs, most computational methods have also been designed to identify HMPs from proteomic data and predict the structural features of putative HMPs.

However, apart from being critical for the normal functioning of both prokaryotic and eukaryotic cells, as discussed in sections 1.3 and 1.5, TMBs also play a role in multi-drug resistance, bacterial virulence and are potential targets for the development of antimicrobial drugs and vaccines. Moreover, the identification of TMBs in endo-symbiotically derived organelles such as mitochondria provides an insight into the relation between mitochondrial morphology and its protein assembly mechanism [4–7]. Even though TMBs perform a wide array of functions, their oligomerization, folding and insertion mechanism is not yet fully understood [8–12]. Given the wide implications of TMBs, for example in the field of channel engineering, bacterial pathogenicity [13], antibiotic resistance [14] and mutational analysis, it is imperative to develop computational methods for predicting their structural and physico-chemical properties [15, 16].

The work is organized into five chapters. Chapter 1 introduces the TM proteins and provides information on the basic structure and functions of TMBs. The chapter also summarizes the main differences between the soluble, HMPs and TMB proteins. The differences in the physico-chemical properties of the inner membrane (IM) and the outer membrane (OM) are also discussed. Chapter 3 deals with the determination of propensity scales for the TMB residues at the membrane core and the membrane-water interface regions to be exposed to the lipid bilayer. To show the physical relevance of the established propensity scale, a statistical method based on ridge regression is developed to predict the exposure status of TMBs (section 3.7). In chapter 4, BTMX, the novel method established by us to predict the exposure status of TMB residues with a higher accuracy is discussed. BTMX is a two-stage, sliding window method that employs positional specific scoring matrices (PSSM) to predict the exposure status of TMB residues. The BTMX method outperforms the method established by Yuan *et al.* [17]. A web service for the BTMX method has also been implemented, which provides sequence-based, exposure status prediction

for residues in the membrane region of putative TMB sequences. The residues exposed to the bilayer are further analysed (section 4.8) and various statistically significantly different physico-chemical properties are identified between the TMB strands at the oligomeric and non-oligomeric interfaces. To the best of our knowledge, these identified physico-chemical properties have not yet been reported in the literature. These physico-chemical properties can be used to develop a prediction method for the identification of the oligomeric interfaces of the TMBs.

BTMX method relies on PROFtmb standalone program for the prediction of TMB strands. In chapter 5 we discuss a hidden markov model established by us to identify the TMB strands and predict the structural topology of TMB proteins. The TMBHMM method employs the positional frequency profile of the 20 amino acids obtained from a multiple sequence alignment (MSA) to predict the structural topology of the given putative TMB sequence. The method is novel in the respect that it not only identifies the TMB strands, but also provides the exposure status of the residues predicted to be in the membrane region. TMBHMM prediction accuracy has been compared to two known methods from the literature and it is shown that TMBHMM is at least as good as the best known methods (section 5.5) in terms of TMB strand identification. TMBHMM method has also been made available as a web service. The conclusions are presented in chapter 6. Furthermore, certain ideas that can be used to extend the work presented here are also proposed in section 6. We also point out certain physico-chemical properties that have so far only been reported for HMPs. We believe that the determination of these physico-chemical properties based on known TMB 3D structures can be useful in developing better computational methods for the structural prediction and identification of TMBs.

1.2 The outer membrane lipid bilayer

The outer membrane (OM) of gram-negative bacteria is crucial for the survival of bacteria in different environments and functions as a selective barrier by controlling the influx and efflux of solutes. The OM has been extensively studied in terms of its biochemistry and recently more details about the lipopolysaccharide (LPS) and outer membrane protein (OMP) factors have been made available [18, 19]. Figure 1 shows the diagrammatic representation of the general structure of the *Escherichia coli* cell envelope. As shown in figure 1, the OM is highly asymmetric with the inner leaflet composed of phospholipids and the outer leaflet composed mostly of LPS [18, 19]. As described by Ruiz *et al.* [18], LPS is composed of *lipid A*, a core oligosaccharide and an O-antigen polysaccharide. Divalent cations intercalated between LPS molecules are crucial for OM Structure, at the prevent repulsion between the negatively charged phosphate groups of adjacent LPS molecules. The strong interaction between fatty-acid chains and between the sugar components and the stabilization of negative charges by divalent cations allow LPS molecules to be highly compacted, giving the OM an almost gel-like appearance that is crucial for its barrier function [18]. Furthermore, the OM contains other lipopolysaccharides and can also serve as the anchor for surface organelles such as pilli that have a crucial role in pathogenesis.

The synthesis of the components of the OM, namely LPS, phospholipids and OMPs takes place in the cytoplasm and inner leaflet of the inner membrane (IM)

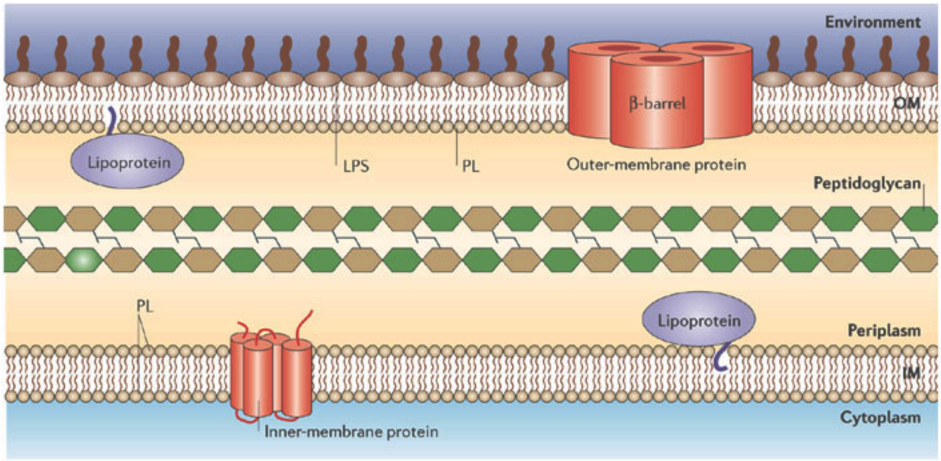


Figure 1: Diagrammatic representation of the general structure of the *Escherichia coli* cell envelope. Figure taken from Ruiz *et al.* [18].

and these components need to be transported to the OM after synthesis [18]. The OMPs are translocated across the IM with the aid of SecYEG translocon [18]. The transport of LPS from the inner leaflet to the outer leaflet of IM is mediated by the ATP-binding cassette transporter MsbA, which flips LPS from one leaflet to the other. As described by Ruiz *et al.* [18], the three possible scenarios for the transit of proteins, phospholipids and LPS from the IM to the OM via the periplasm are as follows:

- Vesicle-mediated transit
- Transit at contact sites between the two membranes
- Chaperone-mediated transit

In addition to lipids, OM consists of lipoproteins and integral OMPs. As reported [18], about 90% of all lipoproteins are located at the inner leaflet of the OM. The main focus of this study are the OMPs that occur in the OM. These OMPs are known as TMBs and consist of anti-parallel, amphipathic β-strands spanning the OM that adopt a barrel like conformation.

1.3 Transmembrane proteins

Transmembrane (TM) proteins can be broadly classified as either helical membrane proteins (HMPs) or transmembrane beta barrel proteins (TMBs), and play crucial roles in diverse physiological processes, including energy generation, signal transduction, transport of solutes across the membrane, and maintenance of ionic and proton gradients [21]. The HMPs are located in the cell membranes of both prokaryotic and eukaryotic organisms and perform a variety of biologically important functions [22]. As shown in figure 2, the HMPs

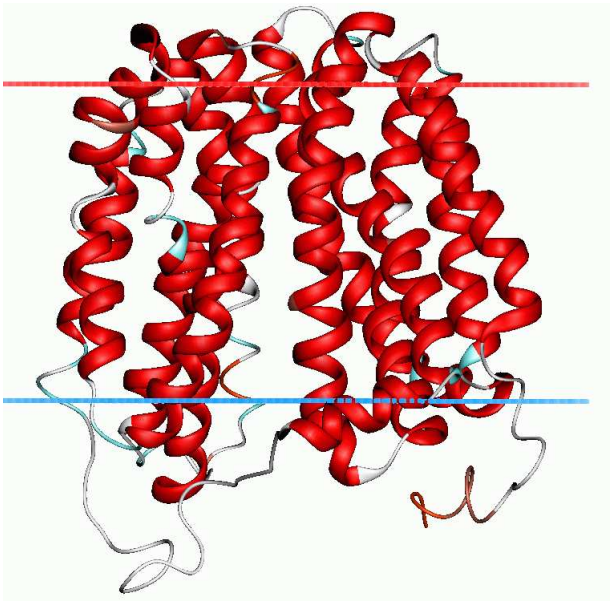


Figure 2: An example of the 3D structure and topology of a HMP (PDB-ID 2cfq). The cytoplasmic and extracellular membrane boundaries, obtained from the OPM database [20] are marked by red and blue lines, respectively.

have alpha helices as their membrane spanning regions, which mainly consists of consecutive hydrophobic amino acid residues [23] and follow the "positive-inside rule" [24]. HMPs are more abundant in both complete genomes as well as more 3D crystal structures have been determined for HMPs as compared to TMBs. These two reasons has made HMPs extensively studied class of proteins than TMBs for many years. As a consequence, there are many more prediction methods and online web services available for predicting structural properties of HMPs as compared to TMBs. Statistical computational methods such as neural networks or Hidden markov models have been demonstrated to be highly accurate in identifying membrane spanning region of HMPs [25].

While HMPs are found in all types of biological membranes including outer membranes (OM), TMBs are primarily found in the OM of gram-negative bacteria [26]. More specifically, TMBs are located in the OM of gram-negative bacteria, chloroplast and mitochondria [27–35]. As shown in figure 3, their membrane spanning regions are formed by anti-parallel beta strands, creating a channel in the form of a barrel that spans the outer membrane [13]. More specifically, TMBs perform a variety of functions including active ion transport, passive nutrient uptake, membrane anchoring, selective maltose and sucrose transport, and act as membrane-bound enzymes [13, 27, 36–38]. TMBs also function as membrane-bound enzymes [13,27,36–38] and also play a role in bacterial virulence and are potential targets for the development of antimicrobial drugs and vaccines [38–43]. TMBs are also known to act as mediators in the protein translocation across or insertion into membranes [44].

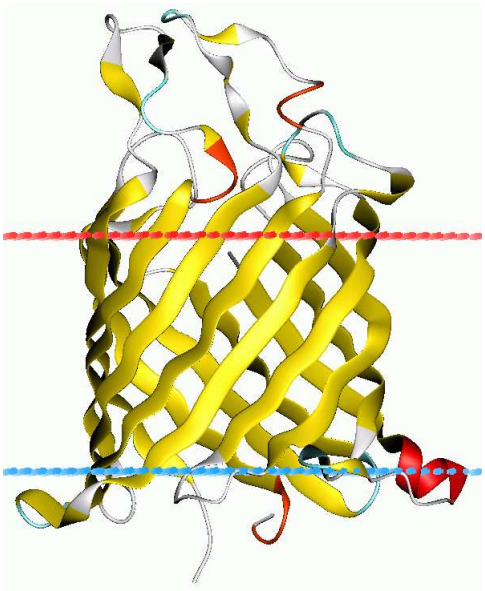


Figure 3: An example of the 3D structure and topology of a TMB (PDB-ID 1tly). The periplasmic and extracellular membrane boundaries, obtained from the OPM database [20] are marked by blue and red lines, respectively.

1.4 Structure of transmembrane beta barrel proteins

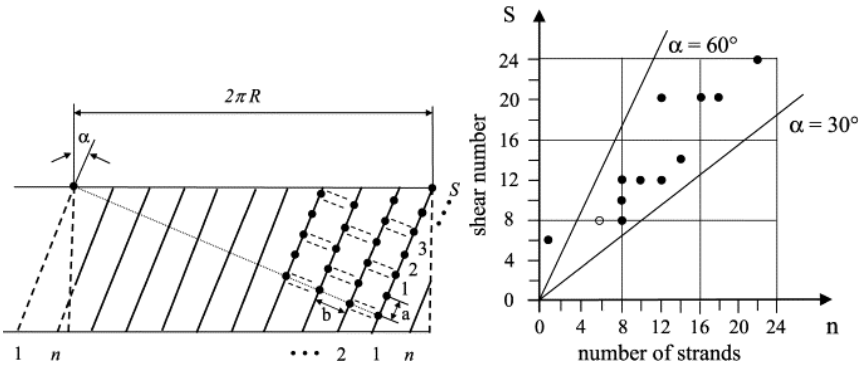
The beta barrel construction is described by the number of strands and shear number, which is defined as the inclination of the beta strands w.r.t. to the barrel axis [13]. In the experimentally known TMB 3D structures, the number of beta-strands ranges from 8 to 22. The construction principles of TMBs are illustrated in figures 4(a) and 4(b). Figure 4(a) shows the general architecture of a beta barrel with $n = 10$ beta strands and is assumed to be circular. As described by Schulz [13], in figure 4(a), the barrel is cut where the first strand reaches the upper end and then flattened out. All beta strands are assumed to run with the same tilt angle α, $a = 3.3\mathring{A}$ and $b = 4.4\mathring{A}$ are the beta pleated sheet parameters that refer to parallel, anti-parallel or mixed sheets. Thus, the relationship between the number of strands n, the shear number S and tilt angle α is given by:

$$R = \frac{[(Sa)^2 + (nb)^2]^{0.5}}{2\pi} \tag{1}$$

$$tan\alpha = \frac{Sa}{nb} \tag{2}$$

$$R = \frac{nb}{2\pi cos\alpha} \tag{3}$$

The shear number S comes with a sign and in canonical beta barrels, is positive, even and ranges between n and $n+4$. Figure 4(b) depicts the relationship between S, n and α for beta barrels. As shown (4(a)), the observed beta



(a) General architecture of beta barrel

(b) Relation between shear and strand number

Figure 4: General geometrical constraints on the architecture of TMBs [13]

barrels concentrate at tilt angles α between 30° and 60°, the shear number S is always positive in completely anti-parallel strands and radius R of the barrel increases with the number of strands n and shear number S. The open circle in figure 4(b) corresponds to the two distorted six-stranded beta barrels in the water-soluble enzyme chymotrypsin.

As describe by Schulz [13], in addition to the construction principles dictated by the beta barrel geometry, the following rules are also followed by the TMBs:

1. All β-strands are anti-parallel and locally connected to their next neighbors.

2. Both the N- and C-termini are at the periplasmic barrel end restricting the strand number n to even values.

3. On trimerization, a nonpolar core is formed at the molecular threefold axis of the porins so that the central part of the trimer resembles a water-soluble protein.

4. The external β-strand connections are long loops named L1, L2, etc., whereas the periplasmic strand connections are generally minimum-length turns named T1, T2, etc.

5. Cutting the barrel as shown in figure 4(a) and placing the periplasmic end at the bottom, the chain runs from the right to the left.

6. In all porins, the constriction at the barrel center is formed by an inserted long loop L3.

7. The β-barrel surface contacting the nonpolar membrane interior is coated with aliphatic side chains forming a nonpolar ribbon. The two rims of this ribbon are lined by girdles of aromatic side chains.

8. The sequence variability in transmembrane β-barrels is higher than in water-soluble proteins and exceptionally high in the external loops.

Furthermore, all known single chain TMBs can be described by a simple grammar [27, 45] consisting of N-terminal signal sequence, M repeats of (upward strand, extra-cellular loop, downward strand, periplasmic hairpin), and possibly a C-terminal region [46]. Given so many rules, it seems that the prediction of structural topology and identification of TMBs from sequence should be highly accurate. However, the TMBs are known to be slightly less hydrophobic than the HMPs, hence the task of identifying the membrane spanning regions is more difficult in TMBs than in HMPs [47]. This difficulty is further complicated due to the lack of a clear pattern in their membrane spanning strands, such as the stretch of 15-30 consecutive hydrophobic residues or the positive-inside rule which is followed by the HMPs. Furthermore, it is more difficult to discriminate between transmembrane strands and beta barrel structures of water soluble proteins because both of them share some common features, such as amphipathicity [25].

1.5 Function of transmembrane beta barrel proteins

Channels are required for the transport of nutrients and ions across the OM of the gram-negative bacteria which forms a protective permeability barrier around the cells, and serves as a molecular filter of hydrophilic substances [27, 40]. Depending on the mode of transport, these channels can be classified into three categories:

1. General and substrate-specific porins
2. substrate-specific transporters
3. active transporters

As described by Galdiero *et al.*, the general porins are passive pores that do not bound to their substrate. These porins usually form trimeric, water-filled pores, through which relatively small (¡ 600 Da) solutes diffuse. The transport across porins is driven by concentration gradient [40]. For nutrients that are present in μM quantities in the extracellular environment, passive transport is no longer feasible, and transport occurs via substrate-specific porins, or via substrate-specific or active transporters. In contrast to general porins, the active transporters such as FepA and FhuA bind their substrates with high affinity and active transport occurs against the concentration gradient. The energy for this transport is, for example provided by the IM proteins such as TonB [40]. The substrate-specific porins (such as LamB and ScrY) and transporters (Tsx and FadL) contain low-affinity substrate-binding sites that are saturable and allow efficient diffusion of substrates at low concentrations [40].

Furthermore, certain virulence factors act as pore forming toxins (PTFs) and aid in the spreading of the bacterium. PTFs can be classified as alpha-PFTS and beta-PTFs. As the name suggests, alpha-PTFs form pores through the insertion of amphipathic alpha helices while beta-PTFs form pores through the insertion of amphipathic beta-hairpins, which give rise to beta barrels [40]. Briefly, as described by Galdiero *et al.*, beta-PTFs are released by bacteria as water-soluble monomeric proteins that undergo a series of conformational changes upon interacting with the membrane of the target cell, where they form hydrophilic pores [40]. Binding of the toxins to the membrane aids in the

oligomerization of the toxin molecules, that leads to the formation of membrane beta barrels.

1.6 Folding, insertion and biogenesis of transmembrane beta barrel proteins

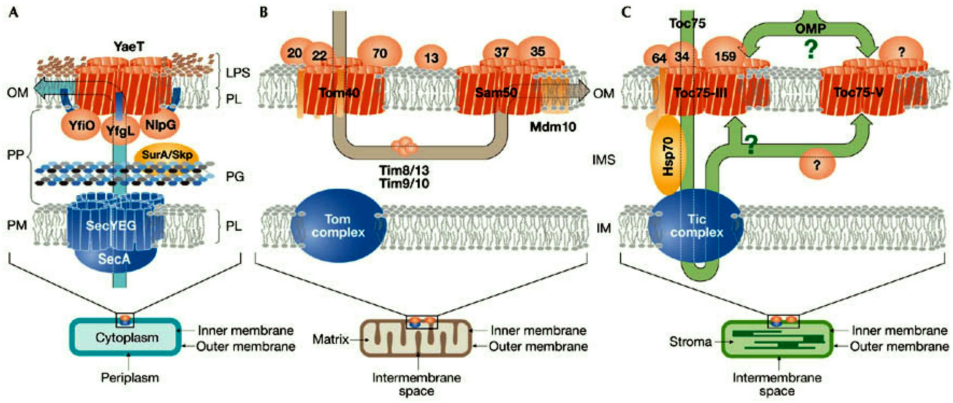


Figure 5: Beta barrel outer membrane protein insertion. Diagrammatic representation of the translocation system for bacteria, mitochondria and chloroplasts taken from Schleiff *et al.* [6].

In the recent years, our understanding of the insertion mechanism of the TMBs into the OM of endosymbiotically derived organelles has increased rapidly [6]. It is known that TMBs are translocated and inserted into the OM of bacteria, chloroplasts and mitochondria by pre-existing translocation machineries. As investigated by Eppens *et al.*, TMBs have been suggested to at least partially unfold before their insertion into the OM. The evidence for this comes from the observation that disulphide bond formation catalysed by periplasmic proteins precedes the insertion of some proteins [6, 10]. In addition, in the case of the prokaryotic as shown in figure 5 the periplasmic chaperone *Skp* and the periplasmic peptidyl-prolyl *cis-trans* isomerase *SurA* in *Escherichia coli* have also been implicated in the TMB insertion and translocation in bacteria [6, 48]. Furthermore, many TMBs such as Omp85 (from *Neisseria meningitidis*) and YaeT, its homologue in *Escherichia coli* have been shown to aid in the translocation of TMBs. Thus, as described by Schlieff *et al.*, these and other TMBs involved in the translocation of TMBS to the OM suggests that the prokaryotic machinery for OMP assembly itself consists of at least one TMB, which is associated with at least three outer membrane lipoproteins putatively involved in the folding of incoming TMBs [6].

In the case of Eukaryotic cells, it is known that endoysmbiotically derived organelles import most of their proteinaceous components from the cytosol [6]. Further as described by Schlieff *et al.*, two pore-forming proteins - Tom40 and Toc75 have been shown to be involved in the translocation of proteins across the OM in mitochondria and plastids, respectively. Interestingly, the mitochondrial channel Tom40 shares its ancestral roots with porins, while Toc75 belongs to the Omp85 family [6]. Moreover, a novel translocation system for the insertion

of TMBs has also been identified in the mitochondrial OM. As shown in figure 5, the main component of this complex is Sam50, which belongs to the Omp85 class. In the case of chloroplasts, Toc75, the central component in the translocation mechanism has been suggested to have prokaryotic origins [6]. Thus, the insertion and assembly pathway of TMBs in prokaryotic and endosymbiotically derived organelles seems to be evolutionarily conserved.

2 Statistical methods employed in this thesis

The statistical methods employed in this study are Linear regression (LR), Ridge regression (RR), Support vector machine for regression and classification (SVR and SVC), K-nearest neighbors (kNN) and Hidden Markov models (HMM). All the aforementioned methods have been applied in the realm of a supervised learning paradigm, which means that the methods learn the underlying relationship between the inputs and outputs by example [49]. Before describing the statistical models, this chapter introduces the Statistical Decision Theory (SDT) in section 2.1, which serves as the basis for developing those models.

The statistical methods employed here are divided into two major classes, namely Regression and Classification methods. The Regression methods are discussed in section 2.2 and comprise of Linear regression, Ridge regression and Support vector regression. Support vector classification, kNN and HMM constitute the classification methods and are described in section 2.3. Throughout this study, X and Y denote the input and output to the prediction-making process, respectively. The individual instances of the input and output variables are represented by x and y. For example, in section 3.10.4, the input parameters such as frequency profile, conservation index and PSSMs act as X and exposure status the amino acid residues acts as Y.

2.1 Statistical Decision Theory

In general, we seek a function $\hat{Y} = f(X)$ for predicting Y given the values of the input X [50, p.19–20]. The most common loss function $L(Y, f(X)))$ employed by this theory is the squared error: $(Y - f(X))^2$, which is the sum of squared differences between the predicted and the observed output value Y. The optimal decisions are obtained by choosing a function $f(X)$ that minimizes the loss function. The average or the expected squared error loss, is then given by [51, p.46–47]:

$$\begin{aligned} EPE(f) &= E(f(X) - Y)^2 && (4) \\ &= \iint (f(x) - y)^2 p(x, y)\, dx\, dy && (5) \end{aligned}$$

The aim here is to choose $f(x)$ such that $EPE(f)$ is minimized, which can be formally done by using the calculus of variations to give:

$$\frac{\delta\{EPE(f)\}}{\delta\{f(x)\}} = 2 \int (f(x) - y) p(x, y)\ dy = 0 \qquad (6)$$

Solving for $y(x)$, and using the sum and product rules of probability, we obtain the conditional expectation, also known as the regression function.

$$f(x) = EPE[Y|X = x] \qquad (7)$$

Thus, the best prediction of Y at any point $X = x$ is the conditional mean when the optimal value is measured by average squared error.

2.2 Regression methods

2.2.1 Linear regression

A Linear regression model assumes that the regression function $f(x)$ is linear or approximately linear in its p arguments [50, p. 19]. The linear regression model has the form

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j \tag{8}$$

where β_j's are unknown coefficients and the input variables X_j can come from different sources.

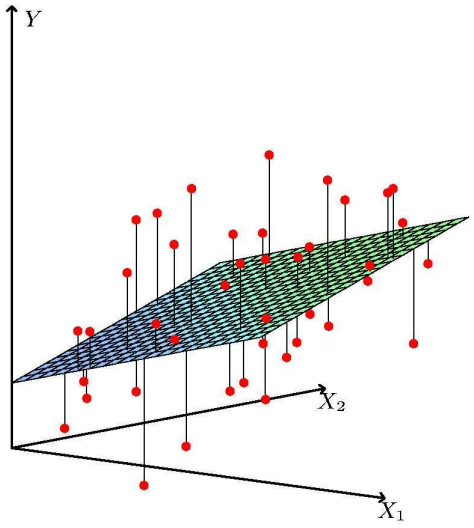


Figure 6: Linear least square fitting: We seek the linear function X (here the shaded plane) that minimizes the sum of squared residuals from Y. Figure taken from Friedman *et al.* [50, p. 43]

The most popular estimation method for the values of the β coefficients is least squares, in which the β coefficients are picked such that the residual sum of squares (RSS) is minimized (as shown in figure 6) [50, p. 42-44].

$$\begin{aligned} RSS(\beta) &= \sum_{i=1}^{n} (y_i - f(x_i))^2 \\ &= \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \end{aligned} \tag{9}$$

Let X denote the $N \times (p+1)$ matrix with each row an input vector and let y be the N-vector of outputs in the training set in equation 9. Then the RSS can be written as:

$$RSS(\beta) = (y - X\beta)^T (y - X\beta) \tag{10}$$

Differentiating with respect to β and assuming that X is nonsingular and hence $X^T X$ is positive definite, we set the first derivative to zero to obtain the unique solution:

$$\hat{\beta} = (X^T X)^{-1} X^T y \tag{11}$$

The fitted values given the training inputs are given by

$$\hat{y} = X\hat{\beta} \tag{12}$$

2.2.2 Ridge regression

Ridge regression constricts the regression coefficients by imposing a penalty on their size. This could be helpful as it avoids the cancelation of a large positive variance by the large negative correlation of its correlated partner. The ridge coefficients minimize a penalized RSS.

$$\hat{\beta}^{ridge} = \arg\min_{\boldsymbol{\beta}} \{\sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{P} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{P} \beta_j^2\} \tag{13}$$

This can be rewritten as:

$$\hat{\beta}^{ridge} = (X^T X + \lambda I)^{-1} X^T y \tag{14}$$

where I is a $p \times p$ identity matrix [50, p. 59-64].

Ridge regression is applied in this thesis to determine the propensity scales in chapter 3. The main advantage of linear regression methods is that linear processes and processes that can be approximated as linear over short ranges can be well-approximated by a linear model. Also, the theory associated with linear regression is well established and studied and thus allows for construction of different types of easily interpretable prediction models and optimizations. The main disadvantages of linear least squares are that linear models generally have poor extrapolation properties and are sensitive to outliers.

2.2.3 Support vector regression

The Support vector algorithm is a nonlinear generalization of the Generalized Portrait algorithm [52]. Support vector machines, in general, produce nonlinear boundaries in the feature space by constructing a linear boundary in a large, transformed version of the feature space [50, p. 371].

Suppose we are given training data $(x_1, y_1), (x_2, y_2), \ldots, (x_l, y_l) \subset \chi \times \mathbb{R}$, where χ denotes the space of the input patterns. In ϵ-SVR, the goal is to find a function $f(x)$ that has at most ϵ deviation from the predicted targets y_i for all training data, and at the same time is as flat as possible.

Let us consider the case when f is a linear function:

$$f(x) = \langle w, x \rangle + b \tag{15}$$

where $w \in \chi, b \in \mathbb{R}$ and flatness refers to a small w.

One way to fulfil the constraints is to minimize the norm $\| w \|^2 = \langle w, w \rangle$:

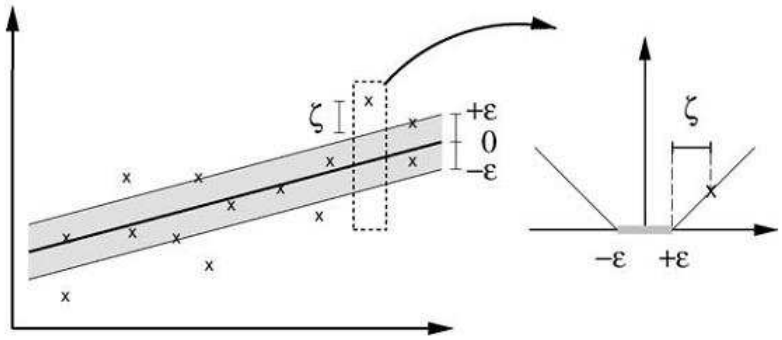


Figure 7: The soft margin loss setting corresponding to a linear support vector machine. Only the points outside the shaded region are penalized for deviations. Figure taken from Smola *et al.* [52]

$$\begin{aligned} & minimize \frac{1}{2} \parallel w \parallel^2 \\ f(x) = & \begin{cases} y_i - \langle w, x_i \rangle - b \leq \epsilon \\ \langle w, x_i \rangle + b - y_i \leq \epsilon \end{cases} \end{aligned} \tag{16}$$

Slack variables ξ_i, ξ_i^* are introduced to overcome the unbounded constraints of the optimization problem:

$$\begin{aligned} & minimize \frac{1}{2} \parallel w \parallel^2 + C \sum_{i=1}^{l} \xi_i + \xi_i^* \\ f(x) = & \begin{cases} y_i - \langle w, x_i \rangle - b \leq \epsilon + \xi_i \\ \langle w, x_i \rangle + b - y_i \leq \epsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases} \end{aligned} \tag{17}$$

The constant $C > 0$ determines the trade-off between the flatness of f and the amount up to which deviations larger than ϵ are tolerated. Figure 7 shows the situation graphically. $< w, x >$ can be defined as a simple product or as a more complicated kernel function. As discussed in chapter 4, SVR method is tested in this thesis for the determination of the positional score in the first stage of the BTMX method. The SVR method is also used for the real-value prediction of the rSASA by BTMX. In general, all SVMs have an advantage that they can be used for analysis even when the data is not regularly distributed or has an unknown distribution. Further, SVMs have an inherent flexibility in the threshold separating the different classes because of the introduction of the kernel. In addition, since the kernel may contain non-linear transformations, no assumptions need to be made about the linearity of the underlying data. A disadvantage of the SVR methods is the selection of the kernel function parameters - ϵ and σ and that the results are not transparent, as SVMs can not represent scores of all data points as a simple parameteric function.

