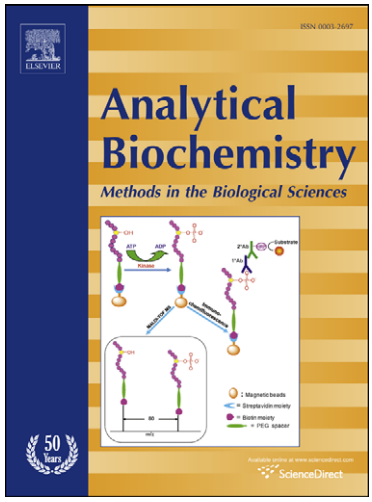


GPCR prediction using pseudo amino acid composition and multi-scale energy representation of different physiochemical properties

Zia-ur-Rehman, Asifullah Khan

PII: S0003-2697(11)00067-4
DOI: 10.1016/j.ab.2011.01.040
Reference: YABIO 10329

To appear in: *Analytical Biochemistry*

Received Date: 31 October 2010
Revised Date: 26 January 2011
Accepted Date: 27 January 2011

Please cite this article as: Zia-ur-Rehman, A. Khan, GPCR prediction using pseudo amino acid composition and multi-scale energy representation of different physiochemical properties, *Analytical Biochemistry* (2011), doi: 10.1016/j.ab.2011.01.040

GPCR prediction using pseudo amino acid composition and multi-scale 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), P.O. 45650, Nilore, Islamabad, Pakistan.

Phone: 92-51-2207381-3, Ext 3159.

Fax: 92-51-2208070

asif@pieas.edu.pk

Short title: Prediction of GPCRs using PseAA and MSE

Appropriate field: Membranes and Receptors

Other categories: Cell biology, Amino acids, Protein structure and analysis

***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 pharmaceutical industry. However, prediction of *GPCRs*, their families and sub families is a challenging task. In this paper, *GPCRs* are classified into families, sub families, and sub-sub families using pseudo amino acid composition and multi-scale energy representation of different physiochemical properties of amino acids. Aim of present research is to assess different feature extraction strategies and to develop a hybrid feature-extraction strategy that can exploit the discrimination capability in spatial as well as transform domain for *GPCR* classification. Support vector machine (*SVM*), nearest neighbor (*NN*) and probabilistic neural network (*PNN*) are used for classification purpose. The overall performance of each classifier is computed individually for each feature extraction strategy. It has been observed that using Jackknife test; the proposed *GPCR-Hybrid* provides 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*.**

Keywords: GPCRs classification, Multi-scale energy, Pseudo amino acid composition, physiochemical properties and Rhodopsin-like receptors.

1. INTRODUCTION

G-protein coupled receptors (*GPCRs*) are known to play an essential role in the coordination of cellular communications and are involved in many physiological processes. They play important role in various mammalian disorders including allergies, cardiovascular dysfunction, depression, cancer, pain, diabetes, and various 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 types of *GPCRs* are known for directly mediating neuronal and endocrine regulation of cardiac and vascular responses. Additionally, many *GPCRs* are known to influence cardiovascular functions. Their role in the development of cancer is becoming apparent 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] such as: Rhodopsin-like receptors (Class A), Secretin receptors (Class B), Metabotropic glutamate receptors (Class C), Fungal mating Pheromone receptors (Class D), Cyclic AMP receptors (Class E) and Frizzled/Smoothened receptors (Class F). The *GPCR* classes A, B, C and F are mostly found in mammals, class D is found only in Fungi and class E *GPCRs* are found in Dictyostelium. Rhodopsin-like receptors is the biggest family of *GPCRs* constituting 80% of all *GPCRs*. It is 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 Cyclic AMP 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 mediation of hedgehog signaling, a key regulator of animal development [9].

Each family is divided into sub-families, and similarly each sub family is further divided into sub-sub families. The classification of *GPCRs* into families, sub-families and sub-sub families is done based on the functionalities performed by each *GPCR* sequence; grouping *GPCR* sequences with similar functionalities in the same family. One amongst several methods for the prediction of *GPCR* sequences is to do sequence similarity searches using pair wise alignment tools [10] e.g. *BLAST* and *FASTA* (Altschul et al., 1997; Pearson, 2000). The second method is to classify *GPCRs* by conducting biological experiments. During the last decade, hundreds of new *GPCRs* have been discovered and it is continuing to grow rapidly. Therefore, their annotation based on the manual experimentation has made it very expensive and almost impossible. Thus, there was a great need of 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 e.g. the Bayes network method [11], support vector machine [2, 12, 13, 14] and the Hidden Markov models [15, 16, 17]. Although these methods classify *GPCRs* with high accuracy but none of these provide hierarchical *GPCRs* classification. The *GPCRs* are hierarchically classified into 4 levels i.e. super family level, family level, sub family level and type level by Gao et al. [18]. The data set used in this method consists of 1406 *GPCRs* sequences and 1406 Globular proteins (*non-GPCRs*). In the first level, *GPCRs* are discriminated from non-*GPCRs*. In the second level, 6 families of *GPCRs* are classified. The Rhodopsin-like family is further classified into sub families in the third level. While in the fourth level, sub-sub families of amine subfamily and olfactory subfamily are predicted. *GPCRs* are classified into three levels i.e. super family, family and specific receptor subtype by Attwood et al. [2]. *GPCRs* are also hierarchically classified into 3 levels by Matthew et al. [19]. In the first level *GPCRs* are classified into 5 families (Class F is ignored), while in the second level, *GPCRs* are classified into 40 sub families. Finally, in the third level, *GPCRs* are classified into 108 sub-sub families. They have also developed an online *GPCRs* prediction server, which is available at [20]. Both of these hierarchical classification methods provide good overall accuracies.

