Amino Acids (2005) 28: 395–402
DOI 10.1007/s00726-005-0189-6

Amino Acids
Printed in Austria

Using string kernel to predict signal peptide cleavage site based on subsite coupling model

M. Wang[1,2], **J. Yang**[1], and **K.-C. Chou**[1,3,4]

[1] Institute of Image Processing and Pattern Recognition, Shanghai Jiaotong University, Shanghai, China
[2] Microsoft Research Asia, Beijing, China
[3] Institute of Bioinformatics and Drug Discovery, Tianjin, China
[4] Gordon Life Science Institute, San Diego, California, U.S.A.

Received February 16, 2005
Accepted February 18, 2005
Published online April 21, 2005;

Summary. Owing to the importance of signal peptides for studying the molecular mechanisms of genetic diseases, reprogramming cells for gene therapy, and finding new drugs for healing a specific defect, it is in great demand to develop a fast and accurate method to identify the signal peptides. Introduction of the so-called {−3,−1, +1} coupling model (Chou, K. C.: *Protein Engineering*, 2001, 14–2, 75–79) has made it possible to take into account the coupling effect among some key subsites and hence can significantly enhance the prediction quality of peptide cleavage site. Based on the subsite coupling model, a kind of string kernels for protein sequence is introduced. Integrating the biologically relevant prior knowledge, the constructed string kernels can thus be used by any kernel-based method. A Support vector machines (SVM) is thus built to predict the cleavage site of signal peptides from the protein sequences. The current approach is compared with the classical weight matrix method. At small false positive ratios, our method outperforms the classical weight matrix method, indicating the current approach may at least serve as a powerful complemental tool to other existing methods for predicting the signal peptide cleavage site.

The software that generated the results reported in this paper is available upon requirement, and will appear at http://www.pami.sjtu.edu.cn/wm.

Keywords: Signal peptide – Chou's subsite coupling approach – Probabilistic model – String kernels – Support vector machine

I Introduction

With the avalanche of protein sequences in the post-genomic era, in order to timely use their information to stimulate the development of medical science and expedite the course of drug design, many new techniques in computational biology have been developed, such as structural bioinformatics (see, e.g., Chou, 2004), protein sequence cleavage site prediction (see, e.g., Chou, 1993, 1996), enzyme active site prediction (see, e.g., Chou and Cai, 2004a), enzyme family class prediction (Chou, 2005; Chou and Cai, 2004b, 2004c), G-protein coupled receptor type prediction (Chou and Elrod, 2002), protein subcellular location prediction (see, e.g., Nakai, 2000; Chou, 2001d; Chou and Cai, 2002, 2003), and signal peptide prediction (see, e.g., Nakai 2000; Chou, 2002).

The discovery of signal peptide has made it possible for pharmaceutical scientists to produce more effective drugs by genetically modifying bacteria, plants and animals (Chou, 2002). The contemporary gene technology has allowed us to generate the gene of the desired protein with sequences coding for transport signals. Thus the knowledge of protein signals can be used to reprogram cells in a specific way for future cell and gene therapy. However, in order to effectively use the knowledge of signal peptides, one has to first identify the signal peptide and predict its cleavage site. Due to the number of nascent protein sequence entering databanks increasing at an unprecedented speed, it is time consuming and costly to identify the signal peptides solely by experiments. Thus, a strong interest in the automated identification of signal sequences and predictions of their cleavage sites has been evoked.

Although N-terminal signal peptides generally share some common features, i.e., a positively charged n-region, a central hydrophobic h-region and a neural but polar c-region (Claros et al., 1997), the extreme variations in length and sequence has posed a difficulty for formulating

a general algorithm to predict the signal peptides. Most of existing methods are mainly based on neural networks (Claros et al., 1997; Nielsen et al., 1999; Nakai, 2000). Although these techniques are always 'readily available' and 'often successful in practice', they are short of physical explanation and statistically poorly characterized (King, 1996). The subsite coupling approach originally proposed by Chou (2001b, 2002) is a completely different approach with which some very encouraging results were observed. According to the approach, although the signal peptides are of extreme variation in both the sequence order and length, some intrinsic coupling might exist among their subsites. Here, on the basis of the subsite coupling model, we would like to propose the probabilistic model for characterizing the sequence of signal peptides and use the support vector machine (SVM) (Vapnik, 1995, 1998) to predict the singular peptide and its cleavage site.

