Nature Methods

SignalP 4.0: discriminating signal peptides from transmembrane regions

Thomas Nordahl Petersen, Søren Brunak, Gunnar von Heijne & Henrik Nielsen

Nature Methods: doi:10.1038/nmeth.1701

Supplementary Methods

Sequence data

We extracted datasets for eukaryotes, Gram-positive bacteria and Gram-negative bacteria from the Uniprot Knowledgebase release 2010_05[1]. Only reviewed entries (i.e. from UniProtKB/SwissProt) were used, and hypothetical proteins were not included. We discarded sequences shorter than 30 amino acids. Fragments were allowed as long as their N-terminal 70 amino acids were complete (i.e. they had no “FT NON_TER” line pointing to position 1 and no “FT NON_CONS” line affecting the first 70 positions). Gram-positive bacteria were defined as *Firmicutes* plus *Actinobacteria*. We did not include *Tenericutes* (*Mycoplasma* and related genera), since they do not seem to have a type I signal peptidase at all.[2] Gram-negative bacteria were defined as all other bacteria.

We only considered positive data (signal peptide sequences) which had experimental evidence for the cleavage site. Lipoproteins (cleaved by Lipoprotein signal peptidase)[3] and signal peptides annotated as being Tat-dependent[4] were excluded from the bacterial sets. Signal peptides from eukaryote and Gram-negative organisms were restricted to have a cleavage site between positions 15 and 45. For Gram-positive bacteria, which tend to have longer signal peptides, the upper limit was extended to allow a cleavage site at position 50.

We extracted two different negative sets. The first consisted of proteins with a subcellular location annotated as cytosolic (cytosolic and/or nuclear in eukaryotes) with experimental evidence (note that UniProt uses the term “cytoplasm” for cytosol). Proteins with known multiple locations (except for cytosolic and nuclear) were not included. The second negative set consisted of transmembrane proteins with an annotated transmembrane region within the first 70 positions. These had experimental evidence for being single- or multi-pass membrane proteins (according to the subcellular location comment), but did not necessarily have experimental evidence for the exact position of the transmembrane region, as this criterion would have left too few sequences in the set.

All sequences were shortened to the 70 N-terminal amino acids. We then carried out homology reduction using algorithm 2 of Hobohm *et al.*[5] and cut-off criteria as described previously[6], *i.e.* a local alignment could include up to 17 identical amino acids for eukaryotic sequences, while for bacteria up to 21 identities were allowed (**Table A**).

In order to benchmark SignalP 4.0 against the previous version of SignalP, we assigned a subset of each of the positive and negative datasets as “comparison dataset” by performing a homology reduction to sequences that were used to train SignalP 3.0 (**Table A**). This means that no sequence in the comparison sets is homologous (according to the aforementioned cut-off criteria) to any sequence used to train SignalP 3.0. The comparison datasets are also used for calculating performances of other methods. For those that are newer than SignalP 3.0 (Philius, SPOCTOPUS, the two MEMSAT versions, Signal-CF and Signal-3L), this can lead to slight overestimations of performance, as these predictors to some extent have been trained on sequences identical

	Signal peptides (SP)			Cytosolic or Nuclear			Transmembrane (TM)		
	Total	Train	Comp.	Total	Train	Comp.	Total	Train	Comp.
Euk	3,595(4)	1,640(2)	606(0)	8,956	5,131	1,000*	2,329	989	939
Gram+	277(1)	208(1)	48(1)	2,914	359	212	186	118	118
Gram–	612(6)	423(4)	104(1)	7,243	908	559	1,054	527	524

Table A. Sequences included in datasets for eukaryotes (Euk), Gram-positive bacteria (Gram+) and Gram-negative bacteria (Gram–). "Total" is the number before homology reduction, "Train" is the homology reduced dataset used for training, "Comp." is the dataset homology reduced against the SignalP 3.0 data, used for comparison to SignalP 3.0 and other methods. The numbers in parentheses are sequences that include both a signal peptide and a transmembrane region within the first 70 positions. The number marked with (*) was artificially reduced to make a more balanced comparison set.

or similar to the sequences in our comparison datasets. The data used to train and test SignalP 4.0 are available at http://www.cbs.dtu.dk/services/SignalP/ under the heading "Data."

Neural Networks

We trained standard feed forward neural networks by the use of a back-propagation procedure[7] to recognize the signal peptide and the cleavage site in protein sequences. A sliding window of up to 41 amino acids was fed into the first layer of the neural network and predictions were made for the central position in the window. Amino acids were sparsely encoded as an orthogonal vector of length 21, where the last entry represented a position outside the protein sequence. In general we trained neural networks for two different tasks, one to predict positions within the signal peptide (SP) and another to predict the cleavage site in each sequence (CS). The output representing 'signal peptide' in the SP networks is called the S-score, while the output representing 'cleavage site' in the CS networks is called C-score (**Figure A**).

