Prediction of signal peptides and signal anchors by a hidden Markov model

Henrik Nielsen and Anders Krogh
Center for Biological Sequence Analysis
Technical University of Denmark
Building 206, 2800 Lyngby, Denmark
hnielsen@cbs.dtu.dk and krogh@cbs.dtu.dk

Abstract

A hidden Markov model of signal peptides has been developed. It contains submodels for the N-terminal part, the hydrophobic region, and the region around the cleavage site. For known signal peptides, the model can be used to assign objective boundaries between these three regions. Applied to our data, the length distributions for the three regions are significantly different from expectations. For instance, the assigned hydrophobic region is between 8 and 12 residues long in almost all eukaryotic signal peptides. This analysis also makes obvious the difference between eukaryotes, Gram-positive bacteria, and Gram-negative bacteria. The model can be used to predict the location of the cleavage site, which it finds correctly in nearly 70% of signal peptides in a cross-validated test—almost the same accuracy as the best previous method. One of the problems for existing prediction methods is the poor discrimination between signal peptides and uncleaved signal anchors, but this is substantially improved by the hidden Markov model when expanding it with a very simple signal anchor model.

Introduction

The general secretory pathway is a mechanism for protein secretion found in both eukaryotic and prokaryotic cells. The entry to the general secretory pathway is controlled by the signal peptide, an N-terminal peptide typically between 15 and 40 amino acids long, which is cleaved from the mature part of the protein during translocation across the membrane, see Figure 1.

The most characteristic common feature of signal peptides is a stretch of hydrophobic amino acids called the *h-region*. The region between the initiator Met and the h-region, the *n-region*, is typically one to five amino acids in length, and normally carries positive charge. Between the h-region and the cleavage site is the *c-region*, which consists of three to seven polar, but mostly uncharged, amino acids. Close to the cleavage site a more specific pattern of amino acids is found: the residues at positions -3 and -1 (relative to the cleavage site) must be small and neutral for cleavage to occur correctly (von Heijne 1985).

Translocation takes place via a multiprotein complex known as the translocon or translocation apparatus (Rapoport, Jungnickel, & Kutay 1996). The signal peptide is recognised by at least three steps in the process: targeting to the membrane by cytoplasmic factors, binding to the translocon, and cleavage by signal peptidase. During translocation, the signal peptide adopts a hairpin-like structure with the N-terminus remaining on the cytoplasmic side, and the cleavage site close to the trans side of the membrane. In eukaryotes translocation is co-translational, with the translation taking place on ribosomes bound to the ER membrane. In bacteria, translocation is predominantly post-translational.

The properties of signal peptides are known to differ between various types of organisms: bacterial signal peptides are longer than their eukaryotic counterparts, and those of Gram-positive bacteria are longer than those of Gram-negative bacteria (which have an outer membrane in addition to the cytoplasmic membrane). The charge difference is much more prominent in bacteria than in eukaryotes. The composition of the h-region also shows some difference: eukaryotic h-regions are generally more hydrophobic with Leu as the most common residue, while bacterial h-regions are slightly less hydrophobic and contain approximately equal amounts of Ala and Leu (von Heijne & Abrahmsén 1989; Nielsen *et al.* 1997). In bacterial signal peptides, the positive charge in the n-region is often balanced by a negative net charge in the c-region or in the first few residues of the mature protein (von Heijne 1986a).

Some proteins have sequences that initiate translocation in the same way as signal peptides do, but are not cleaved by signal peptidase (von Heijne 1988). As the rest of the polypeptide chain is translocated through the membrane, the resulting protein remains anchored to the membrane by the hydrophobic region, with a short N-terminal cytoplasmic domain, see Figure 1. The uncleaved signal peptide is known as a signal anchor, and the resulting protein is known as a type II membrane protein. Signal anchors differ from signal peptides in other respects than the cleavage sites: they have h-regions longer than those of cleaved signal peptides—the length is typically the same as that of a transmembrane α-helix—and the n-regions can also be much longer, up to more than 100 residues. Interestingly, experiments have shown that it is possible to convert a cleaved signal peptide to a signal anchor merely by lengthening the h-region (Chou & Kendall 1990; Nilsson, Whitley, & von