2.3 Classification methods

Classification methods take an input vector x and assign it to one of K discrete classes C_k where $k = 1, \ldots, K$. The classes are mostly taken to be disjoint, so that each input is assigned to one and only one class and the input space is divided into decision regions. These boundaries are called decision boundaries or decision surfaces. Linear models for classification result in decision surfaces that are linear functions of the input vector x and hence are defined by $(D-1)$-dimensional hyperplanes within the D-dimensional input space. Data sets whose classes can be separated exactly by linear decision surfaces are said to be linearly separable [51, p. 179].

For a two class problem, a mistake occurs when an input vector that belongs to class C_1 is assigned to class C_2 or vice versa [51, p. 39]. The probability of this occurring is given by:

$$\begin{aligned} p(mistake) &= p(x \in R_1, C_2) + p(x \in R_2, C_1) \\ &= \int_{R_1} p(x, C_2)dx + \int_{R_2} p(x, C_1)dx \end{aligned} \tag{18}$$

Here R_k is a *decision region* and C_1 and C_2 are the two decision classes. To minimize $p(mistake)$, each x should be assigned to whichever class has the smaller value of the integrand in eqn. 18. Thus, if $p(x, C_1) > p(x, C_2)$ for a given value of x, then we should assign this x to class C_1. From the product rule of probability we have:

$$p(x, C_k) = p(C_k|x)p(x) \tag{19}$$

and thus the minimum probability of making a mistake is obtained if each value of x is assigned to the class for which the posterior probability $p(C_k|x)$ is largest. To account for mistakes of misclassification, a *loss function* is used to measure the overall loss incurred in taking any of the available decisions [51, p. 40-41]. For a value of x, let the true class be C_k and the predicted value classifies x to class C_j (where j may or may not be equal to k), the loss function can be denoted by L_{kj}, which can be viewed as the k, j element of a loss matrix. For the optimal solution the loss function should be minimized. However, the loss function depends on the true class, which is unknown. For a given input vector x, the uncertainty in the true class is expressed through the joint probability distribution $p(x, C_k)$ and so it suffices to minimize the average loss, where the average is computed with respect to the distribution given by:

$$\mathbb{E}[L] = \sum_k \sum_j \int_{R_j} L_{kj} p(x, C_k)dx \tag{20}$$

After having eliminated the common factor of $p(x)$, the decision rule that minimizes the expected loss is the one that assigns each new x to the class j for which the quantity

$$\sum_k L_{kj} p(x, C_k)dx \tag{21}$$

is minimum. Thus the classification problem can be decomposed into two parts, the *inference stage* in which the training data is used to learn a model for

$p(C_k|x)$, and the subsequent *decision stage* in which these posterior probabilities are used to make optimal class predictions. There are three distinct approaches to solving decision problems. As described [51, p. 43], in the decreasing order of complexity, these are given by:

1. Generative models: First solve the inference problem of determining the class-conditional densities $p(x|C_k)$ for each class C_k individually and separately infer the prior class probabilities $p(C_k)$. Then use Bayes' theorem in the form

$$p(C_k|x) = \frac{p(x|C_k)p(C_k)}{p(x)} \tag{22}$$

   to find the posterior class probabilities $p(C_k|x)$. The denominator in Bayes' theorem can be found in terms of the quantities appearing in the numerator, because

$$p(x) = \sum_k p(x|C_k)p(C_k). \tag{23}$$

2. Discriminative models: First solve the inference problem of determining the posterior class probabilities $p(C_k|x)$, and then subsequently use decision theory to assign each new x to one of the classes.

3. Find a function $f(x)$, called a discriminant function, which maps each input x directly onto a class label. In the case of two-class problems, $f(\cdot)$ might be binary valued and such that $f = 0$ represents class C_1 and $f = 1$ represents class C_2. It should be noted that in this case probabilities play no role.

For a detailed description of the relative merits of these three alternatives refer to Bishop *et al.* [51, p. 43].

2.3.1 Support vector classification

As described by Friedman *el al.* [50, p. 371-373], let our training data consists of N pairs $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$, with $x_i \in \mathbb{R}^P$ and $y_i \in \{-1, 1\}$. Define a hyperplane by

$$\{x : f(x) = x^T\beta + \beta_0 = 0\}, \tag{24}$$

where β is a unit vector: $\| \beta \| = 1$. A classification rule induced by $f(x)$ is:

$$G(x) = sign[x^T\beta + \beta_0] \tag{25}$$

$f(x)$ in eqn. 24 gives the signed distance from a point x to the hyperplane. Figure 8 shows the cases when the classes are separable or not. For the case when the classes are separable, we can find a function $f(x) = x^T\beta + \beta_0$ with $y_i f(x_i) > 0 \ \forall \ i$. The hyperplane that creates the biggest margin between the training points for class 1 and -1 can be obtained by solving the optimization problem:

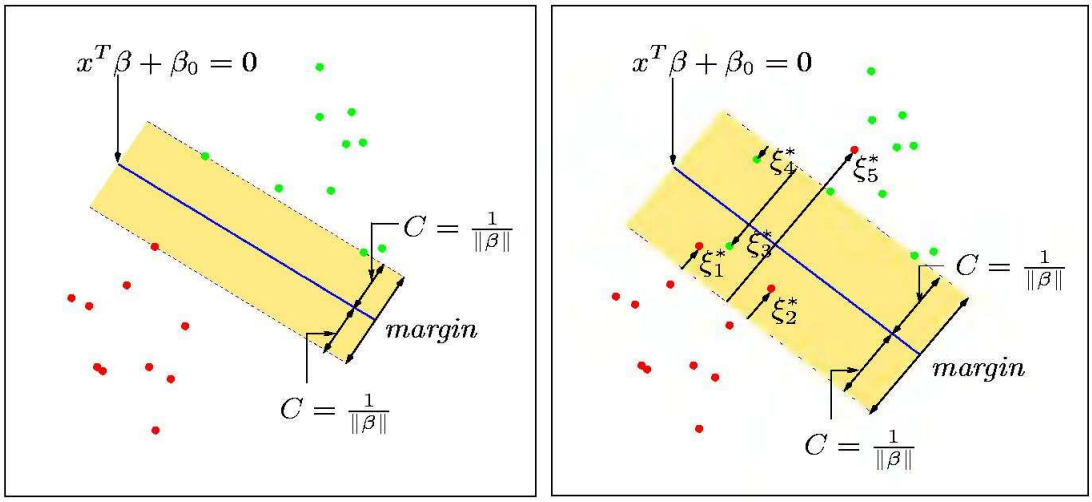


Figure 8: The left panel shows the separable case. The decision boundary is the solid line. The right panel shows the overlap case. The points labeled ξ_i^* are on the wrong side of the margin by an amount $\xi_i^* = C\xi_i$, points on the correct side have $xi_i^* = 0$. Figure taken from Friedman *et al.* [50, p. 372].

$$\max_{\boldsymbol{\beta},\boldsymbol{\beta_0},\|\boldsymbol{\beta}\|=\mathbf{1}} C \tag{26}$$

which can be reformulated as

$$\min_{\boldsymbol{\beta},\boldsymbol{\beta_0}} \| \beta \| \tag{27}$$

subject to $y_i(x_i^T\beta + \beta_0) \geq 1, i = 1, \ldots, N$,

For the case when the classes overlap in feature space [50, p. 373], one way to deal with the overlap is to maximize $\| C \|$, but allow for some points to be on the wrong side of the margin. As in section 2.2.3, we define slack variables $\xi = (\xi_1, \xi_2, \ldots, \xi_n)$ and modify the constraint in eqn. 26 as follows:

$$y_i(x_i^T\beta + \beta_0) \geq C(1 - \xi_i) \tag{28}$$

for $\forall i, \xi_i \geq 0, \sum_{i=1}^N \xi_i \leq constant$.

The value ξ_i is the proportional amount by which the prediction $f(x_i) = x_i^T\beta + \beta_0$ is on the wrong side of the margin. Misclassifications occur when $\xi_i > 1$. By dropping the norm constraint on β, defining $C = 1/ \| \beta \|$, we can write eqn. 27 as:

$$min \| \beta \| \quad subject \quad to \begin{cases} y_i(x_i^T\beta + \beta_0) \geq 1 - \xi_i \forall i \\ \xi_i \geq 0, \sum \xi_i \leq constant \end{cases} \tag{29}$$

which is the usual way the support vector classifer is defined. The SVC method is employed in this thesis in chapter 3 to classify whether the TM

residues belong to the core region of the membrane or to one of the interface regions. In a similar manner, the SVC method is also employed for the discrimination of residues present in beta-strand regions from non beta-strand regions (refer to section 3.8). The advantages and disadvantages of SVMs were discussed above in section 2.2.3.

2.3.2 k-nearest neighbors (kNN)

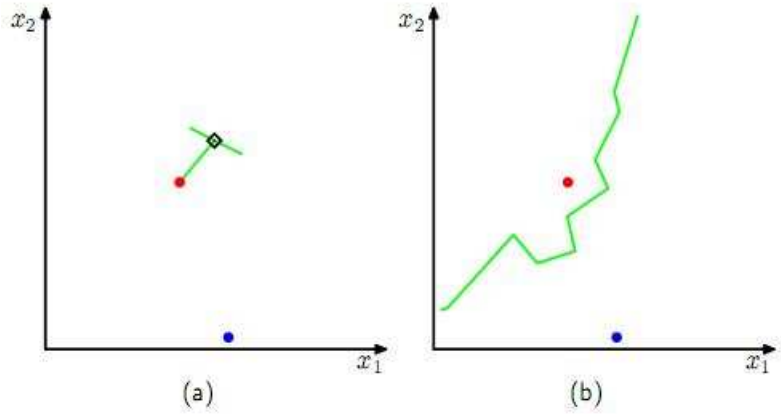


Figure 9: A new point, shown by the black diamond, is classified according to the majority class membership of the k closest training data points, in this case $k = 3$. (b) When $k = 1$, it is called the nearest neighbor rule, because a test point is simply assigned to the same class as the nearest point from the training set. Figure taken from Bishop *et al.* [51, p. 126].

Nearest neighbor methods are an example of *nonparametric approaches* to density estimation that make few assumptions about the form of the distribution. The probability density estimate can be represented as:

$$p(x) = \frac{k}{NV} \tag{30}$$

Where k is the number of points that lie in the region $\mathcal{R}$, N is the number of points being drawn and V is the volume of $\mathcal{R}$ [51, p. 122]. The k-nearest neighbor technique arises when we fix k and determine the value of V from the data. To do this, a small sphere centered on the point x at which we wish to estimate the density $p(x)$ is considered, and the radius of the sphere is allowed to grow until it contains precisely k data points. The estimate of the density $p(x)$ is then given by eqn. (30) with V set to the volume of the resulting sphere. For classification based on kNN, we apply the density estimation technique to each class separately and then make use of Bayes' theorem [51, p. 125]. Let us suppose that we have a dataset of N points, and N_k points belong to class C_k, so that $\sum_k N_k = N$. To classify a new point x, we draw a sphere centered on x containing precisely K points irrespective of their class. Suppose this sphere has volume V and contains K_k points from class C_k. The unconditional density

is given by eqn. (30) and the density assosiated with each class is then given by:

$$p(x|C_k) = \frac{K_k}{N_k V} \tag{31}$$

The class priors are given by:

$$p(C_k) = \frac{N_k}{N} \tag{32}$$

Combining eqn. 30, 31 and 32 using Bayes' theorem, we obtain the posterior probability of class membership:

$$p(C_k|x) = \frac{p(x|C_k)p(C_k)}{p(x)} \tag{33}$$

The probability of misclassification is minimized by assigning the test point to the class having the largest posterior probability, corresponding to the largest value of $\frac{K_k}{K}$. Thus to classify a new point, we identify the K nearest points from the training data set and then assign the new point to the class having the largest number of representatives amongst this set. Ties can be broken at random [51, p. 126]. In section 3.8, the kNN method is employed to classify the TM residues into core/interface and beta-strand/non beta-strand regions, respectively. An advantage of kNN method is that its classification decision is based on a small neighborhood of similar objects. So a good prediction accuracy can be obtained even if the target class consists of objects whose independent variables have different characteristics for different subsets. A drawback of the similarity measure used in kNN is that it uses all features equally in computing similarities. However, this point can be rectified and is addressed below in 2.3.3.

2.3.3 Weighted k-Nearest Neighbors (kwKNN)

In the case with kNN, all the k-nearest neighbors influence the prediction with equal weights. The kwKNN extension is based on the idea that observations which are particularly close to the new observation should get a higher weight in the decision than such neighbors that are far away from the new observation [53]. The kwKNN method is one of the methods that is tested for the discrimination of TM residues into core/interface and beta-strand/non beta-strand regions, respectively (refer to section 3.8).

2.3.4 Hidden Markov models

A hidden Markov model (HMM) is a statistical model in which the system being modeled is assumed to be a Markov process with unobserved state. In a regular Markov model, the state is directly visible, and therefore the state transition probabilities are the only parameters. In a HMM, the states are not directly visible, but the observable output is dependent on the emission probabilities. Figure 10 shows the basic architecture of a HMM. As shown, X_1, X_2, X_3 represent the unique states, a_{ij} represents the state transition probabilities from state i to state j, y_o represents the output symbols and b_{io} represents the observation symbol probability, for example, b_{14} denotes the probability of emitting y_4 from state X_1. In the realm of protein topology predictions, HMMs have been

reported to be the most successful method [47]. Apart from having high prediction accuracy, HMMs have an added advantage of being highly interpretable because of their architecture. A major drawback of HMM methods is the assumption that individual states are independent of each other. In this thesis, a HMM is used to predict the topology of TMBs. The derivation of the HMM architecture and its implementation to predicting the topology of TMB residues is discussed in chapter 5. As described by Lawrence *et al.*, the elements, problems and solutions of HMMs are discussed below [54].

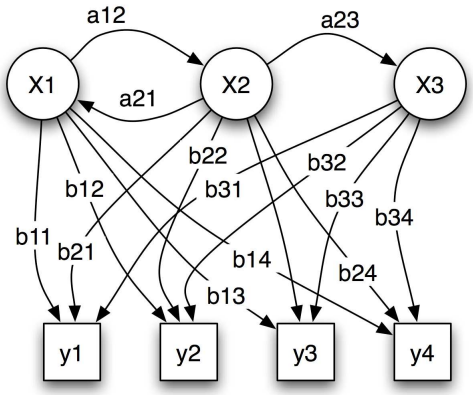


Figure 10: The figure shows the basic architecture of a HMM. Figure taken from *http* : *//en.wikipedia.org/wiki/Hidden_Markov_model*

Elements of HMM:

1. N, the number of states in the model. We denote the individual states as $S = \{S_1, S_2, \ldots, S_N\}$, and the state at time t as q_t.

2. M, the number of distinct observation symbols per state denoted as $Y = \{y_1, y_2, \ldots, y_M\}$.

3. The state transition probability $A = \{a_{ij}\}$ where

$$a_{ij} = p[q_{t+1} = S_j | q_t = S_i]$$
$$1 \leq i, j \leq N$$

4. The observation symbol probability distribution in state $j, B = \{b_j(k)\}$, where

$$b_j(k) = p[v_k \quad at \quad t | q_t = S_j]$$
$$1 \leq j \leq N$$
$$1 \leq k \leq M$$

5. The initial state distribution $\pi = \pi_i$, where

$$\pi_i = p[q_1 = S_i]$$
$$1 \leq i \leq N$$

Given appropriate values of N, M, A, B and π, the HMM can be used to generate an observation sequence

$$O = O_1 O_2 \ldots O_T \tag{34}$$

(where each observation O_t is one of the symbols from Y, and T is the number of observations in the sequence), as follows:

1. Choose an initial state $q_1 = S_i$ according to the initial state distribution π.
2. Set $t = 1$.
3. Choose $O_t = y_k$ according to the symbol probability distribution in state S_i.
4. Transit to a new state $q_{t+1} = S_j$ according to the state transition probability distribution for state S_i.
5. Set $t = t + 1$: repeat if $t < T$: else terminate.

Three basic problems for HMM: The three basic problems for HMM that need to be solved for the model to be useful in real-world applications are as follows:

1. **Evaluation problem**: Given are observation sequence $O = O_1 O_2 \ldots O_T$, and a model $\lambda = (A, B, \pi)$, Given the model how can one compute $p(O|\lambda)$, which is the probability of the observation sequence?
2. **Decoding problem**: Given the observation sequence $O = O_1 O_2 \ldots O_T$, and the model λ, how can one construct a corresponding state sequence $Q = q_1 q_2 \ldots q_T$ which explains the observations?
3. **Learning**: How do we adjust the model parameters $\lambda = (A, B, \pi)$ to maximize $p(O|\lambda)$?

Solutions to the three basic problems for HMM:

1. Solution to problem 1 (The Forward-Backward algorithm):

   We wish to calculate the probability of the observation sequence, $O = O_1 O_2 \ldots O_T$, given the model $\lambda = (A, B, \pi)$, i.e. $p(O|\lambda)$. Consider the forward variable $\alpha_t(i)$ defined as (refer to figure 11):

$$\alpha_t(i) = p(O_1 O_2 \ldots O_t, q_t = S_i|\lambda) \tag{35}$$

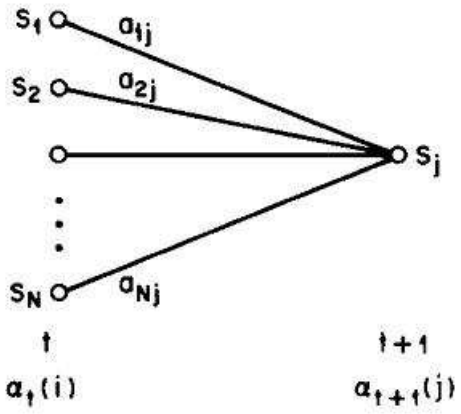


Figure 11: Computation of the forward variable $\alpha_{t+1}(j)$. Figure taken from Rabiner *et al.* [54]

i.e. the probability of the partial observation sequence, $O_1 O_2 \dots O_t$, and state S_i at time t, given the model λ. In a similar manner, we can consider a backward variable $\beta_t(i)$ defined as:

$$\beta_t(i) = p(O_{t+1} O_{t+2} \dots O_T | q_t = S_i, \lambda) \tag{36}$$

i.e. the probability of the partial observation sequence from $t+1$ to the end, given state S_i at time t and the model λ. Since the *Forward* and the *Backward* algorithm are similar in concept, we only discuss the *Forward* algorithm here.

(a) Initialization:

$$\alpha_t(i) = \pi_i b_i(O_1) \tag{37}$$

Where $1 \leq i \leq N$. Step 1 initializes the forward probabilities as the joint probability of state S_i and initial observation O_1.

(b) Induction:

$$\alpha_{t+1}(j) = [\sum_{i=1}^{N} \alpha_t(i) a_{ij}] b_j(O_{t+1}) \tag{38}$$

Where $1 \leq t \leq T-1$ and $1 \leq j \leq N$. Step 2 shows how state S_j can be reached at time $t+1$ from the N possible states at time t. $\alpha_t(i) a_{ij}$ is the probability of the joint event that $O_1 O_2 \dots O_t$ are observed, and state S_j is reached at time $t+1$ via state S_i at time t. Summing this product over all the N possible states results in

the probability of S_j at time $t+1$ with all the accompanying partial observations.

(c) Termination:

$$p(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i) \tag{39}$$

Finally, step 3 gives the $p(O|\lambda)$ as the sum of the terminal forward variables $\alpha_T(i)$'s. This is the case since:

$$\alpha_T(i) = p(O_1 O_2 \dots O_T, q_T = S_i|\lambda) \tag{40}$$

and hence $P(O|\lambda)$ is just the sum of the $\alpha_T(i)$'s.

2. Solution to problem 2 (The Viterbi algorithm):

There is no exact solution to the problem of finding the "optimal" state sequence associated with the given observation sequence because of the different possible definitions of what is "optimal". One possible optimality criterion is to choose the states q_t which maximizes the expected number of correct individual states. To implement this solution, the following variable γ is defined as follows:

$$\gamma_T(i) = P(q_t = S_i|O, \lambda) \tag{41}$$

i.e. the probability of being in state S_i at time t, given the observation sequence O, and the model λ

The single best state sequence $Q = \{q_1 q_2 \dots q_T\}$ for the given observation sequence $O = \{O_1 O_2 \dots O_T\}$ can be obtained based on the Viterbi algorithm.

We further define:

$$\delta_t(i) = \max_{\boldsymbol{q_1, q_2, \dots, q_t}} P[q_1 q_2 \dots q_t = i, O_1 O_2 \dots O_t|\lambda] \tag{42}$$

i.e., $\delta_t i$ is the score with the highest probability along a single path, at time t.

The complete procedure is as follows:

(a) Initialization:

$$\delta_1(i) = \pi_i b_i(O_1) \tag{43}$$

where $1 \leq i \leq N$

$$\psi_1(i) = 0 \tag{44}$$

(b) Recursion:

$$\delta_t(j) = \max_{\mathbf{1 \le i \le N}} [\delta_{t-1}(i) a_{ij}] b_j(O_t) \tag{45}$$

where $2 \le t \le T$ and $1 \le j \le N$

$$\psi_t(j) = \arg \max_{\mathbf{1 \le i \le N}} [\delta_{t-1}(i) a_{ij}] \tag{46}$$

where $2 \le t \le T$ and $1 \le j \le N$

(c) Termination:

$$P^* = \max_{\mathbf{1 \le i \le N}} [\delta_T(i)] \tag{47}$$

$$q_T^* = \arg \max_{\mathbf{1 \le i \le N}} [\delta_T(i)] \tag{48}$$

(d) Sequence backtracking:

$$q_t^* = \psi_{t+1}(q^*{}_{t+1}) \tag{49}$$

$t = T-1, T-2, \ldots, 1$

3. Solution to problem 3 (The Baum-Welch algorithm):

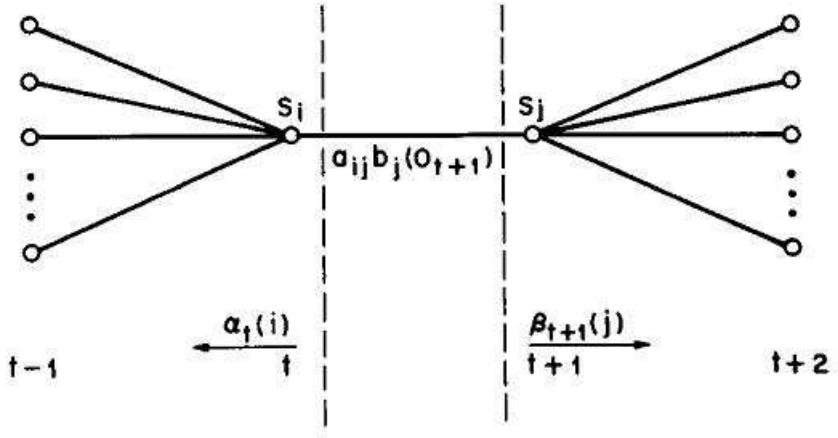


Figure 12: Computation of the joint event that the system is in state S_i at time t and state S_j at time $t+1$. Figure taken from Rabiner *et al.* [54]

Problem 3 deals with the determination of the model parameters (A, B, π) to maximize the probability of the observation sequence given the model. There is no optimal way of estimating the model parameters. However,

the model parameters could be choosen such that $P(O|\lambda)$ is locally maximized. We discuss here one such iterative procedure called the Baum-Welch algorithm that is commonly used to determine model parameters. First, the probability of being in state S_i at time t and state S_j at time $t+1$ is defined by $\xi_t(i,j)$ (refer to figure 12):

$$\xi_t(i,j) = P(q_t = S_i, q_{t+1} = S_j|O,\lambda) \tag{50}$$

which can be rewritten as:

$$\xi_t(i,j) = \frac{\alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)}{p(O|\lambda)} \tag{51}$$

Hence γ can be related to ξ as follows:

$$\gamma_t(i) = \sum_{j=1}^{N} \xi_t(i,j) \tag{52}$$

Summing over time index t, we get:

$$\text{Expected number of transitions from } S_i = \sum_{t=1}^{T-1} \gamma_t(i) \tag{53}$$

$$\text{Expected number of transitions from } S_i \text{ to } S_j = \sum_{t=1}^{T-1} \xi_t(i,j) \tag{54}$$

Thus, the reestimation formulas for π, A and B are:

$$\bar{\pi_i} = \text{expected frequency in state } S_i \text{ at time } t = 1 = \gamma_1(i) \tag{55}$$

$$\begin{aligned} \bar{a}_{ij} &= \frac{\text{expected number of transitions from state } S_i \text{ to state } S_j}{\text{expected number of transitions from state } S_i} \\ &= \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \end{aligned} \tag{56}$$

$$\begin{aligned} \bar{b}_j(k) &= \frac{\text{expected number of times in state } j \text{ and observing symbol } v_k}{\text{expected number of times in state } j} \\ &= \frac{\sum_{t=1}^{T} \gamma_t(j)_{s.t.O_t=V_k}}{\sum_{t=1}^{T} \gamma_t(j)} \end{aligned} \tag{57}$$

The reestimation formulas (eqn. 55-57) can be seen as an implementation of the Expectation Maximization (EM) algorithm, in which the E-step is the calculation of $Q(\lambda, \bar{\lambda})$ and the M-step is the maximization over $\bar{\lambda}$ [54].

2.4 Appendix

2.4.1 Expectation

As described by Bishop *et al.* [51, p.19–20], the average value of some function $f(x)$ under a probability distribution $p(x)$ is called the expectation of $f(x)$ and is denoted by EPE[f]. For a discrete distribution, it is given by:

$$EPE[f] = \sum_x p(x)f(x) \tag{58}$$

In the case of continuous variables, expectations are expressed in terms of an integration with respect to the corresponding probability density

$$EPE(f) = \int p(x)f(x)dx \tag{59}$$

Expectations of functions of several variables are denoted by:

$$EPE_x[f(x,y)] \tag{60}$$

where the subscript indicates which variable is being averaged. It should be noted that $EPE_x[f(x,y)]$ is a function of y [51, p.19–20].

2.4.2 Principal component analysis

Principal Components Analysis (PCA) is a method that reduces data dimensionality by performing a covariance analysis between factors. As defined by Bishop *et al.* [51, p. 561-563], PCA can be defined as the orthogonal projection of the data onto a lower dimensional linear space, known as the principal subspace, such that the variance of the projected data is maximized. It can also be defined as the linear projection that minimizes the average projection cost, defined as the mean squared distance between the data points and their projections.

Maximum variance formulation: We consider a data set of observations $\{\mathrm{x_n}\}$ where $n = 1, \ldots, N$, and $\{\mathrm{x_n}\}$ is a Euclidean variable with dimensionality D. The goal of PCA is to project the data onto a space having dimensionality $M < D$ while maximizing the variance of the projected data. Let us assume that $M = 1$, i.e. the projection space is one dimensional. We can define the direction of this space using a D-dimensional unit vector $\mathrm{u_1}$. Each data point $\mathrm{x_n}$ is then projected onto a scalar value $\mathrm{u_1^T x_n}$. The mean of the projected data is $\mathrm{u_1^T \bar{x}}$ and the variance of the projected data is given by:

$$\frac{1}{N}\sum_{n=1}^{N}(\mathrm{u_1^T \bar{x}} - \mathrm{u_1^T x_n})^2 = \mathrm{u_1^T S u_1} \tag{61}$$

Where S is the covariance of the data defined by

$$\mathrm{S} = \frac{1}{N}\sum_{n=1}^{N}(\mathrm{x_n} - \bar{\mathrm{x}})(\mathrm{x_n} - \bar{\mathrm{x}})^T \tag{62}$$

To maximize the projected variance $u_1^T S u_1$ with respect to u_1 we have to enforce the constraint to prevent $\| u_1 \| \rightarrow \infty$. To enforce this constraint, we introduce a Lagrange multiplier λ_1 and then make an unconstrained maximization of

$$u_1^T S u_1 + \lambda_1 (1 - {u_1}^T u_1) \tag{63}$$

By setting the derivative with respect to u_1 equal to zero, the stationary point is obtained when:

$$S u_1 = \lambda_1 u_1 \tag{64}$$

which says that u_1 must be an eigenvector of S. On multiplying by ${u_1}^T$ and making use of ${u_1}^T u_1 = 1$, we see that the variance is given by

$$u_1^T S u_1 = \lambda_1 \tag{65}$$

So, the variance will be a maximum when u_1 is set equal to the eigenvector having the largest eigenvalue λ_1. This eigenvector is known as the first principal component. Additional principal components can be defined in a similar fashion by choosing each new direction to be that which maximizes the projected variance amongst all possible directions orthogonal to those already considered. For the general case of M-dimensional projection space, the optimal linear projection is defined by M eigenvectors $u_1 u_2 \dots u_M$ of data covariance matrix S corresponding to the M largest eigenvectors $\lambda_1 \lambda_2 \dots \lambda_M$ [51, p. 561-563]. In this thesis, PCA is used in section 3.8 to determine the principal components of the physico-chemical scales identified to be statistically significantly different for the residues belonging to the beta-strands and residues at the non beta-strand regions.

