Identification of Catalytic Residues Using a Novel Feature that Integrates the Microenvironment and Geometrical Location Properties of Residues

Lei Han[1], Yong-Jun Zhang[2], Jiangning Song[3,4], Ming S. Liu[5]*, Ziding Zhang[1]*

1 State Key Laboratory of Agrobiotechnology, College of Biological Sciences, China Agricultural University, Beijing, People's Republic of China, **2** State Key Laboratory for Biology of Plant Diseases and Insect Pests, Institute of Plant Protection, Chinese Academy of Agricultural Sciences, Beijing, People's Republic of China, **3** National Engineering Laboratory for Industrial Enzymes and Key Laboratory of Systems Microbial Biotechnology, Tianjin Institute of Industrial Biotechnology, Chinese Academy of Sciences, Tianjin, People's Republic of China, **4** Department of Biochemistry and Molecular Biology, Faculty of Medicine, Monash University, Melbourne, Victoria, Australia, **5** CSIRO - Mathematics, Informatics and Statistics, Clayton, Victoria, Australia

Abstract

Enzymes play a fundamental role in almost all biological processes and identification of catalytic residues is a crucial step for deciphering the biological functions and understanding the underlying catalytic mechanisms. In this work, we developed a novel structural feature called MEDscore to identify catalytic residues, which integrated the microenvironment (ME) and geometrical properties of amino acid residues. Firstly, we converted a residue's ME into a series of spatially neighboring residue pairs, whose likelihood of being located in a catalytic ME was deduced from a benchmark enzyme dataset. We then calculated an ME-based score, termed as MEscore, by summing up the likelihood of all residue pairs. Secondly, we defined a parameter called Dscore to measure the relative distance of a residue to the center of the protein, provided that catalytic residues are typically located in the center of the protein structure. Finally, we defined the MEDscore feature based on an effective nonlinear integration of MEscore and Dscore. When evaluated on a well-prepared benchmark dataset using five-fold cross-validation tests, MEDscore achieved a robust performance in identifying catalytic residues with an AUC1.0 of 0.889. At a $\leq$10% false positive rate control, MEDscore correctly identified approximately 70% of the catalytic residues. Remarkably, MEDscore achieved a competitive performance compared with the residue conservation score (e.g. CONscore), the most informative singular feature predominantly employed to identify catalytic residues. To the best of our knowledge, MEDscore is the first singular structural feature exhibiting such an advantage. More importantly, we found that MEDscore is complementary with CONscore and a significantly improved performance can be achieved by combining CONscore with MEDscore in a linear manner. As an implementation of this work, MEDscore has been made freely accessible at http://protein.cau. edu.cn/mepi/.

Citation: Han L, Zhang Y-J, Song J, Liu MS, Zhang Z (2012) Identification of Catalytic Residues Using a Novel Feature that Integrates the Microenvironment and Geometrical Location Properties of Residues. PLoS ONE 7(7): e41370. doi:10.1371/journal.pone.0041370

Editor: Morten Nielsen, Technical University of Denmark, Denmark

Received March 22, 2012; **Accepted** June 20, 2012; **Published** July 19, 2012

Funding: This work was supported by grants from the National Key Basic Research Program of China (2009CB918802 and 2012CB1141004), the National Natural Science Foundation of China (30700137), the National Health and Medical Research Council of Australia (NHMRC), the Hundred Talents Program of the Chinese Academy of Sciences (CAS), the Knowledge Innovative Program of CAS (KSCX2-EW-G-8) and Tianjin Municipal Science and Technology Commission (10ZCKFSY05600). JS is an NHMRC Peter Doherty Fellow and the Recipient of the Hundred Talents Program of CAS. MSL thanks China Agricultural University for hosting his visits and acknowledges the support from CSIRO- Advanced Materials TCP and Computational and Simulation Sciences-TCP. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: ming.liu@csiro.au (MSL); zidingzhang@cau.edu.cn (ZZ)

