Available online at www.sciencedirect.com
ScienceDirect
Biochemical and Biophysical Research Communications 364 (2007) 53–59
BBRC
www.elsevier.com/locate/ybbrc

EzyPred: A top–down approach for predicting enzyme functional classes and subclasses

Hong-Bin Shen [1], Kuo-Chen Chou *

Gordon Life Science Institute, San Diego, CA 92130, USA
Institute of Image Processing & Pattern Recognition, Shanghai Jiaotong University, Shanghai 200030, China

Received 15 September 2007
Available online 2 October 2007

Abstract

Given a protein sequence, how can we identify whether it is an enzyme or non-enzyme? If it is, which main functional class it belongs to? What about its sub-functional class? It is important to address these problems because they are closely correlated with the biological function of an uncharacterized protein and its acting object and process. Particularly, with the avalanche of protein sequences generated in the Post Genomic Age and relatively much slower progress in determining their functions by experiments, it is highly desired to develop an automated method by which one can get a fast and accurate answer to these questions. Here, a top–down predictor, called **EzyPred**, is developed by fusing the results derived from the functional domain and evolution information. **EzyPred** is a 3-layer predictor: the 1st layer prediction engine is for identifying a query protein as enzyme or non-enzyme; the 2nd layer for the main functional class; and the 3rd layer for the sub-functional class. The overall success rates for all the three layers are higher than 90% that were obtained through rigorous cross-validation tests on the very stringent benchmark datasets in which none of the proteins has $\geqslant 40\%$ sequence identity to any other in a same class or subclass. **EzyPred** is freely accessible at http://chou.med.harvard.edu/bioinf/EzyPred/, by which one can get the desired 3-level results for a query protein sequence within less than 90 s.

Keywords: Enzyme functional class; Evolution; Functional domain; Fusion approach; **EzyPred** web-server

For a newly-found protein sequence the most interesting thing people wish to know is about its biological function, and hence the following questions are often asked: Is the query protein an enzyme or non-enzyme? If it is, which main functional class does it belong to? Or going further deeper, what about its sub-functional class? Although the answers to these questions can be found by conducting various biochemical experiments, it is both time-consuming and costly to do so solely by experimental approaches. With the explosion of newly-found protein sequences entering into databanks in the Post Genomic Age, it has become a major challenge to bridge the gap between the number of newly generated sequence entries and the number of functionally characterized protein entries. Actually, some efforts were made in this regard [1,2]. However, the investigation in [1] was limited within the scope of oxidoreductases while that in [2] limited among the main enzyme classes only. Particularly, no web-server was provided in either [1] or [2]. The present study was initiated in an attempt to develop a top–down approach to solve all these problems and make it accessible to the vast majority of experimental scientists by providing a user-friendly web-server.

Materials and methods

Materials

The ENZYME database at http://www.expasy.org/enzyme/ (released on 01-May-2007) was used to construct the benchmark datasets for the enzyme main functional classes and their subclasses.

* Corresponding author. Present address: Gordon Life Science Institute, San Diego, CA 92130, USA. Fax: +1 858 484 1018.
E-mail addresses: hbshen@crystal.harvard.edu (H.-B. Shen), kcchou@gordonlifesciene.org (K.-C. Chou).
[1] Present address: BCMP, Harvard Medical School, Boston, MA 02115, USA.

0006-291X/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.bbrc.2007.09.098

Main functional classes. According to their main functions or enzyme commission (EC) numbers [3], enzymes are classified into the following six main classes: (1) oxidoreductase (EC.1), (2) transferase (EC.2), (3) hydrolase (EC.3), (4) lyase (EC.4), (5) isomerase (EC.5), and (6) ligase (EC.6). To get the high quality benchmark dataset, the data were curated strictly according to the following procedures. *Step 1*: to avoid fragment data, those enzymes whose sequences were annotated with "fragment" or had less than 50 amino acids were excluded. *Step 2*: for the uniqueness, those enzymes that occur in two or more classes were excluded. *Step 3*: to reduce the homology bias, a redundancy cutoff was operated by an in-house program to winnow those sequences which have $\geqslant 40\%$ sequence identity to any other in a same functional class. Finally, 9832 enzyme sequences classified into six main functional classes were obtained. Meanwhile, to construct a non-enzyme benchmark dataset, 9850 non-enzyme protein sequences were randomly collected from Swiss-Prot at http://www.ebi.ac.uk/swissprot/ (version 52.0 released on 7-Mar-2007); these non-enzyme proteins were also subject to the same screening procedures to exclude the fragment and redundancy sequences. For the convenience of formulation, suppose the benchmark dataset thus obtained is denoted by $\mathbb{S}$, which consists of the enzyme dataset $\mathbb{S}^{\text{ezy}}$ and the non-enzyme dataset $\mathbb{S}^{\text{non-ezy}}$; i.e.,

