According to genome-wide estimations, roughly 20–30% of the TM proteins is difficult to attain, both experimentally and by online.

Supplementary information: Supplementary data are available at OCTOPUS is the first topology predictor to fully integrate modeling of reentrant/membrane-dipping regions and transmembrane hairpins. OCTOPUS predicts the correct topology for 94% of the sequences. In particular, of hidden Markov models and artificial neural networks, OCTOPUS provides an excellent template for further experimental studies, both in silico and in the laboratory. It is also very useful for functional and structural classification of protein sequences on a genomic level.

In general, topology prediction can be divided into four components: residue representation, prediction of residue preference, grammatical modeling and global topology prediction. First, residue representation refers to how the information of each residue is formulated. The most common input is the amino acid sequence itself, but this information can be preprocessed to include for example an alignment of homologous sequences or a sequence of derived hydrophobicity values. Second, all methods define (implicitly or explicitly) a way to estimate the preference of each residue to be situated in the membrane (M), on the inside (cytoplasmic side, i) or on the outside (non-cytoplasmic side, o). Third, methods define a topological grammar describing which of all combinations of residue label assignments (i, o, M) will constitute a valid topology. For example, grammars include definitions of allowed lengths of TM regions and minimum inter-TM loop lengths, as well as place consecutive loop regions on opposite sides of the membrane. Fourth, there is a procedure for calculating the final topology, given the preference scores for the individual residues and the topological grammar.

During the last couple of decades, a wide variety of prediction methods for TM protein topology has been developed, using conceptually different strategies, with respect to the implementation of these components. The earliest methods were solely based on the fact that TM regions generally are more hydrophobic than the rest of the protein, and subsequently used hydrophobicity profiles to predict the location of TM regions (Argos et al., 1982; Kyte and Doolittle, 1982). Advancements on both the grammatical and global levels came with Toppred (Claros and von Heijne, 1994; von Heijne, 1992), which integrated estimated residue preferences of being on the inside/outside of the membrane, based on the positive inside rule (von Heijne, 1986), with the propensity of being in a TM region, based on hydrophobicity, to produce a full 3-state (i, o, M) topology prediction.

Further improvement on the algorithmic level arrived with MEMSAT (Jones et al., 1994), which was the first method to apply an algorithm, which guaranteed that the globally most likely topology (with respect to the underlying residue preference values) was found. At roughly same time, PHDhtm (Rost et al., 1995, 1996) provided novelties on the residue representation level, both by using
The main dataset used in this study consists of 124 protein chains with known structures where 115 are from the Orientation of proteins in membranes (OPM) database (Lomize et al., 2006). Topology annotations were performed automatically using the coordinates and membrane border estimations from OPM. For proteins located in the mitochondrial outer membrane, OPM defines the inter-membrane space as 'inside' and the cytoplasm as 'outside'. To get a labeling that is consistent with that of the other membranes, we have redefined the inter-membrane space as 'outside' and the cytoplasm as 'inside'. For nine additional sequences in PDB that are not in OPM, the biological unit PDB structures were rotated and translated to two separate sets of rules.

2 METHODS

2.1 Sequence database

The main dataset used in this study consists of 124 protein chains with known three-dimensional structures where 115 are from the Orientation of proteins in membranes (OPM) database (Lomize et al., 2006). Topology annotations were performed automatically using the coordinates and membrane border estimations from OPM. For proteins located in the mitochondrial outer membrane, OPM defines the inter-membrane space as 'inside' and the cytoplasm as 'outside'. To get a labeling that is consistent with that of the other membranes, we have redefined the inter-membrane space as 'outside' and the cytoplasm as 'inside'. For nine additional sequences in PDB that are not in OPM, the biological unit PDB structures were rotated and translated as described by Tushuk et al. (2005). This dataset is homology reduced at 40% sequence identity.

