Stockholm University

This is an accepted version of a paper published in *Proteomics*. This paper has been peer-reviewed but does not include the final publisher proof-corrections or journal pagination.

Citation for the published paper:

Tsirigos, K., Hennerdal, A., Kall, L., Elofsson, A. (2013)

"A guideline to proteome-wide alpha-helical membrane protein topology predictions"
Proteomics, 12(14): 2282-2294

Access to the published version may require subscription.

Permanent link to this version:

<http://urn.kb.se/resolve?urn=urn:nbn:se:su:diva-81726>

<http://su.diva-portal.org>

A guideline to proteome-wide α -helical membrane protein topology predictions.

Konstantinos D. Tsirigos*, Aron Hennerdal*, Lukas Käll, Arne Elofsson

September 20, 2011

*Authors contributed equally.

Stockholm Bioinformatics Center, Center for Biomembrane Research, Department of Biochemistry and Biophysics, Science for Life Laboratory, Swedish E-science Research Center, Stockholm University, SE-106 91 Stockholm, Sweden Email: arne@bioinfo.se Fax: +46-8-15 3679 Phone: +46-8-16 4672

Keywords: membrane proteins, genome analysis, topology predictors, membrane, α -helical

Abstract

For current state-of-the-art methods, the prediction of correct topology of membrane proteins has been reported to be above 80%. However, this performance has only been observed in small and possibly biased datasets obtained from protein structures or biochemical assays. Here, we add four “genome-scale” datasets to the arsenal of benchmark sets, including one recent large set of experimentally validated human membrane proteins with glycosylated sites. The set of glycosylated proteins is also used to examine the ability of prediction methods to separate membrane from non-membrane proteins. The results show that methods utilizing multiple sequence alignments are overall superior to methods that do not. The best performance is obtained by TOPCONS, a consensus method which combines several of the other prediction methods and that the best methods to distinguish membrane from non-membrane proteins belong to the Phobius family of predictors. Further, we observe that the reported high accuracies of up to 90% observed in the smaller benchmark sets are not maintained in larger scale benchmarks. In particular, the performance of the best prediction methods for eukaryotic membrane proteins is most likely between 60% and 70%. The low agreement between predictions from different methods puts earlier estimates about the global properties of the membrane proteome into doubt. In the end, we suggest a pipeline to estimate these properties using a combination of the best predictors that could be applied in large-scale proteomics studies of membrane proteins.

1 Introduction

Integral membrane proteins constitute an important subset of all proteins in a proteome. Their roles vary from cellular recognition and transportation of molecules to secretion, signal transduction or enzymatic activity. They often act as receptors that control pathways interior to the cell, due to this many drugs target these proteins. They can be divided into two structural categories, α -helical and β -barrels. α -helical membrane proteins can be found in both prokaryotic and eukaryotic cells and can be classified according to the number of membrane segments and the location of the N-termini. Both single-spanning proteins, with large extra- or intra-cellular domains, and multi-spanning proteins, with short loops protruding from the membrane, exist. Their membrane-spanning segments form α -helices that consist mostly of hydrophobic residues [1]. In contrast to that, β -barrel outer membrane proteins are exclusively found in Gram-negative bacteria, mitochondria and chloroplasts [2–4]. Here, we focus on the more prevalent class of α -helical proteins.

For many years, there has been a growing effort from experimentalists to study the structure and topology of membrane proteins in order to better understand their biological functions and the mechanisms that guide their topogenesis in the membrane. The fundamental problems in that effort derive from the hydrophobic nature of these proteins, which makes them more difficult to work with in comparison to soluble proteins. This fact is reflected in the Protein Data Bank [5], where less than 1% of all deposited structures represent membrane proteins.

The first approaches towards the prediction of transmembrane (TM) segments date back almost 20 years [6, 7]. The identification of a hydrophobic stretch of about 15-25 residues long is usually enough to produce a successful prediction of putative TM segments. If this is also combined with the “positive inside rule” [8], the basic core of a prediction method is formulated. These rules (hydrophobicity, “positive inside rule”) apply to virtually all biological membranes, including mitochondrial [9] and chloroplast [10], as long as “inside” is defined to be the cytoplasm. In principle, all prediction algorithms need to find an optimal balance between the hydrophobicity and the “positive inside rule”, and it has been shown that it is possible to obtain state-of-the-art

performance by solely optimizing the weights between these two factors [11]. However, on small benchmark sets it is also possible to obtain similar performance for a large set of quite different weights.

Many computational methods based on various techniques, ranging from empirical algorithms using sliding windows across the sequence [12] and dynamic programming [13] to more advanced machine learning techniques, such as Artificial Neural Networks (ANNs) [14, 15], Support Vector Machines (SVMs) [16], Hidden Markov Models (HMMs) [11, 17–23] and more generic Dynamic Bayesian Networks (DBNs) [24], have been developed. Methods that combine several machine learning techniques have also been presented [25, 26]. Many of the algorithms mentioned above make use of evolutionary information by introducing multiple sequence alignments (MSAs) in the prediction procedure [11, 15, 20, 22], while others allow the incorporation of prior knowledge concerning the topology of certain regions of the sequence to be examined (constrained prediction) [11, 17, 19–21, 25, 27]. Recently, consensus algorithms that combine the outputs from different predictors have also been made available to the public [27–29].

One difficulty for the correct prediction of the topology (i.e. the number and location of the TM segments) of membrane proteins is the similarities between signal peptides and the N-terminal TM regions. The hydrophobic core of the signal peptide can be wrongly assigned as a putative first TM segment or vice versa [30, 31]. This fact is often overlooked when testing prediction methods on small datasets, but becomes a major problem in proteomic analyses since here it is crucial not to confuse cleaved signal peptides and TM regions. To tackle this difficulty, a few predictors contain sub-models for both signal peptides and TM segments [16, 21–24, 26]. Another approach commonly followed is to scan the sequence with an independent signal peptide predictor and consequently remove the signal peptide (if predicted) or constrain the N-terminal of the protein to be in the extracellular space before predicting the topology.

The reported performance ranges from about 70% for older methods to over 80%, for the latest methods, in benchmark data either from topologies classified by biochemical assays or from proteins of known structure. The ability to separate between TM and non-TM proteins has been

reported to be over 99% [31]. However, it has not been shown that these high accuracies can be maintained on a genomic scale, as the benchmark sets are small and the inclusion of proteins is certainly not without bias [23]. Anyhow, these methods have been applied to genomes and have estimated the fraction of TM proteins to be between 20-30% [32]. Further, a preference for an even number of TM regions with both the N and C-termini on the inside has been predicted [31]. It has also been reported that some methods perform better on prokaryotic than eukaryotic proteins [15].

