Amino Acids (2012) 43:793–804
DOI 10.1007/s00726-011-1133-6

ORIGINAL ARTICLE

Classification of G proteins and prediction of GPCRs-G proteins coupling specificity using continuous wavelet transform and information theory

Zhanchao Li · Xuan Zhou · Zong Dai · Xiaoyong Zou

Received: 24 May 2011 / Accepted: 20 October 2011 / Published online: 16 November 2011

Abstract The coupling between G protein-coupled receptors (GPCRs) and guanine nucleotide-binding proteins (G proteins) regulates various signal transductions from extracellular space into the cell. However, the coupling mechanism between GPCRs and G proteins is still unknown, and experimental determination of their coupling specificity and function is both expensive and time consuming. Therefore, it is significant to develop a theoretical method to predict the coupling specificity between GPCRs and G proteins as well as their function using their primary sequences. In this study, a novel four-layer predictor (GPCRsG_CWTIT) based on support vector machine (SVM), continuous wavelet transform (CWT) and information theory (IT) is developed to classify G proteins and predict the coupling specificity between GPCRs and G proteins. SVM is used for construction of models. CWT and IT are used to characterize the primary structure of protein. Performance of GPCRsG_CWTIT is evaluated with cross-validation test on various working dataset. The overall accuracy of the G proteins at the levels of class and family is 98.23 and 85.42%, respectively. The accuracy of the coupling specificity prediction varies from 74.60 to 94.30%. These results indicate that the proposed predictor is an effective and feasible tool to predict the coupling specificity between GPCRs and G proteins as well as their functions using only the protein full sequence. The establishment of such an accurate prediction method will facilitate drug discovery by improving the ability to identify and predict protein–protein interactions. GPCRsG_CWTIT and dataset can be acquired freely on request from the authors.

Z. Li · X. Zhou · Z. Dai · X. Zou (✉)
School of Chemistry and Chemical Engineering, Sun Yat-Sen University, Guangzhou 510275, People's Republic of China
e-mail: ceszxy@mail.sysu.edu.cn

Z. Li
School of Chemistry and Chemical Engineering, Guangdong Pharmaceutical University, Guangzhou 510006, People's Republic of China

Keywords G protein-coupled receptors · G proteins · Support vector machine · Continuous wavelet transform · Information theory

Introduction

G protein-coupled receptors (GPCRs), also known as 7 α-helices transmembrane receptors due to their characteristic structure of 7 transmembrane α-helices separated by alternating intracellular and extracellular loops, comprise the largest superfamily of membrane proteins of pharmacological interest. GPCRs transmit extracellular signals into the cell through their interaction with a broad range of ligands such as ions, pheromones, hormones, neurotransmitters and proteins. When GPCRs are activated by these ligands, the receptors change their conformation and lead to associations of the receptors with the guanine nucleotide-binding proteins (G proteins). G proteins are composed of α, β and γ subunits and can be divided into four subtypes (G_s, $G_{i/o}$, $G_{q/11}$ and $G_{12/13}$) based on the structure and function of α-subunits (Guo et al. 2006a; Simon et al. 1991). The G_s class is involved in activating adenylyl cyclase, the $G_{i/o}$ in inhibiting adenylyl cyclase and regulating ion channels, the $G_{q/11}$ in activating phospholipase C and $G_{12/13}$ in activating the Na^+/H^+ exchanger pathway (Sreekumar et al. 2004).

The associations between G proteins and GPCRs trigger the exchange of the guanosine diphosphate (GDP) bound on the α-subunit of the G proteins with guanosine triphosphate (GTP) and the dissociation of the G proteins into α-GTP and β-γ complexes (Elefsinioti et al. 2004; Sgourakis et al. 2005b). The dissociated α-subunit can activate or inhibit several effector proteins, such as adenylyl cyclase 1–9, PLCβ 1–4, tyrosine kinases, phosphodiesterases, phosphoinositide 3-kinase, ion channels and molecules of the mitogen-activated protein kinase pathway, resulting in a variety of cellular functions and physiological responses that depend on the biological specificity of the dissociated subunits (Cabrere-Vera et al. 2003; Elefsinioti et al. 2004; Pierce et al. 2002; Sgourakis et al. 2005b; Theodoropoulou et al. 2008).

As the coupling of GPCRs and a specific class of G proteins plays an extremely important role in signal transduction, it has been a central theme in biology for the elucidation of coupling specificity between GPCRs and G proteins as well as the determination of their function. Until now, various experimental methods in biochemistry have been developed to determine the function and elucidate the coupling mechanism between GPCRs and G proteins. Although many years of intensive research, the coupling mechanism between GPCRs and G proteins is still poorly understood, and the function of many GPCRs and G proteins has not been experimentally determined. In contrast, more and more GPCRs and G proteins amino acid sequences are known with the rapid accumulation of data of new protein sequences produced by the high-throughput sequencing technology. In view of this situation, it is vitally important to develop a computational method for fast and accurate prediction of function and coupling specificity of GPCRs and G proteins.

Actually, many methods have been proposed and successfully used to predict the function of GPCRs (Bhasin and Raghava 2004, 2005; Chou 2005; Chou and Elord 2002; Elrod and Chou 2002; Eo et al. 2007; Gao and Wang 2006; Guo et al. 2006b; Karchin et al. 2002; Li et al. 2010; Papasaikas et al. 2004; Qian et al. 2003; Qiu et al. 2009). But, until now no method is reported and used to predict the function of G proteins. In addition, a series of important computational methods have been developed to predict the coupling specificity of GPCRs and G proteins. These methods can be categorized into three groups according to the technique utilized. The first one is similarity searches with sequence alignment tools such as BLAST (Altschul et al. 1990) and CLUSTALW (Thompson et al. 1994). However, the methods have been proven to be unsuccessful (Wess 1998) due to the following reasons: (1) some homologous GPCR pairs with the same ligands bind to different kinds of G protein; (2) those pairs coupling to the same type of G protein bind to different ligands; (3) some GPCR pairs bind to both the same ligand and the same G protein even though they show sequence similarity is of <25% (Gaulton and Attwood 2003; Yabuki et al. 2005). The second point above is primarily based on site-directed mutagenesis studies (Greasley et al. 2001), correlated mutation analysis (Horn et al. 2000; Oliveira et al. 1999) and kinetic modeling (Kukkonen et al. 2001). However, the conclusions from these studies offer either an insight into a specific receptor or receptor subfamily, or otherwise do not assess the specificity of coupling between a receptor and multiple G proteins (Cao et al. 2003). The third point above is based on statistical and machine-learning method, including support vector machines (SVM) (Guo et al. 2006b; Yabuki et al. 2005), hidden Markov model (Qian et al. 2003; Sgourakis et al. 2005a, b; Sreekumar et al. 2004), Naïve Bayes model (Cao et al. 2003) and other techniques (Moller et al. 2001). However, most of these methods have only used the intracellular domain, which drastically diminishes their accuracy due to limitations in membrane topology prediction algorithms (Moller et al. 2001; Sgourakis et al. 2005a). In addition, although the extracellular domains, transmembrane segments and α subunits of G proteins do not make physical contact with each other, they may also hold the information of coupling specificity. Thus, the effective integration of such information may be able to improve the prediction accuracy and provide useful clues to analyze the coupling mechanism. More importantly, these methods only predict the coupling preference of GPCRs to specific classes of G proteins (i.e. classify GPCRs into $G_{i/o}$, $G_{q/11}$, G_s and $G_{12/13}$ class based on their G protein coupling preference) and cannot predict coupling between arbitrary combinations of GPCRs and G proteins from different classes.

