Proc. Indian Natl. Sci. Acad. **B71** No. 1 pp 19-41

Prediction of β-barrel Outer Membrane Proteins

PANTELIS G. BAGOS, THEODORE D. LIAKOPOULOS, VASILIS J. PROMPONAS* and STAVROS J. HAMODRAKAS

Department of Cell Biology and Biophysics, Faculty of Biology, University of Athens, Panepistimiopolis, 15701, Greece

(Received on 28 June 2004; Accepted after revision on 8 September 2004)

We attempt to summarize sequence and structural features of β-barrel transmembrane proteins, as they have recently been exploited in order to devise efficient computational methods for the discrimination of these proteins in a genomic context and the prediction of the topology of membrane spanning β-strands. We review a series of prediction methods, ranging from empirical computational schemes, developed in the first days of protein sequence analysis, to modern state-of-the-art machine-learning bioinformatics algorithms, from both a historical and a practical perspective. Furthermore, we discuss common pitfalls and inefficiencies in current methods, at both the initial discrimination step and at the topology prediction stage, suggesting future improvements and perspectives in this emerging research field.

Key Words: β-barrel transmembrane proteins, Prediction, Discrimination, Hidden Markov Model.

Introduction

Biological membranes may be considered both as barriers of individual cells (or even whole organisms in the case of unicellular organisms) or cellular compartments, as well as those structural assemblies that enable each cell and compartment to interact with its environment. A number of key cellular functions, such as signalling under various environmental stimuli (Klare et al. 2004), chemotaxis (Cochran et al. 2001), solute transport (Saier 2000), cell and molecular recognition (Wheelock & Johnson 2003), and immune response (Kurucz et al. 2003), are triggered or exclusively performed by membrane proteins. These proteins may be membrane-associated or integral membrane proteins i.e. proteins having segments that span the lipid bilayer one or more times. As a consequence of their importance in living organisms, transmembrane proteins are often molecules of outmost pharmaceutical and medical importance (Axelson 2004). Actually, it has been estimated that 39 of the top 100 marketed drugs currently in use act through activation or blockade of members of a single family of transmembrane receptors, the *G-protein coupled receptors* (Menzaghi et al. 2002).

Transmembrane (TM) proteins may be grossly classified according to the secondary structure adopted by the membrane spanning segments, namely α-helices (isolated or bundled) and β-pleated sheets in the form of anti-parallel closed barrels (figure 1). Proteins in each class possess distinct characteristics, apparently related to the three-dimensional structures adopted by the transmembrane segments and the underlying folding process. Some of their structural features reflect the biogenesis of membrane proteins and the respective membranes, as well as the corresponding translocation machineries and the environmental constraints posed by the specific physicochemical properties of distinct types of lipid bilayers.

β-helical transmembrane proteins appear to be abundant in all cellular membranes, whereas β-barrel transmembrane proteins have been observed so far only in proteins of the outer membrane of Gram-negative bacteria. Actually, all bacterial outer membrane proteins discovered up to now are thought to belong to this class, constituting a substantial fraction of the outer membrane mass. Sequence similarity and further computational analysis (often combined with low-resolution experimental

Corresponding address: *Email: vprobon@biol.uoa.gr, vprobon@cc.uoa.gr; Tel*: +30-210-7274868; *Fax*: +30-210-7274742

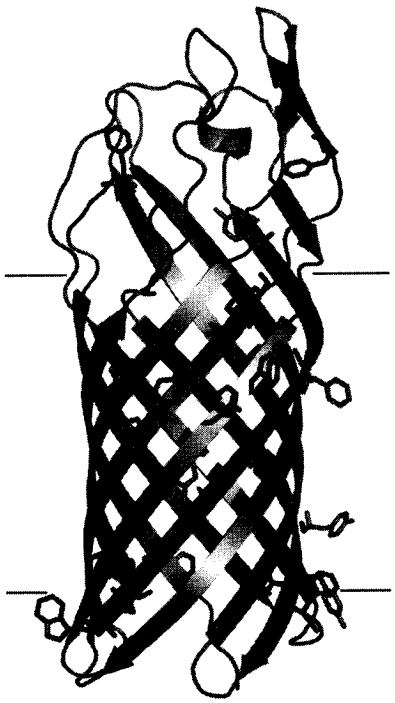

Figure 1. A ribbon diagram of the structure of the OpcA Outer Membrane Adhesin/Invasin from *Neisseria meningitidis* (PDB ID: 1K24; Prince et al. 2002). Aromatic side chains are represented as rods to illustrate the aromatic belts. The horizontal lines indicate the approximate position of the lipid bilayer boundaries. The diagram was drawn using the PyMol molecular graphics package (DeLano, 2003).

evidence) indicate that transmembrane β-barrels are also present in the structures of eukaryotic organellar (mitochondrial or chloroplast) outer membrane proteins. These findings are in accordance with the theory of endosymbiosis; nevertheless, no high-resolution structure of such a protein has yet been reported to the Protein Data Bank (PDB; Berman et al. 2002), in support of this suggestion.

Early experiments (Unwin 1993) provided initial evidence for the existence of mixed-folds composed both of membrane spanning α-helices and β-strands. However, recent work (Miyazawa et al. 2003) shows that this early suggestion is not valid. The presence of a mixture of α-helices and β-strands at the interface with the lipid bilayer could not easily be explained, since stabilizing hydrogen bonding patterns on these secondary structure elements are not complementary (Schulz 2002, Schulz 2003). Thus, experimental data available today indicate that known transmembrane proteins belong exclusively to the aforementioned structural classes.

Known Transmembrane Protein Structures

Knowing the structure of any protein is a major step towards understanding its biological function. High-resolution structures are available for a wide variety of globular water-soluble proteins, whereas the number of unique three-dimensional structures for transmembrane proteins solved at atomic resolution to date is relatively small. Some excellent publicly available resources, namely
http://blanco.biomol.uci.edu Membrane_Proteins_xtal.html,
http://www.mpibp-frankfurt.mpg.de/michel/public memprotstruct.html, and
http://www.enzim.hu/PDB_TM
provide up-to-date information about structural data regarding transmembrane proteins in the Protein Data Bank. In particular, the Protein Data Bank of Transmembrane proteins (PDB_TM; Tusnady et al. 2004), contains not only a collection of transmembrane proteins with known structure, but also annotations for their transmembrane segments computed by a geometrical algorithm that uses as input only the atomic coordinates on the crystal structure.

Despite the tremendous progress witnessed in targeted gene expression, protein purification and crystallization techniques and the advent of the Structural Genomics era, it is expected that deciphering the molecular structure of transmembrane proteins at high atomic resolution will remain a challenging issue in Structural Molecular Biology (Kyogoku et al. 2003, Loll 2003, Walian et al. 2004). Computational studies (Pasquier et al. 2001, Chen & Rost 2002) have already provided more or less accurate estimates that α-helical transmembrane proteins constitute a substantial fraction (ranging between 10-30%) of putative gene products, as deduced from completely sequenced genomes from organisms in all domains of life.

These facts, combined with the availability of an ever-increasing number of complete genomes, highlight the importance of the development of reliable discrimination and classification computational methods to detect and classify transmembrane proteins. Consequently, accurate algorithms to predict the positions of the membrane spanning regions and their topology relative to the lipid bilayer could provide invaluable information for further biochemical, structural or pharmaceutical studies.

Several prediction schemes for α-helical transmembrane proteins have been reported in the literature since the first relevant publications (Argos et al. 1982, Kyte & Doolittle 1982), and several thorough reviews have already been published. In general, even using the simple assumption that sufficiently long (approximately > 15 residues) amino acid stretches of outmost hydrophobicity are putative transmembrane α-helices, a naive predictor might be built. Further utilization of statistical

preferences observed in known transmembrane proteins (Pasquier et al. 1999) and machine learning approaches often combined with evolutionary information (Rost et al. 1996) coming from multiple sequence alignments result in reasonably selective and sensitive prediction methods.

The Repertoire of Transmembrane β-barrel Protein Function

Remarkable advances have been recently made towards the understanding of bacterial β-barrel forming transmembrane protein structure and function. Their functional roles and the Biological processes they are involved in are diverse and may differ between organisms. Long mobile loops resistant to proteolysis (OmpA; Morona et al. 1985) or rigid extensions of the barrel-forming β-strands (OmpX, Vogt & Schulz 1999) in the extracellular space are known to provide molecular recognition sites. Porins are known to mediate the passive transport of small molecules under different environmental conditions (OmpF; Danelon et al. 2003, PhoE; Cowan et al. 1992) or active translocation of larger molecules (FhuA; Braun et al. 2000, FepA; Zhou et al. 1995). In specific cases, they participate in secretion pathways of bacterial exoproteins or type IV pili and flagellar proteins (secretins; Bitter 2003) and virulence through adhesion to host cells (OpcA; Prince et al. 2002).

In the type V secretion pathway (autotransporters, NalP; Oomen et al. 2004), a C-terminal β-barrel domain is necessary to form the pore in the outer membrane, in order to allow the translocation of the secreted mature protein (passenger domain). Furthermore, β-barrel transmembrane proteins have been reported to exhibit key enzymatic activities, either as extracellular proteases (OmpT; Vandeputte-Rutten et al. 2003) or phospholipases (OmpLA, Snijder et al. 1999). Several of these proteins have been shown to function as monomers, but there are known cases where oligomerisation is required for their proper function. Well-known examples for the latter case are bacterial porins, which function after a homotrimer is assembled (Tamm et al. 2001). In some cases, large complexes of outer membrane proteins (>1MDa), both integral and membrane associated, have been reported, for example the secreton of *Klebsiella oxytoca* (Nouwen et al. 1999).

Proteins of the outer membrane of mitochondria and chloroplast outer envelope predicted to belong into this structural class are involved in the major protein translocation complexes (Tom40; Paschen et al. 2003, Toc75; Schleiff et al. 2003) of the respective organelles, mediate the transport of small molecules (porin/VDAC; Mannella 1997), or are key factors determining organelle shape (Mdm10; Sogo & Yaffe 1994). It is noteworthy that high-throughput proteomic analyses (Paschen et al. 2003, Schleiff et al. 2003) have already started to provide additional information in a large scale, which may be further examined by the bioinformatics approaches described in the following sections. There is also experimental evidence suggesting the existence of an anion non-specific porin placed in the peroxisomal membranes. This protein exhibits different channel properties than the already characterised porins of mitochondria and chloroplasts (Reumann et al. 1995). Elucidation of the structural features of these proteins will also provide answers to the speculated endosymbiotic origin of peroxisomes (Borst 1989).

Structural Features of Transmembrane β-barrel Proteins

Any β-barrel may be considered as a β-sheet that twists and coils to form a closed barrel-shaped structure, stabilized by main chain hydrogen bonds formed between the sheet edges (first and last strands). Concerning the connections of the individual strands, different topologies might be associated with β-barrels, such as a simple meander with antiparallel β-strands, where neighboring strands in the sequence are adjacent in the barrel structure, or the more complicated Greek-key arrangement, with relatively long connecting loops on either side of the barrel. Different types of Greek key motifs have been identified in several structures of globular β-barrel proteins of diverge functions (Zhang & Kim 2000). Such cases are the structures of Staphylococcal nuclease (PDB ID: 1STY; Keefe et al. 1993), mitochondrial Elongation factor TU (PDB ID: 1D2E; Andersen et al. 2000), and RNA polymerase subunit RBP8 (PDB ID: 1A1D; Krapp et al. 1998).

Observed transmembrane β-barrels preferentially lay their axis along the membrane normal. All known transmembrane β-barrels are exclusively composed of meandering all-next-neighbor antiparallel β-strands (up and down barrels), suggesting a repeating β-hairpin structural motif. They are described by those parameters, namely the number of β-strands n and the shear number S, that are used to describe all types of β-barrels (McLachlan 1979, Murzin et al. 1994a). S is a measure of the stagger of the strands in the sheet. In an early study (McLachlan 1979), McLachlan showed that n and S determine both the mean radius of the resulting barrel and the relative tilt of strands with respect to the barrel's axis. Fifteen years later, under the light of further experimental

evidence, theoretical analysis (Murzin et al. 1994a) combined with available three-dimensional structures (Murzin et al. 1994b) proved that these two parameters determine all other features of the β-barrel. Currently, available high-resolution structures of transmembrane β-barrel proteins include β-barrels of varying features, with $8 \leq n \leq 22$ and $8 \leq S \leq 24$ (Schulz 2003). It is worth mentioning that all transmembrane α-barrels observed so far consist of an even number of strands.