In a recent study, a draft of the predicted human membrane proteome based on the results from various membrane prediction algorithms was generated [33]. In this work, it was shown that the number of TM regions is predicted to be identical for all methods only for 12% of the predicted membrane proteins, indicating that the topology accuracies obtained on small test sets are not necessarily transferable to complete genomes.

A key post-translational modification of membrane proteins is the N-linked glycosylation. This involves the covalent addition of oligosaccharides to Asn residues, in the sequence motif Asn-X-Thr/Ser and is a common feature of both eukaryotic and archaeal membrane proteins. The aforementioned pattern is usually found on the biggest loop of the protein, which is exposed at the ER lumen [34, 35], or in other loops as well. Given that the modification of glycosylation sites depends on their subcellular location, this characteristic can be used in studies of membrane protein topology. The presence of a glycosylation site indicates that the respective amino acid residue is located in the periplasmic space. Recently, a large-scale effort to identify glycosylated sites in membrane proteins has been published [36]. Here, for the first time, to our knowledge, this dataset is used to evaluate the performance of the different predictors.

By using a number of topology predictors and testing them both on traditional topology benchmark sets, and on four significantly larger, but less detailed datasets, we try to improve our understanding of how the state-of-the-art topology prediction methods would perform on whole genomes. In addition to the glycosylation dataset, we use two earlier datasets of experimentally verified C-terminal locations [37, 38] and a set of GPCRs of presumed identical topology.

Finally we use this set of identified membrane proteins from the glycosylation set to examine

the capacity of all predictors to accurately separate membrane from cytosolic and extracellular proteins.

2 Materials and Methods

2.1 Topology prediction methods

The full list of the software tools for topology prediction used in this study is shown in Table 1. In addition to the fundamental differences in the technical aspects of the methodologies, they can be divided into three classes: 1) methods that have a separate module for signal peptide prediction, 2) methods that use multiple sequence alignments, and 3) methods that allow incorporation of prior topological information of the sequences (i.e. constrained predictions). Many different datasets have been used for the training of these methods, but a general trend is that older methods have been trained on data from proteins whose topology had been determined by biochemical assays (low resolution experiments, noted by “lowres” in Table 1), while the newer methods tend to use data from 3D-structures of membrane proteins (“highres”).

2.2 Benchmark datasets

In Table 2 we have listed some important facts relating to the benchmark datasets used in this study. Notable is that the fraction of proteins predicted to have at least one (two) TM regions in average from all predictors varies between at least 42% (21%) and at most 99% (99%)

2.2.1 Full topology datasets

The first dataset (highres) consists of 124 protein chains with known three-dimensional structures, originating from the Orientations of Proteins in Membranes (OPM) database [39] and was used in an earlier study [25]. In this case, the topologies were assigned based on the coordinates of the $C\alpha$ atoms. The second dataset (lowres) contains 163 proteins with known topology that was

compiled by combining the datasets used to evaluate TOPCONS [27]. Although different datasets are used for training of the different predictors, the overlap with either the lowres or highres sets is substantial for all methods.

2.2.2 Datasets from large-scale experimental studies

Two more datasets, referred to as E.coli-cterm and yeast-cterm in Table 2, consist of sequences with experimentally determined C-terminal locations from membrane proteins in *Escherichia coli* [37] (613 sequences) and *Saccharomyces cerevisiae* [38] (546 sequences), respectively. In these experimental studies, only proteins where TMHMM had predicted two or more TM regions were included.

Finally, a novel dataset, which to our knowledge is introduced for the first time in this type of benchmarking studies, was used. This set consists of proteins with experimentally verified cell surface N-linked glycosylation sites that were identified by Wollscheid et al. [36]. In this work, the authors implemented cell surface capturing (CSC) technology, which enables comprehensive and quantitative analysis of the cell surface glycoproteins at very high specificity through the detection of tryptic peptides. These regions are located in the extracellular domains of the glycoproteins. In this dataset (glyco) the total number of unique protein sequences was 787 with more than 2000 determined glycosylation sites. The glyco dataset is quite different from the earlier datasets not only regarding its size, but also because it was obtained from a higher eukaryote.

2.2.3 GPCR dataset

From the study of Fagerberg et al [33], it was noted that there is a relative agreement regarding the number of GPCRs that are expected in the human proteome. Therefore, we created an additional test set, consisting of 1405 proteins from the Uniprot database [40] using two criteria: first the inclusion of “G-protein coupled receptor” in the keyword (KW) line of the protein entry and second the requirement of having a significant hit on at least one of the 36 Pfam family members of the GPCR_A clan (CL0192: G protein-coupled receptor-like superfamily) of the Pfam [41] database

(GPCR dataset).

2.2.4 Subcellular localizations dataset

We also analyzed two datasets of proteins with known subcellular localization, along with the datasets we have for TM proteins. The first contains proteins that are targeted to the cytoplasm (the “cyto” dataset), while the second contains secreted proteins (that are targeted to the extracellular space; the “extra” dataset). In order to create these datasets we relied on the most recent version of Uniprot (2011_8) and, particularly, the Swissprot database, which is manually annotated and reviewed. We retrieved all entries which listed “cytoplasm” and “secreted” respectively under the “subcellular location” subsection of the “General annotation (Comments)” section that provides information on the location and the topology of the mature protein in the cell. Initially, for the cyto dataset we came up with 13879 and for the extra dataset with 10234 protein sequences. We discarded sequences with less than 20 amino acids and filtered the remaining ones using Cd-hit [42] at a 30% identity cutoff, leaving us with 1925 and 2024 sequences, respectively.

2.3 Performance measurements

For the full-topology datasets (highres and lowres), the accuracy of the methods was measured for completely correct topologies as done in several earlier studies [20, 21, 31]. The overall topology prediction accuracy assessment was based on correct prediction of three components: the N- and C-terminal location, the number of TM helices and the TM helix locations based on an overlap of at least 5 residues with boundary definitions. For the E.coli-cterm and yeast-cterm datasets, we evaluated only the fraction of correctly located C-termini, since the full topology was not available.

In the glyco dataset only “outside” positions identified by Wollscheid et al [36] are known, so a different strategy for measuring the performance was needed. In practical terms, we required the glycosylation sites reported in this paper to be assigned to the outer loops of the sequence for them to be regarded as correctly predicted. However, the above step raises a question regarding the rate of false positives occurring purely by chance. A hypothetical predictor outputting 100% “o”

for a given protein sequence, would yield 100% accuracy, which is obviously incorrect. In order to address this issue, we performed a permutation test by randomly assigning the glycosylation points in the initial sequences. In this way, we obtained a representative set of proteins to estimate the efficiency of the methods if they were given a random sequence. The ranking of the predictors in that case was done based on the Z-scores they produced. The Z (standard) score indicates how many standard deviations an observation is above or below the mean. It is calculated by subtracting the population mean from an individual raw score and then dividing the difference by the population standard deviation.

$$Z = \frac{x - \mu}{\sigma},$$

where x is a raw score to be standardized, μ is the mean of the population and σ is the standard deviation of the population.

