Biochemical and Biophysical Research Communications 439 (2013) 303–308

Contents lists available at ScienceDirect

Biochemical and Biophysical Research Communications

journal homepage: www.elsevier.com/locate/ybbrc

Classifying G-protein-coupled receptors to the finest subtype level

Qing-Bin Gao, Xiao-Fei Ye, Jia He *

Department of Health Statistics, Second Military Medical University, Shanghai 200433, China

ARTICLE INFO

Article history:
Received 31 July 2013
Available online 20 August 2013

Keywords:
G-protein-coupled receptor
Hierarchical classification
Pseudo amino acid composition
Subtype level
Support vector machines

ABSTRACT

G-protein-coupled receptors (GPCRs) constitute a remarkable protein family of receptors that are involved in a broad range of biological processes. A large number of clinically used drugs elicit their biological effect via a GPCR. Thus, developing a reliable computational method for predicting the functional roles of GPCRs would be very useful in the pharmaceutical industry. Nowadays, researchers are more interested in functional roles of GPCRs at the finest subtype level. However, with the accumulation of many new protein sequences, none of the existing methods can completely classify these GPCRs to their finest subtype level. In this paper, a pioneer work was performed trying to resolve this problem by using a hierarchical classification method. The first level determines whether a query protein is a GPCR or a non-GPCR. If it is considered as a GPCR, it will be finally classified to its finest subtype level. GPCRs are characterized by 170 sequence-derived features encapsulating both amino acid composition and physicochemical features of proteins, and support vector machines are used as the classification engine. To test the performance of the present method, a non-redundant dataset was built which are organized at seven levels and covers more functional classes of GPCRs than existing datasets. The number of protein sequences in each level is 5956, 2978, 8079, 8680, 6477, 1580 and 214, respectively. By 5-fold cross-validation test, the overall accuracy of 99.56%, 93.96%, 82.81%, 85.93%, 94.1%, 95.38% and 92.06% were observed at each level. When compared with some previous methods, the present method achieved a consistently higher overall accuracy. The results demonstrate the power and effectiveness of the proposed method to accomplish the classification of GPCRs to the finest subtype level.

1. Introduction

G-protein-coupled receptors (GPCRs) are integral membrane proteins that possess seven membrane-spanning domains or transmembrane helices. Many physiological processes in mammals depend on GPCRs, which are also a major targets for pharmaceutical industry as is reflected by the fact that more than a quarter of all FDA approved drugs act on a GPCR [1]. Despite intensive academic and industrial research efforts over the past three decades, little is known about the structural basis of GPCR function [2]. Understanding the conformational changes and function of these receptors should facilitate the development of potential drugs with fewer side effects and more favorable pharmacological properties. Although some advanced biotechnologies such as NMR allow us to detect the three dimensional structure of proteins, the experimental approaches are generally very time-consuming and costly. Fortunately, with the rapid accumulation of biological data generated by many large-scale genome sequencing projects, it is becoming possible to develop a reliable theoretical method to predict the structure and function of GPCRs by their primary sequences [3].

GPCRDB is a molecular class-specific information system that collects, combines, validates and disseminates large amounts of heterogeneous data on GPCRs [4]. The main way to obtain the data is via the hierarchical list of GPCR families, which is accessible via http://www.gpcr.org/7tm/proteinfamily/. According to the latest release of GPCRDB (March, 2011), GPCRs are grouped into five families or classes based on the pharmacological nature of their ligand and sequence similarity. The five families are (1) Class A rhodopsin like, (2) Class B secretin like, (3) Class C metabotropic glutamate/pheromone, (4) vomeronasal receptors (V1R and V3R), and (5) taste receptors T2R. By the hierarchical list, GPCRs in the first four families are further divided into various subfamilies, sub-subfamilies, *etc.*, until the final subtypes are reached. Users can traverse the GPCR family tree and view or download the data for a selected family. Since the families of GPCRs are closely correlated with their structure and function, it would be very useful to develop a powerful computational method to distinguish GPCRs from genome sequences and, then classify them into a particular GPCR subtype for the purpose of understanding their biological function. The top-down hierarchical structure of GPCRs is shown in Fig. 1. In this structure each layer is called a level, same as Refs. [5–7]. In this study, we deal with the problem of GPCR classification at seven levels and try to assign each receptor to a specific subtype level.

