**ISSN** 1672-9145 **Acta Biochimica et Biophysica Sinica** 2005, **37**(11): 759–766 **CN** 31-1940/Q

# Fast Fourier Transform-based Support Vector Machine for Prediction of G-protein Coupled Receptor Subfamilies

Yan-Zhi GUO, Meng-Long LI*, Ke-Long WANG, Zhi-Ning WEN, Min-Chun LU, Li-Xia LIU, and Lin JIANG

*College of Chemistry, Sichuan University, Chengdu 610064, China*

**Abstract** Although the sequence information on G-protein coupled receptors (GPCRs) continues to grow, many GPCRs remain orphaned (i.e. ligand specificity unknown) or poorly characterized with little structural information available, so an automated and reliable method is badly needed to facilitate the identification of novel receptors. In this study, a method of fast Fourier transform-based support vector machine has been developed for predicting GPCR subfamilies according to protein's hydrophobicity. In classifying Class B, C, D and F subfamilies, the method achieved an overall Matthew's correlation coefficient and accuracy of 0.95 and 93.3%, respectively, when evaluated using the jackknife test. The method achieved an accuracy of 100% on the Class B independent dataset. The results show that this method can classify GPCR subfamilies as well as their functional classification with high accuracy. A web server implementing the prediction is available at http://chem.scu.edu.cn/blast/Pred-GPCR.

**Key words** G-protein coupled receptor; subfamily; fast Fourier transform; support vector machine; prediction

G-protein coupled receptors (GPCRs) constitute a superfamily of cell surface receptor proteins characterized by seven transmembrane segments. The N-terminus is always located extracellularly and the C-terminus extends into the cytoplasm, which makes these proteins capable of transducing signals into the cell by the heterotrimeric G-protein [1]. GPCRs play a key role in cellular signaling networks that regulate various basic physiological processes, such as neurotransmission, cell metabolism, secretion, cell differentiation and growth, inflammatory and immune responses, smell, taste and vision [2]. More than 50% of drugs now available on the market act through GPCRs [3]. Although there have been methods developed to build the structural models of GPCRs [4,5], the structure of only one GPCR, bovine rhodopsin, has been solved experimentally.

The identification of novel GPCRs will greatly facilitate the target validation process and automatically provide a possible compound-screening assay [6]. In the past, many strategies have been used to identify novel GPCRs. The simplest and most frequently used method is to search a sequence database using sequence alignment tools, such as BLAST and FASTA [7–9]. Several pattern databases (e.g. PRINTS) have been built [10,11]. However, they are not always successful when the query proteins have no significant sequence similarity to the sequences in the database. The Pfam classifier based on the profile-hidden Markov model has been developed [12–15], but on the class level. To overcome these limitations, the support vector machine (SVM)-based methods have been used to classify the families and subfamilies, even sub-subfamilies, of GPCRs [3,16,17]. Another method using binary topology pattern has also been used to identify eukaryotic GPCRs [18].

The main goal of this work is to develop a method to determine GPCRs' function at the subfamily level. A new method was developed for classifying subfamilies belonging to Class B, C, D and F GPCRs. This method couples fast Fourier transform (FFT) with SVM on the basis of the hydrophobicity of amino acid sequences. The performance

Received: April 11, 2005 Accepted: August 24, 2005

*Corresponding author: Tel, 86-28-89005151; Fax, 86-28-85412356; E-mail, liml@scu.edu.cn

DOI: 10.1111/j.1745-7270.2005.00110.x

of this method was validated by the jackknife test and evaluated by the independent dataset test.

## Methods

### Dataset

To collect the sequences used for this study, all of the sequences belonging to Class B, C, D and F in GPCRDB (March 2005 release 9.0) (http://www.gpcr.org/7tm/) [19] were picked out, then all orphan/putative sequences and fragments were removed. None of the sequences was identical to another in the dataset. The subfamilies that contained less than 10 sequences were dropped out. For Class B GPCRs, the sequences marked as "new" were excluded as the independent dataset; and because the other three classes were relatively small, all the eligible sequences were used. The final dataset contained 403 sequences belonging to 17 different subfamilies. The number of sequences for each different subfamily is listed in **Table 1**.

