Rapid membrane protein topology prediction – Supplementary information

Aron Hennerdal and Arne Elofsson

The TOPCONS algorithm

The TOPCONS algorithm uses a Hidden Markov Model to combine the results of an arbitrarily large number of topology predictors into one consensus prediction. It is described by Bernsel *et al.*, 2009: The original

> "As input TOPCONS uses a set of topology predictions which are combined into a topology profile by letting each residue be represented by three values representing the fraction of methods that predict that residue to be situated in the membrane (M), on the inside of the membrane (i) or on the outside of the membrane (o). This topology profile is used as input to a dynamic programming algorithm similar to a hidden Markov model that has an alphabet consisting of the characters 'M', 'i' and 'o'. The final topology corresponds to the highest scoring state path through this model using a Viterbi-like algorithm. In each state, the emission score for the structural category modeled by that state (i, o or M) is equal to 1.0 and for all other structural categories it equals 0.0. All transition probabilities are equal to 1.0. Thus, the final prediction equals the state path with the highest geometric mean score with respect to the topology profile and the grammar of the model."

The original implementation uses five topology prediction methods: SCAMPI-single, SCAMPI-multi, PRO-TMHMM, PRODIV-TMHMM and OCTOPUS of which the last four requires BLAST-results as input (Figure 1). The algorithm itself can use predicted topologies from any number of sources.

Prediction accuracy

All figures of accuracy in the paper are the proportion of true results, i.e.

$$\frac{TP + TN}{TP + FP + TM + FN},$$

where TP is the number of true positives, TN the number of true negatives, FP the number of false positives, and FN the number of false negatives.

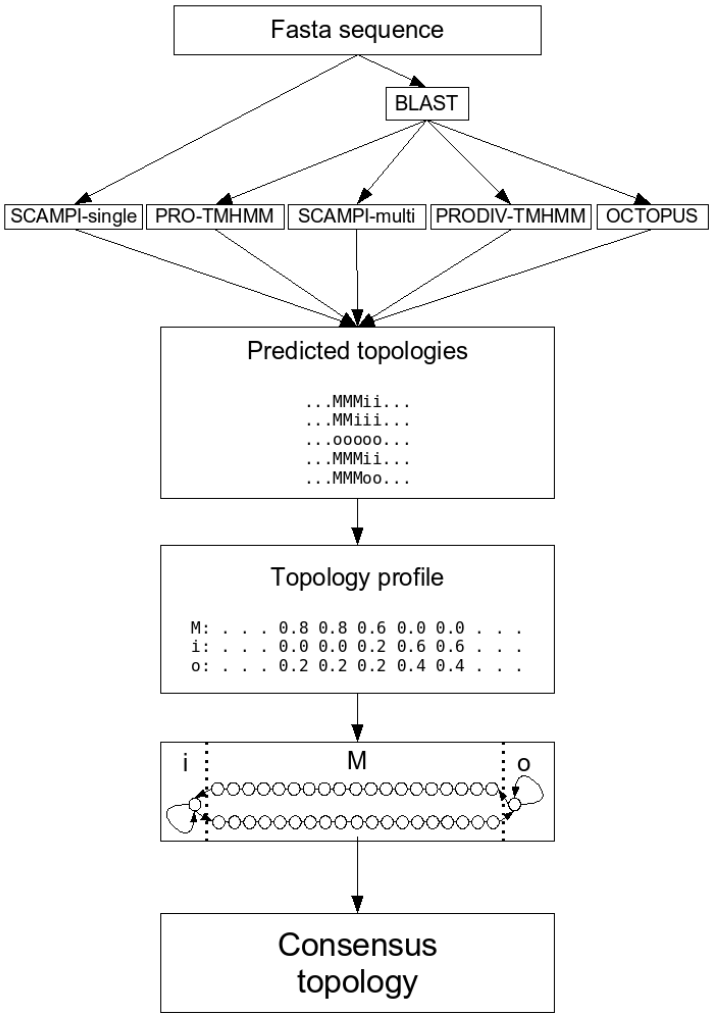

Figure 1: TOPCONS work flow: four of the topology predictors make use of multiple sequence information and require a sequence profile as input, created using BLAST (18), whereas the fifth method (SCAMPI- single) only requires the protein sequence. The topology predictions are used to construct a topology profile, which is fed into the TOPCONS hidden Markov model (Bernsel *et al.*, 2009)

Correct prediction

We use the definition of a correct topology prediction of topology of Krogh *et al*, 2001. A prediction is considered correct if:

- the number of transmembrane helices is correct,
- each predicted helix overlaps by at least 5 residues with corresponding real

helix,

- each helix is oriented correctly, i.e. N-terminals and C-terminals are at the same side of the membrane as respective terminals in the real helix.