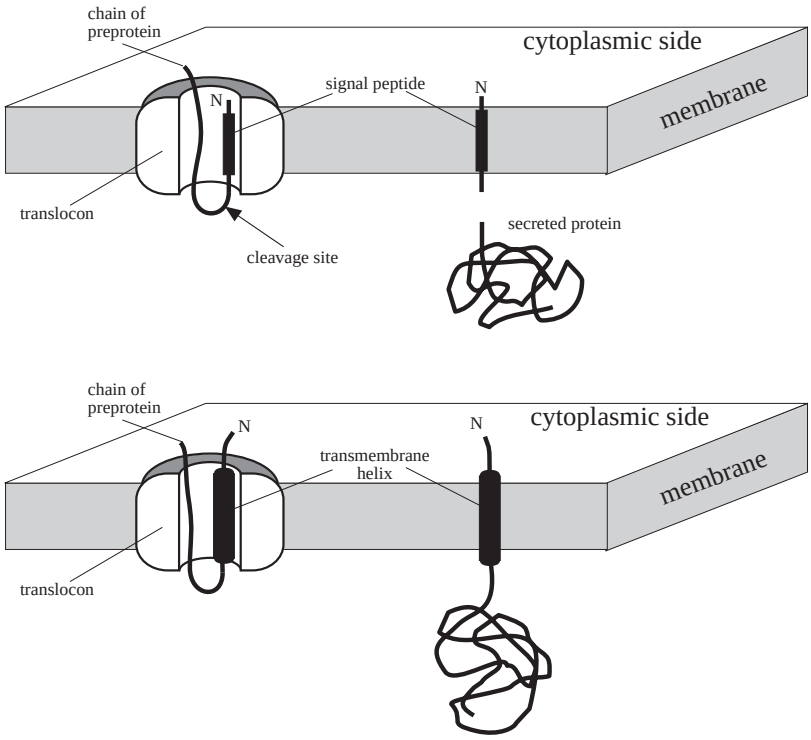

Figure 1: Cartoons of a signal peptide (above) and a signal anchor (below), and how they are translocated by the translocon. After translocation the signal peptide is cleaved off and the mature protein released, whereas the signal anchor is not cleaved off and the protein is anchored to the membrane.

Heijne 1994).

Signal peptide prediction involves two tasks: (1) Given that the sequence is a signal peptide, locate the cleavage site; and (2) discriminate between secretory proteins with signal peptides and non-secretory proteins. Prediction of the cleavage site has been performed with a weight matrix (von Heijne 1986b) and by a neural network method, SignalP (Nielsen *et al.* 1997), which also performs the discrimination task. SignalP has been available as a WWW and mail server since 1996 and is very widely used.

In this paper we apply a hidden Markov model (HMM) for both prediction tasks. An HMM for proteins consists of a number of states that are connected by transition probabilities. Associated with each state is a distribution over the 20 amino acids. It is often useful to think of HMMs as generative models that can 'emit' protein sequences by randomly going from state to state, and in each state emit an amino acid according to the distribution for that state. For a given sequence one can calculate for instance the most probable way this sequence was generated by the model, or the total probability that it was generated by the model at all. Because it is a probabilistic model, one can use standard methods like maximum likelihood to determine the model parameters. Introductions to HMMs can be found in (Rabiner 1989; Krogh 1998; Durbin *et al.* 1998). In computational biology the most commonly used HMM type is probably the profile HMM (Krogh *et al.* 1994; Eddy 1996), which has a structure inspired by profiles (Gribskov, McLachlan, & Eisenberg 1987). However, HMMs are more general, and the model structures used in this work are not of the profile type.

One of the advantages of HMMs is that it is usually very easy to build biological knowledge into the model in an intuitive way—in contrast to *e.g.* neural networks. For the signal peptides we design the model so that it has parts corresponding to each of the three regions of a signal peptide and such that reasonable length constraints are hard-wired in the model. Another advantage of the HMM approach is that the HMM can easily be extended by adding other modules to the model. In this work we combine the signal peptide model with a model of signal anchors, in order to make a model that is good at discriminating between signal peptides and anchors. There are very few known examples of signal anchors, and therefore it is hard to make good models of these. For this situation, the HMMs have another big advantage: it is very easy to control the model complexity by making the model simple enough to be estimated from the amount of data available.

Methods

Data sets

Data were extracted from SWISS-PROT version 35 (Bairoch & Apweiler 1997). Data sets were made for four types of proteins: signal peptides, signal anchors, cytoplasmic, and (for eukaryotes) nuclear. All sets were grouped in subsets for eukaryotes, Gram-positive bacteria, and Gram-negative

	Signal peptides		Cytoplasmic proteins		Nuclear proteins		Signal anchors	
	tot.	red.	tot.	red.	tot.	red.	tot.	red.
Euk	2477	1137	1614	461	2060	990	164	67
G_{neg}	498	356	697	335	–	–	–	–
G_{pos}	222	172	280	151	–	–	–	–

Table 1: The number of sequences in the data sets before (**tot.**) and after (**red.**) redundancy reduction. The organism groups are: Eukaryotes (**Euk**), Gram-negative bacteria ($\mathbf{G_{neg}}$), and Gram-positive bacteria ($\mathbf{G_{pos}}$).

bacteria. Details of the extraction criteria can be found in (Nielsen *et al.* 1996).

The signal peptides were extracted according to feature table annotation. Virus and phage genes, proteins encoded by organellar genes (in eukaryotes), and lipoprotein signal peptides cleaved by signal peptidase II (in prokaryotes) were discarded. In this work, signal peptides annotated to be shorter than 15 or longer than 50 amino acids were excluded from the data set, because if they are correct, we regard these as so atypical that they hardly can be modeled. Signal anchors were identified by transmembrane regions marked "Signal anchor" or "type II membrane protein". Multi-spanning membrane proteins were not included. Entries where the suggested signal anchor region was 70 amino acids or longer (measured from the N-terminal to the C-terminal end of the specified transmembrane region) were also discarded, because these would hardly be mistaken for cleavable signal peptides. Prokaryotic signal anchors were ignored as well, since only one good example was found. The N-terminal parts of cytoplasmic and (for the eukaryotes) nuclear proteins were extracted by searching for comment lines in SWISS-PROT specifying the subcellular location as "cytoplasmic" or "nuclear".

