Refinement of The Prediction Methods of Signal Peptides for The Genome Analyses of *Saccharomyces cerevisiae* and *Bacillus subtilis*

Kenta Nakai
nakai@imcb.osaka-u.ac.jp

Institute for Molecular and Cellular Biology, Osaka University
1-3 Yamada-oka, Suita 565 Japan

Abstract

Since signal peptides play a crucial role for specifying the in-vivo *fate of proteins, prediction of their existence is important for the characterization of ORFs of unknown function. To make such predictions as reliable as possible, the features of signal peptides of two important model organisms,* Saccharomyces cerevisiae *and* Bacillus subtilis, *were examined and the accuracy of current prediction methods was refined using these data. Direct optimization of the threshold values of existing methods significantly raised the predictability but the variables that were most effective for improvement were different in these two organisms. In yeast, the maximum hydrophobicity value of an 8-residue segment mainly contributed to raising the predictability to 98.5% when estimated by the cross validation procedure. In* Bacillus *species, the length of uncharged segment and the charges in the N-terminal region (net charge and negative charge) were combined to give a prediction accuracy of 98.2% although the data size was relatively small in this case.*

1 Introduction

In recent years, the entire genome sequences of several micro-organisms have been determined. In these organisms, in yeast particularly, the systematic characterization of their candidate gene-products has emerged as a next enterprise. The information of the subcellular localization sites of these gene products is one of its important issues because it can be useful for deducing their cellular function. We have developed, and have released through the Internet, a prediction system for protein localization sites, named PSORT [12, 13]. In our current efforts to finalize the new version of PSORT ([7]; K. Nakai and P. Horton, manuscript in preparation), I report here our progress on just one topic: the prediction of signal peptides (leader sequences). In

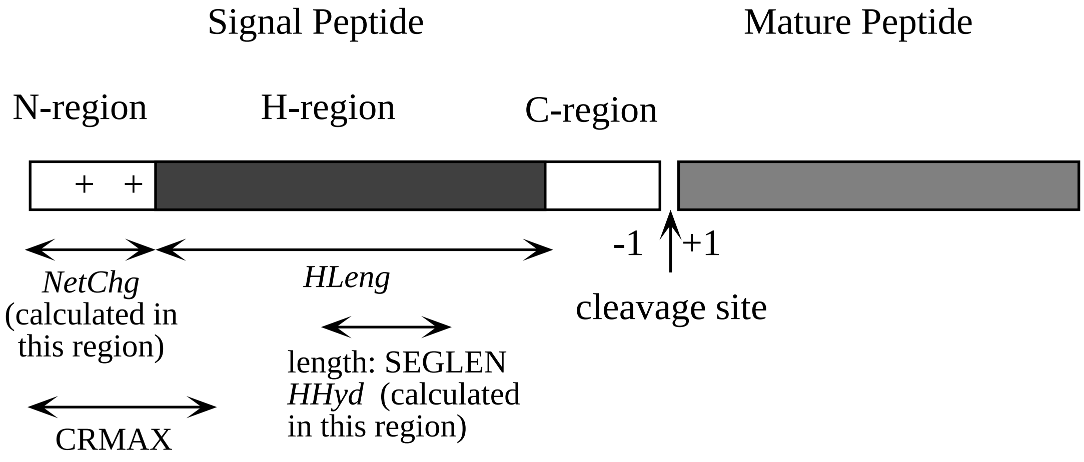

Figure 1: Typical structure of signal peptides

eukaryotic cells, accurate prediction of signal peptides is especially important because signal peptides are required as an essential signal for its first sorting step into several localization sites. In prokaryotic cells also, the signal peptide is a major sorting signal. Although the basic sequence features of both signal peptides are similar, there exist some differences between species. For example, differences between the signal peptides of *Escherichia coli* and those of *Bacillus subtilis,* have been noted [11, 14]. Thus, to detect these signals more accurately, one should refine one's prediction method to reflect such species-specific features.