Start (end) residues of TM regions were defined so that the first (last) residue of a TM region is the first (last) residue in a sequence of consecutive residues, with their C-atoms located inside the defined membrane borders, emerging on the opposite side of the membrane. Consecutive stretches of membrane residues, where both ends emerge at the same side of the membrane are defined as either inside/outside (i/o), transmembrane (M), TM hairpin (H), reentrant (R) or membrane dip (D) according to the criteria described in Figure 1. According to these definitions, our dataset contains 44 TM, 7 hairpin, 28 reentrant and 21 dip regions. A hydrophobicity value for each TM region was calculated using the Goldman, Engelman and Steitz scale by averaging the hydrophobicity values for each residue based on the above definition (Engelman et al., 1986). The hydrophobicity of each TM region was used to approximately divide these regions into 'hard' (hydrophobicity value <0.5) and 'easy' (hydrophobicity value >0.5) targets.

A second dataset consisting of 163 sequences with known topologies, which was homology reduced at 30% sequence identity both internally and with respect to the main dataset, was used for additional testing. Results are supplied in Supplementary Table S1. A third dataset consisting of 1087 globular proteins from Swissprot, originally compiled by Käll et al. (2004), was used to test discrimination between TM and globular proteins.

2.2 Training data

For parameter optimization, each residue was labeled twice, according to two separate sets of rules. First, each residue was assigned to one or two of four structural categories: membrane (M), interface (I), loop (L) and globular (G). These four categories were defined using Z-coordinate cutoffs, where 'globular' was defined by a residue being situated >23 Å away from the membrane center, 'loop' as 13–23 Å from the membrane center, 'membrane' as <13 Å from the membrane center and interface as 11–18 Å from the membrane center. Notice also that these definitions may differ slightly from the actual topology definitions described in the previous section. This is due to the fact that while topologies are defined according to variable membrane borders, OCTOPUS models a generic membrane. This is a necessary simplification, since for an uncharacterized sequence, its actual target membrane is usually unknown.

Second, residues situated in loop regions ≤12 positions from the membrane were defined as either inside (i) or outside (o) depending on which side of the membrane they were located on. These definitions were used as input to optimization of the networks that predict inside/outside preference.

Evaluation of prediction accuracy for OCTOPUS was performed using 10-fold cross-validation. All proteins were structurally aligned on a chain basis using Stuctal (Gerstein and Levitt, 1998) and divided into subsets so that any homologous proteins were placed in the same subset. Two protein chains are considered homologous if they belong to the same superfamily and have a structural alignment with a P-value lower than 10−5. When testing
of the sliding window (columns 3–4) and the per residue accuracy, represented by an

sequence one raw frequency profile was created by running BLAST with

2.3.1 Residue-level representation and preference estimation

For each

The full architecture of the OCTOPUS–HMM is shown in Figure 3.

2.3.3 Global algorithm

Emission scores are calculated using the input values for the two alphabets, the predicted preference values for

by BLAST, which was converted back to amino acid frequencies using the logistic function $1/(1+e^{-x})$ (Jones, 1999).

Two separate sets of neural networks were optimized to predict explicit residue preferences with respect to membrane/non-membrane and inside/outside location.

The first set consists of four separate ANNs that predict the residue preference for M, L, I, and G, respectively. Input data to each network is a sliding window of PSSM-profile columns of 29 residues and each network has one hidden layer with eight nodes and a single output node. The MCCs for residue separation for all networks are shown in Table 1. It can be noted that prediction of globular and membrane residues is considerably easier than predicting reentrant and loop residues.

To attain a smoother output for M/G predictions, a sliding window over 39/51 residues of the output values was used as input to a second network for each of these structural categories. The number of hidden nodes in these networks was eight and there was a single output node. The respective MCC-values after applying this optimization are also shown in Table 1. Finally, network preference values were transformed using the formula new value $= \text{original value} * 0.11/1.1$ to limit the influence of proportionally large differences between small values.

