DIAPO 1

Leer titulo y presentarme

DIAPO 2

As stated in the title of the paper, BOCTOPUS is a predictor of transmembrane beta barrels, so let’s see the motivation behind studying these structures.

Transmembrane β barrel proteins (TMBs) is one of the two types of TM proteins that exist. They are found in the outer membrane of Gram-negative bacteria, chloroplast and mitochondria. Although they are not as recurrent as the other TM proteins, alpha helical proteins, they are very important because they play a major role in the translocation machinery, transport of molecules, pore formation or ion exchange among other features. TMBs are also promising targets for antimicrobial drugs and vaccines.

However, the number of solved TMB structures in PDB is limited, since they are difficult to crystalize; consequently, computational methods to identify TMBs and predict the topology of TMBs are important.

DIAPO 3

The computational methods in the realm of TMBs can be divided into two parts (i) methods that aim to identify TMBs from genomic data and (ii) methods that predict the TMB topology assuming that the given sequence is a putative TMB. The first group consists of a variety of methods including K-nearest neighbor methods (Hu and Yan, 2008), SVMs (Park et al., 2005), Neural Networks (Gromiha and Suwa, 2006; Gromiha et al., 2004), Hidden Markov Models (HMM) (Deng et al., 2004; Martelli et al., 2002), etcetera.

Methods aiming at the prediction of topologies include HMM based methods such as PRED-TMBB (Bagos et al., 2004), TMBHMM (Singh et al., 2011) and PROFtmb (Bigelow and Rost, 2006), SVM-based methods such as TMBETAPRED-RBF (Ou et al., 2010), neural network-based methods such as TMBpro (Randall et al., 2008) and methods based on statistical potentials such as transFold (Waldispuhl et al., 2006).

A comparison and evaluation carried out indicated that HMM-based methods outperform methods based on other types of machine learning. And in this group is where BOCTOPUS is included

DIAPO 4

BOCTOPUS uses a combination of SVMs to predict the local structural preferences for a residue, and a HMM model to create a topology model for a protein. The primary use of BOCTOPUS is topology prediction of TMBs with the assumption that all input sequences are TMBs.

DIAPO 5

The input feature for BOCTOPUS is a position specific scoring matrices (PSSMs) obtained using PSI-BLAST version 2.2.18 (Altschul et al., 1997). Here, default parameters and three iterations of searching the non-redundant nr-database, obtained from the NCBI website was used.

Three SVMs were trained to determine the preference of each residue to be in the ‘I’, ‘O’ or “M” regions. All residues in the dataset were annotated as either ‘I’ (inner-loop), ‘O’ (outer-loop) or ‘M’ (transmembrane β strand) based on the coordinate of the Cα atoms and membrane boundaries obtained from the OPM database. Here, residues located within the membrane boundaries but do not belong to a transmembrane β-strand are labeled as ‘I’ or ‘O’ based on the location of the initial residue.

Radial basis and linear kernels, different windows sizes in the range of 1– 31 were tried. The optimal window size was determined based on the highest Matthews correlation coefficient MCC values such that no statistically significant improvement was gained on further increasing the window size. Based on this criteria, window size of 31, 19 and 21 was chosen for i, o and M SVMs, respectively.

‘IOM-profile’ generated from the probabilities produced by the three SVMs was used as the input for training different combinations of HMM parameters. The HMM describing the global topology consists of a pre-barrel stage (P) describing the region before the first transmembrane β-strand is detected. Further, a TMB is defined by four different states each representing the inner-loop, outer loop and the up and the down strands. The up and down strand states can handle β-strands in the range of 6–15 residues.

The emission scores for the states are the probabilities obtained from the respective SVMs. Based on the emission scores, the most likely topology is predicted using the Viterbi algorithm.

The best performing HMM parameters were chosen based on the correct number of predicted strands on the training set.

\*\* However, to accommodate for the variable length of large outer-loops, small inner-loops and pre-barrel (defined as the region before the first transmembrane beta-strand) part of the sequence, we found that it was necessary to optimize the three states with self-loops. These three states are shown in bold letters I, P and O with a self-loop.

DIAPO 6

As mentioned above, for each round, training was performed on nine sets, and the remaining set was used for testing. First ‘IOM-profiles’ were generated for proteins in the test-set using SVMs trained only on the training sets. However, as thousands of parameters provided identical (and perfect) results on the training set, the topologies of the proteins in the test-set were then determined by using a subset of these top performing HMMs.

For the final evaluation, 10 000 HMMs were randomly selected from a pool of the top performing HMMs to predict the topology of the proteins in the test sets.

TMB identification. The ability to identify TMBs was tested on a non-redundant (at sequence identity ≤50%) representative dataset of 14 232 PDB entries obtained from Freeman and Wimley (2010). Here, a protein was assigned as a TMB when the number of strands predicted by BOCTOPUS is larger than a given number (typically 8).

DIAPO 7

Segment overlap (SOV), Q2 and Q3. Q2 is defined as the two-state (membrane/not membrane) prediction accuracy. Q3 is defined as the three-state prediction accuracy for i, M, o states. In addition, the number and location of the predicted strands was used to evaluate the performance per-protein. A protein was defined to have a correct predicted topology when the number of predicted strands is correct and each predicted strand overlaps with at least two residues with the observed strand. It should be noted that all results for BOCTOPUS are based on the 10-fold cross.