**Table 1 Number of sequences belonging to each G-protein coupled receptor (GPCR) subfamily**

| Class | GPCR subfamily | *n* |
|---|---|---|
| Class B | Calcitonin | 20 |
| | Corticotropin releasing factor | 23 |
| | Glucagon | 12 |
| | Growth hormone-releasing hormone | 13 |
| | Parathyroid hormone | 17 |
| | PACAP | 11 |
| | Vasoactive intestinal polypeptide | 14 |
| | Latrophilin | 20 |
| | Methuselah-like proteins | 21 |
| Class C | Metabotropic glutamate | 46 |
| | Calcium-sensing like | 18 |
| | GABA-B | 23 |
| | Taste receptors | 12 |
| Class D | Fungal pheromone A-factor like | 16 |
| | Fungal pheromone B like | 32 |
| Class F | Frizzled | 94 |
| | Smoothened | 11 |

*n*, number of sequences.

### Quantitative description of proteins

The quantitative description of amino acid sequences is crucial. Here, three principal properties of proteins, the hydrophobicity, bulk and electronic property, were taken into account. The hydrophobicity model, *c*-*p*-*v* model [20] and electron-ion interaction potential (EIIP) model [21] were selected. The hydrophobicity determines the structure and function of proteins, especially for the transmembrane proteins. Three different hydrophobicity scales, KDHΦ [22], MHΦ [23] and FHΦ [24], were selected and optimized. The *c*-*p*-*v* model includes the composition (*c*), polarity (*p*) and molecular volume (*v*). The EIIP model describes the average energy states of all valence electrons of amino acid sequences. These numerical series are normalized to zero mean and unit standard deviation, as defined in **Equation 1**:

$$x_{ij}' = \frac{x_{ij} - \overline{x}_j}{s_j} \qquad 1$$

where $x_{ij}$ is some property value of the *i*th amino acid residue in the *j*th sequence, $\overline{x}_j$ is the mean property value of the *j*th sequence, and $s_j$ is the standard deviation of the *j*th sequence.

### FFT

The Fourier transform changes the signal from time-based to frequency-based, as shown in **Equation 2**:

$$\hat{f}(\omega) = \int_{-\infty}^{+\infty} f(t) e^{-i\omega t} dt \qquad 2$$

FFT has been applied to protein sequence comparison [25] and rapid multiple sequence alignment [26]. FFT is defined in **Equation 3**:

$$X(k) = \sum_{j=1}^{N} x(j) \omega_N^{(j-1)(k-1)} \qquad 3$$

where $\omega = e^{(-2\pi i)/N}$ is an *N*th root of Unity. *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. A plot of power versus frequency is called the power spectrum or power spectral density. The power at each frequency point was taken as the input feature of SVMs. The numerical sequences of variable lengths are transformed to fixed length vectors in this way.

### SVM

SVM [27,28] is a kind of learning machine based on statistical learning theory. The most attractive characteristics of SVM are the absence of local minima, the sparseness of the solution, and the use of the kernel-induced feature spaces. The SVM training process always seeks a global optimized solution and avoids over-fitting, so it has the ability to handle a large number of features and a relatively small dataset.

The basic ideas behind SVM can be introduced as follows. For a two-class problem, there are a series of samples described by the feature vectors $x_i(i=1,2,\ldots,l)$ (**Equation 4**) with corresponding labels $y_i=\{+1,-1\}(i=1, 2,\ldots,l)$ (**Equation 5**). To classify the two classes of samples, SVM maps the input vectors into a higher dimensional feature space, then constructs the maximal margin hyperplane (MMH), which maximizes the distance of the closest vectors belonging to the two classes to the hyperplane. The MMH can be obtained by solving the following convex quadratic programming problem:

$$\text{Maximize } \sum_{i=1}^{l} a_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} a_i a_j y_i y_j K(x_i, x_j) \qquad 4$$

$$\text{subject to } \sum_{i=1}^{l} a_i y_i = 0\,, \qquad 0 \le a_i \le C \qquad 5$$

where $C$ is a regularization parameter that controls the trade-off between the margin and classification error.

