Protein Engineering, Design & Selection vol. 21 no. 5 pp. 295–302, 2008
Published online February 20, 2008 doi:10.1093/protein/gzn003

An improved prediction of catalytic residues in enzyme structures

Yu-Rong Tang[1], Zhi-Ya Sheng[1,2], Yong-Zi Chen and Ziding Zhang[3]

Bioinformatics Center, College of Biological Sciences, China Agricultural University, Beijing 100094, China

[1]Both authors contributed equally to this work.

[2]Present address: National Institute of Biological Sciences, No. 7 Science Park Road, Beijing 102206, China

[3]To whom correspondence should be addressed. E-mail: zidingzhang@cau.edu.cn

The protein databases contain a huge number of function unknown proteins, including many proteins with newly determined 3D structures resulted from the Structural Genomics Projects. To accelerate experiment-based assignment of function, *de novo* prediction of protein functional sites, like active sites in enzymes, becomes increasingly important. Here, we attempted to improve the prediction of catalytic residues in enzyme structures by seeking and refining different encodings (i.e. residue properties) as well as employing new machine learning algorithms. In particular, considering that catalytic residues can often reveal specific network centrality when representing enzyme structure as a residue contact network, the corresponding measurement (i.e. closeness centrality) was used as one of the most important encodings in our new predictor. Meanwhile, a genetic algorithm integrated neural network (GANN) was also employed. Thanks to the above strategies, our GANN predictor demonstrated a high accuracy of 91.2% in the prediction of catalytic residues based on balanced datasets (i.e. the 1:1 ratio of catalytic to non-catalytic residues). When the GANN method was optimally applied to real enzyme structures, 73.9% of the tested structures had the active site correctly located. Compared with two existing methods, the proposed GANN method also demonstrated a better performance.

Keywords: catalytic residues/closeness centrality/genetic algorithm/neural network/prediction

Introduction

Providing functional annotation is one of the major tasks in the field of protein bioinformatics nowadays, given the considerable accumulation of protein sequence and structure data (Shapiro and Harris, 2000; Gutteridge *et al.*, 2003; Ofran *et al.*, 2005). For a query enzyme, the identification of catalytic residues is one of the most important steps towards understanding its biological roles and exploring its applications. In particular, the identified catalytic residues can greatly help in performing enzyme-targeted drug design, understanding the catalytic mechanism of enzyme reactions and constructing metabolic pathways (Bartlett *et al.*, 2002; Chou and Cai, 2004; Porter *et al.*, 2004).

Sequence and structural similarity based methods are two classical bioinformatics strategies 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. Subsequently, catalytic residues in an identified homolog can be transferred to the query sequence. However, in some cases such method can be misleading due to the fact that enzyme functions are less conserved (Todd *et al.*, 2001; Rost, 2002; Tian and Skolnick, 2003). The structural similarity based method is also able to identify catalytic residues even when no clear sequence similarity is detectable, provided that the 3D structure for the query enzyme is available (Orengo *et al.*, 1999). By mapping catalytic residues of a structural homolog into the query enzyme, such 'structure-based functional annotation' can offer in-depth insight by often highlighting 3D structural arrangements of catalytic residues. Even so, the power of structure-based annotation is often weakened by the fact that a similar fold does not necessarily imply a similar function (Nagano *et al.*, 2002).

It has been well accepted that proteins without detectable sequence or structural similarity may have the same configuration of active sites for catalyzing similar reactions (i.e. convergent evolution) (Torrance *et al.*, 2005; Zhang and Grigorov, 2006; Zhang and Tang, 2007). Complementary to sequence or structural similarity based methods, therefore, several methods focusing only on the local pattern of active sites and recognizing catalytic residues by comparing query structures with active site templates of known enzymes have been developed (Torrance *et al.*, 2005; Goyal *et al.*, 2007). With the accumulated enzyme structures deposited in the PDB database (Berman *et al.*, 2000), sequence and structural characters of catalytic residues have been intensively investigated (Bartlett *et al.*, 2002; Amitai *et al.*, 2004; Bate and Warwicker, 2004; Ben-Shimon and Eisenstein, 2005; del Sol *et al.*, 2006; Chea and Livesay, 2007). Meanwhile, *de novo* prediction methods (i.e. strategies independent of sequence alignment, structural comparison, or active-site matching) have also been developed to identify catalytic residues in enzyme structures. For example, some methods based on sequence or structural properties have been reported to achieve quite high accuracy (Chou and Cai, 2004; Ko *et al.*, 2005), although these methods have only been tested on a specific enzyme family or a small number of proteins.

With the advantage of incorporating different sequence or structural properties into a predictor, 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 heterogeneous enzymes (Gutteridge *et al.*, 2003; Petrova and Wu, 2006; Youn *et al.*, 2007). Compared with other machine learning based prediction tasks in the field of protein bioinformatics, this important topic is relatively less addressed and there is still enough room to improve.

In the present study, we focused our efforts to improve the prediction of catalytic residues based on the following two strategies. First, many available encoding schemes were evaluated to refine a subset of useful encodings. In particular, we transformed each enzyme structure into a residue interaction network (Greene and Higman, 2003), in which catalytic residues reveal specific closeness centrality (Amitai *et al.*, 2004; del Sol *et al.*, 2006). To our best knowledge, such information has not been incorporated in previously published machine learning based predictors. Secondly, in addition to SVM algorithm, a genetic algorithm integrated neural network (GANN) was also employed. The central idea of GANN is to use a genetic algorithm (GA) for optimizing the connection weights within neural networks. Compared with ANN trained with the standard back propagation algorithm, GANN generally can achieve a better performance in many applications (Cho, 1999; Fish *et al.*, 2004; Tang *et al.*, 2007). In our recent publication about the prediction of protein phosphorylation sites, GANN can even reveal a better performance than SVM (Tang *et al.*, 2007). In this paper, we report in detail about how the above two strategies are considered together to improve the prediction of catalytic residues.

