BetAware-Deep: An Accurate Web Server for Discrimination and Topology Prediction of Prokaryotic Transmembrane β-barrel Proteins

Giovanni Madeo [a,2], Castrense Savojardo [a,2], Pier Luigi Martelli [a*] and Rita Casadio [a,b]

a - ***Biocomputing Group,*** *Department of Pharmacy and Biotechnology, University of Bologna, Italy*

b - ***Institute of Biomembranes,*** *Bioenergetics and Molecular Biotechnologies, National Research Council (CNR), Bari, Italy*

Correspondence to Pier Luigi Martelli: *pierluigi.martelli@unibo.it (P.L. Martelli)*

https://doi.org/10.1016/j.jmb.2020.166729
Edited by Michael Sternberg

Abstract

TransMembrane β-Barrel (TMBB) proteins located in the outer membranes of Gram-negative bacteria are crucial for many important biological processes and primary candidates as drug targets. Structure determination of TMBB proteins is challenging and hence computational methods devised for the analysis of TMBB proteins are important for complementing experimental approaches. Here, we present a novel web server called BetAware-Deep that is able to accurately identify the topology of TMBB proteins (i.e. the number and orientation of membrane-spanning segments along the protein sequence) and to discriminate them from other protein types. The method in BetAware-Deep defines new features by exploiting a non-canonical computation of the hydrophobic moment and by adopting sequence-profile weighting of the White&Wimley hydrophobicity scale. These features are processed using a two-step approach based on deep learning and probabilistic graphical models. BetAware-Deep has been trained on a dataset comprising 58 TMBBs and benchmarked on a novel set of 15 TMBB proteins. Results showed that BetAware-Deep outperforms two recently released state-of-the-art methods for topology prediction, predicting correct topologies of 10 out of 15 proteins. TMBB detection was also assessed on a larger dataset comprising 1009 TMBB proteins and 7571 non-TMBB proteins. Even in this benchmark, BetAware-Deep scored at the level of top-performing methods. A web server has been developed allowing users to analyze input protein sequences and providing topology prediction together with a rich set of information including a graphical representation of the residue-level annotations and prediction probabilities. BetAware-Deep is available at https://busca.biocomp.unibo.it/betaware2.

Introduction

Transmembrane β-barrel (TMBB) proteins are integral membrane proteins composed by β-strands spanning the membrane phase and forming a structural motif that resembles a barrel.[1] TMBBs are found in outer membranes of Gram-negative bacteria, mitochondria and chloroplasts.

In this work, the attention is focalized on Gram-negative TMBB proteins, which have peculiar structural characteristics.[2] Firstly, they include an even number (between 4 and 36 per chain) of transmembrane β-strands, with the N- and C-terminus always localized in the periplasmic space. Secondly, all β-strands are antiparallel and adjacent in the sequence (all-next-neighbors). Thirdly, the connections between β-strands at the periplasmic side are typically short turns, while in the external side they include long loops. Finally, residues in membrane-crossing β-strands show the alternation of lipid- and pore-facing residues, in a dyad-repeat pattern.[3]

TMBB proteins perform a wide range of functions,[4] including membrane anchoring, peptidase activity, cell adhesion, lipase activity, autotransporter, signal transduction, general and specific diffusion of molecules and ions (porins), efflux pumps. According to the wide range of functions performed by TMBB proteins, they are encoded by as many as 2–3% of all genes in most Gram-negative bacteria,[3] which confirms their relevance in these organisms. These features make TMBB proteins attractive for drug discovery, research on drugs (such as antibiotics) and vaccines. However, despite their relevance and the medical interest, TMBB proteins are largely underrepresented in the Protein Data Bank (PDB).[5] This is mainly due to technical difficulties in the expression and the crystallization process. Thus, it is of crucial importance to have reliable computational methods able to discriminate TMBB proteins in large datasets of proteins, such as whole genomes (discrimination task), and to predict the organization of the protein in the membrane space (protein topology i.e. the number and location of membrane-spanning segments along the sequence and their orientation with respect to the membrane).

Methods available in literature adopt different types of machine-learning approaches (e.g. support vector machines, neural networks, hidden Markov models and conditional random fields) to tackle the topology prediction and the discrimination tasks.[6–16] Tools are typically implemented and released either as web servers or standalone executables.