$$\begin{cases} \mathbb{S} = \mathbb{S}^{\text{ezy}} \cup \mathbb{S}^{\text{non-ezy}} \\ \mathbb{S}^{\text{ezy}} = \mathbb{S}_1^{\text{ezy}} \cup \mathbb{S}_2^{\text{ezy}} \cup \mathbb{S}_3^{\text{ezy}} \cup \mathbb{S}_4^{\text{ezy}} \cup \mathbb{S}_5^{\text{ezy}} \cup \mathbb{S}_6^{\text{ezy}} \end{cases} \quad (1)$$

where $\cup$ is the symbol for union in the set theory, $\mathbb{S}_1^{\text{ezy}}$ is the EC.1 subset containing 1618 oxidoreductase sequences, $\mathbb{S}_2^{\text{ezy}}$ the EC.2 subset of 3450 transferase sequences, $\mathbb{S}_3^{\text{ezy}}$ the EC.3 subset of 2791 hydrolase sequences, $\mathbb{S}_4^{\text{ezy}}$ the EC.4 subset of 679 lyase sequences, $\mathbb{S}_5^{\text{ezy}}$ the EC.5 subset of 518 isomerase sequences, and $\mathbb{S}_6^{\text{ezy}}$ the EC.6 subset of 776 ligase sequences. The 9832 enzyme sequences classified into six subsets as well as the 9850 non-enzyme sequences are provided in Online Supporting Information A.

Sub functional classes. To reflect the enzyme functions at a deeper level, each of the aforementioned six main enzyme classes has been further classified into many sub-classes [3]. For instance, the oxidoreductase family of EC.1 has 18 sub-classes denoted by EC.1.1, EC.1.2, EC.1.3,...,EC.1.18 and their functions are briefly described in a highly condensed Fig. 1. To construct the subclass benchmark dataset for each of

Fig. 1. A schematic drawing to use tree branches to classify enzyme and non-enzyme as well as the six main functional classes of enzymes and their subclasses.

the six main enzyme families, the same procedures in the "*Main functional classes*" section were used. However, if the number of enzyme sequences thus obtained for a subclass was less than 10, the subclass and the sequences therein were left out because of lacking statistical significance. Similar to Eq. (1), the benchmark datasets thus obtained can be formulated as

$$\begin{cases} \mathbb{S}_1^{\text{ezy}} = \mathbb{S}_{1.1}^{\text{ezy}} \cup \mathbb{S}_{1.2}^{\text{ezy}} \cup \mathbb{S}_{1.3}^{\text{ezy}} \cup \mathbb{S}_{1.4}^{\text{ezy}} \cdots \cup \mathbb{S}_{1.18}^{\text{ezy}} \\ \mathbb{S}_2^{\text{ezy}} = \mathbb{S}_{2.1}^{\text{ezy}} \cup \mathbb{S}_{2.2}^{\text{ezy}} \cup \mathbb{S}_{2.3}^{\text{ezy}} \cup \mathbb{S}_{2.4}^{\text{ezy}} \cdots \cup \mathbb{S}_{2.8}^{\text{ezy}} \\ \mathbb{S}_3^{\text{ezy}} = \mathbb{S}_{3.1}^{\text{ezy}} \cup \mathbb{S}_{3.2}^{\text{ezy}} \cup \mathbb{S}_{3.4}^{\text{ezy}} \cup \mathbb{S}_{3.5}^{\text{ezy}} \cup \mathbb{S}_{3.6}^{\text{ezy}} \\ \mathbb{S}_4^{\text{ezy}} = \mathbb{S}_{4.1}^{\text{ezy}} \cup \mathbb{S}_{4.2}^{\text{ezy}} \cup \mathbb{S}_{4.3}^{\text{ezy}} \cup \mathbb{S}_{4.4}^{\text{ezy}} \cup \mathbb{S}_{4.6}^{\text{ezy}} \cup \mathbb{S}_{4.99}^{\text{ezy}} \\ \mathbb{S}_5^{\text{ezy}} = \mathbb{S}_{5.1}^{\text{ezy}} \cup \mathbb{S}_{5.2}^{\text{ezy}} \cup \mathbb{S}_{5.3}^{\text{ezy}} \cup \mathbb{S}_{5.4}^{\text{ezy}} \cup \mathbb{S}_{5.5}^{\text{ezy}} \cup \mathbb{S}_{5.99}^{\text{ezy}} \\ \mathbb{S}_6^{\text{ezy}} = \mathbb{S}_{6.1}^{\text{ezy}} \cup \mathbb{S}_{6.2}^{\text{ezy}} \cup \mathbb{S}_{6.3}^{\text{ezy}} \cup \mathbb{S}_{6.4}^{\text{ezy}} \cdots \cup \mathbb{S}_{6.6}^{\text{ezy}} \end{cases} \tag{2}$$

where $\mathbb{S}_{1.1}^{\text{ezy}}$ represents the EC.1.1 subset with the function acting on the CH–OH group of donors (Fig. 1), and so forth. Note that in Eq. (2) some subsets such as $\mathbb{S}_{3.3}^{\text{ezy}}$ and $\mathbb{S}_{4.5}^{\text{ezy}}$ are missing because the numbers of their sequences obtained through the above procedures were less than 10. All the sequences for each of the subsets in Eq. (2) are provided in Online Supporting Information B.