For the GPCR dataset, we measured the fraction of proteins with a correctly identified number of TM helices (7) and N-terminal (out). To do this, we added two variants of assessing prediction performance. For the methods that predict signal peptides, the prediction was also regarded as correct if it contained 7 TM helices and a predicted signal peptide. In addition to that, for all predictors, a prediction output with 8 TM helices and the N-terminal on the inside was also counted as correct. This was done on terms of “equal treatment” of the predictors, since some of them predict signal peptides and there therefore is a high probability that the first predicted TM region by other methods could be a signal peptide, which would mean that the topology would be shifted and the N-terminal would be set to “out”.

2.3.1 Comparison of topology predictions

The predictions from different predictors can be compared in several different ways. The predicted topology for each individual residue or protein can be compared, or alternatively, the distribution of topologies predicted in a dataset. For per-protein (and per-residue) comparisons, the fraction of

predictions that agreed was used to compare similarities between predictors and over datasets (Table 5). In addition to that, a number of criteria were calculated and used to evaluate the predictors. These include: the fraction of single- and multi-spanning proteins and the average number of TM regions, (Table 2).

As an estimation of the fraction of non-TM (globular) proteins, we calculated the fraction of them in the datasets, but also the fraction of single-spanning proteins with a predicted signal peptide, using an independent signal peptide predictor. The second approach was carried out because the topology predictors aim mostly at the detection of the boundaries of the TM segments rather than defining the topology of globular proteins. Therefore, when a predictor did not categorize a protein as TM, if there was a signal peptide predicted, the label of this particular protein was set to “o”, while in the opposite case it was set to “i”.

2.4 Constrained predictions and filters

The implementation of an independent signal peptide predictor for scanning protein sequences in order to detect putative signal peptides prior to the topology prediction is a common approach that is followed in many cases. We have used SignalP 3.0 [43], one of the currently most used methods for the prediction of signal peptides. The results of this run can be used in the methods allowing constrained predictions, which means using this information as a filter before running the topology predictor. For instance, if there is a predicted signal peptide with a length of 20 amino acids, we would constrain the first 20 amino acids of the sequence to be in the extracellular space (“o” label).

Another crucial feature of an algorithm is the ability to avoid predicting TM regions in globular proteins. This is particularly important if a method is applied on genome-wide studies. A simple but effective solution is the employment of a pre-filter specifically to decide whether the target protein is likely to contain TM segments or not. If the respective protein is likely to be globular the algorithm is not applied. Such filters exist for MEMSAT3, MEMSAT-SVM, SCAMPI and SCAMPI-single.

3 Results

In this study we examine 18 topology predictors, see Table 1 for an overview. Of these predictors, 10 use multiple sequence alignments and five contain a specific module to handle signal peptides. In addition, nine methods can use the prediction from a specific signal peptide predictor, such as SignalP, and constrain the prediction depending on its output. Further, a few methods apply a filter to distinguish between TM and non-TM proteins before the topology is determined. Most of the older predictors were trained mainly on “low-resolution” data, but several of the newer ones were trained exclusively on topologies from crystal structures of membrane proteins.

We have used six different datasets to evaluate the performance of the 18 methods, see Table 2. Two sets contain experimentally verified topologies from membrane proteins; one high-resolution set (highres), where the topologies come from crystal structures, and one low-resolution set (lowres), where the topologies come from biochemical assays. In another two sets (E.coli-term, yeast-term), only the C-terminal locations are known for the proteins. The fifth, glycosylation set (glyco), contains information about one or a few residues that are located on the extracellular side, while in the final set of G-protein Coupled Receptors (GPCRs), it is known that all proteins have their N-terminal on the outside of the membrane, and that the proteins contain seven TM helices. However, GPCRs might also contain a cleaved signal peptide.

3.1 Prediction performance on datasets with known topologies

The first step of our analysis was to test the topology predictors on the datasets of experimentally verified topologies, highres and lowres. The results are shown in Figure 1A; for detailed results please refer to the Supplementary information. In general, the performance is relatively high in both sets, with over 80% correct predictions for the best predictors and an overall average of about 75% for all predictors. This is in general in agreement with the performances reported during the development of the predictors. There is a tendency for methods performing better on one of the sets to perform better also on the other set. However, there is also a trend that methods trained

on high-resolution topologies, with labels in red in Figure 1A, perform relatively better on the highres dataset than on lowres. The reverse also applies, i.e. that the methods trained on low-resolution topologies perform better on the lowres dataset. A further observation is that methods using multiple sequence alignments, with labels in bold, perform better than the methods that do not, as seen in the development of PRO and PRODIV [20]. The best performance on both highres and lowres is obtained by TOPCONS with 89% and 86% correctly predicted topologies, respectively.

3.2 Datasets of C-terminal location

Next, we turn to the larger datasets restricted to information on C-terminal location. The prediction accuracies achieved are slightly higher in E.coli-cterm than in yeast-cterm, see Figure 1B and Supplementary information, indicating that it might be easier to correctly predict topologies among prokaryotic membrane proteins, as suggested before [15]. In general, the methods that perform well on the topology datasets (highres and lowres), i.e. methods using multiple sequence alignments, perform well also here. An exception is MEMSAT-SVM (72% on E.coli-cterm, 73% on yeast-cterm), which in our hands never performs very well. Another is MEMSAT3 on yeast-cterm (75%), where it (only) performs on par with the best single-sequence methods. The best performance is obtained by TOPCONS and PRODIV on the yeast-cterm set (87%) and by SCAMPI, SPOCTOPUS and OCTOPUS on the E.coli-cterm set (92%).

3.3 The glycosylation dataset

For the glycosylation dataset (glyco), the performance is evaluated on the grounds of how many of the identified glycosylation sites are predicted to be on the outside. However, the expected fraction of correctly identified glycosylation points is strongly correlated with the fraction of residues predicted to be on the outside, which complicates the evaluation. Figure 2 shows the correlation between the “o”-bias of the predictors and their performance. We can see that methods that are

aimed at predicting both the presence of a signal peptide and the TM regions (Phobius, PolyPhobius, Philius, SPOCTOPUS) together with two older methods (HMMTOP, TMHMM) predict the most glycosylation sites correctly, as well as the most residues on the outside.

In an attempt to estimate the performance of different predictors using this dataset, a line was fitted to the fraction of positions predicted to be on the outside versus the fraction of correctly identified glycosylation points. It can be seen that all methods perform between 10 and 20 percentage points better than random in identifying glycosylation points. This highlights the surprisingly large differences in behavior of predictors recently reported by Fagerberg *et al.* [33].