Here we present BetAware-Deep, a new web server for both TMBB detection and topology prediction. The method underlying BetAware-Deep consists of a two-step procedure based on deep learning (Long Short-Term Memory, LSTM) followed by the application of probabilistic graphical models for sequence labelling. Moreover, the method includes the computation of a profile-weighted hydrophobic moment, a feature designed to effectively capture the dyad-repeat pattern, which has been proven to be helpful to improve TMBB topology prediction.[15] Benchmarks were performed to assess BetAware-Deep performance with respect to other state-of-the-art methods using a new blind dataset specifically designed for this purpose. In this stringent benchmark, our web server reported the best performance in topology prediction, outperforming recent approaches designed for the same task. A comparative benchmark has been also performed to assess the ability to discriminate TMBB from non-TMBB proteins. Even in this case, BetAware-Deep achieved performances that are comparable to other methods at the state-of-the-art.

The BetAware-Deep web server provides a user-friendly interface and allows performing TMBB detection and topology prediction of a user-submitted query protein. Prediction results include many different types of information and the possibility of browsing predicted topology through an interactive feature viewer. The BetAware-Deep web server it is freely accessible to the scientific community at https://busca.biocomp.unibo.it/betaware2.

Materials and methods

Datasets

The dataset used for training BetAware-Deep includes two parts: a positive dataset, comprising bacterial TMBB proteins, and a negative one, comprising bacterial all-beta, non-TMBB proteins used to improve the predictor discrimination capabilities.

The positive dataset, comprising 58 TMBB proteins, was derived from data deposited at the MPstruc database (https://blanco.biomol.uci.edu/mpstruc/). Redundancy among these proteins was reduced allowing at the most 25% sequence identity on more than 90% coverage on both sequences. One representative was then selected for each cluster. Then, the resulting 58 proteins were grouped in 10 subsets for cross-validation. To avoid any residual redundancy between training/validation/testing sets, sequences were further clustered using a threshold of 25% sequence identity at 50% coverage and requiring proteins clustered together to be in the same subset.

The negative dataset was obtained from the set of bacterial all-beta proteins, as annotated in SCOPe.[1] A final set of 69 proteins was obtained after reducing internal redundancy at 25% sequence identity on a 50% coverage.

To test the method performances in topology prediction, and to compare them with those obtained with other state-of-the-art methods, a blind test set was also compiled. This dataset comprises 15 TMBB proteins also extracted from the MPstruc database. Again, internal redundancy was reduced to 25% sequence identity on a 50% coverage. Similarly, redundancy was reduced (using the same thresholds) with respect to BetAware-Deep training set as well as with respect to the datasets used for training the two most recent approaches available in literature, BOCTOPUS2[15] and PRED-TMBB2.[16] This allowed us to build an unbiased independent dataset on which we could compare the different approaches. We did not consider a splitting based on structural or distant homology classification because beta-barrel membrane proteins are separated in few classes that include proteins with very different topologies and functions.

Sequences in both training and testing sets were downloaded from UniProt[17] rather than PDB, to perform training and predictions on the complete protein sequences. For each protein, a multiple sequence alignment (MSA) was built running PSI-BLAST[18] for 2 iterations against the UniRef90 database[19] (release 2018_03). A sequence profile representing the frequency of each residue type at each MSA column was then computed from the PSI-BLAST output.

Residue-level labeling was performed according to data available at the Orientations of Proteins in Membranes (OPM) database,[20] which provided information about the localization of residues with respect to the membrane plane. Moreover, in order to characterize complete beta-strands, we also computed secondary structures using DSSP[21] and combined strand assignments with OPM-annotated TM strands.

According to our procedure, five different labels can be assigned to each residue:

- n, non-barrel region, namely the portion of the protein which precedes or succeeds the TMBB beta-barrel;
- i, inner or periplasmic region;
- o, outer or extracellular region;
- T, trans-membrane region, namely residues embedded in membrane as indicated by OPM;
- E, portion of the beta-strands not classified as transmembrane by OPM but contiguous to the OPM-annotated segment.

The explicit introduction of the extended strand class allows to better model the uncertainty around TM-segment edges using DSSP annotations. Residues of proteins belonging to the negative set are all labelled with the n label (see Supplementary Materials Figure S1 for a graphical representation of the labelling strategy adopted).

