doi:10.1016/j.jmb.2004.03.016 J. Mol. Biol. (2004) 338, 1027–1036

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

A Combined Transmembrane Topology and Signal Peptide Prediction Method

Lukas Käll[1], Anders Krogh[2] and Erik L. L. Sonnhammer[1]*

[1]*Center for Genomics and Bioinformatics, Karolinska Institutet, SE-17 177 Stockholm, Sweden*

[2]*The Bioinformatics Center University of Copenhagen Universitetsparken 15 DK-2100 Copenhagen Denmark*

**Corresponding author*

An inherent problem in transmembrane protein topology prediction and signal peptide prediction is the high similarity between the hydrophobic regions of a transmembrane helix and that of a signal peptide, leading to cross-reaction between the two types of predictions. To improve predictions further, it is therefore important to make a predictor that aims to discriminate between the two classes. In addition, topology information can be gained when successfully predicting a signal peptide leading a transmembrane protein since it dictates that the N terminus of the mature protein must be on the non-cytoplasmic side of the membrane. Here, we present Phobius, a combined transmembrane protein topology and signal peptide predictor. The predictor is based on a hidden Markov model (HMM) that models the different sequence regions of a signal peptide and the different regions of a transmembrane protein in a series of interconnected states. Training was done on a newly assembled and curated dataset. Compared to TMHMM and SignalP, errors coming from cross-prediction between transmembrane segments and signal peptides were reduced substantially by Phobius. False classifications of signal peptides were reduced from 26.1% to 3.9% and false classifications of transmembrane helices were reduced from 19.0% to 7.7%. Phobius was applied to the proteomes of *Homo sapiens* and *Escherichia coli*. Here we also noted a drastic reduction of false classifications compared to TMHMM/SignalP, suggesting that Phobius is well suited for whole-genome annotation of signal peptides and transmembrane regions. The method is available at http://phobius.cgb.ki.se/ as well as at http://phobius.binf.ku.dk/

Keywords: transmembrane protein; signal peptide; topology prediction; hidden Markov model; machine learning

Introduction

A well-known weakness of currently available transmembrane (TM) helix predictors is the frequent false classifications of signal peptides (SPs) as TM helices.[1–3] Conversely, SP predictors have a tendency of falsely classifying TM helices as SPs.[4–6] These frequent false classifications are a consequence of the fact that both predictions are primarily looking for a stretch of hydrophobic residues as the main recognition pattern. It therefore seems natural to try to resolve this lack of discrimination by constructing a joint TM topology and SP predictor.

Predicting transmembrane protein topology

TM protein topology prediction is a classical problem in bioinformatics. Since the structure of TM proteins is difficult to determine by experimental means, it has been a rewarding task to predict their topologies computationally. It may seem easy to recognize an α-helical TM segment since it normally consists of a 15–30 amino acid residues long region with an overrepresentation of hydrophobic residues. However, it is complicated by the fact that many TM helices in multispanning TM proteins are partially or completely shielded by other TM helices. Since they are not entirely exposed to the lipid bilayer they constitute

Abbreviations used: HMM, hidden Markov model; TM, transmembrane; SPs, signal peptides.

E-mail address of the corresponding author: erik.sonnhammer@cgb.ki.se

amphipatic helices. Long stretches of hydrophobic residues also exist in other types of protein moieties, e.g., buried within globular domains or in SPs, which could be falsely predicted as TM helices. The task to make TM topology predictions, i.e. to localize all TM segments as well as determine the location (inside the cytoplasm or outside) of the loops turns out to be far from trivial.

Early TM helix prediction methods were based on experimentally determined hydropathy indices of hydrophobic properties for each amino acid. For the examined protein, a hydropathy plot was calculated by adding the hydropathy indexes over a window with a fixed length. A heuristically determined cut-off value was then used to indicate possible TM segments.[7,8] An important improvement to this strategy was the observation that there is an overrepresentation of positively charged amino acid residues in the cytoplasmic loops of TM proteins.[9] This gave a hint about the location of the loops and led to the development of the first automated full TM topology prediction methods e.g. TOPPred.[10] The method first scans the sequence for certain and putative TM segments and then selects the most likely topology, including none, some or all of the putative segments, based on the charge of the loops. Instead of calculating hydropathy plots there are methods letting a sequence profile (DAS[11]) or an Artificial Neural Network (PHDhtm[12]) detect potential TM segments.

Instead of scanning the sequence for TM segments and then sorting out the topology as a second step, the search for TM segments can be integrated with the evaluation of possible topologies in one step. The amino acid distribution of the investigated sequence is compared to precalculated expected amino acid distributions in each type of topologically distinct region (TM helices and cytoplasmic and non-cytoplasmic loops) of a TM protein. Given the correlation measurements between the amino acid distributions of the examined protein and the expected amino acid distributions in different topological regions, the most likely topology can be predicted. A nice feature of this approach is the ability to model all parts of the protein so that all topogenic signals are weighted properly, which is preferable to giving priority to the hydrophobic signal. This was first done by expectation maximization in the method Memsat.[13] Probabilistic approaches to the problem have been taken as well. A commonly used probabilistic framework for such tasks is the hidden Markov model (HMM).[14] Some popular HMM-based predictors are TMHMM[1,15] and HMMTOP.[2]

β-Barrel TM proteins seem to be hard to predict with the classical TM prediction methods since their TM segments generally are shorter and with a different amino acid composition than α-helical TM segments. Lately some methods to predict such structures have been published.[16,17] We have chosen not to include β-barrel TM proteins in this study and we restrict our efforts to model α-helical TM segments.