In view of these facts, a novel four-layer predictor (GPCRsG_CWTIT) is presented to classify G proteins and predict coupling specificity between GPCRs and G proteins based on SVM, continuous wavelet transform (CWT) and information theory (IT). In every layer, the primary amino acid sequences are firstly translated into numerical sequences by various physicochemical properties of amino acids. After that, the numerical sequences are transformed by CWT and then a new feature vector is extracted from the coefficient of CWT based on IT. Finally, two SVM models are constructed to predict whether the query protein is GPCR or G protein in the first layer. Three SVM models in the second layer and 6 SVM models in the third layer are used to classify G proteins at level of class (G-α, G-β and G-γ) and family (G_s, $G_{i/o}$, $G_{q/11}$ and $G_{12/13}$), respectively. Four SVM models in the fourth layer are utilized to predict whether the coupling is between a GPCR protein and a G protein. Compared with our previous work (Li et al. 2010), the current work is not devoted to GPCRs function prediction but G protein function and the coupling specificity between GPCRs and G proteins. And more importantly, based on continuous wavelet transform and information theory, a novel method of feature extraction is proposed to characterize the primary structure of protein. The prediction quality evaluated on various non-redundant dataset by the cross-validation test exhibited the state-of-the-art performance of the current method.

Materials and methods

Data collection and dataset construction

To construct high-quality benchmark dataset, primary sequences of GPCRs and G proteins are collected from the database of GPCRs, G proteins, effectors and their interactions (gpDB) at http://biophysics.biol.uoa.gr/gpDB/ (Elefsinioti et al. 2004; Theodoropoulou et al. 2008). The gpDB is a publicly accessible database that collects and combines information about GPCRs and G proteins, as well as information concerning known coupling specificity between these proteins. In the database, the GPCRs and the G proteins are classified according to a hierarchy of different classes, families and sub-families, based on extensive literature searches (Elefsinioti et al. 2004). For example, G proteins can be classified into G-α class, G-β class and G-γ class. G-α class can be divided into four families (G_s, $G_{i/o}$, $G_{q/11}$ and $G_{12/13}$) and each family can also be further subdivided into different subfamilies and types. GPCRs are classified into the following six classes: class A (rhodopsin like), class B (secretin like), class C (metabotropic glutamate/pheromone), class D (fungal pheromone receptor), class E (cyclic nucleotide receptor) and frizzled/smoothened. Each class is grouped into different families and each family is further grouped into different subfamilies.

To ensure quality, data are screened according to the following protocol. Firstly, all of the incomplete sequences of GPCR and G protein (i.e. fragments) are removed. Secondly, all of the proteins belonging to G-α class but without further classification are discarded. Thirdly, a redundancy cutoff is performed to winnow those sequences with more than 90% sequence identity to any other in the same dataset. Taking into account the fact that some GPCRs with high sequence similarity can bind to different kinds of G protein (Gaulton and Attwood 2003; Yabuki et al. 2005), the redundancy cutoff of 90% is operated by CD-HIT program (Li et al. 2001). After strictly following the above procedures, we finally obtained 1376 GPCR proteins and 113 G proteins. In the 113 G proteins, 48 belongs to G-α class, 39 to G-β class and 26 to G-γ class. In the 48 proteins of G-α class, 10 belongs to family of G_s, 18 to $G_{i/o}$, 17 to $G_{q/11}$ and 3 to $G_{12/13}$. Based on these GPCR proteins and G proteins, 421 coupled pairs of GPCRs-G_s, 63 coupled pairs of GPCRs-$G_{i/o}$, 69 coupled pairs of GPCRs-$G_{q/11}$ and 26 coupled pairs of GPCRs-$G_{12/13}$ can be obtained. Since the coupling specificity of GPCRs-G proteins is not a one-to-one function (i.e. a particular G protein may couple to more than one GPCR and vice versa), the number of these coupled pairs is larger than that of G proteins at level of family.

To distinguish G proteins from non-G proteins or GPCR proteins from non-GPCR proteins, a non-G protein and non-GPCRs dataset are also constructed by randomly collecting 113 non-G proteins and 1376 non-GPCR proteins from the UniProtKB database at http://www.uniprot.org/. Within the 113 randomly selected non-G proteins, except the sequence similarity of only two non-G proteins is 44%, the sequence similarity between any two non-G proteins is lower than 40%. Within the 1,376 non-GPCR proteins, except 33 pairwise sequence similarities are higher 40%, the sequence similarity between any two non-GPCR proteins is lower than 40%. And no proteins simultaneously exist in the dataset of G proteins and non-G proteins or the dataset of GPCRs and non-GPCRs. Non-coupled pairs are necessary to build models for distinguishing coupled pairs from non-coupled pairs. So, the non-coupled pairs are generated by randomly pairing proteins from the dataset of G proteins and GPCRs. The method must meet the following three requirements: (1) G proteins appeared in coupled pairs and corresponding non-coupled pairs belong to the same family of G proteins; (2) GPCR proteins contained in non-coupled pairs have no coupling information for G proteins; (3) the number of non-coupled pairs is equal to that of coupled pairs. Please note that the dataset of non-coupled pairs generated using the method mentioned above may contain some false-negative samples due to the incompletion of the coupled pairs contained in the gpDB database. However, this problem can be solved with the accumulation of data about coupling specificity between GPCRs and G proteins.

Continuous wavelet transform and information theory

CWT has become a popular signal analysis tool in various fields since 1980s. Unlike Fourier transform, CWT possesses the ability to elucidate simultaneously both spectral and temporal information within the signal. In mathematics, CWT of a digital signal can be described as follows:

$$W_f(a,b) = \frac{1}{\sqrt{a}} \int f(t)\psi\left(\frac{t-b}{a}\right)\mathrm{d}t \qquad (1)$$

where, a and b are the scale factor and shift factor respectively, $a, b \in \mathrm{R}$ and $a > 0$, $\psi(t)$ is wavelet core, $W_f(a, b)$ is the result of inner product operation between signal and wavelet core. After CWT, the signal can be decomposed into many groups of coefficients in different scales. The coefficients in low scales represent the fine-scale characters in the digital signal, and can be used to analyze the local features of the signal. Contrarily, the coefficients in high scales represent the coarse-scale characters in the digital signal, and can be used to study the global features of the signal.

