"J. Mol. Biol., to appear 2004."

Improved prediction of signal peptides — SignalP 3.0

Jannick Dyrløv Bendtsen[1], Henrik Nielsen[1], Gunnar von Heijne[3] and Søren Brunak[1*]

[1]Center for Biological Sequence Analysis
BioCentrum-DTU
Building 208
Technical University of Denmark
DK-2800 Lyngby, Denmark

[3]Stockholm Bioinformatics Center
Department of Biochemistry and Biophysics
Stockholm University
SE-106 91 Stockholm, Sweden

* To whom correspondence should be addressed (email: brunak@cbs.dtu.dk)

Keywords: Signal peptide, signal peptidase I, neural network, hidden Markov model, SignalP

Running title: Signal peptide prediction by SignalP

We describe improvements of the currently most popular method for prediction of classically secreted proteins, SignalP. SignalP consists of two different predictors based on neural network and hidden Markov model algorithms, where both components have been updated. Motivated by the idea that the cleavage site position and the amino acid composition of the signal peptide are correlated, new features have been included as input to the neural network. This addition, combined with a thorough error-correction of a new data set, have improved the performance of the predictor significantly over SignalP version 2. In version 3, correctness of the cleavage site predictions have increased notably for all three organism groups, eukaryotes, Gram-negative and Gram-positive bacteria. The accuracy of cleavage site prediction has increased in the range from 6-17% over the previous version, whereas the signal peptide discrimination improvement is mainly due to the elimination of false positive predictions, as well as the introduction of a new discrimination score for the neural network. The new method has also been benchmarked against other available methods. Predictions can be made at the publicly available web server `http://www.cbs.dtu.dk/services/SignalP/`.

Introduction

Numerous attempts to predict the correct subcellular location of proteins using machine learning techniques have been developed[1–9]. Computational methods for prediction of N-terminal signal peptides were published around 20 years ago, initially using a weight matrix approach[1,2]. Development of prediction methods shifted to machine learning algorithms in the mid 1990's[10,11], with a significant increase in performance[12]. SignalP, one of the currently most used methods, predicts the presence of signal peptidase I cleavage sites. For signal peptidase II cleavage sites found in lipo-proteins the LipoP predictor has been constructed[13]. SignalP produces both classification and cleavage site assignment, while most of the other methods classifies proteins as secretory or non-secretory.

A consistent assessment of the predictive performance requires a reliable benchmark data set. This is particularly important in this area where the predictive performance is approaching the performance calculated from interpretation of experimental data, which is not always perfect. Incorrect annotation of signal peptide cleavage sites in the databases stems not only from trivial database errors, but also from peptide sequencing where it may be hard to control the level of post-processing of the protein by other peptidases, after the signal peptidase I has made its initial cleavage. Such post-processing typically leads to cleavage site assignments shifted downstream relative to the true signal peptidase I cleavage site.

In the process of training the new version of SignalP we have generated a new, thoroughly curated data set based on the extraction and redundancy reduction method published earlier[14]. Other methods were used for cleaning the new data set, and we found a surprisingly high error rate in Swiss-Prot, where, for example, in the order of 7% of the Gram-positive entries had either wrong cleavage site position and/or wrong annotation of the experimental evidence. Also, we found many errors in a previously used benchmark set[12] (stemming from automatic extraction from Swiss-Prot), and it appears that some programs are in fact better than the performance reported (predictions are correct, while feature annotation is incorrect). For comparison, we made use of this independent benchmark data set that was initially used for evaluation of five different signal peptide predictors[12].

In the new version of SignalP we have introduced novel amino acid composition units as well as sequence position units in the neural network input layer in order to obtain better performance. Moreover, we have slightly changed the window sizes compared to the previous version. We have used fivefold cross-validation tests for direct comparison to the previous version of SignalP[10]. In the previous version of SignalP a combination score, Y, was created from the cleavage site score, C, and the signal peptide score, S, and used to obtain a better prediction of the position of the cleavage site. In the new version, we also use the C-score to obtain a better discrimination between secreted and non-secreted sequences, and have constructed a new D-score for this classification task. The architecture of the hidden Markov model SignalP has not changed, but the models have been retrained on the new data set, and have also significantly increased their performance.

Results and discussion

Generation of data sets

As the predictive performance of the earlier SignalP method was quite high, assessment of potential improvements is critically dependent on the quality of the data annotation. We generated a new positive signal peptide data set from Swiss-Prot[15] release 40.0, retaining the negative data set extracted from the previous work. The method for redundancy reduction was the same as in the previous work[14], and was based on the reduction principle developed by Hobohm *et al.*[16]. Our final positive signal peptide data sets contain 1192, 334 and 153 sequences for eukaryotes, Gram-negative and Gram-positive bacteria, respectively.

In the previous work, we found many errors by detailed inspection of hard-to-learn examples during training and wrongly predicted examples. Nevertheless, we were quite sure that even after careful examination in this manner, the data set would probably still contain errors obtained from incorrect database annotation and wrongly interpreted laboratory results.

Therefore, we developed a new feature based approach where abnormal examples can be detected by inspecting rare amino acid occurrences and outlier physical-chemical properties of signal peptides. In the following, we show that the isoelectric point of signal peptides can help in finding possible annotation errors and other errors, where these errors may be due to the fact that some (long) signal peptides annotated in Swiss-Prot actually include probable propeptides. In such cases, convertase cleavage sites are mixed together with signal peptidase I cleavage sites.

Removal of spurious cleavage site residues

Experimental assessment of the effect of certain amino acids in the cleavage site region has shown that rare residues do not allow for efficient cleavage[17,18]. Examination of amino acids around the signal peptidase I cleavage site in the data set revealed a number of sequences containing amino acids, which very rarely appear at the cleavage site.

In the eukaryotic data set we found and removed seven sequences containing lysines (K) and 13 sequences containing arginines (R) at the -1 position. All sequences with either a lysine or an arginine at position -1 were investigated manually. All of them except one had a predicted cleavage site upstream of the annotated one. Most of these sequences probably undergo N-terminal maturation by different proteases, either in the Trans Golgi Network (TGN) or after release from the cell as mentioned below in the section on propeptide analysis. In one clear case we found an obvious error in the Swiss-Prot entry NPAB_LOCMI. According to the annotation the cleavage site is located between residues 24-25 (arginine in position -1), but in the original paper the authors identified the cleavage to occur between amino acids 22-23. In this case, the two amino acids, ER, are removed by a dipeptidase[19].

Furthermore, we removed sequences where other amino acids appeared at position -1 in very few of the sequences. For the eukaryotic data set, the only allowed residues at position -1 were alanine (A), cysteine (C), glycine (G), leucine (L), proline (P), glutamine (Q), serine (S) and threonine (T). By allowing only the latter amino acids we might have removed a few true, unusual sequences. For instance, tyrosine (Y) and histidine (H) at position -1 were found only in one case each in the entire eukaryotic data set. We removed eight sequences with aspartic acid (D) and eight with phenylalanine (F), seven each with glutamic acid (E) and asparagine (N), respectively. Five with methionine (M), three containing isoleucine (I) and two sequences containing tryptophan (W) at position −1 were also removed. Some of these are in fact provable errors, in one of the aspartic acid examples, `CLUS_BOVIN`[20], the N-terminal peptide sequencing in the paper reports the cleavage as `MKTLLLLMGLLLSWESGWA---ISDKELQEMST` ⋯, while Swiss-Prot annotates the sequence as being cleaved between D and K, thereby changing a common position −1 amino acid, alanine, into a rare one. Interestingly, SignalP predicts the cleavage site as reported in the paper.