II Subsite coupling approach

For readers' convenience, let us first give a brief introduction about the subsite coupling approach. For a detailed description, the readers are suggested to refer to the original papers (Chou, 2001a, b, c). The subsite approach was developed based on the sequence-encoded algorithm (Chou, 2001b) and the scaled window approach (Chou, 2001c). The rationale is as follows: Although the signal peptides are of extreme variation in both the sequence order and length, some intrinsic coupling might exist among their subsites. For example, it has been observed that, for the 1939 secretory protein sequence (Nielsen et al., 1997), the amino acid residues at the subsites -3, -1 and $+1$ are mostly occupied by Ala (Fig. 1), while the occurrence frequencies of the other 19 amino acids at these subsites are relatively much lower. This indicates a highly special match between the signal peptides and the secretory protein at the subsites -3, -1 and $+1$ is required during the cleavage process. Based on such a finding, some special terms that reflect the coupling among these subsites have been incorporate into the prediction algorithm.

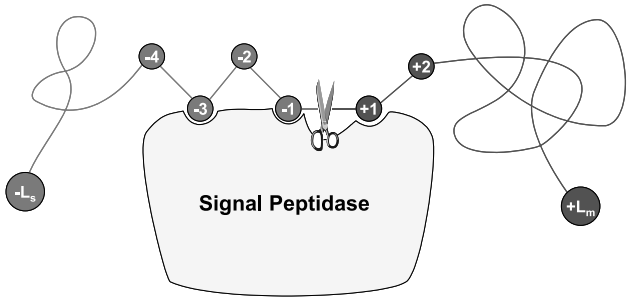

Fig. 1. A schematic drawing to show the $[-3,-1,+1]$ subsite coupling mechanism. During the cleaving process, a highly special match is required between the residues at subsites -3, -1, and $+1$ of the secretory protein and their counterparts in the signal peptides. Based on such a model, the $[-3,-1,+1]$ subsite coupling model was proposed (Chou, 2001b). Reproduced from Chou (2001b) with permission

III A probability kernels based on subsite coupling probability model

1. SVMs and probability kernels

Support vector machines (SVMs) are based on the *Structural Risk Minimization* principle from computational learning theory (Vapnik, 1995, 1998; Christianini and Shawe-Taylor, 2000). The most remarkable characteristics of SVMs are the absence of local minima, the sparseness of the solution, and the use of the kernel-induced feature spaces. The basic idea of applying SVMs to the pattern classification can be outlined as follows. First, map the input vectors into a feature space (possible with a higher dimension), either linearly or non-linearly, which is relevant to the selection of the kernel function. Then, within the feature space, seek an optimized linear division; i.e., construct a hyper-plane which can separate two classes (this can be extended to multi-classes) with the least errors and maximal margin. The SVMs training process always seeks a global optimized solution and avoids over-fitting, so it has the ability to deal with a large number of features. SVMs have been used to deal with protein fold recognition (Ding and Dubchak, 2001), protein–protein interactions prediction (Bock and Gough, 2001), protein subcellular location prediction (Chou and Cai, 2002), and membrane protein type prediction (Cai et al., 2003).

Given a set of N samples, i.e., a series of input training samples

$$\mathbf{x}_k \in \mathbf{X}(k = 1, \ldots, N), \tag{1}$$

where $\mathbf{x}_k$ can be regarded as the k th training example, and $\mathbf{X}$ is called the input space. Since the multi-class identification problem can always be converted into a two-class identification problem, without loss of the generality the formulation below is given for the two-class case only. Suppose the output derived from the learning machine is expressed by $y_k \in \{+1, -1\}(k = 1, \ldots, N)$ where the indexes -1 and $+1$ are used to stand for the two classes concerned, respectively. The goal here is to construct one binary classifier or derive one decision function from the available samples that has a small probability of misclas-

sifying a future sample. The resulting classifier is thus based on the decision function:

$$f(\mathbf{x}) = \sum_{i=1}^{N} \lambda_i K(\mathbf{x}_i, \mathbf{x}) \tag{2}$$

where $\mathbf{x}$ is any new object to be classified, $K(\mathbf{x}_i, \mathbf{x})$ is a so-called kernel function and the coefficients $\{\lambda_1, \lambda_2, \ldots, \lambda_N\}$ are determined during the process of training. The kernel K can be considered as a dot product between the image of the objects after a mapping to a high-dimensional feature space, i.e.,

$$K(\mathbf{x}, \mathbf{z}) = <\phi(\mathbf{x}), \phi(\mathbf{z})> \tag{3}$$

where $\phi(\mathbf{x})$ is the mapping of the original object in the feature space. As a result, the kernel defines a measure of the 'similarity' between any two objects in the feature space.

