SHORT COMMUNICATION

Prediction of Protein Signal Sequences and Their Cleavage Sites

Kuo-Chen Chou*
Computer-Aided Drug Discovery, Pharmacia, Kalamazoo, Michigan

Abstract **Protein signal sequences play a central role in the targeting and translocation of nearly all secreted proteins and many integral membrane proteins in both prokaryotes and eukaryotes. The knowledge of signal sequences has become a crucial tool for pharmaceutical scientists who genetically modify bacteria, plants, and animals to produce effective drugs. However, to effectively use such a tool, the first important thing is to find a fast and effective method to identify the "zipcode" entity; this is also evoked by both the huge amount of unprocessed data available and the industrial need to find more effective vehicles for the production of proteins in recombinant systems. In view of this, a sequence-encoded algorithm was developed to identify the signal sequences and predict their cleavage sites. The rate of correct prediction for 1,939 secretory proteins and 1,440 nonsecretory proteins by self-consistency test is 90.14% and that by jackknife test is 90.13%. The encouraging results indicate that the signal sequences share some common features although they lack similarity in sequence, length, and even composition and that they are predictable to a considerably accurate extent. Proteins 2001;42:136–139.** **© 2000 Wiley-Liss, Inc.**

Key words: topogenic signals; prokaryotic sequence; eukaryotic sequence; discriminant function

INTRODUCTION

Signal sequences of proteins, also called topogenic signals or signal peptides, play a central role in the targeting and translocation of nearly all secreted proteins and many integral membrane proteins in both prokaryotes and eukaryotes.[1,2] The signal peptides from various proteins generally consist of the following three structurally, and, possibly, functionally distinct regions: **(a)** an N-terminal positively charged n-region, **(b)** a central hydrophobic h-region, and **(c)** a neutral but polar c-region.[3] The knowledge of protein signal sequences has become a crucial tool for pharmaceutical scientists who genetically modify bacteria, plants, and animals to produce effective drugs.[4] By adding a specific tag to the desired proteins, one can, for instance, tag them for excretion, making them much easier to harvest. However, to effectively use such a tool, the first important thing is to identify the signal peptides. Because the number of protein sequences entering into data banks has been rapidly increasing, it is time-consuming and costly to identify the signal peptides entirely by experiments. For example, the yearly increment of sequence entries in SWISS-PROT[5] in 1987 was 1,266, and that in 1988 was 3,497, but that in 1997 was already 10,092. In view of this, it is highly desirable to develop a fast and accurate algorithm to identify signal peptides and predict their cleavage sites. However, the existing methods in this area are based mostly on the use of neural networks.[3] Although the results obtained by neural networks were sometimes successful in practice, it was scientifically disappointing that no physical explanation for the possible prediction success was provided. This is because, as pointed out by King[6], the neural networks methods have "very poor explanatory power," "little use of chemical or physical theory," and "are statistically rather poorly characterized." In addition, although the computational costs for training the networks were considerably higher, the prediction accuracy thus obtained was not higher (and sometimes even lower) than the analytical methods. The current study was initiated in an attempt to develop an automated method to identify signal peptides and predict their cleavage sites. The method is expected to have the following features: **(a)** it can fast and accurately predict the desired results, **(b)** it is user-friendly, and **(c)** it is explicitly based on a rational biophysical model that will provide useful insights for understanding the inner workings of such a marvelous protein-sorting machinery at a deeper level.

MATERIALS AND METHODS

Signal peptides comprise the N-terminal part of the amino acid chain (Fig. 1) and are cleaved off while the protein is translocated through the membrane. The length of signal sequence is varied for different proteins. For the data set used in Nielsen et al.,[7] which will be reconsidered here, the shortest signal sequence is 8 amino acids long

*Correspondence to: Kuo-Chen Chou, Computer-Aided Drug Discovery, Pharmacia, Kalamazoo, MI 49007-4940. E-mail: kuo-chen.chou@am.pnu.com

Received 1 June 2000; Accepted 29 August 2000

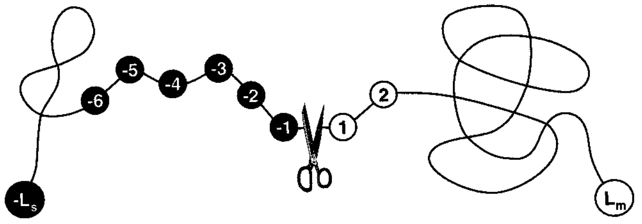

Fig. 1. A schematic drawing to show the signal sequence of a protein and its cleavage site. An amino acid in the signal sequence is depicted as a black circle with a white number to indicate its sequence position, whereas that in the mature protein is depicted as an open circle with a black sequential number. The cleavage site is at the position (−1, +1), i.e., between the last residue of the signal sequence and the first residue of the mature protein.