Introduction

Enzymes play a fundamental role in fulfilling diverse biochemical functions and are essentially required for almost all cellular processes. Although the catalytic mechanisms of certain enzymes have been well characterized [1], it remains a difficult and challenging task to rationalize the sequence-structure-function relationship and unravel the biological functions of the majority of enzymes. Owing to structural genomics efforts [2,3], a considerable number of protein structures have been determined. Solving the three-dimensional structure of an enzyme can further deepen our understanding of its catalytic mechanism at the atomic level. However, it is still a challenging task to establish the linkage between the given protein structures and their catalytic mechanisms, reflected by the vast number of functionally uncharacterized enzyme structures generated from the structural genomics projects [4]. As catalytic residues are directly involved in catalytic processes, their identification is the first crucial step to characterize the catalytic mechanism and function of an enzyme. Since experimental determination of catalytic residues from large-scale proteome data is a costly and daunting task, computational methods that are capable of identifying catalytic residues from enzyme sequence and/or structure information play an increasingly important role in complementing the experimental efforts and supporting the functional annotation. Apart from providing critical insights into the rules that govern enzymatic catalysis, the identification of catalytic residues also has important applications in the areas of drug design [5], protein engineering, metabolic pathway analysis and synthetic biology [6].

In the past few decades, intensive efforts have been dedicated to identifying catalytic residues in proteins and many features or parameters have been exploited to characterize the properties of catalytic residues. These features can be generally divided into two categories: sequence- and structure-based. Amino acid residues have different propensities to be catalytic residues in nature. For example, it was previously observed that roughly 65% of catalytic residues were charged (H, R, K, E, D), 27% were polar (Q, T, S, N, C, Y, W), and 8% were hydrophobic (G, F, L, M, A, I, P, V) [7]. Therefore, amino acid residue type is probably the simplest but one of the most efficiently used sequence features to identify catalytic residues. In addition, residue conservation, derived from the multiple sequence alignment (MSA) of a query sequence, has also proved to be one of the most powerful singular features in predicting catalytic residues [8–13]. The state-of-the-art residue conservation algorithms include the Shannon entropy-based method [14], Jensen-Shannon divergence method [15], Rate4site algorithm [16] and other methods [17–20]. More recently, researchers have found that co-evolutionary features could be commonly derived from the neighboring residues surrounding functionally important sites [21] and such information could be utilized to facilitate the identification of catalytic residues [21–23].

Given that enzymes perform their biological functions on the basis of specific three-dimensional structures, a variety of simple structural features have been proposed to characterize catalytic residues. For example, it has been shown that catalytic residues prefer to be located in the geometric centers of the protein structures [24]. They also tend to be located in a large cleft on the protein structure surface [7,25]. Therefore, the distance of the cleft to the center of protein structure can provide quantitative information for identifying catalytic residues [26]. In addition, as most catalytic residues act as either acceptors or donors in the catalytic process, the hydrogen bonds in protein structures can also be used to discriminate catalytic from non-catalytic residues [8,27]. Other important structural and dynamic properties, such as solvent accessibility [7,27], flexibility of loop regions [7,28] and B-factors [7,29] have also been used as features or descriptors for predicting catalytic residues.

Recently, more complicated structure-based features have been developed to distinguish catalytic from non-catalytic residues. It has been established in protein engineering that mutations of the active site residues usually lead to an increased stability. Therefore, the properties that describe the destabilizing effects of residues were employed to identify catalytic residues [25,30]. Likewise, since the electrostatic property is important for an enzyme to maintain its function, the electrostatic property-based features derived from the titration curves of residues [31,32] and the electrostatic energy of residues [33] have proved useful in predicting catalytic residues. Bryliński et al. observed that the regions with significant irregular hydrophobicity in enzyme structures tend to be functionally important and thus developed a novel feature based on the Fuzzy Oil Drop model to predict active sites [34]. Sacquin-Mora et al. proposed a force constant-based feature to quantify the easiness of moving a given residue relative to the rest in a protein based on the fact that catalytic residues are generally more rigid than others. They further employed it as an informative feature to detect catalytic residues [35]. In summary, most of these structural features are developed based on physicochemical properties of amino acid residues which typically require intensive dynamics and/or energy calculations. This has greatly limited their high-throughput applications.

