Protein Engineering vol.10 no.1 pp.1–6, 1997

SHORT COMMUNICATION

Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites

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

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

[1]Present address: Novo Nordisk A/S, Scientific Computing, Building 9M1, Novo Alle, DK-2880 Bagsværd, Denmark

We have developed a new method for the identification of signal peptides and their cleavage sites based on neural networks trained on separate sets of prokaryotic and eukaryotic sequence. The method performs significantly better than previous prediction schemes and can easily be applied on genome-wide data sets. Discrimination between cleaved signal peptides and uncleaved N-terminal signal-anchor sequences is also possible, though with lower precision. Predictions can be made on a publicly available WWW server.

Keywords: cleavage sites/protein sorting/secretion/signal peptide

Introduction

Signal peptides control the entry of virtually all proteins to the secretory pathway, both in eukaryotes and prokaryotes (Gierasch, 1989; von Heijne, 1990; Rapoport, 1992). They comprise the N-terminal part of the amino acid chain and are cleaved off while the protein is translocated through the membrane. The common structure of signal peptides from various proteins is commonly described as a positively charged n-region, followed by a hydrophobic h-region and a neutral but polar c-region. The (–3,–1) rule states that the residues at positions −3 and −1 (relative to the cleavage site) must be small and neutral for cleavage to occur correctly (von Heijne, 1983, 1985).

A strong interest in the automated identification of signal peptides and the prediction of their cleavage sites has been evoked not only by the huge amount of unprocessed data available, but also by the industrial need to find more effective vehicles for the production of proteins in recombinant systems. The most widely used method for predicting the location of the cleavage site is a weight matrix which was published in 1986 (von Heijne, 1986). This method is also useful for discriminating between signal peptides and non-signal peptides by using the maximum cleavage site score. The original matrices are commonly used today, even though the amount of signal peptide data available has increased since 1986 by a factor of 5–10.

Here, we present a combined neural network approach to the recognition of signal peptides and their cleavage sites, using one network to recognize the cleavage site and another network to distinguish between signal peptides and non-signal peptides. A similar combination of two pairs of networks has been used with success to predict the intron splice sites in pre-mRNA from humans and the dicotelydoneous plant *Arabidopsis thaliana* (Brunak *et al*., 1991; S.Hebsgaard, P.Korning, J.Engelbrecht, P.Rouźe and S.Brunak, submitted). Artificial neural networks have been used for many biological sequence analysis problems (Hirst and Sternberg, 1992; Presnell and Cohen, 1993). They have also been applied to the twin problems of predicting signal peptides and their cleavage sites, but until now without leading to practically applicable prediction methods with significant improvements in performance compared with the weight matrix method (Arrigo *et al*., 1991; Ladunga *et al*., 1991; Schneider and Wrede, 1993).

Materials and methods

The data were taken from SWISS-PROT version 29 (Bairoch and Boeckmann, 1994). The data sets were divided into prokaryotic and eukaryotic entries and the prokaryotic data sets were further divided into Gram-positive eubacteria (*Firmicutes*) and Gram-negative eubacteria (*Gracilicutes*), excluding *Mycoplasma* and *Archaebacteria*. Viral, phage and organellar proteins were not included. In addition, two single-species data sets were selected, a human subset of the eukaryotic data and an *Escherichia coli* subset of the Gram-negative data.

The sequence of the signal peptide and the first 30 amino acids of the mature protein from the secretory protein were included in the data set. The first 70 amino acids of each sequence were used from the cytoplasmic and (for the eukaryotes) nuclear proteins. In addition, a set of eukaryotic signal anchor sequences, i.e. N-terminal parts of type II membrane proteins (von Heijne, 1988), were extracted (see Figure 1).

As an example of a large-scale application of the finished method, we used the *Haemophilus influenzae* Rd genome—the first genome of a free-living organism to be completed (Fleischmann *et al*., 1995). We have downloaded the sequences of all the predicted coding regions in the *H.influenzae* genome from the World Wide Web (WWW) server of the Institute for Genomic Research at http://www.tigr.org/. Only the first 60 positions of each sequence were analysed.