* Corresponding author.
E-mail address: hejia63@yahoo.com (J. He).

0006-291X/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.bbrc.2013.08.023

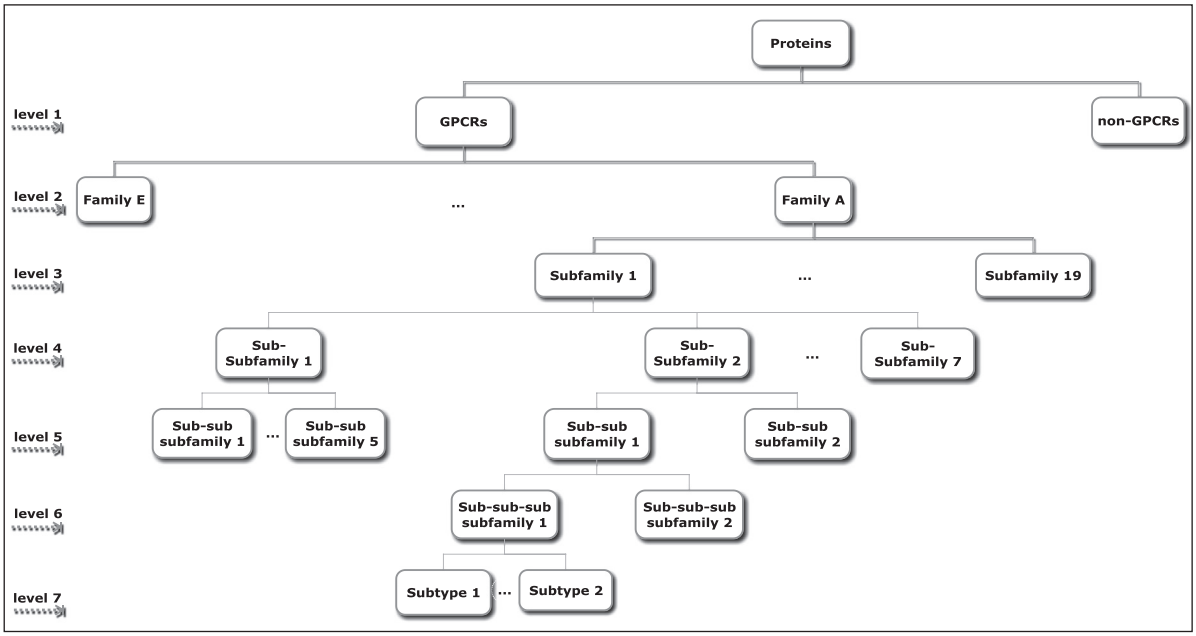

Fig. 1. Hierarchical classification structure of GPCRs.

A number of methods have been proposed for automatic classification of GPCRs in the past. The most obvious and straightforward method to characterizing a protein is to run a standard basic local alignment search tool (BLAST) [8]. BLAST searches have been used to identify novel GPCR proteins in cases where there has been moderate yet detectable sequence similarity to known GPCR sequences. This, however, makes the technique of limited use for the GPCR superfamily where there is a low degree of sequence similarity between the families [9]. Therefore, a lot of statistical and machine learning methods have been proposed. Most of them play the emphasis on the feature extraction techniques of proteins, typically including amino acid composition [10,11], dipeptide composition [6,12–15], pseudo-amino acid composition [16–19], Fisher score vectors [20], and others [21–23]. Based on a specific feature extraction technique, protein sequences of varying length are transformed into numerical vectors of fixed length, which can be used directly as inputs of various classifiers, such as covariant discriminant [10,11,22,24], support vector machines [12,14,16,17,20], nearest neighbor [6,16,17], and intimate sorting [7]. However, as far as we know, there are no methods that can completely predict GPCRs at seven levels. For instance, methods described in [10,11,14,24] can only predict GPCRs at a single level, methods presented in [13] just predict GPCRs at two levels, and methods in [5–7] considered prediction at three, four and five levels, respectively. This lead to some GPCRs cannot be classified to the subtype level. On the other hand, academic and industrial researchers are more interested in the functional roles of GPCRs at the finest subtype level. This is mainly because each subtype can demonstrate its own characteristic ligand binding property, coupling partners of trimeric G-proteins, and interaction partners of oligomerization [25]. Besides, some drugs that act on GPCRs cause therapeutic problems as a result of their failure to differentiate between subtypes. Therefore, prediction of GPCRs at the finest subtype level is obviously significant in the effort to decipher GPCRs. This is undoubtedly a hard and challenging work. Fortunately, more and more GPCR sequences are now being accumulated in the GPCRDB database, which makes it possible to develop a classifier to predict GPCRs at the finest subtype level.