Discrimination of transmembrane α-barrel proteins is in principle harder than the prediction of α-helical transmembrane segments. Despite the fact that transmembrane β-strands in available high-resolution structures are placed with relatively large angles with respect to the normal to the lipid bilayer, they are significantly shorter than transmembrane α-helices due to their extended conformation, their lengths being typically between six and twenty-two residues. A β-strand of between seven and nine residues length might be sufficiently long to span the hydrophobic core of the membrane. Additionally, transmembrane β-strands face different environments (the hydrophobic exterior of the β-barrel opposed to the aqueous pore interior), often resulting in alternating hydrophobic-hydrophilic residues. This alternation is not always exact, since residues on the outer surface of the barrel (facing the apolar lipidic environment) tend to be hydrophobic, whereas residues pointing to the barrel interior are not always polar. Even though hydrophobicity peaks in a classical hydropathy plot coinciding with amphipathic peaks and β-strand predictions are well correlated with the location of transmembrane β-strands (Zhai & Saier 2002), their average hydrophobicity is significantly lower than those of transmembrane α-helical segments. This fact should be related with the underlying translocation mechanism, since in the opposite case, outer membrane proteins might be trapped in the inner membrane during the translocation process. Additionally, oligomerisation of β-barrel domains inside the lipid bilayer weakens the necessity for a hydrophobic barrel exterior, since polar side-chains may provide favourable interactions at the interaction interface.

Summarising the above factors, the sequence signal to be detected is rather weak. Furthermore, common structural features with globular water-soluble proteins with a β-barrel in their three-dimensional structures might lead to a large number of undesired false positives. Nevertheless, if the amino acid sequence of such a protein is carefully examined, several structural characteristics, for example the predomination of aromatic residues at the interfacial positions, might accurately reveal the location of transmembrane β-strands (for excellent reviews see Schulz 2002, Schulz 2003).

Atypical Cases

We briefly go through some unusual cases of transmembrane β-barrel forming proteins that intentionally were not used in the evaluation of the methods presented in this review.

A class of proteins excluded from our review consists of those possessing transmembrane β-barrels formed by more than one amino acid chain. A protein belonging to this class is *Escherichia coli* TolC (PDB ID: 1EK9; Koronakis et al. 2000). TolC is a mixed β-barrel and α-helical protein, which spans both the outer membrane and the periplasmic space of gram-negative bacteria. Three TolC protomers assemble to form a continuous, solvent accessible conduit, a "channel-tunnel" over 140 Å long. Each monomer of the trimer contributes 4 β-strands to the 12-strand β-barrel. Another protein belonging to this class is β-haemolysin from *Staphylococcus aureus* and other microbial toxins such as aerolysin and the anthrax-protective antigen. In the case of α-haemolysin, it has been shown (PDB ID: 7AHL; Song et al. 1996) that it is active as a transmembrane heptamer, where the transmembrane domain is a 14-stranded antiparallel β-barrel, in which two strands are contributed by each monomer. This endotoxin causes disease by forming pores on the infected cell membrane leading to cell lysis or to the destruction of small molecule concentration gradients.

Recently, the structure of a Mycobacterial (Gram-positive) outer membrane channel has been determined at atomic resolution (MspA, Faller et al. 2004). This structure has not been considered in any of the studies mentioned hereinafter, since Mycobacterial mycolate-rich outer membranes are considered atypical. Actually, these are the thickest biological membranes known to date, and present a decreased fluidity toward the periplasmic side of the membrane as opposed to the outer membrane of Gram-negative bacteria (Liu et al. 1995).

In addition, it is well known that apart from integral outer membrane proteins, Gram-negative bacteria possess a number of lipoproteins covalently attached to the outer membrane by means of N-terminally attached lipids. Recent work (Juncker et al. 2003, Brokx et al. 2004) provides evidence that high-throughput experiments might improve the refinement of the few existing predictive methods.

Concepts used for the Prediction of Transmembrane β-strands

The β-barrel outer membrane proteins share some unique characteristic structural features that may be used for predicting their structure. These are:

(1) The transmembrane β-strands are mainly amphipathic showing an alternation of hydrophobic and polar residues. The hydrophobic residues interact with the hydrophobic lipid chains, whereas the polar residues face toward the barrel interior, hence interacting with the aqueous environment of the pore.

(2) The aromatic residues have a greater tendency to be located in the interfaces with the polar heads of the lipids, thus forming the so-called "aromatic belts" around the perimeter of the barrel.

(3) Both the N-terminal and the C-terminal of the proteins are located in the periplasmic space (inside with respect to the outer membrane). In some cases, the N-, and C-terminal tails of the protein may be formed by more than 100 residues-long stretches.

(4) The segments connecting the transmembrane strands that are located in the periplasmic space (inside loops) are generally shorter than those of the extracellular space (outside loops). The periplasmic loops have a length no longer than twelve residues, whereas the extracellular loops may be significantly longer, with lengths exceeding thirty residues. This observation is possible due to the meander arrangement observed in currently available structures. If transmembrane β-barrels adopted a Greek-key topology, longer loops on both sides of the barrel would be present.

(5) The length of the transmembrane strands varies according to the inclination of the strand with respect to the lipid bilayer, and ranges between six and twenty-two residues. However, in some cases only a small portion of the strand is embedded in the lipid bilayer, and the rest of it protrudes far away from the membrane, to the extra-cellular space, forming flexible hairpins.

(6) β-barrel outer membrane proteins show great sequence variability in their amino acid sequences. This, in general, is larger than that of the globular proteins, and it is even larger in the extracellular loops, which often function as antigenic epitopes.

(7) Adjacent strands are connected by a network of hydrogen bonds, stabilizing the barrel.

Prediction Methods Based on Hydrophobicity Analysis

The alternation of hydrophobic and polar residues in the membrane spanning β-strands was used quite early to assist prediction of β-barrel membrane proteins. Vogel and Jahnig (Vogel & Jahnig 1986) introduced the use of a sliding window that averages the mean amphipathicity of every second residue along the sequence. They used a window of seven residues, centered around the residue *i*. Thus, the mean amphipathicity *H*, for an amino acid *i* was defined as:

$$H(i) = [h(i-2) + h(i) + h(i+2) + h(i+4)]/4$$

where *h(k)*, is the hydrophobic index of amino acid *k* according to the Eisenberg hydrophobicity scale (Eisenberg et al. 1984). Vogel and Jahnig combined their analyses with experimental evidence derived from Raman spectroscopy, and they were able to predict correctly the majority of the membrane spanning strands of OmpA, Porin and Maltoporin, proteins with three-dimensional structures not available at that time. Jeanteur and colleagues (Jeanteur et al. 1991) combined amphipathicity with sequence alignments of members of the porin family, and concluded that porins possess a 16-stranded β-barrel. Schirmer and Cowan, (Schirmer & Cowan 1993) extended the approach of Vogel and Jahnig, heuristically setting the hydrophobicity of residues (i-2) and (i+4) to 1.6, if they were found to be aromatic. Doing so, they tuned the method to identify more accurately the aromatic belt of the transmembrane strands, and they were able to verify the correct location of the membrane strands for the recently solved structures of Porin from *Rhodobacter capsulatus* and *E. coli*, as well as that of Osmoporin from *E. coli*. They also managed to predict the membrane strands of the Maltoporin from *E. coli*, of which a high-resolution structure was not available at that time. Rauch and Moran (Rauch & Moran 1994), applied a modified version of this algorithm. They used a window of five residues, and subtracted from the hydrophobicity of each residue a value corresponding to the average hydropathy. Afterwards, in each given window in the sequence, they evaluated the total fraction of oscillations around zero, which they called "fraction of period detected". This way, segments with a fraction close to 1 would be probable transmembrane β-strands. Utilizing this approach, they performed prediction of the membrane spanning strands of the mitochondrial outer membrane proteins VDAC and OM38 from several eukaryotic species, which putatively possess β-barrel structures. In their following study, they extended their method, using similar hydropathy profiles to predict both α-helical membrane segments and transmembrane β-strands (Rauch & Moran 1995). Gromiha and Ponnuswamy (Gromiha & Ponnuswamy 1993) derived the concept of surrounding hydrophobicity that does not depend only on the amphipathic features of the β-strands. They constructed their scale and performed predictions on several bacterial porins with unknown three-dimensional structures.

The Beta Barrel Finder (BBF) program developed by Zhai and Saier (Zhai & Saier 2002) combined hydrophobicity, amphipathicity, predicted secondary

structure and the presence of a signal peptide to identify putative β-barrel outer membrane proteins in prokaryotic genomes. The authors, by performing analyses on the known structures of outer membrane proteins, deduced a set of rules, indicating that putative transmembrane β-strands should be recognized as segments in which a peak in amphipathicity should coincide with a peak in hydrophobicity and a secondary structure prediction for a β-strand. Furthermore, the presence of the signal peptide was considered a strong indication of the protein's localization to the outer membrane, since all the outer membrane proteins are known to possess such a sequence, essential for their translocation through the bacterial inner membrane. With the use of BBF, the authors conducted a search in all the predicted ORFs from the *E. coli* genome sequence, identifying 118 putative β-barrel outer membrane proteins. BBF was one of the first methods applied to entire genomes (see below), however it does not explicitly predict the transmembrane topology, and the results were not evaluated statistically in order to allow reliable conclusions about the rate of false positive or false negative predictions.

However, we should point out that predictive methods based on hydrophobicity analysis and/or secondary structure prediction have inherent limitations. In their recent study of β-sheet folding in membranes, Bishop and colleagues (Bishop et al. 2001) show that sheet forming propensities routinely used for secondary structure prediction are not correlated to their experimental model, possibly reflecting the different underlying folding mechanism between β-sheets in water-soluble and integral membrane proteins. Additionally, they show that hydrophobicity scales based on non-polar core environments (i.e. GES; Engelman et al. 1986, RW; Radzicka & Wolfenden 1988) are perfectly correlated to the sheet forming preferences, whereas scales based on more polar environments (i.e. WW bilayer; Wimley & White 1996, WCW octanol; Wimley et al. 1996) have a poorer correlation (Bishop et al. 2001). Such issues should be seriously taken into account when developing empirical predictive methods for β-barrel integral membrane proteins, since they may lead to inaccurate results.

Statistical Approaches

Soon, it became clear that features of these proteins other than the hydropathy profiles should also provide useful information for predicting the transmembrane strands. Gromiha and associates (Gromiha et al. 1997) derived a set of conformational parameters and the associated rules that helped them to predict the transmembrane strands of the porins known at atomic resolution at that time. In this approach, the authors exploit the hydrophobic and amphipathic character of the sequence, incorporating additional propensities for the amino acids to be parts of a transmembrane strand, as derived from analyses of known structures. Furthermore, they introduced specific rules, derived from expert knowledge of known structures. Combining all these features they achieved a (per-residue) accuracy prediction of 82%, which was the highest achieved until that time. Neuwald and colleagues (Neuwald et al. 1995) applied a method based on the statistical formulation of the Gibbs sampler in order to find and align specific motifs characterizing a set of distantly related (non homologous) bacterial outer membrane proteins. The Gibbs sampler discovered such a repetitive motif, which discriminates outer membrane proteins, with an exceptionally high statistical significance. The motif was present in the transmembrane strands of the porins known at atomic resolution, and more precisely in the strands that form the exterior side of the trimeric pore, suggesting potential structural and functional roles. Later, Manella and associates (Mannella et al. 1996) used the same approach to search for mitochondrial proteins, with a significant match to the motif. They found that only the two hypothesized â-barrel outer membrane mitochondrial proteins, VDAC and Tom40, matched significantly to the motif, a fact that has strengthened the belief that these proteins are indeed transmembrane β-barrels.

Gnanasekaran and colleagues (Gnanasekaran et al. 2000) proposed the use of structure based sequence alignments in order to find specific patterns discriminating β-barrel outer membrane proteins. After superposition of the structures of 5 different bacterial porins, they deduced a multiple sequence alignment that helped them to identify profiles from structurally conserved regions (pSCRs), corresponding to the 16 transmembrane strands occurring in bacterial porins. Using these profiles, they report significant hits to a database consisting of 82 α-helical proteins, 68 β-barrel membrane proteins and 45 unidentified/non-membrane proteins, with a false-positive rate of ~10-20%. However, the authors did not propose either an effective way to combine the individual motifs, which they have ranked in order of discriminative power, or a strategy useful for scanning large databases or genomes.

Wimley developed a scale-based method to identify putative β-barrel outer membrane proteins based on a statistical analysis of 15 known structures (Wimley 2002). By aligning the structures with respect to the hypothesized lipid bilayer plane, he derived statistical frequencies of the residues belonging to a transmembrane strand (pointing either to the barrel

interior or to the exterior) and the residues belonging to non-membrane parts (loops). By observing the fact, that the repetitive structural motif in β-barrel proteins is the β-hairpin (two strands connected by a short periplasmic turn), he developed an algorithm that sums the individual amino acid propensities in a given window (with a period of 2) and produces a α-strand score, capable of identifying the majority of transmembrane strands. Furthermore, by summing the individual segment predictions and averaging for the sequence length he produced a β-barrel score useful for the final classification of the protein. Applying this algorithm to the genome of *E. coli*, taking into consideration only the 200 top-scoring proteins, Wimley concluded that this set includes the majority of the known outer membrane proteins, and a large number of putative or potential outer membrane proteins. However, the method of Wimley, like that of Zhai and Saier, does not report explicit topology predictions, and was not validated statistically on an independent dataset. This means that simple evaluation of the score produced by this method does not suffice to declare a candidate protein as outer membrane protein with a known confidence, as proved later in a research conducted by another group (see below).