In this paper, we have classified *GPCRs* into three levels. First, we have classified *GPCRs* into 5 families, then into 40 sub families and finally into 108 sub-sub families as done by Matthew et al. [19]. Frizzled/Smoothened receptors family is ignored as it contains too few sequences from which to developing an accurate classification algorithm. Three feature extraction strategies are used. The first one is the Pseudo amino acid composition (*PseAA*) [21], which is used in two ways i.e. using two/three physiochemical properties of *GPCRs*. In the second feature extraction strategy, a hybrid feature vector (*MSE-PseAA*) is formed by combining wavelet based multi-scale energy (*MSE*) and

PseAA based features [22]. Third one is also a hybrid feature vector (*MSE-AA*) formed by the combination of Amino acid composition and *MSE* features. We have used three classifiers and 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*). 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 spatial as well as transform domain. We have developed a web predictor (*GPCR-Hybrid*), which takes unknown *GPCR* sequence as input and classifies it, first into family, then into sub family and finally, into sub-sub family. 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 Figure 1.

Figure 1 comes here

The *GPCRs* dataset that we have used is taken from *BIAS-PROFS* website [20]. The overall performance of our proposed approach is better than the existing hierarchical *GPCR* classification methods.

2. MATERIALS AND METHODS

2.1. Data Sets

The dataset that we have mainly used for the training and assessment of our classification approach was downloaded from the *BIAS-PROFS* website [20] developed by Matthew et al. in 2007. *GPCR* sequences for the dataset were identified using the Entrez search and retrieval system [23]. The Text based searching was used to identify all sequences within each sub-sub family of the hierarchy. *GPCR* sequences shorter than 280 amino acids in length were also removed. Finally, all the identical sequences within the dataset were removed to avoid redundancy. Generally, a homology bias is avoided using a cutoff threshold of 25% [24]. However, in this study, the dataset by Mathew et al. has not been put to such a stringent criterion, as the numbers of GPCRs for some of the classes would be too few in number to have statistical significance. Hence, we have used the dataset by Mathew et al.[19] as it is and consequently have not applied any additional processing. The dataset consist of 8354 *GPCR* sequences. Out 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 have also used three other benchmark datasets for comparison with existing methods. These datasets were constructed using older versions of GPCRDB and it is reported that they avoid homology bias largely. For

simplicity, they are referred to as D167, D566 and D365 containing 167, 566 and 365 GPCR sequences, respectively. The GPCRs in the dataset D167 [25] (belonging to the sub-sub family level) are classified into four sub-sub families i.e. (1) acetylcholine (2) adrenoceptor (3) dopamine and (4) serotonin. The dataset D566 [26] (belonging to sub-sub family level) contains GPCRs belonging to seven sub-sub families i.e. (1) Adrenoceptor (2) Chemokine (3) dopamine (4) Neuropeptide (5) Olfactory type (6) Rhodopsin (7) serotonin. The last dataset D365 [27] (belonging to the family level) contains GPCRs belonging to six families: (1) Rhodopsin-like (2) Secretin-like (3) Metabotrophic glutamate pheromone (4) Fungal pheromone (5) cAMP receptor and (6) Drizzled/smoothened family. Chou and Elrod [25, 26, 27] reported that all 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.

2.2. Sequence representation

2.2.1. Amphiphilic Pseudo amino acid composition (PseAA)

To avoid losing much important information hidden in protein sequences, the pseudo amino acid composition (PseAAC) was proposed [28] to replace the simple amino acid composition (AAC) for representing the sample of a protein. PseAAC has been widely used to study various problems in proteins and protein-related systems, such as: predicting sub-cellular location of proteins [29] and GPCR types [30]. However, to the best of our knowledge, so far PseAAC has not been used for predicting GPCRs and their types in conjunction with the approach of multi-scale energy representation of different physiochemical properties. The present study was devoted to do so, and quite encouraging results have been obtained.

The conventional amino acid composition uses only the frequency of occurrence of each amino acid in the *GPCRs* sequence. Unlike the conventional amino acid composition, the *PseAA* composition approach as used in [22, 29] is adopted in the current study. It preserves sequence order and sequence-length information. The *GPCRs* sequence *R,* with *L* amino acids, where *L* represents the length of protein sequence, can be represented as shown by the Eq. (1).

$$\boldsymbol{R} = R_1, R_2 \ldots R_L \qquad (1)$$

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

$$\boldsymbol{PseAA} = P_1, P_2 \ldots P_{20} \ldots P_{\Lambda} \qquad (2)$$

where $\Lambda = 20 + n*\lambda$ (λ is the numbers 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 have selected $\lambda = 21$ and analyzed the performance by using different combination of physiochemical properties. We have taken $\lambda = 21$ because it is giving 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 1st-teir 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 present research, we have used three physiochemical properties i.e. hydrophobicity, electronic and bulk properties. The word hydrophobic literally means afraid of water. It is obvious that hydrophobic residues prefer to be in a non-aqueous environment such as a lipid bilayer. Biological molecules can contain large non-polar regions. These non-polar regions may also be described as hydrophobic region. Hydrophobicity of proteins is one of the most important factors in determining a GPCR's structure and function. However, with different experimental conditions, different organic solvents and computing approaches, hydrophobicity value per amino acid will be different. Various scales of hydrophobicity are employed such as KDH, MH, and FH, etc. However, the FH hydrophobicity scale [31] has been proved the most discriminative out of these hydrophobicity measures [32]. Hence, we have used FH scale for hydrophobicity measure in present research. Electronic property has been modeled using electron ion interaction pseudopotential (EIIP) model [33]. EIIP value describes the average energy states of all valence electron of amino acids. Electrons delocalized from the particular amino acid have the strongest impact on the electronic distribution of the whole protein. Hence, we have chosen EIIP model for electronic property measurement. Finally, the bulk property has been modeled using composition, polarity, and molecular volume model (CPV) [34]. Polarity and volume (size) are known to have a great impact on the folding of the protein. Hence, CPV model has been used in the present research to model bulk property.