Finally, to assess performance in the discrimination task, we adopted a large dataset including 1009 TMBB proteins and 7571 non-TMBB proteins, previously introduced to evaluate performance of Pred-TMBB2[16] in the same task.

All datasets can be downloaded from the BetAware-Deep web server at: https://busca.biocomp.unibo.it/betaware2/datasets.

Overview of BetAware-Deep approach

BetAware-Deep implements a new approach for tackling both TMBB detection and topology prediction tasks. Compared to our previous release (BetAware, based on extreme learning machines and conditional random fields[14]), BetAware-Deep introduces several different improvements. These include: (i) the application, for the first time in this field, of a deep recurrent network to scan input sequences; (ii) the definition of a new feature based on profile-weighted hydrophobic moment to capture the typical TMBB dyad-repeat pattern; (iii) the adoption of a extended labelling approach which takes into consideration non-barrel regions as well as ambiguity around borders of transmembrane β-strand segments; (iv) two brand-new and updated datasets for training and independently testing the method.

The approach implemented in BetAware-Deep consists of two cascading prediction steps (see Supplementary Figure S3 for a graphical representation of BetAware-Deep workflow). In the first step, a deep learning architecture is implemented to predict the probability for each residue of the query protein sequence of being localized into one of the five possible compartments: non-barrel region (n), periplasmic side (i), extracellular side (o), transmembrane beta-strand segment (T) and extended transmembrane beta strand (E). In the second step, these probabilities are processed in order to predict the final protein topology using a probabilistic sequence labelling approach (a detailed description is in the following and for additional details refer to Supplementary Material).

One of the main novelties introduced in BetAware-Deep is the adoption of a sequence profile-weighted hydrophobic moment capturing the dyad-repeat pattern. The hydrophobic moment[22] measures the alternance between hydrophobic and hydrophilic residues in a short protein segment, given a specific angle separating sidechains along the backbone. Fixing this angle to 180° and considering a segment of 5 residues, the hydrophobic moment is computed according to Eq. (1):

$$\gamma = \left| \sum_{n=1}^{5} \mathrm{H}[R_n](-1)^n \right| \qquad (1)$$

where $\mathrm{H}[R_n]$ is the hydrophobicity score for residue R in position n observed in the protein sequence. Here, we adopted the White&Wimley hydrophobic scale for the transfer of unfolded chains into octanol.[23]

This canonical definition of the hydrophobic moment (here, referred to as Unweighted Hydrophobic Moment [UHM]) is extended to include evolutionary information contained in a sequence profile: instead of using the hydrophobicity score for each residue in the primary sequence, all residues observed in a specific column of a multiple sequence alignment are taken into consideration. In mathematical terms, this means applying a simple weighting scheme as follows:

$$\gamma = \left| \sum\nolimits_{n=1}^{5} \sum_{R \in \{A,C,D,\ldots,Y\}} P[R_n]\mathrm{H}[R_n](-1)^n \right| \qquad (2)$$

where the inner summation is taken over all possible residues R, and $\mathrm{H}[R_n]$ and $P[R_n]$ are, respectively, the hydrophobicity score and the frequency for residue R in position n given the sequence profile. Here, this non-canonical computation of the hydrophobic moment is referred to as Profile-Weighted Hydrophobic Moment (PWHM). Since it is likely that the moment value is higher at the center of a transmembrane segment and lower at the edges, we assign to each position the maximum moment value in the 3-residue window centered on the position. This choice allows smoothing the hydrophobic moment signal along the sequence, possibly increasing the value at the TM segment edges.

The hydrophobic moment calculated as described above is then concatenated to the sequence profile itself (ending up with a 21-dimensional encoding of each residue) and it is given as input to the first step method. This step is implemented using a Bidirectional Long Short-Term Memory (BLSTM) model.[24] The Long Short-Term Memory (LSTM)[25] is a type of recurrent neural network well-suited for analysing sequential data. What distinguishes a LSTM from other types of recurrent networks is the ability to better handle vanishing gradient issues. Indeed, thanks to the special gated architecture, the LSTM learns to neglect sequence positions not relevant for the problem at hand, allowing the gradient to flow unchanged along these positions.[25] BLSTM represents a further improvement of this model. It belongs to the family of bidirectional RNNs,[26] whose basic idea is to duplicate the recurrent layer, the first scanning the input sequence left-to-right and right-to-left, respectively. This model has been successfully applied for the first time in speech recognition.[24] The specific BLSTM architecture adopted in this work is described in detail in Supplementary Materials and graphically shown in Figure S4.

