An Overview of the *De Novo* Prediction of Enzyme Catalytic Residues

Ziding Zhang*[,a], Yu-Rong Tang[a], Zhi-Ya Sheng[b] and Dongbin Zhao[a,c]

[a]*State Key Laboratory of Agrobiotechnology, College of Biological Sciences, China Agricultural University, Beijing 100193, China;* [b]*National Institute of Biological Sciences, No. 7 Science Park Road, Beijing 102206, China;* [c]*Current address: Atotech Deutschland GmbH, Zweigniederlassung Basel (F+E)R-1059.P, Postfach 4002 Basel, Switzerland*

Abstract: The identification of catalytic residues of an enzyme is one of the most important steps towards understanding its biological roles and exploring its applications. Thus far, a range of catalytic residue prediction methods have been developed, which play an increasingly important role in complementing the experimental characterization of enzymatic functions. The available approaches can be split into two broad categories: i) similarity-based catalytic residue annotation and ii) *de novo* catalytic residue prediction. In this article, we review the existing research strategies, recently developed bioinformatics tools, and future perspectives in the topic of *de novo* catalytic residue prediction. In particular, we review the various residue properties that have been used to distinguish catalytic and non-catalytic residues. We also detail how these residue properties can be combined into a prediction system with the assistance of different statistical or machine learning methods. Since in many respects *de novo* prediction of catalytic residues is still in its infancy, in this review we also propose some hints that are likely to result in novel prediction methods or increased performance.

Keywords: Bioinformatics, catalytic residues, machine learning methods, prediction.

1. INTRODUCTION

Providing functional annotation for vast amounts of protein sequence and structural data generated by high-throughput technologies is one of the major tasks in the post-genomic era [1-4]. Experimental determination of protein function is challenging, and performing assays to determine the function of all uncharacterized proteins is impossible. Thus, computational tools can play important roles in such a demanding task.

Enzymes are key proteins that are in charge of diverse biochemical functions and catalyze the chemical reactions related to the metabolism of all living organisms [5, 6]. Representing a significant fraction of a proteome, enzymes have long been categorized according to the Enzyme Commission (EC) system, a hierarchical classification that assigns unique four-number codes to different enzymatic reactions. The first number represents the general class of catalyzed reaction: i) oxidoreductases, ii) transferases, iii) hydrolases, iv) lyases, v) isomerases, and vi) ligases. The second and third number (i.e., sub-class and sub-subclass) further defines the catalyzed reaction, and the final number defines the substrate specificity. Prediction of the EC number for a query enzyme is a basic step towards understanding the enzymatic function and many efforts have been directed toward this prediction task in the past decade [6-8]. The identification of catalytic residues within a query enzyme is a further important step towards understanding the biological roles and applications of that enzyme, particularly since only a small number of residues within an enzyme molecule directly participate in catalysis, and the spatial arrangement as well as physicochemical properties of these residues (*i.e.*, catalytic residues) determine the chemical reaction catalyzed by the enzyme. The identified catalytic residues can provide useful information regarding the catalytic mechanism of enzymes, the construction of metabolic pathways, and enzyme-targeted drug discovery [9-11].

Catalytic residue prediction can be classified into sequence- and structure-based methods. The sequence-based method allows for prediction of the catalytic residues of the enzyme directly from the primary sequence. Provided that the predicted or experimentally determined 3D structure of a query enzyme is available, the structure-based method is used to predict catalytic residues based on its primary sequence as well as its 3D structure. The application of the structure-based method is somehow limited, since only a small fraction of proteins have known structures. In the past several years, structural genomics projects have provided a sharp increase in the number of structures of functionally unknown proteins. Therefore, algorithms capable of predicting catalytic residues within a structure are becoming increasingly useful. Moreover, elucidation of how protein structural information can be used to predict catalytic residues is also of theoretical interest.

In addition to the prediction of catalytic residues, efforts have also been focused on predicting active sites of enzymes. Generally, the active site of an enzyme consists of the corresponding catalytic residues and structurally neighboring residues. In some cases the active site can be defined as a sphere, in which the corresponding catalytic residues are centered (Fig. **1**). For an enzyme structure, a correct active site prediction means that the overlap between a predicted active site and the corresponding known site is above a certain threshold (e.g., 50%)[3]. Comparatively, the prediction of an active site is less challenging than that of catalytic residues. In this article, we focus on reviewing the prediction of catalytic residues. As a broader topic, protein functional site (residue) prediction tools have been widely reported in the

*Address correspondence to this author at the State Key Laboratory of Agrobiotechnology, College of Biological Sciences, China Agricultural University, Beijing 100193, China; Fax: +86-10-6274376;
Email: zidingzhang@cau.edu.cn

1574-8936/09 $55.00+.00

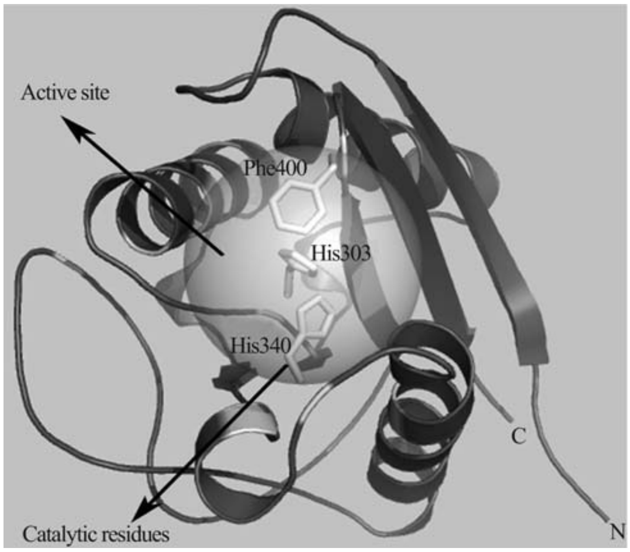

Fig. (1). Catalytic residues and active site in an enzyme structure (beta-ketoacyl-acyl carrier protein synthase II, PDB entry: 1kas). Three catalytic residues (His303, His340, and Phe400) are shown as sticks, while the active site is shown as a sphere.

literature. These tools may also be used to predict catalytic residues, since catalytic residues belong to an important type of functional residues. However, because the prediction of catalytic residues is not a major task of these functional site prediction tools, review of these tools is not included in this article.

Sequence and structural similarity-based methods are two classical bioinformatics strategies that are widely used to identify catalytic residues in a query enzyme. The sequence similarity-based method requires the identification of homologous enzyme sequences with known catalytic residues. Based on the sequence alignment of the query enzyme and an identified homolog, catalytic residues in the query sequence can be inferred from the identified homolog. Relying mainly on this approach, Mistry *et al.* (2008) performed catalytic residue annotation in the whole Pfam database [12]. If no homolog is available, the query sequence can be further searched against sequence motif databases (e.g., ProSite [13]). Once the query sequence contains some active site-related motifs, the corresponding catalytic residues in the query sequence may also be inferred. Although the above sequence similarity searching- or sequence motif matching-based annotations are relatively straightforward, such methods can result in false positive predictions in some cases, due to the fact that enzyme functions are less conserved [14-16].