For all data sets we avoided entries containing hints in the annotation that the feature or subcellular location was not experimentally verified. We also removed proteins that did not start with Met, unless they had an annotation concerning a removed initiator Met. Proteins in all the sets were truncated after 70 residues, which is the region we have chosen to model, because almost all signal peptides are shorter than 70.

All the data sets were then homology reduced, so that no two sequences were homologous within a set (see (Nielsen *et al.* 1996) for details). This was done for two reasons, firstly to limit the bias of the HMM towards over-represented families, and secondly to allow the sets to be used for cross-validation. The sizes of the data sets before and after homology reduction are shown in Table 1. Finally each set was divided into five parts of approximately equal size for cross-validation.

Model structure

To get an idea of the length and amino acid distributions of the three different regions in a signal peptide, we initially assigned tentative n-, h-, and c-regions after the following very simple procedure: (1) Place a pointer at the -1 position (immediately before the cleavage site), set the assignment to c-region, and scan the sequence upstream towards the N-terminus; (2) move the pointer three positions upstream (assigning -1 through -3 as a minimal c-region); (3) set the assignment to h-region at the first occurrence of at least two consecutive hydrophobic residues (Ala, Ile, Leu, Met, Phe, Trp, or Val); (4) move the pointer six positions upstream; (5) set the assignment to n-region at the first occurrence of either a charged residue or at least three consecutive non-hydrophobic residues; (6) if the N-terminal end of the h-region is not a hydrophobic residue, move the pointer back downstream, changing the assignment to n-region until a hydrophobic residue is found.

This set of rules gives h-regions that necessarily have a hydrophobic residue in the N-terminus and two consecutive hydrophobic residues in the C-terminus. All the signal peptides in our data set were assigned an h-region, ranging from 6 to 20 in length with very few exceptions. By the above procedure, the c-region is by definition at least three residues long, whereas the n-region is typically between two and seven long, but can be significantly longer.

The regions defined in this way were used while designing the model shown in Figure 2. It implements an explicit modeling of the length distribution of the h-region with an array of 20 states, where there is a transition from the first state directly to each of the following 15 states, which means that the minimum length of the h-region is 6 and the maximum 20. All these states are *tied*, which means that they have the same amino acid distribution. We placed hard limits on the length of the h-region in accordance with experimental findings: h-regions shorter than 6 amino acids are not able to promote translocation (Bird, Gething, & Sambrook 1990), while the transition from cleavable to non-cleavable seems to occur for h-region lengths between 17 and 20 amino acids (Chou & Kendall 1990; Nilsson, Whitley, & von Heijne 1994).

The n-region is modeled by an array of 8 states, of which the last seven are also tied to each other (but use another distribution than the h states). The first state has probability one for Met, because all the proteins in our datasets begin with Met. From this state there are transitions to all the other n-states as well as the first h-state. The second n-state has a transition to itself, and therefore this part of the model can model the explicit length distribution between one and 8, and then an exponentially decaying length distribution (a geometric distribution) for longer n-regions.

The c-region is modeled by an array of six states prior to the cleavage site, in which each state has a specific distri-

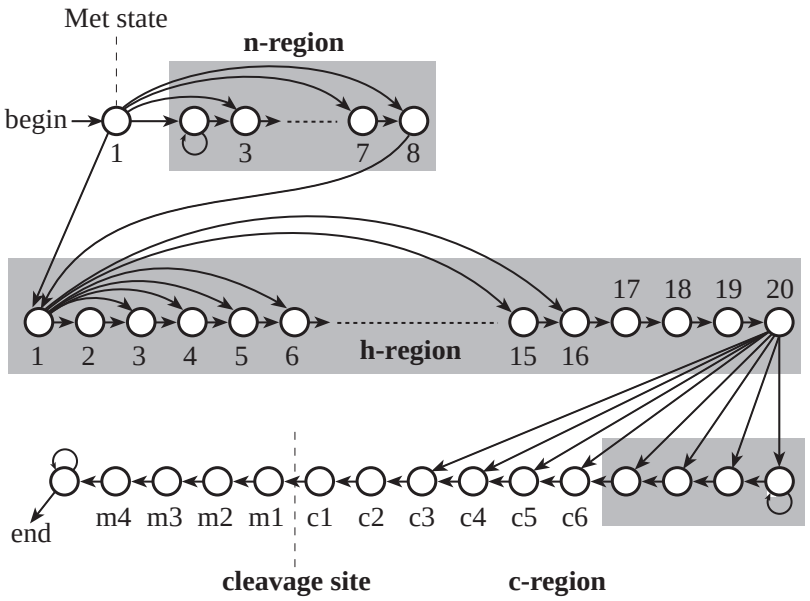

Figure 2: The model used for signal peptides. The states in a shaded box are tied to each other.

