Fast Fourier Transform-based Support Vector Machine for Subcellular Localization Prediction Using Different Substitution Models

Zhimeng WANG, Lin JIANG, Menglong LI*, Lina SUN, and Rongying LIN

College of Chemistry, Sichuan University, Chengdu 610064, China

Abstract There are approximately 10^9 proteins in a cell. A hotspot in bioinformatics is how to identify a protein's subcellular localization, if its sequence is known. In this paper, a method using fast Fourier transform-based support vector machine is developed to predict the subcellular localization of proteins from their physicochemical properties and structural parameters. The prediction accuracies reached 83% in prokaryotic organisms and 84% in eukaryotic organisms with the substitution model of the c-p-v matrix (c, composition; p, polarity; and v, molecular volume). The overall prediction accuracy was also evaluated using the "leave-one-out" jackknife procedure. The influence of the substitution model on prediction accuracy has also been discussed in the work. The source code of the new program is available on request from the authors.

Keywords protein subcellular localization; prediction; substitution model; fast Fourier transform; support vector machine

The number of new or complete protein sequences has dramatically increased over recent years [1] and this has created the need to functionally annotate this data. Subcellular localization is a key functional characteristic of protein [2]. Prediction of protein subcellular localization is an important step in understanding the biochemical function of proteins.

Because proteins with similar functions have sequential identical amino acid sequences to some extent, sequence-scale-based measurement is a feasible prediction method [3]. According to amino acid similarity, amino acid composition, sequence signals, N-terminal amino acid sequence and so on, subcellular localization of the proteins which is only known by sequence has been predicted [4–19]. Many programs, such as the support vector machine (SVM) [6,12,16,18,19], k-nearest-neighbor method [14, 15], neural network model [7,9,17], hidden Markov model [11,17], co-variant discriminant algorithm [10] and discrete wavelet transform [4,5] are widely used. For example, PSLpred [16] is a localization prediction tool for Gram-negative bacteria that uses SVM and PSI-BLAST (http://www.ncbi.nlm.nih.gov/BLAST/) to generate predictions for five localization sites. SubLoc [6] uses SVM to assign a prokaryotic protein to the cytoplasmic, periplasmic, or extracellular sites, and a eukaryotic protein to the cytoplasmic, mitochondrial, nuclear, or extracellular sites. SignalP [17] predicts traditional N-terminal signal peptides in both prokaryotic and eukaryotic proteins based on neural network and hidden Markov model algorithms. HSLPred [18] is a localization prediction tool for human proteins that uses SVM and PSI-BLAST to generate predictions for four localization sites. TargetP [9] predicts the presence of signal peptides, chloroplast transit peptides and mitochondrial targeting peptides for plant proteins, and the presence of signal peptides and mitochondrial targeting peptides for eukaryotic proteins. pSLIP [19] uses SVM and multiple physicochemical properties of amino acids to assign a eukaryotic protein to one of six localization sites. The detailed introduction of localization prediction tools for both prokaryotes and eukaryotes is available at http://www.psort.org. But higher predictive accuracy is still an

Received: March 18, 2007 Accepted: June 20, 2007

The work was supported by a grant from the State Key Laboratory of Chemo/Biosensing and Chemometrics, Hunan University

*Corresponding author: Tel, 86-28-85413330; Fax, 86-28-85412356; E-mail, liml@scu.edu.cn

DOI: 10.1111/j.1745-7270.2007.00326.x

aim for these algorithms.

In this paper, a new method for predicting protein subcellular localization is introduced. This method couples fast Fourier transform (FFT) with SVM on the basis of different physicochemical properties and structural parameters of amino acid sequences. The input sequences first are transformed into numerical signals by the substitution model. Then the FFT changes the signals into equal lengths, from time-based to frequency-based. Finally, SVM is used to model with these signals. The influence of three substitution models to the result is discussed. The performance of the model was also evaluated by the jackknife test.