In contrast to the earlier versions of SignalP, where target values of 0 and 1 were used, SignalP 4.0 uses target values of 0.1 and 0.9 for incorrect and correct outputs, respectively. This creates a visual difference in the output compared to that of SignalP 3.0 (**Figure A**).

A number of different architectures were tried. For the SP networks, the symmetrical input window was varied from 29 to 41. For the CS networks, the input window was asymmetrical with 16 to 21 positions before the putative cleavage site and 1 to 6 positions after. The size of the hidden layer was varied from 2 to 20 units (**Table B**). Not all intermediate values have been used. In addition to the sequence window inputs, two other input types were used: the relative position of the input window in the sequence, encoded as a number between 0 and 1, and the composition of the entire input sequence (up to 70 residues), expressed as 20 numbers between 0 and 1. For the SP networks, networks both with and without these additional inputs were tried, while for the CS networks, they were always included, since preliminary tests showed them to be advantageous (results not shown).

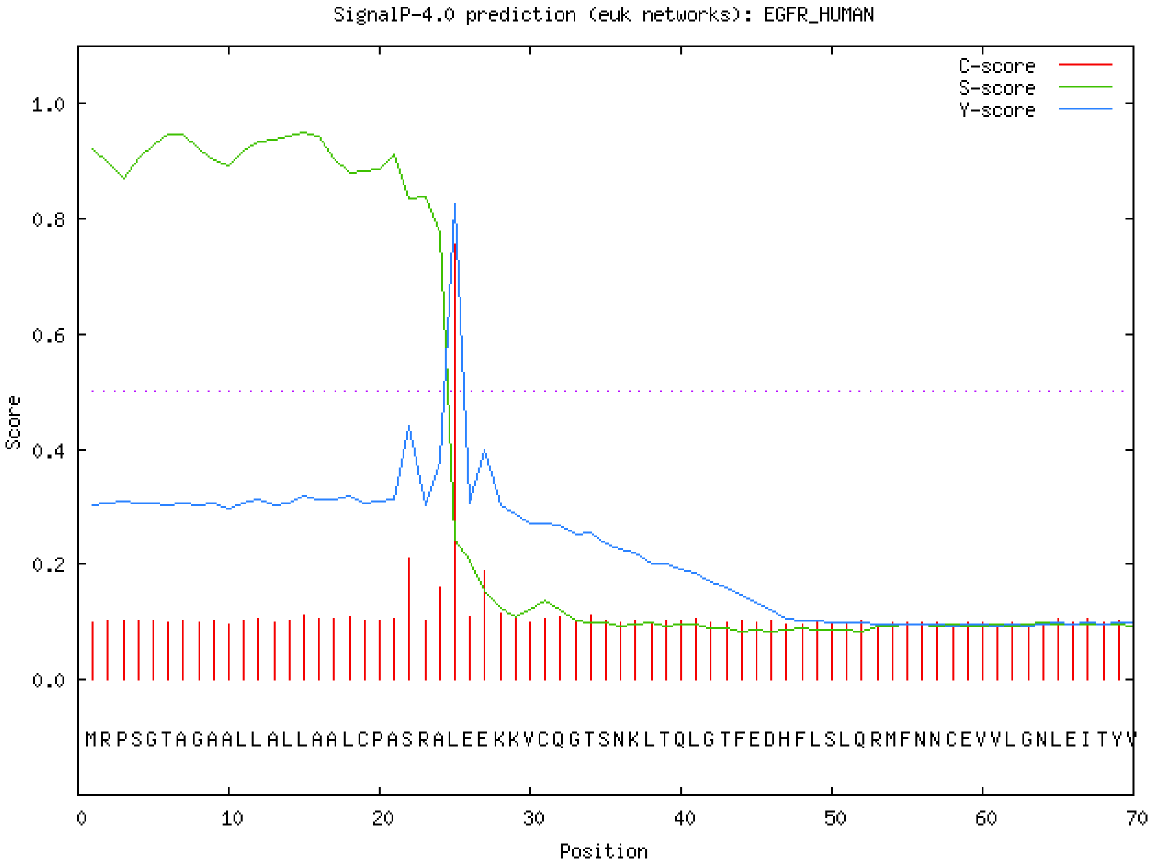

Figure A. Output from the SignalP 4.0 web-server for the sequence EGFR_HUMAN. 'C-score' is the predicted cleavage site value, 'S-score' is the predicted signal peptide value and 'Y-score' is a combination of C- and S-scores as described in Methods. The predicted cleavage site is between position 24 and 25 with the local sequence context "SRA-LE." In a summary line below the plot we show the calculated D-score, the associated cutoff value and which networks were used to make the prediction, in this case: "D=0.868 D-cutoff=0.450 Networks=SignalP-noTM"