3 Statistical propensities of transmembrane amino acid residues to be exposed to the lipid bilayer

3.1 Overview

Based on the concept of biological evolution, it is commonly assumed that protein sequences and 3D structures are shaped by random mutations and by selection due to biological and physico-chemical constraints [55]. Therefore, statistical analysis of representative samples often helps in revealing the underlying physico-chemical principles that control e.g. protein folding, stability, solubility in various solvents, or protein function [56]. If thorough associations can be made, they may reveal so far unknown biological principles about, for example, protein-protein interactions, localization to particular cellular compartments etc [57]. Here, we focus on the analysis of 3D atomic structures of integral membrane proteins. We separately derived the propensities of amino acid residues to be exposed to the lipid bilayer at the hydrophobic core and at the interface regions, respectively. The differences found between the propensities of TMB and HMP residues and between the hydrophobic core region of the membrane and the bilayer interfaces are related to the structural principles of TMBs and HMPs. As an application of the novel scales developed here, we show that these propensities can be used to predict the exposure status of TMB residues with an accuracy of 77.91% and 80.42% for the hydrophobic core and interface regions, respectively.

In the case of HMPs, it has previously been shown that sequence conservation patterns have a strongly negative correlation with exposure patterns of amino acid residues of the TM domains [58, 59]. Further analysis addressed the differences between the propensities of HMP residues at the hydrophobic core and at the interface regions of the inner membrane (IM) and their structural preferences [60]. This information was vital in further understanding the effects of environmental differences of the regions where these residues reside. The exposure patterns of HMP residues can be predicted from the amino acid sequence of the protein based on knowledge-based methods [61]. When consequently combined with sequence conservation patterns, the exposure status of HMP residues in the TM segments can be predicted with an accuracy of about 80% [62]. Incorporating additional information about helix-helix interactions increased the prediction accuracy to 88% [63]. Recent prediction studies in the realm of TMBs have addressed the prediction of TM secondary structure, topology and side chain orientation [17, 64, 65]. Recent studies have also identified characteristic amino acid preferences for different regions of TMBs [39, 66–68]. However, as of now, no quantitative propensity scale exists that accounts for the propensities of TMB residues to be exposed to the lipid bilayer. Moreover, a shortcoming of many machine learning and statistical approaches is that they do not provide insight into their working mechanism and into the governing forces or principles for the classifications made. In this respect, the analytical approach introduced in a recent study [62] provides a straight-forward strategy for deriving propensity scales that allow insightful interpretations within the regime of linear regression.

The aim of this study is to characterize the differences in the exposure status propensities (i.e. if the residues prefer to be exposed to the lipid bilayer

or not) of TM residues. Here, we do not distinguish between the interaction of the nascent protein chains with the translocon or transport machineries that lead to the integration of HMPs and TMBs into the lipid bilayer membranes *in vivo* and the interaction of these assembled TM proteins with the lipid bilayer. To this end, we first derived propensity scales for transmembrane amino acid residues based on known 3D structures of TMBs and HMPs. In contrast to the hydrophobic region, much less is known about the interface region, which forms a boundary at the lipid membrane and the aqueous solvent and has different physico-chemical properties than the rest of the membrane [69]. Hence, we implemented individual propensity scales for the interface and the hydrophobic core region of lipid bilayer membrane to better understand the differences in residue propensities in the two regions. The TM residues were henceforth classified as belonging to one of the following four categories: HMP residues residing at the hydrophobic core (HTMC: for helix in transmembrane core), or at the lipid-water interface of the inner membrane (HTMI) and TMB residues residing at the hydrophobic core (BTMC), or at the lipid-water interface of the outer membrane (BTMI). Three novel propensity scales were then derived for the residues belonging to BTMC, HTMI and BTMI categories. The propensities for residues belonging to the HTMC class were taken from a related work by two of the authors [70], where this scale was termed MO scale.

Here, we first introduce the respective scales in sections 3.3.1, 3.3.2 and 3.3.3 and then compare the propensities of residues to exposed to the bilayer in the different regions 3.3.4. The derived scales are then compared with the known physico-chemical scales in section 3.4 to determine which energy-based, structural and hydrophobic scales correlate well with the propensity scales derived here. The BTMC scale captures the propensity of residues in the core region of the OM to be exposed to the bilayer. Furthermore, to investigate the relation between the exposure status propensity and the oligomerization state of TMBs, two novel propensity scales called BTMC_{mono} and BTMC_{oligo} are then derived based on two separate data sets exclusively consisting of monomeric and oligomeric TMBs (refer to section 3.5.1). BTMC_{mono} and BTMC_{oligo} scales have a correlation coefficient of 0.86. The correlation with known physico-chemical scales obtained from the AAIndex database [71] is discussed in section 3.6. As an application of the derived propensities, a computational method to predict the exposure status of the TM residues at the membrane core and the interface regions has been implemented. The prediction accuracy of the computational method is 77.91% and 80.42%, for the core and the interface regions, respectively. The derivation of the prediction accuracy of the exposure status prediction method is discussed in section 3.7.

3.2 Relation between hydrophobicity and residue exposure

Residues in the BTMC and BTMI data sets were labelled as either being "buried" or "exposed" to the lipid bilayer based on their respective rSASA values. The cutoff was empirically chosen to be 0.03 for the hydrophobic core region and 0.0 for the interface regions. As described in the methods section 3.10.2, these rSASA cutoff values were chosen such that the respective data sets for the core and the interface regions had equal number of residues labelled as buried and exposed. Table 1 shows the per-amino acid exposed and buried

count and the respective average conservation indices for the residues in the core region. The corresponding values for the residues in the interface regions are shown in table 2. As shown in tables 1 and 2, the exposed/buried is more than 1.0 for the hydrophobic residues PHE, ILE, TRP, LEU, VAL, TYR and ALA at both the core and the interface regions, respectively. This shows that most neutral and hydrophilic residues are buried in the protein structure and are not exposed to the lipid bilayer. In the case of MET residues in the interface region (table 2), the exposed/buried ratio reveals that although being hydrophobic, more MET residues at the interface regions are buried in the protein structure. This could be due to the shielding of MET residues by the bulkier aromatic residues present in the interface regions. Also, as shown in table 1, a higher percentage of PRO residues are found to be exposed to the lipid bilayer, which could be due to tendency of PRO residues to occur at protein "turns".

Further, tables 1 and 2 also show the average conservation indices of the residues labelled as exposed or buried at the core and the interface regions, respectively. In the case of the core region, the exposed residues are, in general found to be less conserved than the buried residues. The two anomalies to this rule are the MET residues labelled as buried and PRO residues exposed to the bilayer. As shown in table 1, buried MET residues show a low conservation as compared to the other buried residues, while the exposed PRO residues are found to be highly conserved. The exposed ASP residues also show high conservation, but in this case, it could be attributed to the low residue count. In the case of the residues at the interface regions, on average, the exposed LEU, VAL, MET, TYR, GLY and ASP residues are found to be more conserved than their buried counterparts. In all the cases except TYR, this could again be due to less number of data points available. However, in the case of TYR, 61 exposed residues are, found to have an average conservation index of 0.71, which is significantly higher than the other exposed amino acid residues. In total 869 and 856 residues in the membrane core region are labelled as exposed and buried, respectively. The average conservation indices calculated for all the exposed residues are found to be 0.15 and 0.14, for the core and the interface regions, respectively. The corresponding average conservation indices for the residues buried in the protein structure are 0.50 and 0.27. This shows that, in general residues exposed to the bilayer are less conserved than the buried residues.

3.3 Analysis of the derived propensity scales

The derived propensity scales are shown in table 3. The TMB scales (namely BTMC, BTMI) were derived based on the TM portions of the 20 3D structures of TMBs obtained from the OPM database [20] (refer to section 3.10.1). The HTMI propensity scale for residues in the interface regions of HMPs was derived from a data set of 41 non-homologous HMPs as described by Park *et al.* [70]. The HTMC propensity scale for residues in the hydrophobic core region of HMPs was taken from our previous work [70]. Note that the amplitudes of the four different scales vary due to their mathematical construction (Table 3). Also, the division into predominantly buried or exposed propensities is relative and is generally not equal to zero. The propensity scale values for TMB residues can be roughly grouped based on the amino acid polarities. Mostly, lower values were obtained for polar or charged residues, which have a lower propensity to be exposed to the lipid bilayer. On the other hand, higher values were assigned

core region					
Amino acid	Exposed		Buried		Ratio
	count	avg. CI	count	avg. CI	Exposed/Buried
PHE	65	-0.03	15	0.64	4.33
ILE	63	-0.20	17	0.01	3.71
TRP	21	0.24	9	0.89	2.33
LEU	170	0.08	39	0.06	4.36
VAL	124	-0.10	35	0.21	3.54
MET	26	-0.54	13	-0.29	2.00
TYR	66	0.76	56	0.97	1.18
ALA	115	-0.08	68	0.33	1.69
THR	47	-0.13	71	0.12	0.66
HIS	6	-0.07	8	-0.11	0.75
GLY	102	0.74	129	1.02	0.79
SER	16	-0.16	96	0.21	0.17
GLN	4	0.29	43	0.27	0.09
ARG	3	0.74	60	0.89	0.05
LYS	5	0.38	30	0.42	0.17
ASN	8	0.30	62	0.12	0.13
GLU	1	0.51	63	0.89	0.02
PRO	22	1.50	1	0.13	22.00
ASP	5	1.29	41	0.59	0.12
Total	869	0.15^a	856	0.50^a	

Table 1: Amino acid distribution at the core region. ^{a}average conservation index for all the residues.

to residues that show a higher propensity to be exposed to the bilayer. As discussed, these amino acids are predominantly hydrophobic. These findings are in concert with experimental results [13], which illustrate that a slightly less hydrophobic exterior of TMBs (as compared to HMPs) is necessary for their translocation via the IM [5, 8, 10, 72–74]. The derived scales are described in sections 3.3.1, 3.3.2 and 3.3.3 and the propensities of residues to exposed to the bilayer in the different regions are compared in section 3.3.4. To determine which energy-based, structural and hydrophobic scales correlate well with the propensity scales derived here, the derived scales are then compared with the known physico-chemical scales in section 3.4.

3.3.1 BTMC propensity scale

Table 3 shows BTMC, the propensity scale for the amino acid residues in the core region be exposed to the bilayer. A numerically larger value reflects the tendency of an amino acid to be exposed to the bilayer. Thus, as shown in table 3, hydrophobic residues such as PHE, ILE, TRP, LEU, ALA and VAL have a higher tendency to be exposed to the lipid bilayer as compared to the neutral and hydrophilic residues. Interestingly, the higher propensity of PRO residues to be exposed to the bilayer, as suggested by the propensity scale value of 0.194, is in accord with the anomalous behavior of PRO residues to be exposed to the bilayer (refer table 1). As discussed in section 3.2, this could be due to the presence of PRO residues at the beta turns [68, 75]. Furthermore, the propensity values for MET and TYR residues at -0.359 and -0.191, respectively,

interface regions					
Amino acid	Exposed		Buried		Ratio
	count	avg. CI	count	avg. CI	Exposed/Buried
PHE	33	-0.12	11	0.60	3.00
ILE	13	-0.21	7	-0.11	1.86
TRP	39	0.07	5	1.09	7.80
LEU	30	0.02	11	-0.08	2.73
VAL	32	-0.17	8	-0.40	4.00
MET	3	-0.45	9	-0.66	0.33
TYR	61	0.71	12	0.30	5.08
ALA	21	-0.17	14	-0.03	1.50
THR	4	-0.98	38	-0.06	0.11
HIS	5	0.34	8	0.34	0.63
GLY	16	0.89	36	0.69	0.44
SER	1	-0.34	35	-0.05	0.03
GLN	10	-0.35	8	0.47	1.25
ARG	3	-0.03	21	1.05	0.14
LYS	1	-0.14	14	-0.11	0.07
ASN	2	0.88	12	0.15	0.17
GLU	2	-1.21	18	0.72	0.11
PRO	6	0.25	1	0.90	6.00
ASP	3	0.44	19	0.17	0.16
Total	285	0.14^a	287	0.25^a	

Table 2: Amino acid distribution at the interface regions. ^{a}average conservation index for all the residues.

reflects their lower tendency to be exposed to the bilayer. In the case of TYR, this could be due to the high conservation index of TYR residues labelled as buried as shown in table 1. In the case of MET, the residues are found to have a very low conservation irrespective of the exposure status. This behavior goes against the general trend that buried residues have a higher conservation index than exposed residues [76].

3.3.2 BTMI propensity scale

BTMI is the derived propensity scale for the residues in the interface regions of the outer membrane. As described in section 3.10.1, the interface regions are defined the region that spans ± 65.0% to the membrane thickness. The differences in the physico-chemical properties of the membrane core and the interface regions have been attributed to the increased interaction of the lipids with the water at the lipid-water interface [60, 69]. For the interface region, the hydrophobic residues such as PHE, ILE, TRP, LEU, VAL, MET and TYR show a higher propensity to exposed to the bilayer. However, the propensity value for ILE and MET residues to be exposed to the bilayer is 0.271 and 0.236, respectively, which is lower than the propensity values of the other hydrophobic residues in the interface regions (table 3). This could be due to the fact that, like in the membrane core region, ILE and MET residues in the interface regions, on average have a low conservation index irrespective of the exposure status (refer table 1). Interestingly, PRO residues in the interface regions show a higher propensity to be exposed to the bilayer as well.

Amino acid	BTMC scale for core region	BTMI scale for interface regions	HTMC scale for core region	HTMI scale for interface regions
PHE	0.000	0.414	-0.01	0.017
ILE	0.100	0.271	0.05	0.034
TRP	-0.025	0.470	-0.03	0.050
LEU	0.445	0.390	0.02	0.010
VAL	0.118	0.468	0.02	-0.002
MET	-0.359	0.236	-0.23	-0.009
TYR	-0.191	0.421	-0.15	-0.021
ALA	-0.009	0.150	-0.09	0.000
THR	-0.376	0.067	-0.18	-0.006
HIS	-0.328	0.151	-0.24	-0.011
GLY	-0.319	0.144	-0.18	0.032
SER	-0.563	0.049	-0.19	-0.018
GLN	-0.517	0.144	-0.22	-0.020
ARG	-0.485	0.101	-0.21	-0.016
LYS	-0.455	0.120	-0.10	-0.012
ASN	-0.472	0.107	-0.23	0.022
GLU	-0.495	0.083	-0.20	-0.019
PRO	0.194	0.396	-0.10	0.000
ASP	-0.451	0.133	-0.27	-0.023

Table 3: Propensity scales

3.3.3 HTMI propensity scale

Table 3 also includes an analogous propensity scale for the HMP residues present at the interface regions of the inner membrane. This HTMI scale is an extension to the HTMC (MO) scale previously established for HMP residues at the hydrophobic core region of the IM [70]. As shown in table 3, hydrophobic HMP residues, like their TMB counterparts, have a higher tendency to be exposed to the bilayer as compared to the polar residues (as indicated by numerically higher propensity values in the derived scales). This yields an overall hydrophobic exterior surface, which plays a major role in the folding and insertion mechanism of both TMBs and HMPs [11]. Like TMB residues, the HMP aromatic residues, PHE and TRP present at the inner membrane interfaces have a higher propensity to be exposed to the lipid bilayer, whereas TYR residues show a tendency to remain hidden from the surrounding environment. The propensity values for GLY and ASN residues are found to be high, suggesting that these residues have a higher tendency to be exposed to the bilayer membrane at the interface regions. Interestingly, it has already been shown that irregular structures that are enriched in GLY, PRO, ASN and SER residues are more likely to be exposed to the bilayer at the interface regions (Table 3: HTMI scale) of the IM [60].

3.3.4 Comparison of the derived propensity scales

As shown in table 3, for the BTMI scale, the small neutral GLY residues have a low propensity to be exposed to the membrane, however the GLY propensity value for the HTMI scale is relatively high (w.r.t. to the other residues in the HTMI scale), signifying a high propensity of the GLY residues present in the interface regions of the inner membrane to be exposed to the membrane. As

suggested before in section 3.3.3, this could be due to the presence of GLY residues in the interface helices. Further, in TMBs, MET residues have a low propensity to be exposed to the lipid bilayer in both at the core and the interface regions (refer to table 3). MET residues are hydrophobic and are therefore expected to be exposed to the bilayer in both BTMC and BTMI scales. Besides its anomalous propensity to be hidden from the bilayer, MET residues are also weekly conserved, both when exposed and buried as shown in tables 1 and 2. In the case of TMBs, the hydrophilic PHE residues at the core and at the interface regions are found to have a high propensity to be exposed to the lipid membrane with a propensity scale value of 0.194 and 0.396, respectively. TRP residues, with propensity scale values of -0.025 and 0.470 for the BTMC and BTMI scales, respectively, are different tendencies to be exposed to lipid membrane based on the distance from membrane center. As shown in table 3, TRP residues have a propensity to be buried in the protein structure when present at the membrane core region, however, when present at the interface regions, TRP residues prefer to be exposed to the bilayer. Interestingly, TRP residues are capable of forming a hydrogen bond with the NH group, but at the same time also have the largest non-polar surface and hence are known to be prone to be exposed to the lipid bilayer [60].

Finally, TYR residues at the interface regions of TMBs also have a high propensity to be exposed to the bilayer while the TYR residues in the core region show a tendency to be not accessible to the membrane. This could be due to the fact that the terminal hydroxyl group may not easily find hydrogen-bonding partners. This preferential exposure of TYR residues at the membrane interface regions is likely caused by a combination of snorkelling effect and expulsion of bulky residues from the interior. Briefly, snorkelling is defined as the tendency of polar side chains to orient themselves away from the hydrophobic membrane core and point to the interfacial or aqueous regions [77]. It is known that TYR residues show snorkelling behavior because their OH groups tend to form hydrogen bonds with polar groups, including water at the interface [69].

Scale	BTMC scale for core region	BTMI scale for interface regions	HTMC scale for core region	HTMI scale for interface regions
BTMC	1.0	0.80	0.84	0.48
BTMI	-	1.0	0.71	0.41
HTMC	-	-	1.0	0.56
HTMI	-	-	-	1.0

Table 4: Correlation of the derived propensity scales.

Correlation analysis of the derived scales was performed to measure the overall similarities and differences between the derived scales. As shown in table 4, the HTMC scale, is highly correlated with both BTMC and BTMI scales. The correlation coefficient between HTMC and BTMC was found to be 0.84. The corresponding value of the correlation coefficient between HTMC and BTMI scale was 0.71. Also the BTMC and BTMI scales show significant correlation (0.80). This finding agrees with prior observations that the exposure of non polar residues and hence the net exterior hydrophobicity plays a major role in the insertion and folding of TM proteins [18]. HTMI scale shows a weak correlation with all the three propensity scales. This means that, although the

membranes in which HMP and TMB proteins reside have different physico-chemical properties [18, 78], the lipid-exposed surface patches of TMBs at the membrane core and interface regions have similar physico-chemical properties as the lipid-exposed HMP surface patches at the membrane core of the bilayer. Whereas the interface regions of HMPs seem to have different physico-chemical properties. It is noteworthy to mention that this could also be due to the presence of irregular structures and interface helices that tend to have more exposed residues to the lipid membrane. More importantly, as discussed in section 3.4, the TMB scales are only marginally correlated with the established hydrophobicity scales, suggesting that this correlation can not be justified based on only residue polarity. Instead, one must also consider other factors such as transfer energy, residue bulkiness and packing density, to relate the different behaviors of membrane regions and forces of biological evolution.

3.4 Correlation with physico-chemical scales from the literature

To determine which physico-chemical features determine the lipid exposure of amino acids located in different regions of the membrane, the derived propensity scales were correlated with established physico-chemical scales obtained from the AAIndex database [71]. Tables 5, 6 and 7 show the correlation of the four propensity scales derived here, with energy-based, hydrophobicity-based and structural scales obtained from the AAIndex database [71], respectively. It is to be noted that the primary focus of this analysis was to find physico-chemical properties that correlate well with the TMB scales. The corresponding correlation values for the HMP scales are presented for comparison purposes.

3.4.1 Correlation of propensity scales with energy-based scales

Scale	BTMC scale for core region	BTMI scale for interface regions	HTMC scale for core region	HTMI scale for interface regions
MIYS850101	0.66	0.72	0.65	0.50
MIYS990102	-0.73	-0.75	-0.73	-0.47
MIYS990103	-0.66	-0.70	-0.62	-0.42
MIYS990104	-0.65	-0.74	-0.63	-0.45
MIYS990105	-0.69	-0.76	-0.64	-0.52
RADA880101	0.77	0.68	0.74	0.55
RADA880102	0.65	0.75	0.66	0.54
RADA880104	0.71	0.59	0.61	0.43
EISD860101	0.71	0.78	0.65	0.49
PLIV810101	0.77	0.82	0.74	0.47
NOZY710101	0.60	0.81	0.66	0.52
BULH740101	-0.74	-0.84	-0.76	-0.39
ROBB790101	0.62	0.75	0.65	0.56

Table 5: A. Correlation with (Energy based) physico-chemical scales.

Table 5 shows the correlation of the propensity scales derived here with known energy-based scales. As shown, the BTMI scale has a stronger inverse correlation with scales MIYS990101, MIYS990102, MIYS990103, MIYS990104

and MIYS990105 than the BTMC scale. Briefly, these scales are related to the pair-wise contact energies of amino acid residues [79]. The BTMC scale has a higher correlation with the RADA880101 and RADA880104 scales, which represent the transfer free energy from the cyclo-hexane (chx) to water and octanol, respectively, while the BTMI scales is better correlated to the RADA880102 scale, which represents the transfer free energy from octanol to water [80]. EISD860101 scale for the solvation free energy [81] shows a correlation of 0.78 and 0.71 with the BTMI and BTMC scales, respectively. Scales PLIV810101 (partition coefficient) [82], NOZY710101 (Transfer energy, organic solvent/water) [83] and ROBB790101 (Hydration free energy) [84] show a high positive correlation with the BTMI scale. Interestingly, the BULH740101 scale [85] for transfer free energy to surface (in aqueous conditions) shows a strong inverse correlation with both the BTMC and BTMI scales.

3.4.2 Correlation of propensity scales with hydrophobicity-based scales

Scale	BTMC scale for core region	BTMI scale for interface regions	HTMC scale for core region	HTMI scale for interface regions
GRAR740102	-0.76	-0.78	-0.75	-0.50
HOPT810101	-0.63	-0.73	-0.58	-0.57
ROSM880101	-0.82	-0.76	-0.75	-0.59
ROSM880102	-0.81	-0.71	-0.76	-0.59
PONP930101	0.65	0.66	0.72	0.49
CIDH920105	0.73	0.82	0.75	0.50
JOND750101	0.71	0.86	0.73	0.46
EISD840101	0.73	0.66	0.68	0.56
BLAS910101	0.81	0.83	0.80	0.54
MANP780101	0.71	0.70	0.76	0.46
FAUJ830101	0.78	0.82	0.74	0.58

Table 6: B. Correlation with (Hydrophobicity based) physico-chemical scales.

The correlation of the derived scales with the known hydrophobicity-based scales in shown in table 6. Both BTMC and BTMI scales show a strong inverse correlation with scales GRAR740102 [86] and HOPT810101 [87], which are associated with the polarity and the hydropilicity values of the amino acids, respectively. The hydropathy scales ROSM880101 and ROSM880102 [88] are derived based on the transfer of solutes from water to alkane solutions and account for the reduction in hydrophilicity of polar amino acid side-chains by the flanking peptide bonds. In general, the BTMC and BTMI propensity scales derived here, show a strong correlation with the hydrophobicity scales represented by PONP930101 [89], CIDH920105 [90], JOND750101 [91], EISD840101 [92], BLAS910101 [93], MANP780101 [94] and FAUJ830101 [95]. Interesting the HTMC and BTMC scales show comparable correlation with these hydrophobicity scales. This is in accord with the higher hydrophobicity of residues exposed to the lipid membrane.

Scale	BTMC scale for core region	BTMI scale for interface regions	HTMC scale for core region	HTMI scale for interface regions
VINM940101	-0.62	-0.71	-0.57	-0.44
BULH740102	0.74	0.69	0.83	0.27
ZIMJ680102	0.64	0.74	0.70	0.23
JANJ790101	0.72	0.56	0.75	0.59
FAUJ880108	-0.66	-0.51	-0.72	-0.56

Table 7: C. Correlation with (Structure based) physico-chemical scales.

3.4.3 Correlation of propensity scales with structure-based scales

The correlational analysis of the derived scales with known structural scales was carried out to determine which structural properties correlate well with the derived propensity scales. Table 7 shows the scales that have a strong correlation with the derived scales. As shown, the partial specific volume, average accessible area and localized electrical effect, represented by BULH740102 [85], JANJ790101 [96] and FAUJ880108 [97] scales, respectively show a strong correlation with the derived scales. The "Bulkiness" scale represented by ZIMJ680102 [98] shows a high correlation of 0.74 with the BTMI scale, which could be due to the presence of bulkier aromatic residues at the lipid-water interface regions.

3.5 Exposure status propensity scales for the oligomeric and non-oligomeric data sets

3.5.1 Analysis of BTMC$_{mono}$ and BTMC$_{oligo}$ scales

As discussed in section 3.3.1, the BTMC scale captures the propensity of amino acid residues in the membrane core to be exposed to the lipid bilayer. BTMC scale is derived from a data set of 20 non-redundant TMB 3D structures as described in section 3.10.4. Since it known that many TMBs occur as oligomers [99], it can be argued that the propensities of residues to be exposed to the bilayer might be different based on the oligomeric state. Further, various morphological, physico-chemical and evolutionary properties have been identified and employed to predict the oligomeric status of the HMPs [100–103]. However, such properties for TMBs have not yet been identified. Hence, BTMC$_{mono}$ and BTMC$_{oligo}$ scales were derived based on the two segregated data sets of monomeric and oligomeric TMBs. The data set for monomeric TMBs consists of 15 proteins with a total of 1253 residues in the membrane core region. The corresponding number for residues in the 8 non-redundant oligomeric TMBs is 680 (refer to section 3.10.1). In both the cases and rSASA cutoff value of 0.03 was employed to equi-partition the data set in buried and exposed labels.

Table 8 shows the derived BTMC$_{mono}$ and BTMC$_{oligo}$ propensity scales for the monomeric and oligomeric TMBs, respectively. As in the case of the original BTMC scale, the BTMC$_{mono}$ and BTMC$_{oligo}$ scales show that more hydrophilic residues have a lower propensity to be exposed to the bilayer, while hydrophobic residues have a higher propensity to be exposed to the lipid bilayer. The correlation coefficient between the BTMC$_{mono}$ and BTMC$_{oligo}$ propensity scales is 0.86. However a comparison of the two scales (BTMC$_{mono}$ and BTMC$_{oligo}$), separately derived for the monomeric and oligomeric data sets, respectively reveals

Scale	BTMC$_{mono}$ scale monomeric TMBs	BTMC$_{oligo}$ scale oligomeric TMBs
PHE	-0.025	0.123
ILE	0.070	0.516
TRP	0.088	-0.158
LEU	0.449	0.444
VAL	0.127	0.075
MET	-0.356	-0.393
TYR	-0.196	-0.201
ALA	0.047	-0.137
THR	-0.428	-0.283
HIS	-0.306	-0.468
GLY	-0.381	-0.125
SER	-0.584	-0.411
GLN	-0.551	-0.552
ARG	-0.490	-0.549
LYS	-0.465	-0.533
ASN	-0.478	-0.440
GLU	-0.481	-0.581
PRO	0.229	0.068
ASP	-0.440	-0.529

Table 8: Propensity scales for exclusively monomeric and oligomeric data sets.

differences in the propensities of PRO, ALA, TRP and PHE residues. Interestingly, in the case of monomeric TMBs, PHE residues with a small propensity scale value of -0.025 shows a relatively low tendency to be exposed to the membrane. In contrast to this, the PHE residues in the oligomeric TMBs have a higher propensity of being exposed to the lipid bilayer membrane with a propensity scale value of 0.123. The situation is reversed in the case of TRP residues, with the TRP residues in monomeric TMBs having a high propensity (0.088), while the TRP residues in the oligomeric TMBs showing a low propensity (-0.158) to be exposed to the lipid bilayer. Furthermore, ALA residues in the oligomeric TMBs have a low propensity value to be exposed to the bilayer (table 8). Moreover, the PRO residues in the monomeric TMBs show a higher propensity to be exposed to the bilayer than the PRO residues in the oligomeric data set. As discussed in section 3.3.1 for the non-redundant TMB data set, PRO residues are also found to have a high propensity to be exposed to the bilayer in both the monomeric and oligomeric TMB data sets.