Materials and Methods

Datasets

The dataset used in the present work was generated by Reinhardt and Hubbard [7], which was also used to model the NNPSL [7], the Markov chain model [11], SubLoc [6] and ESLpred [12]. These sequences were extracted from release 33.0 of the SWISS-PROT database [20], all appeared complete and had what appeared to be reliable location annotations coming directly from experiments. Transmembrane proteins, plant sequences and redundant proteins with more than 90% sequence identity were excluded [7,21]. As shown in **Table 1**, there are 2427 protein sequences from eukaryotic species classified into four location groups, cytoplasmic, extracellular, nuclear and mitochondrial. In the prokaryotic species, 997 sequences belong to three location categories, cytoplasmic, extracellular and periplasmic. The continual updating of the SWISS-PROT release means there might be some limitations in the current work. But in order to compare with published work, the number of sequences was not increased.

Table 1 Number of sequences within each subcellular location group [7]

Species	Subcellular localization	No. of sequences
Eukaryotic	Nuclear	1097
	Cytoplasmic	684
	Mitochondrial	321
	Extracellular	325
Prokaryotic	Cytoplasmic	688
	Periplasmic	202
	Extracellular	107

Substitution model

The sensitivity of most protein sequence classification methods depends strongly on the quality of the substitution matrices used. These matrices, which assign weights or similarity scores to every possible amino acid pair, are used to differentiate among the various possible alignments of two or more sequences [22]. In this paper we discuss three amino acid matrices of biochemical properties and structural parameters, that is, the *c-p-v* matrix [23], the *EIIP* model [4] and the *hybrid* model [24]. The *c-p-v* matrix takes into account the three main properties of amino acids: the composition (*c*), polarity (*p*) and molecular volume (*v*) [23]. The *EIIP* model substitutes amino acids with the value of the electron-ion interaction potential (*e*), which describes the average energy states of all valence electrons in particular amino acids [4]. The *hybrid* model [24] combines three physicochemical properties with three structural parameters of amino acids, including polarity (*p*), volume (*v*), composition (*c*), electron-ion interaction potential (*e*), hydrophobicity (*h*) [25] and relative stability of α-helix conformation (*s*) [26]. First, according to frequencies and probabilities, the values for each amino acid property were calculated using the different models [4,23,24]. Then all amino acids in sequences were represented by the value of these matrices (**Table 2**). Finally, these numerical series were normalized to zero mean and unit standard deviation, calculated as **Equation 1**:

$$X_{ij}{}' = \frac{X_{ij} - \bar{X}_j}{S_j} \quad \begin{cases} j = 1,\ 2,\ 3 & c\text{-}p\text{-}v \text{ matrix} \\ j = 1 & EIIP \text{ model} \\ j = 1,\ 2, \ldots, 6 & hybrid \text{ model} \end{cases} \qquad 1$$

where i=amino acid, Ala to Val, 20; X_{ij} is the jth character of the ith one of 20 amino acids; $\bar{X}_j$ is the average value of the jth character in one sequence; S_j is the standard deviation in the jth character in one sequence, and $X_{ij}{}'$ is the normal form of the jth property value of the ith amino acid. Each amino acid in sequences is represented as **Equation 2**:

$$AA_i = \sum_{j=1}^{n} X_{ij}{}' \qquad n\text{=1, 3 or 6} \qquad 2$$

where AA_i is the substitution value of each amino acid [24].

With modern computing technology the digital implementation of the Fourier transform (FT) is widely available, mostly in the form of the FFT [27,28]. FT has become a basic tool for analysis of many biological signals.