We have assessed the performance of the classifier by first considering two physiochemical properties i.e. the electronic and the bulk property. Then, the third property (Hydrophobicity) is also included, which has slightly enhanced the overall performance. The PseAA using two physiochemical properties is termed as PseAA2. The length of feature vector in PseAA2 is $\Lambda = 20 + 2*21$ i.e. 62. The PseAA using three physiochemical properties is termed as PseAA3. The length of feature vector in PseAA3 is 83.

2.2.2. Wavelet based multi-scale energy (MSE) and Pseudo amino acid composition (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, we have used Mallat's Fast algorithm in the present 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 have 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 have used *FH* scale for computing Hydrophobicity values. The significance of Hydrophobicity property is discussed in section 2.2.1. Each of the amino acid is simply replaced with its corresponding value in the *FH* scale [31]. The resulting numeric form is homologous to a digital signal. Next, the wavelet (Haar) transform of this digital signal is taken. Then 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 of 8000 amino acids, then the decomposition levels for that sequence would be 13. The length of sequences may not be same; hence, zero padding is performed in case of shorter sequences to keep consistency in the size of feature vector. The overall feature vector formed in this way is termed as MSE [35]. Hence, the MSE-feature vector of (m+1)-Dimensions is formed as given in the Eq. (3):

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

where $k = 1, 2, \ldots N$, N is total number of *GPCR* sequences, $d_j^{\,k}$ is the root mean square energy of wavelet detail coefficients in the corresponding j^{th} scale and $a_m^{\,k}$ is the root mean square energy of wavelet approximation coefficients in m^{th} scale as shown by Eq. (3) and (5), respectively.

$$d_j{}^k = \sqrt{\frac{1}{N_j}\sum_{n=0}^{N_j-1}\{u_j{}^k(n)\}^2} \tag{4}$$

$$a_m{}^k = \sqrt{\frac{1}{N_j}\sum_{n=0}^{N_m-1}\{V_m{}^k(n)\}^2} \tag{5}$$

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 n^{th} detail coefficient in the j^{th} scale and $v_m^k(n)$ is the n^{th} approx coefficient in the m^{th} scale. The scale here means the decomposition level.

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

$$\textbf{\textit{MSE-PseAA}} = [P_1, P_2 \ldots P_{20} \ldots P_\Lambda, \lambda_1^k, \lambda_2^k \ldots \lambda_{m+1}^k] \quad (6)$$

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

2.2.3. Wavelet based multi-scale energy (MSE) and amino acid composition (AA) based hybrid feature extraction method

In this feature extraction strategy, first the *GPCR* sequences are converted into the numeric form using *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, which is combined with *MSE* features to form a hybrid feature vector (*MSE-AA*) as given by the Eq. (7).

$$\boldsymbol{X}_k = [P_1^k, P_2^k \ldots P_{20}^k, \lambda_1^k, \lambda_2^k \ldots \lambda_{m+1}^k] \quad (7)$$

where the first twenty features (P_1^k to 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 *MSE* feature extraction strategy.

2.3. Nearest Neighbor

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

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

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

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

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

2.4. Support vector machines

The *SVM* classifier inherently is a binary classifier, but it can be tailored for multi-classification 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 hyper plane such that the classification error for new test samples is minimized. For linearly separable sample points, hyper plane is determined by maximizing the distance between the support vectors [36, 37, 38].

As our problem is a multi-class problem, so we have adopted the *one-vs-all* strategy, while using the *LIBSVM 2.88-1* software (Chang and Lin 2008). We have evaluated the performance of *SVM* using four different types of kernel i.e Linear (Lin-SVM), Polynomial (Poly-SVM), Radial Basis Function (RBF-SVM) and Sigmoidal (Sig-SVM). The *LIBSVM 2.88-1* solves *SVM* problem using Nonlinear Quadratic Programming technique. During parameter optimization of *SVM* models, the average accuracy of *SVM* model is maximized.

2.5. Probabilistic Neural Network

The Probabilistic Neural Network (*PNN*) is developed in 1990 by Specht [39] 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; an input, pattern, summation, and decision layers. The input layer has *N* nodes 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 Pattern layer. Here m_k Gaussian functions are calculated for each class k ($1 \leq k \leq c$) as given by the Eq. (10).

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

where μ_j^k is the mean of 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 the Eq. (11).

$$\Phi_k(x) = \sum_{j=1}^{m_k} \pi_j^k \, p_j^k(x) \quad (11)$$

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

$$p_k(x) = \sum_{l=1}^{c} v_l^k \, \alpha_l \, \Phi_l(x) \quad (12)$$

where α_l indicates the prior probability and v_l^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 that need to be optimized is the smoothing factor, which controls the deviations of Gaussian functions. The optimized range of smoothing factor, in our case varies from 0.01 to 5.

2.6. Performance Measures

In statistical prediction, three cross-validation methods are often used to examine a predictor for its effectiveness in practical application i.e. independent dataset test, sub-sampling test and jackknife test. However, among the three cross-validation methods, the jackknife test is deemed the most objective that can always yield a unique result for a given benchmark dataset, and hence has been increasingly used by investigators to examine the accuracy of various predictors. Accordingly, the jackknife test was also adopted here to examine the quality of the present 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 for 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, Mathew Correlation Coefficient (*MCC*) and F-measure. The *TP* (true positive) and *TN* (true negative) are the number of correctly predicted positive and negative samples. The *FP* (false positive) and *FN* (false negative) are the number of incorrectly predicted positive and negative samples.

