Available online at www.sciencedirect.com
ScienceDirect
Biochemical and Biophysical Research Communications 363 (2007) 297–303
BBRC
www.elsevier.com/locate/ybbrc

Signal-3L: A 3-layer approach for predicting signal peptides

Hong-Bin Shen [a,b,*], Kuo-Chen Chou [a]

[a] *Gordon Life Science Institute, San Diego, CA 92130, USA*
[b] *Institute of Image Processing & Pattern Recognition, Shanghai Jiaotong University, Shanghai 200030, China*

Received 19 July 2007
Available online 31 August 2007

Abstract

Functioning as an ''address tag'' that directs nascent proteins to their proper cellular and extracellular locations, signal peptides have become a crucial tool in finding new drugs or reprogramming cells for gene therapy. To effectively and timely use such a tool, however, the first important thing is to develop an automated method for rapidly and accurately identifying the signal peptide for a given nascent protein. With the avalanche of new protein sequences generated in the post-genomic era, the challenge has become even more urgent and critical. In this paper, we have developed a novel method for predicting signal peptide sequences and their cleavage sites in human, plant, animal, eukaryotic, Gram-positive, and Gram-negative protein sequences, respectively. The new predictor is called Signal-3L that consists of three prediction engines working, respectively, for the following three progressively deepening layers: (1) identifying a query protein as secretory or non-secretory by an ensemble classifier formed by fusing many individual OET-KNN (optimized evidence-theoretic *K* nearest neighbor) classifiers operated in various dimensions of PseAA (pseudo amino acid) composition spaces; (2) selecting a set of candidates for the possible signal peptide cleavage sites of a query secretory protein by a subsite-coupled discrimination algorithm; (3) determining the final cleavage site by fusing the global sequence alignment outcome for each of the aforementioned candidates through a voting system. Signal-3L is featured by high success prediction rates with short computational time, and hence is particularly useful for the analysis of large-scale datasets. Signal-3L is freely available as a web-server at http://chou.med.harvard.edu/bioinf/Signal-3L/ or http://202.120.37.186/bioinf/Signal-3L, where, to further support the demand of the related areas, the signal peptides identified by Signal-3L for all the protein entries in Swiss-Prot databank that do not have signal peptide annotations or are annotated with uncertain terms but are classified by Signal-3L as secretory proteins are provided in a downloadable file. The large-scale file is prepared with Microsoft Excel and named ''Tab-Signal-3L.xls'', and will be updated once a year to include new protein entries and reflect the continuous development of Signal-3L.
Published by Elsevier Inc.

Keywords: 3-Layer predictor; Global alignment; $\{-3,-1,+1\}$ coupling; Fusion; Pseudo amino acid composition; PseAA server

A signal peptide is a short sequence chain, which controls the entry of virtually all proteins to the secretory pathway, both in eukaryotes and prokaryotes [1]. Knowledge of signal peptides is very useful for developing novel strategies for drug discovery, as well as for revealing the molecular mechanisms of some genetic diseases (see, e.g., a review [2]). Actually, signal peptides have become a crucial tool for pharmaceutical scientists who genetically modify bacteria, plants, and animals to produce effective drugs [3]. For instance, by adding a specific tag to the desired proteins, one can tag them for excretion, making them much easier to harvest. In view of this, it is highly desired to develop a high throughput tool for fast and accurately identifying the signal peptides of nascent proteins. With the explosion of new protein sequences emerging in the post-genomic era, such a challenge has become even more critical and urgent. Actually, during the last two decades many efforts have been made in this regard [4–21].

In a recent paper [22], we have reported a 2-layer approach called Signal-CF for predicting signal peptides in eukaryotic proteins as well as Gram-positive and

* Corresponding author. Present address: BCMP, Harvard Medical School, Boston, MA 02115, USA.
E-mail addresses: hbshen@crystal.harvard.edu (H.-B. Shen), kcchou@gordonlifescience.org (K.-C. Chou).

0006-291X/$ - see front matter Published by Elsevier Inc.
doi:10.1016/j.bbrc.2007.08.140

Gram-negative proteins. Signal-CF was developed by incorporating the subsite coupling effects along a protein sequence and by fusing the results derived from many width-different scaled windows through a voting system. To further increase the prediction power and the coverage scope, here we are to introduce a 3-layer predictor, called Signal-3L.

Materials

A statistical predictor usually consists of two constituents: one is a prediction engine operating according to some rules to deal with the input and output; and the other a dataset with known information to train the prediction engine.

To obtain a high-quality benchmark dataset, protein sequences were collected from the most recent version of Swiss-Prot database (version 50.7, released on September 19, 2006) by strictly following the following steps. (1) Selected for the secretory protein dataset were those marked with "signal" in the FT (Feature Table) line; while selected for the non-secretory protein dataset were those marked with "cytoplasm" and "nucleus" in the CC (Comment) line. (2) Remove those entries obtained from the above step that were annotated with uncertain terms such as "by similarity", "probable", or "potential". (3) Remove those entries that were either annotated with "fragment" or containing less than 50 amino acids. (4) If several protein entries had the same first 100 resides along their sequences, only one of them was kept to avoid redundancy. Compared with the benchmark dataset constructed for PrediSi [21] that allowed some identical sequences within the first 100 residues (e.g., between P01551 and Q06110; P05618 and P09121; P06654 and P19909; and P09850 and P18429) and the dataset constructed for SignalP [20] that allowed many identical sequences within the first 70 residues as pointed out by Liu et al. [23], the dataset constructed according to the current criteria is much more rigorous and stringent.