Method

To develop a top–down predictor, a novel technique was introduced by fusing the FunD (Functional Domain) approach and the Pse-PSSM (Pseudo Position-Specific Scoring Matrix) approach.

Functional domain (FunD) composition. Proteins often contain several modules or domains, each with a distinct evolutionary origin and function. Based on such a fact, several FunD databases were developed, such as SMART [4], COG [5], KOG [5], CDD [6]. Pfam database is a large collection of multiple sequence alignments and hidden Markov models currently covering 8958 common protein domains and families [7]. With each of the 8958 domain sequences as a vector-base, a given protein sample can be defined as an 8958-D (dimensional) vector according to the following procedures. *Step 1*: use RPS-BLAST (Reverse PSI-BLAST) program [8] to compare the protein sequence with each of the 8958 domain sequences in Pfam database. *Step 2*: if the significance threshold value (expect value) is $\leqslant 0.01$ for the ith profile in Pfam meaning a "hit" is found, then the ith component of the protein in the 8958-D space is assigned 1; otherwise, 0. *Step 3*: the protein sample $\mathbf{P}$ in the FunD space can thus be formulated as

$$\mathbf{P}_{\text{FunD}} = [\,\mathbb{D}_1 \quad \mathbb{D}_2 \quad \cdots \quad \mathbb{D}_i \quad \cdots \quad \mathbb{D}_{8958}\,]^{\mathbf{T}} \tag{3}$$

where $\mathbf{T}$ is the transpose operator, and

$$\mathbb{D}_i = \begin{cases} 1, & \text{when a hit is found for } \mathbf{P} \text{ in the } i\text{th profile of Pfam} \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

Pseudo position-specific scoring matrix (Pse-PSSM). To incorporate the evolution information of proteins, the PSSM (Position-Specific Scoring Matrix) [8] was used; i.e., according to the concept of PSSM, the sample of a protein $\mathbf{P}$ can be represented by:

$$\mathbf{P}_{\text{PSSM}} = \begin{bmatrix} \mathbb{V}_{1\to1} & \mathbb{V}_{1\to2} & \cdots & \mathbb{V}_{1\to20} \\ \mathbb{V}_{2\to1} & \mathbb{V}_{2\to2} & \cdots & \mathbb{V}_{2\to20} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbb{V}_{i\to1} & \mathbb{V}_{i\to2} & \cdots & \mathbb{V}_{i\to20} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbb{V}_{L\to1} & \mathbb{V}_{L\to2} & \cdots & \mathbb{V}_{L\to20} \end{bmatrix} \tag{5}$$

where $\mathbb{V}_{i\to j}$ represents the score of the amino acid residue in the ith position of the protein sequence being changed to amino acid type j during the evolution process. Here, the numerical codes 1, 2,...,20 are used to denote the 20 native amino acid types according to the alphabetical order of their single character codes. The $L \times 20$ scores in Eq. (5) were generated by using PSI-BLAST [8] to search the Swiss-Prot database (version 52.0 released on 6-March-2007) through three iterations with 0.001 as the E-value cutoff for multiple sequence alignment against the sequence of the protein $\mathbf{P}$, followed by a standardization procedure given below:

$$\mathbb{V}_{i\to j} = \frac{\mathbb{V}_{i\to j}^0 - \langle \mathbb{V}_i^0 \rangle}{\text{SD}(\mathbb{V}_i^0)} \quad (i = 1,2,\ldots,L;\ j = 1,2,\ldots,20) \tag{6}$$

where $\mathbb{V}_{i\to j}^0$ represent the original scores directly created by PSI-BLAST [8] that are generally shown as positive or negative integers; $\langle \mathbb{V}_i^0 \rangle$ the mean of $\mathbb{V}_{i\to j}^0$ over 20 native amino acids; $\text{SD}(\mathbb{V}_i^0)$ the standard deviation of $\mathbb{V}_{i\to j}^0$. The standardized scores will have a zero mean value over the 20 amino acids and will remain unchanged if going through the same conversion procedure again. The positive score means that the corresponding mutation occurs more frequently in the alignment than expected by chance, while the negative one means just the opposite. However, according to the PSSM descriptor (Eq. (5)), proteins with different lengths will correspond to row-different matrices. To make the PSSM descriptor become a size-uniform matrix, one possible approach is to represent a protein sample $\mathbf{P}$ by

$$\bar{\mathbf{P}}_{\text{PSSM}} = [\,\bar{\mathbb{V}}_1 \quad \bar{\mathbb{V}}_2 \quad \cdots \quad \bar{\mathbb{V}}_{20}\,]^{\mathbf{T}} \tag{7}$$

