A Novel Approach to Predict Active Sites of Enzyme Molecules

Kuo-Chen Chou[1,2*] **and Yu-dong Cai**[3]
[1]*Gordon Life Science Institute, San Diego, California*
[2]*Tianjin Institute of Bioinformatics and Drug Discovery (TIBDD), Tianjin, China*
[3]*Biomolecular Sciences Department, UMIST, Manchester, United Kingdom*

ABSTRACT **Enzymes are critical in many cellular signaling cascades. With many enzyme structures being solved, there is an increasing need to develop an automated method for identifying their active sites. However, given the atomic coordinates of an enzyme molecule, how can we predict its active site? This is a vitally important problem because the core of an enzyme molecule is its active site from the viewpoints of both pure scientific research and industrial application. In this article, a topological entity was introduced to characterize the enzymatic active site. Based on such a concept, the covariant discriminant algorithm was formulated for identifying the active site. As a paradigm, the serine hydrolase family was demonstrated. The overall success rate by jackknife test for a data set of 88 enzyme molecules was 99.92%, and that for a data set of 50 independent enzyme molecules was 99.91%. Meanwhile, it was shown through an example that the prediction algorithm can also be used to find any typographic error of a PDB file in annotating the constituent amino acids of catalytic triad and to suggest a possible correction. The very high success rates are due to the introduction of a covariance matrix in the prediction algorithm that makes allowance for taking into account the coupling effects among the key constituent atoms of active site. It is anticipated that the novel approach is quite promising and may become a useful high throughput tool in enzymology, proteomics, and structural bioinformatics. Proteins 2004;55:77–82.**

Key words: catalytic triad; serine hydrolase; topological entity; covariance matrix; Mahalanobis distance; discriminant function; structural bioinformatics

INTRODUCTION

The importance of enzymes is well known; particularly, many recent evidences indicate that enzymes are critical in cellular signaling cascades (e.g., see Refs.1 and 2). The core of an enzyme molecule is its active site. To reveal the structural and functional mechanism of an enzyme molecule, we need to know its active site; to conduct structure-based drug design by targeting an enzyme molecule, we need to know its active site as well. Therefore, after the three-dimensional (3D) structure has been determined, the next most important task is to identify the key components of its active site.[3] Because the number of protein 3D structures entering into databanks has been rapidly increasing, it is both time-consuming and costly to approach the problem entirely by conducting various experiments. In view of this, it is highly desirable to develop a fast and automated method to complement the purely experimental approaches in this regard.

To detect the active site of a protein whose 3D structure is available, a straightforward approach is to search the structural analogs of the known active sites. Currently, there are many protein structure comparison approaches based on the geometric hashing algorithm[4] or the graph-theoretic algorithm[5] that can match the user-defined structure against a whole protein structure.[6,7] With the enzyme active-site templates provided by PROCAT database[8] (http://www.biochem.ucl.ac.uk/bsm/PROCAT/PROCAT.html), one can use these approaches to compare the available active-site templates with a target protein to deduce its possible active site. However, the speed of structural comparison process is usually very slow. The present study was initiated in an attempt to develop a different automated approach that can be used for fast and accurate identification of active sites.

MATERIALS AND METHODS

To make the new method simple and easy to understand, we use the serine hydrolase family as a paradigm. The biophysical conception and mathematical formulation introduced here can be easily extended to cover the entire family of enzymes. The active sites of the serine hydrolase family are characterized with a catalytic triad formed by Ser, His, and Asp, whose spatial position and orientation are closely related to the formation of a genuine active site. Particularly those atoms that have a potential to form a hydrogen bond with substrates play a key role in this

The Supplementary Materials Referred to in this article can be found at http://www.interscience.wiley.com/jpages/0887-3585/suppmat/index.html

*Correspondence to: Kuo-Chen Chou, Gordon Life Science Institute, 13784 Torrey Del Mar Drive, San Diego, CA 92130. E-mail: lifescience@san.rr.com