2.6.1. Accuracy

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

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} * 100 \quad (13)$$

2.6.2 Sensitivity

$$Sensitivity = \frac{TP}{TP+FN} * 100 \quad (14)$$

2.6.3. Specificity

$$Specificty = \frac{TN}{FP+TN} * 100 \quad (15)$$

2.6.4. MCC

MCC is takes values in the interval of [−1, 1]. A value of 1 means that the classifier never makes any mistakes and a value -1 means that the classifier always makes mistakes. *MCC* is given by the Eq. (20) as:

$$MCC = \frac{(TP * TN) - (FP * FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (16)$$

2.6.5 F-measure

F-measure is a measure of the accuracy of a test, which considers both the precision and the 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 worst score at *0*.

$$F - measure = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (17)$$

$$Precision = \frac{TP}{TP + FP} \quad (18)$$

$$Recall = \frac{TP}{TP + FN} \quad (19)$$

2.7. Proposed GPCR-Hybrid method

The GPCR-Hybrid is a web predictor, which can efficiently classify unknown *GPCR* sequence, first into Family, then into sub family and finally into sub-sub family. The performance of each classifier is assessed individually for each of the feature extraction strategy. At first, the *GPCR-Hybrid* program asks for the input *GPCR* sequence using a graphical user interface as shown in the Figure 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 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 sub family level. The sub family class of input sequence is then predicted using the best performing classifier of sub family level. The *MSE-PseAA* is selected by *GPCR-Hybrid* for feature extraction at sub family level and *SVM* is used to predict sub family class. Finally, the sub-sub family 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 Figure 1.

Figure 1 comes here

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

Figure 2 comes here

The proposed *GPCR-Hybrid* is available at *http://111.68.99.218/GPCR*.

3. RESULTS AND DISCUSSION

In the *GPCR-Hybrid* method, the hierarchical classification task is performed into three stages. First stage predicts the family of the *GPCR* sequence; second stage predicts the sub family and finally in the third stage sub-sub family of the sequence is predicted. We have used three feature extraction strategies for the sake of sequence conversion into numeric form. First feature extraction strategy is the *PseAA*, which is used in two ways. The second feature extraction strategy is named as *MSE-PseAA*, which is a hybrid feature vector formed by combining Pseudo amino acid with wavelet based Multi scale energy (*MSE*) features. The third feature extraction strategy is named as *MSE-AA*. It is also a Hybrid feature vector formed by combining Multi scale energy based features with Amino acid composition (*AA*) based features. The details of these feature extraction strategies are given in section 2.2. We have used three classifiers to assess the performance for each feature extraction strategy.

At each stage, the best performing classifier and 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.

3.1. Classification at Family level

We have classified *GPCRs* into five families. 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 have used overall accuracy, Sensitivity, Specificity, Mathew Correlation Coefficient (*MCC*) and F-measure. The formulas of all these measures have been described in the section 2.6. The performance measurements using *NN*, *PNN* and *SVM* for each of the feature extraction strategy are described below.

3.1.1. Classifiers performance using PseAA2

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

3.1.2. Classifiers performance using PseAA3

The overall accuracies obtained by using the *NN*, *PNN* and *SVM* classifiers for *PseAA3* are: 97.58%, 97.74% and 93.66%, respectively. The optimal smoothing factor for *PNN* is chosen as 0.6. The *MCC* measures using *NN*, *PNN* and *SVM* are: 0.94, 0.94 and 0.85, respectively. The Specificity measures using *NN*, *PNN* and *SVM* are: 96.96%, 97.16% and 89.83%, respectively. The Sensitivity measures using *NN*, *PNN* and *SVM* are: 98.41%, 98.52% and 98.04%, respectively. The F-measures using *NN*, *PNN* and *SVM* are: 0.96, 0.96 and 0.90, respectively.

3.1.3. Classifiers performance using MSE-PseAA

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

3.1.4. Classifiers performance using MSE-AA

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

For family level classification, *PseAA2* using *SVM* is giving the best performance. It has best Accuracy, *MCC*, sensitivity and F-measure values, while Specificity measure is 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.

Table 1 comes here

In the table I, the performance metrics i.e. Accuracy, *MCC*, Specificity, Sensitivity and F-measure are given as column wise. Their measurements using each of the classifier and feature extraction strategy are given row wise. It is clearly shown that *RBF-SVM* has shown the best performance using *PseAA2*.

3.2. Classification at sub Family level

We have classified *GPCRs* into 40 sub families. The performance measures used at the sub family level are: overall accuracy, Sensitivity and Specificity. These performance measurements using *NN*, *PNN* and *SVM* classifiers are described in the sections given below.

3.2.1. Classifiers performance using PseAA2

The overall accuracies obtained by using the *NN*, *PNN* and *SVM* classifiers for *PseAA2* strategy are: 81.02%, 82.13% and 81.58%, respectively. The Specificity measures using *NN*, *PNN* and *SVM* for sub family level are: 80.99%, 82.10% and 81.55%, respectively. The Sensitivity measures using *NN*, *PNN* and *SVM* are: 80.55%, 81.30% and 81.15%, respectively.

3.2.2. Classifiers performance using PseAA3

The overall accuracies achieved by using the *NN*, *PNN* and *SVM* classifiers for *PseAA3* are: 81.88%, 83.47% and 79.02%, respectively. The Specificity measures using *NN*, *PNN* and *SVM* are: 81.85%, 83.42% and 78.98%, respectively. The Sensitivity measures using *NN*, *PNN* and *SVM* are: 81.52%, 83.18% and 78.85%, respectively.