Methods

Dataset

To facilitate a comparison of different predictors, the enzyme dataset originally compiled by Petrova and Wu (2006) was also used in the present study. Containing 79 protein domains, this dataset covered all 6 top level enzyme classifications (78 unique EC numbers) and 77 SCOP families. Since sequence redundancy was removed, there was no significant sequence similarity for any sequence pair within this dataset (Petrova and Wu, 2006). The corresponding PDB files for these 79 structures were retrieved from the SCOP database (Murzin *et al.*, 1995) (release 1.71, http://scop.mrc-lmb.cam.ac.uk/scop/), and active site annotation was from the Catalytic Site Atlas (Porter *et al.*, 2004) (http://www.ebi.ac.uk/thornton-srv/databases/CSA/), including 240 catalytic residues in all.

Encoding of residue properties

To construct a machine learning based catalytic residue predictor, residue properties must be converted into input feature vectors (i.e. encodings). Residue properties evaluated here covered residue type, sequence conservation, network centrality, relative position, hydrogen bonding, solvent accessibility, flexibility, and secondary structure. Properties represented by characters or strings including residue type, relative position, and secondary structure were converted into binary codes, while the rest real-number scores were directly used as the input of a predictor. More details about these encodings are described as follows.

Residue type Different amino acids evidently have different propensities to be catalytic residues (Bartlett *et al.*, 2002). Two encodings were used to represent this property. The first encoding is named *AA_Type20*, in which each of the 20 amino acids was encoded with 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), C (0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0),. . .,Y (0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1), etc. The second encoding called *AA_Type3* was based on a three-type classification of 20 amino acids, in which charged (DEKHR), polar (CNQSTY), and hydrophobic residues (AFGILMPVW) were encoded as (0 0), (0 1), and (1 0), respectively.

Sequence conservation One of the most important characteristics of catalytic residues is that they are highly conserved. Generally, they are more conserved not only than the average residues, but also than other functional residues, such as the ones involved in binding substrates (Bartlett *et al.*, 2002; Porter *et al.*, 2004). To compute the conservation score for a residue, a BLAST searching (Altschul *et al.*, 1997) for the corresponding sequence was performed against the NCBI non-redundant protein sequence database (the version of 09-03-2007) with a 10^{-5} *E*-value cut-off to obtain a multiple sequence alignment (MSA). The MSA was then submitted to the Scorecons server (Valdar, 2002) (http://www.ebi.ac.uk/thornton-srv/databases/cgi-bin/valdar/scorecons_server.pl) to score residue conservation with default parameters. Finally, the conservation score called *cons* was used as the sequence conservation based encoding. In some cases, the number of hits resulted from BLAST searching was more than 400. To accelerate the processing of the Scorecons server, Cd-hit (Li and Gozik, 2006) was used to filter these hits with an adjustable cut-off of sequence identity until the remained hits was less than 400. In other cases, the number of hits might be less than 10. To include enough sequences for a more reliable calculation of conservation, a three-iteration PSI-BLAST searching (Altschul *et al.*, 1997) was run with an *E*-value cut-off of 10^{-20} to include sequence in the position specific scoring matrix model.

Network centrality Residues in or directly contacting with active site usually have more interactions with other residues, so centrality values of catalytic residues in the network representation of enzyme structures are typically high, especially the closeness centrality (Amitai *et al.*, 2004; del Sol *et al.*, 2006; Chea and Livesay, 2007). To materialize these network centrality based encodings, each structure was transformed into an undirected residue interaction graph. Residues were modeled as vertices in the graph, and an edge was added between a pair of vertices if the shortest distance between any pair of atoms from two residues was no more than 5.0 Å. In other words, residues i and j were considered to have an edge if at least one atom from residue i was at a distance of $\leq$5.0 Å to an atom from residue j. With the established network, three encodings (*NCC_nw*, *NCC_ww*, and *NDC*) were employed to measure network centrality, which are briefly described as follows.

Firstly, the closeness centrality score CC_nw_i for any residue i within a network was calculated as

$$CC\_nw_i = \frac{n-1}{\sum_{i \neq j} d_{ij}} \quad (1)$$

where n was the total number of vertices in the graph and d_{ij} was the shortest path distance between vertices i and j, calculated using the Dijkstra algorithm (del Rio *et al.*, 2001).

The score was normalized over the entire structure as

$$NCC\_nw_i = \frac{CC\_nw_i - \overline{CC\_nw}}{\sigma(CC\_nw)} \quad (2)$$

where NCC_nw_i was the normalized closeness centrality score for residue i, $\overline{CC\_nw}$ was the average value of closeness centrality over all residues, and $\sigma(CC\_nw)$ was the standard deviation. In the above calculation, no weight was assigned for any edge within the network graph, i.e. d_{ij} was equal to the number of edges on the shortest path from vertex i to j. Therefore, the above *NCC_nw* encoding means the normalized closeness centrality score without weight. Meanwhile, we also weighted each edge by the shortest distance between the two corresponding residues to construct the *NCC_ww* encoding, i.e. the normalized closeness centrality score with weight, which was calculated using similar equations except for that d_{ij} was equal to the sum of weights on edges along the shortest path from vertex i to j.

Additionally, the *NDC* encoding, i.e. the normalized degree centrality score, was also derived. For each residue i, NDC_i was defined as

$$NDC_i = \frac{DC_i - \overline{DC}}{\sigma(DC)} \quad (3)$$

where DC_i was the degree centrality, defined as the number of edges connecting to vertex i, $\overline{DC}$ was the average value of degree centrality over all residues, and $\sigma(DC)$ was the corresponding standard deviation.

