J Membrane Biol (2015) 248:1033–1041
DOI 10.1007/s00232-015-9815-8

iCataly-PseAAC: Identification of Enzymes Catalytic Sites Using Sequence Evolution Information with Grey Model GM (2,1)

Xuan Xiao[1,2,3] · Meng-Juan Hui[1] · Zi Liu[1] · Wang-Ren Qiu[1]

Received: 23 March 2015 / Accepted: 6 June 2015 / Published online: 16 June 2015

Abstract Enzymes play pivotal roles in most of the biological reaction. The catalytic residues of an enzyme are defined as the amino acids which are directly involved in chemical catalysis; the knowledge of these residues is important for understanding enzyme function. Given an enzyme, which residues are the catalytic sites, and which residues are not? This is the first important problem for in-depth understanding the catalytic mechanism and drug development. With the explosive of protein sequences generated during the post-genomic era, it is highly desirable for both basic research and drug design to develop fast and reliable method for identifying the catalytic sites of enzymes according to their sequences. To address this problem, we proposed a new predictor, called iCataly-PseAAC. In the prediction system, the peptide sample was formulated with sequence evolution information via grey system model GM(2,1). It was observed by the rigorous jackknife test and independent dataset test that iCataly-PseAAC was superior to exist predictions though its only use sequence information. As a user-friendly web server, iCataly-PseAAC is freely accessible at http://www.jci-bioinfo.cn/iCataly-PseAAC. A step-by-step guide has been provided on how to use the web server to get the desired results for the convenience of most experimental scientists.

Keywords Catalytic active sites · Pseudo amino acid composition · Grey system model · Web server · iCataly-PseAAC

Introduction

Enzyme has been attracting the interests of most experimental scientists because that enzyme is one of the most important biological catalysts. The catalytic residues of an enzyme are defined as the amino acids which are directly involved in chemical catalysis reaction. The knowledge of these residues would be very helpful to understand enzyme function because they are closely related to the function of the enzyme as well as its specificity and molecular mechanisms. The catalytic residues are more conserved than other residues during evolution (Dou et al. 2011). Various biochemical experiments can identify the catalytic active sites. The results from experimental methods have not only provided reliable catalytic sites but also indicated that the catalytic sites were closely correlated with the local downstream and upstream residues from the catalytic sites of their center, respectively. Unfortunately, even if the number of local residues was limited at $\xi = 10$, 11, or 12 for both downstream and upstream, it is by no means easy to determine all the catalytic sites. This is because the

Electronic supplementary material The online version of this article (doi:10.1007/s00232-015-9815-8) contains supplementary material, which is available to authorized users.

✉ Xuan Xiao
jdzxiaoxuan@163.com

Meng-Juan Hui
huimengjuan@163.com

Zi Liu
liuzi189836@163.com

Wang-Ren Qiu
qiuone@163.com

[1] Computer Department, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen 333046, China

[2] Information School, ZheJiang Textile & Fashion College, NingBo 315211, China

[3] Gordon Life Science Institute, 53 South Cottage Road, Belmont, MA 02478, USA

number of possible peptide sequence N thus formed from 20 amino acids runs into

$$N = 20^{2\xi} = 10^{2\xi \log(20)} = \begin{cases} 1.0486 \times 10^{26}, & \xi = 10 \\ 4.1943 \times 10^{28}, & \xi = 11 \\ 1.6777 \times 31^{31}, & \xi = 12 \end{cases} \quad (1)$$

which is an astronomical figure for any of the above three cases! This would be exhausting to purely utilize the experimental approaches to determine the large-scale catalytic sites. With the explosive of protein sequences generated during the post-genomic era, it is highly desired to developed computational methods by which one can identify the catalytic active sites in enzymes.

In fact, during the last decade, many attempts have been made in this regard. For instance, Bartlett et al. derived the specific criteria to define a catalytic residue, and used to build an enzyme catalytic residue dataset (2002). Protein Data Bank (PDB) also provides the catalytic sites annotation for enzymes (Berman et al. 2002; Berman et al. 2000). The sequence information (Dou et al. 2011; Ota et al. 2003; Zhang et al. 2008; Dou et al. 2010; Chea and Livesay 2007; Fischer et al. 2008)and the structure information (Bartlett et al. 2002; Torrance et al. 2005; Zvelebil and Sternberg 1988) have been used in predicting catalytic sites, respectively, or together. Chien et al. developed a structure-based method based on residue side chain orientations and backbone flexibility of enzyme structure (2012). The contribution of structure information is usually less than that of sequence information in existing methods (Chien and Huang 2013). The feature selection methods also have been used to predict catalytic sites (Gao et al. 2013). Besides, the prediction engine or classification algorithm is also a key for prediction accuracy of predictor, the most common algorithms are *K*-nearest neighbor rule, neural work methods, and Support Vector Machine (Tong et al. 2008; Gutteridge et al. 2003). Although the above-mentioned approaches have its own virtue and played an important role in provoking the development in this area, they all need improvement from one or more of the following aspects: (i)The benchmark dataset used by the previous investigators needs to be updated by incorporating some new and experiment-confirmed data, or improved by removing redundancy and duplicate sequences; (ii) Further enhancing the prediction quality by introducing the state-of-the-art machine learning techniques; (iii) Making the formulation of all the statistical samples purely based on the sequence information alone because some of the existing methods needed the structural information that was not always available and hence would unavoidably suffer from some limitation; (iv) Establishing user-friendly and freely accessible web server to improve the service efficiency of the predictor.