Liu and associates (Liu et al. 2003a) took a different approach in order to discriminate β-barrel outer membrane proteins from all β-globular proteins. By analysing the amino acid frequencies of residues occurring in β-strands of both globular and transmembrane β-barrel proteins of known structure, they concluded that certain residues occur statistically more frequently in one or the other group, and thus can be used for discrimination. Considering the predicted secondary structure performed by PSI-PRED (and ignoring sequences with less than 4 predicted strands), they used a linear discriminant function to classify an independent set of outer membrane proteins and globular all-β proteins, with a success rate of 85.5% for outer membrane and 92.5% for globular ones. In the classifier used in this study, six amino acids were selected as having the greatest discriminative power, these are Glycine (G) and Asparagine (N), showing a preference for being part of a transmembrane strand, and Valine (V), Isoleucine (I), Lysine (K) and Cysteine (C) which show preference to participate in strands of water soluble proteins.

Our group (Bagos et al. 2004a) proposed a method based on the Markov Chain Model, in order to perform the task of discrimination of β-barrel outer membrane proteins. The 1st order Markov Chain Model states that the probability of observing a particular residue depends on the occurrence of its immediate predecessor (Durbin et al. 1998). Thus, by obtaining the individual parameters of the model corresponding to the 400 amino acid pairs, the model was able to identify clearly the alternation of hydrophobic/polar residues, frequently occurring in β-barrel proteins, and produce a log-odds score per protein useful for discrimination. Using a set of well-annotated outer membrane proteins and globular proteins with known structures, the authors achieved a correct classification rate of 89.2% for outer membrane proteins and 92.5% for globular ones, in a jackknife test.

The BOMP method (Berven et al. 2004), uses a combination of regular expression patterns, the β-barrel score of Wimley, and a post processing step to filter false positives based on the overall amino acid composition. In particular, the first method applied is based on the presence of a pattern characterizing the most C-terminal β-strand of the barrel. This pattern is:

.{100} [^C] [YFWKLHVITMAD] [^C] [YFWKLHVITMAD] [^C] [YFWKLHVITMAD] [^C] [YFWKLHVITMAD] [^C] [FYW]

There is evidence that the occurrence of an aromatic amino acid, most often phenylalanine, in the last position of the most C-terminal β-strand of the barrel, is important for the assembly of the protein and the insertion into the lipid bilayer (Struyve et al. 1991). This pattern is also flexible in allowing the occurrence of amino acids YFWKLHVITMAD in the remaining positions pointing towards the membrane, and allowing all amino acids except Cysteine in the positions pointing inwards the membrane. We should mention here that Cysteine, is not present in any of the transmembrane strands of β-barrel outer membrane proteins known at atomic resolution, and has a higher propensity for globular proteins as previously reported (Liu et al. 2003a).

The second method applied is the β-barrel score proposed by Wimley, with a threshold empirically obtained from reference sets compiled by the authors. Additionally, there is a filtering procedure necessary to remove false positives. This is based on identifying residues occurring more frequently in the β-barrel outer membrane proteins than in the globular proteins, as confirmed statistically by Principal Components Analysis (PCA). This classifier considers the relative abundances of two amino acids, namely Asparagine (N) and Isoleucine (I) as they gave the best separation between true and false positives in the reference set. In agreement with previous work (Liu et al. 2003a), Asparagine was detected to be more abundant in β-barrel outer membrane proteins,

whereas Isoleucine was more abundant in water soluble proteins.

BOMP achieves an overall recall [i.e. *true positives/(true positives + false negatives)*] of 88% with 80% precision [i.e. *true positives/(true positives + false positives)*], as measured in the well-annotated outer membrane proteins of *E. coli* and *Salmonella typhimurium*, found in SwissProt after removing homologues with similarity above 40%. These correspond to a rate of true positives around 88.1% and true negatives 98.8%. In general, even though BOMP does not utilize any new algorithmic techniques, it performs very well with its main priority of avoiding over predictions, since it has the lowest false positive error rate reported so far.

Machine Learning Methods

As the number of crystallographically solved three-dimensional structures continued to grow, it became obvious that the issue of predicting β-barrel outer membrane proteins was more complicated than the simple detection of alternation of hydrophobic-polar residues. Furthermore, during the '90s an explosion in bioinformatics techniques occurred, where Machine Learning approaches (such as the Artificial Neural Networks, ANNs, and the Hidden Markov Models, HMMs) were adopted to solve well-known biological problems. Such problems were: prediction of protein secondary structure (Qian & Sejnowski 1988, Asai et al. 1993, Rost & Sander 1993), prediction of β-helical transmembrane segments (Rost et al. 1995, Sonnhammer et al. 1998, Pasquier & Hamodrakas 1999, Krogh et al. 2001), prediction of signal peptides (Nielsen et al. 1997, Nielsen & Krogh 1998, Nielsen et al. 1999), gene finding (Demeler & Zhou 1991, Farber et al. 1992, Krogh et al. 1994), protein structural classification (Pasquier et al. 2001), subcellular location prediction (Reinhardt & Hubbard 1998), constructing profiles for sequence families (Eddy 1998) and multiple sequence alignment (Eddy 1995). These methods are, in general, more capable of finding the non-linear correlations of amino acids in protein sequences, and perform better than simple statistical analyses and heuristic methods based on physicochemical parameters and amino acids composition. Furthermore, the mathematical foundations of these methods are sounder, providing a safe starting point for their use.

The first attempt to apply a machine learning approach for predicting the topology of β-barrel outer membrane proteins was conducted by Diederichs and colleagues (Diederichs et al. 1998). They used an Artificial Neural Network (Bishop 1995) for predicting the relative position of the Cα atom of each amino acid residue of bacterial porins with respect to the lipid bilayer. To perform the training, they used seven structures of bacterial porins known at atomic resolution. They aligned the structures belonging to the training set with their pores along the z-axis in order to establish a relationship between the z-coordinates of the Cα and the transmembrane topology. This way, the outer membrane lies in the xy-plane, and the network was trained to predict the z-coordinate of Cα atoms, such that low values of z-coordinate for a given residue indicate the probability of a periplasmic turn, medium values that of a transmembrane β-strand, and higher values an extracellular loop. The network that was used had a standard feed-forward architecture with one hidden layer, trained by the back-propagation algorithm. The authors reported a correlation coefficient (Baldi et al. 2000a) of 0.58 in the per-residue accuracy on the training set. Furthermore, they applied the method to several outer membrane (non porins) proteins, for which three-dimensional high-resolution structures were not available, including OmpA, Omp32, FepA and FhuA. However, the predictions performed for these proteins were proved inaccurate, and this became apparent when additional three-dimensional structures of outer membrane proteins became available.

As the number of crystallographically solved structures continued to rise, one should expect that more refined methods with a better performance would be developed. Indeed, Jacoboni and associates (Jacoboni et al. 2001) proposed the use of a similar feed-forward Neural Network (B2TMPRED), trained on the structures of eleven outer membrane β-barrel proteins deposited in PDB until 2001. The main novelties of this method were the use of evolutionary information derived from multiple alignments made by PSI-BLAST (Altschul et al. 1997) instead of using single sequence information. A post-processing step was introduced, involving a dynamic programming algorithm (Jones et al. 1994, Fariselli et al. 2003) in order to locate correctly the transmembrane strands when a given output of the neural network is obtained. Incorporation of Multiple Sequence Alignments are reported to significantly improve the accuracy of all kinds of secondary structure prediction algorithms (Przybylski & Rost 2002). Additionally, the dynamic programming step, when implemented in accordance to the constraints imposed by the known structures (such as the length of the strands or that of the loops), is a very powerful tool for obtaining a reasonable prediction using the output of the neural network (Fariselli et al. 2003). These two features, along with the fact that the training set comprised eleven non homologous

sequences, allowed the method to achieve a per residue accuracy of 78% and a correlation coefficient of 0.56 in the jackknife test, whereas for the self-consistency the same measures were 89% and 0.77 respectively, much better than those obtained by the neural network of Diederichs. Furthermore, the authors claimed that their method had the ability to predict the protein's full topology, by counting the lengths of the loops and assigning the smaller loops to the periplasmic space. On these grounds, these authors reported the number of correctly predicted topologies (where all strands and loop orientation are correctly predicted) to be 8 out of the 11 proteins of the training set. A Neural Network with a similar architecture, based solely on the amino acid sequence was presented much later by the Gromiha group (Gromiha et al. 2004). This method does not use either evolutionary information or the dynamic programming for the post-processing step, but instead it applies a heuristic that tries to correct the outputs of the network (i.e. to eliminate predicted strands with two or three residues). The method was trained on thirteen non-homologous β-barrel outer membrane proteins, and the authors report a per residue accuracy of 73% and a correlation coefficient of 0.46, results clearly inferior compared to the method proposed by Jacoboni, where evolutionary information was used.

Very recently, a new method (TBBPred), which combines Neural Networks and Support Vector Machines was introduced (Natt et al. 2004), trained on a larger non-redundant dataset of 16 β-barrel outer membrane proteins. The Neural Network part of the method is conceptually similar to that developed by Jacoboni, using profiles derived from PSI-BLAST alignments as the input. This NN (in the jackknife testing procedure) correctly predicts 80.5% of the residues, with a correlation coefficient of 0.63 and correctly locates the number of the transmembrane segments for 7 out of the 16 proteins. The SVM method uses as input the sequence along with 32 features derived from it, such as hydrophobicity etc. It achieves a per residue accuracy of 78.5%, with a correlation coefficient 0.55, whereas the number of proteins with correctly located strands is 10 out of the 16 in the jackknife test. Combining the two methods, the authors report a per residue accuracy of 81.8%, a correlation coefficient of 0.64, whereas the total number of proteins with correctly predicted transmembrane strands equals to 9 out of the 16. This method is the only one until now that exploits the power of the statistical learning theory incorporated in the SVMs (Vapnik 1998). The observation that the combined prediction increases the accuracy reflects the fact that SVMs are capable of capturing different sequence characteristics essential for the prediction, better than the Neural Networks do. Furthermore, the authors, using information derived from the number and the lengths of the predicted strands, report a 88.8% of correct classification for outer membrane proteins and 92.3% for globular proteins.

The other major class of machine learning techniques widely applied to bioinformatics problems is the Hidden Markov Models (HMMs). The HMMs are stochastic models, defining a regular grammar on the amino acid sequence (Rabiner 1989, Durbin et al. 1998). Their mathematical formalism allows the design of elegant algorithms for training these models and perform the predictions (Baum 1972, Durbin et al. 1998). The first method based on a HMM to predict the transmembrane strands of β-barrel outer membrane proteins was the HMM-B2TMR method (Martelli et al. 2002). This method was trained on a non-redundant set of 12 outer membrane proteins, obtaining input from PSI-BLAST derived profiles. This method introduced different states in the HMM architecture corresponding to the structural characteristics of the alternating hydrophobic-polar residues in the transmembrane strands, the aromatic belt, the periplasmic turns and the extracellular loops. HMM-B2TMR was trained according to a modified version of the Baum-Welch algorithm for HMMs with labelled sequences (Krogh 1994), aiming to incorporate the profile as the input instead of the raw sequence, whereas at the decoding stage the posterior decoding method was used, along with an additional post-processing step involving the same dynamic programming algorithm used by Jacoboni and colleagues. This method reached a rather high per residue accuracy (83%) with a correlation coefficient of 0.65 and the number of proteins with correctly determined topology during the jackknife testing procedure was 7 out of the 12. Furthermore, this method was also capable of discriminating between outer membrane proteins and water-soluble proteins, with a correct classification rate of 84% for outer membrane proteins and 90% for water-soluble ones. The method was retrained on a larger dataset of 15 non-homologous outer membrane proteins, improving further the prediction accuracy (Fariselli et al. 2003).

Later, a similar HMM-based method was introduced by Liu and associates (Liu et al. 2003b) for performing the same task. This method was trained on a dataset of 11 outer membrane proteins according to the standard Baum-Welch algorithm (Baum 1972) and accepts single sequence information as input. The decoding was performed with the

standard Viterbi algorithm for HMMs (Durbin et al. 1998), yielding a per segment accuracy of 97% (167 out of the 172 transmembrane strands in the training set) and predicting correctly the topology of 7 out of the 11 proteins, in the jackknife test. The authors do not report the per residue accuracy and correlation coefficient. Furthermore, this method, in contrast to that of Martelli, was not capable of performing discrimination between outer membrane proteins and water-soluble proteins.