where

$$\bar{\mathbb{V}}_j = \frac{1}{L}\sum_{i=1}^{L} \mathbb{V}_{i\to j} \quad (j = 1,2,\ldots,20) \tag{8}$$

where $\bar{\mathbb{V}}_j$ represents the average score of the amino acid residues in the protein $\mathbf{P}$ being changed to amino acid type j during the evolution process. However, if $\bar{\mathbf{P}}_{\text{PSSM}}$ of Eq. (7) was used to represent the protein $\mathbf{P}$, all the sequence-order information during the evolution process would be lost. To avoid complete loss of the sequence-order information, the concept of the pseudo amino acid (PseAA) composition as originally proposed in [9,10] was adopted; i.e., instead of Eq. (7), let us use the pseudo position-specific scoring matrix (Pse-PSSM) as given by

$$\mathbf{P}_{\text{Pse-PSSM}}^{\xi} = [\,\bar{\mathbb{V}}_1 \quad \bar{\mathbb{V}}_2 \quad \cdots \quad \bar{\mathbb{V}}_{20} \quad \Phi_1^{\xi} \quad \Phi_2^{\xi} \quad \cdots \quad \Phi_{20}^{\xi}\,]^{\mathbf{T}} \tag{9}$$

to represent the protein $\mathbf{P}$, where

$$\Phi_j^{\xi} = \frac{1}{L-\xi}\sum_{i=1}^{L-\xi} [\mathbb{V}_{i\to j} - \mathbb{V}_{(i+\xi)\to j}]^2 \quad (j = 1,2,\ldots,20;\ \xi < L) \tag{10}$$

meaning that Φ_j^1 is the correlation factor by coupling the most contiguous PSSM scores along the protein chain for the amino acid type j; Φ_j^2 that by coupling the second-most contiguous PSSM scores; and so forth. Note that, as mentioned in the Material section, the length of the shortest protein sequence in the benchmark dataset is $L = 50$, and hence the value allowed for ξ in Eq. (10) must be smaller than 50. When $\xi = 0$, Φ_j^{ξ} becomes a naught element and Eq. (9) is degenerated to Eq. (7).

Optimized evidence-theoretic k nearest neighbor (OET-KNN) classifier. The OET-KNN classifier is a very powerful classification engine as demonstrated by its role in enhancing the success rates of predicting protein subcellular localization [11], where a detailed mathematical formulation for OET-KNN was also provided in the Appendix B. Here, we just give a brief description of how to use it to identify enzyme, its main-class and subclass. First of all, let us consider the top-level problem, i.e., to identify a protein as enzyme or non-enzyme with the benchmark dataset $\mathbb{S} = \mathbb{S}^{\text{ezy}} \cup \mathbb{S}^{\text{non-ezy}}$(Eq. (1)). Suppose the process in identifying the query protein $\mathbf{P}$ among the two classes by OET-KNN is formulated as

$$\textbf{OET-KNN} \triangleright \mathbf{P} = \begin{cases} \textbf{OET-KNN} \triangleright \mathbf{P}_{\text{FunD}} = \Lambda_1(K,i), & \text{for FunD frame} \\ \textbf{OET-KNN} \triangleright \mathbf{P}_{\text{Pse-PSSM}}^{\xi} = \Lambda_2(K,\xi,i), & \text{for Pse-PSSM} \end{cases} \tag{11}$$

where $\triangleright$ represents an action operator, $\Lambda_1(K,i)$ the creditability score for the query protein believed in the ith class when it is defined in the FunD frame (Eq. (3)), K is the parameter selected for the OET-KNN classifier [11], $\Lambda_2(K,\xi,i)$ the corresponding creditability score when the prediction is operated in the Pse-PSSM frame (Eq. (9)), and ξ the parameter selected for defining $\mathbf{P}_{\text{Pse-PSSM}}^{\xi}$ (Eqs. (9) and (10)). Accordingly, using different

descriptors to represent protein samples may lead to different results; even if the same descriptor is adopted, selecting different parameters may lead to different results as well. In order to get a unique result, the fusion approach is introduced as formulated below.

Fusion approach. The parameter K in Eq. (11) is the number of the nearest proteins counted against the query protein during the prediction process [12]. Generally speaking, for most training datasets, when $K > 10$ the success rate drops down remarkably and hence we can narrow the scope of K from 1 to 10. Also, the parameter ξ must be smaller than 50, the number of amino acids for the shortest protein sequence in the benchmark dataset. Therefore, the final predicted result should be determined by a fusion approach through the following voting mechanism. According to Eq. (11), the voting score for the query protein **P** belonging to the ith class is given by

$$\Pi_i = \sum_{K=1}^{10} w_K^1 \Lambda_1(K,i) + \sum_{K=1}^{10}\sum_{\xi=0}^{49} w_{K,\xi}^2 \Lambda_2(K,\xi,i), \quad (i=1,2) \tag{12}$$

where $i=1$ is for enzyme and $i=2$ for non-enzyme, w_K^1 and $w_{K,\xi}^2$ are the weight factors and were set at 1 for simplicity, thus the query protein **P** is predicted belonging to the class or subset for which the score of Eq. (12) is the highest; i.e.,

$$\mu = \arg \max_i \{\Pi_i\}, \quad (i=1,2) \tag{13}$$

where μ is the argument of i that maximize Π_i. If there is a tie, then the final predicted result will be randomly assigned to one of their corresponding subsets although this kind of tie case rarely happens and actually was not observed in the current study.