Each network was trained for a maximum of 200 training epochs. Optimal architecture and training cycle were selected by the optimization set (see next subsection) according to the MCC (Matthews Correlation Coefficient)[8], where the true and false positives and negatives were counted on the single position level (*i.e.* each single position was counted as a true or false positive or negative).

Nested cross-validation

When prediction methods based on machine learning algorithms are trained and tested, it is customary to use a cross-validation procedure, where the data set is divided into *k* partitions and for each partition a version of the method is trained on *k*-1 partitions, while one partition is left out as a test set for measuring the performance. This is adequate as long as the test set data are not used in any way to optimize the parameters of the method, but if the test set performance is used for "stopping" the network training (*i.e.* selecting the best training cycle) or for selecting the optimal network architecture, the test set performance is no longer an unbiased estimate of how the method will perform on new data, *i.e.* data that have not been used in its development. This may lead to an

Euk	SignalP-TM		SignalP-noTM	
	SP	CS	SP	CS
Window size	41	17+1 - 20+5	41	20+3 - 20+5
Hidden units	16 - 20	2 - 8	8 - 20	2 - 8
Relative position	95%	Always on	55%	Always on
Composition	100%	Always on	95%	Always on
Gram+	SignalP-TM		SignalP-noTM	
	SP	CS	SP	CS
Window size	29 - 37	17+2 - 21+6	29 - 33	18+1 - 20+5
Hidden units	8 - 20	2 - 8	18 - 20	2 - 8
Relative position	80%	Always on	60%	Always on
Composition	90%	Always on	90%	Always on
Gram–	SignalP-TM		SignalP-noTM	
	SP	CS	SP	CS
Window size	33 - 41	16+1 - 21+6	31 - 41	17+1 - 21+6
Hidden units	8 - 20	2 - 8	8 - 20	2 - 8
Relative position	95%	Always on	60%	Always on
Composition	100%	Always on	90%	Always on

Table B. Parameters selected for training of the neural networks for the eukaryote (Euk), Gram-positive (Gram+) and Gram-negative (Gram–) datasets. The columns SP and CS show the optimal parameters for neural network training of signal peptide and cleavage site prediction, respectively. 'Window size' is the number of input positions used in the first layer of the neural networks for sparse encoding of the amino acids. For the CS networks, the input window is given as number of positions before the cleavage site plus number of positions after the cleavage site. 'Hidden units' is the number of neurons used in the second layer. For 'Window size' and 'Hidden units', the numbers show the range in the 20 networks trained during nested 5-fold cross-validation. 'Relative position' is central position in the sliding window, relative to the length of the protein sequence. 'Composition' is the amino acid composition of the entire sequence. For 'Relative position' and 'Composition', the numbers show the percentage of the 20 networks that were selected to have the corresponding input option on.

overestimation of performance and is a problem with many published methods, including our own SignalP version 3.0.[9]

To overcome this problem, we used a nested cross-validation procedure, where the data were divided into homology-screened training, optimization and evaluation sets. Data were divided into five partitions that are all used in each of the three roles. When partition 1 was used as evaluation set, it was kept out of the process while a full four-fold cross-validation was performed on partitions 2 through 5. In each of these runs, three partitions were used for training the networks, while the optimization set was used for selecting the optimal training cycle and network architecture, and the evaluation set could then be used to obtain an unbiased estimate of the performance. The whole process was then repeated with partition 2 as evaluation set, and so on until all partitions had been used. This gave in total 20 versions of each network.

	SignalP-TM			SignalP-noTM		
	Euk	Gram+	Gram–	Euk	Gram+	Gram–
Range (*d*)	13-20	13	12	20-24	12-15	12
Weight (*w*)	0.58-0.60	0.61	0.51-0.63	0.41-0.54	0.45	0.45-0.53
D_{cut}	0.50-0.51	0.45	0.51-0.53	0.36-0.51	0.44-0.57	0.56-0.59

Table C. Parameters optimized on test datasets for the integration of predictions from signal peptide and cleavage site neural networks. 'Range' is the number of amino acid positions before and after a potential cleavage site from which the mean signal peptide likeliness is calculated. 'Weight' is the term used in the linear combination of Smean and Ymax. 'D_{cut}' is the optimized threshold for the prediction of a potential signal peptide.

