Available online at www.sciencedirect.com
ScienceDirect
Methods 41 (2007) 460–474
METHODS
www.elsevier.com/locate/ymeth

Membrane protein prediction methods

Marco Punta [a,b], Lucy R. Forrest [a,b], Henry Bigelow [a,b], Andrew Kernytsky [a,b], Jinfeng Liu [a,b,c], Burkhard Rost [a,b,c,*]

[a] *Department of Biochemistry and Molecular Biophysics, Columbia University, 1130 St. Nicholas Ave., New York, NY 10032, USA*
[b] *Columbia University Center for Computational Biology and Bioinformatics (C2B2), 1130 St. Nicholas Ave., New York, NY 10032, USA*
[c] *North East Structural Genomics Consortium (NESG), Columbia University, 1130 St. Nicholas Ave., New York, NY 10032, USA*

Accepted 5 July 2006

Abstract

We survey computational approaches that tackle membrane protein structure and function prediction. While describing the main ideas that have led to the development of the most relevant and novel methods, we also discuss pitfalls, provide practical hints and highlight the challenges that remain. The methods covered include: sequence alignment, motif search, functional residue identification, transmembrane segment and protein topology predictions, homology and *ab initio* modeling. In general, predictions of functional and structural features of membrane proteins are improving, although progress is hampered by the limited amount of high-resolution experimental information available. While predictions of transmembrane segments and protein topology rank among the most accurate methods in computational biology, more attention and effort will be required in the future to ameliorate database search, homology and *ab initio* modeling.

Keywords: Membrane proteins; Protein structure prediction; Protein function prediction; Alignments; Transmembrane segment prediction; Homology modeling; *ab initio* modeling

1. Introduction

1.1. The fold space of membrane proteins is relatively small…

Integral membrane proteins (IMPs) are polypeptide chains embedded into biological membranes. In this review, we focus exclusively on polytopic membrane proteins, i.e. proteins that span the membrane at least once. We will not consider monotopic membrane proteins, which despite being anchored to the membrane do not cross it (anchors include amphipathic helices and intra-membranous molecules covalently attached to the protein).

Biological membranes are phospholipid bilayers. The polar heads of the phospholipids face the aqueous solution on both sides of the membrane, while the lipid tails form a thick (~30 Å) hydrophobic core. The sequence and structure of IMPs reflects an extreme effort of adaptation to this environment. Since the interactions of polar groups with the hydrophobic core of the membrane are energetically unfavorable, IMPs have evolved so as to minimize them. Inside the membrane, IMP chains form long regular secondary structure elements that cross the entire bilayer, i.e. are transmembrane (TM) segments. In this way, they satisfy the hydrogen-bond potential of the backbone amide and carbonyl groups. These secondary structure elements can either be TM helices (TMH) or TM β-strands (TMB) arranged into β-sheets; helices and strands are connected by regions that generally extend beyond the membrane core, into the aqueous solution. Also, the assembly of TM elements is strongly influenced by the presence of the bilayer. α-Helices form bundles, with each helix strongly oriented towards the normal to the membrane plane [1]. β-strands are arranged in barrel-like structures, with the barrel axis normal to the membrane plane. While helical

* Corresponding author. Fax: +1 212 305 7932.
E-mail address: rost@columbia.edu (B. Rost).
URL: http://www.rostlab.org/ (B. Rost).

1046-2023/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.ymeth.2006.07.026

IMPs are practically ubiquitous (only being absent in the outer membrane of Gram-negative bacteria), β-barrel IMPs are only found in the outer membrane of Gram-negative bacteria, mitochondria and chloroplasts. Proteins with both integral TMHs and TMBs have not been observed. Due to the strong compositional biases imposed by the bilayer onto the amino-acid sequence, predicting the location of TM segments in the protein sequence turns out to be a relatively easy task. In fact, in order to cross the membrane, TMHs need to be at least 15 residues long and composed predominantly of hydrophobic amino acids. TMBs are generally longer than 10 residues and are comprised of alternating hydrophobic and polar amino acids (hydrophobic side chains face the lipids while polar residues face the water- or protein-filled interior of the barrel). Also, regions connecting TM segments that are not translocated across the bilayer ("inside" or cytoplasmic regions) are enriched in positively charged amino acids, following the so-called 'positive-inside rule' [2,3]. This simplifies the prediction of the way in which TM segments cross the membrane (inside–out or outside–in).

In conclusion, in comparison to water-soluble proteins, IMP chains are able to sample only a limited number of folds [4]. This reduced conformational space may be easier to search through traditional sampling algorithms (e.g. molecular dynamics and Monte Carlo). As we said, the number, location and cross-membrane direction of TM segments can be predicted rather accurately (see Section 2). It is, therefore, reasonable to believe that 3D[1]-structure prediction for IMPs is within the reach of computational methods.

1.2. … but membrane proteins still exhibit remarkable structural variability

Notwithstanding the previously described constraints and biases, presently no method can accurately predict the 3D structure of any IMP from sequence alone. On one hand, this is due to the very limited number of high-resolution structures available, giving us a myopic view of the IMP structural space; on the other, IMP architectures may not be as simple as initially thought. Indeed, as more experimental structures have become available, IMPs have revealed an unexpected level of structural diversity. Constraints on TMH length and tilt angles are not as strict as previously hypothesized [5]. Also, analysis of known structures in the Protein Data Bank [6] (PDB) indicated that about 50% of TMHs contain non-canonical elements (i.e. kinks, 3_{10}-helix and π-helix turns) [7] and 5% of all TMHs cross the membrane only partially, i.e. form half TMHs [8]. Stable loop regions have been found inside the hydrophobic core of the membrane [9,10]. In some cases, membrane domains contain crevices and cavities that can accommodate numerous water molecules and other ligands, thus forming a more diverse environment with which the protein can interact. There is also evidence that TMH topology is not solely determined by the protein sequence, but depends on complex interactions with the translocon, the machinery responsible for inserting TMHs into the membrane [11]. Consequently, topology may depend on the organism in which the protein is expressed. Finally, functional IMPs are very often the result of the assembly of several chains, posing the problem of finding the correct protein–protein docking solution, a notoriously difficult task, despite recent progress [12]. For example, receptor binding and activity in G-protein coupled receptors have been reported to be linked to oligomerization [13]. Thus, the IMP structural space, although smaller than that sampled by water-soluble proteins, is still intricate enough to make computational structure prediction a very challenging task.

1.3. Two main prediction tasks: identify and characterize

In this paper, we review computational methods that can be applied to study IMPs. Some of these approaches have been developed specifically for IMPs, others have been adapted or simply transferred from the techniques originally devised for water-soluble proteins. Computational methods have been developed to address two different goals: to identify novel IMPs or to characterize the functional and structural features of a protein experimentally known to be inside the membrane.

The first task (identification of IMPs) can be solved by sequence or profile alignment techniques that detect relatives of known IMPs. Alternatively, we can search motif databases (e.g. PROSITE [14], PRINTS [15]) to identify functional motifs that are characteristic of IMPs and use these motifs to extend the database searches, or we can use prediction methods that locate TM segments. Note that the latter two approaches do not require homology to a known IMP and that the described methods produce different and often complementary information. For example, Ruta et al. [16] used database searches and sequence analysis to detect a previously uncharacterized voltage-gated potassium channel in the archae bacterium *Aeropyrum pernix* and later confirmed the prediction experimentally. On the other hand, prediction of TM segments has been widely used for estimating the fraction of membrane proteins in various genomes and kingdoms of life.

Also the second task (functional and structural characterization) can be achieved by different means, depending on the type of information that is available (Fig. 1). If the high-resolution structure of a protein (template) homologous to the target is known, we can use homology modeling to obtain a prediction for the target structure. The quality of the resulting model typically depends on the sequence similarity between template and target [17]. Once built, the model can prove instrumental for studying the function of the protein. If low-resolution structural information (e.g.

[1] *Abbreviations used:* 3D, three-dimensional; GPCR, G-protein coupled receptor; HMM, Hidden Markov Model; IMP, integral membrane protein; NN, neural network; PDB, Protein Data Bank; SVM, Support Vector Machine; TM, transmembrane; TMH, transmembrane helix; TMB, transmembrane β-strand.

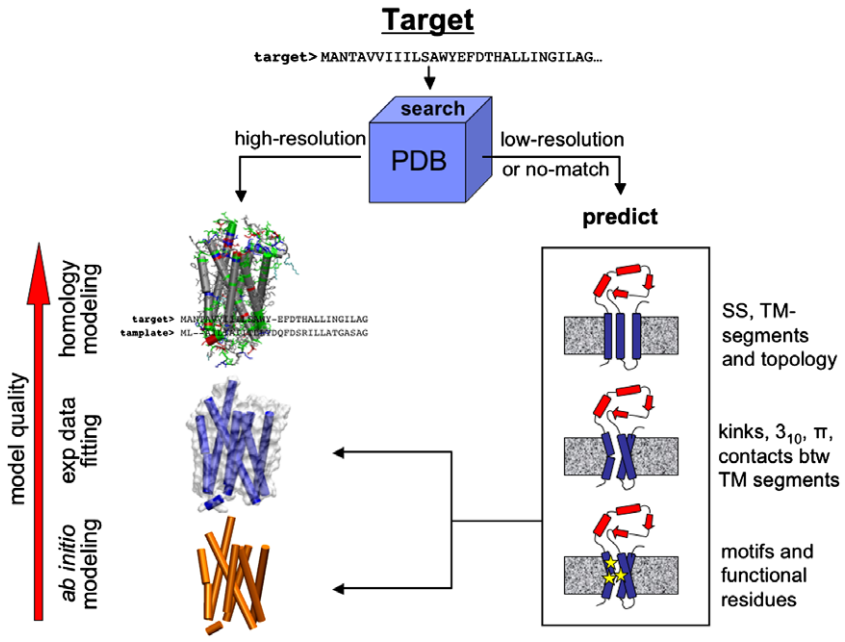

Fig. 1. Predicting structure and function for a protein experimentally known to be an IMP. First, the target sequence will be searched against the PDB [6], looking for homologs of known 3D structure. If this search returns at least one good match (template) with a high-resolution structure, it will be possible to apply homology modeling techniques to obtain a model for the target protein whose resolution will in general depend on the similarity with the template. If the search is either unsuccessful or returns a low-resolution structure, further analysis will instead be needed. Prediction of TM segments, kinks, functional residues and motifs can all help in elucidating the target structural and functional features, either in combination with low-resolution structural information (e.g. from cryo-electron microscopy) or with *ab initio* modeling techniques. When predicting function it will be useful to search not only the PDB but also other databases, such as SWISS-PROT [146], Interpro [147] and Pfam [50]. Note that database searches are performed through alignment methods using either substitution matrices or HMMs.