After the above four steps, we finally obtain the following six datasets

$$\begin{cases} \mathbb{S}_{\text{Human}} = \mathbb{S}^{+}_{\text{Human}} \cup \mathbb{S}^{-}_{\text{Human}} \\ \mathbb{S}_{\text{Plant}} = \mathbb{S}^{+}_{\text{Plant}} \cup \mathbb{S}^{-}_{\text{Plant}} \\ \mathbb{S}_{\text{Animal}} = \mathbb{S}^{+}_{\text{Animal}} \cup \mathbb{S}^{-}_{\text{Animal}} \\ \mathbb{S}_{\text{Euk}} = \mathbb{S}^{+}_{\text{Euk}} \cup \mathbb{S}^{-}_{\text{Euk}} \\ \mathbb{S}_{\text{Gpos}} = \mathbb{S}^{+}_{\text{Gpos}} \cup \mathbb{S}^{-}_{\text{Gpos}} \\ \mathbb{S}_{\text{Gneg}} = \mathbb{S}^{+}_{\text{Gneg}} \cup \mathbb{S}^{-}_{\text{Gneg}} \end{cases} \quad (1)$$

where $\mathbb{S}_{\text{Human}}$ represents the dataset for human proteins, $\mathbb{S}^{+}_{\text{Human}}$ the subset for the human secretory proteins, $\mathbb{S}^{-}_{\text{Human}}$ for the human non-secretory proteins; $\mathbb{S}_{\text{Euk}}$ for eukaryotic proteins except those already occurring in the datasets for human ($\mathbb{S}_{\text{Human}}$), plant ($\mathbb{S}_{\text{Plant}}$), and animal ($\mathbb{S}_{\text{Animal}}$); $\mathbb{S}_{\text{Gpos}}$ and $\mathbb{S}_{\text{Gneg}}$ the datasets for Gram-positive and Gram-negative, respectively; the symbol $\cup$ represents the union in the set theory. The number of proteins in each of these subsets is given in Table 1, and the corresponding accession numbers as well as their amino acid sequences are given in Online Supporting Information A.

Table 1
Numbers of the secretory and non-secretory proteins in each of the six different organism datasets

Organism	Number of secretory proteins $\mathbb{S}^{+}$	Number of non-secretory proteins $\mathbb{S}^{-}$	Total
Human $\mathbb{S}_{\text{Human}}$	894	1129	2203
Plant $\mathbb{S}_{\text{Plant}}$	338	559	897
Animal $\mathbb{S}_{\text{Animal}}$	1435	1762	3197
Eukaryotic[a] $\mathbb{S}_{\text{Euk}}$	635	785	1420
Gram-positive $\mathbb{S}_{\text{Gpos}}$	269	356	625
Gram-negative $\mathbb{S}_{\text{Gneg}}$	613	721	1334

The corresponding accession numbers and sequences are given in Online Supporting Information A.

[a] Including all the eukaryotic proteins obtained by following steps 1–4 in Materials except those already occurring in the datasets for human, plant, and animal (see Eq. (1)).

Methods

The best known signal sequences are N-terminal extensions although they can also be located within a protein or at its C-terminal end (e.g., for "tail-anchored" membrane proteins [24]). All secreted proteins are synthesized with N-terminal signal peptides, which are the focus of the present study. Typically 15–30 amino acids long, the N-terminal signal peptides are cleaved off by signal peptidase during their translocation across the membrane [1,25]. To predict the signal peptide sequence and its cleavage site for a query protein sequence **P**, the first step is to identify whether it is a secretory or non-secretory. When, and only when, the protein is secretory, further action is needed to determine its signal peptide sequence. Therefore, the current approach consists of the following steps.

Discrimination of secretory proteins from non-secretory

Suppose a dataset $\mathbb{S}$ of N proteins $(\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_N)$ that can be separated into two subsets: $\mathbb{S}^{+}$ consists of the secretory proteins only; while $\mathbb{S}^{-}$, the non-secretory proteins only. Now, for a query protein **P** consisting of L amino acids, how can we identify which of the two subsets it belongs to?

According to its amino acid sequence, the query protein **P** can be expressed as

$$\mathbf{P} = \mathrm{R}_1\mathrm{R}_2\mathrm{R}_3\mathrm{R}_4\mathrm{R}_5\mathrm{R}_6\mathrm{R}_7\cdots\mathrm{R}_L \quad (2)$$

where R_1 represents the 1st residue of the protein **P**, R_2 the 2nd residue, and so forth. Considering the fact that most signal peptides are within the first 50 N-terminal residues, to simplify the problem, the formulation for **P** can be simplified as

$$\mathbf{P}' = \mathrm{R}_1\mathrm{R}_2\mathrm{R}_3\mathrm{R}_4\mathrm{R}_5\mathrm{R}_6\mathrm{R}_7\cdots\mathrm{R}_{50}. \quad (3)$$