When measuring the performance on the evaluation set, the output values of the four networks in the inner loop of the nested cross-validation were arithmetically averaged to give a final score. This tended to give a better result than each of the networks alone (**Table D** in **Supplementary Results**). In the final version of SignalP 4.0 made available on the website, the score is an average of the output values of all twenty versions in the nested cross-validation.

The division into training, optimization and evaluation sets is independent of the assignment of part of the data as comparison set. When calculating the comparison performances for SignalP 4.0 (**Fig. 1** in the main text and **Table E** in **Supplementary Results**), we basically used the evaluation performances, but reduced each evaluation set to those sequences belonging to the comparison sets.

Calculating results at sequence level
After training, the C-score will typically be high at the cleavage site, while the S-score drops from a high value before the cleavage site to a low value after the cleavage site. If the C-score shows several peaks, the correct cleavage site can often be inferred by choosing the peak that occurs where the slope of the S-score is highest. This is formalized by the Y-score introduced in the first SignalP version, which is defined as the geometric average between the C-score and a smoothed derivative of the S-score:

$$Y_i = \sqrt{C_i \Delta_d S_i} \text{ where } \Delta_d S_i = \frac{1}{d}\left(\sum_{j=i-d}^{i-1} S_j - \sum_{j=i}^{i+d-1} S_j \right)$$

The cleavage site is then predicted to be at the position where the Y-score is maximal (see also **Fig. A**).

The mean S-score is calculated as the arithmetic average of the S-score from position 1 to the position of the maximal Y-score. The prediction of whether the entire sequence contains a signal peptide or not is then done by the D-score introduced in the third SignalP version,[9] which is a weighted average of the mean S-score and the maximal Y-score:

$$D = wY_{\text{max}} + (1-w)S_{\text{mean}}$$

A signal peptide is predicted if the D-score is larger than a cut-off value, D_{cut}.

The values of the parameters *d*, *w* and D_{cut} were optimized to yield the highest MCC, where true and false positives and negatives are calculated per sequence instead of per position. This optimization was done in a standard five-fold cross-validation (**Table C**).

Methods for comparison

In addition to the previous version of SignalP (3.0), ten methods were selected for comparison of predictive performances on novel data. The methods Philius,[10] MEMSAT3[11] and MEMSAT-SVM[12] were downloaded and run locally on our computers. Phobius,[13] Signal-BLAST,[14] SPOCTOPUS[15] and PrediSi[16] were used directly on their respective websites. For the methods SPEPlip,[17] Signal-CF[18] and Signal-3L[19] the situation was more complicated, since their websites only allow one sequence to be submitted; in these three cases, we wrote Perl scripts to automate the process of submitting a sequence and collecting the results.

Supplementary Results

As mentioned in the main text, SignalP 4.0 uses two different kinds of negative data: the first, corresponding to the negative data in earlier versions of SignalP, consists of cytosolic and, for the eukaryotes, nuclear proteins; the second consists of sequences without signal peptides but with one or more transmembrane helices within the first 70 positions of the sequence.

The Neural Networks have been trained on two types of input data; one method, SignalP-noTM, has been trained with only the first negative set (*i.e.* cytosolic and nuclear sequences) while the other method, SignalP-TM, is trained with both negative sets. The SignalP-noTM method is similar to the previous versions of SignalP where two output neurons are used both for the SP (signal peptide or not) and the CS networks (cleavage site or not). The SignalP-TM method is different as the SP networks have three output states, representing signal peptide, transmembrane region, or anything else.

We found that the two methods SignalP-TM and SignalP-noTM were to some extent complementary, *i.e.* SignalP-TM did not yield as good results as SignalP-noTM when there were no transmembrane sequences involved (**Table E**). Since the overall goal is to make one combined predictor that can be used to discriminate between signal peptide and transmembrane regions while still having a high performance on any kind of input sequence, the best generic predictor is the one that can decide which of the two methods should be used for any possible sequence.