bution to capture the pattern of amino acids just before the cleavage site (states c1 to c6 in Figure 2). To allow for longer c-regions, four more states (c7 to c10) are added, which are tied to each other in order to capture the over-all amino acid distribution of c-regions longer than six. One of these states has a transition to itself so long c-regions are modeled by a geometric distribution. From the last h-state there are transitions to all the c-states except the two just before the cleavage site, making the minimum length of c-regions equal to three. After the cleavage site, four states model the position specific amino acid distributions before a transition is made to the final state with an amino acid distribution equal to a standard background distribution. The six states prior to the cleavage site plus the four states after the cleavage site correspond approximately the weight matrix used earlier for signal peptide prediction (von Heijne 1986b). The difference is that the states c4 through c6 can be skipped, which means that the weight matrix-like part does not have to model hydrophobic residues of signal peptides with very short c-regions.

Models were estimated from the training data by the Baum-Welch algorithm (Rabiner 1989; Durbin *et al.* 1998), which is a maximum likelihood procedure that iteratively increases the total likelihood of the training data. The training was done with the labeled data, such that the cleavage site was always correctly positioned during training, but the model was left to find out for itself where to put the boundaries between n-, h-, and c-regions. However, to help the model find a sensible partition into regions, we initialized the models: for each of the three regions, the initial distributions were set to the amino acid frequencies in the regions as assigned by the simple procedure described above. Pseudocounts (Krogh *et al.* 1994; Durbin *et al.* 1998) were also added, which were obtained by multiplying the same amino acid frequencies by 100. The size of this number is not critical. Each distribution is obtained from more than 1000 amino acids, so the pseudocounts are relatively small.

To predict the cleavage site for a new sequence, the most probable path through the trained model is found by the standard Viterbi algorithm (Rabiner 1989). The most probable path was also used for assigning a region to each amino acid in the sequence.

To discriminate between signal peptides, signal anchors and soluble non-secretory proteins, the model was augmented by a model of anchors as shown in Figure 3. The structure of this model is like the model for signal peptides, but the n- and h-regions are simpler and the c-region is of course omitted. The whole model was now trained from all types of sequences (signal peptides, anchors, cytoplasmic and nuclear). The most likely path through the combined model yields a prediction of which of the three classes the protein belongs to.

Neural network method

The neural network method implemented in the SignalP server is described in detail elsewhere (Nielsen *et al.* 1997). In the present work, we made no modifications to the architecture of the networks, the training scheme, or the output interpretation; we merely retrained the networks on the new data set (the present version of SignalP is based on SWISS-PROT release 29).

In the context of this work, it is important to note that SignalP combines two types of network: the *C-score* (raw cleavage site score) is the output from a network trained solely on signal peptide sequences to recognize cleavage sites from non-cleavage sites; while the *S-score* (signal peptide score) is the output from a network trained to recog-

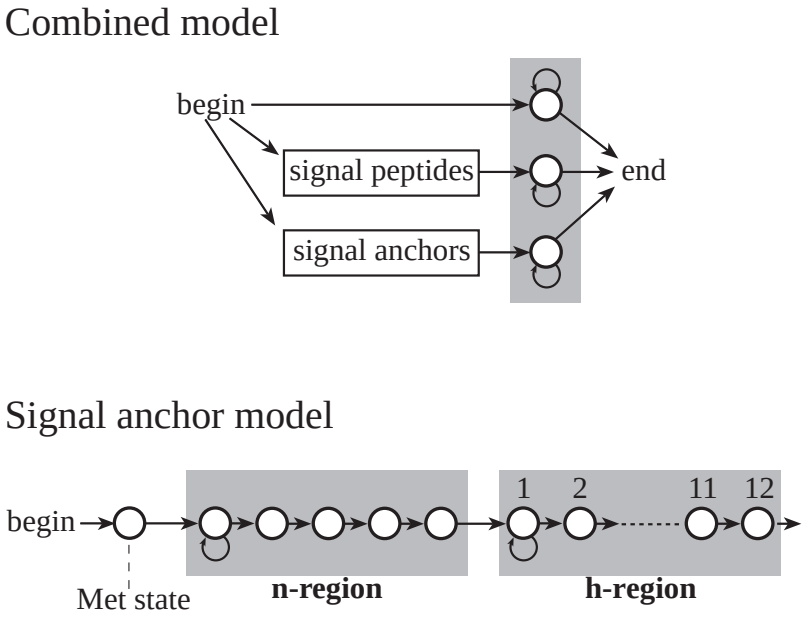

Figure 3: The block diagram (top) shows how the combined model is put together from the signal peptide model and the anchor model. The final states shown in the shaded box are tied to each other, and model all residues not in a signal peptide or an anchor. The model of signal anchors (bottom) has only two types of states (grouped by the shaded boxes) apart from the Met state.

nize windows within signal peptides from windows after the cleavage site and windows in non-secretory proteins. The prediction of cleavage site location is optimized by observing where the C-score is high *and* the S-score changes from a high to a low value. This is formally implemented by the *Y-score* (combined cleavage site score), a geometric average of the C-score and a smoothed derivative of the S-score.