We have attempted to avoid signal peptides where the cleavage sites are not experimentally determined, but we are not able to eliminate them completely, since many database entries simply lack information about the quality of the evidence. The details of the data selection are described in the WWW server and in an earlier paper (Nielsen *et al*., 1996a).

Redundancy in the data sets was avoided by excluding pairs of sequences which were functionally homologous, i.e. those that had more than 17 (eukaryotes) or 21 (prokaryotes) exact matches in a local alignment (Nielsen *et al*., 1996a). Redundant sequences were removed using an algorithm which guarantees that no pairs of homologous sequences remain in the data set (Hobohm *et al*., 1992). This procedure removed 13–56% of the sequences. The numbers of non-homologous sequences remaining in the data sets are shown in Table I. Redundancy

Table I. Data and performance values

Source	Data					
	(Number of sequences)		Network architecture (window/hidden units)		Performance	
	Signal peptides	Non-secretory proteins	*C*-score	*S*-score	Cleavage site location (% correct)	Signal peptide discrimination (correlation)
Human	416	251	15+4/2	27 / 4	68.0 (67.9)	0.96 (0.97)
Eukaryote	1011	820	17+2/2	27 / 4	70.2	0.97
E.coli	105	119	15+2/2	39 / 0	83.7 (85.7)	0.89 (0.92)
Gram–	266	186	11+2/2	19 / 3	79.3	0.88
Gram+	141	64	21+2/0	19 / 3	67.9	0.96

Data: the number of sequences of signal peptides and non-secretory (i.e. cytoplasmic or nuclear) proteins in the data sets after redundancy reduction. The organism groups are eukaryotes, human, Gram-negative bacteria ('Gram–'), *E.coli* and Gram-positive bacteria ('Gram+'). The human data are subsets of the eukaryotic data and the *E.coli* data are subsets of the Gram-negative data. The signal anchor and *H.influenzae* data are not shown in the table. *Network architecture*: the size of the input window and the number of hidden computational units ('neurons') in the optimal neural networks chosen for each data set. *C*-score networks have asymmetrical input windows. *Performance*: the percentage of signal peptide sequences where the cleavage site was predicted to be at the correct location according to the maximal value of the Y-score (see Figure 2). The ability of the method to distinguish between the signal peptides and the N-terminals of non-secretory proteins (based on the mean value of the *S*-score in the region between position 1 and the predicted cleavage site position) is measured by the correlation coefficients (Mathews, 1975). Both performance values are measured on the test sets (the average of five cross-validation tests). The values given in parentheses indicate the performance for the human sequences when using networks trained on all eukaryotic data and for the *E.coli* sequences when using Gram-negative networks respectively.

reduction was not applied to the signal anchor data or the *H.influenzae* data, since these were not used as training data.

Neural network algorithms

The signal peptide problem was posed to the neural networks in two ways: (i) recognition of the cleavage sites against the background of all other sequence positions and (ii) classification of amino acids as belonging to the signal peptide or not. In the latter case, negative examples included both the first 70 positions of non-secretory proteins and the first 30 positions of the mature part of secretory proteins.

The neural networks were feed-forward networks with zero or one layer of two to 10 hidden units, trained using back-propagation (Rumelhart *et al*., 1986) with a slightly modified error function. The sequence data were presented to the network using sparsely encoded moving windows (Qian and Sejnowski, 1988; Brunak *et al*., 1991). Symmetric and asymmetric windows of a size varying from five to 39 positions were tested.

Based on the numbers of correctly and incorrectly predicted positive and negative examples, we calculated the correlation coefficient (Mathews, 1975). The correlation coefficients of both the training and test sets were monitored during training and the performance of the training cycle with the maximal test set correlation was recorded for each training run. The networks chosen for inclusion in the WWW server have been trained until this cycle only.