Predicting signal peptides

Similar to the TM segment, one of the strongest indications of an SP is a hydrophobic α-helical region. This is called the h-region of the SP. However, the hydrophobic region is generally shorter for an SP (approximately 7–15 residues) than for a TM helix. The h-region is near the N-terminal of the protein but it is preceded by a slight positively charged n-region with high variability in length (approximately 1–12 amino acid residues). Between the h-region and the cleavage site, a somewhat polar and uncharged 3–8 amino acid residues long c-region is situated. Another clear motif on the SP is the presence of small, neutral residues at the −3 and −1 relative to the cleavage site.[18,19]

Most available SP prediction methods use weight matrices,[20] Artificial Neural Networks (e.g. SignalP[4]), HMMs (e.g. SignalP-HMM[5]) or Support Vector Machines.[21,22] An evaluation[23] showed that the very popular method, SignalP V2.0.b2, is more sensitive than the other methods, and predicts cleavage sites more accurately, but includes many false positive predictions.

Combined models

Alongside its SP model, SignalP-HMM[5] uses a model of a signal anchor, i.e. a TM protein with one TM segment near the N-terminal of the protein, to help discriminate against false positives. Similarly, LipoP[24] models N-terminal TM helices, SPs and lipoprotein signal peptides in Gram-negative bacteria to improve discrimination between these categories. However, as far as we know, nobody has yet constructed a joint TM topology and SP predictor.

An additional reason for including an SP model when predicting TM topology is, apart from improved SP/TM discrimination, that the presence of an SP indicates that the N terminus of the mature TM protein is on the non-cytoplasmic side of the protein. In that case, the TM topology prediction problem is reduced to finding the correct TM helices, since the orientation of the protein is given by the SP prediction.

Here we describe a new method, Phobius, based on HMM, aiming to predict both TM topology of a protein and the presence of an SP in the protein. The choice of the HMM framework as prediction technique is natural because it has successfully been used for both prediction types separately, and a combination of the model types is relatively straightforward. The main strength of Phobius lies in the ability to discriminate TM segments from SPs. This makes it more accurate on mixed TM/SP proteins than the best TM-only and SP-only predictors. For SP-only proteins, it is more

conservative than SignalP, i.e. has a lower false positive rate but also a higher false negative rate.

Results

Model architecture

The model architecture of Phobius can be regarded as a combination of the models made in TMHMM and SignalP-HMM, with a transition from the last state of the SP model in SignalP-HMM to the outer loop state in the TMHMM model. However, several modifications were made to both models. Different combinations of these modifications were then compared against each other (data not shown). The final Phobius model, which is the architecture with the best performance, is shown in Figure 1(a). In contrast to a profile HMM,[25] where each state has its own individual emission probability, we have tied the states within the various parts of the model, i.e. emission probabilities of tied states are identical. A run of tied states in the model is referred to as a compartment.

The TM helix submodel (Figure 1(b)) consists of three compartments. A four-residue helix cytoplasmic end is followed by a 7–26 residue helix core, which is followed by a four-residue helix non-cytoplasmic end. TM segments can thus be