Received 28 June 2003; Accepted 22 August 2003

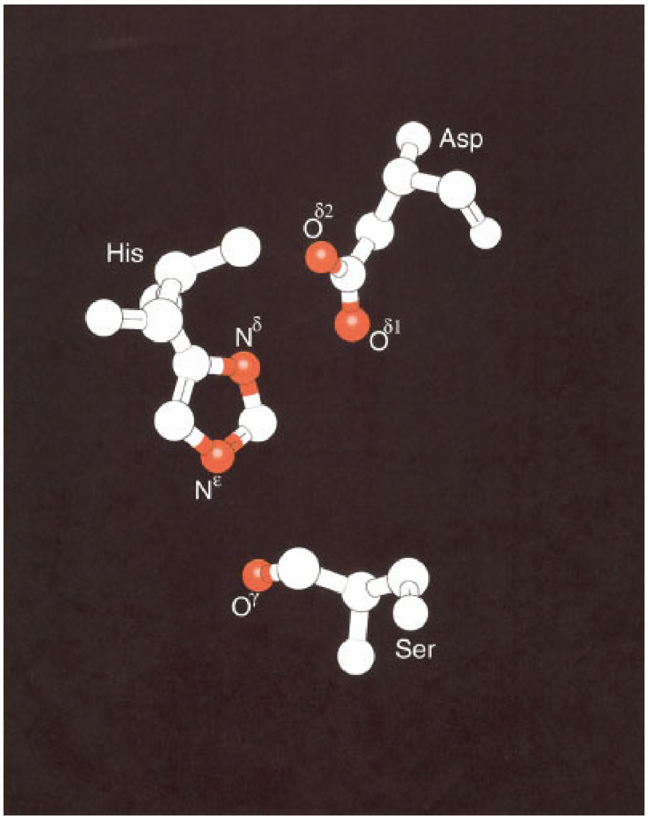

Fig. 1. The active sites (catalytic triad) of the serine hydrolase family are characterized by His-Asp-Ser, particularly the following five atoms (in red): N^{δ} and N^{ϵ} of His, $O^{\delta 1}$ and $O^{\delta 2}$ of Asp, as well as O^{γ} of Ser.

regard. Thus, the catalytic triad of hydrolase can be characterized by the distances among the following five atoms: N^{δ} and N^{ϵ} of His, $O^{\delta 1}$, and $O^{\delta 2}$ of Asp, as well as O^{γ} of Ser (Fig.1). There are totally $C_5^2 = 5!/[(5 - 2)!2!] = 10$ distances; however, the distance between His-N^{δ} and His-N^{ϵ} and the distance between Asp-$O^{\delta 1}$ and Asp-$O^{\delta 2}$ need not to be considered because they are within a ring and a rotator structure, respectively. Accordingly, the catalytic triad is actually characterized by eight pairwise distances between 1) His-N^{δ} and Asp-$O^{\delta 1}$, 2) His-N^{δ} and Asp-$O^{\delta 2}$, 3) His-N^{δ} and Ser-O^{γ}, 4) His-N^{ϵ} and Asp-$O^{\delta 1}$, 5) His-N^{ϵ} and Asp-$O^{\delta 2}$, 6) His-N^{ϵ} and Ser-O^{γ}, 7) Asp-$O^{\delta 1}$ and Ser-O^{γ}, and 8) Asp-$O^{\delta 2}$ and Ser-O^{γ} (Fig. 1). A set of such distances forms a topological entity, which can be represented by an 8D vector, as formulated below:

$$\mathbf{A} = \begin{bmatrix} d_1 \\ d_2 \\ d_3 \\ d_4 \\ d_5 \\ d_6 \\ d_7 \\ d_8 \end{bmatrix}, \tag{1}$$

where d_1 represents the distance between His-N^{δ} and Asp-$O^{\delta 1}$, d_2 between His-N^{δ} and Asp-$O^{\delta 2}$, d_3 between His-N^{δ} and Ser-O^{γ}, d_4 between His-N^{ϵ} and Asp-$O^{\delta 1}$, d_5 between His-N^{ϵ} and Asp-$O^{\delta 2}$, d_6 between His-N^{ϵ} and Ser-O^{γ}, d_7 between Asp-$O^{\delta 1}$ and Ser-O^{γ}, and d_8 between Asp-$O^{\delta 2}$ and Ser-O^{γ} (Fig. 1).