In this paper, we attempt to develop a hierarchical classification method for the purpose of classifying GPCRs to their finest subtype level. Each GPCR is characterized by a 170 dimensional vector feature which consists of 20 amino acid composition features and 150 physicochemical features. Support vector machine (SVM) is employed as the classifier to predict GPCRs because it has been proved to be a powerful tool in multiple areas of biological analysis. Based on the latest version of GPCRDB, our method is able to fulfill the hierarchical classification of GPCRs at seven levels. GPCR proteins assigned to this level cannot be subdivided any more. 5-Fold cross-validation test was used to assess the performance of the present method on a newly-built non-redundant dataset. The high overall accuracy of 99.56%, 93.96%, 82.81%, 85.93%, 94.1%, 95.38% and 92.06% were observed from level 1 to level 7, respectively. The results demonstrate the efficiency and effectiveness of the proposed method. This is the first time that GPCRs are completely classified at seven levels. It is anticipated that this method would play a helpful role in the high quality prediction and classification of GPCRs at the subtype level.

2. Materials and methods

2.1. Dataset

To evaluate the performance of the proposed method, a non-redundant dataset was built from the latest version of GPCRDB [4]. GPCR sequences in GPCRDB are organized in six levels. Thus, plus the level of discriminating GPCRs from non-GPCRs, GPCR proteins in our dataset can be classified at seven levels (see Fig. 1). To reduce the homology bias of prediction, a redundancy reduction procedure was performed on this dataset by the program CD-HIT [26]. Sequences with high degree of similarity to the other sequences in the dataset were removed. In order to guarantee there are enough sequences in the training dataset, different sequence identity cut-offs were used for different levels. As in Ref. [7], cut-off 0.4, 0.7 and 0.8 were applied to the family, subfamily and sub-subfamily level, respectively, while 0.9 was used for the other four levels. After such a removal process, only classes with more than 10 sequences were remained in the dataset for training and testing.

In addition, to estimate the ability of the present method indiscriminating GPCRs from non-GPCRs, a negative dataset of

non-GPCRs was also collected. All the non-GPCRs were derived from ASTRAL SCOP 1.75A (March, 2012) domain sequence subset based on PDB SEQRES records [27], with less than 40% identity to each other. We also ran CD-HIT program to remove homologous sequences in the negative dataset with sequence identity cut-off 0.4, and non-GPCR proteins with sequence length less than 32 was eliminated. GPCRs in the family level are used as the positive dataset for training and evaluation. We randomly selected 2978 on-GPCR proteins from the negative dataset, so that the number of positive and negative prototypes in the dataset was equal. Thus, we totally obtained 5956 proteins in the superfamily level for subsequent discrimination analysis. In the end, the total number of GPCR proteins in each level is 5956, 2978, 8079, 8680, 6477, 1580 and 214, respectively. For convenience, we call this dataset GDSL (GPCR dataset in seven levels), which can be downloaded from http://stat.smmu.edu.cn/bioinfo/.

