journal homepage: www.intl.elsevierhealth.com/journals/cmpb

Prediction of G-protein coupled receptors and their subfamilies by incorporating various sequence features into Chou's general PseAAC

Arvind Kumar Tiwari *

GGSCMT, Kharar, SAS Nagar, Punjab, India

ARTICLE INFO

Article history:
Received 19 February 2016
Received in revised form
27 May 2016
Accepted 1 July 2016

Keywords:
G-protein coupled receptors
Weighted k-nearest neighbor
Minimum redundancy maximum relevance
Sequence derived properties
Matthew's correlation coefficient

ABSTRACT

Background and objective: The G-protein coupled receptors are the largest superfamilies of membrane proteins and important targets for the drug design. G-protein coupled receptors are responsible for many physiochemical processes such as smell, taste, vision, neurotransmission, metabolism, cellular growth and immune response. So it is necessary to design a robust and efficient approach for the prediction of G-protein coupled receptors and their subfamilies.

Methods: In this paper, the protein samples are represented by amino acid composition, dipeptide composition, correlation features, composition, transition, distribution, sequence order descriptors and pseudo amino acid composition with total 1497 number of sequence derived features. To address the issue of efficient classification of G-protein coupled receptors and their subfamilies, we propose to use a weighted k-nearest neighbor classifier with UNION of best 50 features, selected by Fisher score based feature selection, ReliefF, fast correlation based filter, minimum redundancy maximum relevancy, and support vector machine based recursive elimination feature selection methods to exploit the advantages of these feature selection methods.

Results: The proposed method achieved an overall accuracy of 99.9%, 98.3%, 95.4%, MCC values of 1.00, 0.98, 0.95, ROC area values of 1.00, 0.998, 0.996 and precision of 99.9%, 98.3% and 95.5% using 10-fold cross-validation to predict the G-protein coupled receptors and non-G-protein coupled receptors, subfamilies of G-protein coupled receptors, and subfamilies of class A G-protein coupled receptors, respectively.

Conclusions: The high accuracies, MCC, ROC area values, and precision values indicate that the proposed method is better for the prediction of G-protein coupled receptors families and their subfamilies.

1. Introduction

G-protein coupled receptors (GPCRs) are seven-transmembrane domain receptors that sense molecules outside the cell and activate inside signal transduction pathways for cellular responses. These are called seven-transmembrane receptors because they pass through the cell membrane seven times. G-protein coupled receptors can be grouped into six classes based on sequence homology and functional similarity; these are Class A (Rhodopsin-like), Class B (Secretin like), Class C (Metabotropic glutamate), Class D (cyclic AMP), Class E (Taste),

* E-mail address: arvind.tiwari1980@gmail.com.

http://dx.doi.org/10.1016/j.cmpb.2016.07.004

and Class F (Vomeronasal) receptors. A larger number of G-protein coupled receptors are available in humans. In these, some of their function, such as growth factors, light, hormones, amines, neurotransmitters, and lipids, etc., have been identified. However, a large number of G-protein coupled receptors found in the human genome have unknown functions, and so it is necessary to design an efficient approach to predict families and subfamilies of G-protein coupled receptors for a new drug discovery.