When the 3D structure for the query enzyme is available, the structural similarity-based method is also able to identify catalytic residues, even when the sequence similarity-based method is not executable [17]. The potential catalytic residues in the query enzyme can be predicted through mapping catalytic residues of a structural homolog into the query enzyme. Generally, such structural similarity-based method can offer in-depth insight by highlighting 3D structural arrangements of catalytic residues. However, the power of structural similarity-based annotation is often weakened by the fact that a similar structure does not necessarily imply a similar function [18]. Furthermore, proteins without detectable sequence or structural similarity may have the similar spatial arrangements of active sites for catalyzing similar reactions (i.e., convergent evolution) [19-21]. Complementary to structural similarity-based methods, therefore, several methods focusing on the local pattern of active sites (*i.e.*, the active site structure motifs) and recognizing catalytic residues by searching query structure against active site templates of known enzymes have also been developed [21-23].

The above two classical strategies are further illustrated in Fig. (**2**). Due to the fact that the known enzyme active site sequence (structure) motifs are limited, the motif matching based methods are comparatively less useful. For practical use, a combination of these methods is strongly suggested to

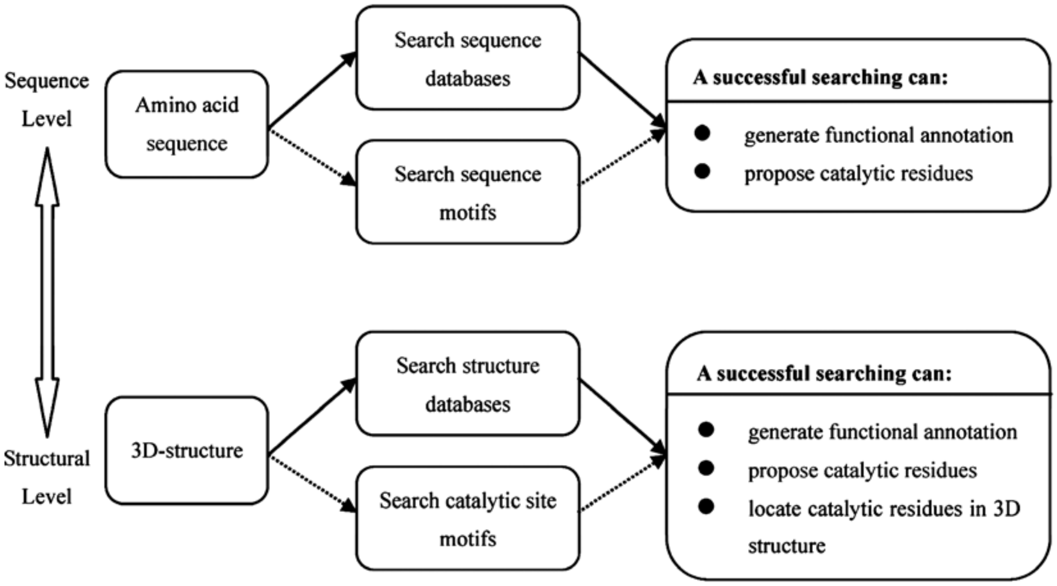
Sequence Level
Amino acid sequence
Search sequence databases
Search sequence motifs
A successful searching can:
● generate functional annotation
● propose catalytic residues
Structural Level
3D-structure
Search structure databases
Search catalytic site motifs
A successful searching can:
● generate functional annotation
● propose catalytic residues
● locate catalytic residues in 3D structure

Fig. (2). Flowchart of the classical catalytic residue detection methods. The active site sequence (structure) motif matching based methods (broken arrows) are comparatively less powerful than the sequence (structural) similarity searching based methods (solid arrows), since the known enzyme active site sequence (structure) motifs are still limited.

obtain a comprehensive understanding of the catalytic residues of a query enzyme. However, all the aforementioned methods may fail to predict catalytic residues for a query enzyme. Therefore, the development of *de novo* prediction methods (i.e., strategies independent of sequence alignment, sequence motif matching, structure comparison, or active site matching) is extremely important.

With the accumulated enzyme structures deposited in the PDB database [24], sequence and structural characteristics of catalytic residues have been extensively investigated [11, 25-32]. Meanwhile, *de novo* prediction methods have also been developed to identify catalytic residues in enzyme sequences and structures. Fig. (**3**) shows an overall flowchart for the development of a *de novo* prediction method. With the advantage of incorporating different sequence or structural properties into a predictor, statistical methods and machine learning algorithms, such as Artificial Neural Network (ANN) and Support Vector Machine (SVM), have also been used for the *de novo* prediction of catalytic residues in enzymes [3, 33-35]. Compared with other prediction tasks in the field of protein bioinformatics, the *de novo* prediction of catalytic residues is emerging as a hot topic and the published prediction algorithms have flourished in the past few years.

In this article, we review the existing research strategies, recently developed bioinformatics tools, and future perspectives concerning *de novo* catalytic residue prediction. In particular, the residue properties that have been reported for distinguishing catalytic and non-catalytic residues are intensively reviewed. We also detail how these residue properties can be combined into a prediction system with the assistance of statistical or machine learning methods. Since *de novo* prediction of catalytic residues may still be considered a relatively nascent methodology, novel and more effective methods are likely to appear in the very near future. Throughout this review, we will indicate those techniques that are likely to result in the development of novel tools or increased performance.

2. METHOD DEVELOPMENT OF *DE NOVO* CATALYTIC RESIDUE PREDICTION

2.1. Enzyme Databases

A breadth of information regarding enzymes, including sequences, structures, functions, catalytic residues, kinetics, binding affinity, enzyme reaction mechanisms, and metabolic pathways, has been continuously accumulated in the primary literature and compiled into different enzyme specific databases (e.g., IntEnz [36], CSA [10], BRENDA [37], EzCatDB [38], SFLD [39], and MACiE [40]). CSA (http://www.ebi.ac.uk/thornton-srv/databases/CSA/), which stands for the Catalytic Site Atlas, is a database that documents enzyme active sites and catalytic residues in enzymes with known 3D structure. The current CSA (version 2.2.10) contains 23,265 entries based on 968 literature entries. Following the rules established by Bartlett *et al.* (2002) [11], assignment of a catalytic residue in CSA includes: i) direct involvement in the catalytic mechanism; ii) effects exerted on residues or water molecules directly involved in catalysis; iii) stabilization a transient intermediate; and iv) the interaction with a substrate or cofactor that helps catalysis. To develop *de novo* catalytic residue predictors, the literature entries in CSA have been widely used and the definition of a catalytic residue in CSA is often regarded as the "gold standard". To facilitate training and testing of a prediction method, the sequences of CSA enzymes are often filtered to remove redundant structures. The annotated catalytic sites for each CSA enzyme serve as positive controls (i.e., catalytic residues), while all other residues are regarded as nega-

