Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

Discrimination of acidic and alkaline enzyme using Chou's pseudo amino acid composition in conjunction with probabilistic neural network model

Zaheer Ullah Khan [a], Maqsood Hayat [a,*], Muazzam Ali Khan [b]

[a] Department of Computer Science, Abdul Wali Khan University Mardan, Mardan, KP, Pakistan
[b] College of Electrical and Mechanical Engineering (NUST), Islamabad, Pakistan

HIGHLIGHTS

- We develop an accurate and high throughput predictor for discrimination of acidic and alkaline.
- PseAA composition and SAAC are used as feature extraction schemes.
- Various classification algorithms are utilized.
- Two datasets were evaluated using 10-fold cross validation test.
- Best results are reported so far in the literature.

ARTICLE INFO

Article history:
Received 2 July 2014
Received in revised form
9 September 2014
Accepted 11 October 2014
Available online 22 October 2014

Keywords:
SVM
PNN
DT
K-Nearest neighbor
SAAC

ABSTRACT

Enzyme catalysis is one of the most essential and striking processes among of all the complex processes that have evolved in living organisms. Enzymes are biological catalysts, which play a significant role in industrial applications as well as in medical areas, due to profound specificity, selectivity and catalytic efficiency. Refining catalytic efficiency of enzymes has become the most challenging job of enzyme engineering, into acidic and alkaline. Discrimination of acidic and alkaline enzymes through experimental approaches is difficult, sometimes impossible due to lack of established structures. Therefore, it is highly desirable to develop a computational model for discriminating acidic and alkaline enzymes from primary sequences. In this study, we have developed a robust, accurate and high throughput computational model using two discrete sample representation methods Pseudo amino acid composition (*PseAAC*) and split amino acid composition. Various classification algorithms including probabilistic neural network (*PNN*), *K*-nearest neighbor, decision tree, multi-layer perceptron and support vector machine are applied to predict acidic and alkaline with high accuracy. 10-fold cross validation test and several statistical measures namely, accuracy, *F*-measure, and area under *ROC* are used to evaluate the performance of the proposed model. The performance of the model is examined using two benchmark datasets to demonstrate the effectiveness of the model. The empirical results show that the performance of *PNN* in conjunction with *PseAAC* is quite promising compared to existing approaches in the literature so for. It has achieved 96.3% accuracy on dataset1 and 99.2% on dataset2. It is ascertained that the proposed model might be useful for basic research and drug related application areas.

1. Introduction

Enzymes are biological catalysts, which are proteinaceous in structure. They can only work within a narrow range of temperature and pH. The pH value of underline environment greatly affects the enzyme activity. Whenever, the enzyme pH value is optimum then it is most effective. However, enzymes have a significant role in industrial applications as well as in medical areas, due to profound specificity, selectivity, and catalytic efficiency. Various factors like pH and temperature have a crucial effect on the enzymatic efficiency (Nakhil Nair et al., 2010). However, most of enzymes endure high activity in the pH range between 6 and 8. Many of acidic and alkaline enzymes derived from acidophilic and alkaliphiles make these organisms in order to survive in high acidic (usually at pH 2.0 or below) or

[*] Corresponding author.
E-mail addresses: zaheerkhan.cs@gmail.com (Z.U. Khan), m.hayat@awkum.edu.pk (M. Hayat).

http://dx.doi.org/10.1016/j.jtbi.2014.10.014
0022-5193/© 2014 Elsevier Ltd. All rights reserved.

alkaline conditions (with pH 9–11). Acidophiles and alkaliphiles have more contribution in biotechnology and industrial applications (Jordan et al., 1996; Sarethy et al., 2011).

The stability of acidic and alkaline enzymes has been studied in the biophysical and biotechnological related literature. Therefore, the stability of acidic and alkaline enzymes is essential because instability at extreme pH is one of the main bottlenecks in extending their applications (Dubnovitsky et al., 2005; Geierstanger et al., 1998; Kelch et al., 2007).

In this regard, a series of efforts have been carried out to discriminate acidic and alkaline enzymes. However, the theoretical methods have been achieved considerable success on the basis of primary and secondary enzyme sequences information at amino acids composition level, where sequences of enzymes and particular amino acids are correlated with the external environments of enzymes (Geierstanger et al., 1998; Shirai et al., 1997)

In a sequel, Zhang et al. (2009) have proposed a computational method for predicting acidic and alkaline enzymes. They have utilized random forest algorithm in conjunction with secondary structure amino acid composition. Likewise, Fan et al. (2013) have introduced a new approach for discriminating acidic and alkaline enzymes. Similarly, Lin et al. (2013) have developed a sequence-based method to discriminate acidic enzymes from alkaline enzymes. In this model, the ANOVA was applied to select the high discriminative features derived from g-gap dipeptide compositions and support vector machine was utilized to establish the prediction model. In addition, Chou (2011) has suggested a comprehensive review and published a series of publications for establishing a computational biological predictor (Chen et al., 2014c.; Fan et al., 2014; Guo et al., 2014; Liu et al., 2014; Qiu and Xiao, 2014; Xu et al., 2014). According to the comprehensive review, the first step is to construct or select a valid benchmark datasets. The second step is to formulate the biological samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the target to be predicted. The third step is to introduce or develop a powerful algorithm to operate the prediction; and finally to perform cross-validation tests to objectively evaluate the anticipated accuracy of the predictor.