3.6 Correlation of BTMC$_{mono}$ and BTMC$_{oligo}$ scales with physico-chemical scales from the literature

3.6.1 Correlation of BTMC$_{mono}$ and BTMC$_{oligo}$ scales with energy-based scales

Table 9 shows the correlation of the derived BTMC$_{mono}$ and BTMC$_{oligo}$ scales with known energy-based scales. Only scales that show a strong correlation are discussed here. As shown, both scales have a marginally strong inverse correlation with MIYS990101, MIYS990102, MIYS990103, and MIYS990105 scales. Briefly, these scales are related to the pair-wise contact energies of

Scale	$BTMC_{mono}$ scale monomeric TMBs	$BTMC_{oligo}$ scale oligomeric TMBs
RADA880101	0.75	0.81
RADA880104	0.69	0.74
JANJ790102	0.62	0.73
VHEG790101	-0.61	-0.72
GUYH850105	-0.60	-0.73
EISD860101	0.71	0.69
PLIV810101	0.77	0.76
BULH740101	-0.72	-0.76
MIYS850101	0.65	0.70
MIYS990102	-0.71	-0.77
MIYS990103	-0.64	-0.70
MIYS990105	-0.68	-0.73

Table 9: A. Correlation with (Energy based) physico-chemical scales.

amino acid residues [79]. In all cases, $BTMC_{oligo}$ seems to have a stronger negative correlation. As in the case with the BTMC and BTMI scales, the BULH740101 scale [85] for transfer free energy to surface (in aqueous conditions) shows a strong inverse correlation with the $BTMC_{mono}$ and $BTMC_{oligo}$ scales. PLIV810101 (partition coefficient) [82] and the EISD860101 scale for the solvation free energy [81] also show a strong positive correlation. Interestingly, JANJ790102 scale for the transfer free energy [96] shows a stronger correlation with the $BTMC_{oligo}$ scale (0.73) than the $BTMC_{mono}$ (0.62). Further, scales VHEG790101 and GUYH850105 for the transfer free energy to lipohilic phase and the apparent partition energies, respectively, show a strong inverse correlation only with the $BTMC_{oligo}$ scale. Both $BTMC_{mono}$ and $BTMC_{oligo}$ scales show a strong correlation with RADA880101 and RADA880104 scales, which represent the transfer free energy from the chx to water and octanol, respectively [80].

3.6.2 Correlation of $BTMC_{mono}$ and $BTMC_{oligo}$ scales with hydrophobicity-based scales

The correlation of hydrophobicity-based scales with the $BTMC_{mono}$ and $BTMC_{oligo}$ scales is shown in table 10. Only scales that show a strong correlation are discussed here. As shown in table 10, scales related to hydrophobicity, such as CIDH920105 [90], BLAS910101 [93], MANP780101 [94] and FAUJ830101 [95] scales show an equally strong correlation with both $BTMC_{mono}$ and $BTMC_{oligo}$ scales. Scales related to hydrophilicity, such as GRAR740102 [86], ROSM880101 and ROSM880102 [88] expectedly show strong negative correlation with the $BTMC_{mono}$ and $BTMC_{oligo}$ scales. Interestingly, PONP930101, PONP930102 and PONP930103 [89], EISD840101 [92] and JOND750101 [91] scales show a stronger correlation with the $BTMC_{oligo}$ scale than the $BTMC_{mono}$ scale. ENGD860101 scale for hydrophobicity [104] shows a stronger negative correlation with the $BTMC_{oligo}$ scale than the $BTMC_{mono}$ scale.

Scale	BTMC$_{mono}$ scale monomeric TMBs	BTMC$_{oligo}$ scale oligomeric TMBs
GRAR740102	-0.74	-0.80
ROSM880101	-0.81	-0.82
ROSM880102	-0.79	-0.84
ENGD860101	-0.61	-0.72
JOND750101	0.74	0.64
EISD840101	0.71	0.80
PONP930101	0.64	0.73
PONP800102	0.63	0.70
PONP800103	0.61	0.68
CIDH920105	0.74	0.71
BLAS910101	0.79	0.84
MANP780101	0.69	0.75
FAUJ830101	0.77	0.79

Table 10: B. Correlation with (Hydrophobicity based) physico-chemical scales.

Scale	BTMC$_{mono}$ scale monomeric TMBs	BTMC$_{oligo}$ scale oligomeric TMBs
DESM900102	0.56	0.69
JANJ790101	0.68	0.82
JANJ780101	-0.59	-0.71
JANJ780103	-0.58	-0.69
JANJ780102	0.66	0.78
BULH740102	0.73	0.68

Table 11: C. Correlation with (Structure based) physico-chemical scales.

3.6.3 Correlation of BTMC$_{mono}$ and BTMC$_{oligo}$ scales with structure-based scales

The most interesting differences in the correlational analysis of BTMC$_{mono}$ and BTMC$_{oligo}$ scales are observed with known structure-based scales. As shown in table 11, JANJ790101 (ratio of buried and accessible molar fractions) [96] and JANJ780102 (percentage of buried residues) [105] are more strongly correlated to the BTMC$_{oligo}$ than the BTMC$_{mono}$ scale, while JANJ780101 (average accessible surface area) [105] and JANJ780103 (percentage of exposed residues) [105] are more strongly inversely correlated to the BTMC$_{oligo}$ scale. On the other hand, the BULH740102 scale for the partial specific volume [85] shows a stronger correlation with the BTMC$_{mono}$ scale.

3.7 Prediction of the exposure status of TMB residues based on the derived propensity scales

Prediction accuracy per protein: As an application of the derived scales, we now present predictions about the exposure status of TMB residues that were made by a computational method based on the TMB scales namely, BTMC and BTMI derived here. The training data set was labelled as described in section 3.10.2. For a given labelled training data set, corresponding TMB scale values were used as coefficients for the respective ridge regression model to obtain a

	Observed			
	core region		interface regions	
Predicted	Buried	Exposed	Buried	Exposed
Buried	676 (78.97%)	201 (23.13%)	238 (82.93%)	63 (22.11%)
Exposed	180 (21.03%)	668 (76.87%)	49 (17.07%)	222 (77.89%)

Table 12: Prediction accuracy for the residues in the membrane core and interface regions is 77.91% and 80.42%, respectively.

Protein	residue count	core region ACC (%)	residue count	interface regions ACC (%)
1thq_A	53	73.58	17	58.82
1a0s_P	103	68.93	37	81.08
1e54_A	88	17.05[a]	28	32.14
1p4t_A	45	73.33	17	100.0
1qj8_A	45	73.33	17	94.12
2f1v_A	51	86.27	18	61.11
1qd6_C	68	85.29	24	79.17
2erv_A	49	91.84	21	66.67
1xkw_A	121	88.43	47	91.49
1qjp_A	52	78.85	20	100.0
1xkh_A	108	84.26	18	88.89
1fep_A	134	84.33	44	93.18
2mpr_A	112	77.68	29	86.21
1t16_A	83	87.95	29	82.76
1i78_A	66	75.76	21	85.71
2j1n_A	101	71.29	32	65.62
1qfg_A	129	83.72	42	76.19
1kmo_A	122	86.06	43	83.72
1tly_A	76	72.37	27	77.78
2gsk_A	119	87.39	41	90.24
total	-	77.91	-	80.42

Table 13: Protein-wise prediction accuracy for the residues in the membrane core and interface regions.[a]discussed below in main text.

positional score for each residue. A support vector classifier (SVC) [50] with default parameters [106] implemented in R [107] was then employed to make exposure status predictions based on these positional scores. A leave-one-out test was conducted to measure the performance of the preliminary prediction method described here. As shown in table 12, the prediction accuracy of the method to distinguish buried and exposed residues is 77.91% and 80.42% for the core and interfaces regions, respectively. Table 13 shows per-protein prediction accuracy. As shown, 1e54_A has the lowest prediction accuracy of 17.05% and 32.14% in the both the core and the interface regions, respectively. Analysis of residues in the case of 1e54_A revealed that in contrast to other proteins in the data set, many exposed residues had a low hydrophobicity which could be due to the large spacing between the beta strands of 1e54_A, which could be attributed to the lesser number of strands (16) in a single 1e54_A monomer chain. Thus rendering the prediction of residues in 1e54_A inaccurate. In the case of interface regions, proteins 1thq_A, 2f1v_A, 2erv_A and 2j1n_A have a low

prediction accuracy of 58.82%, 61.11%, 66.67% and 65.62%, respectively.

Amino Acid	residue count	core region ACC (%)	residue count	interface regions ACC (%)
TRP	30	63.33	44	86.36
PHE	80	82.50	44	70.45
TYR	122	54.10	73	83.56
MET	39	74.36	12	66.67
LEU	209	80.38	41	78.05
ILE	80	81.25	20	80.00
VAL	159	77.99	40	87.50
ALA	183	78.14	35	65.71
GLY	231	64.50	52	71.15
PRO	23	95.65	7	85.71
THR	118	77.97	42	90.48
SER	112	84.82	36	97.22
ASN	70	87.14	14	85.71
GLN	47	89.36	18	50.00
ASP	46	89.13	22	77.27
GLU	64	98.44	20	90.00
HIS	14	64.29	13	69.23
ARG	63	95.24	24	87.50
LYS	35	85.71	15	93.33

Table 14: per amino acid prediction accuracy for the residues in the membrane core and interface regions.

Prediction accuracy per amino acid: Table 14 shows the per amino acid prediction accuracy for residues in the core and the interface regions. While most amino acids can be predicted with a reasonably high accuracy, the prediction accuracy for TRP (63.33%), TYR (54.10%), GLY (64.50%) and HIS (64.29%) residues in the core region is relatively low. This could be due to the exceptionally high conservation indices of the TRP, TYR and GLY residues labelled as exposed. As shown in table 1, the average conservation indices for TRP, TYR and GLY are 0.24, 0.76 and 0.74, respectively. In the interface regions, the prediction accuracy of MET, ALA and GLN residues is 66.67%, 65.71% and 50.0%, respectively, which is significantly lower than the overall prediction accuracy of 80.42%. In the case of MET and ALA, this could be due to the low conservation indices of the residues labelled as buried in the training data set. Residues exposed to the bilayer environment are generally known to be less conserved [76]. As described in the section 3.10.4, the scales are derived such that positional scores derived from the given profiles are maximally correlated with the rSASA (relative solvent accessible surface area) values and a discrepancy between the rSASA and conservation could be reason for the low prediction accuracy for the residues mentioned above. It is to be noted that the scale based prediction method described here has been implemented to show the practical utility of the derived propensity scales. We expect that an even higher accuracy can be achieved by using a two-stage sliding window method as has been employed in the study on HMPs previously described [62].

z	Residue count	Correct predictions	Prediction accuracy [%]
0.0 - 0.1	130	101	77.69
0.1 - 0.2	147	115	78.23
0.2 - 0.3	121	86	71.07
0.3 - 0.4	146	119	81.51
0.4 - 0.5	129	102	79.07
0.5 - 0.6	136	103	75.74
0.6 - 0.7	57	36	63.16

Table 15: Prediction accuracy w.r.t. relative z coordinate for residues in the membrane core region.

z	Residue count	Correct predictions	Prediction accuracy [%]
0.6 - 0.7	58	46	79.31
0.7 - 0.8	114	86	75.44
0.8 - 0.9	77	63	81.82
0.9 - 1.0	23	17	73.91

Table 16: Prediction accuracy w.r.t. relative z coordinate for residues in the membrane interface regions.

Prediction accuracy w.r.t. distance from membrane center: Tables 15 and Tab:zscale-inter show the prediction accuracy w.r.t. the membrane center. As shown in table 15, the prediction accuracy at the edge of the membrane core region is 63.16%, which is the lowest when compared to the regions that completely lie in the membrane core region. As discussed above in section 3.3.4, the BTMX and BTMI scales have a correlation coefficient of 0.80. Thus this low accuracy at the membrane-core/membrane-interface region boundary could be due to the change in the physico-chemical environment [60]. In the case of the interface regions, the prediction accuracy at the lipid-water interface regions is 73.01%, which is lower than the overall average of 80.42%. This lower accuracy at the interfaces could be attributed to the different physico-chemical properties at these interfaces.

3.8 Classification of TMB residues to be in the beta-strand/non beta-strand regions

The three input parameters tested for determining optimal prediction accuracy are conservation index (CI), Frequency profile (Freq) and the physico-chemical scales that showed a statistically significant (p-value of ≤ 0.01) difference between the average values for the residues in the different regions (refer to table 17). As described in section 2.4.2, a principal component analysis (PCA) was carried out on the identified scales for the beta-strand/non beta-strand regions. Figure 13 shows the general scheme of residue classification. As shown, to avoid over-representation of the class with more number of data points, the major class is divided into smaller sub-classes such that the resulting prediction models have equal number of data points from each class. Similar consensus-based approaches have previously been reported in the literature for the prediction

of the topology of TMBs [47] and protein classification [108]. Table 18 shows the average prediction accuracy results over the n models along with the standard deviation based on 10 runs of the prediction scheme for the classification of residues as belonging to beta-strand and non beta-strand regions, respectively. In the case of classifying a residue as belonging to a beta-strand/non beta-strand, the highest prediction accuracy (77.32±0.66) was achieved when conservation index and frequency profile were employed as input to a kwKNN with the Minkowski distance parameter $q = 1$. Similar aproach was applied to the classification of residues into core/interface regions, the highest prediction accuracy (69.99±0.59) was obtained when the kNN method with $k = 3$ was employed and frequency along with the identified phyico-chemical scales were used (results not shown).

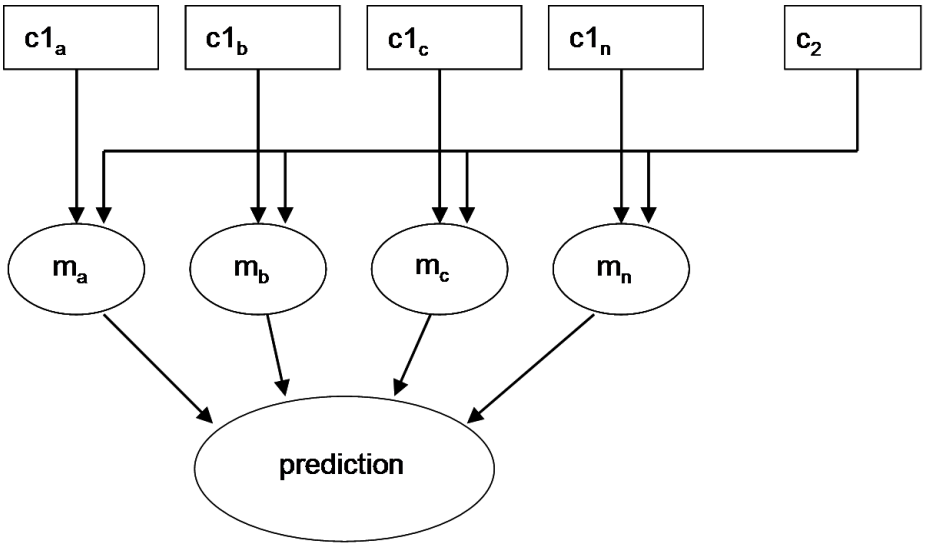


Figure 13: The prediction scheme avoids over-representation of the majority class c_i by undersampling. The prediction accuracy of individual prediction unit m_i is averaged and reported in the main text.

beta-strand/non beta-strand discrimination			
Scales	Avg. value for residues in beta-strands	Avg. value for residues in non beta-strand regions	p-value
BLAS910101	5.98e-01	5.50e-01	4.49e-04
DESM900102	5.78e-01	5.25e-01	2.54e-05
EISD840101	7.14e-01	6.74e-01	2.05e-04
EISD860101	6.09e-01	5.70e-01	4.25e-04
ENGD860101	2.46e-01	2.97e-01	4.31e-05
FASG760102	6.24e-01	5.65e-01	3.86e-09
FAUJ830101	4.83e-01	4.49e-01	5.44e-03
FAUJ880104	4.20e-01	4.51e-01	4.05e-03
FAUJ880105	5.79e-01	6.08e-01	6.40e-03
FAUJ880111	7.14e-02	1.04e-01	7.71e-03
FAUJ880113	7.64e-01	7.39e-01	2.07e-03
GRAR740102	3.68e-01	4.12e-01	7.74e-04
GUYH850105	2.95e-01	3.25e-01	3.95e-03
HOPT810101	4.46e-01	5.04e-01	5.32e-07
JANJ780101	2.91e-01	3.30e-01	2.63e-04
JANJ780102	5.47e-01	5.12e-01	2.46e-03
JANJ780103	2.85e-01	3.29e-01	3.26e-05
JANJ790102	6.73e-01	6.41e-01	1.35e-03
KRIW790101	4.73e-01	5.14e-01	9.21e-05
MIYS990103	4.24e-01	4.69e-01	6.06e-04
MIYS990104	4.07e-01	4.56e-01	2.21e-04
MIYS990105	3.97e-01	4.46e-01	8.34e-05
MUNV940103	3.97e-01	4.39e-01	2.62e-04
MUNV940105	1.56e-01	1.97e-01	5.98e-06
NOZY710101	2.82e-01	2.30e-01	1.08e-05
PONP800103	4.79e-01	4.52e-01	9.46e-03
PONP800104	4.93e-01	4.44e-01	1.20e-04
PONP930101	4.56e-01	4.19e-01	3.76e-03
PRAM900101	2.45e-01	2.96e-01	3.64e-05
RADA880101	7.32e-01	6.90e-01	2.16e-04
RADA880102	5.39e-01	4.96e-01	7.30e-05
ROSM880101	2.89e-01	3.30e-01	1.36e-03
VHEG790101	2.71e-01	3.20e-01	5.99e-06
VINM940101	4.10e-01	4.56e-01	1.34e-04
VINM940103	5.24e-01	5.71e-01	2.13e-05
ZIMJ680103	1.48e-01	2.09e-01	7.72e-05

Table 17: The scales that should a statistically significant difference between residues present at the beta-strand and at the non beta-strand regions. All the TM residues irrespective of their DSSP annotation were considered.

		Input parameters						
Model	Variable	Freq	CI	Scales	CI+Scales	Freq+Scales	CI+Freq	All3
SVC	linear	61.18±0.54	57.3±0.79	55.47±0.81	58.89±1.02	62.68±0.57	62.57±0.39	64.77±0.47
SVC	radial	66.05±0.38	58.12±0.74	56.9±0.49	59.99±0.92	69.37±0.68	68.59±0.70	72.01±0.71
kwKNN	q = 1	71.41±1.14	68.85±1.56	54.41±1.35	69.01±1.87	71.63±1.77	**75.30±2.72**	72.71±2.38
kwKNN	q = 2	70.93±1.45	68.85±1.56	54.51±1.41	68.94±1.82	73.02±1.73	73.29±1.27	74.06±0.89
kNN	k = 3	69.63±0.65	64.52±0.95	56.66±0.62	67.62±0.66	70.05±0.93	72.03±0.46	69.99±0.62
kNN	k = 5	67.46±0.60	63.01±0.76	56.66±0.59	64.54±0.62	68.02±0.94	69.54±0.81	66.71±0.54
kNN	k = 7	66.99±0.60	61.29±0.77	56.60±0.56	63.82±0.95	67.23±0.73	68.43±0.72	64.99±0.82
kNN	k = 9	66.37±0.52	60.33±0.93	56.66±0.49	62.80±0.49	65.81±1.08	67.58±0.75	64.16±0.79
kNN	k = 11	66.18±0.56	60.11±0.85	56.68±0.59	61.63±0.74	64.36±1.14	66.65±0.52	63.67±0.75
kNN	k = 13	65.94±0.53	60.05±0.82	56.57±0.64	61.25±0.81	63.91±1.08	66.39±0.49	63.50±0.83
kNN	k = 15	65.46±0.41	59.16±1.02	56.71±0.63	60.98±0.83	63.17±0.84	65.93±0.64	63.31±0.55

Table 18: Classification of residues in the transmembrane region to be in a beta-strand or a non beta-strand region. The table shows the average prediction accuracy and the standard deviation obtained after 10 separate cross-validation runs. "Scales" refers to the first 6 principal components obtained after PCA.

3.9 Conclusions

Three novel propensity scales that capture the propensity of transmembrane residues to be exposed to the bilayer have been derived and compared with know scales obtained from the AAIndex database [71]. Strong correlation between the BTMC and HTMC scales reveals that similar hydrophobicity constraints have shaped the sequence diversity of the surfaces and interiors of HMPs and TMBs. Interesting differences, however, were revealed between monomeric and oligomeric TMBs. The scale derived for the interface regions of HMPs is least correlated with the scales derived for the other three regions. We speculate that this may be either due to currently unknown protein-protein interactions or due to constraints set by manoeuvring out of the translocon machinery [6]. More experimental data is needed to be able to make definite conclusions along such lines. We show that based on the propensity scales and CI, the exposure status of the TM residues can be predicted with an accuracy of 77.91% and 80.42% for the residues in the membrane core and interface regions, respectively. We also present a prediction scheme that can classify a residue into beta-strand/non beta-strand regions. The average prediction accuracy of the prediction scheme based on a leave one out test for the beta-strand/non beta-strand regions is 75.30% $\pm$2.72. Knowledge of the exposure status of TM residues can be employed to study the interaction between the exposed residues and the lipids of the membrane that have been suggested to act as active folding catalysts [109], in genome wide TMB identification, channel engineering [38], drug design and in mutational studies [110]. Further, the method of deriving statistical propensities as implemented here and correlating them with established scales that capture physico-chemical properties of amino acids appears as a promising path to discovering unknown facts about TM proteins, their environment and interactions. As compared to state of the art prediction methods for HMPs, our position based prediction method for the exposure status does not require labelling the secondary structure elements in the test data. The propensity scales derived here for HMPs and TMBs will likely have biological implications such as allowing the prediction of the oligomeric states and solubility of TM proteins in various membranes.

3.10 Methods

3.10.1 Training and test data sets

A non-redundant data set of known TMB structures was compiled primarily based on the following databases:

- `http://pdbtm.enzim.hu/` [111]
- `http://blanco.biomol.uci.edu/Membrane_Proteins_xtal.html`
- `http://bioinfo.si.hirosaki-u.ac.jp/~TMPDB/`
- `http://www.mpibp-frankfurt.mpg.de/michel/public/memprotstruct`
- `http://opm.phar.umich.edu/` [20]

From this data set, we removed those protein sequences for which less than 20 homologous sequences were found or where the average pair-wise sequence

identity of the aligned retrieved sequences was greater than 80%. The final data set for TMBs comprises of 20 protein chains (table 19) with 1725 and 572 TM residues in the hydrophobic core and interface regions, respectively (see table 1). It is to be noted that only residues that belonged to a beta strand were taken into account. The hydrophobic region, defined as the region with zero probability for the occurrence of the hydration waters of the lipid head-groups, was derived from the OPM database [20]. Only the residues within the range from +0.65 to -0.65 units of their respective hydrophobic thickness as defined in OPM were considered as residing in the hydrophobic core region. This range was 0.65 to $zThickness$ and -0.65 to $-zThickness$ for the residues from the interface regions at the periplasmic and the extra-cellular end, respectively. When splitting the data set into segregated data sets based on the number of strands, the cutoff for the number of strands was chosen as 14 to have a statistically significant number of data points.

The data set for HMP was taken from a recent study and consisted of 1248 residues residing at the interface region [70]. For the derivation of the HTMI scale for the interface of the HMP proteins, residues belonging to irregular structures such as re-entrant loops were not distinguished from the regular secondary structures and hence were included in the training data sets. Also included in the interface region of HMP proteins were interfacial helices, the parts of the transmembrane helices located in the interface region, and irregular structures that mainly comprise of GLY, PRO, ASN and SER residues.

PDB ID	Protein	Functional state	β-strands
1thq_A	Outer membrane Lipid A acylase PagP	Monomer	8
1a0s_P	Sucrose specific porin ScrY	Trimer	18
1e54_A	Anion-selective porin	Trimer	16
1p4t_A	Outer membrane protein NspA	Monomer	8
1qj8_A	Outer membrane protein OmpX	Monomer	8
2f1v_A	Outer membrane protein OmpW	Monomer	8
1qd6_C	Outer membrane phospholipase OmpLA	Dimer	12
2erv_A	Outer membrane enzyme PagL	Monomer	8
1xkw_A	Fe(III)-pyochelin receptor	Monomer	22
1qjp_A	Outer membrane protein OmpA	Monomer	8
1xkh_A	Pyoverdine outer membrane receptor FpvA	Monomer	22
1fep_A	Ferric enterobactin receptor FepA	Monomer	22
2mpr_A	Maltoporin	Trimer	18
1tl6_A	Long-chain fatty acid transporter FadL	Monomer	14
1i78_A	Outer membrane protease OmpT	Monomer	10
2j1n_A	Outer membrane Osmoporin OmpC	Trimer	16
1qfg_A	Ferric hydroxamate uptake receptor FhuA	Monomer	22
1kmo_A	Outer membrane transporter FecA	Monomer	22
1tly_A	Nucleoside-specific channel-forming protein Tsx	Monomer	12
2gsk_A	Outer membrane cobalamin transporter BtuB, complex with TonB	Monomer	22

Table 19: Training and cross validation data set consists of 20 non-redundant TMBs that have a sequence identity of $\leq 30\%$.

PDB ID	Protein	Functional state	β-strands
1a0s_P	Sucrose specific porin ScrY	Trimer	18
1e54_A	Anion-selective porin	Trimer	16
1qd6_C	Outer membrane phospholipase OmpLA	Dimer	12
2j1n_A	Outer membrane Osmoporin OmpC	Trimer	16
2mpr_A	Maltoporin	Trimer	18
2o4v_A	Porin OprP	Trimer	16
3prn_A	Porin	Trimer	16
2por_A	Porin	Trimer	16

Table 20: Oligomeric TMB data set consists of 8 non-redundant TMBs that have a sequence identity of $\leq 25\%$.

3.10.2 Calculation of observed input and output parameters from the data set

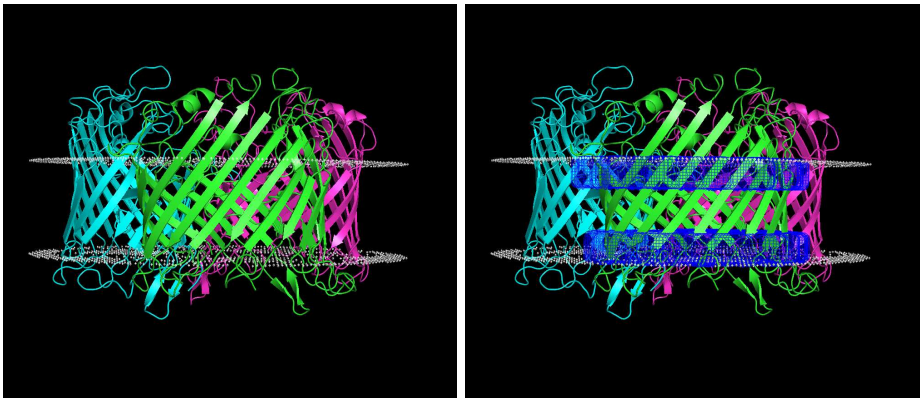
Positional frequency profiles and conservation indices were obtained using the AL2CO program suite [112] as described before [70]. The training data set employed for predicting the burial status was labelled based on the SASA value of the residues in the protein crystal structure. The SASA value was calculated using the VOLBL program suite [113, 114] and a probe radius of 2.2 $\mathring{A}$, which has been suggested to be an appropriate size for the effective radius of the CH2 group of hydrocarbon chains [70]. For correct assessment of the SASA value of the residues present at oligomeric interfaces, only the functional oligomeric forms of the protein were considered [20]. SASA values were normalized to generate relative SASA (rSASA) values by dividing them by the SASA values for each amino acid X in the context of the tri-peptide G-X-G. The tri-peptides employed for normalizing TMBs and HMPs had a flat beta sheet type and a perfect alpha helical backbone conformation, respectively.

3.10.3 Capping of the TMB structures to determine the rSASA value

TMBs, as the name suggests, contain a huge internal cavity in the shape of a barrel. To prevent any internal residues from being labelled as exposed, the proteins were capped at both the periplasmic and exo-cytoplasmic barrel ends by adding a layer of dummy atoms at both the ends. Upon capping, the internal cavity is made inaccessible to the probe and the residues are hence not falsely assigned a high SASA value. The capping at the core region is depicted in figure 14. The top view of the protein capped at the core region is shown in figure 15. The residues labelled as buried and exposed residues are colored in yellow and red, respectively are shown in figure 16. An example of a protein capped at the interface regions is shown in figure 17.