In the context of predicting signal peptide cleavage site, the objects for our analysis are the protein sequences generated by sliding the scaled window $[-\xi_1, +\xi_2]$ (Chou, 2001). The first step is to develop a probabilistic model $p(\mathbf{x})$ on these sequences, and then define a mapping $\phi(\mathbf{x})$ from the original sequence to the feature space, finally the kernel $K(\mathbf{x}, \mathbf{y})$ is obtained according to Eq. 3 and the efficient kernel-based method (such as SVMs) can thus be employed. As mentioned above, an efficient probability model has been introduced that took into account the coupling among three key subsites (Chou, 2001b). The following critical problem is how to define the feature map $\phi(\mathbf{x})$ based on the subsite coupling model. Before formally giving the feature map and the kernel, it is necessary to introduce the concept of probability kernels at first. The probability kernel is defined on the product space $\mathbf{X}\times\mathbf{X}$, and satisfies:

$$\begin{cases} \forall(\mathbf{x}, \mathbf{y}) \in \mathbf{X}\times\mathbf{X},\ 0 \le K(\mathbf{x}, \mathbf{y}) \le 1 \\ \sum_{(\mathbf{x},\mathbf{y}) \in \mathbf{X}\times\mathbf{X}} K(\mathbf{x}, \mathbf{y}) = 1 \end{cases} \tag{4}$$

Two typical probability kernels are the so-called product kernel and diagonal kernel. The product kernel is defined as

$$K_{prod}(\mathbf{x}, \mathbf{y}) = p(\mathbf{x})p(\mathbf{y}) \tag{5}$$

Its resulting classifier can be easily obtained from Eq. 2 as

$$f(x) = a \cdot p(\mathbf{x}) + b \tag{6}$$

with $a = \sum_{i=1}^{N} \lambda_i p(\mathbf{x}_i)$. In fact, the feature space defined by the feature mapping $\phi(\mathbf{x}) = p(\mathbf{x})$ is a one-dimensional line, each sample point $\mathbf{x}_i$ is represented as a single point $p(\mathbf{x}_i)$ along the line. The resulting classifier (Eq. 6) classifies a new point $\mathbf{x}$ by judging if $p(\mathbf{x})$ is above or below the threshold $-b/a$. Under such a mapping, two objects are close if they have close probabilities.

The diagonal kernel is defined as

$$K_{diag}(\mathbf{x}, \mathbf{y}) = p(\mathbf{x})\delta(\mathbf{x}, \mathbf{y}) \tag{7}$$

where $\delta(\mathbf{x}, \mathbf{y}) = 1$ if $\mathbf{x} = \mathbf{y}$, $\delta(\mathbf{x}, \mathbf{y}) = 0$ if $\mathbf{x} \neq \mathbf{y}$. The images of the training set form an orthogonal basis of the feature space. The resulting classifier assigns the new object to the most probable class simply by checking if this object has appeared in the training set. Two objects are close if they are the same, and thus no learning is performed using the diagonal kernel.

Obviously, these two kernels are extremes of the kernels that can be defined on the probability model $p(\mathbf{x})$. To construct a kernel from a probability model, the product kernel is an idea start point since its resulting classifiers classify new sequence to a class according to the probability of that sequence. In order to improve the ability of SVMs to discriminate between two protein sequences, we introduce a generalized probability kernel to reflect some notion of 'closeness' by interpolating between these two typical kernels.

2. A probability kernel based on the subsite coupling model

Some notations are introduced to formularize this idea. Let Σ be a finite alphabet. A string is a finite sequence of characters from Σ, including empty sequence. Let S be a finite set (usually $\{0, 1, \ldots, N\}$ for sequences), and $X = (X_s)_{s \in S}$ a family of random variables defined on a probability space $(\Omega, \mathbb{F}, \mathrm{P})$ and indexed by the elements of S with values in Σ^S. For any subset $T \subset S$ we note $X_T = (X_s)_{s \in T}$. For any subset $T \subset S$ and realization $x_T \in \Sigma^T$ we note $p_T(x_T) = P(X_T = x_T)$. If there is no ambiguity the $p(x_T)$ used instead of $p_T(x_T)$ for simplicity. We define similarities $p(x_T, y_U) = P(X_T = x_t, X_U = y_U)$ and $p(x_T|y_U) = P(X_T = x_T|X_u = y_u)$ for any two subsets $T \subset S$ and $U \subset S$ and realizations $x_T \in \Sigma^T$ and $y_U \in \Sigma^U$. Finally let $P(S)$ be the power set of S, i.e., the set of subsets of S, and $\upsilon \subset P(S)$ be a particular set of subsets. With these notations, a probability kernel can thus be defined as follows.

Definition 1. For any probability density p on χ and any set of subsets $\upsilon \subset P(S)$ we define the (p, υ)-common subset kernel $K_{p,\upsilon}$ by the formula:

$$K_{p,\upsilon} = \frac{p(x)p(y)}{|\upsilon|} \sum_{T \in \upsilon} \frac{\delta(x_T, y_T)}{p(x_T)} \tag{8}$$

for any two realization $(x, y) \in \Sigma^{2S}$, where $\delta(x_T, y_T)$ is 1 if $x_T = y_T$, and 0 if $x_T \neq y_T$. Here $|\upsilon|$ denotes the cardinality of the set υ.