Suppose there are N topological entities derived according to Eq. 1 by searching the atomic coordinates extracted from the PDB files of the serine hydrolase family. Of the **N** entities, $n_1$entities correspond to a real catalytic triad and are assigned to the active subset S^1, and n_2 to a fake catalytic triad and assigned to the non-active subset S^2. The topological entity in the former is denoted as $\mathbf{A}^1$, whereas that in the latter is denoted as $\mathbf{A}^2$. Suppose the kth topological entity in the subset S^1 or S^2 is represented by

$$\mathbf{A}_k^{\mu} = \begin{bmatrix} d_{k,1}^{\mu} \\ d_{k,2}^{\mu} \\ d_{k,3}^{\mu} \\ d_{k,4}^{\mu} \\ d_{k,5}^{\mu} \\ d_{k,6}^{\mu} \\ d_{k,7}^{\mu} \\ d_{k,8}^{\mu} \end{bmatrix}, \quad (\mu = 1 \text{ or } 2) \tag{2}$$

where $d_{k,\tau}^{\mu}$ $(\tau = 1, 2, \ldots, 8)$ has the same meaning as d_{τ} of Eq. 1 but is associated with $\mathbf{A}_k^{\mu}$ instead of $\mathbf{A}$. The standard vector of the subset S^{μ} is defined by

$$\bar{\mathbf{A}}_k^{\mu} = \begin{bmatrix} \bar{d}_{k,1}^{\mu} \\ \bar{d}_{k,2}^{\mu} \\ \bar{d}_{k,3}^{\mu} \\ \bar{d}_{k,4}^{\mu} \\ \bar{d}_{k,5}^{\mu} \\ \bar{d}_{k,6}^{\mu} \\ \bar{d}_{k,7}^{\mu} \\ \bar{d}_{k,8}^{\mu} \end{bmatrix}, \quad (\mu = 1 \text{ or } 2) \tag{3}$$

where

$$\bar{d}_{\tau}^{\mu} = \frac{1}{n_{\mu}} \sum_{k=1}^{n_{\mu}} d_{k,\tau}^{\mu}, \quad (\tau = 1, 2, \ldots, 8;\ \mu = 1 \text{ or } 2), \tag{4}$$

in which n_{μ} is the number of the total topological entities in subset S^{μ} $(\mu = 1, 2)$.

Suppose $\mathbf{A}$ is a topological entity derived from a protein according to Eq. 1. We determine below whether it forms a real catalytic triad or a fake one. According to the similarity principle, if $\mathbf{A}$ has a higher similarity to $\bar{\mathbf{A}}^1$ than $\bar{\mathbf{A}}^2$, then the topological entity is predicted as a catalytic triad; otherwise, not. Now the problem is how to effectively define the similarity between the query topological entity $\mathbf{A}$ and the standard vectors $\bar{\mathbf{A}}^{\mu}$ $(\mu = 1$ and $2)$. Algorithms with various criteria were proposed, as follows.

Least Hamming Distance Algorithm[9]

The algorithm was originally proposed by P.Y.Chou[9] for predicting the protein structural class. The hypothesis was that the similarity of any two proteins could be measured by their Hamming distance or city-block metric[10] defined

TABLE I. List of the PDB Codes for the 88 Enzyme Molecules in the Training Data Set