In spite of tremendous enhancements have been made by pattern recognition and machine learning based approaches to solve this problem, still there exists some room of improvement, which demands for more attention and exploration.

In this concern, we propose a promising computational model to discriminate acidic and alkaline enzymes. In this model, two discrete protein sample representation methods namely *PseAAC* and split amino acid composition (*SAAC*) are used to extract numerical descriptors. *PseAAC* not only computes relative frequency of amino acid but also calculates correlation factors among amino acids. Various classification algorithms are investigated to select the best classification algorithm for this model on the same datasets. A 10-fold cross validation test is applied to assess the performance of proposed model.

The remaining paper is organized as follows: Section 2 describes materials and methods, Section 3 represents results and discussion and finally, conclusion has been drawn in Section 4.

2. Materials and methods

2.1. Benchmark datasets

In order to develop a quite promising computational model, a valid benchmark datasets are required to train the model. In this regards, we have used two benchmark datasets, the first one was originally used by Zhang et al. (2013), who collected and extracted the protein annotation information and sequences from enzyme database BRENDA (Lin et al., 2013) at http://www.brenda-enzyme.info/. In this dataset, enzymes were selected on different criteria, for acidic enzymes the optimal pH below 5.0 and for alkaline enzymes with optimal pH above 9.0. So the original dataset contains 217 enzymes including 105 of acidic enzymes and 112 of alkaline enzymes. Latter, 25% CDHit was applied to remove those sequences from dataset whose identity is more than 25%. Consequently, the second benchmark dataset contains 54 acidic and 68 alkaline enzymes, of total 122 enzymes (Fan et al., 2013; Lin et al., 2013).

2.1.1. Sample representation

In order to extract salient features from protein sequences, one best solution is to formulate or represent all the sequences with an effective strong mathematical expression that enable the sequences to exploit the key correlation with the target to be predicted. Several discrete sample representation methods namely, amino acid composition, dipeptide composition, evolutionary sample representation methods such as position specific scoring matrix, gene ontology, structure representation methods and physicochemical properties of amino acids are used for proteins formulation.

2.2. Pseudo amino acid composition

In this study, we have used pseudo amino acid composition (*PseAAC*) to extract numerical descriptors from enzymes sequences. Primary structure of protein is a polymer of 20 amino acids. However, simple amino acid composition only exhibits the occurrence frequency of each amino acid. As a result only 20 discrete attributes are extracted.

$$P_i = \frac{n_i}{L} \tag{1}$$

$$\sum_{i=1}^{u} p_i = 1 \tag{2}$$

where P_i represents the frequency of each amino acid, i indicates amino acid and L is the length of sequence.

However, each amino acid performs distinct role in formation of protein secondary structure. Therefore, information regarding the location of amino acids and sequence order is essential for discriminating acidic and alkaline enzymes. In order to incorporate correlation factors and sequence order information with simple amino acid composition Chou has introduced the concept of pseudo amino acid composition (*PseAAC*) (Chou, 2001, 2005, 2011). The concept of *PseAAC* has been adopted into almost all the fields of computational proteomics (Du et al., 2014). In addition, it has also penetrated into the area of computational genomics, such as using the pseudo *K*-tuple nucleotide composition (*PseKNC*) to formulate *DNA*/*RNA* sequences (Chen et al., 2012, 2014a, 2014b, 2014c; Guo et al., 2014; Qiu and Xiao, 2014). Likewise, it was used for other biological samples representation (Huang et al., 2012; Li et al., 2012). Recently three different powerful open access web-servers, called '*PseAAC*-Builder' (Du et al., 2012), 'propy' (Cao et al., 2013), and '*PseAAC*-General' (Du et al., 2014) were established to generate various modes of Chou's *PseAAC*. It can be formulated as

$$P = [p_1, \ldots p_{20}, p_{20+1}, \ldots p_{20+\lambda},]^T \tag{3}$$

where $p_1 \ldots\ p_{20}$ are the relative frequencies of 20 native amino acids and the remaining are the correlation factors of amino acids determined on the basis of hydrophobicity, hydrophilicity, charge