The second set of networks, for predicting inside/outside residue preference, consists of only one network with eight hidden nodes and two output nodes. The input to this network is a sliding window of 31 residues, where each residue position contains two values, giving a total of 62 input nodes. The values in each position are based on the raw frequency profiles, and consist of the average fraction over a 25 residue window of the amino acids Arg+Lys and Tyr+Trp, respectively, divided by the maximal such average fraction in the entire sequence. This translates to the input values being a series of values, ranging from 0 to 1, of relative accumulation of Arg+Lys/Tyr+Trp.

Combined, the outputs of these two sets of networks are used to attain residue preference scores as defined in Figure 3.

2.3.2 Topological grammar

The grammar of topology in OCTOPUS is defined using a HMM. The model consists of 10 state compartments (subsets of interconnected states with the same state label), inTM, outTM, inHairpin, outHairpin, inLoop, outLoop, inGlob, outGlob, inReent/Dip and outReent/Dip. The standard definition of an HMM was changed so that transition probabilities are set to 1.0. The original reason for this was to make the final topology depend only on the network output values and the model architecture, and not on the distribution of topologies in the dataset. For practical reasons two exceptions were made to this rule. The transitions from a globular to a loop state was set to 0.001 and transitions from a reentrant/dip state to a loop state was set to 0.2. The reason for this was to limit the minimum length of globular and reentrant/dip sequence stretches without increasing the number of states in the model.

The inTM and outTM compartments allow two TM regions lengths, 21 and 31 residues. The 21 residue track models a ‘typical’ TM helix. In our dataset, the 13 to 17 part of TM helices actually vary between 15 and 33 residues, but >90% of the TM regions are between 17 and 23 residues, with an average length of 20.2. The 31 residue track models long TM helices and is balanced by the hairpin model, which is also 31 residues long. Consequently, the choice between predicting a long TM helix or a hairpin depends only on inside/outside preferences.

The inReent/Dip and outReent/Dip compartments consist of one state each, using a lower transition value out from these compartments to restrict the minimum length of these regions. To enter a reentrant/dip state there is an entry and an exit path consisting of three loop states. This is to avoid predicting reentrants/dips too close to a predicted TM region. Glob and Loop compartments (except for the reentrant/dip entry/exit) on each side of the membrane consist of one state.

The full architecture of the OCTOPUS–HMM is shown in Figure 3.

The performance for each subset, parameter optimization was performed on the remaining 9/10 of the data. The complete dataset including the subset division is available as Supplementary Material and can be downloaded from http://octopus.cbr.su.se.

### Table 1. Training data classification and per residue prediction accuracy

<table>
<thead>
<tr>
<th>Region</th>
<th>Borders</th>
<th>Size1</th>
<th>Size2</th>
<th>MCC1</th>
<th>MCC2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Membrane (M)</td>
<td>(±) 0-13</td>
<td>29</td>
<td>39</td>
<td>0.72</td>
<td>0.76</td>
</tr>
<tr>
<td>Interface (I)</td>
<td>(±) 11-18</td>
<td>29</td>
<td>X</td>
<td>0.27</td>
<td>X</td>
</tr>
<tr>
<td>Loop (L)</td>
<td>(±) 13-23</td>
<td>29</td>
<td>X</td>
<td>0.41</td>
<td>X</td>
</tr>
<tr>
<td>Globular (G)</td>
<td>(±) 23-29</td>
<td>29</td>
<td>51</td>
<td>0.70</td>
<td>0.73</td>
</tr>
<tr>
<td>Inside (i)</td>
<td>1-12</td>
<td>31</td>
<td>X</td>
<td>0.63</td>
<td>X</td>
</tr>
<tr>
<td>Outside (o)</td>
<td>1-12o</td>
<td>31</td>
<td>X</td>
<td>0.63</td>
<td>X</td>
</tr>
</tbody>
</table>

Column 2 shows which Z-coordinate interval (M,L,G) or loop residues (i,o), denoted by distance (in residues) to the nearest M-region, that are used for optimization of the neural network parameters for the first network layer. For each neural network of the first (1) and second (2) layer of the OCTOPUS network architecture (Fig. 2), the size of the sliding window (columns 3–4) and the per residue accuracy, represented by an MCC (Matthews correlation coefficient) value (columns 5–6), are shown. MCC values are calculated as $TP/(TP+FN+FP+TN)$, with a cutoff value of 0.5 for deciding if the neural network output is considered a true positive (TP)/false negative (FN) or false positive (FP)/true negative (TN).