$K(x_i,x_j)$ is the kernel function. In this paper, the radial basis function was selected as the kernel function (**Equation 6**):

$$K(x_i, x) = \exp\left\{-\frac{1}{2\sigma^2}\left\|x - x_i\right\|^2\right\} \qquad 6$$

where $\sigma$ is the kernel width parameter.

The decision function implemented by SVM can be written as **Equation 7**:

$$f(x) = Sgn\left\{\sum_{i=1}^{l} y_i a_i K(x_i, x) = b\right\} \qquad 7$$

The prediction of GPCR subfamilies is a multi-class classification problem. In this paper, $n$ SVMs were constructed for $n$-class classification. The $i$th SVM was trained with all samples in the $i$th subfamily 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 [29]. All the kernel parameters were kept constant except for $C$ and $\sigma$. All the programs of this method were written in Matlab 7.0 programming language.

### Performance evaluation

The models for all the subfamilies were validated by the jackknife test. For cross-validation, the jackknife test is deemed more effective and objective than the independent dataset test and sub-sampling test [30,31]. Chou and Zhang [32] have given a comprehensive discussion, and Mardia *et al.* [33] has explained the mathematical principle behind it. During the process of jackknifing, each receptor was singled out in turn as a test receptor with the remaining receptors used to train SVM.

Four indices, the accuracy (ACC) (**Equation 8**), Matthew's correlation coefficient (MCC) (**Equation 9**) [34], total ACC (**Equation 10**) and total MCC (**Equation 11**), were calculated for the assessment of the prediction system.

$$ACC(i) = \frac{p(i)}{\exp(i)} \qquad 8$$

$$MCC(i) = \frac{p(i)n(i) - u(i)o(i)}{\sqrt{[p(i)+u(i)][p(i)+o(i)][n(i)+u(i)][n(i)+o(i)]}} \qquad 9$$

$$ACC\text{total} = \frac{\sum_{i=1}^{k} p(i)}{N} \qquad 10$$

$$MCC\text{total} = \frac{\sum_{i=1}^{k} \exp(i) MCC(i)}{N} \qquad 11$$

Here, $i$ is the any subfamily, $N$ is the total number of sequences, $k$ is the subfamily number, exp($i$) is the number of sequences observed in subfamily $i$, $p(i)$ is the number of correctly predicted sequences of subfamily $i$, $n(i)$ is the number of correctly predicted sequences not of subfamily $i$, $u(i)$ is the number of under-predicted sequences, and $o(i)$ is the number of over-predicted sequences.

## Results and Discussion

### Selecting principal property for SVMs with the best performance

FHΦ, one of the hydrophobicity scales, was used in the hydrophobicity model. The hydrophobicity, *c*-*p*-*v* and EIIP models transformed the amino acid sequences into numerical sequences separately, which were then transformed to input feature vectors using FFT of 512 frequency points for SVMs. The performance of SVMs based on the three models was validated using the jackknife test, as shown in **Table 2**. **Table 2** shows that the performance based on the hydrophobicity model (FHΦ) is better than that based on the *c*-*p*-*v* model or EIIP model, achieving the highest total ACC and MCC of 91.6% and 0.94, respectively. The results indicate that hydrophobicity is the most important property of proteins and can preferably substitute the amino acid sequences quantitatively.

