OPEN ACCESS Freely available online

PLOS ONE

# AcalPred: A Sequence-Based Tool for Discriminating between Acidic and Alkaline Enzymes

**Hao Lin[1]*, Wei Chen[2]*, Hui Ding[1]**

**1** Key Laboratory for NeuroInformation of Ministry of Education, Center of Bioinformatics, School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu, China, **2** Department of Physics, School of Sciences, Center for Genomics and Computational Biology, Hebei United University, Tangshan, China

## Abstract

The structure and activity of enzymes are influenced by pH value of their surroundings. Although many enzymes work well in the pH range from 6 to 8, some specific enzymes have good efficiencies only in acidic (pH<5) or alkaline (pH>9) solution. Studies have demonstrated that the activities of enzymes correlate with their primary sequences. It is crucial to judge enzyme adaptation to acidic or alkaline environment from its amino acid sequence in molecular mechanism clarification and the design of high efficient enzymes. In this study, we developed a sequence-based method to discriminate acidic enzymes from alkaline enzymes. The analysis of variance was used to choose the optimized discriminating features derived from *g*-gap dipeptide compositions. And support vector machine was utilized to establish the prediction model. In the rigorous jackknife cross-validation, the overall accuracy of 96.7% was achieved. The method can correctly predict 96.3% acidic and 97.1% alkaline enzymes. Through the comparison between the proposed method and previous methods, it is demonstrated that the proposed method is more accurate. On the basis of this proposed method, we have built an online web-server called AcalPred which can be freely accessed from the website (http://lin.uestc.edu.cn/server/AcalPred). We believe that the AcalPred will become a powerful tool to study enzyme adaptation to acidic or alkaline environment.

**Citation:** Lin H, Chen W, Ding H (2013) AcalPred: A Sequence-Based Tool for Discriminating between Acidic and Alkaline Enzymes. PLoS ONE 8(10): e75726. doi:10.1371/journal.pone.0075726

**Editor:** Vladimir N. Uversky, University of South Florida College of Medicine, United States of America

**Received** June 19, 2013; **Accepted** August 16, 2013; **Published** October 9, 2013



**Funding:** This work was supported by the National Natural Science Foundation of China (61202256, 61100092) and the Fundamental Research Funds for the Central Universities (ZYGX2012J113). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist.

* E-mail: hlin@uestc.edu.cn (HL); greatchen@heuu.edu.cn (WC)

## Introduction

An enzyme is able to multiply the speed of a chemical reaction by lowering the activation energy of participant molecules without any physical or chemical change. Due to high selectivity and catalytic efficiency, enzymes have been widely used in industry, medicine and environment management. Improving catalytic efficiency of enzymes has become the most important task of enzyme engineering. Although rational design and directional evolution can make the designed enzymes work better, environmental conditions also influence their activities. Solubility, temperature and pH value significantly influence enzyme activity [1]. Protein solubility is the basic condition in most biochemical experiments [2]. Enzyme activity increases with temperature rise because the heat enhances the kinetic energy of both substrates and enzymes, which results in more contact between them [3]. Catalytic efficiency is also largely influenced by pH value of their surroundings as the charge of amino acids varies with pH value [4]. In general, an enzyme has an optimum pH. Although most enzymes remain high activity in the pH range between 6 and 8, some specific enzymes work well only in extremely acidic (i.e. pH <5.0) or alkaline (i.e. pH >9.0) conditions. Some acidic and alkaline enzymes derived from acidophiles and alkaliphiles make these organisms survive in high acidic or alkaline conditions [5]. These enzymes also have great potentials in industrial applications. Thus, determination of the favorable pH value of an enzyme is important in academic study and industrial application.

Although the optimized environmental conditions can be obtained by the biochemical experimental approaches, the wet experimental technique is time-consuming and high-cost. Hence, it is highly desirable to develop theoretical methods for predicting appropriate environment of enzymes. The properties in primary sequences of enzymes correlate with their surrounding factors [2,6,7]. According to the correlation, machine learning methods have been proposed to predict soluble proteins [8–10] and thermophilic proteins [3,11–16] with the information derived from primary sequence. However, few successful cases were reported to predict acidic and alkaline enzymes based on their sequences because it was difficult to collect enough sequence and structure information about acidic and alkaline enzymes [6]. The growing experimental-confirmed proteins in recent years provide a chance to establish bioinformatics methods for accurate discriminating acidic enzymes from alkaline enzymes. Acidic and alkaline enzymes have some particular amino acids [17]. Based on these findings, Zhang et al. [6] presented a random forest model to distinguish acidic enzymes from alkaline enzymes by using sequence and structure information. The model can achieve the overall accuracy of 90.7% in the 10-fold cross-validation. However, the accuracy is still far from satisfaction. Besides, some high homologous sequences in their benchmark datasets result in the overestimation of accuracy. Furthermore, they did not provide a web server so that their method cannot be easily used to obtain desired data by the experimental scientists. Recently, Fan et al.