The present study was initiated with an attempt to develop a new predictor for identifying enzyme catalytic sites by focusing on the above-mentioned four aspects.

According to a view (Xiao et al. 2012), to establish an efficient statistical predictor, we need to adopt the following procedures: (1) construct a valid benchmark dataset; (2) set an effective mathematical expression to formulate the statistical samples; (3) a powerful algorithm (or engine) to operate the prediction; (4) properly perform jackknife test to objectively evaluate the anticipated accuracy of the predictor. Below, let us describe how to realize these procedures one by one.

Materials and Methods

Benchmark Dataset

In order to develop a statistical predictor, it is fundamentally important to establish a reliable and stringent benchmark dataset to train and test the predictor. If the benchmark dataset contains some errors, the predictor trained by it must be unreliable and the accuracy tested by it would be completely meaningless.

In this study, the all active site information of benchmark dataset ware derived from the Catalytic Site Atlas (CSA) (version 2.2.12) (Porter et al. 2004), we also download the complete protein sequences which containing these active sites from the website http://www.uniprot.org/ UniProt (2007). In order to construct a valid benchmark dataset, the protein sequences with less than 50 amino acid residues were removed because they may be just fragments (Xiao et al. 2011), sequences are composed of non-standard amino acids are also removed. To avoid any homology bias, a redundancy cutoff was imposed using the program CD-Hit (Fu et al. 2012) to winnow those sequences which have $\geq$40 % pairwise sequence identity to any other in the dataset.

According to Chou's peptide formulation that was used for studying other sites (Chou 1996, 2001), a peptide with catalytic sites located at its center (Fig. 1) can be expressed as

$$P(R) = R_{-\xi}R_{-(\xi-1)} \cdots R_{-2}R_{-1}RR_{+1}R_{+2} \cdots R_{+(\xi-1)}R_{+\xi}, \quad (2)$$

where the subscript ξ is an integer (cf. Eq. 1), $R_{-\xi}$ represents the ξ-th downstream amino acid residue from the center, $R_{+\xi}$ the ξ-th upstream amino acid residue, and so forth (Fig. 1).

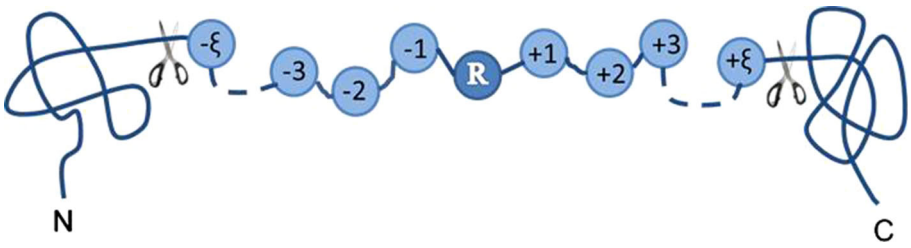

Fig. 1 An illustration to show scheme for a peptide of (2ξ+1) residues with catalytic site at the center

The ((2ξ + 1)-tuple) peptides $P_\xi(R)$ can be further classified into the following categories:

$$P_\xi(R) \in \begin{cases} P_\xi^+(R), & \text{if its center is a methylation site} \\ P_\xi^-(R), & \text{otherwise} \end{cases} \tag{3}$$

where $\in$ represents "a member of" in the set theory.

It is need to separate a benchmark dataset into training dataset and testing dataset for examining the performance of prediction method. Thus, the benchmark dataset for the current study can be formulated as

$$\begin{cases} S_R = S_R^+ \cup S_R^- \\ S_K = S_K^+ \cup S_K^- \end{cases} \tag{4}$$

where S_{R} is the benchmark dataset for training, S_{K} the benchmark dataset for testing, $\cup$ the symbol for "union" in set theory, S_R^+ contains the samples for the catalytic sites peptide only, S_R^- the sample for the non-catalytic sites peptide only, and so forth.

Since the length of the peptide $P_\xi(R)$ is (2ξ + 1) (cf. Eq. 2), the benchmark dataset with different values of ξ will contain peptides of different numbers of amino acid residues, as formulated by

$$P_\xi \text{ contains the peptides of} \begin{cases} 19 \text{ residues}, & \text{when } \xi = 9 \\ 21 \text{ residues}, & \text{when } \xi = 10 \\ 23 \text{ residues}, & \text{when } \xi = 11 \\ \vdots & \end{cases} \tag{5}$$

The detailed procedures to construct P_ξ are as follows. After considering the treatment by the previous investigators, we chose $\xi = 10$ (cf. Eq. 2) to construct the samples for the benchmark dataset S_{R} and S_{K}. If the downstream or the upstream in a protein was less than 10, the lacking residues were complemented by the way of circulation. The peptide samples thus obtained were subject to a screening procedure to winnow those that were identical to any other.

Finally, 811 proteins containing 1433 active sites were obtained. We randomly selected 611 protein chains as training dataset for S_{R} and the rest of 200 protein chains as testing dataset for S_{K}. For the S_{R}, which is a total of 1086 positive samples for S_R^+, these peptides are given in Online Supporting Information S1, and we randomly selected 3258 without active sites as negative samples for S_R^-, which are given in Online Supporting Information S2. The testing dataset contains 347 positive samples for S_K^+, which are given in Online Supporting Information S3 and 1041 randomly selected negative samples for S_K^-, which are given in Online Supporting Information S4.