On the other hand, in order to make comparisons with other existing methods directly, four benchmark datasets from previous studies were evaluated in this study as well. For the sake of simplicity, they are denoted as GDFL, D1238, D365 and D566, respectively. The dataset GDFL [7] was used to develop a hierarchical classification predictor for classifying GPCR proteins at five levels, which contains 3178, 1589, 4772, 4924 and 2741 GPCRs at the superfamily, family, subfamily, sub-subfamily and subtype level, respectively. The dataset D1238 consists of 1238 GPCR proteins from three families: (1) rhodopsin like, (2) secretin like, and (3) metabotrophic/glutamate/ pheromone [10]. The dataset D365 was used to build a two-level classifier, which includes 365 GPCRs from six families: (1) rhodopsin-like (2) secretin-like (3) metabotrophic/glutamate/pheromone, (4) fungal pheromone (5) cAMP receptor, and (6) frizzled/smoothened [22]. Moreover, 365 non-GPCR sequences are collected to serve as a negative dataset against 365 GPCR sequences. While the dataset D566 has 566 proteins in seven sub-subfamilies: (1) adrenoceptor, (2) chemokine, (3) dopamine, (4) neuropeptide, (5) olfactory type, (6) rhodopsin, and (7) serotonin [24].

2.2. Support vector machine

Support vector machine (SVM) is a popular machine learning algorithm based on structural risk minimization for pattern classification [28]. It has been widely used in the community of biological sequence analysis. Different kernel functions define different SVMs. Two typical kernel functions are

$$K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i \cdot \mathbf{x}_j + 1)^d, \quad (1)$$

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma ||\mathbf{x}_i - \mathbf{x}_j||^2), \quad (2)$$

Eq. (1) is the polynomial kernel function of degree d when $d = 1$ it is the linear kernel function. Eq. (2) is the Radial Basic Function (RBF) kernel, where $\gamma = 1/\sigma$ and σ is called the width of the kernel. The SVM classifier is inherently a binary classifier, but it can be tailored for multi-classification. In this paper, the program used to implement SVMs was LibSVM [29]. Empirical studies have shown that RBF kernel outperforms linear kernel and polynomial kernel. Therefore, in this paper SVMs with RBF kernel were employed to construct the classifiers. The grid-search based on 5-fold cross-validation test was used to obtain the optimal values of parameter C and γ. When a classifier is trained, a query protein can be inputted into the classifier to predict its class label directly.

2.3. Sequence representation

To develop a classification model of SVMs, each protein sequence in the training dataset should be represented by a feature vector, which is usually constituted by some protein features. In addition to the conventional amino acid composition, which has been widely used for the characterization of protein sequences, in this study we use the pseudo amino acid composition (PseAAC) to accomplish the prediction of GPCRs. PseAAC is now a very popular approach used for improving the prediction quality of diverse protein attributes by incorporating sequence order and structure information of proteins [30,31]. In this paper, PseAAC of proteins were derived directly via the web server PseAAC [32], which is freely available at http://chou.med.harvard.edu/bioinf/PseAA/. The PseAAC web server can generate two kinds of pseudo amino acid composition, namely type I and type II. It has been demonstrated by some studies that type II PseAAC (series correlation types) is superior to the type I PseAAC (parallel correlation types) in prediction of protein attributes [33,34]. Six physicochemical characters of amino acids were used to calculate the correlations between amino acids at different positions along protein sequence. They are hydrophobicity, hydrophilicity, side chain mass, pK of the $\alpha - COOH$ group, pK of the $\alpha - NH_3^+$ group, and pI at 25 °C. The dimension the output of type II PseAAC is $(20 + \xi \times \lambda)$. Here, λ is a non-negative integer smaller than the length of the input sequence representing the rank of correlation of amino acids along a protein sequence, and ξ is the number of amino acid characters selected by the user. Particularly, when $\lambda = 0$, the PseAAC degenerated to the conventional amino acid composition. There are two parameters w and λ for generating PseAAC. The weighting factor w is designed for users to put weight on the additional PseAAC with respect to the conventional amino acid components. It has been demonstrated that λ and w have an effect on the classification performance of SVMs. In this paper, we use type II PseAAC to represent proteins and set $\lambda = 25$ and $w = 0.5$, which is proved to be able to lead to a higher prediction accuracy.

The hydrophobicity and hydrophilicity of amino acids in a protein are closely correlated with its structure and functions, including its folding, its interaction with the environment and other molecules, and its catalytic mechanism. Previous studies have shown that these two characters can be used to effectively and partially reflect the sequence order effects through the amphiphilic PseAA composition [35]. In this study, all the six physicochemical characters were used to generate PseAA composition ($\xi = 6$). Thus, the dimension of the protein feature vector is $20 + 6 \times 25 = 170$, which is subsequently used as a descriptor to characterize protein sequences. This input vector is expected to be able to encapsulate more sequence order and structural information of proteins, and its prediction quality using SVMs is discussed in the following sections. The values of each element of PseAA composition were normalized between 0 and 1 using a standard conversion formula before it was inputted into the prediction engine of SVMs.