According to definition 1, it is easy to check when υ only contains the full set S, $K_{p,\{S\}} = K_{diag}$, and when υ only contains the empty set $\emptyset$, $K_{p,\{\emptyset\}} = K_{prod}$, showing that the kernel $K_{p,\upsilon}$ interpolates between the diagonal kernel and the product kernel. Eq. 8 introduced the correlations between sequences through their common substrings indexed by υ and the contribution of a given common substring is inversely proportional to its probability. Thus, if two sequences sharing the rarer common substring, their similarity should increase correspondingly according to the definition of the proposed kernel.

There isn't a universal efficient way to compute the kernel $K_{p,\upsilon}$ and the computation becomes prohibitive as the set υ becomes large. To deal with this problem, the probability kernel derived from the subsite coupling probability model (Chou, 2001a) is factorized and computed in linear time with respect to $|S|$ as given in proposition 1. The proof is given in the appendix.

Proposition 1. Let $\{p_i, i \in S\}$ be the family of probability densities on Σ and let p be the subsite coupling distribution on Σ^S (Chou, 2001), i.e.,

$$\forall \mathbf{x} \in \Sigma^S,\ p(\mathbf{x}) = p_{-\xi_1}(x_{-\xi_1}) \ldots p_{-3}(x_{-3}) p_{-2}(x_{-2}) \\ p_{-1}(x_{-1}|x_{-3}) p_{+1}(x_{+1}|x_{-1}) p_{+2}(x_{+2}) \ldots p_{+\xi_2}(x_{+\xi_2}) \quad (9)$$

Then the kernel derived from p when $\upsilon = P(S)$ is the set of all subsets of S can be computed in linear time with respect to $|S|$ by:

$$K_{p,\upsilon}(\mathbf{x}, \mathbf{y}) = \frac{1}{2^{|S|}} \prod_{i \in S} \phi_i(x_i, y_i) \quad (10)$$

with:

$$\phi_i(x_i, y_i) = \begin{cases} p_i(x_i) + p_i(x_i)^2 \text{ if } x_i = y_i \text{ and i} \neq -1, +1 \\ p_i(x_i) p_i(y_i) \text{ if } x_i \neq y_i \text{ and i} \neq -1, +1 \\ p_{-1}(x_{-1}|x_{-3}) + p_{-1}(x_{-1}|x_{-3})^2 \text{ if } x_i = y_i \text{ and i} = -1 \\ p_{-1}(x_{-1}|x_{-3}) p_{-1}(y_{-1}|y_{-3}) \text{ if } x_i \neq y_i \text{ and i} = -1 \\ p_1(x_1|x_{-1}) + p_1(x_1|x_{-1})^2 \text{ if } x_i = y_i \text{ and i} = 1 \\ p_1(x_1|x_{-1}) p_1(x_1|x_{-1}) \text{ if } x_i \neq y_i \text{ and i} = 1 \end{cases} \quad (11)$$

IV Results and discussion

To demonstrate the power of our algorithm, both a good dataset that is available to the public and the comparison with the benchmark algorithm are definitely necessary. The dataset constructed by Nielsen et al. is an appropriate candidate for it is publicly retrievable from an FTP server at ftp://virus.cbs.dtu.dk/pub/signalp. The dataset contains 1418 non-redundant secretary protein sequences. The secretary proteins contain 1011 eukaryote, 266 Gram negative and 141 Gram positive proteins. For the secretory proteins, both the sequence of the signal peptide and the first 30 amino acids of the mature protein were included in the data set. To compare the prediction quality at an equivalent condition, we used the same data set as used by Nielsen et al. (1997). The simple weight matrix method (Heijne, 1986) is regarded as an efficient way to recognize cleavage sites. The weight matrix method has incorporated the relevant biology prior knowledge, e.g., the residues at positions -3 and -1 relative to the cleavage site are usually small and neutral. The weight matrix method is essentially computing the probability of a sequence under an independent model. In this paper, weight matrix method is adopted as a benchmark algorithm.

Prediction was performed with the scaled window $[-8, +2]$, resulting in 1418 positive windows, and 65,216 negative windows. This dataset was randomly split into a training set (80% of the windows) and a test set (20%). It should be pointed out that the sequences that include 'X' symbols are excluded to facilitate our analysis. Then both the weight matrix classifier and our proposed classifiers were constructed:

- The weight matrix: in which $w_i(x_i) = \log p_i^+ - \log p_i^{total}(x_i)$, where $p_i^+(x_i)$ is the probability that amino acid x_i occurs at position i estimated from the positive training set, and p_i^{total} is the probability that amino acid x_i occurs at position i estimated from the total training set.
- The SVM classifier based on the subsite coupling probability model (see Eq. 9) and the classifier is trained on the training set.