Sample Representation

To develop a powerful predictor for identifying enzyme catalytic sites according to the sequence information, one of the key is to formulate the peptide samples with an effective mathematical expression that can truly reflect sequence information. To realize this, the pseudo amino acid composition (PseAAC) was proposed to avoid completely losing the sequence-order information, and replace the simple amino acid composition (AAC) for representing the sample of a protein (Nakashima et al. 1986; Chou and Zhang 1994; Schaffer et al. 2001).

The PseAAC of a protein is actually a set of discrete numbers that is derived from its amino acid sequence; the PseAAC for a peptide P can be generally formulated as

$$P = [\Psi_1, \Psi_2, \cdots, \Psi\mu, \cdots, \Psi_\Omega]^T, \tag{6}$$

where the T is a transpose operator and the subscript Ω is an integer and its value as well as the components Ψ_1, Ψ_2,…Ψ_Ω will depend on how to extract the desired information from the protein or peptide sequence of P.

In this paper, we incorporated the AAC and protein sequential evolution information in order to capture as much useful information from a protein or peptide sequence as possible.

Amino Acid Composition

Among the discrete models, the simplest one is the AAC widely used to transform peptide sequences into 20-D (dimensional) numerical vectors as defined by

$$P_{\mathrm{AAC}} = [\, f_1 \quad f_2 \quad \cdots \quad f_{20} \,]^T, \tag{7}$$

where $f_u(u = 1,2,\ldots,20)$ are the normalized occurrence frequencies of the 20 native amino acid in peptide P (Nakashima et al. 1986; Chou and Zhang 1994).

Grey-PSSM Approach

Sequential evolution information shows the course of biological species development. Many similarities between initial and resultant amino acid sequences are gradually eliminated after a long period of development time, but the corresponding proteins may share many common attributes. So the sequential evolution information is very effective for predicting various protein properties, such as protein function and subcellular location.

To extract the sequential evolution information and use it to define the components of Eq. 6, as described below.

According to (Schaffer et al. 2001), the sequence evolution information for a peptide with 21 amino acid residues can be represented by a 21 × 20 matrix as given by

$$P_{\mathrm{PSSM}}^{(0)} = \begin{bmatrix} m_{1\to1}^{(0)} & m_{1\to2}^{(0)} & \cdots & m_{1\to20}^{(0)} \\ m_{2\to1}^{(0)} & m_{2\to2}^{(0)} & \cdots & m_{2\to20}^{(0)} \\ \cdots & \cdots & \cdots & \cdots \\ m_{21\to1}^{(0)} & m_{21\to2}^{(0)} & \cdots & m_{21\to20}^{(0)} \end{bmatrix}, \tag{8}$$

where $m_{i\to j}^{(0)}$ represents the original score of amino acid residue in the i-th (i = 1,2,3,…,21) sequence position of peptide that is being charged to amino acid type j (j = 1,2,3,…,20) during the evolution process. Here the numerical codes 1,2,…,20 are used to denote the native amino acid types according to the alphabetical order of their single character codes (Chou et al. 2012). The 21 × 20 score in Eq. 8 is gained by PSI-BLAST (Schaffer et al. 2001) to search the UniPortKB/Swiss-Prot database (Release 2010_05) through three interactions with 0.001 as the E value cutoff for multiple sequence alignment against the sequence of the peptide P. In order to make every element in Eq. 8 within the range of 0–1, a conversion was performed through the standard sigmoid function to make it become

$$P_{\mathrm{PSSM}}^{(1)} = \begin{bmatrix} m_{1\to1}^{(1)} & m_{1\to2}^{(1)} & \cdots & m_{1\to20}^{(1)} \\ m_{2\to1}^{(1)} & m_{2\to2}^{(1)} & \cdots & m_{2\to20}^{(1)} \\ \vdots & \vdots & \vdots & \vdots \\ m_{21\to1}^{(1)} & m_{21\to2}^{(1)} & \cdots & m_{21\to20}^{(1)} \end{bmatrix}, \tag{9}$$

where

$$m_{i\to j}^{(1)} = \frac{1}{1+\mathrm{e}^{-m_{i\to j}^{(1)}}} \quad (1 \le i \le 21,\ 1 \le j \le 20). \tag{10}$$

We can directly extract $\Omega = 20$ elements features from the Eq. 9, described as

$$P_{\mathrm{PSSM}} = [\ \ell_1 \quad \ell_2 \quad \cdots \quad \ell_j \quad \cdots \quad \ell_{20}\ ], \tag{11}$$

where

$$\ell_j = \frac{1}{21} \times \sum\nolimits_{k=1}^{21} m_{k\to j}^{(1)} \quad (j = 1, 2, \cdots, 20). \tag{12}$$

Next, let us use the grey model approach to extract more useful information from Eq. 8 to define some additional components in Eq. 6. The grey model is particularly useful for solving complicated problems that lack sufficient information, or need to process uncertain information and to reduce random effects of acquired data (Deng 1989). In the grey system theory, an important and generally used model is called GM (1, 1) (Deng 1989). It is quite effective for monotonic series, with good simulating effect and small error, as reflected by the fact that using the GM (1,1) model has remarkably improved the success rates in predicting protein structural classes (Xiao et al. 2008). However, if the series concerned are not monotonic, the simulating effect of the GM (1,1) model would not be good and its error might be quite large. To overcome such a shortcoming, we use a different grey system model called GM (2,1) (Deng 1989), which can be effectively used to deal with the oscillation series.