Outside (o) 1–12o 31 X 0.63 X

Inside (i) 1–12i 31 X 0.63 X

Fig. 2. Flowchart of OCTOPUS. BLAST (Altschul et al., 1990) is used to generate a multiple sequence alignment from which a raw sequence profile and a sequence profile based on the PSSM are extracted. These are used as input to a set of neural networks that predict inside/outside and membrane/non-membrane residue preferences. The network outputs are used as input to the OCTOPUS–HMM (Fig. 3) and the Viterbi algorithm is used to calculate the final topology, based on emission scores as defined in Figure 3.

H.Viklund and A.Elofsson
the exception that a few transition values are not equal to 1.0.

Fig. 3. HMM-architecture of OCTOPUS. State compartment labels: G, inGlob, outGlob; L, iLoop, oLoop, M, inTM, outTM; H, inHairpin, outHairpin; Hi, H0, in/out state at the bottom of the hairpin compartment; R/D, in/outDip, o/inDip, H1, first/last state in TM/hairpin compartments; Mo, first/last state in TM/ourTM compartments. The box labelled ‘State scores’ describes the calculation of emission scores in the different states. P(X) represents the estimated score for the likelihood of a residue to be situated in this region. For M, L and G, this score is equal to the output of the respective neural network. For H, the P(X) equals the output from the M network, while for R/D it is a combination of the output from M and L networks. For Mi and Mo, the score is calculated as the average i/o value (i for Mi and o for Mo) of 16 residues. Given that the residue for which the score is calculated is emitted in a particular Mi/Mo state, these 12 residues are the residue itself, the three adjacent residues that would be emitted in the neighboring states of the inTM/ourTM compartment and the eight adjacent residues on the opposite side. With some flexibility to accommodate slight shifts in the positioning of the TM helix, this interval is designed to correlate with the training data, where the 12 loop residues closest to each membrane border are used. For Hi and H0, the score is calculated in the same way as for Mi and Mo, but based on the residue itself and the two adjacent residues on either side.

membrane/non-membrane and inside/outside, which are the outputs from the neural networks.

The first alphabet consists of the letters M, L, G and I and is used for calculation of membrane/non membrane preference. For compartments M, L and G, the model emission probabilities for this alphabet are set to 1.0 for each letter in its corresponding state compartment and 0.0 in all other state compartments. For Hi compartments, emission probabilities are set to 1.0 for letter M, while for R/D compartments, the emission score is P(I) = 0.7 + P(M). The second alphabet consists of the letters i and o and is used for the calculation of inside/outside preference. This alphabet is utilized in the start/end states of TM compartments (inTM, outTM, inHairpin and outHairpin). In iTM and oTM, the average value over 16 residues (4 residues towards the membrane side and 12 residues towards the loop side) of the output from the i/o network, depending on which side that state is situated, is added to the score of the first alphabet. For the bottom states of the hairpin compartments (Hi and H0), the calculation is similar, but based on five residues, namely the residue itself and the two adjacent to it on either side. In all other states, the generic value 0.5 is added.

Based on these emission scores, the most likely topology is calculated using the Viterbi algorithm. Intuitively, the final prediction corresponds to the state path that has the highest geometric mean of emission scores, with the exception that a few transition values are not equal to 1.0.

