**Protein Engineering** vol.14 no.2 pp.75–79, 2001

**COMMUNICATION**

# Using subsite coupling to predict signal peptides

**Kuo-Chen Chou**

Computer-Aided Drug Discovery, Pharmacia and Upjohn, Kalamazoo, MI 49007-4940, USA. E-mail: kuo-chen.chou@am.pnu.com

**Given a nascent protein sequence, how can one predict its signal peptide or 'Zipcode' sequence? This is a first important problem for scientists to use signal peptides as a vehicle to find new drugs or to reprogram cells for gene therapy. Based on a model that takes into account the coupling effect among some key subsites, the so-called {–3, –1, +1} coupling model, a new prediction algorithm is developed. The overall rate of correct prediction for 1939 secretory proteins and 1440 non-secretary proteins was over 92%. It has not escaped our attention that the new method may also serve as a useful tool for helping investigate further many unclear details regarding the molecular mechanism of the ZIP code protein-sorting system in cells.**

*Keywords*: {–3, –1, +1} coupling/non-secretory proteins/secretory proteins/'Zipcode' sequence

## Introduction

The knowledge of protein signals can be used to reprogram cells in a specific way for future cell and gene therapy. Protein signals have become a crucial tool for researchers to construct new drugs that are targeted to a particular organelle to correct a specific defect. For example, by adding a specific tag to the desired proteins, one can tag them for excretion, making them much easier to harvest (Hagmann, 1999). To use such a tool successfully, first one has to identify the signal sequences. Since the number of nascent protein sequences entering databanks has been rapidly increasing, it is time consuming and costly to identify the signal peptides entirely by experiments. Thus, a strong interest in the automated identification of signal sequences and prediction of their cleavage sites has been evoked. The importance of predicting protein signal peptides has also been elaborated recently in an excellent review by Nakai (2000).

The existing methods in this area are based mostly on the use of neural networks (Claros *et al.*, 1997; Nielsen *et al.*, 1999; Nakai, 2000). They are actually the application of machine learning techniques. As pointed out by King (1996), the advantages of neural network prediction methods are that they are 'readily available' and 'often successful in practice'. He also pointed out that the disadvantages are that 'there is little use of chemical or physical theory', the methods have 'very poor explanatory power—a Hinton diagram means nothing to a protein chemist' and 'they are statistically rather poorly characterized'. Besides, 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 based on the sub-site coupling principle that can be used to identify signal peptides faster and more accurately.

## Materials and methods

Signal peptides comprise the N-terminal part of the secretory protein chain. They control the entry of virtually all proteins to the secretory pathway, in both eukaryotes and prokaryotes (Gierasch, 1989; Rapoport, 1992) and are cleaved off by signal peptidase while the protein is translocated through the membrane. As shown in Figure 1, the cleavage site is at (-1, +1), i.e. the location between residues –1 and +1 or between the last residue of the signal peptide and the first residue of the mature protein. Accordingly, the prediction of the signal peptide of a nascent protein is immediately correlated with the prediction of its cleavage site by the signal peptidase. The length of signal peptides is varied for different secretory proteins. As shown in Figure 2, of the 1939 signal peptides studied by Nielsen *et al.* (1997), the shortest one contains eight amino acid residues and the longest contains 90 residues while the majority have a length within 18–25 residues. The extreme variation in length and sequence has posed a difficulty for formulating a general algorithm to predict the signal peptides. To deal with this kind of situation, let us consider a window with a scale of $\xi_1$, ..., –3, –2, –1, +1, +2, ..., $\xi_2$ (Figure 3). Such a window is called a 'scaled window' and symbolized as $[-\xi_{1,}\ +\xi_2]$. When sliding the scaled window $[-\xi_{1,} +\xi_2]$ along a sequence of *n* residues, one can consecutively highlight $n - (\xi_1 + \xi_2) + 1$ different sequences. Note that for the current study the identification 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 identified at (–2, –1) or $(+1, +2)_,$ then the corresponding signal peptide thus derived will be one residue shorter or longer than the actual one (Figure 1). Therefore, of the sequence segments highlighted by the scaled window, only the one with the residue at the scale –1 being the very last residue of the signal sequence is regarded as the secretion-cleavable segment (Figure 3a); while all the other segments regarded as non-secretion-cleavable (see, e.g., Figure 3b and c). In this way, if sliding the scaled window $[-\xi_{1,}\ +\xi_2]$ along a protein sequence of *n* residues, one can generate one, and only one, secretion-cleavable segment and $n - (\xi_1 + \xi_2)$ non-secretion-cleavable segments if the protein is secretory, but $n - (\xi_1 + \xi_2) + 1$ non-secretion-cleavable segments if it is non-secretory. All the secretion-cleavable segments form a cleavable or positive set denoted by $S^+$ and all the non-secretion-cleavable segments form a non-cleavable or negative set $S^-$.