As mentioned above, the output provided by the BLSTM model for each position in the input sequence is a set of five Per-Residue Probabilities (PRPs) of being localized into one of the possible compartments relative to the membrane (n,i,o,T and E). PRPs are then positionally combined to the sequence profile (leading to a 25-dimensional encoding of each residue) and passed to the second BetAware-Deep predictive step. At this stage, we apply Grammatical-Restrained Hidden Conditional Random Fields (GRHCRFs), a discriminative probabilistic model already used in the first version of BetAware and fully described in general terms by Fariselli and co-workers.[27] The GRHCRF model, depicted as a finite-state automaton in Figure S5, allows to predict the protein topology in agreement with a regular grammar defined over structural constraints known for TMBB proteins. The prediction phase in the GRHCRF model consists in the identification of the most probable path along the model given the input sequence. This is achieved using the Posterior-Viterbi dynamic-programming algorithm.[27]

As a final step, the five-class topology (labels: n,i, o,T and E) predicted by the GRHCRF model is mapped to a canonical three-class scheme (labels: i, o and T) by mapping label n to i and label E to T.

Discrimination of TMBB from non-TMBB proteins in BetAware-Deep is based on the number of predicted transmembrane segments. Specifically, BetAware-Deep predicts a protein as a TMBB if the number of TM segments is at least 4. This choice builds on the fact that the number of segments in known Gram-negative bacteria TMBBs ranges from 4 to 36.

Model selection and evaluation procedure

Optimal input encodings, parameters and hyperparameters of the two machine-learning methods included in BetAware-Deep were estimated using a cross-validation procedure. Each run of cross validation was carried out using eight subsets for training, one for validation and one for testing. BLSTM training procedure was stopped when the validation loss ceased to decrease for at least 20 epochs. Analogously, the validation set was used to establish the number of iterations used to train the GRHCRFs. Once optimized over the validation set, prediction was performed over the testing set.

Blind test predictions were carried out training both machine-learning steps using the complete training set and using a small fraction of it for validation (specifically used for early stopping of the BLSTM model).

Topology prediction has been evaluated using the following scoring measures:

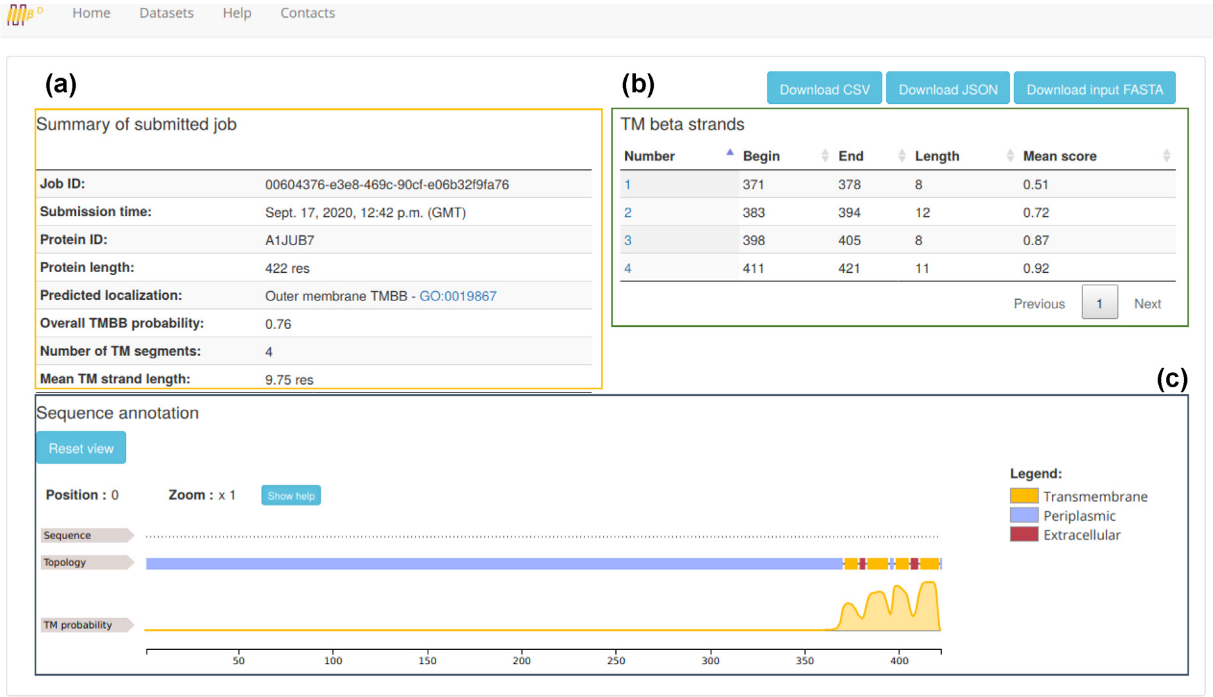
- three-class accuracy (Q_3);
- the Segment OVerlap measure (SOV), evaluating how well predicted segments in the three classes cover observed segments;[28]
- the Protein OVerlap measure (POV), namely the number of proteins with completely correct predicted topology. Here, a topology prediction is considered correct if the number of TM segments is correct and the overlap between observed and corresponding predicted segments is above a given threshold;
- N_{TM}, defined as the number of proteins with the correct number of predicted TM segments.