The GM (2,1) model can be expressed by the following 2nd-order grey differential equation with three variables

$$x^{(1)}m_{k,j}^{(1)} + x_1^j m_{k,j}^{(1)} + x_2^j z^{(1)}(k) = b^j \quad (k = 2, 3, \cdots, 21; j = 1, 2, \cdots, 20), \tag{13}$$

where

$$x^{(1)}m_{k,j}^{(1)} = m_{k,j}^{(1)} - m_{k-1,j}^{(1)} \tag{14}$$

and

$$z^{(1)}(k) = \sum_{i=1}^{k-1} m_{i,j}^{(1)} + 0.5_{k,j}^{(1)}. \tag{15}$$

In Eq. 13, the x_1^j and x_2^j are the developing coefficients, and y^j the influence coefficient. Actually, x_1^j, x_2^j and y^j can be expressed as the components of a 3D vector given by

$$\begin{bmatrix} x_1^j \\ x_2^j \\ y^j \end{bmatrix} = (B_j^T B_j)^{-1} B_j^T U_j \quad (j = 1, 2, \cdots, 20), \tag{16}$$

where

$$B_j = \begin{bmatrix} -m_{2\to j}^{(1)} & -m_{1\to j}^{(1)} - 0.5m_{2\to j}^{(1)} & 1 \\ -m_{3\to j}^{(1)} & -\sum_{i=1}^{2} m_{i\to j}^{(1)} - 0.5m_{3\to j}^{(1)} & 1 \\ \vdots & \vdots & \vdots \\ -m_{L\to j}^{(1)} & -\sum_{i=1}^{L-1} m_{i\to j}^{(1)} - 0.5m_{L\to j}^{(1)} & 1 \end{bmatrix} \tag{17}$$

and

$$U_j = \begin{bmatrix} m^{(1)}_{2\to j} - m^{(1)}_{1\to j} \\ m^{(1)}_{3\to j} - m^{(1)}_{2\to j} \\ \vdots \\ m^{(1)}_{L\to j} - m^{(1)}_{L-1\to j} \end{bmatrix}. \quad (18)$$

Therefore, when using the grey-PSSM formulation, we can extract a length of $\Omega = 3\times 20 = 60$ quantities, given by

$$P_{GreyPSSM} = [\ \Psi_1 \quad \Psi_2 \quad \cdots \quad \Psi_{60}]^T, \quad (19)$$

where

$$\begin{cases} \Psi_{3j-2} = w_1 f_j x_1^j \\ \Psi_{3j-1} = w_2 f_j x_2^j \quad (j = 1, 2, \cdots, 20), \\ \Psi_{3j} \quad = w_3 f_j y^j \end{cases} \quad (20)$$

where $f_j (j = 1,2,\ldots,20)$ are the occurrence frequencies of the 20 different types of amino acids in the peptide sample concerned, w_1, w_2 and w_3 are the weight factors that will be determined by optimizing the performance of predictor. Finally, we obtained a total of 100 feature elements, of which 20 are from $\boldsymbol{P}_{\mathrm{AAC}}$, 20 from P_{PSSM} and 60 from P_{GreyPSSM}. Thus according to the general formulation of PseAAC in Eq. 6, a peptide sample can be formulated as a 100-D vector given by

$$P = [\ \Psi_1 \quad \Psi_2 \quad \cdots \quad \Psi_{100}\ ]^T, \quad (21)$$

where $\Psi_1,\Psi_2,\ldots,\Psi_{20}$ are the same as those in Eq. 7, $\Psi_{21},\Psi_{22},\ldots,\Psi_{40}$ in Eq. 11, $\Psi_{41},\Psi_{42},\ldots,\Psi_{100}$ in Eq. 19.

Operation Engine

The *K*-nearest neighbor algorithm (Wang et al. 2011) is one of the powerful methods for performing nonparametric classification. According to the *K-NN* rule, given a sample which was unknown the labels, its labels are assigned according to the labels of its K-nearest neighbors in the training sample.

Fuzzy *K-NN* classifier is a special variation of the *K-NN* classification family. Suppose $\{\mathbf{P}_1, \mathbf{P}_1, \cdots, \mathbf{P}_N\}$ is a set of vectors representing N proteins fragment in the training dataset which has been classified into two classes: $\{C_1, C_2\}$, where C_1represents the catalytic sites, and C_2 represents not-catalytic sites. Thus, for a query protein P, its fuzzy membership value for the i-th class is given by

$$\mu_i(p) = \frac{\sum_{j=1}^{K} \mu_i(P_j) d(P, P_j)^{-2/(m-1)}}{\sum_{j=1}^{K} d(P, P_j)^{-2/(m-1)}}, \quad (22)$$

where K is the number of the nearest neighbors counted, $\mu_i(P_j)$ is the fuzzy membership value of the peptide P_j to the i-th class, d(P,P_j) is the distance between the query peptide sample P and its j-th nearest peptide sample P_j in the training dataset; and m (>1) is the fuzzy coefficient for determine how heavily the distance is weighted when calculating each nearest neighbor's contribution to the membership value. There various metrics can be chosen as d(P,P_j), for example Hamming distance, Euclidean distance, and Mahalanobis distance (Chou 1995; Chou and Zhang 1995). In this paper, the Euclidean metric was used. The values of m and K will be mentioned later. After calculating all the memberships for a query peptide, it is assigned to the class with which it has the highest membership value, i.e., the predicted class for the query peptide P should be

$$\Lambda_u = \arg\max_i \{\mu_i(P)\}, \quad (23)$$

where $\boldsymbol{u}$ is the argument of i that maximizes $\mu_i(P)$.

