Protein Engineering vol.12 no.1 pp.3–9, 1999

Review

Machine learning approaches for the prediction of signal peptides and other protein sorting signals

Henrik Nielsen[1], Søren Brunak and Gunnar von Heijne[2]

Center for Biological Sequence Analysis Department of Biotechnology,
The Technical University of Denmark, DK-2800 Lyngby, Denmark and
[2]Department of Biochemistry, Arrhenius Laboratory, Stockholm University,
S-106 91 Stockholm, Sweden

[1]To whom correspondence should be addressed

Prediction of protein sorting signals from the sequence of amino acids has great importance in the field of proteomics today. Recently, the growth of protein databases, combined with machine learning approaches, such as neural networks and hidden Markov models, have made it possible to achieve a level of reliability where practical use in, for example automatic database annotation is feasible. In this review, we concentrate on the present status and future perspectives of SignalP, our neural network-based method for prediction of the most well-known sorting signal: the secretory signal peptide. We discuss the problems associated with the use of SignalP on genomic sequences, showing that signal peptide prediction will improve further if integrated with predictions of start codons and transmembrane helices. As a step towards this goal, a hidden Markov model version of SignalP has been developed, making it possible to discriminate between cleaved signal peptides and uncleaved signal anchors. Furthermore, we show how SignalP can be used to characterize putative signal peptides from an archaeon, *Methanococcus jannaschii*. Finally, we briefly review a few methods for predicting other protein sorting signals and discuss the future of protein sorting prediction in general.

Introduction

Subcellular protein sorting, i.e. the processes through which proteins are routed to their proper final destination within a cell, is a fundamental aspect of cellular life. In many cases, sorting depends on ‘signals’ that can already be identified by looking at the primary structure of a protein. Thus, targeting to the secretory pathway, to mitochondria and to chloroplasts normally depends on an N-terminal presequence or targeting peptide that can be recognized by receptors on the surface of the appropriate organelle. After targeting, membrane-embedded translocation machineries ensure the delivery of the protein to the interior of the organelle.

By definition, the cell can recognize all kinds of protein sorting signals with almost 100% selectivity and specificity—the level of mis-sorting *in vivo* appears to be very low, although this aspect of the problem has not been studied in detail. Given that the sorting signals mentioned above seem to be, at least to a good approximation, defined by a linear, N-terminal stretch of the polypeptide, it would appear that we should be able to devise sequence-based methods that can recognize these signals with an efficiency approaching that of the cell itself. If such methods can be developed, they will clearly be of major use for genome analysis and automatic database annotation; at the same time, these massive data analysis tasks necessitate very accurate prediction methods.

While prediction of sorting signals has a long history, started by the early work on secretory signal peptides (von Heijne, 1983; McGeoch, 1985; von Heijne, 1986b), it is only with the application of modern machine learning techniques, such as neural networks (NNs) and hidden Markov models (HMMs), that we seem to be approaching the necessary levels of accuracy (Baldi and Brunak, 1998; Durbin *et al.*, 1998). Machine-learning techniques are ideally suited for pattern recognition tasks where relatively large amounts of data are present and where the patterns are ‘noisy’ and not easily described by a compact set of rules. The fundamental idea behind these approaches is to learn to discriminate automatically from the data, using experimentally verified examples, which most often are extracted from large public sequence and structure databases. While HMMs are best at recognizing, in an ‘elastic’ fashion, the sequential pattern in the amino acids or nucleotides, the NN algorithms are better at handling sequence features correlated over a longer range, especially if there is some degree of conservation in the positioning of the relevant features. Together, the NN and HMM methods can therefore handle a very substantial part of the sequence diversity created by evolution that is characteristic for many complex biological mechanisms. Thus, there now exist quite reliable machine learning-based methods for the identification of both secretory signal peptides (SPs), mitochondrial targeting peptides (mTPs) and chloroplast transit peptides (cTPs).

In this review, we will concentrate on the present status and future perspectives of SP prediction—in particular the developments and applications of our own method, SignalP, since it was published in Protein Engineering two years ago (Nielsen *et al.*, 1997a). Several NN-based methods for prediction of SPs have been developed (Ladunga *et al.*, 1991; Schneider and Wrede, 1993), but only SignalP is publicly available. SignalP has been used extensively since it was made available over the internet, but the first version has some important shortcomings that necessitate further development and integration with other prediction methods. In addition, we will review a couple of methods for predicting other protein sorting signals, and discuss some general aspects of sorting signal prediction.

Constructing the training set for machine learning methods

While different algorithms within the broad range of machine learning methods available will have different advantages in terms of their pattern recognition abilities, they are all driven by the data used to train them. The selection of the training set is arguably the most important part in the construction of a prediction method. No matter how sophisticated the algorithm, with poor training data one will get poor results. In the cases discussed here, SWISS-PROT (Bairoch and Apweiler, 1997) is the natural primary source of sequence data, but even in a well-curated database such as this, one cannot take all the sequence annotations at face value.

