Process Biochemistry 44 (2009) 654–660

Contents lists available at ScienceDirect

Process Biochemistry

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

Discriminating acidic and alkaline enzymes using a random forest model with secondary structure amino acid composition

Guangya Zhang, Hongchun Li, Baishan Fang *

Department of Biotechnology and Bioengineering, Huaqiao University, Xiamen, Fujian, 361021, China

ARTICLE INFO

Article history:
Received 4 November 2008
Received in revised form 17 January 2009
Accepted 13 February 2009

Keywords:
Acidic enzyme
Alkaline enzyme
Adaptation mechanism
Secondary structure amino acid composition
Feature extraction
Random forests

ABSTRACT

Understating the adaptation mechanism of enzymes to pH extremes and discriminating them is a challenging task and would help to design stable enzymes. In this work, we have systematically analyzed the secondary structure amino acid compositions of 105 acidic and 111 alkaline enzymes, respectively. We found that the propensity of the individual residues to participate in different secondary structures might be a general stability mechanism for their adaptation to pH extremes. Based on it, we present a secondary structure amino acid composition method for extracting useful features from sequence, and a novel ensemble classifier named random forest was used. The overall prediction accuracy evaluated by the 10-fold cross-validation reached 90.7%. Comparing our method with other feature extraction methods, the improvement of the overall prediction accuracy ranged from 5.5% to 21.2%. The random forests algorithm also outperformed other machine learning techniques with an improvement ranging from 3.2% to 19.9%.

1. Introduction

Acidophiles and alkaliphiles are organisms that thrive under highly acidic (usually at pH 2.0 or below) or alkaline condition (with a pH of 9–11). During the past two decades, studies have focused on the physiology and molecular genetics of them to elucidate their mechanisms of adaptation to acidic or alkaline environment [1]. Industrial applications of them have also been investigated, and some commercial acidic and alkaline enzymes from them have brought great advantages to industry [2–4]. Thus, it is clear that acidophiles and alkaliphiles are quite important and interesting not only academically but also industrially. However, the internal pH values of acidophiles and alkaliphiles are close to neutrality, it is clear that their extracellular enzymes must be stable and active at the appropriate pH extreme [5].

The stability of acidic and alkaline enzymes has been studied in the biophysical and biotechnological research areas [6–8], because enzyme instability at pH extreme is one of the main bottlenecks in extending the application of it. Dubnovitsky et al. [6] determined the crystal structure of the phosphoserine aminotransferase from the obligatory alkaliphiles *Bacillus alcalphilus* at 1.08 Å resolution and compared to the other two neutrophilic homologs, they found that the alkaliphilic representatives possessed a set of distinctive structural features. Kelch et al. [7] analyzed the unfolding behavior and determined the structure of *Nocardiopsis alba* Protease A (NAPase), an acid-resistant, kinetically stable protease, and compared these results with a neutrophilic homolog, α-lytic protease (αLP). Although NAPase and αLP had the same number of acid-titratable residues, kinetic studies revealed that the height of the unfolding free energy barrier for NAPase was less sensitive to acid than that of αLP, thereby accounting for NAPase's improved tolerance of low pH. On the other hand, it has been investigated whether acidophily and alkaliphily can be detected at the amino acid level [8–10]; such studies have detected some preferences of acidic and alkaline enzymes for particular amino acids. However, to our present knowledge, there have been few parallel progresses with respect to theoretical predictions about acidic (or alkaline) enzyme stability, although it has been commonly applied to predict protein thermostabilization [11–13]. One of the main reasons is that it is difficult to collect the sequence and structure information about acidic and alkaline enzymes. As mentioned above, the internal pH values of acidophiles and alkaliphiles are close to neutrality, the proteome information of some acidophiles (*e.g.*: *Ferroplasma acidarmanus* [14]) and alkaliphiles (*e.g.*: *Bacillus halodurans* [15]) cannot be directly used, although their genome information was available. On the other hand, some non-acidophiles and non-alkaliphiles can also produce acidic or alkaline enzymes [16–17].

In the present work, we aim to design optimal predictors to discriminate acidic and alkaline enzymes based on sequence and structure information. As we know, many machine learning methods, such as neural network and support vector machine

* Corresponding author. Fax: +86 595 2269 1095.
E-mail address: zhgyghh@hqu.edu.cn (B. Fang).

1359-5113/$ – see front matter.
doi:10.1016/j.procbio.2009.02.007

have been successfully applied in many fields for data classification. However, they are definitely time consuming to find the appropriate function and optimal-free parameters for very large training datasets. So it is significant to find a new algorithm that is fast and robust. Here, we will make use of random forests (RFs), a novel tree-based ensemble approach, developed by late Breiman [18]. Researchers have shown that RF performed better than other machine learning methods in genomics and proteomics studies [19,20], as well as in protein–protein interaction prediction [21], gene expression data analysis [22] and more recently in prediction of DNA-binding residues in proteins [23].

In this paper, we propose a novel method for predicting acidic and alkaline enzymes using the RF algorithm in conjunction with a feature named secondary structure amino acid composition (ssAAC). We got 216 sequences with less 25% identity to each other and investigated the stability mechanism systematically. And, the predicting result was quite encouraging; it could achieve an overall accuracy of 90.7% with Matthew's correlation coefficient of 0.82, and with an area under ROC curve (AUC) of 0.958.