Reliability scores

TOPCONS-single

The reliability score for the TOPCONS-single predictions is the same as introduced by Bernsel et al, 2009. The TOPCONS algorithm produces a consensus profile on the same format as the input topology profiles, Figure 1. Each of the topology states 'i' (cytosol), 'o' (ER-lumen/extra-membrane space) and 'M' (in the membrane) are represented by a number between zero and one for each position in the predicted topology. A higher value for one of the states compared to the others can be interpreted as higher certainty in the prediction for that position. The reliability measure is then computed as described by Bernsel et al :

> "A reliability value is calculated for each residue in a sequence by taking the average over a 21 position window of the topology profile value for the consensus prediction of that position (i, o, M). A reliability score on the protein level is calculated by taking the minimum value as calculated above."

HMMTOP, MEMSAT-1.0, TopPred

These reliability scores are the same as in the work of Melen *et al*, 2003. Each of the three methods produces a number of suggestions for topology prediction from which the best one according to a method-specific score is chosen as the reported prediction. The idea for a reliability score is that a large difference in this score between the best (chosen) prediction and the second best prediction indicates that the prediction chosen is more certain and gets a correspondingly higher reliability score.

HMMTOP produces a value of entropy for the best path through its Hidden Markov Model and an entropy for the whole model. The reliability score is defined as the difference between them:

$$\text{rel. score} = \textit{entropy}(\text{best path}) - \textit{entropy}(\text{model}).$$

MEMSAT-1.0 gives scores for all possible topologies starting with one helix and increasing until a certain score-threshold is reached. The reliability score is defined as the difference between the highest score and the second highest score:

$$\text{rel. score} = \textit{score}(\text{best prediction}) - \textit{score}(\text{second best topology}).$$

TopPred relies on a hydrophobicity scale to define "certain" and "putative" helices. Alternative topologies are generated including all helices designated "certain" and different combinations of the "putative" helices. The prediction with the largest difference in the number of positively charged amino acids between the two sides of the membrane is chosen as the final prediction. Ranking

the predictions in this way, the reliability score is defined as the difference between the charge-differences of the two predictions of highest rank:

rel. score = Δpositive charges(best topology)$-\Delta$positive charges(second best topology).

SCAMPI-single, S-TMHMM, PHOBIUS

All three methods are implemented using the MODHMM-package (http://www.topcons.net/index.php?about=download), which yields output in a common format with a number of scores based on the posterior probabilities of the respective Hidden Markov Model. The package also allows the fixing of certain positions in the input sequence as being in a particular topology state. As a reliability score, we have used the difference in normalized log likelihood for the actual prediction and a prediction made after fixing the N-terminal of the input sequence to the opposite side of the membrane as compared to the actual prediction, i.e.

$$\text{rel. score} = \begin{cases} nll(\text{pred. top.}) - nll(\text{pred. top.}_{Nin}), & \text{if pred. top. is Nout} \\ nll(\text{pred. top.}) - nll(\text{pred. top.}_{Nout}), & \text{if pred. top. is Nin} \end{cases}$$

where *nll* is the normalized log likelihood, Nin refers to a topology with the N-terminal in the cytosol, and Nout similarly means a topology with the N-terminal in the ER-lumen/extra-membrane space.

A higher score means a larger difference in prediction score between the prediction made and a prediction very close to it, and thus indicates a higher confidence in the prediction.

Agreement between input methods

The performance of TOPCONS-single as a function of the underlying methods' internal agreement was investigated, and is shown in Figure 2. As a measure of "agreement", we have used the definition of correct topology prediction as described by Krogh *et al*, 2001 (see above), but used to compare two predicted topologies. Two predicted topologies are "identical" if they have the same number of transmembrane helices, their N- and C-terminals are on the same respective side of the membrane, and each helix in a prediction is overlapping by at least five residues with its counter-part in the other prediction.

Over- and under-predictions

A break-down of the prediction results into different categories is shown in tables 1, 2, and 3. "Over" refers to over-prediction of the number of helices; "Under" similarly means under-prediction of helices; "Inverse" refers to inverted topologies, i.e. the number of helices is predicted correctly, but the N-terminal is predicted to be on the opposite side of the membrane; "Shift" designate predictions where the number of helices is correct, but one or more of them doesn't overlap with 5 residues or more with a real helix. See also the definition of a correct prediction above.

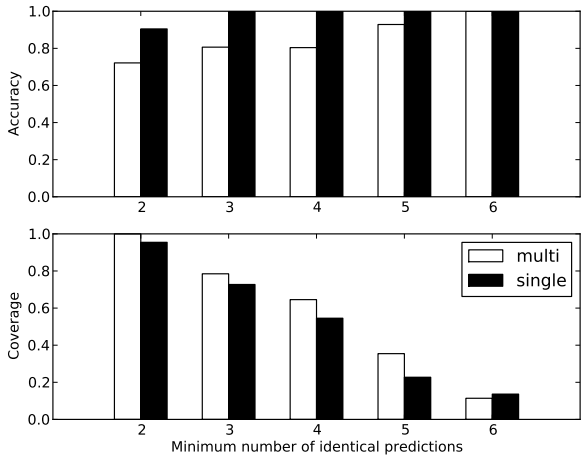

Figure 2: Performance of TOPCONS-single for different amount of agreement between individual methods. The x-axis lists the number of agreeing methods and the y-axis shows predictive performance in the form of accuracy on the left and the fraction of proteins falling into this category on the right.