Discrimination between signal peptides and non-secretory proteins is done by using either the maximal value of the Y-score or the mean value of the S-score, averaged from position 1 to the most likely cleavage site.

Results and discussion

Performance of the HMM method

The performances of the trained hidden Markov model and neural networks are shown in Table 2. All the results reported are obtained by five-fold cross validation. For cleavage site location, the neural networks are slightly better than the HMM. The observation that the neural networks—even using only the C-score—are able to locate the cleavage site a few percent more precisely than the HMM suggests that there might be a weak non-linear feature involved in the cleavage site recognition.

Discrimination between signal peptides and soluble non-secretory proteins is performed with a version of the HMM where the anchor model is omitted. If the three-module HMM including the signal anchor model is used instead, a few signal peptides are falsely classified as signal anchors, bringing the correlation coefficient for eukaryotic sequences down by 0.02. The simple neural network (the C-score network alone) is poorer than the HMM for discrimination, which is not remarkable, since the non-secretory proteins were not used in the training of this network. The combination of C-score and S-score networks has a discrimination performance comparable to that of the HMM: for eukaryotes the networks are slightly better, while for Gram-negative bacteria the HMM is slightly better.

The neural network performances were in general comparable to those obtained with the data from SWISS-PROT release 29 as reported in (Nielsen *et al.* 1997), but the cleavage site location was two percent better for eukaryotes and four percent better for Gram-negative bacteria. Since the number of signal peptide sequences extracted has not grown very much, this suggests that the quality of signal peptide annotations has improved.

Discrimination between cleaved signal peptides and uncleaved signal anchors is shown in the rightmost column of Table 2. The HMM correlation coefficient of 0.74 corresponds to a sensitivity of 71% and a specificity of 81%—a far better performance for this problem than hitherto reported.

For the neural network, uncleaved signal anchors can to some degree be identified by intermediate values of the mean S score, but even when the threshold is optimized specifically for this task, the correlation coefficient does not exceed 0.4. Interestingly, the cleavage site scores provided an even worse discrimination between signal peptides and signal anchors, suggesting that cryptic cleavage sites are not uncommon in signal anchors. These results should not be taken as a claim that the neural network method is unable to

Task	Cleavage site location			Discrimination			
				sig/non-sec			sig/anc
Method	Euk	G_{neg}	G_{pos}	Euk	G_{neg}	G_{pos}	Euk
HMM	69.5%	81.4%	64.5%	0.94	0.93	0.96	0.74
NN (simple)	71.8%	81.7%	66.9%	(0.87)	(0.71)	(0.71)	(0.18)
NN (combined)	72.4%	83.4%	67.5%	0.97	0.89	0.96	(0.39)

Table 2: Performance of the hidden Markov model (**HMM**) compared to the neural network (**NN**) for eukaryotes (**Euk**), Gram-negative bacteria (**G_{neg}**), and Gram-positive bacteria (**G_{pos}**). Cleavage site location is given as percentage of signal peptide sequences where the cleavage site was placed correctly. Discrimination values between sequence types are given as correlation coefficients (Mathews 1975). The sequence types are signal peptides (**sig**), soluble non-secretory—*i.e.* cytoplasmic or nuclear—proteins (**non-sec**), and signal anchors (**anc**). Results are given for two versions of the NN method: one network trained on cleavage sites *vs.* non-cleavage sites (**simple**), and a combination of this with a network trained on signal peptides *vs.* non-signal peptides (**combined**). Discrimination values in parentheses are obtained without including both categories in the training set. All values are five-fold cross-validated.

solve the signal anchor problem, since the signal anchors were not included as training data in the neural network model as was the case for the HMM.

If the n- and h-states of the signal anchor submodel are tied to the corresponding states in the signal peptide submodel, performance drops slightly. This shows that discrimination between signal peptides and signal anchors does not rely solely on the presence or absence of a cleavage site pattern, but involves the differences in amino acid composition of the n- and h-regions (see Table 3).

We have also tested a more detailed modeling of the n-to-h and h-to-c region boundaries by introducing extra untied states. This could *e.g.* model a possible preference for positive charges in long n-regions to occur close to the h-region, or the need for helix-breaking residues immediately after the h-region. However, these modifications offered no significant increase in performance. Accordingly, inspection of alignments by the n-to-h boundary or the h-to-c boundary in the regions assigned by the model does not show any special composition of any position within the n- and h-region.

One position which does show a clear deviation from the background distribution is position 2 in the non-secretory proteins (immediately after the initiator Met). This reflects what is known as the N-end rule (Varshavsky 1996) which states that certain amino acids in this position make the protein stable and other make the protein subject to rapid degradation by the ubiquitination system. We tried to incorporate this into the model by one additional explicit state for the non-secretory proteins, but it had no effect on the discrimination performance.