[18] built a free web server called Pred-enzyme to predict acidic and alkaline enzymes. The predictor can achieve an overall accuracy of 94.01% in 10-fold cross-validation. However, their predictor needs gene ontology (GO) information. Our statistical results show that most of proteins have no GO information (<50%). If a query protein has not been annotated in GO database and no homologous can be found in GO database, the prediction with the model is not available.

To overcome these disadvantages, we developed an effective method to discriminate acidic enzymes from alkaline enzymes based on their sequence information alone. A feature selection technique was used to pick out a number of informative features. On the basis of these features, the support vector machine (SVM) was performed to establish prediction model. Jackknife cross-validation was used to evaluate the performance of the proposed method. Prediction results demonstrate that the proposed method is reliable. Based on this method, a free online server called AcalPred was built to provide a useful tool for basic academic study and industrial application of acidic and alkaline enzymes.

## Materials and Methods

### Benchmark Dataset

The original dataset used in this study was obtained from Zhang et al. [6] who extracted the protein annotation information and sequences from enzyme database BRENDA [19] at http://www.brenda-enzymes.info/. In this dataset, only the acidic enzymes with optimal pH below 5.0 and alkaline enzymes with optimal pH above 9.0 were selected. Enzymes with sequence length less than 100 amino acids have been removed. This original dataset contains 105 acidic enzymes and 111 alkaline enzymes. It is well known that high similarity data can lead to erroneous estimation of the performance of the methods. To reduce homologous bias and redundancy, the program PISCES [20] was used to remove those enzymes that have more than 25% pairwise sequence identity to any other. Finally, the benchmark dataset contains 54 acidic enzymes and 68 alkaline enzymes. The 122 enzymes can be freely downloaded from our website (http://lin.uestc.edu.cn/server/AcalPred/data).

### The g-gap Dipeptide Composition

In pattern recognition, one of the key points is to generate a set of informative parameters. It has become a challenge in protein prediction to formulate proteins with an effective mathematical expression for truly reflecting the intrinsic properties of proteins. In the past two decades, various sequence parameters such as amino acid composition (AAC) [3], pseudo amino acid composition (PseAAC) [18] and position-specific scoring matrix (PSSM) [18] have been successfully employed to predict protein structure and function. Because the proximate dipeptide compositions can be used to describe the correlation between two proximate residues, they have been widely applied in protein prediction [21,22]. However, the intrinsic properties of protein sequences may be deposited in higher tier correlation of residues because of the hydrogen bonding in secondary structure [23,24]. Thus, we extended the proximate dipeptide composition to the *g*-gap dipeptide composition which can be used to describe the correlation between two residues.

Suppose a protein sequence $\boldsymbol{P}$ with $L$ amino acid residues as follows:

$$P = R_1 R_2 R_3 R_4 R_5 \ldots R_{L-3} R_{L-2} R_{L-1} R_L \quad (1)$$

where $R_1$ represents the amino acid residue at the sequence position 1, $R_2$ represents the amino acid residue at position 2 and so on. For each $g$ of the $g$-gap dipeptide, the feature vector of the protein sequence contains $20\times20 = 400$ components and can be formulated as:

$$P = \left[f_1^g, f_2^g, \cdots, f_\lambda^g, \cdots, f_{400}^g\right]^{\mathrm{T}} \quad (2)$$

where the symbol $\mathbf{T}$ denotes the transposition of the vector; $f_\lambda^g$ denotes the frequency of the λ-th g-gap dipeptide and is defined as:

$$f_\lambda^g = n_\lambda^g / \sum_{\lambda=1}^{400} n_\lambda^g = n_\lambda^g/(L-g) \quad (3)$$

where $n_\lambda^g$ denotes the number of the λ-th g-gap dipeptide. g = 0 indicates the correlation of two proximate residues; g = 1 describes the correlation between two residues with one residue interval; g = 2 indicates the correlation between two residues with the interval of two residues and so forth.

### Feature Selection Technique

Generally, the high dimension vector in feature set would cause the following three problems [25]: one is over-fitting which results in low generalization ability and overestimation of prediction model; another is information redundancy or noise which results in bad prediction accuracy and error description of intrinsic properties; the other is dimension disaster which results in a handicap for the computation or increase of computational time. To overcome these advantages and improve the prediction quality, it is necessary to pick out informative parameters with feature selection techniques to gain deeper insights into the intrinsic properties of protein sequences. Obviously, the best feature combination can be surely achieved by examining the performance of all kinds of feature sets. However, the computation time is so long that we cannot complete it. For economizing runtime and computational resource, a wise strategy is to use algorithm to find the optimal features.

