BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 4 2013, pages 504–505
doi:10.1093/bioinformatics/bts728

Structural bioinformatics

Advance Access publication January 6, 2013

BETAWARE: a machine-learning tool to detect and predict transmembrane beta-barrel proteins in prokaryotes

Castrense Savojardo[1,2], Piero Fariselli[1,2,*] and Rita Casadio[1]

[1]Biocomputing Group, CIRI-Health Science and Technology/Department of Biology, University of Bologna, 40126 Bologna and [2]Department of Computer Science and Engineering, University of Bologna, 40127 Bologna, Italy

Associate Editor: Anna Tramontano

ABSTRACT

Summary: The annotation of membrane proteins in proteomes is an important problem of Computational Biology, especially after the development of high-throughput techniques that allow fast and efficient genome sequencing. Among membrane proteins, transmembrane β-barrels (TMBBs) are poorly represented in the database of protein structures (PDB) and difficult to identify with experimental approaches. They are, however, extremely important, playing key roles in several cell functions and bacterial pathogenicity. TMBBs are included in the lipid bilayer with a β-barrel structure and are presently found in the outer membranes of Gram-negative bacteria, mitochondria and chloroplasts. Recently, we developed two top-performing methods based on machine-learning approaches to tackle both the detection of TMBBs in sets of proteins and the prediction of their topology. Here, we present our BETAWARE program that includes both approaches and can run as a standalone program on a linux-based computer to easily address in-home massive protein annotation or filtering.

Availability and implementation: http://www.biocomp.unibo.it/~savojard/betawarecl

Contact: piero.fariselli@unibo.it

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on October 26, 2012; revised on December 13, 2012; accepted on December 23, 2012

1 INTRODUCTION

Transmembrane β-barrel proteins (TMBBs) are important proteins that cross the lipid bilayer with a series of β-strands arranged in a cylindrical geometry and forming a structure that resembles a barrel (Schulz, 2000). Although all living organisms have transmembrane proteins organized into all-α helical bundles, TMBBs are presently found in the outer membranes of Gram-negative bacteria, mitochondria and chloroplasts. Despite their relevance, few TMBB structures are available at atomic resolution from Gram-negative organisms [about 0.1% of all the structures from Gram-negative organisms in the protein database (PDB)].

Several computational methods have been developed to predict TMBBs from protein sequences. Two prediction problems can be addressed: (i) TMBB detection in genomes (Bigelow *et al*., 2004; Casadio *et al*., 2003; Freeman and Wimley, 2010; Gromiha and Suwa, 2006; Hayat and Elofsson, 2012; Remmert *et al*., 2009; Savojardo *et al*., 2011), and (ii) the prediction of protein topology (Bagos *et al*., 2004, 2005; Bigelow *et al*., 2004; Fariselli *et al*., 2009; Hayat and Elofsson, 2012; Martelli *et al*., 2002).

Here, we present BETAWARE, a tool based on machine-learning approaches to detect TMBBs in large sets of proteins and predict their topology. BETAWARE is available under GPL license as a standalone program, which is particularly well-suited for large-scale genome analyses. For the prediction of TMBB topology, it has been shown on comparative analyses that approaches based on grammatical modelling are the best performing (Fariselli *et al*., 2009; Hayat and Elofsson, 2012), while for the detection of TMMBs in a set of proteins, the best-performing method is based on N-to-1 Extreme Learning Machines (ELM) (Huang *et al*., 2006; Mooney *et al*., 2011; Savojardo *et al*., 2011).

2 METHODS

Detecting TMBBs in large sets of proteins is like finding a needle in a haystack. To address this problem, we used the previously developed method based on N-to-1 network encoding (Mooney *et al*., 2011) and ELM training algorithm (Huang *et al*., 2006). This method is the best performing on this task as compared with the most recent approaches (Savojardo *et al*., 2011 and Supplementary Table S1). Once a protein sequence is predicted as putative TMBB, we applied a Grammatical Restrained Hidden Conditional Random Field (GRHCRF) model to predict the protein topology (Fariselli *et al*., 2009). The model used is the same as previously described (Fariselli *et al*., 2009), but here it has been retrained on a larger set of proteins.

2.1 Usage and program requirements

BETAWARE is written in pure Python to allow high compatibility. However, it has been explicitly designed to run on Unix/Linux systems (although it would be not too difficult to modify it for other operating systems). The BETAWARE program requires that the following packages are installed in the system: (i) python v2.x (tested on 2.6 or later), python argparse library, python numpy and scipy libraries (under Linux debian/ubuntu you have just to type on a single line: `sudo apt-get install python-numpy python-scipy python-argparse`). Once the software is downloaded and uncompressed, it is possible to run it directly moving into the BETAWARE root directory and typing the command:

```
$> ./predBeta.sh FASTA_FILE PROFILE_FILE
```

where `predBeta.sh is a simple bash script,` which takes as arguments a file containing the sequence to predict in FASTA format and its corresponding sequence profile in a separate file. An example of a

*To whom correspondence should be addressed.

