Bioinformatics, 32, 2016, i665–i671
doi: 10.1093/bioinformatics/btw444
ECCB 2016

PRED-TMBB2: improved topology prediction and detection of beta-barrel outer membrane proteins

Konstantinos D. Tsirigos[1,2], Arne Elofsson[1] and Pantelis G. Bagos[2,*]

[1]Department of Biochemistry and Biophysics, Science for Life Laboratory, Swedish E-Science Research Center, Stockholm University, 17121 Solna, Sweden and [2]Department of Computer Science and Biomedical Informatics, University of Thessaly, 35100 Lamia, Greece

*To whom correspondence should be addressed.

Abstract

Motivation: The PRED-TMBB method is based on Hidden Markov Models and is capable of predicting the topology of beta-barrel outer membrane proteins and discriminate them from water-soluble ones. Here, we present an updated version of the method, PRED-TMBB2, with several newly developed features that improve its performance. The inclusion of a properly defined end state allows for better modeling of the beta-barrel domain, while different emission probabilities for the adjacent residues in strands are used to incorporate knowledge concerning the asymmetric amino acid distribution occurring there. Furthermore, the training was performed using newly developed algorithms in order to optimize the labels of the training sequences. Moreover, the method is retrained on a larger, non-redundant dataset which includes recently solved structures, and a newly developed decoding method was added to the already available options. Finally, the method now allows the incorporation of evolutionary information in the form of multiple sequence alignments.

Results: The results of a strict cross-validation procedure show that PRED-TMBB2 with homology information performs significantly better compared to other available prediction methods. It yields 76% in correct topology predictions and outperforms the best available predictor by 7%, with an overall SOV of 0.9. Regarding detection of beta-barrel proteins, PRED-TMBB2, using just the query sequence as input, achieves an MCC value of 0.92, outperforming even predictors designed for this task and are much slower.

Availability and Implementation: The method, along with all datasets used, is freely available for academic users at http://www.compgen.org/tools/PRED-TMBB2.

Contact: pbagos@compgen.org

1 Introduction

Beta-barrel outer membrane proteins (OMPs) are localized in the outer membrane of Gram-negative bacteria and in the outer membranes of plastids and mitochondria. Their membrane-spanning segments are formed by short amphipathic beta-strands that create a closed structure resembling a barrel (Schulz, 2003). The difficulty in obtaining crystals suitable for high-resolution studies of OMPs has resulted in their under-representation in the Protein Data Bank (Rose *et al.*, 2013).

Given these difficulties, and because many beta-barrel OMPs attract an increased medical interest, several approaches have been made towards the development of topology prediction algorithms for this type of proteins. These methods are based grossly on hydrophobicity analysis (Zhai and Saier, 2002), statistical preferences of amino acids (Wimley, 2002), remote homology detection (Remmert *et al.*, 2009), Hidden Markov Models (Bagos *et al.*, 2004a, b; Bigelow *et al.*, 2004; Hayat *et al.*, 2016; Martelli *et al.*, 2002; Savojardo *et al.*, 2013), feed-forward Neural Networks (Gromiha *et al.*, 2004; Jacoboni *et al.*, 2001) and radial basis function Neural Networks (Ou *et al.*, 2008, 2010). PRED-TMBB (Bagos *et al.*, 2004,b,c) was introduced in 2004 and is still one of the most widely used methods for topology prediction and discrimination of beta-barrel outer membrane proteins. It was based on a HMM architecture and was one of the first methods to perform well in both tasks. To date, BOCTOPUS2 (Hayat *et al.*, 2016), which is the successor to BOCTOPUS (Hayat and Elofsson, 2012) method, is the

Downloaded from https://academic.oup.com/bioinformatics/article/32/17/i665/2450774 by guest on 25 April 2024

© The Author 2016. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

most accurate in topology predictions. The main improvement in BOCTOPUS2 is the exploitation of the dyad-repeat pattern of lipid and pore-facing residues in bacterial beta-barrel proteins. A previously presented benchmark study of several topology prediction methods showed that HMMs are the most reliable predictors for beta-barrels (Bagos *et al.*, 2005). The same also holds for alpha-helical membrane proteins, as shown in (Viklund and Elofsson, 2004) and (Tsirigos *et al.*, 2015).

Other existing methods aim specifically at the identification of beta-barrel proteins and discrimination of them from other classes of proteins in proteome-wide analyses. The most popular are BetAware (Savojardo *et al.*, 2013), BOMP (Berven *et al.*, 2004), the Freeman–Wimley beta-Barrel Analyzer (Freeman and Wimley, 2010), HHomp (Remmert *et al.*, 2009), PSORTb (Yu *et al.*, 2010), SSEA-OMP (Yan *et al.*, 2011), TMB-Hunt (Garrow *et al.*, 2005a, b), SOSUIgramN (Imai *et al.*, 2008) and TMBETADISC-RBF (Ou *et al.*, 2008). Some of the aforementioned tools make use of evolutionary information in the form of multiple sequence alignments (MSAs), which are a bottleneck in large-scale analyses.

Finally, special purpose biological databases that include families of beta-barrel proteins have also been available to the public. These include TCDB (Saier *et al.*, 2006), PDBTM (Kozma *et al.*, 2013), TOPDB (Tusnady *et al.*, 2008), PSORTdb (Yu *et al.*, 2011), OPM (Lomize *et al.*, 2006) and Mptopo (Jayasinghe *et al.*, 2001), and the most recent ones, OMPdb (Tsirigos *et al.*, 2011) and TMBB-DB (Freeman and Wimley, 2012).