from cryo-electron microscopy or mutation analysis) is available either for the target or for a homolog, we can analyze sequence conservation and co-variation to identify functional residues, or, in combination with molecular modeling, to provide a better guess of the target structure. When no experimental structural information is available, we can apply *de novo* and *ab initio* methods. *De novo* methods predict TM segments and topology through knowledge-based approaches, while *ab initio* methods attempt to model structural features from first principles. Both *de novo* and *ab initio* methods may help in finding optimal packing through the application of energy-based scoring functions. In general, the less is known experimentally, the lower is the expected accuracy of the models, with homology modeling being, when applicable, by far the most reliable modeling technique.

1.4. Goals and standards of this review

While we present empirical observations and theoretical background when needed to understand computational approaches, our main aim is to highlight the state-of-the-art methods that readers can rely upon for the study of IMPs. As far as the quality of the predictions is concerned, it is not always possible to report independent estimates of performance. In fact, despite the important progress witnessed in the last few years in membrane protein structure determination [18], the small number of high-resolution structures available has so far prevented experiments of the type devised for the assessment of prediction methods for water-soluble proteins, such as CASP [19], CAFASP [20], EVA [21] and LIVEBENCH [22]. Thus, when no independent assessment data are available, we refrain from reporting self-assessed performance measures, while attempting to point out the strengths and weaknesses of the methods. Finally, since in many respects computational analysis of IMPs may still be considered in its infancy, new and more effective methods are likely to emerge in the very near future. Throughout the paper, we will try to point to those techniques that are likely to break new ground and lead to novel tools or increased performance.

2. Methods

2.1. Databases

Several databases have been built as comprehensive repositories of IMP sequences or structures and for the purpose of helping computational biologists to develop and test their prediction methods (Table 1). PDB_TM [23] and Stephen White's database (http://blanco.biomol.uci.edu/Membrane_Proteins_xtal.html) contain all IMPs of known structure, with links to the PDB [6] and PubMed (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db = PubMed) entries. MPtopo [24] additionally includes a list of proteins

Table 1
Membrane proteins databases

Database	Description/URL
GPCRDB [29], *KchannelDB and others*	Several receptor databases http://www.receptors.org
OPM [26]	Database reporting predictions for the orientation of IMPs within the membrane http://opm.phar.umich.edu/
PDB_TM [94]	Database of known membrane protein structures http://pdbtm.enzim.hu/
MPtopo [24]	Database of experimentally determined protein topologies http://blanco.biomol.uci.edu/mptopo/
Stephen White's database	Database of known membrane protein structures http://blanco.biomol.uci.edu/Membrane_Proteins_xtal.html
PRNDS [30]	Database of porins http://gene.tn.nic.in/PRNDS
TCDB [28]	Transport classification database http://www.tcdb.org/
TMDET [27]	Web server for predicting the orientation of a query membrane protein structure http://www.enzim.hu/TMDET

of unknown 3D structure, the topology of which has been experimentally annotated through low-resolution techniques such as gene fusion or proteolytic degradation. The reliability of such annotations is typically much lower than that of annotations from 3D structures [25]. The PDB_TM and OPM databases [26] contain predictions for the orientation of IMPs of known structure relative to the hydrophobic core of the membrane. The web server TMDET (a companion to PDB_TM) [27] can also be used to predict the membrane orientation of model structures, e.g. models obtained by homology. The transport classification database (TCDB) [28] is a comprehensive classification of membrane transporters (analogous to the Enzyme Classification system for enzymes) that incorporates both functional and phylogenetic information. It classifies about 3000 transporter sequences into more than 550 families. Interestingly, the TCDB website also reports a list of transporters that have been associated with disease. Finally, there are several databases that collect information on specific IMP families, such as G-protein coupled receptors (GPCR) (the database is GPCRDB [29]) and potassium channels (KchannelDB, see 'usage notice' at http://www.receptors.org/KCN/htmls/consortium.html), or structural classes, such as porins (PRNDS [30]).

2.2. Sequence alignments, motifs and functional residues

Sequence alignment methods are ubiquitous tools for the prediction of structure and function; they are primarily used to identify related sequences via database searches and to detect template structures needed for the construction of homology models [17]. Different applications may require different alignment approaches since so far no single method combines the sensitivity and efficiency necessary for a database search with the accuracy required for the construction of a homology model. Thus, we first consider developments that relate to all types of alignment approaches, and then address separately methods for database searching and for the purpose of building homology models.

2.2.1. A general consideration: substitution matrices

Pair-wise sequence alignment methods generate many possible matches between two protein sequences. Matches are typically scored as a sum over the probabilities for each observed amino-acid substitution (match), as defined in the so-called substitution matrix, and the highest-scoring alignment is reported. Thus, when selecting a sequence alignment method there are two major considerations: the substitution matrix and the algorithm for optimizing the scores. For water-soluble proteins, the two most widely used substitution matrices are PAM [31] and BLOSUM [32]. However, since the amino-acid composition of TM domains differs from that of water-soluble proteins [33], several groups have introduced amino-acid substitution matrices specific to helical IMPs, including JTT [34], PHAT [35] and SLIM [36]. Of these, SLIM was reported to have the highest accuracy for detecting remote relationships between sequences in the manually curated GPCR database [36]. To date, however, the membrane protein-specific matrices have not been independently assessed on large-scale IMP data sets, and as such are not recommended for general use.

2.2.2. Database searching for detection of sequence relations

For the detection of related sequences in databases, novel methods have been developed that use the signal from the long hydrophobic stretches of TMHs, as well as more complex approaches (Table 2). These methods involve: matching sequences with an equivalent number of predicted TMHs [37–39]; per-residue alignment of hydropathy profiles [40]; matching patterns of peaks and troughs extracted from hydropathy profiles [41] or of TM topology (GPCRHMM) [42]; matching of functional motifs (PRED-GPCR) [43] or of amino-acid composition [42]; and favoring matches of predicted TM residues within a standard PSI-BLAST search (TM-PSI) [44]. For these approaches, success was often measured by the identification of novel family members not found by sequence alone, although TM-PSI, PRED-GPCR and GPCRHMM were assessed

Table 2
Sequence alignment programs, family HMM profiles and motifs databases[a]

Method	Description/URL
BLAST or PSI-BLAST [148]	Database searching http://www.ncbi.nlm.nih.gov/BLAST/
ClustalW [149]	Multiple-sequence alignment http://www.ebi.ac.uk/clustalw/
GPCRHMM [42]	Identification and classification of GPCRs http://gpcrhmm.cgb.ki.se
HMAP [57]	Profile-to-profile alignment including secondary and tertiary structure information http://trantor.bioc.columbia.edu/hmap/
Hydropathy-profile alignment	Direct alignment of hydropathy profiles for database search http://bioinformatics.weizmann.ac.il/hydroph/hydroph.html
Hydropathy pattern matching PHAT search [35]	Conversion of representative profile into pattern of peaks/troughs for database search http://blocks.fhcrc.org/sift/PHAT_submission.html
PSI-BLAST with compositional bias [148]	Database searching allowing for differences in composition in membrane proteins http://www.ncbi.nlm.nih.gov/blast/
PFAM [50]	Database of protein families including HMMs and sequence alignments http://www.sanger.ac.uk/Software/Pfam/
PRED-GPCR [43]	Identification and classification of GPCRs http://bioinformatics.biol.uoa.gr/PRED-GPCR
PRINTS [15]	Database of protein family fingerprints (motifs) http://umber.sbs.man.ac.uk/dbbrowser/PRINTS/
PROSITE [14]	Database of protein families including patterns and profiles http://www.expasy.ch/prosite/
T-Coffee [56]	Advanced multiple-sequence alignment http://www.igs.cnrs-mrs.fr/Tcoffee/tcoffee_cgi/index.cgi
TM-PSI [44]	Matching of predicted TM regions from TMHMM within PSI-BLAST profiles for database search

[a] Code for other methods mentioned in the text is usually available from the authors.

for their ability to detect GPCR sequence relationships [42–44]. No systematic comparisons of all these methods exist making it impossible to name the most effective approach. However, matching numbers of TMHs are likely to be the least sensitive method, whereas the family specific methods may prove to be the most sensitive. Hydropathy-profile alignments may be useful for classification of novel protein families [45,46], and are reported to be appropriate for detecting close homologies [41]. Hydropathy-profile pattern matching, on the other hand, which may be more sensitive for remote-homology detection, has the disadvantage that it requires manual selection of representative patterns [41]. TM-PSI is probably the most general and easily automated method, but it requires that a TMH prediction is made for all sequences in the database before a search can proceed, and is not currently available through a web server. It will be interesting to compare these methods with general sequence profile-based or Hidden Markov Models (HMM) methods used for water-soluble proteins [47], which are not limited to one fold type or family.