Anyhow, in an attempt to rank the predictors, we decided to use the improvement over the expected number of correctly predicted residues given a certain “o” bias, as a measure. See Materials and Methods for a detailed description. The Z-scores are shown in the inset in Figure 2. The best performance using this measure is obtained by TOPCONS, SCAMPI and PolyPhobius with Z-scores of 24. In contrast to the other sets, a few methods using multiple sequence alignments, e.g. OCTOPUS, perform worse than several single-sequence methods. In particular, the single-sequence methods Phobius and Philius perform quite well here (Z-scores 21 and 22) indicating that accurate prediction of signal peptides is more important in this dataset than in the other sets.

3.4 GPCRs

We also collected a dataset containing the human GPCRs in order to further evaluate the performance of the predictors. The topology of GPCRs can be challenging to predict, as several of the TM helices can be uncharacteristically hydrophilic [44]. In addition, some GPCR sequences contain a cleaved signal peptide while others do not. The results of the benchmark are summarized in Table 3. Clearly, the best performance obtained for GPCRs is that of PolyPhobius, which predicts 85% of the proteins to be 7TM-*Nout*, and of these, 7% are predicted to contain a signal peptide. This performance is followed by HMMTOP (77% 7TM-*Nout*) and PRODIV (76% 7TM-*Nout*). If we accept that a cleaved signal peptide might be incorrectly classified as a TM helix and relax the criterion for correct topology accordingly, i.e. also accept 8TM-*Nin* predictions to be correct, the

same three predictors remain on top.

When comparing with the tests previously described, the list of best performing methods is relatively unperturbed, with the exception of PRO that is ranked lower when using the looser definition of correctness. Notable is that methods such as Phobius, HMMTOP and in particular PolyPhobius, perform much better in this set (and also in the glyco set) than indicated in the topology datasets. An interesting observation showing how much multiple sequence alignments help is the performance of SCAMPI-single (37% correct) versus that of SCAMPI (73%). However, some single-sequence methods such as HMMTOP (81%) and Phobius (76%) perform very well.

3.5 Using constrained predictions

Next, we examined the effect that constraining prediction with a separate predictor of signal peptides has. It should be noted that in all sets except glyco, relatively few proteins actually contain a signal peptide. SignalP predicts a signal peptide in less than a third of the proteins in the E.coli-term and yeast-term datasets, while in the glyco dataset, a signal peptide is predicted in 70% of the proteins. In the two topology datasets (highres and lowres), it is clear that the inclusion of constraints through SignalP decreases the prediction accuracy significantly, see Supplementary information. Only in yeast-term is a similar prediction accuracy obtained as for unconstrained predictions, see Figure 3.

On the other hand, when we examine the performance of predictors with an incorporated module for signal peptide prediction, we see that they in general perform well. This indicates that SignalP is not ideal to separate between signal peptides and TM regions and therefore constraining the predictions does not help the overall topology prediction capability.

In the glyco dataset, where the fraction of predicted signal peptides is twice (or more) as big as in the other datasets, constrained predictions do improve the prediction performance of “outside” points drastically. This is not surprising, since there is a strong correlation between the fraction of correctly identified glycosylation points and the % “o”-bias. Therefore, when constraining the N-terminal to be “outside”, the final predicted topology tends to contain more residues predicted

as “outside”, with a corresponding increase in performance. When compensating for the “o”-bias using Z-scores as described before, it is apparent that the prediction performance is the same or even worse also for the glyco set, see Supplementary information.

4 Discussion

In Table 4, we summarize the performance of topology predictors over the six different datasets. For all sets, a Z-score based on the average performance of all predictors on a set is calculated for each method. All Z-scores higher than 1 are marked in bold. The last two columns show the average rank and average Z-score of the different prediction methods. From this analysis, the general trend is that methods using multiple sequence alignments are better than methods that do not, since eight (of nine) multiple sequence alignments methods are on top. The best performance is obtained by TOPCONS (Z-score of 1.33), a consensus method utilizing the predictions from five of the other methods (SCAMPI, SCAMPI-single, PRO, PRODIV, OCTOPUS). Actually, TOPCONS is on top in four out of six tests. Surprisingly, the second best method is PRODIV, a method developed seven years ago. The best performing single-sequence method is Philius, followed by TOPCONS-single.

4.1 Larger disagreements on predictions “in the wild”

The large variations in predictions that we observed in the glyco set led us to the assumption that there must be differences in the way the prediction methods treat datasets that differ from traditional small-scale sets. We therefore analyzed our results on the glyco set in more detail to assess possible similarities and sources of disagreement. Table 2 shows the main properties of the dataset together with those of the other benchmark sets used.

As a measurement of the heterogeneousness in predictions, we devised a simple scheme for comparing the predictions within each dataset. For each protein in a dataset, we went through all pair-wise combinations of prediction methods and noted whether the two methods in the pair agreed in their topology prediction or not. We used a definition of “agreement” analogous to the

definition of a correct prediction, namely that the predicted TM regions match with an allowed offset of no more than five residues and that the orientation is the same. In Table 5, the “Agreeing, all” column lists the measurement just described, i.e. the fraction of all pairs of prediction comparisons that agree. “Agreeing, top 6” gives the same measurement, but limited to include the methods PRODIV, SCAMPI, SPOCTOPUS, PolyPhobius, MEMSAT3, and PRO. This list of six methods contains the top eight predictors from the ranking procedure, see Table 4, minus TOPCONS, which is a consensus of other methods, and OCTOPUS, which is closely related to SPOCTOPUS. The “Consensus, all” column contains the fraction of proteins predicted equally by all methods, and “Consensus, top 6” is the fraction of proteins where PRODIV, SCAMPI, SPOCTOPUS, PolyPhobius, MEMSAT3, and PRO predicted equally.

Interestingly, we can not confirm the result showing larger disagreements for single-spanning than for multi-spanning proteins reported by Fagerberg *et al.* [33]. When grouping proteins in the glyco set according to their averaged number of predicted TM segments, we see a pairwise agreement of 33% for single-spanning proteins (average between 0.5 and 1.5 predicted TM segments) and 32% for multi-spanning proteins (average of 1.5 TM segments or above).

From the agreement between predictors it is possible to make an estimate of the overall accuracy in prediction on the dataset, assuming that the errors are randomly distributed. Although this assumption is most likely incorrect, it could be used as an upper limit estimate on the average accuracies. Using the “Consensus, all” column of Table 5, this is calculated by $\sqrt[18]{c_{all}}$, where c_{all} is the decimal fraction of agreeing predictions. This yields values for the expected accuracy of 77% on the glyco set and 90% on the highres and lowres sets, noted in parenthesis in Table 5. Similarly, a better estimate of the expected accuracy could be calculated from the “Consensus, top 6” column as $\sqrt[6]{c_6}$. Here, the numbers are 66% for the glyco set, 86% for highres, and 89% for lowres. For the two “Agreeing” columns, the comparisons are pairwise and the expression used is $\sqrt[2]{a}$, where a is the agreement. All expected accuracies are found within parenthesis in Table 5.

