Neurocomputing 154 (2015) 110–118

Contents lists available at ScienceDirect

Neurocomputing

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

An ensemble classifier based prediction of G-protein-coupled receptor classes in low homology

Quan Gu [a,b,*], Yong-Sheng Ding [a,c], Tong-Liang Zhang [a]

[a] College of Information Sciences and Technology, Donghua University, Shanghai 201620, China
[b] MRC—University of Glasgow Centre for Virus Research, Glasgow G11 5JR, United Kingdom
[c] Engineering Research Center of Digitized Textile & Fashion Technology, Ministry of Education, Shanghai 201620, China

ARTICLE INFO

Article history:
Received 15 August 2014
Received in revised form
29 October 2014
Accepted 7 December 2014
Communicated by Yang Tang
Available online 8 January 2015

Keywords:
G-protein-coupled receptors
Approximate entropy
Predicted secondary structural features
Ensemble classifier
Fuzzy *K*-nearest neighbor classifier

ABSTRACT

G-protein-coupled receptors (GPCRs) play an important role in physiological processes which are the targets of more than 50% of marketed drugs. In this research, we use a hybrid approach of predicted secondary structural features (PSSF) and approximate entropy (ApEn) as the feature selection method for predicting G-protein-coupled receptors in low homology. The low homology dataset is used to validate the proposed method for its objectivity. The classification model based on the fuzzy *K*-nearest neighbor classifier has been utilized on the classification of membrane proteins data. In order to enhance the prediction accuracies, here we propose an ensemble classifier as the prediction engine. Compared with the previous best-performing method, the success rate is encouraging. The reliable results also demonstrate the proposed method could contribute more to the characterization of various proteomes and further utilized in neuroscience.

1. Introduction

Prediction of native conformation of proteins from their primary structure is one of the most challenging problems in post genetic era. With the splendid development of the protein sequence data in recent years, there is a great demand for advanced method of observing the protein structure information.

G protein-coupled receptors (GPCRs), also known as G protein-linked receptors, seven trans-membrane receptors or serpentine receptors, encompass one of the largest eukaryotic integral membrane protein families in the human genome [10]. The membrane receptors modulate biological function by initiating cellular signaling in response to chemically diverse agonists [52]. As the largest, most ubiquitous and most versatile family of membrane receptors [48], GPCRs are major contributors to the information flow into cells and, as such, are associated with a multitude of diseases that make members of the family important pharmacological become targets [44].

At present, more than 50% of marketed drugs for human therapeutics act through GPCRs [43] which also mediate the actions of certain medications for treating some disorders (e.g. cardiovascular disease, drug dependency, and mental illness), with annual revenues exceeding $40 billion. As pointed out in a recent review, the structure of GPCRs is crucial for guiding the development of more specific drugs and can be combined with traditional chemical screening methods to improve and accelerate drug discovery [52].

Despite all the previous efforts, very few crystal GPCR structures have been determined until now. It is a challenging task to predict the membrane receptors structure by experimental method, which is mainly due to the difficulties of crystallizing or obtaining good X-ray diffraction data in membrane proteins [63]. Although the current high resolution nuclear magnetic resonance (NMR) spectroscopy is proficient in determining the 3-dimensional (3D) structures of membrane proteins, it is time-consuming and costly [47]. On the other hand, the amino acid sequences of more than 1000 GPCR-related proteins have been investigated [29]. On account of an enormous amount of data and the paramount importance of GPCRs, a highly accurate, sequence-based prediction of GPCR function has considerable practical value for both research biologists and pharmaceutical companies [1]. Since the 3D structural information is necessary for exploring protein function but as not all GPCRs are amenable to experimental structure determination, hence the computational prediction methods have become essential alternatives.

The pattern recognition has been widely used in bioinformatics [61]. With these considerations in mind, several rational pattern recognition approaches have been employed for the GPCRs categories prediction. In the field of classification, support vector machines (SVM) [1,17,24], hidden Markov model [53], simple alignment-free [59],

* Corresponding author at: Donghua University, Shanghai 201620, China. Tel.: +86 2167792301.
E-mail address: supergu0925@gmail.com (Q. Gu).

http://dx.doi.org/10.1016/j.neucom.2014.12.013
0925-2312/

probabilistic neural network (PNN) [63], fuzzy *K*-nearest neighbor (FKNN) [65], intimate sorting [7,45], AdaBoost [21], and covariant-discriminant [66] have been adopted for the prediction.

In the area of feature selection, the methods involving amino acids sequence information are exploited such as (but not limited to) amino acids composition (AA) [6] and fast Fourier transform (FFT) [23]. To avoid a complete loss of the sequence-order information, Chou proposed a pseudo amino acid (PseAA) composition [5] for sequence representation with the combination of the AA and additional information from protein sequence, which has been successfully applied to the representation of amino acid sequence in our previous works [68-70].