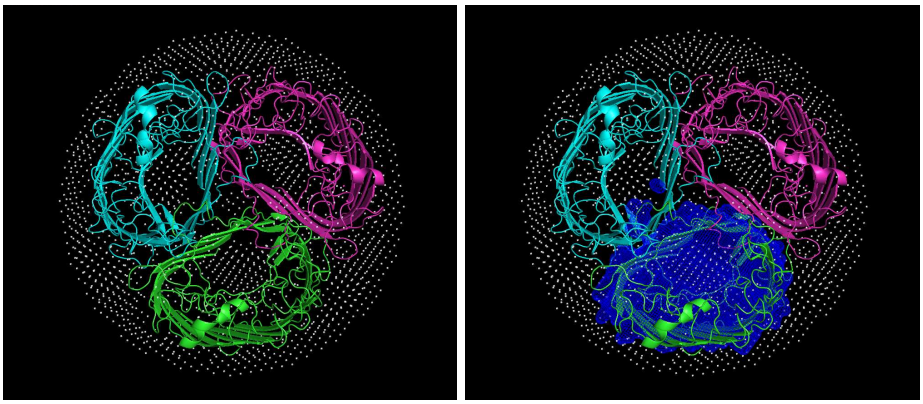
3.10.4 Computation method for the determination of the propensity scales

It has been shown previously that polar residues tend to be buried inside and hence are less exposed to the bilayer in the hydrophobic core region of OM [9]. In this study, we employed the method previously established by us for HMP



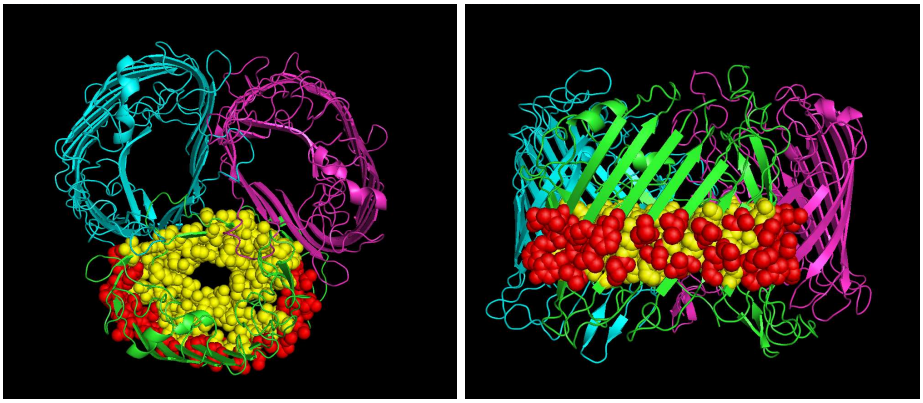
(a) uncapped TMB (front view) (b) capped TMB (front view)

Figure 14: Front view of the capping of a TMB at the core region (PDB-ID 1a0s). The three beta barrels are colored according to the chain ID. The while dotted line represents the membrane boundary as described by [20]. Only chain P is capped.



(a) uncapped TMB (top view) (b) capped TMB (top view)

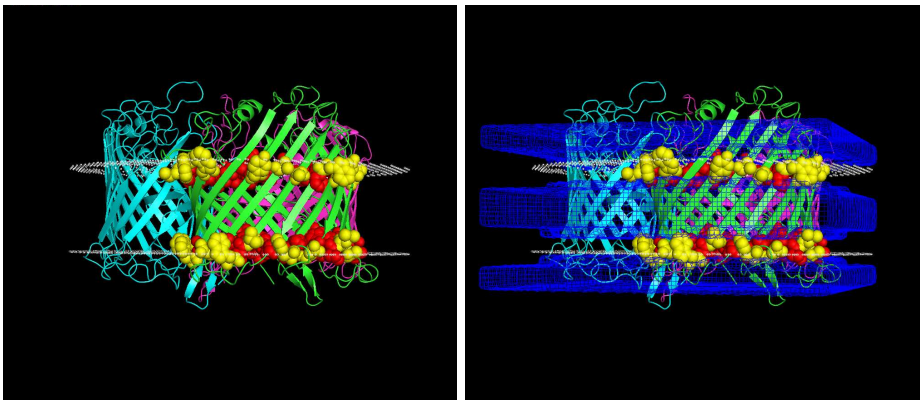
Figure 15: Top view of the capping of a TMB at the core region (PDB-ID 1a0s). The three beta barrels are colored according to the chain ID. The while dotted line represents the membrane boundary as described by [20]. Only chain P is capped.



(a) Top view of the labelled exposed/buried residues (core region)

(b) Front view of the labelled exposed/buried residues (core region)

Figure 16: Top and front view of the residues labelled as buried/exposed in the membrane core region (PDB-ID 1a0s). The three beta barrels are colored according to the chain ID. The while dotted line represents the membrane boundary as described by [20]. Only chain P is capped. Buried and exposed residues are colored in yellow and red, respectively.



(a) Residues at the interface regions (front view)

(b) uncapped TMB (top view)

Figure 17: Front view of the capping of a TMB at the interface regions (PDB-ID 1a0s). The three beta barrels are colored according to the chain ID. The while dotted line represents the membrane boundary as described by [20]. Only chain P is capped. Buried and exposed residues are colored in yellow and red, respectively.

residues [70] to generate propensity scales for TMB residues to be exposed to the lipid bilayer at the hydrophobic core and interface region of OM, respectively. Frequency profiles of the 20 amino acids generated from the multiple sequence alignment of a given protein sequence by the program ClustalW [115] were employed as the input, while the relative solvent accessible surface area (rSASA) value determined using the VOLBL program suite [113, 114] acted as the dependent variable. Briefly, our method tries to maximize the correlation coefficient between the observed exposure patterns and the positional scores derived from a given frequency profile.

The propensity scale should capture the affinities of the 20 naturally occurring amino acids to preferentially interact with the lipid bilayer as reflected in experimental TMB structures. To this end, the scale was derived such that the positional scores derived from the given profiles are maximally correlated with the observed rSASA values. A straightforward solution is provided by minimizing the residual sum of squares error (RSS):

$$RSS(\beta) = \sum (y_i - f(x_i))^2 \tag{66}$$

This can be re-written as:

$$RSS(\beta) = (y - X\beta)^T (y - X\beta) \tag{67}$$

Differentiating with respect to β and rearranging the equation gives us the solution:

$$\beta = (X^T X)^{-1} X^T y \tag{68}$$

where X is an n by 21 matrix comprising of frequency profiles and an intercept value, y is a column vector of size n and β is a column vector of size 21, representing the derived propensities for the 20 amino acids and an intercept value, respectively. Ridge regression, which has widely been used as an alternative to the least squares estimate for ill-conditioned matrices was employed with a complexity parameter $\lambda = 0.00001$ for both the core and interface regions, respectively. Ridge regression was performed to penalize the very large coefficient values obtained as:

$$RSS(\lambda) = (y - X\beta)^T (y - X\beta) + \lambda\beta\beta^T \tag{69}$$

$$\beta^{ridge} = (X^T X + \lambda I)^{-1} X^T y \tag{70}$$

A leave-one-out test with known 3D structures yielded the correlation coefficient between the observed rSASA and calculated positional score as 0.64 and 0.52 for the residues in the core and the interface regions of the OM, respectively.

3.10.5 Classification of residues to be in the beta-strand/non beta-strand regions

For the classification of residues as beta-strand/non beta-strand, all the residues within the transmembrane region were considered unlike the case when only residues in beta-strands were considered while deriving the propensity scales. Thus 2297 residues belonging to the beta-strands and 757 belonging to non beta-strand region were taken into consideration. It should be noted that the class

distribution is imbalanced and hence, as discussed in the literature, a straightforward classifier can not be applied as it will lead to an over-fitted model that will have poor performance for the minority class [108]. To overcome this problem, we divided the majority class C, into n equal parts, $C_1, C_2, \ldots, C_n$ where each part has atleast m data points (as shown in Figure 1), where m is the number of residues in the minority class. Thus the non beta-strand residues class is divided into 3 equal parts that consist of 757 residues each and $n = 3$ classifiers are trained and tested. Each classifier model m is trained on a subset of the majority class, and the whole of the minority class and the prediction accuracy of the model is determined based on a test data set that comprises of equal number of data points from the rest of the majority classes that were not employed while training, and the data points from the minority class. Thus, for testing, 500 residues were randomly chosen from the minority class and and equal number of residues was randomly chosen from the unused bins of the majority class.

Three statistical classification methods namely support vector classification (SVC) [116], k-nearest neighbors (kNN) classification [117] and weighted k-nearest neighbors (kwKNN) methods [118] were tested to determine optimal prediction accuracy. Physico-chemical scales that showed statistically significantly difference between the values averaged over the two classes were identified. A pricipal component analysis was carried out on these identified physico-chemical scales and the chosen principal components along with the frequency profile and conservation index obtained from a multiple sequence alignment were employed to classify the residues into beta-strand/non beta-strand regions, respectively. A total of 36 physico-chemical scales with a p-value of ≤ 0.01 were identified for the beta-strand/non beta-strand regions (refer to table 17). Principal component analysis (PCA) was performed to reduce the number of dimensions of the identified scales and the first 6 principal components were chosen for the beta-strand/non beta-strand classification.

The R implementation of kNN, kwKNN [119] and SVC classification methods was employed for the actual classification. The prediction accuracy of a given set of parameters was averaged over all the n classifiers and the scheme was run 10 times. In each run, the training and the test data sets were obtained after random distribution of data points in the majority and minority classes. The prediction accuracy results for each classifier, along with their respective standard deviation are shown in table 18.

4 Prediction of the exposure status of transmembrane beta barrel residues from protein sequence

4.1 Overview

Established methods for discrimination of TMBs from globular and HMP proteins, genome-wide identification and topology prediction of TMBs have been classified and evaluated in the literature [47,110] and it has been suggested that (Hidden Markov Method) HMM-based methods are the most accurate ones. Methods that combine Neural Networks (NNs) and Support Vector Machines (SVMs) such as TBBPred [120] have also been reported. Interstrand pairing patterns in TMBs have been discussed in a study by Liang and co-workers [39]. Recently, Naveed *et. al* reported a prediction method for the identification of weakly stable regions, oligomerization states and protein-protein interfaces in TMBs [121]. To the best of our knowledge, transFold [122] and TMBPro [64] are the only prediction methods that can provide the side-chain orientation of TMB residues in contact. transFold [122] employs multi-tape S-attribute grammars to describe all potential structural conformations and then uses dynamic programming to determine the global minimum of the secondary structure. The method performs well on small proteins, but it puts an upper limit on the length of input sequence. Moreover, the method considers an input sequence as comprising of one large domain and hence can not be used for multimeric protein sequences. These methods can successfully predict interstrand residue contacts, however, they do not provide any information about the exposure status of the TMB residues.

In this respect, Yuan *et al.* developed a method (hereafter called the YU method) for predicting the relative solvent accessible surface area (rSASA) of transmembrane (TM) proteins [17]. Based on this numeric value, one may predict the exposure status of the TM residues [17]. Here, we re-implemented the YU method and show that the prediction accuracy of BTMX for classification as buried or exposed residues is significantly higher than the YU method. We note that although alternating residues in beta strands are known to strictly follow an in/out pattern, this does not necessarily mean that outwards pointing residues are exposed to the lipids. We show that such residues may often be covered by neighboring side chains. Thus, prediction of the exposure status of TMB residues in addition with the topology models of TMBs can provide additional insights into the folding and insertion mechanism of TMBs [123].

We then discuss the use of the predicted exposure status in detecting the strands at the oligomeric interfaces of TMBs. Interestingly, Seshadri *et al.* have analysed the difference in the information content of the exposed residues at the oligomeric interface and have proposed the use of information content in detecting the oligomeric interface of TMBs [67]. Moreover, it is known that not one but a combination of many different morphological, physico-chemical and evolutionary properties play a role in the folding and oligomerisation of soluble proteins [100, 101]. Using multiple physico-chemical properties was also beneficial for the detection of protein-protein interaction interfaces in the realm of soluble proteins [103]. Such differences in the physico-chemical properties of the strands at the oligomeric interface of TMBs have not yet been reported.

We present here BTMX, a new method for predicting the exposure status of TMB residues. BTMX uses positional specific scoring matrices (PSSM) [124] as an input factor and is motivated by two recent computational studies conducted for TM proteins [17, 62]. Unlike some other state of the art prediction methods for TMBs [47], BTMX also provides a confidence score of the predictions made. Moreover, unlike other statistical methods such as the ones based on HMM, where the biological relevance of the mathematical model is not explicit, BTMX is derived so that the method parameters, such as the size of the sliding window and the input parameters employed have an apparent biological significance. We have also identified physico-chemical properties such as hydrophobicity, size, length and width of the side chain, screening coefficients γ_{local} and $\gamma_{non-local}$, localized electric effect and free energy that show statistically significant differences in the mean values of the oligomeric and non-oligomeric strands in TMBs. We propose that these physico-chemical properties of the residues predicted to be exposed to the bilayer can be employed to develop a prediction method for the identification of oligomeric interfaces of TMBs [67, 125].

4.2 Alternate in/out dyad repeat pattern of amino acid side chains

Amino Acid	Out-pointing + Exposed	Out-pointing +Buried	In-pointing +Exposed	In-pointing +Buried
ASP	5	20	4	53
SER	13	19	32	76
ASN	8	12	9	61
GLN	3	9	12	37
LYS	1	8	6	37
THR	48	21	12	77
PRO	23	19	2	8
HIS	5	3	0	11
PHE	70	6	7	15
ALA	117	31	16	61
ILE	61	25	2	27
LEU	180	35	6	41
ARG	0	14	9	87
TRP	22	5	4	8
VAL	133	31	4	45
GLU	0	12	11	65
TYR	41	9	17	55
MET	26	6	5	17
Total	756	285	158	781

Table 21: The in/out pattern of exposed/buried residues.

Hydrogen bonding of the anti-parallel beta strands present in the TMBs bestows a unique alternating in/out pattern to the side chains of the residues in beta strands. This alternating in/out pattern along with the residue exposure has previously been employed to identify TMBs in genomic data [66]. Table 21 shows the in/out pattern of the residues in the TMBs. It is to be noted that only the residues that belong to a beta strand were taken into account while

determining the in/out status. As expected, the orientation of $C_\alpha - C_\beta$ vectors shows that this alternating in/out pattern is strictly followed by all the residues in the beta strands of the training and cross-validation data set. However, an out-pointing side chain (i.e. C_β pointing away from the barrel axis) is not necessarily exposed to the lipid membrane (see table 21). As observed in the 3D structures in the data set, approximately 27% of the residues were found to be pointing out but were still shielded from the lipid membrane by the side chains of the adjoining residues. Similarly, roughly 17% in-pointing residues were found to be slightly exposed to the lipid membrane. Furthermore, since the rSASA of the residues was calculated in the functionally active state of the given protein, some out-pointing residues at the interface with other oligomeric chains were not exposed to the lipid membrane. This may be of significance with respect to the conservation and physico-chemical properties of those residues. Interestingly, some side chains pointing towards the barrel axis (in-pointing) were found to be slightly exposed to the membrane. In total, 1930 residues were found to be belonging to a beta strand. GLY residues were excluded from this analysis. Out of the 781 residues that were identified as pointing in, 158 were exposed to the membrane, while 285 from a total of 1041 out-pointing residues were classified as buried.

4.3 Determination of optimal input parameters for BTMX predictions

Input	IVS	C-value	ϵ-SVR Radial(WS1) ACC[%]	LR(WS1) ACC[%]
CI	1	7	60.7(15)	50.9(3)
FP	20	1	74.3(11)	72.0(9)
CI+FP	21	1	74.5(11)	62.2(3)
PSSM	20	1	83.3(15)	80.3(9)
PSSM+CI	21	1	82.8(11)	79.6(9)
PSSM+FP	40	5	80.9(7)	78.4(5)
PSSM+CI+FP	41	3	80.8(7)	77.9(5)

Table 22: Prediction accuracy based on different parameter sets. Only the highest prediction accuracies obtained with different input parameters, window sizes and C values are shown. Both linear regression (LR) and $\epsilon - SVR$ with a radial kernel were tested for generating the positional score in the first stage. The window size of the positional scores in the second stage was set to 1. All data points in a given window were taken into account while calculating the accuracy based on a jack-knife test. PSSM = positional specific scoring matrix, CI = conservation index, FP = frequency profile, WS1 = Window size of the input data in the first stage, IVS = Input vector size.

Table 22 shows the heuristically determined most accurate predictions obtained for a given set of parameters. In almost all cases, slightly lower prediction accuracies were attained when linear regression was used instead of ϵ-SVR with a radial kernel for obtaining the positional scores in the first stage (table 22). The highest prediction accuracy (83.3%) was obtained when PSSMs were employed as the input parameter and ϵ-SVR with a radial kernel was used in the first stage. Moreover, in the case of PSSMs, window sizes of 5 to 15 were also

found to have higher prediction accuracies (see supplementary). In the case of PSSMs, the use of ϵ-SVR with a linear kernel in the first stage was also tested but yielded lower prediction accuracies for all the given window sizes (see supplementary). Thus, based on the high prediction accuracy, a $C - value$ of 1, window size of 5 to 15 and PSSMs along with the use of ϵ-SVR were chosen for Fisher's analysis. Interestingly, the use of PSSMs in discriminating outer membrane proteins from other folding architectures has been discussed in a recent study [126].

The rationale behind testing conservation indices as input data was that for HMPs, the residues that are exposed to the environment are less conserved than the residues that are buried in the protein structure [58, 59]. Hence, the rSASA value should be inversely correlated to the conservation index for a given position. Interestingly however, as shown in table 22, for all the tested input parameters, conservation index alone had the lowest prediction accuracy at 60.7% and 50.9% for ϵ-SVR and linear regression, respectively. The average conservation index of residues labelled as exposed revealed that GLY, TYR, PRO, ARG, ASP, LYS and GLN residues are highly conserved even when exposed to the environment (see supplementary), thereby rendering the prediction of exposure status of TMB residues based on conservation index difficult. A higher prediction accuracy (74.3% and 72.0% for ϵ-SVR and linear regression, respectively) was obtained when a frequency profile based on the positional frequency obtained from the MSA of the given protein sequence was employed as the input factor. In the case when ϵ-SVR was employed, there was only a marginal increase in prediction accuracy (0.14%) when both conservation index and frequency profile were employed together, whereas a decrease in prediction accuracy to 62.2% was observed in the case of linear regression. Further, when PSSMs obtained from the PSI-BLAST alignment of a given sequence were employed as the input, a significant improvement in prediction accuracy was attained (83.3% and 80.3%, for ϵ-SVR and linear regression, respectively). No further improvement in the prediction accuracy was obtained when different combinations of PSSM with conservation index and frequency profile were tested (table 22). A similar study conducted for HMPs revealed that the highest prediction accuracy was obtained when a combination of conservation indices and frequency profile was employed as the input factor [62]. This could be due to the differences in the structural topology and physico-chemical properties of the TM residues in the HMP and TMB proteins.

4.4 Optimization of window size

It has been previously pointed out that smaller window sizes are noisier than intermediate window spans and that employing window spans of less than five residues is generally unsatisfactory [127]. Long spans on the other hand are known to miss small consistent features. To reduce the number of residues taken into account and identify key data points affecting the prediction accuracy, Fisher's indices [128] were calculated for all residues in the range from 5% to 100% (steps of 5%) for the window sizes ranging from 5 to 15 (steps of 2), centered at the target residue. Briefly, the Fisher's index represents the ability of a given element to maximize the distance between the centroids of the two given classes and minimize their overlap. It was employed here for feature selection due to its high interpretability. Table 27 shows the prediction accu-

racies obtained in a leave-one-out test when Fisher's analysis was applied to the chosen input data (i.e. PSSMs) in the first stage. In the case when ϵ-SVR was employed in the first stage, based on Fisher's analysis it was found that when a population comprising of only 40% of the highest ranking data points in a window size of 9 and 13 was used, the prediction accuracy increased to 82.7% and 83.2%, respectively (Table 27). Since the difference amongst the prediction accuracies thus obtained was statistically insignificant, all the tested window sizes along with their respective population size were further tested for the optimization of the second stage. Tables 23, 24 and 25 show the different prediction accuracies while optimizing the window size in the first stage. Table 26 shows the change in the prediction accuracy with different population sizes based on Fisher's analysis in the first stage, while table 28 shows the prediction accuracies for various window sizes in the second stage.

The size of the sliding window in the second stage was optimized in a similar way. Window sizes of positional scores obtained from the first stage (centered at the target residue) were progressively tested starting from a window size of 1 to 15 in steps of 1. As shown in table 27, an overall accuracy of 84.2 was achieved when the positional scores were used in a window size of 3 in the second stage. The top three indices thus identified were found to be the central residue and the two amino acids at a distance of +/- 2 residues on either side of the central residue for which the prediction was being generated. Thus, interestingly the sliding window comprises of three residues whose $C_\alpha - C_\beta$ vectors point in the same direction based on the alternate dyad repeat pattern observed in the beta strands.

4.5 Analysis of BTMX predictions

Table 29 shows the prediction accuracy for each protein in the training and cross-validation data set. The prediction accuracy for Omp32, the anion-selective porin (PDB: 1e54) was found to be exceptionally low at 38.4%. Analysis of 1e54 residues employed in the training and cross-validation data set revealed that a large portion of residues was labelled as exposed (74.5%). As has been suggested in the literature [70], residues exposed to the environment are, in general, found to be less conserved than the buried counterparts (see supplementary). However, for 1e54, the average conservation index of the residues labelled as exposed was found to be 0.4. This value is significantly higher as compared to the other proteins in the data set (see supplementary). Such a discrepancy in observed conservation indices of exposed residues could explain the difficulty in predicting the exposure status of the 1e54 residues. The per amino acid prediction accuracy is reported in the supplementary information. As expected, a higher fraction of apolar residues was found to be exposed to the lipid bilayer. Such predictions are in concert with the fact that apolar residues have a higher propensity to be exposed to the hydrophobic OM lipid bilayer. At 56.8%, GLY was predicted with a significantly lower prediction accuracy than the other amino acids. This anomaly could be due to the 'aromatic rescue' of GLY residues. The shielding of GLY residues from the protein exterior by aromatic residues is called 'aromatic rescue' of GLY and is well documented in the literature [129]. The percentage of correctly predicted buried and exposed residues is X and Y, respectively. The Sensitivity and specificity are found to be 79.6% and 94.4%, respectively and the overall prediction accuracy (ACC) of BTMX is found to be 84.2%.

SVR Radial kernel						
Window Size[WS1]	C-value	ACC	TN	TP	FN	FP
11	1.00	82.61	92.52	71.50	7.48	28.50
13	1.00	82.79	93.28	71.02	6.72	28.98
15	1.00	83.28	93.45	71.88	6.55	28.12
1	1.00	78.47	90.99	64.44	9.01	35.56
3	1.00	81.35	92.35	69.02	7.65	30.98
5	1.00	82.47	93.28	70.35	6.72	29.65
7	1.00	82.65	92.86	71.21	7.14	28.79
9	1.00	82.74	92.69	71.59	7.31	28.41
11	3.00	81.75	91.67	70.64	8.33	29.36
13	3.00	81.98	92.09	70.64	7.91	29.36
15	3.00	82.61	92.35	71.69	7.65	28.31
1	3.00	78.52	91.16	64.35	8.84	35.65
3	3.00	81.26	92.26	68.92	7.74	31.08
5	3.00	81.66	92.52	69.49	7.48	30.51
7	3.00	81.57	91.84	70.07	8.16	29.93
9	3.00	81.98	91.92	70.83	8.08	29.17
11	5.00	81.57	91.75	70.16	8.25	29.84
13	5.00	81.80	92.01	70.35	7.99	29.65
15	5.00	82.65	92.35	71.78	7.65	28.22
1	5.00	78.74	91.50	64.44	8.50	35.56
3	5.00	80.94	92.26	68.26	7.74	31.74
5	5.00	81.35	92.09	69.30	7.91	30.70
7	5.00	81.21	91.50	69.69	8.50	30.31
9	5.00	81.53	91.75	70.07	8.25	29.93
11	6.00	81.44	91.67	69.97	8.33	30.03
13	6.00	81.80	92.09	70.26	7.91	29.74
15	6.00	82.65	92.35	71.78	7.65	28.22
1	6.00	78.79	91.58	64.44	8.42	35.56
3	6.00	80.49	92.09	67.49	7.91	32.51
5	6.00	81.17	91.84	69.21	8.16	30.79
7	6.00	80.94	91.24	69.40	8.76	30.60
9	6.00	81.35	91.50	69.97	8.50	30.03
11	7.00	81.39	91.67	69.88	8.33	30.12
13	7.00	81.80	92.09	70.26	7.91	29.74
15	7.00	82.61	92.35	71.69	7.65	28.31
1	7.00	78.79	91.33	64.73	8.67	35.27
3	7.00	80.09	91.75	67.02	8.25	32.98
5	7.00	81.03	91.84	68.92	8.16	31.08
7	7.00	80.90	91.16	69.40	8.84	30.60
9	7.00	81.26	91.58	69.69	8.42	30.31

Table 23: Optimization of the window size in the first stage. SVR with a radial kernel was employed. Window size (WS1) were varied from 1 to 15, in steps of 2. $c - value$ was tested in the range of 1 to 7.

SVR Linear kernel						
Window Size[WS1]	C-value	ACC	TN	TP	FN	FP
15	1.00	80.99	90.39	70.45	9.61	29.55
15	1.00	80.94	91.24	69.40	8.76	30.60
15	1.00	81.26	90.48	70.92	9.52	29.08
15	1.00	81.21	91.50	69.69	8.50	30.31
15	1.00	79.33	87.07	70.64	12.93	29.36
15	1.00	80.90	90.14	70.54	9.86	29.46
15	1.00	80.81	89.29	71.31	10.71	28.69
15	1.00	81.21	90.73	70.54	9.27	29.46
15	1.00	81.30	91.75	69.59	8.25	30.41
15	1.00	82.02	92.69	70.07	7.31	29.93
15	1.00	81.35	92.01	69.40	7.99	30.60
15	1.00	80.94	91.24	69.40	8.76	30.60
15	1.00	81.17	91.41	69.69	8.59	30.31
15	1.00	80.94	91.67	68.92	8.33	31.08
15	1.00	80.54	91.07	68.73	8.93	31.27
15	1.00	80.58	90.39	69.59	9.61	30.41

Table 24: Testing SVR with a linear kernel in the first stage. Window size (WS1) and $c - value$ are set at 15 and 1, respectively.

Linear regression					
Window Size[WS1]	ACC	TN	TP	FN	FP
1	82.20	91.50	71.78	8.50	28.22
11	82.56	90.31	73.88	9.69	26.12
13	82.47	89.80	74.26	10.20	25.74
15	82.38	89.37	74.55	10.63	25.45
3	82.88	90.73	74.07	9.27	25.93
5	82.61	90.39	73.88	9.61	26.12
7	82.61	89.88	74.45	10.12	25.55
9	82.65	90.14	74.26	9.86	25.74

Table 25: Optimization of the window size in the first stage when Linear regression was employed.

Population size[PS]	window size [ws1]	ACC	window size [ws1]	ACC
0.05	13	79.24	15	79.87
0.15	13	81.53	15	81.53
0.10	13	81.48	15	81.44
0.25	13	82.56	15	82.56
0.20	13	81.93	15	82.38
0.35	13	82.83	15	83.24
0.30	13	82.43	15	82.65
0.45	13	82.61	15	83.01
0.40	13	83.15	15	82.74
0.55	13	83.10	15	83.15
0.50	13	82.74	15	83.24
0.65	13	83.01	15	83.10
0.60	13	82.83	15	83.06
0.75	13	82.65	15	83.19
0.70	13	82.61	15	83.15
0.85	13	82.92	15	83.28
0.80	13	82.83	15	83.15
0.95	13	82.79	15	83.15
0.90	13	82.88	15	83.28
1.00	13	82.79	15	83.28

Table 26: Optimization of the population size in the first stage based on Fisher analysis. $c - value$ was set to 1.0 and an SVR with radial kernel was employed.

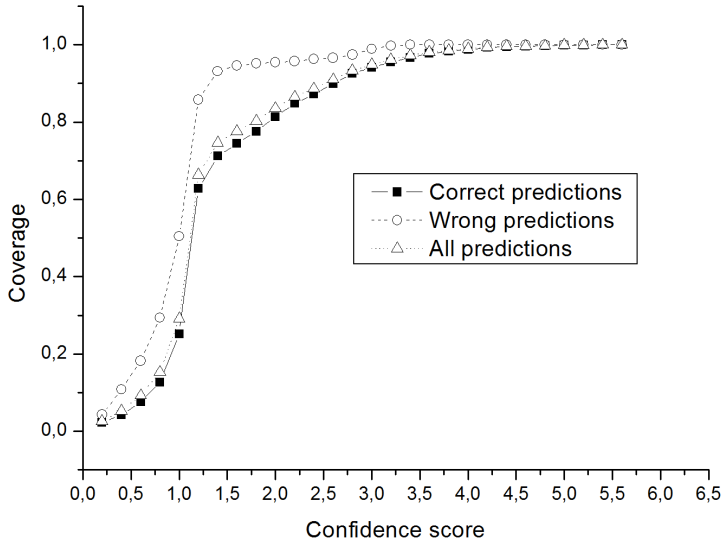


Figure 18: Confidence score coverage for predictions made for the residues in the training and cross validation data set. Only residues predicted to be in the TM region were considered.

WS1	PS [%]	WS2	ACC[%]	$WS2^a$	$ACC[\%]^a$
5	0.95	1	82.7	15	83.8
7	0.90	1	83.2	12	84.1
9	0.40	1	82.7	12	83.8
11	0.65	1	82.9	10	83.9
13	0.40	1	83.2	3	84.2
15	0.90	1	83.3	15	84.0

Table 27: Effect of population size on the prediction accuracy. Optimization of population size of input data of the first stage and window size of the second stage. PSSMs with window size 5 to 15 and $\epsilon - SVR$ with a radial kernel were tested for generating the positional score in the first stage with WS2 of the second stage set to 1. After finding out the optimal population size based on higher accuracy, the size of the sliding window (WS2) of the second stage was tested from 1 to 15 neighbouring residues. Only the top PS% of all data points in a given window were taken into account while calculating the accuracy based on a jack-knife test. Only the highest prediction accuracies from all the tested parameters are shown. WS1 = Window size of the input data of the first stage, WS2= window size of positional scores in the second stage, PS = Population size. $WS2^a$ = optimal window size of the second stage. ACC^a = Final prediction accuracy based on the optimal WS1, population size and WS2.