Discrimination of TMBB from non-TMBB proteins was assessed using standard binary classification measures, including: Matthew's Correlation Coefficient (MCC), Specificity (Spe) and Sensitivity (Sen).

The BetAware-Deep web server

BetAware-Deep is released as a public web server accessible at https://busca.biocomp.unibo.it/betaware2. The server home page provides a very simple interface allowing the user to either paste a protein sequence in FASTA format or to upload an external FASTA file. The server accepts a single protein sequence per job.

After submission, the user is automatically redirected to the page where results will appear. After job completion, the server output is shown. In Figure 1 the BetAware-Deep result page is



Summary of submitted job

Job ID:	00604376-e3e8-469c-90cf-e06b32f9fa76
Submission time:	Sept. 17, 2020, 12:42 p.m. (GMT)
Protein ID:	A1JUB7
Protein length:	422 res
Predicted localization:	Outer membrane TMBB - GO:0019867
Overall TMBB probability:	0.76
Number of TM segments:	4
Mean TM strand length:	9.75 res

TM beta strands

Number	Begin	End	Length	Mean score
1	371	378	8	0.51
2	383	394	12	0.72
3	398	405	8	0.87
4	411	421	11	0.92

Previous 1 Next

Figure 1. The BetAware-Deep result page. **(a)** This panel reports summary information about the submitted job and aggregated prediction results including predicted protein localization, overall TMBB probability, number of TM segments and their average length. **(b)** This panel reports detailed information about predicted TM β-strands, including begin and end positions, length and average predicted probability scores. **(c)** This panel reports detailed topology annotation using an interactive feature viewer. Users can zoom-in and visualize areas of interest along the protein sequence. All results can be downloaded in JSON and CSV formats.

shown for the input protein Adhesin YadA from Yersinia enterocolitica (UniProt: A1JUB7).

BetAware-Deep output page is divided into three parts. On the panel placed on the top-left corner (panel (a) in Figure 1), general information about the submitted job is reported, including the unique **Job ID** internally assigned by BetAware-Deep, **submission time**, protein **accession** and **length** as extracted from the input FASTA. In the same table, summary prediction results are also reported, including BetAware-Deep **predicted localization** (**Outer membrane TMBB** when the number of predicted TM segments predicted by BetAware-Deep is greater or equal to 4, **Other, non-TMBB** otherwise) and corresponding overall **prediction probability**. This probability is internally computed by BetAware-Deep as the average probability assigned to each predicted membrane-spanning residue by the GRHCRF model (see Supplementary Materials for details). The **number of predicted TM segments** and their **average length** is reported.

On the panel on the top-right corner (panel (b) in Figure 1), the **list of predicted TM segments** is shown along with details about individual strands, including **begin and end positions** in the sequence, **segment length** and **mean TM-strand probability score**.

On the bottom of the page (panel (c) in Figure 1), **residue-level topology annotation** is reported by means of an interactive feature viewer. In particular, the protein primary sequence is shown along with two annotation tracks: the **protein topology track**, reporting the alternation of periplasmic, transmembrane and extracellular segments along the sequence and the **TM probability track**, reporting per-residue transmembrane probabilities. The user can highlight and zoom-in specific regions along the sequence by selecting an area of interest. Moreover, clicking on the strand number reported in the table placed top-right, the visualization is automatically centered on the selected strand.