Here, we present a new method, PRED-TMBB2, which shows superior predictive ability over previously developed algorithms. Apart from the use of a larger training set, this is achieved by incorporating some novel features into the model itself, by applying newly derived decoding algorithms, by using a modified technique that optimizes the labels of the training sequences and, finally, by incorporating evolutionary information in the form of MSAs. We also show that PRED-TMBB2 can efficiently differentiate between beta-barrel and non-beta-barrel proteins using only single sequences as input, which makes it ideal for scanning large datasets.

2 Methods

2.1 Training and test set

For training, we initially retrieved all outer membrane proteins with known three-dimensional structures deposited in PDB (Rose *et al.*, 2013), using the boundary definitions of the beta strands as deposited in the PDBTM database (Kozma *et al.*, 2013). Some non-canonical beta-barrel structures, such as TolC (Koronakis *et al.*, 2000), α-hemolysin from *Staphylococcus aureus* (Song *et al.*, 1996) and the mycobacterial (Gram-positive) outer membrane channel MspA (Faller *et al.*, 2004), as well as the mitochondrial porin (Bayrhuber *et al.*, 2008), were removed from the set.

In the next step, we used the 2nd algorithm of Hobohm *et al.* (1992) to remove sequences having more than 30% sequence similarity in a BLAST (Altschul *et al.*, 1997) alignment in a length of more than 80 residues. This procedure resulted in 49 outer membrane proteins that constitute our final training set. In order to reduce the risk of over-training, we further divided the set according to the family classification of OMPdb (Tsirigos *et al.*, 2011). OMPdb is based on the Pfam classification (Finn *et al.*, 2016) but includes additional families of outer membrane proteins that are not deposited in Pfam. This way, the 49 proteins were divided into 30 families, and each set was used in a strict cross-validation procedure, where members of one family were removed from the set, the method was trained using the proteins of the remaining families, and the whole process was repeated.

Since many of the proteins in our training set were also used in BOCTOPUS2 training set, we decided to perform our benchmark using the 42 proteins used for training BOCTOPUS2 (Hayat *et al.*, 2016) dataset, so that both tools would be evaluated fairly. Out of the 42 proteins, 28 were already present in PRED-TMBB2's training set, whereas, for the remaining 14, we used the cross-validated models with respect to the corresponding family in which each one belongs.

For testing PRED-TMMB2 in terms of discriminating capability between beta-barrel and non-beta-barrel proteins, we used a positive and a negative dataset. The latter is a non-redundant (20% sequence similarity) PDB dataset, with 8858 sequences (Freeman and Wimley, 2010). We decided to use the respective full-length sequences (as we did in the training of PRED-TMBB2) and removed some proteins that were included more than once and some that were not, in fact, beta-barrels. We further performed an additional redundancy reduction at a cut-off of 30% in the full sequences, using CD-HIT (Huang *et al.*, 2010). We chose to use full sequences in both topology prediction and detection of beta-barrel proteins, since such a procedure would resemble a real-life situation, where a whole proteome would be scanned for outer membrane proteins. In the end, 7571 proteins remained. The positive dataset is comprised of the proteins which belong to the seed alignments of the 92 families that OMPdb currently has, on which we performed a redundancy reduction following the same principles as for training of PRED-TMBB2. This procedure left us with 1009 protein sequences, which we consider, with a good confidence, to be transmembrane beta-barrel proteins.

2.2 The HMM architecture

The HMM architecture consists of three sub-models that correspond to periplasmic loops, TM beta-strands and extracellular loops and is similar to the initial model, with several modifications (Fig. 1). First of all, we explicitly added an end state to the model, which is particularly useful in correct modeling of the beta-barrel domain. It is well known that in all the available structures of outer membrane proteins from Gram-negative bacteria, both N- and C-terminal are located in the periplasmic space (Schulz, 2003). In the first version of the method (Bagos *et al.*, 2004b), this feature was partially exploited by fixing the N-terminal part, whereas now, with the addition of the end state, the same is accomplished for the C-terminal as well. By allowing transitions to the end state only from states of the cytoplasmic loops, we achieve better correspondence between the mathematical formalism of the model and the features of the known beta-barrels.

Furthermore, we now treat the emission probabilities of the strands having their N-terminal to the periplasmic space differently as compared to those having the opposite orientation. This asymmetry in the distribution of amino acids is something that has been exploited earlier in HMM prediction methods of beta-barrels (Bigelow *et al.*, 2004), and has been validated statistically (Chamberlain and Bowie, 2004; Slusky and Dunbrack, 2013). Another modification in the emission probabilities was performed in order to model explicitly the so-called 'positive-outside rule', that is, the preference of positively charged residues to localize in the extracellular loops (Jackups and Liang, 2005).