Segments generated by sliding the scaled window $[-\xi_{1,}\ +\xi_2]$ along protein sequences can be generally expressed as

$$R_{-\xi_1}R_{-(\xi_1-1)} \cdots R_{-3}R_{-2}R_{-1}R_{+1}R_{+2}\cdots R_{+(\xi_2-1)}R_{+\xi_2} \qquad (1)$$

where $R_{-\xi_1}$ represents the residue at the scale $-\xi_1$, $R_{-1}$ the residue at the scale –1, $R_{+1}$ the residue at the scale +1 and so forth.

If the amino acid residue at each of the segment subsites (Equation 1) can be treated as an independent element, i.e.

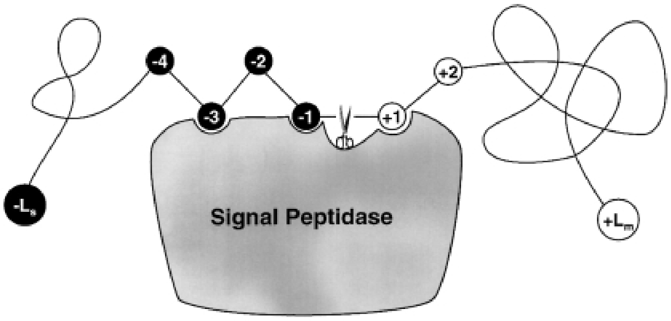

**Fig. 1.** A schematic drawing to show the signal sequence of a protein and how it is cleaved by the signal peptidase. An amino acid in the signal part is depicted as a black circle with a white number to indicate its sequential position, while that in the mature protein depicted as an open circle with a black 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. During the cleavage process, a highly special fit is required between the amino acid residues at the subsites –3, –1 and +1 of the secretory protein and their counterpart of the enzyme (cf. Figure 4).

there is no coupling at all among these subsites, then its attribute to the cleavable set $S^+$ and that to the non-cleavable set $S^-$ can be formulated, respectively, as

$$\begin{aligned}\Psi_0^+(\mathrm{R}_{-\xi_1}\cdots\mathrm{R}_{-3}\mathrm{R}_{-2}\mathrm{R}_{-1}\mathrm{R}_{+1}\mathrm{R}_{+2}\cdots\mathrm{R}_{+\xi_2)}\\ =P^+_{-\xi_1}(\mathrm{R}_{-\xi_1})\cdots P^+_{-3}(\mathrm{R}_{-3})P^+_{-2}(\mathrm{R}_{-1})\\ P^+_{+1}(\mathrm{R}_{+1})P^+_{+2}(\mathrm{R}_{+2})\cdots P^+_{+\xi_2}(\mathrm{R}_{+\xi_2})\end{aligned}\tag{2a}$$

and

$$\begin{aligned}\Psi_0^-(\mathrm{R}_{-\xi_1}\cdots\mathrm{R}_{-3}\mathrm{R}_{-2}\mathrm{R}_{-1}\mathrm{R}_{+1}\mathrm{R}_{+2}\cdots\mathrm{R}_{+\xi_2)}\\ =P^-_{-\xi_1}(\mathrm{R}_{-\xi_1})\cdots P^-_{-3}(\mathrm{R}_{-3})P^-_{-2}(\mathrm{R}_{-1})\\ P^-_{+1}(\mathrm{R}_{+1})P^-_{+2}(\mathrm{R}_{+2})\cdots P^-_{+\xi_2}(\mathrm{R}_{+\xi_2})\end{aligned}\tag{2b}$$