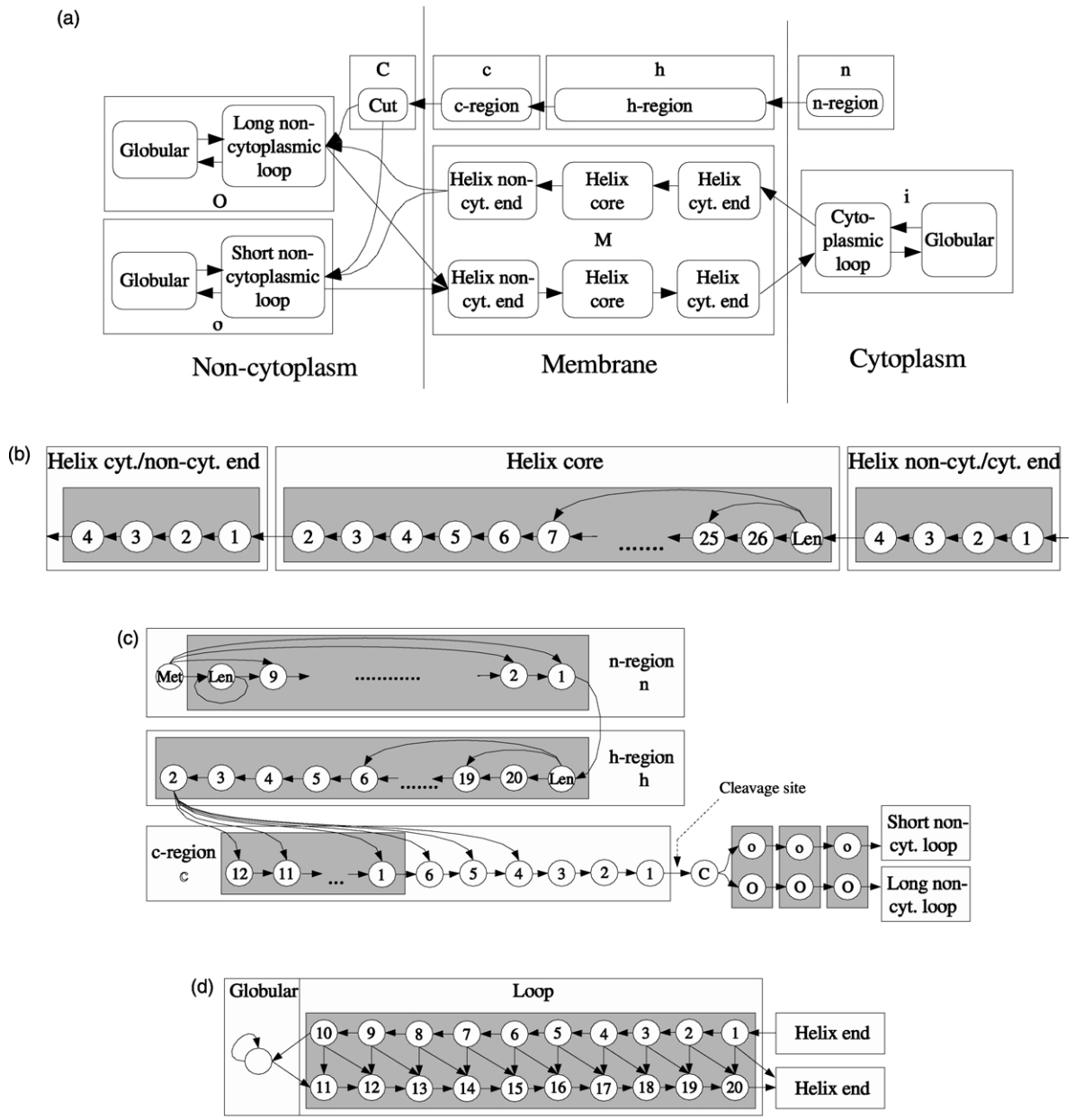

Figure 1. The layout of the HMM model. States within grayed boxes have tied emission probabilities. (a) Overview of the model. (b) The TM helix submodel. (c) The signal peptide submodel. (d) The cytoplasmic and short non-cytoplasmic loop submodels.

Table 1. Accuracy of transmembrane topology predictions by Phobius, TMHMM 2.0, HMMTOP 2.1, a combination of SignalP-HMM and TMHMM, and a combination of SignalP-HMM and HMMTOP, measured for different data sets

Proteins contain	Both-TM-and-SP		TM-only		SP-only	Neither-TM-nor-SP
Test set sequences	New	All	New	All	All	All
Phobius	94.1%	91.1%	53.9%	63.6%	96.1%	98.2%
TMHMM 2.0	70.6%	71.1%	44.5%	65.2%	73.9%	98.7%
HMMTOP 2.1	52.9%	51.1%	50.8%	66.8%	37.2%	85.0%
TMHMM-SignalP	88.2%	86.7%	39.5%	58.7%	98.0%	99.2%
HMMTOP-SignalP	88.2%	82.2%	45.0%	59.1%	89.6%	86.0%

A prediction was counted as correct when all the predicted TM helices overlap all the annotated TM helices of the protein over a stretch of at least five residues and the location of the loops were correct. Incorrectness in predicting SPs was not regarded. For the test sets not containing TM helices a correct TM topology prediction corresponds to a prediction that does not contain any TM segments. The New test set refers to the subset of proteins that excluded the TMHMM and the HMMTOP training data. All Phobius measurements were cross-validated, while TMHMM and HMMTOP were not.

between 15 and 34 amino acid residues long in total. The emission probabilities of all three compartments are tied between the inward and outward going parts of the model.

The SP submodel (Figure 1(c)) is split into n, h, c and post-cleavage site regions. The n-region begins with a methionine state followed by an n-compartment. The n-compartment contains ten states, where one of the states contains a self-transition, enabling n-regions of arbitrary lengths with a probability distribution tailing off exponentially. The transitions are arranged such that the minimum length of the n-region is two residues. The h-region consists of a 6–20 residue compartment. The c-region consists of a 12 state compartment followed by six untied states. The transitions are such that the length of the c-region can be anywhere between 4 and 18 residues. In contrast to the SignalP-HMM model, no self-transition that allows the c-region to be infinitely long is included in the model, as no experimental evidence supports this. The post-cleavage site region consists of four untied states. The region surrounding the cleavage site thereby contains ten untied states, making this part of the model similar to a weight matrix.

The loop submodels (Figure 1(d)) consist first of a 20 state compartment that allow any loop length between 1 and 20 residues. For longer loops, a self-looping globular state is connected between states 10 and 11. There are three different loop models: The cytoplasmic loop, the short non-cytoplasmic loop and the long non-cytoplasmic loop. The reason for having separate short and long non-cytoplasmic loop compartments is that long loops are likely to contain globular domains, making the loops sufficiently different to warrant separate treatment. One could imagine that globular domains have to be transported across the membrane by a special mechanism. It is hard to say exactly the loop length at which the line has to be drawn; we have defined it arbitrarily at 100 residues as in TMHMM.[1] Only using one non-cytoplasmic loop model reduced the correct TM predictions on the TM-only set (see below) from 63.6% to 61.5%. The long non-cytoplasmic loop has a fixed zero-probability transition between the incoming and the outgoing states, in order to force the sequence to pass the globular state. The globular compartment of the long non-cytoplasmic loop contains three states (not shown in Figure 1) with tied self-transitions while the other loops only contain one. The aim of this is to produce a length model of the long non-cytoplasmic loop that

Table 2. Errors in signal peptide predictions made by Phobius and SignalP V2.0.2b

Proteins contain	Both-TM-and-SP	TM-only	SP-only		Neither-TM-nor-SP
Error type	False negatives	False positives	False negatives		False positives
Test set sequences	All	All	New	All	All
Phobius	4.4%	7.7%	2.4%	3.5%	2.3%
SignalP-NN	2.2%	42.9%	2.2%	2.3%	4.8%
SignalP-HMM	0.0%	19.0%	0.6%	1.4%	4.0%

In the test sets containing signal peptides the values correspond to false negatives, while for the ones not containing signal peptides the values correspond to false positives. The New test set sequences refers to measurements done on the test set where proteins potentially used for training SignalP were removed, i.e. removing proteins added to SWISS-PROT before release 35. They were included when training Phobius, but using cross-validation. There were no proteins containing both TM and SP that we could exclude from being a part of the SignalP training data. Therefore no value is presented for the "Both-TM-and-SP" with "New" test set sequences. The values in the SignalP-NN column were obtained when only taking in account the "mean S" score flag of the prediction, while the values in the row SignalP-HMM were obtained by only taking in account the final SignalP-HMM prediction. SignalP was executed with the kingdom-specific (i.e. eukaryote, Gram-positive bacteria or Gram-negative bacteria) version according to the annotation of the test set. All Phobius measurements were cross-validated, while SignalP was not.