New features based on improved representations of protein structures have been proposed in recent years. For example, a protein structure can be represented as a residue interaction network where each residue is represented as a node and two interacting residues are connected by an edge [36]. A network parameter, i.e. the Closeness centrality (also called Closeness), has been demonstrated to be an informative feature in detecting catalytic residues [37–39]. The concept of microenvironment (ME) has been previously proposed to describe a residue's local structural environment [40], extracted from the physical and chemical properties of the residue and its structurally neighboring residues at the residue/atom level. The ME-related features have been widely used to recognize catalytic residues in protein structures [41–46].

To further improve the prediction performance, some features have been integrated into different predictors using either statistical (e.g. logistic regression [30,47] and maximum likelihood models [48]) or machine learning methods (e.g. support vector machines [8,43,44,49,50] and neural networks [10,27,51]). In the past few years, we have witnessed the flourish of such integrative predictors [52,53]. In summary, statistical methods can yield an improved performance based on an efficient integration of individual (which are largely independent) descriptors to simple and interpretable models. In comparison, although machine learning methods can usually lead to a more competitive performance through the use of much larger feature sets, they have certain disadvantages. For example, they are often criticized for lacking biological interpretation of the trained 'black box' models and thereby difficult for biologists to readily deploy and understand the predictions of such models.

In this study, we developed a novel structural feature to identify catalytic residues by integrating the ME and geometrical properties of residues. Firstly, we converted the ME of a residue into a series of spatially neighboring residue pairs, whose propensities to occur in the catalytic ME were deduced from a pre-built enzyme dataset. Then, we proposed a new feature called MEscore to characterize the ME of a residue. To the best of our knowledge, this is the first endeavor to represent the ME of a residue using a group of residue pairs. We then proposed and validated another feature called Dscore that quantifies the centrality of a residue to the whole protein structure. As MEscore and Dscore are largely complementary to each other, we further integrated these two features to a novel feature named MEDscore. Remarkably, MEDscore revealed a competitive performance compared with the residue conservation score. In this work, we describe and discuss the construction of MEscore and MEDscore as well as the overall performance assessment of different features in detail. Specially, the fundamental mechanism by which MEscore and MEDscore are informative for catalytic residue recognition is also discussed.

Results and Discussion

Propensities of residues in the microenvironment (ME) surrounding the catalytic residues

We systematically analyzed the amino acid compositions of catalytic residues and their spatially neighboring residues based on a well-prepared enzyme dataset consisting of 223 catalytic domains. More details about this enzyme dataset can be found in the 'Materials and Methods' Section. As shown in Figure 1, catalytic residues tend to be either charged or polar residues, which is in accordance with previous studies [7]. Interestingly, we also found that the neighboring residues in the ME surrounding the catalytic residues exhibit remarkably different propensities: some residues (C, M, H, S, T, W, Y, F, G) prefer to be located in the neighborhood of catalytic residues, while others (E, K, R, D, A, L, P, V, Q) are seldom observed to be around catalytic residues.