For Gram-positive and Gram-negative bacteria, only four residues were allowed at position −1. These residues were alanine (A), glycine (G), serine (S) and threonine (T)[17,18]. For the Gram-positive data set, this approach removed four sequences containing arginines (R), three containing valines (V), two containing lysines (K) and one sequence each of glutamic acid (E), leucine (L), asparagine (N), glutamine (Q), threonine (T) and tyrosine (Y). In the Gram-negative data set, we removed two sequences containing valine (V) at position −1 and one sequence for each of the following amino acids, glutamic acid (E), lysine (K), leucine (L), asparagine (N), glutamine (Q).

Isoelectric point calculations

Previous studies have shown differences in amino acid composition between signal peptide and mature protein[21,22]. Thus, we examined to what extent the isoelectric point (pI) could be used as a unique feature of signal peptides.

We calculated the pI for all signal peptides and the corresponding mature proteins in the data set and presented this in three scatter plots (Figure 1). In the scatter plot for Gram-positive bacteria two very distinct clusters appear. Only three signal peptide outliers were found and by manual inspection of the corresponding Swiss-Prot entries, we found that these proteins most likely were either not carrying signal peptides, or were annotated wrongly.

These outliers having pI values below 8 had the following Swiss-Prot ID's `CWLA_BACSP`, `IAA2_STRGS`, `COTT_BACSU`. The three entries have annotated signal peptides, but it is doubtful whether the annotation is correct. According to the prediction from SignalP

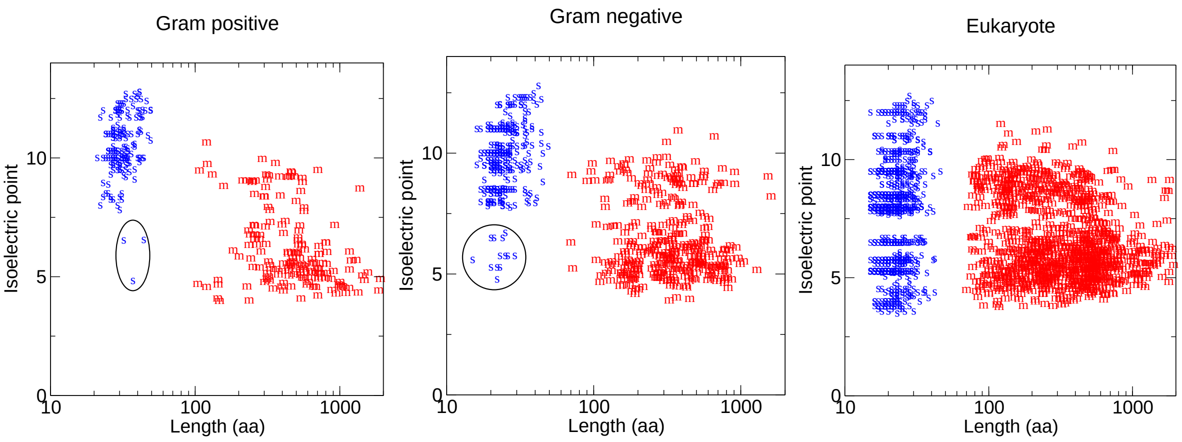


Figure 1: **Isoelectric point calculations.** Calculations of isoelectric point of signal peptide and mature protein, indicated by s and m, respectively. Clusters of outlier examples for bacteria are indicated on the two plots.

and PSORT, CWLA_BACSP does not carry a signal peptide. CWLA_BACSP was in the paper described as a "putative" signal peptide[23] and later it was indicated that *cwlA* is part of an ancestral prophage, still remnant in the *Bacillus subtilis* genome[24]. All phage and virus sequences were initially removed from the SignalP training set, which could result in the negative prediction for this prophage sequence.

The cleavage site in the alpha-amylase inhibitor IAA2_STRGS turns out not to be experimentally verified. It is predicted to have a cleavage site at position 26 (SignalP) or 24 (PSORT). Calculation of pI using the SignalP predicted signal peptide length gave a new result of 8.66, closer to the average for Gram-positive bacteria. The paper proposes two other cleavage site positions, but none of these have been verified experimentally[25].

The last entry COTT_BACSU is a spore coat protein from *B. subtilis*[26,27] and no BLAST homologs in Swiss-Prot were found to contain an experimentally verified signal peptide. CotT is proteolytically processed from a 10kD precursor protein and is localized to spore coat where it controls the assembly. By N-terminal sequencing the N-terminus of the mature and processed protein was identified, although nowhere in the two papers is an SPase I cleavage site indicated, thus no signal peptide is mentioned[26,27]. With the current knowledge about spore coats, spore coat assembly does not involve translocation of coat protein across any membrane[28–30]. Hence, it is very unlikely for CotT to carry an N-terminal signal peptide as annotated in Swiss-Prot.

The average isoelectric point of signal peptides and mature proteins in the entire Gram-positive data set was 10.59 and 6.24, respectively. This is consistent with the fact that Gram-positive bacteria are known to have the longest signal peptides that carry more basic residues (K/R) in the n-region, than Gram-negatives and eukaryotes[11].

When inspecting the scatter plot for Gram-negative bacteria, we find the same overall clustering as observed for the Gram-positive bacteria, although not as distinct. Here the major group of signal peptides have pIs between 8 and 13, although the variation is larger than in the Gram-positive scatter plot. A few sequence entries with acidic signal peptides were investigated in detail. Sequence entry SFMA_ECOLI having a pI of 4.78 was found to

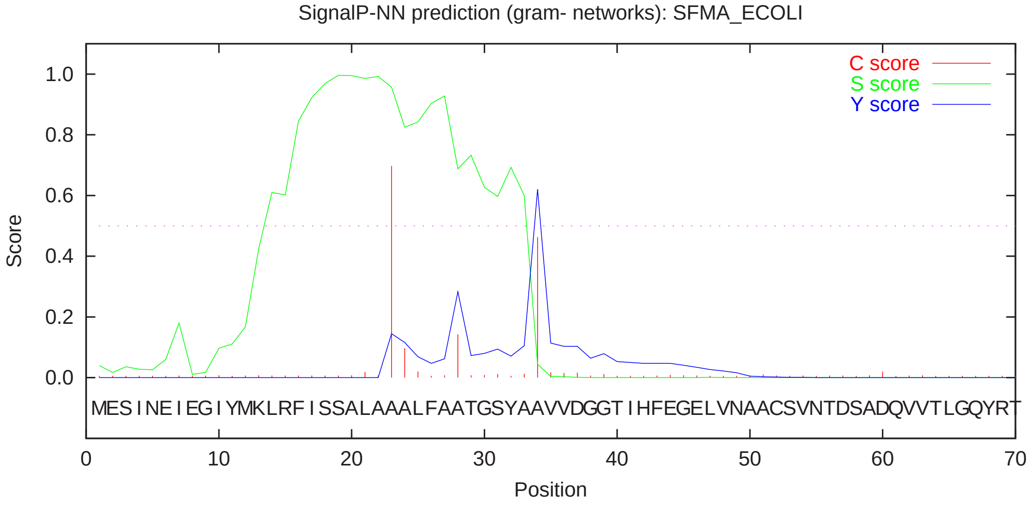


Figure 2: **Alternative start codon assignment.** The graphical output from SignalP strongly indicates erroneous annotation of the signal peptide from Swiss-Prot entry SFMA_ECOLI. Further investigation showed a wrong annotation of the start codon (see text for details). C, S, and Y-score indicate cleavage site, "signal peptide-ness" and combined cleavage site predictions, respectively.

be an obvious erroneous annotation in Swiss-Prot. This entry had an annotated cleavage site at position 22, but a predicted cleavage site at position 34. As seen from Figure 2 we found an internal methionine at position 12. Since the signal peptide-ness is very low until position 12 we assumed that this was an incorrectly annotated start codon. If the initial 11 amino acids until the internal methionine were removed, SignalP correctly predicted the cleavage to be at position 22 and the pI of the signal peptide increased from 4.78 to 9.99. Indeed, in release 41.0 of Swiss-Prot this entry was corrected and the signal peptide marked "POTENTIAL".