Characteristics of signal peptides of eukaryotes and bacteria

The trained hidden Markov model is capable of assigning n-, h-, and c-regions to a signal peptide, and n- and h-regions to a signal anchor. We have performed this assignment with a model trained and tested on the whole data set, forcing the model to use the correct classification. During assignment of the signal peptide regions, we furthermore forced the annotated cleavage site to use the cleavage site state, so that the total lengths of the signal peptides correspond to those given in the database. However, no substantial differences were seen in the statistics if the predicted classifications and signal peptide lengths were used instead, or if the cross-validation models and test sets were used (data not shown).

We have computed the length distribution (see Figure 4) and amino acid composition (see Table 3) for all the assigned regions. Note that the cleavage site consensus region (position -3 through -1)—conventionally considered a part of the c-region—was not included when calculating the amino acid composition of the c-region. This was done to avoid the amino acid bias from the -3 and -1 positions.

The lengths of the HMM-assigned h- and c-regions show a quite surprising distribution which is substantially different from the length distribution of the tentative regions assigned by the simple rule described earlier (data not shown). While the tentative h- and c-regions had a rather wide distribution, remarkably few lengths are represented in the HMM-assigned regions. The eukaryotic h-regions are practically only found in lengths 8 through 12, even though the architecture of the model allows a range of 6 through 20. The c-regions are even more peaked, with more than 70% having a length of 5 in eukaryotes and 6 or 7 in bacteria.

More remarkable still is that the length distribution of the h-region is two-peaked, with modes at 8 and 11 for eukaryotes, 9 and 12 for Gram-negative bacteria, and 14 and 17 for Gram-positive bacteria. Even though the lengths are very different, the distance between the peaks is three positions in all three data sets.

The h- and c-regions of the signal anchors show a wider length distribution (Figure 4 bottom), and they are clearly longer than those of the eukaryotic signal peptides. The h-region lengths show almost no overlap between signal peptides and signal anchors.

The sharply one-peaked and two-peaked distributions of the signal peptide h- and c-regions did not persist if they were modeled by a simple loop structures like that of the h-region of the signal anchor (results not shown), but in that case performance—both in cleavage site location and sequence type discrimination—went significantly down. These results suggest that the h-region length does indeed play an important part in discrimination between signal peptides and signal anchors.