To provide an intuitive picture, a flowchart is provided in Fig. 2 to illustrate the prediction process of iCataly-PseAAC.

Evaluation of Prediction Performance

In literature, the following four metrics are often used for examining the performance quality of a predictor (Chen et al. 2013; Xu et al. 2013)

$$\begin{cases} \mathrm{Sn} = 1 - \dfrac{N_-^+}{N^+}, & 0 \le Sn \le 1 \\ \mathrm{Sp} = 1 - \dfrac{N_+^-}{N^-}, & 0 \le Sp \le 1 \\ \mathrm{ACC} = 1 - \dfrac{N_-^+ + N_+^-}{N^+ + N^-}, & 0 \le ACC \le 1 \\ \mathrm{MCC} = \dfrac{1 - \left(\dfrac{N_-^+}{N^+} + \dfrac{N_+^-}{N^-}\right)}{\sqrt{\left(1 + \dfrac{N_+^- - N_-^+}{N^+}\right)\left(1 + \dfrac{N_-^+ - N_+^-}{N^-}\right)}}, & 0 \le MCC \le 1 \end{cases} \quad (24)$$

where N^+ is the total number of the catalytic sites peptides investigated, while N_-^+ the number of the peptides incorrectly predicted as the non-catalytic sites peptides; N^- the total number of the non-catalytic sites investigated, while N_+^- the number of the non-catalytic sites incorrectly predicted as the catalytic sites peptides; Sn, the sensitivity; Sp, the specificity; Acc, the accuracy; and MCC, the Mathew's correlation coefficient.

Now, it is crystal clear from Eq. 24 that when $N_-^+ = 0$ meaning none of the catalytic sites peptides were incorrectly predicted to be a non-catalytic sites peptide, we have the sensitivity Sn = 1. When $N_-^+ = N^+$ meaning that all the catalytic sites peptides were incorrectly predicted to be the non-catalytic sites peptide, we have the sensitivity Sn = 0. Likewise, when N_+^- meaning none of the non-catalytic sites peptides were incorrectly predicted as the catalytic sites peptide, we have the specificity Sp = 1; whereas $N_+^- = N^-$ meaning all the non-catalytic sites peptides were incorrectly predicted as the catalytic sites peptides, we have the specificity Sp = 0. When $N_-^+ = N_+^- = 0$ meaning that all the none of catalytic sites peptides in the positive dataset and none of the non-catalytic sites peptides in the negative dataset were incorrectly predicted, we have the overall accuracy ACC = 1 and MCC = 1; when $N_-^+ = N^+$ and $N_+^- = N^-$ meaning that all the catalytic sites peptides in the positive dataset and all the non-catalytic sites peptides in the negative dataset were incorrectly predicted, we have the overall accuracy ACC = 0 and MCC = −1; whereas when $N_-^+ = N^+/2$ and $N_+^- = N^-/2$ we have ACC = 0.5 and MCC = 0 meaning to better than random prediction. As we can see from the above discussion based on Eq. 24, the meanings of sensitivity, specificity, overall accuracy, and Mathew's correlation coefficient have become much more intuitive and easier to understand.

Web Server and User Guide

For the convenience of the vast majority of biological scientists, here let us provide a step-by-step guide to show how the users can easily get the desired result by means of the web server for iCataly-PseAAC without following the complicated mathematical equations presented in this paper for the process of developing the predictor and its integrity.

Step 1. Open the web server at the site http://www.jci-bioinfo.cn/iCataly-PseAAC and you will see the top page of the predictor on your computer screen, as shown in Fig. 3. Click on the *Read Me* button to see a brief introduction about iCataly-PseAAC predictor and the caveat when using it.

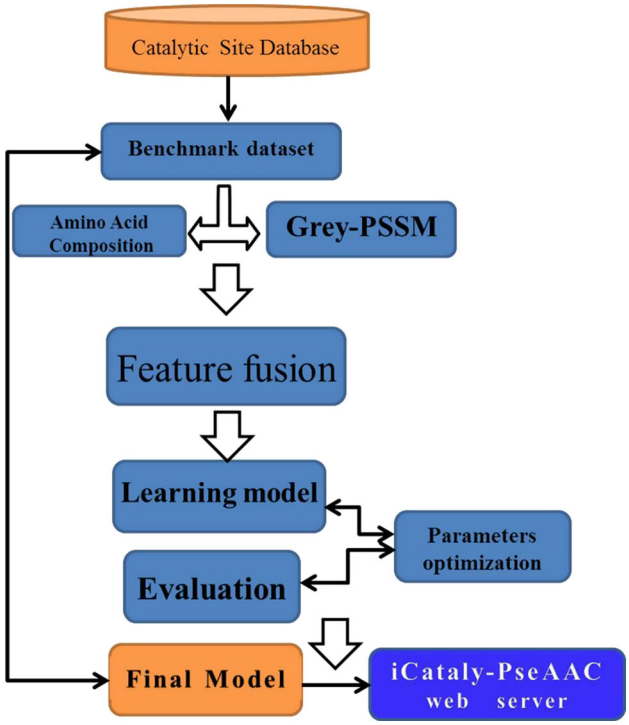

Fig. 2 A flowchart was provided to show the process of prediction

Step 2. Either type or copy/paste the query pairs into the input box at the center of Fig. 3. The input should be in FASTA format. Examples for the query pairs input and the corresponding output 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. For example, if you use the query amino acids sequences in the Example window as the input, you will see on your screen that the status of your job. When the job was done, the result will be displayed in the page.