Finally, the transition probabilities were modified in the periplasmic loop with the addition of a self-transitioning state in order to accommodate some recently solved structures with long periplasmic loops (the initial model allowed periplasmic loops with a maximum

Downloaded from https://academic.oup.com/bioinformatics/article/32/17/i665/2450774 by guest on 25 April 2024

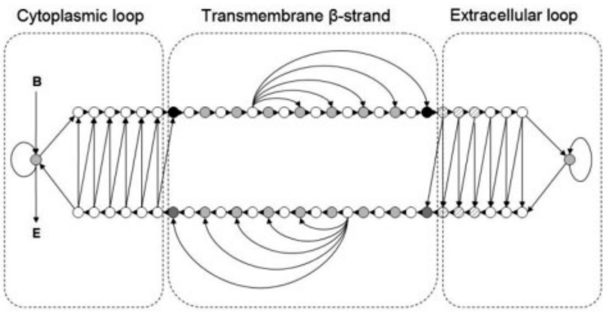

Fig. 1. Schematic representation of the HMM architecture used in PRED-TMBB2. Within each sub-model that corresponds to the labels (cytoplasmic, TM, extracellular), states with the same shading share the same emission probabilities. B represents the start state and E the end state. In total, the model contains 62 states with 171 freely estimated emission probabilities and 42 freely estimated transitions, yielding a total of 213 free estimated parameters

length of 12 residues). Similarly to the initial model, states that are believed to possess the same properties, share the same emission probabilities (Fig. 1). Thus, the HMM is still very parsimonious since it contains only 171 freely estimated emission probabilities and 42 freely estimated transitions, yielding a total of 213 free parameters.

2.3 Algorithms

For training, we used the conditional maximum likelihood (CML) approach for labeled sequences (Krogh, 1997). The CML approach has several advantages over traditional maximum likelihood, but it suffers mainly from two disadvantages: first, it lacks a simple and globally convergent algorithm for parameter estimation (such as the Baum–Welch algorithm); and second, it requires good-quality labels in order to work efficiently. Towards this end, the model was trained according to a modified gradient descent algorithm which offers robustness and fast convergence (Bagos *et al.*, 2004a). Moreover, another improvement consists of redesigning the training procedure in order to optimize the predictive accuracy as much as possible. Instead of just using the labeled sequences derived from the three-dimensional structures, we now performed an extra optimization step, re-defining the labels as proposed in (Krogh *et al.*, 2001). Briefly, a model was initially estimated using the Baum-Welch algorithm for labeled sequences (Krogh, 1994). Afterwards, the labels of the sequences were deleted in a region flanking three residues in each direction of the end of a membrane-spanning strand, and predictions were performed with the Viterbi algorithm using the model estimated from step 1. The final model was estimated using the labels derived from step 2 with the modified gradient-descent method for CML training.

For decoding, apart from the standard Viterbi algorithm (Forney, 1973) and the N-best decoder (Krogh, 1997), we also considered the Posterior–Viterbi algorithm (Fariselli *et al.*, 2005) and the optimal accuracy posterior decoder (OAPD) (Kall *et al.*, 2005), that both optimize the path derived from the posterior probabilities. We also evaluated a previously developed dynamic programming algorithm, but the initial results suggested that the optimal accuracy posterior decoder performs slightly better in all tests (data not shown). Taking into account the fact that OAPD is also used in the subsequent steps with the incorporation of information from homologs, the OAPD was chosen as the default decoding method and we do not report results from other decoders.

Furthermore, we have implemented previously presented versions of all the above mentioned decoding algorithms that can impose constraints on the prediction arising from any kind of prior knowledge concerning the protein at hand. For instance, if experimentally derived information regarding the localization of a particular segment exists, this could be incorporated to the prediction, outputting better results. This feature is unique among the predictors of beta-barrel membrane proteins, since until now it was available only for alpha-helical membrane proteins predictors (Bagos *et al.*, 2006; Kall *et al.*, 2005; Tusnady and Simon, 2001) and can be used in various ways. For example, the obvious way is to provide the user with the option of constraining some part(s) of the sequence to a predefined localization according to experimentally derived information (from experiments with antibodies, cysteine-scanning mutagenesis, proteolysis, gene fusions, etc.). Even though this practice is widely used for alpha-helical membrane proteins, there are several cases of outer membrane proteins in which it could be beneficial as well. Moreover, some outer membrane proteins are known to possess long N-, and C-terminal regions extending far away from the transmembrane beta-barrel domain (the OmpA family, the Autotransporter family, the Initimin/Invasin family and so on). These regions contain known Pfam domains that can be used for imposing constraints, using the modified algorithms described above. This technique was applied for the first time in alpha-helical membrane proteins (Bernsel and Von Heijne, 2005), and in order to use it for outer membrane beta-barrels, we use the Pfam phmm collection from which we have removed the models that we have identified as characteristic for beta-barrel proteins according to OMPdb. Each query sequence is scanned using the *hmmscan* package from HMMER3 (Eddy, 2011) against these models prior to the actual submission to the server.

2.4 Multiple sequence alignments

In many protein structure prediction problems, a significant gain in prediction accuracy can be obtained by incorporating evolutionary information in the form of multiple sequence alignments (MSAs). This is also the case for prediction of transmembrane beta-barrels. The majority of the existing methods use such information in the form of profiles generated by PSI-BLAST (Altschul *et al.*, 1997). Since PRED-TMBB was based on standard HMM that uses single sequences, we used a modified version of the method developed by Käll and coworkers (Kall *et al.*, 2005). Briefly, given a query sequence and a MSA of its homologs, predictions with the single sequence method are obtained on each of the homologs. Then, the predicted labels (I, M and O, for Intracellular, Membrane and Outer loops, respectively) are mapped on the alignment and averaged for each position of the query sequence. This creates a 'posterior label probability' (PLP) table for the query sequence that contains information from the MSA. In the last step, the OAPD is applied and the final prediction is obtained.

