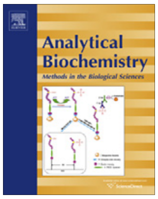
Analytical Biochemistry 412 (2011) 173–182

Contents lists available at ScienceDirect

Analytical Biochemistry

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



G-protein-coupled receptor prediction using pseudo-amino-acid composition and multiscale energy representation of different physiochemical properties

Zia ur-Rehman, Asifullah Khan *

Department of Computer and Information Sciences, Pakistan Institute of Engineering and Applied Sciences (PIEAS), Nilore, Islamabad 45650, Pakistan

ARTICLE INFO

Article history:
Received 31 October 2010
Received in revised form 26 January 2011
Accepted 27 January 2011
Available online 2 February 2011

Keywords:
GPCR classification
Multiscale energy
Pseudo-amino-acid composition
Physiochemical properties
Rhodopsin-like receptors

ABSTRACT

G-protein-coupled receptors (GPCRs) are the largest family of cell surface receptors that, via trimetric guanine nucleotide-binding proteins (G-proteins), initiate some signaling pathways in the eukaryotic cell. Many diseases involve malfunction of GPCRs making their role evident in drug discovery. Thus, the automatic prediction of GPCRs can be very helpful in the pharmaceutical industry. However, prediction of GPCRs, their families, and their subfamilies is a challenging task. In this article, GPCRs are classified into families, subfamilies, and sub-subfamilies using pseudo-amino-acid composition and multiscale energy representation of different physiochemical properties of amino acids. The aim of the current research is to assess different feature extraction strategies and to develop a hybrid feature extraction strategy that can exploit the discrimination capability in both the spatial and transform domains for GPCR classification. Support vector machine, nearest neighbor, and probabilistic neural network are used for classification purposes. The overall performance of each classifier is computed individually for each feature extraction strategy. It is observed that using the jackknife test the proposed GPCR–hybrid method provides the best results reported so far. The GPCR–hybrid web predictor to help researchers working on GPCRs in the field of biochemistry and bioinformatics is available at http://111.68.99.218/GPCR.

G-protein-coupled receptors (GPCRs)[1] are known to play an essential role in the coordination of cellular communications and are involved in many physiological processes. They play important roles in various mammalian disorders, including allergies, cardiovascular dysfunction, depression, cancer, pain, diabetes, and central nervous system disorders. They consist of seven transmembrane alpha helices, an intracellular C terminal, an extracellular N terminal, three intracellular loops, and three extracellular loops. They can activate signaling pathways that control gene expression and cell proliferation, serving as crucial mediators for various cellular signal transduction events that provide the means for cells, tissues, organs, and organisms to react properly to the changing environment [1]. They are also widely expressed in the central nervous system, where they mediate and modulate synaptic transmission in the brain and spinal cord. GPCRs play an important role in drug discovery. The location of GPCRs on a cell makes them readily accessible to drugs. More than 50% of the current drug targets are focused on GPCRs [1,2]. At least 55 GPCR types are known for directly mediating neuronal and endocrine regulation of cardiac and vascular responses. In addition, many GPCRs are known to influence cardiovascular functions. Their role in the development of cancer is becoming apparent, and that is why GPCRs are the emerging targets for therapeutic interventions to treat cancer.

GPCRs consist of different amino acid sequences, and based on the sequence homology, these are divided into six families [3]: rhodopsin-like receptors (class A), secretin-like receptors (class B), metabotropic glutamate receptors (class C), fungal mating pheromone receptors (class D), cyclic AMP (cAMP) receptors (class E), and frizzled/smoothened receptors (class F). The GPCR classes A, B, C, and F are found mostly in mammals, class D GPCRs are found only in fungi, and class E GPCRs are found in *Dictyostelium*. Rhodopsin-like receptors are the biggest family of GPCRs, constituting 80% of all GPCRs. They are used to bind peptides, biogenic amines, or lipid-like substances [4]. The secretin receptors bind large peptides such as secretin, parathyroid hormone, glucagon, vasoactive intestinal peptide growth hormone-releasing hormone, and pituitary adenylyl cyclase-activating protein [5]. The metabotropic glutamate receptors are activated through an indirect metabotropic process [6]. Fungal mating pheromone receptors are used for chemical communication in various organisms [7]. Similarly, the cAMP receptors form a part of the chemotactic signaling system of slime molds [8]. On the other hand, frizzled/smoothened receptors are necessary for Wnt binding and the

* Corresponding author. Fax: +92 51 2208070.
E-mail addresses: asif@pieas.edu.pk, khan.asifullah@gmail.com (A. Khan).

[1] *Abbreviations used:* GPCR, G-protein-coupled receptor; cAMP, cyclic AMP; PseAA, pseudo-amino-acid composition; MSE, multiscale energy; AA, amino acid composition; SVM, support vector machine; NN, nearest neighbor; PNN, probabilistic neural network; EIIP, electron ion interaction pseudopotential; CPV, composition, polarity, and molecular volume; DWT, discrete wavelet transform; RBF, radial basis function; *MCC*, Matthews correlation coefficient; PCA, principal component analysis; CA, cellular automaton.

0003-2697/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.ab.2011.01.040

mediation of hedgehog signaling, a key regulator of animal development [9]. Each family is divided into subfamilies, and each subfamily is further divided into sub-subfamilies. The classification of GPCRs into families, subfamilies, and sub-subfamilies is done based on the functionalities performed by each GPCR sequence, grouping GPCR sequences with similar functionalities in the same family. One method among several methods for the prediction of GPCR sequences is to do sequence similarity searches using pairwise alignment tools [10] such as BLAST and FASTA. The second method is to classify GPCRs by conducting biological experiments. During the past decade, hundreds of new GPCRs have been discovered, and they are continuing to grow rapidly. Therefore, their annotation based on the manual experimentation has made it very expensive and nearly impossible. Thus, there was a great need for fast, reliable, and efficient systems that can exploit different properties of GPCRs to annotate their functions automatically. Several statistical and machine learning methods have been proposed in this regard, including the Bayes network method [11], support vector machine [2,12–14], and hidden Markov models [15–17]. Although these methods classify GPCRs with high accuracy, none of them provides hierarchical GPCR classification. The GPCRs are hierarchically classified into four levels—super family, family, subfamily, and type—by Gao [18]. The data set used in this method consists of 1406 GPCR sequences and 1406 globular proteins (non-GPCRs). In the first level, GPCRs are discriminated from non-GPCRs. In the second level, six GPCR families are classified. The rhodopsin-like family is further classified into subfamilies in the third level. In the fourth level, sub-subfamilies of amine subfamily and olfactory subfamily are predicted. GPCRs are classified into three levels—super family, family, and specific receptor subtype—by Attwood and coworkers [19]. GPCRs are also hierarchically classified into three levels by Davies and coworkers [20]. In the first level, GPCRs are classified into 5 families (class F is ignored), whereas in the second level, GPCRs are classified into 40 subfamilies. Finally, in the third level, GPCRs are classified into 108 sub-subfamilies. Davies and coworkers have also developed an online GPCR prediction server (see Ref. [21]). Both of these hierarchical classification methods provide good overall accuracy.