Although CWT can reveal many important characters of digital signal, it is still an important problem to accurately extract these characters and formulate them as a set of features. Here, IT approach is introduced to deal with the problem. IT, as a branch of applied mathematics, was developed by Shannon (Shannon 1948) to find fundamental limits on signal processing operations. In the theory, entropy as a measure of information can be used to quantify the uncertainty involved in a random variable. For example, for a discrete random variable X with N states, entropy can be defined as:

$$H(X) = -\sum_{i=1}^{N} p_i \log p_i \tag{2}$$

where, p_i is the probability of X with ith state, $0 < p_i < 1$ and $\sum_{i=1}^{N} p_i = 1$. More explanation about the calculation of probability can be found in reference (Li et al. 2010). Similarly, if there are two discrete random variables X and Y with N states, relative entropy can be computed according to Eq. 3:

$$RH(X|Y) = -\sum_{i=1}^{N} \frac{p_i(x)}{p_i(y)} \log \frac{p_i(x)}{p_i(y)} - \sum_{i=1}^{N} \frac{p_i(y)}{p_i(x)} \log \frac{p_i(y)}{p_i(x)} \tag{3}$$

The relative entropy can be used to describe the differences between the two variables. Obviously, $\mathrm{RH}(X|Y) = 0$ if the variable X is equal to the variable Y.

Characterization of protein sequence and coupled pair based on CWT and IT

The interaction between GPCRs and G proteins involves many physicochemical properties. To reflect the properties and convert protein sequence into corresponding digital signal, we select 7 physicochemical properties, including hydrophobicity, hydrophilicity, volumes of side chains of amino acids, polarity, polarizability, solvent-accessible surface area and net charge index of side chains of amino acids. These properties have been successfully used to predict protein–protein interactions (Guo et al. 2008).

Here, the protein with ID number of GPR0005 in the gpDP database is chosen as an example to describe protein primary structure representation by CWT and IT. The protein GPR0005 with 355 residues belongs to G-α class, $\mathrm{G_{i/o}}$ family. Firstly, the amino acid sequence of protein is converted into a digital signal based on the hydrophobic values of amino acids. Secondly, Meyer wavelet is chosen and scale vector is selected in the range of 1–100 for the step of 1 to perform CWT based on Eq. 1. The results are shown in Fig. 1, in which x, y and z axes represent the residue position of the amino acids in primary structure, the decomposition scales and the coefficients in different scales, respectively. Thirdly, the maximum and minimum values are found in all coefficients, and the difference between the maximum and minimum values is divided into L zones (L is a positive integer greater than 1). Therefore, L data zones are obtained as follows: [Min, Min + d], [Min + d, Min + 2d], …, [Min + (L − 1)d, Max], where $d = \frac{\mathrm{Max}-\mathrm{Min}}{L}$, Max and Min are the maximum and minimum values, respectively. Fourthly, coefficients at the level of scale belonging to the L data zones are counted. Thus, we obtain a series of number in each scale vector j (j = 1, 2, …, 100): m_{j1}, m_{j2}, …, m_{jL}. Where m_{j1} represents the number in the zone of [Min, Min + d] at the scale j, m_{j2} represents the number in the zone of [Min + d, Min + 2d] at the scale j, etc. And the probability (p_i) of any coefficient belongs to ith data zone can be computed. Finally, the entropy in each scale can be obtained using the Eq. 2 and can be used to characterize the protein.

Similarly, we can also calculate the p_i of any coefficient belongs to ith data zone to GPCR protein that can couple to the protein GPR0005 based on the method mentioned above. The relative entropy between the GPCR protein and the G protein can be obtained using the Eq. 3 and can be used to characterize the coupled pair. Finally, a protein or coupled pair is characterized by concatenating the entropy or relative entropy that is derived from various physicochemical properties.

Construction of the prediction system and assessment of performance

To the current study, publicly available Libsvm software (Chan and Lin 2001) is utilized to construct the classifier with the radial basis function as kernel function. The strategy of grid search is adopted to optimize the accuracy of tenfold cross-validation test. In the tenfold cross-validation, the dataset is divided randomly into 10 equally sized subsets. The training and testing are carried out ten times, each time using one distinct subset for testing and the remaining nine subsets for training. For the current study, a total of 15 SVM classifiers are constructed: 1 for identifying G proteins from non-G proteins, 1 for discriminating between GPCRs and non-GPCRs, 3 for classifying G proteins at level of class (G-α, G-β and G-γ), 6 for classifying G-α proteins at level of family ($\mathrm{G_s}$, $\mathrm{G_{i/o}}$, $\mathrm{G_{q/11}}$ and $\mathrm{G_{12/13}}$), 4 for predicting the coupling between GPCRs and G proteins (GPCRs-$\mathrm{G_s}$, GPCRs-$\mathrm{G_{i/o}}$, GPCRs-$\mathrm{G_{q/11}}$ and GPCRs-$\mathrm{G_{12/13}}$).

Recognizing G proteins from non-G proteins or GPCRs from non-GPCRs and the prediction of coupling between GPCRs and G proteins can be formulated as two-class

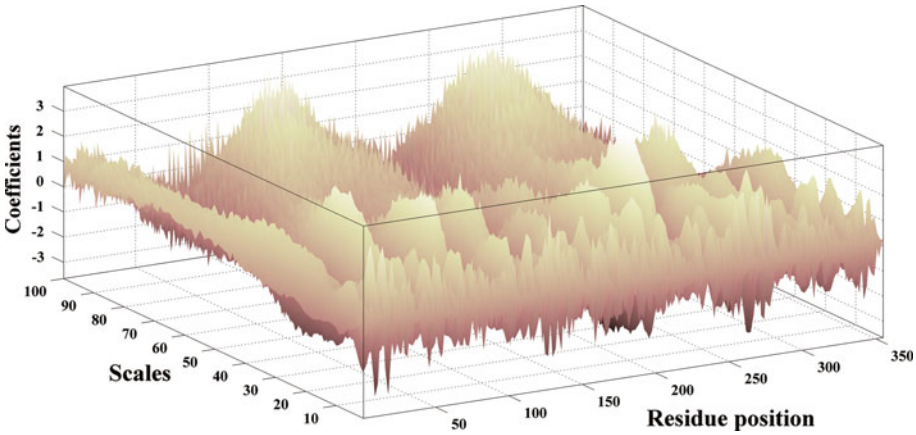

Fig. 1 The results of CWT for primary structure of protein GPR0005

classification problem, namely each protein can be classified as G protein/non-G protein, GPCR/non-GPCR or two proteins can be predicted as coupled or non-coupled. So, the performance of classifier can be evaluated using accuracy (Acc), sensitivity (Sen), specificity (Spe):

$$\mathrm{Acc} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{4}$$

$$\mathrm{Sen} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{5}$$

$$\mathrm{Spe} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \tag{6}$$

where, TP, TN, FP and FN are the numbers of true positives, true negatives, false positives and false negatives, respectively. The classification of G proteins at level of class and family is a multi-class classification problem, namely a given protein can be classified into specific class or family. We adopted the one-versus-one strategy to transfer it into a series of two-class problems. The overall accuracy (OA) and accuracy (A_i) for each class or family calculated for assessment of the classifier are given by Eqs. 7–8.

$$\mathrm{OA} = \frac{\sum_{i=1}^{k} p(i)}{\mathrm{Nu}} \tag{7}$$

$$Ai = \frac{p(i)}{\mathrm{obs}(i)} \tag{8}$$

where, Nu is the total number of proteins, obs(i) is the number of proteins observed in class i or family i, $p(i)$ is the number of correctly predicted proteins of class i or family i.