In summary, these numbers agree quite well with the performance of the best predictors found in the smaller benchmark sets. Similar numbers (78-89%) are found in the E.coli-cterm set in-

dicating that the performance of 80-90% correct topologies for the top predictors might not be unrealistic to see on prokaryotic genomes. However, for the eukaryotic sets yeast-cterm (65-82%) and in particular glyco (60-77%) the agreement is lower than would be expected from predictors with an accuracy of over 80%. It seems therefore likely that, as proposed before [15], it is more difficult to predict the topology of eukaryotic membrane proteins. The average agreement decreased with increasing length of the proteins, while there was no such trend observed for proteins with few or many TM regions, see Supplementary information.

4.2 How to detect the membrane proteome

When predicting the topology of proteins in complete genomes as opposed to in controlled datasets, an additional difficulty arises. In a genome, it is important to be able to sort out soluble proteins and only predict TM segments in actual membrane proteins. To address this, we assembled two datasets of mammalian globular proteins, one containing proteins located in the cytoplasm (cyto), and one containing proteins exported to the extracellular space (extra), see Table 2.

In Figure 4, we show the number of TM regions predicted by the 18 methods in the two sets of globular (non-TM) proteins, and also in the glyco set as comparison. Proteins in the glyco set should contain at least one glycosylated residue and therefore be located outside the cell. They should also be bound to the membrane. However, it is possible that they are just attached to the membrane and not true TM proteins. Anyhow, the vast majority of these proteins should be TM. From the results in the cyto and extra we can see that the “Phobius”-group of predictors (PolyPhobius, Phobius, Philius) are superior to all other methods in distinguishing between membrane and non-membrane proteins. The other two methods with specific modules for signal peptides (SPOCTOPUS, MEMSAT-SVM) are not performing as well. When tracing the predictions, it is clear that the reason for this is the way the signal peptide predictions are implemented in SPOCTOPUS and MEMSAT-SVM. In MEMSAT-SVM it is actually impossible to predict a signal peptide and no TM helix. In SPOCTOPUS, the reason is that when it does not predict a signal peptide, it uses the OCTOPUS algorithm and OCTOPUS clearly overpredicts TM helices in non-membrane proteins.

The only method that performs almost on par with the “Phobius”-group is SCAMPI when using SignalP-constrained predictions, Figure 4d. Worth noting in the case of SCAMPI (and MEMSAT3), is that it uses a filter to distinguish between TM and non-TM proteins. This makes it possible to use the observation that the first TM helix needs to be more hydrophobic than later helices [45]. In the extra set, out of 2024 proteins in total, the globular filter used in SCAMPI predicts 1073 of them to be non-TM proteins (53%), while the numbers for MEMSAT3, MEMSAT-SVM and SCAMPI-single filters are 1214 (60%), 1619 (80%) and 1275 (63%) respectively. Corresponding numbers for filtering the cyto set (1925 proteins) are 1867 (97%) by SCAMPI, 1829 (95%) by MEMSAT3, 1617 (84%) by MEMSAT-SVM and 1810 (94%) by SCAMPI-single. The importance of the filter can further be exemplified with TOPCONS, whose MCC value for distinguishing TM and cytoplasmic proteins increases from 0.61 to 0.85 when the SCAMPI filter is applied.

An interesting finding is that constraining does to some extent help the separation between TM and non-TM proteins. This is particularly evident in the extra dataset, where more than half of the proteins are predicted to have a signal peptide. From Figure 5 we can see that the methods that allow constrained predictions (with labels in *italics*) have much higher MCC values when the constraints are applied, for instance the MCC for TOPCONS increases from 0.36 to 0.85. However, it still does not perform as well as Phobius *et al.*

5 Conclusions

We have applied 18 topology prediction methods on α -helical membrane proteins in six datasets. The datasets contain information on the full topology (the highres and lowres sets), the location of the C-termini (E.coli-cterm, yeast-cterm), and the location of a number of glycosylation sites (glyco). The sixth dataset contains members of the GPCR family of proteins, known to have 7 TM regions and their N-termini on the outside.

Of the six datasets, the glyco set is to our knowledge new in these kinds of benchmarks. We find that simply using the number of glycosylation points correctly predicted as “outside” biases

the results to favor methods predicting a large fraction of points as “outside”. Instead we describe a better-than-random approach, which we use in the ranking of prediction methods. Further, we note that the results on this set show large variations, which we have quantified and use to argue that they indicate an overall worse performance on this set. Similar estimates over all six datasets indicate that previous reports of topology prediction accuracies of up to 90% is most probably not transferrable to larger, proteome-wide datasets.

We developed a ranking scheme for the 18 predictors, based on the performance on the six datasets measured by Z-scores. The method with the overall best performance was found to be TOPCONS. In general, methods using evolutionary information in the form of multiple sequence alignments aggregate at the top of the ranking in all six datasets.

We have also attempted the commonly used practice of employing a separate detection method for signal peptides to use as input in the topology prediction. Using SignalP, we note a performance loss in our benchmark datasets for prediction methods capable of incorporating this kind of information. We do, however, see that methods implementing their own signal peptide detection module fare comparably well, in particular on the glyco set.

To predict topology in whole genome studies, the ability to separate non-TM (globular) proteins from TM proteins is of particular importance. We tested the 18 methods on their ability to correctly identify cytoplasmic and extracellular proteins as non-TM proteins, and found that three related methods, Phobius, PolyPhobius and Philius, do this particularly well. All three incorporate signal peptide detection, but outperform the other methods also when signal peptides identified by SignalP are used to constrain predictions.

In conclusion, a recommended strategy for the topology prediction of proteins in whole genomes would include using the most recent method of the “Phobius”-group, Philius, to filter non-membrane proteins, followed by TOPCONS for the topology prediction of the remaining membrane proteins.

6 Acknowledgement

We thank Dr. Bernd Wollscheid for the glycosylation dataset.

This work was supported by grants from the Swedish Research Council (VR-NT 2009-5072, VR-M 2010-3555), SSF (the Foundation for Strategic Research) , the EU 7th Framework Program by support to the EDICT project, contract No: FP7-HEALTH-F4-2007-201924.