The FT of a function $g(t)$ is defined as **Equation 3**:

Table 2 Properties of amino acid

aa	Property					
	c	*p*	*v*	*e*	*h*	*s*
Ser	1.42	9.2	32.0	0.0829	−0.18	−1.71
Arg	0.65	10.5	124.0	0.0959	−1.37	−2.84
Leu	0.00	4.9	111.0	0.0000	1.06	−2.59
Pro	0.39	8.0	32.5	0.0198	0.12	12.54
Thr	0.71	8.6	61.0	0.0941	−0.05	−0.46
Ala	0.00	8.1	31.0	0.0373	0.62	−3.22
Val	0.00	5.9	84.0	0.0057	1.08	−0.59
Gly	0.74	9.0	3.0	0.0050	0.48	0.00
Ile	0.00	5.2	111.0	0.0000	1.38	−0.96
Phe	0.00	5.2	132.0	0.0946	1.19	−1.88
Tyr	0.20	6.2	136.0	0.0516	0.26	−0.71
Cys	2.75	5.5	55.0	0.0829	0.29	−1.13
His	0.58	10.4	96.0	0.0242	−0.40	−0.25
Gln	0.89	10.5	85.0	0.0761	−0.78	−1.46
Asn	1.33	11.6	56.0	0.0036	−0.85	−0.29
Lys	0.33	11.3	119.0	0.0371	−1.35	−2.72
Asp	1.38	13.0	54.0	0.1263	−1.05	−0.63
Glu	0.92	12.3	83.0	0.0058	−0.87	−1.38
Met	0.00	5.7	105.0	0.0823	0.64	−2.09
Trp	0.13	5.4	170.0	0.0548	0.81	−1.88

Composition (*c*), polarity (*p*), molecular volume (*v*), electron-ion interaction potential (*e*), hydrophobicity (*h*) and relative stability of α-helix conformation (*s*) [4,23,24]. The values of *c*, *p*, *v*, *e*, *h* and *s* were obtained with different models. aa, amino acid.

$$G(\omega) = \int_{-\infty}^{\infty} g(t) e^{-i\omega t} dt \qquad 3$$

The inverse is shown as **Equation 4:**

$$g(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} G(\omega) e^{i\omega t} d\omega \qquad 4$$

The FFT is a fast algorithm of discrete Fourier transform (DFT) (**Equation 5**):

$$X(k) = \sum_{j=1}^{N} x(j) \omega_N^{(j-1)(k-1)} \qquad 5$$

where $\omega = e^{(-2\pi i)/N}$ is an *N*th root of unity.

Because the FFT changes the signal from time-based to frequency-based, *N* is the number of frequency points. In this work, 512 frequency points were set and the power spectrum, a measurement of the power at each frequency, was used. Given by *X*(*k*)=*Y*, the power spectrum is represented as **Equation 6:**

$$Pyy = Y.*conj(Y)/N \qquad 6$$

The power at each frequency point was taken as the input feature of the SVMs.

SVM

SVM [29,30] is a kind of learning machine based on statistical learning theory. The SVM is particularly attractive to biological sequence analysis due to its ability to handle noise, large datasets and large input spaces [31]. For a two-class problem, samples are described by the feature vectors x_i (*i*=1, 2,..., *k*) with corresponding labels y_i={+1,−1} (*i*=1, 2,..., *k*), where +1 and −1 are used to stand for the two classes. To classify them, SVM maps the input vector into feature space using radial Gaussian kernel and constructs the maximum margin hyperplane in the feature space. The maximum margin hyperplane is given by solving the following convex quadratic programming problem (**Equation 7**):

$$\text{Maximize} \sum_{i=1}^{l} a_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} a_i a_j y_i y_j K\left(x_i, x_j\right) \qquad 7$$

$$\text{subject to} \sum_{i=1}^{l} a_i y_i = 0 \qquad (0 \le a_i \le \theta) \qquad 8$$

θ is a regularization parameter that controls the trade-off between margin and classification error. $K\left(x_i, x_j\right)$ is the kernel function. In this paper, the radial basis function (RBF) was selected as the kernel function (**Equation 9**):

$$K\left(x, x_i\right) = \exp\left\{-\frac{1}{2\sigma^2} \| x - x_i \|^2\right\} \qquad 9$$

where, σ is the kernel width parameter.