Table 2. Prediction accuracies and different types of prediction errors

<table>
<thead>
<tr>
<th>Method</th>
<th>Corr. topo.</th>
<th>Seqs w. FN</th>
<th>Seqs w. FP</th>
<th>Inv. topo.</th>
<th>DR FP / all FP</th>
</tr>
</thead>
<tbody>
<tr>
<td>OCTOPUS</td>
<td>94</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>100</td>
</tr>
<tr>
<td>MEMSAT</td>
<td>87</td>
<td>7</td>
<td>6</td>
<td>56</td>
<td></td>
</tr>
<tr>
<td>PRODIV</td>
<td>81</td>
<td>5</td>
<td>11</td>
<td>5</td>
<td>83</td>
</tr>
<tr>
<td>PKO</td>
<td>77</td>
<td>12</td>
<td>10</td>
<td>3</td>
<td>92</td>
</tr>
<tr>
<td>HMMTOP_multi</td>
<td>74</td>
<td>7</td>
<td>15</td>
<td>6</td>
<td>68</td>
</tr>
<tr>
<td>HMMTOP</td>
<td>73</td>
<td>9</td>
<td>15</td>
<td>5</td>
<td>58</td>
</tr>
<tr>
<td>Topped II</td>
<td>69</td>
<td>13</td>
<td>10</td>
<td>9</td>
<td>53</td>
</tr>
<tr>
<td>PolyPhbias</td>
<td>66</td>
<td>13</td>
<td>9</td>
<td>13</td>
<td>77</td>
</tr>
<tr>
<td>PhDlmm</td>
<td>63</td>
<td>21</td>
<td>10</td>
<td>6</td>
<td>69</td>
</tr>
<tr>
<td>Phobus</td>
<td>63</td>
<td>16</td>
<td>11</td>
<td>12</td>
<td>73</td>
</tr>
<tr>
<td>TMHMM2.0</td>
<td>60</td>
<td>17</td>
<td>7</td>
<td>17</td>
<td>88</td>
</tr>
<tr>
<td>MEMSAT</td>
<td>57</td>
<td>16</td>
<td>12</td>
<td>15</td>
<td>61</td>
</tr>
<tr>
<td>PhDlmm</td>
<td>52</td>
<td>27</td>
<td>15</td>
<td>6</td>
<td>62</td>
</tr>
</tbody>
</table>

Column 1: Method—Prediction methods. The predictions for each method were obtained using default parameter settings. In the case where multiple sequence information needed to be provided as input to the algorithm (HMMTOP_multi, PhDlmm, multi), the homolog/multiple sequence alignments found with BLAST using an e-value cutoff of 10^-3 were used. Column 2: Corr. topo.—Topology prediction accuracy. A topology prediction is correct if the number of predicted TM helices is accurate as well as their location with respect to the helices as defined in the experimental structures. The exact criterion is an overlap of at least five residues between a predicted and target TM helix. Further, the protein must be correctly oriented with respect to the membrane. For sequences with reentrant regions, the predictions of the reentrant regions themselves do not have to be correct, since regular topology predictors do not model these regions. But if a mispredicted reentrant region induces errors in other parts of the sequence (such as inverted loop regions), the topology is considered incorrect. Column 3: Seqs w. FN—Sequences with false negatives. The fraction of sequences for which a method misses at least one TM region (including false merges, where two real TMs are falsely predicted as one). Column 4: Seqs w. FP—Sequences with false positives. The fraction of sequences for which a method overpredicts at least one TM region (including false splits, where two real TM regions are falsely predicted as two). Column 5: Inv. topo.—Inverted topologies. The fraction of sequences for which a method predicts the TM regions correctly, but the orientation wrongly. Column 6: DR FP / all FP—The fraction of all false positive TM predictions that correspond to an actual reentrant or dip region.

2.3.4 Post-processing OCTOPUS applies two post-processing steps. First, if no TM region is predicted, any predicted reentrant or dip region is also removed. Second, to limit over-prediction of TM regions in globular domains, OCTOPUS removes TM regions predicted to be more than 60 residues from another TM region if they also have an uncertain M-preference compared to its G- and L-preferences. Specifically, if 1.35 × \(\left(\prod (M_m) - 0.61 < \prod (G_m + L_m)\right)^{1/3}\), the M-values of this TM region are set to zero and the Viterbi algorithm is run again. As representations in the predicted TM region and \(n\) the number of residues. M, G, and L are the respective structure preference values.