The test performances have been calculated by cross-validation: each data set was divided into five approximately equal-sized parts and then every network run was carried out with one part as test data and the other four parts as training data. The performance measures were then calculated as an average over the five different data set divisions.

For each of the five data sets, one signal peptide/non-signal peptide network architecture and one cleavage site/non-cleavage site network architecture was chosen on the basis of the test set correlation coefficients. We did not pick the architecture with absolutely the best performance, but instead the smallest network that could not be significantly improved by enlarging the input window or adding more hidden units.

The trained networks provide two different scores between zero and one for each position in an amino acid sequence. The output from the signal peptide/non-signal peptide networks, the *S*-score, can be interpreted as an estimate of the probability of the position belonging to the signal peptide, while the output from the cleavage site/non-cleavage site networks, the *C*-score, can be interpreted as an estimate of the probability of the position being the first in the mature protein (position +1 relative to the cleavage site).

If there are several *C*-score peaks of comparable strength, the true cleavage site may often be found by inspecting the *S*-score curve in order to see which of the *C*-score peaks coincides best with the transition from the signal peptide to the non-signal peptide region. In order to formalize this and improve the prediction, we have tried a number of linear and non-linear combinations of the raw network scores and evaluated the percentage of sequences with correctly placed cleavage sites in the five test sets. The best measure was the geometric average of the C-score and a smoothed derivative of the S-score, termed the Y-score:

$$Y_i = \sqrt{\mathrm{C}_i \Delta_d S_i}, \tag{1}$$

where $\Delta_d S_i$ is the difference between the average *S*-score of *d* positions before and *d* position after position *i*:

$$\Delta_d S_i = \frac{1}{d}\left(\sum_{j=1}^{d} S_{i-j} - \sum_{j=0}^{d-1} S_{i+j} \right) \tag{2}$$

In Figure 2(A), examples of the values of the C-, S- and Y-scores are shown for a typical signal peptide with a typical cleavage site. The C-score has one sharp peak that corresponds to an abrupt change in the S-score from a high to low value. Among the real examples, the C-score may exhibit several peaks and the S-score may fluctuate. We define a cleavage site as being correctly located if the true cleavage site position corresponds to the maximal Y-score (combined score).

For a typical non-secretory position, the values of the C-, S- and Y-scores are lower, as shown in Figure 2(B). We found the best discriminator between signal peptides and non-secretory