where $P_i{}^+(\mathrm{R}_i)$ is the probability of amino acid $\mathrm{R}_i$ occurring at the subsite $i$ ( = $-\xi_1$, ..., –3, –2, –1, +1, +2, ..., $+\xi_2$) for the secretion-cleavable segments and $P_i{}^-(\mathrm{R}_i)$ the corresponding probability for the non-secretion-cleavable segments. The values of the former can be derived from a positive training data set $S_0{}^+$ consisting of only secretion-cleavable segments and the values of the latter can be derived from a negative training data set $S_0^+$ consisting of only non-cleavable segments. The subscript 0 of $\psi$ indicates that the attribute function is formed by independent probabilities in which no coupling effect between subsites is included, as shown by the right-hand side of Equation 2. However, in reality the protein subsites are often coupled with one another. Therefore, it is instructive to conduct a statistical analysis for the 1939 secretory protein sequences retrieved from Nielsen *et al.* (1997). The result thus obtained is illustrated in Figure 4, from which we can see that the amino acid residues at the subsites –3, –1 and +1 are mostly occupied by Ala. Furthermore, according to the detailed numbers generated through the statistical analysis, of the 1939 protein sequences, the occurrence frequencies of Ala at the subsites –3, –1 and +1 are 667, 1084 and 397, respectively, while the occurrence frequencies of the other 19 amino acids at these subsites are relatively much lower. Besides, all these three subsites are very close to the cleavage site (Figure 1). This suggests that a highly special match between the signal peptidase and the secretory protein at the subsites –3, –1 and +1 is required during the cleavage process. Accordingly, to establish a powerful method for predicting the signal peptides, the coupling among these three key subsites, i.e. the {–3, –1, +1} coupling, must be taken into account. Thus, Equations 2a and 2b should be modified to

$$\begin{aligned}\Psi^+(\mathrm{R}_{-\xi_1}\cdots\mathrm{R}_{-3}\mathrm{R}_{-2}\mathrm{R}_{-1}\mathrm{R}_{+1}\mathrm{R}_{+2}\cdots\mathrm{R}_{+\xi_2})\\ =P^+_{-\xi_1}(\mathrm{R}_{-\xi_1})\cdots P^+_{-3}(\mathrm{R}_{-3})P^+_{-2}(R_{-2})P^+_{-1}(\mathrm{R}_{-1}|\mathrm{R}_{-3})\\ P^+_{+1}(\mathrm{R}_{+1}|\mathrm{R}_{-1})P^+_{+2}(\mathrm{R}_{+2})\cdots P^+_{+\xi_2}(\mathrm{R}_{+\xi_2})\end{aligned}\tag{3a}$$

and

$$\begin{aligned}\Psi^-(\mathrm{R}_{-\xi_1}\cdots\mathrm{R}_{-3}\mathrm{R}_{-2}\mathrm{R}_{-1}\mathrm{R}_{+1}\mathrm{R}_{+2}\cdots\mathrm{R}_{+\xi_2})\\ =P^-_{-\xi_1}(\mathrm{R}_{-\xi_1})\cdots P^-_{-3}(\mathrm{R}_{-3})P^-_{-2}(R_{-2})P^-_{-1}(\mathrm{R}_{-1}|\mathrm{R}_{-3})\\ P^-_{+1}(\mathrm{R}_{+1}|\mathrm{R}_{-1})P^-_{+2}(\mathrm{R}_{+2})\cdots P^-_{+\xi_2}(\mathrm{R}_{+\xi_2})\end{aligned}\tag{3b}$$

respectively, where $P_i{}^+(\mathrm{R}_i)$ and $P_i{}^-(\mathrm{R}_i)$ are the same as those in Equation 2. $P^+_{-1}(\mathrm{R}_{-1}|\mathrm{R}_{-3})$ is the probability of amino acid $\mathrm{R}_{-1}$ occurring at the subsite –1, given that $\mathrm{R}_{-3}$ has occurred at the subsite –3; $P^+_{+1}(\mathrm{R}_{+1}|\mathrm{R}_{-1})$ is the probability of amino acid $\mathrm{R}_{+1}$ occurring at the subsite +1, given that $\mathrm{R}_{-1}$ has occurred at the subsite –1. Their values can be derived from a positive training data set $S_0^+$ consisting of only secretion-cleavable peptides. Also, $P^-_{-1}(\mathrm{R}_{-1}|\mathrm{R}_{-3})$ and $P^-_{+1}(\mathrm{R}_{+1}|\mathrm{R}_{-1})$ have the same meaning as $P^+_{-1}(\mathrm{R}_{-1}|\mathrm{R}_{-3})$ and $P^+_{+1}(\mathrm{R}_{+1}|\mathrm{R}_{-1})$ except that they are derived from a negative training data set $S_0{}^-$ consisting of only non-cleavable peptides.