In contrast to the significant number of novel methods developed for database searching of helical IMPs, the analysis of β-barrel IMPs has generally focused on assigning function through the prediction of the number of TMBs (see TM segment prediction section, below), or has relied on standard BLAST searches for the detection of related sequences.

A recent development in database searching is the adjustment of the scoring in BLAST to reflect the composition of the sequences [48,49], which theoretically should improve results for membrane protein searches. However, this method has not yet been assessed specifically on membrane proteins, and thus should be used with caution. Note also that during BLAST or PSI-BLAST searches, it is recommended not to apply the "low-complexity filter", since this is likely to remove hydrophobic regions from the target sequence [44] (see also http://ca.expasy.org/tools/blast/blast_help.html).

2.2.3. Detecting relations through motifs and patterns

When database searching fails to find homologs, it is still possible to retrieve useful information about the target by searching databases that contain motifs or patterns relating to specific protein families. Such databases, for example PFAM [50], PROSITE [14] and PRINTS [15] (Table 2), usually cover water-soluble proteins as well as IMPs. Although two proteins that share a common motif may not have common ancestry, a match can often provide very useful information about the function of the target.

2.2.4. Sequence alignments

The accuracy of the alignment between two related membrane protein sequences is another important consideration (Table 2). The bipartite alignment method for helical IMPs uses a TM substitution matrix in the (predicted) membrane regions of a sequence, and a standard matrix in the remaining regions (e.g. STAM) [35,51,52]. This can be achieved by separating out the predicted TMHs from other residues, aligning them independently and finally reassembling the full sequence (e.g. [52,53]). Often, known functionally

important motifs are matched in order to guide the alignment [54,55]. Alternatively, one can simply adopt a standard alignment algorithm but apply substitution probabilities from different matrices according to whether the position is predicted to be in the membrane or not. Using this approach with the PHAT matrix however does not provide an improvement over the BLOSUM matrix [157]. Of all these methods, STAM [52] is possibly the most user-friendly, since it incorporates a graphical user interface, and allows manual adjustment of the definitions of the ends of the TM segments. Most of these methods are not currently widely available, and, in the absence of more thorough benchmarks, are not recommended for the casual user. Advanced multiple-sequence alignment and profile-to-profile alignment methods, however, have been shown to work well for water-soluble proteins, and can in fact be used for membrane proteins [157]. Multiple-sequence alignment methods such as T-Coffee [56] are also very accurate at high sequence identities (>40%), despite using a single generic substitution matrix [157]. At lower sequence identities, the use of profile-based methods, such as HMAP [57], which incorporate structural information, can improve the alignment accuracy significantly [157]. For a review of the best methods for water-soluble proteins, and their availabilities, see [58].

2.2.5. Marking functional residues

The residues that are most relevant for the structure and function of a protein are not always recognizable as part of a global sequence similarity pattern or as motifs of neighboring conserved positions on the sequence. Sometimes conservation patterns are irregular and sparse; in other cases, the relevant signal is not the conservation but rather the co-variation of two or more sequence positions. Specific methods have been developed to deal with these situations and they can be very useful for predicting protein structural and functional features, either in combination with the previously described approaches, or in cases where these approaches are not applicable. Unfortunately, most of these methods have not yet been implemented as web servers.

Amino-acid conservation at a specific position within a family (i.e. an ensemble of proteins sharing a common ancestor) is considered to be an indication of strong evolutionary pressure at that site. Hence, conserved residues, easily identified through multiple-sequence alignments (Table 2), represent potential targets for structural and functional studies. Also, within a family it is often possible to identify several subfamilies that exhibit different functional characteristics (for example, binding different ligands or substrates). When this is the case, functional variability is usually encoded in subfamily specific conservation at particular sites (dubbed tree determinants). In two recent papers, the Valencia and Lichtarge groups reviewed these methods and attempted an evaluation of their performance in predicting functionally and structurally important residues [59,60]. Another approach that aims at the identification of functionally relevant residues is co-variation analysis. In this case, it is assumed that proteins can accommodate a potentially deleterious mutation at a given position if this is compensated by concerted mutations at other sites. Co-variation is usually considered to be an indication of spatial proximity [61] and can hence be used for predicting intra-protein contacts [61,62] or protein–protein interactions [63].

In general, all methods based on sequence conservation within families and subfamilies or co-variation can hardly distinguish between different evolutionary constraints. For this reason, interpretation of a specific signal often relies on some additional experimental information. In the realm of membrane proteins, conservation and co-variation studies performed on GPCRs have taken advantage of the high-resolution structure of rhodopsin [64] to identify putative oligomerization interfaces and G-protein interaction sites [65]; knowledge of the crystal structure of a voltage-gated potassium channel from *Aeropyrum pernix* [66] prompted the use of correlated mutation analysis to identify gating-related residues [67]; recently, conservation and co-variation analysis has been applied in combination with molecular modeling to the gap junction channel, starting from a low-resolution cryo-electron microscopy structure [68].

2.3. Prediction of TM segments

There are two goals of methods that predict TM segments. The first, whole-protein prediction, identifies IMPs among a set of protein sequences for which structures and membrane/non-membrane localization are unknown. The second is per-residue prediction, which labels the residues in a protein according to whether they span the membrane or not. Whole-protein predictions are frequently used in structural genomics projects when defining the target lists, with the objective to either exclude or to specifically select IMPs. In contrast, per-residue predictions are routinely used by experimentalists, e.g. as a guide for engineering mutants.

2.3.1. Early approaches

Early predictions of TM segments for helical IMPs (Fig. 2) generally used the following four-step procedure: (1) derive a 'propensity scale', a set of 20 numbers corresponding to properties or statistics of the 20 amino acids when found in TM regions. (2) Generate a plot of propensity values along the query sequence. (3) Smooth the plot by taking the average propensity value in a window of N residues and plot the average at the center of the window (i.e. a sliding-window average). (4) Identify TM stretches on the smoothed plot using some propensity threshold.

Early studies derived propensity scales from biophysical or chemical measurements such as water/vapor transfer free energy. A well-known early example was the 'hydropathy index' of Kyte and Doolittle [69]. As more experimental structures have become available, propensity scales have been expanded to reflect more than the original two

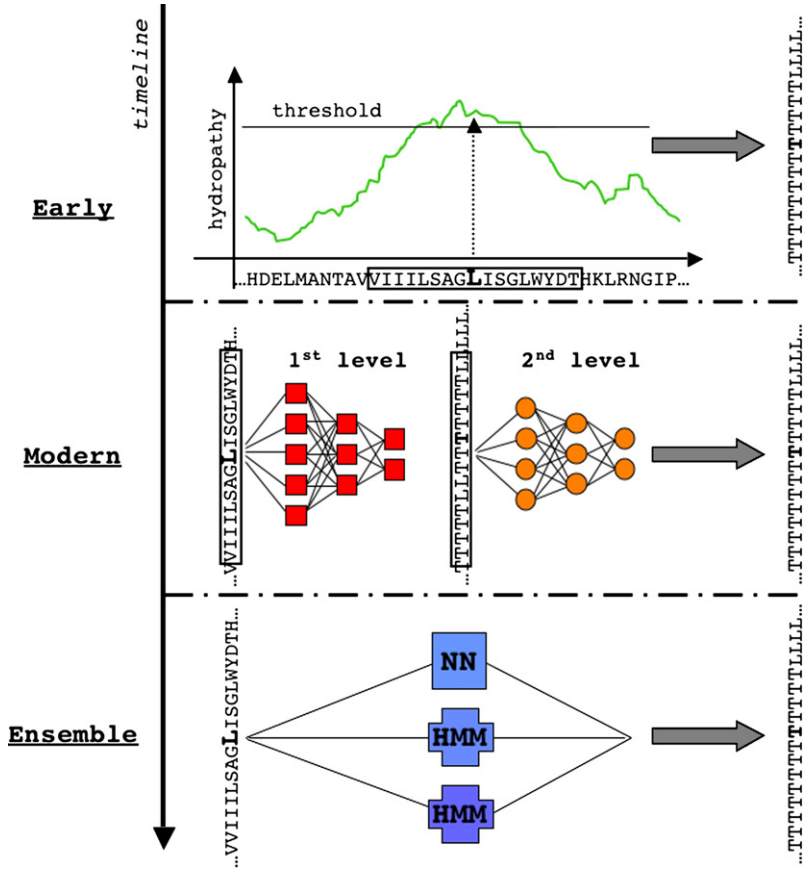

Fig. 2. Timeline of TMH prediction methods. Early methods used per-residue hydropathy scales and a window around each residue to produce a smoothed sequence profile. A threshold, usually manually adjusted, was then introduced to predict TMHs. Modern approaches use machine-learning algorithms such as NNs (illustrated), HMMs or SVMs to predict each residue in e.g. two states, TM (T) or non-TM (L). In this figure, we schematically show a common NN architecture used for predicting TMHs. As discussed in more detail in the text, there are two NN levels: the first (also called sequence-to-structure NN) produces a per-residue score (analogous to the per-residue value from hydropathy scales); the second (structure-to-structure) takes the output of the first NN and smoothes its values by taking into account the first level predictions for the neighboring residues (similar to the window used in early approaches). Finally, ensemble approaches combine several different methods (in this example, one NN and two HMMs) to produce a consensus prediction.

(TM/non-TM) structural states. Careful inspection of 3D structures has led to scales derived from 'snorkeling preferences' (polar amino acids enriched at TMH N-terminal ends, due to chain-direction dependent rotamer preferences) [70], helical hairpins [71], or residue orientation in helix bundles [72].