Owing to the development of probability and statistics, some techniques such as principal component analysis (PCA) [26], minimal-redundancy-maximal-relevance (mRMR) [27] and diffusion maps [28] have been presented in sequence analysis and prediction. This study proposed a statistics-based algorithm called the analysis of variance (ANOVA) to score each of the features. The principle of ANOVA is to calculate the ratio ($F$ value) of features between groups and within groups for measuring feature variances [21]. Then the $F$ value ($F(\lambda)$) of the $\lambda$-th $g$-gap dipeptide in benchmark dataset is defined by:

$$F(\lambda) = \frac{s_B^2(\lambda)}{s_W^2(\lambda)} \quad (4)$$

where $s_B^2(\lambda)$ and $s_W^2(\lambda)$ are the sample variance between groups (also called Means Square Between, MSB) and sample variance within groups (also called Mean Square Within, MSW), respectively. They are given by:

$$s_B^2(\lambda) = SS_B(\lambda)/df_B \quad (5)$$

$$s_W^2(\lambda) = SS_W(\lambda)/df_W \quad (6)$$

where $df_B = K-1$ and $df_W = M-K$ are degrees of freedom for MSB and MSW, respectively. $K$ and $M$ represent the number of groups and total number of samples, respectively. $SS_B(\lambda)$ and $SS_W(\lambda)$ are the sums of squares of the $\lambda$-th feature between groups and within groups, respectively, which can be calculated by:

$$SS_B^2(\lambda) = \sum_{i=1}^{K} m_i \left( \frac{\sum_{j=1}^{m_i} f_\lambda^g(i,j)}{m_i} - \frac{\sum_{i=1}^{K}\sum_{j=1}^{m_i} f_\lambda^g(i,j)}{\sum_{i=1}^{K} m_i} \right)^2 \quad (7)$$

$$SS_W^2(\lambda) = \sum_{i=1}^{K}\sum_{j=1}^{m_i} \left( f_\lambda^g(i,j) - \frac{\sum_{i=1}^{K}\sum_{j=1}^{m_i} f_\lambda^g(i,j)}{\sum_{i=1}^{K} m_i} \right)^2 \quad (8)$$

where $f_\lambda^g(i,j)$ denotes the frequency of the $\lambda$-th $g$-gap dipeptide of the $j$-th sample in the $i$-th group. The $m_i$ denotes the number of samples in the $i$-th group. Thus we have $M = \sum_{i=1}^{K} m_i$.

The $F(\lambda)$ value in Eq.(4) reveals the correlation between the $\lambda$-th feature and the group variables. The $F(\lambda)$ will become large as the MSB becomes increasingly larger than the MSW. In the absence of differences between groups, the $F(\lambda)$ will be near to 1. In other words, the features with a larger $F(\lambda)$ indicate that it is a more highly relevant one for the target to be predicted. Hence, the features can be initially ranked according to $F$ value in Eq.(4). Subsequently, the incremental feature selection (IFS) is used to determine the optimal number of features. The IFS procedure includes the following steps: starting with one feature with the highest score in the feature set, adding the second feature with the second high score, adding the third feature with the third high score and repeating this process until all candidate features are added. Thus, for each gap $g$, there are 400 feature subsets consisted of 400 ranked $g$-gap dipeptides. Thus, the $t$-th feature subset is composed of $t$ ranked $g$-gap dipeptides and can be expressed as:

$$P_t^g = \left[ f_1^g, f_2^g, \cdots, f_t^g \right]^{\mathrm{T}} \quad 1 \le t \le 400,\ g \ge 0 \quad (9)$$

For 400 feature sets, the prediction accuracy was examined on the benchmark dataset by using jackknife cross-validation. Then we obtained the IFS curve in a 2D Cartesian coordinate system with index $t$ as its abscissa (or $X$-coordinate) and the overall accuracy as its ordinate (or $Y$-coordinate). When the $g$ was selected from 0 to $g_0$, there are $g_0$ IFS curves. With the peaks (or maximum accuracies) of these curves and comparison results of these accuracies, the optimal feature subset with parameters $t_0$ and $g_0$ can be obtained and expressed as:

$$P_{t_0}^{g_0} = \left[ f_1^{g_0}, f_2^{g_0}, \cdots, f_{t_0}^{g_0} \right]^{\mathrm{T}} \quad (10)$$

which can provide the maximum accuracy. Then the high-dimensional data can be projected into a low-dimensional space. The final classifier model was built by the optimal feature subset.