2.4. Prediction evaluation

In statistical prediction, n-fold cross-validation test and jackknife test are often used to examine a classifier for its effectiveness in practical application. The jackknife test is deemed the most objective and rigorous one that can always yield a unique result for a given benchmark dataset, and hence has been increasingly used by investigators to examine the accuracy of various predictors. However, performing jackknife test with SVMs might take a long time in practical applications, particularly for a large dataset. So the 5-fold cross-validation is also widely accepted in the performance evaluation of various prediction methods. In this paper, we used the 5-fold cross-validation test to evaluate the quality of the present method on dataset GDSL. As for the comparison with other methods on dataset GDFL, D1238, D365 and D566, the jackknife test is used. The performance metrics used for evaluating classifiers are overall accuracy and accuracy. They are defined by

Table 1
Overall prediction accuracy of the present method on each level by 5-fold cross-validation test performed on GDSL dataset.

Level	No. of proteins	Correctly predicted	Overall accuracy (%)
1st	5956	5930	99.56
2nd	2978	2798	93.96
3rd	8079	6690	82.81
4th	8680	7459	85.93
5th	6477	6095	94.10
6th	1580	1507	95.38
7th	214	197	92.06

$$\text{overall accuracy} = \frac{\sum_{i=1}^{k} p(i)}{N}, \tag{3}$$

$$\text{accuracy} = \frac{p(i)}{\text{obs}(i)}, \tag{4}$$

where N is the total number of sequences in a level, k is the number of classes in that level, $p(i)$ is the number of correctly predicted sequences of class i obs(s) is the number of sequences observed in class i.

3. Results and discussion

3.1. Predicting GPCRs at seven levels

Our work tries to classify GPCRs at seven levels by a top-down hierarchical classification structure as shown in Fig. 1. At the first level, a query protein is predicted to be either a GPCR or a non-GPCR. If it is identified as a GPCR, it will be further classified into one of the five families at the second level. Then, the third level will determine which subfamily the protein belongs to and pass it down to the next level. This decision process continues until the protein reaches a level that cannot be further divided. At present, all the GPCRs in GPCRDB database can reach a specific subtype level by this seven-level hierarchical classification method. The 5-fold cross-validation test performed on the dataset GDSL showed that the overall accuracies of our method achieved 99.56%, 93.96%, 82.81%, 85.93%, 94.1%, 95.38% and 92.06% at the seven levels, respectively (see Table 1). Detailed prediction results for the individual classes of each level are presented in the supplementary file 1. From the prediction results we can see that the overall accuracy of each level is very high. However, the accuracies for some individual classes are still very low. This is probably due to the smaller number of training samples in those classes. With the accumulation of more GPCR proteins, the accuracy can be improved significantly.

3.2. Comparison with other methods

In order to explain the competitive performance of the present method, we made comparisons with several previous methods based on various benchmark datasets. The details of each comparison are described as follows.

3.3. Comparison with PCA-GPCR at five levels

We first made comparison with PCA-GPCR developed on dataset GDFL [7]. PCA-GPCR is the first classifier aims at predicting GPCRs at five levels. The dataset GDFL can be downloaded at http://www1.spms.ntu.edu.sg/~chenxin/PCA_GPCR. The comparison results are presented in Table 2. From Table 2 we can see that the present method resulted in an overall accuracy of 99.7%, 94.3%, 85.9%, 87.3% and 93.9% on the five levels of GDFL dataset, respectively. The results indicate that the proposed method of this paper outperforms PCA-GPCR at all the five levels consistently.

3.4. Comparison with GPCR-CA at two levels

We also compared our method with GPCR-CA [22] and PCA-GPCR [7] on dataset D365, which was originally developed for predicting GPCRs at the first two levels. The prediction results at the first level are shown in Table 3. From Table 3 we know that the present method achieved an overall accuracy of 97.12%, higher than 91.64% and 95.21% of the other two methods. As for the classification of GPCR families at the second level, Table 4 shows that the present method reached an overall accuracy of 93.15, greatly higher than 83.56% of GPCR-CA and slightly higher than 92.6% of PCA-GPCR.