Another problem is that a sequence database always contains numerous examples of genes belonging to gene families and homologous genes from various organisms. This can lead to statistical results that are biased for the over-represented sequences, and the performance of prediction methods will be overestimated if the test set contains sequences closely related to those used in the training. Thus, after selecting an initial set of sequences from SWISS-PROT, one has to remove homologous sequences (unless the training algorithm can deal with redundant data sets) using, for example, the Hobohm redundancy reduction method (Hobohm *et al.*, 1992). The question of when two sequences are 'too closely related' to be kept within the reduced data set is far from trivial. For the SignalP data set, the similarity threshold is found from the principle that if it is possible to infer the position of the cleavage site in one SP by alignment to another SP, the sequences are too similar. Another approach, which uses the statistical theory of local alignments (Altschul and Gish, 1996), is to fit the alignment scores to an extreme value distribution and choose a threshold value above which there are more observations than expected from the distribution (Pedersen and Nielsen, 1997).

Unless the remaining set at this point is prohibitively large, it should be checked by hand against the primary publications. In our experience, features like cleavage sites for sorting signals are not always correctly annotated: sites not listed as 'putative' may in fact be based only on an informed guess (or even an existing prediction method), and experimentally verified sites are sometimes incorrectly entered into the database (database 'typos'). In a recent study of chloroplast transit peptides (O.Emanuelsson, H.Nielsen and G.von Heijne, manuscript submitted), we had to remove around 10% of the sequences in our homology-reduced data set for such reasons. Even experimentally verified data may be wrong if the interpretation of the results has been faulty. The most relevant example in this context is that an N-terminus of a mature protein, confirmed by amino acid sequencing, might derive not from cleavage by the signal peptidase but from a subsequent cleavage by another protease in the secretory pathway.

If the data set is too large to allow for manual inspection of all entries, some suspicious looking examples may be identified by automated methods. One possibility is to use alignments of the unreduced set to single out pairs of sequences that show a very high similarity but discrepancies in assignment of subcellular location or cleavage site position (Nielsen *et al.*, 1996). Another method is to use the training algorithm itself to pick out cases which are more difficult to learn than others (Brunak, 1993). Both these approaches are necessarily biased; the first will never be able to pick up errors in sequences with no matching homologues, and both can fail to recognize systematic errors that occur in several entries. Still, experience has shown that machine learning methods can serve as extremely useful tools for data set validation; in several cases, NNs have been able to detect errors caused both by simple misprints and by incorrect interpretation of experiments (Brunak *et al.*, 1990a,b).

Another aspect of the choice of training set is whether sequences from all, some subset of, or only a single organism should be included. If there is enough data, organism-specific methods should be expected to perform better than more general ones, but in most cases it is not possible to be this restrictive.

In the SignalP work, we trained two species-specific versions on human and *Escherichia coli* SPs, and concluded that there was no significant gain in performance when testing with networks trained on a single-species data set relative to networks trained on larger groups (Nielsen *et al.*, 1997a). This result is not definitive, however. The reason why the *E.coli*-specific network did not show an improvement compared with one trained on a larger set of Gram-negative SPs might simply be that the *E.coli* set at that time was too small to achieve the same relative performance. Regarding the human-specific network, one should note that the eukaryotic set is dominated by mammals, i.e. rather close relatives to humans; and we cannot exclude the possibility that signal peptides from, for example, yeast (which are relatively underrepresented in the data set), are significantly different from those of mammals. Nevertheless, genomic sequencing opens up the possibility of constructing species-specific versions of the basic algorithm, perhaps by a bootstrapping procedure where a more general version trained on, for example, all eukaryotic sequences, is used to extract an initial set of reliably predicted sequences from, for example, yeast, which is then used to iteratively train a species-specific version.

Current status of the SignalP method

SignalP is a typical example of a NN-based method, and three versions trained on different data sets (eukaryotes, Gram-negative and Gram-positive bacteria) are available. These three versions reflect significant differences in the characteristics of signal peptides from these groups of organisms, and each gives a better performance than a method trained on all groups together. They also provide the opportunity to test the efficiency of a given signal peptide sequence in a non-native host. For example, a human sequence can be analysed by the Gram-positive version of the method and thus give an indication of how effective the sequence will appear in a production organism, say, *Bacillus subtilis*. If it appears to have a low degree of 'signal peptide-ness' in the new host, it can subsequently be engineered such that the SP sequence will optimally match the N-terminus of the mature protein.

SignalP combines two different NNs, one that has been trained to classify each residue in the sequence as either belonging or not belonging to a SP (S-score), and one that has been trained only to recognize the site at the C-terminal end of the SP that is cleaved by the signal peptidase enzyme after targeting (C-score). Cleavage-site prediction performance is significantly enhanced by penalizing C-score peaks that are far away from the transition region between the SP and the mature polypeptide identified by the S-score. This is formalized by using the 'Y-score', a geometric average of the C-score and a numerical derivative of the S-score. In the example shown in Figure 1, the C-score has two peaks, where the upstream one is slightly higher but the downstream one occurs in the transition zone of the S-score and therefore has a higher Y-score.