### Selecting input feature vectors for SVMs from the FFT transformed signals

The numerical sequences based on the hydrophobicity model with FHΦ as the scale were transformed with FFT to the input feature vectors for SVMs in three ways: (1)

**Table 2 Performance of support vector machines based on the hydrophobicity model (FHΦ), composition, polarity and molecular volume (*c-p-v*) model or electron-ion interaction potential (EIIP) model respectively, using fast Fourier transform of 512 frequency points, as validated by the jackknife test**

| Class | GPCR subfamily | Hydrophobicity model* | | *c-p-v* model | | EIIP model | |
|---|---|---|---|---|---|---|---|
| | | ACC | MCC | ACC | MCC | ACC | MCC |
| Class B | Calcitonin | 95.0% | 0.97 | 85.0% | 0.91 | 95.0% | 0.97 |
| | Corticotropin releasing factor | 100.0% | 1.00 | 95.7% | 0.97 | 95.7% | 0.97 |
| | Glucagon | 91.7% | 0.95 | 91.7% | 0.95 | 58.3% | 0.75 |
| | Growth hormone-releasing hormone | 84.6% | 0.91 | 76.9% | 0.87 | 69.2% | 0.82 |
| | Parathyroid hormone | 76.5% | 0.86 | 58.8% | 0.75 | 52.9% | 0.71 |
| | PACAP | 90.9% | 0.95 | 81.8% | 0.90 | 90.9% | 0.95 |
| | Vasoactive intestinal polypeptide | 85.7% | 0.92 | 71.4% | 0.83 | 57.1% | 0.74 |
| | Latrophilin | 100.0% | 1.00 | 95.0% | 0.97 | 95.0% | 0.97 |
| | Methuselah-like proteins | 61.9% | 0.76 | 57.1% | 0.73 | 47.6% | 0.66 |
| | Total | 87.4% | 0.92 | 80.1% | 0.88 | 75.5% | 0.84 |
| Class C | Metabotropic glutamate | 91.3% | 0.92 | 82.6% | 0.85 | 91.3% | 0.92 |
| | Calcium-sensing like | 66.7% | 0.79 | 61.1% | 0.75 | 61.1% | 0.75 |
| | GABA-B | 95.7% | 0.97 | 65.2% | 0.77 | 65.2% | 0.77 |
| | Taste receptors | 91.7% | 0.95 | 91.7% | 0.95 | 66.7% | 0.80 |
| | Total | 87.8% | 0.91 | 75.8% | 0.83 | 76.8% | 0.84 |
| Class D | Fungal pheromone A-Factor like | 87.5% | 0.91 | 93.8% | 0.95 | 50.0% | 0.63 |
| | Fungal pheromone B like | 100.0% | 1.00 | 100.0% | 1.00 | 100.0% | 1.00 |
| | Total | 95.8% | 0.97 | 97.9% | 0.98 | 83.3% | 0.88 |
| Class F | Frizzled | 100.0% | 1.00 | 100.0% | 1.00 | 100.0% | 1.00 |
| | Smoothened | 90.9% | 0.95 | 90.9% | 0.95 | 90.9% | 0.95 |
| | Total | 99.0% | 0.99 | 99.0% | 0.99 | 99.0% | 0.99 |
| Total | | 91.6% | 0.94 | 86.0% | 0.91 | 82.7% | 0.88 |

For each subfamily, all the negative samples were correctly predicted. ACC, accuracy; GPCR, G-protein coupled receptors; MCC, Matthew's correlation coefficient.
* FHΦ scale was used.

using all the 512 frequency points; (2) using 256 odd frequency points extracted from the FFT signals; and (3) using 256 even frequency points extracted from the FFT signals. The performances of the three groups of feature vectors were compared using the jackknife test, as shown in **Table 3**. **Table 3** shows that adopting 256 even frequency points as input vectors can get the highest overall ACC and MCC. For each GPCR subfamily, the total ACC of adopting 256 even frequency points is equal or higher than that of adopting the other two groups of frequency points. So, in the later experiments, the input feature vectors of SVMs were the 256 even frequency points for all the subfamilies.

## Selecting one hydrophobicity scale with the best performance

Adopting the 256 even frequency points as the feature vectors for each subfamily, the performances of SVMs based on the different hydrophobicity scales were compared using the jackknife test. The results are listed in **Table 4**. From **Table 4**, we can see that the performance of SVMs based on KDHΦ and MHΦ is not as good as that based on FHΦ. The SVM based on FHΦ achieves the highest total ACC and MCC of 93.3% and 0.95, respectively. This method can classify Class B with a total ACC of 90.7%; Class C, 87.9%; Class D, 95.8%; and Class F, 100%. The results prove that SVM based on FHΦ can classify GPCR subfamilies with the highest accuracy.