By changing ($i=1,2$) to ($i=1,2,\ldots,6$) and working on the benchmark dataset $\mathbb{S}^{ezy}$ (Eq. (1)), Eqs. (11)–(13) can be automatically used to solve the 2nd-level problem; by changing to ($i=1,2,\cdots,18$) and working on $\mathbb{S}_1^{ezy}$ (Eq. (1)), solve the 1st problem at the 3rd-level; and so forth. Such a procedure is the so-called top–down approach, and the entire predictor called **EzyPred**.

The above fusion approach not only can incorporate both the functional domain information and the protein evolution information but will also automatically solve the problem caused by the incompleteness of the FunD database. For example, if a query protein has no hit whatsoever when searching the Pfam database, it will correspond to a naught vector according to Eq. (3). The creditability score for a naught vector is zero by default (i.e., $\Lambda_1(K,i)=0$) according to Eq. (11), and the creditability score will be solely determined by $\Lambda_2(K,\xi,i)$ derived from the Pse-PSSM frame.

To provide an intuitive picture, a flowchart to show how to fuse the FunD approach and Pse-PSSM approach is given in Fig. 2A, and that to show the top–down approach process of the 3-layer predictor is given in Fig. 2B.

Table 1
Success rates by the jackknife test in identifying the enzyme proteins and non-enzyme proteins

Protein type	Number of proteins	Number of correct predictions	Success rate (%)
Enzyme	9832	9089	92.4
Non-enzyme	9850	8875	90.1
Overall	19,682	17,964	91.3

Table 2
Success rates by the jackknife test in identifying enzyme main functional classes

Enzyme main functional class	Number of proteins	Number of correct predictions	Success rate (%)
EC.1: Oxidoreductase	1618	1478	91.4
EC.2: Transferase	3450	3260	94.5
EC.3: Hydrolase	2791	2711	97.1
EC.4: Lyase	679	578	85.1
EC.5: Isomerase	518	433	83.6
EC.6: Ligase	776	749	96.5
Overall	9832	9209	93.7

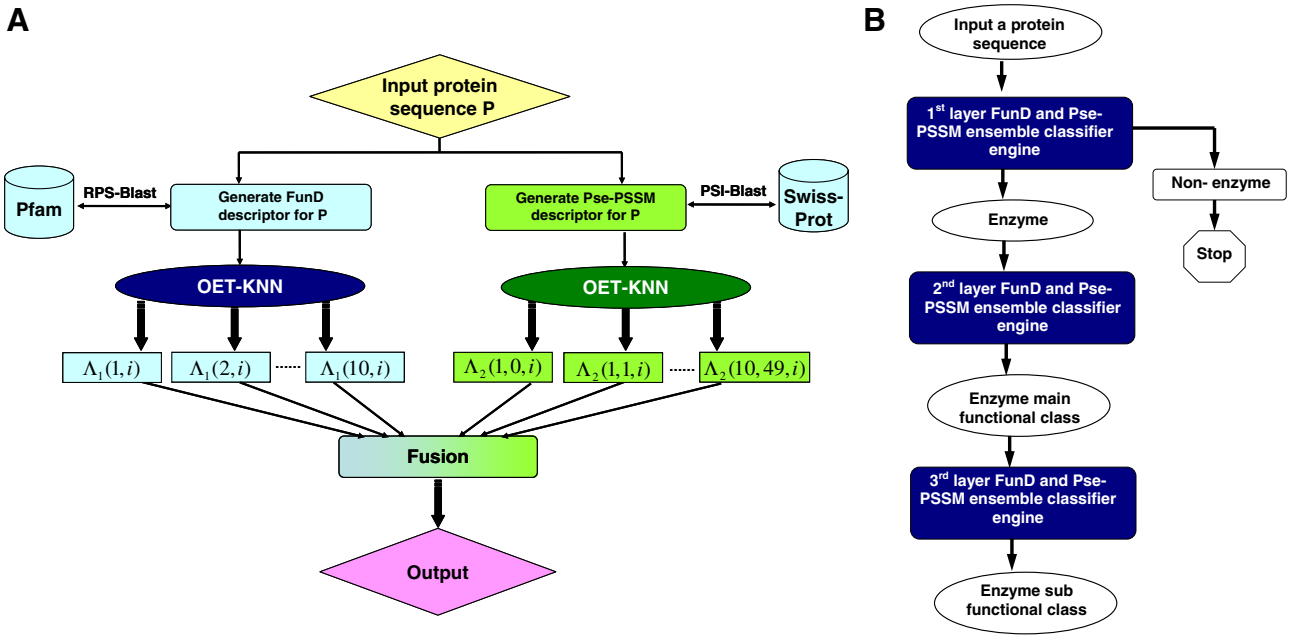

Fig. 2. A flowchart to show (A) how to fuse the FunD approach and Pse-PSSM approach into a prediction engine, and (B) how the top–down approach of the 3-layer predictor works.