The solution was to use the ability of SignalP-TM to predict TM-regions, such that if a certain number of positions or more were predicted as part of a TM-region, then the SignalP-TM predictor should be used, with the SignalP-noTM predictor as the default. The number 4 was found to give the best performance for the combined predictor for the eukaryotic and Gram-negative bacterial data (this means that if SignalP-TM predicts a

	Euk.		Gram+		Gram–	
	S-corr.	C-corr.	S-corr.	C-corr.	S-corr.	C-corr.
Training	0.908	0.736	0.935	0.903	0.909	0.908
Optimization	0.872	0.654	0.825	0.631	0.857	0.793
Evaluation (individual)	0.870	0.628	0.806	0.543	0.850	0.747
Evaluation (averaged)	0.881	0.648	0.833	0.579	0.866	0.766

Table D. The effects of the nested cross-validation and of averaging on the SignalP-TM networks. ‘S-corr.’ denotes Matthews Correlation Coefficient (MCC) for the S-score (signal peptide score in the SP networks), while ‘C-corr.’ denotes MCC for the C-score (cleavage site score in the CS networks). Evaluation performance is calculated in two different ways: for each optimization network individually or after averaging the scores of the four optimization networks.

transmembrane helix that is shorter than 4 residues, SignalP-noTM is used). The final predictor SignalP 4.0 uses this selection scheme for these two organism groups, while for Gram-positive bacterial data no improvement was found after applying the selection scheme, so in this case the output is always that of SignalP-TM.

During training, the data were divided into three parts (see **Supplementary Methods** for details regarding nested cross-validation): training (for adjusting the weights of the neural networks), optimization (for selecting the optimal training cycle and neural network architecture), and independent evaluation (for measuring the final performance). The performances of the SignalP-TM networks on the three parts of the data sets are shown as Matthews Correlation Coefficients (MCC)[8] with true and false positives and negatives counted at single position level; the same values that were used for optimization (**Table D**). The S-correlation is a measure of how good the SP networks are at distinguishing positions within signal peptides from positions after the cleavage site and positions in negative sequences. The C-correlation is a measure of how good the CS networks are at distinguishing cleavage site positions from all other positions.

It is evident that the performance measured on the evaluation data were not much lower than that measured on the optimization data (except for C-correlation for the Gram-positive data). The difference would probably be larger if the data had not been subjected to the strict homology reduction as described in **Supplementary Methods**.

The effect of averaging the output scores of the four different optimization networks before calculating the performance on each evaluation set is also shown (**Table D**). This gives a quite remarkable rise in performance – measured by S-correlation, the evaluation performance using averaging even exceeds the optimization performance. This averaging assumes that each network has the optimal architecture parameters for its corresponding optimization set.

The optimal parameter settings for the neural networks are shown in **Supplementary Methods** (**Table B**). For most of the data sets, there was quite a large variation in the optimal parameters over the twenty versions of the nested cross-validation, both in input window size and hidden layer size. The extra inputs (relative position in the sequence and amino acid composition) were selected to be included in most of the SP networks. Note, however, that relative position was more important for SignalP-TM than for SignalP-noTM, indicating that this information helps in discriminating between signal peptides and transmembrane helices.

We benchmarked SignalP 4.0 against the previous version and ten other signal peptide prediction methods (**Fig. 1** in the main text and **Table E i-iii**). Performance values for the best prediction methods have been highlighted in bold in the table. The first column (signal peptide correlation for the entire data set) corresponds to the values shown in the figure. Benchmarking is done using the comparison datasets (**Table A**). For the methods developed in this project (SignalP-noTM, SignalP-TM and SignalP 4.0), the results are evaluation performances, *i.e.* for each version of the network they are calculated on the corresponding evaluation part of the comparison dataset. Signal peptide performance is given as the MCC,[8] where true and false positives and negatives are counted at the sequence level (*i.e.* each sequence was counted as a true or false positive or negative). Cleavage site performance is given as two measures: sensitivity, *i.e.* the percentage of signal peptides that are predicted as a signal peptide and have the correct cleavage site assigned, and precision, *i.e.* the percentage of cleavage site predictions that are correct. In addition, we show the false positive rate specifically for the negative set with transmembrane proteins, and the signal peptide correlation for the case where no transmembrane proteins were included in the negative set (the situation for which SignalP-noTM is optimized).

SignalP 4.0 is best at the prediction of signal peptides, measured by correlation coefficient on the entire data set, for all three organism types. However, the cleavage site sensitivity is not as high as that of SignalP 3.0. In general, we believe that the much lower level of false positives from transmembrane regions is more important than the few percent worse cleavage site sensitivity. However, we have chosen to give the user an opportunity of running SignalP-noTM instead of SignalP 4.0 by telling the server that the submitted sequences do not contain transmembrane regions.

SignalP-noTM should be directly comparable to SignalP 3.0 NN (the neural network part of SignalP 3.0). For the bacterial datasets, SignalP-noTM is the best method according to correlation coefficient when there are no transmembrane sequences in the data, while the eukaryotic SignalP-noTM is slightly worse than SignalP 3.0 NN. When measured by cleavage site sensitivity, SignalP-noTM is better than SignalP 3.0 NN only for Gram-positive bacteria.