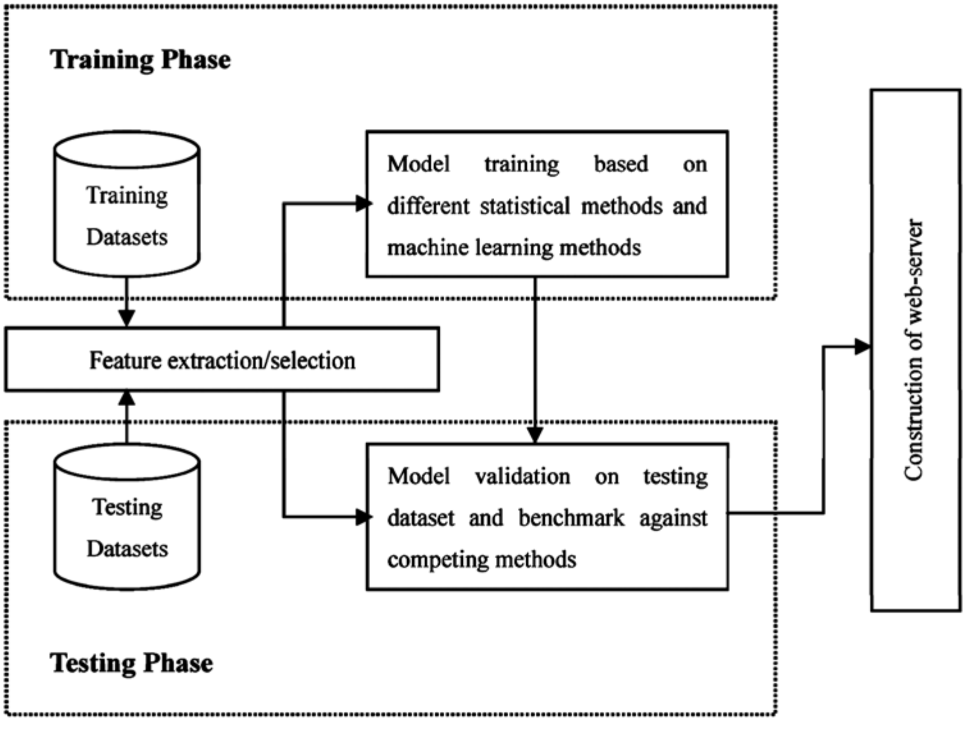
Training Phase
Training Datasets
Model training based on different statistical methods and machine learning methods
Feature extraction/selection
Testing Datasets
Model validation on testing dataset and benchmark against competing methods
Testing Phase
Construction of web-server

Fig. (3). Workflow of developing a *de novo* catalytic residue predictor.

tives (i.e., non-catalytic residues). It should be emphasized that there are much more non-catalytic than catalytic residues in enzyme sequences. A prediction model based on the original ratio of catalytic and non-catalytic residues in CSA enzymes would inevitably result in a strong bias toward prediction of all residues as non-catalytic [3, 41]. To avoid such bias, the ratio of catalytic to non-catalytic residues is optimally selected to create the training dataset and the original ratio is kept for the testing dataset.

2.2. Different Residue Properties Used in Developing *de Novo* Predictors

To construct a *de novo* catalytic residue predictor, residue properties that can be used to distinguish catalytic and non-catalytic residues must be explored and converted into feature vectors (also known as encodings or descriptors). Residue properties reviewed herein cover residue type, sequence conservation, network centrality, relative position, hydrogen bonding, solvent accessibility, flexibility, secondary structure information, electrostatic property, and structural stability score. Rather than quantitatively evaluating the performance of these residue properties for distinguishing catalytic and non-catalytic residues, we instead emphasize discussion of the physicochemical implications of these residue properties and review some necessary bioinformatics tools to obtain the corresponding encodings. It is also worth mentioning that some residue properties are directly inferred from protein sequences, while the calculation of some other residue properties requires the information of protein 3D structures. More details on these encodings are described herein.

2.2.1. Residue Type

Different amino acids have different propensities to be catalytic residues [11]. For catalytic residues, approximately 65% are charged residues (H, R, K, E, D), 27% of catalytic residues are polar residues (Q, T, S, N, C, Y, W), and 8% are composed of hydrophobic residues (G, F, L, M, A, I, P, V), as reported by Bartlett *et al.* [11]. Two encodings were frequently used to represent this property of different residue types. The first encoding is the standard binary encoding, which entails each of the 20 amino acids encoded as a 20-dimensional binary vector, e.g., A (1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0), C (0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0), …, and Y (0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1). Considering that the 20 amino acids can be grouped into different types based on physicochemical properties, the second encoding was utilized. For instance, the implemented residue type encoding was based on a three-type classification of the 20 amino acids, in which charged, polar, and hydrophobic residues were encoded as (0 0), (0 1), and (1 0), respectively, as reported by Petrova and Wu [33].

2.2.2. Sequence Conservation

One of the most important characteristics of catalytic residues is the high degree of conservation [10, 11]. Therefore, the sequence conservation-related encoding is the most important component in the current *de novo* catalytic residue predictors. To compute the conservation score for a residue, an iterated PSI-BLAST searching [42] for the corresponding sequence was performed against the NCBI non-redundant protein sequence database to obtain a multiple sequence alignment (MSA). The MSA was then used to infer the conservation score for each residue within a sequence. Thus far, a series of conservation scoring methods have been developed, such as Shannon entropy, von Neumann entropy, and relative entropy. Please refer to Valdar (2002) [43] for a more complete discussion of these scoring schemes and the evolution of these methods. To score residue conservation, some well-maintained servers are available to the community. For example, the Scorecons server (http://www.ebi.-ac.uk/thornton-srv/databases/cgi-bin/valdar/scorecons_server.-pl) is widely used. These diverse conservation scoring methods may yield different performances. In 2008, Capra and Singh [44] evaluated different conservation scoring methods and determined that the Jensen-Shannon divergence-based score is the most informative for detection of the catalytic residues.

In addition to sequence conservation, information regarding evolutionary conservation was also introduced through phylogenetic analysis. Wang *et al.* (2008) proposed a novel score, called the state to step ratio score (SSR), for measuring evolutionary conservation [45]. The maximum parsimony tree can be constructed based on a given MSA. Then, the variation patterns from the root of the tree (theoretical ancestral sequence) to the leaf of the tree (sequences in MSA) are used to create a score (i.e., SSR) for each residue. The SSR score has been established as a simple, yet effective evaluation for measuring evolutionary conservation and has been demonstrated as a useful descriptor for distinguishing catalytic and non-catalytic residues. It should be mentioned that some query proteins may fail to find sufficiently diverse homologues. In this case, the power of the above sequence conservation or evolutionary conservation is relatively limited.

2.2.3. Network Centrality

Each protein structure can be transformed into a network (i.e., an undirected residue interaction graph (RIG)), with residues modeled as vertices and residue interactions as edges [46]. A series of network topology parameters were explored within the established network. In the past several years, investigations utilizing the RIGs of enzyme structures have demonstrated that residues within or directly contacting active site usually have more interactions with other residues, and the centrality values of catalytic residues in the network are typically high, especially the closeness centrality [26, 27, 30]. As the most informative network topology parameter, the closeness centrality has thus been employed to distinguish catalytic and non-catalytic residues [35].

The application of RIG has also been explored in other topics. Based on the RIGs, Li *et al.* reported a simple method to detect the folding nucleus in a protein structure [47]. Recently, David-Eden and Mandel-Gutfreund [48] performed such network analysis on ribosomal structures, elucidating that the major functional sites of the ribosome exhibit significantly high centrality measures. In the above calculation, no weight was assigned for any edge within the network graph, i.e., the strength of each interaction is disregarded in this network. Blundell and co-workers (2008) also used a similar network analysis to predict the structural effects caused by non-synonymous single nucleotide polymorphisms (nsSNPs) [49]. Furthermore, the network analysis was extended to predict disease-associated nsSNPs with high-quality performance, which may be ascribed to the introduced weightings of edges. Likewise, suitable weighting schemes need to be considered to further explore the application of RIG in catalytic residue prediction.