For finding the homologs and performing the alignments, we chose to use the *jackhmmer* program from the HMMER3 package. *jackhmmer* finds the homologs and simultaneously performs the multiple alignment with a sensitivity that is comparable (if not better) to PSI-BLAST. We managed to reduce the running time using OMPdb as a reference database. Thus, a query sequence is first scanned against the profiles contained in OMPdb and, if a significant hit is found, the *jackhmmer* search is performed only against the members of the respective family. For a potential newly found protein (i.e. a protein that does not belong to any known family), a

Downloaded from https://academic.oup.com/bioinformatics/article/32/17/i665/2450774 by guest on 25 April 2024

similar search is performed against the nr90 database. This step can be time-consuming but we anticipate it will rarely happen.

2.5 Detection of OMPs

We finally investigated the ability of PRED-TMBB2 to discriminate transmembrane beta-barrels from other classes of proteins (globular and alpha-helical inner membrane proteins). The initial version of PRED-TMBB used the length-normalized log-probability of a sequence, in order to achieve detection rates at a sensitivity of approximately 89% (using a small benchmark dataset). Here, we also used the same metric, but we additionally explored the use of several other metrics. First of all, we used the log-odds scores derived from comparing the probability under the model, compared to the probability of a null model (a single state with emissions derived from Uniprot (Magrane and Consortium, 2011)). We also used the length of the sequence, the number of predicted TM strands, the presence of a signal peptide, the reliability of the prediction and, finally, two indicator variables; these correspond to a hit in one of the characteristic non-beta-barrel Pfam domains we identified and to a hit in OMPdb's pHMMs in an *hmmscan* search. The last metric was chosen since we observed that, in many cases, proteins belonging to one family of OMPdb have detectable hits in the models of other families, even though these are with a much lower score as compared to the family's trusted cut-off. For proteins that belong to a known family, only the putative second (insignificant) hit was counted in order to mimic a situation when the predictor will encounter a member of a previously unseen family. These eight metrics were evaluated with logistic regression. In order to avoid overfitting, the proteins of the positive test set that belong to a family with known structure, were submitted to the HMM predictor in a cross-validation mode (i.e. they were tested using the model that was built excluding the members of the particular family).

2.6 Evaluation criteria

We used a number of metrics to evaluate PRED-TMBB2; the model's accuracy in topology prediction was estimated in a strict family-wise cross-validation procedure, as described in the Section 2.1 (Training and test set). We evaluated the correctly predicted residues in a three-state mode (Q_3), the segments overlap measure (SOV) (Zemla, *et al.*, 1999), the number of proteins with correctly determined topologies and the number of proteins with correctly predicted number of strands. For all comparisons, we used the annotated strands present in PDBTM (Kozma *et al.*, 2013) as reference (i.e. observed transmembrane strands).

Regarding the discrimination performance, PRED-TMBB2, along with all methods that were evaluated, were tested based in terms of sensitivity (the proportion of TMBBs positively identified in the datasets of known TMBBs), specificity (the proportion of non-TMBBs eliminated in the datasets with known non-TMBBs) and the Matthews correlation coefficient (MCC), a metric of overall efficiency of a prediction algorithm (Matthews, 1975). In cases like constrained predictions, or detection of beta-barrels, another useful metric that was employed was the reliability of the prediction, as described in Melen *et al.* (2003).

3 Results

In Table 1, we present the cross-validated results on the training dataset, both for the single- and the multi-sequence version of PRED-TMBB2. The results clearly show that the incorporation of evolutionary information increases the prediction accuracy greatly.

Table 1. Cross-validated results on the training dataset (49 proteins)

Method	Q3	Correct #TM	Correct top	SOV
PRED-TMBB2-MSA	0.880	46	38	0.900
PRED-TMBB2-single	0.850	31	19	0.828

Out of the 49 proteins, PRED-TMBB2 predicts the correct topology in 78% of them, whereas it manages to predict the correct number of strands in most of the proteins (94%).

We also tried to estimate the contribution of the changes in the model architecture, so we compared PRED-TMBB2 with and without evolutionary information against the old version of PRED-TMBB on the training set. For the same task, we also used the old model and performed a retraining, in order to evaluate the impact of the larger datas*et al*one. The experimentation was performed in the self-consistency phase and, after the model architecture and the other settings were chosen, we evaluated the performance in the cross-validation. The incorporation of information from homologs resulted in a significant increase in all measures (Q3, SOV and correctly predicted topologies) even for the old predictor.

The larger training s*et al*one resulted in no significant increase of Q3 and SOV, but improved the number of correctly predicted topologies. The changes in the model architecture were proven to be beneficial in all measures. As shown in Figure 2, by combining all the enhancements (larger set, modifications in model architecture and use of alignments), PRED-TMBB2 achieved the highest performance in all measures, thus justifying our choice.

The benchmark results on the cross-validation set and the comparison against the other predictors (on the BOCTOPUS2 dataset of 42 proteins) are shown in Table 2. Even from the results of the cross-validation test, it is obvious that PRED-TMBB2 performs significantly better than the previous version of the method (PRED-TMBB), and also better compared to other available methods. BOCTOPUS2 ranks second in the comparison in terms of number of correctly predicted topologies by 7%, but is better than PRED-TMBB2 in terms of Q_3 and SOV. The performance of all other methods is much lower, even though they actually contain some of the proteins in their training sets.