A prediction for the existence of a SP can be made by the maximal value of the C-, S- and Y-scores, or the mean S-score between the N-terminus and the predicted cleavage site. Of these, the maximal Y-score or the mean S-score give the best discrimination performance, but all four values are reported in the output. A more thorough description of the SignalP

Table I. Performances of SignalP in the neural network (NN) and hidden Markov model (HMM) versions

	Task	Cleavage site location			Discrimination			
Method	Data (release)	Euk	G_{neg}	G_{pos}	SP/non-sec Euk	G_{neg}	G_{pos}	SP/SA Euk
NN	29	70.2%	79.3%	67.9%	0.97	0.88	0.96	(0.39)
NN	35	72.4%	83.4%	67.5%	0.97	0.89	0.96	(0.39)
HMM	35	69.5%	81.4%	64.5%	0.94	0.93	0.96	0.74

The column labeled ‘Data’ refers to the SWISS-PROT release number, so that the first line (NN, 29) show the performance of the original SignalP (Nielsen *et al.*, 1997a). Data sets are divided into 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, and discrimination values between sequence types are given as correlation coefficients (Mathews, 1975). The sequence types are signal peptides (SP), soluble non-secretory—i.e. cytoplasmic or nuclear—proteins (non-sec), and signal anchors (SA). For SignalP-NN, cleavage site location is predicted by maximal Y-score, and discrimination performed using mean S-score; discrimination values for signal anchors are in parentheses because signal anchors were not included as negative examples in the NN training set. All values are averages over five cross-validation sets.

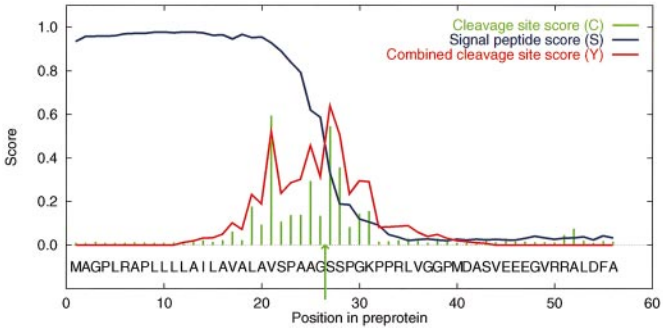

Fig. 1. An example of a prediction for a protein with a known signal peptide, human cystatin C precursor. The values of the C-score (output from cleavage site networks), S-score (output from signal peptide networks) and Y-score (combined cleavage site score, $Y_\mathrm{i} = \sqrt{C_\mathrm{i}\Delta_d S_i}$) are shown for each position in the sequence, and the true cleavage site is marked with an arrow. In this example with two C-score peaks, the cleavage site would be incorrectly predicted when relying on the C-score alone, but the combined Y-score is able to predict it correctly. (Note: the C-score is defined to be high for the position immediately *after* the cleavage site, i.e. the first position in the mature protein.)

architecture and the definition of the various measures can be found elsewhere (Nielsen *et al.*, 1997b).

The performance values of SignalP are shown in Table I, both for the original version and for a version retrained on a new data set, based on SWISS-PROT release 35 instead of 29. Note that the performance for cleavage site location has improved. Since the old and new data sets are extracted by the same method, and the sizes have changed only slightly, the most probable explanation for the improvement is that the quality of SWISS-PROT annotations concerning SPs are better in the newer version.

There are two important points to be made about the performance values. One the one hand, they should be regarded as minimal, because they are test set performances (averaged over five cross-validation sets), where the homology reduction of the data has assured that the similarity between training and test sets is so low that the correct cleavage sites cannot be found by alignment (Nielsen *et al.*, 1996). These performance values should therefore be expected for a protein unrelated to anything in the data sets, while prediction accuracy on sequences with some similarity to the sequences in the data sets will in general be much higher. For example, the accuracy of cleavage site location (original release 29 version) goes up to 76.8, 85.0 and 76.6% (for eukaryotes, Gram-positive and Gram-negative bacteria, respectively) when the data sets are tested on the full ensemble.

On the other hand, the performance values given in Table I are calculated under two limiting assumptions: that the correct N-terminus of the protein in question is known, and that the sequence does not contain an N-terminal transmembrane helix. The data sets on which SignalP is trained and tested contain only the N-terminal part (up to 70 amino acids) of each protein, and transmembrane proteins were not included in the negative set. The decision to use only the N-terminal part of each protein was based on the idea that SignalP should reproduce the recognition task met by the cell *in vivo*, where SP cleavage takes place only within a certain range from the N-terminus. The reason for the lack of transmembrane helices in the negative set is more practical: it is very hard to ensure that there is experimental evidence for absence of cleavage of a transmembrane protein. For a subset of transmembrane proteins, however, we have a reliable set: eukaryotic signal anchors (see below).