2.2.4. Relative Position

Catalytic residues for almost all enzymes tend to reside in a large cleft on the molecular surface [11, 50, 51]. As observed by Bartlett *et al.* (2002) [11] , this tendency is particularly striking for the largest cleft, and is also significant for the second and third largest clefts. The difference is not significant for clefts that are smaller than the third largest [11]. As a representation of the relative position of a given residue within a protein structure, the cleft related encoding was thus employed in several prediction methods [3, 33, 35]. All clefts for a given structure can be detected through the use of computational programs (e.g., SURFNET [52] and CASTp [53]), and then the cleft encoding for each residue can be assigned. For instance, (1 0 0 0) denotes those residues in the largest cleft, (0 1 0 0) represents the second or third largest cleft, (0 0 1 0) indicates the fourth to ninth largest cleft, and (0 0 0 1) corresponds to none of the above clefts, as described by Gutteridge *et al.* (2003) [3]. Based on a similar strategy, some other cleft encoding variants were also proposed by Petrova and Wu [33]. In 2005, Ben-Shimon and Eisenstein found that catalytic residues are very often located among the 5% of residues closest to the centroids of enzyme molecules [29]. Moreover, this property of catalytic residues was implemented in a predictor called EnSite for locating the active sites of enzymes. In contrast to cleft encoding, the property proposed by Ben-Shimon and Eisenstein also reflects the relative position of residues and can be easily derived from a query protein's 3D structure, since searching and defining all the clefts on the enzyme molecular surface is not necessary.

2.2.5. Hydrogen Bonds

Most catalytic residues act as donors or acceptors in at least one hydrogen bond [11]. Therefore, hydrogen bond related information could be incorporated into different catalytic residue predictors [3, 33, 35]. Generally, the hydrogen bond information for a residue can be assigned from the query protein's 3D structure, through using a range of well established software (e.g., HBPLUS [54]). In our previous work [35], the following three parameters were used to represent this property: i) the number of hydrogen bonds from a main-chain atom in a given residue to any other atom in a protein (*NmHB*); ii) the number of hydrogen bonds from a side-chain atom in a given residue to any other atom in a protein (*NsHB*); and iii) the total number of hydrogen bonds involving any atom in a given residue (*tNHB*). Comparatively, *NsHB* is the most informative among these hydrogen bond related encodings [33, 35].

2.2.6. Relative Solvent Accessibility

Catalytic residues are generally more exposed to solvent than other residues. For a query protein structure, the relative solvent accessibility (RSA) for each residue can be computed via some software (e.g., NACCESS [55]), and then some RSA based encodings can be constructed. As reported in the literature [33, 35], five RSA based encodings were explored, including the RSA of all atoms (*AaRSA)*, the RSA of all side chain atoms including alpha carbons (*AsRSA*), the RSA of non-polar side chain atoms (*NpRSA*), the RSA of all polar side chain atoms (*ApRSA*), and the RSA of all main chain atoms (*McRSA*). Comparatively, *AsRSA* was found to be the most informative encoding.

2.2.7. Structural flexibility

Catalytic residues tend to be more rigid than average ones in an enzyme structure [11, 56]. As reported in the literature [3, 33, 35], the encodings based on the B-factors were frequently used to measure residue flexibility. One caveat is that this encoding may only be suitable for those protein structures determined by X-ray crystallography.

2.2.8. Secondary Structure

Catalytic residues are more inclined to locate in coil regions [11]. Therefore, the secondary structure state (SSS) of a residue can be employed as a useful encoding in the prediction of catalytic residues. For a sequence based prediction, the SSS of a residue can be predicted via secondary structure prediction methods (e.g., PSIPRED[57]); for a structural based prediction, the SSS can be assigned from the corresponding protein structure through using secondary structure assignment methods (e.g., DSSPcont [58]).

2.2.9. Computed Electrostatic Properties

Some complicated residue properties (e.g., computed electrostatic properties) were also explored to predict catalytic residues. Ondrechen *et al.* reported a computational method, namely theoretical microscopic titration curves (THEMATICS), for the identification of active sites in protein structures [59]. To perform the calculation of THEMATICS, the Poisson-Boltzmann (P-B) equations must first be solved, and then the proton occupations of the ionizable residues are computed as functions of pH values. A small proportion of the ionizable residues in proteins were reported as perturbed, exhibiting non-Henderson-Hasselbalch (H-H) titration behavior [60]. Since the residues with perturbed titration behavior are more likely to appear in the active site of an enzyme [27, 59-62], THEMATICS is able to detect these perturbed residues and predict the location of the active site. Based on the computed electrostatic properties, Bate and Warwicker (2004) [28] also developed a method to identify a point near the active site using the peak of the electrostatic potential in the solvent space above the protein structure.

2.2.10. Structural Stability Score

There is accumulated evidence that catalytic residues are conserved to maintain the enzymatic function at the cost of stability [51]. Mutations of active site residues usually result in increased stability. Therefore, the property reflecting the destabilizing effect of a residue could be employed to distinguish catalytic and non-catalytic residues, which have been implemented in several catalytic residue prediction methods [45, 51]. For instance, Wang *et al.* (2008) [45] used residue-specific all atom probability discriminatory function (RAPDF) based scores to quantify the structural stability of a residue. Furthermore, these two RAPDF based scores were integrated into their catalytic residue prediction system called MFS [45]. Similar to the calculation of electrostatic properties, a relatively complicated calculation is required to obtain such structural stability scores. To derive these two RAPDF based scores, each residue in a query protein structure was mutated into one of the 19 alternative amino acids, and the generated new structures were further refined for the optimization of topology and the maximization of stability [45]. Finally, the two RAPDF based scores were obtained from the set of 20 conformations via complex energy calculations.

2.2.11. Other Residue Properties

In addition to the aforementioned encodings, other residue properties have also been reported in the literature. For instance, Zhang *et al.* (2008) [41] used the average cumulative hydrophobicity of a residue as a feature representation. Meanwhile, some frequently occurring catalytic residue pairs in known enzymes were used to construct a feature vector. As reported by Pugalenthi *et al.* (2008) [63], a series of physicochemical properties calculated from each residue and the corresponding spatial neighbors were also used as a type of feature encoding. A total of 264 atom-based structural properties were calculated using S-BLEST [64] and employed for an important protein structure based encoding. Notably, extraction of a novel and effective residue property is becoming more challenging. In fact, some newly developed descriptors frequently contain overlapped information with the previously reported residue properties. For instance, the physicochemical properties used by Pugalenthi *et al.* [63] overlap with the residue type encoding reviewed in section 2.2.1 to some extent.

2.3. Prediction Algorithms

2.3.1. Prediction Algorithms Based on Some Individual Descriptors

As reviewed in the previous section, many residue properties have been explored to distinguish catalytic and noncatalytic residues. Based on these descriptors, a series of *de novo* catalytic residue prediction algorithms were developed [3, 33-35, 41]. A few predictors rely heavily on some individual descriptors. For instance, the network centrality and RSA properties have been combined to detect catalytic residues in protein structures [27, 30]. Developed by Ondrechen *et al.* (2001)[59], THEMATICS is merely based on the computed electrostatic properties to predict catalytic residues in protein structures. Conceptually, such methods are most likely the best predictors, since the physicochemical meanings of employed descriptors are easily interpreted.