1A10	1A46	1A5G	1A61	1AB9	1AD8	1AF4	1AHT	1AUT	1AY6
1B5G	1BA8	1BCU	1BH6	1BMA	1BRA	1BTU	1C4U	1CA8	1CSE
1DIT	1DWB	1E34	1ELA	1ELC	1ESA	1ESB	1FAX	1FPC	1GCD
1GMH	1H4W	1HAG	1HB0	1HCG	1HGT	1HJA	1HPG	1HUT	1HXE
1IHS	1IHT	1K22	1MEE	1NRN	1PEK	1PTK	1QGF	1QIX	1QJ1
1QL9	1S02	1SBC	1SCA	1SGP	1SIB	1SUE	1TBZ	1TGN	1THM
1TMB	1TOM	1TRY	1YJA	1YYY	2EST	2GCH	2GMT	2HAT	2HGT
2HNT	2HPP	2SEC	2SGP	2SIC	2SNI	2TGD	3GCT	3HAT	3SIC
3TGI	4HTC	5GDS	5SIC	6EST	7KME	8GCH	8KME		

in an amino acid-composition space. Now, based on the current 8D space introduced for characterizing the catalytic triad (cf. Eqs.1 and 3), the Hamming distance between $\mathbf{A}$ and $\bar{\mathbf{A}}^{\mu}$ should be formulated as

$$D_{\mathrm{H}}(\mathbf{A}, \bar{\mathbf{A}}^{\mu}) = \sum_{\tau=1}^{8} |d_{\tau} - d_{\tau}^{\mu}|, \quad (\mu = 1, 2) \tag{5}$$

Thus, the prediction rule was given by

$$D_{\mathrm{H}}(\mathbf{A}, \bar{\mathbf{A}}^{\mu}) = \mathbf{Min}\{D_{\mathrm{H}}(\mathbf{A}, \bar{\mathbf{A}}^{1}), D_{\mathrm{H}}(\mathbf{A}, \bar{\mathbf{A}}^{2})\}, \tag{6}$$

where μ can be 1 or 2, and the operator **Min** means taking the least one among those in the brackets, and the superscript μ is the subset predicted for the query topological entity $\mathbf{A}$ to belong to. If there is a tie, μ is not uniquely determined, but cases like that rarely occur.

Least Euclidean Distance Algorithm

Rather than Hamming distance, Nakashima et al.[11] used the square Euclidean distance as a scale to measure the similarity for predictiing protein structural class. Thus, instead of Eqs. 5 and 6, the similarity between A and $\bar{\mathbf{A}}^{\mu}$ should be defined by

$$D_{\mathrm{E}}^{2}(\mathbf{A}, \bar{\mathbf{A}}^{\mu}) = \sum_{\tau=1}^{8} (d_{\tau} - d_{\tau}^{\mu})^{2}, \quad (\mu = 1, 2) \tag{7}$$

and the prediction rule given by

$$D_{\mathrm{E}}^{2}(\mathbf{A}, \bar{\mathbf{A}}^{\mu}) = \mathbf{Min}\{D_{\mathrm{E}}^{2}(\mathbf{A}, \bar{\mathbf{A}}^{1}), D_{\mathrm{E}}^{2}(\mathbf{A}, \bar{\mathbf{A}}^{2})\}. \tag{8}$$

Covariant Discriminant Algorithm

It is instructive to point out that in the above geometric algorithms, the eight pairwise distances used to characterize a catalytic triad were treated completely independently without taking into account any coupling effect among them. However, according to the reality in biochemical and biophysical process, the key constituent atoms of active site are highly coupled with one another, as reflected by many internal motion effects deduced theoretically (e.g., see Ref. 12) and observed experimentally (e.g., see Ref. 13), such as induced fitting effect and allosteric transition effect. Besides, of the topological entities derived from the atomic coordinates of Ser, His, and Asp according to Eq. 1, most would belong to the nonactive subset S^{2} because an enzyme molecule usually contains only one real catalytic triad. Therefore, we have $S^{2} \gg S^{1}$ or $n_{2} \gg n_{1}$ How to cope with such highly uneven subsets is also a key to enhance the prediction quality. To deal with these two critical issues, below we introduce the covariant discriminant algorithm. According to the algorithm, the similarity between $\mathbf{A}$ and $\bar{\mathbf{A}}^{\mu}$ is measured by the following function