These two points constitute a problem for the application of SignalP to genome and EST data. As an illustration of this, the scanning of the *Haemophilus influenzae* genome which we reported in the SignalP paper (Nielsen *et al.*, 1997a) produced a remarkably large variation in the estimate of the proportion of proteins with SPs: from 14% if using the maximal Y-score as discriminator, to 28% when using the maximal S-score, even though all these measures give high discrimination performances when used on the SignalP data set. This means that the performance of (at least) one of these measures is considerably lower when applied to genome data; and that SignalP, when used for this purpose, should ideally be combined with a transmembrane helix prediction and a start codon prediction.

SignalP-HMM: distinguishing signal peptides from signal anchors

Some proteins have sequences that initiate translocation in the same way as SPs 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. The uncleaved signal peptide is known as a signal anchor (SA), and the resulting protein is known as a type II membrane protein. SAs differ from SPs in other respects than the cleavage sites: they

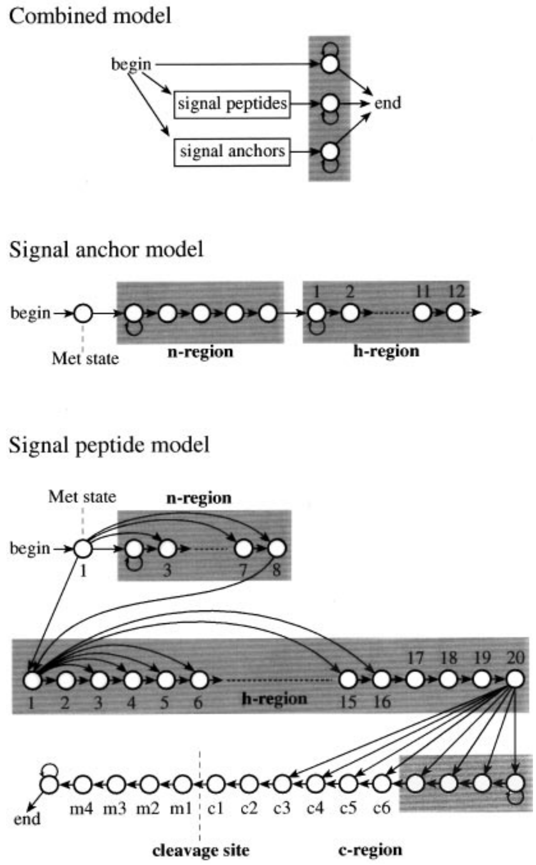

Fig. 2. The architecture of the hidden Markov model for signal peptide and signal anchor prediction (SignalP-HMM). (**Top**) The diagram shows how the combined model is put together from a signal peptide model, an anchor model and a null model representing non-secretory proteins. The model of signal anchors (**Center**) has only two types of states (n- and h-region), while the signal peptide model (**Bottom**) additionally contains a model of the c-region and the cleavage site. The states in a shaded box are tied to each other, i.e. are forced to have the same amino acid distribution.

have longer hydrophobic stretches—the length is typically the same as that of a transmembrane α-helix—and the region N-terminal of the hydrophobic stretch can also be much longer. Interestingly, experiments have shown that it is possible to convert a cleaved SP into an uncleaved SA merely by lengthening the hydrophobic region (Chou and Kendall, 1990; Nilsson *et al.*, 1994).

The discrimination between SAs and SPs has proved to be very difficult for the neural network: approximately 50% of the SAs are predicted as SPs according to the mean S-score. Since both the C-score and the S-score are calculated from sequence windows of a limited width, a feature such as region length is difficult to represent in the input. To solve this problem, we have developed SignalP-HMM, a HMM architecture for SPs and SAs (Figure 2).

The advantage of the HMM method in this context is that it does not use windows of a fixed width, but threads an entire sequence through a trained model. An HMM is a chain of 'states', each with a characteristic amino acid distribution, with transitions that specify possible orders of states. Thus, a HMM can model sequences of varying length by transitions that skip or repeat states. By assigning states to known regions of the signal to be modeled, biological knowledge can be built into the HMM.

Secretory signal peptides have three distinct regions—an N-terminal positively charged n-region, a central hydrophobic h-region, and a C-terminal c-region encompassing the signal peptidase cleavage site (von Heijne, 1985). Each of these is represented by a separate part of the model: the n- and h-regions are modeled in a simple way, with all states having the same amino acid frequencies, while the region around the cleavage sites is modeled in more detail (essentially like a weight matrix). Signal anchors have both an n- and an h-region, and no cleavage site. By having two parallel submodels of the HMM, it is possible to represent differences in both length distribution and amino acid frequencies between the n- and h-regions of SPs and SAs. A third branch (actually, just a shortcut) is added to represent those sequences that are neither SPs nor SAs. When threading a sequence through this model, one of the three branches is chosen, and this serves as the prediction of protein type. Additionally, this method provides an objective way to delineate the n-, h- and c-regions in a SP, and it may thus be used to compare the overall design of SPs from different organisms.