In this article, we have classified GPCRs into three levels. First we classified GPCRs into 5 families, then into 40 subfamilies, and finally into 108 sub-subfamilies, as was done by Davies and coworkers [20]. The frizzled/smoothened receptor family is ignored because it contains too few sequences from which to develop an accurate classification algorithm. Three feature extraction strategies are used. The first one is pseudo-amino-acid composition (PseAA) [22], which is used in two ways: using either two or three physiochemical properties of GPCRs. In the second feature extraction strategy, a hybrid feature vector is formed by combining wavelet-based multiscale energy (MSE) and PseAA-based features (MSE–PseAA) [23]. The third one is also a hybrid feature vector formed by the combination of amino acid composition (AA) and MSE features (MSE–AA). We have used three classifiers, and the jackknife test is used to evaluate the performance of the classifiers for each feature extraction strategy. These three classifiers are support vector machine (SVM), nearest neighbor (NN), and probabilistic neural network (PNN). The aim of this research is to assess different feature extraction strategies and to develop hybrid feature extraction strategies that can exploit the discrimination capability in both spatial and transform domains. We have developed a web predictor, GPCR–hybrid, which takes an unknown GPCR sequence as input and classifies it first into family, then into subfamily, and finally, into sub-subfamily. At each level, the proposed GPCR–hybrid method selects the best performing feature extraction strategy and the classifier to predict the class of the test sequence, as shown in Fig. 1.

The GPCR dataset that we used was taken from the BIAS-PROFS website (http://www.cs.kent.ac.uk/projects/biasprofs) [21]. The overall performance of our proposed approach is better than that of the existing hierarchical GPCR classification methods.

Materials and methods

Data sets

The dataset that we mainly used for the training and assessment of our classification approach was downloaded from the BIAS-PROFS website [21] developed by Davies and coworkers in 2007. GPCR sequences for the dataset were identified using the Entrez search and retrieval system [24]. Text-based searching was used to identify all sequences within each sub-subfamily of the hierarchy. GPCR sequences shorter than 280 amino acids in length were also removed. Finally, all of the identical sequences within the dataset were removed to avoid redundancy. In general, a homology bias is avoided by using a cutoff threshold of 25% [25]. However, in this study, the dataset by Davies and coworkers [20] was not required to adhere to such a stringent criterion because the numbers of GPCRs for some of the classes would be too small to have statistical significance. Hence, we used the dataset by Davies and coworkers as it is and did not apply any additional processing. The dataset consists of 8354 GPCR sequences, of which 5526 sequences belong to rhodopsin-like, 625 belong to secretin-like, 2172 belong to metabotropic glutamate, 13 belong to fungal pheromone, and 18 belong to cAMP receptors family.

In addition, we also used three other benchmark datasets for comparison with existing methods. These datasets were constructed using older versions of GPCRDB, and it has been reported

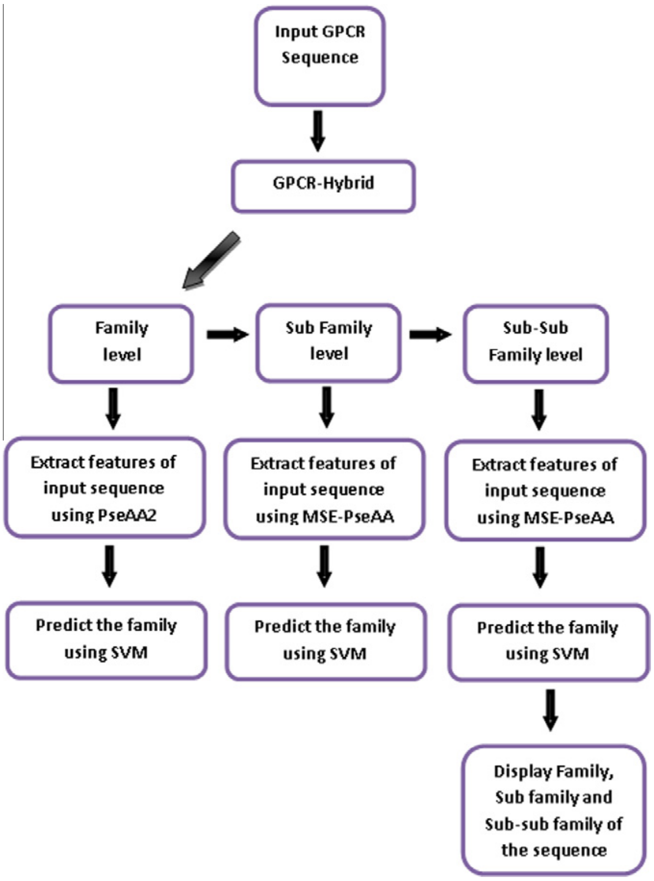

Input GPCR Sequence
GPCR-Hybrid
Family level
Sub Family level
Sub-Sub Family level
Extract features of input sequence using PseAA2
Extract features of input sequence using MSE-PseAA
Extract features of input sequence using MSE-PseAA
Predict the family using SVM
Predict the family using SVM
Predict the family using SVM
Display Family, Sub family and Sub-sub family of the sequence

Fig. 1. GPCR–hybrid method.

that they largely avoid homology bias. For simplicity, they are referred to as the D167, D566, and D365 datasets containing 167, 566, and 365 GPCR sequences, respectively. The GPCRs in the D167 dataset [26] (belonging to the sub-subfamily level) are classified into four sub-subfamilies: (i) acetylcholine, (ii) adrenoceptor, (iii) dopamine, and (iv) serotonin. The D566 dataset [27] (belonging to the sub-subfamily level) contains GPCRs belonging to seven sub-subfamilies: (i) adrenoceptor, (ii) chemokine, (iii) dopamine, (iv) neuropeptide, (v) olfactory type, (vi) rhodopsin, and (vii) serotonin. The D365 dataset [28] (belonging to the family level) contains GPCRs belonging to six families: (i) rhodopsin-like, (ii) secretin-like, (iii) metabotropic glutamate pheromone, (iv) fungal pheromone, (v) cAMP receptor, and (vi) drizzled/smoothened family. Chou and Elrod [26–28] reported that all of the receptor sequences in the above-mentioned datasets were generally lower than 40%, according to their definition of the average sequence identity percentage between two protein sequences.

Sequence representation

Amphiphilic PseAA

To avoid losing much important information hidden in protein sequences, PseAA was proposed [29] to replace the simple AA for representing the sample of a protein. PseAA has been widely used to study various problems in proteins and protein-related systems such as: predicting subcellular location of proteins [30] and GPCR types [31]. However, to the best of our knowledge, so far PseAA has not been used for predicting GPCRs and their types in conjunction with the approach of MSE representation of different physiochemical properties. The current study was devoted to doing so, and quite encouraging results were obtained.