The diagrammatic view of the prediction system is shown in Fig. 2, and the steps are described as follows:

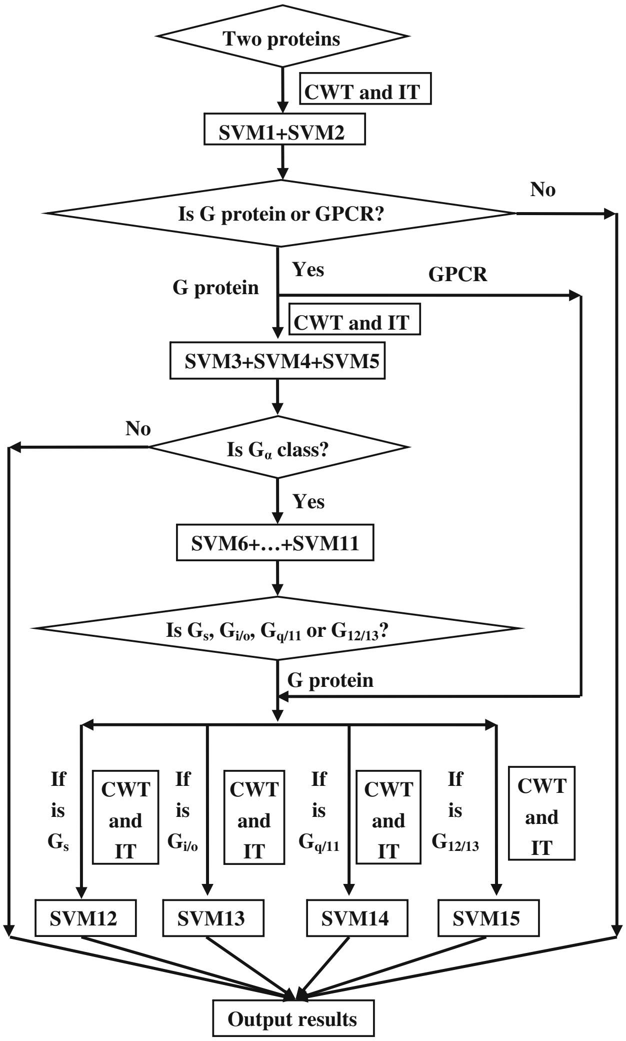

Fig. 2 The diagrammatic view of the prediction system

Step 1. Two feature vectors that represent two query proteins are constructed based on the CWT and IT.

Step 2. The two vectors are entered in SVM1 and SVM2, respectively. Predict whether the two query proteins belong to GPCR and G protein. If the two proteins are classified into GPCR and G protein, go to the next step. Otherwise, the process is stopped with the output of results.

Step 3. The vector of G protein is fed to SVM3, SVM4 and SVM5, respectively. Predict whether the G protein belongs to the G-α class. If the protein is classified into G-α class, go to the next step. Otherwise, the process is stopped with the output of results.

Step 4. The vector of G protein is entered in SVM6, SVM7,…, SVM10 and SVM11, respectively. Predict which family the protein belongs to: If the protein belongs to the G_s family, go to step 5. If it belongs to the $G_{i/o}$ family, go to step 6. If it belongs to the $G_{q/11}$ family, go to step 7. If it belongs to the $G_{12/13}$ family, go to step 8.

Step 5. A new feature vector that represents coupled pair of GPCR-G_s is constructed based on the CWT and IT. The vector is fed to SVM12, and classify as coupled or non-coupled. The process is stopped with the output of results.

Step 6. A new feature vector that represents coupled pair of GPCR-$G_{i/o}$ is constructed. The vector is fed to SVM13, and predict as coupled or non-coupled. The process is stopped with the output of results.

Step 7. A new feature vector that represents coupled pair of GPCR-$G_{q/11}$ is constructed. The vector is fed to SVM14, and identify as coupled or non-coupled. The process is stopped with the output of results.

Step 8. A new feature vector that represents coupled pair of GPCR-$G_{12/13}$ is constructed. The vector is entered in SVM15, and identify as coupled or non-coupled. The process is stopped with the output of results.

Results

Selection of wavelet function

It is well known that the ability of CWT to elucidate information within the digital signal is highly dependent on the selection of the mother wavelet. To investigate the effect of the different wavelet on prediction accuracy, 18 wavelet functions are selected and used to test, including bior1.1, bior1.5, bior2.2, bior2.4, bior2.6, bior3.1, bior3.5, bior3.9, db1, db5, db9, dmey, haar, mexh, meyr, morl, sym2 and sym6. Firstly, the primary structure of protein is mapped to a digital signal based on the various physicochemical properties. Secondly, 18 wavelet functions are selected and CWT is performed for the digital signal, respectively. Thirdly, the entropy and relative entropy are calculated according to the description in the part of characterization of protein sequence and coupled pair based on CWT and IT. Finally, the entropy and relative entropy are entered in the corresponding model and the predictive results are obtained for the coupling specificity between GPCRs and G proteins. The results based on the different wavelet function and tenfold cross-validation test are illustrated in Fig. 3.

As can be seen from the Fig. 3a, the Acc, Sen and Spe for recognizing G proteins from non-G proteins using the bior1.1 wavelet reach 88.94, 92.04 and 85.84%, respectively. When the morl wavelet is adopted, the Acc, Sen and Spe for discriminating GPCRs from non-GPCRs are 96.15, 97.24 and 95.06%, respectively. When the meyr wavelet is used to classify G proteins at level of class, the highest OA of 98.23%, A_α of 100.0%, A_β of 97.44% and A_γ of 96.15% are obtained. When using the meyr wavelet, the OA for classifying G proteins at level of family is 85.42%. And the classification accuracies for G_s, $G_{i/o}$, $G_{q/11}$ and $G_{12/13}$ families are 90.00, 88.89, 82.35 and 66.67%, respectively. Please note that these results are from threefold cross-validation, because the $G_{12/13}$ family contains only three proteins. The classification accuracy of $G_{12/13}$ family is significantly lower than that of G_s, $G_{i/o}$ and $G_{q/11}$ family, indicating that the number of proteins contained in the family is too few to have statistical significance.

It is very clear from Fig. 3e that the performance of the predictor is the best using the meyr wavelet, and the Acc of 94.30%, Sen of 95.72% and Spe of 92.87% are obtained to predict whether the coupling is between a GPCR protein and a G_s protein. When using the bior3.9 wavelet, Acc of 74.60%, Sen of 76.19% and Spe of 73.02% are obtained to distinguish whether the coupling is between a GPCR protein and a $G_{i/o}$ protein. When the morl wavelet is used to identify whether the coupling is between a GPCR protein and a $G_{q/11}$ protein, the highest Acc of 92.03%, Sen of 95.65% and Spe of 88.41% are obtained. The performance of the classifier based on sym2 wavelet achieved the highest Acc of 90.38%, Sen of 96.15% and Spe of 84.62% for coupling between a GPCR protein and a $G_{12/13}$ protein. These results indicate that the performance of bior1.1, bior3.9, morl, meyr and sym2 wavelet is better than that of other wavelet functions. Therefore, these wavelets of bior1.1, bior3.9, morl, meyr and sym2 are selected to classify and predict coupling specificity between GPCRs and G proteins.

Selection of optimal L in IT

In this study, entropy or relative entropy is computed based on the L data zone between the maximum and minimum values. Therefore, the L may have various optimal values for different prediction or classification models. We follow the standard that the sum of Acc, Sen and Spe reaches the highest value to select the optimal L. The standard is rarely used in existing publications. Using our constructed working dataset, various models are built based on the different values of L and the results derived from the tenfold cross-validation test are illustrated in Fig. 4.