3.5. Comparison at a single level

Many existing methods just classified GPCRs at a single level. In order to make comparisons with these methods directly, we used the benchmark datasets of D566 and D1238 to evaluate our method at a specific level. Dataset D566 comprises GPCRs from the fourth level, while dataset D1238 is composed of GPCRs from the second level. The prediction results about dataset D566 are shown in Table 5. From Table 5 we know that the present method achieved an overall accuracy of 98.23%, higher than 92.05% from Chou and Elrod [24] and 97.88% from PCA-GPCR [7]. Table 6 shows the comparison results with Chou [10] and PCA-GPCR [7] on dataset D1238. From Table 6 we know that the present method reached an overall accuracy of 99.84%, slightly higher than the other two methods. It is noticeable that there are more individual accuracies that reached 100% by our method.

The comparisons above indicate that our method has achieved a higher overall accuracy on the four benchmark datasets than some previous methods. This demonstrates that the 170 dimensional feature vector used in this paper contains enough information for characterization of GPCR proteins.

4. Discussion

In this paper, we performed a pioneer work to predict GPCRs at seven levels for the purpose of classifying GPCRs to their finest subtype level. This is the first time that GPCRs are predicted at seven levels by a top-down hierarchical classification structure. A 170 dimensional feature vector derived from Chou's PseAAC was utilized to characterize protein sequences, which is anticipated to be able to capture some sequence order and structure information of proteins. A new large GPCR dataset was built to evaluate the performance of the present method using support vector machines. By the 5-fold cross-validation test, a high overall accuracy of 99.56%, 93.96%, 82.81%, 85.93%, 94.1%, 95.38% and 92.06% were reached at seven levels, respectively. We then compared our method with some previous methods based on four benchmark datasets widely used in the literature. When compared with a five-level method

Table 2
Comparison with PCA-GPCR [7] on GDFL dataset by the jackknife test.

Level	PCA-GPCR	This paper
Superfamily	99.5	**99.7**
Family	88.8	**94.3**
Subfamily	80.5	**85.9**
Sub-subfamily	80.3	**87.3**
Subtype	92.3	**93.9**

The highest overall accuracy is shown in bold.

Table 3
Comparison with GPCR-CA [22] and PCA-GPCR [7] on dataset D365 at the first level by the jackknife test ($C = 8$, $\gamma = 0.5$).

Protein type	GPCR-CA	PCA-GPCR	This paper
GPCR	92.33	96.99	97.53
Non-GPCR	90.96	93.42	96.71
Overall	91.64	95.21	**97.12**

The highest overall accuracy is shown in bold.
C and γ: parameters for RBF kernel.

Table 4
Comparison with GPCR-CA [22] and PCA-GPCR [7] on dataset D365 at the second level by the jackknife test ($C = 32$, $\gamma = 0.0625$).

Family	GPCR-CA	PCA-GPCR	This paper
Rhodopsin-like	96.55	95.69	99.14
Secretin-like	74.36	87.18	82.05
Meta/glut/pher	81.82	88.64	90.91
Fungal pheromone	8.70	95.65	82.61
CAMP receptor	60	100	100
Frizzled/smoothened	47.06	64.71	52.94
Overall	83.56	92.60	**93.15**

The highest overall accuracy is shown in bold.
C and γ: parameters for RBF kernel.

Table 5
Comparison with Chou and Elrod [24] and PCA-GPCR [7] on dataset D566 by the jackknife test ($C = 8$, $\gamma = 0.0625$).

Subfamily	Chou and Elrod	PCA-GPCR	This paper
Adrenoceptor	90.91	98.48	96.97
Chemokine	92.39	97.83	96.74
Dopamine	69.77	93.02	90.70
Neuropeptide	61.29	96.77	96.77
Olfactory	97.62	100	100
Rhodopsin	99.45	98.36	100
Serotonin	90.03	97.01	100
Overall	92.05	97.88	**98.23**

The highest overall accuracy is shown in bold.
C and γ: parameters for RBF kernel.