The conventional AA uses only the frequency of occurrence of each amino acid in the GPCR sequence. Instead of conventional AA, the PseAA approach as used in Refs. [23,30] was adopted in the current study. It preserves sequence order and sequence-length information. The GPCR sequence *R* with *L* amino acids, where *L* represents the length of the protein sequence, can be represented as shown by Eq. (1):

$$\mathbf{R} = R_1, R_2, \ldots, R_L, \tag{1}$$

where R_1 represents the amino acid at position 1 and R_L is the amino acid at position *L* in the sequence *R*. Its respective PseAA representation is given in Eq. (2):

$$\mathbf{PseAA} = P_1, P_2, \ldots, P_{20}, \ldots, P_\Lambda, \tag{2}$$

where $\Lambda = 20 + n^{*\lambda}$ (λ is the number of tiers used in PseAA, $\lambda = 0,1,\ldots,m$, and n is the number of physiochemical properties used for each GPCR sequence). The value of λ and the optimal selection physiochemical properties can influence the classification performance. In our case, we selected $\lambda = 21$ and analyzed the performance by using different combinations of physiochemical properties. We took $\lambda = 21$ because it gives the best results in our case. The first 20 elements (i.e., $P_1,P_2,\ldots,P_{20}$) are the occurrence frequencies of the 20 amino acids. The remaining $P_{21},P_{22},\ldots,P_\Lambda$ elements are first-tier to λ-tier correlation factors of amino acid sequences in the GPCR chain. These elements are determined based on physiochemical properties. There are many physiochemical properties. In our current research, we used three physiochemical properties: hydrophobicity, electronic, and bulk. The word "hydrophobic" literally means afraid of water. It is obvious that hydrophobic residues prefer to be in a nonaqueous environment such as a lipid bilayer. Biological molecules can contain large nonpolar regions. These nonpolar regions may also be described as hydrophobic regions. Hydrophobicity of proteins is one of the most important factors in determining a GPCR's structure and function. However, with different experimental conditions and different organic solvents and computing approaches, hydrophobicity values per amino acid will be different. Various scales of hydrophobicity are employed, including KDH, MH, and FH. However, the FH hydrophobicity scale [32] has been shown to be the most discriminative of these hydrophobicity measures [33]. Hence, we used the FH scale for the hydrophobicity measure in the current research. The electronic property has been modeled using the electron ion interaction pseudopotential (EIIP) model [34]. The EIIP value describes the average energy states of all valence electrons of amino acids. Electrons delocalized from the particular amino acid have the strongest impact on the electronic distribution of the whole protein. Hence, we chose the EIIP model for electronic property measurement. Finally, the bulk property has been modeled using the composition, polarity, and molecular volume (CPV) model [35]. Polarity and volume (size) are known to have a great impact on the folding of the protein. Hence, the CPV model was used in the current research to model bulk property.

We assessed the performance of the classifier by first considering two physiochemical properties: electronic and bulk. Next, the third property (hydrophobicity) was also included, slightly enhancing the overall performance. The PseAA using two physiochemical properties is termed as PseAA2. The length of the feature vector in PseAA2 is 62 ($\Lambda = 20 + 2 * 21$). The PseAA using three physiochemical properties is termed as PseAA3. The length of the feature vector in PseAA3 is 83.

Wavelet-based MSE and PseAA-based hybrid feature extraction method

The discrete wavelet transform (DWT) is a representation of signal using an orthonormal basis consisting of countably infinite set of discrete wavelets. There are several methods for implementing DWT, and we used Mallat's fast algorithm in the current method. The basic idea of the fast algorithm is to represent the mother wavelet as a set of high-pass and low-pass filter banks. The signal is passed through the filter banks and decimated by a factor of 2. The outputs of the low-pass filter are wavelet approximation coefficients, and those of the high-pass filter are wavelet detail coefficients. We focused on low-frequency components because the high-frequency components are noisier. This is just like the case of protein internal motions where the low-frequency components are functionally more important.

In this feature extraction strategy, first the GPCR sequences are converted into the numeric form using hydrophobicity values. We used the FH scale for computing hydrophobicity values. The significance of the hydrophobicity property was discussed in the previous section. Each of the amino acids is simply replaced by its corresponding value in the FH scale [32]. The resulting numeric form is homologous to a digital signal. Next, the wavelet (Haar) transform of this digital signal is taken. After that, the approximation and detailed coefficients are calculated. The decomposition level for a sequence is taken as Log_2 (length of sequence). For example, if a sequence length is 8000 amino acids, the decomposition levels for that sequence would be 13. The length of sequences might not be same; hence, zero padding is performed in the case of shorter sequences to keep consistency in the size of the feature vector. The overall feature vector formed in this way is termed as MSE [36]. Hence, the MSE feature vector of $(m + 1)$ – dimensions is formed as given in Eq. (3):

$$\mathbf{MSE}(k) = [d_1^k, d_2^k, \ldots, d_m^k, a_m^k], \tag{3}$$

where $k = 1,2,\ldots,N$, N is the total number of GPCR sequences, d_j^k is the root mean square energy of wavelet detail coefficients in the corresponding jth scale, and a_m^k is the root mean square energy of wavelet approximation coefficients in the mth scale, as shown by Eqs. (4) and (5), respectively:

$$d_{\mathrm{j}^{\mathrm{k}}} = \sqrt{\frac{1}{N_j}\Sigma_{n=0}^{N_j-1}\left\{u_j^k(n)\right\}^2}, \tag{4}$$

$$a_{m^k} = \sqrt{\frac{1}{N_j}\Sigma_{n=0}^{N_m-1}\left\{V_m^k(n)\right\}^2}, \tag{5}$$

where N_j is the number of wavelet detail coefficients, N_m is the number of wavelet approximation coefficients, $u_j^k(n)$ is the nth detail coefficient in the jth scale, and $V_m^k(n)$ is the nth approximate coefficient in the mth scale. The scale here means the decomposition level.

Finally, MSE features are combined with PseAA3 to form the MSE–PseAA feature vector as given by Eq. (6):

$$\textbf{MSE}-\textbf{PseAA} = \left[P_1, P_2, \ldots, P_{20}, \ldots, P_\Lambda, \lambda_1^k, \lambda_2^k, \ldots, \lambda_{m+1}^k\right], \tag{6}$$

where P_1,P_2,…,P_{20},…,t P_Λ are the PseAA features and the remaining ($\lambda_j^k = d_j^k$ and $\lambda_{m+1}^k = a_m^k$) are given by the MSE feature extraction strategy.

Wavelet-based MSE and AA-based hybrid feature extraction method

In this feature extraction strategy, first the GPCR sequences are converted into the numeric form using the FH scale. The amino acid composition calculates the frequency of occurrence of each amino acid in the GPCR sequence. There are 20 amino acids. Hence, a 20-dimensional feature vector is formed, and this is combined with MSE features to form a hybrid feature vector (MSE–AA) as given by Eq. (7):

$$\mathbf{X}_k = \left[P_1^k, P_2^k, \ldots, P_{20}^k, \lambda_1^k, \lambda_2^k, \ldots, \lambda_{m+1}^k\right], \tag{7}$$

where the first 20 features ($P_1^k - P_{20}^k$) are given by amino acid and the remaining ($\lambda_j^k = d_j^k$ and $\lambda_{m+1}^k = a_m^k$) are given by the MSE feature extraction strategy.