For eukaryotes on the other hand, we were not able to distinguish the pI of the signal peptide and the mature protein. Eukaryotes have the shortest signal peptides and the amount of basic residues is much lower than for bacteria.

Propeptide or signal peptide?

For the eukaryotic data we examined whether annotated signal peptides could possibly include propeptides. In secreted proteins, propeptides are often found immediately downstream of the signal peptidase I cleavage site and their cleavage site is defined by a conserved set of basic amino acids. Propeptides can be hard to detect by N-terminal Edmann degradation, as the propeptides are cleaved off in the TGN before the release of the mature protein to the surroundings[31].

We used a new propeptide predictor, ProP, to predict propeptide cleavage sites[32] in the eukaryotic data set. In ten sequences we found a predicted cleavage site for a propeptide at the same position where a signal peptidase I cleavage site was annotated in Swiss-Prot. In all ten cases SignalP predicted a shorter signal peptide than annotated, thus making room for a short propeptide between the predicted signal peptide and the mature protein. The ten sequences, AMYH_SACFI, CRYP_CRYPA, FINC_RAT, GUX2_TRIRE, LIGC_TRAVE, MDLA_PENCA, RNMG_ASPRE, RNT1_ASPOR, XYN2_TRIRE, XYNA_THELA, were reassigned according to the prediction of SignalP version 2.0. This is an exceptional case where we tend to rate the computational analysis higher than experimental evidence, which must be considered weak, as the propeptide processing takes place before the proteins have been subjected to experimental, N-terminal peptide sequencing.

After the signal peptide in these cases had been reassigned, we got marginally higher correlation coefficients when retraining the neural network on the reassigned data set (data not shown).

Optimization of window sizes

As in the earlier SignalP approach, the signal peptide discrimination and the signal peptidase I cleavage site prediction were handled using two different types of neural networks[10,33].

We used a brute force approach to optimize the window sizes for the neural networks by calculating single position correlation coefficients for all possible combinations of symmetric and asymmetric windows. Using this approach we trained approximately 6500 neural networks for window optimization for a single organism group. This was furthermore done for different combinations where amino acid composition and position information was included in the input to network or not, leading to approximately 27000 neural networks being tested in all.

For eukaryotes, these data are shown in Figure 3. It is clear that optimal signal peptide discrimination prediction requires symmetric (or nearly symmetric) windows, whereas cleavage site training needs asymmetric windows with more positions upstream of the cleavage site included in the input to the network. The optimal window size for cleavage site prediction for the eukaryote network included 20 positions upstream and 4 positions downstream of the cleavage site. The window sizes for the Gram-positive networks were retained as previously found[10], whereas the Gram-negative cleavage site network included one more position downstream of the cleavage site, resulting in a window of 11 positions upstream and 3 positions downstream of the cleavage site. The eukaryote discrimination network performs best when using a symmetric window of 27 positions. For both Gram-positive and Gram-negative bacteria the discrimination network is based on a symmetric window of 19 positions. This brute force approach changed the optimal window sizes of the cleavage site network slightly from those used in SignalP 2.0[10, 33].

Network performance

We have evaluated the performance of SignalP version 3.0 using the same performance measures as used for the previous two versions of SignalP, see Table 1. The performance values were calculated using five fold cross-validation, *i.e.* testing on sequences not present in the training set (all data split into five subsets of approximately the same size). The most significant performance increase was obtained for the cleavage site prediction as seen in Table 1. A performance increase of 6-17% for all three organism classes was obtained. We were able to optimize the signal peptide discrimination performance by introducing a new score, termed the D-score, replacing the earlier used mean S-score quantifying the "signal peptide-ness" of a given sequence segment. In the earlier versions of SignalP the scores from the two types of networks were combined for cleavage site assignment, and not for the task of discrimination. In the new version 3, the D-score is calculated as the average of the mean S-score and the maximal Y-score, and the two types of networks are

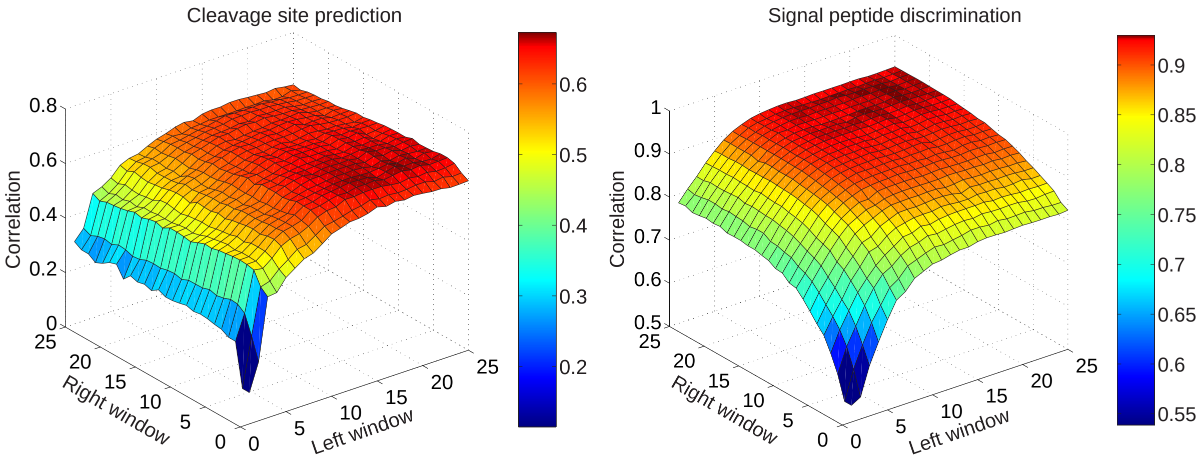


Figure 3: **Window optimization.** These plots show single position level correlation coefficients for all combinations of window sizes for the signal peptide cleavage and discrimination networks used for eukaryotic signal peptide prediction. The optimal window size for cleavage site for the eukaryotic network included 20 positions to the left and 4 positions to the right of the cleavage site. For reasons of computational efficiency we have selected a discrimination network with a symmetric window of 27 amino acids, although networks with larger windows have slightly higher single position level correlation coefficients.

then used for both purposes (see Material and Methods for details).

Version	Cleavage site (Y-score)			Discrimination (SP/non-SP)		
	Euk	Gram−	Gram+	Euk	Gram−	Gram+
SignalP 1 NN	70.2	79.3	67.9	0.97	0.88	0.96
SignalP 2 NN	72.4	83.4	67.4	0.97	0.90	0.96
SignalP 2 HMM	69.5	81.4	64.5	0.94	0.93	0.96
SignalP 3 NN	79.0	92.5	85.0	0.98	0.95	0.98
SignalP 3 HMM	75.7	90.2	81.6	0.94	0.94	0.98

Table 1: **Performances of three different SignalP versions.** The most significant improvement was for the cleavage site predictions. Cleavage site performances are presented as % and discrimination values (based on D-score) as correlation coefficients. NN and HMM indicate neural network and hidden Markov model, respectively. Results are based on five-fold cross validation for all SignalP versions

.