2.3.2. Modern approaches

Many popular modern approaches (Fig. 2), based on HMMs or neural networks (NNs), are related to sliding-window hydropathy plot methods. The most successful NN approaches use the concept of combining two consecutive NNs, first introduced by PHDhtm [73,74], to predict TMHs (Fig. 2). The first (dubbed 'sequence-to-structure' network) receives a window of the amino-acid sequence as input and outputs a score representing the propensity of the central residue in that window to be within the membrane. When sliding the window over the entire sequence, the NN produces a succession of propensities (analogous to the propensity plot mentioned above). The second NN (dubbed 'structure-to-structure' network) smoothes the propensities generated from the first one. These NNs are capable of capturing subtle patterns, such as amphipathicity in helices, or TMH length preferences, which may evade simple hydropathy plot methods.

HMMs are more specialized models. Like hydropathy plots and NN-based methods, they capture both 'sequence-to-structure' and 'structure-to-structure' relationships. The difference is that in HMMs the two steps are merged into one integrated model. This has advantages and disadvantages, and HMMs are often complementary to NNs. For example, HMMs can take advantage of global patterns in the structure, such as the repeated strand-periplasmic loop-strand-extracellular loop pattern in β-barrel IMPs. NNs have no such "global view" of proteins. However, NNs can recognize local sequence patterns involving several residues, which HMMs cannot. The best performing HMM- and NN-based methods will be discussed in a later section (see also Tables 3 and 4).

More recently, Support Vector Machines (SVM) have also been applied to this problem. An SVM is a general algorithm used to classify patterns into two groups (known

as the 'binary classification problem'). Yuan et al. [75] use an SVM for per-residue prediction of helical IMPs with a sliding window, similar to 'sequence-to-structure' NN or early propensity value methods. SVMs (like NNs) are capable of learning complex relationships among the amino acids comprised in the local windows with which they are trained.

Ultimately, selection and preparation of training data affects performance of each model, and it is unclear whether one kind of model is clearly better than others.

2.3.3. Ensemble methods

An ensemble method consists of taking the output of individual predictors and combining them by a weighted vote (Fig. 2). This idea has been extensively tested in statistics, and it is known that, if the predictors have roughly equal accuracy and loosely correlated errors, the ensemble will yield a better prediction than each individual method [76]. The majority vote tends to cancel the errors, choosing the correct prediction among the individuals [77] and thus combining the advantages of the different methods. Martelli and Casadio combined a NN and two different HMM predictors for predicting helical IMPs [78]. Taylor and coworkers [79] combined five methods for predicting helical IMP topology. Nilsson and coworkers also used five topology prediction methods to predict partial membrane topologies [80]. Bagos et al. [81] created an ensemble predictor for β-barrel IMPs from several available methods.

2.3.4. Detecting membrane proteins in genomes (whole-protein predictions)

While some methods provide whole-protein predictions, fewer are suitable for whole-proteome screening, either because they have not been specifically evaluated for this task, or because the websites do not accept multiple-sequence submissions, making screening impractical. The whole-protein prediction methods are invariably applicable to only one of the two membrane protein fold types (helical or β-barrel IMPs, Tables 3 and 4, respectively). One important observation is that most methods using only hydrophobicity indices incorrectly identify TMHs in over 50% of all proteins [25]. In contrast, the very best methods confuse fewer than 2–4% of the globular proteins for IMPs [25]. Particularly successful are: DAS-TMFilter [82], SOSUI [83], TMHMM [84] and PHDhtm [25,74]. Note that these error rates assume predictions based on the mature protein sequence, i.e. without signal and transit peptides. Error rates are significantly higher than the values quoted if we include the confusion between signal peptides and TMHs [25,85].

2.3.5. Assessing TMH predictions

Most TMH prediction methods use the evolutionary information from multiple alignments either directly or indirectly. For these methods, poor predictions often result from incorrect or uninformative (no homologs found) alignments. Some servers flag such predictions as less reliable; others do not. Another common error is to predict

Table 3
Recommended TMH prediction methods[a]

Method	URL	Service
DAS-TMfilter [82]	http://mendel.imp.ac.at/sat/DAS/DAS.html	PR2, WP
HMMTOP 2.0 [86]	http://www.enzim.hu/hmmtop/	PR3
MEMSAT3 [150]	http://bioinf.cs.ucl.ac.uk/psipred/	PR3
MINNOU [151]	http://minnou.cchmc.org/	PR2
OrienTM [152]	http://o2.biol.uoa.gr/orienTM/	PR3
PHDhtm [87]	http://rostlab.org/predictprotein/submit_adv.html	PR3, WP
Phobius [100]	http://phobius.cgb.ki.se/	PR3 + SP
SOSUI [83]	http://sosui.proteome.bio.tuat.ac.jp/sosuiframe0.html	PR2
Split4 [88]	http://split.pmfst.hr/split/4/	PR2
THUMBU	http://phyyz4.med.buffalo.edu/service.htm	PR2, WP
TMAP [90]	http://bioinfo.limbo.ifm.liu.se/tmap/index.html	PR3
TMHMM [84]	http://www.cbs.dtu.dk/services/TMHMM/, http://www.sbc.su.se/~erikgr/tmhmm/index.html	PR3, WP, database
TOP-PRED [153]	http://bioweb.pasteur.fr/seqanal/interfaces/toppred.html	PR3

[a] The methods in this Table are recommended either because they stood out in independent assessments or because they are otherwise interesting. PR2, two-state TM/non-TM per-residue prediction; PR3, three-state TM/inside/outside per-residue prediction, i.e. implicitly a topology prediction; SP, N-terminal signal-peptide prediction; WP, suitable for whole-proteome screening. Results are returned in seconds or minutes by most servers.

Table 4
Recommended TMB prediction methods[a]

Method	URL	Service
B2TMR, HMM-B2TMR [95]	http://gpcr.biocomp.unibo.it	PR3
BOMP [154]	http://www.bioinfo.no/tools/bomp	WP
PRED-TMBB [91]	http://bioinformatics2.biol.uoa.gr/PRED-TMBB	PR3
PROFtmb [93]	http://rostlab.org/services/proftmb	PR3, WP
TMB-HUNT [155,156]	http://www.bioinformatics.leeds.ac.uk/	WP

[a] The methods in this Table are recommended either because they stood out in independent assessments or because they are otherwise interesting.

TMHs in globular proteins or in place of signal peptides. Both signal peptides, which comprise a hydrophobic region, and long hydrophobic stretches in globular proteins, may be easily mistaken for TMHs. The simplest defense against the mis-prediction of water-soluble proteins is only consider predictions with multiple TMHs. N-terminal signal peptides can also be excluded from predictions by a similar rule, i.e. by excluding cases where a single TMH is predicted to be within a given number of residues from the N-terminus. Because of a variety of reasons, few servers implement such rules. Another pitfall that many prediction methods have succumbed to involves setting hard limits for the minimum and maximum lengths of TMHs. When few high-resolution structures were known, low-resolution experiments found TMHs to range between 17 and 25 residues in length [25]. Many methods used this information to divide what would otherwise have been predicted to be, for example, a 36-residue helix into two 18-residue helices, since it was clear that a 36-residue helix fell outside of the experimentally observed range. However, as high-resolution structures became more numerous and diverse, it became clear that membrane helices are in fact extremely diverse, and range from 10 (half-TMHs) [8] to 40 residues in length [25]. As a result, the prediction of long TMHs as well as of half helices is generally inaccurate. In addition to servers enforcing strict limits on the lengths of TMHs, some servers predict all TMHs to be exactly the same length (e.g. 20 residues). This may be a way of effectively "hedging ones bets" (you do not terribly overshoot or undershoot the boundary) at the TM region boundaries and scoring better on a per-residue basis (see Fig. 4 in [8]).

Methods that predict TM segments have been subjected to several assessment studies and they have been evaluated on both high-resolution and low-resolution datasets. While high-resolution data, including crystal and NMR structures, provide structural details at an atomic level, low-resolution experiments, mainly of C-terminal fusions with indicator proteins and antibody-binding studies, only identify which portions of an IMP are inside and outside of the membrane [85]. Unfortunately, low-resolution experiments have been suggested to be no more accurate than the best prediction methods [25]. Two recent studies have examined the accuracy of TMH prediction [8,25]. These studies used a wide variety of measures to gauge performance, including prediction accuracy measured on a per-residue, per-segment (a segment is correctly predicted if most of its residues are correctly predicted), and per-protein basis (a protein is correctly predicted if all of its TM segments are identified, see previous paragraph), as well as the accuracy of segment boundary and topology prediction. Methods were further evaluated on their ability to avoid false positive predictions of TMHs in signal-peptide regions and in globular proteins. No single method was able to perform best according to all of the various measures. As a result, these studies did not designate a best method, but rather reported which methods consistently ranked very high across all of the measures. The top performing programs reported in the earlier study [25] were HMMTOP2 [86] and PHDhtm [87], while SPLIT4 [88], TMHMM2 [89], HMMTOP2 and TMAP [90] topped the rankings in the more recent study [8]. The difference between the two lists can largely be attributed to the inclusion of more recent methods in the later study. When looking at these results, it has to be considered that although the assessments were performed under controlled conditions, the paucity of the available data forced the assessors to use the very same structures on which the methods had been trained. In other words, it is often difficult to guess from these data how the methods would perform on previously unseen examples. In this view, the fact that HMMTOP2 appears in both lists of top predictors gives some indication that it may perform well on proteins it has not seen in training. Note that the ensemble methods, most of them reporting higher accuracies than individual methods, have not yet been independently assessed.

