Rapid membrane protein topology prediction
Aron Hennerdal\(^1\) and Arne Elofsson\(^1\)\(^*\)

\(^1\)Department of Biochemistry and Biophysics, Stockholm Bioinformatics Center, Center for Biomembrane Research, Swedish e-science Research Center; Stockholm University; 106 91 Stockholm; Sweden

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Summary: State-of-the-art methods for topology of \(\alpha\)-helical membrane proteins are based on the use of time-consuming multiple sequence alignments obtained from PSI-BLAST or other sources. Here, we examine if it is possible to use the consensus of topology prediction methods that are based on single sequences to obtain a similar accuracy as the more accurate multiple sequence-based methods. Here, we show that TOPCONS-single performs better than any of the other topology prediction methods tested here, but about 6\% worse than the best method that is utilizing multiple sequence alignments.

Availability and Implementation: TOPCONS-single is available as a web-server from http://single.topcons.net/ and is also included for local installation from the website. In addition, consensus based topology predictions for the entire IPI is available from the web-server and will be updated at regular intervals.

Supplementary information: Benchmark data and the prediction results for all combinations is available as supplementary information.

Contact: arne@bioinfo.se

1 INTRODUCTION

Today only 268 unique \(\alpha\)-helical membrane protein structures are known according to the Orientation of Proteins in Membranes database (OPM, http://opm.phar.umich.edu/). The “topology” of such proteins has proven to be a convenient concept. In essence, the topology specifies the number of transmembrane \(\alpha\)-helices of the protein together with the location of the N-terminal end of the chain, i.e. whether it is in the cytosol (“in”) or in the ER-lumen or extra-membrane space (“out”).

The TOPCONS algorithm (Bernsel et al., 2009) computes consensus predictions of membrane protein topology using a Hidden Markov Model (HMM) and input from several topology predictors. The original method is available as a web-server (http://topcons.net/) and is based on five state-of-the-art topology prediction methods and typically takes a couple of minutes to run. The bulk of that time is spent running a PSI-BLAST (Altschul et al., 1997) search against a sequence database to obtain evolutionary information that is then used by the underlying predictors. This approach is quite accurate, but woefully inappropriate when running predictions for many sequences, e.g. in studies of whole genomes.

Table I. The accuracy of different predictors on different datasets. Homology reduced to 30\% sequence identity. The numbers in parenthesis denote the number of protein sequences in the set. “Time” is the time it takes to process the set of 101 protein sequences.

<table>
<thead>
<tr>
<th>Topology</th>
<th>Predictor</th>
<th>Time (s)</th>
<th>all (101)</th>
<th>multi (79)</th>
<th>single (22)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SCAMPI-single</td>
<td>2</td>
<td>62%</td>
<td>62%</td>
<td>64%</td>
<td></td>
</tr>
<tr>
<td>HMMTOP</td>
<td>10</td>
<td>57%</td>
<td>53%</td>
<td>73%</td>
<td></td>
</tr>
<tr>
<td>PHOBIUS</td>
<td>26</td>
<td>52%</td>
<td>56%</td>
<td>41%</td>
<td></td>
</tr>
<tr>
<td>S-TMHMM</td>
<td>10</td>
<td>51%</td>
<td>53%</td>
<td>45%</td>
<td></td>
</tr>
<tr>
<td>MEMSAT-1.0</td>
<td>18</td>
<td>56%</td>
<td>54%</td>
<td>64%</td>
<td></td>
</tr>
<tr>
<td>TOPPRED</td>
<td>2</td>
<td>33%</td>
<td>30%</td>
<td>41%</td>
<td></td>
</tr>
<tr>
<td>TOPCONS-single</td>
<td>64</td>
<td>73%</td>
<td>68%</td>
<td>91%</td>
<td></td>
</tr>
<tr>
<td>TOPCONS</td>
<td>4483</td>
<td>79%</td>
<td>77%</td>
<td>86%</td>
<td></td>
</tr>
</tbody>
</table>

2 DEVELOPMENT OF TOPCONS-SINGLE

Here, we have benchmarked the TOPCONS algorithm (Bernsel et al., 2009) using six different topology prediction methods that do not use any homology information, i.e. do not require BLAST to be run. Six individual methods were tested: SCAMPI-single (Bernsel et al., 2008) S-TMHMM (Viklund and Elofsson, 2004), HMMTOP (Tusnády and Simon, 2001), TopPred (von Heijne, 1992, Claros and Heijne, 1994), MEMSAT-1.0 (Jones et al., 1994) and PHOBIUS (Käll et al., 2004).

The methods were benchmarked using a modified version of the data-set used in SCAMPI (Bernsel et al., 2008). The original set consisted of two subsets stemming from high-resolution structures (123 sequences) and from structures of lower resolution (146 sequences). This set was homology-reduced to 30\% sequence identity using the method proposed and implemented by Holm and Sander, 1998. The reduced set contain 101 sequences and was further divided into multi-spanning (79 sequences) and single-spanning (22 sequences) proteins resulting in three sets labeled ‘all’, ‘multi’ and ‘single’, respectively.

All possible combinations of three or more topology predictors were used as input to the TOPCONS algorithm and the results were evaluated. The best combination – the one scoring the highest accuracy over the dataset – is listed in Table I. Accuracy is the proportion of correct predictions, and correct topology predictions are defined as by Krogh et al., 2001. All definitions and the full list of all method combinations are available in the supplementary information. To enable comparison, the performance of the original TOPCONS server based on homology information is listed, as well as the individual performance for the six single-sequence methods.

\(^{*}\)to whom correspondence should be addressed
Previously described (Bernsell et al., 2008), we have constructed a consensus predictor for predicting the number of transmembrane helices in the same extent as the single-sequence methods (Supplementary information). TOPCONS-single performs especially well on single-spanning membrane proteins in our dataset (Table I) mainly by not over-predicting the number of transmembrane helices in the same extent as the single-sequence methods (Supplementary information).

A possible caveat to our approach is the use of benchmark sets where at least subsets have been previously used to train the underlying single-sequence methods. We judge this to be less influential since the authors of said prediction methods have taken steps to avoid over-training on their respective sets.

The best-performing version of TOPCONS-single, using four individual methods (Table I), is available as an easy-to-use web-based prediction server at http://single.topcons.net/. It uses the globular protein filter of SCAMPI to weed out non-membrane proteins and then proceeds to run the rest of the predictors – HMMTOP, MEMSAT-1.0 and S-TMHMM – on the remaining set. The output consists of text-files with well-defined formats for easy parsing.

**ACKNOWLEDGMENT**

This work was supported by grants from the Swedish Research Council (VR-NT 2009-5072, VR-M 2007-3065), SSF (The Foundation for Strategic Research), the EU 6’Th Framework Program by support to the EMBRACE project, contract No: LSHG-CT-2004-512092 and the EU 7’Th Framework Program by support to the EDICT project, contract No: FP7-HEALTH-F4-2007-201924. We also thank Dr. Håkan Viklund for writing the modhmm code used in TOPCONS.

**REFERENCES**