Improvement by position information and composition features

In order to improve the performance of the neural network version of SignalP, we introduced two new features into the network input: information about the position of the sliding window as well as information on the amino acid composition of the entire sequence. This information was encoded by additional input units in the neural network. The new position information units were found to be important for both the cleavage site and discrimination networks, whereas the amino acid composition information only improved the discrimination network. The idea of including compositional information is based on the observation that the composition of secreted and non-secreted proteins differ[21,22].

The average length of signal peptides range from 22 (eukaryotes) and 24 (Gram-negatives) to 32 amino acids for Gram-positives, and the new network encoding the position of the sliding window uses these averages to penalize prediction of extremely long or short signal peptides. Therefore, twin arginine signal peptides often receive a below threshold D-score as they tend to be quite long (average 37 amino acids)[34,35]. This also means that a few cases of ordinary signal peptides with extreme length are not predicted correctly by the neural networks. The HMM is also in its structure penalizing long signal peptides, and similarly the SignalP3 HMM is not able to predict these cases correctly. One example[36] is the (`NUC_STAAU`) with a 63 amino acid long signal peptide that is not predicted correctly by any of the SignalP3 models. SignalP3 does not always fail to predict long signal peptides correctly, *e.g.* the 56 amino acids long signal peptide of `CYGD_BOVIN`[37] is handled correctly by the neural network version, both in terms of cleavage site and discrimination. However, great care should be taken when interpreting the scores for long potential signal peptides.

From Figure 4 the importance of the new approach where position and amino acid composition information is included can be assessed. Including information of the position of the sliding window during training, increased the neural network cleavage site prediction performance slightly (left panel of the figure). Composition information did not increase the performance of the cleavage site prediction, therefore it is excluded from the left panel in Figure 4. But composition information did increase the performance of the discrimination network slightly (right panel of the figure), whereas information of the

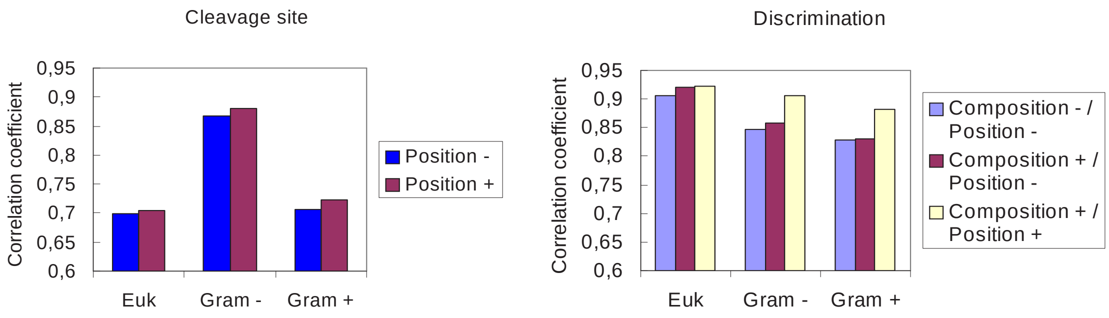


Figure 4: **Improvement of the neural network by introducing length and composition features.** Position of the sliding window in the neural network input increased cleavage site prediction performance slightly (left panel). Amino acid composition information together with information of the position of the sliding window improved the discrimination network significantly as seen in the right panel. The performance improvement was evaluated as single position level correlations during training on the individual networks for cleavage and discrimination, respectively.

position of the sliding window together with composition increased the discrimination significantly (right panel). Another improvement of the discrimination stems from the new D-score (see Table 2). The final prediction method uses both position and composition information.

Effect of the new discrimination score

In SignalP version 3.0 we have introduced a new discrimination score for the neural network, termed the D-score. Based on the mean S-score and maximal Y-score it was found to give increased discriminative performance over the mean S-score, used in SignalP version 2.0. In Table 2, the D-score shows superior performance over the mean S-score for the novel part of the benchmark set defined by Menne et al. (see below).

Dataset	sensitivity	specificity	accuracy	cc
Eukaryotes	0.99 (0.98)	0.85 (0.84)	0.93 (0.93)	0.87 (0.86)
Gram−	0.94 (0.93)	0.88 (0.81)	0.95 (0.93)	0.88 (0.82)
Gram+	0.98 (0.98)	0.98 (0.98)	0.98 (0.98)	0.96 (0.95)

Table 2: **D-score outperforms the mean S-score for discrimination of signal peptide versus non-signal peptide.** Using the novel part of the Menne test set[12], we tested the D-score for discrimination compared to the mean S-score. The mean S-score performances are shown in parentheses.

The above mentioned 56 amino acid long signal peptide in `CYGD_BOVIN` is an example where the D-score leads to a correct classification, while the mean S-score is below the threshold. In this case the strong cleavage site score adds to a weaker signal peptide-ness in the C-terminal part of the leader sequence.

Performance comparison to other prediction methods

As described in a recent review of signal peptide prediction methods it is hard to find an ideal benchmark set, as methods have been frozen at different times[12]. The data used to train a method is in general "easier" than genuine test sequences that are novel to a particular method. Since we have used a more recent version of Swiss-Prot than did

Menne *et al.* in their assessment, we have merely retained Menne set sequences that are not present in the SignalP version 3.0 training set. In this manner, we do not give an advantage to SignalP, as some of these sequences possibly have been included in the training set for other methods.

We did not test the performance of the weight matrix-based methods SigCleave or SPScan as the earlier report shows that these are outperformed by machine learning methods[12]. SigCleave is based on von Heijne's weight matrix[2] from 1986. SPScan is also based on the weight matrix from von Heijne, but in addition to this it uses McGeoch's criteria for a minimal, acceptable signal peptide[1].

We have tested other methods which are made available, one problem being that they do not necessarily predict the same organism classes, e.g. the PSORT-B method[8] does only predict on Gram-negative data, and not on the two other SignalP organism classes.

The comparative results are given in Table 3. For the PSORT-II method[38,39] which predicts on eukaryotic sequences, the subcellular localization classes "endoplasmic reticulum (ER)", "extracellular" and "Golgi" were merged into one category of secretory proteins, whereas the rest "cytoplasmic", "mitochondrial", "nuclear", 'peroxisomal" and "vacuolar" were merged into a single "non-secretory" category. The performance reported in the paper is 57% correct for all categories. In Table 3 it can be seen that SignalP3 outperforms PSORT-II on this particular set with a significant margin. PSORT-II does not assign cleavage sites, and we have therefore only compared the discrimination performance. We believe that the minor decrease in discrimination performance of SignalP3 on this set, when compared to the cross-validation performance reported above in Table 1, is a result of errors in the Menne set (originating from Swiss-Prot) together with its redundancy (see below), but more importantly, the presence of transmembrane helices within the first 60 amino acids in more than 10% of the novel negative test sequences from this set (when analyzed by TMHMM[40]).