Thus, for a given peptide sequence as defined in Equation 1, if its attribute function to the positive training set $S_0{}^+$ is greater than that to the negative training set $S_0{}^-$, i.e. $\psi^+ > \psi^-$, then the sequence is predicted to be secretion-cleavable; otherwise, it is predicted to be non-secretion-cleavable. We define a discriminant function $\Delta$, given by

$$\begin{aligned}\Delta(\mathrm{R}_{-\xi_1}\cdots\mathrm{R}_{-3}\mathrm{R}_{-2}\mathrm{R}_{-1}\mathrm{R}_{+1}\mathrm{R}_{+2}\cdots\mathrm{R}_{+\xi_2}) =\\ w^+\Psi^+(\mathrm{R}_{-\xi_1}\cdots\mathrm{R}_{-3}\mathrm{R}_{-2}\mathrm{R}_{-1}\mathrm{R}_{+1}\mathrm{R}_{+2}\cdots\mathrm{R}_{+\xi_2})\\ -w^-\Psi^-(\mathrm{R}_{-\xi_1}\cdots\mathrm{R}_{-3}\mathrm{R}_{-2}\mathrm{R}_{-1}\mathrm{R}_{+1}\mathrm{R}_{+2}\cdots\mathrm{R}_{+\xi_2}\end{aligned}\tag{4}$$

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 predicting the seretion-cleavability for a given peptide sequence can be formulated as follows:

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

During the training process, the parameters $\xi_1$ and $\xi_2$ can be changed so as to find the optimal prediction quality. Once a secretion-cleavable peptide is predicted, the corresponding cleavage site and signal peptide are automatically obtained as described above (cf. Figures 1 and 3a).

## Results and discussion

To show the power of the key-subsites-coupled algorithm, the following two criteria should be followed: (1) using a good data set that is accessible to the public and (2) comparison with the best result reported in the literature. The data set investigated by Nielsen *et al.* (1997) satisfies the first criterion; it can be retrieved from an FTP server at ftp://virus.cbs.dtu.dk/pub/signalp. They consist of 1939 secretory proteins and 1440