In the current PSORT program, two complementary methods are mainly used for the detection of signal peptides: McGeoch's method (abbreviated as McG; [9]) and von Heijne's method (abbreviated as GvH; [16]). The former method tries to detect their internal structure (Figure 1); That is, the most N-terminal region, where a few positively-charged residues are often observed (N-region), the central uncharged region, rich in hydrophobic residues (H-region), and the most C-terminal region, where a weak consensus sequence for signal peptidase exists (C-region; [15]). The McG method has been implemented in PSORT using the discriminant analysis by Nakai and Kanehisa [12]. The other method, GvH, is a variation of the so-called weight-matrix method in which the features of both the H-region and the C-region are detected. In our previous analysis, the prediction accuracy of McG appeared to be superior but the information of cleavage sites predicted by GvH was rather accurate and useful. In this work, I collected a number of amino acid sequences of two important model organisms, *Saccharomyces cerevisiae* and *Bacillus* species, with and without signal peptides. By including their species-specific features, both the McG and GvH methods were refined to give higher reliability.

2 Data and Methods

2.1 Yeast Data

Sequence data with signal-peptide information were collected based on the annotation of SWISS-PROT (Rel. 33, [2]) and YPD (Rel. 5.0, [5]). From SWISS-PROT, 39 *Saccharomyces* *cerevisiae* entries in which the length of their signal peptide is described were selected. In principle, the information noted with the keywords such as "probable" or "potential" was not included but the information suggested "by similarity" was used. We call this positive data set *Pos1.* The second data set, *Pos2,* were derived by merging and eliminating the redundancy of both YPD and SWISS-PROT information. Note that the YPD does not explicitly distinguish between experimentally-derived information and information based on prediction. To remove redundant information, one member of any similar pair with more than 50% identical residues was discarded (a global alignment program ALIGN [10] was used). *Pos2* contains 70 sequences. Using the SWISS-PROT database, the contamination of cleavage-site information for propeptides were checked when possible. For negative data, *i.e.*, data not containing the signal peptides, yeast SWISS-PROT sequences whose localization sites are either nucleus, cytoplasm, mitochondrion, or peroxisome were collected based on the YPD annotation. Using the BLASTP [1] and the ALIGN programs, similar (more than 50% identical) sequences were discarded. This data set, *Neg,* contained 1089 sequences.

2.2 *Bacillus* Data

The data for signal peptides from *Bacillus* species, including *B. subtilis,* were collected both from the SWISS-PROT database (Rel. 33) and from the literature [11, 14]. These data include "potential" information especially on the positions of cleavage sites. The sequences of *Bacillus* cytoplasmic proteins were collected from the SWISS-PROT annotation. In these positive and negative data, redundant (more than 50% identical) sequences are checked and short (less than 20 residues) sequences were eliminated but some fragment sequences of cytoplasmic proteins were retained. The data size was 36 for the positive non-redundant set, *Pos1,* and 73 for the positive redundant set, *Pos2,* while the number of negative data, *Neg,* was 97. Four lipoprotein sequences were also collected.

2.3 Conventional Methods

For the discriminant analysis, the DISCRIMINANT program in the SPSS program package (6.1J for the Macintosh) was used. Each variable was removed from or added to the calculation when the stepwise option is used. The optimum threshold value was determined to minimize the total number of errors.

In the cross-validation procedure, the data are divided into 10 subgroups of nearly equal size; their ratios of positive to negative data are also taken to be nearly equal. When each subgroup is used as testing data, the remainders are used as training data, *i.e.,* data for optimizing the parameter set. The obtained parameters are used for predicting the testing data for each subgroup and these 10 predictability values are averaged.

2.4 The McG algorithm (PSORT version)

In McG, the N-region (see Introduction and Figure 1) of signal sequences is defined as a region from the N-terminus to the most C-terminal charged residue (K, R, H, D, or E) within the position specified by the CRMAX parameter (originally set to 11). The net charge, *i.e.,* the number of positively-charged residues (K, R, and H) subtracted by the number of negatively-charged residues (E and D), within this region is used as a variable (*NetChg*) for later calculation. The border between the H-region and the following C-region is often unclear because a cleavage site sometimes exists in the midst of an uncharged segment that characterizes the H-region. Thus, McGeoch simply measured the length of the H-region from the next residue from the N-region to the first downstream charged residue (variable: *HLeng*). Another variable that characterizes the H-region is the maximum hydrophobicity value of sequence segments of length 8 (specified by the SEGLEN parameter) within the N-terminal 30 residues (variable: *HHyd*). McGeoch originally showed that the combination of *HLeng* and *HHyd* is effective to discriminate the sequences with signal peptides. In the PSORT implementation, a linear combination of *NetChg*, *HLeng*, and *HHyd* were optimized using discriminant analysis. The derived coefficients turned out to be applicable for the prediction of both prokaryotic and eukaryotic signal peptides [12, 13].