Nearest neighbor

The NN algorithm is a method for classifying objects based on nearest training examples in the feature space. A point in the space is assigned to class C if its Euclidean distance to class C is the minimum. Euclidian distance is calculated using Eq. (8):

$$S(\mathbf{X}, \mathbf{X}_i) = 1 - \frac{\mathbf{X}.\mathbf{X}_i}{||\mathbf{X}||\,||\mathbf{X}_i||}(i = 1, 2, \ldots, N). \tag{8}$$

The minimum Euclidean distance is calculated using Eq. (9):

$$S(\mathbf{X}, \mathbf{X}_i) = Min\{S(\mathbf{X}, \mathbf{X}_1), (\mathbf{X}, \mathbf{X}_2), \ldots S(\mathbf{X}, \mathbf{X}_N)\}, \tag{9}$$

where **X**,**X**$_i$ is the dot product of vectors **X** and **X**$_i$, and $\|\mathbf{X}\|$ and $\|\mathbf{X}_i\|$ are their respective Norm. The sample under question is assigned the category corresponding to the training sample **X**$_k$.

Support vector machines

The SVM classifier is inherently a binary classifier, but it can be tailored for multiclassification as well. The SVM model finds a decision surface that has maximum distance to the closest points in the training set. The classification problem is solved as a quadratic optimization problem. The training principle of SVM is to find an optimal linear hyperplane such that the classification error for new test samples is minimized. For linearly separable sample points, hyperplane is determined by maximizing the distance between the support vectors [37–39].

Because our problem is a multiclass problem, we adopted the one-versus-all strategy while using LIBSVM 2.88-1 software. We evaluated the performance of SVM using four different types of kernel: linear (Lin–SVM), polynomial (Poly–SVM), radial basis function (RBF–SVM), and sigmoidal (Sig–SVM). LIBSVM 2.88-1 solves SVM problems using the nonlinear quadratic programming technique. During parameter optimization of SVM models, the average accuracy of the SVM model is maximized.

Probabilistic neural network

The PNN was developed in 1990 by Specht [40] and is based on Bayes theory. It estimates the likelihood of a sample being part of a learned category. The PNN consists of four layers: input, pattern, summation, and decision. The input layer has N nodes, with each corresponding to one independent variable. These input nodes are then fully connected to the M nodes of the pattern layer. The PNN receives n dimensional feature vector as input (i.e., $x_i = x_1, x_2, \ldots, x_n$). This input vector is applied to the input neurons and is passed to the neurons in the pattern layer. Here m_k Gaussian functions are calculated for each class k ($1 \leqslant k \leqslant c$) as given by Eq. (10):

$$P_j^k(x) = \frac{1}{2\pi^{n/2}\left|\Sigma_j^k\right|^{-1/2}} e - \frac{1}{2}\left(x - mu_j^k\right)^T\left(\Sigma_j^k\right)^{-1}\left(x - \mu_j^k\right), \tag{10}$$

where μ_j^k is the mean of the distribution and Σ_j^k is the covariance matrix of the distribution. The summation layer computes the approximation of the class probability functions as given in Eq. (11):

$$\Phi_k(x) = \sum_{j=1}^{mk} \pi_j^k P_j^k(x), \tag{11}$$

where π_j^k is the within-class mixing proportion and $\Sigma_{j=1}^{mk}\pi_j^k = 1$ for $k = 1, 2, \ldots, c$. The decision layer computes the risk as given in Eq. (12):

$$P_k(x) = \sum_{l=1}^{c} V_l^k a_l \Phi_l(x), \tag{12}$$

where α_1 indicates the prior probability and v_1^k is the weight of class l. Hence, the test sample is assigned the label of class for which risk is the minimum.

The PNN calculates most of the terms from the training data. The only one that needs to be optimized is the smoothing factor, which controls the deviations of Gaussian functions. The optimized range of the smoothing factor in our case varies from 0.01 to 5.

Performance measures

In statistical prediction, three cross-validation methods are often used to examine a predictor for its effectiveness in practical application: independent dataset test, subsampling test, and jackknife test. However, among the three cross-validation methods, the jackknife test is deemed as the most objective and can always yield a unique result for a given benchmark dataset; hence, it has been increasingly used by investigators to examine the accuracy of various predictors. Accordingly, the jackknife test was also adopted here to examine the quality of the current predictor. In the jackknife test, one of the sequence patterns is considered as the test sample and the remaining N – 1 sequences are taken as the training patterns. The label of the test sequence is predicted using the rest of the N – 1 training sequences. The process is repeated N times, and the label of each sample is predicted. The performance metrics used for the evaluation of the classifiers are overall accuracy, sensitivity, specificity, Matthews correlation coefficient (*MCC*), and *F* measure. *TP* (true positive) and *TN* (true negative) are the numbers of correctly predicted positive and negative samples, respectively. *FP* (false positive) and *FN* (false negative) are the numbers of incorrectly predicted positive and negative samples, respectively.

Accuracy

Accuracy assesses the overall effectiveness of the algorithm. It is given by Eq. (13):

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} * 100. \tag{13}$$

Sensitivity

$$\text{Sensitivity} = \frac{TP}{TP + FN} * 100. \tag{14}$$

Specificity

$$\text{Specificity} = \frac{TN}{FP + TN} * 100. \tag{15}$$

Matthews correlation coefficient

The *MCC* takes values in the interval of [−1, 1]. A value of 1 means that the classifier never makes any mistakes, and a value of −1 means that the classifier always makes mistakes. The *MCC* is given by Eq. (16):

$$MCC = \frac{(TP * TN) - (FP * FN)}{\sqrt{[TP + FP][TP + FN][TN + FP][TN + FN]}}. \tag{16}$$

F measure

The *F* measure is a measure of the accuracy of a test that considers both the precision and recall of the test to compute the score. The *F* measure can be interpreted as a weighted average of the precision and recall, where an *F* measure reaches its best value at 1 and its worst score at 0:

$$F \text{ measure} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{17}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{18}$$

$$\text{Recall} = \frac{TP}{TP + FN}. \tag{19}$$

Proposed GPCR–hybrid method

The GPCR–hybrid is a web predictor that can efficiently classify unknown GPCR sequences—first into family, then into subfamily, and finally into sub-subfamily. The performance of each classifier is assessed individually for each of the feature extraction strategies. At first, the GPCR–hybrid program asks for the input GPCR sequence using a graphical user interface as shown in Fig. 2. The input GPCR sequence should be in capital letters. As soon as the user clicks on the Submit button, it extracts features of input sequence using the best performing feature extraction strategy of the family level and applies the best performing classifier for predicting the family class. For family level, the best performing feature extraction strategy is PseAA2 and the best performing classifier is SVM. Hence, PseAA2 and SVM are selected by GPCR–hybrid for predicting family class of the test GPCR sequence. Once the family class is predicted, features are extracted again using the best performing feature extraction strategy of the subfamily level. The subfamily class of input sequence is then predicted using the best performing classifier of the subfamily level. MSE–PseAA is selected by GPCR–hybrid for feature extraction at the subfamily level and SVM is used to predict subfamily class. Finally, the sub-subfamily of input sequence is predicted. The sequence is converted into numeric form using MSE–PseAA, and its class is predicted using SVM. The algorithm of GPCR–hybrid is shown in Fig. 1.

After the prediction of family-, subfamily-, and sub-subfamily-level classes, the names of the classes are displayed as shown in Fig. 2.

The proposed GPCR–hybrid is available at http://111.68.99.218/GPCR.