$$\mathcal{T}(\mathbf{A}, \bar{\mathbf{A}}^{\mu}) = D_{\mathrm{M}}^{2}(\mathbf{A}, \bar{\mathbf{A}}^{\mu}) + \ln|\mathbf{C}^{\mu}|, \quad (\mu = 1, 2) \tag{9}$$

where

$$D_{\mathrm{M}}^{2}(\mathbf{A}, \bar{\mathbf{A}}^{\mu}) = (\mathbf{A} - \bar{\mathbf{A}}^{\mu})^{\mathrm{T}}\mathbf{C}_{\mu}^{-1}(\mathbf{A} - \bar{\mathbf{A}}^{\mu}) \tag{10}$$

is the squared Mahalanobis distance[14–16] between $\mathbf{A}$ and $\bar{\mathbf{A}}^{\mu}$, $\mathbf{T}$ is the transposition operator, and

$$\mathbf{C}_{\mu} = \begin{bmatrix} c_{1,1}^{\mu} & c_{1,2}^{\mu} & \cdots & c_{1,8}^{\mu} \\ c_{2,1}^{\mu} & c_{2,2}^{\xi} & \cdots & c_{2,8}^{\mu} \\ \vdots & \vdots & \ddots & \vdots \\ c_{8,1}^{\mu} & c_{8,2}^{\mu} & \cdots & c_{8,8}^{\mu} \end{bmatrix} \tag{11}$$

the covariance matrix for the subset S^{μ} and its elements given by

$$c_{i,j}^{\mu} = \frac{1}{n_{\mu} - 1} \sum_{k=1}^{n_{\mu}} [d_{k,i}^{\mu} - \bar{d}_{i}^{\mu}][d_{k,j}^{\mu} - \bar{d}_{j}^{\mu}],$$
$$(i, j = 1, 2, \ldots, 8;\ \mu = 1, 2), \tag{12}$$

while $\mathbf{C}_{\mu}^{-1}$ and $|\mathbf{C}_{\mu}|$ are the inverse matrix and the determinant of the matrix $\mathbf{C}_{\mu}$, respectively. $\mathcal{T}(\mathbf{A}, \bar{\mathbf{A}}^{\mu})$ as defined in Eq. 9 is called the covariant discriminant function: the smaller the value of the function, the higher the similarity between $\mathbf{A}$ and $\bar{\mathbf{A}}^{\mu}$. Accordingly, the prediction rule should now be expressed by

$$\mathcal{T}(\mathbf{A}, \bar{\mathbf{A}}^{\mu}) = \mathbf{Min}\{\mathcal{T}(\mathbf{A}, \bar{\mathbf{A}}^{1}), \mathcal{T}(\mathbf{A}, \bar{\mathbf{A}}^{2})\}, \tag{13}$$

where μ and **Min** have the exactly same meanings as in Eq. 6.

As we can see from the right side of Eq. 8, the first term incorporates the coupling effect among different pairwise distances,[17] whereas the second term makes allowance for modulating the effect caused by the unevenness of different subsets in size.[18]

RESULTS AND DISCUSSION

To construct a training dataset, 88 serine hydrolase entries in enzyme class E.C. 3.4.21 were selected from

TABLE II. List of the PDB Codes for the 50 Enzyme Molecules in the Testing Data Set

1A1Y	1A4W	1ABI	1ABJ	1AFQ	1BRB	1BRC	1C4V	1C4Y	1DWC
1DWD	1DWE	1ELB	1ELD	1ELE	1ELF	1ELG	1ELV	1GCT	1HAH
1HAI	1HAO	1HAP	1HAX	1HAZ	1HXF	1K1I	1NRO	1NRP	1NRQ
1NRR	1NRS	1SBH	1SBI	1SBN	1SCB	1SCD	1SCN	1SGQ	1SGR
1SGT	1SUP	1THR	1THS	1TMT	1TMU	2GCT	2HPQ	3TGK	7EST

TABLE III. Success Identification Rate for the 88 Enzymes of Table I by Jackknife Test