3 RESULTS AND DISCUSSION

3.1 Overall prediction accuracy

Table 2 shows the prediction performance of OCTOPUS on a database of 124 sequences, along with the results for a number of existing topology prediction methods. The results presented for OCTOPUS include a 10-fold, sequence-based, cross-validation. For the remaining methods, no cross-validation was performed, but since many of the structures in our dataset are newer than most methods,
it is likely that only relatively few of these sequences were present in each method’s training data.

A correctly predicted topology includes predicting the correct number and location of all TM regions, as well as the correct orientation of the protein with respect to the membrane. According to this definition, the topology prediction accuracy of OCTOPUS in this test is roughly 7% units better than the next method, which is MEMSAT3 (Jones, 2007). Common for both these methods is the strategy of using ANNs to make explicit predictions of residue preferences with sequence profiles as input in combination with an algorithm that finds the highest scoring topology based on the predicted residue preferences.

In total, five TM regions are missed by OCTOPUS. Three of these correspond to hydrophilic TM helices with hydrophobicity values (calculated using the Goldman, Engelman and Steitz scale (Engelman et al., 1986)) between −2.5 (PDB chains 1ym6B, helix 1) and 1yeWB, helices 6) and −0.7 (1txhA). In particular, the first two helices are exceptionally difficult targets, which none of the methods in our test can detect. Both contain a large fraction of charged residues and with current understanding of the mechanisms for helix insertion, it cannot be explained how these regions are inserted into the membrane. The fourth missed helix (1p59A, helix 1) has a hydrophobicity value of 0.7, but when aligned to its BLAST hits, the hydrophobicity calculated from the profile becomes much lower (<-0.2). The last missed helix (1yeWB, helix 4) is reasonably hydrophobic (1.4), and is likely missed due to a combination of shortness (16 residues) and inside/outside preference compensation (since helix 6 is also missed).

In 1p90C, OCTOPUS falsely predicts a membrane dip as a TM region, and in 1txhA, two reentrant regions are mistaken for TM helices. In 1otsA, three of its four TM hairpins are predicted as a single TM region.

The remaining error made by OCTOPUS consists of one protein (1kgB), where the orientation is inverted.

In addition to the results presented above, performance was also tested using a second independent dataset consisting of 163 sequences with known topologies. Here, OCTOPUS is also top ranked, closely followed by PRODIV-TMHMM (Viklund and Elofsson, 2004) and MEMSAT3 (Table S1). If taking into account that most of the older methods (like PRODIV-TMHMM) to some extent used these sequences in their training data, we interpret these results as qualitatively similar to those presented in Table 2.

3.2 ANNs and relative residue propensities provide higher sensitivity for TM detection

In theory, the advantage of using ANNs is that they have the ability to pick up position-specific properties, possibly unrelated to average amino acid distributions. An effect of this is that (compared to methods where the residue preference score is based on similarity with a fixed amino acid composition), both OCTOPUS and MEMSAT3 detect more of the TM regions with low hydrophobicity (hydrophobicity value <0.5 according to the GES-scale (Engelman et al., 1986)). Out of 39 such TM regions in our dataset, OCTOPUS correctly detects 92% (36) and MEMSAT3 85% (33). This can be compared to methods based on amino acid composition (TMHMM2.0 (Krogh et al., 2001), Phobius (Käll et al., 2004), PRO-TMHMM (Viklund and Elofsson, 2004) and PolyPhobius (Käll et al., 2005)), which all detect <72% of these TM regions.

Compared to the other HMM-based methods, HMMTOP, HMMTOP_multi (Tusnady and Simon, 1998) and PRODIV-TMHMM and employ a different strategy for calculating residue preferences. These methods re-estimate emission (and transition) parameters based on each target sequence. In practise, the effect of this is that, based on similarity with the original amino acid distributions for the different structural regions (i, o and M), any sequence is divided into segments corresponding to these regions, so that the similarity within and diversity between each such segment category is maximized.