Thus two competitive classifiers are obtained to predict signal peptide cleavage site: the score function of the weight matrix method:

$$s(x) = \sum_{i=-8}^{2} w_i(x_i), \quad (12)$$

and the SVM's classification function:

$$f(x) = \sum_{x^{(j)} \in \text{training set}} \lambda^{(j)} K_{p^+,\upsilon}(\mathbf{x}^{(j)}, \mathbf{x}), \quad (13)$$

where $\lambda^{(j)}$ is determined during the training process. For a given threshold, each of the function assigns a new data as

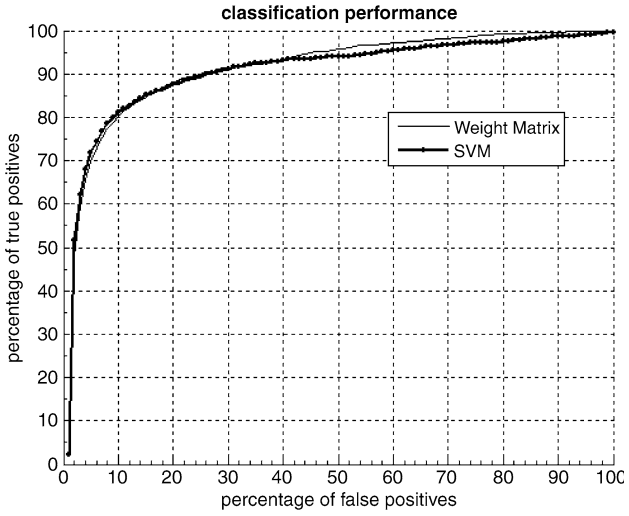

Fig. 2. Classification performance of the weight matrix method and the SVM method with window $[-8, +2]$

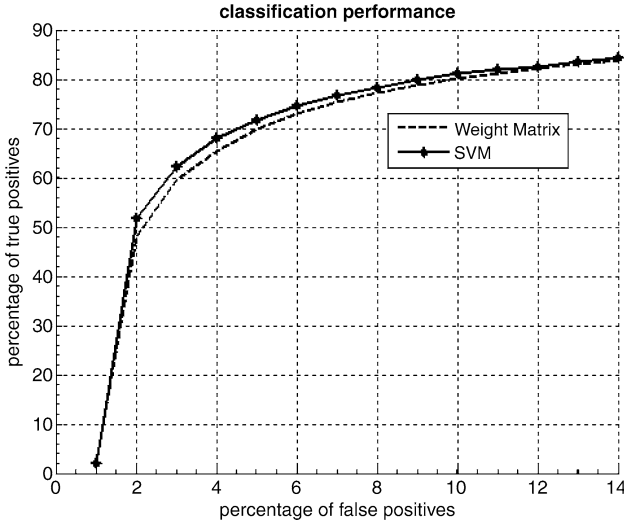

Fig. 3. Classification performance of the weight matrix method and the SVM method for small false positive rates with window $[-8, +2]$

positive or negative example by comparing if the function value is above or below the threshold. Varying the threshold and classifying the data in the test set, a curve of true positive versus false positives for each function (averaged over 10 fold cross-validation tests) are drawn in Figs. 2 and 3. It is easy to see in the interesting area, i.e., from 2% of false positive to 14% of false positive where both the false positive is not too high and the true positive is acceptable, our method outperforms the weight matrix method. For example, when we hope to have a 2% of false positive, the weight matrix method would retrieve on average 48.1% of true positive and the proposed kernel method would retrieve 51.7% of true positive. In other word, an increase of 7.4% in terms of true positive retrieval is obtained, which indicates the proposed method has more discriminant power compared with the weight matrix method. However, at the area of below 2% of false positive, these two methods have similar performance; when above 14% of false positive, the performance of weight matrix method is better. The results obtained by varying the parameter ξ_1 are also drawn in Fig. 4. It can be observed that until the windows size grows up to $[-12, +2]$, the trends of these two curves remain similar, indicating the performance of our method is robust while changing the window's size. When the windows size becomes even larger, the curves of these two methods are almost the same, which may be partially explained by the fact: the effects of the local property of subsite coupling 'decay' with the expansion of the windows. When the scaled window is small, the common substrings for a couple sequences usually occur at the right end of each protein sequence according to the subsite coupling model. However, with the increase the windows size, the chances of the common substrings appearing in other positions will grow accordingly, which in fact 'damps' the effects of the subsite coupling model. In other words, when the small windows are taken, the performance of the proposed method is better than that of the weight matrix method in the low false positive region. When the window size grows larger than $[-12, +2]$, the performance of the proposed method degenerates to the classical weight matrix method.