3 Analysis of Yeast Data

3.1 Basic Features

Table 1: Basic statistics of yeast signal peptides

	Pos1				*Pos2*			
Item	*mean*	*s.d.*	*min*	*max*	*mean*	*s.d.*	*min*	*max*
signal peptide length	20.6	4.3	17	42	22.3	5.2	10	42
N-region length	3.7	2.9	1	10	4.5	2.8	1	11
N-region net charge (*NetChg*)	+0.9		0	+3	+1.1		-1	+3
H-region length (*HLeng*)	17.3	4.7	11	36	17.9	6.8	10	37

In table 1, some statistical values are listed. The average signal length is 20.6 for *Pos1* (22.3 for *Pos2*). The minimum length is 10 (SWISS-PROT accession number: P14020) but it seems to be an exceptional case. If excluding this case, the minimum length is 17 and the maximum is 42.

The McG procedure for locating the N-region was checked whether it can correctly include the N-terminal positively-charged residues of all yeast signal peptides. It succeeded in all sequences of *Pos1* but it failed in two cases of *Pos2*. Of these, one was rescued if histidine residue is not considered to be charged while the other (P27614), whose N-region length is 19, was rather unusual . Thus the current value of CRMAX seems reasonable; if it is set to too large a value, it may detect a charged-residue in the C-region. McGeoch described that in the N-region net charges were clustered between -1 to +2. In our data, they are between 0 to +3 (average +0.9 for *Pos2*) if we exclude the two exceptional cases mentioned above. It should be noted that in the 'confirmed' data (*Pos1*), no negatively-charged residues are observed in this region at all.

If measured following to McG's definition, the length of the central uncharged segment (*HLeng*) varies from 11 to 36 in *Pos1* and from 2 to 37 in *Pos2* (when the above two cases were excluded). Again, if histidine is not regarded as a charged residue, the minimum length 2 is restored to 22 and the new minimum value becomes 10.

In summary, the previous parameters of McG still seem to be suitable for the analysis of yeast signal peptides but histidine should be omitted from the list of charged residues in McG's algorithm.

3.2 Cleavage Site Consensus

Table 2: Frequent amino acids around the cleavage site of yeast signal peptides

-5		-4		-3		-2		-1		+1	
S	20.0%	L/T	15.7%	V	28.6%	L/S	18.6%	A	61.4%	A	14.3%
N	12.9%	V	14.3%	A	21.4%	V	17.1%	G	15.7%	T	11.4%
L/A	11.4%	S	11.4%	I/S	11.4%	T/I/A	5.7%	P	4.3%	S/L	10.0%

the *Pos2* data were used.

In table 2, the frequencies of the major amino acids around the cleavage site are listed. The numerals in the first line represent the relative position from the cleavage site. For example, this table shows that both the leucine (L) and serine (S) residues occupy 18.6% at the -2 position. It is well established that there are strong amino acid tendencies both in the sites -3 and -1 [15]. This (-3, -1)-rule also holds with our data although isoleucine seems to be more preferable at the -3 site than in the original report. In addition, there seems to be a weak amino acid preference at the -2 site.

The GvH method can be used for predicting not only the existence of a signal peptide but also the position of its cleavage site. Of the 39 sites in *Pos1,* 32 (82.1%) are predicted correctly. In *Pos2*, the accuracy is 43/70 (61.4%) but the cleavage site information itself is not guaranteed to be accurate in this set.