2.3.6. Assessing TMB predictions

To our knowledge, Bagos et al. [81] have provided the only independent evaluation of TMB predictors. They evaluated four HMM-based, four NN-based and two SVM-based methods in terms of several statistics focused on per-residue, per-segment and topology prediction accuracy. Two issues affect accuracy assessment for TMBs. First, many methods were originally evaluated on the sequences taken from PDB structures. β-Barrel IMPs, needing to be translocated to the outer membrane of Gram-negative bacteria and organelles, and in contrast with most helical IMPs, invariably carry a cleavable signal peptide. However, sequences extracted from the PDB are almost always mature, signal-peptide cleaved forms of the original sequence, and in some cases, are missing N- or C-terminal domains. Bagos et al. found that when signal peptides were present in the query protein sequence, NN- and SVM-based methods in particular tended to mistake it for a TMB, while HMMs did not. Thus, practically speaking, removing known signal peptides from query sequences before submitting them to a TMB predictor will yield slightly improved predictions. The second important issue is that β-strands in β-barrel IMPs often extend well past the membrane, mostly on the extracellular side. Some methods are designed to predict the TM regions of the strands, such as PRED-TMBB [91,92], while others predict the full extent of the β-strands, such as PROFtmb [93]. Bagos et al. assessed the ability of the methods to predict the location of the TM regions of β-strands, as determined by PDB_TM [94]. With the above factors as *caveats*, the study made one major conclusion: the top three methods namely, *HMM-B2TMR* [95], *PROFtmb* [93] *and PRED-TMBB* [91], *are* HMM-based and far outperform NN- and SVM-based methods in predicting topology and number of strands correctly.

2.3.7. Topology prediction for helical IMPs

Topology prediction identifies non-TM regions as either non-translocated (inside) or translocated (outside). Membrane

insertion and topogenesis of helical IMPs is conserved between bacteria and eukaryotes, and the mechanism is complex and actively researched (reviewed in [96–98]). Future insights into these mechanisms will be invaluable for improving topology prediction. Helical IMPs generally traverse the membrane in unbroken helices, creating an alternating translocated/non-translocated pattern of extra-membrane loops. Thus, when the exact number of TMHs is known and assuming that the protein contains only integral TMHs (i.e. no half-TMHs), the topology is uniquely determined by the location of the TM segments and of the N-terminus (that is, either inside or outside).

Current predictors rely on two topogenic signals in the protein sequence: the distribution of positively charged residues in extra-membrane loops, known as the "positive-inside rule" [2] and the existence of N-terminal signals, cleaved and uncleaved. The "positive-inside rule", introduced by von Heijne and Gavel [2], reflects the observation that Lys and Arg are enriched in non-translocated loops and depleted in translocated loops (although the bias is strong only for loops shorter than 60 residues). Later, Andersson and von Heijne demonstrated in the inner membrane of *Escherichia coli* that SecA ATPase was required for translocation of loops in a fashion linearly dependent on loop length and the number of positive charges, hypothesizing that the dependence was due to the requirement of several rounds of ATP hydrolysis to translocate positive charges against the membrane potential [99]. More recently, the basis of the "positive-inside rule" has been proposed to be the interaction with the translocon itself [97]. While many methods use a simplified version of the "positive-inside rule" that ignores the lengths of the loops, TMHMM [89] explicitly models short and long loops using separate HMM architecture modules, thus accounting for the length dependence. The other important topogenic signal is the N-terminal signal. Phobius [100] is the first topology model to combine signal-peptide cleavage site detection with TMH prediction in an integrated probabilistic model, which seeks to compensate for mis-predictions of cleavable signal peptides. It also makes optimal use of the fact that a cleaved signal corresponds to a topology with the N-terminus outside. Recently, Bernsel and von Heijne [101] improved topology predictions of helical IMPs using searches of SMART [102] database families to identify water-soluble proteins that have compartment specific localization and are homologous to the IMP non-TM domains.

2.3.8. Prediction of kinks and other non-helix conformations in TMHs

A significant fraction of TMHs exhibits kinks and other deviations from standard α-helical conformations. These features are likely to play an important role in determining the structural and functional diversity of proteins with the same topology [103]. Therefore, they are an important target for computational predictions, particularly as a complement to homology modeling (see below). Recently, Bowie and coworkers [103] proposed a simple algorithm to predict kinks in TMHs. The method consists of producing a multiple-sequence alignment of homologs of the target protein, and identifying the positions at which more than 10% of the residues are prolines. While self-assessed performances are high, the authors do not specify the number and evolutionary-distance distribution of homologs that are required in order to produce a robust prediction. Another method, based on sequence pattern descriptors, predicts kinks as well as other non-canonical helical conformations (3_{10} and π turns) [104]. Although the method is reported to perform well in identifying non-alpha conformation in TMHs of known structure, it is unclear how it would perform when evaluated on previously unseen proteins, since pattern recognition methods can be very accurate but are often unable to generalize well.

2.3.9. Genomics of helical IMPs

The explosion of genome sequencing projects since the last decade has enabled systematic computational study of IMPs within the genome. Studies using several different methods generally agreed that in most genomes, about 20–30% of proteins have at least one TMH [89,140–144]. An early study on 14 genomes found that larger genomes contained a higher fraction of membrane proteins than smaller genomes [143], which led to the hypothesis that more complex multi-cellular organisms may need more membrane proteins to carry out inter-cellular communications. Several more recent studies based on more genomes and more complete sequence data failed to reproduce the same results [89,142,144]. Detailed topology prediction, although not as accurate as the simple membrane vs. non-membrane classification, suggested that within an individual genome, proteins with fewer TMHs are more abundant than proteins with more helices. Proteins with seven TMHs seem to be much more abundant in higher eukaryotes, presumably due to the expansion of the GPCR family, and proteins with six or twelve TMHs are overrepresented in bacteria, reflecting the abundance of small molecule transporters [142]. It has also been observed that there is an interesting correlation between protein length and the number of TMHs. That is, most helical IMPs either have many TMHs with short connecting loops or have only 1–2 helices with large extra-membranous domains [142,143]. Recently, it was shown that topology prediction improves significantly when the periplasmic or cytoplasmic location of the C-terminus is known [145]. By constraining a topology prediction algorithm (TMHMM [84]) with large-scale experimental data on locations of the C-termini, high-quality topology models for more than 600 IMPs in *E. coli* were constructed [146], providing an elegant example of the power of combining both computational and experimental approaches.

2.4. 3D-structure prediction

2.4.1. Homology modeling and fold recognition for membrane proteins

Homology modeling (or comparative modeling) is the construction of a model for the structure of a target protein

using an experimentally known 3D structure of a related protein as a template. That is, once the template has been selected, and an alignment between template and target is generated (see Alignments), the non-conserved side chains are replaced, and insertions (regions with no template structure) are modeled as 'loop' regions by *de novo* or *ab initio* techniques. Of the many modeling programs available, those that produce the highest quality models (i.e. the best stereochemistry) are SegMod/ENCAD [105] and Modeller [106], whereas Nest [107] makes models that are the most similar to the native structure of the target [108] (Table 5). However, none of the commonly used homology modeling programs have been modified for membrane proteins (except for the SWISS-MODEL [109] 7TM interface, see *ab initio* methods): thus, care must be taken to ensure that the hydrophobic protein–lipid interface region of the protein does not incorrectly contain polar side chains, which may occur depending on the constraints imposed by the sequence alignment and the template structure. Side chain prediction algorithms such as SCWRL [110] or SCAP [111], which are certainly likely to improve the accuracy of the model within the core of the protein, may suffer from the same problem at the lipid-interacting surface.

Note that, as for globular proteins, the accuracy of a homology model is strongly dependent on the identity between the two sequences, even if the 'correct' alignment is obtained [157]. Specifically, below 30% sequence identity, the accuracy of the model decreases rapidly, since the difference in structure between the template and the true native structure is higher. Furthermore, there is significant variation between the extra-membranous loop regions of related IMPs, so that these regions of a homology model are likely to be particularly unreliable. On the other hand, *de novo*/*ab initio* predictions for loop regions shorter than 12 residues tend to be reliable [112,113], if constructed using reasonable physical–chemical force fields that incorporate descriptions of the aqueous solution rather than a vacuum environment [114,115]. Existing loop-prediction programs include Loopy [116], RAPPER [113] and PLOP [112] (Table 5); of these, PLOP is reportedly the most accurate, although it requires the most computational time, whereas Loopy is the fastest. Other strategies have been reported for loop modeling in GPCRs, including simulated annealing with Monte Carlo [115] and coarse-grained backbone dihedral sampling [117], although larger benchmarks are required to confirm their general applicability. To date, no methods have solved the significant problem of predicting multiple interacting loops simultaneously.

Table 5
Homology modeling, loop prediction and side-chain optimization methods[a]

Method	Description/URL
Modeller 7v7 [106]	Homology modeling, alignment, loop prediction etc http://salilab.org/modeller/
SegMod/ENCAD [105]	Homology modeling http://csb.stanford.edu/levitt/segmod/
Jackal	Homology modeling (Nest), side chain prediction (SCAP) and loop prediction (Loopy) http://trantor.bioc.columbia.edu/
PLOP [112]	Homology modeling, side chain and loop prediction http://francisco.compbio.ucsf.edu/~jacobson/
RAPPER [113]	Loop prediction http://raven.bioc.cam.ac.uk/
SCWRL 3.0 [110]	Side chain prediction http://dunbrack.fccc.edu/SCWRL3.php

[a] A selection of available tools for the construction and manipulation of homology models.