SignalP-HMM is able to discriminate between SPs and SAs with a correlation coefficient of 0.74 (see Table I)—far from perfect, but much better than with the NNs. In a sense, this comparison is not quite fair, because the SAs were not used explicitly as negative examples during training of the NN, but this would have been problematic given the small size of the SA set. With the HMM, it is easy to take this limitation into account by using a simpler submodel (with a smaller number of free parameters) in the SA branch than in the SP branch. Regarding the identification of SPs versus soluble non-secretory proteins, the HMMs perform on a par with the NNs—and for Gram-negative bacteria even better—but they are less accurate for cleavage site prediction, see Table I.

Type II membrane proteins constitute only a minor fraction of transmembrane proteins. When scanning genome data, it is desirable to distinguish SPs not only from SAs, but also from other types of transmembrane helices. It is advisable to combine SignalP with one of the available prediction methods for transmembrane helices, e.g. PHDhtm (Rost *et al.*, 1996) or TopPred (von Heijne, 1992). Of course, it would be preferable, both for usage on large data sets and from a theoretical point of view, to obtain one prediction of the presence and location of both SPs and transmembrane helices in the sequence. To this end, we plan to build an integrated HMM architecture based on SignalP-HMM and an HMM-based transmembrane helix prediction method, TMHMM (Sonnhammer *et al.*, 1998).

Start codon prediction

A difficulty for prediction of SPs—or any other N-terminal sorting signals—is that the position of the N-terminus in the preprotein is rarely known experimentally. This is particularly troublesome when using genomic data, where protein coding regions are predicted by gene finding algorithms containing numerous potential sources of error. Wrong start codon assignments can produce false negatives, since the resulting sequence may either contain only a partial SP sequence, or a SP plus a stretch of irrelevant amino acid sequence (derived from DNA which is untranslated *in vivo*) without SP characteristics.

For expressed sequence tags (ESTs) the problem can be even worse, since it is very difficult to decide whether a given sequence includes the start codon at all—it might be entirely untranslated, or correspond to an internal stretch of a protein. The last case can also produce false positive predictions, since non-cytoplasmic ends of transmembrane helices are often rather similar to SP cleavage sites, and the SignalP networks have never been trained to avoid SPs here.

Therefore, it would be desirable to have a method which, given a nucleotide sequence, would provide a prediction of both ends of a SP, i.e. the start codon and the cleavage site. Such a method does not exist yet, but a partial solution would be a score describing the probability that any given triplet is the start codon. To this end, we have developed a NN-based method for start codon prediction in eukaryotes, NetStart (Pedersen and Nielsen, 1997). It is trained to recognize the start codon AUG against all other AUG triplets in the mRNA sequence. It performs this task by using both local context—the Kozak box (Kozak, 1984)—and long-range context in the form of implicit reading frame detection. NetStart is designed to work with EST or cDNA data; for use with genomic DNA, the possible occurrence of introns shortly downstream of the start codon could be detrimental to the prediction.

Statistical analyses (A.G.Pedersen *et al.*, manuscript in preparation) have shown that the local start codon context varies widely between different systematic groups of eukaryotes. The current NetStart 1.0 contains only two organism-specific versions, for vertebrates and *Arabidopsis thaliana*, but more will be added in future releases. Although NetStart 1.0 should be regarded as a ‘first attempt’ at this problem, it does show test set performances, measured by correlation coefficient, of 0.62 for vertebrates and 0.71 for *A.thaliana*.

Signal peptides of Archaea

Secretory SPs from eukaryotes and bacteria are well described, but only very few experimental examples are known from the third domain of life, the archaea (formerly known as archaebacteria). Although being prokaryotic, they show greater similarity in many respects to eukaryotes than to bacteria, especially concerning informational cellular processes such as replication and translation (Olsen and Woese, 1997). Furthermore, their membranes exhibit very specialized properties not found in other organisms. It is therefore not clear which, if any, of the three current organism-specific SignalP versions is valid for identification of archaeal SPs.

We used a ‘consensus’ between the three SignalP versions in a first attempt at characterizing the SPs of *Methanococcus jannaschii*, the first archaeon to be completely sequenced (Bult *et al.*, 1996). SPs should indeed be expected in this organism: a signal peptidase has been identified by homology in the genome, and it shows greater homology to its eukaryotic than to its bacterial counterpart. The underlying idea is that if we are able to find sequences in the genome which could function as SPs in all other domains of life (i.e. in eukaryotes and both groups of bacteria), they would presumably function as signal peptides in *M.jannaschii* as well.

Methanococcus jannaschii SPs might have been predicted by alignment to known SPs from other organisms, if significant matches to experimentally verified secretory proteins including the SP region could be found. We made local pairwise alignments between all the predicted *M.jannaschii* protein sequences and all sequences in the SignalP data set, but found only insignificant matches. Even the best pairwise alignment scores were considerably lower than the threshold required for using a local alignment of two SP sequences to predict the location of the cleavage site (Nielsen *et al.*, 1996). This shows that we cannot expect to find *M.jannaschii* SPs by alignment—a prediction method is indeed necessary for this task.