In order to incorporate the sequence-order effects, instead of the conventional amino acid (AA) composition, let us adopt the pseudo amino acid (PseAA) composition to treat the problem. According to the PseAA composition discrete model [26], $\mathbf{P}'$ of Eq. (3) can be formulated as

$$\mathbf{P}' = \left[p_1, p_2, \ldots, p_{20}, p_{20+1}, \ldots, p_{20+\lambda}\right]^{\mathbf{T}}, \quad (\lambda < 50) \quad (4)$$

where **T** is a transpose operator, $p_1, p_2, \ldots, p_{20}$ are associated with the conventional AA composition reflecting the occurrence frequencies of the 20 native amino acids in $\mathbf{P}'$ [27], and $p_{20+1}, p_{20+2}, \ldots, p_{20+\lambda}$ are the λ correlation factors that reflect the first tier, second tier, ..., and the λ-th tier sequence order correlation patterns. Given a real amino acid sequence for $\mathbf{P}'$ of Eq. (3), the $(20+\lambda)$ elements in Eq. (4) can be easily derived by the **PseAA** web-server at http://chou.med.harvard.edu/bioinf/PseAA/ or Eqs. (2)–(6) of [26]. It is the additional λ factors that approximately incorporate the sequence-order effects. Likewise, we can also have the corresponding expression for each of $(\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_N)$ in $\mathbb{S}$.

Using the PseAA composition discrete model to represent protein samples can significantly improve the prediction quality for various protein attributes as demonstrated by a series of recent publications [28–40].

Based on the PseAA composition discrete model as formulated in Eq. (4), the optimized evidence-theoretic K nearest neighbor (OET-KNN) classifier, a very powerful classification engine as presented in Appendix B of [41], was utilized to identify the query protein between secretory and non-secretory.

However, the predicted result by the OET-KNN classifier will depend on the selection of the parameter K, the number of the nearest neighbors to the query protein **P**, and the value of λ, the number of the correlation factors in Eq. (4). Therefore, the OET-KNN classifier should be expressed as $\mathbb{C}(K, \lambda)$, as done in [41]. To get the optimal success rate, one has to test the results by using different numbers of K and λ one by one. However, it

is both time-consuming and tedious to do so. To solve such a problem, the following two-dimensional fusion process is developed to generate an ensemble classifier by fusing many individual $\mathbb{C}(K,\lambda)$ classifiers.

Preliminary tests indicate that the success rate obtained by the $\mathbb{C}(K,\lambda)$ trained with the current dataset was gradually lower when $K>10$, $\lambda<20$, or $\lambda>30$ and hence we can reduce the scope by just considering

$$\{K\}=\{1,2,\ldots,10\};\{\lambda\}=\{20,21,\ldots,30\}. \tag{5}$$

Thus, the ensemble classifier obtained by the two-dimensional fusion process can be formulated as

$$\mathbb{C}=\mathbb{C}(1,20)\oplus\mathbb{C}(1,21)\oplus\cdots\oplus\mathbb{C}(10,29)\oplus\mathbb{C}(10,30) \tag{6}$$

where the symbol $\oplus$ denotes the fusion operator, and $\mathbb{C}$ the ensemble classifier formed by fusing the $10\times11=110$ basic individual classifier $\mathbb{C}(1,20),\mathbb{C}(1,21),\ldots,\mathbb{C}(10,30)$ according to the following procedures. Suppose the predicted classification results for the query protein **P** by the 110 individual classifiers in Eq. (6) are

$$\begin{Bmatrix} C_{1,20} & C_{1,21} & \cdots & C_{1,30} \\ C_{2,20} & C_{2,21} & \cdots & C_{2,30} \\ \vdots & \vdots & \ddots & \vdots \\ C_{10,20} & C_{10,21} & \cdots & C_{10,30} \end{Bmatrix} \in \{\mathbb{S}^+,\mathbb{S}^-\} \tag{7}$$

where $\in$ is a symbol in the set theory meaning "member of", $\mathbb{S}^+$ is the subset for the secretory proteins, $\mathbb{S}^-$ is the subset for the non-secretory proteins, and the voting score for the query protein **P** belonging to the θ-th subset is defined by

$$Y^\theta=\sum_{i=1}^{10}\sum_{j=20}^{30}w_{i,j}\Delta(C_{i,j},\mathbb{S}^\theta),\quad(\theta=+\text{ or }-) \tag{8}$$

where $w_{i,j}$ is the weight and was set at 1 for simplicity, the delta function in Eq. (8) is given by

$$\Delta(C_{i,j},\mathbb{S}^\theta)=\begin{cases}1, & \text{if } C_{i,j}\in\mathbb{S}^\theta\\ 0, & \text{otherwise}\end{cases} \tag{9}$$

thus the query protein **P** is predicted belonging to the secretory protein subset if $Y^+>Y^-$, or vice versa. For the case of $Y^+=Y^-$, the query protein would be randomly assigned to one of the two subsets but this kind of tie case never happened in the current study. If the query protein **P** is identified as secretory, it will be continued to predict the cleavage site of its signal peptide.

Selecting cleavage sites candidates

The cleavage site of an N-terminal signal peptide is at the position between the last residue of the signal peptide and the first residue of the mature protein. Accordingly, once the cleavage site is identified, the corresponding signal peptide is automatically known. The scaled window approach [15] was adopted for this study. Symbolized as $[-\xi_1,+\xi_2]$, the scaled window is marked consecutively with $-\xi_1,\ldots,-3,-2,-1,+1,+2,\ldots,+\xi_2$ to define the position of amino acids of a protein sequence within the window (Fig. 1). Thus, when sliding the scaled window $[-\xi_1,+\xi_2]$ along the sequence of a secretory protein with n residues, one can consecutively highlight $n-(\xi_1+\xi_2)+1$ segments, of which only the one with the residue at the scale -1 being the very last residue of the signal sequence (or the residue at the scale $+1$ being the first residue of the mature sequence) is regarded as the secretion-cleavable segment (Fig. 1A), while all the other segments regarded as non-secretion-cleavable (see, e.g., Fig. 1B and C).