References

- [1] von Heijne, G. Recent advances in the understanding of membrane protein assembly and structure. *Quart.Rev.Biophys* 2000, 285–307.
- [2] Cavalier-Smith, T. Membrane heredity and early chloroplast evolution. *Trends Plant Sci.* 2000, 5:174–82.
- [3] Elofsson, A. & von Heijne, G. Membrane protein structure: prediction versus reality. *Annu Rev Biochem* 2007, 76:125–140.
- [4] Gray, M. W., Burger, G. & Lang, B. F. Mitochondrial evolution. *Science* 1999, 283:1476–81.
- [5] Westbrook, J., Feng, Z., Jain, S., Bhat, T. *et al.* The Protein Data Bank: unifying the archive. *Nucleic Acids Res* 2002, 30:245–248.
- [6] Eisenberg, D., Weiss, R. M. & Terwilliger, T. C. The hydrophobic moment detects periodicity in protein hydrophobicity. *Proc. Natl. Acad. Sci. USA.* 1984, 81:140–4.
- [7] Kyte, J. & Doolittle, R. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* 1982, 157:105–132.
- [8] von Heijne, G. Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule. *J. Mol. Biol.* 1992, 225:487–494.
- [9] Rojo, E. E., Guiard, B., Neupert, W. & Stuart, R. A. N-terminal tail export from the mitochondrial matrix. Adherence to the prokaryotic "positive-inside" rule of membrane protein topology. *J. Biol. Chem.* 1999, 274:19617–22.
- [10] Houben, E., JW, D. G. & KJ, V. W. Insertion of leader peptidase into the thylakoid membrane during synthesis in a chloroplast translation system. *Plant Cell.* 1999, 11:1553–64.
- [11] Bernsel, A., Viklund, H., Falk, J., Lindahl, E. *et al.* Prediction of membrane-protein topology from first principles. *Proc Natl Acad Sci U S A* 2008, 105:7177–7181.
- [12] Claros, M. & von Heijne, G. TopPred II: an improved software for membrane protein structure predictions. *Comput Appl Biosci* 1994, 10:685–686.
- [13] Jones, D., Taylor, W. & Thornton, J. A model recognition Approach to the Prediction of All-Helical Membrane Protein Structure and Topology. *Biochemistry* 1994, 33:3038–3049.
- [14] Rost, B., Fariselli, P. & Casadio, R. Topology prediction for helical transmembrane proteins at 86% accuracy. *Protein Sci.* 1996, 5:1704–1718.
- [15] Jones, D. Improving the accuracy of transmembrane protein topology prediction using evolutionary information. *Bioinformatics* 2007, 23:538–544.

- [16] Nugent, T. & Jones, D. T. Transmembrane protein topology prediction using support vector machines. *BMC bioinformatics* 2009, 10:159.
- [17] Tusnady, G. & Simon, I. The HMMTOP transmembrane topology prediction server. *Bioinformatics*. 2001, 17:849–850.
- [18] Sonnhammer, E., von Heijne, G. & Krogh, A. A hidden Markov model for predicting transmembrane helices in protein sequences. In *Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology*. Edited by Glasgow J, Littlejohn T, Major F, Lathop R, Sankoff D, Sensen C, Menlo Park, CA: AAAI Press 1998:175–182.
- [19] Bagos, P. G., Liakopoulos, T. D. & Hamodrakas, S. J. Algorithms for incorporating prior topological information in HMMs: application to transmembrane proteins. *BMC bioinformatics* 2006, 7:189.
- [20] Viklund, H. & Elofsson, A. Best alpha-helical transmembrane protein topology predictions are achieved using hidden Markov models and evolutionary information. *Protein Sci.* 2004, 13:1908–1917.
- [21] Kall, L., Krogh, A. & Sonnhammer, E. A combined transmembrane topology and signal peptide prediction method. *J Mol Biol* 2004, 338:1027–1036.
- [22] Kall, L., Krogh, A. & Sonnhammer, E. An HMM posterior decoder for sequence feature prediction that includes homology information. *Bioinformatics* 2005, 21 Suppl 1:i251–7.
- [23] Kall, L., Krogh, A. & Sonnhammer, E. L. L. Advantages of combined transmembrane topology and signal peptide prediction—the Phobius web server. *Nucleic Acids Res.* 2007, 35:W429–32.
- [24] Reynolds, S. M., Kall, L., Riffle, M. E., Bilmes, J. A. & Noble, W. S. Transmembrane topology and signal peptide prediction using dynamic bayesian networks. *PLoS. Comput Biol.* 2008, 4:e1000213.
- [25] Viklund, H. & Elofsson, A. OCTOPUS:Improving topology prediction by two-track ANN-based preference scores and an extended topological grammar. *Bioinformatics* 2008, 24:1662–1668.
- [26] Viklund, H., Bernsel, A., Skwark, M. & Elofsson, A. SPOCTOPUS: a combined predictor of signal peptides and membrane protein topology. *Bioinformatics* 2008, 24:2928–2929.
- [27] Bernsel, A., Viklund, H., Hennerdal, A. & Elofsson, A. TOPCONS: consensus prediction of membrane protein topology. *Nucleic Acids Res.* 2009, 37:W465–8.
- [28] Arai, M., Mitsuke, H., Ikeda, M., Xia, J. *et al.* ConPred II: a consensus prediction method for obtaining transmembrane topology models with high reliability. *Nucleic Acids Res.* 2004, 32:W390–3.
- [29] Klammer, M., Messina, D. N., Schmitt, T. & Sonnhammer, E. L. L. MetaTM - a consensus method for transmembrane protein topology prediction. *BMC bioinformatics* 2009, 10:314.
- [30] Lao, D. M., Arai, M., Ikeda, M. & Shimizu, T. The presence of signal peptide significantly affects transmembrane topology prediction. *Bioinformatics* 2002, 18:1562–6.
- [31] Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* 2001, 305:567–580.
- [32] Wallin, E. & von Heijne, G. Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. *Protein Sci.* 1998, 7:1029–1038.
- [33] Fagerberg, L., Jonasson, K., Heijne, G. V., Uhlen, M. & Berglund, L. Prediction of the human membrane proteome. *Proteomics* 2010, 10:1141–9.