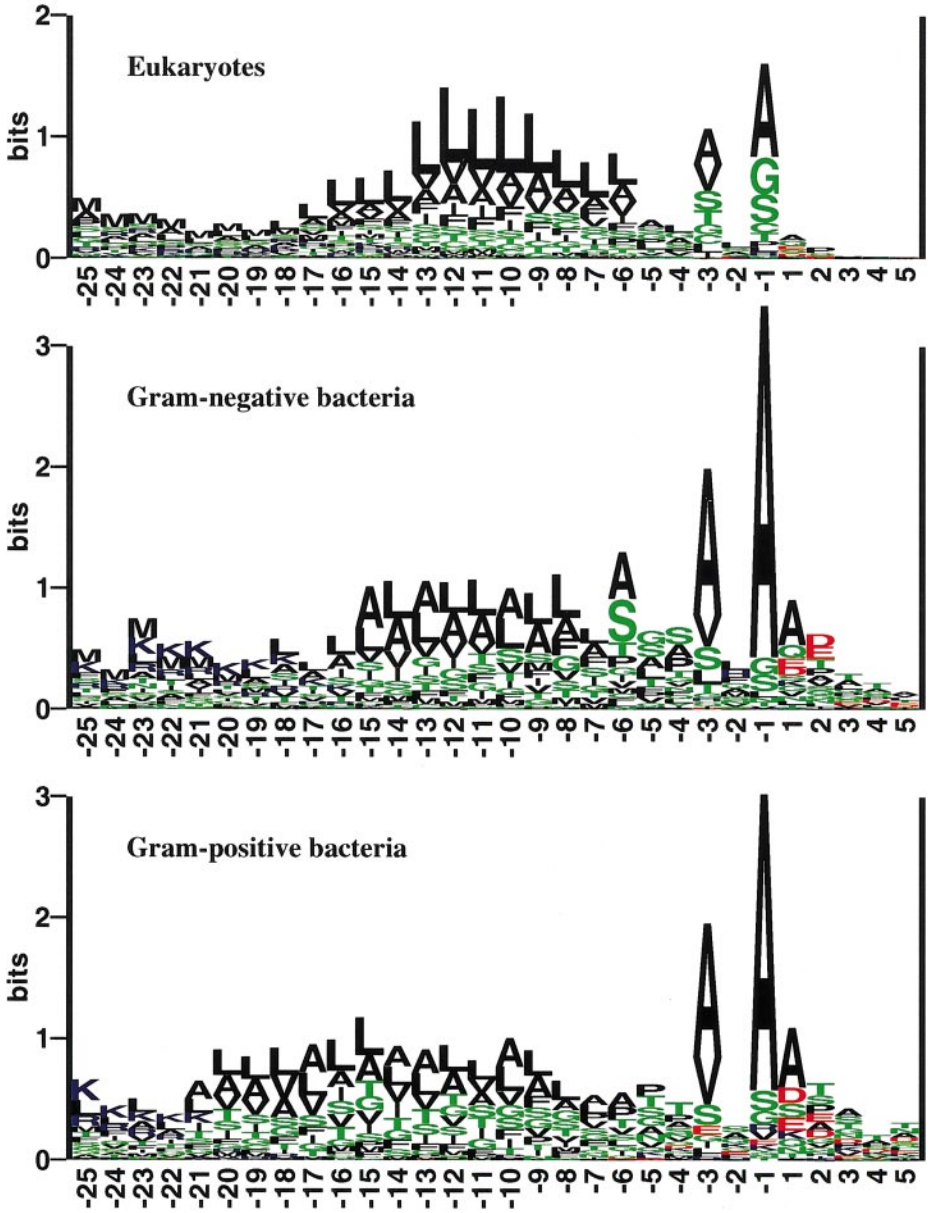

Fig. 1. Sequence logos (Schneider and Stephens, 1990) of signal peptides, aligned by their cleavage sites. The total height of the stack of letters at each position shows the amount of information, while the relative height of each letter shows the relative abundance of the corresponding amino acid. The information is defined as the difference between the maximal and actual entropy (Shannon, 1948): $I_j = H_{\max} - H_j = \log_2 20 + \Sigma_\alpha\, n_j(\alpha)/N_j \log_2 n_j(\alpha)/N_j$, where $n_j(\alpha)$ is the number of occurrences of the amino acid α and N_j is the total number of letters (occupied positions) at position j. Positively and negatively charged residues are shown in blue and red respectively, while uncharged polar residues are green and hydrophobic residues are black.

proteins to be the average of the S-score in the predicted signal peptide region, i.e. from position 1 to the position immediately before the position where the Y-score has a maximal value. If this value—the mean S-score—is greater than 0.5, we predict the sequence in question to be a signal peptide (cf. Figure 3).

The relationship between the various performance measures and their development during the training process is described in detail elsewhere (Nielsen *et al.*, 1997).

Results and discussion

The optimal network architecture and corresponding predictive performance for all the data sets are shown in Table I. The C-score problem is best solved by networks with asymmetric windows, i.e. windows including more positions upstream than downstream of the cleavage site. This corresponds well with the location of the cleavage site pattern information which is shown as sequence logos (Schneider and Stephens, 1990) in Figure 1. The S-score problem, on the other hand, is best solved by symmetric or approximately symmetric windows.