Table 3
Success rates by the jackknife test in identifying sub-classes of the six main functional classes

	Number of proteins	Number of correct predictions	Success rate (%)
Subclass of oxidoreductases (*EC.1*)			
EC.1.1: Acting on the CH–OH group of donors	449	440	98.0
EC.1.2: Acting on the aldehyde or oxo group of donors	158	138	87.3
EC.1.3: Acting on the CH–CH group of donors.	149	101	67.8
EC.1.4: Acting on the CH–NH_2 group of donors	72	56	77.8
EC.1.5: Acting on the CH–NH group of donors	117	92	78.6
EC.1.6: Acting on NADH or NADPH	207	186	89.9
EC.1.7: Acting on other nitrogenous compounds as donors	35	17	48.6
EC.1.8: Acting on a sulfur group of donors	76	65	85.5
EC.1.9: Acting on a heme group of donors	69	66	95.7
EC.1.10: Acting on diphenols and related substances as donors	42	34	81.0
EC.1.11: Acting on a peroxide as acceptor	71	68	95.8
EC.1.12: Acting on hydrogen as donor	17	14	82.4
EC.1.13: Acting on single donors with incorporation of molecular oxygen	47	32	68.1
EC.1.14: Acting on paired donors, with incorporation or reduction of molecular oxygen	173	157	90.8
EC.1.15: Acting on superoxide as acceptor	25	23	92.0
EC.1.16: Oxidizing metal ions	20	14	70.0
EC.1.17: Acting on CH or CH_2 groups	67	58	86.6
EC.1.18: Acting on iron–sulfur proteins as donors	26	16	61.5
Overall	1820	1577	86.7
Subclass of transferases (*EC.2*)			
EC.2.1: Transferring one-carbon groups	529	512	96.8
EC.2.2: Transferring aldehyde or ketone residues	34	34	100
EC.2.3: Acyltransferases	324	294	90.7
EC.2.4: Glycosyltransferases	467	443	94.9
EC.2.5: Transferring alkyl or aryl groups (other than methyl groups)	277	267	96.4
EC.2.6: Transferring nitrogenous groups	114	112	98.3
EC.2.7: Transferring phosphorous-containing groups	1039	1007	96.9
EC.2.8: Transferring sulfur-containing groups	63	57	90.5
Overall	2847	2726	95.8
Subclass of hydrolases (*EC.3*)			
EC.3.1: Acting on ester bonds	1228	1214	98.9
EC.3.2: Glycosylases	464	446	96.1
EC.3.4: Acting on peptide bonds (peptide hydrolases)	486	446	91.8
EC.3.5: Acting on carbon–nitrogen bonds other than peptide bonds	436	408	93.6
EC.3.6: Acting on acid anhydrides	665	632	95.0
Overall	3279	3146	95.9
Subclass of lyases (*EC.4*)			
EC.4.1: Carbon–carbon lyases	340	329	96.8
EC.4.2: Carbon–oxygen lyases	365	350	95.9
EC.4.3: Carbon–nitrogen lyases	62	50	80.7
EC.4.4: Carbon–sulfur lyases	31	23	74.2
EC.4.6: Phosphorus–oxygen lyases	56	56	100
EC.4.99: Other lyases	38	34	89.5
Overall	892	842	94.4
Subclass of isomerases (*EC.5*)			
EC.5.1: Racemases and epimerases	111	102	91.9
EC.5.2: *cis-trans*-Isomerases	110	109	99.1
EC.5.3: Intramolecular oxidoreductases	207	186	89.9
EC.5.4: Intramolecular transferases (mutases)	139	133	95.7
EC.5.5: Intramolecular lyases	11	5	45.5
EC.5.99: Other isomerases	61	61	100
Overall	639	596	93.3
Subclass of ligases (*EC.6*)			
EC.6.1: Forming carbon–oxygen bonds	496	493	99.4
EC.6.2: Forming carbon–sulfur bonds	36	34	94.4

(*continued on next page*)

Table 3 (*continued*)

	Number of proteins	Number of correct predictions	Success rate (%)
EC.6.3: Forming carbon–nitrogen bonds	364	358	98.4
EC.6.4: Forming carbon–carbon bonds	13	11	84.6
EC.6.5: Forming phosphoric ester bonds	46	44	95.7
EC.6.6: Forming nitrogen–metal bonds	10	9	90.0
Overall	965	949	98.3

Results and discussion

In statistical prediction the independent dataset test, sub-sampling test, and jackknife test are often used in literatures for examining the accuracy of a predictor. Among these three, the jackknife test is deemed the most rigorous and objective [13], and hence has been increasingly adopted by investigators in examining the quality of various prediction methods (see, e.g., [14–32] as well as a recent review [33] in this regard).