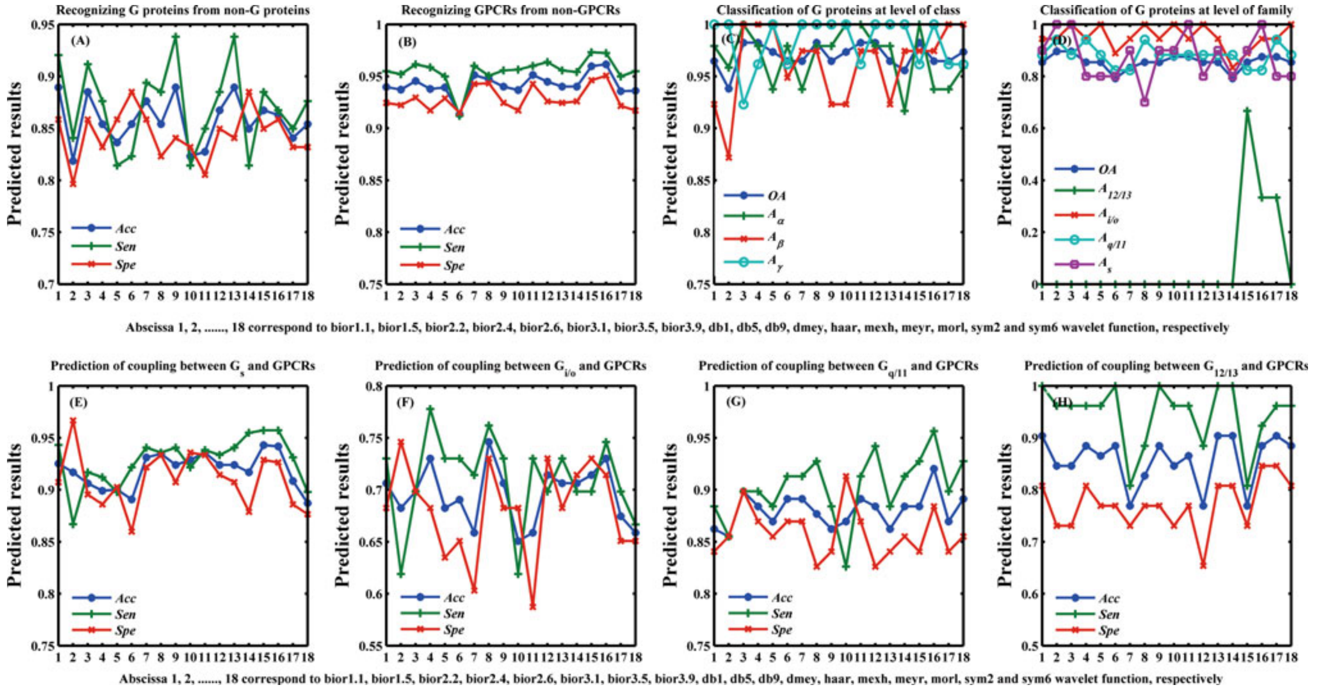

Fig. 3 The results of different wavelet function for various predictors (In each sub-figure, Abscissa and ordinate of each point correspond to wavelet function and predicted result, respectively. Abscissa 1, 2, …, 18 correspond to bior1.1, bior1.5, bior2.2, bior2.4, bior2.6, bior3.1, bior3.5, bior3.9, db1, db5, db9, dmey, haar, mexh, meyr, morl, sym2 and sym6 wavelet function, respectively)

As shown in Fig. 4a–c, when L is equal to 3, the sum of Acc, Sen and Spe reach the highest value for identifying G proteins from non-G proteins/GPCRs from non-GPCRs and classifying G protein at level of class respectively. The best prediction results can be obtained when L is equal to 16, which are derived from classification of G proteins at level of family and prediction of coupling between GPCRs and $G_{i/o}$. As shown in Fig. 4e, the prediction accuracies increase when L increases from 3 to 13, and the best results are obtained when L is equal to 14. The results in Fig. 4g indicate that the sum of Acc, Sen and Spe increases when L increases from 2 to 4, and fluctuates when L increases from 8 up to 16. The sum reaches the highest value when L is equal to 7. We can see from Fig. 4h, the sum of Acc,

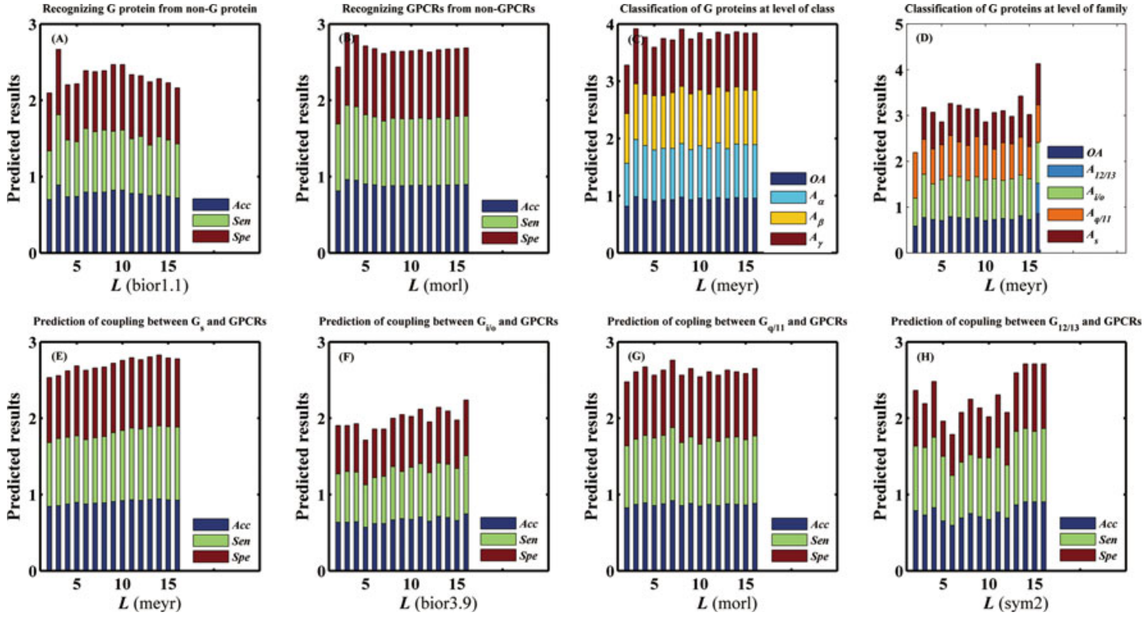

Fig. 4 The effect of different L value on various predictors

Sen and Spe for prediction of coupling specificity between GPCRs and $G_{12/13}$ is equal and reaches the highest value when L is equal to 14, 15 and 16. These results suggest that the value of L has great effect on the prediction accuracies.

Effect of negative dataset to model performance

In this work, the number of negative sample is equal to the number of positive sample to facilitate the SVM modeling. However, it is not always consistent with the real situation where negative sample is far more abundant than positive sample. Therefore, the negative sample selected only once may be not sufficient to randomly sample. To test the robustness of the random selection of negative sample, the process of sampling is repeated nine times and the statistical average results based on the tenfold cross-validation test using the optimal wavelet and L are listed in Table 1.