## Assigning a reliability index to the prediction

It is important to know the prediction reliability when using machine-learning techniques to assign subfamilies of GPCRs. The reliability index (*RI*) was assigned according to the difference (noted as *diff*) between the highest and the second-highest output score of SVMs in a multi-

**Table 3 Performance of support vector machines based on the hydrophobicity model (FHΦ) adopting 512, 256 odd or 256 even frequency points respectively, as validated by the jackknife test**

| Class | GPCR subfamily | Frequency points used | | | | | |
|---|---|---|---|---|---|---|---|
| | | 512 points | | 256 odd points | | 256 even points | |
| | | ACC | MCC | ACC | MCC | ACC | MCC |
| Class B | Calcitonin | 95.0% | 0.97 | 95.0% | 0.97 | 95.0% | 0.97 |
| | Corticotropin releasing factor | 100.0% | 1.00 | 100.0% | 1.00 | 100.0% | 1.00 |
| | Glucagon | 91.7% | 0.95 | 100.0% | 1.00 | 100.0% | 1.00 |
| | Growth hormone-releasing hormone | 84.6% | 0.91 | 84.6% | 0.91 | 84.6% | 0.91 |
| | Parathyroid hormone | 76.5% | 0.86 | 82.4% | 0.90 | 82.4% | 0.90 |
| | PACAP | 90.9% | 0.95 | 90.9% | 0.95 | 90.9% | 0.95 |
| | Vasoactive intestinal polypeptide | 85.7% | 0.92 | 85.7% | 0.92 | 85.7% | 0.92 |
| | Latrophilin | 100.0% | 1.00 | 95.0% | 0.97 | 100.0% | 1.00 |
| | Methuselah-like proteins | 61.9% | 0.76 | 61.9% | 0.76 | 71.4% | 0.83 |
| | Total | 87.4% | 0.92 | 88.1% | 0.93 | 90.7% | 0.94 |
| Class C | Metabotropic glutamate | 91.3% | 0.92 | 93.5% | 0.94 | 93.5% | 0.94 |
| | Calcium-sensing like | 66.7% | 0.79 | 66.7% | 0.79 | 66.7% | 0.79 |
| | GABA-B | 95.7% | 0.97 | 87.0% | 0.91 | 91.3% | 0.94 |
| | Taste receptors | 91.7% | 0.95 | 91.7% | 0.95 | 91.7% | 0.95 |
| | Total | 87.9% | 0.91 | 86.9% | 0.91 | 87.9% | 0.91 |
| Class D | Fungal pheromone A-Factor like | 87.5% | 0.91 | 81.3% | 0.86 | 87.5% | 0.91 |
| | Fungal pheromone B like | 100.0% | 1.00 | 100.0% | 1.00 | 100.0% | 1.00 |
| | Total | 95.8% | 0.97 | 93.8% | 0.95 | 95.8% | 0.97 |
| Class F | Frizzled | 100.0% | 1.00 | 100.0% | 1.00 | 100.0% | 1.00 |
| | Smoothened | 90.9% | 0.95 | 90.9% | 0.95 | 100.0% | 1.00 |
| | Total | 99.0% | 0.99 | 99.0% | 0.99 | 100.0% | 1.00 |
| Total | | 91.6% | 0.94 | 91.3% | 0.94 | 93.3% | 0.95 |

For each subfamily, all the negative samples were correctly predicted. ACC, accuracy; GPCR, G-protein coupled receptors; MCC, Matthew's correlation coefficient.

class classification [3,27]. The reliability score in this work has been computed using **Equation 12**:

$$RI = \begin{cases} INT(\frac{diff \times 5}{2}) + 1 & ,0 \le diff < 2.0 \\ 5 & ,diff \ge 2.0 \end{cases} \qquad 12$$

The expected prediction accuracy and the number of sequences for each given *RI* were calculated, as shown in **Fig. 1**. **Fig. 1** shows that the model predicted 81.9% (330/403) sequences with *RI*≥5. Three hundred and thirty sequences (*RI*≥5) were nearly 100% correctly predicted, and only one sequence is under-predicted (the accuracy is 329/330=99.7%). These results suggest that our model can predict GPCR subfamilies with high reliability.