Relative position Active sites in almost all enzymes reside in clefts (Bartlett *et al.*, 2002; Tseng and Liang, 2007). Therefore, cleft environment was used here to present the relative position of a given residue. First, all clefts for a given structure were assigned by SURFNET (Laskowski, 1995). As described by Gutteridge *et al.* (Gutteridge *et al.*, 2003), the relative position of a residue was then divided into four categories according to the size of the cleft in which it located. Finally, the *Cleft* encoding for a residue was assigned, i.e. lying in the largest cleft (1 0 0 0), the second or third largest (0 1 0 0), the fourth to ninth largest (0 0 1 0) or none of the above clefts (0 0 0 1).

Hydrogen bonding Most catalytic residues act as donor or acceptor in at least one hydrogen bond. In particular, hydrogen bonds from main chain atoms to other residues in a protein are important in maintaining the conformation of these catalytic residues (Bartlett *et al.*, 2002). Hydrogen bonds were calculated using HBPLUS (McDonald and Thornton, 1994), and the following three parameters were used to represent this property. *NmHB* is the number of hydrogen bonds from a main-chain atom in a given residue to any other atom in a protein, *NsHB* denotes the number of hydrogen bonds from a side-chain atom in a given residue to any other atom in a protein, and *tNHB* indicates the total number of hydrogen bonds involving any atom in a given residue.

Relative solvent accessibility It has been well established that catalytic residues are generally more exposed to solvent than non-catalytic residues. Accordingly, we calculated relative solvent accessibility (RSA) for residues via NACCESS (Hubbard and Thornton, 1993), and five RSA based encodings were constructed: *AaRSA* means the RSA of all atoms; *TsRSA* is the RSA of all side chain atoms, including alpha carbons; *NpRSA* stands for the RSA of non-polar side chain atoms (i.e. all non-oxygens and non-nitrogens in the side chain); *ApRSA* is the RSA of all polar side chain atoms (i.e. all oxygen and nitrogen in the side chain); and *McRSA* is the RSA of all main chain atoms.

Structural flexibility Catalytic residues are often more rigid than average ones in an enzyme structure (Bartlett *et al.*, 2002; Yuan *et al.*, 2003). Here, two normalized B-factors based encodings (*NBf_RES* and *NBf_CA*) were calculated to measure residue flexibility. *NBf_RES* was the normalized B-factor of a residue, which was given by

$$NBf\_RES = \frac{B_{\mathrm{RES}} - \overline{B_{\mathrm{RES}}}}{\sigma(RES)} \quad (4)$$

where B_{RES} was the average B-factor over all atoms in a residue, $\overline{B_{\mathrm{RES}}}$ was the average B_{RES} over all residues, and σ(RES) was the corresponding standard deviation. *NBf_CA* was the normalized B-factor of C_α atom, which was defined as

$$NBf\_CA = \frac{B_{\mathrm{CA}} - \overline{B_{\mathrm{CA}}}}{\sigma(CA)} \quad (5)$$

where B_{CA} was the B-factor of C_α atom in a residue, $\overline{B_{\mathrm{CA}}}$ was the average value over C_α atoms from all residues, and σ(CA) was the corresponding standard deviation.

Secondary structure It is well known that catalytic residues are more inclined to locate in coil regions (Bartlett *et al.*, 2002). Therefore, secondary structure information may be helpful in catalytic residue prediction. DSSPcont (Carter *et al.*, 2003) was used to assign secondary structure state. The structural categories generated by DSSPcont include 3_{10}-helix (G), α-helix (H), π-helix (I), β-strand (E), isolated β-bridge (B), turn (T), and bend (S) and other. In this paper, these eight states were simplified to helix = {G, H, I}, sheet = {E, B}, and coil = {T, S and other}. For each residue, the *SS3* (Three-State Secondary Structure) based encoding was assigned, i.e. (0 0) for helix, (0 1) for sheet, and (1 0) for coil.

Training and testing

Testing based on balanced datasets To validate the performance based on different encodings as well as different machine learning methods, the ratio of positive instances (i.e. catalytic residues) to negative instances (non-catalytic residues) was initially set as 1:1. A 10-fold cross-validation was performed. Since the number of available non-catalytic residues is much larger than that of catalytic residues, five different negative sets were randomly selected to train and test a predictor for a reliable assessment.

First, an integrated SVM program named LIBSVM (Chang and Lin, 2001) was used for evaluating each encoding with default parameters. The applied kernel function here is the radial basis function. Secondly, the feature selection tool (Chang and Lin, 2001) based on LIBSVM was

employed to find the optimal subset of properties, which turned out to be eight encodings with a dimension of 30 in this work.

Moreover, we passed the best property subset to GANN, in which a GA was performed to optimize the connection weights of an ANN over the training dataset. The current GANN contains one input layer, one hidden layer, and one output layer. To obtain the optimized connection weights, a four-step genetic process was applied. First, an initial population of chromosomes is randomly created in the first generation. Each chromosome is used to encode a weight vector of the neural network. Secondly, a fitness value is assigned to each chromosome in the current generation. The fitness function (f) for GA is defined as the Matthews correlation coefficient (MCC). Thirdly, three operators (SELECTION, CROSSOVER, and MUTATION) are applied to the chromosomes of the current generation to obtain the new chromosomes of the next generation. In the fourth step, the above iterative training procedures are carried out to obtain the newer generation until fulfilling a terminal condition. At the end of training, the best chromosome with the highest fitness from the last generation is selected to create an ANN prediction model that can be used to perform a feed-forward computation to obtain the prediction output over a test dataset. After preliminary optimization, in this work parameters used in the GANN algorithm were set as follows: (i) the number of input nodes: 30; (ii) the number of hidden nodes: 5; (iii) the number of output nodes: 1; (iv) maximum generation number: 1420; (v) population size: 100; (vi) crossover probability: 0.95; (vii) initial mutation probability: 0.015; (viii) threshold of the fitness value: 0.9. For more details about the GANN algorithm and corresponding configurations, please refer to our recent publication related to GANN based protein phosphorylation site predictor (Tang *et al.*, 2007).