Then the problem of classifying a new data vector *x* is simply solved by the following (**Equation 10**):

$$f\left(x\right) = Sgn\left\{\sum_{i=1}^{l} y_i a_i K\left(x_i, x\right) + b\right\} \qquad 10$$

If *f(x)*≥0, it means that the functional group is likely to be present, or it means that the functional group is likely to be absent.

Measuring prediction performance

To compare with other models, accuracy [32] was calculated to assess the predictive performance. *tp*, *fn*, *fp* and *tn* are the number of true positives, false negatives, false positives and true negatives, respectively, and *N* is the total number of sequences.

Accuracy (*Acc*, proportion of correct predictions to one localization group prediction) is calculated as **Equation 11**:

$$Acc = \frac{(tp + tn)}{(tp + tn + fn + fp)} = \frac{(tp + tn)}{N} \qquad 11$$

Overall accuracy (proportion of correct predictions of all the sequences) is calculated as **Equation 12**:

$$\text{Overall } Acc = \frac{\sum_{i=1}^{k} Acc(i)}{k} \qquad 12$$

where k is the class number, and $Acc(i)$ is the proportion of correct predictions of location i, containing positive and negative samples.

Results and Discussion

Comparing different substitution models

Three substitution models are compared in this work: the *c-p-v* matrix, the *EIIP* model and the *hybrid* model. Because of the computer's technological ability, two-fold tests were used. This means that half of the data are used as training, and the others are test, randomly. This is called Sample 1 in this paper. Then the training and test sets are turned, which is called Sample 2. The performance of the three models is shown in **Tables 3** and **4**.

Except for the cytoplasmic location of eukaryotic sequences, the results of the *c-p-v* matrix is best (**Tables 3** and **4**). The overall *Acc* of Sample 1 and Sample 2 with the *c-p-v* matrix are 0.84 and 0.83, respectively. We suggest that the *EIIP* model only reflects the electron-ion interaction potential of the protein and can not represent its full characteristics. The *hybrid* model causes the germination of the information. But the *c-p-v* model appears to contain the most sequence information. It shows that periodicity of polarity is one of the good indicators for the signal peptide. The individual polarity values were used to build the substitution model (**Tables 3** and **4**). From the results, the accuracies of the polarity are the highest in all. The polarity is one of the essential forces for localization. So the *c-p-v* matrix was selected to transform the amino acid

Table 3 **Performance comparisons of three models for eukaryotic sequences and polarity values model results**

Model	Sample	Cytoplasmic			Extracellular			Mitochondrial			Nuclear			Overall *Acc* (%)
		θ	σ	*Acc* (%)	θ	σ	*Acc* (%)	θ	σ	*Acc* (%)	θ	σ	*Acc* (%)	
c-p-v	Sample 1	3.0	7.0	82	4.0	6.5	90	3.5	6	88	4.0	4.5	76	84
	Sample 2	6.5	6.5	80	5.0	6.5	90	4.0	6	88	2.0	5.0	75	83
EIIP	Sample 1	3.0	4.5	89	6.0	4.5	88	5.0	4	87	4.5	5.5	70	84
	Sample 2	3.0	4.5	90	5.0	4.5	88	5.0	4	87	5.5	5.5	68	83
hybrid	Sample 1	2.5	7.0	77	4.5	6.0	89	4.5	6	87	3.5	3.5	70	81
	Sample 2	5.0	5.5	78	4.0	5.5	88	4.0	7	88	1.0	3.5	67	80
p	Sample 1	5.0	4.5	81	6.5	3.5	91	3.0	5	88	4.0	3.5	78	84
	Sample 2	4.5	3.5	82	7.0	3.5	91	4.5	5	88	4.5	3.5	77	85

θ is a regularization parameter that controls the trade-off between margin and classification error. σ is the kernel width parameter. *Acc*, accuracy.