PRED-TMBB2 performs also very well in discriminating OMPs from globular and alpha-helical transmembrane proteins. The key advantage of PRED-TMBB2 compared to other methods that can

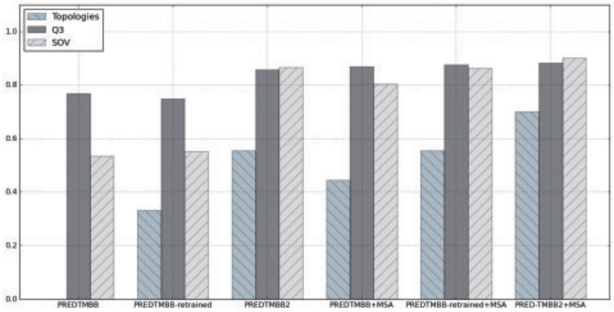

Fig. 2. The comparison of PRED-TMBB, PRED-TMBB retrained and PRED-TMBB2, with and without evolutionary information. The version of PRED-TMBB2 which combines all enhancements (larger set, modifications in model architecture and use of alignments) achieved the highest performance in all measures (Q3, SOV and correctly predicted topologies). The most remarkable improvement is in the fraction of correctly predicted topologies

Downloaded from https://academic.oup.com/bioinformatics/article/32/17/i665/2450774 by guest on 25 April 2024

Table 2. Benchmark results on the training dataset of BOCTOPUS2 (42 proteins)

Method	Q3	Correct #TM	Correct top	SOV
PRED-TMBB2-MSA	0.892	39	32	0.905
PRED-TMBB2-single	0.868	22	14	0.840
BOCTOPUS2	0.914	35	29	0.925
PROFtmb (8)	0.840	24	18	0.751
HMM-B2TMR (18)	0.839	25	18	0.783
PRED-TMBB (16)	0.826	21	12	0.678
BetAware (38)	0.851	23	10	0.725
TMBETAPRED-RBF (26)	0.851	19	8	0.559

Denoted in parentheses is the number of proteins (out of 42) that were present in the respective training set of each method. PRED-TMBB2 and BOCTOPUS2 results are reported based on a cross-validation test. Concerning topology, PRED-TMBB2-MSA performs the best of all methods tested, while BOCTOPUS2 shows the highest SOV.

discriminate OMPs is that it does that with a very high accuracy and without the need for homologous proteins. After all, the inclusion of MSAs benefits mainly the topology prediction and not the classification process. Using the eight metrics described earlier, a logistic regression classifier achieves 91.87% sensitivity and 99.14% specificity on the 1009 OMPs from OMPdb (positive set) and the 7571 non-OMPs from the set of Wimley (negative set), resulting in a MCC value of 0.92. These results are much better compared to all currently available methods used for detection of OMPs (Table 3), both the single-sequence- and the multiple-sequence-based ones, with the exceptions of HHomp and BOCTOPUS2, which, however, are the slowest of all methods and cannot be used to scan entire proteomes.

Our method is capable of correctly excluding non-OMPs with high reliability, something that is desirable in proteome-wide applications. There are also cases of programs that show high success rates for one measurement at the expense of the other; for example BetAware has a very high specificity but its sensitivity is quite low. The opposite goes for PROFtmb which ranks last in excluding non-beta-barrel proteins in our benchmark but has a very high sensitivity for beta-barrel proteins. All in all, PRED-TMBB2 shows the best balance between sensitivity and specificity, without the need for obtaining a MSA. This allowed our method to scan all 8580 sequences in a couple of hours on our server, using just 1 processor out of the 20, whereas HHomp and BOCTOPUS2 required almost a whole month each in order to run on the same machine using all available processors.

Table 3. Benchmark results on the discrimination datasets

Method	MSA	Sensitivity	Specificity	MCC
PRED-TMBB2	N	91.87	99.14	0.92
BOMP	N	75.22	98.18	0.77
F-W β-Barrel Analyzer	N	97.62	90.97	0.72
PSORTb 3.0	N	59.66	98.89	0.70
TMBETADISC-RBF	N	88.90	92.22	0.69
SOSUIgramN	N	65.11	95.25	0.60
PRED-TMBB (v1)	N	69.38	92.27	0.56
TMBHunt	N	76.11	89.54	0.55
HHomp	Y	97.73	99.95	0.98
BOCTOPUS2	Y	98.12	98.81	0.93
BetAware	Y	67.29	99.87	0.80
BOMP-MSA	Y	78.20	98.18	0.79
SSEA-OMP	Y	96.04	88.57	0.66
PROFtmb	Y	98.12	84.97	0.62

PRED-TMBB2 results are reported based on a cross-validation test. For MSA-based methods, four rounds of PSI-BLAST and an E-value of 10e−3 were used. Other parameters were set to default. For BOCTOPUS2, as stated in the respective publication, a protein is considered a beta-barrel if at least three beta strands are predicted. HHomp achieves the highest performance, while PRED-TMBB2 performs equally good as BOCTOPUS2, however it is orders of magnitude faster than both of them.

It is important to state here that the ‘blind’ (cross-validated) test presented on the discrimination table for PRED-TMBB2 will rarely be used due to the regular Pfam and OMPdb updates. This way, if the sequence at hand has a significant hit in one of the collection of characteristic beta-barrel models, it is automatically assigned as being a beta-barrel protein (thus the actual sensitivity would be actually 100%).