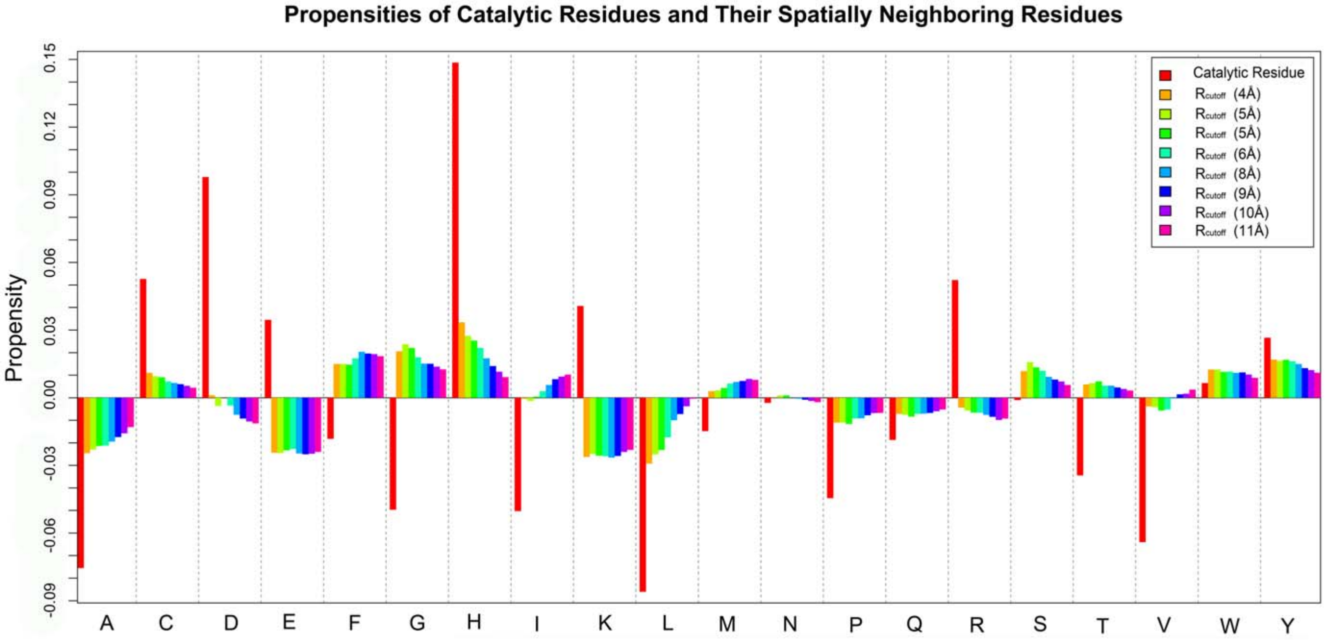

Figure 1. Propensities of 20 amino acids in their roles as catalytic residues and their spatially neighboring residues. The catalytic propensity of any residue is defined as its frequency been a catalytic residue minus its corresponding background frequency. Likewise, the propensity of any residue as catalytic residues' neighbor is defined as its frequency in the MEs of catalytic residues minus the corresponding background frequency. A positive bar means that the residue is enriched, while a negative bar means that the residue is depleted. The distance cutoff (R_{cutoff}) values, ranging from 4 to 11 Å at an interval of 1Å, were used to calculate the structural neighbors of catalytic residues. All the residues in the enzyme dataset were used to calculate the background frequency.
doi:10.1371/journal.pone.0041370.g001

When different distance cutoff (R_{cutoff}) values (ranging from 4.0 to 11.0 Å) were applied to assign the structural neighbors of the catalytic residues, the corresponding trends of amino acid compositions in the ME of catalytic residues remain largely unchanged (Figure 1). Such different residue propensities of ME indicate that it is possible to develop a feature that represents ME at the residue level to distinguish catalytic from non-catalytic residues.

If we consider possible biochemical mechanisms and driving forces, the propensities of 20 amino acids in the ME of catalytic residues might reflect the intrinsic requirement for catalytic reactions. For instance, the enrichment of aromatic residues (i.e. F, Y, and W) may be attributed to the fact that their side chains are required to form the cation-π interactions with the charged catalytic residues and/or charged substrates, which is helpful for stabilizing the transition state. Similarly, the enrichment of residue G in the ME of catalytic residues may reduce the steric effects and facilitate the conformational change of catalytic sites, given that conformational changes are needed for substrate binding, protons and/or electrons transport, and product release [54]. We could envisage that the use of these properties will be very helpful for deciphering and understanding ME, which will be discussed in the following sections.

Statistical analysis of MEscore

Based on the observation that catalytic residues have unique ME features, we further considered converting the ME of a query residue into a series of spatially neighboring residue pairs and proposed a scoring function called MEscore to measure the potential of a query residue being catalytic. To achieve this, we first constructed a 400-dimensional weight coefficient vector, $\mathbf{W_{ME}}$, to quantify the likelihood of each residue pair in the ME of catalytic residues, inferred from a pre-built enzyme dataset (the training dataset). To obtain the MEscore for a query residue, we summed over the corresponding coefficients of all residue pairs related to the query residue within a distance threshold. Generally, the higher the MEscore, the higher the probability for the query residue to be catalytic. Details regarding the definitions and calculations of $\mathbf{W_{ME}}$ and MEscore can be found in the 'Materials and Methods' Section.