The new version of PSORT (PSORT-B) has been trained on five subcellular localization classes in Gram-negative bacteria and was reported to obtain a 97% specificity and 75% sensitivity[8]. PSORT-B was optimized for specificity over sensitivity. Another recent method, SubLoc[5] predicts three subcellular compartments for prokaryotes and four com-

Data set / Method	sensitivity	specificity	accuracy	cc
Eukaryotes SignalP3-NN	0.99	0.85	0.93	0.87
Eukaryotes PSORT-II	0.65	0.75	0.80	0.56
Eukaryotes SubLoc	0.58	0.70	0.77	0.47
Gram− SignalP3-NN	0.92	0.88	0.95	0.87
Gram− PSORT-B	0.99	0.64	0.75	0.58
Gram− Subloc	0.90	0.79	0.91	0.78
Gram+ SignalP3-NN	0.95	0.93	0.97	0.92
Gram+ PSORT	0.86	0.80	0.91	0.77
Gram+ SubLoc	0.82	0.92	0.86	0.76

Table 3: **Performance measures for signal peptide discrimination.** Using the novel part of the Menne *et al.* test set[12] we obtained the results shown in the table. Note that the values for PSORT-B is calculated on the part of the data set where PSORT-B produces a classification. Around 55% of the sequences were classified as "Unknown", and the actual performance is therefore much lower than indicated here. For a given organism class the relevant version of PSORT has been used to make the predictions and calculated the performance.

partments for eukaryotes. For SubLoc the total prediction accuracy was reported to be 91.4% for the three subcellular locations in prokaryotes and 79.4% for the four locations in eukaryotes. None of the latter two methods, PSORT-B and SubLoc, reports a predicted cleavage site, but is only designed to perform discrimination. Also, none of these methods were compared to the SignalP version 2.0 in their publications.

Of the 289 negative test set sequences for Gram-negative bacteria in the novel part of the Menne set, 191 received an "Unknown" classification by PSORT-B, and similarly 22 sequences in the positive test set were classified as unknown. The "Unknown" sequences were discarded in the calculation of the PSORT-B performance shown in Table 3. PSORT-B classifies 55% of the submitted sequences as "Unknown". From Table 3 it can be seen that SignalP3-NN is significantly better than both SubLoc and PSORT-B on the novel part of the Menne set, even when excluding these "unknown" sequences. SignalP3-HMM has a similar performance on this set. All PSORT-B categories except cytoplasmic were regarded as secretory. In comparison to SignalP3 we have merged the two categories "periplasmic" and "extracellular" predicted from SubLoc to one category for secretory proteins.

The original version of PSORT was used for predicting signal peptides in Gram-positive bacteria[3]. We merged the output categories of "cleaved signal peptide" and "uncleaved signal peptide" into one category, "secretory". Sequences with a negative N-terminal signal peptide prediction were regarded as cytoplasmic. Again, the performance of SignalP3 is higher than PSORT. As the amount of data used to train this version of PSORT was quite small, the performance is surprisingly good.

Sigfind[41], another (eukaryotic only) method based on neural networks, has a limitation of four sequences per host per day. We have submitted 50 randomly chosen negative and 50 randomly chosen positive test sequences from the novel Menne set. Sigfind reported two false positive within this negative test set and 1 false negative in the positive test set. When running the same sequences on SignalP3, we obtained no false positives, but the same false negative as Sigfind.

Manual inspection of the one false negative prediction of `APL_HUMAN` by both Sigfind and SignalP revealed that this particular Swiss-Prot has been updated (with new identifier `APL1_HUMAN`) after the development of the Menne set, which was based on release 38.0 of Swiss-Prot. The sequence has now been extended by 15 amino acids at the N-terminus which results in a 27 amino acid long signal peptide and not a 12 amino acid signal peptide as earlier reported. Taken this change into consideration, both Sigfind and SignalP3 correctly classifies this protein as being secreted.

For this limited part of the novel test set by Menne, Sigfind obtained a correlation coefficient of 0.96, compared to the perfect correlation coefficient of 1.0 for SignalP3.

Wondering about the true performance of Sigfind, we chose to submit eukaryotic secretory, cytoplasmic and nuclear protein sequences initially created in Swiss-Prot release 42.0. Neither SignalP or the Sigfind method have been trained or tested on these new sequences. We were able to extract 54 signal peptide containing sequences, 86 cytoplasmic, and 119 nuclear sequences. When submitting those Sigfind correctly classifies all new signal peptide containing sequences as secretory, but classifies four of the 86 cytoplasmic and five of the nuclear sequences as secretory (false positives). SignalP3 correctly classifies all new eukaryotic secretory and cytoplasmic proteins correctly, but make two false positive predictions for the nuclear sequences. For discrimination of secretory and non-secretory proteins newly entered into Swiss-Prot, the Sigfind method obtains a correlation coefficient of 0.91, whereas SignalP again obtains a better correlation of 0.98. It appears that the Sigfind method quite strongly overpredicts signal peptide containing sequences, and this means that on a normal data set (either the one used to train SignalP or a full proteome), where the non-secretory proteins greatly outnumber the secretory proteins, the actual performance in terms of specificity will be much lower than on this more balanced set.

A neural network based method (NNPSL) for prediction of subcellular localization in eukaryotes and prokaryotes were published a few years back[4]. Unfortunately, the online prediction method is capable of handling only a couple of sequences per submission, which made it hard to compare to SignalP.

Another prediction method SPEPlip[7] was recently published, but we were not able to perform a comparison as the server did not function for a duration of four weeks in which we checked it. However, we are quite sceptical in relation to the generalization ability of this method, as it was trained and tested on the full version of the highly redundant Menne set. The set has not been redundancy reduced before training and testing of SPEPlip[7]. Consequently, the test part of the data set will contain many sequences highly similar to training set sequences. For such sequences with high similarity, the cleavage site position can easily be found by alignment, and the inclusion of such sequences leads to a significant overestimation of the predictive performance.

A recently published method for cleavage site prediction (not available for test) based on support vector machines[6] reports a performance increase of 47% in terms of true positive predictions at a false positive rate of 3% when compared to the original weight matrix method by von Heijne[2]. Unfortunately, the support vector machine method was not compared to SignalP. The SVM method finds 68% true positive cleavage sites at a false positive rate of 3%. This method does not distinguish between eukaryotic and bacterial sequences.

Very recently a new method, Phobius, designed to improve transmembrane helix topology predictions by integrating topology and signal peptide predictions became available[42]. Often the first transmembrane helix can be mistaken for a signal peptide and vice versa. This method was trained on data collected from Swiss-Prot release 41. While the performance values are not easy to compare as e.g. the negative data set has been extracted from PDB using different similarity criteria than those used to develop SignalP[14], we made an evaluation of Phobius, using the same novel sequences from release 42 of Swiss-Prot, as used in the test of Sigfind described above (note that the comparison between Phobius and SignalP in the paper was made using the old SignalP 2.0 version). Out of 205 negative test examples, Phobius generated 4 false positive predictions, whereas SignalP generated 2 false positive predictions. Both methods were able to correctly classify all signal peptide containing sequences. For discrimination this results in a correlation coefficient of 0.96 for Phobius and 0.98 for SignalP. As also reported in the Phobius paper, the cleavage site prediction accuracy is below the accuracy of the SignalP method. For this set from Swiss-Prot rel. 42, Phobius could correctly predict the position of the cleavage site in 75% of the sequences, while SignalP version 3.0 is able to correctly predict the cleavage site position in 87% of the sequences. When tested on a very small set of eleven novel, experimentally verified Gram+ and Gram- sequences from Swiss-Prot release 42, we found that Phobius predicted 64% of those correctly, while SignalP3 was correct in 82% of the cases. Thus, for these novel sequences found in the Swiss-Prot database, SignalP performs better, both in terms of discrimination and cleavage site prediction. Nevertheless, Phobius is indeed superior over SignalP when it comes to prediction of transmembrane helices close to the

N-terminus, which are easier to confuse with signal peptides.

Misuse of SignalP

We have noticed that users of SignalP in some cases interpret a positive prediction as meaning that the protein is extracellular. As many proteins with signal peptides are retained e.g. in ER/Golgi this is not always the case. Eukaryotic proteins may be retained in the ER if the protein holds an “ER retention signal”, which is found at the C-terminus of the mature protein. SignalP does not take such signals into account. It is hard to assess how often wrong interpretations are made due to lack of experimental data, but for the data set used to train SignalP3 we found eleven cases with retention signals (based on Swiss-Prot annotation). The true level of retention is presumably higher.