Prediction in entire structures Since there are much more non-catalytic than catalytic residues in real enzymes, the predictor trained with balanced datasets is not suitable. To construct a better predictor for the prediction in entire structures, the ratio must be optimized. In this test, 79 enzymes were divided into 10 roughly equal groups by structure, i.e. each group contained seven or eight intact structures. Again, a 10-fold cross-validation was performed. For every testing group, we constructed five training sets by varying the non-catalytic residues. Performance was averaged over results based on all the tests.

Using LIBSVM based on the selected eight properties, it was found that the best ratio of catalytic to non-catalytic residues in the training sets was 1:6, which is consistent with Gutteridge *et al.*'s work (Gutteridge *et al.*, 2003). However, at such a relatively large proportion of non-catalytic residues, some structural properties might bring in more noise than useful information, so we tried to re-optimize the selected subset by removing that kind of properties. Since the dimensionality of the input feature vector changed, the number of hidden nodes and the terminal condition in the GANN algorithm were also re-optimized.

Performance measure

Four measurements, i.e. accuracy (AC), true positive rate (TPR), false positive rate (FPR), and MCC, were used to evaluate the prediction performance with definition as follows:

$$\mathrm{AC} = \frac{\mathrm{tp} + \mathrm{tn}}{\mathrm{tp} + \mathrm{fn} + \mathrm{tn} + \mathrm{fp}} \tag{6}$$

$$\mathrm{TPR} = \frac{\mathrm{tp}}{\mathrm{tp} + \mathrm{fn}} \tag{7}$$

$$\mathrm{FPR} = \frac{\mathrm{fp}}{\mathrm{fp} + \mathrm{tn}} \tag{8}$$

$$\mathrm{MCC} = \frac{\mathrm{tp} \times \mathrm{tn} - \mathrm{fp} \times \mathrm{fn}}{\sqrt{(\mathrm{tp} + \mathrm{fp}) \times (\mathrm{tp} + \mathrm{fn}) \times (\mathrm{tn} + \mathrm{fn}) \times (\mathrm{tn} + \mathrm{fp})}}. \tag{9}$$

where tp, fp, fn, and tn denote true positives, false positives, false negatives, and true negatives. When the numbers of positive and negative data are different, MCC should be more suitable for assessing the overall prediction accuracy. The value of MCC ranges from -1 to 1, and higher MCC means better prediction performance.

When a prediction is performed on an entire structure, it is also important to know if the active site can be correctly identified. Based on the predicted catalytic residues, the following procedures described in Gutteridge *et al.*'s paper (Gutteridge *et al.*, 2003) were employed to locate the predicted active sites. First, two predicted catalytic residues were clustered together if the shortest distance between them was no more than 4.0 Å, and each cluster was represented by a sphere whose centre was the geometric centroid of C_β atoms of all component residues (C_α atom in glycine) and radius was equal to the distance from the farthest C_β atom to the centre. Secondly, single residues were added to an existing cluster if this would not increase its radius to over 20.0 Å. Thus, several clusters could be constructed and each cluster was considered as a predicted active site. Known active sites were also defined as spheres as above, and a radius of 3.0 Å was assigned for a single-residue site. For each enzyme structure, a correct active site prediction means the overlap between a predicted active site and the corresponding known site is greater than 50% of the volume of the known one; a partially correct prediction means the overlap is less than 50%; and an incorrect prediction means no overlap at all.

Results and discussion

Results based on balanced datasets

In this work, 18 residue property based encodings were individually evaluated with the assistance of LIBSVM program. As shown in Fig. 1A, sequence conservation based encoding (i.e. *cons*) remained the most informative encoding, which is in line with previous studies (Gutteridge *et al.*, 2003; Petrova and Wu, 2006). Interestingly, closeness centrality based encoding (*NCC_ww* and *NCC_nw*) appeared to be the second discriminative feature. In addition, performance based on closeness centrality seemed relatively steady over different datasets (cf. Fig. 1A). Comparatively, *NCC_ww* was a little bit more powerful than *NCC_nw* (cf. Fig. 1A), probably due to that *NCC_ww* could consider the intensity of interaction between residues to some extent. Also in accordance with previous study (Amitai *et al.*, 2004), *NDC* encoding was not useful. A plausible reason is that degree