3.2.3. Classifiers performance using MSE-PseAA

The overall accuracies obtained by using the *NN*, *PNN* and *SVM* classifiers for *MSE-PseAA* strategy are: 80.73%, 80.36% and 84.97%, respectively. The Specificity measures using *NN*, *PNN* and *SVM* are: 80.69%, 80.27% and 84.94%, respectively. The Sensitivity measures using *NN*, *PNN* and *SVM* are: 80.72%, 81.24% and 84.08%, respectively.

3.2.4. Classifiers performance using MSE-AA

The overall accuracies obtained by using the *NN*, *PNN* and *SVM* classifiers for *MSE-PseAA* strategy are: 78.55%, 78.29% and 80.96%, respectively. The Specificity measures using *NN*, *PNN* and *SVM* are: 78.51%, 78.21% and 81.90%, respectively. The Sensitivity measures using *NN*, *PNN* and *SVM* are: 78.51%, 78.79% and 81.95%, respectively.

For the sub family-level classification, *SVM* is performing best using *MSE-PseAA* feature extraction strategy. The values of all the three performance metrics i.e. Accuracy, Specificity and sensitivity are the best. Hence, the *MSE-PseAA* and *RBF-SVM* are selected by *GPCR-Hybrid* for *GPCRs* sub family level classification. The results for sub family-level classification are summarized in Table 2.

Table 2 comes here

Three performance metrics are used for the performance evaluation. It is clearly shown in the Table 2 that the values of Accuracy, Specificity and Sensitivity are the highest for *SVM* classifier with *MSE-PseAA* feature extraction strategy. The best values of performance metrics are shown in bold letters.

3.3. Classification at sub-sub Family level

We have classified *GPCRs* into 108 sub-sub families. The performance metrics used at the sub-sub family level are: overall accuracy, Sensitivity and Specificity. The details of the performance measurements are described in the sections given below.

3.3.1. Classifiers performance using PseAA2

The overall accuracies obtained by using the *NN*, *PNN* and *SVM* classifiers for *PseAA2* are: 72.95%, 72.88% and 72.65%, respectively. The Specificity measures using *NN*, *PNN* and *SVM* for sub family level are: 73.01%, 72.94% and 72.70%, respectively. The Sensitivity measures using *NN*, *PNN* and *SVM* are: 69.02%, 67.77% and 67.08%, respectively.

3.3.2. Classifiers performance using Pseudo PseAA3

The overall accuracies achieved by using the *NN*, *PNN* and *SVM* classifiers for *PseAA3* are: 73.67%, 74.29% and 68.78%, respectively. The Specificity measures using *NN*, *PNN* and *SVM* are: 73.72%, 74.35% and 68.81%, respectively. The Sensitivity measures using *NN*, *PNN* and *SVM* are: 69.71%, 69.82% and 68.96%, respectively.

3.3.3. Classifiers performance using MSE-PseAA

The overall accuracies obtained by using the *NN*, *PNN* and *SVM* classifiers for *MSE-PseAA* feature extraction strategy are: 72.48%, 71.10% and 75.60%, respectively. The Specificity measures using *NN, PNN* and *SVM* are: 72.53%, 71.15% and 70.32%, respectively. The Sensitivity measures using *NN, PNN* and *SVM* are: 69.01%, 67.67% and 75.67%, respectively.

3.3.4. Classifiers performance using MSE-AA

The overall accuracies obtained by using the *NN, PNN* and *SVM* classifiers for *MSE-PseAA* are: 69.75%, 69.53% and 73.45%, respectively. The Specificity measures using *NN, PNN* and *SVM* are: 69.80%, 68.58% and 73.59%, respectively. The Sensitivity measures using *NN, PNN* and *SVM* are: 66.32%, 65.01% and 69.89%, respectively.

For the sub-sub family-level classification, *MSE-PseAA* with *SVM* classifier is performing the best and hence, selected by *GPCR-Hybrid* for sub-sub family level classification of any test *GPCR* sequence. The values of all the

three performance metrics i.e. Accuracy, Specificity and sensitivity are the best. The results for sub family-level classification are summarized in Table 3.

Table 3 comes here

For sub-sub family level, we have three performance metrics i.e. Accuracy, Specificity and Sensitivity, as shown in Table 3. The best performance is given by *RBF-SVM* classifier using *MSE-PseAA* feature extraction strategy; shown bold in Table 3.

3.4. Comparison with other hierarchical GPCRs classification methods

3.4.1. Comparison with Selective top down method

As we have shown in the sections: 3.2 and 3.3 that *SVM* has outperformed using *MSE-PseAA* feature extraction strategy as compared to the other classifiers at sub family and sub-sub family level. While at family level, *SVM* classifier with *PseAA2* is performing the best as explained in section 3.1. Hence, we have used the results of *SVM* in comparison with Selective top down approach [19]. The performance metric used in Selective top down approach is the overall accuracy. Hence, we have compared the accuracy of our approach with that of Selective top down approach. The results achieved by our approach are slightly better than the Selective top down approach. The best overall Accuracy achieved in Selective top down approach, for family level is 95.87%, while *GPCR-Hybrid* has achieved an overall accuracy equal to 97.86%. For sub family level, the Selective top down approach has an overall accuracy equal to 80.77% and *GPCR-Hybrid* has achieved an accuracy of 84.97%. Finally, for sub-sub family level, Selective top down approach has accuracy equal to 69.98% and the accuracy achieved by *GPCR-Hybrid* is equal to 75.60%. At sub family and sub-sub family level, there is much improvement in the performance, which is because of hybrid feature-extraction strategy. The *GPCR-Hybrid* has performed better than the Selective top down method at all of the *GPCR* classification levels. The comparison of results is shown in Table 4.