Another more rare type of wrong use happens when negative predictions are interpreted incorrectly. A negative classification by SignalP does not necessarily imply that the protein is indeed a non-secreted protein, as some protein enters the extracellular space by non-classical and leaderless pathways. We have dealt with the issue of non-classical secretion elsewhere[43], and have developed the SecretomeP server for this purpose (see below).

Conclusion

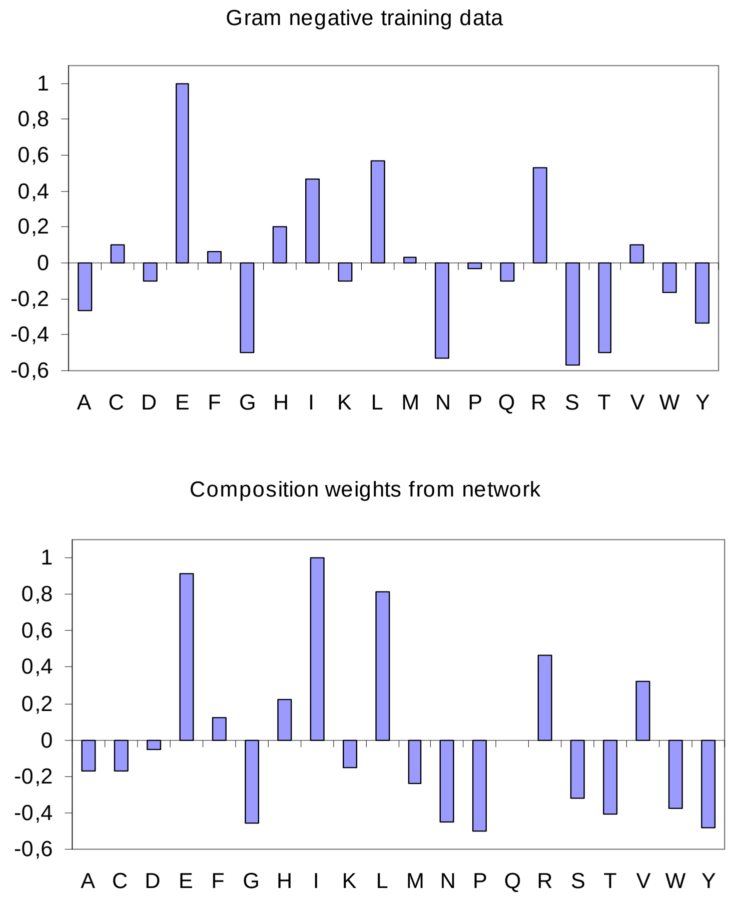
We present new versions of SignalP, based on an expanded, highly curated data set. The architecture of the hidden Markov model based version was unchanged, while the neural network scheme was improved by including information of the amino acid composition of the precursor protein as well as the position of the sliding window. Furthermore, we optimized the window sizes by testing all possible combinations of asymmetric and symmetric input windows up to a total input of 51 amino acids. These were changed slightly compared to the earlier SignalP version.

For all organism groups, we obtained an improvement of the cleavage site predictions based on the maximal Y-score as defined earlier[10]. Discrimination between signal peptide containing sequences and non-secretory sequences was improved by introducing a new score, the D-score, replacing the mean S-score used for discrimination in earlier SignalP versions.

On an independent test set (limited to sequences not used for SignalP3 construction) we achieved better sensitivity, specificity, accuracy and correlation coefficient for the cleavage site predictions for all three organism groups. Moreover, we obtained better signal peptide discrimination in most cases.

The 3.0 version of SignalP that we present here shows in all cases a performance increase over version 2.0 (see Table 1). This was observed both for the cleavage site prediction and the signal peptide discrimination. The improved performance of discrimination predictions for the new version of SignalP is partly due to the introduction of composition units in the neural network input layer.

The improved performance obtained when simultaneously using position and composition information proves that these features are indeed correlated. We were not able to figure out in detail how the cleavage site position and the protein composition actually are correlated, but when inspecting the neural network composition input unit weights, it



Gram negative training data

Composition weights from network

Figure 5: **Amino acid composition information.** Over- and underrepresentation of amino acids in the Gram-negative training set and corresponding weights from the composition units in the network to the hidden units. The unit for neural network weights is arbitrary. Negative numbers indicate overrepresentation of the particular amino acid in secretory proteins.

became clear that the networks readily learn how secreted proteins differ in composition from non-secretory proteins. Figure 5 shows the over- and underrepresentation of the twenty amino acids in secretory proteins from Gram-negative bacteria, and the sizes of the corresponding weights connecting the input composition units and the hidden units. With the exception of methionine, the sign of these weights reflects directly the over- or underrepresentation of the amino acids, explaining why this additional input is able to improve the predictive performance of SignalP3. The amino acid bias in the data is in agreement with earlier observations[21]. The picture was similar for eukaryotes and Gram-positive bacteria (data not shown).

The prediction quality of the new method was significantly better than other methods performing classification. SignalP3 was compared to other methods made available, with the exception of older weight matrix approaches. It should be noted that the independent test set we have used (from Menne *et al.*), has not been cleaned in the same thorough manner as the training set. We are aware of several annotation errors in the test that have not been removed, and therefore the performance of all methods is in practice better than estimated from this set.

As the discrimination task is quite close to being correct for most sequences, the most prominent problem in signal peptide prediction is the prediction of the correct cleavage site. We have thoroughly investigated the data set for cleavage site annotation errors and many of these were indeed found. One source of error of signal peptide annotation is the neglect of maturation proteases that act on the protein after the SPase I cleavage. By cleaning for obvious cleavage site annotation errors, we improved the cleavage site performance without retraining. This meaning that the old versions indeed had a better performance than assessed by the old incorrect data.

The method described is made available at `http://www.cbs.dtu.dk/services/SignalP/`. The SecretomeP server mentioned above can be found at `http://www.cbs.dtu.dk/services/SecretomeP/`

Materials and methods

Data set extraction

All sequence data were extracted from Swiss-Prot[15] release 40.0. A total of 12975 entries with the keyword "SIGNAL" were found. The data set was split into three species specific groups which are Eukaryotic, Gram-negative prokaryotes and Gram-positive prokaryotes. We excluded all archaeal sequences. Non experimentally verified signal peptides which had "POTENTIAL" or "HYPOTHETICAL" stated in the keyword line were removed. Furthermore, any phage, viral or eukaryote organelle encoded proteins were excluded. Lipoproteins were also removed. Similarly, we excluded entries with more than one cleavage site and entries with non verified N-terminus if indicated in the keyword line. This reduced the number of signal peptide carrying sequences in the data set to 3902.

In order to remove any bias in the data set, the set was redundancy reduced by the scheme previously developed, excluding pairs of sequences that were functionally homologous[14]. Subsequently no two sequences in the data set have more than 17 (eukaryotes) or 21 (prokaryotes) identical amino acids in a local alignment. Approximately half of the remaining sequences were thus removed. This scheme for redundancy reduction ensures that, the method does not transfer functional information to one sequence (from a set of experimentally characterized examples) by mere sequence similarity in the usual sense as detected by alignment. The SignalP data set preparation paper[14] determine how dissimilar two sequences should be in order to prompt the application of a “discrimination” method, rather than alignment, for the specific case of signal peptide prediction. Thus, the performance values reported here correspond to a situation where the inference cannot reliably be made by alignment.

A few of the remaining sequences were removed from the data set due to their extreme length. We removed sequences with signal peptides with less than 15 amino acids for all three organism groups, sequences with more than 45 amino acids for eukaryotes and Gram-negatives and 50 amino acids for Gram-positive. We did not specifically remove Tat signal peptides, but some of them were removed due to their length.