	Success identification rate		
Algorithm	Active site (A^+)	Nonactive site (A^-)	Overall (A)
Least Hamming distance[9]	$\frac{84}{88} = 95.45\%$	$\frac{5009}{6189} = 80.93\%$	$\frac{5093}{6277} = 81.14\%$
Least Euclidean distance[11]	$\frac{84}{88} = 95.45\%$	$\frac{5217}{6189} = 84.29\%$	$\frac{5301}{6277} = 84.45\%$
Covariant discriminant	$\frac{84}{88} = 95.45\%$	$\frac{6188}{6189} = 99.98\%$	$\frac{6272}{6277} = 99.92\%$

Enzyme Database,[19] and their PDB codes are given in Table I. Based on the atomic coordinates extracted from the corresponding PDB files, an active subset S^1 and a nonactive subset S^2 were formed for the training data set. The former contains $n_1 = 88$ vectors of $\mathbf{A}^1$ and the latter $n_2 = 6{,}189$ vectors of $\mathbf{A}^2$. Actually, according to the definition in the last section, 143,099 $\mathbf{A}^2$ vectors were found in S^2. However, after truncating those containing $d_\tau > 18$ Å ($\tau = 1, 2, \cdots$, or 8), the number of $\mathbf{A}^2$ vectors in S^2 was reduced to 6,189. The elements of the 88 $\mathbf{A}^1$ vectors in S^1 and the elements of the 6,189 $\mathbf{A}^2$ vectors in S^2, along with the corresponding sequence positions of His, Asp, and Ser, are given in the Online Supplemental Materials I. Similarly, an independent testing data set was constructed. It consists of 50 serine hydrolases. Their PDB codes are listed in Table II, and the corresponding vector elements and His-Asp-Ser sequence positions provided in the Online Supplemental Materials II.

The rates of correct prediction for the samples in the active subset S^1 and nonactive subset S^2 are defined by

$$\begin{cases} \Lambda^+ = \dfrac{N^+ - m^+}{N^+}, & \text{for active sites} \\ \Lambda^- = \dfrac{N^- - m^-}{N^-}, & \text{for nonactive sites} \end{cases} \tag{14}$$

where N^+ represents the total number of entities (samples) in the active subset S^1, and m^+ the number of catalytic triads missed in prediction; N^- is the total number of entities in the nonactive subset S^2, and m^- the number of noncatalytic triads incorrectly predicted as catalytic triads. Thus, the overall rate of correct prediction concerned is given by

$$\Lambda = \frac{\Lambda^+ N^+ + \Lambda^- N^-}{N^+ + N^-} = 1 - \frac{m^+ + m^-}{N^+ + N^-} \tag{15}$$

As is well known, the independent data set test, subsampling test and jackknife test are the three methods often used for cross-validation to examine the prediction quality in statistical prediction. Among these three, however, the jackknife test is deemed as the most effective and objective one (e.g., see Chou and Zhang[20] for a comprehensive discussion about this and Mardia et al.[10] for the mathematical principle). Jackknife test is particularly useful for checking the cluster-tolerant capacity[21] and, hence, was often used for the case when the training datasets were far from complete yet (e.g., see Refs. 22–24). During jackknifing, each protein in the data set is in turn singled out as a tested sample and all the rule parameters are calculated on the basis of the remaining samples. The predicted results thus obtained for the 88 proteins of Table I are given in Table III, from which we can see that the overall jackknife success rate by the current covariant discriminant algorithm is as high as 99.92%, which is about 15–20% higher than the simple geometry algorithms.[9,11] We also can see that the covariant discriminant algorithm is particularly effective in reducing the cases of incorrectly overpredicting nonactive sites as active ones.