Table 4 comes here

3.4.2. Comparison with other existing methods

We have also performed comparisons using three existing datasets; D167, D566 and D365. As these datasets represent GPCRs sequences belonging only to one level, the comparison with all of the three datasets is performed at only one level. In addition, the performance measurement used for comparison is overall accuracy. We have computed results on each of these datasets using SVM classifier with four different kernels i.e. Lin-SVM, Poly-SVM, RBF-SVM and Sig-SVM. The best of these four kernels have been chosen for classification.

The dataset D167 has been used by many researcher to test their methods. We have compared our method with six such methods [25, 40, 41, 42, 43, 44]. One of these six methods, which is termed as PCA-GPCR [44] is reported in 2010. We have observed that the overall accuracy achieved by our method is higher than these methods. The comparison with these six methods is shown in Table 5.

Table 5 comes here

We have compared our method with two existing methods on dataset D365. First method is termed as GPCR-CA [45] and second one is named as PCA-GPCR [44]. The overall accuracies achieved by GPCR-CA and PCA-GPCR method are 83.56% and 92.60% respectively. The overall accuracy achieved by the proposed GPCR-Hybrid method is 91.72%, which is almost 9% higher than GPCR-CA method and is comparable to PCA-GPCR method. The comparison on D365 is shown in Table 6.

Table 6 comes here

Finally, on D566 dataset, we have compared our method with PCA-GPCR method. The overall accuracy achieved by PCA-GPCR method is 97.88, while the accuracy achieved by our proposed method is 97.91%. The comparison on D365 is shown in Table 7.

Table 7 comes here

The improvement in performance of GPCR-Hybrid method over the existing methods is because of using the hybrid combination 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 usage of proper kernel for a dataset has also contributed in the improved performance of GPCR-Hybrid.

3. CONCLUSIONS

In this work, we have hierarchically classified GPCRs into three level i.e. family, sub family and sub-sub family levels. We have developed a web predictor, which 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 have observed that using hybrid feature-extraction strategy, the overall performance of the *GPCRs* predictor can be improved. It is shown that by using hybrid feature-extraction strategy, which exploits both the spatial and transform domain variation of amino acid composition, the different types of *GPCRs* can be discriminated in a better way and consequently, high prediction performance can be achieved. We have also observed that *SVM* performs better than that of *PNN* and *NN*

for *GPCR* classification at any level. The performance of *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.

ACKNOLEDGMENT

This work was supported by the Higher Education Commission (*HEC*) under indigenous PhD program (074-1844-PS4-406).

REFRENCES

[1] K.H. Lundstrom and M.L. Chiu, G- protein coupled receptors in drug discovery, CRC Press, Taylor & Francis Group 6000 Broken Sound Parkway NW, Suite 300 Boca Raton, FL 33487-2742, 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] SF. Altschul, W. Gish, W. Miller, EW. Myers, DJ. Lipman, Basic local alignment search tool, J Mol Biol. 215 (1990), 403–410.

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

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

[6] S.S. Das, and 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 and K. Touhara, Insect sex-pheromone signals mediated by specific combinations of olfactory receptors, Science 307 (2005), 1638–1642.

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

[9] SM. Foord, S. Jupe and 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 and JE. 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 and 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 and G.P.S. Raghava, GPCRsclass: a web tool for the classification of amine type of Gprotein-coupled receptors, Nucleic Acids 33 (2005), 143-147.

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

[14] JX. Wang, P. Qin, QL. Liu, HY. Yang , YZ. Fan, Yu JK, S. Zheng, Detection and Significance of Serum Protein Marker of Hirschsprung Disease, Protein Eng. 120 (2007), e56-e60.

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

[16] PK. Papasaikas, PG. Bagos, ZI. Litou, SJ. Hamodrakas, A novel method for GPCR recognition and family classification from sequence alone using signatures derived from profile hidden Markov models, SAR and QSAR Environmental Research 14 (2003), 413-420.

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

[18] QB. Gao, Classification of G-Protein coupled receptors at four levels, Protein engineering, Design & Selection 19 (2006), 511-516.

[19] M.N. Davies, A. Secker, A.A. Freitas, M. Mendao, J. Timmis and D.R. Flower, On the Hierarchical classification of G-Proteon coupled receptors, Bioinformatics 23 (2007), 3113-3118.

[20] GPCRs dataset, http://www.cs.kent.ac.uk/projects/biasprofs/

[21] KC. Chou, Prediction of protein cellular attributes using pseudo-amino-acid-composition, Proteins 43 (2001), 246-255.

[22] A. Khan, M.F. Khan, T. Choi, Proximity Based GPCRs Prediction in Transform Domain, Biochemical and Biophysical Research Communications 371 (2008), 411–415.

[23] 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 and E. Yaschenko Database resources of the national center for biotechnology information, Nucleic Acids Res. 35 (2007), D5–D12.

[24] K.C. Chou, H.B. Shen, Review: Recent progresses in protein subcellular location prediction, Analytical Biochemistry 370 (2007), 1-16.

[25] 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.

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

[27] K.c. Chou: Prediction of G-protein-coupled receptor classes, J Proteome Res 4 (2005), 1413-1418.

[28] K.C. Chou, Prediction of protein cellular attributes using pseudo amino acid composition. PROTEINS: Structure, Function, and Genetics (Erratum: ibid., 2001, Vol.44, 60) 43 (2001), 246-255.

[29] S.W. Zhang, Q. Pan, H.C. Zhang, Z.C. Shao , J.Y. Shi, Prediction Protein Homo-oligomer Types by Pesudo Amino Acid Composition: Approached with an Improved Feature Extraction and Naive Bayes Feature Fusion, Amino Acids 30 (2006), 461-468.

[30] 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, Analytical Biochemistry 390 (2009), 68-73.

[31] Fauche`re J-L , Plis˘kaV, 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.