(L_S=8), and the longest is 90 amino acids (L_S=90). Most of them are within 18–25 amino acids long. Suppose a signal peptide and its cleavage site can be statistically characterized by a dummy sequence symbolized as $[-L_1, +L_2]$, in which L_1 is the number of amino acid residues belonging to the signal part and L_2 the number of residues to the mature part of a protein, and the cleavage site is at the location between residues −1 and +1 (Fig.1). Such a segment can serve as a "benchmark window" to search the secretion-cleavable site along a sequence and deduce its signal peptide accordingly. Without loss of generality, let us consider L_1=6 and L_2=2; i.e., the benchmark window is [−6, +2]. The algorithm thus derived can be easily extended to cover any values of L_1 and L_2. A [−6, +2] sequence can be generally expressed as

$$R_{-6}R_{-5}R_{-4}R_{-3}R_{-2}R_{-1}R_{+1}R_{+2}, \tag{1}$$

where R_{-6} represents the amino acid residue at the nascent protein sequence position −6, R_{-5} the residue at the position −5, and so forth (Fig. 1). The site (−1, +1), i.e., the location between R_{-1} and R_{+1} of the sequence, is the cleavage site during the secretion process and hence all the residues ahead of this site in the nascent protein constitute the signal part. For the sequence as defined by Eq.1, suppose its attributes to the secretion-cleavable set and non-secretion-cleavable set are expressed by $\boldsymbol{\Psi}^+$ and $\boldsymbol{\Psi}^-$, respectively; i.e.,

$$\begin{aligned}\boldsymbol{\Psi}^+&(R_{-6}R_{-5}R_{-4}R_{-3}R_{-2}R_{-1}R_{+1}R_{+2})\\ &= P^+_{-6}(R_{-6})P^+_{-5}(R_{-5})P^+_{-4}(R_{-4})P^+_{-3}(R_{-3})P^+_{-2}(R_{-2})\\ &\qquad P^+_{-1}(R_{-1})P^+_{+1}(R_{+1})P^+_{+2}(R_{+2})\end{aligned} \tag{2}$$

$$\begin{aligned}\boldsymbol{\Psi}^-&(R_{-6}R_{-5}R_{-4}R_{-3}R_{-2}R_{-1}R_{+1}R_{+2})\\ &= P^-_{-6}(R_{-6})P^-_{-5}(R_{-5})P^-_{-4}(R_{-4})P^-_{-3}(R_{-3})P^-_{-2}(R_{-2})\\ &\qquad P^-_{-1}(R_{-1})P^-_{+1}(R_{+1})P^-_{+2}(R_{+2})\end{aligned} \tag{3}$$

where $P^+(R_i)$ is the probability of amino acid R_i occurring at the subsite i (= −6, −5, ..., −1, +1, +2) for the sequences with a secretion-cleaved site at (−1, +1), and P_- (R_i) the corresponding probability for the sequences without any secretion-cleaved site or for those with a secretion-cleaved site located at a position other than (−1, +1). The values of the former can be derived from a positive training data set consisting of only those sequences that have a secretion-cleaved site between R_{-1} and R_{+1}, and the values of the latter can be derived from a negative training data set consisting of only those sequences that have no secretion-cleaved site at all or have one but its location is at any position but (−1, +1). Note that for the current study the location of cleavage site is very important because it is directly correlated with a correct prediction of the signal peptide. For example, instead of the site (−1, +1), if the cleavage site is found at (−2, −1) or (+1, +2), then the corresponding signal peptide thus derived will be one residue shorter or longer than the actual one. Therefore, for brevity hereafter only those sequences with a cleavage site at (−1, +1) are called secretion-cleavable; otherwise, they are non-secretion-cleavable. According to above definition, if a sequence is really secretion-cleavable, the value of its $\boldsymbol{\Psi}^+$ should be greater than that of $\boldsymbol{\Psi}^-$.