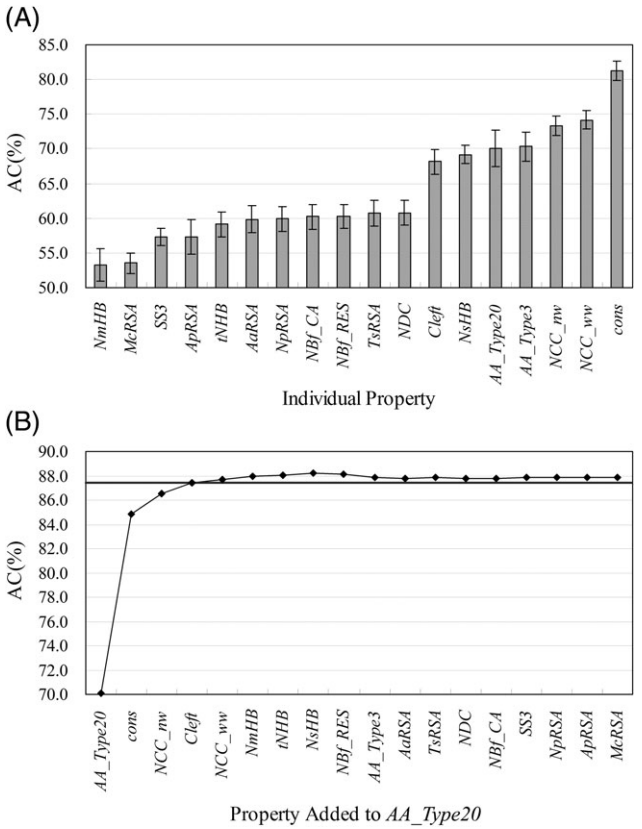

Fig. 1. Property evaluation and selection using LIBSVM. (**A**) Accuracy based on individual property. Error bars indicated standard deviations. (**B**) The prediction accuracy when other properties were added to *AA_type20* step by step. The bold solid line indicated the performance of Petrova and Wu's method (Petrova and Wu, 2006).

centrality only described the local environment around a residue, while closeness centrality was inclined to characterize a residue by its relationship with all residues in the structure, which was more helpful to decide what role this residue played in the entire enzyme.

Moreover, the feature selection tool based on LIBSVM was employed to select the optimized subset of properties. As shown in Fig. 1B, eight encodings with a dimension of 30 jointly contributed to an optimal performance in predicting catalytic residues, which fell into five categories, i.e. residue type, sequence conservation, network centrality, relative position, and hydrogen bonding. It is interesting to mention that *NCC_nw* contributed more than *NCC_ww* in the optimized subset of residue properties (cf. Fig. 1B), which might be due to the fact that *NCC_ww* is affected by the size of side chains in different residues that enlarges its overlap with *AA_Type20*. The three hydrogen-bonding-based encodings (*NmHB*, *NsHB*, and *tNHB*) were not so powerful when testing alone, but they helped when added to the first five encodings (cf. Fig. 1B), owing to the fact that hydrogen bonding presents another aspect of residue property, i.e. conformational freedom, which is not covered in the first five encodings. Meanwhile, other properties could not help much when added, because to some extent, they may have overlap with encodings in the optimal subset, e.g. *AA_Type3* with *AA_Type20*, RSA with *Cleft*, structural flexibility with hydrogen bonding, etc. Detailed analyses of these optimal properties were illustrated in Fig. 2. It was suggested that catalytic and non-catalytic residues did differ in these characteristics. In particular, it was clear that catalytic residues tended to have both high conservation scores and high closeness centrality values (cf. Fig. 2C).

The same datasets and inputs were used to train and test GANN based algorithm. It turned out that GANN can achieve a better performance than LIBSVM (cf. Table I). Compared with LIBSVM, the average accuracy was increased by 3.0%. This might indicate that GANN was more applicable for this kind of data than SVM.

Results of prediction in entire structures

As a matter of fact, the ratio of catalytic to non-catalytic residues is quite different from 1:1 in real enzymes. Therefore, our method was also tested on entire enzyme structures. To conduct such a prediction, the ratio of catalytic to non-catalytic residues in the training datasets was optimally set as 1:6, and the testing was performed against entire enzymes. Based on the same datasets, performance of LIBSVM and GANN based algorithms was compared in Table II. On the whole, GANN still showed its advantage over the LIBSVM based method. With only three properties (*AA_Type20*, *cons*, and *NCC_nw*), GANN could achieve an MCC of 0.364, while LIBSVM needed seven attributes (*AA_Type20*, *cons*, *NCC_nw*, *Cleft*, *NCC_ww*, *NmHB*, and *tNHB*) to reach an MCC of 0.342. In addition, the most noteworthy superiority was that GANN was much more sensitive when handling data with such a large portion of negative instances. In comparison to LIBSVM, GANN could increase TPR from 57.8 to 73.2%, without notable increase in FPR (only from 2.6 to 3.8%) (cf. Table II).

We also tried to locate active sites in enzyme structures according to the predicted catalytic residues. Spheres containing clusters of predicted catalytic residues were used to represent predicted active sites as described in the method section. As shown in Table III, based on the prediction of GANN, 73.9% of the enzymes had the active site correctly located, thanks to GANN's increased sensitivity, and in another 20.9% the locating was partially correct. Actually, predicted sites often lay close to the known active sites, as only 5.2% of the tested enzymes had no predicted active sites overlapping with the known ones.

To intuitively show the difference resulted from different ratios of positive and negative data in training datasets, the catalytic residue prediction of an enzyme structure (i.e. aspartylglucosaminidase, PDB entry: 1apy) was exemplified. As shown in Fig. 3, when using a balanced training set (1:1), all catalytic residues could be successfully identified, but false positive rate was also quite high. When using a 1:6 training set, much fewer non-catalytic residues were incorrectly predicted as catalytic yet true positive rate fell significantly at the same time. Actually all residues would be predicted as non-catalytic when the proportion of negative instances kept growing in the training sets. Compared with the prediction based on a 1:1 training dataset, it is interesting to mention that most of the false positives located close to catalytic residues, which indicated that they were quite likely to be involved in the binding of substrates or the stabilization of products. Due to the relationship between false and true positives, location of the active sites in most enzymes including 1apy can be correctly detected, although the identification of catalytic residues were not as precise as in the 1:1