The 400-dimensional weight coefficient vector $\mathbf{W_{ME}}$, after being converted into a 20×20 matrix, is shown in Figure 2 and Table S1. Different residue pairs exhibit scaled propensities in the ME of catalytic residues, thereby providing important insights into the molecular mechanism of enzymatic catalysis. For instance, the residue D is frequently observed in the ME of catalytic H and R (Figure 2 and Table S1). It has been previously shown that the *pK*a value of the catalytic residue H would increase when there was a structurally neighboring residue D in local structures, as residue D could help the catalytic residue H perform its function as an acid-base [55]. A similar finding is associated with the catalytic residue R. In catalytic processes, the residue R usually plays a stabilizing role [56]. Its spatially neighboring residue D, which has opposite charge to residue R, helps to stabilize charge concentration [55]. Moreover, the salt bridges or hydrogen bonds formed between the catalytic residue R and its spatially neighboring residue D also facilitate the stabilization. In contrast to the amino acid compositions of catalytic residues and their spatially neighboring residues, $\mathbf{W_{ME}}$ clearly quantifies the preference of a residue pair in the ME of catalytic residues (Figure 2). This implies that the transformation of ME into the combination of residue-residue pairs should be more informative in distinguishing catalytic from non-catalytic residues.

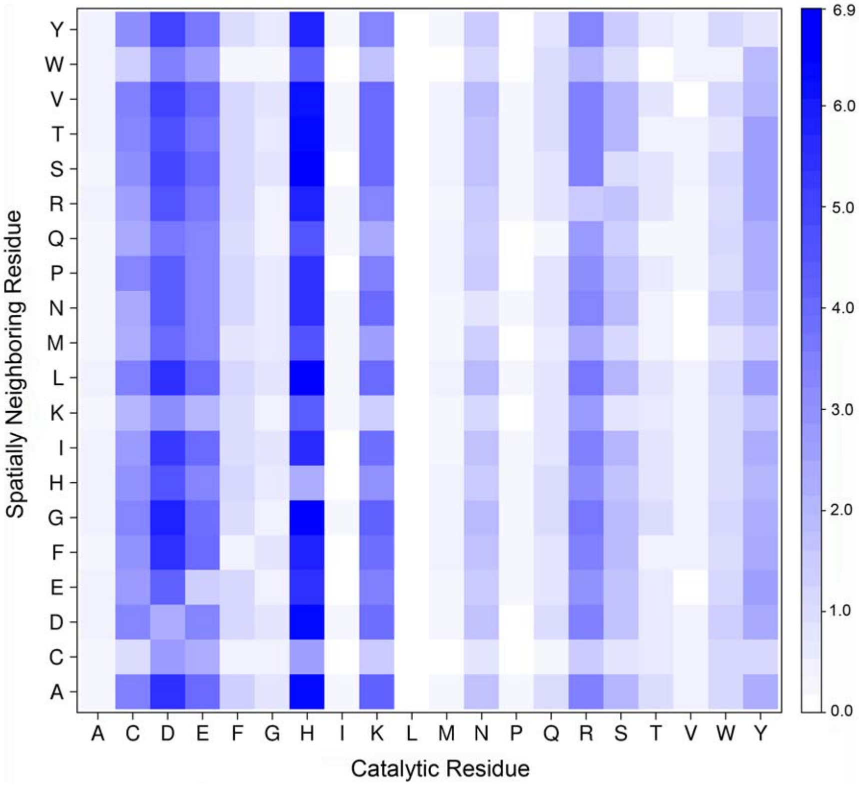

Figure 2. The weight coefficients of spatially neighboring residue pairs in the MEs of catalytic residues. The *x*-axis denotes different catalytic amino acids and the *y*-axis represents the corresponding neighboring residue types occurring in the MEs of catalytic residues. A weight coefficient close to maximum value is color-coded in blue, and it varies continuously to white color as equal to 0.0. Note that the weight coefficients were derived from the whole enzyme dataset (i.e. the 223 enzymes used in this work).
doi:10.1371/journal.pone.0041370.g002