Thus, for a given sequence, or an octapeptide as defined in Eq.1, if its attribute function to the secretion-cleavable set is greater than that to the non-secretion-cleavable set, i.e. $\boldsymbol{\Psi}^+ > \boldsymbol{\Psi}^-$, then the sequence is predicted to be secretion-cleavable; otherwise, it is predicted to be non-secretion-cleavable. We define a discriminant function $\boldsymbol{\Delta}$, given by

$$\begin{aligned}\boldsymbol{\Delta}&(R_{-6}R_{-5}R_{-4}R_{-3}R_{-2}R_{-1}R_{+1}R_{+2})\\ &= w^+\boldsymbol{\Psi}^+(R_{-6}R_{-5}R_{-4}R_{-3}R_{-2}R_{-1}R_{+1}R_{+2})\\ &\qquad - w^-\boldsymbol{\Psi}^-(R_{-6}R_{-5}R_{-4}R_{-3}R_{-2}R_{-1}R_{+1}R_{+2})\end{aligned} \tag{4}$$

where w^+ and w^- are the weight factors for the attribute functions derived from the positive training data set S_0^+ and negative training data set S_0^-, respectively. If there is no special reason, they are generally set to be one, i.e., $w^+ = w^- = 1$. Thus, the criterion of the secretion-cleavable peptide prediction for a given sequence can be formulated as follows:

$$\begin{cases}\text{The peptide is secretion-cleavable,} & \text{if its } \Delta > 0\\ \text{The peptide is non-secretion-cleavable,} & \text{otherwise}\end{cases} \tag{5}$$

Note that although the above algorithm was formulated on the basis of an octapeptide segment [−6, +2], it can be straightforwardly extended to any length of segment $[-L_1, +L_2]$.

RESULTS AND DISCUSSION

To calculate the attribute function $\boldsymbol{\Psi}^+$ and $\boldsymbol{\Psi}^-$ for a given sequence, we have to first find the values of $P_i^+(R_i)$ and $P_i^-(R_i)$ (i = ..., −2, −1, +1, +2). These can be derived from a positive training data set S_0^+ and negative training data set S_0^-, respectively. The former consists of only the secretion-cleavable peptides, and the latter only the non-secretion-cleavable peptides. The data sets constructed in Ref. 7 contain 1,939 secretory proteins and 1,440 nonsecretory proteins. Redundant sequences were removed to guarantee that no pairs of homologous sequences exist in the data sets. For the secretory proteins, the sequence of the signal peptide and the first 30 amino acids of the mature protein were included in the data set, whereas for the nonsecretory proteins, the first 70 amino acids of each sequence were included. To compare the performance of prediction under the equivalent condition, we use the same data structure as used in Ref. 7. By sliding the octapeptide benchmark window along each of these sequences, we can generate the desired peptides for the training data sets S_0^+ and S_0^-, respectively. The number of the non-secretion-cleavable peptides thus obtained will be much larger than that of the secretion-cleavable peptides. For example, for a 50 amino acids long secretory protein sequence we can only generate one secretion-cleavable octapeptide but 50 − 8 non-secretion-cleavable octapeptides. For a 70 amino acids long nonsecretory protein sequence we can generate 70 − 8 + 1 non-secretion-cleavable peptides but no secretion-cleavable octapeptide at all. For the current data structure, we have 1,939 secretion-cleavable octapeptides generated for data set S_0^+ and 179,435 non-secretion-cleavable octapeptides for data set S_0^-. Increasing the length of the training peptides will gradually reduce the total number in the training data set.

The rates of correct prediction for the secretion-cleavable peptides and non-secretion-cleavable peptides are given by

$$\begin{cases} \Lambda^+ = \dfrac{N^+ - m^+}{N^+}, & \text{for secretion-cleavable peptides} \\ \Lambda^- = \dfrac{N^- - m^-}{N^-}, & \text{for non-secretion-cleavable peptides} \end{cases} \tag{6}$$

where N^+ represents the total number of secretion-cleavable peptides, and m^+ the number of secretion-cleavable peptides missed in prediction; N^- the total number of non-secretion-cleavable peptides, and m^- the number of non-secretion-cleavable peptides incorrectly predicted as cleavable. The overall rate of correct prediction for the cleavage site and hence the signal peptide concerned is given by

$$\Lambda = \frac{\Lambda^+ N^+ + \Lambda^- N^-}{N^+ + N^-} = 1 - \frac{m^+ + m^-}{N^+ + N^-} \tag{7}$$