For discrimination of secretory proteins versus cytoplasmic proteins we used the old data set from SignalP version 2.0 of cytoplasmic, nuclear and signal anchor sequences, as we saw no reason to doubt these. We found and removed one error, though.

Additionally, we cleaned the redundancy reduced data set by additional techniques as described below.

Our final eukaryotic data set contains 1192 secretory, 990 nuclear 459 cytoplasmic and 67 signal anchors sequences. The data set for Gram-negative bacteria contains 334 secretory and 358 cytoplasmic sequences. Finally, the data set for Gram-positive bacteria contains 153 secretory and 151 cytoplasmic sequences.

Further cleaning of the extracted data set

Propeptides

Recently, a eukaryotic propeptide convertase predictor — ProP was developed[32]. In conjunction with SignalP version 2.0, we reassigned sequences from the eukaryotic data set, which seemed to include a propeptide annotated as a signal peptide (see text).

Spurious cleavage site residues

In addition to cleaning the eukaryotic data set for potential propeptide cleavage sites, we also removed any sequence that contained the basic residues Lysine (K) or arginine (R) at position -1. Furthermore, we removed all sequences that had residues in position -1 occurring only in a small minority of sequences. For the eukaryotic data set the only allowed residues at position -1 were, alanine (A), cysteine (C), glycine (G), leucine (L), proline (P), glutamine (Q), serine (S) and threonine (T). For Gram-positive and Gram-negative bacteria only alanine (A), glycine (G), serine (S) and threonine (T) were allowed at position -1, according to the cleavage site.

Database annotation errors

After initial training of our method, the data set was manually cleaned for errors that could be related to annotation errors. These erroneous annotations were identified as predictions made by the neural network, which did not correspond with the annotation from Swiss-Prot. Only errors that could be identified manually and confirmed from the corresponding papers were removed or reannotated.

List of sequence entries for Gram-negative bacteria that were either reannotated or removed from the training set: PGL2_ERWCA, YBCL_ECOLI, OMLA_PSEAE, CBPG_PSES6, BLP2_PSEAE, GUNC_PSEFL, HLYB_PROMI, FPTA_PSEAE, CY1_PARDE.

List of sequence entries for Gram-positive bacteria that were either reannotated or removed from the training set: XYNC_STRLI, CHOD_STRSQ, HYSA_PROAC, BLAF_MYCFO, AGAR_STRCO, CHOD_BREST, CHOD_STRSQ, IMD_ARTGO, ALDC_BACBR, BLAC_STRAU, GUNG_CLOTM, GUNH_CLOTM, TACY_STRPY, TEE6_STRPY, THI1_PANTH, XYN1_BACST, AMY_BACSU, CHI1_BACCI.

This approach has not been carried out on the eukaryote training set.

All errors found in the training sets have been reported to Swiss-Prot.

Independent test set

During previous work for an evaluation of signal peptide predictors, an independent test was created[12]. The positive test set was divided into three groupings corresponding to eukaryotes, Gram-negative and Gram-positive bacteria. Sequences already found in the SignalP3 training set were removed to prevent artificially high performance measure for the SignalP method when testing. This resulted in 557 eukaryotic, 100 Gram-negative and 42 Gram-positive sequences which were used as a positive test set. In the eukaryotic negative test set, sequences carrying mitochondrial transit peptides were removed. The negative test set consisted of 1056 eukaryotic, 289 Gram-negative and 129 Gram-positive sequences.

The three negative test sets were investigated using TMHMM[40] for presence of transmembrane helices within the first 60 amino acid of each sequence. For the 129 Gram-positive cytoplasmic sequences we found 12 sequences with one transmembrane helix within the first 60 amino acids. 32 sequences out of the 289 Gram-negative cytoplasmic sequences carry a (predicted) transmembrane helix within the first 60 amino acids, where 123 out of 1056 eukaryotic cytoplasmic sequences carry a helix within the first 60 amino acids of each sequence.

Neural network architecture

We used two different neural networks for coping with the signal peptide prediction problem. One network for recognition of the cleavage site, and one network for determining whether a given amino acid belongs to the signal peptide or not. A more thorough description of the neural network has been presented in a previous paper[33]

Furthermore, we improved the neural network by introducing new input features. These features were position of the sliding window as a parameter together with the amino acid composition of the entire sequence.

Position of the sliding window was used as an input neuron in the neural network for cleavage site prediction and for the discrimination of signal peptides. Composition neurons were only introduced as additional neurons into the neural network for discrimination of signal peptides versus non-signal peptides, as they did not improve the cleavage site prediction.

Also, we optimized the window sizes by testing all combinations of symmetric and asymmetric windows varying from 3 to 51 positions.

Measuring prediction performance

Performance of the prediction method was carried out on different data sets. Firstly, we used the same measure of performance as used in earlier versions of SignalP *i.e.* cross-validation on the training set, meaning that the data set was split into five equally sized parts and trained on four and tested on one for a total of five times until all sequences had been used for training and testing, respectively.

Performance of the prediction method on the novel part of the Menne *et al.* test set[12] was measured as the sensitivity, giving the fraction of positive examples truly predicted as positive.

$$\text{sensitivity} = \frac{tp}{tp + fn}. \tag{1}$$

The specificity gives the fraction of all positive predictions that are true positives,

$$\text{specificity} = \frac{tp}{tp + fp}. \tag{2}$$

Accuracy gives the fraction of all true predictions, both true positive and true negative,

$$\text{accuracy} = \frac{tp + tn}{tp + tn + fp + fn}. \tag{3}$$

The Matthews correlation coefficient[44] is defined as,

$$\text{cc} = \frac{tp \cdot tn - fp \cdot fn}{\sqrt{(tp + fn)(tp + fp)(tn + fp)(tn + fn)}} \tag{4}$$

and was used in calculation of optimal network performance during the training.
Here, *tp* = true positive and *tn* = true negative and *fp* = false positive and *fn* = false negative.

New discrimination score

Previous versions of SignalP showed the best discrimination of signal peptides versus non-signal peptides, by using the mean S-score calculated as the average of the S-score in the predicted signal peptide region. We have implemented a new score for better discrimination called the D-score. The D-score is simply an average of the mean S-score and the maximal Y-score, where the Y-score is defined as,

$$Y_i = \sqrt{C_i \Delta_d S_i}, \tag{5}$$

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

$$\Delta_d S_i = \frac{1}{d}\left(\sum_{j=1}^{d} S_{i-j} - \sum_{j=0}^{d-1} S_{i+j}\right). \tag{6}$$

Acknowledgements

We thank Anders Krogh for use of his HMM software. This work was supported by grants from the Danish National Research Foundation, the Danish Natural Science Research Council, the Danish Center for Scientific Computing, and by a grant from Novozymes A/S (to JDB).

References

1. McGeoch, D. J. (1985). On the predictive recognition of signal peptide sequences. *Virus Res.* **3**, 271–286.

2. von Heijne, G. (1986). A new method for predicting signal sequence cleavage sites. *Nucl. Acids Res.* **14**, 4683–4690.

3. Nakai, K. and Kanehisa, M. (1991). Expert system for predicting protein localization sites in Gram-negative bacteria. *Proteins,* **11**, 95–110.

4. Reinhardt, A. and Hubbard, T. (1998). Using neural networks for prediction of the subcellular location of proteins. *Nucl. Acids Res.* **26**, 2230–2236.

5. Hua, S. and Sun, Z. (2001). Support vector machine approach for protein subcellular localization prediction. *Bioinformatics,* **17**, 721–728.