The confidence score coverage of the predictions made for the training and test data are shown in figure 18. As shown, the confidence score of the incorrect predictions is lower than the correct predictions. As discussed in section 4.7, this feature of the BTMX method can be used to filter incorrect predictions.

4.6 Comparison with YU method

The YU method [17] was re-implemented here to predict the rSASA value of the proteins in the training and cross-validation data set. For the YU method, the correlation between the observed and the predicted rSASA was found to be 0.70 and 0.66 for the BTMX data set and the YU data set, respectively. When classifying predicted rSASA values as buried or exposed based on the rSASA cutoff criteria such that the exposed residue abundance was close to 50% [17], the corresponding prediction accuracy based on a leave-one-out test was found to be 54.3% and 53.2%, respectively. The prediction accuracy of the BTMX method at 84.2% is significantly higher than the YU method. This is likely due to the fact that BTMX employs a SVC in the second stage while a constant numerical cut-off is used to report the accuracy of the YU method. In order to compare the YU and the BTMX method w.r.t. correlation between the observed and predicted rSASA values, a SVM for regression was incorporated in the second stage of the BTMX method to generate real value rSASA predictions. The correlation was found to be 0.70, which is comparable to the YU method (see supplementary). This correlation goes up to 0.77 when Omp32, the anion-selective porin (PDB: 1e54) is excluded from the training and cross validation data set. No additional attempt was made to further optimize the BTMX method for predicting real value rSASAs. All the C-Values and window sizes mentioned in the YU method were tested and only the highest accuracy is reported.

Window Size[WS2]	ACC (WS1 = 13, PS=0.40)	ACC (WS1 = 15, PS=0.90)
10	84.04	83.46
11	84.00	83.55
12	84.04	83.69
13	84.09	83.64
14	84.09	83.69
15	84.04	84.04
1	83.15	83.28
2	83.28	83.15
3	84.22	83.64
4	84.18	83.69
5	84.09	83.78
6	84.13	83.69
7	84.13	83.64
8	84.09	83.73
9	84.04	83.78

Table 28: Optimization of the window size in the second stage.

4.7 Analysis of BTMX predictions for the test data set and the TMBs involved in transport of hydrophobic compounds

The average prediction accuracy for the three proteins in the test data set is 80.2%. In total, the exposure status of 207 out of 261 beta strand residues in the TM region of 1k24_A, 2POR_A and 1PRN_A is predicted correctly. Figure 19 (top) shows the confidence score coverage for the predictions made by BTMX on the test data set. For the test data set, the average confidence score for wrongly predicted residues was found to be 1.2, 0.7 and 0.9 for 1k24_A, 2POR_A and 1PRN_A, respectively. The corresponding confidence score for the correctly predicted residues for the same proteins was 3.4, 3.0 and 3.1, respectively. As shown in figure 19-A, approx. 60% of the incorrectly classified residues have a confidence score of ≤ 1. The corresponding number for correct predictions is roughly 35%. We propose the use of confidence score generated for each prediction to filter the prediction results.

Recently, Hearn *et al.* experimentally showed that the inward pointing kink in β-strand S3 creates an opening that aids in the lateral diffusion of hydrophobic substrates through the outer membrane long-chain fatty acid transporter FadL in *Escherichia coli* [130]. We obtained the crystal structures of the wild type FadL and various mutants created to demonstrate the lateral diffusion mechanism from the PDB and predicted the exposure status of the TM residues (see supplementary). We also predicted the burial status of TM residues in the FadL homologue (PaFadL) from *Pseudomonas aeruginosa* (PDB: 3DWO), which has a low (20%) sequence identity to *E. coli* FadL. Figure 19-B shows the confidence score coverage for the predictions made for 3DWN residues. As shown, with an average value of 1.4, incorrect predictions have a lower confidence score. The corresponding value for correct predictions is 3.0. The vertical line plotted at confidence score value = 1 shows that almost 40% of the incorrect predictions have a confidence score of ≤ 1. The corresponding value for correct predictions

Protein	Chain	Total count	Correct predictions	Specificity	Sensitivity	PCN	ACC[%]
1qd6	C	77	65	0.72	1.00	0.74	84.4
1p4t	A	49	40	0.64	1.00	0.73	81.6
1a0s	P	137	115	0.80	0.93	0.73	83.9
1e54	A	133	51	0.41	0.58	0.19	38.4^a
1fep	A	195	178	0.90	0.97	0.83	91.3
1i78	A	68	57	0.69	1.00	0.75	83.8
1kmo	A	178	160	0.89	0.96	0.78	89.9
1qj8	A	43	32	0.52	1.00	0.65	74.4
1qjp	A	55	44	0.61	1.00	0.71	80.0
1t16	A	108	94	0.88	0.90	0.82	87.0
1xkh	A	173	155	0.85	0.99	0.77	89.6
1xkw	A	179	159	0.86	0.96	0.78	88.8
2erv	A	51	40	0.65	0.89	0.72	78.4
2gsk	A	165	153	0.91	0.98	0.85	92.7
2mpr	A	137	120	0.79	1.00	0.76	87.6
1thq	A	52	40	0.69	0.82	0.73	76.9
2j1n	A	116	93	0.71	0.98	0.63	80.2
1tly	A	76	72	0.95	0.95	0.95	94.7
2f1v	A	51	44	0.74	1.00	0.77	86.3
1qfg	A	182	162	0.87	0.97	0.75	89.0

Table 29: Prediction accuracy for proteins in the training and cross-validation data set. Using PSSM as the input factor, leave-one-out test was conducted on the training and cross-validation data set with window size of the first stage set to 13. Only the top 40% of the total 180 neighbouring data points were employed to obtain positional scores in the first stage. The window size of the second stage was set to 3. *a* is discussed in section 4.5.

is roughly 20%. Further, the prediction accuracy for BTMX predictions in this case was 81.1% over 159 residues. The accuracy goes up to 86.2% when residues with rSASA value ≤ 0.01 are considered as buried. The exposure status of the TM residues of PaFadL is predicted with an accuracy of 74.5%. The accuracy goes up to 81.0% when residues with rSASA value ≤ 0.01 are considered as buried. The average confidence score for correct and wrong predictions is 3.1 and 0.9, respectively.

4.8 Comparison of physico-chemical properties of oligomeric and non-oligomeric strands

Various physico-chemical and evolutionary properties of protein interfaces have been analysed and used to identify the interfaces of globular proteins [102, 103]. Profiles of propensities of interfacial residues were employed by Dong *et al.* to predict binding sites of proteins [125]. More specifically, Elcock *et al.* suggested that the protein oligomerization states can be identified by interface conservation [131]. However, in a later study on a data set of 64 proteins, Caffrey *et al.* found that protein-protein interfaces are only rarely significantly more conserved than the rest of the protein surface. These slightly more conserved interface residues mostly belonged to an enzyme active site [132]. It is notewor-

Amino Acid	Total	Exposed[%]	ACC[%]	Sensitivity	Specificity
ALA	223	59.19	87.90	0.82	0.91
GLU	88	12.50	86.40	0.87	0.98
ASP	81	11.11	91.40	0.92	0.99
GLY	271	56.45	56.80^a	0.50	0.85
PHE	97	78.35	84.50	0.59	0.95
ILE	114	53.51	94.70	0.98	0.91
HIS	19	26.31	89.50	0.87	1
LYS	50	14.00	90.00	0.89	1
MET	53	58.49	90.60	0.84	0.95
LEU	260	71.54	91.90	0.86	0.85
ASN	90	18.89	85.60	0.86	0.98
GLN	60	21.67	81.70	0.81	1
PRO	50	46.00	92.00	0.90	0.96
SER	140	32.14	73.60	0.74	0.95
ARG	109	8.25	89.90	0.93	0.98
THR	158	37.97	88.60	0.85	0.99
TRP	37	64.86	86.50	0.75	0.92
VAL	210	64.28	93.30	0.92	0.89
TYR	115	45.22	81.70	0.76	0.98
All	2225	47	84.20	0.79	0.94

Table 30: Prediction accuracy for proteins in the training and cross-validation data set. Using PSSM as the input factor, leave-one-out test was conducted on the training and cross-validation data set with window size of the first stage set to 13. Only the top 40% of the total 180 neighbouring data points were employed to obtain positional scores in the first stage. The window size of the second stage was set to 3. *a* is discussed in the section 4.5.

thy that in the data set employed in this study, the conservation index and the information content [133] values for the oligomeric interfaces were not found to be statistically significantly different from the rest of the protein surface.

As described in section 4.11.1, the data set employed here for the analysis of the oligomeric interfaces consisted only of 8 oligomeric TMBs and is slightly different from the data set employed to develop the BTMX methods. Oligomeric interface residues and strands were defined as described in the Methods section. The physico-chemical and evolutionary properties of the exposed, outpointing residues at the oligomeric and non-oligomeric strands were analysed at the strand level and the residue level (see section 4.11.1). Tables Tab:04-strandwise and Tab:04-residuewise show the mean value for a given physico-chemical property at the oligomeric and the non-oligomeric region at the strand and the residue level, respectively. Only the statistically highly significant properties (p-value ≤ 0.05) are enlisted.

Table 31 shows that the parameters related to the morphology of the side chain of an amino acid, such as the width, length, size and volume [97] prominently have a higher mean value for the strands at the oligomeric interface as compared to the non-oligomeric strands. Furthermore, the overall size, volume [134] and bulkiness [98] of the residues in the oligomeric strands are also larger than the corresponding values at the non-oligomeric strands. As shown in table 31, oligomeric strands have bulkier side chains than the non-oligomeric

Strand-wise analysis				
scale	description	mean value oligo	mean value non-oligo	p-value
AVBF000102	Screening coefficients gamma, non-local	0.78	0.62	5.88e-06
AVBF000101	Screening coefficients gamma, local	0.69	0.55	7.015e-05
FAUJ880105	STERIMOL minimum width of the side chain	0.69	0.56	0.0002
FAUJ880102	Smoothed upsilon steric parameter	0.77	0.65	0.0009
FAUJ880113	pKa(RCOOH)	0.79	0.70	0.0011
MUNV940105	Free energy in beta-strand region	0.09	0.14	0.0011
FAUJ880108	Localized electrical effect	0.18	0.09	0.0024
FAUJ880104	STERIMOL length of the side chain	0.42	0.34	0.0032
FAUJ880101	Graph shape index	0.64	0.54	0.0050
DAWD720101	Size	0.64	0.54	0.0074
EISD840101	Consensus normalized hydrophobicity scale	0.83	0.76	0.0261
RADA880102	Transfer free energy from oct to wat	0.67	0.60	0.0266
ZIMJ680102	Bulkiness	0.77	0.68	0.0266
KRIW790103	Side chain volume	0.53	0.45	0.0281
FAUJ880103	Normalized van der Waals volume	0.45	0.38	0.0337
ZIMJ680104	Isoelectric point	0.39	0.36	0.0459
GOLD730102	Residue volume	0.52	0.45	0.0476

Table 31: Strand-wise analysis of physico-chemical properties of beta strands at the oligomeric interfaces and at the rest of the protein surface. Strands were defined as belonging to the oligomeric interface region when more than 70% of the residues in the strand showed a decrease in rSASA value in the oligomeric form as compared to the monomeric form. All out-pointing residues that lie in the hydrophobic core region of the membrane and are exposed to the bilayer were used to calculate the mean values.

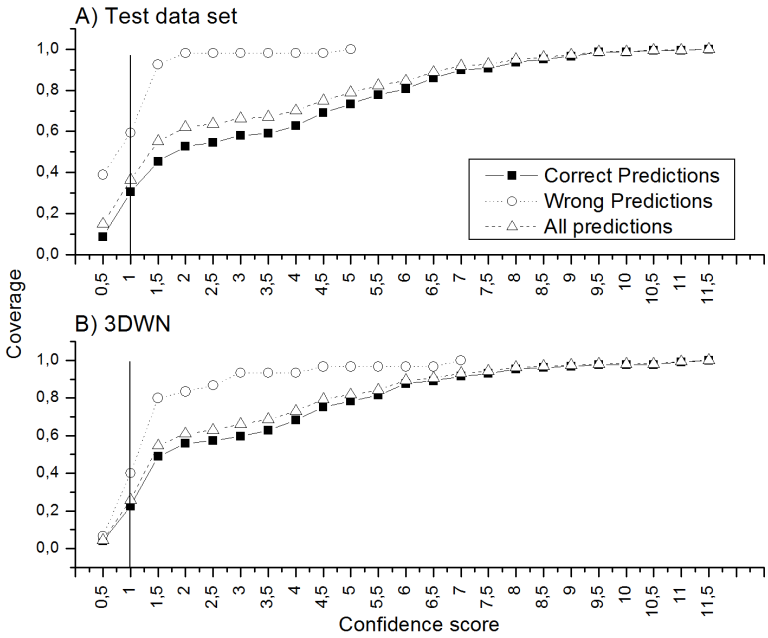


Figure 19: Confidence score analysis - BTMX outputs a confidence score for each prediction. The vertical line at x = 1 shows a probable threshold for filtering BTMX predictions based on the confidence score A) confidence score coverage for all the residues in the test data set. B) confidence score coverage of BTMX predictions for FadL mutant 3DWN. Only residues predicted to be in the TM region were considered. The average confidence scores for correct and incorrect predictions are 3.0 and 1.4, respectively.

strands. Further, the steric effect of the side chain estimated using the Graph shape index [97] and the Upsilon steric parameter [97] showed that the side chains in the oligomeric strands exert a higher steric effect. Similar trends in the differences between the mean values were observed for the screening coefficients, side chain width, localized electric parameter and free energy in beta-strand region at the residue level (Table 32).

The physical significance of the differences in the morphological parameters is reflected in the beta sheet propensity [135] and hydrophobicity [80, 92] of the residues in the oligomeric and non-oligomeric strands. In the following, we will discuss these two parameters in more detail. Munoz *et al.* have reported the beta strand propensities of amino acid residues based only on the dihedral angles such that neither the identity nor the conformation of the neighbouring residues was not taken in account. As reported, the empirical pseudo-energies derived from the beta strand propensities are comparable to experimental free energies necessary for the transition of a residue from a free to the defined dihedral state [135]. As shown in table 31, the mean free energy value of the exposed out-pointing residues in the oligomeric beta strands (0.09) is slightly lower than the value of non-oligomeric strands (0.14). This suggests that these residues in the oligomeric beta strands have a higher beta strand propensity than the residues in the non-oligomeric strands. Interestingly, it is also known

Residue-wise analysis				
scale	description	mean value oligo	mean value non-oligo	p-value
FASG760102	Melting point	0.68	0.76	0.0003
FAUJ880108	Localized electrical effect	0.16	0.10	0.0080
AVBF000102	Screening coefficients gamma, non-local	0.75	0.68	0.0125
MUNV940105	Free energy in beta-strand region	0.11	0.16	0.0220
FAUJ880105	STERIMOL minimum width of the side chain	0.67	0.60	0.0227
RADA880105	Transfer free energy from vap to oct	0.80	0.84	0.0362
RADA880104	Transfer free energy from chx to oct	0.86	0.89	0.0466
AVBF000101	Screening coefficients gamma, local	0.66	0.60	0.0475

Table 32: Residue-wise analysis of physico-chemical properties of beta strands at the oligomeric interface and at the rest of the protein surface. Residues were defined as belonging to the oligomeric interface region when they showed a decrease in rSASA value in the oligomeric form as compared to the monomeric form. All out-pointing residues that lie in the hydrophobic core region of the membrane and are exposed to the bilayer were used to calculate the mean values.

from the analysis of soluble proteins that sterically bulky amino acid side chains increase the beta sheet propensity [136]. Here, the mean value of the transfer free energy [80] from octanol to water is higher for oligomeric strands (0.67) than for the non-oligomeric strands (0.60). This signifies that the oligomeric interface is slightly more hydrophobic than the non-oligomeric surface of the protein. This observation is further corroborated by the higher hydrophobicity value [92] of the oligomeric interface (0.83) as compared to the non-oligomeric surface (0.76).

4.9 Web server

We have developed the BTMX web server for predicting the exposure status of TMB residues. The web server takes input in the form of a FASTA sequence or a MSA and calculates the predicted exposure status along with the confidence score for each residue. Since BTMX is trained using only the residues in the hydrophobic core region of the proteins in the data set, the standalone version of PROFtmb [137] is employed to predict the TM region. BTMX then generates snake-plots with annotated exposure status information for each residue predicted to be in the TM region as part of its output as shown in figure 20. BTMX also reports the predicted numeric rSASA value for each residue in the hydrophobic core region. The BTMX web server is available under the BTMX tab at `http://service.bioinformatik.uni-saarland.de/tmx-site`

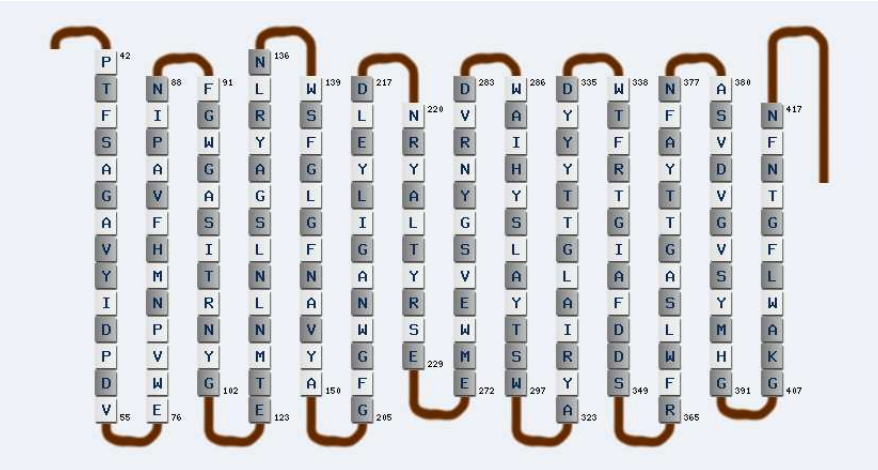


Figure 20: BTMX web server output - As part of its output, the BTMX web server generates snake-plots annotated with the exposure status prediction. The strands in the TM region are predicted using the PROFtmb standalone program. Dark colour represents residues predicted as buried while light colour represents exposed residues.

4.10 Conclusions

BTMX predicts the exposure status of TMB residues in the training and cross-validation data set of 2225 residues with an accuracy of 84.2% and also generates a confidence score for the predictions. The positional scores in the first stage of the BTMX method are generated so that the observed rSASA value is maximally correlated to the input PSSMs. Further, the predicted exposure status of the target residue is most accurate when the sliding window in the second stage consists of the target residue and one residue on either side of the target residue at a distance of +/- 2 residues. In such a case, all three residues have their respective $C_\alpha - C_\beta$ vector in the same direction. The average prediction accuracy of BTMX on a non redundant test data set was found to be 80.1%. Analysis of the prediction results for the FadL mutants shows that BTMX is sensitive to single residue mutations and the predictions in combination with the reported confidence score can be used as an aid while designing mutational experiments. Further, we have identified several physico-chemical properties that can be used to develop a computational method to differentiate between oligomeric and non-oligomeric TMB strands. The BTMX web server employs the standalone PROFtmb program [137] to predict the beta strands and generates a colored snake-plot annotated with the predicted exposure status of each residue. Given the dearth of methods focusing on the prediction of the exposure status of TMB residues, and the fact that only a few TMB crystal structures are available, the predicted exposure status should be helpful in understanding the structural organization of TMBs and can be used as an additional parameter in predicting the topology, oligomeric state and in identifying TMBs from genome wide data [138].

4.11 Methods

4.11.1 Generation of benchmark data set

The non-redundant data set of known TMB structures with sequence identity $\leq$ 30% was compiled based on the literature. TMB structures aligned to the membrane normal and with their respective hydrophobic thickness, defined as the region for which the probability of occurrence of the hydration waters of the lipid head-groups is 0.0, were retrieved from the OPM database [20]. Only the residues within the range +0.65 to -0.65 units of their respective hydrophobic thickness were considered as residing in the hydrophobic core region. As described elsewhere [70], protein sequences for which we could not retrieve more than 20 homologous sequences or where the average pair-wise sequence identity of the aligned retrieved sequences was greater than 80%, were excluded from the training and cross-validation data set. The training and cross-validation data set comprising of 20 protein chains with 2225 TMB residues is as follows: 1qd6_C, 1p4t_A, 1a0s_P, 1e54_A, 1fep_A, 1i78_A, 1kmo_A, 1qj8_A, 1qjp_A, 1t16_A, 1xkh_A, 1xkw_A, 2erv_A, 2gsk_A, 2mpr_A, 1thq_A, 2j1n_A, 1tly_A, 2f1v_A, 1qfg_A (see supplementary). A separate test data set comprising of three non redundant protein chains (1k24_A, 1prn_A, 2por_A) was used to test the accuracy of the final BTMX model. These protein chains were excluded from the training and cross-validation data set as not enough (14, 19 and 3, respectively) diverse homologous sequences could be retrieved after MSA. The non-redundant data set for oligomeric interface analysis consists of 8 oligomeric TMBs namely 1a0s_P, 2o4v_A, 3prn_A, 2por_A, 1qd6_C, 2j1n_A, 2mpr_A, 1e54_A. The sequence identity of this data set is $\leq$ 25%. As mentioned above, some oligomeric TMBs included in the oligomeric data set were excluded from the BTMX training and cross-validation data set as enough diverse homologous sequence could not be obtained for those proteins.

Frequency profiles were estimated using a modified AL2CO [112] program suite, as previously described [70]. Conservation indices were calculated from the multiple sequence alignment of a given protein sequence by employing the variance based method implemented in AL2CO. PSSMs were generated using the blastpgp program obtained from the NCBI website. blastpgp was run in the PSI-BLAST mode and PSSMs were built after three iterations of scanning the target sequence against the non-redundant reference data set.

The classification of a residue as being buried or exposed to the lipid bilayer was based on its rSASA value, which was used as the output parameter. The SASA values were calculated with the VOLBL program suite [113,114] employing a probe radius of 2.2Å. As previously discussed [70], this probe radius approximates the effective radius of the CH_2 group of hydrocarbon chains of phospholipids. Only the functional oligomeric forms of the proteins were considered and, when necessary, the two faces of the TM region (the cytoplasmic and exoplasmic faces) were capped with dummy atoms before computing SASA values to avoid internal residues from being labelled as exposed [70]. SASA values were then normalized to generate rSASA values by dividing them by the SASA values for each amino acid X in the context of the tri-peptide G-X-G. The tri-peptides employed for normalizing TMBs had a flat beta sheet type conformation. Based on their rSASA value and employing a cutoff of 0.0, 1176 (53%) and 1049 (47%) residues were labelled as buried and exposed, respectively. For

comparing the BTMX and the YU method, the rSASA values were also calculated using the MSMS program [139] as described by Yuan *et al.* [17]. The per-residue correlation between the SASA values generated by the two methods was found to be 0.92. The corresponding correlation for the average SASA per amino acid and the reference values employed in the BTMX and YU method was found to be 0.99 and 0.98, respectively.

The residues in a beta strand that showed a decrease in the rSASA value in the oligomeric state as compared to the monomeric state were considered to be at the oligomeric interface. A strand was defined to be at the oligomeric interface if the number of residues classified to be at the oligomeric interface in that strand was more than a given cutoff. Such a technique of employing surface patches instead of individual residues in order to predict protein-protein interaction sites has been reported in the realm of globular proteins [102, 103]. The values for the various physico-chemical and morphological properties were obtained from the AAIndex database [71]. The information content of each position in a given amino acid sequence was calculated as described by Schneider *et al.* [133].

4.11.2 Estimation of the in/out dyad repeat pattern based on the $C_\alpha - C_\beta$ orientation

The barrel axis was determined as the geometric center of the structural coordinates obtained from the OPM database. The $C_\alpha - C_\beta$ vector orientation was then determined based on the planar distance of the C_α and C_β atoms from the barrel axis. A residue was classified as out (i.e. pointing away from the barrel axis), if its C_β distance from the barrel axis was greater than the corresponding C_α distance.

4.11.3 Performance evaluation

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

$$cc = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{[n\sum {x_i}^2 - (\sum x_i)^2][n\sum {y_i}^2 - (\sum y_i)^2]}} \tag{71}$$

Prediction accuracy (ACC), sensitivity, selectivity and ratio of correct exposed predictions (PCN) was calculated as follows:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{72}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{73}$$

$$Specificity = \frac{TP}{TP + FP} \tag{74}$$

$$PCN = \frac{TN}{TN + FP} \tag{75}$$

where TP = number of residues correctly predicted as buried, TN = number of residues correctly predicted as exposed, FP = number of residues wrongly predicted as buried, and FN = number of residues wrongly predicted as exposed.

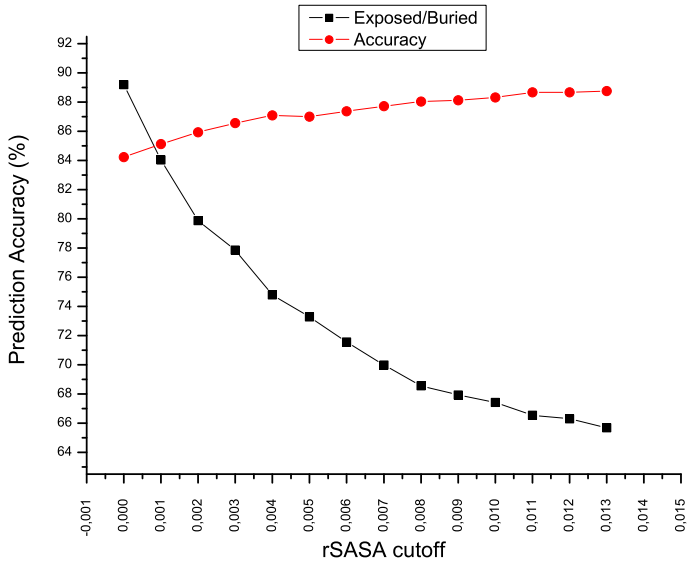


Figure 21: Prediction accuracy increases with increasing threshold. More importantly, the number of residues labeled as buried in the training and cross validation data set increases as well and the data set becomes highly biased. Ideally the data set should have equal numbers of residues labeled as buried and exposed and the ratio of both populations should be close to 1.0.

4.11.4 Derivation of BTMX

BTMX is an extension of the TMX method previously described by us [62] for HMPs. Three potential sources of input data namely, conservation indices, positional frequency profiles and PSSMs were identified from the literature [17, 59,70,140] and all possible combinations were tested for the derivation of BTMX. These input factors were generated from the given protein sequence as described above. An rSASA value of 0.0 was used as the cutoff to label the residues as buried or exposed in the training and cross validation data set. This value was chosen so that both the classes (i.e. buried or exposed) are equally populated. Figure 21 shows that although the prediction accuracy of the BTMX method increases with an increasing cutoff value during the labelling of the data set, the number of residues labelled as buried increases and introduces a bias in the training and cross validation data set.

BTMX is a two stage classifier. In the first stage, a sliding window (centered at the target residue) consisting of the input factor was employed to obtain positional scores using a SVM for regression ($\epsilon - SVR$) with a radial kernel. In the second stage, to again incorporate the contextual information for individual positional score, a sliding window consisting of positional scores obtained from the first stage was employed as the input for a Support Vector Classifier (c-SVC) with a linear kernel to predict the exposure status for each residue. Sliding windows of size ranging from 1 to 15 residues were tested based on the fact that

beta strands that span the OM (with a tilt of $20-45\,^{\circ}$) mostly consist of 9 to 11 residues [26]. For the first stage, a leave-one-out test was conducted to optimize the C value and window sizes ranging from 1 to 7 and 1 to 15 (in steps of 2), respectively, based on higher prediction accuracy (see supplementary). Fisher's analysis was then conducted on the parameters yielding the highest accuracies. The size of the sliding window in the second stage was optimized in a similar way. It is to be noted that in the first stage, linear regression and a SVR with a linear kernel were also tested. The R implementation of support vector classifier (SVC)/support vector regression (SVR) [107,116] was used for the current work. The use of multiple stages to incorporate contextual information is widespread in the literature [62, 64, 141].

5 TMBHMM: A frequency-profile based HMM for predicting the topology of transmembrane beta barrel proteins and the exposure status of transmembrane residues

5.1 Overview

Acknowledgements: The software developed and the results presented in this chapter represent joint work by Aaron Goodman, a summer student from the University of Pennsylvannia who I supervised during a 2-month internship in summer 2008, by Nitesh Kumar Singh who I supervised during his Master thesis in bioinformatics from June 2009 to January 2010, and by myself. Besides supervising both students, my own contribution in this work was coming up with the initial idea and stating the problem statement. I generated the training data set and guided the aforementioned students to look at relevant literature. I reviewed the TMBHMM software during its implentation stagee and along with Nitesh and Aaron analysed the results generated by the TMBHMM program.

The existing sequence-based computational methods in the realm of TMBs can be classified into two main categories. The computational methods in the first category aim at identifying the TMBs in a given proteome based on the sequence [1,22,25,137,142,143]. The computational methods in the other category focus on determining the structural topology of the given sequence [64,137,144]. There are also methods that combine both features by providing the structural topology of the identified TMBs [1, 25, 142]. In contrast to the extensively studied globular proteins following the pioneering work of Rost *et al.* [145], the problem of predicting exposed/buried residues has remained untouched for TMB proteins. In addition to predicting membrane spanning regions and the structural topology, prediction of the exposure status is of interest due to its implied applications in channel engineering and site-specific mutational studies [146, 147]. As discussed above, by exposed residues we mean those residues that are in contact with the membrane lipids. In contrast, buried residues are hidden in the protein structure. To the best of our knowledge, so far no method gives the exposure status of the residues predicted to be in the transmembrane region of the putative TMBs.