We have recently presented a HMM method (PRED-TMBB;Bagos et al. 2004b, Bagos et al. 2004c) quite similar in principle to the previously mentioned HMM-based methods, but with some major practical and theoretical improvements. Whereas the overall architecture of the model was conceptually similar to HMM-B2TMR and the method of Liu, and designed in order to specifically fit the limitations imposed by the known structures, the method was trained and decoded following a completely different philosophy. In contrast to the previous methods that relied for training on the Baum-Welch algorithm that performs Maximum Likelihood (ML) estimation, the PRED-TMBB method was trained according to the Conditional Maximum Likelihood (CML) criterion (Krogh 1997) using a gradient-descent method (Krogh & Riis 1999). Whereas the ML criterion maximizes the probability of the sequences given the model, the CML approach maximizes the probability of the correct prediction given the sequences and the model. Hence, even though using CML is computationally more intensive than the Baum-Welch (ML) approach, the predictive ability is expected to be better, given that we have data with good quality of labelling. Toward this end, the authors performed meticulous manual sequence labelling (the assignment of each amino acid in one of the three classes to be predicted, transmembrane strand, periplasmic turn and extracellular loop), by observing directly the three-dimensional structures and not relying on the annotation of the PDB entries, as was done by all previous investigators. Following this approach, we were able to precisely locate the aromatic belt of the barrel, the residues facing the barrel interior and exterior, and more importantly not to include for training as transmembrane those parts of the strands protruding far away from the lipid bilayer to the extracellular space. For the decoding step, we did not rely on the Viterbi algorithm, routinely used for HMMs. Along these lines, we introduced for the first time, for outer membrane proteins, the N-Best algorithm (Krogh 1997), which is a heuristic that seeks to find the most probable labelling of the sequence and not just the most probable path of states, as the Viterbi algorithm does. We, optionally, also used posterior decoding, with a dynamic programming algorithm for the post-processing step, which differed in some aspects from that used by B2TMPRED and HMM-B2TMR. With these approaches, PRED-TMBB reached in the jackknife test a per-residue accuracy of 84.2% and a correlation coefficient of 0.72. Moreover, the method correctly predicted the topologies for 10 out of the 14 proteins in the training set (in the jackknife test), showing that PRED-TMBB locates correctly the transmembrane β-barrels better than any other method, even if it is based on single sequence information. Additionally, the method also discriminates outer membrane proteins from globular, water-soluble ones, reaching a correct rate of 89% for both classes. The method has recently been retrained in order to include some recently solved three-dimensional structures. This way, the algorithm reached, in the jackknife test, an accuracy of 87.5% and a correlation coefficient of 0.74, with 12 out of the 16 proteins having their topologies correctly predicted (Bagos et al. 2004c).

The latest addition to the family of HMM-based predictors is the ProfTMB method introduced by the Rost research group (Bigelow et al. 2004). This method also uses input derived from PSI-BLAST profiles and training is obtained by a modified version of the Baum-Welch algorithm and decoding using the standard Viterbi algorithm. The model architecture is once again similar to that of the previously mentioned methods, and for the training set 8 non-homologous proteins from PDB were used. The main novelty of the method is the use of different model parameters (emission and transition probabilities) to model the strands with direction from the periplasmic space to the extracellular matrix (Up strands), and different ones for the strands of the opposite direction (Down strands). Furthermore, it uses different states to model explicitly the different structural types of periplasmic loops (turns, hairpins etc). ProfTMB performs equally well compared to the previously mentioned HMM-based methods in terms of per-residue accuracy (83%) and correlation coefficient (0.70). However, the author's choice to use a different (log-odds) score for discrimination purposes implicates the discrimination capability of the method. In particular, the authors report that a log-odds score of 12 should be used as a cut-off for the discrimination; with this as a threshold, the method achieves 100% positive predictive value (called accuracy in the paper, i.e. does not predict a single false positive) but the percentage of correctly predicted OMPs is 40% (called coverage in the paper). It is obvious that the method, using a cutoff score of 12, is very specific (does not predict OMPs falsely) but not quite sensitive (misses a lot of

true positives), an issue raising implications discussed in the following sections. The authors also report that the same method with a threshold of 3 achieves 50% coverage and 80% accuracy, whereas for a threshold of 0 it reaches an equal accuracy and coverage of 55%.

Alternative Formulations

Subcellular Location Prediction: Another way to formulate the problem of outer membrane protein prediction is to address it in the context of subcellular location prediction. The PSORT-B algorithm was designed specifically to predict the subcellular location of proteins belonging to Gram-negative bacteria (Gardy et al. 2003). In order to achieve this, the algorithm combines several independent modules, i.e. prediction of transmembrane β-helices, prediction of a signal peptide etc. One of the modules used in this system, the "OMP motif", was built by searching for frequent motifs found in outer membrane proteins. A total number of 279 motifs were found and a query sequence is scanned for the presence of 3 or more such motifs, in order to be classified as an outer membrane protein. The individual specificity of this module was found to be 100%, whereas the sensitivity was 23.6%. However, combining the individual modules with the use of a Bayesian Network classifier (Cowell et al. 1999), the authors finally report 98.8% specificity and 90.3% sensitivity for the classification of outer membrane proteins. Some authors of the same group extended their approach by data-mining frequent subsequences of outer membrane proteins (She et al. 2003), and used them to build a classifier based on Support Vector Machines (SVM; Vapnik 1998). Since the primary objective (see below) was mainly to make good predictions of outer membrane proteins, they report the precision (positive predictive value – the fraction of correctly predicted OMPs among the totally predicted as OMPs) and the recall (sensitivity – the fraction of correctly predicted OMPs among the actually true OMPs) to be 98% and 81%, respectively. The main disadvantage of the PSORT-B algorithm is that it cannot discriminate the β-barrel outer membrane proteins from outer membrane lipoproteins.

Secondary Structure Prediction: As mentioned in previous sections, secondary structure prediction methods have been used in combination with hydrophobicity analyses (Zhai & Saier 2002) or coupled with discriminant analysis utilizing differences in the distribution of specific amino acids (Liu et al. 2003a). On the other hand, it has been clearly stated that the transmembrane β-strands of outer membrane proteins differ significantly in their physico-chemical properties and amino-acid composition, compared to the β-strands occurring in the water-soluble proteins (Schulz 2000, Schulz 2002, Schulz 2003, Wimley 2003).

However, molecular biologists repeatedly use general-purpose secondary structure algorithms, such as PSI-PRED (McGuffin et al. 2000), PHD (Rost 1996, Rost & Liu 2003) or JPRED (Cuff et al. 1998), combined with hydrophobicity analyses to manually locate the putative membrane-spanning strands of newly sequenced outer membrane proteins. This strategy has been used for years and seems to work, judging from the relevant publications. Using such approaches, along with sequence alignments of outer membrane proteins of known structures, Rodriguez-Maranon and colleagues (Rodriguez-Maranon et al. 2002) concluded that the Major Outer Membrane Protein (MOMP) of Chlamydia is a porin with 16 transmembrane strands. Paquet and associates (Paquet et al.2000) combined several different algorithms to perform predictions on the *Brucela abortus* Omp2a and Omp2b porins. Along these lines, Zhang and colleagues (Zhang et al. 2000) performed prediction on the *Cambylobacter jejuni* MOMP.

Genome Scale Analysis

As we have noted earlier, predicting the transmembrane strands and outer membrane protein discrimination in large datasets are two entirely different problems. Thus, when it comes to genome analysis there are some important issues that have to be carefully taken into account.

In table 1, we cite all the available prediction algorithms, listing the corresponding capabilities regarding transmembrane strand prediction and discrimination power. Some algorithms predict both transmembrane strands and discriminate β-barrel proteins, while others perform only one of the aforementioned tasks. Moreover, the performance of each algorithm is highly variable, since each one of them is oriented toward a different end. Thus, some algorithms are more specific i.e. do not misclassify outer membrane proteins, with the disadvantage of giving a lot of false positive results, whereas others are more sensitive i.e. they do not produce many false positives, with the cost of misclassifying some true β-barrels.

The algorithms' sensitivity and specificity become important in genome analysis due to the fact that β-barrel outer membrane proteins constitute only a small fraction (<4%) of the genome of Gram-negative bacteria (Casadio et al. 2003a, Berven et al. 2004, Bigelow et al. 2004). Given this, even an algorithm with a 5% error classification rate will produce a large number of false positives. The first methods applied

Table 1. *Methods for discriminating/predicting β-barrel integral membrane proteins.*
The available predictors, used for discriminating and/or predicting the transmembrane strands of β-barrel integral membrane proteins.
(1) HMM-B2TMR, is available only as a commercial demo.

	Method	Reference	TM Strands	TM Strands + Orientation	Discrimination	URL
Outer membrane predictors	B2TMPRED	Jacoboni et al.2001	x	-	-	http://gpcr.biocomp.unibo.it/cgi/predictors/outer/pred_outercgi.cgi
	BOMP	Berven et al. 2004	-	-	x	http://www.bioinfo.no/tools/bomp
	HMM-B2TMR (1)	(Martelli et al. 2002)	x	x	-	http://gpcr.biocomp.unibo.it/biodec/ (1)
	MCMBB	Bagos et al. 2004a	-	-	x	http://bioinformatics.biol.uoa.gr/mcmbb
	OM_Topopredict (2)	Diederichs et al. 1998	x	x	-	http://strucbio.biologie.uni-konstanz.de/~kay/om_topo_predict2.html (2)
	PRED-TMBB	Bagos et al. 2004b, Bagos et al. 2004c	x	x	x	http://bioinformatics.biol.uoa.gr/PRED-TMBB/
	ProfTMB	Bigelow et al. 2004	x	x	x	http://cubic.bioc.columbia.edu/services/proftmb/
	PSORT-B	Gardy et al. 2003	-	-	x	http://www.psort.org
	TBBpred	Natt et al. 2004	x	-	x	http://www.imtech.res.in/raghava/tbbpred/
	TM-BETA	Gromiha et al. 2004	x	-	-	http://psfs.cbrc.jp/tmbeta-net/
	Wimley	Wimley 2002	-	-	x	Under construction
Secondary Structure Predictors	PHD	Rost & Liu 2003	-	-	-	http://www.predictprotein.org/
	PSI-PRED	McGuffin et al. 2000	-	-	-	http://bioinf.cs.ucl.ac.uk/psipred/

to whole genomes with the aim of finding β-barrel outer membrane proteins on a genomic scale were those of Wimley (Wimley 2002) and the BBF program (Zhai & Saier 2002). As we stated earlier, both methods were not statistically validated in advance, but the proteins that are more likely to be β-barrels were reported instead. Wimley, reports 200 potential β-barrels, while the BBF method predicts 118 β-barrels in the genome of *E. coli*. Both methods classified correctly the verified â-barrel proteins of *E. coli*, but the question whether the remaining of the predicted proteins are true β-barrels or false positives remains unanswered. The PSORT-B algorithm (Gardy et al. 2003), when applied to 77 genomes of Gram-negative bacteria (including *E. coli*), predicted 255 outer membrane proteins, without distinguishing however β-barrels from lipoproteins.

The first method, based on statistically validated tools, was the Hunter suite of programs (Casadio et al. 2003a). Hunter combines the NN and the HMM predictor previously developed by the same group (Jacoboni et al. 2001, Martelli et al. 2002) coupled with two NN predictors, one for α-helical membrane proteins and another for signal peptides. This method makes the useful assumption that the β-barrel outer membrane proteins possess a signal peptide for their translocation across the inner membrane, applying a pre-processing filter to exclude α-helical transmembrane proteins. By this approach, Hunter minimizes the candidates presented to the β-barrel predictor, thus minimizing the probability of having false positives. The method finally clusters the genome into three categories: β-barrel membrane proteins, α-helical membrane proteins and soluble proteins, with

a correct classification rate of 95.6% for the three classes and 84% for β-barrels, as tested on the well-annotated subset of *E. coli* proteins. The authors utilized Hunter to perform predictions on 9 Gram-negative bacterial genomes, and concluded that the β-barrel content ranges from 1.5% to 2.4%, with the *E. coli* genome having 78 β-barrels (1.5% of the total genome). In conclusion, Hunter seems to be the more sophisticated method for predicting the β-barrel proteins in entire genomes, yet it misses at least 16% of the true β-barrels. Furthermore, it is not publicly available.

The ProfTMB method was also used for scanning 72 genomes of Gram-negative bacteria (Bigelow et al. 2004). As we already mentioned, profTMB is rather specific, hence the total number of β-barrels that it reports is relatively small, and as expected it predicts 70 β-barrel outer membrane proteins in *E. coli*, and 164 novel outer membrane proteins (with no known homologs) in all the genomes analyzed. Finally, the BOMP program (Berven et al. 2004) predicted the total genome content of 10 Gram-negative bacteria to range between 1.8% and 3%. Particularly, the *E. coli* genome was predicted to encode for 103 â-barrel membrane proteins. Since all predictions point to similar estimates, these seem to be quite accurate and close to reality. However, the issue of algorithms' specificity and sensitivity has to be addressed thoroughly in the near future. Improvement of the methods is necessary for predicting more reliably the β-barrel genome content of Gram-negative bacteria.