### Support Vector Machine

Support vector machine (SVM), as a powerful machine learning method, has been widely and successfully applied in protein bioinformatics [29–31]. The basic idea of SVM is to map data of samples into a high dimensional Hilbert space and use kernel function to seek a decision boundary that is able to separate two training data. The decision boundary is a hyperplane which can maximize the margin between the two sets in the feature vector space [32].

In this study, the software LibSVM designed by Lin's lab was used to implement SVM [33]. In this software, four kinds of kernel functions of linear function, polynomial function, sigmoid function and radial basis function (RBF), can be used to perform prediction. Empirical studies have demonstrated that the RBF outperforms the other three kinds of kernel functions in nonlinear classification. Thus the RBF kernel function was used in the current work. The regularization parameter $C$ and the kernel width parameter $\gamma$ were optimized via an optimization procedure according to a grid search approach. In grid research, the search spaces for parameter $C$ and $\gamma$ are from $2^{15}$ to $2^{-5}$ and from $2^{-5}$ to $2^{-15}$ with the steps of $2^{-1}$ and 2, respectively. The jackknife cross-validation was adopted in this search.

### Performance Assessment

The predictive capability and reliability of the method is estimated by four parameters: sensitivity ($S_n$), specificity ($S_p$), correlation coefficient ($CC$) and overall accuracy ($Ac$) that are defined as follows:

$$S_n = TP/(TP+FN) \quad (11)$$

$$S_p = TN/(TN+FP) \quad (12)$$

$$CC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP+FP) \times (TN+FN) \times (TP+FN) \times (TN+FP)}} \quad (13)$$

$$Ac = (TP+TN)/(TP+TN+FP+FN) \quad (14)$$

where $TP$ denotes the numbers of the correctly recognized alkaline enzymes; $FN$ denotes the numbers of the alkaline enzymes recognized as acidic enzymes; $FP$ denotes the numbers of the acidic enzymes recognized as alkaline enzymes; $TN$ denotes the numbers of correctly recognized acidic enzymes.

## Results and Discussion

In statistical prediction, the following three cross-validation methods are often used to evaluate the performance of a predictor: independent dataset test, subsampling (K-fold cross validation) test, and jackknife test [34,35]. Among the three cross-validation methods, the jackknife test is the least arbitrary and the most objective because it can yield a unique result for a given benchmark dataset, and hence has been increasingly used by investigators to examine the quality of various predictors. Accordingly, we adopted the jackknife cross-validation in this study to examine the anticipated success rates of the predictor.

### Predictive Accuracy

The correlation between two arbitrary amino acids with a distance of $g$ amino acids can be reflected by the frequencies of the $g$-gap dipeptides (Eq.(3)). For each gap $g$, we must find out the best feature subset which can achieve the best result. Here, taking 2-gap dipeptides as an example, we show the way to achieve the anticipated result. At first, the 400 2-gap dipeptides were ranked according to their $F$ values as defined by Eq. (4). The ranked 2-gap dipeptide with a higher $F$ value suggests that it is a more highly

relevant one for the discrimination between acidic and alkaline enzymes. Subsequently, based on the ranked 2-gap dipeptides, we can build 400 individual predictors for the 400 sub-feature sets by adding the ranked 2-gap dipeptides one by one from higher to lower ranks. It is well known that the sub-feature sets with high $F$ value can give more reliable information for classification. However, the number of the selected features is too small to afford enough information, which results in the poor prediction accuracy. For example, the 30$^{th}$ predictor can only produce the overall accuracy of 89.3% in jackknife test. On the contrary, the high dimension sub-feature sets contain enough information. However, the reduction of cluster-tolerant capacity of prediction model will lead to a bad prediction in cross-validation. An example is that the jackknife cross-validated accuracy of 400$^{th}$ predictor is only 82.0%. Therefore, the third step is to investigate the prediction performance for each of the 400 predictors with jackknife cross-validation and then plot the IFS curve. According to the IFS curve shown in Fig. 1, the overall accuracy reached its peak (Ac = 96.7%) when the top ranked 62 2-gap dipeptides were used. These dipeptides have the $F$ score more than 6.39 ($P$-value$<$0.0128). The successful prediction rates were 96.3% and 97.1% for acidic and alkaline enzymes, respectively.