In this work, we have developed a comprehensive computational method (TMBHMM) based on a Hidden Markov Model to predict the structural topology of TMBs by employing only the frequency profiles of the amino acids in a given sequence as input. The novelty of the method is that it also predicts the exposure status of the transmembrane residues. The prediction accuracy of TMBHMM has been compared with PRED-TMBB [25], which has been reported to have one of the highest reported prediction accuracies [142] and we show that TMBHMM is at least as good as PRED-TMBB in terms of strand prediction accuracy. We have also established the TMBHMM web server that accepts amino acid sequence or multiple sequence alignment as input and predicts the structural topology of the given amino acid sequence annotated with the exposure status. The training of the TMBHMM was performed on a non-redundant data set of 19 TMBs. The self consistency test yielded Q_2 accuracy of 0.87, Q_3 accuracy of 0.83, Matthews correlation coefficient of 0.74 and SOV

for beta strand of 0.95. In self consistency test the method predicted 83.0% of transmembrane residues with correct exposure status. The jack-knife test yielded Q_2 accuracy of 0.86, Q_3 accuracy of 0.83, MCC of 0.72 and SOV for beta strand of 0.92. TMBHMM predicts the exposure status of the correctly predicted transmembrane residues with an accuracy of 83.21%.

5.2 The HMM architecture

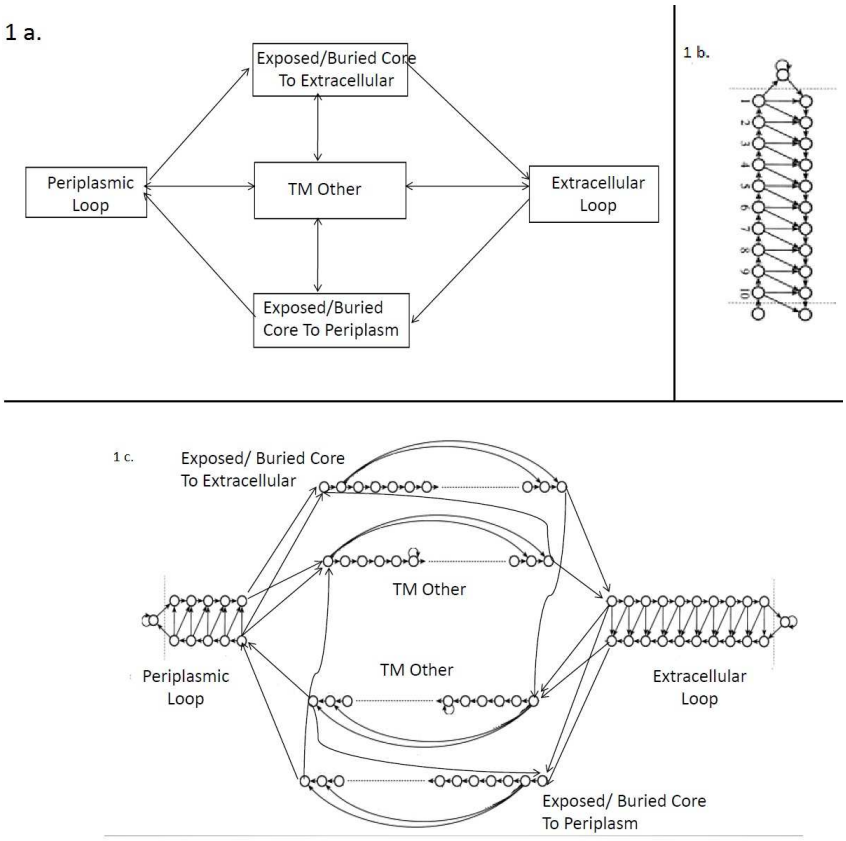


Figure 22: HMM architecture.

The HMM architecture employed in TMBHMM is shown at different levels of detail in figure 22. Figure 22-(a) shows the general overview of the HMM. Boxes in figure 22-(a) correspond to the membrane, the periplasmic and the cytoplasmic regions of a TMB, respectively. The 'TMOther' region consists of residues that are in the transmembrane region but do not belong to any beta strand. Each region represented by a box in figure 22-(a) corresponds to a sub-model depicting an array of states which share similar transition probabilities. The structure of the sub-model is shown in figure 22-(b). Each region in figure 22-(a) can be expanded into its constituent array of states as shown in figure 22-(b) except for the membrane region which lacks the self loop state. The overall architecture of the HMM model is shown in figure 22-(c). The HMM architecture

is similar to the previously established methods [46] except for the membrane region. In general, existing HMM based methods have states for the extracellular and the periplasmic loop and the membrane regions. In the case of TMBHMM, the membrane region has been further divided into two states to capture the topological signal for exposed and buried residues. Thus, unlike upward and downward strands in the membrane region [46], TMBHMM has five different arrays of states. These five arrays of states include exposed-membrane to extracellular, buried-membrane to extracellular, exposed-membrane to periplasm, buried-membrane to periplasm and TMOther. As an example, the 'exposed-membrane to extracellular' state consists of residues in the membrane region that are exposed to the bilayer and the strand on which they are located traverses towards the extracellular region. Corresponding definitions apply to the other four arrays of states. The minimum and the maximum allowed lengths for the transmembrane state arrays are 3 and 15, respectively. The minimum length of the loop and the TMOther state arrays is 1 and there is no maximum due to the presence of the self-loop state.

5.3 Prediction of the structural topology of TMBs

5.3.1 Estimation of rSASA threshold value

rSASA cutoff	Q_2	Q_3	MCC	Core accuracy	Exposure accuracy	Exposed/Buried
0.01	87.0	83.0	0.74	91.0	80.0	1.12
0.02	86.0	83.0	0.73	90.0	81.0	1.10
0.03	87.0	83.0	0.74	91.0	83.0	1.05
0.04	86.0	82.0	0.73	90.0	83.0	1.01
0.05	86.0	83.0	0.73	90.0	82.0	0.95

Table 33: Accuracy measures at different rSASA thresholds. Exposed/Buried is the ratio of observed exposed and buried residues at each rSASA threshold.

In the training data set (refer to section 5.8.2), the relative Solvent Accessible Surface Area (rSASA) value was used to label the transmembrane residues as exposed or buried. Since there is no consensus on which rSASA value is to be used as threshold [70], we explored all reasonable rSASA values from 0.01 to 0.05 in steps of 0.01 as threshold. While selecting a rSASA threshold we also have to consider the random probability of an amino acid being exposed or buried. Hence, a rSASA threshold giving equi-partition of the data set as buried or exposed is more acceptable. The various accuracy measure employed in this study have been defined in section 5.8.1. The results with different rSASA thresholds are shown in table 33. As shown (table 33), all threshold values give comparable prediction accuracies. Thus, based on the slightly higher prediction accuracy, an rSASA cutoff value of 0.03 was chosen as the threshold for labelling the residues in the training data set as buried or exposed. Hence, residues with an rSASA value of $\leq$ were labelled as buried, while the transmembrane residues with an rSASA value > 0.03 are labelled as exposed.

5.3.2 Determination of optimal labelling feature in the membrane spanning region

Wimley *et al.* have shown that the in/out dyad repeat pattern is strictly followed by TMBs [66]. However, as a result of shielding by the side chains of the neighboring residues, it can be argued that not all out-pointing residues are necessarily exposed to the lipid bilayer and that even some in-pointing residues can be exposed to the bilayer. The in/out status of residues in the membrane region of TMBs was determined as described in the Methods section. Table 21 in section 4.2 shows the comparison of the in/out and the buried/exposed status of residues in the membrane region. As discussed above in section 4.2, approximately 27% of a total of 1041 out-pointing residues are found to be hidden in the protein structure. Also, 17% of a total of 781 in-pointing residues are found to be exposed to the lipid bilayer.

Label	Q_2	Q_3	MCC	Core accuracy	Exposure accuracy
Exposed/Buried	87.0	83.0	0.74	91.0	83.0
C_{in}/C_{out}	85.0	79.0	0.69	82.0	75.0
No labelling	85.0	81.0	0.70	86.0	-

Table 34: Comparison of accuracy with no labelling, Exposed/buried labelling and alternate dyad repeat (C_{in}/C_{out}) labelling in the membrane spanning region.

It is noteworthy that state of the art computational methods in the realm of TMBs employ the regular in/out dyad repeat pattern in determining TMB strands and identifying TMBs from genomic data [66, 148]. A shortcoming of such methods is that they can not account for the shielding effect of neighboring residues. Moreover, the in/out dyad repeat pattern has to be empirically determined for these methods and if the starting residue in the membrane is assigned the wrong status, the error is propagated throughout the strand. To circumvent this, TMBHMM employs the exposure status of membrane residues instead of the fixed in/out dyad repeat pattern. TMBHMM prediction accuracy for the case when the membrane residues were labelled based on the in/out instead of the exposure status are shown in table 34. As shown in table 34, the exposure status prediction accuracy is 83.0% when the membrane spanning region is labelled w.r.t. its buried/exposed status. The corresponding accuracy for C_{in}/C_{out} labelling is 75.0%. The core accuracy for exposed/buried and C_{in}/C_{out} labelling is 91.0% and 82.0%, respectively. Thus, as shown in table 34, TMBHMM prediction accuracy is marginally higher for all the accuracy measures, when the exposure status labelling instead of in/out labelling is applied to the residues in the membrane region. We also calculated the TMBHMM prediction accuracy when the training was performed without labelling the residues in the membrane core region. As shown in table 34, all the accuracy measures are slightly higher when the exposure status of the membrane residues was incorporated in the TMBHMM architecture. Thus, the inclusion of the exposure status of the residues in the membrane region resulting in separate states for membrane residues, enables TMBHMM to predict the exposure status of query sequences and this additional number of states in the TMBHMM architecture does not adversely affect the prediction accuracy. In future, a combination of exposed/buried and in/out labelling can be used to extract more information

about the membrane residues.

5.3.3 Prediction accuracy of the TMBHMM method

Prediction accuracy for the training data set						
Test	Q_2	Q_3	MCC	Core	SOV	Exposure
Self-consistency	87.0	83.0	0.74	91.0	95.0	83.0
Jack-knife	86.0	83.0	0.72	88.0	92.0	83.0

Table 35: Accuracy measures of training data set for self-consistency test and jack-knife test.

The prediction accuracy measures for both the self-consistency and the jack-knife test for TMBHMM are given in table 35. As shown, the leave-one-out Q_2 accuracy of TMBHMM is 86.0%, while the Q_3 accuracy of determining the periplasmic, cytosolic and the membrane region is 83.0%. The core accuracy of accurately predicting the residues in the membrane region is 88.0%. The exposure status prediction accuracy for the residues correctly predicted to be in the membrane region is 83.0%.

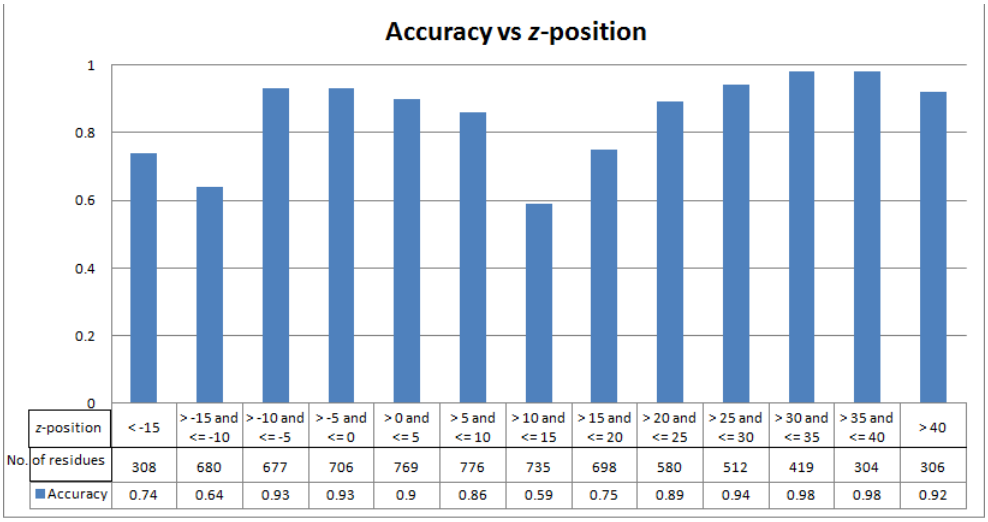


Figure 23: Q_2 accuracy as a function of z-coordinate.

The Q_2 accuracy for strand/non-strand predictions was further analysed as a function of the z-position relative to the membrane center. As shown in figure 23, there is a drop in Q_2 accuracy at z-position between -10 Åand -15 Åand between 10 Åand 15 Å. It is noteworthy that the average half length of bilayer thickness for our training data set, determined based on the PDB structures obtained from the OPM database [20] is 11.78 Å. For the training of TMBHMM, this value is used as the boundary for classifying a residue as a membrane or a non-membrane residue. Hence, we can conclude that most misclassifications were observed at the boundaries, which is a common problem with the TMB prediction methods [46]. These misclassifications can be attributed to the

	One neighbor correct	Both neighbors correct	One neighbor wrong	Both neighbors wrong
For Q_2 prediction				
Correct prediction	86.0	97.0	NA	NA
Wrong prediction	NA	NA	66.0	95.0
For exposure prediction				
Correct prediction	85.0	93.0	NA	NA
Wrong prediction	NA	NA	53.0	90.0

Table 36: The probability of a residue being predicted correctly or wrongly if its one or both neighbors are predicted correctly or wrongly was calculated. This was done to see if there is tendency for correct or wrong predictions in clusters.

difference in physico-chemical properties of the membrane core and the membrane/water interface regions [60]. Moreover, inherent errors in the theoretical and experimental methods in determining the membrane boundary could also be responsible for the higher rate of misclassifications at the membrane/water interface region. Further analysis was done to check for the tendency of correct or wrong predictions to occur together. The probability of true/false prediction was calculated given the case when one or both the neighbors were predicted correctly/wrongly. Table 36 shows that the probability of a residue to have a correct strand/non-strand assignment (Q_2) is 97.0%, when the regions of both the neighboring residues have also been correctly predicted. More interesting, when both the neighboring residues are assigned wrongly predicted regions, the target residues is also misclassified with a probability of 95.0%. Thus, as shown in table 36, there is a tendency of true/false predictions to occur in clusters.

5.4 Analysis of statewise prediction accuracy of TMBHMM

The data set consists of 7470 residues. These residues were labeled as belonging to one of the five states namely, non beta-strand transmembrane region (TMOther), periplasmic side (periSide), beta-strand residues buried and exposed in the membrane region (buriedCore and exposedCore) and the extracellular side (extraSide). The total number of residues in each state was found to be 899, 723, 1262, 1321 and 3265, respectively. As shown in table 37, the overall accuracy for the 5-state TMBHMM predictions is 76.18%. The prediction accuracy for 1thq and 1e54 at 42.18% and 40.79%, respectively is lower than the overall average prediction accuracy. Table 38 shows the misclassification of residues into a wrong state. The diagonal represents the correct predictions. As shown, the lowest prediction accuracy is obtained for the residues labeled as TMOther. It should be noted that all non beta-strand residues in the TM region are labeled as "TMOther" and the relative exposure of these residues to the space inside the beta barrel is not taken into account, which could lead to a higher misclassification rate. However, as shown in table 38, 22.03% TMOther residues are misclassified as belonging to the beta-strands in membrane region. When the buriedCore and exposedcore states are merged with the TM-other state, a total of 3062 out of 3482 (87.94%) residues belonging to the TM region are correctly classified. The overall average 3-state accuracy of classification when TMOther state is merged with buriedCore and exposeCore regions

Protein	correct	wrong	ACC[%]
1thq	62	85	42.18
1qfg	560	147	79.21
1kmo	542	119	82.00
1a0s	282	131	68.28
1xkw	547	108	83.51
2gsk	465	125	78.81
1qjp	111	26	81.02
2f1v	157	25	86.26
1xkh	518	169	75.40
2erv	116	34	77.33
1qd6	187	53	77.92
2mpr	315	106	74.82
1tly	185	66	73.71
1i78	234	63	78.79
1qj8	129	19	87.16
1fep	554	126	81.47
1e54	135	196	40.79
1t16	331	96	77.52
2j1n	261	85	75.43
Total	5691	1779	76.18

Table 37: The prediction accuracy per protein. The table shows the 5-state prediction accuracy. The prediction accuracy for the case when all TM states are merged into one (3-state prediction accuracy) is discussed in the main text below.

	TM-other	periSide	buriedCore	exposedCore	extraSide
TM-other	52.17	7.23	9.79	12.24	18.58
periSide	4.56	62.66	11.48	19.78	1.52
buriedCore	0.48	1.51	77.81	15.06	5.15
exposedCore	0.00	0.76	13.85	78.27	7.12
extraSide	8.27	0.40	3.71	3.31	84.32

Table 38: The state-wise per residue prediction accuracy. Diagonal shows the accurate predictions.

is 83.91%. The per amino acid prediction accuracy is shown in table 39. At 67.21%, the prediction accuracy for HIS residues is the lowest.

5.5 Comparison of TMBHMM structural topology predictions with PRED-TMBB

We compared the TMBHMM accuracy with an existing method for TMB strand prediction. We followed the work done by Bagos *et al.*, where they assessed the performance of different methods available for topology predictions of TMBs and have shown that HMM based methods perform better than Neural network or Support vector machine based methods [47]. It was also shown that amongst the HMM-based methods, PRED-TMBB [144] has the best performance [47]. Thus, we compared the accuracy of the TMBHMM method with the PRED-TMBB method. It is, however, to be noted that since none of these methods

Amino acid	correct	wrong	ACC[%]
CYS	6	1	85.71
GLN	230	91	71.65
ILE	219	69	76.04
SER	407	129	75.93
VAL	316	95	76.89
GLY	574	237	70.78
PRO	207	64	76.38
LYS	253	45	84.90
THR	425	119	78.13
PHE	208	114	64.60
ALA	421	134	75.86
HIS	82	40	67.21
MET	95	23	80.51
ASP	426	94	81.92
GLU	253	57	81.61
LEU	393	137	74.15
ARG	281	81	77.62
TRP	145	49	74.74
ASN	369	112	76.72
TYR	381	88	81.24

Table 39: The per amino acid residue prediction accuracy.

Method	Q_2	MCC	Core	SOV
TMBHMM (OPM)	84.0	0.67	84.0	0.91
PRED-TMBB (OPM)	85.0	0.69	82.0	0.90
TMBHMM (PDB)	76.0	0.53	69.0	0.86
PRED-TMBB (PDB)	76.0	0.54	67.0	0.84

Table 40: The performance of the method was compared with the performance of PRED-TMBB. Test data set was used for the comparison. The first two rows are the accuracy when strand regions were considered from OPM database while last two rows are the accuracy when strand regions from PDB database were considered.

including PRED-TMBB predicts the exposure status of residues in the membrane region, we can only compare the methods in terms of strand prediction. The non-redundant data set used by Bagos *et al.* [47] was employed to compare the prediction accuracies of the different methods. The results are shown in table 40. As shown in table 40, the Q_2, MCC, $Core$ and SOV prediction accuracies for the TMBHMM method are 0.84, 0.67, 0.84 and 0.91, respectively. The corresponding accuracies for the PRED-TMBB method are 0.85, 0.69, 0.82 and 0.90, respectively. Lower prediction accuracies were obtained when raw PDB structures were employed instead of the membrane oriented protein structures obtained from the OPM database [20]. However, in this case as well, the prediction accuracy of the TMBHMM method is at least as good as the PRED-TMBB method. Thus, the TMBHMM prediction accuracy is at least as good as the PRED-TMBB method. An advantage of TMBHMM method described here is that it can predict the exposure status of the residues predicted to be in the beta-strand in the TM region.

5.6 Web server

The TMBHMM web server for predicting the structural topology and exposure status of TMB residues is available under the TMBHMM tab at `http://service.bioinformatik.uni-saarland.de/tmx-site`. The web server takes input in the form of a FASTA sequence or a MSA and predicts the membrane spanning beta strands along with the exposure status of the residues in the membrane spanning region. As described in section 5.2, TMBHMM is trained using all the residues in the training data set and can thus predict the membrane, the periplasmic and the cytoplasmic regions of a TMB with high accuracy.

5.7 Conclusions

We presented TMBHMM, a computational method based on Hidden Markov Model, for the prediction of the structural topology of the TMBs. TMBHMM employs evolutionary information in the form of frequency profiles. The leave one out $Q2$ and $Q3$ accuracies for the TMBHMM method on the traing data set are 86.0% and 83.0%, respectively. We have shown that the accuracy of TMBHMM is at least as high as the best available method PRED-TMBB. In addition, the TMBHMM method also predicts the exposure status of the residues in the beta strands at the membrane spanning region. To the best of our knowledge, this is the first HMM based method for prediction of the structural topology of TMBs along with the exposure status of the transmembrane residues. TMBHMM method has been implemented as a web service that take a protein sequence of a multiple sequence alignment as input and predicts the structural topology for the given sequence along with the exposure status of the residues in the predicted membrane spanning beta strands. As part of its output, TMBHMM web server generates colored snake-plots of the predicted structural topology annotated with the exposure status of the transmembrane residues. In future, TMBHMM method can be extended for the discrimination of TMBs from proteomic data.

5.8 Methods

5.8.1 Accuracy measures

Several accuracy measures have been employed for evaluating the performance of TMBHMM. Q_2 is defined as the two state accuracy measure where the amino acid residues are classified as either belonging to a strand or a non-strand region. Similarly, Q_3 is the prediction accuracy when the amino acid residues are classified into three regions namely, periplasmic, extracellular or the membrane region. We have also used Matthews correlation coefficient (MCC) [142] and Segment overlap measure (SOV) for beta strands as defined by Zemla *et. al* [149] as accuracy measures. Core accuracy is defined as the percentage of residues that are correctly predicted to be in the membrane region by TMBHMM. The exposure accuracy gives the percentage of correctly predicted membrane residues with correct exposure status.

Training and Decoding

The initial parameters for the HMM were generated from the training data set using the Baum-welch algorithm [150,151]. For decoding, the Viterbi algorithm [152] was employed. The "jhmm" java module, which is freely available at `http://code.google.com/p/jhmm/` was used to develop the hidden markov model for the TMBHMM method.

5.8.2 Training and test data sets

As described in section 3.10.1, initially a non-redundant set of TMB 3D structures was obtained from the OPM database such that the sequence identity was $< 30\%$. From this set, some sequences were further removed based on the criteria previously described by Park *et al.* [70]. Briefly, for a given query sequence, a maximum of 1000 homologous sequences were retrieved from the non-redundant database using BLAST. Initial MSAs were built using ClustalW. Sequences that are more than 80% identical to the query sequence were removed. The remaining sequences were realigned using ClustalW to yield a final MSA, which was used to obtain profiles. The training data set thus obtained consists of 19 non-redundant TMBs. Furter, unlike the PDB database, the OPM database provides membrane boundaries which were useful in classifying the amino acid residues [20]. The residues within these boundaries were classified as membrane residues and residues outside were classified as non-membrane residues. Thus the final 3D structures were obtained from the OPM database. The test data set was taken from Bagos *et al.* [47]. It includes a set of 20 non-redundant proteins. We have used this data set for testing our method (TMBHMM) and comparing results with the existing PRED-TMBB web-server [144].

5.8.3 Computation of rSASA

Residues in the membrane spanning region can be classified as either buried in the barrel interior or as exposed to the lipid membrane. The residues have been classified as being buried or exposed based on their relative solvent-accessible surface area (rSASA) value. The same method was used by Park *et al.* [70]. The probe radius of $2.2\mathring{A}$ for calculating the rSASA value was also taken from Park *et al.* [70]. The two faces of the transmembrane region are capped to prevent the probe from entering the core and hence, misclassifying buried residue as exposed. The VOLBL program suite [113, 114] was employed for actual computation of SASA values.

5.8.4 Computation of frequency profile

A frequency profile is a vector of size 20, which holds the frequency of 20 amino acids in a MSA at the position of a particular residue. For generating the frequency profiles, first of all a blast search was performed using the blastp program available from the ncbi website. The blast search was performed against a non-redundant protein sequence database with default parameters. A total of up to 1000 homologous sequences obtained after blast were taken into consideration while performing the MSA. The lower limit of homologous sequences for generating MSA is 20 while the upper limit is 1000. Then we generated a multiple

sequence alignment using ClustalW for each training sequence and their homologues. Sequences having less than 25% similarity with the query sequence were removed. Also, sequences with length shorter than 80% of the query sequence were removed. In the final step, AL2CO [112] was used to generate a frequency profile from the MSA using a modified method of Henikoff and Henikoff [153].

6 Conclusions and outlook

The prediction of the structural features and topology of transmembrane proteins along with their identification from proteomic data has been an active field of research for the last decades. Transmembrane proteins can be mainly classified as helical membrane proteins and transmembrane beta barrel proteins. In the recent years, with an improvement in the experimental structural determination methods, many novel transmembrane protein structures have been determined, however, the 3D structures of transmembrane proteins are still highly under-represented as compared to the soluble proteins. Transmembrane proteins are known to play a major role in the normal functioning of the cell and many transmembrane proteins act as a drug target and are hence of utmost importance to the pharmaceutical industry. Given the important roles played by transmembrane proteins, it is imperative to develop novel computational methods to elucidate the structure and function of these proteins.

Most computational methods in the realm of transmembrane protein structure prediction deal with helical membrane proteins. Moreover, several structural rules and physico-chemical features have been described in the literature, which has made it possible to identify and predict the structural topology of these proteins with a high accuracy. On the other hand, only a few computational methods for the identification and prediction of the structural features and topology of the putative transmembrane beta barrel proteins are available at present. In general, transmembrane beta barrel proteins are known to have a simple structure that can be represented by a simple grammar and comprises of anti-parallel beta strands, long extracellular loops and short periplasmic loops. In spite of the relatively simpler structure than the helical membrane proteins, the identification of transmembrane beta barrel proteins has proven to be more difficult mainly because of the lack of a long stretch of hydrophobic residues in the membrane region. The less hydrophobic exterior and the fact that the soluble beta barrels share similar structural features with transmembrane ones has made the task of identifying transmembrane beta barrel proteins more challenging. In the recent years, some computational methods for the identification of transmembrane beta barrel have emerged with reasonable accuracy but they still do not provide a wide range of structural feature predictions as in the field of helical membrane proteins.

Various rules regarding the structural topology and geometric constraints have been proposed for the transmembrane proteins. These rules have been employed in the prediction methods to predict putative transmembrane beta barrel proteins. On such rule is the dyad repeat pattern of residues in beta strands, where alternate residues are known to point either towards the lipid membrane or the barrel core. The aim of this study is to develop novel methods of the prediction of structural features of transmembrane beta barrel proteins. One such feature is the exposure status of transmembrane beta barrel residues. The exposure status conveys if a particular residue is exposed to the bilayer or hidden in the proteins structure. We show that it is possible for out-pointing residues to be hidden from the membrane due to the shielding by neighbouring residues. Furthermore, we also show that it is also possible for in-pointing residues to be slightly exposed to the bilayer. We first determine the propensity of transmembrane beta barrel residues to be exposed to the bilayer. The propensities are derived such that the exposure status of the amino acid residues is maximally

correlated with the frequency profile of a given amino acid residue. To the best of our knowledge, the propensity of transmembrane beta barrel residues to be exposed to the bilayer has never been dealt before. Furthermore, in this study we differentiate between the residues in the core region of the membrane from the lipid-water interface regions and thus derive separate propensity scales for the residues in the membrane region of transmembrane beta barrel proteins. We then compare the derived scales with similar scales for helical membrane proteins and discuss the differences and similarities between the propensities of amino acid residues to be exposed to the bilayer in the two transmembrane proteins. The derived scales for transmembrane beta barrel proteins are then compared with known physico-chemical scales from the literature. The differences in the propensities of residues in the membrane core region to be exposed to the bilayer are then discuss for a separate non-redundant data set comprising of only oligomeric transmembrane beta barrel proteins. To show the practical utility of the derived scales, a preliminary prediction method based on ridge regression is proposed to predict the exposure status of transmembrane beta barrel residues from sequence.

Further, we have developed a more sophisticated two-stage sliding window method called BTMX to predict the exposure status of transmembrane beta barrel residues. BTMX employs positional specific scoring matrices obtained from the multiple sequence alignment of a given protein sequence and predicts the exposure status of the given protein sequence. The BTMX method has also been made available as a web server, which generates colored snake-plots representing beta strands annotated with the predicted exposure status of their residues. The BTMX method has been compared with another method from the literature and it has been shown that BTMX clearly outperforms the other method in terms of higher prediction accuracy. Moreover, BTMX generates a confidence score for the made predictions, which is especially advantageous for the practical utility of the predictions.

To the best of our knowledge, the prediction of the oligomeric state of a given transmembrane proteins is also an open problem in the field of transmembrane beta barrel proteins. Also no physico-chemical properties have so far been identified to distinguish beta strands at the oligomeric interfaces from the rest of the beta strands. To this end, we have analysed relevant physico-chemical properties for the out-pointing, exposed residues and identified several physico-chemical properties that can be used to develop a oligomeric strand identification method in the future.