Table 4 **Performance comparisons of three models for prokaryotic sequences and polarity values model results**

Model	Sample	Cytoplasmic			Extracellular			Periplasmic			Overall *Acc* (%)
		θ	σ	*Acc* (%)	θ	σ	*Acc* (%)	θ	σ	*Acc* (%)	
c-p-v	Sample 1	5.0	6.5	75	5.5	4.5	90	5.0	7.0	81	82
	Sample 2	3.5	7.5	76	4.5	5.5	90	4.5	7.0	82	83
EIIP	Sample 1	5.0	4.0	71	1.0	0.5	89	4.5	3.5	81	80
	Sample 2	5.0	4.5	71	4.0	2.5	90	3.5	3.5	81	81
hybrid	Sample 1	3.0	5.0	73	5.0	6.0	90	5.0	7.0	81	81
	Sample 2	4.0	6.0	74	3.5	4.5	90	5.0	4.5	81	82
p	Sample 1	5.0	4.5	77	6.5	4.5	90	6.5	3.5	83	83
	Sample 2	6.5	6.5	78	3.0	6.5	90	6.5	4.5	82	83

θ is a regularization parameter that controls the trade-off between margin and classification error. σ is the kernel width parameter. *Acc*, accuracy.

sequences into numerical sequences. The model has also been used in some protein prediction systems [5,24,28], which reflects its popularity in practical applications.

Comparison of the prediction programs

The NNPSL [7], the Markov chain model [11], SubLoc [6], ESLpred [12] and the co-variant discriminant algorithm [10], which were all tested by the Reinhardt and Hubbard dataset, were used for comparisons with our programs. The NNPSL, SubLoc and the co-variant discriminant algorithm are all based on amino acid composition alone, and ESLpred is based on different features such as amino acid composition, dipeptide composition and physico-chemical properties. The prediction results for eukaryotic and prokaryotic sequences are summarized in **Tables 5** and **6**, respectively. The results of the co-variant discrimination, the Markov model, the SubLoc and our method were obtained by the jackknife test. The performance of all the ESLpred modules was evaluated through 5-fold cross-validation, and the neural network method results were evaluated using 6-fold cross-validation. Because all the comparative programs calculated the accuracy, *Acc* results were used for comparison. For eukaryotic sequences in **Table 5**, the overall accuracy of our method was 18%, 11% and 5% higher than that of the NNPSL, Markov chain and SubLoc, respectively and 4% lower than that of the ESLpred. The prediction accuracies for mitochondrial and extracellular sequences were the highest in the five comparative methods.

For prokaryotic sequences (**Table 6**), the overall accuracy of our method was only approximately 3% higher than that of the neural network method and lower than that of the other four methods. But for periplasmic and extracellular localizations of prokaryotic organisms our method had the best performance of 82.1% and 90.4%, respectively, much higher than that of any other method.

The accuracies for mitochondrial and extracellular proteins in eukaryotes and extracellular proteins in prokaryotes are clearly higher than those obtained by other algorithms. Note that the number of these data is small, proving that our method adapts to small samples.

Discussion about optimal parameter selection of the method

FFT is the universal method in signal processing and it is mature in the analysis of many biological signals. We have used FFT, with 512 points of FFT co-efficients, to predict G protein-coupled receptors [24], nuclear receptors [33] and voltage-gated potassium channels [34] on the basis of protein power spectrum, and high accuracies were reached. In our work, the protein power spectrum is need, so the FFT is suitable and used for protein localization. Also 512 points were chosen in the system directly, which simulated the frequency wave of each sequence. So all the different-length signals were changed into equal-length signals with 512 points. Because SVM can only study the equal-length signals, all of the 512 points were put into SVM models, a very important step. Other conventional

Table 5 Performance comparisons for the eukaryotic sequences (%)

	NNPSL	Markov chain	SubLoc	ESLpred	SVM (this work)
Nuclear	55.0	78.1	76.9	95.3	77.0
Cytoplasmic	75.0	62.2	80.0	85.2	81.3
Mitochondrial	61.0	69.2	56.7	68.2	88.7
Extracellular	72.0	74.1	87.4	88.9	90.7
Overall accuracy	66.0	73.0	79.4	88.0	84.4