Although our method is able to locate cleavage sites and discriminate signal peptides from non-secretory proteins with a reasonably high reliability, the accuracy of the cleavage site location is lower than that reported for the original weight matrix method (von Heijne, 1986): 78% for eukaryotes and

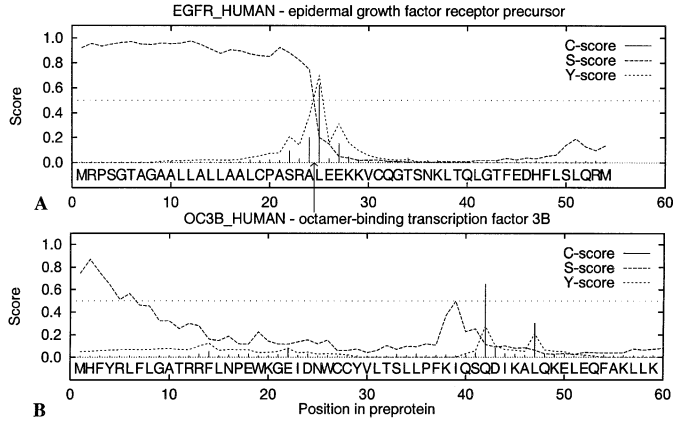

Fig. 2. Examples of network output. The values of the C- (output from cleavage site networks), S- (output from signal peptide networks) and Y-scores (combined cleavage site score, $Y_k = \sqrt{C_i \Delta_d S_i}$) are shown for each position in the sequence. The C- and S-scores are averages over five networks trained on different parts of the data. Note: the C- and Y-scores are high for the position immediately after the cleavage site, i.e. the first position in the mature protein. (A) A successfully predicted signal peptide. The true cleavage site is marked wih an arrow. (B) A non-secretory protein. For many non-secretory proteins, all three scores are very low throughout the sequence. In this example, there are peaks of the C- and S-scores, but the sequence is still easily classified as non-secretory, since the C-score peak occurs far away from the S-score decline and the region of the high S-score is far too short.

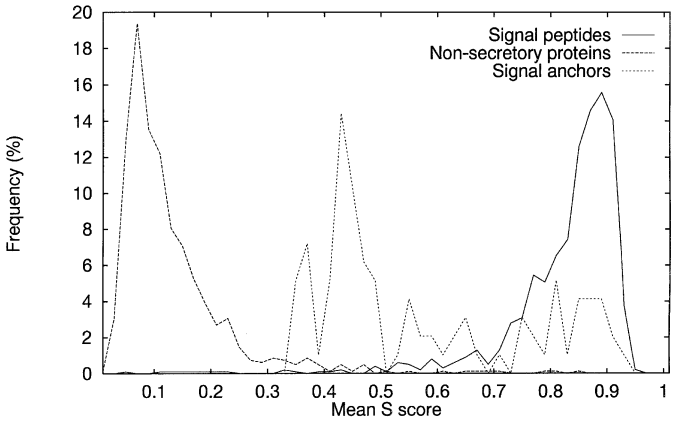

Fig. 3. Distribution of the mean signal peptide score (S-score) for signal peptides and non-signal peptides (eukaryotic data only). 'Non-secretory proteins' refer to the N-terminal parts of cytoplasmic or nuclear proteins, while 'signal anchors' are the N-terminal parts of type II membrane proteins. The mean S-score of a sequence is the average of the S-score over all positions in the predicted signal peptide region (i.e. from the N-terminal to the position immediately before the maximum of the Y-score). The bin size of the distribution is 0.02.

89% for prokaryotes (not divided into Gram-positive and -negative). When the original weight matrix is applied to our recent data set, however, the performance is much lower. This suggests a larger variation in the examples of the signal peptides found since then. It may, of course, also reflect a higher occurrence of errors in our automatically selected data than in the manually selected 1986 set.