6. Vert, J. P. (2002). Support vector machine prediction of signal peptide cleavage site using a new class of kernels for strings. vol. pp. 649–660, World Scientific.

7. Fariselli, P. and Finocchiaro, G. and Casadio, R. (2003). SPEPlip: the detection of signal peptide and lipoprotein cleavage sites. *Bioinformatics,* **19**, 2498–2499.

8. Gardy, J. L. and Spencer, C. and Wang, K. and Ester, M. and Tusnady, G. E. and Simon, I. and Hua, S. and deFays, K. and Lambert, C. and Nakai, K. and Brinkman, F. S. (2003). PSORT-B: improving protein subcellular localization prediction for Gram-negative bacteria. *Nucl. Acids Res.* **31**, 3613–3617.

9. Zhang, Z. and Wood, W. I. (2003). A profile hidden Markov model for signal peptides generated by HMMER. *Bioinformatics,* **19**, 307–308.

10. Nielsen, H. and Brunak, S. and 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.

11. Nielsen, H. and Krogh, A. (1998). Prediction of signal peptides and signal anchors by a hidden Markov model. vol. pp. 122–130, AAAI Press, Menlo Park, CA.

12. Menne, K. M. and Hermjakob, H. and Apweiler, R. (2000). A comparison of signal sequence prediction methods using a test set of signal peptides. *Bioinformatics,* **16**, 741–742.

13. Juncker, A. S. and Willenbrock, H. and Von Heijne, G. and Brunak, S. and Nielsen, H. and Krogh, A. (2003). Prediction of lipoprotein signal peptides in Gram-negative bacteria. *Protein Sci.* **12**, 1652–1662.

14. Nielsen, H. and Engelbrecht, J. and von Heijne, G. and Brunak, S. (1996). Defining a similarity threshold for a functional protein sequence pattern: the signal peptide cleavage site. *Proteins,* **26**, 165–177.

15. Bairoch, A. and Apweiler, R. (2000). The Swiss-Prot protein sequence database and its supplement TrEMBL in 2000. *Nucl. Acids Res.* **28**, 45–48.

16. Hobohm, U. and Scharf, M. and Schneider, R. and Sander, C. (1992). Selection of representative protein data sets. *Protein Sci.* **1**, 409–417.

17. Karamyshev, A. L. and Karamysheva, Z. N. and Kajava, A. V. and Ksenzenko, V. N. and Nesmeyanova, M. A. (1998). Processing of *Escherichia coli* alkaline phosphatase: role of the primary structure of the signal peptide cleavage region. *J. Mol. Biol.* **277**, 859–870.

18. Paetzel, M. and Karla, A. and Strynadka, N. C. and Dalbey, R. E. (2002). Signal peptidases. *Chem. Rev.* **102**, 4549–4580.

19. Lagueux, M. and Kromer, E. and Girardie, J. (1992). Cloning of a *Locusta* cDNA encoding neuroparsin A. *Insect Biochem. Mol. Biol.* **22**, 511–516.

20. Palmer, D. J. and Christie, D. L. (1990). The primary structure of glycoprotein III from bovine adrenal medullary chromaffin granules. Sequence similarity with human serum protein-40,40 and rat Sertoli cell glycoprotein. *J. Biol. Chem.* **265**, 6617–6623.

21. Cedano, J. and Aloy, P. and Perez-Pons, J. A. and Querol, E. (1997). Relation between amino acid composition and cellular location of proteins. *J. Mol. Biol.* **266**, 594–600.

22. Chou, K. C. (2001). Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins,* **43**, 246–255.

23. Potvin, C. and Leclerc, D. and Tremblay, G. and Asselin, A. and Bellemare, G. (1988). Cloning, sequencing and expression of a *Bacillus* bacteriolytic enzyme in *Escherichia coli. Mol. Gen. Genet.* **214**, 241–248.

24. Takemaru, K. and Mizuno, M. and Sato, T. and Takeuchi, M. and Kobayashi, Y. (1995). Complete nucleotide sequence of a skin element excised by DNA rearrangement during sporulation in *Bacillus subtilis. Microbiology,* **141**, 323–327.

25. Nagaso, H. and Saito, S. and Saito, H. and Takahashi, H. (1988). Nucleotide sequence and expression of a *Streptomyces griseosporeus* proteinaceous alpha-amylase inhibitor (HaimII) gene. *J. Bacteriol.* **170**, 4451–4457.

26. Aronson, A. I. and Song, H. Y. and Bourne, N. (1989). Gene structure and precursor processing of a novel *Bacillus subtilis* spore coat protein. *Mol. Microbiol.* **3**, 437–444.

27. Bourne, N. and FitzJames, P. C. and Aronson, A. I. (1991). Structural and germination defects of *Bacillus subtilis* spores with altered contents of a spore coat protein. *J. Bacteriol.* **173**, 6618–6625.

28. Driks, A. (1999). Bacillus subtilis spore coat. *Microbiol. Mol. Biol. Rev.* **63**, 1–20.

29. Driks, A. (2002). Maximum shields: the assembly and function of the bacterial spore coat. *Trends. Microbiol.* **10**, 251–254.

30. Chada, V. G. and Sanstad, E. A. and Wang, R. and Driks, A. (2003). Morphogenesis of bacillus spore surfaces. *J. Bacteriol.* **185**, 6255–6261.

31. Thomas, G. (2002). Furin at the cutting edge: from protein traffic to embryogenesis and disease. *Nat. Rev. Mol. Cell Biol.* **3**, 753–766.

32. Duckert, P. and Brunak, S. and Blom, N. (2004). Prediction of proprotein convertase cleavage sites. *Protein Eng., Design and Sel.* **17**, 107–112.

33. Nielsen, H. and Engelbrecht, J. and Brunak, S. and von Heijne, G. (1997). A neural network method for identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Int. J. Neural Syst.* **8**, 581–599.

34. Berks, B. C. and Sargent, F. and Palmer, T. (2000). The Tat protein export pathway. *Mol. Microbiol.* **35**, 260–274.

35. Palmer, T. and Berks, B. C. (2003). Moving folded proteins across the bacterial cell membrane. *Microbiology,* **149**, 547–556.

36. Miller, J. R. and Kovacevic, S. and Veal, L. E. (1987). Secretion and processing of staphylococcal nuclease by *Bacillus subtilis. J. Bacteriol.* **169**, 3508–3514.

37. Kristensen, T. and Ogata, R. T. and Chung, L. P. and Reid, K. B. and Tack, B. F. (1987). cDNA structure of murine C4b-binding protein, a regulatory component of the serum complement system. *Biochemistry,* **26**, 4668–4674.

38. Horton, P. and Nakai, K. (1997). Better prediction of protein cellular localization sites with the k nearest neighbors classifier. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **5**, 147–152.

39. Nakai, K. and Horton, P. (1999). PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem. Sci.* **24**, 34–36.

40. Krogh, A. and Larsson, B. and von Heijne, G. and Sonnhammer, E. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* **305**, 567–580.

41. Reczko, M. and Fiziev, P. and Staub, E. and Hatzigeorgiou, A. (2002). Finding signal peptides in human protein sequences using recurrent neural networks. pp. 60–67, Springer-Verlag, Heidelberg, Germany.

42. Käll, L. and Krogh, A. and Sonnhammer, E. L. (2004). A combined transmembrane topology and signal peptide prediction method. *J. Mol. Biol.* **338**, 1027–1036.

43. Bendtsen, J. D. and Jensen, L. J. and Blom, N. and von Heijne, G. and Brunak, S. (2004). Feature based prediction of non-classical protein secretion. *Submitted,* .

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