The prediction of the membrane spanning beta strands from the given protein sequence is another problem that has been addressed in this study. We have established a prediction method based on a hidden markov model to predict structural topology of putative transmembrane beta barrel proteins. The method named TMBHMM is the most comprehensive computational method for the prediction of the structural topology of the given putative transmembrane beta barrel sequence in terms of the number of states predicted. TMBHMM, like BTMX also predicts the exposure status of the residues predicted to be in the membrane spanning region. The TMBHMM web server has also been made available.

In the future, the information about the exposure status of the residues in the transmembrane beta strands can be used to identify putative transmembrane beta barrels from proteomic data. Furthermore the exposure status can also be

used to determine novel structural motifs in transmembrane beta barrel proteins that could aid in their identification from proteomic data. The exposure status prediction can also be extended to predict the pore region of transmembrane beta barrel proteins from sequence, which could in turn be used in the function prediction of putative transmembrane beta barrels.

References

[1] C.P. Chen and B. Rost. State-of-the-art in membrane protein prediction. *Applied bioinformatics*, 1(1):21–36, 2002.

[2] M.S. Weiss, T. Wacker, J. Weckesser, W. Welte, and G.E. Schulz. The three-dimensional structure of porin from Rhodobacter capsulatus at 3 A resolution. *FEBS Lett*, 267(2):268–72, 1990.

[3] M. Bannwarth and G.E. Schulz. The expression of outer membrane proteins for crystallization. *BBA-Biomembranes*, 1610(1):37–45, 2003.

[4] D. Rapaport. Finding the right organelle. Targeting signals in mitochondrial outer-membrane proteins. *EMBO Rep*, 4(10):948–952, 2003.

[5] N. Pfanner, N. Wiedemann, C. Meisinger, and T. Lithgow. Assembling the mitochondrial outer membrane. *Nature Structural & Molecular Biology*, 11(11):1044–1048, 2004.

[6] E. Schleiff and J. Soll. Membrane protein insertion: mixing eukaryotic and prokaryotic concepts. *EMBO Rep*, 6(11):1023–1027, 2005.

[7] C. Meisinger, N. Wiedemann, M. Rissler, A. Strub, D. Milenkovic, B. Schonfisch, H. Muller, V. Kozjak, and N. Pfanner. Mitochondrial Protein Sorting: Differentiation of beta-barrel assembly by Tom7-mediated segregation of Mdm10. *Journal of Biological Chemistry*, 281(32):22819–22826, 2006.

[8] J. Tommassen, M. Struyvé, and H. Cock. Export and assembly of bacterial outer membrane proteins. *Antonie van Leeuwenhoek*, 61(2):81–85, 1992.

[9] L.K. Tamm, H. Hong, and B. Liang. Folding and assembly of β-barrel membrane proteins. *BBA-Biomembranes*, 1666(1-2):250–263, 2004.

[10] E.F. Eppens, N. Nouwen, and J. Tommassen. Folding of a bacterial outer membrane protein during passage through the periplasm. *The EMBO Journal*, 16:4295–4301, 1997.

[11] C.J. Lazdunski and H. Benedetti. Insertion and translocation of proteins into and though membranes. *FEBS letters*, 268(2):408–414, 1990.

[12] K. Verner and G. Schatz. Protein translocation across membranes. *Science*, 241(4871):1307–1313, 1988.

[13] G.E. Schulz. The structure of bacterial outer membrane proteins. *BBA-Biomembranes*, 1565(2):308–317, 2002.

[14] I. Broutin, H. Benabdelhak, X. Moreel, M.B. Lascombe, D. Lerouge, and A. Ducruix. Expression, purification, crystallization and preliminary X-ray studies of the outer membrane efflux proteins OprM and OprN from Pseudomonas aeruginosa. *Acta Crystallographica Section F: Structural Biology and Crystallization Communications*, 61(3):315–318, 2005.

[15] E. Durham, B. Dorr, N. Woetzel, R. Staritzbichler, and J. Meiler. Solvent accessible surface area approximations for rapid and accurate protein structure prediction. *Journal of molecular modeling*, 15(9):1093–1108, 2009.

[16] MS Sansom and ID Kerr. Transbilayer pores formed by beta-barrels: molecular modeling of pore structures and properties. *Biophysical Journal*, 69(4):1334–1343, 1995.

[17] Z. Yuan, F. Zhang, M.J. Davis, M. Boden, and R.D. Teasdale. Predicting the solvent accessibility of transmembrane residues from protein sequence. *J Proteome Res*, 5(5):1063–1070, 2006.

[18] N. Ruiz, D. Kahne, T.J. Silhavy, et al. Advances in understanding bacterial outer-membrane biogenesis. *Nature Reviews: Microbiology*, 4(1):57–66, 2006.

[19] G. Van Meer, D.R. Voelker, and G.W. Feigenson. Membrane lipids: where they are and how they behave. *Nature Reviews Molecular Cell Biology*, 9(2):112–124, 2008.

[20] M.A. Lomize, A.L. Lomize, I.D. Pogozheva, and H.I. Mosberg. OPM: orientations of proteins in membranes database. *Bioinformatics*, 22(5):623–625, 2006.

[21] E. Wallin and G. Von Heijne. Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. *Protein Science: A Publication of the Protein Society*, 7(4):1029, 1998.

[22] P.L. Martelli, P. Fariselli, A. Krogh, and R. Casadio. A sequence-profile-based HMM for predicting and discriminating *beta*-Barrel membrane proteins. *Bioinformatics*, 18(Suppl 1):S46, 2002.

[23] G. von Heijne. Recent advances in the understanding of membrane protein assembly and structure. *Quarterly Reviews of Biophysics*, 32(04):285–307, 1999.

[24] G. Von Heijne. Membrane protein structure prediction: hydrophobicity analysis and the positive-inside rule. *Journal of Molecular Biology*, 225(2):487–494, 1992.

[25] P.G. Bagos, T.D. Liakopoulos, I.C. Spyropoulos, and S.J. Hamodrakas. A Hidden Markov Model method, capable of predicting and discriminating β-barrel outer membrane proteins. *BMC bioinformatics*, 5(1):29, 2004.

[26] W.C. Wimley. The versatile β-barrel membrane protein. *Current Opinion in Structural Biology*, 13(4):404–411, 2003.

[27] G.E. Schulz. β-Barrel membrane proteins. *Current Opinion in Structural Biology*, 10(4):443–447, 2000.

[28] M.W. Gray, G. Burger, and B.F. Lang. Mitochondrial Evolution. *Science*, 283(5407):1476–1481, 1999.

[29] T. Vellai. A New Aspect to the Origin and Evolution of Eukaryotes. *Journal of Molecular Evolution*, 46(5):499–507, 1998.

[30] D. Moreira, H. Le Guyader, and H. Philippe. The origin of red algae and the evolution of chloroplasts. *Nature*, 405(6782):69–72, 2000.

[31] T. Cavalier-Smith. Membrane heredity and early chloroplast evolution. *Trends in Plant Science*, 5(4):174–182, 2000.

[32] E. Blachly-Dyson, S. Peng, M. Colombini, and M. Forte. Selectivity changes in site-directed mutants of the VDAC ion channel: structural implications. *Science*, 247(4947):1233–1236, 1990.

[33] K. Fischer, A. Weber, S. Brink, B. Arbinger, D. Schunemann, S. Borchert, H.W. Heldt, B. Popp, R. Benz, and T.A. Link. Porins from plants. Molecular cloning and functional characterization of two new members of the porin family. *Journal of Biological Chemistry*, 269(41):25754–25760, 1994.

[34] G. Báthori, I. Parolini, I. Szabó, F. Tombola, A. Messina, M. Oliva, M. Sargiacomo, V. De Pinto, and M. Zoratti. Extramitochondrial Porin: Facts and Hypotheses. *Journal of Bioenergetics and Biomembranes*, 32(1):79–89, 2000.

[35] R. Benz. Permeation of hydrophilic solutes through mitochondrial outer membranes: review on mitochondrial porins. *Biochim Biophys Acta*, 1197(2):167–196, 1994.

[36] R. Koebnik, K.P. Locher, and P. Van Gelder. Structure and function of bacterial outer membrane proteins: barrels in a nutshell. *Molecular Microbiology*, 37(2):239–253, 2000.

[37] A. Pautsch and G.E. Schulz. Structure of the outer membrane protein A transmembrane domain. *Nature Structural Biology*, 5:1013–1017, 1998.

[38] G.E. Schulz. Porins: general to specific, native to engineered passive pores. *Current Opinion in Structural Biology*, 6(4):485–490, 1996.

[39] R. Jackups and J. Liang. Interstrand Pairing Patterns in β-Barrel Membrane Proteins: The Positive-outside Rule, Aromatic Rescue, and Strand Registration Prediction. *Journal of Molecular Biology*, 354(4):979–993, 2005.

[40] S. Galdiero, M. Galdiero, and C. Pedone. β-Barrel Membrane Bacterial Proteins: Structure, Function, Assembly and Interaction with Lipids. *Current Protein and Peptide Science*, 8(1):63–82, 2007.

[41] R. Pajón, D. Yero, A. Lage, A. Llanes, and C.J. Borroto. Computational identification of beta-barrel outer-membrane proteins in Mycobacterium tuberculosis predicted proteomes as putative vaccine candidates. *Tuberculosis*, 86(3-4):290–302, 2006.

[42] M.H. Saier, Jr. Families of Proteins Forming Transmembrane Channels. *Journal of Membrane Biology*, 175(3):165–180, 2000.

[43] R.J.C. Gilbert. Pore-forming toxins. *Cellular and Molecular Life Sciences (CMLS)*, 59(5):832–844, 2002.

[44] D.M. Walther, D. Rapaport, and J. Tommassen. Biogenesis of β-barrel membrane proteins in bacteria and eukaryotes: evolutionary conservation and divergence. *Cellular and Molecular Life Sciences*, 66(17):2789–2804, 2009.

[45] J.E.W. Meyer, M. Hofnung, and G.E. Schulz. Structure of maltoporin from Salmonella typhimurium ligated with a nitrophenyl-maltotrioside. *Journal of molecular biology*, 266(4):761–775, 1997.

[46] H.R. Bigelow, D.S. Petrey, J. Liu, D. Przybylski, and B. Rost. Predicting transmembrane beta-barrels in proteomes. *Nucleic acids research*, 32(8):2566, 2004.

[47] P.G. Bagos, T.D. Liakopoulos, and S.J. Hamodrakas. Evaluation of methods for predicting the topology of β-barrel outer membrane proteins and a consensus prediction method. *BMC Bioinformatics*, 6(1):0–7, 2005.

[48] A.E. Rizzitello, J.R. Harper, and T.J. Silhavy. Genetic evidence for parallel pathways of chaperone activity in the periplasm of Escherichia coli. *Journal of Bacteriology*, 183(23):6794, 2001.

[49] R. Nowak. Basic Elements of Statistical Decision Theory and Statistical Learning Theory. 2009.

[50] J. Friedman, T. Hastie, and R. Tibshirani. The elements of statistical learning, 2001.

[51] C.M. Bishop et al. *Pattern recognition and machine learning*. Springer New York:, 2006.

[52] A.J. Smola and B. Schölkopf. A tutorial on support vector regression. *Statistics and Computing*, 14(3):199–222, 2004.

[53] K. Hechenbichler and KP Schliep. Weighted k-nearest-neighbor techniques and ordinal classification. Technical report, Citeseer, 2004.

[54] L.R. Rabiner. A tutorial on hidden Markov models and selected applications inspeech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[55] J.U. Bowie. Solving the membrane protein folding problem. *Nature*, 438(7068):581–589, 2005.

[56] B. Rost and C. Sander. Third generation prediction of secondary structures. *Methods in molecular biology (Clifton, NJ)*, 143:71–96, 2000.

[57] B. Rost, G. Yachdav, and J. Liu. The predictprotein server. *Nucleic acids research*, 32(Web Server Issue):W321, 2004.

[58] D. Donnelly, J.P. Overington, S.V. Ruffle, J.H.A. Nugent, and T.L. Blundell. Modeling {alpha}-helical transmembrane domains: The calculation and use of substitution tables for lipid-facing residues. *Protein Science*, 2(1):55–70, 1993.

[59] Y. Park and V. Helms. How Strongly do Sequence Conservation Patterns and Empirical Scales Correlate with Exposure Patterns of Transmembrane. *Biopolymers*, 83:389–399, 2006.

[60] E. Granseth, G. von Heijne, and A. Elofsson. A Study of the Membrane–Water Interface Region of Membrane Proteins. *Journal of Molecular Biology*, 346(1):377–385, 2005.

[61] T. Beuming and H. Weinstein. A knowledge-based scale for the analysis and prediction of buried and exposed faces of transmembrane domain proteins, 2004.

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

[63] L. Adamian and J. Liang. Prediction of transmembrane helix orientation in polytopic membrane proteins. *BMC Structural Biology*, 6(1):13, 2006.

[64] A. Randall, J. Cheng, M. Sweredoski, and P. Baldi. TMBpro: secondary structure,{beta}-contact and tertiary structure prediction of transmembrane {beta}-barrel proteins. *Bioinformatics*, 24(4):513–520, 2008.

[65] J. Waldispuhl, B. Berger, P. Clote, and J.M. Steyaert. transFold: a web server for predicting the structure and residue contacts of transmembrane beta-barrels. *Nucleic Acids Research*, 34(Web Server issue):W189, 2006.

[66] W.C. Wimley. Toward genomic identification of ß-barrel membrane proteins: Composition and architecture of known structures. *Protein Science*, 11:301–312, 2002.

[67] K. Seshadri, R. Garemyr, E. Wallin, G. von Heijne, and A. Elofsson. Architecture of {beta}-barrel membrane proteins: Analysis of trimeric porins. *Protein Science*, 7(9):2026, 1998.

[68] M.B. Ulmschneider and M.S.P. Sansom. Amino acid distributions in integral membrane protein structures. *BBA-Biomembranes*, 1512(1):1–14, 2001.

[69] J. Liang, L. Adamian, and R. Jackups. The membrane–water interface region of membrane proteins: structural bias and the anti-snorkeling effect. *Trends in Biochemical Sciences*, 30(7):355–357, 2005.

[70] Y. Park and V. Helms. On the derivation of propensity scales for predicting exposed transmembrane residues of helical membrane proteins. *Bioinformatics*, 23(6):701–708, 2007.

[71] S. Kawashima, P. Pokarowski, M. Pokarowska, A. Kolinski, T. Katayama, and M. Kanehisa. AAindex: amino acid index database, progress report 2008. *Nucleic Acids Research*, 36(Database issue):D202, 2008.

[72] J.H. Kleinschmidt. Membrane protein folding on the example of outer membrane protein A of Escherichia coli. *Cellular and Molecular Life Sciences (CMLS)*, 60(8):1547–1558, 2003.

[73] S.H. White and W.C. Wimley. Membrane protein folding and stability: Physical Principles. *Annual Review of Biophysics and Biomolecular Structure*, 28(1):319–365, 1999.

[74] G. von Heijne. Principles of membrane protein assembly and structure. *Progress in Biophysics and Molecular Biology*, 66(2):113–139, 1996.

[75] K. Gunasekaran, HA Nagarajaram, C. Ramakrishnan, and P. Balaram. Stereochemical punctuation marks in protein structures: glycine and proline containing helix stop signals. *Journal of Molecular Biology*, 275(5):917–932, 1998.

[76] S. Konishi, T. Nishio, and T. Shimizu. Inner Residues in the Transmembrane Helix Bundle are More Conservative. *Genome Informatics Series*, pages 549–550, 2003.

[77] A.K. Chamberlain, Y. Lee, S. Kim, and J.U. Bowie. Snorkeling preferences foster an amino acid composition bias in transmembrane helices. *Journal of molecular biology*, 339(2):471–479, 2004.

[78] M. Gimpelev, L.R. Forrest, D. Murray, and B. Honig. Helical Packing Patterns in Membrane and Soluble Proteins. *Biophysical Journal*, 87(6):4075–4086, 2004.

[79] S. Miyazawa and R.L. Jernigan. Self-consistent estimation of inter-residue protein contact energies based on an equilibrium mixture approximation of residues. *Proteins Structure Function and Genetics*, 34(1):49–68, 1999.

[80] A. Radzicka and R. Wolfenden. Comparing the polarities of the amino acids: side-chain distribution coefficients between the vapor phase, cyclohexane, 1-octanol, and neutral aqueous solution. *Biochemistry*, 27(5):1664–1670, 1988.

[81] D. Eisenberg and A.D. McLachlan. Solvation energy in protein folding and binding. 1986.

[82] V. Pliska, M. Schmidt, and J.L. Fauchère. Partition coefficients of amino acids and hydrophobic parameters [pi] of their side-chains as measured by thin-layer chromatography. *Journal of Chromatography A*, 216:79–92, 1981.

[83] Y. Nozaki and C. Tanford. The solubility of amino acids and two glycine peptides in aqueous ethanol and dioxane solutions. Establishment of a hydrophobicity scale. *Journal of Biological Chemistry*, 246(7):2211, 1971.

[84] B. Robson and D.J. Osguthorpe. Refined models for computer simulation of protein folding:: Applications to the study of conserved secondary structure and flexible hinge points during the folding of pancreatic trypsin inhibitor. *Journal of Molecular Biology*, 132(1):19–51, 1979.

[85] H.B. Bull and K. Breese. Surface tension of amino acid solutions: a hydrophobicity scale of the amino acid residues. *Archives of Biochemistry and Biophysics*, 161(2):665–670, 1974.

[86] R. Grantham. Amino acid difference formula to help explain protein evolution. *Science*, 185:862–864, 1974.

[87] TP Hopp and KR Woods. Prediction of protein antigenic determinants from amino acid sequences. *Proceedings of the National Academy of Sciences of the United States of America*, 78(6):3824, 1981.

[88] MA Roseman. Hydrophilicity of polar amino acid side-chains is markedly reduced by flanking peptide bonds. *Journal of molecular biology*, 200(3):513, 1988.

[89] PK Ponnuswamy. Hydrophobic characteristics of folded proteins. *Progress in Biophysics and Molecular Biology*, 59(1):57–103, 1993.

[90] H. Cid, M. Bunster, M. Canales, and F. Gazitua. Hydrophobicity and structural classes in proteins. *Protein engineering*, 5(5):373–375, 1992.

[91] D.D. Jones. Amino acid properties and side-chain orientation in proteins: a cross correlation approach. *Journal of theoretical biology*, 50(1):167–183, 1975.

[92] D. Eisenberg, R.M. Weiss, T.C. Terwilliger, and W. Wilcox. Hydrophobic moments and protein structure. In *Faraday Symposia of the Chemical Society*, volume 17, pages 109–120. Royal Society of Chemistry, 1982.

[93] S.D. Black and D.R. Mould. Development of hydrophobicity parameters to analyze proteins which bear post-or cotranslational modifications. *Analytical biochemistry*, 193(1):72–82, 1991.

[94] P. Manavalan and PK Ponnuswamy. Hydrophobic character of amino acid residues in globular proteins. *Nature*, 275:673–674, 1978.

[95] J.L. Fauchere and V. Pliska. Hydrophobic parameters pi of amino-acid side chains from the partitioning of N-acetyl-amino-acid amides. *Eur. J. Med. Chem*, 18(4):369–375, 1983.

[96] J. Janin. Surface and inside volumes in globular proteins. *Nature*, 277:491–492, 1979.

[97] JL Fauchere, M. Charton, L.B. Kier, A. Verloop, and V. Pliska. Amino acid side chain parameters for correlation studies in biology and pharmacology. *Int J Pept Protein Res*, 32(4):269–278, 1988.

[98] JM Zimmerman, N. Eliezer, and R. Simha. The characterization of amino acid sequences in proteins by statistical methods. *Journal of Theoretical Biology*, 21(2):170, 1968.

[99] G. Meng, R. Fronzes, V. Chandran, H. Remaut, and G. Waksman. Protein oligomerization in the bacterial outer membrane (Review). *Molecular membrane biology*, 26(3):136–145, 2009.

[100] H. Ponstingl, T. Kabir, D. Gorse, and J.M. Thornton. Morphological aspects of oligomeric protein structures. *Progress in Biophysics and Molecular Biology*, 89(1):9–35, 2005.

[101] M.H. Ali and B. Imperiali. Protein oligomerization: how and why. *Bioorganic and Medicinal Chemistry*, 13(17):5013–5020, 2005.

[102] S. Jones and J.M. Thornton. Analysis of protein-protein interaction sites using surface patches. *Journal of Molecular Biology*, 272(1):121–132, 1997.

[103] S. Jones and J.M. Thornton. Prediction of protein-protein interaction sites using patch analysis. *Journal of Molecular Biology*, 272(1):133–143, 1997.

[104] DM Engelman, TA Steitz, and A. Goldman. Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. *Annual Review of Biophysics and Biophysical Chemistry*, 15(1):321–353, 1986.

[105] J. Janin, S. WodakMichael, and B. Maigret. Conformation of amino acid side-chains in proteins. *Journal of molecular biology*, 125(3):357–386, 1978.

[106] C.W. Hsu and C.J. Lin. A comparison of methods for multiclass support vector machines. *IEEE transactions on Neural Networks*, 13(2):415–425, 2002.

[107] R.D.C. Team. R: A language and environment for statistical computing. *Vienna, Austria: R Foundation for Statistical Computing*, 2006.

[108] X.M. Zhao, X. Li, L. Chen, and K. Aihara. Protein classification with imbalanced data. *Proteins: Structure, Function, and Bioinformatics*, 70(4):1125–1132, 2007.

[109] M. Bogdanov and W. Dowhan. Lipid-assisted protein folding. *Journal of Biological Chemistry*, 274(52):36827, 1999.

[110] I.K. Valavanis, P.G. Bagos, and I.Z. Emiris. β-Barrel transmembrane proteins: Geometric modelling, detection of transmembrane region, and structural properties. *Computational Biology and Chemistry*, 30(6):416–424, 2006.

[111] G.E. Tusnady, Z. Dosztanyi, and I. Simon. PDB_TM: selection and membrane localization of transmembrane proteins in the protein data bank. *Nucleic acids research*, 33(Database Issue):D275, 2005.

[112] J. Pei and N.V. Grishin. AL2CO: calculation of positional conservation in a protein sequence alignment. *Bioinformatics*, 17(8):700–712, 2001.

[113] H. Edelsbrunner, M. Facello, P. Fu, and J. Liang. Measuring proteins and voids in proteins. In *Proceedings of the Hawaii International Conference on System Sciences*, volume 28, pages 256–264. IEEE Institute of Electrical and Electronics, 1995.

[114] H. Edelsbrunner. The union of balls and its dual shape. *Discrete and Computational Geometry*, 13(1):415–440, 1995.

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

[116] A. Karatzoglou, D. Meyer, and K. Hornik. Support vector machines in R. *Journal of Statistical Software*, 15(9):1–28, 2006.

[117] T. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967.

[118] K. Hechenbichler and K. Schliep. Weighted k-nearest-neighbor techniques and ordinal classification. *Discussion Paper 399, SFB 386*, 2006.

[119] K. Schliep and K. Hechenbichler. kknn: Weighted k-Nearest Neighbors, 2007. *R package version*, pages 1–0.

[120] N.K. Natt, H. Kaur, and G.P. Raghava. Prediction of transmembrane regions of beta-barrel proteins using ANN-and SVM-based methods. *Proteins*, 56(1):11–18, 2004.

[121] H. Naveed, R. Jackups, and J. Liang. Predicting weakly stable regions, oligomerization state, and protein–protein interfaces in transmembrane domains of outer membrane proteins. *Proc. Natl. Acad. Sci. USA*, 106(31):12735–12740, 2009.

[122] J. Waldispühl, B. Berger, P. Clote, and J.M. Steyaert. Predicting transmembrane β-barrels and interstrand residue interactions from sequence. *Proteins*, 65(1):61–74, 2006.

[123] L.K. Tamm, A. Arora, and J.H. Kleinschmidt. Structure and assembly of beta-barrel membrane proteins. *Journal of Biological Chemistry*, 276(35):32399–32402, 2001.

[124] M. Gribskov, AD McLachlan, and D. Eisenberg. Profile analysis: detection of distantly related proteins. *Proc. Natl. Acad. Sci. USA*, 84(13):4355–4358, 1987.

[125] Q. Dong, X. Wang, L. Lin, and Y. Guan. Exploiting residue-level and profile-level interface propensities for usage in binding sites prediction of proteins. *BMC Bioinformatics*, 8(1):147, 2007.

[126] Y.Y. Ou, M.M. Gromiha, S.A. Chen, and M. Suwa. TMBETADISC-RBF: Discrimination of β-barrel membrane proteins using RBF networks and PSSM profiles. *Computational Biology and Chemistry*, 32(3):227–231, 2008.

[127] J. Kyte and R.F. Doolittle. A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology*, 157(1):105–132, 1982.

[128] V.N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 2000.

[129] J.S. Merkel and L. Regan. Aromatic rescue of glycine in β sheets. *Folding and Design*, 3(6):449–456, 1998.

[130] E.M. Hearn, D.R. Patel, B.W. Lepore, M. Indic, and B. Van den Berg. Transmembrane passage of hydrophobic compounds through a protein channel wall. *Nature*, 458(7236):367–370, 2009.

[131] A.H. Elcock and J.A. McCammon. Identification of protein oligomerization states by analysis of interface conservation. *Proc. Natl. Acad. Sci. USA*, 98(6):2990, 2001.

[132] D.R. Caffrey, S. Somaroo, J.D. Hughes, J. Mintseris, and E.S. Huang. Are protein–protein interfaces more conserved in sequence than the rest of the protein surface? *Protein Science*, 13(1):190, 2004.

[133] TD Schneider, GD Stormo, L. Gold, and A. Ehrenfeucht. Information content of binding sites on nucleotide sequences. *Journal of Molecular Biology*, 188(3):415, 1986.

[134] DE Goldsack and RC Chalifoux. Contribution of the free energy of mixing of hydrophobic side chains to the stability of the tertiary structure of proteins. *Journal of Theoretical Biology*, 39(3):645, 1973.

[135] V. Muñoz and L. Serrano. Intrinsic Secondary Structure Propensities of the Amino Acids, Using Statistical ϕ-ψ Matrices: Comparison With Experimental Scales. *Proteins: Structure, Function, and Genetics*, 20:301–311, 1994.

[136] T.S. Niwa and A. Ogino. Multiple regression analysis of the beta-sheet propensity of amino acids. *Journal of Molecular Structure: THEOCHEM*, 419(1-3):155–160, 1997.

[137] H. Bigelow and B. Rost. PROFtmb: a web server for predicting bacterial transmembrane beta barrel proteins. *Nucleic Acids Research*, 34(Web Server issue):W186, 2006.

[138] M.M. Gromiha, Y. Yabuki, and M. Suwa. TMB Finding Pipeline: Novel Approach for Detecting-Barrel Membrane Proteins in Genomic Sequences. *J. Chem. Inf. Model*, 47(6):2456–2461, 2007.

[139] M.F. Sanner, A.J. Olson, and J.C. Spehner. Reduced surface: an efficient way to compute molecular surfaces. *Peptide Science*, 38(3):305–320, 1996.

[140] D.T. Jones. Protein secondary structure prediction based on position-specific scoring matrices. *Journal of Molecular Biology*, 292(2):195–202, 1999.

[141] M.N. Nguyen and J.C. Rajapakse. Two-stage support vector regression approach for predicting accessible surface areas of amino acids. *Proteins: Structure, Function, and Bioinformatics*, 63:542–550, 2006.

[142] P. Baldi, S. Brunak, Y. Chauvin, C.A.F. Andersen, and H. Nielsen. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, 16(5):412, 2000.

[143] T.C. Freeman Jr and W.C. Wimley. A Highly Accurate Statistical Approach for the Prediction of Transmembrane {beta}-Barrels. *Bioinformatics*, 2010.

[144] P.G. Bagos, T.D. Liakopoulos, I.C. Spyropoulos, and S.J. Hamodrakas. PRED-TMBB: a web server for predicting the topology of {beta}-barrel outer membrane proteins. *Nucleic acids research*, 32(Web Server Issue):W400, 2004.

[145] B. Rost and C. Sander. Conservation and prediction of solvent accessibility in protein families. *Proteins: Structure, Function, and Bioinformatics*, 20:216–226, 1994.

[146] G.E. Schulz. Porins: general to specific, native to engineered passive pores. *Current Opinion in Structural Biology*, 6(4):485–490, 1996.

[147] H. Bayley. Designed membrane channels and pores. *Current opinion in biotechnology*, 10(1):94–103, 1999.

[148] O. Mirus and E. Schleiff. Prediction of-barrel membrane proteins by searching for restricted domains. *BMC Bioinformatics*, 6:254, 2005.

[149] A. Zemla, C. Venclovas, K. Fidelis, and B. Rost. A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment. *Proteins Structure Function and Genetics*, 34(2):220–223, 1999.

[150] L.E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics*, 41(1):164–171, 1970.

[151] L.R. Welch. Hidden markov models and the baum-welch algorithm. *IEEE Information Theory Society Newsletter*, 53(4):1–10, 2003.

[152] G.D. Forney. The viterbi algorithm. *proc. IEEE*, 61(3):268–278, 1973.

[153] S. Henikoff and J.G. Henikoff. Position-based sequence weights. *Journal of Molecular Biology*, 243(4):574–578, 1994.