Furthermore, as a demonstration of a practical application, predictions were also performed for the 50 independent enzymes in Table II using the rule parameters trained from the 88 enzymes of Table I. The predicted results are that, for the 50 active sites, 46 were correctly predicted, whereas for the 3128 non-active site entities, only one was incorrectly overpredicted as active site. Accordingly, the overall success prediction rate for the 50 independent enzymes is 99.84%. It is interesting to point out that, of the four enzymes whose active sites were missed in prediction, one is 1K1I. According to the annotation of its PDB file, the catalytic triad is formed by His-40, Asp-102, and Ser-195. On the basis of this finding, the following active topological entity was derived as given by (cf. Online Supplemental Materials II)

$$\mathbf{A}^1 = \begin{bmatrix} d_1 \\ d_2 \\ d_3 \\ d_4 \\ d_5 \\ d_6 \\ d_7 \\ d_8 \end{bmatrix} = \begin{bmatrix} 15.345 \\ 15.175 \\ 10.064 \\ 14.689 \\ 14.247 \\ 8.556 \\ 8.009 \\ 6.795 \end{bmatrix}. \tag{16}$$

The above equation means that the distances between His-N$^\delta$ and Asp-O$^{\delta 1}$, between His-N$^\delta$ and Asp-O$^{\delta 2}$, between His-N$^\epsilon$ and Asp-O$^{\delta 1}$, and between His-N$^\epsilon$ and Asp-O$^{\delta 2}$ are greater than 14 Å in the catalytic triad of 1K1I. It is highly unlikely according to the common sense in biochemistry. The problem with such unreasonably large internal distances for a catalytic triad might be due to some typographic error in annotating the His-Asp-Ser sequence positions. According to our prediction for the enzyme 1K1I, it was found that the topological entity for His-57, Asp-102, and Ser-195 had the following values for its matrix elements

$$\begin{bmatrix} d_1 \\ d_2 \\ d_3 \\ d_4 \\ d_5 \\ d_6 \\ d_7 \\ d_8 \end{bmatrix} = \begin{bmatrix} 3.465 \\ 2.703 \\ 4.993 \\ 5.438 \\ 4.587 \\ 3.106 \\ 8.009 \\ 6.795 \end{bmatrix}, \tag{17}$$

indicating that the catalytic triad in 1K1I is actually formed by His-57, Asp-102, and Ser-195 rather than His-40, Asp-102, and Ser-195. In other words, in the prediction for the 50 independent enzymes, one active site, which was originally thought missed, is not really missed, and one nonactive entity, which was thought incorrectly overpredicted as an active site, is not really incorrect. Accordingly, the real predicted results for the independent data set should be that, for the 50 active sites, 47 were correctly predicted; for the 3128 non-active site entities, none was incorrectly overpredicted as active site. The overall success rate was 99.91%.

Although the above demonstration is, as a paradigm, performed for the prediction of only catalytic triads in the serine hydrolase family, the algorithm and concept presented here can be used to any other enzyme families as well. However, when using the current method to a different enzyme family, such as aspartic protease family, a different training data set should be constructed because of the difference in the key components of the active site, leading to a different definition of topological entity (cf. Eq. 1) accordingly.

CONCLUSION

The atoms that constitute the active site of an enzyme molecule are highly coupled with each other during the process of performing its catalytic function. To reflect such an important biochemical and biophysical effects, the covariance matrix was incorporated in the prediction algorithm that made allowance for coupling interaction. This is the essence of why the covariant discriminant algorithm can yield much higher success rates in predicting the active site than the simple geometry algorithms. In addition, the novel algorithm is established on the basis of the equations derived from the purely analytical mathematics; hence, it is computationally much more efficient than the neural networks[25] and support vector machines[26] because no convergence requirement was involved during computation. For example, it took less than 3 seconds of CPU time by an SGI Octane workstation (300 MHz, R12000 processor) to complete the prediction of the catalytic triads for the 50 independent enzyme molecules in Table II. Therefore, in addition to the high accuracy, the computational speed would also be a remarkable advantage.