SVM has been widely used in current bioinformatics (Ding and Dubchak, 2001; Bock and Gough, 2001; Chou and Cai, 2002; Cai et al., 2003) as a novel classifier. Another important issue in applying SVM is to construct the specific kernel that can incorporate the prior knowledge for given research objects (Vapnik, 1995; 1998). It should be pointed out that the goal of this study is not to determine the possible upper limit of the success rate for predicting signal peptide cleavage site, but to propose a novel and different approach to incorporate the various biological prior knowledge, which is often in the form of different probability model, into the construction of the kernels used in the more powerful SVM classifier and other kernel machines. Although our experiments have demonstrated its feasibility and shown that it has its own advantage under curtain conditions, much work is left to be done along such a line. For example, it is still a big challenge how to contrive an efficient way to compute the kernel that can incorporate more complex probability models.

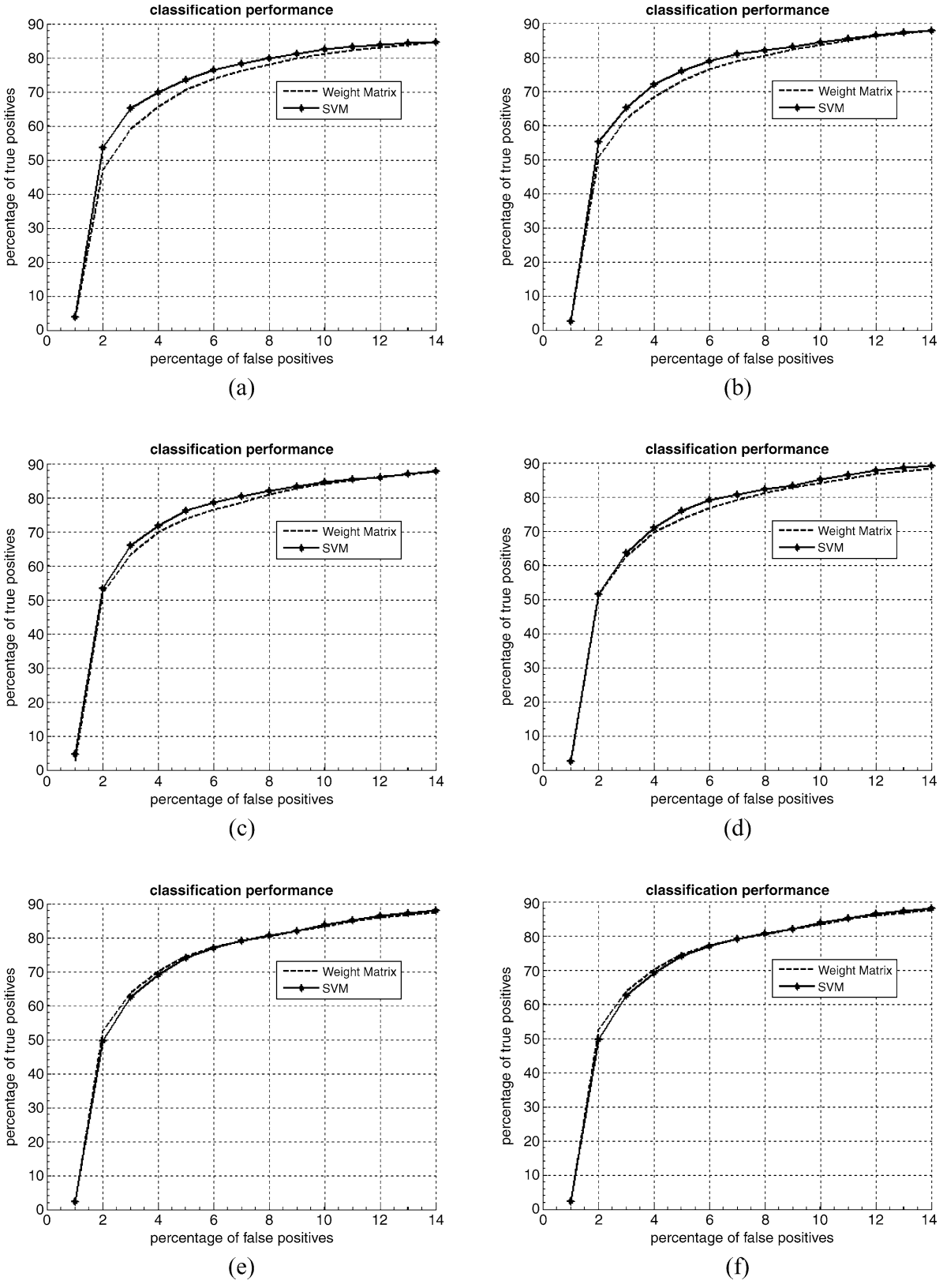

Fig. 4. Classification performance of the weight matrix method and the SVM method for small false positive rates with different windows: (a) $[-9, +2]$, (b) $[-10, +2]$, (c) $[-11, +2]$, (d) $[-12, 2]$, (e) $[-13, +2]$, and (f) $[-14, +2]$