Apart from short loops, attempts to 'refine' an entire homology model to try to make it more native-like, and less like the template, are presently not recommended, neither for membrane proteins nor for water-soluble proteins, since no method has been shown to consistently improve the accuracy of homology models [159]. An exception may be when significant amounts of experimental information are available to provide constraints for refinement, such as for GPCRs [54].

2.4.2. Ab initio and de novo modeling of membrane proteins

When little or no experimental information about a protein or its homologs is available, it is necessary to resort to *ab initio* or *de novo* approaches. These approaches include prediction of interactions between pairs of TM segments as well as full 3D-structure modeling. While such predictions for water-soluble proteins can sometimes be used to obtain indications about 3D structure [118], methods developed for IMPs are still of limited practical relevance and, to our knowledge, none are presently available as web servers. Nonetheless, this is an area that is expected to grow considerably in the years ahead and we will therefore briefly review it in order to give the reader a glimpse of things to come.

Predicting TMH interaction is of crucial interest in helical IMP structure prediction for at least two reasons: firstly, the internal packing of helical IMPs is mostly determined by interactions between TMHs; secondly, oligomerization of helical IMPs appears to be mediated by inter-helical interactions between the different monomers (e.g. in GPCRs [119,120]). A handful of motifs have been identified that are often observed in TMH–TMH interfaces. In the GxxxG motif, two glycine residues are found on the same side of the α-helix. Due to their small size, they facilitate interaction with other helices, by maximizing van der Waals contacts between a larger number of residues and by establishing inter-helical hydrogen bonds [121]. The GxxxG motif was first identified as a principal effector of glycophorin A dimerization [122,123] only to be later discovered in internal interfaces and other oligomers [124,125]. Note, however, that recent studies dispute the relevance of this pattern for α-helix dimerization [126,127]. Analogous motifs involving small residues, such as alanine and serine, and longer motifs referred to as glycine zippers [128] are thought to play a similar stabilizing role. More

generally, helix–helix interfaces appear to be enriched in small amino acids, both in membrane proteins [129] and in globular proteins [5,130]. The simple "small residues go inside" rule was implemented with reasonable success in a TMH–TMH interaction prediction algorithm by Fleishman and Ben-Tal [132]. So far, the method has been tested on a carefully selected set of helix pairs, and it is not immediately applicable to multiple-TMH bundles. Polar residues are also often located at helix–helix interfaces. Although rare in TMHs, they seem to have a role in driving helix association [133].

Several methods have been developed for predicting global internal helix assembly or oligomerization of helical IMPs [134]. Most have been tested on very few structures, with the glycophorin A homodimer often serving as a model system. Ponder and coworkers [135] used a potential smoothing and search algorithm to predict the structure of the glycophorin A dimer. Jones and coworkers [136] implemented a knowledge-based method, FILM, which uses the energy function of a previous globular protein fragment-assembly program (FRAGFOLD) modified by the introduction of a potential term that mimics the presence of the membrane bilayer. For the moment, the method can only predict the structure of small, less than 100 amino acid-long proteins (i.e. proteins for which the conformational space to search is small). Bowie and coworkers [137] predicted the structure of IMP oligomers using knowledge of the oligomer symmetry. They use a simple softened van der Waals potential and Monte Carlo minimization to pack ideal α-helices. The method has been applied to homo-oligomers constituted of monomers with only one TMH. This builds on earlier work by Sansom and coworkers [158]. Methods for predicting 7-TMH IMPs, proteins that alone account for about 50% of all drug targets, have exploited a combination of computational modeling and experimental constraints, the latter ranging from mutation data to low-resolution cryo-electron microscopy data [138]. However, most of these methods have not been extended to model helical IMPs with a different number of TM segments. At the border of homology modeling and *ab initio/de novo*, is the SWISS-MODEL [109] 7TM interface (http://swiss-model.ncifcrf.gov/cgi-bin/sm-gpcr.cgi), explicitly designed to assist in the modeling of 7-TMH proteins. SWISS-MODEL 7TM allows homology modeling based on experimental and computationally designed templates, with the 'computational templates' generated by taking into account constraints from low-resolution experimental data. The user needs only to provide information about the location of the TMHs in the query sequence and to select the most appropriate template from those available.

Very few groups have addressed the *ab initio/de novo* modeling of β-barrel IMPs. The task may be easier than with helical IMPs, since in all known β-barrel IMP structures adjacent TMBs are found to be in contact. Hence, once TMBs have been correctly located in the protein sequence, the only goal is to identify the contacting residues. Jackups and Liang [139] recently published a method that predicts the contacting residues on paired strands. While the reported accuracy is not very high, this seems to be a first important step toward modeling of β-barrel IMP structures.

3. Conclusions

Predicting the structure and function of integral membrane proteins appears to be easier than predicting water-soluble globular proteins. However, our knowledge and understanding of membrane protein structural features has so far been hampered by the limited experimental high-resolution information available. While recently solved structures have shown that membrane proteins are more diverse than initially thought, we are probably still unable to see the full complexity of the problem.

We have presented a survey of the state-of-the-art computational methods developed for predicting structural and functional features in membrane proteins. In the past, for the reasons mentioned, progress in many areas has been slow. However, some approaches have matured and reached high levels of performance and reliability. Transmembrane helix, transmembrane strand and protein topology predictions are among those. Alignment methods specific to membrane proteins are now being developed; however, many questions, such as which substitution matrices are more appropriate for use in transmembrane regions, remain unsolved. Homology modeling has been widely used on membrane proteins in the attempt to fill the gap between the number of known sequences and structures. Still, in most of the cases, models are built using programs originally developed for water-soluble proteins. As the number of available templates increases, is seems likely that new methods will emerge, which are tailored specifically to membrane proteins.

Structural biologists solve more membrane protein structures today than ever before. Structural genomics projects targeting integral membrane proteins are now under way. If experimentalists continue to expand our knowledge of the membrane protein world, it will be possible for computational biologists to develop new and better methods, that may eventually confirm that, yes, membrane protein structure is easy to predict.

Acknowledgments

Thanks to Hans-Erik G. Aronson (Columbia) for computer assistance; thanks to Gunnar von Heijne, Henrik Nielsen, Jannick Bendtsen and Lukas Käll for very generous email clarification and discussion. This work was supported by the Grant RO1-LM07329-01 from the National Library of Medicine (NLM) and the Grants RO1-GM64633-01 and U54-GM75026-01 from the National Institutes of Health (NIH). Last, not least, thanks to all

those who deposit their experimental data in public databases, and to those who maintain these databases.