If the values for SignalP 3.0 are compared to those given in the SignalP 3.0 article,[9] it must be taken into account that we have changed the way cleavage site performance is calculated in order to be able to benchmark SignalP against other methods. In the SignalP 3.0 article, it was the percentage of cleavage sites predicted by the position of the maximal Y-score, regardless of whether the D-score was above threshold or not (see

i: *Eukaryotic sequences*					
Method	All Sequences			Only TM	No TM
	SP corr.	CS sens. (%)	CS prec. (%)	FP-rate (%)	SP corr.
SignalP 4.0	**0.874**	68.3	65.9	6.1	0.919
SignalP-TM	0.871	66.2	**68.2**	**3.3**	0.909
SignalP-noTM	0.674	71.3	44.2	38.1	0.960
SignalP 3.0 NN	0.686	**72.9**	45.1	39.5	**0.968**
SignalP 3.0 HMM	0.762	72.1	51.8	23.2	0.932
PrediSi	0.561	66.0	37.6	52.6	0.909
SPEPlip	0.717	66.5	46.4	30.7	0.953
Signal-CF*	0.475	62.3	28.0	74.8	0.878
Signal-3L*	0.432	53.3	24.5	73.7	0.821
SignalBlast SP1	0.690	29.4	19.1	27.3	0.851
SignalBlast SP2	0.739	30.0	24.3	16.0	0.843
SignalBlast SP3	0.736	24.6	26.5	7.2	0.774
SignalBlast SP4	0.546	29.5	14.7	51.3	0.794
Phobius	0.811	62.4	50.8	15.3	0.931
Philius	0.800	65.8	54.3	14.0	0.902
MEMSAT3	0.252	0.2	0.2	26.7	0.468
MEMSAT-SVM	0.381	0.8	1.0	16.6	0.488
SPOCTOPUS	0.776	34.7	30.2	13.8	0.880

ii: *Gram-positive bacterial sequences*					
Method	All Sequences			Only TM	No TM
	SP corr.	CS sens. (%)	CS prec. (%)	FP-rate (%)	SP corr.
SignalP 4.0	**0.851**	72.9	**83.3**	**2.6**	0.894
SignalP-TM	**0.851**	72.9	**83.3**	**2.6**	0.894
SignalP-noTM	0.556	77.1	36.3	47.9	**0.948**
SignalP 3.0 NN	0.618	72.9	41.7	35.0	0.934
SignalP 3.0 HMM	0.558	77.1	38.1	43.5	0.895
PrediSi	0.607	60.4	38.7	28.2	0.880
SPEPlip	0.488	60.4	28.7	50.4	0.866
Signal-CF*	0.415	**78.3**	23.4	29.8	0.814
Signal-3L*	0.420	34.8	10.5	84.6	0.799
SignalBlast SP1	0.575	22.9	11.8	33.3	0.804
SignalBlast SP2	0.684	22.9	15.9	19.7	0.868
SignalBlast SP3	0.711	22.9	16.9	17.1	0.881
SignalBlast SP4	0.419	22.9	7.3	62.4	0.659
Phobius	0.704	60.4	43.9	18.8	0.894
Philius	0.726	52.1	41.7	12.8	0.854
MEMSAT3	0.101	0.0	0.0	10.3	0.261
MEMSAT-SVM	0.675	0.0	0.0	11.1	0.812
SPOCTOPUS	0.625	35.4	34.0	14.5	0.807

iii: *Gram-negative bacterial sequences*					
	All Sequences			Only TM	No TM
Method	SP corr.	CS sens. (%)	CS prec. (%)	FP-rate (%)	SP corr.
SignalP 4.0	**0.848**	65.4	70.8	1.5	0.882
SignalP-TM	0.815	61.5	**75.3**	**1.1**	0.839
SignalP-noTM	0.497	71.2	26.1	35.8	**0.948**
SignalP 3.0 NN	0.542	74.0	30.8	28.5	0.925
SignalP 3.0 HMM	0.477	**76.9**	26.1	39.2	0.931
PrediSi	0.479	75.0	27.2	35.6	0.901
SPEPlip	0.429	70.2	21.4	45.1	0.891
Signal-CF*	0.288	73.1	13.8	78.1	0.698
Signal-3L*	0.287	73.1	13.5	81.1	0.714
SignalBlast SP1	0.530	39.4	14.6	25.4	0.767
SignalBlast SP2	0.252	18.3	3.2	72.8	0.543
SignalBlast SP3	0.642	34.6	22.8	11.5	0.836
SignalBlast SP4	0.387	39.4	9.4	46.1	0.635
Phobius	0.586	73.1	33.6	23.3	0.920
Philius	0.639	**76.9**	26.1	15.7	0.872
MEMSAT3	0.084	0.0	0.0	17.8	0.312
MEMSAT-SVM	0.497	1.0	0.6	16.4	0.780
SPOCTOPUS	0.510	33.7	18.6	20.5	0.848