Table 3. Number of correctly predicted signal peptide cleavage sites with Phobius and SignalP V2.0.2b

Proteins contain	SP-only	
Test set sequences	New	All
Phobius	75.9%	73.4%
SignalP-NN	84.8%	81.9%
SignalP-HMM	81.8%	80.2%

The New test set sequences refers to measurements done on the test set where proteins potentially used for training SignalP were removed, i.e. removing proteins added to SWISS-PROT before release 35. They were included when training Phobius, but using cross-validation. All Phobius measurements were cross-validated, while SignalP was not.

favors longer loops. Using only one globular state reduced the correct TM predictions on the TM-only set from 63.6% to 62.3%. The emission probabilities of all globular compartment states in different loop models were set identical.

Training was done without dividing the sequences by kingdom, i.e. eukaryota, archea, Gram-positive bacteria, and Gram-negative bacteria. We tried training by separate kingdoms, but we found no performance increase in doing so (data not shown).

Comparison with other methods

The performance of Phobius was measured by tenfold cross-validation and compared to the performance of TMHMM ver. 2.0[1] and HMMTOP 2.1 Tusnady9769220, reported in Table 1, and SignalP V2.0.b2,[26] reported in Tables 2 and 3. The test sets are described in Table 5. Because the reference methods results were not cross-validated, we also tested them using only data that had not been used during their training (column New). In the case of SignalP V2.0.b2, for which the explicit training set is not available, this was done by excluding sequences that had been reported in SWISS-PROT[27] up to release 35, which is the release that the training set of SignalP V2.0.b2 was extracted from. The measurements on Phobius in the New column were obtained by training on the full cross-validation sets, but only testing against the parts of the cross-validation sets not used in the other methods' training.

The comparison shows that Phobius is successful in making fewer misclassifications of TM helices as SPs and fewer misclassifications of SPs as TM helices with respect to the compared methods. On the other hand, Phobius is less sensitive when predicting SPs and less accurate in predicting cleavage sites than SignalP, but this is well compensated for by the reduction of false positive predictions.

In most tests, Phobius was more accurate than the other TM prediction methods. The exceptions are on the Neither-TM-nor-SP dataset where Phobius is marginally less accurate (0.5%) than TMHMM, and on the complete TM-only dataset (1.6%). Given the much lower accuracy of TMHMM on the New part of the TM-only set, the high value on the All set could be due to the overlap between the TMHMM training set and the test set (i.e. lack of cross-validation). To investigate this we retrained TMHMM on the All dataset and measured the accuracy with tenfold cross-validation. This resulted in a drop in accuracy of 2.9%, which makes it less accurate than Phobius.

The comparison also shows that HMMTOP performs better than TMHMM on the TM-only data set, but has clear problems when running on data containing SPs or trying to sort out soluble proteins.

A way to improve TM predictors that do not handle SPs is to first remove any SP detected by a separate SP predictor before running the TM predictor.[1,28] To investigate the behavior of such a predictor, we removed SPs detected by SignalP-HMM from our test sets and reran TMHMM and HMMTOP. As can be seen in Table 2, this resulted in a clear increase in performance on the datasets containing SPs, but there was also a drop in performance on the TM-only set.

All the investigated TM prediction methods have a surprisingly low TM topology accuracy on the TM-only dataset. It is even lower when removing the original TMHMM training set. This suggests that the training set of TMHMM is more easily predicted than the other TM sequences. This observation is well in line with a previously drawn conclusion that the TMHMM training dataset is much easier to predict than genomic datasets.[29,30]

The much higher accuracy for the TM predictors in the Both-TM-and-SP category should be read in the light of the fact that it is a very small data set containing only 45 sequences in the whole test set and 17 sequences in the New part. The test set is probably biased towards topologies that are easy to predict. Thus, although the difference in accuracy between Phobius and other methods not taking SPs into account is undisputable, the absolute level of accuracy is perhaps an overestimate.

Our results indicate that SignalP-HMM is both more sensitive and selective in detecting SPs than SignalP-NN, but that SignalP-NN has higher accuracy in predicting correct cleavage sites than SignalP-HMM. A possible explanation for the low accuracy of Phobius in predicting cleavage sites is that we took all cleavage site annotations in SWISS-PROT when we gathered the training data. Given that the used discriminative training procedure (see Materials and Methods) is rather sensitive to bad training data, there is a risk that incorrect cleavage sites, even if just a few, have biased the model towards false sites.

Application to genomic data

Given the much lower rate of false classifications produced by Phobius, it should be more reliable

Table 4. Examination of difference in behavior of different prediction methods

	Phobius SP + TM	Phobius TM only	Phobius SP only	TMHMM	SignalP-HMM	SignalP-NN
A. *The 26,309 sequences in the* H. sapiens *genome*						
Phobius TM and SP	1572	0	0	1441	1508	1499
Phobius TM only	0	4763	0	4050	785	1411
Phobius SP only	0	0	2630	479	2491	2280
TMHMM	1441	4050	479	6030	2525	3193
SignalP-HMM	1508	785	2491	2525	5614	4527
SignalP-NN	1499	1411	2280	3193	4527	5696
B. *The 4289 sequences in the* E. coli *genome*						
Phobius TM and SP	143	0	0	139	104	130
Phobius TM only	0	811	0	762	203	482
Phobius SP only	0	0	599	128	505	542
TMHMM	139	762	128	1037	422	725
SignalP-HMM	104	203	505	422	841	786
SignalP-NN	130	482	542	725	786	1359

The values above represent the number of sequences in intersections between the sets of sequences containing predicted SPs and/or TM segments by Phobius, TMHMM 2.0 and SignalP V2.0.2b.

than other methods for whole-genome annotation of SP and TM-containing proteins. To investigate this, we applied Phobius, SignalP and TMHMM to the *Homo sapiens*[31] (26,309 proteins) and *Escherichia coli*[32] (4289 proteins) genomes. The sizes of the sequence sets predicted by the different methods to contain SPs and/or TM segments are found on the diagonal in Table 4. The sizes of the intersections between the sets are found off the diagonal in the same Table.