It is worth noting the differences between eukaryotes and

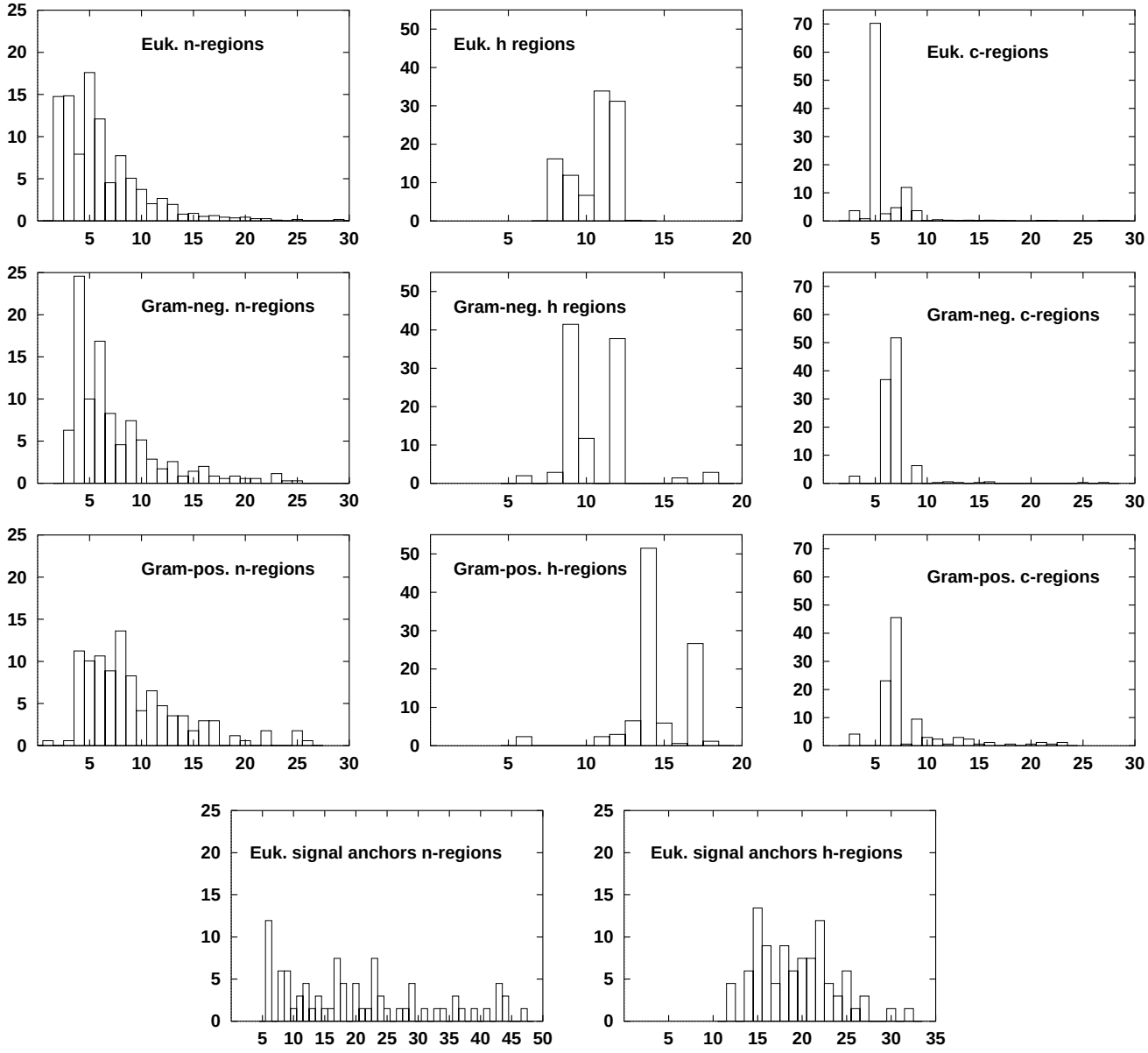

Figure 4: The length distributions of the n-, h-, and c-regions of signal peptides, and n- and h-regions of signal anchors, as assigned by the trained HMM models. The x-axis is length, and the histograms display the number of sequences in percent.

	n-region				h-region				c-region		
	anc	signal peptide			anc	signal peptide			signal peptide		
	Euk	Euk	G_{neg}	G_{pos}	Euk	Euk	G_{neg}	G_{pos}	Euk	G_{neg}	G_{pos}
A	6.1	10.2	4.1	4.1	9.7	11.6	23.9	18.8	13.3	17.5	15.8
C	2.7	1.5	0.4	0.3	4.0	4.0	2.4	1.5	2.8	1.6	0.0
D	3.3	1.3	1.0	0.9	0.1	0.1	0.0	0.1	2.3	0.7	2.4
E	6.2	2.3	1.2	1.1	0.0	0.1	0.0	0.0	3.6	0.5	2.4
F	3.4	3.7	4.7	4.1	9.2	8.0	5.1	5.5	2.8	7.0	2.9
G	6.1	7.2	2.3	3.5	5.3	2.9	6.2	7.4	10.7	8.9	7.1
H	2.5	2.2	1.9	1.8	0.1	0.3	0.0	0.0	2.4	1.2	0.9
I	3.0	2.4	6.7	5.0	11.2	7.1	6.9	7.8	2.6	3.6	3.6
K	7.5	7.9	20.6	22.7	0.0	0.0	0.0	0.0	1.8	0.9	2.6
L	9.3	8.6	7.2	6.5	27.1	39.5	27.3	23.4	8.0	5.3	5.0
M	1.5	2.5	2.2	2.4	2.2	1.7	3.1	3.3	1.6	4.1	1.9
N	3.2	2.7	6.3	5.8	0.1	0.3	0.3	0.2	2.1	2.9	5.6
P	7.5	8.0	3.0	3.6	0.6	1.1	1.4	1.0	9.8	7.8	15.7
Q	4.1	3.8	3.7	2.9	0.6	0.6	0.1	0.1	5.1	3.1	5.1
R	12.2	10.4	12.7	17.0	0.0	0.1	0.2	0.1	3.2	1.2	1.4
S	9.7	10.8	7.2	6.9	4.1	4.8	7.8	8.7	11.9	19.3	9.1
T	4.3	5.3	7.2	5.3	5.3	3.7	5.4	8.5	6.9	8.0	11.2
V	3.2	4.5	3.8	3.7	15.9	11.7	9.0	12.8	6.2	4.8	6.5
W	2.1	2.9	1.4	0.4	1.2	1.6	0.4	0.5	1.6	0.4	0.2
Y	2.2	1.7	2.1	1.9	3.1	0.8	0.3	0.3	1.5	1.2	0.7

Table 3: Amino acid distributions in the n-, h-, and c-regions of signal peptides assigned by the trained HMM. Results are shown for eukaryotes (**Euk**), Gram-negative bacteria (**G_{neg}**), and Gram-positive bacteria (**G_{pos}**). For the eukaryotes, n- and h-regions of signal anchors (**anc**) are also included (the concept of c-regions does not apply to signal anchors). The c-regions do *not* include the cleavage site consensus (position -3 through -1).

bacteria in Table 3: the positive charge in the h-region is more dominant in bacteria (up to 40% Lys+Arg for the Gram-positives), while eukaryotes have the most hydrophobic h-region with almost 40% Leu. In the c-region, the most conspicuous feature is the high occurrence of Gly and Pro—again, the Gram-positives stand out as the most extreme group with almost 16% Pro.

Note also the difference between eukaryotic signal anchors and signal peptides: the n-regions of signal anchors are more tolerant to the negatively charged residues Asp and Glu; and the h-region is less dominated by Leu, allowing higher proportions of other hydrophobic residues such as Ile and Val.

Conclusion

In terms of accuracy of the cleavage site prediction, the neural network-based SignalP is slightly better than the hidden Markov model described here. However, the HMM can be used to label the three different regions of a signal peptide, which yields quite surprising results. It was also demonstrated that the HMM can discriminate well between signal peptides, signal anchors, and other proteins. Because of the small number of known signal anchors, it is not likely that a neural network could be trained to discriminate so well.