Results and discussion

In the GPCR–hybrid method, the hierarchical classification task is performed in three stages. The first stage predicts the family of the GPCR sequence, the second stage predicts the subfamily of the sequence, and finally the third stage predicts the sub-subfamily of the sequence. We used three feature extraction strategies for the sake of sequence conversion into numeric form. The first feature extraction strategy is PseAA, which is used in two ways. The second feature extraction strategy is called MSE–PseAA, which is a hybrid feature vector formed by combining PseAA-based features with wavelet-based MSE features. The third feature extraction strategy is named MSE–AA, which is also a hybrid feature vector formed by combining MSE-based features with AA-based features. The details of these feature extraction strategies are given in the "Sequence representation" section of Materials and Methods. We used three classifiers to assess the performance for each feature extraction strategy.

At each stage, the best performing classifier with the feature extraction strategy is selected by the GPCR–hybrid program. The details of the prediction results for each level are described in the following sections.

Classification at family level

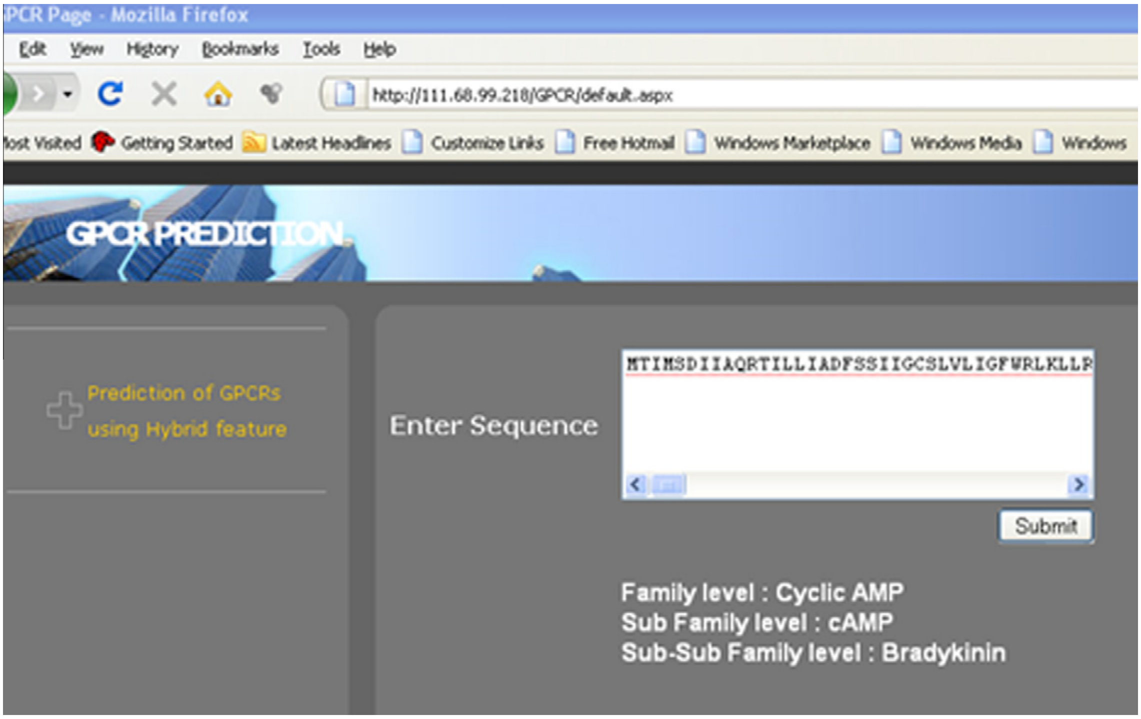
We classified GPCRs into five families. The frizzled and smoothened receptors family (class F) is ignored because current protein databases do not have enough sequences belonging to this family. For performance measurements, we used overall accuracy, sensitivity, specificity, *MCC*, and *F* measure. The formulas of all these measures were described in the "Performance measures" section of Materials and Methods. The performance measurements using NN, PNN, and SVM for each of the feature extraction strategies are described below.

Classifier performance using PseAA2

The overall accuracies achieved by using the NN, PNN, and SVM classifiers for the PseAA2 feature extraction strategy were 97.22%, 97.38%, and 97.86%, respectively. The optimal smoothing factor for PNN was chosen as 1. The *MCC* measures using NN, PNN, and SVM were 0.93, 0.94, and 0.95, respectively. The specificity measures using NN, PNN, and SVM were 96.50%, 96.72%, and 96.89%, respectively. The sensitivity measures using NN, PNN, and SVM were 98.13%, 98.22%, and 98.95%, respectively. The *F* measures using NN, PNN, and SVM were: 0.96, 0.96, and 0.97, respectively.

Classifier performance using PseAA3

The overall accuracies obtained by using the NN, PNN, and SVM classifiers for the PseAA3 strategy were 97.58%, 97.74%, and 93.66%, respectively. The optimal smoothing factor for PNN was chosen as 0.6. The *MCC* measures using NN, PNN, and SVM were 0.94, 0.94, and 0.85, respectively. The specificity measures using NN, PNN, and SVM were 96.96%, 97.16%, and 89.83%, respectively. The sensitivity measures using NN, PNN, and SVM were 98.41%, 98.52%, and 98.04%, respectively. The *F* measures using NN, PNN, and SVM were 0.96, 0.96, and 0.90, respectively.



Enter Sequence
MTIMSDIIAQRTILLIADFSSIIGCSLVLIGFWRLKLLR

Family level : Cyclic AMP
Sub Family level : cAMP
Sub-Sub Family level : Bradykinin

Fig. 2. Graphical user interface of GPCR–hybrid.

Classifier performance using MSE–PseAA

The overall accuracies obtained by using the NN, PNN, and SVM classifiers for the MSE–PseAA strategy were 96.89%, 96.98%, and 97.41%, respectively. The *MCC* measures using NN, PNN, and SVM were 0.92, 0.93, and 0.94, respectively. The specificity measures using NN, PNN, and SVM were 96.01%, 96.16%, and 96.58%, respectively. The sensitivity measures using NN, PNN, and SVM were 97.97%, 98.01%, and 98.43%, respectively. The *F* measures using NN, PNN, and SVM were 0.96, 0.96, and 0.90, respectively.

Classifier performance using MSE–AA

The overall accuracies obtained by using the NN, PNN, and SVM classifiers for the MSE–PseAA strategy were 96.22%, 96.28%, and 97.06%, respectively. The *MCC* measures using NN, PNN, and SVM were 0.91, 0.91, and 0.93, respectively. The specificity measures using NN, PNN, and SVM were 95.08%, 95.22%, and 96.06%, respectively. The sensitivity measures using NN, PNN, and SVM were 97.59%, 97.57%, and 98.23%, respectively. The *F* measures using NN, PNN, and SVM were 0.94, 0.94, and 0.95, respectively.

For family-level classification, PseAA2 using SVM gave the best performance. It had the best accuracy, *MCC*, sensitivity, and *F* measure values, whereas the specificity measure was also comparable. Hence, PseAA2 is used for family-level feature extraction and SVM is used for family-level class prediction. The results for family-level classification are summarized in Table 1.

In Table 1, the performance metrics (accuracy, *MCC*, specificity, sensitivity, and *F* measure) are given in the columns. Their measurements using each of the classifier and feature extraction strategies are given in the rows. It is clearly seen that RBF–SVM showed the best performance using PseAA2.