Table 6 Performance comparisons for the prokaryotic sequences (%)

	NNPSL	Co-variant discriminant	Markov chain	SubLoc	SVM (this work)
Cytoplasmic	80.0	91.6	93.6	97.5	78.3
Periplasmic	85.0	72.3	79.7	78.7	82.1
Extracellular	77.0	80.4	77.6	75.7	90.4
Overall accuracy	81.0	86.5	89.1	91.4	83.6

methods for inversion might only include part information, whereas the FFT analyzes frequency-based information, solves the problem skillfully and lessens the loss of information. This also means that our methods are not limited by the length of sequences. This step can firstly guarantee the improvement of the precision.

The prediction of subcellular localization is a multi-classification problem. In this paper, seven SVMs were constructed for seven class classifications. They were divided into two categories: eukaryotic and prokaryotic species. The *i*th SVM was trained with all samples in the *i*th localization with the label "+1" and all other samples with the label "−1". The SVMs trained in this way were referred to as one-versus-rest SVMs [6]. The kernel function parameters RBF and linear, and the POLY kernel function were all tried; RBF is suitable for FFT representation. The kernel function parameters RBF and optimizer Gradient were selected. All the kernel parameters were kept constant except for θ and σ. The fixed-dimensions vector was obtained based on setting a fixed number of frequency points from the power spectrum of the FFT transformed signals. All the programs of this method were written in Matlab 7.0 programming language.

In the paper, both the jackknife method and two-fold test were used. When different substitution models were compared, the two-fold test was used. Because the dataset was big, the selection of parameters θ and σ was involved, and considering the ability of the computer, the two-fold test was used. When comparing the different prediction programs, parameters θ and σ had been made certain, and most of the compared prediction programs used the jack-knife test. Considering the persuasion of results, the jack-knife test was used.

A new method for protein subcellular localization prediction is established and it can classify protein sequences with high accuracy. FFT reveals more information than other methods and guarantees long sequence prediction. SVM is particularly effective to sequence analysis due to its ability to handle noise, large datasets and large input spaces. The *c-p-v* matrix is used to substitute the sequences and contains more information than the other two substitution models. Research into substitution models will be the important part of future work. If the substitution model well represents the full-scale information of sequences, the prediction performance can be improved dramatically.

References

1 Bult CJ, White O, Olsen GJ, Zhou L, Fleischmann RD, Sutton GG, Blake JA *et al*. Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. Science 1996, 273: 1058–1073

2 Eisenhaber F, Bork P. Wanted: Subcellular localization of proteins based on sequence. Trends Cell Biol 1998, 8: 169–170

3 Eisenhaber F, Bork P. Evaluation of human-readable annotation in biomolecular sequence databases with biological rule libraries. Bioinformatics 1999, 15: 528–535

4 de Trad CH, Fang Q, Cosic I. Protein sequence comparison based on the wavelet transform approach. Protein Eng 2002, 15: 193–203

5 Jiang L, Li M, Wen Z, Wang K, Diao Y. Prediction of mitochondrial proteins using discrete wavelet transform. Protein J 2006, 25: 241–249

6 Hua S, Sun Z. Support vector machine approach for protein subcellular localization prediction. Bioinformatics 2001, 17: 721–728

7 Reinhardt A, Hubbard T. Using neural networks for prediction of the subcellular location of proteins. Nucleic Acids Res 1998, 26: 2230–2236

8 Fujiwara Y, Asogawa M. Prediction of subcellular localizations using amino acid composition and order. Genome Inform 2001, 12: 103–112

9 Emanuelsson O, Nielsen H, Brunak S, von Heijne G. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J Mol Biol 2000, 300: 1005–1016