Step 4. Click on the *Citation* button to find the relevant paper that documents the detailed development and algorithm of iCataly-PseAAC.

Step 5. Click on the *Data* button to download the benchmark dataset used to train and test the iCataly-PseAAC predictor.

Results and Discussion

The following three methods are often used in literatures: independent dataset test, subsampling (cross validation) test, and jackknife test (Chou 2005). However, as elucidated by a comprehensive review (Chou 2011), among

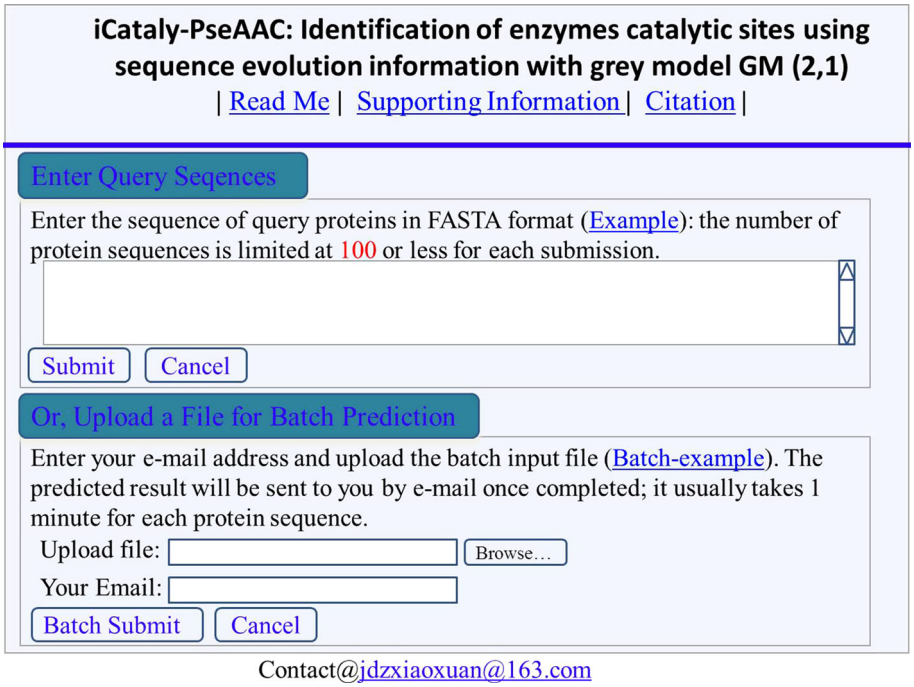

Fig. 3 A semiscreenshot to show the top page of iCataly-PseAAC

the three methods, the jackknife test was deemed the least arbitrary and most objective because it could always yield a unique result for a given benchmark dataset, and hence has been increasingly used and widely recognized by investigators to examine the accuracy of various predictor. Therefore, in this study, we also adopt the jackknife test to examine the predictor quality of the iCataly-PseAAC predictor.

The jackknife test rate achieved by iCataly-PseAAC for the catalytic sites system is given in Table 1, respectively. The value of m and K used in Eq. 22 was determined by optimizing the overall jackknife success rate through 2D research. The result was obtained in Fig. 4, from which we obtain when $m = 1.21$ and $k = 10$ the predictor reaches its optimized status.

To further demonstrate its power, we also used a Receiver Operating Characteristic (ROC) curve (Fawcett 2004; Davis and Goadrich 2006), which plots the true positive rate as function of the false positive rate for all possible thresholds. Furthermore, the overall performance of iCataly-PseAAC can also be quantified by the corresponding area under the ROC curve (AUC). Generally, the closer the AUC value is to 1, the better the performance is.

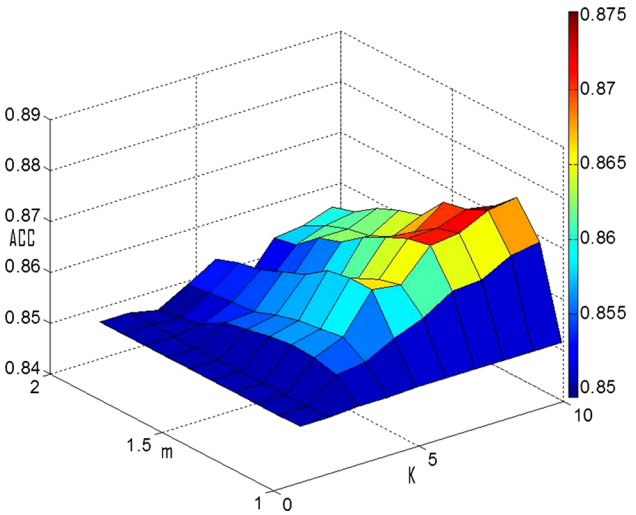

Fig. 4 The 3D graph to show the success rate by jackknife test were different value of m and K in the FKNN. The results were obtained for the independent testing prediction