For testing the newly developed method in the presence of experimentally derived information, we also used several outer membrane proteins with reliable experimental information derived from the literature as test cases. We decided to focus on proteins that do not have a determined 3D-structure and belong to a family with no known structure. Such proteins include the MOMP of *C. trachomatis* (Findlay *et al.*, 2005; Yen *et al.*, 2005), HopE of *H. pylori* (Bina *et al.*, 2000) and PorT of *P. gingivalis* (Nguyen *et al.*, 2009). It is of importance to note here that, in such cases, we cannot evaluate the prediction directly, but we can draw useful conclusions (indirectly) from the reliability score. Of course, this would only be the case if PRED-TMBB2 does not predict the topology accurately without the use of experimental information. We have to note here, that the respective publications informed us about the location of specific residues in the sequence with respect to the membrane, i.e. the full topology was not known. Thus, we initially tested whether PRED-TMBB2 and BOCTOPUS2 could assign the correct localization to the residues with experimentally determined topology and afterwards, whether PRED-TMBB2 could obtain more reliable overall predictions by incorporating this information.

The three proteins (PorT, MOMP and HopE), were submitted to a blind prediction using PRED-TMBB2 and subsequently the same was done incorporating the available experimental information and information for signal peptides. As we can observe in Table 4, PRED-TMBB2 and BOCTOPUS2 predict the same number of TM strands for two out of the three proteins. Interestingly, BOCTOPUS2 fails to predict any TM segments in the case of MOMP and classifies it as a non-beta-barrel protein (Table 4). When the experimental information is taken into account, the predicted location of the TM strands by PRED-TMBB2 changes for MOMP and HopE, whereas the reliability is increased for all of them. BOCTOPUS2 predicts the correct localization of the experimentally verified segments only for PorT protein. We need to mention that, prior experimental information could not aid the topology prediction in the case of BOCTOPUS2, since this method cannot perform constrained predictions like PRED-TMBB2.

4 Discussion

We presented a HMM-based method (PRED-TMBB2) that is able to predict the topology of transmembrane beta-barrels and discriminate them from other proteins with improved accuracy. During the development of PRED-TMBB2, we used new model architecture and decoding method, while the training was performed using newly developed algorithms in order to optimize the labels of the training sequences. The non-redundant dataset used for training includes all

Downloaded from https://academic.oup.com/bioinformatics/article/32/17/i665/2450774 by guest on 25 April 2024

Table 4. Results of the blind test in the three selected OMPs

Protein (Uniprot)	PRED-TMBB2 (TM/Correct/Reliability)	PRED-TMBB2^{EXP} (TM/Correct/Reliability)	BOCTOPUS2 (TM/Correct)
PorT (F5HG90)	8/YES/0.836	8/YES/0.889	8/YES
MOMP (Q46409)	12/NO/0.839	12/YES/0.895	0/NO
HopE (Q9ZLD5)	8/NO/0.844	8/YES/0.869	8/NO

We used PRED-TMBB2 in MSA-mode; here, PRED-TMBB2^{EXP} denotes the predictions obtained by incorporating experimental information. For both methods, we report the number of predicted TM strands and whether the location of the experimentally verified segments is correctly predicted or not. For PRED-TMBB2, we additionally report the reliability of the prediction.

recent representative 3D-structures. Finally, the method now not only allows the incorporation of evolutionary information in the form of multiple alignments, but also the prior knowledge of topology of specific regions in the protein sequence, a feature which is unique among the beta-barrel predictors.

The results of the strict, family-wise, cross-validation procedure, showed that PRED-TMBB2 performs significantly better than other available prediction methods. We evaluated all currently available methods for both topology prediction as well as for detection of beta-barrel proteins, where PRED-TMBB2 was compared even against predictors designed specifically for this task. One additional advantage of PRED-TMBB2 is the fact that it operates in single-sequence mode and thus can be used effectively to scan entire proteomes in a reasonable time even with the use of a personal computer, which is practically impossible for methods based on multiple alignments. Finally, we showed that the incorporation of experimental information (which up to now was only possible for alpha-helical TM proteins), can be valuable in newly discovered proteins.

The method, along with the datasets used for training and testing, is freely available for academic users at http://www.compgen.org/tools/PRED-TMBB2. The server can accept either one protein sequence at a time, when using the MSA-based version of the tool, or a batch submission in single-sequence mode. There is also the option of including signal peptide prediction [using the PRED-TAT (Bagos *et al.*, 2010) algorithm] as well as prior screening of the input sequence(s) with the collection of Pfam domains, which are characteristic for beta-barrel and non-beta-barrel regions. Finally, the user can specify the topology for certain parts of the query sequence and perform constrained predictions. We plan to make PRED-TMBB2 source code available for download in the near future.

Acknowledgements

The authors gracefully acknowledge NGP-net COST Action (BM1405) for support and the four anonymous reviewers for valuable comments that improved the quality of the manuscript.

Funding

This study was supported by from the Swedish Research Council (VR-NT 2012-5046 to A.E.) and from SeRC (the Swedish E-science Research Community).

Conflict of Interest: none declared.

References

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Bagos,P.G. *et al.* (2004a) Faster gradient descent conditional maximum likelihood training of hidden Markov models, using individual learning rate adaptation. In Paliouras, G. and Sakakibara, Y. (eds), ***ICGI** 2004, LNAI.* Spinger-Verlag, Athens, pp. 40–52.