Eukaryotic β-barrel Membrane Proteins

As we noted earlier in the text, eukaryotic organisms presumably possess a fraction of β-barrel membrane proteins, located in the outer membrane of the semi-autonomous organelles such as mitochondria and chloroplasts, a fact explained by the theory of endosymbiosis. However, there is not available up to now a three-dimensional structure, of any of these proteins. As we already mentioned, early attempts were made in order to predict the putative transmembrane strands of mitochondrial porins using hydrophobicity analyses (Rauch & Moran 1994), and the Gibbs sampler (Mannella et al. 1996). Currently, some of the machine learning approaches discussed here, were shown capable of predicting plausible topologies for the mitochondrial VDAC and Tom40 (Liu et al. 2003b). Furthermore, Casadio and coworkers, using their NN-based predictor B2TMPRED, and performing a threading approach, were able to provide a three-dimensional model of the VDAC, using as template a bacterial porin (Casadio et al. 2002). A similar approach was followed to obtain a three-dimensional model of the voltage-independent and cation- selective DmPorin2 of the fruit fly *Drosophila melanogaster* (Aiello et al. 2004).

Recently, a semi-automatic strategy was proposed for finding the β-barrel outer membrane proteins of plant's chloroplasts (Schleiff et al. 2003). The method, named by the authors as BITS, combined the β-barrel score (BBS) of Wimley (Wimley 2002), isoelectric point calculations, the TargetP system (Emanuelsson et al. 2000) for predicting N-terminal signal sequences, and manual annotations. With this method, the authors collected a large set of putative chloroplast β-barrel outer membrane proteins of *Arabidopsis thaliana*, and finally proposed reliable topological models for four of them that have never been implicated with chloroplast outer membrane localization. This analysis shows that the proposed evolutionary relation between the semi-autonomous organelles (chloroplasts and mitochondria) and bacterial species may also imply common strategies for analyzing their outer membrane content.

Interestingly, there were early experimental sources of evidence suggesting the existence of an anion non-specific porin placed in the peroxisomal membranes (Reumann et al.1995). These proteins differ from the other previously characterized porins in the outer membrane of mitochondria or chloroplasts. They show a relatively small single-channel conductance and are strongly anion selective, thus, they were proposed to be essential for the passage of small metabolites through the membrane (Reumann et al. 1996). Although such porin-like activity has been observed in several plant peroxisomal membranes (Reumann et al. 1997), there is currently no evidence that this protein adopts a â-barrel structure. There is a hypothesis that the peroxisomes do not form de novo but grow and divide like mitochondria and chloroplasts, even though they do not possess a genome of their own. These speculations have lead to the development of an endosymbiotic theory for peroxisomes as well (Borst 1989). Further studies, both utilizing biochemical and bioinformatical approaches, are needed in order to investigate the properties of the putative porin-like proteins of the peroxisomal membrane and elucidate their structural characteristics and possible evolutionary relationships.

However, in the luck of any experimental data at atomic resolution, only weak hypotheses can be drown about the structure of eukaryotic β-barrel integral membrane proteins as predicted with methods tailored on their bacterial counterparts. Such methods should be revisited as soon as the first high-resolution three-dimensional structures become available.

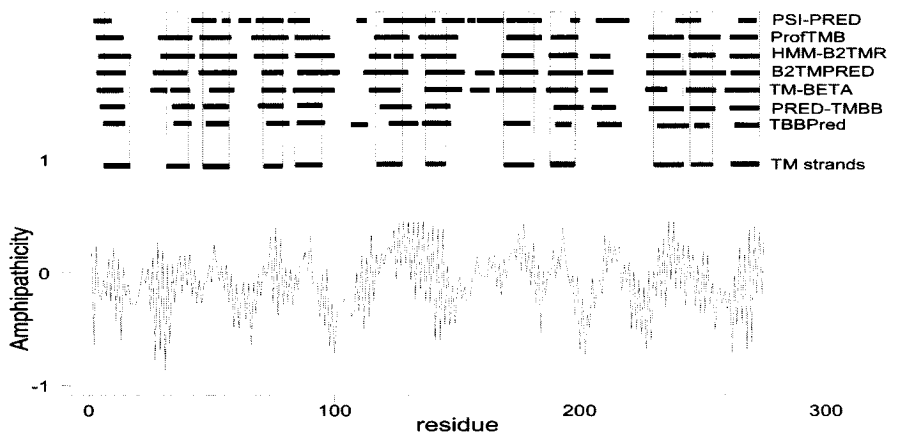

Figure 2. Graphical representation of the different predictions obtained on the translocator β-barrel domain of NalP from *N. meningitidis* (Oomen et al. 2004). The transmembrane β-strands obtained from PDB_TM (Tusnady et al. 2004) and the amphipathicity plot along the sequence, using the method of Vogel and Jahnig (Vogel & Jahnig 1986), are also illustrated. For the names of the different methods, see Table 1. We observe that most of the peaks in the plot correspond to TM strands, but also to a false positive strand, near residue 220, which is predicted by all methods except ProfTMB.

Evaluation of the Performance of Individual Methods

In order to evaluate the performance of the different methods presented so far, we have compiled two datasets. The first dataset consists of four proteins with recently solved three-dimensional structures that were not included (neither these nor a close homologue) in the training sets of the methods presented here. This dataset consists of the Translocator Domain Of Autotransporter Nalp from *Neisseria meningitidis* (PDB ID: 1UYN; Oomen et al. 2004), the Neisserial Surface Protein A (Nspa) of *N. meningitidis* (PDB ID: 1P4T; Vandeputte-Rutten et al. 2003), the Outer Membrane Enzyme Pagp of *E. coli* (PDB ID: 1MM4; Hwang et al. 2002) and the Outer Membrane Cobalamin Transporter (Btub) from *E. coli* (PDB ID: 1NQE;

Table 2. *Results of the blind test on transmembrane β-barrel topology prediction.*

Overall measures of accuracy of the different web-predictors, tested on the set of four newly crystallographically solved structures. The comparison is made against the observed transmembrane strands deposited in PDB_TM (Tusnady et al. 2004).
Q_β: Percentage of correctly predicted residues.
C_β: Matthews Correlation Coefficient.
SOV: Segment Overlap measure.
Correctly Predicted Topologies: Proteins with correctly predicted strand localization and loop orientation.
Correct Barrel Size: Proteins with correctly predicted number of transmembrane strands, allowing the inclusion of one shifted strand prediction per protein.

	Qb	Cb	SOV	Correctly Correct Barrel	Predicted Size
B2TMPRED	0.705	0.449	0.727	1	2
HMM-B2TMR	0.802	0.643	0.884	3	3
PRED-TMBB	0.843	0.689	0.91	2	4
ProfTMB	0.766	0.569	0.837	3	3
TBBPred	0.754	0.509	0.69	0	0
TM-BETA	0.669	0.363	0.681	0	0
PSI-PRED	0.727	0.486	0.678	0	0

Chimento et al. 2003). On these proteins we performed a blind test, predicting their transmembrane β-strands, using the following methods: PRED-TMBB, B2TMPRED, HMM-B2TMR, TM-BETA, ProfTMB, TBBPred and PSI-PRED. Detailed results for the *N. meningitidis* Translocator Domain Of Autotransporter Nalp are illusrated in figure 2.

For the transmembrane strand predictions, we evaluated the methods using the well-known SOV (measure of the segment's overlap), which is considered to be the most reliable measure for evaluating the performance of secondary structure prediction methods (Zemla et al. 1999). We also used the total number of correctly predicted topologies, i.e. when both the strands' localization and the loops' orientation have been predicted correctly, and the correctly predicted barrel size i.e. when the correct number of strands has been predicted, with no more than one mismatch (shifted prediction). It should be noted that only the HMM-based predictors report the full topology, hence for the NN-based predictors we count as correctly predicted topology a prediction where all the strands are correctly located. As measures of the per residue accuracy (Baldi et al. 2000a), we used both the total fraction of the correctly predicted residues (Q_β) in a two-state model (transmembrane versus non-transmembrane) and the well known Matthews Correlation Coefficient (C_β). The results of this test are summarized in table 2. The comparison is performed against the transmembrane strands that are reported in the PDB_TM database entries (Tusnady et al. 2004). From table 2, it is obvious that HMM-based methods (HMM-B2TMR, PRED-TMBB and ProfTMB) perform better than the NN- and SVM-based methods (TM-BETA, TBBPred and B2TMPRED). Furthermore, the refined all-purpose secondary structure prediction methods such as PSI-PRED perform comparably to the β-barrel specific NN-based methods, even though the former sometimes predict strands with non-realistic lengths.

The superiority of the PRED-TMBB method based on the measures of per-residue and per-segment accuracy is justified, since PRED-TMBB is the only method trained on the transmembrane part of the strands, while all the other methods are trained according to the PDB annotations. Thus, it is trained to find only the transmembrane part of the strand and not the whole strands, which in some cases protrude far away to the extracellular matrix. ProfTMB and HMM-B2TMR provide better predictions for the overall topology of the proteins (3 out of 4 cases). However, PRED-TMBB predicts the correct barrel size better (4 out of 4), even though it uses single sequence information.

When it comes to proteins with newly solved structures, a strong bias may arise mainly for two reasons. Firstly, the presence of a protein that declines from those used for training in some of the structural features (e.g. a protein with a very small shear number corresponding to sheets unusually longer), and secondly, the presence of a protein with a strong bias in its amino acid composition. It seems that this is not the case here; nevertheless, we will try to address this issue later (see Historical Prospective Study section).

The second independent test set consists of 68 recently characterized outer membrane proteins, which were collected after an extensive literature search. This set consists of sequences with low sequence similarity to those used to train the different methods, and is used to test the ability of the methods to correctly discriminate novel outer membrane proteins encoded in the genomes of Gram-negative bacteria. These proteins were identified by performing searches in articles indexed in PUBMED (for the last 2-3 years period), containing keywords such as "novel porin", "novel outer membrane protein", "porin activity", "cloning", "characterisation" of "outer membrane protein" or "porin". By reading the relevant abstracts or papers, we were able to retrieve the corresponding sequences in public databases such as GENBANK or TREMBL. Sequence fragments and clear cases of outer membrane lipoproteins were discarded. Using this reference set, we tested the following methods: PRED-TMBB, ProfTMB (with a cutoff of 10), BOMP, MCMBB, PSORT-B and the method of Wimley.

Table 3. Assessment of different predictors in the task of β-barrel integral outer membrane protein discrimination.

Evaluation of the different available methods used for discrimination of the β-barrel integral outer membrane proteins. BLAST results, correspond to a significant hit, to the database of annotated β-barrel outer membrane proteins used in Berven et al. 2004.

Method	Correctly Predicted OMPs (%)
BOMP	41 (65.1%)
MCMBB	49 (77.8%)
PRED-TMBB	56 (88.9%)
ProfTMB	43 (68.3%)
PSORT-B	22 (34.9%)
Wimley	46 (73.0%)
BLAST	13 (20.6%)

From the initial set of 68 protein sequences, we removed after careful manual inspection 5 sequences that clearly did not correspond to β-barrel membrane proteins. For example, the SwissProt/TrEMBL (Boeckmann et al. 2003) entry with Accession Number (AC) P72122, which is incorrectly annotated as "Outer membrane protein C", is predicted to contain α-helical transmembrane segments. The former annotation is valid for the sequence entry with AC P27121. After 'cleaning' the data-set, we ended up with a set of 63 putative outer membrane β-barrel proteins. For the evaluation of the different methods, we report the total number and the fraction of the correctly predicted outer membrane proteins, and the results are listed in table 3. Keeping in mind that even with the reduction of the set some false positives may still be present, we can draw some general conclusions as follows. PSORT-B is the most conservative in its predictions, since it detects a significant lower fraction of β-barrels. On the other hand, PRED-TMBB seems to predict the largest number of β-barrels, but it is known that when this method is applied without the filtering steps mentioned earlier, it is prone to over-prediction. The rest of the methods seem to perform comparably, but they do not predict the same set of proteins as β-barrels.

In order to perform an efficient and more accurate analysis, there is a need for well-annotated sets of positive examples (β-barrel proteins) and negative examples (globular proteins). However, at this point the task of finding well-annotated sets of both β-barrel proteins and globular proteins that did not participate (neither them nor a close homologue) in the sets used for training any of the above-mentioned algorithms is a difficult task, and out of the scope of the current work. Thus, this blind test is appropriate only to uncover general trends in the performance of the algorithms used for discrimination. More thorough statistical analysis may be feasible in the near future when large datasets will be annotated using both algorithmic and biochemical methods.

Future Directions

We have seen so far that the increase in the number of available structures is followed by an increase in the predictive accuracy of the methods. As the number of the available crystallographically solved

Table 4. *The non-redundant data set used in the "historical prospective" study.*

This non-redundant data set consists of 18 outer membrane proteins. Proteins are ranked by the year of publication of the corresponding high-resolution three-dimensional structure.