An important application for the signal peptide HMM will be analysis of whole genomes and other large datasets derived from single species. Here, we have only considered differences between three large groups of organisms, but it is conceivable that further differences can be found within these groups. Statistical analysis suggests a difference between mammalian and plant signal peptides (von Heijne & Abrahmsén 1989), and there is experimental evidence that a yeast signal peptide can be non-functional in mammalian cells (Bird, Gething, & Sambrook 1987). The HMM can be used to divide the signal peptides into regions and thereby facilitate comparisons between these regions.

Archaea represent a special problem, since very few signal peptides are known experimentally from this domain of life, and therefore it is not clear which, if any, of the SignalP versions will apply. An analysis of signal-peptide like sequences from *Methanococcus jannaschii* suggests that its signal peptides differ from both their eukaryotic and bacterial counterparts (manuscript in preparation).

When analysing unknown sequences, it is important to note that the type II membrane proteins addressed in this work comprise only a small fraction of the transmembrane proteins. In particular, we have not tested the performance of neither the HMM nor the NN method on N-terminal parts of multispanning (type IV) transmembrane proteins. A combined model of signal peptides, signal anchors, and other transmembrane helices is clearly needed.

Finally, it has not escaped our notice that the two-peaked length distributions of h-regions might be correlated to a difference in translocation mechanism for two classes of signal peptides; but this question demands further investigation before anything definitive can be said.

Acknowledgments

We thank Gunnar von Heijne and Erik Sonnhammer for helpful discussions. This work was supported by the Danish

National Research Foundation.

References

Bairoch, A., and Apweiler, R. 1997. The SWISS-PROT protein sequence data bank and its supplement TrEMBL. *Nucleic Acids Res.* 25:31–36.

Bird, P.; Gething, M.-J.; and Sambrook, J. 1987. Translocation in yeast and mammalian cells: not all signal sequences are functionally equivalent. *J. Cell Biol.* 105:2905–2914.

Bird, P.; Gething, M.-J.; and Sambrook, J. 1990. The functional efficiency of a mammalian signal peptide is directly related to its hydrophobicity. *J. Biol. Chem.* 265:8420–8425.

Chou, M. M., and Kendall, D. A. 1990. Polymeric sequences reveal a functional interrelationship between hydrophobicity and length of signal peptides. *J. Biol. Chem.* 265:2873–2880.

Durbin, R. M.; Eddy, S. R.; Krogh, A.; and Mitchison, G. 1998. *Biological Sequence Analysis*. Cambridge University Press. To appear.

Eddy, S. R. 1996. Hidden Markov models. *Current Opinion in Structural Biology* 6:361–365.

Gribskov, M.; McLachlan, A. D.; and Eisenberg, D. 1987. Profile analysis: Detection of distantly related proteins. *Proc. Natl. Acad. Sci. USA* 84:4355–4358.

Krogh, A.; Brown, M.; Mian, I. S.; Sjölander, K.; and Haussler, D. 1994. Hidden Markov models in computational biology: Applications to protein modeling. *J. Mol. Biol.* 235:1501–1531.

Krogh, A. 1998. An introduction to hidden Markov models for biological sequences. In Salzberg, S.; Searls, D.; and Kasif, S., eds., *Computational Methods in Molecular Biology*. Elsevier. chapter 4. To appear.

Mathews, B. 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* 405:442–451.

Nielsen, H.; Engelbrecht, J.; von Heijne, G.; and Brunak, S. 1996. Defining a similarity threshold for a functional protein sequence pattern: The signal peptide cleavage site. *Proteins* 24:165–177.

Nielsen, H.; Brunak, S.; Engelbrecht, J.; and von Heijne, G. 1997. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.* 10:1–6.

Nilsson, I.; Whitley, P.; and von Heijne, G. 1994. The COOH-terminal ends of internal signal and signal-anchor sequences are positioned differently in the ER translocase. *J. Cell Biol.* 126:1127–1132.

Rabiner, L. R. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* 77(2):257–286.

Rapoport, T. A.; Jungnickel, B.; and Kutay, U. 1996. Protein transport across the eukaryotic endoplasmic reticulum and bacterial inner membranes. *Annu. Rev. Biochem.* 65:271–303.

Varshavsky, A. 1996. The N-end rule: functions, mysteries, uses. *Proc. Natl. Acad. Sci. USA* 93:12142–12149.

von Heijne, G., and Abrahmsén, L. 1989. Species-specific variation in signal peptide design. *FEBS Lett.* 244:439–446.

von Heijne, G. 1985. Signal sequences. The limits of variation. *J. Mol. Biol.* 184:99–105.

von Heijne, G. 1986a. Net N–C charge imbalance may be important for signal sequence function in bacteria. *J. Mol. Biol.* 192:287–290.

von Heijne, G. 1986b. A new method for predicting signal sequence cleavage sites. *Nucleic Acids Res.* 14:4683–4690.

von Heijne, G. 1988. Transcending the impenetrable: How proteins come to terms with membranes. *Biochim. Biophys. Acta* 947:307–333.