A predictor based on the subsite-coupled model, or $\{-3,-1,+1\}$ coupling model, was developed by Chou [15] for predicting the secretion-cleavable segment among the $n-(\xi_1+\xi_2)+1$ peptide segments generated by the scaled window $[-\xi_1,+\xi_2]$. However, for a given secretory protein, usually more than one secretion-cleavable segment was predicted by the subsite-coupled predictor, meaning that the results thus obtained might contain false positive. Therefore, the desired cleavage site and hence the signal peptide for a query protein cannot be uniquely identified. Nevertheless, the results from the subsite-coupled predictor can help us wisely

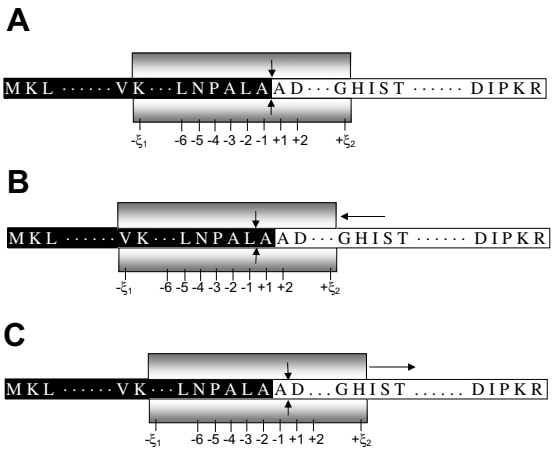

Fig. 1. Illustration to show the sequence segments highlighted by sliding the scaled window $[-\xi_1,+\xi_2]$ along a protein sequence. During the sliding process, the scales on the window are aligned with different amino acids so as to define different peptide segments. When, and only when, the scale -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.

select a set of candidates for uniquely determining the cleavage site and signal peptide, as formulated below.

According to Table 1 of [15], the optimal values for the window parameters are $\xi_1=13$ and $\xi_2=2$, and hence the scaled window $[-13,+2]$ was adopted for the current study as well. Of the $n-(13+2)+1=n-14$ segments, suppose the following μ segments are predicted by the subsite-coupled predictor [15] as secretion-cleavable:

$$\mathbf{P}^i=\mathrm{R}^i_{-\xi_1}\mathrm{R}^i_{-(\xi_1-1)}\cdots\mathrm{R}^i_{-3}\mathrm{R}^i_{-2}\mathrm{R}^i_{-1}\mathrm{R}^i_{+1}\mathrm{R}^i_{+2}\cdots\mathrm{R}^i_{+(\xi_2-1)}\mathrm{R}^i_{+\xi_2},\quad(i=1,2,\ldots,\mu) \tag{10}$$

where the position between R^i_{-1} and R^i_{+1} is the possible cleavage site for protein **P**. Each of the possible cleavage sites in Eq. (10) is taken as a cleavage site candidate, and we have μ candidates that correspond to μ different positions in the sequence of **P**, as formulated by

$$\{k_i\},\quad(i=1,2,\ldots,\mu) \tag{11}$$

with the corresponding credit given by

$$\{\tau_i\},\quad(i=1,2,\ldots,\mu) \tag{12}$$

where τ_i is the value of the discriminant function for $\mathbf{P}^i$ of Eq. (10) as clearly defined by Eq. 4 of [15]: the greater the value of τ_i, the higher the possibility for $\mathbf{P}^i$ to be secretion-cleavable. Below, we are to determine which one of the μ candidates represents the real cleavage site.

Identifying the real cleavage site from the candidates

Suppose $\mathbf{P}(k_i\downarrow30)$ is a hypothetical protein which has exactly the same amino acid sequence with the query protein **P** except for (1) the last amino acid of its signal peptide is located at the sequence position k_i, and (2) only the first 30 amino acids after the position k_i is kept for the mature chain, i.e., the mature chain beyond these amino acids is truncated as treated in PrediSi [21] and SignalP [20]. According to Eq. (11), there are μ such hypothetical proteins; i.e.,

$$\{\mathbf{P}(k_i\downarrow30)\},\quad(i=1,2,\ldots,\mu). \tag{13}$$

To determine the real cleavage site among the μ possible cleavage sites of Eq. (11), we use the Needleman–Wunsch alignment algorithm [42]. Because using different parameters d and e for Needleman–Wunsch algorithm will result in different results, below we shall use $\mathbf{NW}(d,e)$ to denote the algorithm, where d is called the gap-open penalty, e the gap-extension penalty (in this study, we took $d=5$ and $e=2$). Suppose

$$\{S_j\}, \quad (j=1,2,\ldots,N) \tag{14}$$

is a set of N secretory protein sequences each with known signal sequence. The global alignment of the i-th hypothetical protein $\mathbf{P}(k_i\downarrow 30)$ with each of the sequences in Eq. (14) will generate N alignment pairs, as formulated below:

$$\{[\![\mathbf{P}(k_i\downarrow 30), S_j]\!]\}, \quad (j=1,2,\ldots,N). \tag{15}$$

Each of the N alignments will leave a cleavage mark onto $\mathbf{P}(k_i\downarrow 30)$, and hence we have

$$\{\Lambda_{i,j}\}, \quad (i=1,2,\cdots,\mu;\ j=1,2,\cdots,N) \tag{16}$$

where $\Lambda_{i,j}$ is the site in $\mathbf{P}(k_i\downarrow 30)$ that corresponds to the known cleavage site in S_j and is regarded as the deduced cleavage site from the alignment $[\![\mathbf{P}(k_i\downarrow 30), S_j]\!]$.