This strategy is very successful with respect to finding TM helices correctly (Table 2). In particular these methods are also successful in finding TM helices with low hydrophobicity. For all three methods >87% of these helices are detected correctly. It can also be noted that all of HMMTOP, HMMTOP_multi and PRODIV-TMHMM tend to overpredict more helices than other methods. This is most likely a side-effect of these relative preference scores.

3.3 Erroneous predictions of H-, R-, and D-regions are often the cause of incorrect topologies

To achieve accurate topology predictions, sensitivity of TM detection must be complemented by avoiding overpredicting such regions. In this study we have defined three types of uncharacteristic membrane associated regions, namely TM hairpins, reentrant regions and membrane dips (Fig. 1). As can be seen in Table 2, a majority of overpredicted TM regions correspond to either reentrant or dip regions, although these regions constitute <4% of all non-TM residues in our dataset. This is not surprising since these regions are considerably more hydrophobic than an average loop region and can therefore more easily be mistaken for a TM region (Viklund et al., 2006).

In the same fashion, a clear majority of all false merges of TM regions correspond to actual TM hairpins (Supplementary Table S2). This is most likely due to that they are generally much shorter than a full TM helix.

OCTOPUS provides a first attempt both at defining and correcting this type of errors by including modelling of H-, R- and D-regions in the topological grammar (Fig. 3). For the dataset of 124 sequences, this causes one more sequence (1%) to be correctly predicted compared to using an identical grammar as in Figure 3 with the exception of removing the hairpin and reentrant/dip compartments. For the second dataset of 163 sequences, the corresponding difference is seven sequences (4%), which are predicted correctly with hairpin and reentrant/dip compartments and wrongly without them (Supplementary Table S1).

Actual detection of these special regions turned out to be difficult. Overall, OCTOPUS detects roughly 20% of the reentrant and dip regions correctly, while correctly predicting four out of seven TM hairpins as two helices instead of one (Table 3).

3.4 Sensitive methods overpredict TM regions in globular domains

An additional test of method accuracy is the ability to avoid predicting TM regions in globular proteins. This is particularly important if a method is to be useful for genome-wide studies. On a dataset of 1087 globular proteins, all methods with high
As can be seen in Table 2, methods with residue scores based on multiple sequence information and neural networks to explicitly predict residue preferences seems to be the best way thus far to detect atypical helices without including false positives (inside of TM domains).}

This type of error can be corrected by applying a preprocessing filter trained to distinguish globular from TM proteins. For example, MEMSAT3 implements this type of filter (Jones, 2007), and shows high accuracy (<3.5% FPs on our data). However, a problem remains for proteins with multiple domains, where only a subset are integral membrane domains. If a method tends to predict TM regions in globular proteins, it is likely that it will overpredict TM regions in this type of multidomain TM proteins as well. Based on the results for the globular sequences, this is a potential problem for all the methods with high sensitivity with respect to TM region detection.

OCTOPUS implements a simple post-processing filter that removes the weakest TM predictions in domains that are otherwise predicted to be globular (details in Section 2). This causes a slight improvement in prediction accuracy for TM-proteins containing globular domains (4 out of 17 sequences containing an overprediction in a globular domain get corrected), but does not significantly reduce the number of predicted TM regions in our globular data. To overcome these difficulties, more detailed modeling of multi-TM domains and single TM regions in globular domains would be a natural extension to OCTOPUS. However, we find that to be beyond the scope of this study.

### 3.5 What sequence properties determine orientation?

The implementation for predicting orientation is perhaps what differs most between different topology prediction methods, particularly with respect to what sequence information is being used. Two methods in our test, Toppred II (Claro and von Heijne, 1994) and PHDhtm (Rost et al., 1996), rely solely on the fact that positively charged residues are more common in inside loops for this task (von Heijne, 1986, 1992). Both methods simply count the number of positively charged residues in opposite loops to determine orientation.