2.3.2. Prediction Algorithms Based on Simple Statistical Methods

To develop a predictor with improved performance, combination of different descriptors is necessary, which can be developed by employing statistical methods and state-of-the-art machine learning methods. Simple statistical methods can yield an improved performance by efficiently integrating several largely independent descriptors in a simple model. For instance, Fischer *et al.* (2008) [65] used conditional probability density estimation to calculate the probability of each residue to be catalytic given its conservation, the profile amino acid frequencies, and the predicted secondary structure and RSA states. Thus, several descriptors were effectively combined into a simple statistical frame [65]. Recently, Wang *et al.* (2008) used a simple logistic regression model to integrate several descriptors into a predictor called MFS [45]. The employed descriptors in MFS include sequence conservation, evolutionary conservation, structure stability score, and residue type. As indicated by Wang *et al.* (2008)[45], such simple statistical models are conceptually valuable with statistical parameters that are comprehensible.

2.3.3. Prediction Algorithms Based on Machine Learning Approaches

An alternative way to integrate descriptors is through the use of sophisticated machine learning approaches. Two frequently used machine learning methods are ANN and SVM. Based on the same dataset and descriptors, different machine learning methods can have distinctive performances. In general, SVM appears to be more popular in the *de novo* prediction of catalytic residues as well as for other bioinformatics prediction topics. Recently, some free available machine learning software packages (e.g., WEKA (http://www.cs.waikato.ac.nz/ml/weka/)) allowed developers to test different machine learning methods for a given prediction task. One of the cornerstone methods for *de novo* prediction of catalytic residues was developed by Thornton and co-workers using ANN [3] . Petrova and Wu (2006) [33] evaluated 26 different algorithms in the WEKA software package, and reported that a SVM model trained on a set of seven out of 24 residue properties can result in an optimized performance for predicting catalytic residues. Furthermore, Youn *et al.* (2006) [34] tested SVM on 314 different features, demonstrated that the combination of multiple features improves performance, and presented the most highly ranked features. Using SVM, Pugalenthi *et al.* (2008) [63] tested 278 different features for catalytic site prediction and investigated the performance with a refined subset of features. As reported in our previous study [35], a genetic algorithm assisted neural network (GANN) was employed to construct an improved catalytic residue predictor. The core idea of GANN is to use a genetic algorithm (GA) for optimization of the connection weights within neural networks [66]. Based on our dataset, GANN can result in a better performance than SVM [35]. Although machine learning methods can usually lead to improved performance, some of them are often criticized and labeled as “black box” methods, due to a lack of biological interpretation. The most prominent example of “black box” methods is ANN, which is not an appropriate technique if the biological interpretation of the model is desired.

Although machine learning methods are able to combine many descriptors within one prediction system, a refined subset should be selected from all the descriptors under investigation. A refined subset can effectively avoid “the curse of dimensionality”, filter the overlapped information caused from different features, and result in an improved predictive accuracy. Furthermore, a refined descriptor subset can result in a prediction system that is more conceptually concise by highlighting the key descriptors. Based on a benchmarking dataset of 254 catalytic residues, As reported by Petrova and Wu (2006)[33], the wrapper subset selection algorithm was performed on a dataset of 254 catalytic residues and 254 non-catalytic residues. Of the 24 attributes under investigation, seven attributes with a dimension of 26 were selected as an optimal subset of residue properties. The feature selection tool, based on LIBSVM [67], was employed to select the optimized subset of properties in our previous study [35]. Based on a balanced dataset of 480 residues (i.e. 240 catalytic residues and 240 non-catalytic residues), eight properties with a dimension of 30 contributed the optimal performance. In 2008, Zhang *et al.* used the $\chi 2$-statistic algorithm to perform the feature selection [41]. The $\chi 2$-value for each feature was computed on a dataset of 606 catalytic and 3636 non-catalytic residues. Then, the initial 544 features were ranked according to the calculated $\chi 2$-values. Finally, the top 210 ranked features were optimally selected in training the prediction model [41]. To identify the prominent features that separate the positive and negative classes, the information gain algorithm was employed in the prediction method of Pugalenthi *et al.* [63]. Based on a benchmarking dataset of 1500 catalytic residues and 1500 non-catalytic residues, the selected 100 features out of the 278 features under investigation resulted in the best performance. The feature selection is often required to handle multidimensional feature vectors, and many methods have been developed and widely used to address different prediction tasks. In principle, the feature-selection methods that have been used in other topics can also be introduced in the prediction of catalytic residues.

2.4. Performance Assessment of Different Predictors

2.4.1. Training and Cross-Validation

For statistical theory or machine learning approach based prediction algorithms, the dataset is generally divided into training dataset and testing datasets. The prediction model was inferred from the training dataset and cross-validated by the testing dataset. As previously reported, *n*-fold cross-validation was frequently employed. The whole dataset was randomly divided into *n* subgroups of roughly equal size. In each evaluation step, one subgroup was selected for testing, while the other *n*-1 subgroups were used as the training dataset. Finally, the overall performance was averaged over the *n*-fold cross-validation experiments. To obtain prediction models that are suitable for real enzymes, a 1:6 ratio of catalytic to non-catalytic residues was frequently assigned as the best ratio to train the corresponding prediction models [3, 34, 35, 41].

2.4.2. Performance Measures

As reported in the literature, four measurements, i.e., Accuracy (*AC*), True positive rate (*TPR*), False positive rate (*FPR*), and Matthews correlation coefficient (*MCC*), were frequently used to evaluate the prediction performance:

$$AC = \frac{TP + TN}{TP + FN + TN + FP}$$

$$TPR = Sensitivity = \frac{TP}{TP + FN}$$

$$FPR = 1 - Specificity = \frac{FP}{FP + TN}$$

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

where *TP*, *FP*, *FN,* and *TN* denote true positives, false positives, false negatives, and true negatives. *MCC* should be more suitable for assessing the overall prediction accuracy, since the numbers of catalytic and non-catalytic residues are different in real enzymes. The value of *MCC* ranges from -1 to 1, and a higher *MCC* indicates a better prediction performance. In general, *MCC* = 1 conveys the best prediction, *MCC* = -1 indicates the worst prediction, and *MCC* = 0 means a random prediction. In addition, similar parameters to the above four measures were also employed. For instance, the parameter *Precision* (i.e., *Precision=TP*/(*TP*+*FP*)) was utilized by Zhang *et al.* (2008) [41].

Prediction accuracy was also assessed through using the ROC analysis [68, 69]. For a prediction method, the curve of ROC plots true positive rate (i.e., *Sensitivity*) as a function of false positive rate (i.e., 1-*Specificity*) for all possible thresholds. The area under the ROC curve (AUC) was also calculated to provide a comprehensive understanding of the proposed prediction method. Generally, the closer the AUC value is to 1 indicates a better prediction method. It should be emphasized that some ROC curve variants were also frequently invented to assess the catalytic residue predictors [65].