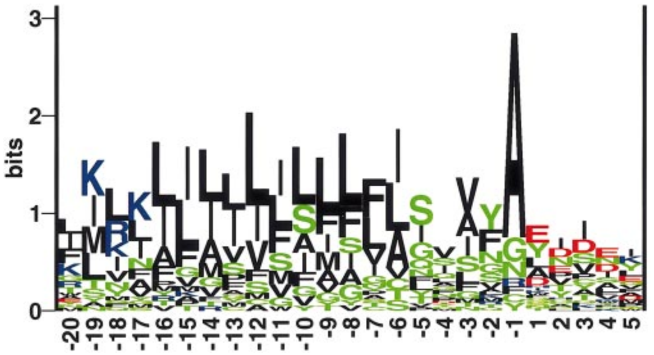

Fig. 3. A sequence logo of 34 predicted signal peptides from *Methanococcus jannaschii*, aligned by their cleavage sites (no gaps). Positively and negatively charged residues are shown in blue and red respectively, while uncharged polar residues are green and hydrophobic residues are black.

We selected sequences where both the maximal Y-score and the mean S-score were above their cut-off values for all three SignalP versions (eukaryotic, Gram-positive and Gram-negative). This is a very conservative criterion: when tested on the SignalP data sets, it accepts 75% Gram-negative, 66% Gram-positive and only 39% of the eukaryotic SPs. Used on the *M.jannaschii* genome, it yielded 34 putative SPs, none of which had a known subcellular location. This number is too small to train a species-specific neural network (it might be used for an HMM but this has not yet been implemented), but it is enough to draw a few tentative conclusions about *M.jannaschii* SPs.

The 34 sequences were divided into n-, h- and c-regions, and the amino acid content compared with that of eukaryotes and bacteria. The *H.influenzae* genome (Fleischmann *et al.*, 1995) served as a reference example of a Gram-negative bacterium. In Figure 3, the 34 putative *M.jannaschii* SPs are represented as a sequence logo, i.e. a sequence of stacked letters, where the total height of the stack at each position shows the amount of information (conservation), while the relative height of each letter shows the relative abundance of the corresponding amino acid (Schneider and Stephens, 1990). When compared with logos of eukaryotic or bacterial SPs (Nielsen *et al.*, 1997a), the following characteristics are observed.

In the n-region, the content of Lys is very high, while Arg is relatively rare. A positively charged n-region is also found in bacterial SPs, but in these Arg and Lys are present in more equal proportions. The Lys content of *M.jannaschii* n-regions is approximately 30% compared with 20% in *H.influenzae*. A very characteristic feature is the high content of Ile in the h-region. This is not limited to signal peptides, as Ile is strongly over-represented in *M.jannaschii* as compared with *H.influenzae* also in transmembrane regions (16 versus 12%) and in the genome as a whole (10.5 versus 7.1%). However, the difference is more drastic for the h-regions (22 versus 11%).

In the c-region, the dominance of Ala at position –1 is typical for both bacterial and eukaryotic signal peptide cleavage sites, whereas the tolerance of other uncharged residues, such as Val, Leu and Ile, at –3 and the short length of the c-region clearly suggest a eukaryotic type of cleavage site. Around the cleavage site, a unique feature is also found: a high occurrence of Tyr (8% of the c-regions as opposed to 2% in *H.influenzae*), particularly visible at positions +1 and –2. This seems to be specific for SPs, since the general Tyr content is only slightly higher in *M.jannaschii* than in *H.influenzae* (4.3 versus 3.3%). Finally, the occurrence of negatively charged residues in the first few positions of the mature protein has previously been noted for bacterial but not for eukaryotic signal peptides (von Heijne, 1986a).

In conclusion, our analysis suggests that SPs from an archaeon have a eukaryotic-looking cleavage site, a bacterial-looking charge distribution and a unique composition of the hydrophobic region. The statistical description is of course to some extent affected by the fact that we use a consensus method, which only finds signal peptides and cleavage sites that would be acceptable in both eukaryotes and bacteria; chances are that signal peptides peculiar to archaea have gone undiscovered. In other words, we have if anything underestimated the unique characteristics of the *M.jannaschii* signal peptides.

Other protein sorting prediction methods

ChloroP is the equivalent of SignalP for predicting chloroplast transit peptides (cTPs), and has been constructed in much the same way (O.Emanuelsson, H.Nielsen and G.von Heijne, manuscript submitted). Two novel aspects are that the yes/no cTP prediction is based on a NN trained on the S-score outputs from the basic NN, and that the cleavage site prediction is not done using a NN but by a simple weight matrix. The weight matrix approach was chosen since a recent experimental study of the cTP processing enzyme stromal processing peptidase (SPP) suggested that the mature N-terminus of chloroplast proteins is often generated by an ill-defined proteolytic removal of one or a few extra residues after the initial SPP cleavage (Richter and Lamppa, 1998). Since the cleavage sites given in SWISS-PROT are based on amino acid sequencing of mature chloroplast proteins, they will, in general, not correspond to the SPP cleavage sites. To get around this problem, we used MEME (Bailey and Elkan, 1994), an automatic motif-finding algorithm that does not require pre-aligned sequences, to construct a weight matrix for the SPP cleavage site. ChloroP can distinguish between cTPs and other proteins with a correlation of 0.76, and it can locate the cleavage site within three residues from the annotated position in about 60% of the cTPs.