## Independent dataset test

Because Class C, D and F have fewer members than Class B, all proteins of Class C, D and F were used for the training dataset. As described in "Methods", the sequences marked as "new" in Class B in GPCRDB (March 2005 release 9.0) that do not exist in the training set were used to check the practical application of this method as the independent dataset. There are three receptors for the calcitonin subfamily, three for the corticotropin-releasing factor subfamily, two for the glucagon subfamily, two for the parathyroid hormone subfamily, four for the PACAP subfamily, three for the vasoactive intestinal polypeptide subfamily and ten for the latrophilin subfamily. The overall ACC is 100%, which shows the method's strong facility for practical application.

## Comparison with SVM based on amino acid composition and dipeptide composition

Artificial intelligence-based techniques such as SVM and the neural network require a fixed number of inputs for training, so it is necessary to find a strategy for

**Table 4 Performance of support vector machines based on different hydrophobicity scales using 256 even frequency points, as validated by the jackknife test**

| Class | GPCR subfamily | KDHΦ | | MHΦ | | FHΦ | |
|---|---|---|---|---|---|---|---|
| | | ACC | MCC | ACC | MCC | ACC | MCC |
| Class B | Calcitonin | 95.0% | 0.97 | 90.0% | 0.94 | 95.0% | 0.97 |
| | Corticotropin releasing factor | 100.0% | 1.00 | 100.0% | 1.00 | 100.0% | 1.00 |
| | Glucagon | 83.3% | 0.91 | 91.7% | 0.95 | 100.0% | 1.00 |
| | Growth hormone-releasing hormone | 84.6% | 0.91 | 76.9% | 0.87 | 84.6% | 0.91 |
| | Parathyroid hormone | 76.5% | 0.86 | 76.5% | 0.86 | 82.4% | 0.90 |
| | PACAP | 90.9% | 0.95 | 90.9% | 0.95 | 90.9% | 0.95 |
| | Vasoactive intestinal polypeptide | 71.4% | 0.83 | 92.9% | 0.96 | 85.7% | 0.92 |
| | Latrophilin | 100.0% | 1.00 | 100.0% | 1.00 | 100.0% | 1.00 |
| | Methuselah-like proteins | 66.7% | 0.80 | 66.7% | 0.80 | 71.4% | 0.83 |
| | Total | 86.1% | 0.92 | 87.4% | 0.93 | 90.7% | 0.94 |
| Class C | Metabotropic glutamate | 95.7% | 0.96 | 91.3% | 0.92 | 93.5% | 0.94 |
| | Calcium-sensing like | 66.7% | 0.79 | 66.7% | 0.79 | 66.7% | 0.79 |
| | GABA-B | 8.7% | 0.26 | 8.7% | 0.26 | 91.3% | 0.94 |
| | Taste receptors | 91.7% | 0.95 | 91.7% | 0.95 | 91.7% | 0.95 |
| | Total | 69.7% | 0.77 | 67.7% | 0.75 | 87.9% | 0.91 |
| Class D | Fungal pheromone A-Factor like | 87.5% | 0.91 | 68.8% | 0.77 | 87.5% | 0.91 |
| | Fungal pheromone B like | 100.0% | 1.00 | 100.0% | 1.00 | 100.0% | 1.00 |
| | Total | 95.8% | 0.97 | 89.6% | 0.92 | 95.8% | 0.97 |
| Class F | Frizzled | 100.0% | 1.00 | 100.0% | 1.00 | 100.0% | 1.00 |
| | Smoothened | 100.0% | 1.00 | 100.0% | 1.00 | 100.0% | 1.00 |
| | Total | 100.0% | 1.00 | 100.0% | 1.00 | 100.0% | 1.00 |
| Total | | 86.9% | 0.91 | 86.4% | 0.90 | 93.3% | 0.95 |

For each subfamily, all the negative samples were correctly predicted. ACC, accuracy; GPCR, G-protein coupled receptor; MCC, Matthew's correlation coefficient.