In Table 4, we can see that Phobius found fewer SPs than SignalP. The human genome was predicted to contain 4202 SPs by Phobius and 5614 SPs by SignalP-HMM. For *E. coli*, Phobius predicted 742 proteins with SPs while SignalP-HMM predicted 841.

Phobius found more TM proteins than TMHMM in the human data, 6335 *versus* 6030. In *E. coli*, however, Phobius found fewer TM proteins than TMHMM, 954 *versus* 1037. For the human proteins, 479 of the TM proteins predicted by TMHMM were predicted to have only an SP by Phobius, and it is likely that the majority of those are not TM proteins, because TMHMM often confuses SPs with TM helices (see Table 1). Ignoring these leaves only 60 proteins predicted as TM proteins by TMHMM but not by Phobius. On the other hand, Phobius predicted 844 TM proteins in human that were not predicted by TMHMM. We see a similar pattern for *E. coli*, for which TMHMM predicted eight TM proteins not predicted by Phobius (if we assume that Phobius was correct in predicting that 128 only contain an SP). Phobius predicted 53 TM proteins that were not predicted by TMHMM.

This behavior is consistent with the observation that when disregarding the SP/TM confusion by TMHMM, Phobius is less specific but more sensitive than TMHMM. We believe that TM proteins found by both methods have a very low error rate indeed, and those predicted by TMHMM or Phobius alone have higher error rates. To get an indication of whether the 844 TM proteins predicted by Phobius alone are false positives, we analyzed the proteins that were annotated by SWISS-PROT and examined how many of them contained the keyword "Transmembrane". For human, 99 of 234 (42%) SWISS-PROT entries contained Transmembrane, compared to 15.4% for all of SWISS-PROT. Although the Phobius unique set appears enriched in true TM proteins some may be false positives.

Table 5. Test set composition

	Both-TM-and-SP	TM-only	SP-only	Neither-TM-nor-SP
Sequence similarity is measured over	Whole sequence	Whole sequence	SP + 6 aa	Whole sequence
Max sequence similarity				
Within subsets	40%	80%	40%	40%
Between subsets	35%	30%	20%	20%
Number of sequences				
Total	45	247	1275	1087
Eukaryotic	37	100	847	414
Prokaryotic	4	133	428	540
Other (viral or Archea)	4	14	0	133
After removal of TMHMM training data	17	128	–	–
After removal of potential SignalP training data	0	–	494	–

To assess the level at which TMHMM and SignalP confuse SPs and TM helices, we looked at how often they predict overlapping TM and SP segments. Even though we do not know the true answer for genomic datasets, such an overlap means that one of them must be a false prediction. If the predictions by TMHMM largely overlap the predictions of SignalP, we can be sure that they have a high false positive rate. In the human genome, Phobius predicted 1641 proteins containing both an SP and one or more TM segments while TMHMM + SignalP-HMM predicted 2525. In *E. coli*, Phobius predicted 143 while TMHMM + SignalP-HMM predicted 422. For 1272 (50% of the 2525) sequences in the human genome and 415 (98% of the 422) sequences in *E. coli* there was an overlap of at least one amino acid residue between the SignalP-HMM-predicted SP and a TMHMM-predicted TM segment. We therefore conclude that the observation on the experimental test set holds true for genomic predictions as well, namely that TMHMM and SignalP have high false positive rates on the "other" type of hydrophobic segment. Phobius is forced to make a choice between these types and therefore produces much fewer false predictions.

Length distribution

Both TMHMM and SignalP-HMM are prone to preferentially predict certain lengths of the hydrophobic region. In the case of TMHMM 2.0, this results in a high representation of predictions with 23, 20 and 18 amino acid residues length of TM segments as shown in Figure 2. It has also been reported that SignalP-HMM favors certain lengths of the h-regions of SPs (8 and 11 for eukaryotes, 9 and 12 for Gram-negatives, and 14 and 17 for Gram-positives).[5]

Does this length preference stem from a biological preference of certain TM helix lengths, or is it an artifact caused by the training procedure of TMHMM and SignalP-HMM? It is impossible to define exactly where the boundaries of TM helices are located, even when the crystal structure is known. The length distribution obtained by TMHMM is optimal for the model choice and estimation procedure, but we do not believe that it reflects a true biological phenomenon.

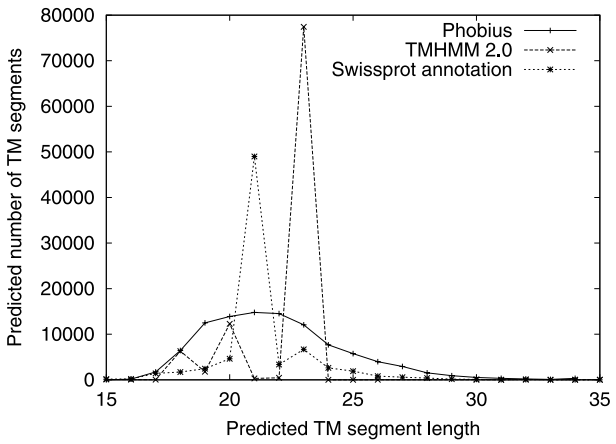

Figure 2. The length distribution of TM helices of the 122,564 sequences in SWISS-PROT release 41.0 as predicted by Phobius, TMHMM 2.0 and the annotation of SWISS-PROT.