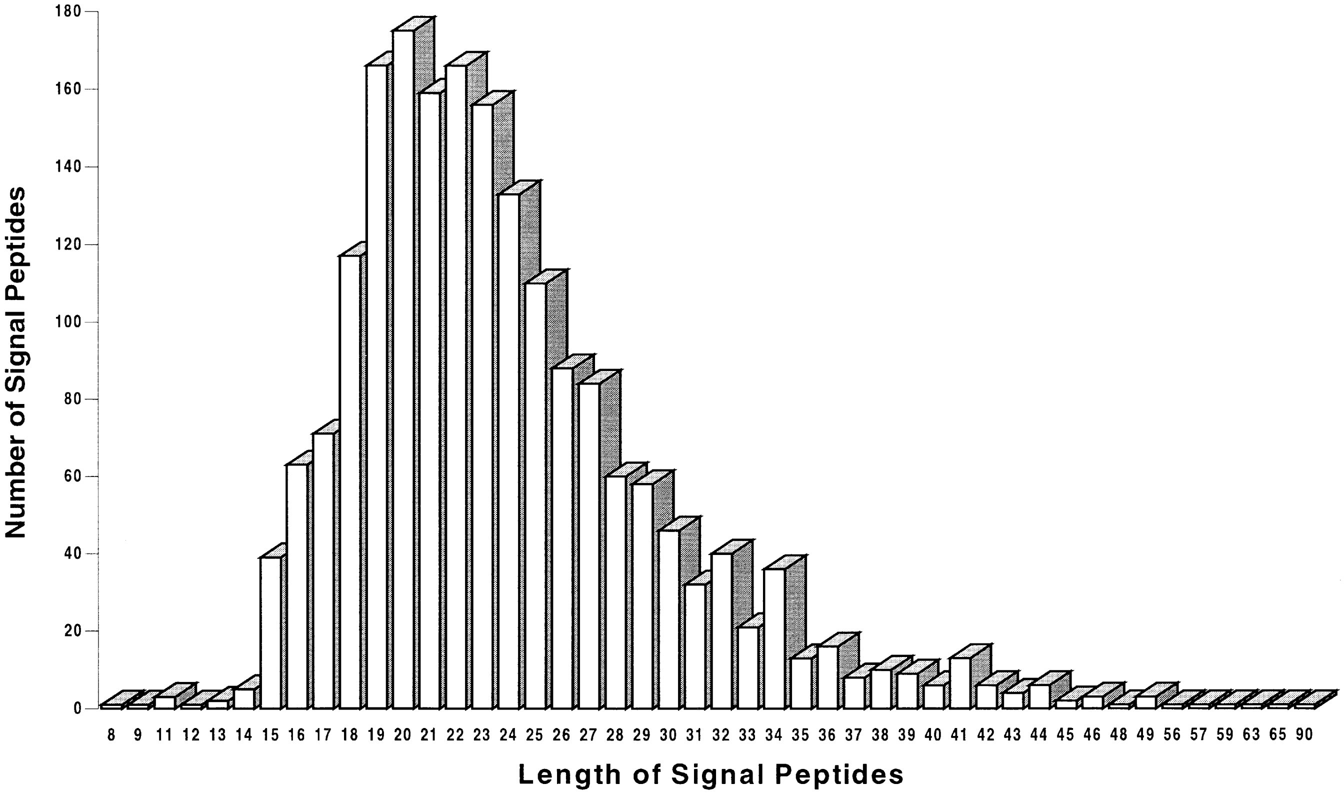

**Fig. 2.** A histogram to show the distribution of signal peptides with their length in the 1939 secretory proteins retrieved from Nielsen *et al.* (1997).

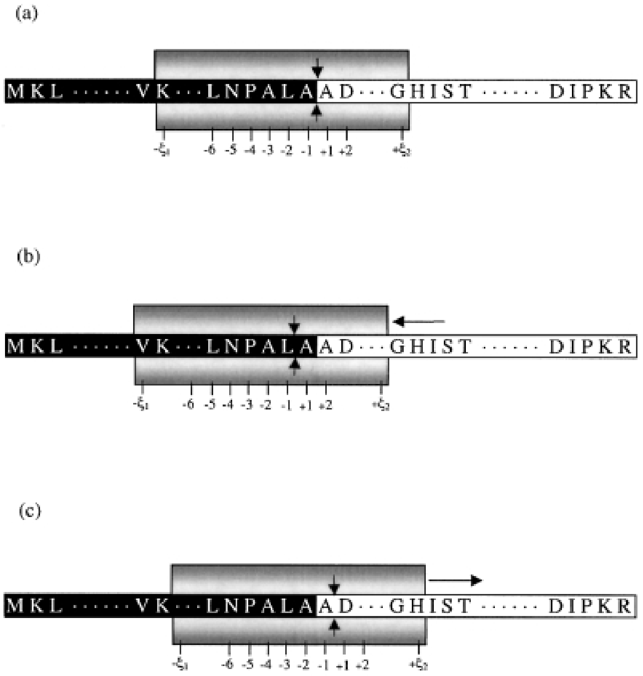
(a)
MKL ······ VK···LNPALA AD···GHIST ······ DIPKR
-ξ1 -6 -5 -4 -3 -2 -1 +1 +2 +ξ2
(b)
MKL ······ VK···LNPALA AD···GHIST ······ DIPKR
-ξ1 -6 -5 -4 -3 -2 -1 +1 +2 +ξ2
(c)
MKL ······ VK···LNPALA AD···GHIST ······ DIPKR
-ξ1 -6 -5 -4 -3 -2 -1 +1 +2 +ξ2

**Fig. 3.** Illustration to show the sequence segments highlighted by sliding the scaled window [$-\xi_1$, $+\xi_2$] along a protein sequence. During the sliding process, the scales on the window are aligned with different amino acids so as to define different peptide segments. When, and only when, the scale –1 is aligned with the last residue of the signal sequence and scale +1 aligned with the first residue of the mature protein as shown in panel (**a**) is the peptide segment seen within the window regarded as secretion-cleavable. Peptides segments seen within the window for all the other cases, such as those shown in panels (**b**) and (**c**), are regarded as non-secretion-cleavable.