Job results can be downloaded in two different formats: a **JSON** file, reporting the complete job result, and a **CSV** file providing residue-level annotation of topology (including per-residue probabilities).

Results

Predictive performance of BetAware-Deep

BetAware-Deep was tested by adopting a 10-fold cross-validation procedure over the positive training set, which includes 58 TMBB proteins, in order to

Table 1 Cross-validation results over the positive training set (58 proteins) obtained with different BLSTM input encodings.

BLSTM input encoding	Q3	SOV	POV	N_{TM}
PROFILE	83%	91%	35	39
PROFILE + UHM	81%	92%	37	40
PROFILE + PWHM	88%	95%	40	46

All models tested use a sequence profile as input. For models **PROFILE + UHM** and **PROFILE + PWHM**, sequence profile has been combined with an unweighted and profile-weighted hydrophobic moment, respectively (see Methods for details). **Q3**: three-class accuracy. **SOV**: Segment Overlap. **POV**: number of correctly predicted topologies. **N_{TM}**: proteins with correct number of predicted transmembrane segments.

compare performances obtained with different input encodings provided to the first-step BLSTM (Table 1). Specifically, we assessed and compared the following models:

- a baseline model, exploiting only the sequence profile (labelled as **PROFILE** in Table 1);
- a model trained using sequence profile combined with unweighted hydrophobic moment (**PROFILE + UHM** in Table 1);
- the final model using sequence profiles and profile-weighted hydrophobic moments (**PROFILE + PWHM** in Table 1).

Both models incorporating the hydrophobic moment feature outperform the baseline model (reaching 35/58 of POV and 39/58 of N_{TM}). However, the highest scores are obtained when the PWHM is included. In fact, this input improves performances up to 40 correct topologies and N_{TM} to 46. Moreover, it reports the highest SOV (95%) and accuracy (88%). These results show that the PWHM feature has the best discrimination ability, as also highlighted by ROC curves obtained using weighted and unweighted moments as direct predictors for TM residues (see Supplementary Materials, Figure S2). Given these results we adopted the model trained on profile and PWHM as the final one.

Table 2 reports a comparative analysis of the performance in topology prediction and discrimination obtained with different available methods. Beside BetAware-Deep, tested methods include BOCTOPUS2,[15] PRED-TMBB2,[16] the previous version of our BetAware[14] and HHomp[10] (for discrimination only). For topology prediction, methods were compared on the blind test set defined in this work and comprising 15 TMBB proteins. For discrimination, a larger dataset from a previous study[16] comprising 8580 proteins (1009 out of which are TMBBs) was adopted.

Comparative results confirm BetAware-Deep ability in correctly predicting protein TM topology. Indeed, BetAware-Deep achieves 10/15 on both POV and N_{TM} as well as high values for SOV (94%) and accuracy (80%). The improvement with respect to the previous version of the method (BetAware) is substantial (6 and 5 proteins in POV and N_{TM}, respectively, and significantly higher SOV and accuracy scores).

These results show that BetAware-Deep is the best-performing method for topology prediction also when compared with recently developed approaches available in literature. Indeed, BetAware-Deep outperforms all other tools assessed for topology prediction in all reported indexes, with the only exception of N_{TM} for which PRED-TMBB2 reports a slightly higher value. The comparative benchmark performed here is somewhat hampered by the limited number of TMBB proteins available. As a consequence of this, only few proteins can be used to effectively and unbiasedly compare the different methods. Moreover, some of the methods are not able to handle specific classes of TMBB (e.g. multimeric ones), which can also lead to underestimation of their performances. In any case, our results highlights that BetAware-Deep well-compares with other tools at the state-of-the-art in topology prediction.

Finally, discrimination results show that BetAware-Deep performs at the level of other state-of-the-art methods, having an MCC of 0.91, slightly lower but comparable to those achieved by

Table 2 Comparative benchmark of different methods in topology prediction and discrimination.