V Conclusion

Based on the subsite coupling probability model (Chou, 2001b), a probability kernel is developed for predicting the signal peptide cleavage site. Using this probability kernel, the statistical characteristics of the protein sequence can be better grasped and represented in the mapped feature space in comparison with the simple score function of the weight matrix method. The computed results show the method proposed in this paper outperforms the classical weight matrix method, demonstrating that the current method can play a complementary role to the existing methods in predicting signal peptides in proteins.

Appendix A. Proof of Proposition 1

For a general density p and a general set υ, it is impossible to find a universal way that avoids summing $|\upsilon|$ times to compute the kernel $K_{p,\upsilon}$. This computation soon becomes prohibitive as the $|\upsilon|$ grows exponentially with respect to the windows' size. For example, if the set $\upsilon = P(S)$ then the cardinality of υ will be $2^{|S|}$. Therefore, for the subsite coupling model (Chou, 2001), an efficient computation method is definitely necessary.

The subsite coupling probability model is given as:

$$\begin{aligned}&\forall \mathbf{x} \in \Sigma^S,\ p(\mathbf{x})\\ &\quad = p_{-\xi_1}(x_{-\xi_1}) \ldots p_{-3}(x_{-3}) p_{-2}(x_{-2}) p_{-1}(x_{-1}|x_{-3}).\\ &p_{+1}(x_{+1}|x_{-1}) p_{+2}(x_{+2}) \ldots p_{+\xi_2}(x_{+\xi_2}) \end{aligned} \tag{A1}$$

For the convenience of analysis and computation, this model can be treated as a product density model $p(\mathbf{x}) = \prod_{i \in S} p_i(x_i)$, where

$$p_i(x_i) = \begin{cases} p_i(x_i) & \mathrm{i} \neq -1, +1 \\ p_{-1}(x_{-1}|x_{-3}) & \mathrm{i} = -1, \\ P_{+1}(x_{+1}|x_{-1}) & \mathrm{i} = +1 \end{cases} \tag{A2}$$

if the probability density terms at position -1, $+1$ is treated equally as those at other positions. For this product density $p(\mathbf{x}) = \prod_{i \in S} p_i(x_i)$, the following holds for any subset $T \subset S$:

$$\forall x_T \in \Sigma^T,\ p(x_T) = \prod_{i \in T} p_i(x_i) \tag{A3}$$

Therefore, we can compute for any $(x, y) \in \Sigma^{2S}$:

$$\frac{p(x)p(y)\delta(x_T, y_T)}{p(x_T)} = \prod_{i \in T} p(x_i)\delta(x_i, y_i) \times \prod_{i \notin T} p(x_i)p(y_i) \tag{A4}$$

Using Eq. 8 and the fact that $|\upsilon| = 2^{|S|}$ we can therefore compute:

$$\begin{aligned} K(x, y) &= \frac{1}{2^{|S|}} \sum_{T \subset S} \frac{p(x)p(y)\delta(x_T, y_T)}{p(x_T)} \\ &= \frac{1}{2^{|S|}} \sum_{T \subset S} \left\{ \prod_{i \in S} p(x_i)\delta(x_i, y_i) \times \prod_{i \notin S} p(x_i)p(y_i) \right\}. \\ &= \frac{1}{2^{|S|}} \prod_{i \in S} \{p(x_i)\delta(x_i, y_i) + p(x_i)p(y_i)\} \end{aligned} \tag{A5}$$