- [34] Landolt-Marticorena, C. & Reithmeier, R. A. Asparagine-linked oligosaccharides are localized to single extracytosolic segments in multi-span membrane glycoproteins. *Biochem J.* 1994, 302 (Pt 1):253–60.
- [35] Tam, L. Y., Landolt-Marticorena, C. & Reithmeier, R. A. Glycosylation of multiple extracytosolic loops in Band 3, a model polytopic membrane protein. *Biochem J.* 1996, 318 (Pt 2):645–8.
- [36] Wollscheid, B., Bausch-Fluck, D., Henderson, C., O'Brien, R. *et al.* Mass-spectrometric identification and relative quantification of N-linked cell surface glycoproteins. *Nat. Biotechnol* 2009, 27:378–86.
- [37] Rapp, M., Drew, D., Daley, D. O., Nilsson, J. *et al.* Experimentally based topology models for E. coli inner membrane proteins. *Protein Sci.* 2004, 13:937–45.
- [38] Kim, H., Melén, K., Osterberg, M. & Heijne, G. V. A global topology map of the *Saccharomyces cerevisiae* membrane proteome. *Proc. Natl. Acad. Sci. USA.* 2006, 103:11142–7.
- [39] Lomize, M. A. OPM: Orientations of Proteins in Membranes database. *Bioinformatics* 2005, 22:623–625.
- [40] Consortium, U. Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res.* 2011, 39:D214–9.
- [41] Finn, R. D., Mistry, J., Tate, J., Coghill, P. *et al.* The Pfam protein families database. *Nucleic Acids Res.* 2010, 38:D211–22.
- [42] Li, W. & Godzik, A. CD-Hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 2006, 22:1658–9.
- [43] Bendtsen, J. D., Nielsen, H., Heijne, G. V. & Brunak, S. Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.* 2004, 340:783–95.
- [44] Wistrand, M., Kall, L. & Sonnhammer, E. A general model of G protein-coupled receptor sequences and its application to detect remote homologs. *Protein Sci* 2006, 15:509–521.
- [45] Hedin, L., Ojemalm, K., Bernsel, A., Hennerdal, A. *et al.* Membrane insertion of marginally hydrophobic transmembrane helices depends on sequence context. *J Mol Biol* 2010, 396:221–229.

7 Tables

Method	Release year	Algorithm	URL	MSA	SP	Constrained	Training data	Filter
HMM-TM	2006	HMM	http://biophysics.bioL.uoa.gr/HMM-TM/	-	-	+	highres	-
HMMTOP	2001	HMM	http://www.enzim.hu/hmmtop/	*	-	+	lowres	-
MEMSAT	1994	ANN + Grammar	http://single.topcons.net/	-	-	-	lowres	+
MEMSAT3	2007	ANN + Grammar	http://bioinf.cs.ucl.ac.uk/psipred/	+	-	-	lowres	+
MEMSAT-SVM	2009	SVM + Model	http://bioinf.cs.ucl.ac.uk/psipred/	+	+	-	highres	+
OCTOPUS	2008	ANN + HMM	http://octopus.cbr.su.se/	+	-	+	highres	-
Philius	2008	DBN	http://www.yeast-ctermrc.org/philius/	+	+	-	lowres	-
Phobius	2004	HMM	http://phobius.sbc.su.se/	-	+	+	lowres	-
PolyPhobius	2005	HMM	http://phobius.sbc.su.se/poly.html	+	+	-	lowres	-
PRO	2004	HMM	http://topcons.net/	+	-	+	lowres	-
PRODIV	2004	HMM	http://topcons.net/	+	-	+	lowres	-
SCAMPI	2008	Hydrophobicity + Model	http://topcons.net/	+	-	+	highres	+
SCAMPI-single	2008	Hydrophobicity + Model	http://scampi.cbr.su.se/	-	-	+	highres	+
SPOCTOPUS	2008	ANN + HMM	http://octopus.cbr.su.se/	+	+	-	highres	-
TMHMM	2001	HMM	http://www.cbs.dtu.dk/services/TMHMM/	-	-	-	lowres	-
TOPCONS	2009	HMM - Consensus	http://topcons.net/	+	-	+	both	-
TOPCONS-single	2011	HMM - Consensus	http://single.topcons.net/	-	-	-	both	+
TopPred II	1994	Hydrophobicity profiles	http://mobyle.pasteur.fr/	-	-	-	lowres	-

Table 1: The membrane topology predictors used in the current study. The key features of each are listed. MSA: the predictor uses multiple sequence alignments as input. *: The original version of HMMTOP is capable of utilizing homologous sequences to improve predictions, however we benchmark the single-sequence version. SP: the predictor also predicts signal peptides. Constrained: the predictor allows constrained predictions, i.e. it allows using (experimentally derived) topology information as input. Training data: the predictor is trained on high-resolution (highres) or low-resolution (lowres) topology information. Filter: the predictor incorporates a filter for non-TM (globular) proteins. +*: The method is used as input to TOPCONS-single, and is pre-filtered with the SCAMPI-single filter for non-TM (globular) proteins.

Dataset	Method	Size	SP	No. TM	multi-spanning	single-spanning	non-TM
highres	3D-structure	124	-	3.7	54%	45%	1%
highres	Predicted	124	32%	3.8 ± 0.1	$55\% \pm 2\%$	$43\% \pm 2\%$	$2\% \pm 2\%$
lowres	Experimental	163	-	5.1	72%	28%	0%
lowres	Predicted	163	25%	5.2 ± 0.2	$74\% \pm 3\%$	$25\% \pm 3\%$	$1\% \pm 2\%$
E.coli-cterm	Predicted	612	20%	7.5 ± 0.2	$99\% \pm 1\%$	$1\% \pm 1\%$	$0\% \pm 0\%$
yeast-cterm	Predicted	546	20%	7.3 ± 0.5	$97\% \pm 3\%$	$2\% \pm 2\%$	$1\% \pm 1\%$
glyco	Predicted	787	70%	2.8 ± 0.4	$53\% \pm 14\%$	$34\% \pm 11\%$	$13\% \pm 9\%$
GPCR	Predicted	1405	34%	6.7 ± 0.2	$100\% \pm 0\%$	$0\% \pm 0\%$	$0\% \pm 0.1\%$
cyto	Predicted	1925	2%	0.3 ± 0.4	5 ± 7	13 ± 16	82 ± 23
extra	Predicted	2024	60%	0.4 ± 0.4	4 ± 5	31 ± 20	65 ± 24

Table 2: Properties of benchmark sets. Values based on predictions are averages over all 18 predictors, with one standard deviation noted after the $\pm$ -sign. SP: fraction of proteins with signal peptides. No. TM: average number of TM regions. multi-spanning : average number of multi-spanning (>1 TM regions) proteins. single-spanning : average number of single-spanning (1 TM region) proteins. non-TM: average number of non-TM (0 TM regions) proteins.

Method	Average No. TM	7TM N_{out} (Sp + 7TM N_{out})	8 TM Nin	“Correct”
HMM-TM	6.5	36% (-)	10%	46%
HMMTOP	6.8	77% (-)	4%	81%
MEMSAT	6.5	53% (-)	12%	65%
MEMSAT3	6.7	63% (-)	10%	73%
MEMSAT-SVM	6.5	52% (15%)	0%	52%
OCTOPUS	7.1	59% (-)	20%	79%
Philius	6.8	58% (6%)	9%	67%
Phobius	6.8	73% (9%)	3%	76%
PolyPhobius	6.9	85% (7%)	1%	86%
PRO	6.6	54% (-)	6%	60%
PRODIV	6.9	76% (-)	9%	85%
SCAMPI	6.5	64% (-)	9%	73%
SCAMPI-single	6.8	22% (-)	15%	37%
SPOCTOPUS	7.0	58% (8%)	22%	80%
TMHMM	6.5	65% (-)	2%	67%
TOPCONS	6.8	71% (-)	8%	79%
TOPCONS-single	6.5	49% (-)	11%	60%
TopPred II	6.6	36% (-)	9%	45%

Table 3: Evaluation on the test set of human GPCRs. 7TM N_{out} includes the predictions with signal peptides.