In order to compare the strength of the neural network approach to the weight matrix method, we recalculated new weight matrices from our new data and tested the performances of these (results not shown). The weight matrix method was comparable to the neural networks when calculating the C-score, but was practically unable to solve the S-score problem

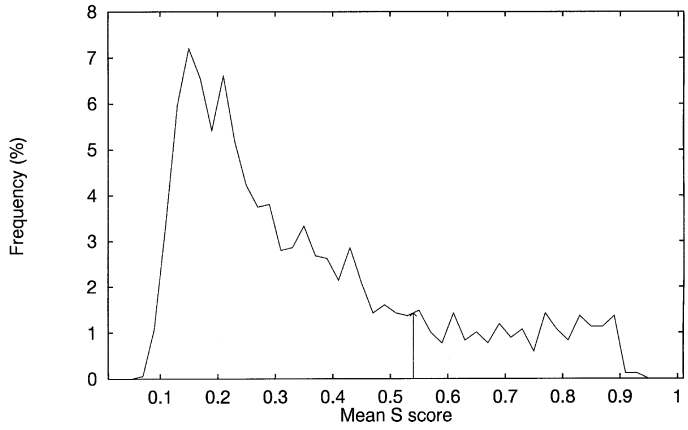

Fig. 4. Distribution of the mean signal peptide score (S-score) for all the predicted *H.influenzae* coding sequences. The mean S-score is calculated using networks trained on the Gram-negative data set. The bin size of the distribution is 0.02. The arrow shows the optimal cut-off for predicting a cleavable signal peptide. The predicted number of secretory proteins in *H.influenzae* (corresponding to the area under the curve to the right of the arrow) is 330 out of 1680 (20%).

and therefore did not provide the possibility of calculating the combined Y-score.

Note that the prediction performances reported here correspond to minimal values. The test sets in the cross-validation have a very low sequence similarity; in fact, the sequence similarity is so low that the correct cleavage sites cannot be found by alignment (Nielsen *et al.*, 1996a). This means that the prediction accuracy on sequences with some similarity to the sequences in the data sets will in general be higher.

The differences between the signal peptides from different organisms are apparent from Figure 1. The signal peptides from Gram-positive bacteria are considerably longer than those of other organisms, with much more extended h-regions, as observed previously (von Heijne and Abrahmsén, 1989). The prokaryotic h-regions are dominated by Leu (L) and Ala (A) in approximately equal proportions and in the eukaryotes they are dominated by Leu with some occurrence of Val (V), Ala, Phe (F) and Ile (I). Close to the cleavage site, the (–3,–1) rule is clearly visible for all three data sets, but while a number of different amino acids are accepted in the eukaryotes, the prokaryotes accept alanine almost exclusively in these two positions. In the first few positions of the mature protein (downstream of the cleavage site) the prokaryotes show certain preferences for Ala, negatively charged (D or E) amino acids, and hydroxy amino acids (S or T), while no pattern can be seen for the eukaryotes. In the leftmost part of the alignment, the positively charged residue Lys (K) [and to a smaller extent Arg (R)] is seen in the prokaryotes, while the eukaryotes show a somewhat weaker occurrence of Arg (barely visible in the figure) and almost no Lys. This corresponds well with the hypothesis that positive residues are required in the n-region where the N-terminal Met is formulated for prokaryotes, but not necessarily for eukaryotes where the N-terminal Met in itself carries a positive charge (von Heijne, 1985).

The difference in structure is reflected in the performances of the trained neural networks (see Table I). Gram-negative cleavage sites have the strongest pattern—i.e. the highest information content—and, consequently, they are the easiest to predict, both at the single-position and at the sequence level. The eukaryotic cleavage sites are significantly more difficult

to predict. Gram-positive cleavage sites are slightly more difficult to predict than the eukaryotic ones, which would not be expected from the sequence logos (Figure 1), since they show nearly as high an information content as the Gram-negative cleavage sites, but the longer Gram-positive signal peptides means that the cleavage sites have to be located against a larger background of non-cleavage site positions. The discrimination of signal peptides versus non-secretory proteins, on the other hand, is better for the eukaryotes than for the prokaryotes. This may be due to the more characteristic leucine-rich h-regions of the eukaryotic signal peptides.