2.4.3. Comparison of Different Prediction Methods

Rather than ranking the performance of different predictors quantitatively, we instead focus on discussing the overall predictive accuracy. With the booming of newly developed predictors, *de novo* catalytic residue prediction has been improved to reach a level of reasonably good accuracy. Of these predictors, most have focused on detecting catalytic residues from protein structures and only a few predict catalytic residues directly from protein sequences. As clearly demonstrated by some structural based predictors [3, 33, 35], residue properties inferred from protein structure do improve the prediction accuracy. It is also interesting to mention that some sequence based predictors also exhibit fully comparable performance [41, 44, 65]. Despite the improvements indicated above, the *MCC* value of some predictors remained below 0.4 in the identification of catalytic residues for entire enzyme molecules, suggesting that the current algorithms were still not suitable for practical use [33, 35]. In other words, the catalytic residue prediction result of a real enzyme may inevitably contain too many false positives (*i.e.*, *Precision* is too low). The catalytic residue prediction of an enzyme structure, which resulted from our previous method [35], was used to demonstrate this low *Precision.* As shown in Fig. (**4**), two of three catalytic residues in beta-ketoacyl-acyl carrier protein synthase II (PDB entry: 1kas) could be successfully identified, but the prediction also resulted in eight false positives (i.e., *Precision* = 20%). Therefore, it is still quite time-consuming and daunting for experimental scientists to characterize the correct catalytic residues, based solely on the results from the currently available predictors.

Due to relatively limited data, the accuracy of different predictors reported by developers may be more or less overestimated, since these methods were predominantly optimized for the corresponding training and testing data. Generally, a newly developed predictor is needed to be intensively benchmarked with some existing methods, before acceptance of this predictor for publication in a peer-reviewed journal. However, such benchmark experiments may still be quite arbitrary, due to the lack of “gold standard” datasets,

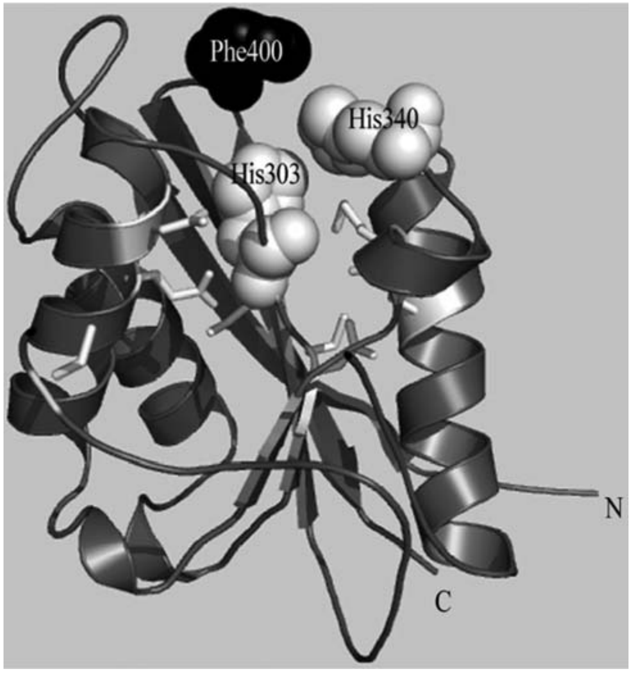

Fig. (4). Predicted catalytic residues in beta-ketoacyl-acyl carrier protein synthase II (PDB entry: 1kas). The prediction was performed by using our previously described method [35], in which the predictive model was trained on a 1:6 ratio of catalytic to non-catalytic residues. White spheres indicate true positives, black spheres indicate false negatives and false positives are shown by side chains in white sticks.

which can sufficiently cover enzyme sequence and structure spaces. With increasing experimental verification of catalytic residues, some standard training and testing datasets should be available in the near future. Thus, different prediction methods can be reliably benchmarked. Meanwhile, some well-established strategies for assessing different protein structure prediction methods (e.g., Live-Bench [70] and EVA [71]) should also be considered for evaluating different catalytic residue predictors.

2.5. Web-Servers for Some Existing Predictors

Thus far, few *de novo* catalytic residue prediction web-servers have been available to the community and these URLs are summarized in Table **1**. Some sequence (structure) motifs- based catalytic site detection servers (i.e., E1DS [72], PAR-3D [22], and Catalytic Site Search [23]) are also listed in Table **1**. However, the machine learning methods-based catalytic residue prediction server is still not available. To develop a bioinformatics tool, providing a web-server is vital for the community as well as developers. Free-accessible web-servers can allow users to experience the power of these algorithms and then maximize applications of the algorithms. In the meantime, feedbacks from users will also urge developers to continuously improve their algorithms. Compared with other areas of bioinformatics, the number of catalytic residue prediction servers is relatively small, which may be ascribed to the following reasons. First, some methods heavily rely on other algorithms to calculate different residue properties. Once the source (binary) codes of such algorithms are not publicly available, establishment of a web-server is difficult. Second, the less impressive performance of *de novo* catalytic residue prediction may also hinder developers from constructing web-servers. Even so, these available web-servers play increasingly important roles to help experimental scientists accelerate the functional characterization of enzyme molecules. Although prediction scores for the predicted catalytic residues are provided, the statistical significances of the prediction scores are not apparent in each prediction server. For this, we may follow the current protein fold recognition servers, in most of which the confident levels for different prediction scores are well defined [70, 73]. By learning from the protein fold recognition community in further, it is also expected that a meta-server could be developed for different catalytic residue prediction methods. Thus, users can take advantage of the results from different methods to make more reliable predictions.

3. FUTURE PERSPECTIVES

In summary, the *de novo* catalytic residue prediction of enzymes is an increasingly important topic in the field of protein bioinformatics, and there have been major improvements in the past several years. The current available web-servers play an important role to help experimental scientists to accelerate the functional characterization of enzyme molecules. Considering the overall performance of different prediction methods, however, the current *de novo* predictors are still not effective for practical use.

To improve the *de novo* prediction of catalytic residues, the following strategies can be used. First, the combination of the current *de novo* algorithms with some classical catalytic residue detection methods (e.g., active site structure

Table 1. A Selection of Catalytic Residue Prediction Web-Servers

Methods	URLs	References
SARIG[a]	http://bioinfo2.weizmann.ac.il/~pietro/SARIG/V3/index.html	[27]
THEMATICS[a]	http://pfweb.chem.neu.edu/	[59, 60]
FRPRED[a]	http://toolkit.tuebingen.mpg.de/frpred	[65]
PAR-3D[b]	http://sunserver.cdfd.org.in:8080/protease/PAR_3D/index.html	[22]
E1DS[b]	http://e1ds.ee.ncku.edu.tw/	[72]
Catalytic Site Search[b]	http://www.ebi.ac.uk/thornton-srv/databases/cgi-bin/CSS/makeEbiHtml.cgi?file=form.html	[23]

[a] *De novo* catalytic residue prediction methods. [b] Sequence (structure) motifs searching-based catalytic residue prediction methods.

motifs-based searching methods) can result in a more reliable catalytic residue prediction. Second, selection of more advanced statistical or machine learning methods may also be possible to achieve better performances. Third, developing some enzyme class specific predictors can yield higher prediction accuracy [9, 22, 74]. However, a *priori* knowledge of the enzyme class that a query enzyme belongs to is required for such predictors. In addition to the above strategies, exploring new properties (encodings) is still the most important direction for development of a better predictor. Recently, some complicated physicochemical attributes inferred from protein structures have been used to predict catalytic sites [75, 76]. To detect functional sites within a protein, developers have been heavily involved in finding new sequence or structural properties (e.g., see Refs. [77-80]). Careful validation on these properties may result in the discovery of encodings suitable for the prediction of catalytic residues, since catalytic residues belong to an important type of functional residues. Through the integration of more physical chemistry in prediction models, we expect that newly identified residue properties will definitely improve the overall performance for prediction of catalytic residues as well as strengthen our basic understanding of the molecular mechanisms of enzymatic reaction.