Thus, the supporting degree from the global alignment for the candidate k_i of Eq. (11) as the real signal peptide cleavage site of the query protein $\mathbf{P}$ is defined by

$$Q_i=\sum_{j=1}^{N}\tau_i\Phi(\mathbf{P}(k_i\downarrow 30), S_j)\Delta(k_i,\Lambda_{i,j}), \quad (i=1,2,\ldots,\mu) \tag{17}$$

where τ_i is the credit for the candidate k_i in Eq. (11) that can be obtained from Eq. (12); $\Phi(\mathbf{P}(k_i\downarrow 30),S_j)$ is the similarity criterion between $\mathbf{P}(k_i\downarrow 30)$ and S_j that is obtained by $\mathbf{NW}(d,e)$ algorithm; and the delta function $\Delta(k_i,\Lambda_{i,j})$ is given by

$$\Delta(k_i,\Lambda_{i,j})=\begin{cases}1, & \text{if } k_i=\Lambda_{i,j}\\ 0, & \text{otherwise.}\end{cases} \tag{18}$$

The final decision is made by assigning k_u of Eq. (11) as the signal sequence cleavage site for the query sequence $\mathbf{P}$ if

$$u=\mathbf{ArgMax}_i\{Q_i\}, \quad (i=1,2,\ldots,\mu) \tag{19}$$

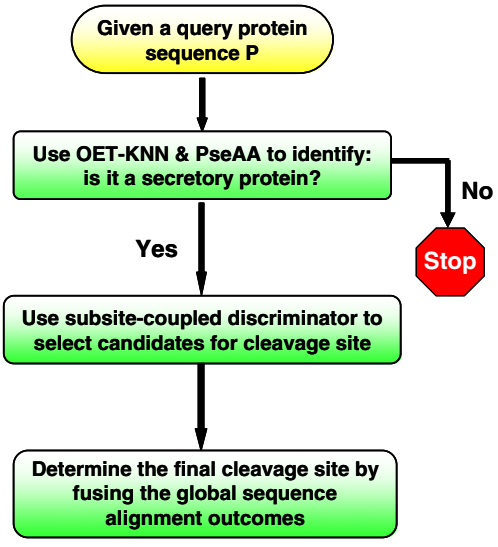

Fig. 2. A flowchart to show the process of how the 3-layer predictor works in identifying a query protein as secretory or non-secretory, selecting the candidates of its signal peptide cleavage site if the protein is secretory, and determining the final cleavage site.

where the operator $\mathbf{ArgMax}_i$ means taking the subscript of Q with which the score function of Eq. (17) is the maximum. If there is a tie among two or more deduced cleavage sites, then the final predicted site will be randomly assigned to one of their corresponding sites although this kind of tie case rarely happens and actually was not observed in the current study.

A flowchart is given in Fig. 2 to show the process of how the 3-layer predictor works in identifying a query protein as secretory or non-secretory, selecting the candidates of its signal peptide cleavage site if the protein is secretory, and determining the final cleavage site.

Results and discussion

In statistical prediction, the methods often used for cross-validating the accuracy of a predictor are the single independent dataset test, sub-sampling test and jackknife test. Of these three, however, the jackknife test is deemed as the most rigorous and objective one, as illustrated by a comprehensive review [43]. Therefore, jackknife test has been increasingly used in literatures [28,31,33,35,44–55] for examining the power of various prediction methods. In the jackknife test, each protein in the benchmark dataset was singled out in turn as a "test protein" and all the rule parameters were calculated from the remaining proteins. In other words, the signal peptide of each protein was predicted by the rules derived using all the other proteins except the one that was being predicted. During the process of jackknife test, both the training and testing datasets were actually open, and a protein was in turn moving from one to the other.

PrediSi [21] and SignalP [20] are two popular web-server predictors developed recently for identifying the signal peptide and its cleavage site. The way SignalP was designed does not allow it to be examined by the jackknife test. However, it is elaborated in [21] that the prediction accuracy of PrediSi is quite compatible with that of SignalP. Here, let us first compare the current Signal-3L with PrediSi [21]. The success rates by the jackknife test with PrediSi and Signal-3L on the newly constructed datasets for human, plant, animal, eukaryotic, Gram-positive, and Gram-negative benchmark datasets (Online Supporting Information A) are listed in Table 2, from which we can see that, in identifying the signal peptide cleavage sites for the secretory proteins in the six organism-different

Table 2
Comparisons of success rates by jackknife test on each of the six newly constructed datasets as given in Online Supporting Information A

Organism	Discrimination of secretory and non-secretory proteins (%)		Prediction of signal peptide cleavage site (%)	
	PrediSi	**Signal-3L**	PrediSi	**Signal-3L**
Human	91.1	**92.3**	68.0	**73.4**
Plant	93.6	**95.8**	70.1	**82.8**
Animal	93.2	**95.7**	71.9	**77.7**
Eukaryotic[a]	92.1	**94.0**	65.7	**76.2**
Gram-positive	94.6	**98.1**	60.2	**78.8**
Gram-negative	91.2	**94.4**	80.3	**88.1**

The results obtained by Signal-3L are in bold-face; those by PrediSi are in normal face.