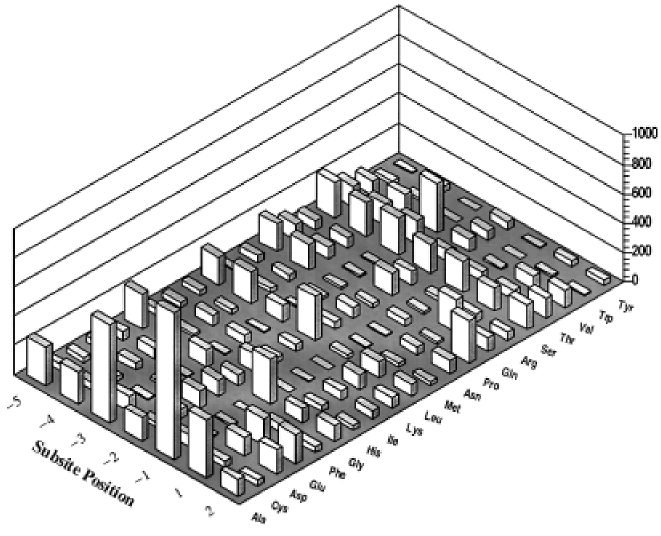

**Fig. 4.** A 3-D histogram to show the frequencies of the 20 native amino acids that occur at the subsites proximal to the cleavage site. As shown, the occurrence frequencies of Ala at the subsites –3, –1 and +1 are overwhelming in comparison with the other 19 amino acids, suggesting a high selectivity of Ala at the three key subsites (cf. Figure 1).

non-secretory proteins. The former contains 416 human, 1011 eukaryote, 105 *Escherichia coli*, 266 Gram negative and 141 Gram positive proteins, and the latter 251 human, 820 eukaryote, 119 *E.coli*, 186 Gram negative and 64 Gram positive proteins. Redundant sequences were removed to guarantee that no pairs of homologous sequences exist in the data set. As treated by Nielsen *et al.* (1997), 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 non-secretory proteins, the first 70 amino acids of each sequence were included. According to their report, the average rate of correct prediction for the cleavage site location by the neural network method was 71.54%. This is the highest success rate so far reported for such a large data set available to the public. Therefore, the result reported by Nielsen *et al.* (1997) also satisfies the second criterion. To compare the prediction quality at an equivalent condition, we used the same data set as used by Nielsen *et al.* (1997).

The rate of correct prediction for the signal peptide set and non-signal peptide set are given by

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

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

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

The prediction quality was examined by the standard testing procedure in statistics (Mardia *et al.*, 1979), that is, a combination of the self-consistency and jackknife tests. In the former, the signal peptide of each protein in a given data set was predicted 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 sub-sampling test often adopted in biology, the jackknife test is considered to be the most effective method for cross-validation in statistics (Mardia *et al.*, 1979). 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 (Chou and Zhang, 1995). As for the sub-sampling 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.* (1977), each data set was divided into five approximately equal size parts and then every network run was carried out with one part as test data and the other four parts as training data. The performance measures were then calculated as an average over the five different data set divisions. Thus, even for the data of only secretory proteins, the number of possible combinations would be $\Phi = \Phi_1 \times \Phi_2 \times \Phi_3 \times \Phi_4 \times \Phi_5$, where $\Phi_1 = 416!/(83!83!83!83!84!)$, $\Phi_2 = 1011!/(202!202!202!202!202!)$, $\Phi_3 = 105!/(21!21!21!21!21!)$, $\Phi_4 = 266!/(53!53!53!53!54!)$ and $\Phi_5 = 141!/(28!28!28!28!29!)$. Of $\Phi_1$, $\Phi_2$, $\Phi_3$, $\Phi_4$ and $\Phi_5$, the smallest is $\Phi_3 \approx 3.1 \times 10^{69}$, implying $\Phi$ would be $\gg 15.5 \times 10^{345}$. It is impossible for any existing computer to handle such a huge number of combinations. In fact in any practical sub-sampling tests as performed by Nielsen *et al.* (1997), only a very small fraction of the possible combinations were investigated and the results thus obtained could not avoid a considerable arbitrariness. Accordingly, the testing procedure adopted here is much more objective and rigorous.