The jackknife cross-validation results by **EzyPred** on the datasets $\mathbb{S}$ and $\mathbb{S}^{ezy}$ (cf. Eq. (1) and Online Supporting Information A) are given in Tables 1 and 2, respectively, from which we can see that the overall success rate in identifying the proteins as enzymes or non-enzymes is 91.3%, and that the overall success rate in identifying the enzymes among their six main functional classes is 93.7%. The corresponding results by **EzyPred** on the datasets $\mathbb{S}_1^{ezy}$, $\mathbb{S}_2^{ezy}$, $\mathbb{S}_3^{ezy}$, $\mathbb{S}_4^{ezy}$, $\mathbb{S}_5^{ezy}$, and $\mathbb{S}_6^{ezy}$ (cf. Eq. (2) and Online Supporting Information B) are given in Table 3, from which we can see that the overall success rates in identifying the subfamily classes of oxidoreductase, transferases, hydrolases, lyases, isomerases, and ligases are 86.7%, 95.8%, 95.9%, 94.4%, 93.3%, and 98.3%, respectively.

It was reported [34] that even for the pair fragments with >50% sequence identity the probability of having a same EC number (enzymatic function) is <30%, meaning that enzyme function is much less conserved than anticipated. However, for the current datasets in which none of enzymes has $\geqslant$40% sequence identity to any others in a same subset, the overall success rates by the **EzyPred** in identifying the main functional classes of enzymes and their subclasses are very high. As is well known, the more the number of classes to be identified, the less the success rate will be. However, even for the oxidoreductase dataset $\mathbb{S}_1^{ezy}$ consisting of 18 subfamily classes, the overall success rate obtained by the **EzyPred** is above 86%, indicating that **EzyPred** is a very powerful predictor in identifying enzymes, their main classes, and their subclasses.

Conclusion

The reason why **EzyPred** predictor can yield so high success rates is because it operates by fusing the FunD approach and Pse-PSSM approach. The former is closely related to the functions of proteins, while the latter can incorporate their evolution information. It is anticipated that with more data available in the ENZYME database, the current top–down **EzyPred** predictor can be extended to cover sub-subclass and sub-sub-subclass of enzymes as well. **EzyPred** is available to the public at the site http://chou.med.harvard.edu/bioinf/EzyPred/.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.bbrc.2007.09.098.

References

[1] K.C. Chou, D.W. Elrod, Prediction of enzyme family classes, Journal of Proteome Research 2 (2003) 183–190.
[2] K.C. Chou, Y.D. Cai, Predicting enzyme family class in a hybridization space, Protein Science 13 (2004) 2857–2863.
[3] A. Bairoch, The ENZYME Database in 2000, Nucleic Acids Research 28 (2000) 304–305.
[4] I. Letunic, R.R. Copley, B. Pils, S. Pinkert, J. Schultz, P. Bork, SMART 5: domains in the context of genomes and networks, Nucleic Acids Research 34 (2006) D257–D260.
[5] R.L. Tatusov, N.D. Fedorova, J.D. Jackson, A.R. Jacobs, B. Kiryutin, E.V. Koonin, D.M. Krylov, R. Mazumder, S.L. Mekhedov, A.N. Nikolskaya, B.S. Rao, S. Smirnov, A.V. Sverdlov, S. Vasudevan, Y.I. Wolf, J.J. Yin, D.A. Natale, The COG database: an updated version includes eukaryotes, BMC Bioinformatics 4 (2003) 41.
[6] A. Marchler-Bauer, J.B. Anderson, M.K. Derbyshire, C. DeWeese-Scott, N.R. Gonzales, M. Gwadz, L. Hao, S. He, D.I. Hurwitz, J.D. Jackson, Z. Ke, D. Krylov, C.J. Lanczycki, C.A. Liebert, C. Liu, F. Lu, S. Lu, G.H. Marchler, M. Mullokandov, J.S. Song, N. Thanki, R.A. Yamashita, J.J. Yin, D. Zhang, S.H. Bryant, CDD: a conserved domain database for interactive domain family analysis, Nucleic Acids Research 35 (2007) D237–D240.
[7] R.D. Finn, J. Mistry, B. Schuster-Bockler, S. Griffiths-Jones, V. Hollich, T. Lassmann, S. Moxon, M. Marshall, A. Khanna, R. Durbin, S.R. Eddy, E.L. Sonnhammer, A. Bateman, Pfam: clans, web tools and services, Nucleic Acids Research 34 (2006) D247–D251.
[8] A.A. Schaffer, L. Aravind, T.L. Madden, S. Shavirin, J.L. Spouge, Y.I. Wolf, E.V. Koonin, S.F. Altschul, Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements, Nucleic Acids Research 29 (2001) 2994–3005.
[9] K.C. Chou, Prediction of protein cellular attributes using pseudo amino acid composition, Proteins: Structure, Function, and Genetics 43 (2001) 246–255 (Erratum: Proteins: Structure, Function, and Genetics 44 (2001) 60).
[10] K.C. Chou, Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes, Bioinformatics 21 (2005) 10–19.
[11] K.C. Chou, H.B. Shen, Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-nearest neigh-