ACKNOWLEDGEMENTS

We are grateful to the anonymous referees whose constructive comments were very helpful in improving the quality of this work. This work was supported by grants from the National Natural Science Foundation of China (30700137), and the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry.

REFERENCES

[1] Ofran Y, Punta M, Schneider R, Rost B. Beyond annotation transfer by homology: novel protein-function prediction methods to assist drug discovery. *Drug Discov Today* **2005**; 10: 1475-82.

[2] Shapiro L, Harris T. Finding function through structural genomics. *Curr Opin Biotechnol* **2000**; 11: 31-5.

[3] Gutteridge A, Bartlett GJ, Thornton JM. Using a neural network and spatial clustering to predict the location of active sites in enzymes. *J Mol Biol* **2003**; 330: 719-34.

[4] Kristensen DM, Ward RM, Lisewski AM*, et al.* Prediction of enzyme function based on 3D templates of evolutionarily important amino acids. *BMC Bioinformatics* **2008**; 9: 17.

[5] Freilich S, Spriggs RV, George RA*, et al.* The complement of enzymatic sets in different species. *J Mol Biol* **2005**; 349: 745-63.

[6] Arakaki AK, Tian W, Skolnick J. High precision multi-genome scale reannotation of enzyme function by EFICAz. *BMC Genomics* **2006**; 7: 315.

[7] Kunik V, Meroz Y, Solan Z*, et al.* Functional representation of enzymes by specific peptides. *PLoS Comput Biol* **2007**; 3: e167.

[8] Cai CZ, Han LY, Ji ZL, Chen YZ. Enzyme family classification by support vector machines. *Proteins* **2004**; 55: 66-76.

[9] Chou KC, Cai YD. A novel approach to predict active sites of enzyme molecules. *Proteins* **2004**; 55: 77-82.

[10] Porter CT, Bartlett GJ, Thornton JM. The catalytic site atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res* **2004**; 32: D129-33.

[11] Bartlett GJ, Porter CT, Borkakoti N, Thornton JM. Analysis of catalytic residues in enzyme active sites. *J Mol Biol* **2002**; 324: 105-21.

[12] Mistry J, Bateman A, Finn RD. Predicting active site residue annotations in the Pfam database. *BMC Bioinformatics* **2007**; 8: 298.

[13] Hofmann K, Bucher P, Falquet L, Bairoch A. The PROSITE database, its status in 1999. *Nucleic Acids Res* **1999**; 27: 215-9.

[14] Rost B. Enzyme function less conserved than anticipated. *J Mol Biol* **2002**; 318: 595-608.

[15] Tian W, Skolnick J. How well is enzyme function conserved as a function of pairwise sequence identity? *J Mol Biol* **2003**; 333: 863-82.

[16] Todd AE, Orengo CA, Thornton JM. Evolution of function in protein superfamilies, from a structural perspective. *J Mol Biol* **2001**; 307: 1113-43.

[17] Orengo CA, Todd AE, Thornton JM. From protein structure to function. *Curr Opin Struct Biol* **1999**; 9: 374-82.

[18] Nagano N, Orengo CA, Thornton JM. One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions. *J Mol Biol* **2002**; 321: 741-65.

[19] Zhang Z, Grigorov MG. Similarity networks of protein binding sites. *Proteins* **2006**; 62: 470-8.

[20] Zhang Z, Tang YR. Genome-wide analysis of enzyme structure-function combination across three domains of life. *Protein Pept Lett* **2007**; 14: 291-7.

[21] Torrance JW, Bartlett GJ, Porter CT, Thornton JM. Using a library of structural templates to recognise catalytic sites and explore their evolution in homologous families. *J Mol Biol* **2005**; 347: 565-81.

[22] Goyal K, Mohanty D, Mande SC. PAR-3D: a server to predict protein active site residues. *Nucleic Acids Res* **2007**; 35: W503-5.

[23] Barker JA, Thornton JM. An algorithm for constraint-based structural template matching: application to 3D templates with statistical analysis. *Bioinformatics* **2003**; 19: 1644-9.

[24] Berman HM, Westbrook J, Feng Z*, et al.* The protein data bank. *Nucleic Acids Res* **2000**; 28: 235-42.

[25] Zvelebil MJ, Sternberg MJ. Analysis and prediction of the location of catalytic residues in enzymes. *Protein Eng* **1988**; 2: 127-38.

[26] del Sol A, Fujihashi H, Amoros D, Nussinov R. Residue centrality, functionally important residues, and active site shape: analysis of enzyme and non-enzyme families. *Protein Sci* **2006**; 15: 2120-8.

[27] Amitai G, Shemesh A, Sitbon E*, et al.* Network analysis of protein structures identifies functional residues. *J Mol Biol* **2004**; 344: 1135-46.

[28] Bate P, Warwicker J. Enzyme/non-enzyme discrimination and prediction of enzyme active site location using charge-based methods. *J Mol Biol* **2004**; 340: 263-76.

[29] Ben-Shimon A, Eisenstein M. Looking at enzymes from the inside out: the proximity of catalytic residues to the molecular centroid can be used for detection of active sites and enzyme-ligand interfaces. *J Mol Biol* **2005**; 351: 309-26.

[30] Chea E, Livesay DR. How accurate and statistically robust are catalytic site predictions based on closeness centrality? *BMC Bioinformatics* **2007**; 8: 153.

[31] Tobi D, Bahar I. Recruitment of rare 3-grams at functional sites: is this a mechanism for increasing enzyme specificity? *BMC Bioinformatics* **2007**; 8: 226.

[32] Meroz Y, Horn D. Biological roles of specific peptides in enzymes. *Proteins* **2008**; 72: 606-12.

[33] Petrova NV, Wu CH. Prediction of catalytic residues using support vector machine with selected protein sequence and structural properties. *BMC Bioinformatics* **2006**; 7: 312.

[34] Youn E, Peters B, Radivojac P, Mooney SD. Evaluation of features for catalytic residue prediction in novel folds. *Protein Sci* **2007**; 16: 216-26.

[35] Tang YR, Sheng ZY, Chen YZ, Zhang Z. An improved prediction of catalytic residues in enzyme structures. *Protein Eng Des Sel* **2008**; 21: 295-302.

[36] Fleischmann A, Darsow M, Degtyarenko K*, et al.* IntEnz, the integrated relational enzyme database. *Nucleic Acids Res* **2004**; 32: D434-7.

[37] Schomburg I, Chang A, Ebeling C*, et al.* BRENDA, the enzyme database: updates and major new developments. *Nucleic Acids Res* **2004**; 32: D431-3.

[38] Nagano N. EzCatDB: the enzyme catalytic-mechanism database. *Nucleic Acids Res* **2005**; 33: D407-12.