References

[1] J.U. Bowie, J. Mol. Biol. 272 (1997) 780–789.
[2] G. von Heijne, Y. Gavel, Eur. J. Biochem. 174 (1988) 671–678.
[3] J. Nilsson, B. Persson, G. von Heijne, Proteins 60 (2005) 606–616.
[4] J.U. Bowie, Protein Sci. 8 (1999) 2711–2719.
[5] M. Gimpelev, L.R. Forrest, D. Murray, B. Honig, Biophys. J. 87 (2004) 4075–4086.
[6] H.M. Berman, T. Battistuz, T.N. Bhat, W.F. Bluhm, P.E. Bourne, K. Burkhardt, Z. Feng, G.L. Gilliland, L. Iype, S. Jain, P. Fagan, J. Marvin, D. Padilla, V. Ravichandran, B. Schneider, N. Thanki, H. Weissig, J.D. Westbrook, C. Zardecki, Acta. Crystallogr. D Biol. Crystallogr. 58 (2002) 899–907.
[7] R.P. Riek, I. Rigoutsos, J. Novotny, R.M. Graham, J. Mol. Biol. 306 (2001) 349–362.
[8] J.M. Cuthbertson, D.A. Doyle, M.S. Sansom, Protein Eng. Des. Sel. 18 (2005) 295–308.
[9] D.A. Doyle, J.M. Cabral, R.A. Pfuetzner, A. Kuo, J.M. Gulbis, S.L. Cohen, B.T. Cahit, R. MacKinnon, Science 280 (1998) 69–77.
[10] D. Fu, A. Libson, L.J. Miercke, C. Weitzman, P. Nollert, J. Krucinski, R.M. Stroud, Science 290 (2000) 481–486.
[11] V. Goder, T. Junne, M. Spiess, Mol. Biol. Cell 15 (2004) 1470–1478.
[12] R. Mendez, R. Leplae, M.F. Lensink, S.J. Wodak, Proteins 60 (2005) 150–169.
[13] R. Maggio, F. Novi, M. Scarselli, G.U. Corsini, FEBS J. 272 (2005) 2939–2946.
[14] N. Hulo, A. Bairoch, V. Bulliard, L. Cerutti, E. De Castro, P.S. Langendijk-Genevaux, M. Pagni, C.J. Sigrist, Nucleic Acids Res. 34 (2006) D227–D230.
[15] T.K. Attwood, P. Bradley, D.R. Flower, A. Gaulton, N. Maudling, A.L. Mitchell, G. Moulton, A. Nordle, K. Paine, P. Taylor, A. Uddin, C. Zygouri, Nucleic Acids Res. 31 (2003) 400–402.
[16] V. Ruta, Y. Jiang, A. Lee, J. Chen, R. MacKinnon, Nature 422 (2003) 180–185.
[17] D. Petrey, B. Honig, Mol. Cell 20 (2005) 811–819.
[18] S.H. White, Protein Sci. 13 (2004) 1948–1949.
[19] J. Moult, Curr. Opin. Struct. Biol. 15 (2005) 285–289.
[20] P.E. Bourne, Methods Biochem. Anal. 44 (2003) 501–507.
[21] I.Y. Koh, V.A. Eyrich, M.A. Marti-Renom, D. Przybylski, M.S. Madhusudhan, N. Eswar, O. Grana, F. Pazos, A. Valencia, A. Sali, B. Rost, Nucleic Acids Res. 31 (2003) 3311–3315.
[22] L. Rychlewski, D. Fischer, Protein Sci. 14 (2005) 240–245.
[23] G.E. Tusnady, Z. Dosztanyi, I. Simon, Bioinformatics 20 (2004) 2964–2972.
[24] S. Jayasinghe, K. Hristova, S.H. White, Protein Sci. 10 (2001) 455–458.
[25] C.P. Chen, A. Kernytsky, B. Rost, Protein Sci. 11 (2002) 2774–2791.
[26] M.A. Lomize, A.L. Lomize, I.D. Pogozheva, H.I. Mosberg, Bioinformatics 22 (2006) 623–625.
[27] G.E. Tusnady, Z. Dosztanyi, I. Simon, Bioinformatics 21 (2005) 1276–1277.
[28] M.H. Saier Jr., C.V. Tran, R.D. Barabote, Nucleic Acids Res. 34 (2006) D181–D186.
[29] F. Horn, E. Bettler, L. Oliveira, F. Campagne, F.E. Cohen, G. Vriend, Nucleic Acids Res. 31 (2003) 294–297.
[30] A.V. Katta, R. Marikkannu, R.V. Basaiawmoit, S. Krishnaswamy, In Silico Biol. 4 (2004) 549–561.
[31] M.O. Dayhoff, R.M. Schwartz, B.C. Orcutt, in: M.O. Dayhoff (Ed.), Atlas of Protein Sequence and Structure, vol. 5, 1978, pp. 345–352.
[32] S. Henikoff, J.G. Henikoff, Proc. Natl. Acad. Sci. USA 89 (1992) 10915–10919.
[33] M. Cserzo, E. Wallin, I. Simon, G. von Heijne, A. Elofsson, Protein Eng. 10 (1997) 673–676.
[34] D.T. Jones, W.R. Taylor, J.M. Thornton, FEBS Lett. 339 (1994) 269–275.
[35] P.C. Ng, J.G. Henikoff, S. Henikoff, Bioinformatics 16 (2000) 760–766.
[36] T. Muller, S. Rahmann, M. Rehmsmeier, Bioinformatics 17 (2001) S182–S189.
[37] Y. Liu, D.M. Engelman, M. Gerstein, Gen. Biol. 3 (2002), research 0054.1–54.12..
[38] Q. Gao, A. Chess, Genomics 60 (1999) 31–39.
[39] S. Takeda, S. Kadowaki, T. Haga, H. Takaesu, S. Mitaku, FEBS Lett. 520 (2002) 97–101.
[40] J.S. Lolkema, D.J. Slotboom, Mol. Membr. Biol. 15 (1998) 33–42.
[41] J.D. Clements, R.E. Martin, Eur. J. Biochem. 269 (2002) 2101–2107.
[42] M. Wistrand, L. Kall, E.L.L. Sonnhammer, Protein Sci. (2006) 15.
[43] P.K. Papasaikas, P.G. Bagos, Z.I. Litou, V.J. Promponas, S.J. Hamodrakas, Nuclic Acids Res. 32 (2004) W380–W382.
[44] M. Hedman, H. Deloof, G. Von Heijne, A. Elofsson, Protein Sci. 11 (2002) 652–658.
[45] J.S. Lolkema, D.J. Slotboom, J. Mol. Biol. 327 (2003) 901–909.
[46] J.S. Lolkema, D.J. Slotboom, Mol. Membr. Biol. 22 (2005) 177–189.
[47] M.A. Marti-Renom, M.S. Madhusudhan, A. Sali, Protein Sci. 13 (2004) 1071–1087.
[48] Y.-K. Yu, J.C. Wootton, S.F. Altschul, Proc. Natl. Acad. Sci. USA 100 (2003) 15688–15693.
[49] S.F. Altschul, J.C. Wootton, E.M. Gertz, R. Agarwala, A. Morgulis, A.A. Schaffer, Y.-K. Yu, FEBS J. 272 (2005) 5101–5109.
[50] A. Bateman, L. Coin, R. Durbin, R.D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E.L. Sonnhammer, D.J. Studholme, C. Yeats, S.R. Eddy, Nucleic Acids Res. 32 (2004) D138–D141.
[51] D.T. Jones, W.R. Taylor, J.M. Thornton, Biochemistry 33 (1994) 3038–3049.
[52] Y. Shafrir, H.R. Guy, Bioinformatics 20 (2004) 758–769.
[53] C. Bissantz, A. Logean, D. Rognan, J. Chem. Inf. Comput. Sci. 44 (2004) 1162–1176.
[54] F. Fanelli, P.G. De Benedetti, Chem. Rev. 105 (2005) 3297–3351.
[55] A. Giorgetti, P. Carloni, Curr. Opin. Chem. Biol. 7 (2003) 150–156.
[56] C. Notredame, D.G. Higgins, J. Heringa, J. Mol. Biol. 302 (2000) 205–217.
[57] C.L. Tang, L. Xie, I.Y.Y. Koh, S. Posy, E. Alexov, B. Honig, J. Mol. Biol. 334 (2003) 1043–1062.
[58] K. Ginalski, N.V. Grishin, A. Godzik, L. Rychlewski, Nuclic Acids Res. 33 (2005) 1874–1891.
[59] A. del Sol Mesa, F. Pazos, A. Valencia, J. Mol. Biol. 326 (2003) 1289–1302.
[60] I. Mihalek, I. Res, O. Lichtarge, J. Mol. Biol. 336 (2004) 1265–1282.
[61] U. Gobel, C. Sander, R. Schneider, A. Valencia, Proteins 18 (1994) 309–317.
[62] F. Pazos, M. Helmer-Citterich, G. Ausiello, A. Valencia, J. Mol. Biol. 271 (1997) 511–523.
[63] F. Pazos, A. Valencia, Proteins 47 (2002) 219–227.
[64] K. Palczewski, T. Kumasaka, T. Hori, C.A. Behnke, H. Motoshima, B.A. Fox, I. Le Trong, D.C. Teller, T. Okada, R.E. Stenkamp, M. Yamamoto, M. Miyano, Science 289 (2000) 739–745.
[65] M. Filizola, H. Weinstein, FEBS J. 272 (2005) 2926–2938.
[66] Y. Jiang, A. Lee, J. Chen, V. Ruta, M. Cadene, B.T. Chait, R. MacKinnon, Nature 423 (2003) 33–41.
[67] S.J. Fleishman, O. Yifrach, N. Ben-Tal, J. Mol. Biol. 340 (2004) 307–318.
[68] S.J. Fleishman, V.M. Unger, M. Yeager, N. Ben-Tal, Mol. Cell 15 (2004) 879–888.
[69] J. Kyte, R.F. Doolittle, J. Mol. Biol. 157 (1982) 105–132.
[70] A.K. Chamberlain, Y. Lee, S. Kim, J.U. Bowie, J. Mol. Biol. 339 (2004) 471–479.
[71] M. Monne, M. Hermansson, G. von Heijne, J. Mol. Biol. 288 (1999) 141–145.
[72] Y. Pilpel, N. Ben-Tal, D. Lancet, J. Mol. Biol. 294 (1999) 921–935.
[73] B. Rost, R. Casadio, P. Fariselli, C. Sander, Protein Sci. 4 (1995) 521–533.
[74] B. Rost, Methods Enzymol. 266 (1996) 525–539.