[a] See footnote 'a' of Table 1.

benchmark datasets, Signal-3L is 5–18% higher than PrediSi [21]. Because SignalP [20] is a predictor with a built-in training dataset covering only three different organisms, to compare it with the current predictor Signal-3L, let us use both SignalP and Signal-3L to deal with the proteins whose signal peptides have been experimentally verified [56]. The outcomes are given in Table 3, from which we can see that many protein signal peptides miss-predicted by SignalP were successfully corrected by Signal-3L. Also, it can be seen from Table 3 that some of the results predicted by SignalP 3.0-NN and SignalP 3.0-HMM, two important signal peptide predictors in the SignalP package, are often inconsistent. For example, the signal peptide of FZD3_HUMAN predicted by SignalP 3.0-NN is 1–17, but that by SignalP 3.0-HMM is 1–22. This kind of inconsistency might cause confusion if no experiment result is timely available. But the predicted result by Signal-3L supports the latter, fully consistent with the experimental observation. However, for a different protein, such as IBP7_HUMAN, Signal-3L supports the result obtained by SignalP 3.0-NN rather than SignalP 3.0-HMM, also fully consistent with the experimental observation.

The above results and discussion indicate that the current Signal-3L is a powerful tool for predicting signal peptides. It can at least play an important complementary role to SignalP [20] and PrediSi [21] widely used in the relevant areas.

Table 3
List of examples showing that signal peptides miss-predicted by SignalP-NN and/or SignalP-HMM are corrected by Signal-3L

Protein[a]	Experimentally verified signal peptide[a]	SignalP 3.0-NN	SignalP 3.0-HMM	Signal-3L
AAF91396.1	**1–40**	1–37	1–37	**1–40**
DKK1_HUMAN	**1–31**	1–22	1–28	**1–31**
MIME_HUMAN	**1–20**	1–19	1–19	**1–20**
NP_057466.1	**1–21**	1–19	1–19	**1–21**
NP_057663.1	**1–35**	1–30	1–46	**1–35**
NP_443122.2	**1–21**	1–22	1–22	**1–21**
NP_443164.1	**1–26**	1–33	1–33	**1–26**
Q6UXL0	**1–28**	1–29	1–29	**1–28**
STC1_HUMAN	**1–17**	1–21	1–18	**1–17**
TRLT_HUMAN	**1–25**	1–24	1–27	**1–25**
CD5L_HUMAN	**1–19**	1–18	**1–19**	**1–19**
EDAR_HUMAN	**1–26**	1–28	**1–26**	**1–26**
FZD3_HUMAN	**1–22**	1–17	**1–22**	**1–22**
IBP7_HUMAN	**1–26**	**1–26**	1–29	**1–26**
KLK3_HUMAN	**1–17**	**1–17**	1–23	**1–17**
NMA_HUMAN	**1–20**	**1–20**	1–26	**1–20**
NP_064510.1	**1–22**	**1–22**	1–23	**1–22**
NP_068742.1	**1–24**	**1–24**	1–25	**1–24**
NTRI_HUMAN	**1–33**	1–30	**1–33**	**1–33**
SY01_HUMAN	**1–23**	**1–23**	1–18	**1–23**
TIE1_HUMAN	**1–21**	**1–21**	1–22	**1–21**
TL19_HUMAN	**1–26**	1–23	**1–26**	**1–26**
TR14_HUMAN	**1–38**	1–36	**1–38**	**1–38**
TR19_HUMAN	**1–29**	**1–29**	1–25	**1–29**
XP_166856	**1–17**	**1–17**	1–20	**1–17**
XP_209141	**1–22**	1–23	**1–22**	**1–22**

[a] Data taken from [56]. The signal peptides experimentally verified and correctly predicted are in bold-face; those incorrectly predicted are in normal face.

Conclusions

Signal-3L is a predictor of three layers. The target of the 1st-layer is to identify a query protein as secretory or non-secretory with the OET-KNN classifier in a PseAA composition space. If the protein is identified as secretory, the process will be automatically continued by entering into the 2nd-layer where a set of candidates for its signal peptide cleavage site are to be selected with a subsite-coupled discriminator by sliding a scaled window along the protein sequence. The 3rd-layer is to finally determine the unique cleavage site by fusing the global sequence alignment outcome for each of the selected candidates through a voting system.

Signal-3L is freely available as a web-server at http://chou.med.harvard.edu/bioinf/Signal-3L/ or http://202.120.37.186/bioinf/Signal-3L. To support the people working in the relevant areas, Signal-3L has been used to predict the signal peptide cleavage sites for all those protein entries in the Swiss-Prot database that have no signal peptide annotations or are annotated with uncertain terms but are classified as secretory proteins by Signal-3L. The results thus obtained have filled the blank area of signal peptide for 4080 human proteins, 3124 plant proteins, 13,527 animal proteins, 6165 other eukaryotic proteins, 5418 Gram-positive proteins, and 13,790 Gram-negative. The large-scale results have been deposited in a downloadable file prepared with Microsoft Excel and named "Tab_Signal-3L.xls". This file, along with Signal-3L as a free web-server, is available at http://chou.med.harvard.edu/bioinf/Signal-3L/ or http://202.120.37.186/bioinf/Signal-3L, and will be updated once a year to include new protein entries and reflect the continuous development of Signal-3L. These large-scale results can serve two purposes: one is that they can be directly used by those who need the signal peptide information immediately; the other is to set a preceding mark to examine the accuracy of our predicted results by the future experimental results.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.bbrc.2007.08.140.