Method	highres	lowres	E.coli-cterm	yeast-cterm	Glyco	GPCR	<Rank>	<Z-score>
TOPCONS	1.56	1.73	0.87	1.43	1.54	0.83	2.5	1.33
PRODIV	0.66	0.98	0.72	1.43	0.41	1.25	4.4	0.91
SCAMPI	1.16	0.58	1.17	0.35	1.48	0.41	4.4	0.86
SPOCTOPUS	0.90	0.17	1.17	0.81	0.52	0.90	4.9	0.75
OCTOPUS	0.98	1.40	1.17	0.66	-0.60	0.83	5.3	0.74
PolyPhobius	-1.33	0.09	0.13	1.27	1.30	1.32	7.2	0.46
MEMSAT3	0.82	0.75	1.02	-0.41	-0.12	0.34	7.3	0.40
PRO	0.32	0.42	0.57	1.27	-0.09	-0.51	8.3	0.33
Philius	0.16	0.17	0.13	0.20	0.71	-0.02	8.9	0.23
TOPCONS-single	0.10	0.46	0.10	0.19	0.25	-0.51	10.4	0.10
Phobius	-1.58	0.01	-0.47	0.35	0.83	0.62	9.9	-0.04
HMMTOP	-0.50	-0.16	-0.62	-1.48	0.45	0.97	11.1	-0.22
SCAMPI-single	0.49	-0.24	0.13	-0.56	-0.26	-2.13	12.0	-0.43
MEMSAT-SVM	0.32	-0.57	-1.81	-0.72	-0.78	-1.07	13.8	-0.77
HMM-TM	0.16	-1.06	0.28	-0.72	-1.99	-1.49	13.7	-0.80
TMHMM	-1.33	-0.81	-1.37	-1.48	-0.70	-0.02	14.8	-0.95
MEMSAT	-1.88	-1.32	-1.22	-1.27	-1.38	-0.16	15.5	-1.20
TopPred II	-1.00	-2.61	-1.96	-1.33	-1.56	-1.56	16.7	-1.67

Table 4: Ranking of the prediction methods using six different benchmark. Column 2 to 7 shows the Z-score (number of standard deviations above/below the average performance) of a predictor for the six different benchmarking datasets. <Rank>: The average rank in these sets. <Z-score>: The average Z-score. All Z-scores larger than 1.0 are marked in bold.

Dataset	Agreeing, “top 6”	Agreeing, all	Consensus, “top 6”	Consensus, all
highres	69% (83%)	65% (81%)	42% (86%)	16% (90%)
lowres	73% (86%)	65% (81%)	50% (89%)	16% (90%)
E.coli-cterm	68% (82%)	61% (78%)	41% (86%)	11% (89%)
yeast-cterm	53% (72%)	42% (65%)	30% (82%)	2.4% (82%)
glyco	37% (60%)	36% (60%)	8.5% (66%)	0.9% (77%)
GPCR	60% (77%)	51% (71%)	31% (82%)	3.2% (83%)

Table 5: Topology predictions are more diverse in the glyco dataset than in the other benchmark datasets. Agreeing, “top 6”: the fraction of pairwise comparisons of topology prediction results that agree, only considering PRODIV, SCAMPI, SPOCTOPUS, PolyPhobius, MEMSAT3, and PRO. Agreeing, all: the fraction of pairwise comparisons of topology prediction results that agree among all 18 predictors. Consensus, “top 6”: the fraction of proteins predicted equally by PRODIV, SCAMPI, SPOCTOPUS, PolyPhobius, MEMSAT3, and PRO. Consensus, all: the fraction of proteins predicted equally by all 18 predictors. Numbers in parenthesis denote the estimated upper limit of achievable prediction accuracy, see text for details.

8 Figures

Figure 1: The highres and E.coli-cterm benchmark sets are easier to predict. Predictors incorporating homology information in the form of multiple sequence alignments have labels in bold. Predictors capable of constraining predictions with experimental information have labels in italics. Predictors trained on high-resolution data are in red, predictors trained on low-resolution data are in blue, and predictors in green used both kinds in training.

Figure 2: Prediction performance on the glyco dataset correlates with the fraction of “outside” points predicted. The solid line shows the linear regression of the data points and the dotted line is the prediction performance of a random predictor. Predictors incorporating homology information in the form of multiple sequence alignments have labels in bold. Predictors capable of constraining predictions with experimental information have labels in italics. Predictors trained on high-resolution data are in red, predictors trained on low-resolution data are in blue, and predictors in green used both kinds in training. Note that all predictors line up close to the regression line and without exception well above the line of the random predictor.

Figure 3: Constraining the N-terminal to be “outside” for proteins with predicted signal sequences does not help prediction performance on the yeast-cterm and E.coli-cterm sets. Predictors not capable of incorporating topology constraints or that predict signal peptides themselves are excluded. Results for the E.coli-cterm dataset are marked with filled circles and results for the yeast-cterm dataset have crosses. Predictors where the prediction was constrained have labels in italics. Predictors incorporating homology information in the form of multiple sequence alignments have labels in bold. Predictors trained on high-resolution data are in red, predictors trained on low-resolution data are in blue, and predictors in green used both kinds in training. Within the E.coli-cterm dataset, SignalP predicted the presence of a signal peptide in 123 out of 613 proteins and in the yeast-cterm dataset the corresponding numbers were 126 out of 546 proteins.

(a) glyco

(b) cyto

(c) extra (not constrained)

(d) extra (constrained)

Figure 4: Predicted number of TM regions in a) the glyco dataset, b) the cyto dataset, c) the extra dataset (not constrained), and d) the extra dataset (constrained).

Figure 5: Performance at separation of TM and non-TM proteins. a) Phobius, PolyPhobius and Philius are the best at separation. PRODIV and TopPred II need pre-filtering of non-TM proteins to be effective. b) Constraining signal peptides identified by SignalP to be “outside” helps separation. Predictors incorporating homology information in the form of multiple sequence alignments have labels in bold. Predictors capable of constraining predictions with experimental information have labels in italics. Predictors trained on high-resolution data are in red, predictors trained on low-resolution data are in blue, and predictors in green used both kinds in training.