The logos for the human and *E.coli* data sets are not shown, since they show no significant differences from those of the eukaryotes or Gram-negative bacteria respectively. Accordingly, the predictive performance was not improved by training the networks on single-species data sets. On the contrary, the *E.coli* signal peptides are predicted even better by the Gram-negative networks than by the *E.coli* networks (probably due to the relatively small size of the *E.coli* data set). In other words, we have found no evidence for species-specific features of the signal peptides of humans and *E.coli*.

Signal anchors often have sites similar to signal peptide cleavage sites after their hydrophobic (transmembrane) region. Therefore, a prediction method can easily be expected to mistake signal anchors for peptides. In Figure 3, the distribution of the mean S-score for the 97 eukaryotic signal anchors is included. It shows some overlap with the signal peptide distribution. If the standard cut-off of 0.5 is applied to the signal anchor data sets, 50% of the eukaryotic signal anchor sequences are falsely predicted as signal peptides (the corresponding figure for the human signal anchors is 75% when using human networks and 68% when using eukaryotic networks). With a cut-off optimized for signal anchor versus signal peptide discrimination (0.62), we were able to lower this error rate to 45% for the eukaryotic data set. The mean S-score still gives a better separation than the maximal C- or Y-score, which indicates that the pseudo-cleavage sites are in fact rather strong.

However, the pseudo-cleavage sites often occur further from the N-terminal than genuine cleavage sites do. If we do not accept signal peptides longer than 35 residues (this will exclude only 2.2% of the eukaryotic signal peptides in our data set), the percentage of false positives among the signal anchors drops to 28% for the eukaryotic and 32% for the human signal anchors (39% when using eukaryotic networks). When taking this into account, our method does provide a reasonably good discrimination between signal peptides and signal anchors. This has not been reported by any of the earlier published methods for signal peptide recognition.

Scanning the Haemophilus influenzae genome

We have applied the prediction method with networks trained on the Gram-negative data set to all the amino acid sequences of the predicted coding regions in the *Haemophilus influenzae* genome. The distribution of the mean S-score (from position 1 to the position with a maximal Y-score) is shown in Figure 4.

When applying the optimal cut-off value found for the Gram-negative data set, we obtained a crude estimate of the number of sequences with cleavable signal peptides in *H.influenzae*: 330 out of 1680 sequences or approximately 20%. If the maximal S-score is used instead of the mean S-score, the estimate comes out as 28% and with the maximal Y-score it is 14% (distributions not shown). If all three criteria are applied together, leaving only 'typical' signal peptides, we obtain 188 sequences (11%).

Some of the sequences predicted to be signal peptides according to the S-score but not according to the Y-score may be signal anchor-like sequences of type II (single-spanning) or type IV (multispanning) membrane proteins. This hypothesis is strengthened by a hydrophobicity analysis of the ambiguous examples (results not shown). If we apply the slightly higher cut-off optimized for the discrimination of signal anchors versus signal peptides in eukaryotes (0.62) to the mean S-score, the estimate is lowered from 20 to 15%.

On the other hand, some of the sequences predicted to be signal peptides according to the maximal Y-score but not the mean S-score may be the effect of the initiation codon of the predicted coding region having been placed too far upstream. In this case, the apparent signal peptide becomes too long and the region between the false and the true initiation codon will probably not have signal peptide character, thereby bringing the mean S-score of the erroneously extended signal peptide region below the cut-off. This is strengthened by the finding that these ambiguous examples are longer than average and contain more methionines.

In conclusion, we estimate that 15–20% of the *H.influenzae* proteins are secretory. However, a whole-genome analysis like this would be more reliable if combined with other analyses, notably transmembrane segment predictions and initiation site predictions.

Method and data publicly available

The finished prediction method is available both via an e-mail server and a WWW server. Users may submit their own amino acid sequences in order to predict whether the sequence is a signal peptide and, if so, where it will be cleaved. We recommend that only the N-terminal part (say 50–70 amino acids) of the sequences is submitted, so that the interpretation of the output is not obscured by false positives further downstream in the protein.