To show the power of the current prediction method, the comparison was made with the best result reported by the previous investigators. According to the report by Nielsen et al.,[7] the average rate of correct prediction for the cleavage site location is 71.54%. This is the highest accuracy rate for the data set reported in the literature. Now for the same data set, we used the discriminant function algorithm to perform prediction. The prediction quality was examined by the standard testing procedure in statistics[8] that consists of the self-consistency and jackknife tests. In the former, the cleavage location for each protein in a given data set was predicted by using the parameters derived from the same data set, the so-called training data set, whereas in the latter, each protein in the training data set was singled out in turn as a "test protein" and all the rule-parameters were derived from the remaining proteins. Compared with the independent data set test and subsampling test often adopted in biology, the jackknife test is thought the most effective method for cross-validation in statistics.[8] This is because in the independent data set test, the selection of a testing data set is arbitrary, and the accuracy thus obtained lacks an objective criterion unless the testing data set is sufficiently large.[9] As for the subsampling test in which a given data set is divided into several subsets, the problem is that the number of possible divisions might be too large to be handled. For example, in the treatment by Nielsen et al.,[7] each data set was divided into five approximately equal size parts, and then every network run was conducted with one part as test data and the other four parts as training data. The performance measures were then calculated as an average over the five different data set divisions. Thus, even for a moderate size (e.g., 105 proteins) data set, the number of possible divisions would be 105!/(21!21!21!21!21!) ≈ 3.1×10^{69}. This is an astronomical figure that is too large to be handled by any existing computers. Hence, in any practical subsampling tests as conducted in Ref.7, only a very small fraction of the possible divisions was investigated, and the results thus obtained would certainly bear a considerable arbitrariness. Accordingly, the test procedure combining self-consistency and jackknife tests as adopted here is statistically much more objective and rigorous.

The predicted results by self-consistency and jackknife tests are given in Table I. Note that the number of miss-predicted secretion-cleavable peptides (i.e., m^+ in Eq. 6) also includes those having length shorter than L_1. For example, when [−18, +2] was used as a benchmark window, 186 secretion-cleavable peptides were missed in prediction because, of the 1,939 secretory proteins, 186 have signal sequences with length $L_s < 18$ amino acid residues (see Fig. 1). When [−13, +2] was used as a benchmark window, only 6 secretion-cleavable peptides were missed in prediction because of the $L_s < L_1$ problem. And for the case of [−8, +2], none of such misses happened because no signal sequence in the 1,939 secretory proteins is shorter than 8 residues. The same situation also occurs for the algorithm of Ref. 7. It can be seen from Table 1 that, when [−13, +2] is used as the sliding benchmark window, the results obtained by both self-consistency and jackknife tests have the highest overall success rates, i.e., 90.14% and 90.13%, respectively. These rates are considerably higher than those reported in Ref. 7 based on the same data set. Therefore, from both the rationality of testing procedure and the success rates of test results, it is justified to introduce the sequence-coded discriminant function algorithm as an additional tool for predicting protein signal sequences and their cleavage sites.

The current algorithm, although much simpler, did yield better results. One might naturally ask why. What are its molecular underpinnings? Or what are the biophysical insights from this study? To address these problems, it is instructive to conduct a data analysis as illustrated below. It was found that, within the sequence range [−13, +2] of the 1,939 secretory proteins, most constituent elements

TABLE I. Performance Values by Using Different Benchmark Windows

Benchmark window	Rate of correct prediction for cleavage site location (%)[a]		
	Signal peptides	Nonsecretory proteins	Overall
$[-L_1, +L_2]$	Λ^+	Λ^-	Λ
Self-consistency test			
[−6, +2]	86.95	86.54	86.54
[−8, +2]	87.93	87.94	87.94
[−10, +2]	90.20	89.20	89.21
[−12, +2]	92.30	89.97	90.00
[−13, +1]	92.63	89.95	89.98
[−13, +2]	92.52	90.11	**90.14**
[−14, +1]	92.99	89.90	89.93
[−14, +2]	92.73	90.03	90.06
[−16, +2]	91.18	89.71	89.73
[−18, +2]	85.09	89.68	89.62
Jackknife test			
[−6, +2]	86.23	86.82	86.82
[−8, +2]	87.26	87.02	87.02
[−10, +2]	89.63	89.08	89.08
[−12, +2]	91.28	89.87	89.89
[−13, +1]	92.21	89.95	89.97
[−13, +2]	92.16	90.11	**90.13**
[−14, +1]	92.32	89.90	89.93
[−14, +2]	92.26	89.94	89.97
[−16, +2]	90.30	89.71	89.72
[−18, +2]	84.32	89.69	89.61

