Classifying G protein-coupled receptors and nuclear receptors on the basis of protein power spectrum from fast Fourier transform

Y.-Z. Guo\textsuperscript{1,2}, M. Li\textsuperscript{1,2}, M. Lu\textsuperscript{1}, Z. Wen\textsuperscript{1,2}, K. Wang\textsuperscript{1}, G. Li\textsuperscript{1}, and J. Wu\textsuperscript{1}

\textsuperscript{1} College of Chemistry, Sichuan University, Chengdu, China
\textsuperscript{2} State Key Laboratory of Chemo/Biosensing and Chemometrics, Hunan University, Changsha, China

Summary. As the potential drug targets, G-protein coupled receptors (GPCRs) and nuclear receptors (NRs) are the focuses in pharmaceutical research. It is of great practical significance to develop an automated and reliable method to facilitate the identification of novel receptors. In this study, a method of fast Fourier transform-based support vector machine was proposed to classify GPCRs and NRs from the hydrophobicity of proteins. The models for all the GPCR families and NR subfamilies were trained and validated using jackknife test and the results thus obtained are quite promising. Meanwhile, the performance of the method was evaluated on GPCR and NR independent datasets with good performance. The good results indicate the applicability of the method. Two web servers implementing the prediction are available at http://chem.scu.edu.cn/Pred-GPCR and http://chem.scu.edu.cn/Pred-NR.

Keywords: G-protein coupled receptors – Nuclear receptors – Hydrophobicity – Fast Fourier transform – Power spectrum – Support vector machine

Introduction

G-protein coupled receptors (GPCRs) belong to the largest superfamily of cell-surface receptors and they are characterized by seven transmembrane segments. They play a key role in the basic cellular processes such as vision, smell, taste, neurotransmission, and metabolism and so on. They are major therapeutic targets of numerous prescribed drugs and more than 50% of all medicines available today act through GPCR (Guemermann et al., 1995). The sequences of thousands of GPCRs have been known (Horn et al., 2003), however many receptors remain orphaned (i.e. with unknown ligand specificity), and to date the crystal structure of only one GPCR (bovine rhodopsin) is solved (Palczewski et al., 2000). So it is highly desirable to develop the computational methods to facilitate the identification and characterization of novel receptors only using sequence information.

Methods have been developed to predict GPCRs. The covariant discriminant algorithm was proposed to predict GPCRs (Chou, 2005a; Chou and Elrod, 2002; Elrod and Chou, 2002), and support vector machines (SVMs) were used to classify GPCRs at family and subfamily level (Bhasin and Raghava, 2004a; Karchin et al., 2002). The methods based on profile-hidden Markov model (HMM) have been developed (Bateman et al., 2004; Papasaikas et al., 2004; Qian et al., 2003), and there has been a method of bagging classification tree for the classification of GPCRs (Huang et al., 2004).

Nuclear receptors (NRs) are important transcription factors involved in many physiological functions like cell growth, differentiation and homeostasis (Gronemeyer and Laudet, 1995; Mangelsdorf et al., 1995). Many of them are important drug targets in designing drugs for diseases such as breast cancer and diabetes (Robinson-Rechavi and Laude, 2003). The simple similarity-based search tools like BLAST and FASTA (Altschul et al., 1990; Pearson and Lipman, 1988) can easily distinguish NRs from the genome sequences, but they are not always successful in classifying the subfamilies of NRs. To overcome this limitation, a SVM based method (Bhasin and Raghava, 2004b) has been developed for NR subfamily classification, but only four subfamilies.

Based on the concept of pseudo amino acid composition (Chou, 2001), the Fourier transform spectra has been used to predict membrane protein type (Liu et al., 2005a; Wang et al., 2004), particularly their low-frequency parts.
(Chou, 1988), have been used to predict membrane protein types. This paper describes a new combination of fast Fourier transform with support vector machine for the classification of all GPCR classes and NR subfamilies based on the hydrophobicity of proteins. The similar method has been successfully used for the prediction of GPCR subfamilies (Guo et al., 2005). The primary amino acid sequences are translated into numerical sequences using the hydrophobicity and then the numerical series are transformed into uniform matrix according to fast Fourier transform. Last, taking the protein power spectrum as input, SVM is used to construct classifiers.

Materials and methods

Data set

On the basis of pharmacological knowledge, the GPCRDB and NucleARDB information systems (Horn et al., 2001) classify GPCRs into six different families and NRs into eight subfamilies respectively. The sequence data of GPCRs were collected from GPCRDB (release 9.0, March 2005) and the data of NRs were obtained from NucleARDB (release 5.0, April 2005). All sequences denoted as ‘putative’, ‘hypothetical’ or ‘orphan’ and fragmental sequences were removed. Meanwhile, it was assured that none of the sequences was identical to others. Next, all sequences are partitioned into two parts, the training dataset and the test dataset. The newly publicized sequences that are marked as ‘new’ in the two databases were used as the independent dataset. All the remaining sequences were used as the training dataset. The final training dataset contained 946 sequences belonging to the six GPCR families and 465 sequences belonging to the eight NR subfamilies. For Class A of GPCRs, we chose 540 sequences randomly through equal interval selection (one in four), but for other classes of GPCRs, all the eligible sequences were selected because of the fewer members. For all the subfamilies of NRs, all the eligible sequences were chosen. Considering the limited amount of data available for some classes, such as GPCR Class E and NR Nerve Growth factor IB like subfamily, the proteins with high identity sequence were not removed in order to provide enough sequences to develop a wide-range predictive system that can be applied to all GPCR families and NR subfamilies. The number of sequences for each GPCR family and NR subfamily is listed in Tables 1 and 2, respectively.

