Computational Biology and Chemistry 53 (2014) 198–203

Contents lists available at ScienceDirect

Computational Biology and Chemistry

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

Inferring biological basis about psychrophilicity by interpreting the rules generated from the correctly classified input instances by a classifier

Abhigyan Nath[a], Karthikeyan Subbiah[b,*]

[a] Bioinformatics Section, Mahila Mahavidyalaya, Banaras Hindu University, Varanasi 221005, India
[b] Department of Computer Science, Banaras Hindu University, Varanasi 221005, India

ARTICLE INFO

Article history:
Received 28 May 2014
Received in revised form 2 September 2014
Accepted 6 October 2014
Available online 17 October 2014

Keywords:
Biologically interpretable rules
Cold adaptation
Amino acid composition patterns
Rotation forest
PART rule induction method

ABSTRACT

Organisms thriving at extreme cold surroundings are called as psychrophiles and they present a wealth of knowledge about sequence adjustments in proteins that had occurred during the adaptation to low temperatures. In this paper, we propose a new cascading model to investigate the basis for psychrophilicity. In this model, a superior classifier was used to discriminate psychrophilic from mesophilic protein sequences, and then the PART rule generating algorithm was applied on the input instances that are correctly classified by the classifier, to generate human interpretable rules. These derived rules were further validated on a structural dataset and finally analyzed to discover the underlying biological basis about the psychrophilicity. In this study, we have used one of the key features of psychrophilic proteins accountable for remaining functional in extreme cold temperature surroundings i.e., global patterns of amino acid composition as the input features. The rotation forest classifier outperformed all the other classifiers with maximum accuracy of 70.5% and maximum AUC of 0.78. The effect of sequence length on the classification accuracy was also investigated. The analysis of the derived rules and interpretation of the analyzed results had revealed some interesting phenomena such as the amino acids *A*,*D*, *G*, *F*, and *S* are over-represented, and T is under-represented in psychrophilic proteins. These findings augment the existing domain knowledge for psychrophilic sequence features.

1. Introduction

Extreme cold or hot temperatures are detrimental to the growth of organisms, yet a number of organisms thrive in these hostile environments. Psychrophiles are among the extremophilic organisms which have optimum growth even at the very low temperature surroundings. These psychrophilic proteins have been evolved through changes at the molecular level to cope up with the continuing threat posed by the low temperature surroundings such as *cold denaturation*, *local flexibility* etc., (Siddiqui and Cavicchioli, 2006) for carrying out normal physiological and other biochemical functions. A better understanding of psychrophilicity will aid us in designing biocatalysts that can remain functional at very low temperatures. Various other fruitful applications of psychrophilic enzymes have been cited in the research papers on biotechnology and biocatalysis (Bell et al., 1995; Margesin and Schinner, 1998; Timmis and Pieper, 1999).

Many classification schemes have been successfully attempted for discriminating mesophilic and thermophilic proteins using both sequence and structural features (Li et al., 2010; Lin and Chen, 2011; Wu et al., 2009; Zhang and Fang, 2006). This is due to the abundant availability of structural data for mesophilic and thermophilic proteins as compared to the scarce availability of structural data for psychrophilic proteins.

Previous studies had revealed significant amino acid compositional differences between mesophilic and psychrophilic protein sequences (Jahandideh et al., 2008, 2007a,b; Metpally and Reddy, 2009). The amino acid composition alone was proved to be a very useful feature for discriminating mesophilic and thermophilic protein sequences as reported in the literature (Gromiha and Suresh, 2008; Nikookar et al., 2012; Zhang and Fang, 2007). In our previous study, we had used a random forest classifier for discriminating psychrophilic protein sequences from mesophilic counterparts using sequence based features of amino acid composition and hydrophobic residue patterns (Nath et al., 2012).

* Corresponding author. Tel.: +91 9473967721.
E-mail addresses: abhigyannath01@gmail.com (A. Nath), karthinikita@gmail.com (K. Subbiah).

http://dx.doi.org/10.1016/j.compbiolchem.2014.10.002
1476-9271/© 2014 Elsevier Ltd. All rights reserved.

The molecular basis of the cold adaptation is yet to be explored in detail. It has been observed that temperature adaptation of proteins is linked with small adjustments that occurred along the amino acid sequences. These adjustments like an insertion/deletion/change of amino acids at one or more positions might have resulted in some distinctive amino acid compositional patterns (De Vendittis et al., 2008).