It is necessary to investigate whether other $g$-gap sub-feature sets can obtain higher accuracies or not. We changed the $g$ from 0 to 10 and repeated the feature selection process to find the maximum accuracy of each $g$-gap dipeptides. For the convenience of observation and comparison, eleven IFS curves ($g$ varying from 0 to 10) were plotted in Fig. 1. These results indicate that the sub-feature set with parameters $t_0 = 62$ and $g_0 = 2$ is the best one among the 4400 (400×11) optimized feature sets. The area under receiver operating characteristic (ROC) curve (AUC) achieves 0.956 in the jackknife cross-validation.

To provide an overall view, the distribution for the $F$ values of the 400 2-gap dipeptides and their roles for the prediction model were given in Fig. 2. The features in blue boxes were positively correlated with acidic enzymes, while those in red boxes were positively correlated with alkaline enzymes. As shown in Fig.2, Arg (R), Leu (L) and Ile (I) are preferred in acidic enzymes and Asp (D), Tyr (Y), Ser (S) and Thr (T) are preferred in alkaline enzymes. The Arg is a basic amino acid with the largest Isoelectric point (10.76) among 20 types of amino acids, whereas the Asp is an acidic amino acid with the smallest Isoelectric point (2.98) among 20 types of amino acids. The pH environment has a major effect on ionic binding, which is essential for enzyme activation and chemical reactions. The Arg δ-guanido moiety can provide more surface area for charged interactions and more easily maintains ion pairs and a net positive charge at elevated pH [36]. Therefore, the reason that acidic or alkaline enzymes contain many basic or acidic amino acids is that they need such specific residues to neutralize with extremely acidic (pH <5.0) or alkaline (pH >9.0) surroundings for executing enzymes' activities.

For demonstrating the prediction capability of the proposed model, we built an independent dataset which contained 20 acidic and 20 alkaline enzymes. These sequences derived from BRENDA [19] can be freely downloaded from our website. The sequence identity between training benchmark dataset and independent set is less 40%. Our model can correctly identify the 19 acidic and 20 alkaline enzymes. Furthermore, we investigated the accuracy of our method on another independent constructed by Fan et al. [18]. Results showed that 17 acidic and 16 alkaline enzymes could be correctly predicted when optimized cutoff was selected.

## Comparison with Other Methods

To further demonstrate the performance of the proposed method, it is necessary to compare it with other existing methods. However, it is not objective and strictly to directly compare the results due to different benchmark datasets. Therefore, we repeated the process of feature selection and prediction on the original dataset (105 acidic and 111 alkaline enzymes). It should be noted that the results reported by Zhang et al. [6] and Fan et al. [18] were derived by 10-fold cross-validation test. As elucidated by Chou [35], their test can not provide unique result. For the current case, the benchmark data set contains 105 acidic and 111 alkaline enzymes. According to the Equations 28 and 29 in [35,37], if one tenth samples are selected from each of the two subsets for conducting the 10-fold cross-validation, the number of possible combinations will be more than $10^{29}$, which is too large to be completed. Therefore, in previous studies [6,18], one of $10^{29}$ possible combinations is randomly picked out to perform the 10-fold cross-validation. To make the comparison between our method and their methods with the same test method, we also randomly picked one of the possible combinations from the same

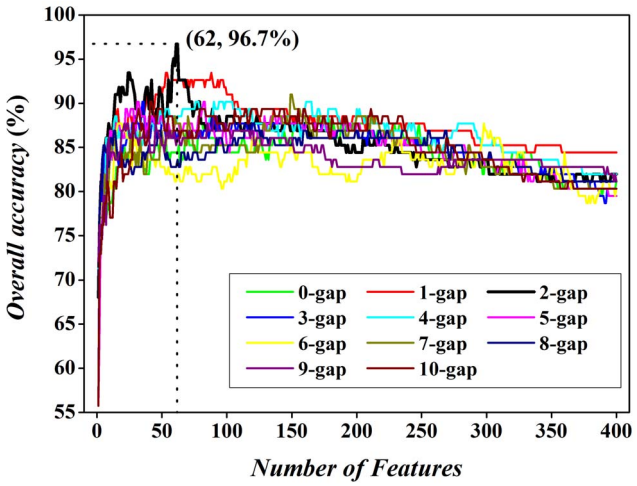


**Figure 1. A plot to show the IFS procedure.** When the top 62 2-gap dipeptides were used to perform prediction, the overall success rate reached its peak of 96.7%.
doi:10.1371/journal.pone.0075726.g001