3.3 Optimum Threshold Value

The simplest way to optimize the existing program to specific data is to modify its threshold (cut-off) value for discriminating two groups. Currently, the value is set to 0.0 for McG and -2.5 for GvH. As shown in table 3, their prediction accuracy with these values is 96.2% (false negative 1/39; false positive 42/1089) and 97.5% (false negative 6/39; false positive 23/1089), respectively (for *Pos1+Neg* data). Optimized threshold value was selected by simply testing a series of values from -5.0 to 5.0 with the interval of 0.1 and choosing the value with least total errors. When multiple threshold values give the same number of least errors, one of them was chosen rather arbitrary. To avoid over-fitting to our rather small data, the cross validation procedure was used. Typical threshold values obtained at each trial are 4.3 (McG) and -1.2 (GvH; the value -2.1 was major in case of *Pos2+Neg*) and they were rather stable during the 10 trials. As shown in table 3, significant improvements were observed for both methods using these new thresholds.

3.4 Further Refinement

Several efforts were made to improve the McG method. Histidine was excluded from the list of charged amino acids. Better parameters for calculating the *HHyd* variable (that is,

Table 3: Prediction accuracy of McG and GvH

Program	*Item*	*Pos1*	*Pos2*
McG	original accuracy	96.2%	96.2%
	optimized accuracy	98.3%	97.2%
	s.d. of opt. accuracy	1.5%	1.2%
	typical threshold value	4.3	4.3
GvH	original accuracy	97.5%	96.8%
	optimized accuracy	98.0%	97.2%
	s.d. of opt. accuracy	0.7%	1.6%
	typical threshold value	-1.2	-2.1

the SEGLEN parameter and a set of hydrophobicity indices for which Kyte-Doolittle's values have been used [8]) were explored; A series of SEGLEN values were tested for both Kyte-Doolittle's and Engelman *et al.*'s parameter sets. The latter set is known to be suitable for the detection of transmembrane helices [4]. For Kyte-Doolittle's parameter with $SEGLEN = 12$ (original value was 8), the least total error number improved from 33 to 29 in the *Pos2+Neg* data. The improvement by employing Engelman *et al.*'s parameter set was more drastic; with $SEGLEN = 8$, the prediction accuracy by using only this variable is estimated to be 98.2% by cross-validation for *Pos2+Neg*. In this case, the original value of SEGLEN marked the best result. As for the *HLeng* variable, some negative sequences showed extraordinary large apparent values (more than 50 while the maximum *HLeng* value for positive data is 40). Then, an *ad-hoc* rule that the values exceeding 60 are converted to 0 was added.

The coefficients to combine the three variables (*NetChg*, *HLeng*, and *HHyd*) were recalculated by discriminant analysis. Surprisingly, the classification accuracy using the obtained discriminant function was only 95.1% for *Pos2+Neg*, which was inferior to the result using the *HHyd* as a single variable. This is explainable by the fact that the discriminant analysis does not minimize the total error number but optimizes another value, Wilks' lambda. For a predictive purpose, the *HHyd* variable is a major indicator while the *NetChg* and the *HLeng* are useful to remove some negative cases. If we exclude the cases with negative net charge value by, say, changing the *HHyd* value to 0, the prediction accuracy is estimated to be 98.5% (s.d. 1.0%) for *Pos2+Neg* and 98.8% (s.d. 0.9%) for *Pos1+Neg*.

4 Analysis of *Bacillus* Data

4.1 Basic Features

In table 4, basic statistical data for *Bacillus* signal peptides are summarized. The average length of the whole signal peptide is 28.7, which is considerably longer than that in yeast. In terms of both the minimum length (23) and the maximum length (41), there seem to be no highly exceptional cases in our data.

The average length of the N-region is 6.8 and its maximum length is 15. Although the original CRMAX value (11) is inappropriate for only two cases, a longer value can be safely

Table 4: Basic statistics of *Bacillus* signal peptides

Item	*mean*	*s.d.*	*min*	*max*
signal peptide length	28.7	4.0	23	41
N-region length	6.8	2.7	3	15
N-region net charge (*NetChg*)	+3.0		+2	+5
H-region length (*HLeng*)	22.9	3.8	15	32

the *Pos1* data were used.

used because it does not falsely include the charged residues in the C-region. As described in the literature, the N-region of *Bacillus* signal peptides is rich in positively-charged residues: the average net charge is +3.0 (histidines were not counted); at least two basic residues always exist in this region while an acidic residue is observed only in one case.