In order to obtain a smoother length distribution, a modified training procedure was used when training Phobius (see Materials and Methods). The length distributions of all TM segments in SWISS-PROT and predicted by Phobius and by TMHMM 2.0 are shown in Figure 2. The overrepresentation of annotated TM segment lengths of 21 amino acid residues in SWISS-PROT is probably due to the frequent use of TOPPred as TM topology predictor, since it sets TM segment length to 21 by default.

Discussion

We have trained and tested a new prediction method, Phobius, that predicts both transmembrane helices and SPs. In handling both types of predictions at the same time it discriminates better between SPs and TM helices. The method is based on a HMM and works without any post processing of the HMM decoding.

We have shown that Phobius is able to reduce cross-prediction errors when analyzing the genome of *H. sapiens* and *E. coli* and thereby giving more accurate figures of TM protein and SP content. We expect this improvement to extend to other genomes as well.

The procedure for pretraining of length models for the TM and SP regions contributes to prediction accuracy. When testing a model without such pretraining the correctly predicted TM topologies decrease from 63.6% to 61.9% on the TM-only-set, and the SP false positives increase from 3.5% to 3.8% on the SP-only-set. Although the performance increase is relatively modest, we believe the model produces results that are more biologically realistic, because the length distribution is smoother (see Figure 2).

Studies have been published on comparisons between different TM prediction methods.[3,33,34] Such comparisons generally turn out to be quite dependent on the test set used; in particular they are sensitive to fraction of proteins having an SP. It is hard to construct an objective dataset. On top of that, the accuracy values reported on the test sets have recently been shown to severely overestimate the expected accuracy in a whole-genome test,[29] so any performance figure from a benchmark test should not be translated to an expected accuracy value when predicting genomic data.

A tempting extension of our method is to incorporate a submodel for mitochondrial targeting peptides, chloroplast transit peptides, lipoprotein signal peptides and/or other protein sorting

signals in the model. Only very few of these sorting signals are misclassified as SPs or TM helices, but they could give a valuable contribution to TM prediction by indicating the location of the loops of the TM protein.

Materials and Methods

Data sets

We have collected and curated four different datasets:

- A set with TM proteins with SPs (Both-TM-and-SP set)
- A set with TM proteins without SPs (TM-only set)
- A set with SPs and no TM helices (SP-only set)
- A set without SPs or TM helices (Neither-TM-nor-SP set)

The sets containing TM helices originate from different sources:

- 146 sequences from the TMHMM "160 dataset"[15]
- 140 sequences from TMPDB ver 6.2[35]
- 2 sequences from the Möller dataset[36]
- 4 sequences of TM proteins with known 3D structure found in SWISS-PROT

The TM helix containing proteins were divided into the Both-TM-and-SP and the TM-only sets based on their annotation.

The SP-only set was collected from SWISS-PROT Release 41.0/TrEMBL Release 23.0 employing the Menne procedure[23] followed by removal of all putative TM proteins based on their annotation.

Finally, the Neither-TM-nor-SP set was collected by extracting sequences from SWISS-PROT 41.0 with a known 3D structure that had no indication of an SP or TM segment in their annotation.

Homology reduction was done in two steps that both employed matches reported by BLAST processed by MSPcrunch[37] 2.0 with the option zero coverage rejection. To assure that the sequence identity was not too high between the sequences a "Hobohm algorithm 2 redundancy reduction"[38] was performed using BLAST sequence identity cut-offs according to Table 5. The resulting sets were sorted into ten cross-validation subsets, so that the identity between the subsets was kept below the cut-offs in Table 5. The cut-offs were for practical reasons different for different datasets, but we never allowed more than 80% identity within a subset or more than 35% between subsets. The cut-offs and sizes of the different data sets are shown in Table 5.

To enable supervised training and relevant testing, each amino acid in the data sets was labelled as cytoplasmic (i), non-cytoplasmic (o), long looped non-cytoplasmic (O), or SP cleavage site (C) according to the annotation of protein or n-region (n), h-region (h), or c-region (c) of an SP using a SignalP-HMM based model.

HMM training

The training procedure for the HMM model was based on the procedure used for training TMHMM 2.0.[1]