Classification at subfamily level

We classified GPCRs into 40 subfamilies. The performance measures used at the subfamily level were overall accuracy, sensitivity, and specificity. These performance measurements using NN, PNN, and SVM classifiers are described in the sections given below.

Classifiers performance using PseAA2

The overall accuracies obtained by using the NN, PNN, and SVM classifiers for the PseAA2 strategy were 81.02%, 82.13%, and 81.58%, respectively. The specificity measures using NN, PNN, and SVM for the subfamily level were 80.99%, 82.10%, and 81.55%, respectively. The sensitivity measures using NN, PNN, and SVM were 80.55%, 81.30%, and 81.15%, respectively.

Classifier performance using PseAA3

The overall accuracies achieved by using the NN, PNN, and SVM classifiers for the PseAA3 strategy were 81.88%, 83.47%, and 79.02%, respectively. The specificity measures using NN, PNN, and SVM were 81.85, 83.42, and 78.98, respectively. The sensitivity measures using NN, PNN, and SVM were 81.52%, 83.18%, and 78.85%, respectively.

Classifier performance using MSE–PseAA

The overall accuracies obtained by using the NN, PNN, and SVM classifiers for the MSE–PseAA strategy were 80.73%, 80.36%, and 84.97%, respectively. The specificity measures using NN, PNN, and SVM were 80.69%, 80.27%, and 84.94%, respectively. The sensitivity measures using NN, PNN, and SVM were 80.72%, 81.24%, and 84.08%, respectively.

Classifier performance using MSE–AA

The overall accuracies obtained by using the NN, PNN, and SVM classifiers for the MSE–PseAA strategy were 78.55%, 78.29%, and 80.96%, respectively. The specificity measures using NN, PNN, and SVM were 78.51%, 78.21%, and 81.90%, respectively. The sensitivity measures using NN, PNN, and SVM were 78.51%, 78.79%, and 81.95%, respectively.

For the subfamily-level classification, SVM performed best using the MSE–PseAA feature extraction strategy. The values of

Table 1
GPCR classification performance for family level.

Classifier	Feature extraction strategy	Accuracy (%)	*MCC*	Specificity (%)	Sensitivity (%)	*F* measure
NN	PseAA2[a]	97.22	0.93	96.50	98.13	0.96
PNN	PseAA2	97.38	0.94	96.72	98.22	0.96
SVM	**PseAA2**	**97.86**	**0.95**	96.89	**98.95**	**0.97**
NN	PseAA3[b]	97.58	0.94	96.96	98.41	0.96
PNN	PseAA3	97.74	0.94	**97.16**	98.52	0.96
SVM	PseAA3	93.56	0.85	89.83	98.04	0.90
NN	MSE–AA[c]	96.22	0.91	95.08	97.59	0.94
PNN	MSE–AA	96.28	0.91	95.22	97.57	0.94
SVM	MSE–AA	97.06	0.93	96.06	98.23	0.95
NN	MSE–PseAA[d]	96.89	0.92	96.01	97.97	0.95
PNN	MSE–PseAA	96.98	0.93	96.16	98.01	0.95
SVM	MSE–PseAA	97.41	0.94	96.58	98.43	0.96

Note. The best values of performance metrics are shown in bold.
[a] PseAA2 is the feature vector formed using pseudo-amino acid by considering two physiochemical properties of amino acids.
[b] PseAA3 is the feature vector formed using pseudo-amino acid by considering three physiochemical properties of amino acids.
[c] MSE–AA is the hybrid feature vector formed by combining amino acid features with wavelet-based multiscale energy features.
[d] MSE–PseAA is the hybrid feature vector formed by combining pseudo-amino-acid features with wavelet-based multiscale energy features.

all three performance metrics (accuracy, specificity, and sensitivity) are the best. Hence, MSE–PseAA and RBF–SVM were selected by the GPCR–hybrid program for GPCR subfamily-level classification. The results for subfamily-level classification are summarized in Table 2.

Three performance metrics were used for the performance evaluation. It is clearly seen in Table 2 that the values of accuracy, specificity, and sensitivity were the highest for the SVM classifier with the MSE–PseAA feature extraction strategy. The best values of performance metrics are shown in bold.

Classification at sub-subfamily level

We classified GPCRs into 108 sub-subfamilies. The performance metrics used at the sub-subfamily level are overall accuracy, sensitivity, and specificity. The details of the performance measurements are described in the sections given below.

Classifier performance using PseAA2

The overall accuracies obtained by using the NN, PNN, and SVM classifiers for the PseAA2 strategy were 72.95%, 72.88%, and 72.65%, respectively. The specificity measures using NN, PNN, and SVM for the subfamily level were 73.01%, 72.94%, and 72.70%, respectively. The sensitivity measures using NN, PNN, and SVM were 69.02%, 67.77%, and 67.08%, respectively.

Classifier performance using PseAA3

The overall accuracies achieved by using the NN, PNN, and SVM classifiers for the PseAA3 strategy were 73.67%, 74.29%, and 68.78%, respectively. The specificity measures using NN, PNN, and SVM were 73.72%, 74.35%, and 68.81%, respectively. The sensitivity measures using NN, PNN, and SVM were 69.71%, 69.82%, and 68.96%, respectively.

Classifier performance using MSE–PseAA

The overall accuracies obtained by using the NN, PNN, and SVM classifiers for the MSE–PseAA feature extraction strategy were 72.48%, 71.10%, and 75.60%, respectively. The specificity measures using NN, PNN, and SVM were 72.53%, 71.15%, and 70.32%, respectively. The sensitivity measures using NN, PNN, and SVM were 69.01%, 67.67%, and 75.67%, respectively.

Classifier performance using MSE–AA

The overall accuracies obtained by using the NN, PNN, and SVM classifiers for the MSE–PseAA strategy were 69.75%, 69.53%, and 73.45%, respectively. The specificity measures using NN, PNN, and SVM were 69.80%, 68.58%, and 73.59%, respectively. The sensitivity measures using NN, PNN, and SVM were 66.32%, 65.01%, and 69.89%, respectively.

For the sub-subfamily-level classification, MSE–PseAA with SVM classifier performed the best and, hence, was selected by GPCR–hybrid for sub-subfamily-level classification of any test GPCR sequence. The values of all three performance metrics (accuracy, specificity, and sensitivity) are the best. The results for subfamily-level classification are summarized in Table 3.

For the sub-subfamily level, we had three performance metrics (accuracy, specificity, and sensitivity), as shown in Table 3. The

Table 2
GPCR classification performance for subfamily level.

Classifier	Feature extraction strategy	Accuracy (%)	Specificity (%)	Sensitivity (%)
NN	PseAA2	81.02	80.99	80.55
PNN	PseAA2	82.13	82.10	81.30
SVM	PseAA2	81.58	81.55	81.15
NN	PseAA3	81.88	81.85	81.52
PNN	PseAA3	83.47	83.42	83.18
SVM	PseAA3	79.02	78.98	78.85
NN	MSE–AA	78.55	78.51	78.51
PNN	MSE–AA	78.29	78.21	78.79
SVM	MSE–AA	81.96	81.90	81.95
NN	MSE–PseAA	80.73	80.69	80.72
PNN	MSE–PseAA	80.36	80.27	81.24
SVM	**MSE–PseAA**	**84.97**	**84.94**	**84.08**

Note. The best values of performance metrics are shown in bold.