Protein name	Number of β-strands	PDB ID	Year of publication	Organism
Porin	16	2POR	1992	*Rhodobacter capsulatus*
Porin	16	1PRN	1994	*Rhodobacter blasticus*
OmpF	16	2OMF	1995	*Escherichia coli*
Sucrose porin	18	1A0S	1997	*Salmonella typhimurium*
Maltoporin	18	2MPR	1997	*Salmonella typhimurium*
FhuA	22	2FCP	1998	*Escherichia coli*
FepA	22	1FEP	1999	*Escherichia coli*
OmpLA	12	1QD5	1999	*Escherichia coli*
OmpX	8	1QJ8	1999	*Escherichia coli*
OmpA	8	1QJP	1999	*Escherichia coli*
Omp32	16	1.00E+54	2000	*Comamonas Acidovorans*
OmpT	10	1I78	2001	*Escherichia coli*
FecA	22	1KMO	2001	*Escherichia coli*
OpcA	10	1K24	2002	*Neisseria meningitidis*
Pagp	8	1MM4	2002	*Escherichia coli*
BtuB	22	1NQE	2003	*Escherichia coli*
NspA	8	1P4T	2003	*Neisseria meningitidis*
Nalp	12	1UYN	2004	*Neisseria meningitidis*

structures continues to increase, we will have even more data that could be used for training the predictors. Since β-barrel membrane proteins share little sequence similarity, solving the structure of a new porin with low sequence similarity with the sequences of porins of already known structure might provide some useful information for the improvement of existing predictive algorithms. On the other hand, solving the structure of a protein belonging to a family of β-barrel outer membrane proteins with no structure currently available would prove to be even more helpful towards predicting the structure of β-barrels. Examples of such cases, have been observed in the past, considering the recently solved structures of NalP (Oomen et al. 2004), NspA (Vandeputte-Rutten et al. 2003), PagP (Hwang et al. 2002) and BtuB (Chimento et al. 2003).

With the large number of completed or ongoing genome sequencing projects and the large effort spent on structural genomics projects, we have reasons to believe that the number of the available structures will continue to increase in the near future. We could speculate that this increase will follow a rate similar to that of water-soluble proteins two decades ago (Rees 2003). This will also be facilitated by the advances in the techniques used for cloning, expressing and crystallizing integral membrane proteins (Bannwarth & Schulz 2003). The increase in the number of available structures will have a direct impact on the quality of algorithms used for prediction, as different classes of algorithms require datasets of different sizes in order to be trained effectively. For example, a Neural Network method trained on 3 structures would be clearly over-fitted since it requires the adjustment of approximately 600 free parameters while the training set will only have about 1200 amino acid residues.

For the sake of argument, we have conducted a "historical prospective" study on the prediction performance on β-barrel membrane proteins. We ranked the structures of the 18 non-redundant protein sequences, according to the year of publication, as shown in table 4. Afterwards, we created 8 virtual datasets corresponding to years 1997 to 2004, including in each dataset all the structures published up to that year. Thus, 5 structures are included in the set for 1997, 6 structures for 1998 and continuing in the same manner, we finally obtained the complete set of 18 representative sequences for the year 2004. For each such set, we trained a HMM with a similar architecture with PRED-TMBB (Bagos et al. 2004b), and we evaluated the performance on the jacknife test, i.e. removing a protein from the training set, training the model with the remaining proteins and performing the test on the protein removed. Given this, and the fact that the sequences do not show any significant similarity (no more than 30% identities in a BLAST alignment; Altschul et al. 1997), the results of the study were, approximately, what would have been observed if such an algorithm was applied at that

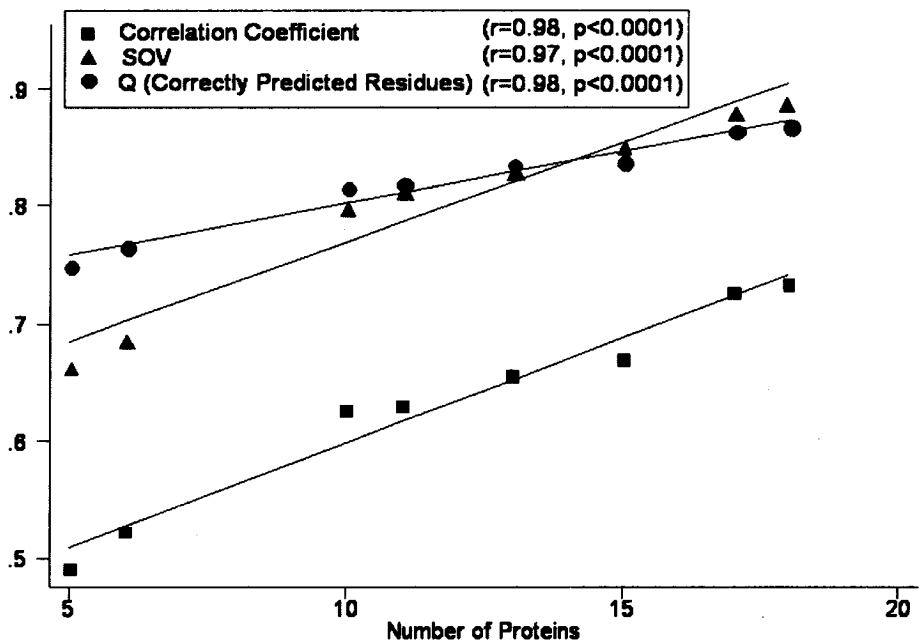

Figure 3. Plot of the measures of predictive performance against the number of sequences used for training in the "historical prospective" study. Squares: Mathews Correlation Coefficient. Triangles: Segment Overlap. Circles: Correctly predicted residues.
r: Pearson's linear correlation coefficient. p: p-value of the regression line.

particular time.

Figure 3 illustrates the results of this historical prospective study, namely the correlation coefficient (C_β), the correctly predicted residues (Q_β) and the segment overlap (SOV) against the structures used for training. By fitting a linear regression line, we observe that the accuracy of the prediction increases constantly with the increase in the number of structures used for training. We can safely assume that the same will continue to happen in the near future, as the number of available structures continues to grow until a plateau is reached.

One of the most effective ways to increase the prediction accuracy is to combine individual predictors. This strategy is well documented in the case of α-helical membrane proteins, where the so-called "Consensus Prediction" methods have been proved superior to the individual predictors. Such approaches have already been proposed by several research groups (Promponas et al. 1999, Nilsson et al. 2000, Taylor et al. 2003, Xia et al. 2004). Another combined approach, slightly different from the Consensus Prediction, is the use of ENSEMBLE learning techniques (Perrone & Cooper 1993, Sollich & Krogh 1996). This kind of machine learning methods are suitable for combining individual predictors, taking advantage of the points that the predictors disagree, instead of looking at where the predictors agree, in this way maximizing the information content of the individual predictors. Such an ENSEMBLE system has been used by Martelli and colleagues for the prediction of α-helical membrane proteins, combining the predictions of two different HMMs and a NN method (Martelli et al. 2003). Considering the fact that we already possess a variety of algorithms for predicting the transmembrane strands of β-barrel outer membrane proteins, it is tempting to speculate that by combining those individual predictors effectively, we might be able to improve further the prediction accuracy.

Finally, as already discussed, given the increased number of available structures, we may face the need to develop novel, entirely different algorithms or models in order to improve the prediction performance. The most obvious way to do this is to exploit structural features of the β-barrels that have not been used before in prediction methods. One such important feature is the coupling of the adjacent β-strands of the barrel, forming hydrogen bonds that stabilize the whole structure. This issue has been addressed in the past in the case of β-sheet containing soluble proteins. Krogh and Riis used a Neural Network method utilizing two–independently moving- sliding windows along the sequence, predicting each time if the two central amino acids form a hydrogen bond or not (Krogh & Riis 1996). With this method, they managed not only to improve the prediction accuracy for β-strands by 1%, but also to predict the coupling of the predicted strands, making a step towards the prediction of tertiary structure from sequence alone. Another, probably more sophisticated use of NNs was proposed by Baldi and colleagues (Baldi et al. 2000b), utilizing the concept of Bidirectional Recurrent Neural Networks (BRNNs). BRNNs have also been used for protein secondary structure prediction with excellent results (Pollastri et al. 2002), and their application for matching the β-sheet partners yielded an accuracy of approximately 84% for soluble proteins.

Perhaps the most advanced computational method that could be used in the future for predicting both the location of the transmembrane strands and their connectivity is that of the Stochastic Context Free Grammars (SCFGs). SCFGs are commonly used in molecular biology for predicting the secondary structure of tRNA (Sakakibara et al. 1994, Lefebvre 1995, Knudsen & Hein 1999, Knudsen & Hein 2003) and rRNA (Brown 2000), where base pairing introduces long range interactions that may not be captured by other machine learning approaches. In the field of protein modelling, SCFGs were first applied by Mamitsuka and Abe (Mamitsuka & Abe 1994), for predicting the location and the connectivity of the β-strands in water-soluble proteins with considerable success. The main disadvantage of SCFGs, compared to the already mentioned NNs and HMMs, is the computational complexity of the algorithms used for training and parsing (testing) these models. However, since the available computational power is increasing rapidly, it would be no surprise to find in the near future reliable predictors for the β-barrel outer membrane proteins based on SCFGs.

Conclusions

We formulated the problem of predicting the transmembrane β-barrel proteins and presented the general framework currently in use for the topology prediction of β-barrel outer membrane proteins. Structural characteristics of β-barrel outer membrane proteins were highlighted, emphasizing on those features that have been proved useful in the implementation of prediction algorithms. We discussed historical aspects of the evolution of different methods, in terms of both the available data and the algorithms used for prediction. We also presented the issue of discrimination of outer membrane proteins encoded in complete genomes, along with the prediction of their topology, and we presented all the available predictors used for both

purposes. After a comparison of the different methods we concluded that the HMM-based predictors (HMM-B2TMR, ProfTMB and PRED-TMBB) perform significantly better than the NN- and SVM-based methods, and also better than the all-purpose secondary structure prediction algorithms, thus their use should be preferred. We emphasized on the need for finding ways to improve the performance of the methods, and we provided evidence that (at least for the time being) newly crystallographically determined structures will continue to improve the performance of the already available methods. Furthermore, we proposed new directions that could be followed to develop new methods to reliably predict the topology of β-barrel integral membrane proteins. As previously stated, threading or homology modelling studies might yield interesting insight into the structural features of a wide range of proteins belonging to this class until hard to obtain experimental data become available (Casadio et al. 2003b).

Large-scale predictions will definitely facilitate the identification of surface exposed regions in predicted outer membrane proteins and complement or guide laboratory analysis of key bacterial target proteins. Such an approach might lead, for example, to more efficient design of microbial cell-surface display systems (Lee et al. 2003). Additionally, bioinformatics methods might be applied in combination with experimental low-resolution structural information (e.g. Infrared Dichroism; Marsh 2000) to yield more accurate models in the absence of detailed atomic structures of outer membrane proteins.

Useful applications of prediction schemes might be to aid in the successful design of channels with desired properties, where porin-like structures may be used as 'Trojan Horses' to control the permeability of selected membranes or as sensitive biosensors. Additionally, predictions might significantly aid in the efficient and selective blockade of relevant channels in cases of pathogenic Gram-negative bacteria.

Apart from the obvious practical applications of computerized methods for the discrimination, topology prediction and modelling of bacterial outer membrane proteins, such methods will surely prove complementary to experimental methods for the elucidation of the biogenesis of outer membrane proteins.

Acknowledgements

The authors would like to thank the PINSA-B editor Professor Muralidhar for his kind invitation and help throughout the submission process, and the anonymous referees for their valuable comments and constructive criticism. We are especially grateful to Dr. William Wimley, for kindly performing the requested predictions on the dataset of newly discovered outer membrane proteins and for useful discussion on the predictions. We also wish to thank all those experimentalists who enable the development of computational methods by depositing their experimental data in publicly available databases, those people responsible for maintenance and curation of these data, and the numerous Bioinformatics groups for developing predictive methods and making them available for public use.