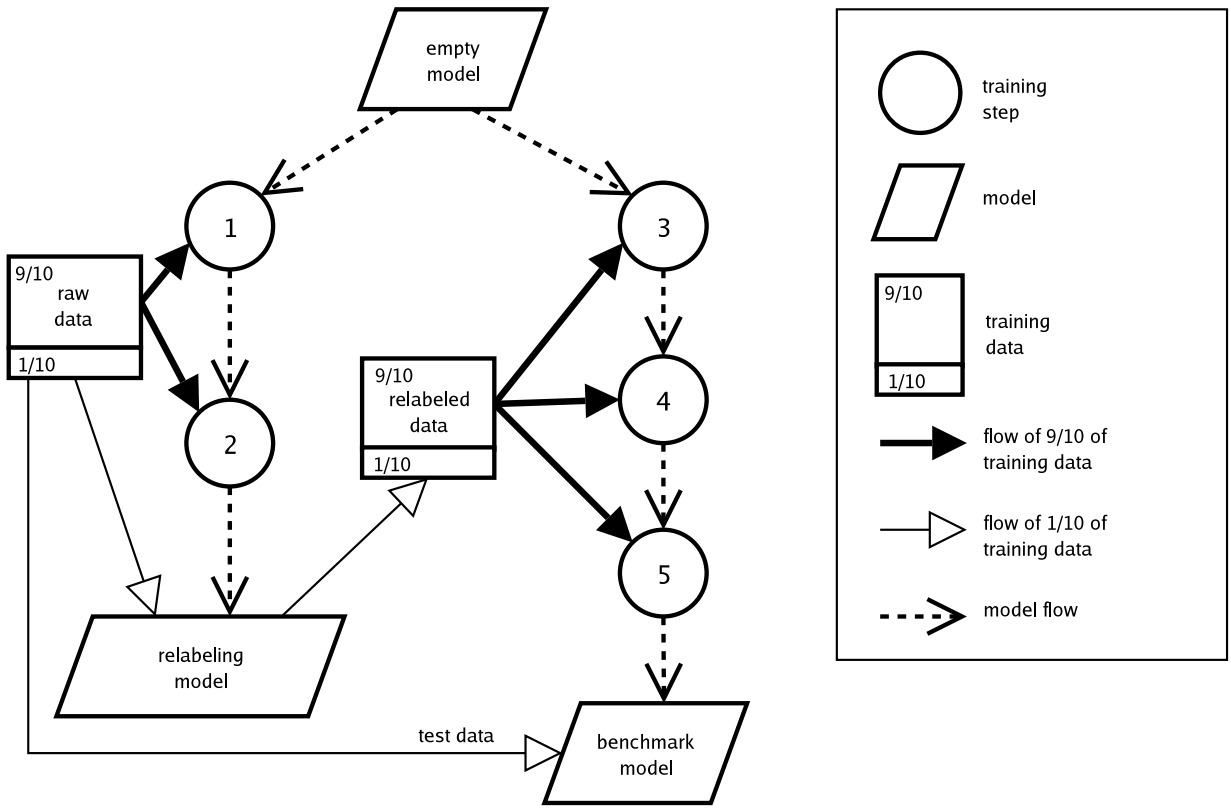

Figure 3. Graphical representation of the training of each of the ten cross-validation models. The third to the fifth training step are depending on that the other nine models are trained in parallel, and thus delivering relabelled data to the training of the tenth model.

However, the training was modified to overcome the described problem with an extreme bias towards certain TM helix and SP region lengths. The training procedure includes five steps and is graphically illustrated in Figure 3.

To test the accuracy of Phobius, tenfold cross-validation was used, i.e. training was performed on nine of the ten training data subsets while the remaining subset was used for testing. The subsets were then rotated so that in total ten models were trained and tested on the take-out set. This avoids any testing on sequences that were used during training.

In the first step, only the transition probabilities of the states in the TM segments and the n-, h- and c-regions were estimated. The distribution of TM segment lengths was fitted to a gamma probability distribution while the different SP regions length distributions were fitted to normal probability distributions. This was done by maximum likelihood estimation of the parameters in the distributions with respect to the measured lengths in the training data. Since the gamma and the normal distributions are continuous functions rather than discrete, the transition probabilities were assigned to the states based on integration of the estimated distributions. In the case of the TM helix the transition probabilities were frozen during the next steps, while the calculated probabilities were used as priors for the SP regions. All other transition probabilities as well as all emission probabilities are unaffected by this step.

The second step aims to correct imperfections in the annotation of the extent of TM helices and SP regions in the data sets. Three amino acid residues in both directions from a border between a loop and a TM helix were "unlabelled"[15] in the training data. The same unlabelling was also done for the n, h and c-regions of SPs, but keeping the cleavage site intact. Then the model from the first step was trained with a noise injected Baum–Welch iterative procedure.[1,25] The output model from this training step (called "relabelling model" in Figure 3) was then used to relabel the training data for the subsequent steps. In the relabelling procedure a freedom of five amino acid residues was allowed in each direction between a loop and a TM helix and full flexibility in assigning the compartments of the SPs. The "relabelling model" was trained on nine of the ten training data subsets and the "relabelling model" relabelled the remaining subset.

The third step was identical with the first except that the relabelled training set was used.

In the fourth step a normal Baum–Welch procedure was used to train the model from the third step with the relabelled training set as input, i.e. the rest of the transition probabilities and the emission probabilities were assigned. This was mainly done to help the discriminative training in the next step to converge.

In the fifth step the model parameters were updated by discriminative training using conditional maximum likelihood.[15,39] The training maximized the probability of correct labelling rather than maximizing the probability of the observed sequences, as in the previous steps. To reduce the dominance of the much larger sets containing no TM helices, only one cross-validation subset of the SP-only and none of the Neither-TM-nor-SP subsets are used for training during this step.

After performance figures for the resulting ten models were measured, the third to fifth stage were redone with all the ten subsets as training data (without cross-validation) so that a single model was obtained.

HMM prediction

The 1-best algorithm[25,39,40] operating on the Phobius model was used to perform the predictions as it is considered the most suitable decoder for models trained by a conditional maximum likelihood procedure†.

Acknowledgements

This work was supported by grants from Pfizer Corporation and from the Swedish Knowledge Foundation. A.K. was supported by EU grant no. QLRI-CT-2001-00015.

References

1. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* **305**, 567–580.
2. Tusnady, G. E. & Simon, I. (1998). Principles governing amino acid composition of integral membrane proteins: application to topology prediction. *J. Mol. Biol.* **283**, 489–506.
3. Lao, D. M., Arai, M., Ikeda, M. & Shimizu, T. (2002). The presence of signal peptide significantly affects transmembrane topology prediction. *Bioinformatics*, **18**, 1562–1566.
4. Nielsen, H., Engelbrecht, J., Brunak, S. & von Heijne, G. (1997). Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.* **10**, 1–6.
5. Nielsen, H. & Krogh, A. (1998). Prediction of signal peptides and signal anchors by a hidden Markov model. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **6**, 122–130.
6. Nielsen, H. (1999). From sequence to sorting: prediction of signal peptides. PhD Thesis, Stockholm University.
7. Kyte, J. & Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157**, 105–132.
8. Argos, P., Rao, J. K. & Hargrave, P. A. (1982). Structural prediction of membrane-bound proteins. *Eur. J. Biochem.* **128**, 565–575.
9. von Heijne, G. (1986). The distribution of positively charged residues in bacterial inner membrane proteins correlates with the trans-membrane topology. *EMBO J.* **5**, 3021–3027.
10. von Heijne, G. (1992). Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule. *J. Mol. Biol.* **225**, 487–494.
11. Cserzo, M., Wallin, E., Simon, I., von Heijne, G. & Elofsson, A. (1997). Prediction of transmembrane alpha-helices in prokaryotic membrane proteins: the dense alignment surface method. *Protein Eng.* **10**, 673–676.
12. Rost, B., Casadio, R., Fariselli, P. & Sander, C. (1995). Transmembrane helices predicted at 95% accuracy. *Protein Sci.* **4**, 521–533.