[75] Z. Yuan, J.S. Mattick, R.D. Teasdale, J. Comput. Chem. 25 (2004) 632–636.
[76] L.K. Hansen, P. Salamon, IEEE Trans. Pattern Anal. Mach. Intell. 12 (1990) 993–1001.
[77] T.G. Dietterich, Lect. Notes Comput. Sci. 1857 (2000) 1–15.
[78] P.L. Martelli, P. Fariselli, R. Casadio, Bioinformatics 19 (Suppl. 1) (2003) i205–i211.
[79] P.D. Taylor, T.K. Attwood, D.R. Flower, Nucleic Acids Res. 31 (2003) 3698–3700.
[80] I. Nilsson, B. Persson, G. von Heijne, FEBS Lett. 486 (2000) 267–269.
[81] P.G. Bagos, T. Liakopoulos, S. Hamodrakas, BMC Bioinformatics 6 (2005) 7.
[82] M. Cserzo, F. Eisenhaber, B. Eisenhaber, I. Simon, Bioinformatics 20 (2004) 136–137.
[83] T. Hirokawa, S. Boon-Chieng, S. Mitaku, Bioinformatics 14 (1998) 378–379.
[84] E.L.L. Sonnhammer, G. von Heijne, A. Krogh, in: ISMB-98, AAAI Press, 1998.
[85] S. Moller, M.D.R. Croning, R. Apweiler, Bioinformatics 17 (2001) 646–653.
[86] G.E. Tusnady, I. Simon, Bioinformatics 17 (2001) 849–850.
[87] B. Rost, P. Fariselli, R. Casadio, Protein Sci. 5 (1996) 1704–1718.
[88] D. Juretic, L. Zoranic, D. Zucic, J. Chem. Inf. Comput. Sci. 42 (2002) 620–632.
[89] A. Krogh, B. Larsson, G. von Heijne, E.L. Sonnhammer, J. Mol. Biol 305 (2001) 567–580.
[90] B. Persson, P. Argos, J. Protein Chem. 16 (1997) 453–457.
[91] P.G. Bagos, T.D. Liakopoulos, I.C. Spyropoulos, S.J. Hamodrakas, Nucleic Acids Res. 32 (2004) W400–W404.
[92] P.G. Bagos, T.D. Liakopoulos, I.C. Spyropoulos, S.J. Hamodrakas, BMC Bioinformatics 5 (2004) 29.
[93] H.R. Bigelow, D.S. Petrey, J. Liu, D. Przybylski, B. Rost, Nucleic Acids Res. 32 (2004) 2566–2577.
[94] G.E. Tusnady, Z. Dosztanyi, I. Simon, Nucleic Acids Res. 33 (2005) D275–D278.
[95] P.L. Martelli, P. Fariselli, A. Krogh, R. Casadio, Bioinformatics 18 (Suppl.1) (2002) S46–S53.
[96] I. Collinson, Biochem. Soc. Trans. 33 (2005) 1225–1230.
[97] J.W. de Gier, J. Luirink, Mol. Microbiol. 40 (2001) 314–322.
[98] S.H. White, G. von Heijne, Curr. Opin. Struct. Biol. 15 (2005) 378–386.
[99] H. Andersson, G. von Heijne, FEBS Lett. 347 (1994) 169–172.
[100] L. Kall, A. Krogh, E.L. Sonnhammer, J. Mol. Biol. 338 (2004) 1027–1036.
[101] A. Bernsel, G. Von Heijne, Protein Sci. 14 (2005) 1723–1728.
[102] J. Schultz, F. Milpetz, P. Bork, C.P. Ponting, Proc. Natl. Acad. Sci. USA 95 (1998) 5857–5864.
[103] S. Yohannan, S. Faham, D. Yang, J.P. Whitelegge, J.U. Bowie, Proc. Natl. Acad. Sci. USA 101 (2004) 959–963.
[104] I. Rigoutsos, P. Riek, R.M. Graham, J. Novotny, Nucleic Acids Res. 31 (2003) 4625–4631.
[105] M. Levitt, J. Mol. Biol. 226 (1992) 507–533.
[106] A. Sali, T.L. Blundell, J. Mol. Biol. 234 (1993) 779–815.
[107] D. Petrey, Z.X. Xiang, C.L. Tang, L. Xie, M. Gimpelev, T. Mitros, C.S. Soto, S. Goldsmith-Fischman, A. Kernytsky, A. Schlessinger, I.Y.Y. Koh, E. Alexov, B. Honig, Proteins 53 (2003) 430–435.
[108] B. Wallner, A. Elofsson, Protein Sci. 14 (2005) 1315–1327.
[109] K. Arnold, L. Bordoli, J. Kopp, T. Schwede, Bioinformatics 22 (2006) 195–201.
[110] A.A. Canutescu, A.A. Shelenkov, R.L. Dunbrack, Protein Sci. 12 (2003) 2001–2014.
[111] Z.X. Xiang, B. Honig, J. Mol. Biol. 311 (2001) 421–430.
[112] M.P. Jacobson, D.L. Pincus, C.S. Rapp, T.J.F. Day, B. Honig, D.E. Shaw, R.A. Friesner, Proteins 55 (2004) 351–367.
[113] P.I.W. de Bakker, M.A. DePristo, D.F. Burke, T.L. Blundell, Proteins 51 (2003) 21–40.
[114] L.R. Forrest, T.B. Woolf, Proteins Struct. Funct. Genet. 52 (2003) 492–509.
[115] E.L. Mehler, X. Periole, S.A. Hassan, H. Weinstein, J. Comput. Aid. Mol. Des. 16 (2002) 841–853.
[116] Z. Xiang, C.S. Soto, B. Honig, Proc. Natl. Acad. Sci. USA 99 (2002) 7432–7437.
[117] G.V. Nikiforovich, G.R. Marshall, Biophys. J. 89 (2005) 3780–3789.
[118] J.J. Vincent, C.H. Tai, B.K. Sathyanarayana, B. Lee, Proteins 61 (Suppl. 7) (2005) 67–83.
[119] J.J. Carrillo, J.F. Lopez-Gimenez, G. Milligan, Mol. Pharmacol. 66 (2004) 1123–1137.
[120] Y. Liang, D. Fotiadis, S. Filipek, D.A. Saperstein, K. Palczewski, A. Engel, J. Biol. Chem. 278 (2003) 21655–21662.
[121] A. Senes, M. Gerstein, D.M. Engelman, J. Mol. Biol. 296 (2000) 921–936.
[122] M.A. Lemmon, J.M. Flanagan, H.R. Treutlein, J. Zhang, D.M. Engelman, Biochemistry 31 (1992) 12719–12725.
[123] K.R. MacKenzie, J.H. Prestegard, D.M. Engelman, Science 276 (1997) 131–133.
[124] W.P. Russ, D.M. Engelman, J. Mol. Biol. 296 (2000) 911–919.
[125] R.A. Melnyk, A.W. Partridge, C.M. Deber, J. Mol. Biol. 315 (2002) 63–72.
[126] A.K. Doura, K.G. Fleming, J. Mol. Biol. 343 (2004) 1487–1497.
[127] F.J. Kobus, K.G. Fleming, Biochemistry 44 (2005) 1464–1470.
[128] S. Kim, T.J. Jeon, A. Oberai, D. Yang, J.J. Schmidt, J.U. Bowie, Proc. Natl. Acad. Sci. USA 102 (2005) 14278–14283.
[129] S. Jiang, I.A. Vakser, Proteins 40 (2000) 429–435.
[130] S. Jiang, I.A. Vakser, Protein Sci. 13 (2004) 1426–1429.
[132] S.J. Fleishman, N. Ben-Tal, J. Mol. Biol. 321 (2002) 363–378.
[133] W.F. DeGrado, H. Gratkowski, J.D. Lear, Protein Sci. 12 (2003) 647–665.
[134] S.J. Fleishman, V.M. Unger, N. Ben-Tal, Trends Biochem. Sci. 31 (2006) 106–113.
[135] R.V. Pappu, G.R. Marshall, J.W. Ponder, Nat. Struct. Biol. 6 (1999) 50–55.
[136] M. Pellegrini-Calace, A. Carotti, D.T. Jones, Proteins 50 (2003) 537–545.
[137] S. Kim, A.K. Chamberlain, J.U. Bowie, J. Mol. Biol. 329 (2003) 831–840.
[138] F. Fanelli, P.G. DeBenedetti, Chem. Rev. (2005).
[139] Ronald J. Jackups, J. Liang, J. Mol. Biol. 354 (2005) 979–993.
[140] R.Y. Kahsay, G. Gao, L. Liao, Bioinformatics 21 (2005) 1853–1858.
[141] J. Liu, B. Rost, Protein Sci. 10 (2001) 1970–1979.
[142] E. Wallin, G. von Heijne, Protein Sci. 7 (1998) 1029–1038.
[143] C.G. Knight, R. Kassen, H. Hebestreit, P.B. Rainey, Proc. Natl. Acad. Sci. USA 101 (2004) 8390–8395.
[144] K. Melen, A. Krogh, G. von Heijne, J. Mol. Biol. 327 (2003) 735–744.
[145] D.O. Daley, M. Rapp, E. Granseth, K. Melen, D. Drew, G. von Heijne, Science 308 (2005) 1321–1323.
[146] B. Boeckmann, A. Bairoch, R. Apweiler, M.C. Blatter, A. Estreicher, E. Gasteiger, M.J. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout, M. Schneider, Nucleic Acids Res. 31 (2003) 365–370.
[147] N.J. Mulder, R. Apweiler, T.K. Attwood, A. Bairoch, A. Bateman, D. Binns, P. Bradley, P. Bork, P. Bucher, L. Cerutti, R. Copley, E. Courcelle, U. Das, R. Durbin, W. Fleischmann, J. Gough, D. Haft, N. Harte, N. Hulo, D. Kahn, A. Kanapin, M. Krestyaninova, D. Lonsdale, R. Lopez, I. Letunic, M. Madera, J. Maslen, J. McDowall, A. Mitchell, A.N. Nikolskaya, S. Orchard, M. Pagni, C.P. Ponting, E. Quevillon, J. Selengut, C.J. Sigrist, V. Silventoinen, D.J. Studholme, R. Vaughan, C.H. Wu, Nucleic Acids Res. 33 (2005) D201–D205.
[148] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, Nucleic Acids Res. 25 (1997) 3389–3402.
[149] J.D. Thompson, D.G. Higgins, T.J. Gibson, Nucleic Acids Res. 22 (1994) 4673–4680.
[150] L.J. McGuffin, K. Bryson, D.T. Jones, Bioinformatics 16 (2000) 404–405.
[151] B. Cao, A. Porollo, R. Adamczak, M. Jarrell, J. Meller, Bioinformatics 22 (2006) 303–309.

[152] T.D. Liakopoulos, C. Pasquier, S.J. Hamodrakas, Protein Eng. 14 (2001) 387–390.
[153] G. von Heijne, J. Mol. Biol. 225 (1992) 487–494.
[154] F.S. Berven, K. Flikka, H.B. Jensen, I. Eidhammer, Nucleic Acids Res. 32 (2004) W394–W399.
[155] A.G. Garrow, A. Agnew, D.R. Westhead, Nucleic Acids Res. 33 (2005) W188–W192.
[156] A.G. Garrow, A. Agnew, D.R. Westhead, BMC Bioinformatics 6 (2005) 56.
[157] L.R. Forrest, C.L. Tang, B. Honig, Biophys. J. 91 (2006) 508–517.
[158] I.D. Kerr, R. Sankararamakrishnan, O.S. Smart, M.S. Sansom, Biophys. J. 67 (4) (1994) 1501–1515.
[159] M. Tress, I. Ezkurdia, O. Grana, G. Lopez, A. Valencia, Proteins 61 (Suppl. 7) (2005) 27–45.