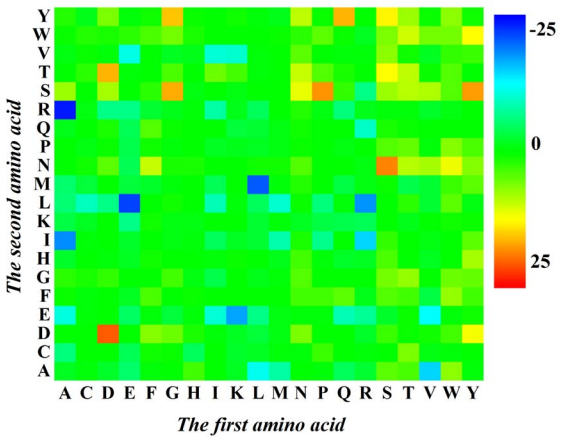


**Figure 2. A chromaticity diagram for the $F$ values of 400 2-gap dipeptides.** The blue boxes were positively correlated with acidic enzymes, while the red boxes were positively correlated with alkaline enzymes.
doi:10.1371/journal.pone.0075726.g002

**Table 1.** Comparing the performance of the proposed method with other existing methods.

| | *Sn*(%) | *Sp*(%) | *CC* | *Ac*(%) | *AUC*(%) |
|---|---|---|---|---|---|
| **Our method** | **94.6** | **94.3** | **0.89** | **94.4** | **0.975** |
| Zhang's method | 88.6 | 92.8 | 0.82 | 90.7 | 0.958 |
| Fan's method | 92.4 | 95.5 | 0.88 | 94.0 | 0.961 |

doi:10.1371/journal.pone.0075726.t001

benchmark data set to perform the 10-fold cross-validation test and the compared results were recorded in Table 1.

According to Table 1, when the top 81 1-gap dipeptides are used, our method can achieve the maximum accuracy of 94.4% with the AUC of 0.975 in 10-fold cross-validation, which is higher than the maximum accuracy obtained with other methods. Although the *Sp* obtained by our method is not the best, the *Sn*, *CC* and *Ac* are dramatically better than those of other methods, suggesting that the proposed method outperforms other published methods.

We noticed that Zhang et al. [6] and Fan et al. [18] also achieved encouraging results. Zhang's [6] proposed to use secondary structure amino acid composition as inputting parameters. This kind of information derived from software Predator program. It should be noted that the accuracy of Predator program was only about 75%. If the secondary structure of a protein chain is not correctly predicted, it will provide wrong information for further acidic/alkaline enzyme description. This is the possible reason that the Zhang's models can not obtain higher accuracies with the predicted secondary structural feature. The parameter sets of Fan et al. [18] model include average chemical shift (acACS) information, Go information and evolutionary (PSSM) information. In fact, their novel feature acACS can only achieve the overall accuracy of 85.7%. Go and PSSM information play an important role in their model construction. It is well known that some proteins don't have the Go annotation. We investigated the number of proteins in Uniprot and found that less than 50% proteins have GO information. Thus, their model can not provide any information for the protein that has not been annotated in GO database. Moreover, the PSSM information also has shortcomings. The generation of PSSM of a protein depends largely on the searching dataset. If no homologous sequence is found in the searching dataset, the PSSM will not give exact description, thus leading to wrong prediction. With primary sequence information, our model can obtain such high accuracy, suggesting that the proposed model is more neat free and efficient.

## Web-Server Guide

For the convenience of the vast majority of experimental scientists, we built a free web server called AcalPred to discriminate acidic enzymes from alkaline enzymes. Below, let us give a step-by-step guide on how to use the AcalPred web server. Then experimental scientists may get the desired results without the complicated mathematic equations. The detailed steps are provided as follows:

**Step 1.** Open the web server at http://lin.uestc.edu.cn/server/AcalPred and you will see the homepage of AcalPred on your computer screen, as shown in Fig. 3. Click on the Read Me button to see a brief introduction about the predictor and the caveat. Users may click on the Data button to download the training set and test set. By clicking on the Citation button, users

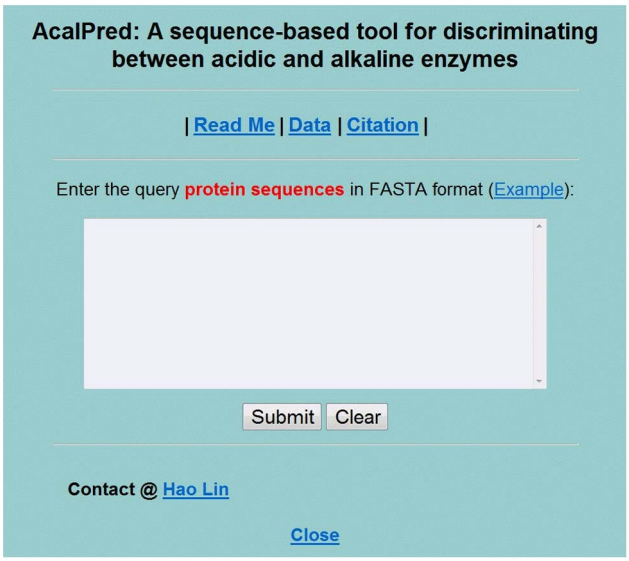


**Figure 3. A semi-screenshot to show the top page of the AcalPred web-server.** Its website address is at http://lin.uestc.edu.cn/server/AcalPred.
doi:10.1371/journal.pone.0075726.g003

may find the relevant papers on the detailed development and algorithm of AcalPred.