As shown in Table 1, the average Acc, Sen and Spe to identify G proteins from non-G proteins are 88.00, 88.79, 87.22% and 95.22, 96.38, 94.04% to identify GPCRs from non-GPCRs, respectively. The average Acc, Sen and Spe are 92.33, 93.53 and 91.13% for prediction of coupling between GPCRs and $G_{s,}$ respectively. These results are close to the optimal results (the results are shown in Fig. 1) of the first sample of negative sample. The average Acc, Sen and Spe derived from GPCRs-$G_{i/o}$ and GPCRs-$G_{q/11}$ are only about 4% lower than those derived from the first sample. In addition, we can see that the standard deviation of these predictions is very low. These results indicate that random sample is reasonable for identification of G proteins from non-G proteins, GPCRs from non-GPCRs, and coupling between GPCRs and G_s, $G_{i/o}$ and $G_{q/11}$. Negative sample selected only once is also sufficient. However, we can also see from Fig. 1 and Table 1 that the average Acc and Spe for prediction of coupling between GPCRs and $G_{12/13}$ are 83.97 and 73.50%, respectively, and about 7 and 11% lower than those of the first sample. And the standard deviation for the prediction is a little high. The results indicate that the negative sample selected only once may not be sufficient for prediction of coupling between GPCRs and $G_{12/13}$. We also note that the average Spe is about 20% lower than the average Sen for the prediction of coupling of GPCRs-$G_{12/13}$, which may be caused by the fact that the negative sample contains a number of false-negative samples. But, this problem can be overcome with the cumulative data.

Comparison with existing methods

To demonstrate the performance of the current method, we perform a comparison with the methods of GRIFFIN (Yabuki et al. 2005) and PRED-COUPLE2 (Sgourakis et al. 2005a). GRIFFIN assumes that the ligands, GPCRs and G proteins form complex. And the method uses structural information about the ligands and GPCRs as well as hidden Markov model and SVM to predict the coupling specificity between GPCRs and G proteins. Similar to our method, PRED-COUPLE2 does not require membrane topology information, but uses artificial neural network (ANN) and hidden Markov model to predict coupling. The two methods are similar to most existing methods (i.e. only predicts the coupling preference of GPCRs to G proteins). However, the current method can not only directly predict coupling between arbitrary combinations of GPCRs and G proteins but also indirectly predict coupling preference.

We retrieve 158 GPCRs from IUPHAR-DB database (Harmar et al. 2009) and download the corresponding sequences from http://www.uniprot.org/. In these GPCRs, 81, 52, 21 and 4 proteins have coupling reference to $G_{i/o}$, $G_{q/11}$, G_s and $G_{12/13}$, respectively. Any one of these proteins is not included in the dataset, which is used to train GRIFFIN, PRED-COUPLE2 and GPCRsG_CWTIT models. These proteins of 158 GPCRs are submitted to two Web servers at http://griffin.cbrc.jp/ and http://athina.biol.uoa.gr/bioinformatics/PRED-COUPLE2/, and predicted results with GRIFFIN and PRED-COUPLE2 are listed in Table 2.

Further, we generated 1,458 (81×18) coupled pairs of GPCRs-$G_{i/o}$, 884 (52 × 17) coupled pairs of GPCRs-$G_{q/11}$, 210 (21 × 10) coupled pairs of GPCRs-G_s and 12 (4×3) coupled pairs of GPCRs-$G_{12/13}$ by pairing these GPCRs and proteins of $G_{i/o}$, $G_{q/11}$, G_s and $G_{12/13}$ class that appeared in the current working dataset. Obviously, if a GPCR protein can be coupled with any one protein that belongs to $G_{i/o}$ or $G_{q/11}$ or G_s or $G_{12/13}$, the GPCR protein has coupling preference to $G_{i/o}$ or $G_{q/11}$ or G_s or $G_{12/13}$ and can be divided into corresponding class. Therefore, these coupled pairs are entered in corresponding SVM model and predicted results are transformed into the results of coupling preference prediction. The final results with GPCRsG_CWTIT are also listed in Table 2 to compare with GRIFFIN and PRED-COUPLED2.

Table 1 The statistical average results based on the optimal wavelet and L

	G proteins/non-G proteins	GPCRs/non-GPCRs	GPCRs-G_s	GPCRs-$G_{i/o}$	GPCRs-$G_{q/11}$	GPCRs-$G_{12/13}$
Acc/std	0.8800/0.0160	0.9522/0.0029	0.9233/0.0095	0.7063/0.0235	0.8824/0.0217	0.8397/0.0385
Sen/std	0.8879/0.0297	0.9638/0.0041	0.9353/0.0085	0.7196/0.0328	0.9002/0.0211	0.9444/0.0513
Spe/std	0.8722/0.0226	0.9406/0.0058	0.9113/0.0121	0.6931/0.0276	0.8647/0.0324	0.7350/0.0621

std standard deviation (n = 10, where n is the number of parallel experiment)

Table 2 Comparison with existing methods

Methods	$A_{i/o}$	$A_{q/11}$	A_s	$A_{12/13}$
GRIFFIN[a]	0.3686	0.8077	0.6190	N/A[c]
GRIFFIN[b]	0.1429	0.5000	0.2500	N/A[c]
PRED-COUPLE2[a]	0.6173	0.4038	0.2857	0.0000
PRED-COUPLE2[b]	0.4286	0.0000	0.0000	0.0000
GPCRsG_CWTIT[a]	0.8765	0.9231	0.8095	0.7500
GPCRsG_CWTIT[b]	0.7143	0.8333	0.2500	0.0000

[a] Results are from the 158 GPCRs

[b] Results are from the 18 GPCRs

[c] GRIFFIN only predict coupling preference to $G_{i/o}$, $G_{q/11}$ and G_s class

As shown in Table 2, the accuracies of $G_{i/o}$, $G_{q/11}$ and G_s derived from the GRIFFIN are 36.86, 61.73, 80.77%, and 40.38, 61.90, 28.57% from PRED-COUPLE2, respectively. The accuracy of coupling preference for $G_{i/o}$ based on the proposed method is 87.65%, which is almost 51 and 26% higher than those of GRIFFIN and PRED-COUPLE2, respectively. The accuracy of $G_{q/11}$ is 92.31%, about 12 and 52% higher than those of GRIFFIN and PRED-COUPLE2. The accuracy by the current method for G_s is 80.95%, which is 21 and 52% improvement over GRIFFIN and PRED-COUPLE2, respectively. Although the accuracy of the $G_{12/13}$ derived from GPCRsG_CWTIT is only 75%, the accuracy by PRED-COUPLE2 is 0.

To further demonstrate the improved results are not from the high sequence similarity, 18 GPCRs are extracted from the 158 GPCRs, and the pairwise identity between the 18 GPCRs and three dataset is always lower than 40%. In the 18 GPCRs, 7, 6, 4 and 1 proteins have coupling preference to $G_{i/o}$, $G_{q/11}$, G_s and $G_{12/13}$, respectively. According to the proposed method, various coupled pairs are generated and entered into corresponding model. And the 18 GPCRs are also submitted to GRIFFIN and PRED-COUPLE2. Then the results are listed in Table 2. It is shown that the accuracies of $G_{i/o}$ and $G_{q/11}$ based on the current method are still higher than those based on GRIFFIN and PRED-COUPLE2. The accuracy of G_s derived from the current approach is equal to that from GRIFFIN and higher than that from PRED-COUPLE2. The results further show that our method is superior to other existing methods, and accuracy improvement does not come from the high sequence similarity.