	Correct	Inverse	Over	Under	Shift
TOPCONS-single	74	3	11	11	2
SCAMPI-single	63	4	21	12	1
HMMTOP	58	2	26	14	1
S-TMHMM	52	13	14	21	1
MEMSAT-1.0	57	3	16	24	1

Table 1: Prediction result category break-down for the 'all' dataset.

	Correct	Inverse	Over	Under	Shift
TOPCONS-single	20	1	1	0	0
SCAMPI-single	14	0	7	1	0
HMMTOP	16	1	4	1	0
S-TMHMM	10	9	1	2	0
MEMSAT-1.0	14	2	6	0	0

Table 2: Prediction result category break-down for the 'single' dataset.

Results for all different combinations of input methods to the TOPCONS algorithm

Tables 4, 5, 6, and 7 show the performance of TOPCONS-single using 6, 5, 4, and 3 input methods respectively.

	Correct	Inverse	Over	Under	Shift
TOPCONS-single	54	2	10	11	2
SCAMPI-single	49	4	14	11	1
HMMTOP	42	1	22	13	1
S-TMHMM	42	4	13	19	1
MEMSAT-1.0	43	1	10	24	1

Table 3: Prediction result category break-down for the 'multi' dataset.

Topology predictors used	all	multi sp.	single sp.
ss, ht, tp, ps, st, ms	0.69	0.65	0.86

Table 4: TOPCONS-single results for 6 input methods. st: S-TMHMM, ss: SCAMPI-single, tp: TopPred, ht: HMMTOP, ps: PHOBIUS, ms: MEMSAT-1.0

Topology predictors used	all	multi sp.	single sp.
ss, ht, tp, st, ms	0.69	0.66	0.82
ss, ht, tp, ps, ms	0.69	0.68	0.73
ss, ht, ps, st, ms	0.65	0.65	0.68
ht, tp, ps, st, ms	0.65	0.65	0.68
ss, ht, tp, ps, st	0.63	0.62	0.68
ss, tp, ps, st, ms	0.62	0.65	0.55

Table 5: TOPCONS-single results for all combinations of 5 input methods. st: S-TMHMM, ss: SCAMPI-single, tp: TopPred, ht: HMMTOP, ps: PHOBIUS, ms: MEMSAT-1.0

Topology predictors used	all	multi sp.	single sp.
ss, ht, st, ms	0.73	0.68	0.91
ss, ht, ps, ms	0.71	0.67	0.86
ss, ht, tp, st	0.67	0.61	0.91
ss, ht, tp, ms	0.65	0.61	0.82
ss, ht, tp, ps	0.65	0.61	0.82
ss, tp, st, ms	0.66	0.65	0.73
ht, tp, st, ms	0.61	0.58	0.73
ss, tp, ps, ms	0.64	0.66	0.59
ht, tp, ps, ms	0.59	0.57	0.68
ss, ht, ps, st	0.60	0.59	0.64
ht, ps, st, ms	0.60	0.59	0.64
ss, ps, st, ms	0.61	0.62	0.59
ht, tp, ps, st	0.56	0.56	0.59
tp, ps, st, ms	0.59	0.62	0.50
ss, tp, ps, st	0.57	0.58	0.55

Table 6: TOPCONS-single results for all combinations of 4 input methods. st: S-TMHMM, ss: SCAMPI-single, tp: TopPred, ht: HMMTOP, ps: PHOBIUS, ms: MEMSAT-1.0

Topology predictors used	all	multi sp.	single sp.
ss, ht, ms	0.70	0.66	0.86
ss, ht, ps	0.65	0.63	0.73
ss, ht, tp	0.62	0.57	0.82
ss, ht, st	0.65	0.65	0.68
ht, ps, ms	0.63	0.61	0.73
ht, st, ms	0.64	0.63	0.68
ss, st, ms	0.64	0.65	0.64
ss, tp, ms	0.59	0.58	0.64
ss, ps, ms	0.63	0.67	0.50
ps, st, ms	0.60	0.62	0.55
ss, ps, st	0.59	0.59	0.59
ss, tp, st	0.59	0.61	0.55
ht, tp, ms	0.54	0.51	0.68
tp, st, ms	0.57	0.58	0.55
ht, ps, st	0.55	0.56	0.55
ht, tp, ps	0.53	0.54	0.50
ht, tp, st	0.52	0.53	0.50
ss, tp, ps	0.54	0.58	0.41
tp, ps, ms	0.51	0.53	0.45
tp, ps, st	0.50	0.54	0.36

Table 7: TOPCONS-single results for all combinations of 3 input methods. st: S-TMHMM, ss: SCAMPI-single, tp: TopPred, ht: HMMTOP, ps: PHOBIUS, ms: MEMSAT-1.0