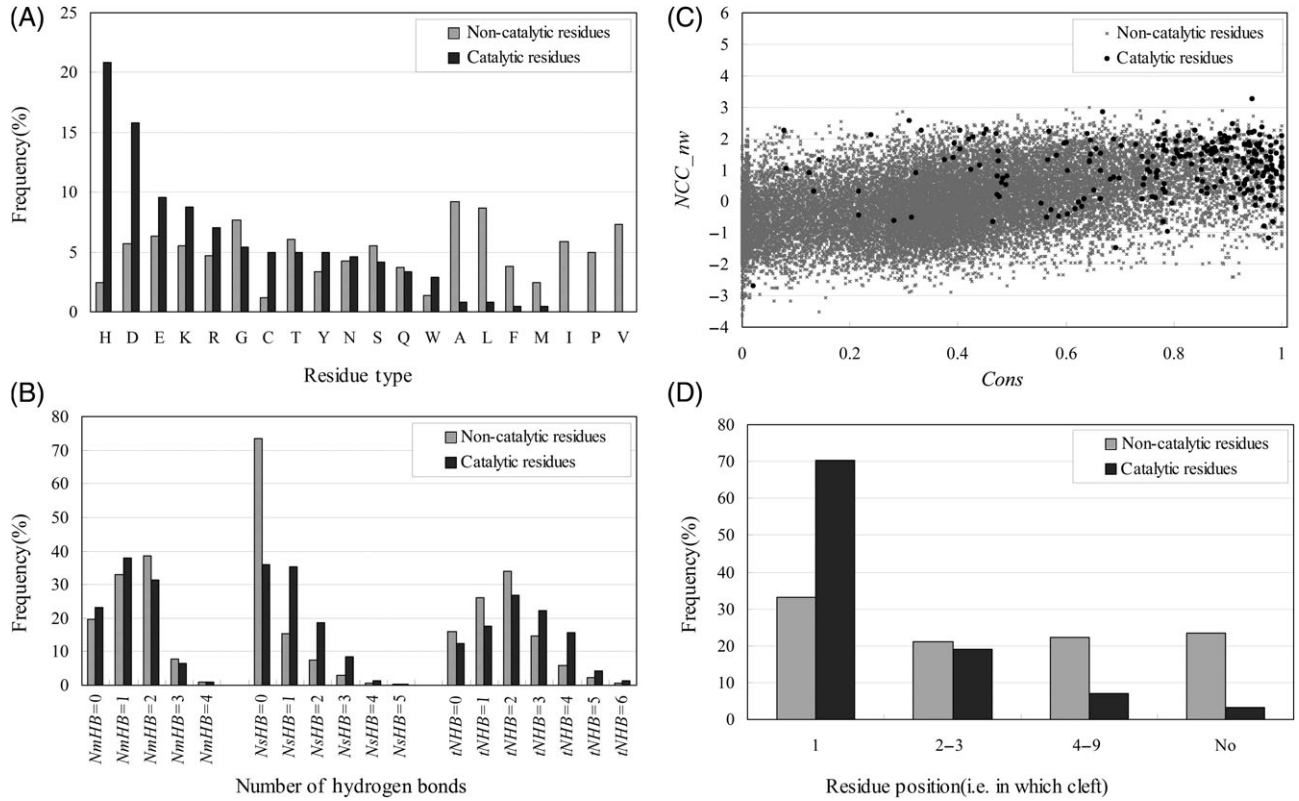

Fig. 2. Analyses of residue properties in the optimal subset. (**A**) Frequency distribution of 20 amino acids in catalytic and non-catalytic residues; (**B**) number of hydrogen bonds in catalytic and non-catalytic residues; (**C**) residue conservation and closeness centrality properties; (**D**) relative position of catalytic and non-catalytic residues.

Table I. Performance of different algorithms based on the balanced training datasets

Algorithms	AC (%)	TPR (%)	FPR (%)	MCC
LIBSVM[a]	88.2 ± 1.3	89.9 ± 0.8	13.4 ± 2.4	0.769 ± 0.024
GANN[a]	91.2 ± 1.2	93.0 ± 1.1	10.6 ± 1.9	0.827 ± 0.023
Petrova and Wu (Petrova and Wu, 2006)	87.4	89	14	0.75

[a]The corresponding measurement was represented as the average value ± standard deviation.

Table II. Performance of different algorithms based on the 1:6 training datasets[a]

Algorithms	TPR (%)	FPR (%)	MCC
LIBSVM[b]	57.8 ± 1.4	2.6 ± 0.2	0.342 ± 0.012
GANN[b]	73.2 ± 2.0	3.8 ± 0.4	0.364 ± 0.008
Gutteridge *et al.*, before clustering (Gutteridge *et al.*, 2003)	56	3.4[c]	0.28
Gutteridge *et al.*, after clustering (Gutteridge *et al.*, 2003)	68	3.6[c]	0.32

[a]All tests in this table were performed against entire enzymes. [b]The corresponding measurement was represented as the average value ± standard deviation. [c]The FPR value was estimated based on the other measures of performance reported by Gutteridge *et al.* (2003).

model. However, how to discriminate catalytic residues and their structural neighbouring residues remains a challenge to improve the accuracy of predicting catalytic residues in enzyme structures.

Comparison of the proposed method with two existing methods

Using the same 79 enzyme structures previously used in Petrova and Wu's method allowed a fair comparison between the performance of their method and ours. Benefited from the network closeness centrality based encoding, only four properties (*AA_Type20*, *cons*, *NCC_nw*, and *Cleft*) were able to achieve the same accuracy as Petrova and Wu's method (cf. Fig. 1B). Furthermore, the LIBSVM algorithm based on the optimal properties slightly surpassed Petrova and Wu's method (cf. Table I). Further empowered by a new machine learning algorithm, the GANN based prediction can result in an even higher accuracy (about +4.0%) (cf. Table I). Considering that only a small fraction of residues in an enzyme structure are catalytic residues, the different choice of negative dataset may have significant impact on the reported accuracy. Either in this study or Petrova and Wu's paper, the prediction accuracy was averaged over several 10-fold cross-validation tests by changing negative datasets (i.e. non-catalytic residues). Thus, the comparison between these two methods should be reliable.