BETAWARE output is:

```
Sequence id     : 1QJ8:A|PDBID|CHAIN|SEQUENCE
Sequence length : 148
Predicted TMBB  : Yes
Topology     : 2-9,23-29,35-46,60-69,78-87,103-114,120-131,135-145
Seq : ATSTVTGGYAQSDAQGQMNKMGGFNLKYRYEEDNSPLGVIGSFTYTEKSRTASSGDYNKN
SS  : iTTTTTTTTooooooooooooTTTTTTiiiiiTTTTTTTTTTTTooooooooooooT
Prob: cbaaaaaaaaaaaaaaaaaaccaaaaaaacaaabccaaaaaaabbdaaaaaaaaacdeb
------------------------------------------------------------------
Seq : QYYGITAGPAYRINDWASIYGVVGVGYGKFQTTEYPTYKNDTSDYGFSYGAGLQFNPMEN
SS  : TTTTTTTTiiiiiiiiTTTTTTTTTooooooooooooooTTTTTTTTTTiiiiiT
Prob: aaaaaaaaabaaaaaabaaaaaaaaaaaaaaaaaaaaaaaadaaaaaaaaabdaaaaae
------------------------------------------------------------------
Seq : VALDFSYEQSRIRSVDVGTWIAGVGYRF
SS  : TTTTTTTTTTooоTTTTTTTTTTiii
Prob: baaaaaaaaaaaaaaaaaaaaaaaba
---------------------------------
```

In the basic setting, BETAWARE works in two steps. First, it predicts with N-to-1 ELM a score for a protein to be a TMBB. Therefore, in case the protein is predicted as TMBB, it assigns the putative topology using the GRHCRF. The output includes the sequence name present in the FASTA file, the sequence length, the TMBB prediction (yes or no) and possibly the topology prediction. When the topology is assigned, the output also displays the posterior probability of the label for each sequence position (from 'a' = highest probability to 'j' = lowest probability, see output example). BETAWARE allows a more advanced control using directly the python script `betaware.py`. In this case, several options are possible, such as forcing the prediction of the topology even if the detection program does not recognize the sequence as a TMBB (-t option), changing the threshold for allowing a more or less stringent detection criterion (-s option), changing the order of the sequence profile columns (the order of the residues in the sequence profile, -a option) and directing the output on a file (-o option).

3 BETAWARE PERFORMANCE

It has been shown that BETAWARE is one of the best-performing methods on the problem of the detection of TMBB in large set of proteins (Savojardo *et al*., 2011). This is also confirmed when it is compared with more recent approaches achieving a Matthews Correlation Coefficient and a F1-score of 0.82 and 0.81, respectively (Supplementary Table S1, Supplementary Material). In particular, BETAWARE appears effective in reducing the number of false positive predictions as compared with other available methods, as indicated by the positive predicted value (PPV) of 0.92 (Supplementary Table S1). The same picture holds also when BETAWARE is applied to genome analyses (Supplementary Table S3, Supplementary Material).

The number of TMMBs available at atomic resolution increased since we introduced the predictor based on GRHCRF (Fariselli, 2009). Here, we retrain the BETAWARE-topology component on a larger set of proteins (for more detail see BTM in Supplementary Material). BTM is a dataset consisting of 54 non-redundant TMBB structures (and topology) obtained by merging OPM (Lomize *et al*., 2006) and PDBTM (Tusnady *et al*., 2005). BETAWARE performance in cross-validation on BTM is reported in Table 1. Compared with previous results on a smaller dataset (Fariselli *et al*., 2009), the overall accuracy is lower. This is probably owing to the fact that BTM contains more difficult topologies, and it is confirmed by the fact that when BETAWARE is trained and tested in cross-validation on a more recent, but smaller, dataset (BOCTOPUSset), its performance is comparable with those obtained by the best-performing methods on this task as indicated by a protein overlap of 0.72 (Supplementary Table S2).

Table 1. BETAWARE performance on the topology prediction on the BTM dataset

POV	SOV(t)	MCC(t)	SN(t)	PPV(t)	Q_2	Q_3
54%	81%	0.59	70%	72%	83%	77%

For the indices see Supplementary Material.

4 CONCLUSION

In this application notes, we present BETAWARE, a standalone program that can detect with high reliability TMBBs in large sets of proteins and that furnishes a state-of-the-art performance in the task of the prediction of TMBB topology. The program can be installed on a Linux-based machine and can be used in-house for wide-screening applications.

ACKNOWLEDGMENTS

R.C. thanks the following grants: PRIN 2009 project 009WXT45Y (MIUR), PON project PON01_02249 and COST BMBS Action TD1101(EU-RTD Programme).

Conflict of Interest: none declared.

REFERENCES

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

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**, 1–13.

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

Casadio,R. *et al*. (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*., **11**, 1158–1168.

Fariselli,P. *et al*. (2009) Grammatical-restrained hidden conditional random fields for bioinformatics applications. *Algorithms Mol. Biol*., **22**, 4–13.

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

Gromiha,M.M. and Suwa,M. (2006) Discrimination of outer membrane proteins using machine learning algorithms. *Proteins*, **63**, 1031–1037.

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

Huang,G.B. *et al*. (2006) Extreme learning machine: theory and applications. *Neurocomputing*, **70**, 489–501.

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

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

Mooney,Y. *et al*. (2011) SCLpred: protein subcellular localization prediction by N-to-1 neural networks. *Bioinformatics*, **27**, 2812–2819.

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

Savojardo,C. *et al*. (2011) Improving the detection of transmembrane β-barrel chains with N-to-1 Extreme Learning Machines. *Bioinformatics*, **27**, 3123–3128.

Schulz,G.E. (2000) β-Barrel membrane proteins. *Curr. Opin. Struct. Biol*., **10**, 443–447.

Tusnady,G. *et al*. (2005) PDB_TM: selection and membrane localization of transmembrane proteins in the protein data bank. *Nucleic Acids Res*., **33**, D275–D278.