The G-protein coupled receptors are transmembrane proteins which, via G-proteins, initiate some of the important signaling pathways in a cell and are involved in various physiological processes. Initially, Bhasin and Raghava [1] proposed an SVM based method by using amino acid composition and dipeptide of amino acids for the prediction of G. protein coupled receptor. Later, Bhasin and Raghava [2] proposed an SVM based method for the classification of amine type of G-protein-coupled receptors by using of amino acid composition and dipeptide composition of proteins. Gao and Wang [3] proposed a nearest neighbor method to discriminate GPCRs from non-GPCRs, and subsequently classify GPCRs at four levels on the basis of amino acid composition and dipeptide composition of proteins. Gu and Ding [4] have proposed a binary particle swarm optimization algorithm to extract effective feature for amino acids pair compositions of GPCRs protein sequence. Then they used ensemble fuzzy k-nearest neighbor classifier to predict GPCRs families. Gu et al. [5] proposed an Adaboost classifier to predict G-protein-coupled receptors by pseudo amino acid composition with approximate entropy and hydrophobicity patterns. Peng et al. [6] proposed a principal component analysis-based method for the prediction of GPCRs, family and their subfamilies by using sequence derived features. Lin and Xiao [7] proposed an ensemble of k-NN based classifier with grey incidence analysis and used pseudo amino acid composition to predict the GPCRs, family and their subfamilies. Xiao and Wang [8] have proposed a covariant discriminant-based approach by using grey level co-occurrence matrix obtained from cellular automata images of pseudo amino acid composition of protein sequence to predict the GPCRs and their functional classes. Xiao and Wang [9] have proposed a fuzzy k-NN based approach by using the combination of two different variation of pseudo amino acid composition. These are functional domain pseudo amino acid composition and low frequency Fourier spectrum pseudo amino acid composition to predict GPCR and their types. Xiao et al. [10] have proposed a fuzzy k-NN based approach to predict the interaction between GPCRs and drug in cellular network by using two dimensional fingerprint of pseudo amino acid composition generated through grey level model. Elrod [11] studied and observed that good and accurate data set is necessary for the prediction of GPCRs and their types. Chou [12] also studied about the coupling interaction between the receptors and G-binding proteins and observed that new therapeutic approached may be designed by manipulating the interaction of receptors and G-binding proteins. Chou [13] have proposed a covariant discriminant-based method by using amino acid composition to predict GPCRs and their classes. Qiu et al. [14] have proposed support vector machine based method by using pseudo amino acid compositions that are generated by using discrete wavelet transform to extract the feature from hydrophobicity scale of amino acid composition to predict the GPCRs classes. Zia-Ur Rehman and Khan [15] have proposed an ensemble classifier based on majority voting by using nearest neighbor, probabilistic neural network, and grey incidence degree and support vector machine and used combination of pseudo amino acid composition, wavelet based multi-scale energy and position-specific scoring matrix to predict the GPCRs and their subfamilies. Xie et al. [16] have proposed an ensemble support vector machine-based approach by using amino acid hydrophobicity-based pseudo amino acid composition to predict GPCRs and their subfamilies. Elord [17] have studied and observed that the amino acid compositions are closely related to GPCR's families. Therefore, it is necessary for good training data and for the identification of GPCR's families and subfamilies. Xiao and Lin [18] have reviewed and summarized the development and future challenges for the prediction of GPCRs and their subfamilies by using sequence derived properties.

Chou [19] studied and observed that five things are necessary for the identification of uncharacterized protein by using sequence-derived properties. These are construction of benchmark dataset, formulation of protein sample by using various sequence derived properties, design and development of computational intelligence-based method, and cross-validation test to measure the performance of the classifier and provide a web server that is easily accessible and available to public. Therefore, considering these points, for the construction of benchmark dataset, the sequence of the G-protein coupled receptors are extracted from GPCRDB [20] and all the non-G-protein coupled receptors proteins are selected from Uniport database with the keyword NOT G-protein coupled receptors. To avoid the homology bias, the CD-HIT [21] server is used to remove the homologous sequence. For the formulation of protein sample, eight feature vectors are extracted from protein sequence; amino acid composition, dipeptide composition, correlation, composition and transition, distribution of physiochemical properties, sequence order descriptors, and pseudo amino acid composition are used. This paper proposes a weighted k-nearest neighbor in which inverse kernel function is applied to calculate weighted distance to improve the prediction performance of G-protein coupled receptors families and their subfamilies by using sequence derived properties. For non-redundant, relevant, robust, and optimal feature subset selection, a feature selection method based on fusion of five supervised filter based methods is proposed. These supervised feature selection methods include Fisher score-based feature selection [22], ReliefF [23], fast correlation-based filter (FCBF) [24], minimum redundancy and maximum relevancy (MRMR) [25], and support vector machine-based recursive feature elimination (SVM-RFE) [26]. If we apply these feature selection methods on the same dataset, then each of them results in different feature subset where features are ranked. Also, the performance of a classifier for each feature subset selected by different method might be different. Therefore here, we address this problem by proposing a method for optimal feature selection by the fusion of feature subsets produced by these methods using union of the selected features by different feature selection algorithms. Further, in this paper, the proposed method used three level strategies to predict G-protein coupled receptors and their subfamilies. First, it is determined