The user is asked to choose between the network ensembles trained on data from Gram-positive, Gram-negative or eukaryotic organisms. We did not include the networks trained on the single-species data sets in the servers, since these did not improve the performance.

The values of the C-, S- and Y-scores are returned for every position in the submitted sequence. In addition, the maximal Y-score, maximal S-score and mean S-score values are given for the entire sequence and compared with the appropriate cut-offs. If the sequence is predicted to be a signal peptide, the position with the maximal Y-score is mentioned as the most likely cleavage site. A graphical plot in postscript format, similar to those in Figure 2, may be requested from the servers. We strongly recommend that a graphical plot is always used for the interpretation of the output. The plot may give hints about, for example, multiple cleavage sites or erroneously assigned initiation, which would not be found when using only the maximal or mean score values.

The address of the mail server is signalp@cbs.dtu.dk. For detailed instructions, send a mail containing the word 'help' only. The WWW server is accessible via the Center for Biological Sequence Analysis homepage at http://www.cbs.dtu.dk/.

All the data sets mentioned in Table I are available from an FTP server at ftp://virus.cbs.dtu.dk/pub/signalp. Retrieve the file README for detailed descriptions of the data and the format.

The FTP server and the mail server can both be accessed directly from the WWW server.

References

Arrigo,P., Giuliano,F., Scalia,F., Rapallo,A. and Damiani,G. (1991) *CABIOS*, **7**, 353–357.
Bairoch,A. and Boeckmann,B. (1994) *Nucleic Acids Res.*, **22**, 3578–3580.
Brunak,S., Engelbrecht,J. and Knudsen,S. (1991) *J. Mol. Biol.*, **220**, 49–65.
Fleischmann,R. *et al.* (1995) *Science*, **269**, 449–604.
Gierasch,L.M. (1989) *Biochemistry*, **28**, 923–930.
Hirst,J.D. and Sternberg,M.J.E. (1992) *Biochemistry*, **31**, 7211–7218.
Hobohm,U., Scharf,M., Schneider,R. and Sander,C. (1992) *Protein Sci.*, **1**, 409–417.
Ladunga,I., Czakó,F., Csabai,I. and Geszti,T. (1991) *CABIOS*, **7**, 485–487.
Mathews,B. (1975) *Biochim. Biophys. Acta*, **405**, 442–451.
Nielsen,H., Engelbrecht,J., von Heijne,G. and Brunak,S. (1996a) *Proteins*, **24**, 165–177.
Nielsen,H., Engelbrecht,J., von Heijne,G. and Brunak,S. (1997) *Int. J. Neural Sys.*, in press.
Presnell,S.R. and Cohen,F.E. (1993) *Annu. Rev. Biophys. Biomol. Struct.*, **22**, 283–298.
Qian,N. and Sejnowski,T.J. (1988) *J. Mol. Biol.*, **202**, 865–884.
Rapoport,T.A. (1992) *Science*, **258**, 931–936.
Rumelhart,D.E., Hinton,G.E. and Williams,R.J. (1986) In Rumelbart,D., McClelland,J. and the PDP Research Groups (eds), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Vol. 1: Foundations*. MIT Press, Cambridge, MA, pp. 318–362.
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.
Shannon,C.E. (1948) *Bell System Technol. J.*, **27**, 379–423, 623–656.
von Heijne,G. (1983) *Eur. J. Biochem.*, **133**, 17–21.
von Heijne,G. (1985) *J. Mol. Biol.*, **184**, 99–105.
von Heijne,G. (1986) *Nucleic Acids Res.*, **14**, 4683–4690.
von Heijne,G. (1988) *Biochim. Biophys. Acta*, **947**, 307–333.
von Heijne,G. (1990) *J. Membrane Biol.*, **115**, 195–201.
von Heijne,G. and Abrahmsén,L. (1989) *FEBS Lett.*, **244**, 439–446.

Received April 19, 1996; revised September 2, 1996; accepted September 12, 1996