[32] 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.

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

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

[35] 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.

[36] A. Khan, S. F. Tahir, A. Majid, T.S. Choi, Machine Learning based Adaptive Watermark Decoding in View of an Anticipated Attack, Pattern Recognition, 41, 2594-2610.

[37] A. Khan, S. F. Tahir, T.S. Choi, Intelligent Extraction of a Digital Watermark from a Distorted Image, IEICE TRANS. INF. & SYST E91-D (2008), 2072-2075.

[38] J. Javed, A. Khan, A. Majid, A. M. Mirza, J.Bashir, Lattice Constant Prediction of orthorhombic ABO3 Perovskites using Support Vector Machines, Computational Materials Science 39 (2007), 627-634.

[39] D.F. Specht, Probablistic neural networks, Neural Networks 3 (1990), 109-118.

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

[41] 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.

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

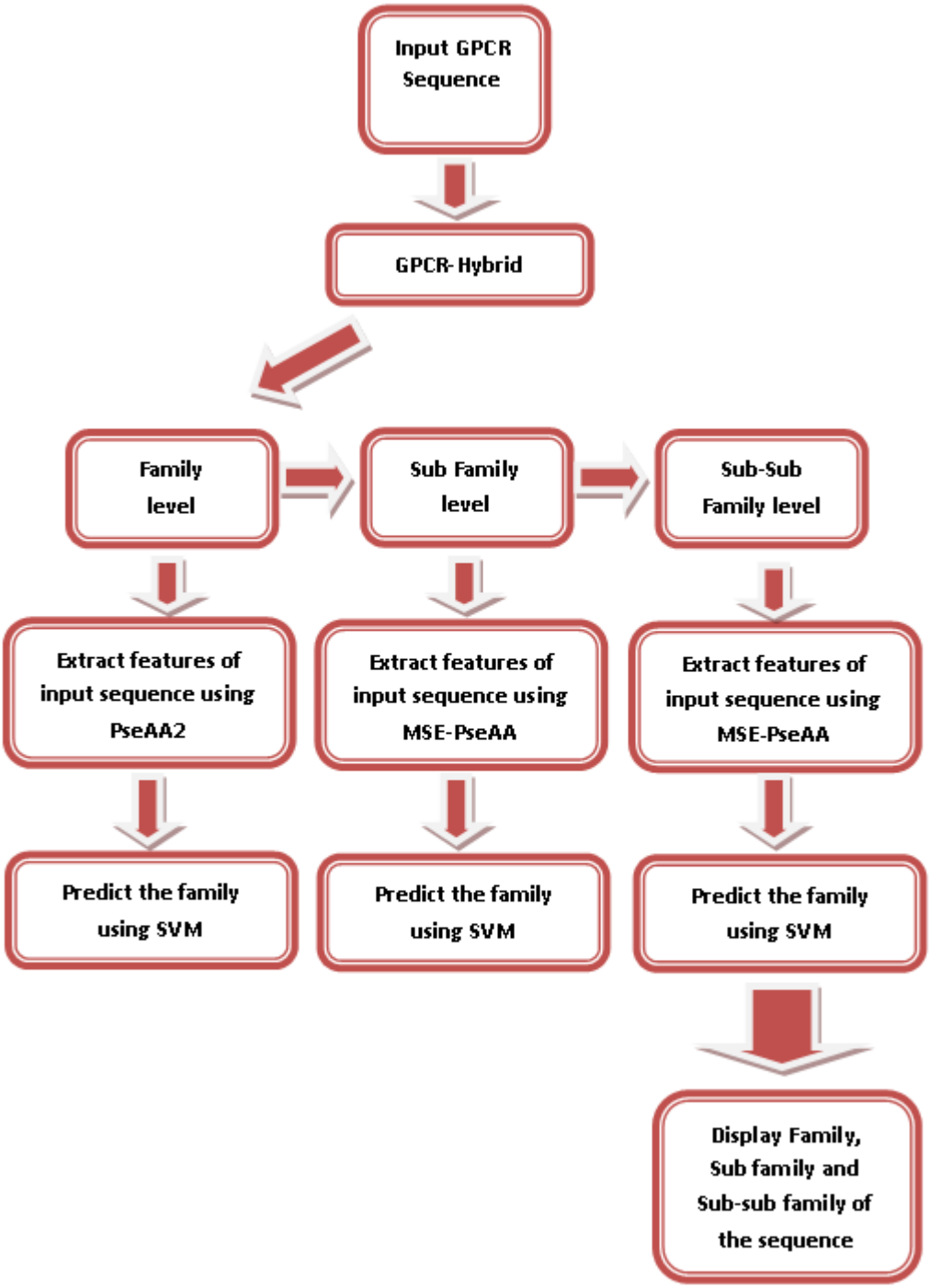
[43] Q.B. Gao, C. Wu, X.Q. Ma, J. Lu, J. He, Classification of amine type G-protein coupled receptors with feature selection, Protein Pept Lett 15 (2008), 834-842.

[44] Z. L. Peng, J. Y. Yang, X. Chen, An improved classification of G-protein-coupled receptors using sequence-derived features, BMC Bioinformatics 11 (2010).

[45] Xiao X, Wang P, Chou KC: GPCR-CA: A cellular automaton image approach for predicting G-protein-coupled receptor functional classes, J Comput Chem 30 (2009), 1413-1423.