**Table I.** Performance values by using the subsite coupling model

| | Rate of correct prediction for cleavage site location (%)[a] | | |
|---|---|---|---|
| | Signal peptides | Non-secretory proteins | Overall |
| Scaled window | Self-consistency test | | |
| $[-\xi_1, +\xi_2]$ | $\Lambda^+$ | $\Lambda^-$ | $\Lambda$ |
| [–6, +2] | 89.84 | 87.44 | 87.47 |
| [–8, +2] | 90.36 | 89.10 | 89.12 |
| [–10, +2] | 92.06 | 90.76 | 90.78 |
| [–12, +2] | 93.66 | 92.11 | 92.13 |
| [–13, +2] | 93.96 | 92.46 | 92.48 |
| [–14, +2] | 93.97 | 92.57 | 92.59 |
| [–15, +2] | 93.97 | 92.67 | 92.69 |
| [–16, +2] | 92.26 | 92.75 | 92.74 |
| [–18, +2] | 86.02 | 93.09 | 92.99 |
| Scaled window | Jackknife test | | |
| $[-\xi_1, +\xi_2]$ | $\Lambda^+$ | $\Lambda^-$ | $\Lambda$ |
| [–6, +2] | 85.25 | 87.69 | 87.66 |
| [–8, +2] | 86.90 | 89.15 | 89.12 |
| [–10, +2] | 87.98 | 90.74 | 90.71 |
| [–12, +2] | 89.12 | 92.10 | 92.06 |
| [–13, +2] | 89.63 | 92.46 | 92.42 |
| [–14, +2] | 89.58 | 92.57 | 92.53 |
| [–15, +2] | 89.94 | 92.66 | 92.63 |
| [–16, +2] | 88.14 | 92.74 | 92.68 |
| [–18, +2] | 81.74 | 93.08 | 92.93 |

[a] See Equations 6 and 7 for the definitions of $\Lambda^+$, $\Lambda^-$ and $\Lambda$.

Prediction was performed by selecting different parameters for the scaled window $[-\xi_1, +\xi_2]$. Preliminary tests indicated that for a given $\xi_1$ the optimal result for $\Lambda^+$ was obtained when $\xi_2 = 2$. The predicted results by both self-consistency and jackknife tests with different values of $\xi_1$ are given in Table I, from which we can see that the overall success rate $\Lambda$ is improved with increase in $\xi_1$. However, if $\xi_1$ is too large, many short signal peptides will be excluded. For example, two signal peptides were excluded when $\xi_1 = 10$, five when $\xi_1 = 12$, six when $\xi_1 = 13$, eight when $\xi_1 = 14$, 13 when $\xi_1 = 15$, 52 when $\xi_1 = 16$ and 186 when $\xi_1 = 18$. Each of these excluded signal peptides was counted as an unsuccessful prediction event, contributing to the reduction of the success rate for the prediction of signal peptides. As a consequence, $\Lambda^+$ was gradually reduced when $\xi_1 \geqslant 16$ (Table I). As a compromise, we select $\xi_1 = 13$, 14 or 15 and $\xi_2 = 2$ as the optimal parameters for the scaled window $[-\xi_1, +\xi_2]$. When $\xi_1$ and $\xi_2$ are within these values, the success rates $\Lambda^+$ (Equation 6) for the signal peptide set are over 93 and 89% by self-consistency and jackknife tests, respectively, while the corresponding success rates $\Lambda^-$ (Equation 6) for the non-signal peptide set are both over 92%. Also, the overall success rates (Equation 7) for the cleavage site location by both self-consistency and jackknife tests are over 92%.

Besides the neural network (NN) method proposed by Nielsen *et al.* (1997), there are some other methods, such as the simple weight matrix method (von Heijne, 1986), the hidden Markov method (Baldi and Brunak, 1998) and the physical sequence analysis method (Ladunga, 1999). Like Nielsen *et al.*'s method, all these methods have played an important role in stimulating the development of this area. The simple weight matrix method is one of the earliest practical approaches for predicting the signal peptide cleavage sites. However, as pointed out by Nielsen *et al.* (1997), if 'the original weight matrix algorithm (von Heijne, 1986) is applied to' the current data set, 'the performance is much lower' in comparison with their NN method. The hidden Markov method (HMM) also belongs to the machine learning approach; the term 'hidden' refers to the invisibility of the underlying random walk between different states. Actually, the HMM method is a different type of artificial neural network method and hence also bears the disadvantages elaborated by King (1996). The physical sequence analysis method, also called PHYSEAN method, was established on the basis of the physical, chemical and biological characteristics of protein sequences. The working data sets for PHYSEAN consists of 2532 preproteins with signal peptides and 1138 cytosolic proteins. As described by Ladunga (1999), three-quarters of the sequences in the data sets were randomly selected to form a training set and the predictions were performed on the remaining one-quarter of sequences. The prediction accuracy was estimated on untrained proteins by five repetitions of cross-validation experiments. The success rate thus obtained for the prediction of cleavage sites was 79.28%. It was not possible to make a direct comparison of the present algorithm with PHYSEAN based on a same data set because, unlike NN (Nielsen *et al.*, 1997), the data sets in PHYSEAN are not accessible to the public. Moreover, as we can see, the cross-validation procedure in PHYSEAN is also of sub-sampling test and hence could not avoid the problem of arbitrariness either. This can be illustrated as follows. Even only for the 2532 preproteins, the number of possible sub-sampling combinations would be $2532!/(633!1899!) \gg 10^{370}$. Compared with such a huge number, five different sub-samplings, although randomly selected, are merely a very tiny fraction of the possible combinations (i.e., the fraction of sub-samplings considered is $\ll 0.5\times10^{-369}$).