Per-residue accuracy comparison. Q2, Q3 and SOV scores for BOCTOPUS and other methods. BOCTOPUS-SVM shows the accuracy measures before the HMM stage. When calculating the accuracy measures using the SVMs alone, each residue is assigned to the region with the highest probability in the ‘IOM-profile’. The Q2, Q3 and SOV scores for BOCTOPUSSVM are 90, 85 and 75%, respectively (Table 1). The Q2 and Q3 scores compete favorably with earlier methods, which have Q2 scores ∼85% and Q3 scores up to 82%. However, the SOV score is much lower than these methods, as most of the strands predicted are too short.

In BOCTOPUS, the SVM predictions are used as input into a HMM-like model to obtain the final prediction. This step increases the per-residue accuracy in particular as measured by Q3 and particularly SOV. The SOV score (92%) of BOCTOPUS is higher than any of the earlier methods. BOCTOPUS

DIAPO 8

Regarding topology prediction, BOCTOPUS predicts the correct number of strands on average for 30.1 out of 36 proteins. No. strands is defined as the number of sequences where the number of predicted strands is equal to the number of observed strands.

Further, BOCTOPUS predicted on average 25.4 of these proteins with correct topology. Here, a topology is defined as correct when the number of predicted strands is equal to the number of observed strands and all predicted strands overlap the observed strand.

\*\* Underpredicted (UP) is defined as sequences where the number of strands underpredicted, i.e. some strands are missed. Overpredicted (OP) is defined as sequences where the number of strands is overpredicted. BOCTOPUS-SVM shows the accuracy measures without the HMM stage. Fraction of strands (FS) is defined as the number of observed strands that are correctly predicted to be at the correction location.

Erroneous topology predictions using BOCTOPUS. (A) In the top six cases, BOCTOPUS predicts the correct number of strands; however, one or more strands does not overlap with observed strands. \*\* In 9 of 12 cases, the errors can be attributed to overpredictions in the pre-barrel state

(B) In the bottom six cases, the number of strands predicted by BOCTOPUS does not match the number of observed strands.

\*\*The number below the pdb id represents the percentage of incorrect outcomes when multiple HMMs were employed from the pool of best-trained HMMs.

DIAPO 9

MULTI-CHAIN TMB TOPOLOGY PREDICTION. Multi-chain TMBs are TMBs whose barrel composes of β-strands from different chains.

A comparison of different prediction methods on multi-chain TMBs shows that none of the methods, including BOCTOPUS, can predict the correct topology for more than one or two of these proteins. Out of six multi-chain TMBs, BOCTOPUS gets the number of strands correct only for Haemophilus influenzae Hia autotransporter ß (3emo\_A).

DIAPO 10

TMB IDENTIFICATION BASED ON BOCTOPUS PREDICTIONS. Although BOCTOPUS is best suited for topology prediction and it initially assumes all input sequences to be putative TMBs, we tested its ability to discriminate between TMBs and non-TMBs in a dataset from Freeman and Wimley (2010). Here, a protein is assigned as TMB if the number of predicted strands is larger than a given number N. Figure 6 shows the TMB identification results for different methods combined with BOCTOPUS.

In Table 5, BOCTOPUS(8) and BOCTOPUS(4) refer to cases where the threshold for being a TMB is set at 8 and 4, respectively. BOCTOPUS (8) alone misclassifies 1374 non-TMBs as TMBs, resulting in an MCC value of 0.14. Secondary structure analysis of non-TMB proteins predicted as TMB shows that the regions predicted asTM β-strands are enriched in β-strands

The methods specialized in identification of TMBs (for example: BOMP, tmbetaNet and PSORTb) are better at identifying TMBs than BOCTOPUS (Table 5). However, when these methods are used together with BOCTOPUS, such that a protein is first predicted by a TMB identification method and then checked by BOCTOPUS if the number of strands is ≥ 4 or 8, the number of false positives is reduced without a large decrease in sensitivity. The highest accuracy is obtained when combining PSORTb with BOCTOPUS(8).

DIAPO 11

Here, we present an improved topology predictor for TMBs named BOCTOPUS that combines local per-residue predictions with global preferences. BOCTOPUS is based on ideas previously implemented for the topology prediction of HMPs where different residue preference scores derived from sequence profiles are combined to predict the global topology (Jones, 2007; Viklund and Elofsson, 2008).

Based on a 10-fold cross validation test, the prediction accuracy of BOCTOPUS is higher than earlier methods both when measured on a per-residue and a per-protein basis. BOCTOPUS predicts the correct number of strands in 30 of 36 (83%) TMBs and obtains the correct topology for 70% TMBs in the dataset. We also show that when BOCTOPUS is combined with dedicated TMB identification methods such as BOMP (Berven et al., 2004), PSORTb (Yu et al., 2010) and tmbetaNet (Gromiha et al., 2005b), it can reduce the false positive detection of TMBs. However, the performance in multi-chain TMBs is far from perfect indicating that the correct prediction and identification of such proteins is not a solved problem.