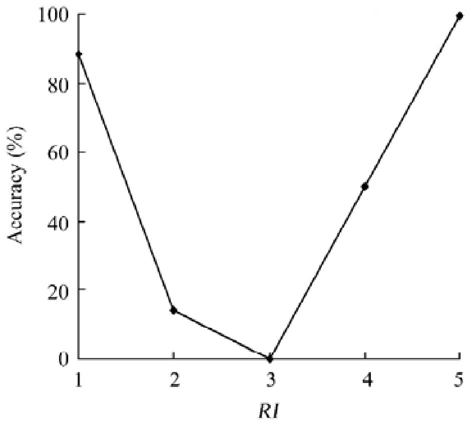

**Fig. 1 Expected prediction accuracy with a given reliability index (*RI*)**

transforming the variable lengths of proteins to fixed length patterns. Amino acid composition [3,29,33] and dipeptide composition [3,35,36] have been used widely in bioinformatics using intelligence techniques. The amino acid and dipeptide compositions of proteins can be used to encapsulate the protein information in a vector of 20 and 400 dimensions, respectively. It is worth comparing our method with the SVM based on amino acid composition and dipeptide composition respectively. But the performances of the SVMs based on the two approaches with the jackknife test indicate that neither of the two approaches can discriminate any subfamily successfully. It may be because Class B, C, D and F do not have enough samples, so that there is no statistical meaning for each subfamily using the two statistical approaches. However, our method can classify GPCR subfamilies with good performance, indicating its powerful ability to deal with relatively small datasets.

## Prediction Web Server

Based on this study, a web server has been set up to allow users to recognize GPCR subfamilies. It is freely available at http://chem.scu.edu.cn/blast/Pred-GPCR. Users can submit the query sequence in the standard format of FASTA. After analysis, the results will show the GPCR subfamily to which the query sequence belongs.

## Conclusion

This paper describes a method of SVM in combination with FFT to classify GPCR subfamilies. During the development of SVMs, three principal properties, hydrophobicity, bulk and electronic property, were compared. The results show that the hydrophobicity of proteins is the most important property in deciding GPCRs' function. Three hydrophobicity scales, KDHΦ, MHΦ and FHΦ, were optimized, and the performance based on FHΦ was best. It was indicated that taking 256 even frequency points of FFT transformed signals as input vectors could achieve the highest accuracy. From these results, it is obvious that the substitution model can affect the performance of the method. We think that the performance of this method will be improved further if a more suitable substitution model is found. We expect to find a hybrid model (related to hydrophobicity) that can integrate other important properties in a reasonable way, rather than using hydrophobicity alone. The establishment of such methods will facilitate drug discovery for many diseases.

## Acknowledgement

This work was supported by the State Key Laboratory of Chemo/Biosensing and Chemometrics, Hunan University (Changsha, China).

## References

1 Attwood TK, Croning MDR, Gaulton A. Deriving structural and functional insights from a ligand-based hierarchical classification of G protein-coupled receptors. Protein Eng 2002, 15: 7–12

2 Hebert TE, Bouvier M. Structural and functional aspects of G protein-coupled receptor oligomerization. Biochem Cell Biol 1998, 76: 1–11

3 Bhasin M, Raghava GPS. GPCRpred: An SVM-based method for prediction of families and subfamilies of G-protein coupled receptors. Nucleic Acids Res 2004, 32: W383–W389

4 Yin YB, Luo JC, Jiang Y. Advances in G-protein coupled receptor research and related bioinformatics study. Chin Sci Bull 2003, 48: 511–516

5 Huang XQ, Jiang HL, Luo XM, Chen KX, Zhu YC, Ji RY, Cao Y. Comparative molecular modeling on 3D-structure of opioid receptor-like 1 receptor. Acta Pharmacol Sin 2000, 21: 529–535