Substitution models

Three kinds of substitution models: hydrophobicity model, electron-ion interaction potential (EIP) model (Cosic, 1994) and c-p-v model (Grantham, 1974), representing three principal properties of hydrophobicity, electronic property and bulk respectively, are used to transform the protein sequences into numerical sequences. Hydrophobicity of proteins is one of the most important factors in determining a protein’s structure and function. However, with different experimental conditions, different organic solvents and computing approaches, hydrophobicity value per amino acid will be different. So, three hydrophobicity scales, including KDDΦ (Kyle and Doolittle, 1982), MΦ (Mandell et al., 1997) and FΦ (Fauchère and Pliaka, 1983) were selected for optimization. EIP value describes the average energy states of all valence electron of amino acids and c-p-v model includes the composition (c), polarity (p) and molecular volume (v) of each amino acid.

Protein power spectrum

The Fourier transform (FT) has been commonly used in bioinformatics (Hiramoto et al., 2002; Katoh et al., 2002; Shepherd et al., 2003; Trad et al., 2002) because the frequency content of signals is often of great importance. It is a good method in capturing the essence of data. In this paper, fast Fourier transform (FFT) was used to transform proteins of variable length into fixed length vectors. The power spectrum or power spectral density, a measurement of the power at various frequencies was taken as the input of SVMs by using 512-point FFT.

Support vector machine

The support vector machine (SVM) is a kind of learning machine based on statistical learning theory. A brief and clear description for how to use SVM to do classification has been given by Chou and Cai (see, e.g., Chou and Cai, 2002; Cai et al., 2003). For a two-class classification problem, only one SVM classifier needs to be constructed, but the classification of GPCRs and NRs is a multi-class problem, so we used the ‘one versus rest’ method (Hua and Sun, 2001) to transfer it into a two-class problem. The radial basis function (RBF) was selected as the kernel function. All the parameters were kept constant except for C (regulatory parameter) and σ (kernel width parameter). In the training process, C and σ were optimized. The fixed length feature vector was obtained using the protein power spectrum with the fixed number of frequency points.

Performance evaluation

The performance of all classifiers was examined by jackknife test because it is the most rigorous and objective way to do cross-validation as elaborated in a comprehensive review (Chou and Zhang, 1995), and nowadays it has been adopted by more and more leading investigators in the area of statistical prediction (see, e.g., Cai and Chou, 2005; Chou, 1995, 2005b; Chou and Cai, 2004; Gao et al., 2005; Liu et al., 2005b; Shen and Chou, 2005a, b; Xiao et al., 2006; Zhou, 1998; Zhou and Assa-Munt, 2001; Zhou and Doctor, 2003). Each receptor is selected as the test receptor and the remaining receptors are used to train the SVMs. The prediction quality was evaluated using accuracy, total accuracy and Matthew’s correlation coefficient (MCC) (Matthews, 1975).

\[
\text{accuracy}(i) = \frac{p(i)}{\exp(i)}
\]

\[
\text{total accuracy} = \frac{\sum_{i=1}^{K} p(i)}{\exp(i)}
\]

\[
\text{MCC} = \frac{p(i)n(i) - o(i)\bar{o}(i)}{\sqrt{(p(i) + u(i))(p(i) + o(i))(n(i) + u(i))(n(i) + o(i))}}
\]

where, \(K\) is the class number, \(N\) is the total number of sequences, \(p(i)\) is the number of sequences observed in class i, \(o(i)\) is the number of correctly predicted sequences of class i, \(n(i)\) is the number of correctly predicted sequences not of class i, \(\bar{o}(i)\) is the number of under-predicted sequences, and \(\bar{n}(i)\) is the number of over-predicted sequences.

The measurement of prediction reliability is absolutely necessary when using the machine learning approaches for prediction. Here the index indicating the reliability of prediction (R) (Novic and Zupan, 1995) was used, as given by:

\[
R(i) = \frac{2(\text{accuracy}(i) - \text{error}(i))}{1 + |\text{accuracy}(i) - \text{error}(i)|}
\]

where, \(\text{error}(i) = \frac{o(i)}{n(i) + o(i)}\).

The reliability value \(R(i)\) ranges from 1 to –1. In the best case, when all the receptors are correctly predicted, \(R(i)\) is maximal (equal to 1), that is