As we can see from Table 1, the accuracy achieved by iCataly-PseAAC for the catalytic sites system was 87.52 % in training dataset. Meanwhile, we can see that the

Table 1 The jackknifing success rates were obtained in identifying the catalytic sites

Dataset	ACC[a] (%)	Sn (%)	Sp (%)	AUC (%)	MCC (%)
Training dataset	87.52	81.26	89.09	89.63	65.02
Testing dataset	83.07	75.00	85.00	82.45	52.26

Sn sensitivity, *Sp* specificity

[a] The parameters used: $m = 1.21$ and $k = 10$ in FKNN operation engine

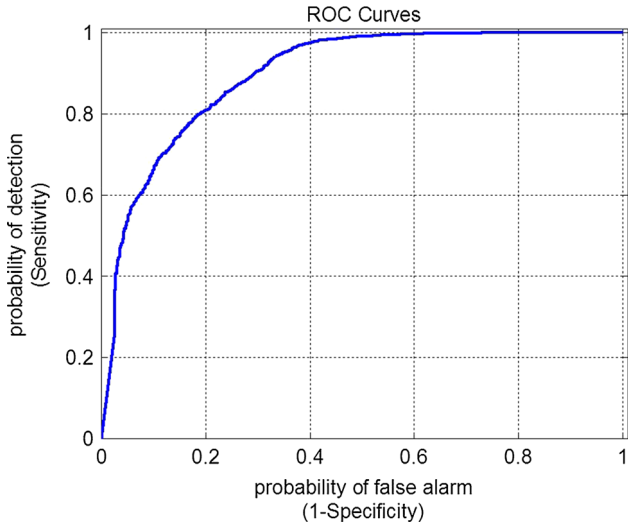

Fig. 5 The ROC curves of iCataly-PseAAC

Table 2 Comparison of iCataly-PseAAC with the existing predictor by the four dataset from (Petrova and Wu 2006; Youn et al. 2007)

Predictor	Acc (%)			
	PW79[c]	EF fold[c]	EF superfamily[c]	EF family[c]
CRpred[a]	83.84	85.21	84.58	86.60
EXIA[b]	85.54	85.68	85.71	83.57
iCataly-PseAAC	88.44	89.56	87.01	87.45

[a] From (Zhang et al. 2008)

[b] From (Chien and Huang 2012)

[c] From (Petrova and Wu 2006; Youn et al. 2007)

corresponding MCC (cf. Eq. 22) were 65.02 %, (Sn = 81.26 %, Sp = 89.09 %). Furthermore, the ROC curve of value of AUC was 89.63 %. The result is given in Fig. 5. For the testing dataset, the result of the accuracy is 83.07 % and the corresponding MCC (cf. Eq. 22) were 52.26 %, (Sn = 75.00 %, Sp = 85.00 %) and the ROC curve of value AUC was 82.45 %.

To further approve its power, let's compare iCataly-PseAAC with the existing predictor in this area. Predictor CRpred (Zhang et al. 2008) is based on sequence predictor, method EXIA (Chien and Huang 2012) is based on the residue side chain structure and sequence conservation method. Because our predictor is only one with a publicly accessible web server, the best way to compare them is though practical application. To realize this, let us construct a set range of multiple benchmark datasets. The datasets included the PW of 79 enzymes selected by (Petrova and Wu 2006) and three benchmark datasets with varying homology level including EF fold, EF family, and EF superfamily (Youn et al. 2007).

Listed in Table 2 are the prediction results of iCataly-PseAAC, CRpred, and EXIA on four datasets. As can be seen from Table 2, the accuracy achieved by iCataly-PseAAC was remarkably higher than those by its counterparts on the four independent datasets. These results have clearly indicated that iCataly-PseAAC is superior to its counterparts in predicting the catalytic site and our predictor can be used online for the convenience of the vast majority of biological scientists.

Conclusion

To acquire the information of the catalytic sites in enzymes is important for in-depth study of enzymes function and for developing a new drug. Our iCataly-PseAAC predictor may become a very useful high throughput tool in this regard. To promote the biological community, a web server of iCataly-PseAAC was constructed, which can be freely accessible at http://www.jci-bioinfo.cn/iCataly-PseAAC.

Acknowledgments This work was partially supported by the National Nature Science Foundation of China (Nos. 31260273, 61261027), Natural Science Foundation of Jiangxi Province, China (Nos. 20114BAB211013, 20122BAB211033, 20122BAB201044, 20122BAB201020), the Department of Education of JiangXi Province (GJJ12490), the LuoDi plan of the Department of Education of JiangXi Province(KJLD12083), and the JiangXi Provincial Foundation for Leaders of Disciplines in Science (20113BCB22008) and the Graduated innovation found of Jingdezhen ceramic institute (JYC1310, JYC201427).