6 Takeshi H, Wataru N,Takeshi K, Norihisa F. Construction of hypothetical three-dimensional structure of P2Y1 receptor based on Fourier transform analysis. J Protein Chem 2002, 21: 537–545

7 Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990, 215: 403–410

8 Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res 1997, 25: 3389–3402

9 Pearson WR. Flexible sequence similarity searching with the FASTA3 program package. Methods Mol Biol 2000, 132: 185–219

10 Lapinsh M, Gutcaits A, Prusis P, Post C, Lundstedt T, Wikberg JES. Classification of G-protein coupled receptors by alignment-independent extraction of principal chemical properties of primary amino acid sequences. Protein Sci 2002, 11: 795–805

11 Sadowski MI, Parish JH. Automated generation and refinement of protein signatures: case study with G-protein coupled receptors. Bioinformatics 2003, 19: 727–734

12 Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S *et al*. The Pfam protein families database. Nucleic Acids Res 2002, 30: 276–280

13 Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A *et al*. The Pfam protein families database. Nucleic Acids Res 2004, 32: D138–D141

14 Papasaikas PK, Bagos PG, Litou ZI, Promponas VJ, Hamodrakas SJ. PRED-GPCR: GPCR recognition and family classification server. Nucleic Acids Res 2004, 32: W380–W382

15 Papasaikas PK, Bagos PG, Litou ZI, Hamodrakas SJ. A novel method for GPCR recognition and family classification from sequence alone using signatures derived from profile hidden Markov models. SAR QSAR Environ Res 2003, 14: 413–420

16 Karchin R, Karplus K, Haussler D. Classifying G-protein coupled receptors with support vector machines. Bioinformatics 2002, 18: 147–159

17 Huang Y, Cai J, Ji L, Li YD. Classifying G-protein coupled receptors with bagging classification tree. Comput Biol Chem 2004, 28: 275–280

18 Inoue Y, Ikeda M, Shimizu T. Proteome-wide functional classification and identification of mammalian-type GPCRs by binary topology pattern. Comput Biol Chem 2004, 28: 39–49

19 Horn F, Vriend G, Cohen FE. Collecting and harvesting biological data: The GPCRDB and NucleaRDB information systems. Nucleic Acids Res 2001, 29: 346–349

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

21 Cosic I. Macromolecular bioactivity: Is it resonant interaction between macromolecules? Theory and applications. IEEE Trans Biomed Eng 1994, 41: 1101–1114

22 Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. J Mol Biol 1982, 157: 105–132

23 Mandell AJ, Selz KA, Shlesinger MF. Wavelet transformation of protein hydrophobicity sequences suggests their memberships in structural families. Physica A 1997, 244: 254–262

24 Fauchére J, Pliška V. Hydrophobic parameters Φ of amino-acid side chains from the partitioning of n-acetyl-amino-acid amides. Eur J Med Chem Chim Ther 1983, 18: 369–375

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

26 Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 2002, 30: 3059–3066

27 Haykin S. Support vector machines. In: Haykin S ed. Neural Networks: A

Comprehensive Foundation. 2nd ed. New York: Prentice Hall Inc. 1999

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

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

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

31 Elrod DW, Chou KC. A study on the correlation of G-protein-coupled receptor types with amino acid composition. Protein Eng 2002, 15: 713–715

32 Chou KC, Zhang CT. Prediction of protein structural classes. Crit Rev Biochem Mol 1995, 30: 275–349

33 Mardia KV, Kent JT, Bibby JM eds. Multivariate Analysis. London: Academic Press 1979

34 Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta 1975, 405: 442–451

35 Bhasin M, Raghava GPS. Classification of nuclear receptors based on amino acid composition and dipeptide composition. J Biol Chem 2004, 279: 23262–23266

36 Reczko M, Bohr H. The DEF data base of sequence based protein fold class predictions. Nucleic Acids Res 1994, 22: 3616–3619

Edited by
**Lu-Hua LAI**