† Prediction servers based on the Phobius method are available at http://phobius.cgb.ki.se/ as well as http://phobius.binf.ku.dk/

13. Jones, D. T., Taylor, W. R. & Thornton, J. M. (1994). A model recognition approach to the prediction of all-helical membrane protein structure and topology. *Biochemistry*, **33**, 3038–3049.
14. Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, **77**, 257–286.
15. Sonnhammer, E. L., von Heijne, G. & Krogh, A. (1998). A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **6**, 175–182.
16. Martelli, P. L., Fariselli, P., Krogh, A. & Casadio, R. (2002). A sequence-profile-based HMM for predicting and discriminating beta barrel membrane proteins. *Bioinformatics*, **18**, S46–S53.
17. Zhai, Y. & Saier, M. H. (2002). The beta-barrel finder (BBF) program, allowing identification of outer membrane beta-barrel proteins encoded within prokaryotic genomes. *Protein Sci.* **11**, 2196–2207.
18. von Heijne, G. (1983). Patterns of amino acids near signal-sequence cleavage sites. *Eur. J. Biochem.* **133**, 17–21.
19. Perlman, D. & Halvorson, H. O. (1983). A putative signal peptidase recognition site and sequence in eukaryotic and prokaryotic signal peptides. *J. Mol. Biol.* **167**, 391–409.
20. von Heijne, G. (1986). A new method for predicting signal sequence cleavage sites. *Nucl. Acids Res.* **14**, 4683–4690.
21. Chou, K. C. (2001). Prediction of protein signal sequences and their cleavage sites. *Proteins: Struct. Funct. Genet.* **42**, 136–139.
22. Vert, J. P. (2002). Support vector machine prediction of signal peptide cleavage site using a new class of kernels for strings. *Pac. Symp. Biocomput.*, 649–660.
23. Menne, K. M., Hermjakob, H. & Apweiler, R. (2000). A comparison of signal sequence prediction methods using a test set of signal peptides. *Bioinformatics*, **16**, 741–742.
24. Juncker, A. S., Willenbrock, H., von Heijne, G., Brunak, S., Nielsen, H. & Krogh, A. (2004). Prediction of lipoprotein signal peptides in Gram-negative Bacteria. *Protein Sci.* in the press.
25. Krogh, A., Brown, M., Mian, I. S., Sjolander, K. & Haussler, D. (1994). Hidden Markov models in computational biology. Applications to protein modeling. *J. Mol. Biol.* **235**, 1501–1531.
26. Nielsen, H., Engelbrecht, J., Brunak, S. & von Heijne, G. (1997). A neural network method for identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Int. J. Neural Syst.* **8**, 581–599.
27. Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M. C., Estreicher, A., Gasteiger, E. *et al.* (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucl. Acids Res.* **31**, 365–370.
28. Arai, M., Ikeda, M. & Shimizu, T. (2003). Comprehensive analysis of transmembrane topologies in prokaryotic genomes. *Gene*, **304**, 77–86.
29. Käll, L. & Sonnhammer, E. L. (2002). Reliability of transmembrane predictions in whole-genome data. *FEBS Letters*, **532**, 415–418.
30. Melen, K., Krogh, A. & von Heijne, G. (2003). Reliability measures for membrane protein topology prediction algorithms. *J. Mol. Biol.* **327**, 735–744.
31. Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J. *et al.* (2001). Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
32. Blattner, F. R., Plunkett, G., III, Bloch, C. A., Perna, N. T., Burland, V., Riley, M. *et al.* (1997). The complete genome sequence of *Escherichia coli* K-12. *Science*, **277**, 1453–1462.
33. Möller, S., Croning, M. D. & Apweiler, R. (2001). Evaluation of methods for the prediction of membrane spanning regions. *Bioinformatics*, **17**, 646–653.
34. Chen, C. P., Kernytsky, A. & Rost, B. (2002). Transmembrane helix predictions revisited. *Protein Sci.* **11**, 2774–2791.
35. Ikeda, M., Arai, M., Okuno, T. & Shimizu, T. (2003). TMPDB: a database of experimentally-characterized transmembrane topologies. *Nucl. Acids Res.* **31**, 406–409.
36. Möller, S., Kriventseva, E. V. & Apweiler, R. (2000). A collection of well characterised integral membrane proteins. *Bioinformatics*, **16**, 1159–1160.
37. Sonnhammer, E. L. & Durbin, R. (1994). A workbench for large-scale sequence homology analysis. *Comput. Appl. Biosci.* **10**, 301–307.
38. Hobohm, U., Scharf, M., Schneider, R. & Sander, C. (1992). Selection of representative protein data sets. *Protein Sci.* **1**, 409–417.
39. Krogh, A. (1994). *Hidden Markov models for labeled sequences* Proceedings of the 12th IAPR International Conference on Pattern Recognition, IEEE Computer Society Press, Los Alamitos, CA pp. 140–144.
40. Schwartz, R. & Chow, Y. (1990). The N-best algorithm: an efficient and exact procedure for finding the N most likely sentence hypotheses. *Proc. ICASSP, Albuquerque, NM*, **1990**, 81–84.

Edited by J. Thornton

(Received 30 October 2003; received in revised form 25 February 2004; accepted 9 March 2004)