References

Aiello R, Messina A, Schiffler B, Benz R, Tasco G, Casadio R and De Pinto V 2004 Functional characterization of a second porin isoform in *Drosophila melanogaster*. DmPorin2 forms voltage-independent cation-selective pores; *J. Biol. Chem.* **279** 25364-25373

Altschul S F, Madden T L, Schaffer A A, Zhang J, Zhang Z, Miller W and Lipman D J 1997 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs; *Nucleic Acids Res.* **25** 3389-3402

Andersen G R, Thirup S, Spremulli L L and Nyborg J 2000 High resolution crystal structure of bovine mitochondrial EF-Tu in complex with GDP; *J. Mol. Biol.* **297** 421-436

Argos P, Rao J K and Hargrave P A 1982 Structural prediction of membrane-bound proteins; *Eur. J. Biochem.* **128** 565-575

Asai K, Hayamizu S and Handa K 1993 Prediction of protein secondary structure by the hidden Markov model; *Comput. Appl. Biosci.* **9** 141-146

Axelson H 2004 The Notch signaling cascade in neuroblastoma: role of the basic helix-loop-helix proteins HASH-1 and HES-1; *Cancer Lett.* **204** 171-178

Bagos P G, Liakopoulos T D and Hamodrakas J S 2004a Finding beta-barrel outer membrane proteins with a Markov Chain model; *WSEAS Transactions on Biology and Biomedicine* **1** 186-189

------,------, Spyropoulos I C and Hamodrakas S J 2004b A Hidden Markov Model method, capable of predicting and discriminating beta-barrel outer membrane proteins; *BMC Bioinformatics* **5**

------,------, ------ and ------ 2004c PRED-TMBB: a web server for predicting the topology of beta-barrel outer membrane proteins; *Nucleic Acids Res.* **32** W400-W404

Baldi P, Brunak S, Chauvin Y, Andersen C A and Nielsen H 2000a Assessing the accuracy of prediction algorithms for classification: an overview; *Bioinformatics* **16** 412-424

------, Pollastri G, Andersen C A and Brunak S 2000b Matching protein beta-sheet partners by feedforward and recurrent neural networks; *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **8** 25-36

Bannwarth M and Schulz G E 2003 The expression of outer membrane proteins for crystallization; *Biochim. Biophys. Acta* **1610** 37-45

Baum L 1972 An inequality and associated maximization technique in statistical estimation for probalistic functions of Markov processes; *Inequalities* **3** 1-8

Berman H M, Battistuz T, Bhat T N, Bluhm W F, Bourne P E, Burkhardt K, Feng Z, Gilliland G L, Iype L, Jain S, Fagan P, Marvin J, Padilla D, Ravichandran V, Schneider B, Thanki N, Weissig H, Westbrook J D and Zardecki C 2002 The Protein Data Bank; *Acta Crystallogr. D Biol. Crystallogr.* **58** 899-907

Berven F S, Flikka K, Jensen H B and Eidhammer I 2004 BOMP: a program to predict integral b-barrel outer membrane proteins encoded within genomes of Gram-negative bacteria; *Nucleic Acids Res.* **32** W394-W399

Bigelow H R, Petrey D S, Liu J, Przybylski D and Rost B 2004 Predicting transmembrane beta-barrels in proteomes; *Nucleic Acids Res.* **32** 2566-2577

Bishop C M 1995 Neural Networks for Pattern Recognition;(Oxford, England, Oxford University Press)

------, Walkenhorst W F and Wimley W C 2001 Folding of beta-sheets in membranes: specificity and promiscuity in peptide model systems; *J. Mol. Biol.* **309** 975-988

Bitter W 2003 Secretins of Pseudomonas aeruginosa: large holes in the outer membrane; *Arch. Microbiol.* **179** 307-314

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

Borst P 1989 Peroxisome biogenesis revisited; *Biochim. Biophys. Acta* **1008** 1-13

Braun V, Braun M and Killmann H 2000 Iron transport in *Escherichia coli*. Crystal structure of FhuA, an outer membrane iron and antibiotic transporter; *Adv. Exp. Med. Biol.* **485** 33-43

Brokx S J, Ellison M, Locke T, Bottorff D, Frost L and Weiner J H 2004 Genome-Wide Analysis of Lipoprotein Expression in Escherichia coli MG1655; *J. Bacteriol.* **186** 3254-3258

Brown M P 2000 Small subunit ribosomal RNA modeling using stochastic context-free grammars; *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **8** 57-66

Casadio R, Jacoboni I, Messina A and De Pinto V 2002 A 3D model of the voltage-dependent anion channel (VDAC); *FEBS Lett.* **520** 1-7

------, Fariselli P, Finocchiaro G and Martelli P L 2003a Fishing new proteins in the twilight zone of genomes: the test case of outer membrane proteins in Escherichia coli K12, Escherichia coli O157:H7, and other Gram-negative bacteria; *Protein Sci.* **12** 1158-1168

------, Fariselli P and Martelli P L 2003b In silico prediction of the structure of membrane proteins: is it feasible?; *Brief Bioinform.* **4** 341-348

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

Chimento D P, Mohanty A K, Kadner R J and Wiener M C 2003 Substrate-induced transmembrane signaling in the cobalamin transporter BtuB; *Nat. Struct. Biol.* **10** 394-401

Cochran J R, Aivazian D, Cameron T O and Stern L J 2001 Receptor clustering and transmembrane signaling in T cells; *Trends Biochem. Sci.* **26** 304-310

Cowan S W, Schirmer T, Rummel G, Steiert M, Ghosh R, Pauptit R A, Jansonius J N and Rosenbusch J P 1992 Crystal structures explain functional properties of two E. coli porins; *Nature* **358** 727-733

Cowell R G, Dawid A P, Lauritzen S L and Spiegelhalter D J 1999 Probabilistic Networks and Expert Systems; (New York, Springer-Verlag)

Cuff J A, Clamp M E, Siddiqui A S, Finlay M and Barton G J 1998 JPred: a consensus secondary structure prediction server; *Bioinformatics* **14** 892-893

Danelon C, Suenaga A, Winterhalter M and Yamato I 2003 Molecular origin of the cation selectivity in OmpF porin: single channel conductances vs. free energy calculation; *Biophys. Chem.* **104** 591-603

DeLano W L 2003 PyMol, DeLano Scientific;

Demeler B and Zhou G W 1991 Neural network optimization for E. coli promoter prediction; *Nucleic Acids Res.* **19** 1593-1599

Diederichs K, Freigang J, Umhau S, Zeth K and Breed J 1998 Prediction by a neural network of outer membrane beta-strand protein topology; *Protein Sci.* **7** 2413-2420

Durbin R, Eddy S R, Krogh A and Mithison G 1998 Biological sequence analysis, probabilistic models of proteins and nucleic acids.;(Cambridge University Press)

Eddy S R 1995 Multiple alignment using hidden Markov models; *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **3** 114-120

------1998 Profile hidden Markov models; *Bioinformatics* **14** 755-763

Eisenberg D, Weiss R M and Terwilliger T C 1984 The hydrophobic moment detects periodicity in protein hydrophobicity; *Proc. Natl. Acad. Sci. USA* **81** 140-144

Emanuelsson O, Nielsen H, Brunak S and von Heijne G 2000 Predicting subcellular localization of proteins based on their N-terminal amino acid sequence; *J. Mol. Biol.* **300** 1005-1016

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

Faller M, Niederweis M and Schulz G E 2004 The structure of a mycobacterial outer-membrane channel; *Science* **303** 1189-1192

Farber R, Lapedes A and Sirotkin K 1992 Determination of eukaryotic protein coding regions using neural networks and information theory; *J. Mol. Biol.* **226** 471-479

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

Gardy J L, Spencer C, Wang K, Ester M, Tusnady G E, Simon I, Hua S, deFays K, Lambert C, Nakai K and Brinkman F S 2003 PSORT-B: Improving protein subcellular localization prediction for Gram-negative bacteria; *Nucleic Acids Res.* **31** 3613-3617

Gnanasekaran T V, Peri S, Arockiasamy A and Krishnaswamy S 2000 Profiles from structure based sequence alignment of porins can identify beta stranded integral membrane proteins; *Bioinformatics* **16** 839-842

Gromiha M M and Ponnuswamy P K 1993 Prediction of transmembrane beta-strands from hydrophobic characteristics of proteins; *Int. J. Pept. Protein Res.* **42** 420-431

------, Majumdar R and Ponnuswamy P K 1997 Identification of membrane spanning beta strands in bacterial porins; *Protein Eng.* **10** 497-500

Gromiha M M, Ahmad S and Suwa M 2004 Neural network-based prediction of transmembrane beta-strand segments in outer membrane proteins; *J. Comput. Chem.* **25** 762-767

Hwang P M, Choy W Y, Lo E I, Chen L, Forman-Kay J D, Raetz C R, Prive G G, Bishop R E and Kay L E 2002 Solution structure and dynamics of the outer membrane enzyme PagP by NMR; *Proc. Natl. Acad. Sci. USA* **99** 13560-13565

Jacoboni I, Martelli P L, Fariselli P, De Pinto V and Casadio R 2001 Prediction of the transmembrane regions of beta-barrel membrane proteins with a neural network-based predictor; *Protein Sci.* **10** 779-787

Jeanteur D, Lakey J H and Pattus F 1991 The bacterial porin superfamily: sequence alignment and structure prediction; *Mol. Microbiol.* **5** 2153-2164

Jones D T, Taylor W R and Thornton J M 1994 A model recognition approach to the prediction of all-helical membrane protein structure and topology; *Biochemistry* **33** 3038-3049

Juncker A S, Willenbrock H, von Heijne G, Brunak S, Nielsen H and Krogh A 2003 Prediction of lipoprotein signal peptides in Gram-negative bacteria; *Protein Sci.* **12** 1652-1662

Keefe L J, Sondek J, Shortle D and Lattman E E 1993 The alpha aneurism: a structural motif revealed in an insertion mutant of staphylococcal nuclease; *Proc. Natl. Acad. Sci. USA* **90** 3275-3279

Klare J P, Gordeliy V I, Labahn J, Buldt G, Steinhoff H J and Engelhard M 2004 The archaeal sensory rhodopsin II/ transducer complex: a model for transmembrane signal transfer; *FEBS Lett.* **564** 219-224

Knudsen B and Hein J 1999 RNA secondary structure prediction using stochastic context-free grammars and evolutionary history; *Bioinformatics* **15** 446-454

------ and Hein J 2003 Pfold: RNA secondary structure prediction using stochastic context-free grammars; *Nucleic Acids Res.* **31** 3423-3428

Koronakis V, Sharff A, Koronakis E, Luisi B and Hughes C 2000 Crystal structure of the bacterial membrane protein TolC central to multidrug efflux and protein export; *Nature* **405** 914-919

Krapp S, Kelly G, Reischl J, Weinzierl R O and Matthews S 1998 Eukaryotic RNA polymerase subunit RPB8 is a new relative of the OB family; *Nat. Struct. Biol.* **5** 110-114

Krogh A 1994 Hidden Markov models for labelled sequences.; *Proceedings of the12th IAPR International Conference on Pattern Recognition* 140-144

------, Mian I S and Haussler D 1994 A hidden Markov model that finds genes in E. coli DNA; *Nucleic Acids Res.* **22** 4768-4778

------ and Riis S K 1996 Predicting beta sheets in proteins; *In* Advances in Neural information Processing Systems 8. D.S. Touretzky, M.C. Mozer and M.E. Hasselmo, ed.: Eds.MIT Press) 917-923

------ 1997 Two methods for improving performance of an HMM and their application for gene finding; *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **5** 179-186

------ and Riis S K 1999 Hidden neural networks; *Neural Comput.* **11** 541-563

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

Kurucz E, Zettervall C J, Sinka R, Vilmos P, Pivarcsi A, Ekengren S, Hegedus Z, Ando I and Hultmark D 2003 Hemese, a hemocyte-specific transmembrane protein, affects the cellular immune response in Drosophila; *Proc. Natl. Acad. Sci. USA* **100** 2622-2627

Kyogoku Y, Fujiyoshi Y, Shimada I, Nakamura H, Tsukihara T, Akutsu H, Odahara T, Okada T and Nomura N 2003 Structural genomics of membrane proteins; *Acc. Chem. Res.* **36** 199-206

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

Lee S Y, Choi J H and Xu Z 2003 Microbial cell-surface display; *Trends Biotechnol.* **21** 45-52

Lefebvre F 1995 An optimized parsing algorithm well suited to RNA folding; *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **3** 222-230

Liu J, Rosenberg E and Nikaido H 1995 Fluidity of the Lipid Domain of Cell Wall From Mycobacterium chelonae; *PNAS* **92** 11254-11258

Liu Q, Zhu Y, Wang B and Li Y 2003a Identification of beta-barrel membrane proteins based on amino acid composition properties and predicted secondary structure; *Comput. Biol. Chem.* **27** 355-361

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

Loll P J 2003 Membrane protein structural biology: the high throughput challenge; *J. Struct. Biol.* **142** 144-153

Mamitsuka H and Abe N 1994 Predicting location and structure of beta-sheet regions using stochastic tree grammars; *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **2** 276-284

Mannella C A, Neuwald A F and Lawrence C E 1996 Detection of likely transmembrane beta strand regions in sequences of mitochondrial pore proteins using the Gibbs sampler; *J. Bioenerg. Biomembr.* **28** 163-169

------ 1997 On the structure and gating mechanism of the mitochondrial channel, VDAC; *J. Bioenerg. Biomembr.* **29** 525-531

