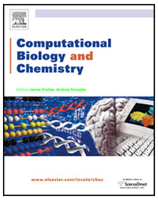
Computational Biology and Chemistry 59 (2015) 101–110

Contents lists available at ScienceDirect

Computational Biology and Chemistry

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



Research article

Maximizing lipocalin prediction through balanced and diversified training set and decision fusion

Abhigyan Nath*, Karthikeyan Subbiah**

Department of Computer Science, Banaras Hindu University, Varanasi 221005, India

ARTICLE INFO

Article history:
Received 2 December 2014
Received in revised form 8 September 2015
Accepted 23 September 2015
Available online 28 September 2015

Keywords:
Lipocalins
Diverse input patterns
Balanced training set
Boosted random forest
KNN
Classifier fusion schemes

ABSTRACT

Lipocalins are short in sequence length and perform several important biological functions. These proteins are having less than 20% sequence similarity among paralogs. Experimentally identifying them is an expensive and time consuming process. The computational methods based on the sequence similarity for allocating putative members to this family are also far elusive due to the low sequence similarity existing among the members of this family. Consequently, the machine learning methods become a viable alternative for their prediction by using the underlying sequence/structurally derived features as the input. Ideally, any machine learning based prediction method must be trained with all possible variations in the input feature vector (all the sub-class input patterns) to achieve perfect learning. A near perfect learning can be achieved by training the model with diverse types of input instances belonging to the different regions of the entire input space. Furthermore, the prediction performance can be improved through balancing the training set as the imbalanced data sets will tend to produce the prediction bias towards majority class and its sub-classes. This paper is aimed to achieve (i) the high generalization ability without any classification bias through the diversified and balanced training sets as well as (ii) enhanced the prediction accuracy by combining the results of individual classifiers with an appropriate fusion scheme. Instead of creating the training set randomly, we have first used the unsupervised Kmeans clustering algorithm to create diversified clusters of input patterns and created the diversified and balanced training set by selecting an equal number of patterns from each of these clusters. Finally, probability based classifier fusion scheme was applied on boosted random forest algorithm (which produced greater sensitivity) and K nearest neighbour algorithm (which produced greater specificity) to achieve the enhanced predictive performance than that of individual base classifiers. The performance of the learned models trained on Kmeans preprocessed training set is far better than the randomly generated training sets. The proposed method achieved a sensitivity of 90.6%, specificity of 91.4% and accuracy of 91.0% on the first test set and sensitivity of 92.9%, specificity of 96.2% and accuracy of 94.7% on the second blind test set. These results have established that diversifying training set improves the performance of predictive models through superior generalization ability and balancing the training set improves prediction accuracy. For smaller data sets, unsupervised Kmeans based sampling can be an effective technique to increase generalization than that of the usual random splitting method.

1. Introduction

Lipocalins are a part of calycin super-family along with FABPs (fatty acid binding proteins), Triabin, avidins and metalloprotease inhibitors (Bo Akerstrom et al., 2006). Lipocalins have significant diversity at the sequence level and perform a wide variety of biological functions. Apart from their diversity at sequence level as well as in functionalities, they are found in a variety of the organisms ranging from unicellular bacteria to multi-cellular plants and animals. Initially they are identified as the transporters of small hydrophobic molecules. Later they are found to be involved in immune-modulation (Logdberg and Wester, 2000) and are used as biomarkers for various diseases (Xu and Venge, 2000). Lipocalins are also found to have important roles in cell regulation and cancer (Bratt, 2000). Some of the animal lipocalins are found to behave as allergens (Virtanen et al., 1999). Artificial lipocalins are known as Anticalins (Skerra, 2008). Anticalins are being

* Corresponding author.
** Corresponding author.
E-mail addresses: abhigyannath01@gmail.com (A. Nath), karthinikita@gmail.com (K. Subbiah).

http://dx.doi.org/10.1016/j.compbiolchem.2015.09.011
1476-9271/© 2015 Elsevier Ltd. All rights reserved.

engineered to have highly specific molecular recognition functionality and offer a profitable technology over the conventional antibodies as promising reagents.

The experimental determination of lipocalins is an expensive and time consuming process. Moreover the detection of putative lipocalins using sequence similarity search methods is far elusive as the members of the lipocalin family share very low sequence similarity (Flower et al., 2000) often below the twilight zone (Rost, 1999). However the crystallographic structure of lipocalins reveals a conserved folding pattern that consists of eight beta strands and three structurally conserved regions (SCRs). This conserved pattern is having a close similarity with the one which is found in FABPs (Flower et al., 1993). Hence the presence of the structurally conserved pattern had inspired researchers to use the machine learning based prediction methods such as SVM for identifying the structurally diverse lipocalins by using sequential and structural features (Pugalenthi et al., 2010; Ramana and Gupta, 2009). Basically the sequence similarity scores are obtained by using the computationally significant comparison methods using sequence alignment algorithm, etc. In a nutshell, these methods use the primary sequences as their inputs. Whereas the machine learning methods use the underlying biological significant features that are extracted from the primary sequences as their inputs. So an intelligent pre-processing of input data for extraction of useful biologically significant sequence features is required. The appropriate choice of the extracted features will dictate the degree of success in solving the problems by applying machine learning methods.

