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

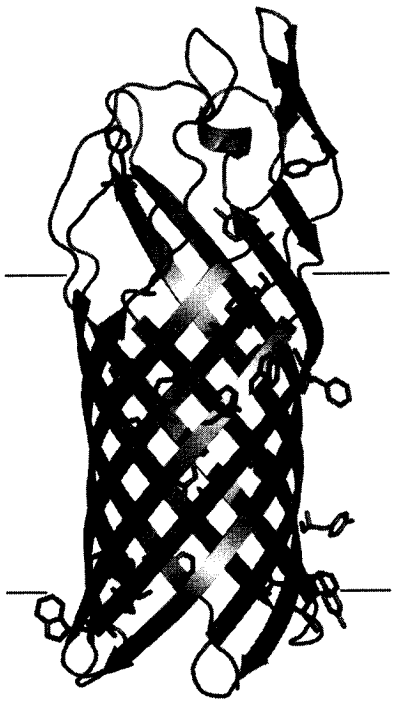

Figure 1. A ribbon diagram of the structure of the OpcA Outer Membrane Adhesin/Invasin from *Neisseria meningitidis* (PDB ID: 1K24; Prince et al. 2002). Aromatic side chains are represented as rods to illustrate the aromatic belts. The horizontal lines indicate the approximate position of the lipid bilayer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