Marsh D 2000 Infrared dichroism of twisted beta-sheet barrels. The structure of E. coli outer membrane proteins; *J. Mol. Biol.* **297** 803-808

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

------, Fariselli P and Casadio R 2003 An ENSEMBLE machine learning approach for the prediction of all-alpha membrane proteins; *Bioinformatics* **19 Suppl 1** i205-211

McGuffin L J, Bryson K and Jones D T 2000 The PSIPRED protein structure prediction server; *Bioinformatics* **16** 404-405

McLachlan A D 1979 Gene duplications in the structural evolution of chymotrypsin; *J. Mol. Biol.* **128** 49-79

Menzaghi F, Behan D P and Chalmers D T 2002 Constitutively activated G protein-coupled receptors: a novel approach to CNS drug discovery; *Curr. Drug. Targets CNS Neurol. Disord.* **1** 105-121

Miyazawa A, Fujiyoshi Y and Unwin N 2003 Structure and gating mechanism of the acetylcholine receptor pore; *Nature* **423** 949-955

Morona R, Kramer C and Henning U 1985 Bacteriophage receptor area of outer membrane protein OmpA of Escherichia coli K-12; *J. Bacteriol.* **164** 539-543

Murzin A G, Lesk A M and Chothia C 1994a Principles determining the structure of beta-sheet barrels in proteins. I. A theoretical analysis; *J. Mol. Biol.* **236** 1369-1381

------, Lesk A M and Chothia C 1994b Principles determining the structure of beta-sheet barrels in proteins. II. The observed structures; *J. Mol. Biol.* **236** 1382-1400

Natt N K, Kaur H and Raghava G P 2004 Prediction of transmembrane regions of beta-barrel proteins using ANN- and SVM-based methods; *Proteins* **56** 11-18

Neuwald A F, Liu J S and Lawrence C E 1995 Gibbs motif sampling: detection of bacterial outer membrane protein repeats; *Protein Sci.* **4** 1618-1632

Nielsen H, Engelbrecht J, Brunak S and von Heijne G 1997 A neural network method for identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites; *Int. J. Neural Syst.* **8** 581-599

------ and Krogh A 1998 Prediction of signal peptides and signal anchors by a hidden Markov model; *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **6** 122-130

------, Brunak S and von Heijne G 1999 Machine learning approaches for the prediction of signal peptides and other protein sorting signals; *Protein Eng.* **12** 3-9

Nilsson J, Persson B and von Heijne G 2000 Consensus predictions of membrane protein topology; *FEBS Lett.* **486** 267-269

Nouwen N, Ranson N, Saibil H, Wolpensinger B, Engel A, Ghazi A and Pugsley A P 1999 Secretin PulD: association with pilot PulS, structure, and ion-conducting channel formation; *Proc. Natl. Acad. Sci. USA* **96** 8173-8177

Oomen C J, Van Ulsen P, Van Gelder P, Feijen M, Tommassen J and Gros P 2004 Structure of the translocator domain of a bacterial autotransporter; *Embo. J.* **23** 1257-1266

Paquet J Y, Vinals C, Wouters J, Letesson J J and Depiereux E 2000 Topology prediction of Brucella abortus Omp2b and Omp2a porins after critical assessment of transmembrane beta strands prediction by several secondary structure prediction methods; *J. Biomol. Struct. Dyn.* **17** 747-757

Paschen S A, Waizenegger T, Stan T, Preuss M, Cyrklaff M, Hell K, Rapaport D and Neupert W 2003 Evolutionary conservation of biogenesis of beta-barrel membrane proteins; *Nature* **426** 862-866

Pasquier C and Hamodrakas S J 1999 An hierarchical artificial neural network system for the classification of transmembrane proteins; *Protein Eng.* **12** 631-634

------, Promponas V J, Palaios G A, Hamodrakas J S and Hamodrakas S J 1999 A novel method for predicting transmembrane segments in proteins based on a statistical analysis of the SwissProt database: the PRED-TMR algorithm; *Protein Eng* **12** 381-385

------, ------ and Hamodrakas S J 2001 PRED-CLASS: cascading neural networks for generalized protein classification and genome-wide applications; *Proteins* **44** 361-369

Perrone M and Cooper L, Eds. (1993). *In* When networks disagree: ensemble methods for hybrid neural networks.; Neural networks for speech and image processing. (London: Chapman and Hall)

Pollastri G, Przybylski D, Rost B and Baldi P 2002 Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles; *Proteins* **47** 228-235

Prince S M, Achtman M and Derrick J P 2002 Crystal structure of the OpcA integral membrane adhesin from Neisseria meningitidis; *Proc. Natl. Acad. Sci. USA* **99** 3417-3421

Promponas V J, Palaios G A, Pasquier C M, Hamodrakas J S and Hamodrakas S J 1999 CoPreTHi: a Web tool which combines transmembrane protein segment prediction methods; *In Silico Biol.* **1** 159-162

Przybylski D and Rost B 2002 Alignments grow, secondary structure prediction improves; *Proteins* **46** 197-205

Qian N and Sejnowski T J 1988 Predicting the secondary structure of globular proteins using neural network models; *J. Mol. Biol.* **202** 865-884

Rabiner L 1989 A tutorial on hidden Markov models and selected applications in speech recognition; *Proc. IEEE* **77** 257-286

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

Rauch G and Moran O 1994 On the structure of mitochondrial porins and its homologies with bacterial porins; *Biochem. Biophys. Res. Commun.* **200** 908-915

------ and Moran O 1995 Prediction of polypeptide secondary structures analysing the oscillation of the hydropathy profile; *Comput. Methods Programs Biomed.* **48** 193-200

Rees D C 2003 Advances in protein chemistry. Preface; *Adv. Protein Chem.* **63** xi-xvi

Reinhardt A and Hubbard T 1998 Using neural networks for prediction of the subcellular location of proteins; *Nucleic Acids Res.* **26** 2230-2236

Reumann S, Maier E, Benz R and Heldt H W 1995 The membrane of leaf peroxisomes contains a porin-like channel; *J. Biol. Chem.* **270** 17559-17565

------ , Maier E, Benz R and Heldt H W 1996 A specific porin is involved in the malate shuttle of leaf peroxisomes; *Biochem. Soc. Trans.* **24** 754-757

Reumann S, Bettermann M, Benz R and Heldt H W 1997 Evidence for the Presence of a Porin in the Membrane of Glyoxysomes of Castor Bean; *Plant Physiol.* **115** 891-899

Rodriguez-Maranon M J, Bush R M, Peterson E M, Schirmer T and de la Maza L M 2002 Prediction of the membrane-spanning beta-strands of the major outer membrane protein of Chlamydia; *Protein Sci.* **11** 1854-1861

Rost B and Sander C 1993 Prediction of protein secondary structure at better than 70% accuracy; *J Mol. Biol.* **232** 584-599

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

------ 1996 PHD: predicting one-dimensional protein structure by profile-based neural networks; *Methods Enzymol.* **266** 525-539

------, Fariselli P and Casadio R 1996 Topology prediction for helical transmembrane proteins at 86% accuracy; *Protein Sci.* **5** 1704-1718

------ and Liu J 2003 The PredictProtein server; *Nucleic Acids Res.* **31** 3300-3304

Saier M H, Jr. 2000 A functional-phylogenetic classification system for transmembrane solute transporters; *Microbiol. Mol. Biol. Rev.* **64** 354-411

Sakakibara Y, Brown M, Hughey R, Mian I S, Sjolander K, Underwood R C and Haussler D 1994 Stochastic context-free grammars for tRNA modeling; *Nucleic Acids Res.* **22** 5112-5120

Schirmer T and Cowan S W 1993 Prediction of membrane-spanning beta-strands and its application to maltoporin; *Protein Sci.* **2** 1361-1363

Schleiff E, Eichacker L A, Eckart K, Becker T, Mirus O, Stahl T and Soll J 2003 Prediction of the plant beta-barrel proteome: a case study of the chloroplast outer envelope; *Protein Sci.* **12** 748-759

Schulz G E 2000 beta-Barrel membrane proteins; *Curr. Opin. Struct. Biol.* **10** 443-447

------ 2002 The structure of bacterial outer membrane proteins; *Biochim. Biophys. Acta* **1565** 308-317

------ 2003 Transmembrane beta-barrel proteins; *Adv. Protein Chem.* **63** 47-70

She R, Chen F, Wang K, Ester M, Gardy J L and Brinkman F S L 2003. Frequent Subsequence-Based Prediction of Outer Membrane Proteins; *In Proceedings of the* 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA

Snijder H J, Ubarretxena-Belandia I, Blaauw M, Kalk K H, Verheij H M, Egmond M R, Dekker N and Dijkstra B W 1999 Structural evidence for dimerization-regulated activation of an integral membrane phospholipase; *Nature* **401** 717-721

Sogo L F and Yaffe M P 1994 Regulation of mitochondrial morphology and inheritance by Mdm10p, a protein of the mitochondrial outer membrane; *J. Cell Biol.* **126** 1361-1373

Sollich P and Krogh A 1996 Learning with ensembles: How overfitting can be useful; *In* Advances in Neural Information Processing Systems 8. D.S. Touretzky, M.C. Mozer and M.E. Hasselmo, Ed.^ Eds.MIT Press): 190-196

Song L, Hobaugh M R, Shustak C, Cheley S, Bayley H and Gouaux J E 1996 Structure of Staphylococcal alpha-Hemolysin, a Heptameric Transmembrane Pore; *Science* **274** 1859-1865

Sonnhammer E L, von Heijne G and Krogh A 1998 A hidden Markov model for predicting transmembrane helices in protein sequences; *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **6** 175-182

Struyve M, Moons M and Tommassen J 1991 Carboxy-terminal phenylalanine is essential for the correct assembly of a bacterial outer membrane protein; *J. Mol. Biol.* **218** 141-148

Tamm L K, Arora A and Kleinschmidt J H 2001 Structure and assembly of beta-barrel membrane proteins; *J. Biol. Chem.* **276** 32399-32402

Taylor P D, Attwood T K and Flower D R 2003 BPROMPT: A consensus server for membrane protein prediction; *Nucleic Acids Res.* **31** 3698-3700

Tusnady G E, Dosztanyi Z and Simon I 2004 Transmembrane proteins in protein data bank: identification and classification; *Bioinformatics* doi:10.1093/bioinformatics/bth1340

Unwin N 1993 Nicotinic acetylcholine receptor at 9 A resolution; *J. Mol. Biol.* **229** 1101-1124

Vandeputte-Rutten L, Bos M P, Tommassen J and Gros P 2003 Crystal structure of Neisserial surface protein A (NspA), a conserved outer membrane protein with vaccine potential; *J. Biol. Chem.* **278** 24825-24830

Vapnik V N 1998 Statistical Learning Theory;(New York, Wiley-Interscience)

Vogel H and Jahnig F 1986 Models for the structure of outer-membrane proteins of Escherichia coli derived from raman spectroscopy and prediction methods; *J Mol. Biol.* **190** 191-199

Vogt J and Schulz G E 1999 The structure of the outer membrane protein OmpX from Escherichia coli reveals possible mechanisms of virulence; *Structure Fold Des.* **7** 1301-1309

Walian P, Cross T A and Jap B K 2004 Structural genomics of membrane proteins; *Genome Biol.* **5** 215

Wheelock M J and Johnson K R 2003 Cadherin-mediated cellular signaling; *Curr. Opin. Cell Biol.* **15** 509-514

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

Wimley W C and White S H 1996 Experimentally determined hydrophobicity scale for proteins at membrane interfaces; *Nat. Struct. Biol.* **3** 842-848

------ 2002 Toward genomic identification of beta-barrel membrane proteins: composition and architecture of known structures; *Protein Sci.* **11** 301-312

------ 2003 The versatile beta-barrel membrane protein; *Curr. Opin. Struct. Biol.* **13** 404-411

Xia J X, Ikeda M and Shimizu T 2004 ConPred elite: a highly reliable approach to transmembrane topology prediction; *Comput. Biol. Chem.* **28** 51-60

Zemla A, Venclovas C, Fidelis K and Rost B 1999 A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment; *Proteins* **34** 220-223

Zhai Y and Saier M H, Jr. 2002 The beta-barrel finder (BBF) program, allowing identification of outer membrane beta-barrel proteins encoded within prokaryotic genomes; *Protein Sci.* **11** 2196-2207

Zhang C and Kim S H 2000 A comprehensive analysis of the Greek key motifs in protein beta-barrels and beta-sandwiches; *Proteins* **40** 409-419

Zhang Q, Meitzler J C, Huang S and Morishita T 2000 Sequence polymorphism, predicted secondary structures, and surface-exposed conformational epitopes of Campylobacter major outer membrane protein; *Infect Immun.* **68** 5679-5689

Zhou X H, van der Helm D and Venkatramani L 1995 Binding characterization of the iron transport receptor from the outer membrane of Escherichia coli (FepA): differentiation between FepA and FecA; *Biometals* **8** 129-136