In this perspective, the inversion error rate for these methods can be seen as a baseline for the question: How much better can orientation be predicted if more sequence information is used? As can be seen in Table 2, methods with residue scores based on average amino acid distributions generally perform no better than the simpler methods based on positive charge bias. Several of these methods even perform slightly worse [Phobius, TMHMM2.0 and MEMSAT (Jones et al., 1994)]. This can be taken as an indication that there are no obvious sequence properties that provide information aiding in this task, which at least not detectable by observing non-position-specific, average distributions.

The strategy applied in OCTOPUS for predicting orientation is based on this observation, and the hypothesis that the sequence information determining inside/outside preference should be treated separately from that determining membrane/non-membrane preference. The main motivation for this hypothesis is that while membrane/non-membrane preference is (fairly) symmetric with respect to the two sides of the membrane, inside/outside preference is not, meaning that if the two types of sequence signals are not independent, it should be beneficial to treat them separately.

Although many types of sequence information was tested, the best results were achieved when using only relative positive charge accumulation in combination with relative accumulation of polar aromatic amino acids (Tyr and Trp) as input to the neural networks. A possible explanation for this is based on two observations: first, polar aromatic residues are overrepresented in the membrane/water interface regions (Granseth et al., 2005; Killian and von Heijne, 2000), and second, the effect of positive charges for both helix insertion and orientation seems to be relative to their position with respect to the helix and the membrane (Hessa et al., 2007). Therefore, it is likely that the combined information of charge occurrence and location with respect to the membrane (estimated from the relative accumulation of Tyr and Trp in that area) can provide better predictions than using charge alone.

### 4 CONCLUSIONS

Based on the results in this and earlier studies, topology can be predicted with high accuracy. Still, no method is perfect in the sense that it can always find the correct topology. According to our analysis, erroneous topology predictions are most often due to one out of five things:

1. Atypical helices (containing kinks, many charged residues, etc.) are missed.
2. Transmembrane and other membrane dipping regions are mistaken for TM helices.
3. Two adjacent short transmembrane helices (hairpins) are falsely merged into one.
4. Transmembrane helices are overpredicted in hydrophobic parts of globular domains.
5. The orientation of a protein is inverted.

With OCTOPUS we have tried to address, in particular, the first three of these problems. Our results indicate that the use of multiple sequence information and neural networks to explicitly predict residue preferences seems to be the best way thus far to detect atypical helices without including false positives (inside of TM domains).

To avoid mispredicting uncharacteristic membrane associated regions such as reentrant-, membrane dipping- and TM hairpin-regions the topological grammar of OCTOPUS includes with compartments for modeling these regions, providing a slight improvement compared to using a generic topological model.

In addition, we see that using positive inside bias (weighted with respect to estimated closeness to the membrane/water interface) to predict orientation seems to provide at least accurate predictions as when additional sequence information is used.

<table>
<thead>
<tr>
<th>Region type</th>
<th>Total</th>
<th>TP</th>
<th>FP</th>
</tr>
</thead>
<tbody>
<tr>
<td>DiR</td>
<td>49</td>
<td>10</td>
<td>2</td>
</tr>
<tr>
<td>H</td>
<td>7</td>
<td>4</td>
<td>0</td>
</tr>
</tbody>
</table>

Since dips and reentrants are modeled by the same compartment, they are treated as one region type in this analysis. Column 1: Region type—DiR/Reentrant and Hairpin. Column 2: Total—Total number of regions in the 124-sequence dataset. Column 3: TP (True positives)—Total number of regions correctly detected. Column 4: FP (False positives)—Total number of over-predicted regions.
We also wish to acknowledge Anni Kauko for valuable contributions to this work.

Funding: This work was supported by grants from the Swedish Natural Sciences Research Council, the Wallenberg Consortium North, the Knut & Alice Wallenberg SSF (the Foundation for Strategic Research). The EU 6th Framework Program is gratefully acknowledged for support to the Embrace Contract No. LSHG-CT-2003-512092.

Conflict of Interest: none declared.

REFERENCES

Käll,L. et al. (2005) An HMM posterior decoder for sequence feature prediction that includes homology information. Bioinformatics, 21 (Suppl. 1), i251–i257.