In the recent years, a series of GPCR researches have validated the effectiveness of PseAA [41,65]. In the present study, we utilize the predicted secondary structural sequence (PSSF) [11] for representing the general contents and spatial arrangements of the secondary structural elements of a given GPCR sequence, which are more reliable than the exploration only based on the protein primary structure. We describe a novel PseAA composition with PSSF and approximate entropy (ApEn) of protein sequence to predict the classification of GPCRs in low homology. ApEn is a non-negative parameter which denotes the complexity of a sequence by measuring the likelihood of pattern occurrence [49], and it has also been successfully applied to the field of representation of various kinds of sequences in our previous works (e.g. amino acid sequence [21], protein coding sequences [26]). As illustrated in Ref. [50], GPCRs share a common structural signature of the seven membrane-spanning helices with an extracellular N-terminus and an intracellular C-terminus. As shown in Fig. 1, the seven-α helix transmembrane structure can be imaged as time series crossing the filter parameters [21]. Therefore, inspired by such heuristic, we use ApEn as the feature of the GPCRs. Sequence homology is one of the main factors that heavily influence the accuracy of the prediction. As illustrated by the previous work [33], researchers often achieve high prediction accuracy owing to high homology or improper dataset. Therefore, a more credible prediction method based on the low homology dataset is more precise for the studies on protein structural prediction.

This paper is organized as follows. We first introduce the dataset of G-protein-coupled receptor data base (GPCRDB) with the low homology for validating the performance of our method. Second we investigate the secondary structural features vector and ApEn for the sequence expression. Then we use the ensemble classifier [20], an advanced algorithm for the prediction engine to improve the prediction accuracy given by basic FKNN classifiers. As shown in the present study, our approach outperforms prediction accuracy ~2% higher than the best previous published method. In general, the proposed approach appears very promising for GPCR prediction. The novelty of feature selection of approximate entropy and predicted secondary structural features could play an important complementary role to the existing methods in prediction of GPCR classes and further be explored in various proteomes. The ensemble classifier with increasing the prediction accuracy of basic classifier could not only be applied in proteomics but also be further applied in neuroscience [62], such as evolutionary computation [60] and neural network optimization [71].

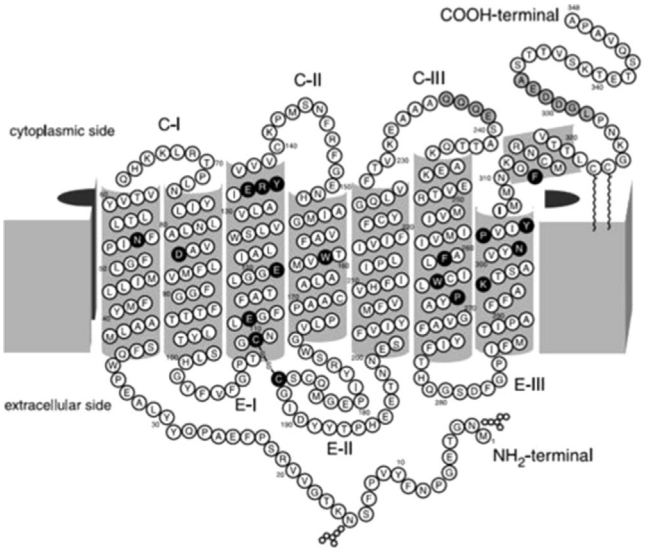

Fig. 1. The two-dimensional model of bovine rhodopsin [43].

2. Feature extraction

2.1. Database

The dataset investigated for this research are collected from the GPCRDB (the latest version) [25], which has been used in our prior work [21]. GPCRs have been grouped into five classes on the basis of sequence conservation, Class A—rhodopsinlike receptor, Class B—secretin like receptor, Class C—metabotrophic/glutamate/pheromone receptor, Class D—fungal pheromone, and Class E—cyclic adenosine monophosphate (cAMP) receptor. Rhodopsinlike receptors are further divided into groups associated with particular ligand specificity, such as the opsin, amine, peptide, cannabinoid, and olfactory receptors [4]. In this research we consider the former four classes, i.e. Class A–D, owing to the limited amount of Class E, which increases the bias of prediction accuracy which is mentioned in Ref. [21].

As illustrated in the literature [33], although the prediction accuracy is decreasing as the declining of the homology, it is more objective to verify the algorithm in the low homology dataset. In the present study, we adopt CD-HIT package [36] to reduce the homology of protein sequences. Compared with the whole dataset (i.e. GPCRDB), the low homology dataset called 40% GPCRDB with identity of protein sequence reduced to 40% (i.e. the lowest cutoff). There are 6620 sequences in the dataset GPCRDB. After processed by the CD-HIT, the number is reduced to 511. The sequences number of each GPCR class on different datasets is depicted in Fig. 2. As shown in the pie graph of the figure, the 40% GPCRDB is more balanced than GPCRDB, which also verifies the objective of the low homology dataset for the decrease of the bias of the prediction accuracy.

2.2. Pseudo amino acid composition

Compared with conventional protein composition (20-D), the concept of pseudo amino acid composition (PseAA) as originally introduced by Chou [5], which is defined in a $(20+\lambda)$-D features space, contains much more sequence-order information. A protein sample can be represented by $(20+\lambda)$-D vectors, where the λ is the number of additional properties of sequence.

As shown in Eq. (1), a protein sample can be represented by $(20+\lambda)$-D feature vectors [8].

$$X=[x_1,x_2,\ldots,x_{20},\ldots,x_{20+\lambda}]^T \tag{1}$$

where

$$x_i=\begin{cases}\dfrac{f_i}{\sum_{i=1}^{20}f_i+\sum_{i=1}^{\lambda}w_ip_i}(1\le i\le 20)\\ \dfrac{w_ip_i}{\sum_{i=1}^{20}f_i+\sum_{i=1}^{\lambda}w_ip_i}(21\le i\le 20+\lambda)\end{cases} \tag{2}$$

where the f_i $(1\le i\le 20)$ in Eq. (2) is the occurrence frequencies of 20 amino acids in sequence, i.e. the AA composition which was