Table E. Benchmarking of signal peptide and cleavage site predictions on the comparison dataset for all three organism groups. ‘SP corr.’ denotes signal peptide correlation, while ‘CS sens.’ denotes cleavage site sensitivity (the percentage of actual cleavage sites that are predicted correctly) and ‘CS prec.’ denotes cleavage site precision (the percentage of predicted cleavage sites that are correct). ‘FP-rate’ (false positive rate) is the percentage of transmembrane sequences that are incorrectly predicted as a signal peptide. ‘No TM’ denotes the test where there were no transmembrane sequences in the data, *i.e.* only the first negative set was used. Note that performance for SignalP 4.0 and SignalP-TM are identical for Gram-positive bacterial sequences, since SignalP 4.0 does not use the combination scheme for this organism group. The methods indicated with a star (*) can only make predictions for sequences longer than 50 aa. For those methods the evaluation sets were reduced by 4, 2, and 22 sequences for the Eukaryote, Gram-positive and Gram-negative sets, respectively.

Supplementary Methods for definitions). In other words, a cleavage site could be scored as correct even though the sequence was not predicted to be a signal peptide. This necessarily gives a higher number than the cleavage site sensitivity (**Table E**). In addition, our use of the nested cross-validation (**Supplementary Methods**) results in a slightly more conservative performance estimation than the plain cross-validation used in the SignalP 3.0 article.

The other methods that were designed to discriminate between signal peptides and transmembrane sequences (Philius, Phobius and SPOCTOPUS) have also high correlation coefficients for eukaryotic data, while they are somewhat worse for the bacterial sequences. This may be due to the fact that these three methods did not divide their training data into different organism groups but pooled them all together, resulting in methods that are optimized for the most abundant organism group in the data, the eukaryotes. The cleavage site performance is rather high for Phobius and Philius, especially regarding cleavage site sensitivity on the Gram-negative bacterial sequences, where Philius shares the winning position with SignalP 3.0 HMM (the Hidden Markov Model part of SignalP 3.0), while SPOCTOPUS is not as good at locating cleavage sites correctly. The number of false-positive signal peptide predictions from transmembrane sequences is lower for these methods than for most of the dedicated signal peptide predictors, but still much higher than for SignalP 4.0.

MEMSAT3 and MEMSAT-SVM do surprisingly bad, especially regarding cleavage sites where they are almost always wrong. In all fairness, it should be said that the authors of these two methods never promoted them as signal peptide prediction tools. However, since they do provide predictions of signal peptides in their output, users may be led into believing that these predictions are useful, which our analysis indicates that they are not.

Signal-CF and Signal-3L have fairly high cleavage site sensitivities (except for Signal-3L for Gram-positive bacteria, where Signal-CF is the winner), but this should be seen in context of their low cleavage site precisions and very high false positive rates on transmembrane regions. On eukaryotic data, these two methods predict a signal peptide for almost 75% of the transmembrane helices. Signal-3L had a peculiar problem with its cleavage site assignment, since it in some cases produced an output stating that there was a signal peptide from position 1 to –1. In these cases, we scored it as a positive prediction with wrong cleavage site. This happened 18 times in the eukaryotic set and 22 times in the Gram-negative set.

Signal-BLAST differs from all the other methods, since it uses alignments to known signal peptides for prediction instead of machine learning or statistical methods. Signal-BLAST can be run in four modes optimized for different situations, and we have tested all four. The best mode according to signal peptide correlation is SP3 (optimized for accuracy and specificity) which almost matches Phobius and Philius in correlation coefficient and has very few false positives. However, the cleavage site performance of all Signal-BLAST versions is rather low. This is an interesting observation, since it suggests that cleavage sites are much less conserved in evolution than signal peptides are.

In contrast to earlier versions of SignalP, we have measured the performance of SignalP 4.0 using nested cross-validation (see **Supplementary Methods** for details), which ensures total separation between the data used for optimizing the neural networks and data used for testing the performance. However, we have shown that the overestimation of performance due to early stopping and architecture selection is not large. On the other hand, our use of the nested cross-validation procedure has shown that averaging over several networks is very efficient for improving performance.