Accordingly, from both the higher success rate and the more rationality in test procedures, it is worth communicating the new algorithm to those working in the area concerned. At least it will play a complementary role to the existing algorithms, stimulating the development of protein signal peptide prediction.

## Conclusion

Since the present model has explicitly incorporated the coupling among the subsites –3, –1 and +1 and all these subsites are very close to the cleavage site, it can be directly used for investigating the protein secretion-cleaved mechanism by signal peptidase. The present model can also serve as a useful vehicle for helping further investigate many unclear details regarding the molecular mechanism of the ZIP code protein-sorting system in cells. Furthermore, since signal peptides are the key in determining the subcellular location of proteins, the {–3, –1, +1} model might have some impact in improving the prediction quality of protein subcellular location (Cedano *et al.*, 1997; Reinhardt and Hubbard, 1998; Chou and Elrod, 1998, 1999a,b; Chou, 2000; Nakai, 2000).

## Acknowledgements

Illuminative discussions with Dr A.P.Elhammer and Dr Jinhe Li are gratefully acknowledged. The author also thanks Raymond B.Moeller, Cynthia A.Ludlow and Wendy Vanderheide for their help in drawing the figures.

## References

Baldi,P. and Brunak,S. (1998) *Bioinformatics: the Machine Learning Approach*. MIT Press, Cambridge, MA.

Cedano,J., Aloy,P., Perez-Pons,J.A. and Querol,E. (1997) *J. Mol. Biol.*, **266**, 594–600.

Chou,K.C. (2000) *Curr. Protein Pept. Sci.*, **1**, 171–208.

Chou,K.C. and Elrod,D.W. (1998) *Biochem. Biophys. Res. Commun.*, **252**, 63–68.

Chou,K.C. and Elrod,D.W. (1999a) *Protein Eng.*, **12**, 107–118.

Chou,K.C. and Elrod,D.W. (1999b) *Proteins: Struct. Funct. Genet.*, **34**, 137–153.

Chou,K.C. and Zhang,C.T. (1995) *Crit. Rev. Biochem. Mol. Biol.*, **30**, 275–349.

Claros,M.G., Brunak,S. and von Heijne,G. (1997) *Curr. Opin. Struct. Biol.*, **7**, 394–398.

Gierasch,L.M. (1989) *Biochemistry*, **28**, 923–930.

Hagmann,M. (1999) *Science*, **286**, 666–666.

King,R.D. (1996) In Sternberg,M.J.E. (ed.), *Protein Structure Prediction: a Practical Approach*. IRL Press, Oxford, pp. 79–97.

Ladunga,I. (1999) *Bioinformatics*, **15**, 1028–1038.

Mardia,K.V., Kent,J.T. and Bibby,J.M. *Multivariate Analysis*. Academic Press, London, 1979, pp. 322 and 381.

Nakai,K. (2000) *Adv. Protein Chem.*, **54**, 277–344.

Nielsen,H., Engelbrecht,J., Brunak S. and von Heijne,G. (1997) *Protein Eng.*, **10**, 1–6.

Nielsen,H., Brunak S. and von Heijne,G. (1999) *Protein Eng.*, **12**, 3–9.

Rapoport,T.A. (1992) *Science*, **258**, 931–936.

Reinhardt,A. and Hubbard,T. (1998) *Nucleic Acids Res.*, **26**, 2230–2236.

von Heijne,G. (1986) *Nucleic Acids Res.*, **14**, 4683–4690.

*Received October 13, 2000; revised November 29, 2000; accepted December 8, 2000*