Table 6
Comparison with Chou [10] and PCA-GPCR on dataset D1238 by the jackknife test ($C = 8$, $\gamma = 0.03125$).

Family	Chou	PCA-GPCR	This paper
Rhodopsin-like	99.00	99.91	100
Secretin-like	88.10	98.81	97.62
Meta/glut/pher	78.43	98.04	100
Overall	97.42	99.76	**99.84**

The highest overall accuracy is shown in bold.
C and γ parameters for RBF kernel.

named PCA-GPCR, our method reached an overall accuracy of 99.7%, 94.3%, 85.9%, 87.3% and 93.9% on the five levels. They are consistently higher than those of PCA-GPCR on all the five levels. When comparisons were made on a two-level dataset, the present method achieved an overall accuracy of 97.12% and 93.15% at the first two levels, higher than the methods under comparison. As for the comparisons with methods predicting GPCRs at a single level, our method also outperformed those methods under comparison.

We can conclude that the present method fulfilled the prediction of GPCRs at seven levels with high accuracy. It can be used to classify GPCRs to the finest subtype level. It is also likely to provide valuable information both in seeking novel receptors in genome data and in the characterization of orphan receptors.

Acknowledgments

This work was supported by the grants from the National Natural Science Foundation of China (Nos. 31371344, 30901243), Leading Talents of Science in Shanghai 2010 (No. 022) and key discipline construction fund of evidence-based public health in Shanghai (No. 12GWZX0602).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.bbrc.2013.08.023.

References

[1] J.P. Overington, B. Al-Lazikani, A.L. Hopkins, How many drug targets are there?, Nat Rev. Drug Discov. 5 (2006) 993–996.
[2] D.M. Rosenbaum, S.G. Rasmussen, B.K. Kobilka, The structure and function of G-protein-coupled receptors, Nature 459 (2009) 356–363.
[3] N. Vaidehi, W.B. Floriano, R. Trabanino, S.E. Hall, P. Freddolino, et al., Prediction of structure and function of G protein-coupled receptors, Proc. Natl. Acad. Sci. USA 99 (2002) 12622–12627.
[4] B. Vroling, M. Sanders, C. Baakman, A. Borrmann, S. Verhoeven, et al., GPCRDB: information system for G protein-coupled receptors, Nucleic Acids Res. 39 (2011) D309–D319.
[5] M.N. Davies, A. Secker, A.A. Freitas, M. Mendao, J. Timmis, et al., On the hierarchical classification of G protein-coupled receptors, Bioinformatics 23 (2007) 3113–3118.
[6] Q.B. Gao, Z.Z. Wang, Classification of G-protein coupled receptors at four levels, Protein Eng. Des. Sel. 19 (2006) 511–516.
[7] Z.L. Peng, J.Y. Yang, X. Chen, An improved classification of G-protein-coupled receptors using sequence-derived features, BMC Bioinformatics 11 (2010) 420.
[8] R. Lopez, V. Silventoinen, S. Robinson, A. Kibria, W. Gish, WU-Blast2 server at the European Bioinformatics Institute, Nucleic Acids Res. 31 (2003) 3795–3798.
[9] M.N. Davies, D.E. Gloriam, A. Secker, A.A. Freitas, M. Mendao, et al., Proteomic applications of automated GPCR classification, Proteomics 7 (2007) 2800–2814.
[10] K.C. Chou, Prediction of G-protein-coupled receptor classes, J. Proteome Res. 4 (2005) 1413–1418.
[11] D.W. Elrod, K.C. Chou, A study on the correlation of G-protein-coupled receptor types with amino acid composition, Protein Eng. 15 (2002) 713–715.
[12] M. Bhasin, G.P. Raghava, GPCRpred: an SVM-based method for prediction of families and subfamilies of G-protein coupled receptors, Nucleic Acids Res. 32 (2004) W383–W389.
[13] Y. Huang, J. Cai, L. Ji, Y. Li, Classifying G-protein coupled receptors with bagging classification tree, Comput. Biol. Chem. 28 (2004) 275–280.
[14] M. Bhasin, G.P. Raghava, GPCRsclass: a web tool for the classification of amine type of G-protein-coupled receptors, Nucleic Acids Res. 33 (2005) W143–W147.
[15] Q.B. Gao, C. Wu, X.Q. Ma, J. Lu, J. He, Classification of amine type G-protein coupled receptors with feature selection, Protein Pept. Lett. 15 (2008) 834–842.
[16] W.Z. Lin, X. Xiao, K.C. Chou, GPCR-GIA: a web-server for identifying G-protein coupled receptors and their families with grey incidence analysis, Protein Eng. Des. Sel. 22 (2009) 699–705.
[17] Z. Ur-Rehman, A. Khan, G-protein-coupled receptor prediction using pseudo-amino-acid composition and multiscale energy representation of different physiochemical properties, Anal. Biochem. 412 (2011) 173–182.
[18] J.D. Qiu, J.H. Huang, R.P. Liang, X.Q. Lu, Prediction of G-protein-coupled receptor classes based on the concept of Chou's pseudo amino acid composition: an approach from discrete wavelet transform, Anal. Biochem. 390 (2009) 68–73.
[19] Q. Gu, Y.S. Ding, T.L. Zhang, Prediction of G-protein-coupled receptor classes in low homology using Chou's pseudo amino acid composition with approximate entropy and hydrophobicity patterns, Protein Pept. Lett. 17 (2010) 559–567.
[20] R. Karchin, K. Karplus, D. Haussler, Classifying G-protein coupled receptors with support vector machines, Bioinformatics 18 (2002) 147–159.
[21] Y.Z. Guo, M. Li, M. Lu, Z. Wen, K. Wang, et al., Classifying G protein-coupled receptors and nuclear receptors on the basis of protein power spectrum from fast Fourier transform, Amino Acids 30 (2006) 397–402.
[22] X. Xiao, P. Wang, K.C. Chou, GPCR-CA: a cellular automaton image approach for predicting G-protein-coupled receptor functional classes, J. Comput. Chem. 30 (2009) 1414–1423.
[23] R. Gupta, A. Mittal, K. Singh, A novel and efficient technique for identification and classification of GPCRs, IEEE Trans. Inf. Technol. Biomed. 12 (2008) 541–548.
[24] K.C. Chou, D.W. Elrod, Bioinformatical analysis of G-protein-coupled receptors, J. Proteome Res. 1 (2002) 429–433.
[25] K. Kristiansen, Molecular mechanisms of ligand binding, signaling, and regulation within the superfamily of G-protein-coupled receptors: molecular