The *HLeng* value (histidines were not counted) is 22.9 on the average, which is comparable with the yeast values. In most cases, the cleavage site was located within these uncharged segments.

The amino acid preference around the cleavage site is well consistent with von Heijne's compilation of prokaryotic sequences (table 5) although our data include not a few cleavage-site information that was only hypothesized from the (-3, -1) rule . Our data also confirmed the previous observation that some amino acids favored at the β-turn conformation (*e.g.*, prolines and glycines) are rich in the positions -5 to -7 [11].

Table 5: Frequent amino acids around the cleavage site of *Bacillus* signal peptides

-5		-4		-3		-2		-1		+1	
T	25.0%	A/S	19.4%	A	63.9%	S	22.2%	A	88.9%	A	38.9%
P/S	13.9%	P/G	11.1%	V	13.9%	F/A	13.9%	S	5.6%	S	11.1%
V	11.1%	T	8.3%	L	8.3%	Q	11.1%	T/G	2.8%	D/E/K/Q	8.3%

the *Pos1* data were used.

4.2 Further Refinement

With the current parameters used in PSORT, the McG and GvH methods could predict the *Pos1+Neg* data with the accuracy of 96.9% and 80.4%, respectively. The former method was refined specific to this data set as follows. Based on the observation above, the CRMAX parameter was set to 15 and histidine was not counted as a charged amino acid. Unlike the yeast case, the effectiveness of Engelman *et al.*'s parameters for discriminating *Bacillus* signal peptides was almost at the same level with Kyte-Doolittle's although the data size used was much smaller in this case; the total error number was 2/133 with the Kyte-Doolittle value averaged over a 12-residue segment while it was 3/133 with the Engelman value over an 8-residue segment. In view of the yeast result, the latter parameters (Engelman, length 8) were adopted for the analyses below. The prediction accuracy for unknown data was estimated to be 96.3% (s.d. 4.9%) using the cross-validation test.

The prediction accuracy of the discriminant analysis method was also explored. Since almost all *Bacillus* signal peptides do not have any negatively-charged residues in their N-region, four variables (*HHyd*, *HLeng*, *NetChg*, and *NegChg* that is the number of negatively-charged residues in the N-region) were included as repertoire variables for the stepwise variable selection. In contrast to the yeast case, all variables except *HHyd* were selected. The obtained formula is:

$$\mathit{score} = 0.92 * HLeng + 2.03 * NetChg + 1.84 * NegChg - 21.3$$

If the score is larger than 0, it is predicted to have a signal peptide. The prediction accuracy estimated by the cross validation was 98.2% (s.d. 3.2%) and it could also discriminate all but one (P38538) positive example in the redundant set, *Pos2*. In addition, the same method could detect all of the 4 type II signals, *i.e.,* signal peptides for lipoproteins, cleaved by signal peptidase II. Though the current data size is totally insufficient, combined use of this method and a sequence motif (PS00013 in PROSITE [3]) could successfully discriminate all of them.

5 Concluding Remarks

In this paper, methods for signal-peptide prediction were refined to give higher reliability using the sequence data of two specific organisms. For the prediction of yeast data, the maximum hydrophobicity value of an 8-residue segment measured by Engelman *et al.*'s scale was the most effective variable while, for *Bacillus* data, the linear combination of other three variables (H-region length, net and negative charges in the N-region) was effective. Does this difference reflect the different nature of signal peptides between yeast and *Bacillus* species? Because the sizes of negative data are totally different (1089 to 97) and because the prediction accuracy seems to have been rather saturated, it is difficult to conclude that this is a significant difference. However, the unique nature of the N-region in *Bacillus* signal peptides is likely to be related to this result. The relatively longer size of *Bacillus* signal peptides also seems to be largely due to this region.

The refined methods will be undoubtedly useful for the characterization of a number of ORFs produced by the genome sequencing projects. One potential difficulty for such optimization may be the difference between the prior probability of a sequence having the signal peptide and the ratio of positive examples in our data. For the case of yeast, the negative data were collected in a large scale and thus the ratio is expected not to differ too much from the 'true' ratio of genes bearing signal peptides in the whole genome. However, a preliminary application to the whole yeast ORFs suggests that a more stringent threshold value seems appropriate. In contrast, for *Bacillus* data, such a 'true ratio' is likely to be significantly different. However, it seems to predict a reasonable number of signal peptides when applied to a large number of ORFs (A. Ogiwara, personal communication).