Since the calculation of MEscore requires a training dataset, we applied five-fold cross validation tests to evaluate the performance of MEscores for all the catalytic residues in the enzyme dataset. Briefly, the enzyme dataset was divided into five subsets with roughly equal numbers of protein domains. At each cross-validation step, four subsets were merged into a training dataset to infer $\mathbf{W_{ME}}$ and the MEscores of the residues in the remaining subset (i.e. test dataset) were calculated using the established$\mathbf{W_{ME}}$ from the training dataset. After completing the five-fold cross validation tests, we obtained the distribution of MEscores for catalytic and non-catalytic residues based on the whole enzyme dataset (Figure 3). The average MEscore for all residues is 0.172. As shown in Figure 3A, a significant portion of catalytic residues (>85%) have MEscore values larger than the average for all residues. In contrast, approximately 35% of the non-catalytic residues have MEscores larger than the average value (Figure 3B). Meanwhile, the average MEscore for the catalytic residues is 0.511, which is in sharp contrast to 0.169 for the non-catalytic residues. Therefore, the MEscores for catalytic residues are significantly higher than non-catalytic residues (Wilcoxon rank-sum test, p-value = 8.07e-197), suggesting that MEscore can serve as a useful feature to discriminate catalytic from non-catalytic residues. Although the five subsets were compiled randomly, they share a reasonably similar distribution of MEscore (Figure S1). This indicates that the MEscore feature is generally robust and should achieve stable performance in each subset.

Performance of MEscore

The performance of MEscore in predicting catalytic residues was evaluated using the receiver operating characteristic (ROC) curves that plot the true positive rate as a function of the false positive rate for all the possible thresholds. Additionally, the prediction performance of MEscore was also quantified by the AUC value (AUC1.0) that represents the corresponding area under the complete ROC curve. In our study, MEscore achieved an AUC1.0 value of 0.846 (Figure 4A). For real-world applications, the ROC curve at a low false positive rate control is more practical. Therefore, the ROC curve at a 10% false positive rate control was plotted and the corresponding AUC value (AUC0.1) was 0.041 (Figure 4B). As listed in Table S2, MEscore does provide a similar performance across five different subsets in the 5-fold cross-validation tests. Note that the above ROC analysis was based on the subset level. That is to say, we generated a ROC curve for each subset and reported the average ROC curve over the generated five ROC curves. We also conducted the ROC analysis on per enzyme basis. Briefly, we generated a ROC curve for each enzyme domain and the resulting ROC curve was averaged over all the 223 domains in the enzyme dataset (Figure S2). Since MEscores were normalized for each enzyme domain, the ROC curves generated in these two different ways are very close (cf. Figure 4 and Figure S2).

To avoid the overestimation of the performance of MEscore, a stringent sequence filter criterion (i.e. the sequence identity between any two sequences should be less than 30%) was applied to compile the enzyme dataset. Considering that a larger dataset may represent more completeness of the known catalytic residues'