modeling and mutagenesis approaches to receptor structure and function, Pharmacol. Ther. 103 (2004) 21–80.
[26] W. Li, A. Godzik, Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences, Bioinformatics 22 (2006) 1658–1659.
[27] SCOP ASTRAL website. <http://scop.berkeley.edu>/.
[28] V. Vapnik, Statistical Learning Theory, Wiley, New York, 1998.
[29] C.C. Chang, C.J. Lin, LIBSVM: a library for support vector machines, ACM Trans. Int. Syst. Technol. 2 (27) (2011) 1–27. Software is available at <http://www.csie.ntu.edu.tw/≃cjlin/libsvm>.
[30] K.C. Chou, Prediction of protein cellular attributes using pseudo-amino acid composition, Proteins 43 (2001) 246–255.
[31] K.C. Chou, Some remarks on protein attribute prediction and pseudo amino acid composition, J. Theor. Biol. 273 (2011) 236–247.
[32] H.B. Shen, K.C. Chou, PseAAC: a flexible web server for generating various kinds of protein pseudo amino acid composition, Anal. Biochem. 373 (2008) 386–388.
[33] Q.B. Gao, Z.C. Jin, X.F. Ye, C. Wu, J. He, Prediction of nuclear receptors with optimal pseudo amino acid composition, Anal. Biochem. 387 (2009) 54–59.
[34] Q.B. Gao, H. Zhao, X. Ye, J. He, Prediction of pattern recognition receptor family using pseudo-amino acid composition, Biochem. Biophys. Res. Commun. 417 (2012) 73–77.
[35] K.C. Chou, Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes, Bioinformatics 21 (2005) 10–19.