The major advantages of the machine learning classifiers are their high generalization ability and graceful performance degradation with respect to increasing errors in the input. But their main disadvantage is their poor interpretability due to black box nature as they do not provide any information regarding the exact relationship between the input instance and its corresponding classification category. The predictions/classifications/recognitions made by various machine learning techniques provide the prediction results without explaining the concrete basis on which the classification decisions were taken. So, the classification results are needed to be analyzed/interpreted to know the basis on which the classification decisions were taken which consequently help us in making the biological interpretation. This has inspired us for developing a novel three stage cascade model for inferring the underlying biological basis of psychrophilicity. Further, the feature ranking algorithms are used to validate whether the global patterns of amino acid composition can effectively be used as a significant input feature for discriminating psychrophilic and mesophilic protein sequences or not. We have also investigated the effect of length on discrimination accuracy.

2. Materials and methods

2.1. Dataset

Two datasets were created in this study. The first dataset was created from the protein sequences of completely sequenced species of psychrophiles and mesophiles as used by (Metpally and Reddy, 2009). A pre-processing was done on these sequences initially by removing those sequences having the keyword putative, predicted, hypothetical, and fragment and then removing the sequences having non-standard residues of '*B*','*J*','*O*','*U*','*X*', or '*Z*'. All the protein sequences belonging to *Lactobaccilus salivarius* had automatically gotten removed during this process. After the pre-processing the dataset is reduced to 19,403 psychrophilic protein sequences and 11,721 mesophilic sequences belonging to six psychrophiles and five mesophiles. The CD-HIT (Li and Godzik, 2006) program was then applied to the reduced dataset to cluster them with less than 40% sequence identity in order to eliminate the classifier bias due to redundancy. The resulting dataset contains 8677 sequences belonging to psychrophilic group and 5455 sequences belonging to mesophilic group. From this, resultant dataset 4000 psychrophilic sequences and 4000 mesophilic sequences were separated for creating training set by a random selection process and the remaining of 4677 psychrophilic and 1455 mesophilic sequences were used as testing set. This first dataset was used for both discriminating psychrophilic proteins by the rotation forest algorithm as well as for investigating the effectiveness of global amino acid composition patterns as an input feature in the discrimination. The correctly predicted instances by the rotation forest algorithm forms the subset of first dataset (called Rule Induction dataset) that is used for extracting rules by applying a rule generating algorithm on them to gain interpretability of the classification mechanism.

The second dataset of (Jahandideh et al., 2007a) which consists of 13 pairs of structurally defined psychrophilic and mesophilic proteins was used for verifying and validating the classification rules generated from the result of first data set.

2.2. Selection of input feature

In this work, we have focussed on exploring the biological basis of psychrophilicity in terms of molecular level changes. So, accordingly, the input feature vectors were generated by encoding each sequence into 20 length feature vectors of amino acid composition. These feature vectors are calculated using the formula:

$$f(i) = \frac{X(i)}{\sum_{i=1}^{20} X(i) \times 100} \qquad (1)$$

where $f\,(i)$ stands for the percentage frequency of ith residue (where 'i' varies from 1 to 20 indicating specific amino acids). $X\,(i)$ stands for the total number of residues of ith type.

2.3. Cascading model

A novel three stage cascading model was designed for discovering the basis for psychrophilicity. The first stage performs classification using a superior classifier for discriminating psychrophiles from mesophiles. The second stage generates the rule for determining psychrophilicity and mesophilicity based on the correctly classified sequences of the first stage by filtering out the falsely classified sequences. The third and final stage is the manual analysis and interpretation of the generated rules. The information flow in our *cascading approach* of this model is given below:

(i) The correctly classified input instances from the output of the first stage classifier become the input to the second stage rule generating PART algorithm.
(ii) The outputs of the second stage i.e., human interpretable rules are used for further analysis and interpretation in order to extract the biological basis of the psychrophilicity.

2.3.1. Classification protocol

In the first stage, ROF is used as a filter to sieve out the negatively predicted sequences from the dataset. This provides a rich set of strong instances for generating the interpretable classification rules. A variety of classifiers was experimented for selecting the best one for generating the classification rules. Both rotation forest and random forest (ensemble learning methods) performs better than all the other classifiers. The rotation forest (Rodriguez et al., 2006) creates the training data by randomly splitting feature vector into *K* subsets and principal component analysis is applied to each subset and trains each base classifier with whole dataset. In random forest (Breiman, 2001), the best among randomly chosen feature is chosen to split a node during the construction of the tree. In discriminating psychrophiles from mesophiiles, the rotation forest algorithm has a slight edge over random forest in the classification performance. So, the rotation forest was chosen for creating strong instances for rule generation.

The rotation forest (ROF) is an ensemble classifier that combines different models to reach an outcome. For a successful ensemble classifier there should be an optimal diversity and accuracy among the different base classifiers. Bagging is one of the prominent ensemble classifier schemes developed by (Breiman, 1996). The idea is to use a group of base classifiers, each independently grown with randomly chosen samples of the same size with replacement (bootstrap aggregation) from the given training set. Decision trees are the most preferable base classifiers for using with bagging as they are susceptible to slight changes in the training data. Diverse classifiers making errors in the different parts of the input space produce better results when they are