REFERENCES

1. Chou JJ, Matsuo H, Duan H, Wagner G. Solution structure of the RAIDD CARD and model for CARD/CARD interaction in caspase-2 and caspase-9 recruitment. Cell 1998;94:171–180.
2. Chou JJ, Li H, Salvessen GS, Yuan J, Wagner G. Solution structure of BID, an intracellular amplifier of apoptotic signalling. Cell 1999;96:615–624.
3. Zvelebil MJ, Barton GJ, Taylor WR, Sternberg MJ. Prediction of protein secondary structure and active sites using the alignment of homologous sequences. J Mol Biol 1987;195:957–961.
4. Nussinov R, Wolfson HJ. Efficient detection of three-dimensional structural motifs in biological macromolecules by computer vision techniques. Proc Natl Acad Sci USA 1991;88:10495–10499.
5. Artymiuk PJ, Poirrette AR, Grindley HM, Rice DW, Willett P. A graph-theoretic approach to the identification of three-dimensional patterns of amino acid side-chains in protein structures. J Mol Biol 1994;243:327–344.
6. Fischer D, Wolfson H, Lin SL, Nussinov R. Three-dimensional, sequence order-independent structural comparison of a serine protease against the crystallographic database reveals active site similarities: potential implications to evolution and to protein folding. Protein Sci 1994;3:769–778.
7. Wallace AC, Laskowski RA, Thornton JM. Derivation of 3D coordinate templates for searching structural databases: application to Ser-His-Asp catalytic triads in the serine proteinases and lipases. Protein Sci 1996;5:1001–1013.
8. Wallace AC, Borkakoti N, Thornton JM. TESS: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites. Protein Sci 1997;6:2308–2323.
9. Chou PY. Prediction of protein structural classes from amino acid composition. In: Fasman GD, editor. Prediction of protein structure and the principles of protein conformation. New York: Plenum Press; 1989. p 549–586.
10. Mardia KV, Kent JT, Bibby JM. Multivariate analysis. London: Academic Press; 1979. p 322 and 381.
11. Nakashima H, Nishikawa K, Ooi T. The folding type of a protein is relevant to the amino acid composition. J Biochem 1986;99:152–162.
12. Chou KC. Review: low-frequency collective motion in biomacromolecules and its biological functions. Biophys Chem 1988;30:3–48.
13. Chou JJ, Li S, Klee CB, Bax A. Solution structure of Ca$^+$-calmodulin reveals flexible hand-like properties of its domains. Nat Struct Biol 2001;8:990–997.
14. Mahalanobis PC. On the generalized distance in statistics. Proc Natl Inst Sci India 1936;2:49–55.
15. Pillai KCS. Mahalanobis D2. In: Kotz S, Johnson NL, editors.

Encyclopedia of statistical sciences. Vol. 5. New York: John Wiley & Sons; 1985. p 176–181.
16. Chou KC. A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space. Proteins 1995;21:319–344.
17. Chou KC, Zhang CT. Predicting protein folding types by distance functions that make allowances for amino acid interactions. J Biol Chem 1994;269:22014–22020.
18. Liu W, Chou KC. Prediction of protein structural classes by modified Mahalanobis discriminant algorithm. J Protein Chem 1998;17:209–217.
19. Bairoch A. The ENZYME database in 2000. Nucleic Acids Res 2000;28:304–305.
20. Chou KC, Zhang CT. Review: prediction of protein structural classes. Crit Rev Biochem Mol Biol 1995;30:275–349.
21. Chou KC. A key driving force in determination of protein structural classes. Biochem Biophys Res Commun 1999;264: 216–224.
22. Zhou GP. An intriguing controversy over protein structural class prediction. J Protein Chem 1998;17:729–738.
23. Zhou GP, Assa-Munt N. Some insights into protein structural class prediction. Proteins 2001;44:57–59.
24. Zhou GP, Doctor K. Subcellular location prediction of apoptosis proteins. Proteins 2003;50:44–48.
25. Cai YD, Liu XJ, Xu XB, Chou KC. Support vector machines for prediction of protein subcellular location by incorporating quasi-sequence-order effect. J Cell Biochem 2002;84:343–348.
26. Cai YD, Chou KC. Using neural networks for prediction of subcellular location of prokaryotic and eukaryotic proteins. Mol Cell Biol Res Commun 2000;4:172–173.