The currently most developed method to predict mTPs is based on a linear combination of a number of sequence characteristics such as amino acid abundance, maximum hydrophobicity and maximum hydrophobic moment that are combined into an overall score (Claros and Vincens, 1996). Preliminary work using the same NN approach as for ChloroP suggests that similar performance levels can be reached using machine learning (our unpublished data).

In addition to the recognition of the sorting signals, prediction of protein sorting can exploit the fact that proteins of different subcellular compartments differ in global properties such as amino acid composition and residue-pair frequencies. While the signal prediction methods are probably closer to mimicking the information processing in the cell, methods based on global properties can complement imperfect signal-based methods, especially on incomplete sequences. Specifically, a composition-based method for recognizing extracellular proteins can be used without knowledge of the N-terminus, and could, for example, give correct predictions for EST-derived protein fragments where the signal peptide has not even been sequenced. The drawback is that such methods will not be able to distinguish between very closely related proteins that differ in the presence or absence of a SP. Most of the work on such methods has been based on traditional statistics (Nakashima and Nishikawa, 1994; Cedano *et al.*, 1997), but machine learning has been employed in the NNPSL method, which uses NNs trained on overall amino acid composition to predict location to three (bacteria) or four (eukaryotes) possible subcellular compartments (Reinhardt and Hubbard, 1998).

The PSORT program (Nakai and Kanehisa, 1992; Horton and Nakai, 1997) is an integrated system of several prediction methods, using both sorting signals and global properties. Some of the components are developed within the PSORT group, others are implementations of methods published elsewhere. PSORT is the only publicly available system that shows this degree of integration, and it includes sorting predictions that are not found elsewhere (e.g. nuclear or peroxisomal targeting). However, it does not include the newest machine-learning methods, which means that PSORT prediction of the more extensively studied protein sorting problems, e.g. SPs or transmembrane helices, is in many cases not the best available.

The future

With the recent advances in prediction methods for protein sorting, the vision of a computer program that is able to predict the subcellular location of almost any given protein with high confidence seems not entirely unrealistic. This would be an integrated system of sorting signal predictors and methods based on overall amino acid composition, and as described above, start codon prediction and transmembrane helix prediction should be included. A major use of such a program would be automatic annotation of sequence databases, including complete genomes.

On the other hand, one big integrated system of all methods may not be the most desirable solution for all users. For automated annotation of very large data sets, integrated prediction systems are of course preferable, but the biologist working on one specific gene might be better off considering comprehensive graphical output from several prediction methods separately, and then deciding which conclusion should be drawn from the possibly conflicting predictions. In some cases (rare but interesting), the biologically correct answer will be something not anticipated by the method builders (e.g. dual targeting, double cleavage, non-standard use of sorting machineries), and uncritical use of a totally integrated prediction system could actually block new discoveries instead of promoting them.

Finally, any given application will require careful consideration of how to strike the best balance between sensitivity and specificity. For gene hunting, one may want high sensitivity (i.e. few false negatives) in order not to miss interesting candidate genes, whereas for database annotation it may be more prudent to ask for high specificity (i.e. few false positives) even if this will leave many sequences unannotated.

The trade-off between sensitivity and specificity illustrates a common aspect in the evaluation of prediction methods. Performances are given as percent correct, correlation coefficients etc., but these depend on the choice of cut-off and the definition of positive and negative data sets. In the signal peptide case, it is quite clear what the positive data sets should be, although it may be argued whether, for example, bacterial lipoproteins should be considered as positive examples. On the other hand, there are many questions to be asked about negative examples: should they comprise only soluble cytoplasmic and nuclear proteins, or include transmembrane and membrane-associated proteins? Should they be limited to N-terminal parts or include entire protein chains? There is no single correct answer to questions like these, which makes comparison of performances of different methods a very tricky business.

Since numerical performance measures are mandatory for deciding whether methods have improved, the task of defining such measures is very important, and much more work is needed within the bioinformatics field in order to arrive at common testing standards for method comparison (Nielsen *et al.*, 1996). However, we feel that the most informative test of the performance and applicability of a sequence-based prediction method is carried out by making it available to the biological community, both in academia and in industry, e.g. by implementing it as a server or a portable program. The feedback from users, either directly, or implicitly via usage and citation statistics, can tell us more about the quality of our bioinformatics work than percentages and correlation coefficients will ever be able to.

Availability of methods

SignalP, TMHMM, NetStart and ChloroP are all available under the prediction server page of Center for Biological Sequence Analysis (http://www.cbs.dtu.dk/services/). For transmembrane helix prediction, two possibilities in addition to TMHMM (our apologies to several others not mentioned here) are PHDhtm (http://www.embl-heidelberg.de/predict-protein/) and TopPred (http://www.biokemi.su.se/server/toppred2/). PSORT is found at http://psort.nibb.ac.jp/, and NNPSL at http://predict.sanger.ac.uk/nnpsl/.