**Step 2.** Input or copy/paste the query protein sequence that you want to predict into the input text area at the center of Fig. 3. The input sequence should be in the FASTA format. A sequence in FASTA format consists of a single initial line beginning with a greater-than symbol ('>') in the first column, followed by lines of sequence data. The words right after the '>' symbol in the single initial line are optional and only used for the purpose of identification and description. All lines should be no longer than 120 characters and usually do not exceed 80 characters. The sequence ends if another line starting with a '>', which indicates the start of another sequence. Example sequences in FASTA format can be seen by clicking on the Example button right above the input box.

**Step 3.** Click on the Submit button to see the predicted result. The probabilities belonging to two classes will be given in the second and third columns. The first column gives the prediction type with prediction probability is above 0.5. For example, if you use the query protein sequences in the Example window as the input, after clicking the Submit button, you will see the following contexts on your screen: the outcome for the first query sample is 'acidic enzyme' because the prediction probabilities of acidic enzyme and alkaline enzyme are respectively 0.903627 and 0.096373; the outcome for the second query sample is 'alkaline enzyme' because the prediction probabilities of acidic enzyme and alkaline enzyme are 0.074938 and 0.925062, respectively.

## Conclusion

In this work, we developed a promising method to discriminate acidic enzymes from alkaline enzymes. The ANOVA-based feature selection technique was utilized to optimize dipeptide compositions for improving the prediction capability of model. An overall accuracy of 96% was achieved, demonstrating that the proposed model is a powerful tool for the study of enzymes in the adaptation to acidic or alkaline environment. For the convenience of experimental scientists, a free web server AcalPred was built to

implement the prediction. A friendly guide was given to describe the way to use the AcalPred web server. We believe that the predictor will be helpful for wet lab scientists who focus on enzyme activity.

## Author Contributions

Conceived and designed the experiments: HL. Performed the experiments: HL WC. Analyzed the data: HL WC HD. Contributed reagents/materials/analysis tools: HL HD. Wrote the paper: HL WC.