When testing the prediction in entire structures, Petrova and Wu's method did not optimize the corresponding ratio of catalytic to non-catalytic residues in the training dataset, and only achieved an MCC of 0.23. We then compared our method with Gutteridge *et al.* As shown in Table II, our GANN method increased MCC to 0.364 (Gutteridge *et al.*: 0.28 before clustering, 0.32 after clustering). Considering the

Table III. Performance of different algorithms in locating active sites based on the 1:6 training datasets[a]

Algorithms	Correct (%)	Partial (%)	Incorrect (%)
LIBSVM[b]	63.1 ± 1.9	29.4 ± 1.1	7.5 ± 1.3
GANN[b]	73.9 ± 7.1	20.9 ± 7.2	5.2 ± 1.1
Gutteridge *et al.* (Gutteridge *et al.*, 2003)	69.2	24.5	6.3

[a]All tests in this table were performed against entire enzymes. [b]The corresponding measurement was represented as the average value ± standard deviation.

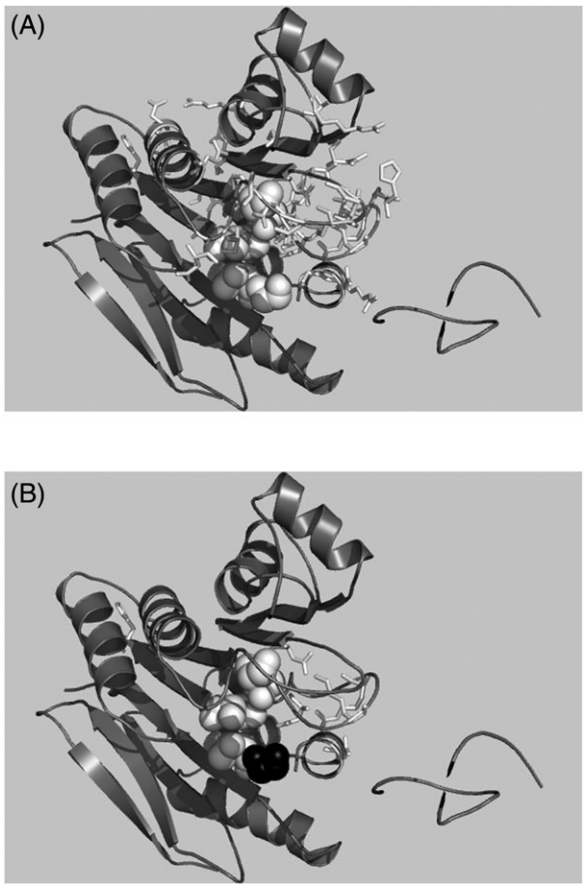

Fig. 3. Predicted catalytic residues in aspartylglucosaminidase (PDB entry: 1apy). (**A**) Prediction based on a 1:1 training dataset; (**B**) prediction based on a 1:6 training dataset. White spheres indicated true positives, black spheres indicated false negatives, and false positives were shown by their side chains in white sticks.

correct location of active sites, our GANN also demonstrated a nearly 5.0% higher accuracy (cf. Table III). Although both the datasets used in Gutteridge *et al.*'s method and ours were extracted from the Catalytic Site Atlas (Porter *et al.*, 2004), noted that the data set used in ours is smaller but more stringent since the redundancy had been removed as reported by Petrova and Wu (2006). Thus, such a comparison is generally reasonable.

Future perspective

By using network closeness centrality as one of the key input features as well as adopting the GANN algorithm, not only a high accuracy was achieved in catalytic residue identification in the balanced model, but also most active sites in real enzymes were successfully located. One immediate application is to combine the current algorithm with active site templates based searching method for a more reliable active site prediction. Therefore, the current algorithm can be useful in the functional annotation of newly determined protein structures from the Structural Genomics Projects (Brenner, 2001).

In spite of the improvement indicated above, the MCC value of our method remained below 0.4 in the identification of catalytic residues in entire structures, suggesting that the current algorithm alone was still not good enough for practical use. To improve the identification of catalytic residues, filtering out some non-catalytic residues before prediction should be helpful. For instance, residues located in the functional surface can be computationally identified first, which have been materialized in several algorithms (Tseng *et al.*, 2007). Thus, difference between catalytic and non-catalytic residues may be even more obvious due to the reduction of noise. Exploring new properties (encodings) also leads to an important direction to develop a better predictor. In this study, closeness centrality plays a more important part than any other structural feature. To some extent, this is due to the fact that closeness centrality characterizes the relationship between a given residue and all other residues in a protein structure, which helped to decide its role in the entire enzyme when catalyzing a reaction. To detect functional sites within a protein, efforts have been increasingly paid on finding some new sequence or structural properties (e.g. see Refs. Bagley and Altman, 1995; Bate and Warwicker, 2004; Liang *et al.*, 2006; Ofran and Rost, 2007), which may further be validated for their suitability in predicting catalytic residues. We expect that newly identified properties will not only improve the accuracy in predicting catalytic residues, but also strengthen our basic understanding in molecular mechanisms of enzymatic reaction.

Funding

The National Natural Science Foundation of China (30700137).