10 Chou KC, Elrod D. Protein subcellular location prediction. Protein Eng 1999, 12: 107–118

11 Yuan Z. Prediction of protein subcellular locations using Markov chain models. FEBS Lett 1999, 451: 23–26

12 Bhasin M, Raghava GP. ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. Nucleic Acids Res 2004, 32: 414–419

13 Feng ZP. An overview on predicting the subcellular location of a protein. In Silico Biol 2002, 2: 297–303

14 Nakai K, Horton P. PSORT: A program for detecting sorting signals in proteins and predicting their subcellular localization. Trends Biochem Sci 1999, 24: 34–36

15 Nakai K, Kanehisa M. A knowledge base for predicting protein localization sites in eukaryotic cells. Genomics 1992, 14: 897–911

16 Bhasin M, Garg A, Raghava GP. PSLpred: Prediction of subcellular localization of bacterial proteins. Bioinformatics 2005, 21: 2522–2524

17 Bendtsen JD, Nielsen H, von Heijne G, Brunak S. Improved prediction of signal peptides: SignalP 3.0. J Mol Biol 2004, 340: 783 –795

18 Garg A, Bhasin M, Raghava GP. Support vector machine-based method for subcellular localization of human proteins using amino acid compositions, their order, and similarity search. J Biol Chem 2005, 280: 14427–14432

19 Sarda D, Chua GH, Li KB, Krishnan A. pSLIP: SVM based protein subcellular localization prediction using multiple physicochemical properties. BMC Bioinformatics 2005, 6: 152

20 Bairoch A, Boeckmann B. The SWISS-PROT protein sequence data bank, recent developments. Nucleic Acids Res 1993, 21: 3093–3096

21 Rost B, Casadio R, Fariselli P, Sander C. Transmembrane helices predicted at 95% accuracy. Protein Sci 1995, 4: 521–533

22 Vogt G, Etzold T, Argos P. An assessment of amino acid exchange matrices in aligning protein sequences: The twilight zone revisited. J Mol Biol 1995, 249: 816–831

23 Grantham R. Amino acid difference formula to help explain protein evolution. Science 1974, 185: 862–864

24 Wang KL, Wen ZN, Nie FS, Li ML. A new hybrid model of amino acid substitution for protein functional classification. Chin Chem Lett 2005, 16: 1133–1136

25 Fauchere JL, Pliska V. Hydrophobic parameters of pi amino-acid side chains from the partitioning of N-acetyl-amino-acid amides. Eur J Med Chem 1983, 18: 369–375

26 O'Neil KT, DeGrado WF. A thermodynamic scale for the helix-forming tendencies of the commonly occurring amino acids. Science 1990, 250: 646–651

27 Harris CM. The Fourier analysis of biological transients. J Neurosci Methods 1998, 83: 15–34

28 Guo YZ, Li ML, Wang KL,Wen ZN, Lu MC, Liu LX, Lin J. Fast Fourier transform-based support vector machine for prediction of G-protein coupled receptor subfamilies. Acta Biochim Biophys Sin 2005, 37: 759–766

29 Haykin S. Neural Networks: A Comprehensive Foundation. 2nd ed. New Jersey: Prentice Hall 1999

30 Vapnik V. Support Vector Machines of Pattern Recognition. Statistical Learning Theory. Peking: Publishing House of Electronics Industry 2004

31 Zavaljevski N, Stevens FJ, Reifman J. Support vector machines with selective kernel scaling for protein classification and identification of key amino acid positions. Bioinformatics 2002, 18: 689–696

32 Kohavi R, Provost F. Glossary. Mach. Learning J 1998, 30: 271–274

33 Guo YZ, Li M, Lu M, Wen Z, Wang K, Li G, Wu J. Classifying G protein-coupled receptors and nuclear receptors on the basis of protein power spectrum from fast Fourier transform. Amino Acids 2006, 30: 397–402

34 Liu LX, Li ML, Tan FY, Lu MC, Wang KL, Wen ZN, Guo YZ *et al.* Local sequence information based support vector machine to classify voltage-gated potassium channels. Acta Biochim Biophys Sin 2006, 38: 363–371

Edited by
Minoru ASOGAWA