## References


1. Nikhil UN, Carl AD, Zhao H (2010) Engineering of Enzymes for Selective Catalysis. Current Organic Chemistry 14: 1870–1882.
2. Diaz AA, Tomba E, Lennarson R, Richard R, Bagajewicz MJ, et al. (2010) Prediction of protein solubility in Escherichia coli using logistic regression. Biotechnol Bioeng 105: 374–383.
3. Lin H, Chen W (2011) Prediction of thermophilic proteins using feature selection technique. J Microbiol Methods 84: 67–70.
4. Dubnovitsky AP, Kapetaniou EG, Papageorgiou AC (2005) Enzyme adaptation to alkaline pH: atomic resolution (1.08 A) structure of phosphoserine aminotransferase from Bacillus alcalophilus. Protein Sci 14: 97–110.
5. Takami H, Horikoshi K (2000) Analysis of the genome of an alkaliphilic Bacillus strain from an industrial point of view. Extremophiles 4: 99–108.
6. Zhang G, Li H, Fang B (2009) Discriminating acidic and alkaline enzymes using a random forest model with secondary structure amino acid composition. Process Biochemistry 44: 654–660.
7. Idicula-Thomas S, Balaji PV (2005) Understanding the relationship between the primary structure of proteins and its propensity to be soluble on overexpression in Escherichia coli. Protein Sci 14: 582–592.
8. Magnan CN, Randall A, Baldi P (2009) SOLpro: accurate sequence-based prediction of protein solubility. Bioinformatics 25: 2200–2207.
9. Smialowski P, Martin-Galiano AJ, Mikolajka A, Girschick T, Holak TA, et al. (2007) Protein solubility: sequence based prediction and experimental verification. Bioinformatics 23: 2536–2542.
10. Idicula-Thomas S, Kulkarni AJ, Kulkarni BD, Jayaraman VK, Balaji PV (2006) A support vector machine-based method for predicting the propensity of a protein to be soluble or to form inclusion body on overexpression in Escherichia coli. Bioinformatics 22: 278–284.
11. Nakariyakul S, Liu ZP, Chen L (2012) Detecting thermophilic proteins through selecting amino acid and dipeptide composition features. Amino Acids 42: 1947–1953.
12. Gromiha MM, Suresh MX (2008) Discrimination of mesophilic and thermophilic proteins using machine learning algorithms. Proteins 70: 1274–1279.
13. Taylor TJ, Vaisman II (2010) Discrimination of thermophilic and mesophilic proteins. BMC Struct Biol 10 Suppl 1: S5.
14. Wang D, Yang L, Fu Z, Xia J (2011) Prediction of thermophilic protein with pseudo amino Acid composition: an approach from combined feature selection and reduction. Protein Pept Lett 18: 684–689.
15. Zhang G (2013) A simple statistical method for discrimination of thermophilic and mesophilic proteins based on amino acid composition. Int J Bioinform Res Appl 9: 41–52.
16. Zhang G, Fang B (2006) Support vector machine for discrimination of thermophilic and mesophilic proteins based on amino acid composition. Protein Pept Lett 13: 965–970.
17. Settembre EC, Chittuluru JR, Mill CP, Kappock TJ, Ealick SE (2004) Acidophilic adaptations in the structure of Acetobacter aceti N5-carboxyaminoimidazole ribonucleotide mutase (PurE). Acta Crystallogr D Biol Crystallogr 60: 1753–1760.
18. Fan G-L, Li Q-Z, Zuo Y-C (2013) Predicting acidic and alkaline enzymes by incorporating the average chemical shift and gene ontology informations into the general form of Chou's PseAAC. Process Biochemistry 48: 1048–1053.
19. Chang A, Scheer M, Grote A, Schomburg I, Schomburg D (2009) BRENDA, AMENDA and FRENDA the enzyme information system: new content and tools in 2009. Nucleic Acids Res 37: D588–592.
20. Wang G, Dunbrack RL Jr. (2005) PISCES: recent improvements to a PDB sequence culling server. Nucleic Acids Res 33: W94–98.
21. Lin H, Ding H (2011) Predicting ion channels and their types by the dipeptide mode of pseudo amino acid composition. J Theor Biol 269: 64–69.
22. Gromiha MM, Suwa M (2005) A simple statistical method for discriminating outer membrane proteins with better accuracy. Bioinformatics 21: 961–968.
23. Ding H, Guo S-H, Deng E-Z, Yuan L-F, Guo F-B, et al. (2013) Prediction of Golgi-resident protein types by using feature selection technique. Chemometrics and Intelligent Laboratory Systems 124: 9–13.
24. Chou KC (2001) Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins 43: 246–255.
25. Ding C, Yuan LF, Guo SH, Lin H, Chen W (2012) Identification of mycobacterial membrane proteins and their types using over-represented tripeptide compositions. J Proteomics 77: 321–328.
26. Ma J, Gu H (2010) A novel method for predicting protein subcellular localization based on pseudo amino acid composition. BMB Rep 43: 670–676.
27. Jia P, Qian Z, Feng K, Lu W, Li Y, et al. (2008) Prediction of membrane protein types in a hybrid space. J Proteome Res 7: 1131–1137.
28. Yin JB, Fan YX, Shen HB (2011) Conotoxin superfamily prediction using diffusion maps dimensionality reduction and subspace classifier. Curr Protein Pept Sci 12: 580–588.
29. Lin H, Ding H, Guo FB, Zhang AY, Huang J (2008) Predicting subcellular localization of mycobacterial proteins by using Chou's pseudo amino acid composition. Protein Pept Lett 15: 739–744.
30. Kumar M, Gromiha MM, Raghava GP (2011) SVM based prediction of RNA-binding proteins using binding residues and evolutionary information. J Mol Recognit 24: 303–313.
31. Chen C, Chen L, Zou X, Cai P (2009) Prediction of protein secondary structure content by using the concept of Chou's pseudo amino acid composition and support vector machine. Protein Pept Lett 16: 27–31.
32. Vapnik V (1998) Statistical learning theory. Wiley-Interscience, New York.
33. Fan RE, Chen PH, Lin CJ (2005) Working set selection using second order information for training support vector machines. J Mach Learn Res 6: 1889–1918.
34. Chou KC, Zhang CT (1995) Prediction of protein structural classes. Crit Rev Biochem Mol Biol 30: 275–349.
35. Chou KC (2011) Some remarks on protein attribute prediction and pseudo amino acid composition. J Theor Biol 273: 236–247.
36. Vieille C, Zeikus GJ (2001) Hyperthermophilic enzymes: sources, uses, and molecular mechanisms for thermostability. Microbiol Mol Biol Rev 65: 1–43.
37. Chen W, Feng PM, Lin H, Chou KC (2013) iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. Nucleic Acids Res 41: e68.