In specific both of the methods (Pugalenthi et al., 2010; Ramana and Gupta, 2009) used the features derived from the predicted secondary structure and evolutionary information in the form of position specific scoring matrices (PSSM) along with other sequence based features. A detailed sequence and structural analysis of lipocalins was carried out to deduce the lipocalin fold and assigned LIR2 to lipocalin family by Adam et al. (Adam et al., 2008).

Previously machine learning methods have been successfully used for annotating the protein sequences belonging to various specific protein families (Chou, 2001; Pugalenthi et al., 2007; Shen and Chou, 2007; Kandaswamy et al., 2013). In this paper, we have attempted to enhance the prediction performance by using protocols for diversifying and balancing the training set as well as by applying classifier fusion schemes. Diversified training data set yields greater generalization ability and balanced training data set provides unbiased prediction performance. The classifier fusion schemes were used to achieve improved prediction accuracy in comparison to that of any individual classifier. The unsupervised Kmeans clustering was used to create the balanced and diverse training set and probability based fusion scheme for combining the results from the classifiers. The results of the experiments using these protocols have established that a balanced and diverse training set facilitates the machine learned models to have the superior generalization ability with unbiased performance as compared to that of a randomly created training set.

2. Materials and methods

2.1. Dataset

We have chosen the Ramana and Gupta datasets (Ramana and Gupta, 2009) for immediate comparison and robust analysis. This dataset consist of two parts, the first consists of 136 lipocalins and 166 non lipocalins for training and testing the models. The second part consists of 42 lipocalins, 25 FABPs and 28 Triabins, and is completely separate and mutually exclusive of the first. This second part of the data set is exclusively used for testing the machine learning models in order to get the unbiased prediction metrics and henceforth referred as the test set II.

2.2. Feature extraction

The selection of apposite input features for any machine learning model plays an important role in accurately classifying the input instances. Discovering the best combination of direct and derived features that are distinctively responsible for accurate classification is an extremely difficult task as there is no standard technique available for it. However, one can try to identify them through trial and error basis. A better combination of input features for a given classification problem can be identified through intelligently experimenting with different combinations of features with the aid of the problem's domain knowledge. For this study, we have used a combination of three sequence-based features, namely: amino acids composition, property group composition and physiochemical n-grams feature. The descriptions about these three features are described as follows:

2.2.1. Amino acid composition

The percentage composition of each of the twenty different amino acid residues (aa) is used as the first component of the input feature vectors and calculated using the formula:

$$\mathrm{PC}_{\mathrm{aa},i} = \frac{C_{\mathrm{aa},i}}{C_{\mathrm{res},i}} \times 100 \tag{1}$$

where aa denotes a specific one of the 20 amino acid residues, $\mathrm{PC}_{\mathrm{aa},i}$ denotes the amino acid percentage composition of specific type 'aa' in the ith sequence. $C_{\mathrm{aa},i}$ denotes the total count of amino acid of specific type aa in the ith sequence. $C_{\mathrm{res},i}$ denotes the total count of all residues in the ith sequence (i.e. sequence length).

Table 1
The amino acids property groups that have been taken for feature creation.

S.No.	Amino acid property group	Amino Acids in the Specific Group
1.	Tiny group	Ala, Cys, Gly, Ser, Thr
2.	Small group	Ala, Cys, Asp, Gly, Asn, Pro, Ser, Thr and Val
3.	Aliphatic group	Ile, Leu and Val.
4.	Non-polar groups	Ala, Cys, Phe, Gly, Ile, Leu, Met, Pro, Val, Trp and Tyr
5.	Aromatic group	Phe, His, Trp and Tyr
6.	Polar group	Asp, Glu, His, Lys, Asn, Gln. Arg, Ser, and Thr.
7.	Charged group	Asp, Glu, His, Arg, Lys
8.	Basic group	His, Lys and Arg
9.	Acidic group	Asp and Glu
10.	Hydrophobic group	Ala, Cys, Phe, Ile, Leu, Met, Val, Trp, Tyr
11.	Hydrophilic group	Asp, Glu, Lys, Asn, Gln,Arg