Acknowledgements

The authors extend their gratitude to Dr James Torrance for providing the fully annotated version of the Catalytic Site Atlas. This research was supported by the National Natural Science Foundation of China (30700137).

References

Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) *Nucleic Acids Res.*, **25**, 3389–3402.
Amitai,G., Shemesh,A., Sitbon,E., Shklar,M., Netanely,D., Venger,I. and Pietrokovski,S. (2004) *J. Mol. Biol.*, **344**, 1135–1146.
Bagley,S.C. and Altman,R.B. (1995) *Protein Sci.*, **4**, 622–635.
Bartlett,G.J., Porter,C.T., Borkakoti,N. and Thornton,J.M. (2002) *J. Mol. Biol.*, **324**, 105–121.
Bate,P. and Warwicker,J. (2004) *J. Mol. Biol.*, **340**, 263–276.
Ben-Shimon,A. and Eisenstein,M. (2005) *J. Mol. Biol.*, **351**, 309–326.
Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) *Nucleic Acids Res.*, **28**, 235–242.
Brenner,S. (2001) *Nature Rev. Genet.*, **2**, 801–809.
Carter,P., Andersen,C.A. and Rost,B. (2003) *Nucleic Acids Res.*, **31**, 3293–3295.

Chang,C.C. and Lin,C.J. (2001) *Computer Program*. Department of Computer Science, National Taiwan University, Taipei, Taiwan.
Chea,E. and Livesay,D.R. (2007) *BMC Bioinformatics*, **8**, 153.
Cho,S.B. (1999) *Fuzzy Sets Syst.*, **103**, 339–347.
Chou,K.C. and Cai,Y.D. (2004) *Proteins*, **55**, 77–82.
del Rio,G., Bartley,T.F., del-Rio,H., Rao,R., Jin,K.L., Greenberg,D.A., Eshoo,M. and Bredesen,D.E. (2001) *FEBS Lett.*, **509**, 230–234.
del Sol,A., Fujihashi,H., Amoros,D. and Nussinov,R. (2006) *Protein Sci.*, **15**, 2120–2128.
Fish,K.E., Johnson,J.D., Dorsey,R.E. and Blodgett,J.G. (2004) *J. Business Res.*, **57**, 79–85.
Goyal,K., Mohanty,D. and Mande,S.C. (2007) *Nucleic Acids Res.*, **35(Web Server issue)**, W503–W505.
Greene,L.H. and Higman,V.A. (2003) *J. Mol. Biol.*, **334**, 781–791.
Gutteridge,A., Bartlett,G.J. and Thornton,J.M. (2003) *J. Mol. Biol.*, **330**, 719–734.
Hubbard,S.J. and Thornton,J.M. (1993) *Computer Program*. Department of Biochemistry and Molecular Biology, University College, London.
Ko,J., Murga,L.F., Wei,Y. and Ondrechen,M.J. (2005) *Bioinformatics*, **21**(Suppl. 1), 258–265.
Laskowski,R.A. (1995) *J. Mol. Graph.*, **13**, 323–330.
Li,W. and Godzik,A. (2006) *Bioinformatics*, **22**, 1658–1659.
Liang,S., Zhang,C., Liu,S. and Zhou,Y. (2006) *Nucleic Acids Res.*, **34**, 3698–3707.
McDonald,I.K. and Thornton,J.M. (1994) *J. Mol. Biol.*, **238**, 777–793.
Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) *J. Mol. Biol.*, **247**, 536–540.
Nagano,N., Orengo,C.A. and Thornton,J.M. (2002) *J. Mol. Biol.*, **321**, 741–765.
Ofran,Y. and Rost,B. (2007) *Bioinformatics*, **23**, e13–e16.
Ofran,Y., Punta,M., Schneider,R. and Rost,B. (2005) *Drug Discov. Today*, **10**, 1475–1482.
Orengo,C.A., Todd,A.E. and Thornton,J.M. (1999) *Curr. Opin. Struct. Biol.*, **9**, 374–382.
Petrova,N.V. and Wu,C.H. (2006) *BMC Bioinformatics*, **7**, 312.
Porter,C.T., Bartlett,G.J. and Thornton,J.M. (2004) *Nucleic Acids Res.*, **32**, D129–D133.
Rost,B. (2002) *J. Mol. Biol.*, **318**, 595–608.
Shapiro,L. and Harris,T. (2000) *Curr. Opin. Biotechnol.*, **11**, 31–35.
Tang,Y.R., Chen,Y.Z., Canchaya,C. and Zhang,Z. (2007) *Protein Eng. Des. Sel.*, **20**, 405–412.
Tian,W. and Skolnick,J. (2003) *J. Mol. Biol.*, **333**, 863–882.
Todd,A.E., Orengo,C.A. and Thornton,J.M. (2001) *J. Mol. Biol.*, **307**, 1113–1143.
Torrance,J.W., Bartlett,G.J., Porter,C.T. and Thornton,J.M. (2005) *J. Mol. Biol.*, **347**, 565–581.
Tseng,Y.Y. and Liang,J. (2007) *Ann. Biomed. Eng.*, **35**, 1037–1042.
Valdar,W.S. (2002) *Proteins*, **48**, 227–241.
Youn,E., Peters,B., Radivojac,P. and Mooney,S.D. (2007) *Protein Sci.*, **16**, 216–226.
Yuan,Z., Zhao,J. and Wang,Z.X. (2003) *Protein Eng.*, **16**, 109–114.
Zhang,Z. and Grigorov,M. (2006) *Proteins*, **62**, 470–478.
Zhang,Z. and Tang,Y.R. (2007) *Protein Pept. Lett.*, **14**, 291–297.

Received October 15, 2007; revised January 4, 2008; accepted January 4, 2008

Edited by Valerie Daggett