References

[1] L.M. Gierasch, Signal sequences, Biochemistry 28 (1989) 923–930.
[2] K.C. Chou, Review: prediction of protein signal sequences, Curr. Protein Pept. Sci. 3 (2002) 615–622.
[3] M. Hagmann, Colleagues say "Amen" to this year's (Nobel Prizes) choice, Science 286 (1999) 666.
[4] D.J. McGeoch, On the predictive recognition of signal peptide sequences, Virus Res. 3 (1985) 271–286.
[5] G. von Heijne, A new method for predicting signal sequence cleavage sites, Nucleic Acids Res. 14 (1986) 4683–4690.

[6] R.J. Folz, J.I. Gordon, Computer-assisted predictions of signal peptidase processing sites, Biochem. Biophys. Res. Commun. 146 (1987) 870–877.
[7] I. Ladunga, F. Czako, I. Csabai, T. Geszti, Improving signal peptide prediction accuracy by simulated neural network, Comput. Appl. Biosci. 7 (1991) 485–487.
[8] P. Arrigo, F. Giuliano, F. Scalia, A. Rapallo, G. Damiani, Identification of a new motif on nucleic acid sequence data using Kohonen's self-organizing map, Comput. Appl. Biosci. 7 (1991) 353–357.
[9] G. Schneider, S. Rohlk, P. Wrede, Analysis of cleavage-site patterns in protein precursor sequences with a perception-type neural network, Biochem. Biophys. Res. Comm. 194 (1993) 951–959.
[10] G. Schneider, P. Wrede, Development of artificial filters for pattern recognition in protein sequences, J. Mol. Evol. 36 (1993) 586–595.
[11] G. Schneider, P. Wrede, Signal analysis of protein targeting sequences, Protein Seq. Data Anal. 5 (1993) 227–236.
[12] H. Nielsen, J. Engelbrecht, S. Brunak, G. von Heijne, Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites, Protein Eng. 10 (1997) 1–6.
[13] H. Nielsen, A. Krogh, Prediction of signal peptides and signal anchors by a hidden Markov model, Intell. Syst. Mol. Biol. 6 (1998) 122–130.
[14] K.C. Chou, Prediction of protein signal sequences and their cleavage sites, Proteins: Struct. Funct. Genet. 42 (2001) 136–139.
[15] K.C. Chou, Using subsite coupling to predict signal peptides, Protein Eng. 14 (2001) 75–79.
[16] K.C. Chou, Prediction of signal peptides using scaled window, Peptides 22 (2001) 1973–1979.
[17] K. Nakai, P. Horton, PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization, Trends Biochem. Sci. 24 (1999) 34–36.
[18] K. Nakai, Protein sorting signals and prediction of subcellular localization, Adv. Protein Chem. 54 (2000) 277–344.
[19] I. Ladunga, Large-scale predictions of secretory proteins from mammalian genomic and EST sequences, Curr. Opin. Biotechnol. 11 (2000) 13–18.
[20] J.D. Bendtsen, H. Nielsen, G. von Heijne, S. Brunak, Improved prediction of signal peptides: SignalP 3.0, J. Mol. Biol. 340 (2004) 783–795.
[21] K. Hiller, A. Grote, M. Scheer, R. Munch, D. Jahn, PrediSi: prediction of signal peptides and their cleavage positions, Nucleic Acids Res. 32 (2004) W375–W379.
[22] K.C. Chou, H.B. Shen, Signal-CF: a subsite-coupled and window-fusing approach for predicting signal peptides, Biochem. Biophys. Res. Commun. 357 (2007) 633–640.
[23] H. Liu, J. Yang, D.Q. Liu, H.B. Shen, K.C. Chou, Using a new alignment kernel function to identify secretory proteins, Protein Pept. Lett. 14 (2007) 203–208.
[24] U. Kutay, G. Ahnert-Hilger, E. Hartmann, B. Wiedenmann, T.A. Rapoport, Transport route for synaptobrevin via a novel pathway of insertion into the endoplasmic reticulum membrane, EMBO J. 14 (1995) 217–223.
[25] T.A. Rapoport, Transport of proteins across the endoplasmic reticulum membrane, Science 258 (1992) 931–936.
[26] K.C. Chou, Prediction of protein cellular attributes using pseudo amino acid composition, Proteins: Struct. Funct. Genet. 43 (2001) 246–255 (Erratum: ibid, 2001, vol. 44, 60).
[27] K.C. Chou, C.T. Zhang, Predicting protein folding types by distance functions that make allowances for amino acid interactions, J. Biol. Chem. 269 (1994) 22014–22020.
[28] Q.B. Gao, Z.Z. Wang, C. Yan, Y.H. Du, Prediction of protein subcellular location using a combined feature of sequence, FEBS Lett. 579 (2005) 3444–3448.
[29] H.B. Shen, K.C. Chou, Using optimized evidence-theoretic *K*-nearest neighbor classifier and pseudo amino acid composition to predict membrane protein types, Biochem. Biophys. Res. Commun. 334 (2005) 288–292.
[30] H.B. Shen, K.C. Chou, Predicting protein subnuclear location with optimized evidence-theoretic *K*-nearest classifier and pseudo amino acid composition, Biochem. Biophys. Res. Commun. 337 (2005) 752–756.