Method	Topology Prediction				Discrimination		
	Q3	SOV	POV	N_{TM}	Sen	Spec	MCC
BetAware-Deep	80%	94%	10	10	98.12%	97.53%	0.91
BOCTOPUS2	65%	68%	8	8	98.12%	98.81%	0.93
PRED-TMBB2	71%	80%	6	11	91.87%	99.14%	0.92
BetAware	60%	55%	4	5	67.29%	99.87%	0.8
HHomp	–	–	–	–	97.73%	99.95%	0.98

Results for topology prediction were generated using a blind test comprising 15 TMBB proteins. For discrimination, a test set taken from a previous study[16] was adopted. For topology prediction, **Q3**: three-state accuracy; **SOV**: Segment OVerlap; **POV**: Protein OVerlap, number of proteins with correctly predicted topology; **N_{TM}**: number of proteins with correct number of predicted transmembrane segments. For discrimination, **Sen**: sensitivity, portion of correctly predicted positive examples; **Spec**: specificity, portion of correctly predicted negative examples; **MCC**: Matthew's Correlation Coefficient. Discrimination results for all methods but BetAware-Deep were taken from 16.

PRED-TMBB2 and BOCTOPUS2 (0.92 and 0.93, respectively). The only method outperforming all others is HHomp. This method is based on a precomputed database of profile HMMs of putative TMBBs. For a new sequence in input, the method builds a profile HMM and then compare it with the database. Hence, HHomp is limited to the discrimination of TMBBs belonging to previously discovered families. This strategy is radically different from the ones pursued by methods such as BetAware-Deep, BOCTOPUS2 or PRED-TMBB2, which are instead grounded on pure machine learning-based predictive approaches.

Conclusions

In this paper we present a web server implementing BetAware-Deep, a new method based on deep-learning approaches for discrimination and topology prediction of prokaryotic transmembrane beta-barrel proteins form sequence. BetAware-Deep takes advantage of evolutionary information and a profile-weighted computation of the hydrophobic moment to capture the distinctive dyad-repeat pattern of TM beta-barrel proteins. When compared with other state-of-the-art approaches on a non-redundant independent dataset our method achieved the best performance in the topology prediction task. In discrimination, our BetAware-Deep performance are comparable to those reported by other approaches.

As other methods available,[15,16] BetAware-Deep is well-suited for analyzing TMBB proteins from Gram-negative bacterial species. This class of proteins does not include other types of TMBB e.g. eukaryotic TMBB (like the Voltage-dependent anion channel) or beta-sheet pore-forming toxins, which have non-canonical topologies.

The BetAware-Deep web server (https://busca.biocomp.unibo.it/betaware2) is freely available for the scientific community. The user interface has been designed having in mind usability and accessibility. The output provided includes all relevant information and can be easily exported to standard interoperable formats like JSON and CSV.

CRediT authorship contribution statement

Giovanni Madeo: Data curation, Methodology, Software, Formal analysis. **Castrense Savojardo:** Conceptualization, Methodology, Software, Formal analysis, Validation, Writing - original draft, Writing - review & editing. **Pier Luigi Martelli:** Conceptualization, Validation, Methodology, Writing - review & editing, Resources, Supervision. **Rita Casadio:** Conceptualization, Writing - review & editing, Resources.

Acknowledgements

CS acknowledges the support by the Italian Ministry of University and Research (MUR), PRIN 2017 project 2017483NH8_002.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jmb.2020.166729.

Received 30 September 2020;
Accepted 30 November 2020;

***Keywords*:**
transmembrane beta-barrel proteins;
protein topology prediction;
discrimination of beta-barrels;
deep learning;
conditional random fields

[2] Equally contributed to this work.