[39] Pegg SC, Brown SD, Ojha S*, et al.* Leveraging enzyme structure-function relationships for functional inference and experimental design: the structure-function linkage database. *Biochemistry* **2006**; 45: 2545-55.

[40] Holliday GL, Bartlett GJ, Almonacid DE*, et al.* MACiE: a database of enzyme reaction mechanisms. *Bioinformatics* **2005**; 21: 4315-6.

[41] Zhang T, Zhang H, Chen K, *et al.* Accurate sequence-based prediction of catalytic residues. *Bioinformatics* **2008**; 24: 2329-38.

[42] Altschul SF, Madden TL, Schaffer AA, *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **1997**; 25: 3389-402.

[43] Valdar WS. Scoring residue conservation. *Proteins* **2002**; 48: 227-41.

[44] Capra JA, Singh M. Predicting functionally important residues from sequence conservation. *Bioinformatics* **2007**; 23: 1875-82.

[45] Wang K, Horst JA, Cheng G, Nickle DC, Samudrala R. Protein meta-functional signatures from combining sequence, structure, evolution, and amino acid property information. *PLoS Comput Biol* **2008**; 4: e1000181.

[46] Greene LH, Higman VA. Uncovering network systems within protein structures. *J Mol Biol* **2003**; 334: 781-91.

[47] Li J, Wang J, Wang W. Identifying folding nucleus based on residue contact networks of proteins. *Proteins* **2008**; 71: 1899-907.

[48] David-Eden H, Mandel-Gutfreund Y. Revealing unique properties of the ribosome using a network based analysis. *Nucleic Acids Res* **2008**; 36: 4641-52.

[49] Cheng TM, Lu YE, Vendruscolo M, Lio P, Blundell TL. Prediction by graph theoretic measures of structural effects in proteins arising from non-synonymous single nucleotide polymorphisms. *PLoS Comput Biol* **2008**; 4: e1000135.

[50] Tseng YY, Liang J. Predicting enzyme functional surfaces and locating key residues automatically from structures. *Ann Biomed Eng* **2007**; 35: 1037-42.

[51] Ota M, Kinoshita K, Nishikawa K. Prediction of catalytic residues in enzymes based on known tertiary structure, stability profile, and sequence conservation. *J Mol Biol* **2003**; 327: 1053-64.

[52] Laskowski RA. SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J Mol Graph* **1995**; 13: 323-30.

[53] Binkowski TA, Naghibzadeh S, Liang J. CASTp: computed atlas of surface topography of proteins. *Nucleic Acids Res* **2003**; 31: 3352-5.

[54] McDonald IK, Thornton JM. Satisfying hydrogen bonding potential in proteins. *J Mol Biol* **1994**; 238: 777-93.

[55] Hubbard SJ, Thornton JM. NACCESS Computer program, Department of Biochemistry and Molecular Biology, University College of London, UK, **1993**.

[56] Yuan Z, Zhao J, Wang ZX. Flexibility analysis of enzyme active sites by crystallographic temperature factors. *Protein Eng* **2003**; 16: 109-14.

[57] Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol* **1999**; 292: 195-202.

[58] Carter P, Andersen CA, Rost B. DSSPcont: Continuous secondary structure assignments for proteins. *Nucleic Acids Res* **2003**; 31: 3293-5.

[59] Ondrechen MJ, Clifton JG, Ringe D. THEMATICS: a simple computational predictor of enzyme function from structure. *Proc Natl Acad Sci USA* **2001**; 98: 12473-8.

[60] Tong W, Williams RJ, Wei Y, *et al.* Enhanced performance in prediction of protein active sites with THEMATICS and support vector machines. *Protein Sci* **2008**; 17: 333-41.

[61] Ko J, Murga LF, Andre P, *et al.* Statistical criteria for the identification of protein active sites using Theoretical Microscopic Titration Curves. *Proteins* **2005**; 59: 183-95.

[62] Ko J, Murga LF, Wei Y, Ondrechen MJ. Prediction of active sites for protein structures from computed chemical properties. *Bioinformatics* **2005**; 21 (Suppl 1): i258-65.

[63] Pugalenthi G, Kumar KK, Suganthan PN, Gangal R. Identification of catalytic residues from protein structure using support vector machine with sequence and structural features. *Biochem Biophys Res Commun* **2008**; 367: 630-4.

[64] Mooney SD, Liang MH, DeConde R, Altman RB. Structural characterization of proteins using residue environments. *Proteins* **2005**; 61: 741-7.

[65] Fischer JD, Mayer CE, Soding J. Prediction of protein functional residues from sequence by probability density estimation. *Bioinformatics* **2008**; 24: 613-20.

[66] Tang YR, Chen YZ, Canchaya CA, Zhang Z. GANNPhos: a new phosphorylation site predictor based on a genetic algorithm integrated neural network. *Protein Eng Des Sel* **2007**; 20: 405-12.

[67] Chang CC, Lin CJ. LIBSVM: a library for support vector machines. Computer Program, Department of Computer Science, National Taiwan University, Taipei, Taiwan, **2001**.

[68] Centor RM. Signal detectability: the use of ROC curves and their analyses. *Med Decis Making* **1991**; 11: 102-6.

[69] Gribskov M, Robinson NL. Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching. *Comput Chem* **1996**; 20: 25-33.

[70] Bujnicki JM, Elofsson A, Fischer D, Rychlewski L. LiveBench-1: continuous benchmarking of protein structure prediction servers. *Protein Sci* **2001**; 10: 352-61.

[71] Koh IY, Eyrich VA, Marti-Renom MA, *et al.* EVA: evaluation of protein structure prediction servers. *Nucleic Acids Res* **2003**; 31: 3311-5.

[72] Chien TY, Chang DT, Chen CY, Weng YZ, Hsu CM. E1DS: catalytic site prediction based on 1D signatures of concurrent conservation. *Nucleic Acids Res* **2008**; 36: W291-6.

[73] Ginalski K, Grishin NV, Godzik A, Rychlewski L. Practical lessons from protein structure prediction. *Nucleic Acids Res* **2005**; 33: 1874-91.

[74] Sterner B, Singh R, Berger B. Predicting and annotating catalytic residues: an information theoretic approach. *J Comput Biol* **2007**; 14: 1058-73.

[75] Brylinski M, Prymula K, Jurkowski W, *et al.* Prediction of functional sites based on the fuzzy oil drop model. *PLoS Comput Biol* **2007**; 3: e94.

[76] Sacquin-Mora S, Laforet E, Lavery R. Locating the active sites of enzymes using mechanical properties. *Proteins* **2007**; 67: 350-9.

[77] Bagley SC, Altman RB. Characterizing the microenvironment surrounding protein sites. *Protein Sci* **1995**; 4: 622-35.

[78] Liang S, Zhang C, Liu S, Zhou Y. Protein binding site prediction using an empirical scoring function. *Nucleic Acids Res* **2006**; 34: 3698-707.

[79] Ofran Y, Rost B. ISIS: interaction sites identified from sequence. *Bioinformatics* **2007**; 23: e13-6.

[80] Ming D, Cohn JD, Wall ME. Fast dynamics perturbation analysis for prediction of protein functional sites. *BMC Struct Biol* **2008**; 8: 5.

Received: January 01, 2009 Revised: March 08, 2009 Accepted: June 14, 2009