Acknowledgements

We would like to thank our co-workers in the protein sorting field: Olof Emanuelsson (ChloroP), Erik Sonnhammer (TMHMM), Anders Krogh (TMHMM, SignalP-HMM) and Anders Gorm Pedersen (NetStart). Figure 2 was made by Anders Krogh. This work was supported by grants from the Danish National Research Foundation to SB and HN, and from the Swedish Natural and Technical Sciences Research Councils to GvH.

References

Altschul,S. and Gish,W. (1996) *Methods Enzymol.*, **266**, 460–480.

Bailey,T. and Elkan,C. (1994) *ISMB*, **2**, 28–36.

Bairoch,A. and Apweiler,R. (1997) *Nucleic Acids Res.*, **25**, 31–36.

Baldi,P. and Brunak,S. (1998) *Bioinformatics: The Machine Learning Approach*. MIT Press, Cambridge.

Brunak,S. (1993) In Soumpasis,D. and Jovin,T. (eds) *Computation of Biomolecular Structures—Achievements, Problems and Perspectives.* Springer-Verlag, Berlin, pp. 43–54.

Brunak,S., Engelbrecht,J. and Knudsen,S. (1990a) *Nature*, **343**, 123.

Brunak,S., Engelbrecht,J. and Knudsen,S. (1990b) *Nucleic Acids Res.*, **18**, 4797–4801.

Bult,C.J., White,O., Olsen,G.J. *et al.* (1996) *Science*, **273**, 1058–1073.

Cedano,J., Aloy,P., Pérez-Pons,J. and Querol,E. (1997) *J. Mol. Biol.*, **266**, 594–600.

Chou,M.M. and Kendall,D.A. (1990) *J. Biol. Chem.*, **265**, 2873–2880.

Claros,M.G. and Vincens,P. (1996) *Eur. J. Biochem.*, **241**, 779–786.

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

Fleischmann,R.D., Adams,M.D., White,O. *et al*. (1995) *Science*, **269**, 496–512.

Hobohm,U., Scharf,M., Schneider,R. and Sander,C. (1992) *Protein Sci.*, **1**, 409–417.

Horton,P. and Nakai,K. (1997) *ISMB*, **5**, 147–152.

Kozak,M. (1984) *Nucleic Acids Res.*, **12**, 857–872.

Ladunga,I., Czakó,F., Csabai,I. and Geszti,T. (1991) *CABIOS*, **7**, 485–487.

Mathews,B. (1975) *Biochim. Biophys. Acta*, **405**, 442–451.

McGeoch,D.J. (1985) *Virus Res.*, **3**, 271–286.

Nakai,K. and Kanehisa,M. (1992) *Genomics*, **14**, 897–911.

Nakashima,H. and Nishikawa,K. (1994) *J. Mol. Biol.*, **238**, 54–61.

Nielsen,H., Brunak,S., Engelbrecht,J. and von Heijne,G. (1997a) *Protein Engng*, **10**, 1–6.

Nielsen,H., Brunak,S., Engelbrecht,J. and von Heijne,G. (1997b) *Int. J. Neural Sys.*, **8**, in press.

Nielsen,H., Engelbrecht,J., von Heijne,G. and Brunak,S. (1996) *Protein*s, **24**, 165–177.

Nilsson,I., Whitley,P. and von Heijne,G. (1994) *J. Cell Biol.*, **126**, 1127–1132.

Olsen,G. and Woese,C. (1997) *Cell*, **89**, 991–994.

Pedersen,A.G. and Nielsen,H. (1997) *ISMB*, **5**, 226–233.

Reinhardt,A. and Hubbard,T. (1998) *Nucleic Acids Res.*, **26**, 2230–2236.

Richter,S. and Lamppa,G. (1998) *Proc. Natl Acad. Sci. USA*, **95**, 7463–7468.

Rost,B., Fariselli,P. and Casadio,R. (1996) *Protein Sci.*, **5**, 1704–1718.

Schneider,G. and Wrede,P. (1993) *J. Mol. Evol.*, **36**, 586–595.

Schneider,T.D. and Stephens,R.M. (1990) *Nucleic Acids Res.*, **18**, 6097–6100.

Sonnhammer,E.L., von Heijne,G. and Krogh,A. (1998) *ISMB*, **6**, 175–182.

von Heijne,G. (1983) *Eur. J. Biochem.*, **133**, 17–21.

von Heijne,G. (1985) *J. Mol. Biol.*, **184**, 99–105.

von Heijne,G. (1986a) *J. Mol. Biol.*, **192**, 287–290.

von Heijne,G. (1986b) *Nucleic Acids Res.*, **14**, 4683–4690.

von Heijne,G. (1988) *Biochim. Biophys. Acta*, **947**, 307–333.

von Heijne,G. (1992) *J. Mol. Biol.*, **225**, 487–494.

Received November 23, 1998; accepted November 24, 1998