Table 3
Classification performance for sub-subfamily level.

Classifier	Feature extraction strategy	Accuracy (%)	Specificity (%)	Sensitivity (%)
NN	PseAA2	72.95	73.01	69.02
PNN	PseAA2	72.88	72.94	67.77
SVM	PseAA2	72.65	72.70	69.08
NN	PseAA3	73.67	73.72	69.71
PNN	PseAA3	74.29	**74.35**	69.82
SVM	PseAA3	68.78	68.81	68.96
NN	MSE–AA	69.75	69.80	66.32
PNN	MSE–AA	68.53	68.58	65.01
SVM	MSE–AA	73.45	73.59	69.89
NN	MSE–PseAA	72.48	72.53	69.01
PNN	MSE–PseAA	71.10	71.15	67.61
SVM	**MSE–PseAA**	**75.60**	70.32	**75.67**

Note. The best values of performance metrics are shown in bold.

best performance was given by the RBF–SVM classifier using the MSE–PseAA feature extraction strategy (shown in bold).

Comparison with other hierarchical GPCR classification methods

Comparison with selective top-down method

As we showed in the "Classification at subfamily level" and "Classification at sub-subfamily level" sections above, SVM using the MSE–PseAA feature extraction strategy outperformed the other classifiers at the subfamily and sub-subfamily levels, whereas at the family level the SVM classifier with PseAA2 performed the best, as explained in the "Classification at family level" section above. Hence, we used the results of SVM in comparison with the selective top-down approach [20]. The performance metric used in the selective top-down approach is the overall accuracy. Hence, we compared the accuracy of our approach with that of the selective top-down approach. The results achieved by our approach were slightly better than those of the selective top-down approach. The best overall accuracy achieved in the selective top-down approach for the family level was 95.87%, whereas the GPCR–hybrid method achieved an overall accuracy of 97.86%. For the subfamily level, the selective top-down approach had an overall accuracy of 80.77%, whereas the GPCR–hybrid method achieved an accuracy of 84.97%. Finally, for the sub-subfamily level, the selective top-down approach had an accuracy of 69.98%, whereas the GPCR–hybrid method achieved an accuracy of 75.60%. At the subfamily and sub-subfamily levels, there was much improvement in the performance because of the hybrid feature extraction strategy. The GPCR–hybrid method performed better than the selective top-down method at all of the GPCR classification levels. The comparison of results is shown in Table 4.

Comparison with other existing methods

We also performed comparisons using three existing datasets: D167, D566, and D365. Because these datasets represent GPCR sequences belonging to only one level, the comparison with all three datasets was performed at only one level. In addition, the performance measurement used for comparison was overall accuracy. We computed results on each of these datasets using the SVM classifier with four different kernels: Lin–SVM, Poly–SVM, RBF–SVM, and Sig–SVM. The best of these four kernels were chosen for classification.

The D167 dataset has been used by many researchers to test their methods. We compared our method with six such methods [26,41–45]. One of these six methods, which is termed as principal component analysis (PCA)–GPCR [45], was reported in 2010. We observed that the overall accuracy achieved by our method was higher than that of these other methods. The comparison with these six methods is shown in Table 5.

We compared our method with two existing methods on the D365 dataset. The first method is termed as GPCR–CA (cellular automaton) [46], and second one is named as PCA–GPCR [45]. The overall accuracies achieved by the GPCR–CA and PCA–GPCR methods were 83.56% and 92.60% respectively. The overall accuracy achieved by the proposed GPCR–hybrid method was 91.72%, which is nearly 9% higher than the GPCR–CA method and is comparable to the PCA–GPCR method. The comparison on D365 is shown in Table 6.

Finally, on the D566 dataset, we compared our method with the PCA–GPCR method. The overall accuracy achieved by the PCA–GPCR method was 97.88%, whereas the accuracy achieved by our proposed method was 97.91%. The comparison on D566 is shown in Table 7.

The improvement in performance of the GPCR–hybrid method over the existing methods is due to using the hybrid combination

Table 4
Comparison with selective top-down method.

GPCR classification level	Selective top-down accuracy (%) [20]	GPCR–hybrid accuracy (%)
Family	95.87	97.86
Subfamily	80.77	84.97
Sub-subfamily	69.98	75.60

Table 5
Comparison with other methods on D167 dataset.

Reference	Overall accuracy (%)
[26]	83.23
[41]	83.20
[42]	96.40
[43]	97.60
[44]	97.60
PCA–GPCR [45]	**98.20**
GPCR–hybrid	98.45

Note. The best value of performance metrics is shown in bold.

Table 6
Comparison with other methods on D365 dataset.

Method	Overall accuracy (%)
GPCR–CA [46]	83.56
PCA–GPCR [45]	92.60
GPCR–hybrid	**92.59**

Note. The best value of performance metrics is shown in bold.

Table 7
Comparison with other methods on D566 dataset.

Method	Overall accuracy (%)
PCA–GPCR [45]	97.88
GPCR–hybrid	**97.91**

Note. The best value of performance metrics is shown in bold.

of MSE- and PseAA-based features. In this way, both the spatial and transform domains are exploited at the same time. In addition, the optimization of SVM parameters and use of the proper kernel for a dataset also contribute to the improved performance of the GPCR–hybrid program.

Conclusions

In this work, we have hierarchically classified GPCRs into three levels: family, subfamily, and sub-subfamily. We developed a web predictor that is able to predict a GPCR sequence with effective accuracy. This web predictor can be very helpful for pharmacists for annotating the unknown GPCRs. Once the input GPCR is categorized, its function can be learned and it can be used in the relevant drug. We observed that by using the hybrid feature extraction strategy, the overall performance of the GPCR predictor can be improved. It was shown that by using the hybrid feature extraction strategy, which exploits both the spatial and transform domains of AA, the different GPCR types can be discriminated in a better way and, consequently, a high prediction performance can be achieved. We also observed that SVM performs better than PNN and NN for GPCR classification at any level. The performance of the SVM classifier seems to be less affected by the curse of dimensionality. In addition, if more physiochemical properties are used while representing GPCR sequences, the overall prediction performance might be further improved.

Acknowledgment

This work was supported by the Higher Education Commission (HEC) of Pakistan under the indigenous Ph.D. program (074-1844-PS4-406).

References

[1] K.H. Lundstrom, M.L. Chiu, G-Protein Coupled Receptors in Drug Discovery, CRC Press, Taylor & Francis Group, Boca Raton, FL, 2006.

[2] M. Bhasin, G.P.S. Raghava, GPCRpred: an SVM-based method for prediction of families and sub-families of G-protein coupled receptors, Nucleic Acids Res. 32 (2004) 383–389.

[3] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (1990) 403–410.

[4] D. Fridmanis, R. Fredriksson, I. Kapa, B.S. Helgi, J. Klovins, Formation of new genes explains lower intron density in mammalian rhodopsin G protein-coupled receptors, Mol. Phylogenet. Evol. 43 (2006) 864–880.

[5] J.C. Cardoso, V.C. Pinto, F.A. Vieira, M.S. Clark, D.M. Power, Evolution of secretin family GPCR members in the metazoa, BMC Evol. Biol. 6 (2006) 108.