To show statistical significance of the current method, we perform bootstrapping by selecting 80% of our testing data and calculate the accuracy on 80% of the data. This process is repeated 1,000 times and the results are listed in Table 3. As illustrated in Table 3, the mean accuracy of the current method is always higher than that of the GRIFFIN and PRED-COUPLE2, and the standard deviation of the proposed method is always lower than that of the GRIFFIN and PRED-COUPLE2, indicating the current method has better performance than other existing method. Meanwhile, we also perform *t* test for a statistical test to further investigate the statistical significance between the proposed method and GRIFFIN/PRED-COUPLE2 by bootstrapping and repeat 1,000 times. Although 18 proteins are not enough data to prove the results, we perform *t* test and the *P* values between the different methods are listed in Table 4. As illustrated in Table 4, *P* value between GPCRsG_CWTIT and GRIFFIN/PRED-COUPLE2 is always lower than 10^{-200}, and the mean accuracy and the standard deviation of the proposed method is higher and lower than that of GRIFFIN and PRED-COUPLED2, respectively. The results indicate that the current method has more statistical significance than GRIFFIN and PRED-COUPLE2 methods. In one word, the results based on the current method are higher than those based on GRIFFIN and PRED-COUPLE2. Therefore, GPCRsG_CWTIT may serve as a powerful complementary tool for the prediction of GPCRs coupling preference.

Application of GPCRsG_CWTIT: two case studies

To demonstrate the capacity of GPCRsG_CWTIT, the coupling between human bradykinin B2 receptor and G proteins as well as human thyrotropin receptor and G proteins is predicted. The coupling information about human bradykinin B2 receptor, human thyrotropin receptor and G proteins is derived from the database of Human-gpDB at http://bioinformatics.biol.uoa.gr/human_gpdb/. The human bradykinin B2 receptor is encoded by the bkdrb2 gene in humans and belongs to the class A, bradykinin receptors family of GPCRs. The receptor couples six different G proteins belonging to $G_{q/11}$ and $G_{12/13}$ family. Three coupled pairs from the six coupled pairs are contained in the current working dataset. GPCRsG_CWTIT correctly classifies the six G proteins at level of family and predicts four coupled pairs. The human thyrotropin receptor, also known as thyroid-stimulating hormone receptor, is a member of the rhodopsin-like GPCR family and plays a most important role in controlling thyroid cell metabolism.

Table 3 The statistical average results based on the various methods

Methods	$A_{i/o}$/std	$A_{q/11}$/std	A_s/std	$A_{12/13}$/std
GRIFFIN[a]	0.3076/0.0507	0.8085/0.0538	0.6138/0.1028	N/A[c]
GRIFFIN[b]	0.1474/0.1753	0.5030/0.2009	0.2627/0.2224	N/A[c]
PRED-COUPLE2[a]	0.6153/0.0542	0.4075/0.0683	0.2852/0.0974	0.0000/0.0000
PRED-COUPLE2[b]	0.4360/0.1842	0.0000/0.0000	0.0000/0.0000	0.0000/0.0000
GPCRsG_CWTIT[a]	0.8772/0.0368	0.9241/0.0366	0.8122/0.0842	0.7542/0.2162
GPCRsG_CWTIT[b]	0.7210/0.1672	0.8328/0.1593	0.2440/0.2203	0.0000/0.0000

std standard deviation ($n = 1{,}000$, where n is the number of parallel experiment)

[a] Results are from the 158 GPCRs

[b] Results are from the 18 GPCRs

[c] GRIFFIN only predict coupling preference to $G_{i/o}$, $G_{q/11}$ and G_s class

Table 4 The results derived from the *t* test

	$A_{i/o}$	$A_{q/11}$	A_s	$A_{12/13}$
GPCRsG_CWTIT versus GRIFFIN[a]	<2.2251E−308	<2.2251E−308	5.1273E−285	N/A[c]
GPCRsG_CWTIT versus GRIFFIN[b]	<2.2251E−308	5.3084E−291	N/A[d]	N/A[c]
GPCRsG_CWTIT versus PRED-COUPLE2[a]	<2.2251E−308	<2.2251E−308	<2.2251E−308	<2.2251E−308
GPCRsG_CWTIT versus PRED-COUPLE2[b]	3.5397E−230	<2.2251E−308	5.6396E−216	N/A[d]

[a] Results are from the 158 GPCRs

[b] Results are from the 18 GPCRs

[c] GRIFFIN only predict coupling preference to $G_{i/o}$, $G_{q/11}$ and G_s class

[d] Because the accuracy based on the different method is equal, *t* test is not performed

The activity of the receptor is mediated by coupling to 13 G proteins that belongs to all four G-α family. Four coupled pairs from 13 coupled pairs are contained in the current working dataset. GPCRsG_CWTIT successfully classifies 12 from the 13 G proteins at level of family and identifies 10 coupled pairs. The results indicate that GPCRsG_CWTIT method can yield high prediction accuracy and is useful for prediction of coupling specificity between GPCRs and G proteins.

Discussion

In this study, a set of physicochemcial properties of amino acids are used to convert primary sequence into digital signal. The CWT and IT are utilized to extract features, which can effectively characterize the information of protein sequence and the coupling specificity between GPCRs and G proteins. After that, a novel predictor called GPCRsG_CWTIT is proposed and used to predict the coupling specificity between GPCRs and G proteins.

The performance of GPCRsG_CWTIT is evaluated against a well-curated dataset, which consisted of 1,376 GPCRs and non-GPCR proteins, 113 G and non-G proteins, 421 coupled and non-coupled pairs of GPCRs-G_s, 63 coupled and non-coupled pairs of GPCRs-$G_{i/o}$, 69 coupled and non-coupled pairs of GPCRs-$G_{q/11}$ as well as 23 coupled and non-coupled pairs of GPCRs-$G_{12/13}$. The results indicate that GPCRsG_CWTIT can accurately predict the coupling specificity between G proteins and GPCRs. The prediction accuracy of GPCRsG_CWTIT is dependent on the selection of mother wavelet function, and the best results for various prediction tasks are not obtained for any kind of wavelet function. The information contained in different protein sequence and coupled pair is different, and it can be precisely expressed with specific wavelet function. In addition, the results of L value optimization indicate that the information included in different class of protein sequence usually can be properly characterized by entropy with L less than 5, and some noise will be introduced by larger L. On the contrary, relative entropy with larger L can be used to describe the information of the coupled pair and non-coupled pair, and smaller L may lose some useful features.

In the real situation, the number of negative sample may be very higher than that of positive sample. Therefore, a large dataset of negative sample should be constructed to reflect the real situation. Unfortunately, a large negative sample dataset usually causes the trained model preferentially to correctly predict negative sample rather than positive sample (i.e. very high specificity and low sensitivity). So, the balanced positive and negative datasets are constructed by random sampling to prevent skewing the prediction of the positive or negative sample. The results of repeated random sampling indicate that the negative data selected only once are sufficient for most of the prediction task.