The current version of PSORT hypothesizes that a sequence which is predicted to have a signal peptide by McG but not by GvH has an uncleavable signal [13]. In this study, this hypothesis could not be tested because sufficient examples of experimentally-confirmed uncleavable signals could not be collected. In addition, the hypothesis that signal peptides whose H-region exceed the normal level of its hydrophobicity or its length may become uncleavable could not be tested. We have been collecting references of experimentally-proven topology information of membrane proteins (T. Shimizu and K. Nakai, unpublished). Such efforts will be clearly useful for the future verification of these hypotheses.

The data sets and programs used in this study will be distributed upon request.

Acknowledgement

I would like to thank to Dr. Naotake Ogasawara for the information on relevant references on *Bacillus* signal peptides and to Mr. Paul Horton for critically reading the manuscript. This work was supported by a Grant-in-Aid (*Genome Science*) for Scientific Research on Priority Areas from The Ministry of Education, Science, Sports and Culture in Japan.

References

[1] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic Local Alignment Search Tool," *J. Mol. Biol.*, Vol. 215, pp.403-410, 1990.

[2] A. Bairoch and R. Apweiler, "The SWISS-PROT protein sequence data bank and its new supplement TREMBL," *Nucl. Acids Res.*, Vol. 24, pp.21-25, 1996.

[3] A. Bairoch, P. Bucher, and K. Hofmann, "The PROSITE database, its status in 1995," *Nucl. Acids Res.*, Vol. 24, pp.189-196, 1996.

[4] D. M. Engelman, T. A. Steitz, and A. Goldman, "Identifying nonpolar transbilayer helices in amino acid ssequences of membrane proteins, " *Annu. Rev. Biophys. Biophys. Chem.*, Vol. 15, pp. 321-353, 1986.

[5] J. I. Garrels, http://www.proteome.com/YPDhome.html.

[6] A. Goffeau, K. Nakai, P. Slonimski, and J.-L. Risler, "The membrane proteins encoded by yeast chromosome III genes," *FEBS Lett.*, Vol. 325, pp. 112-117, 1993.

[7] P. Horton and K. Nakai "A Probabilisitc Classification System for Predicting the Cellular Localization Sites of Proteins," *Proc. Fourth Internat. Conf. Intelligent Systems for Molecular Biology*, AAAI Press, pp. 109-115, 1996.

[8] J. Kyte and R. F. Doolittle, "A simple method for displaying the hydropathic character of a protein," *J. Mol. Biol.*, Vol. 157, pp. 105-132, 1982.

[9] D. J. McGeoch, "On the predictive recognition of signal peptide sequences," *Virus Res.*, Vol. 3, pp. 271-286, 1985.

[10] E. Myers and W. Miller, "Optimal Alignments in Linear Space," *CABIOS*, Vol. 4, pp. 11-17, 1988.

[11] V. Nagarajan, "Protein Secretion," in "*Bacillus subtilis* and other Gram-Positive Bacteria (A. L. Sonenshein, J. A. Hoch, and R. Losick eds)," American Society for Microbiology, Washington, D.C., pp. 713-726, 1993.

[12] K. Nakai and M. Kanehisa, "Expert system for predicting protein localization sites in Gram-negative bacteria," *PROTEINS: Structure, Function, and Genetics*, Vol. 11, pp. 95-110, 1991.

[13] K. Nakai and M. Kanehisa, "A knowledge base for predicting protein localization sites in eukaryotic cells," *Genomics*, Vol. 14, pp. 897-911, 1992.

[14] M. Simonnen and I. Palva, "Protein Secretion in *Bacillus* Species," *Microbiolog. Rev.*, Vol. 57, pp. 109-137, 1993.

[15] G. von Heijne, "Patterns of Amino Scids near Signal-Sequence Cleavage Sites," *Eur. J. Biochem.*, Vol. 133, pp. 17-21, 1983.

[16] G. von Heijne, "A new method for predicting signal sequence cleavage sites," *Nucleic Acids Res.*, Vol. 14, pp. 4683-4690, 1986.