2. Materials and methods

2.1. Datasets

To obtain high-quality and unbiased dataset, the data were strictly screened according to the following procedures. (1) The classification of acidic enzymes (with optimal pH < 5.0) and alkaline enzymes (with optimal pH > 9.0) was based on BRENDA at http://www.brenda-enzymes.info/ [24] and the sequences were also downloaded from it, which came from the databases of UniProt/Swiss-Prot. (2) Sequences which have less than 100 amino acid residues were removed because they might be partial or just be fragments. (3) Sequences which contain three or more consecutive uncertain amino acids (i.e. "XXX," "XXXX," and so on) were removed. (4) To avoid any homologous bias, a redundancy cutoff was imposed by Blastclust [25] to exclude those sequence that have ≥25% sequence identity to any other in the same subset according to Chou and Cai's work [26]. Thus, a total of 216 sequences were generated that consist of 105 acidic enzymes, 111 alkaline enzymes. It could be obtained freely at http://iib.hqu.edu.cn/zhang/download.asp.

2.2. Secondary structure amino acid composition

Secondary structure prediction of the 216 sequences was carried out using the Predator program [27]. It could perform protein secondary structure prediction from a set of sequences, the method is especially appropriate for large-scale sequence analysis efforts [28,29]. The content of amino acid residues in helix, sheet and random coil regions were computed. It defined as

$$\mathrm{Comp}(i,j) = \sum \frac{n_{i,j}}{N_j}, \tag{1}$$

where i stands for amino acid residues, j stands for helix, sheet and coil. $n_{i,j}$ is the number of i in j, N_j is the total number of all 20 amino acids in j. All these calculations were performed by a C^{++} program developed in-house.

2.3. Random forests

Random forests is a novel ensemble classifier; it uses a similar but improved method of bootstrap as bagging. It uses the strategy of a random selection of a subset of m predictors to grow each tree, where each tree is grown on a bootstrap sample of the training set. This number, m, is used to split the nodes and is much smaller than the total number of variables available for analysis. For more detailed information, please see references proposed by Breiman [18].

2.4. Validation check methods

The performance and robustness of the model was evaluated by the self-consistency test and the independent test. Among the independent dataset test, the sub-sampling (n-fold cross-validation) test and jackknife test, which are often used for examining the accuracy of a statistical prediction method [30], the jackknife test was deemed the most objective that can always yield a unique result for a given benchmark dataset, as elucidated in [31] and demonstrated by [32]. Therefore, the jackknife test has been increasingly and widely adopted by investigators to test the power of various prediction methods [33–36]. However, since the jackknife test would take too much computational time, in the current study we adopted the n-fold cross-validation to demonstrate the performance of our method. We have carried out 2-fold, 3-fold, 4-fold, 5-fold, 7-fold, 10-fold, 20-fold and 50-fold cross-validation tests.

2.5. Waikato environment for knowledge analysis (Weka)

We have also approached the discrimination of acidic and alkaline enzymes using other machine learning techniques as alternatives to RF using the same datasets. All the algorithms implementations were achieved using the *Weka* package, which is an open-source collection of machine learning algorithms created at the University of Waikato in New Zealand. It is written in Java and is comprised of a powerful experimenter to run datasets through multiple algorithms [37]. The program is still in active development, so at the time of this manuscript the latest Version 3.5.8 was used.

2.6. Evaluation of the performance

The final performance of our method was determined by measuring the sensitivity (SE), specificity (SP), accuracy (ACC), Matthew's correlation coefficient (MCC) and the receiver operating characteristic (ROC) score. The ROC score is the area under the ROC curve (AUC) and were calculated automatically by the *Weka* software. The SE, SP, ACC and MCC parameters were calculated using equations (2)–(5), respectively.

$$\mathrm{SE} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{2}$$

$$\mathrm{SP} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} \tag{3}$$

$$\mathrm{ACC} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{TN}+\mathrm{FN}} \tag{4}$$

$$\mathrm{MCC} = \frac{\mathrm{TPTN}-\mathrm{FPFN}}{\sqrt{(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TP}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}} \tag{5}$$

where TP are true positives (alkaline enzymes predicted as alkaline); FN are false negatives (alkaline enzymes predicted as acidic); TN are true negatives (acidic enzymes predicted as acidic) and FP are false positives (acidic enzymes predicted as alkaline).

3. Results

3.1. Differences of amino acid composition in acidic and alkaline enzymes

The differences in the contribution of individual amino acid to the overall and the predicted secondary structures between acidic and alkaline enzymes are given in Fig. 1. From this figure, we observed that the amino acids in acidic and alkaline enzymes greatly varied. Here, if $|\mathrm{Comp}_{\mathrm{AK},i} - \mathrm{Comp}_{\mathrm{AC},i}| > 1$, we regard i as the significant amino acids, and they are listed in Table 1. From it, one can see that the overall significant amino acids in alkaline enzymes are Glu, Arg, Leu and Ala, whereas Ser, Thr and Asn in acidic enzymes. As we know, the Arg δ-guanido moiety can provide more surface area for charged interactions, its side chain contains one fewer methylene group than Lys, it has the potential to develop less unfavorable contacts with the solvent, and it more easily maintains ion pairs and a net positive charge at elevated pH [38]. Although our results shared a common tendency for some amino acids in alkaline enzymes, some differences exist among the cases. For example, Glu is found significantly higher in alkaline enzymes in our study; however, an analysis of alkaline M-protease suggested that the alkaline adaptation involved decreasing in Asp, Glu, and Lys residues [10]. On the other hand, Thr and Ser are known as the best residue for interacting with the water surrounding protein structure [39], this might be related to acidic adaptation for enzymes. An observation that has not been reported earlier and deserves mention is the consistent higher usage of Ser and Thr in acidic enzymes. From a structural viewpoint, Ser, Thr and Asn are polar uncharged residues; an increase in the number of polar but uncharged residues would be expected to help maintain the polar-outside/non-polar-inside balance that is critical for a folded protein in an aqueous environment [40].

Our results also show that there are marked, significant amino acid composition differences in the secondary structures of alkaline and acidic enzymes. In alpha-helix, alkaline enzymes