Bagos,P.G. *et al.* (2005) Evaluation of methods for predicting the topology of beta-barrel outer membrane proteins and a consensus prediction method. *BMC Bioinformatics*, **6**, 7.

Bagos,P.G. *et al.* (2006) Algorithms for incorporating prior topological information in HMMs: application to transmembrane proteins. *BMC Bioinformatics*, **7**, 189.

Bagos,P.G. *et al.* (2004b) A hidden Markov model method, capable of predicting and discriminating beta-barrel outer membrane proteins. *BMC Bioinformatics*, **5**, 29.

Bagos,P.G. *et al.* (2004c) PRED-TMBB: a web server for predicting the topology of beta-barrel outer membrane proteins. *Nucleic Acids Research*, **32**, W400–W404.

Bagos,P.G. *et al.* (2010) Combined prediction of Tat and Sec signal peptides with hidden Markov models. *Bioinformatics*, **26**, 2811–2817.

Bayrhuber,M. *et al.* (2008) Structure of the human voltage-dependent anion channel. *Proc. Natl. Acad. Sci. USA*, **105**, 15370–15375.

Bernsel,A. and Von Heijne,G. (2005) Improved membrane protein topology prediction by domain assignments. *Protein Sci.*, **14**, 1723–1728.

Berven,F.S. *et al.* (2004) BOMP: a program to predict integral beta-barrel outer membrane proteins encoded within genomes of Gram-negative bacteria. *Nucleic Acids Res.*, **32**, W394–W399.

Bigelow,H.R. *et al.* (2004) Predicting transmembrane beta-barrels in proteomes. *Nucleic Acids Res.*, **32**, 2566–2577.

Bina,J. *et al.* (2000) Functional expression in *Escherichia coli* and membrane topology of porin HopE, a member of a large family of conserved proteins in *Helicobacter pylori. J. Bacteriol.*, **182**, 2370–2375.

Chamberlain,A.K. and Bowie,J.U. (2004) Asymmetric amino acid compositions of transmembrane beta-strands. *Protein Sci.*, **13**, 2270–2274.

Eddy,S.R. (2011) Accelerated Profile HMM Searches. *PLoS Comput. Biol.*, **7**, e1002195.

Faller,M. *et al.* (2004) The structure of a mycobacterial outer-membrane channel. *Science*, **303**, 1189–1192.

Fariselli,P. *et al.* (2005) A new decoding algorithm for hidden Markov models improves the prediction of the topology of all-beta membrane proteins. *BMC Bioinformatics*, **6 Suppl 4**, S12.

Findlay,H.E. *et al.* (2005) Surface expression, single-channel analysis and membrane topology of recombinant *Chlamydia trachomatis* major outer membrane protein. *BMC Microbiol.*, **5**, 5.

Finn,R.D. *et al.* (2016) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.*, **44**, D279–D285.

Forney,G.D. (1973) The Viterbi algorithm. *Proc. IEEE*, **61**, 268–227.

Freeman,T.C Jr. and Wimley,W.C. (2010) A highly accurate statistical approach for the prediction of transmembrane beta-barrels. *Bioinformatics*, **26**, 1965–1974.

Freeman,T.C Jr. and Wimley,W.C. (2012) TMBB-DB: a transmembrane beta-barrel proteome database. *Bioinformatics*, **28**, 2425–2430.

Garrow,A.G. *et al.* (2005a) TMB-Hunt: a web server to screen sequence sets for transmembrane beta-barrel proteins. *Nucleic Acids Res.*, **33**, W188–W192.

Garrow,A.G. *et al.* (2005b) TMB-Hunt: an amino acid composition based method to screen proteomes for beta-barrel transmembrane proteins. *BMC Bioinformatics*, **6**, 56.

Gromiha,M.M. *et al.* (2004) Neural network-based prediction of transmembrane beta-strand segments in outer membrane proteins. *J. Comput. Chem.*, **25**, 762–767.

Hayat,S. and Elofsson,A. (2012) BOCTOPUS: improved topology prediction of transmembrane beta barrel proteins. *Bioinformatics*, **28**, 516–522.

Downloaded from https://academic.oup.com/bioinformatics/article/32/17/i665/2450774 by guest on 25 April 2024

Hayat,S. *et al.* (2016) Inclusion of dyad-repeat pattern improves topology prediction of transmembrane beta-barrel proteins. *Bioinformatics*, **32**, 1571–1573.

Hobohm,U. *et al.* (1992) Selection of representative protein data sets. *Protein Sci.*, **1**, 409–417.

Huang,Y. *et al.* (2010) CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics*, **26**, 680–682.

Imai,K. *et al.* (2008) SOSUI-GramN: high performance prediction for subcellular localization of proteins in gram-negative bacteria. *Bioinformation*, **2**, 417–421.

Jackups,R Jr. and Liang,J. (2005) Interstrand pairing patterns in beta-barrel membrane proteins: the positive-outside rule, aromatic rescue, and strand registration prediction. *J. Mol. Biol.*, **354**, 979–993.

Jacoboni,I. *et al.* (2001) Prediction of the transmembrane regions of beta-barrel membrane proteins with a neural network-based predictor. *Protein Sci.*, **10**, 779–787.