bor classifiers, Journal of Proteome Research 5 (2006) 1888–1897.
[12] K.C. Chou, H.B. Shen, Hum-PLoc: a novel ensemble classifier for predicting human protein subcellular localization, Biochemical and Biophysical Research Communication 347 (2006) 150–157.
[13] K.C. Chou, C.T. Zhang, Review: prediction of protein structural classes, Critical Reviews in Biochemistry and Molecular Biology 30 (1995) 275–349.
[14] G.P. Zhou, An intriguing controversy over protein structural class prediction, Journal of Protein Chemistry 17 (1998) 729–738.
[15] K.C. Chou, Y.D. Cai, Using functional domain composition and support vector machines for prediction of protein subcellular location, Journal of Biological Chemistry 277 (2002) 45765–45769.
[16] Y. Huang, Y. Li, Prediction of protein subcellular locations using fuzzy k-NN method, Bioinformatics 20 (2004) 21–28.
[17] C. Chen, X. Zhou, Y. Tian, X. Zou, P. Cai, Predicting protein structural class with pseudo-amino acid composition and support vector machine fusion network, Analytical Biochemistry 357 (2006) 116–121.
[18] C. Chen, Y.X. Tian, X.Y. Zou, P.X. Cai, J.Y. Mo, Using pseudo-amino acid composition and support vector machine to predict protein structural class, Journal of Theoretical Biology 243 (2006) 444–448.
[19] P. Du, Y. Li, Prediction of protein submitochondria locations by hybridizing pseudo-amino acid composition with various physicochemical features of segmented sequence, BMC Bioinformatics 7 (2006) 518.
[20] K.C. Chou, H.B. Shen, MemType-2L: a Web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM, Biochemical and Biophysical Research Communication 360 (2007) 339–345.
[21] S. Mondal, R. Bhavna, R. Mohan Babu, S. Ramakumar, Pseudo amino acid composition and multi-class support vector machines approach for conotoxin superfamily classification, Journal of Theoretical Biology 243 (2006) 252–260.
[22] H.B. Shen, K.C. Chou, Virus-PLoc: a fusion classifier for predicting the subcellular localization of viral proteins within host and virus-infected cells, Biopolymers 85 (2007) 233–240.
[23] H. Lin, Q.Z. Li, Predicting conotoxin superfamily and family by using pseudo amino acid composition and modified Mahalanobis discriminant, Biochemical and Biophysical Research Communication 354 (2007) 548–551.
[24] H. Lin, Q.Z. Li, Using pseudo amino acid composition to predict protein structural class: approached by incorporating 400 dipeptide components, Journal of Computational Chemistry 28 (2007) 1463–1466.
[25] H.B. Shen, K.C. Chou, Hum-mPLoc: an ensemble classifier for large-scale human protein subcellular location prediction by incorporating samples with multiple sites, Biochemical and Biophysical Research Communication 355 (2007) 1006–1011.
[26] Y. Cao, S. Liu, L. Zhang, J. Qin, J. Wang, K. Tang, Prediction of protein structural class with Rough Sets, BMC Bioinformatics 7 (20) (2006).
[27] Q.B. Gao, Z.Z. Wang, Classification of G-protein coupled receptors at four levels, Protein Engineering, Design and Selection 19 (2006) 511–516.
[28] K.D. Kedarisetti, L.A. Kurgan, S. Dick, Classifier ensembles for protein structural class prediction with varying homology, Biochemical and Biophysical Research Communication 348 (2006) 981–988.
[29] K.C. Chou, H.B. Shen, Euk-mPLoc: a fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites, Journal of Proteome Research 6 (2007) 1728–1734.
[30] S. Jahandideh, P. Abdolmaleki, M. Jahandideh, E.B. Asadabadi, Novel two-stage hybrid neural discriminant model for predicting proteins structural classes, Biophysical Chemistry 128 (2007) 87–93.
[31] K.C. Chou, H.B. Shen, Signal-CF: a subsite-coupled and window-fusing approach for predicting signal peptides, Biochemical and Biophysical Research Communication 357 (2007) 633–640.
[32] X.B. Zhou, C. Chen, Z.C. Li, X.Y. Zou, Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes, Journal of Theoretical Biology 248 (2007) 546–551.
[33] K.C. Chou, H.B. Shen, Review: recent progresses in protein subcellular location prediction, Analytical Biochemistry 370 (2007) 1–16.
[34] B. Rost, Enzyme function less conserved than anticipated, Journal of Molecular Biology 318 (2002) 595–608.