In contrast to existing computation method, our method only requires primary sequence of a protein. More importantly, the proposed method can not only indirectly predict coupling preference of GPCRs to G proteins, but also classify G proteins at level of class and family as well as predict couplings between any GPCRs and any G proteins. And the work is not reported in any previous publication. The high prediction accuracies indicate that the current method is an effective and feasible tool for prediction of coupling specificity using only the protein full sequence. Compared with the experimental methods, the presented GPCRsG-CWTIT only takes a few seconds to predict whether the two proteins belong to the GPCR or G protein, which class and family the G protein belongs to and whether coupling between the GPCR and G protein takes place.

The establishment of such an accurate prediction method will facilitate drug discovery by improving the ability to identify and predict protein–protein interactions. Of course, the current prediction system also has its own limitation. For example, the current method only classifies G proteins at level of class and family. In fact, each family can be further classified into different subfamilies and types. And knowledge of a G protein about subfamily and type is more important for signal transduction research.

Acknowledgments This work was supported by National Natural Science Foundation of China (No. 20975117; 21175158; 81171666), Natural Science Foundation of Guangdong Province (No. 10151027501000070), Scientific Technology Project of Guangdong Province (No. 2010A040302001; 2011B031800257) and Fundamental research funds for the central universities (No. 10lgzd13).

Conflict of interest The authors declare that they have no conflict of interest.

References

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

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

Bhasin M, Raghava GP (2005) GPCRsclass: a web tool for the classification of amine type of G-protein-coupled receptors. Nucleic Acids Res 33:W143–W147

Cabrere-Vera TM, Vanhauwe J, Thomas TO, Medkova M, Preininger A, Mazzoni MR, Hamm HE (2003) Insight into G protein Structure, function, and regulation. Endocr Rev 24:765–781

Cao J, Panetta R, Yue S, Steyaert A, Young-Bellido M, Ahmad S (2003) A naïve Bayes model to predict coupling between seven transmembrane domain receptors and G-proteins. Bioinformatics 19:234–240

Chan CC, Lin CJ (2001) LIBSVM: a library for support vector machines. http://www.csie.ntu.edu.tw/~cjlin/libsvm

Chou KC (2005) Prediction of G-protein-coupled receptor classes. J Proteome Res 4:1413–1418

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

Elefsinioti AL, Bagos PG, Spyropoulos IC, Hamodrakas SJ (2004) A database for G proteins and their interaction with GPCRs. BMC Bioinform 5:208

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

Eo HS, Choi JP, Noh SJ, Hur GG, Kim W (2007) A combined approach for the classification of G protein-coupled receptors and its application to detect GPCR splice variants. Comput Biol Chem 31:246–256

Gao QB, Wang ZZ (2006) Classification of G-protein coupled receptors at four levels. Protein Eng Des Sel 19:511–516

Gaulton A, Attwood TK (2003) Bioinformatics approaches for the classification of G-protein-coupled receptors. Curr Opin Pharmacol 3:114–120

Greasley PJ, Fanelli F, Scheer A, Abuin L, Nenniger-Tosato M, DeBenedetti PG, Cotecchia S (2001) Mutational and computational analysis of the α(1b)-adrenergic receptor. Involvement of basic and hydrophobic residues in receptors activation and G protein coupling. J Biol Chem 276:46485–46494

Guo Y, Li M, Lu M, Wen Z, Huang Z (2006a) Predicting G-protein coupled receptors-G-protein coupling specificity based on auto-cross-covariance transform. Proteins 65:55–60

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

Guo Y, Yu L, Wen Z, Li M (2008) Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences. Nucleic Acids Res 36:3025–3030

Harmar AJ, Hills RA, Rosser EM, Jones M, Buneman OP, Dunbar DR, Greenhill SD, Hale VA, Sharman JL, Bonner TI, Catterall WA, Davenport AP, Delagrange P, Dollery CT, Foord SM, Gutman GA, Laudet V, Neubig RR, Ohlstein EH, Olsen RW, Peters J, Pin JP, Ruffolo RR, Searls DB, Wright MW, Spedding M (2009) IUPHAR-DB: the IUPHAR database of G protein-coupled receptors and ion channels. Nucleic Acids Res 37:D680–D685

Horn F, van der Wenden EM, Oliveira L, Jzerman AP, Vriend G (2000) Receptors coupling to G proteins: is there a signal behind the sequence? Proteins 41:448–459

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

Kukkonen JP, Nasman J, Akerman KE (2001) Modelling of promiscuous receptor-Gi/Gs-protein coupling and effector response. Trends Pharmacol Sci 22:616–622

Li W, Jaroszewski L, Godzik A (2001) Clustering of highly homologous sequences to reduce the size of large protein databases. Bioinformatics 17:282–283

Li Z, Zhou X, Dai Z, Zou X (2010) Classification of G-protein coupled receptors based on support vector machine with maximum relevance minimum redundancy and genetic algorithm. BMC Bioinform 11:325

Moller S, Vilo J, Croning MD (2001) Prediction of the coupling specificity of G protein coupled receptors to their G proteins. Bioinformatics 17:S174–S181

Oliveira L, Paiva AC, Vriend G (1999) A low resolution model for the interaction of G proteins with G protein-coupled receptors. Protein Eng 12:1087–1095

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

Pierce KL, Premont RT, Lefkowitz RJ (2002) Seven-transmembrane receptors. Nat Rev Mol Cell Biol 3:639–650

Qian B, Soyer OS, Neubig RR, Goldstein RA (2003) Depicting a protein's two faces: GPCR classification by phylogenetic tree-based HMMs. FEBS Lett 554:95–99

Qiu JD, Huang JH, Liang RP, Lu XQ (2009) 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 390:68–73

Sgourakis NG, Bagos PG, Hamodrakas SJ (2005a) Prediction of the coupling specificity of GPCRs to four families of G-proteins using hidden Markov models and artificial neural networks. Bioinformatics 21:4101–4106

Sgourakis NG, Bagos PG, Papasaikas PK, Hamodrakas SJ (2005b) A method for the prediction of GPCRs coupling specificity to G-proteins using refined profile hidden Markov models. BMC Bioinform 6:104

Shannon CE (1948) A mathematical theory of communication. Bell System Technical J 27:379–423

Simon MI, Strathmann MP, Gautan M (1991) Diversity of G proteins in signal transduction. Science 252:802–808

Sreekumar KR, Huang Y, Pausch MH, Gulukota K (2004) Predicting GPCR-G protein coupling using hidden Markov models. Bioinformatics 20:3490–3499

Theodoropoulou MC, Bagos PG, Spyropoulos IC, Hamodrakas SJ (2008) gpDB: a database of GPCRs, G-proteins, effectors and their interactions. Bioinformatics 24:1471–1472

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680

Wess J (1998) Molecular basis of receptor/G-protein-coupling selectivity. Pharmacol Ther 80:231–264

Yabuki Y, Muramatsu T, Hirokawa T, Mukai H, Suwa M (2005) GRIFFIN: a system for predicting GPCR-G-protein coupling selectivity using a support vector machine and a hidden Markov model. Nucleic Acids Res 33:W148–W153