Finally, when $\upsilon = P(S)$ is the set of all subsets of S, the kernel derived from the subsite coupling model can be computed in linear time with respect to $|S|$ by:

$$K_{p,\upsilon}(x, y) = \frac{1}{2^{|S|}} \prod_{i \in S} \phi_i(x_i, y_i) \tag{A6}$$

with:

$$\begin{aligned} &\phi_i(x_i, y_i) \\ &= \begin{cases} p_i(x_i) + p_i(x_i)^2 \text{ if } x_i = y_i \text{ and i} \neq -1, +1 \\ p_i(x_i)p_i(y_i) \text{ if } x_i \neq y_i \text{ and i} \neq -1, +1 \\ p_{-1}(x_{-1}|x_{-3}) + p_{-1}(x_{-1}|x_{-3})^2 \text{ if } x_i = y_i \text{ and i} = -1 \\ p_{-1}(x_{-1}|x_{-3})p_{-1}(y_{-1}|y_{-3}) \text{if } x_i \neq y_i \text{ and i} = -1 \\ p_1(x_1|x_{-1}) + p_1(x_1|x_{-1})^2 \text{ if } x_i = y_i \text{ and i} = 1 \\ p_1(x_1|x_{-1})p_1(x_1|x_{-1}) \text{if } x_i \neq y_i \text{ and i} = 1 \end{cases} \end{aligned} \tag{A7}$$

References

Bock JR, Gough DA (2001) Predicting protein–protein interactions from primary structure. Bioinformatics 17: 455–460

Cai YD, Zhou GP, Chou KC (2003) Support vector machines for predicting membrane protein types by using functional domain composition. Biophys J 84: 3257–3263

Chou KC (1993) A vectorized sequence-coupling model for predicting HIV protease cleavage sites in proteins. J Biol Chem 268: 16938–16948

Chou KC (1996) Review: Prediction of HIV protease cleavage sites in proteins. Anal Biochem 233: 1–14

Chou KC (2001a) Prediction of protein signal sequences and their cleavage sites. Proteins: Structure, Function, and Genetics 42: 136–139

Chou KC (2001b) Using subsite coupling to predict signal peptides. Protein Eng 14/2: 75–79

Chou KC (2001c) Prediction of signal peptides using scaled window. Peptides 22: 1973–1979

Chou KC (2001d) Prediction of protein cellular attributes using pseudo-amino-acid-composition. Proteins: Structure, Function, and Genetics (Erratum: ibid., 2001, Vol. 44, 60) 43: 246–255

Chou KC (2002) Review: Prediction of protein signal sequences. Curr Protein Pept Sci 3: 615–622

Chou KC (2004) Review: Structural bioinformatics and its impact to biomedical science. Curr Med Chem 11: 2105–2134

Chou KC (2005) Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics 21: 10–19

Chou KC, Cai YD (2002) Using functional domain composition and support vector machines for prediction of protein subcellular location. J Biol Chem 277: 45765–45769

Chou KC, Cai YD (2003) A new hybrid approach to predict subcellular localization of proteins by incorporating Gene ontology. Biochem Biophys Res Commun 311: 743–747

Chou KC, Cai YD (2004a) A novel approach to predict active sites of enzyme molecules. Proteins: Structure, Function, and Genetics 55: 77–82

Chou KC, Cai YD (2004b) Predicting enzyme family class in a hybridization space. Protein Sci 13: 2857–2863

Chou KC, Cai YD (2004c) Using GO-PseAA predictor to predict enzyme sub-class. Biochem Biophys Res Commun 325: 506–509

Chou KC, Elrod DW (2002) Bioinformatical analysis of G-protein-coupled receptors. J Proteome Res 1: 429–433

Christianini N, Shawe-Taylor J (2000) An introduction to support vector machines and other kernel-based methods. Cambridge University Press, Cambridge

Claros MG, Brunak S, von Heijne G (1997) Prediction of N-terminal protein sorting signals. Curr Opin Struct Biol 7: 394–398

Ding CH, Dubchak I (2001) Multi-class protein fold recognition using support vector machines and neural networks. Bioinformatics 17: 349–358

Durbin R, Sean R, Eddy RS, Krogh A, Mitchison G (1998) Probabilistic models of proteins and nucleic acids. Cambridge University Press, Cambridge

von Heijne G (1986) A new method for predicting signal sequence cleavage sites. Nucleic Acids Res 14: 4683–4690

Hua SJ, Sun ZR (2001) A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach. J Mol Biol 308: 397–407

King RD (1996) In: Sternberg MJE (ed) Protein structure predictions: a practical approach IRL Press, Oxford, pp 79–97

Nakai K (2000) Protein sorting signals and prediction of subcellular localization. Adv Protein Chem 54: 277–344

Nielsen H, Engelbrecht J, Brunak S, von Heijne G (1997) Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Eng 10: 1–6

Nielsen H, Brunak S, von Heijne G (1999) Machine learning approaches for the prediction of signal peptides and other protein sorting signals. Protein Eng 12: 3–9

Vapnik V (1995) The nature of statistical learning theory. Springer, New York

Vapnik V (1998) Statistical learning theory. Wiley & Sons, New York

Authors' address: Prof. Kuo-Chen Chou, Gordon Life Science Institute, 13784 Torrey Del Mar Drive, San Diego, CA 92130, U.S.A.,
E-mail: kchou@san.rr.com