List of Figures

List of tables



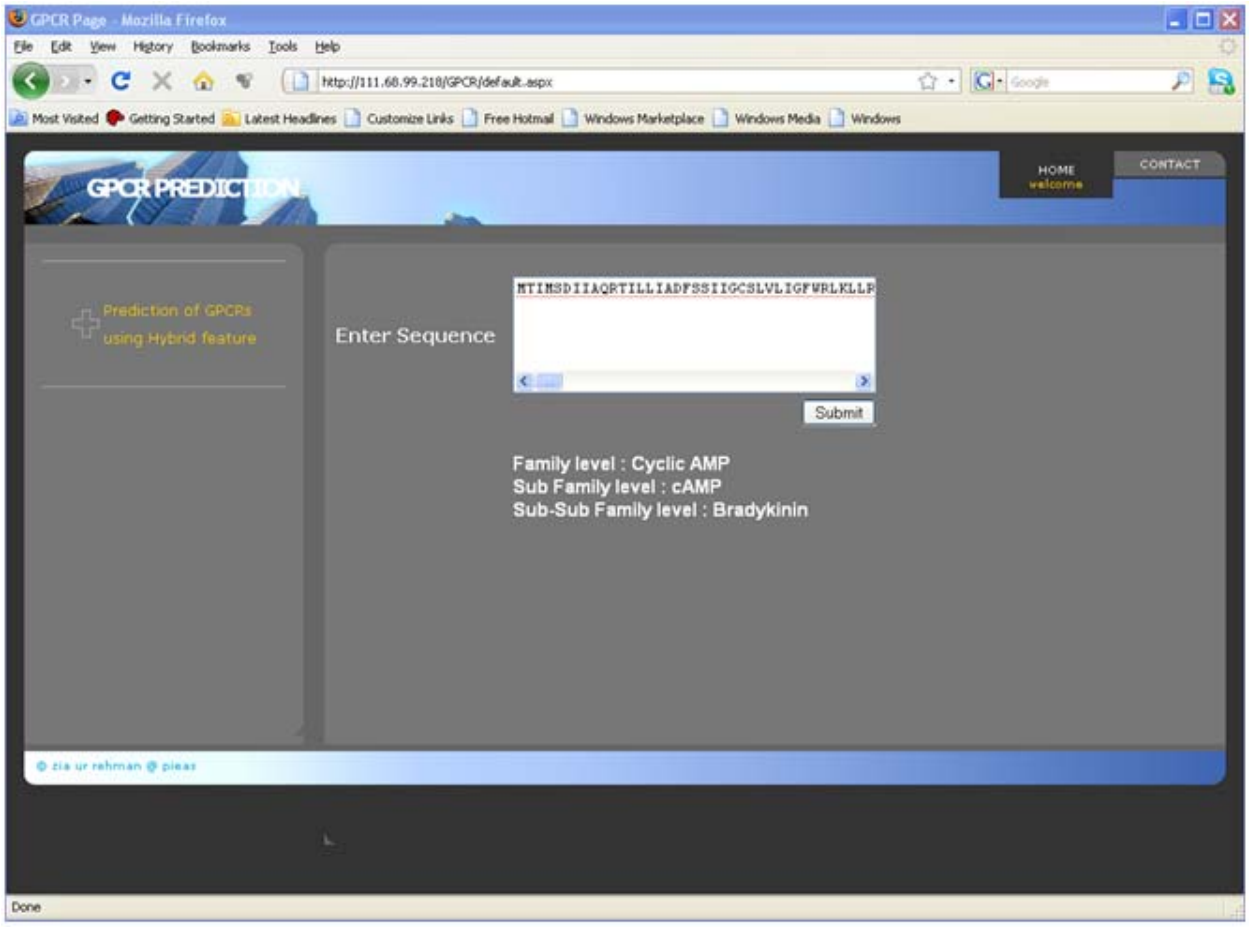

GPCR Page - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://111.68.99.218/GPCR/default.aspx
Most Visited Getting Started Latest Headlines Customize Links Free Hotmail Windows Marketplace Windows Media Windows
GPCR PREDICTION
HOME
welcome
CONTACT
Prediction of GPCRs using Hybrid feature
Enter Sequence
MTIMSDIIAQRTILLIADFSSIIGCSLVLIGFWRLKLLP
Submit
Family level : Cyclic AMP
Sub Family level : cAMP
Sub-Sub Family level : Bradykinin
© zia ur rehman @ pieas
Done

Table 1

GPCR Classification Performance for family level

Classifier	Feature extraction strategy	Accuracy (%)	Mathew's Correlation Coefficient	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

[a] PseAA2 = Feature vector formed using Pseudo amino-acid by considering two Physiochemical properties of amino acids
[b] PseAA3 = Feature vector formed using Pseudo amino-acid by considering three Physiochemical properties of amino acids
[c] MSE-AA = Hybrid feature vector formed by combining amino acid features with wavelet based multi scale energy features
[d] MSE-PseAA = Hybrid feature vector formed by combining Pseudo amino acid features with wavelet based multi scale energy features

Table 2

GPCR Classification Performance for sub family 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**

Table 3

Classification Performance for sub-sub family 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**

Table 4

Comparison with Selective Top down method

GPCRs Classification level	Selective top-down Accuracy (%) [19]	GPCR-Hybrid Accuracy (%)
Family	95.87	97.86
Sub family	80.77	84.97
Sub-Sub Family	69.98	75.60

Table 5

Comparison with other methods on D167 dataset

Reference	Overall Accuracy (%)
[25]	83.23
[40]	83.20
[41]	96.40
[42]	97.60
[43]	97.60
PCA-GPCR[44]	98.20
GPCR-Hybrid	**98.45**

Table 6

Comparison with other methods on D365 dataset

Method	Overall Accuracy (%)
GPCR-CA [45]	83.56
PCA-GPCR [44]	92.60
GPCR-Hybrid	**92.59**

Table 7

Comparison with other methods on D566 dataset

Method	Overall Accuracy (%)
PCA-GPCR [44]	97.88
GPCR-Hybrid	**97.91**