[6] S.S. Das, G.A. Banker, The role of protein interaction motifs in regulating the polarity and clustering of the metabotropic glutamate receptor mGluR1a, J. Neurosci. 26 (2006) 8115–8125.

[7] T. Nakagawa, T. Sakurai, T. Nishioka, K. Touhara, Insect sex-pheromone signals mediated by specific combinations of olfactory receptors, Science 307 (2005) 1638–1642.

[8] Y. Prabhu, L. Eichinger, The *Dictyostelium* repertoire of seven transmembrane domain receptors, Eur. J. Cell Biol. 85 (2006) 937–946.

[9] S.M. Foord, S. Jupe, J. Holbrook, Bioinformatics and type II G-protein-coupled receptors, Biochem. Soc. Trans. 30 (2002) 473–479.

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

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

[12] M. Bhasin, G.P.S. Raghava, GPCRsclass: a web tool for the classification of amine type of G-protein-coupled receptors, Nucleic Acids 33 (2005) 143–147.

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

[14] J.X. Wang, P. Qin, Q.L. Liu, H.Y. Yang, Y.Z. Fan, J.K. Yu, S. Zheng, Detection and significance of serum protein marker of Hirschsprung disease, Protein Eng. 120 (2007) e56–e60.

[15] S. Möller, J. Vilo, M.D. Croning, Prediction of the coupling specificity of G protein coupled receptors to their G-proteins, Bioinformatics 17 (2001) 174–181.

[16] P.K. Papasaikas, P.G. Bagos, Z.I. Litou, S.J. Hamodrakas, A novel method for GPCR recognition and family classification from sequence alone using signatures derived from profile hidden Markov models, SAR QSAR Environ. Res. 14 (2003) 413–420.

[17] P.L. Martelli, P. Fariselli, L. Malaguti, R. Casadio, Prediction of the disulfide bonding state of cysteines in proteins with hidden neural networks, Protein Eng. 15 (2002) 951–953.

[18] Q.B. Gao, Classification of G-protein coupled receptors at four levels, Protein Eng. Des. Sel. 19 (2006) 511–516.

[19] T.K. Attwood, M.D.R. Croning, A. Gaulton, Deriving structural and functional insights from a ligand-based hierarchical classification of G protein-coupled receptors, Protein Eng. 15 (2002) 7–12.

[20] M.N. Davies, A. Secker, A.A. Freitas, M. Mendao, J. Timmis, D.R. Flower, On the hierarchical classification of G-protein coupled receptors, Bioinformatics 23 (2007) 3113–3118.

[21] BIAS-PROFS: Bioinformatics, immunology, and algorithms make short work of protein function classification [GPCR dataset]. Available from: <http://www.cs.kent.ac.uk/projects/biasprofs>.

[22] K.C. Chou, Prediction of protein cellular attributes using pseudo-amino-acid composition, Proteins 43 (2001) 246–255.

[23] A. Khan, M.F. Khan, T. Choi, Proximity based GPCRs prediction in transform domain, Biochem. Biophys. Res. Commun. 371 (2008) 411–415.

[24] D.L. Wheeler, T. Barrett, D.A. Benson, S.H. Bryant, K. Canese, D.M. Church, M. DiCuccio, R. Edgar, S. Federhen, W. Helmberg, D.L. Kenton, O. Khovayko, D.J. Lipman, T.L. Madden, D.R. Maglott, J. Ostell, J.U. Pontius, K.D. Pruitt, G.D. Schuler, L.M. Schriml, E. Sequeira, S.T. Sherry, K. Sirotkin, G. Starchenko, T.O. Suzek, R. Tatusov, T.A. Tatusova, L. Wagner, E. Yaschenko, Database resources of the National Center for Biotechnology Information, Nucleic Acids Res. 35 (2007) D5–D12.

[25] K.C. Chou, H.B. Shen, Review: recent progresses in protein subcellular location prediction, Anal. Biochem. 370 (2007) 1–16.

[26] D.W. Elrod, K.C. Chou, A study on the correlation of G-protein-coupled receptor types with amino acid composition, Protein Eng. Des. Sel. 15 (2002) 713–715.

[27] K.C. Chou, D.W. Elrod, Bioinformatical analysis of G-protein-coupled receptors, J. Proteome Res. 1 (2002) 429–433.

[28] K.C. Chou, Prediction of G-protein-coupled receptor classes, J. Proteome Res. 4 (2005) 1413–1418.

[29] K.C. Chou, Prediction of protein cellular attributes using pseudo amino acid composition, Proteins 43 (2001) 246–255 (Erratum: vol. 44, p. 60).

[30] S.W. Zhang, Q. Pan, H.C. Zhang, Z.C. Shao, J.Y. Shi, Prediction protein homo-oligomer types by pseudo amino acid composition: approached with an improved feature extraction and naive Bayes feature fusion, Amino Acids 30 (2006) 461–468.

[31] 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.

[32] J-L. Fauchere, V. Pliska, Hydrophobic parameters of amino-acid side chains from the partitioning of *n*-acetyl-amino-acid amides, Eur. J. Med. Chem. Chim. Ther. 18 (1983) 369–375.

[33] Y.Z. Guo, M.L. Li, K.L. Wang, Z.N. Wen, M.L. Lu, L.X. Liu, L. Jiang, Fast Fourier transform-based support vector machine for prediction of G-protein coupled receptor subfamilies, Acta Biochim. Biophys. Sin. (Shanghai) 37 (2005) 759–766.

[34] I. Cosic, Macromolecular bioactivity: is it resonant interaction between macromolecules? Theory and applications, IEEE Trans. Biomed. Eng. 41 (1994) 1101–1114.

[35] R. Grantham, Amino acid difference formula to help explain protein evolution, Science 185 (1974) 862–864.

[36] J.Y. Shi, S.W. Zhang, Q. Pan, Y.M. Cheng, J. Xie, Prediction of protein subcellular localization by support vector machines using multi-scale energy and pseudo amino acid composition, Amino Acids 33 (2007) 69–74.

[37] A. Khan, S.F. Tahir, A. Majid, T.S. Choi, Machine learning based adaptive watermark decoding in view of an anticipated attack, Pattern Recogn. 41 (2008) 2594–2610.

[38] A. Khan, S.F. Tahir, T.S. Choi, Intelligent extraction of a digital watermark from a distorted image, IEICE Transact. Inform. Syst. E91-D (2008) 2072–2075.

[39] J. Javed, A. Khan, A. Majid, A.M. Mirza, J. Bashir, Lattice constant prediction of orthorhombic ABO3 perovskites using support vector machines, Comput. Mater. Sci 39 (2007) 627–634.

[40] D.F. Specht, Probabilistic neural networks, Neural Netw. 3 (1990) 109–118.

[41] Y. Huang, J. Cai, L. Ji, Y. Li, Classifying G-protein coupled receptors with bagging classification tree, Comput. Biol. Chem. 28 (2004) 275–280.

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

[43] Q.B. Gao, Z.Z. Wang, Classification of G-protein coupled receptors at four levels, Protein Eng. Des. Sel. 19 (2006) 511–516.

[44] 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.

[45] 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.

[46] 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) 1413–1423.