[a] See Eqs. 6–7 for the definitions of Λ^+, Λ^-, and Λ.

are, in the order of their occurrence frequencies, Leu, Ala, Ser, Val, and Gly. Of the 20 native amino acids, the occurrence frequency of these five residues alone is >62%. Furthermore, each of the five residues has its own special distribution feature in the sequence range. According to the order of occurrence frequency, most leucines are located at the sequence positions −10, −12, −11; most alanines at −1, −3, +1; most serines at −3, −5, −4; most valines at −3, −12, −13; and most glycines at −1, −5, −4. It can be seen from these data that the probabilities of Ala occurring in the neighboring positions to the cleavage site must be quite high, whereas the same is true for the probabilities of Leu occurring in 10–12 positions preceding the cleavage site. For sequences with a segment pattern like that, it is often quite successful to use the simple probability theory to deal with them, as described in this article. In addition, it has been simultaneously elucidated through the above analysis why the optimal sliding benchmark window should be [−13, +2] for the case studied here.

It has not escaped our note that, because the current model is explicitly correlated with the occurrence frequency of each individual amino acid at a given subsite, it will provide a useful vehicle for helping further investigate many unclear molecular details about the molecular mechanism of the ZIP code protein-sorting system in cells, such as how the signal sequences in the nascent protein chains serve as an address tag to guide the traffic destination of proteins in a cell, and how the cell uses a ZIP code system of sorts to deliver thousands of protein to various addresses within the cell. Moreover, the present method will also have an impact in improving the protein subcellular location prediction. As is well known, knowledge of the subcellular location of a protein is vitally important because the function of a protein is closely correlated with its subcellular location. The existing methods in predicting protein subcellular location are all based on the amino acid composition, as described in a number of articles in this area[10–13] and summarized in a recent review article.[14] This is due to lack of the data of signal sequences although they play a central role in sorting newly synthesized proteins and sending them wherever they are needed. Accordingly, an improvement of predicting protein signal sequences will definitely have an impact in improving the prediction quality of protein subcellular location as well. Further work in these aspects is under way.

ACKNOWLEDGMENTS

The author thanks the anonymous referees whose comments were very helpful for strengthening the presentation of this study and Raymond B. Moeller, Cynthia A. Ludlow, and Wendy Vanderheide for their help in drawing the figure.

REFERENCES

1. Gierasch LM. Signal sequences. Biochemistry 1989;28:923–930.
2. Zheng N, Gierasch LM. Signal sequences: the same yet different. Cell 1996;86:849–852.
3. Claros MG, Brunak S, von heijne G. Prediction of N-terminal protein sorting signals. Curr Opin Struct Biol 1997;**7**:394–398.
4. Hagmann M. Colleagues say ‘Amen’ to this year’s (Nobel Prizes) choices. Science 1999;286:666–666.
5. Bairoch A, Apweiler R. The SWISS-PROT protein sequence data bank and its supplement TrEMBL. Nucleic Acids Res 1997;25:31–36.
6. King RD. Prediction of secondary structure. In: Sternberg MJE, editor. Protein structure prediction: a practical approach. Oxford: IRL Press; 1996. Chapter 4, p 79–97.
7. Nielsen H, Engelbrecht J, Brunak S, von Heijne G. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Eng 1997;10:1–6.
8. Mardia KV, Kent JT, Bibby JM. Multivariate analysis. London: Academic Press; 1979. p 322 and 381.
9. Chou KC, Zhang CT. Prediction of protein structural classes. Crit Rev Biochem Mol Biol 1995;30:275–349.
10. Cedano J, Aloy P, Perez-pons JA, Querol E. Relation between amino acid composition and cellular location of proteins. J Mol Biol 1997;266:594–600.
11. Reinhardt A, Hubbard T. Using neural networks for prediction of the subcellular location of proteins. Nucleic Acids Res 1998;26;2230–2236.
12. Chou KC, Elrod DW. Using discriminant function for prediction of subcellular location of prokaryotic proteins. Biochem Biophys Res Commun 1998;252:63–68.
13. Chou KC, Elrod DW. Protein subcellular location prediction. Protein Eng 1999;12:107–118.
14. Chou KC. Review: prediction of protein structural classes and subcellular locations. Curr Prot Peptide Sci, 2000;1:171–208.