Reference

Bartlett GJ, Porter CT, Borkakoti N, Thornton JM (2002) Analysis of catalytic residues in enzyme active sites. J Mol Biol 324:105–121

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN et al (2000) The Protein Data Bank. Nucleic Acids Res 28:235–242

Berman HM, Battistuz T, Bhat TN, Bluhm WF, Bourne PE et al (2002) The Protein Data Bank. Acta Crystallogr D Biol Crystallogr 58:899–907

Chea E, Livesay DR (2007) How accurate and statistically robust are catalytic site predictions based on closeness centrality? BMC Bioinform 8:153

Chen W, Feng P-M, Lin H, Chou K-C (2013) iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. Nucleic Acids Res 41:e68

Chien Y-T, Huang S-W (2012) Accurate prediction of protein catalytic residues by side chain orientation and residue contact density. PLoS One 7:e47951

Chien YT, Huang SW (2013) On the structural context and identification of enzyme catalytic residues. Biomed Res Int 2013:802945

Chou KC (1995) A novel approach to predicting protein structural classes in a (20–1)-D amino acid composition space. Proteins 21:319–344

Chou K-C (1996) Prediction of human immunodeficiency virus protease cleavage sites in proteins. Anal Biochem 233:1–14

Chou K-C (2001) Prediction of signal peptides using scaled window. Peptides 22:1973–1979

Chou KC (2005) Progress in protein structural class prediction and its impact to bioinformatics and proteomics. Curr Protein Pept Sci 6:423–436

Chou K-C (2011) Some remarks on protein attribute prediction and pseudo amino acid composition. J Theor Biol 273:236–247

Chou K-C, Zhang C-T (1994) Predicting protein folding types by distance functions that make allowances for amino acid interactions. J Biol Chem 269:22014–22020

Chou K-C, Zhang C-T (1995) Prediction of protein structural classes. Crit Rev Biochem Mol Biol 30:275–349

Chou K-C, Wu Z-C, Xiao X (2012) iLoc-Hum: using the accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites. Mol Biosyst 8:629–641

Davis J, Goadrich M (2006) The relationship between precision–recall and ROC curves. ACM, New York, pp 233–240

Deng J-L (1989) Introduction to grey system theory. J Grey Syst 1:1–24

Dou Y, Zheng X, Yang J, Wang J (2010) Prediction of catalytic residues based on an overlapping amino acid classification. Amino Acids 39:1353–1361

Dou Y, Geng X, Gao H, Yang J, Zheng X et al (2011) Sequence conservation in the prediction of catalytic sites. Protein J 30:229–239

Fawcett T (2004) ROC graphs: notes and practical considerations for researchers. Mach Learn 31:1–38

Fischer JD, Mayer CE, Soding J (2008) Prediction of protein functional residues from sequence by probability density estimation. Bioinformatics 24:613–620

Fu L, Niu B, Zhu Z, Wu S, Li W (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28:3150–3152

Gao YF, Li BQ, Cai YD, Feng KY, Li ZD et al (2013) Prediction of active sites of enzymes by maximum relevance minimum redundancy (mRMR) feature selection. Mol Biosyst 9:61–69

Gutteridge A, Bartlett GJ, Thornton JM (2003) Using a neural network and spatial clustering to predict the location of active sites in enzymes. J Mol Biol 330:719–734

Nakashima H, Nishikawa K, Tatsuo O (1986) The folding type of a protein is relevant to the amino acid composition. J Biochem 99:153–162

Ota M, Kinoshita K, Nishikawa K (2003) Prediction of catalytic residues in enzymes based on known tertiary structure, stability profile, and sequence conservation. J Mol Biol 327:1053–1064

Petrova NV, Wu CH (2006) Prediction of catalytic residues using Support Vector Machine with selected protein sequence and structural properties. BMC Bioinform 7:312

Porter CT, Bartlett GJ, Thornton JM (2004) The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. Nucleic Acids Res 32:D129–D133

Schaffer AA, Aravind L, Madden TL, Shavirin S, Spouge JL et al (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Res 29:2994–3005

Tong W, Williams RJ, Wei Y, Murga LF, Ko J et al (2008) Enhanced performance in prediction of protein active sites with THEMATICS and support vector machines. Protein Sci 17:333–341

Torrance JW, Bartlett GJ, Porter CT, Thornton JM (2005) Using a library of structural templates to recognise catalytic sites and explore their evolution in homologous families. J Mol Biol 347:565–581

UniProt C (2007) The universal protein resource (UniProt). Nucleic Acids Res 35:D193–D197

Wang P, Xiao X, Chou K-C (2011) NR-2L: a two-level predictor for identifying nuclear receptor subfamilies based on sequence-derived features. PLoS One 6:e23505

Xiao X, Wang P, Chou K-C (2008) Predicting protein structural classes with pseudo amino acid composition: an approach using geometric moments of cellular automaton image. J Theor Biol 254:691–696

Xiao X, Wang P, Chou KC (2011) Quat-2L: a web-server for predicting protein quaternary structural attributes. Mol Divers 15:149–155

Xiao X, Wang P, Chou K-C (2012) inr-physchem: A sequence-based predictor for identifying nuclear receptors and their subfamilies via physical-chemical property matrix. PLoS One 7:e30869

Xu Y, Ding J, Wu L-Y, Chou K-C (2013) iSNO-PseAAC: predict cysteine S-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition. PLoS One 8:e55844

Youn E, Peters B, Radivojac P, Mooney SD (2007) Evaluation of features for catalytic residue prediction in novel folds. Protein Sci 16:216–226

Zhang T, Zhang H, Chen K, Shen S, Ruan J et al (2008) Accurate sequence-based prediction of catalytic residues. Bioinformatics 24:2329–2338

Zvelebil MJ, Sternberg MJ (1988) Analysis and prediction of the location of catalytic residues in enzymes. Protein Eng 2:127–138