References

1. Fox, N.K., Brenner, S.E., Chandonia, J.-M., (2014). SCOPe: Structural classification of proteins—extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Res.*, **42**, D304–D309.
2. Schulz, G.E., (2000). β-Barrel membrane proteins. *Curr. Opin. Struct. Biol.*, **10**, 443–447.
3. Wimley, W.C., (2003). The versatile β-barrel membrane protein. *Curr. Opin. Struct. Biol.*, **13**, 404–411.
4. Galdiero, S., Galdiero, M., Pedone, C., (2007). β-Barrel Membrane Bacterial Proteins: Structure, Function, Assembly and Interaction with Lipids. *Curr. Protein Pept. Sci.*, **8**, 63–82.
5. Berman, H.M., (2000). The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.
6. Bigelow, H.R., Petrey, D.S., Liu, J., Przybylski, D., Rost, B., (2004). Predicting transmembrane beta-barrels in proteomes. *Nucleic Acids Res.*, **32**, 2566–2577.
7. Casadio, R., Fariselli, P., Finocchiaro, G., Martelli, P.L., (2003). Fishing new proteins in the twilight zone of genomes: The test case of outer membrane proteins in Escherichia coli K12, Escherichia coli O157:H7, and other Gram-negative bacteria. *Protein Sci. Publ. Protein Soc.*, **12**, 1158–1168.
8. Freeman, T.C., Wimley, W.C., (2010). A highly accurate statistical approach for the prediction of transmembrane β-barrels. *Bioinformatics*, **26**, 1965–1974.
9. Hayat, S., Elofsson, A., (2012). BOCTOPUS: improved topology prediction of transmembrane β barrel proteins. *Bioinformatics*, **28**, 516–522.
10. Remmert, M., Linke, D., Lupas, A.N., Söding, J., (2009). HHomp—prediction and classification of outer membrane proteins. *Nucleic Acids Res.*, **37** (suppl_2), W446–W451.

11. Savojardo, C., Fariselli, P., Casadio, R., (2011). Improving the detection of transmembrane -barrel chains with N-to-1 extreme learning machines. *Bioinformatics*, **27**, 3123–3128.
12. Bagos, P.G., Liakopoulos, T.D., Spyropoulos, I.C., Hamodrakas, S.J., (2004). PRED-TMBB: a web server for predicting the topology of β-barrel outer membrane proteins. *Nucleic Acids Res.*, **32** (Web Server issue), W400–W404.
13. Martelli, P.L., Fariselli, P., Krogh, A., Casadio, R., (2002). A sequence-profile-based HMM for predicting and discriminating β barrel membrane proteins. *Bioinformatics*, **18** (suppl_1), S46–S53.
14. Savojardo, C., Fariselli, P., Casadio, R., (2013). BETAWARE: a machine-learning tool to detect and predict transmembrane beta-barrel proteins in prokaryotes. *Bioinformatics*, **29**, 504–505.
15. Hayat, S., Peters, C., Shu, N., Tsirigos, K.D., Elofsson, A., (2016). Inclusion of dyad-repeat pattern improves topology prediction of transmembrane β-barrel proteins. *Bioinformatics*, **32**, 1571–1573.
16. Tsirigos, K.D., Elofsson, A., Bagos, P.G., (2016). PRED-TMBB2: improved topology prediction and detection of beta-barrel outer membrane proteins. *Bioinformatics*, **32**, i665–i671.
17. UniProt Consortium, (2019). UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.*, **47**, D506–D515.
18. Altschul, S., (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
19. Suzek, B.E., Wang, Y., Huang, H., McGarvey, P.B., Wu, C. H., (2015). UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, **31**, 926–932.
20. Lomize, M.A., Pogozheva, I.D., Joo, H., Mosberg, H.I., Lomize, A.L., (2012). OPM database and PPM web server: resources for positioning of proteins in membranes. *Nucleic Acids Res.*, **40**, D370–D376.
21. Kabsch, W., Sander, C., (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
22. Eisenberg, D., Weiss, R.M., Terwilliger, T.C., (1984). The hydrophobic moment detects periodicity in protein hydrophobicity. *Proc. Natl. Acad. Sci.*, **81**, 140–144.
23. White, S.H., Wimley, W.C., (1998). Hydrophobic interactions of peptides with membrane interfaces. *Biochim. Biophys. Acta BBA – Rev. Biomembr.*, **1376**, 339–352.
24. Graves, A., Schmidhuber, J., (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Netw.*, **18**, 602–610.
25. Hochreiter, S., Schmidhuber, J., (1997). Long short-term memory. *Neural Comput.*, **9**, 1735–1780.
26. Schuster, M., Paliwal, K.K., (1997). Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.*, **45**, 2673–2681.
27. Fariselli, P., Savojardo, C., Martelli, P.L., Casadio, R., (2009). Grammatical-restrained hidden conditional random fields for bioinformatics applications. *Algorithms Mol. Biol.*, **4**, 13.
28. Zemla, A., Venclovas, Č., Fidelis, K., Rost, B., (1999). A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment. *Proteins Struct. Funct. Bioinf.*, **34**, 220–223.