[31] C. Chen, X. Zhou, Y. Tian, X. Zou, P. Cai, Predicting protein structural class with pseudo-amino acid composition and support vector machine fusion network, Anal. Biochem. 357 (2006) 116–121.
[32] H.B. Shen, K.C. Chou, Ensemble classifier for protein fold pattern recognition, Bioinformatics 22 (2006) 1717–1722.
[33] S. Mondal, R. Bhavna, R. Mohan Babu, S. Ramakumar, Pseudo amino acid composition and multi-class support vector machines approach for conotoxin superfamily classification, J. Theor. Biol. 243 (2006) 252–260.
[34] C. Chen, Y.X. Tian, X.Y. Zou, P.X. Cai, J.Y. Mo, Using pseudo-amino acid composition and support vector machine to predict protein structural class, J. Theor. Biol. 243 (2006) 444–448.
[35] S.W. Zhang, Q. Pan, H.C. Zhang, Z.C. Shao, J.Y. Shi, Prediction protein homo-oligomer types by pseudo amino acid composition: approached with an improved feature extraction and naive Bayes feature fusion, Amino Acids 30 (2006) 461–468.
[36] P. Du, Y. Li, Prediction of protein submitochondria locations by hybridizing pseudo-amino acid composition with various physico-chemical features of segmented sequence, BMC Bioinformatics 7 (2006) 518.
[37] H. Lin, Q.Z. Li, Predicting conotoxin superfamily and family by using pseudo amino acid composition and modified Mahalanobis discriminant, Biochem. Biophys. Res. Commun. 354 (2007) 548–551.
[38] H. Lin, Q.Z. Li, Using pseudo amino acid composition to predict protein structural class: approached by incorporating 400 dipeptide components, J. Comput. Chem. 28 (2007) 1463–1466.
[39] Y.L. Chen, Q.Z. Li, Prediction of apoptosis protein subcellular location using improved hybrid approach and pseudo amino acid composition, J. Theor. Biol. 248 (2007) 377–381.
[40] X.B. Zhou, C. Chen, Z.C. Li, X.Y. Zou, Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes, J. Theor. Biol. 248 (2007) 546–551.
[41] K.C. Chou, H.B. Shen, Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic *K*-nearest neighbor classifiers, J. Proteome Res. 5 (2006) 1888–1897.
[42] S.B. Needleman, C.D. Wunsch, A general method applicable to the search for similarities in the amino acid sequence of two proteins, J. Mol. Biol. 48 (1970) 443–453.
[43] K.C. Chou, C.T. Zhang, Review: prediction of protein structural classes, Crit. Rev. Biochem. Mol. Biol. 30 (1995) 275–349.
[44] G.P. Zhou, K. Doctor, Subcellular location prediction of apoptosis proteins, Proteins: Struct. Funct. Genet. 50 (2003) 44–48.
[45] Y.Z. Guo, M. Li, M. Lu, Z. Wen, K. Wang, G. Li, J. Wu, Classifying G protein-coupled receptors and nuclear receptors based on protein power spectrum from fast Fourier transform, Amino Acids 30 (2006) 397–402.
[46] X.D. Sun, R.B. Huang, Prediction of protein structural classes using support vector machines, Amino Acids 30 (2006) 469–475.
[47] Z. Wen, M. Li, Y. Li, Y. Guo, K. Wang, Delaunay triangulation with partial least squares projection to latent structures: a model for G-protein coupled receptors classification and fast structure recognition, Amino Acids 32 (2006) 277–283.
[48] K.C. Chou, H.B. Shen, Hum-PLoc: a novel ensemble classifier for predicting human protein subcellular localization, Biochem. Biophys. Res. Commun. 347 (2006) 150–157.
[49] Y. Cao, S. Liu, L. Zhang, J. Qin, J. Wang, K. Tang, Prediction of protein structural class with Rough Sets, BMC Bioinformatics 20 (7) (2006).
[50] J. Guo, Y. Lin, X. Liu, GNBSL: a new integrative system to predict the subcellular location for Gram-negative bacteria proteins, Proteomics 6 (2006) 5099–5105.

[51] K.C. Chou, H.B. Shen, Euk-mPLoc: a fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites, J. Proteome Res. 6 (2007) 1728–1734.
[52] Q.B. Gao, Z.Z. Wang, Classification of G-protein coupled receptors at four levels, Protein Eng. Des. Sel. 19 (2006) 511–516.
[53] K.C. Chou, H.B. Shen, Large-scale plant protein subcellular location prediction, J. Cell. Biochem. 100 (2007) 665–678.
[54] H.B. Shen, K.C. Chou, Hum-mPLoc: an ensemble classifier for large-scale human protein subcellular location prediction by incorporating samples with multiple sites, Biochem. Biophys. Res. Commun. 355 (2007) 1006–1011.
[55] K.C. Chou, H.B. Shen, MemType-2L: a Web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM, Biochem. Biophys. Res. Commun. 360 (2007) 339–345.
[56] Z. Zhang, W.J. Henzel, Signal peptide prediction based on analysis of experimentally verified cleavage sites, Protein Sci. 13 (2004) 2819–2824.