As described in the main text, the data were divided into eukaryotic, Gram-positive bacterial, and Gram-negative bacterial sets. Since some differences are also known to occur between mammalian and yeast signal peptides,[20] we tested a further subdivision of the eukaryotic data into animals, fungi, and plants, and trained separate methods for these three groups. However, this did not give any improvement, and performance for all three groups was better when using the method trained on all eukaryotic sequences (results not shown).

When comparing the SignalP-noTM part of SignalP 4.0 with SignalP 3.0 NN, one would expect to see an improvement in performance, since the data set has grown. This is not the case for the eukaryotic data, where both the cleavage site performances and the correlation coefficients are slightly worse. One reason for this might be that we did not discard sequences with rare amino acids at the **–**1 position relative to the cleavage site, as was done in the data set of SignalP 3.0. Our reasoning was that since all signal peptides are annotated as experimentally verified, the method should be able to recognize even those with atypical cleavage sites; but maybe these outliers actually disturb the training and make the method perform slightly worse. This should be thoroughly tested in the next SignalP version.

References

1. The UniProt Consortium The Universal Protein Resource (UniProt) in 2010. *Nucl. Acids Res.* **38**, D142-148 (2010).
2. Fraser, C.M. et al. The minimal gene complement of Mycoplasma genitalium. *Science* **270**, 397-403 (1995).
3. Juncker, A.S. et al. Prediction of lipoprotein signal peptides in Gram-negative bacteria. *Protein Sci*. **12**, 1652-62 (2003).
4. Bendtsen, J.D., Nielsen, H., Widdick, D., Palmer, T. & Brunak, S. Prediction of twin-arginine signal peptides. *BMC Bioinformatics* **6**, 167 (2005).
5. Hobohm, U., Scharf, M., Schneider, R. & Sander, C. Selection of representative protein data sets. *Protein Sci.* **1**, 409-417 (1992).
6. Nielsen, H., Engelbrecht, J., von Heijne, G. & Brunak, S. Defining a similarity threshold for a functional protein sequence pattern: the signal peptide cleavage site. *Proteins* **24**, 165-77 (1996).
7. Rumelhart, D.E., Hinton, G.E. & Williams, R.J. Learning internal representations by error propagation. In *Parallel distributed processing: explorations in the microstructure of cognition, vol. 1: foundations*, 318-362 (MIT Press, Cambridge, MA, 1986).
8. Matthews, B.W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* **405**, 442-451 (1975).

8. Bendtsen, J.D., Nielsen, H., von Heijne, G. & Brunak, S. Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.* **340**, 783-95 (2004).
9. Reynolds, S.M., Käll, L., Riffle, M.E., Bilmes, J.A. & Noble, W.S. Transmembrane Topology and Signal Peptide Prediction Using Dynamic Bayesian Networks. *PLoS Comput. Biol.* **4**, e1000213 (2008).

11. Jones, D.T. Improving the accuracy of transmembrane protein topology prediction using evolutionary information. *Bioinformatics* **23**, 538 -544 (2007).

12. Nugent, T. & Jones, D. Transmembrane protein topology prediction using support vector machines. *BMC Bioinformatics* **10**, 159 (2009).
12. Käll, L., Krogh, A. & Sonnhammer, E.L.L. A Combined Transmembrane Topology and Signal Peptide Prediction Method. *J. Mol. Biol.* **338**, 1027-1036 (2004).
14. Frank, K. & Sippl, M.J. High-performance signal peptide prediction based on sequence alignment techniques. *Bioinformatics* **24**, 2172 (2008).
15. Viklund, H., Bernsel, A., Skwark, M. & Elofsson, A. SPOCTOPUS: a combined predictor of signal peptides and membrane protein topology. *Bioinformatics* **24**, 2928-2929 (2008).
16. Hiller, K., Grote, A., Scheer, M., Munch, R. & Jahn, D. PrediSi: prediction of signal peptides and their cleavage positions. *Nucl. Acids Res.* **32**, W375-379 (2004).
17. Fariselli, P., Finocchiaro, G. & Casadio, R. SPEPlip: the detection of signal peptide and lipoprotein cleavage sites. *Bioinformatics* **19**, 2498-2499 (2003).
17. Chou, K.-C. & Shen, H.-B. Signal-CF: A subsite-coupled and window-fusing approach for predicting signal peptides. *Biochem. Biophys. Res. Commun.* **357**, 633-640 (2007).
18. Shen, H.-B. & Chou, K.-C. Signal-3L: A 3-layer approach for predicting signal peptides. *Biochem. Biophys. Res. Commun.* **363**, 297-303 (2007).
20. Bird, P., Gething, M.J. & Sambrook, J. Translocation in yeast and mammalian cells: not all signal sequences are functionally equivalent. *J. Cell Biol*. **105**, 2905-2914 (1987).