Jayasinghe,S. *et al.* (2001) MPtopo: a database of membrane protein topology. *Protein Sci.*, **10**, 455–458.

Kall,L. *et al.* (2005) An HMM posterior decoder for sequence feature prediction that includes homology information. *Bioinformatics*, **21**, i251–i257.

Koronakis,V. *et al.* (2000) Crystal structure of the bacterial membrane protein TolC central to multidrug efflux and protein export. *Nature*, **405**, 914–919.

Kozma,D. *et al.* (2013) PDBTM: Protein Data Bank of transmembrane proteins after 8 years. *Nucleic Acids Res.*, **41**, D524–D529.

Krogh,A. (1994) Hidden Markov models for labelled sequences. *Proceedings of the12th IAPR International Conference on Pattern Recognition*, pp. 140–144.

Krogh,A. (1997) Two methods for improving performance of an HMM and their application for gene finding. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **5**, 179–186.

Krogh,A. *et al.* (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.*, **305**, 567–580.

Lomize,M.A. *et al.* (2006) OPM: orientations of proteins in membranes database. *Bioinformatics*, **22**, 623–625.

Magrane,M. and Consortium,U. (2011) UniProt Knowledgebase: a hub of integrated protein data. *Database*, **2011**, bar009.

Martelli,P.L. *et al.* (2002) A sequence-profile-based HMM for predicting and discriminating beta barrel membrane proteins. *Bioinformatics*, **18 Suppl 1**, S46–S53.

Matthews,B.W. (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta*, **405**, 442–451.

Melen,K. *et al.* (2003) Reliability measures for membrane protein topology prediction algorithms. *J. Mol. Biol.*, **327**, 735–744.

Nguyen,K.A. *et al.* (2009) Verification of a topology model of PorT as an integral outer-membrane protein in *Porphyromonas gingivalis*. *Microbiology*, **155**, 328–337.

Ou,Y.Y. *et al.* (2010) Prediction of membrane spanning segments and topology in beta-barrel membrane proteins at better accuracy. *J. Comput. Chem.*, **31**, 217–223.

Ou,Y.Y. *et al.* (2008) TMBETADISC-RBF: discrimination of beta-barrel membrane proteins using RBF networks and PSSM profiles. *Comput. Biol. Chem.*, **32**, 227–231.

Remmert,M. *et al.* (2009) HHomp–prediction and classification of outer membrane proteins. *Nucleic Acids Res.*, **37**, W446–W451.

Rose,P.W. *et al.* (2013) The RCSB Protein Data Bank: new resources for research and education. *Nucleic Acids Res.*, **41**, D475–D482.

Saier,M.H., Jr. *et al.* (2006) TCDB: the Transporter Classification Database for membrane transport protein analyses and information. *Nucleic Acids Res.*, **34**, D181–D186.

Savojardo,C. *et al.* (2013) BETAWARE: a machine-learning tool to detect and predict transmembrane beta-barrel proteins in prokaryotes. *Bioinformatics*, **29**, 504–505.

Schulz,G.E. (2003) Transmembrane beta-barrel proteins. *Adv. Protein Chem.*, **63**, 47–70.

Slusky,J.S. and Dunbrack,R.L. Jr. (2013) Charge asymmetry in the proteins of the outer membrane. *Bioinformatics*, **29**, 2122–2128.

Song,L. *et al.* (1996) Structure of Staphylococcal alpha-hemolysin, a heptameric transmembrane pore. *Science*, **274**, 1859–1865.

Tsirigos,K.D. *et al.* (2011) OMPdb: a database of {beta}-barrel outer membrane proteins from Gram-negative bacteria. *Nucleic Acids Res.*, **39**, D324–D331.

Tsirigos,K.D. *et al.* (2015) The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides. *Nucleic Acids Res.*, **43**, W401–W407.

Tusnady,G.E. *et al.* (2008) TOPDB: topology data bank of transmembrane proteins. *Nucleic Acids Res.*, **36**, D234–D239.

Tusnady,G.E. and Simon,I. (2001) The HMMTOP transmembrane topology prediction server. *Bioinformatics*, **17**, 849–850.

Viklund,H. and Elofsson,A. (2004) Best alpha-helical transmembrane protein topology predictions are achieved using hidden Markov models and evolutionary information. *Protein Sci.*, **13**, 1908–1917.

Wimley,W.C. (2002) Toward genomic identification of beta-barrel membrane proteins: composition and architecture of known structures. *Protein Sci.*, **11**, 301–312.

Yan,R.X. *et al.* (2011) Outer membrane proteins can be simply identified using secondary structure element alignment. *BMC Bioinformatics*, **12**, 76.

Yen,T.Y. *et al.* (2005) Characterization of the disulfide bonds and free cysteine residues of the *Chlamydia trachomatis* mouse pneumonitis major outer membrane protein. *Biochemistry*, **44**, 6250–6256.

Yu,N.Y. *et al.* (2011) PSORTdb—an expanded, auto-updated, user-friendly protein subcellular localization database for Bacteria and Archaea. *Nucleic Acids Res.*, **39**, D241–D244.

Yu,N.Y. *et al.* (2010) PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics*, **26**, 1608–1615.

Zemla,A. *et al.* (1999) A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment. *Proteins*, **34**, 220–223.

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

Downloaded from https://academic.oup.com/bioinformatics/article/32/17/i665/2450774 by guest on 25 April 2024