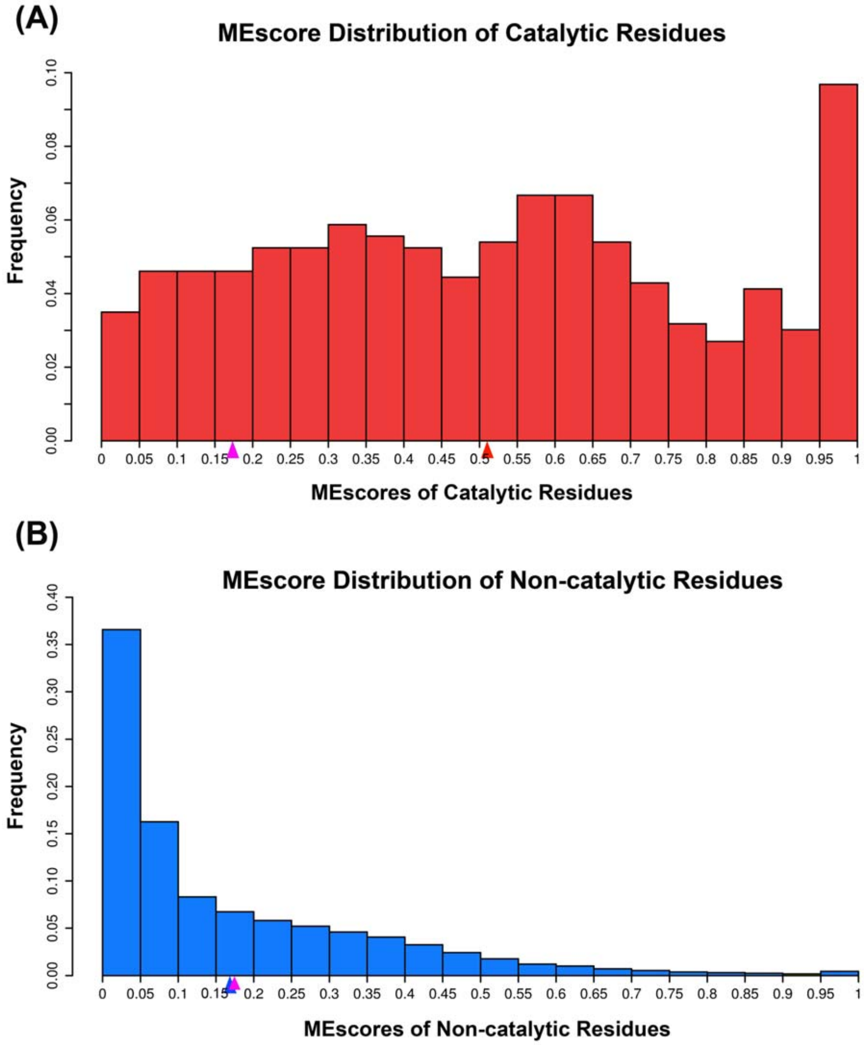

Figure 3. The distribution of MEscores. Panel A and Panel B represent the distributions of MEscore in catalytic and non-catalytic residues, respectively. The pink triangle in *x*-axis represents the average MEscore of all residues in the enzyme dataset, while the red (Panel A) and blue triangle (Panel B) denote the average MEscores of catalytic residues and non-catalytic residues, respectively.
doi:10.1371/journal.pone.0041370.g003

ME information, we also used a looser sequence filter criterion (i.e. 50% sequence identity cutoff) to obtain an enlarged enzyme dataset containing 765 domains and reassessed the performance of MEscore. Interestingly, there was only a slight increase of the overall performance based on this extended dataset (Figure S3), suggesting that the non-redundant enzyme dataset based on 30% sequence identity has already included sufficient information to deduce MEscore.

Compared to the exposed residues, the buried residues generally have a larger number of neighbors (i.e. having more information about ME). To investigate the performance of MEscore for residues in different structural locations, we classified residues into the buried and the exposed according to their relative solvent accessibility (RSA) values calculated by NACCESS [57]. Then, the ROC curves for the buried and the exposed residues were respectively plotted. As shown in Figure S4, the buried catalytic residues could be more accurately identified compared to the exposed ones. We further investigated the performance of MEscore in different structural folds. As shown in Figure S5, the performance of MEscore varies in different folds.

We further benchmarked MEscore against a simple residue type-based predictor, which was implemented via statistical analysis of the catalytic likelihood (CL) of each residue type. As shown in Figure 4, the performance of MEscore is significantly better than that of CL at varying false positive rate controls. For example, MEscore achieved an increase of 6.3% in terms of AUC1.0 and an increase of 64% in terms of AUC0.1, respectively, compared to CL's performance (AUC1.0 = 0.791 and AUC0.1 = 0.025). The quantitative performance comparison between MEscore and CL indicates that MEscore does capture more valuable information beyond the residue type in predicting catalytic residues.

Performance comparison between MEscore and Dscore

Since catalytic residues tend to be located in the center of protein structures, the distance of a residue to the geometrical center of protein has been previously shown to be a powerful structural feature for identifying catalytic residues [24]. In this work, we calculated Dscore for each residue in order to characterize this structural feature and revisited its performance in identifying the catalytic residues based on the benchmark

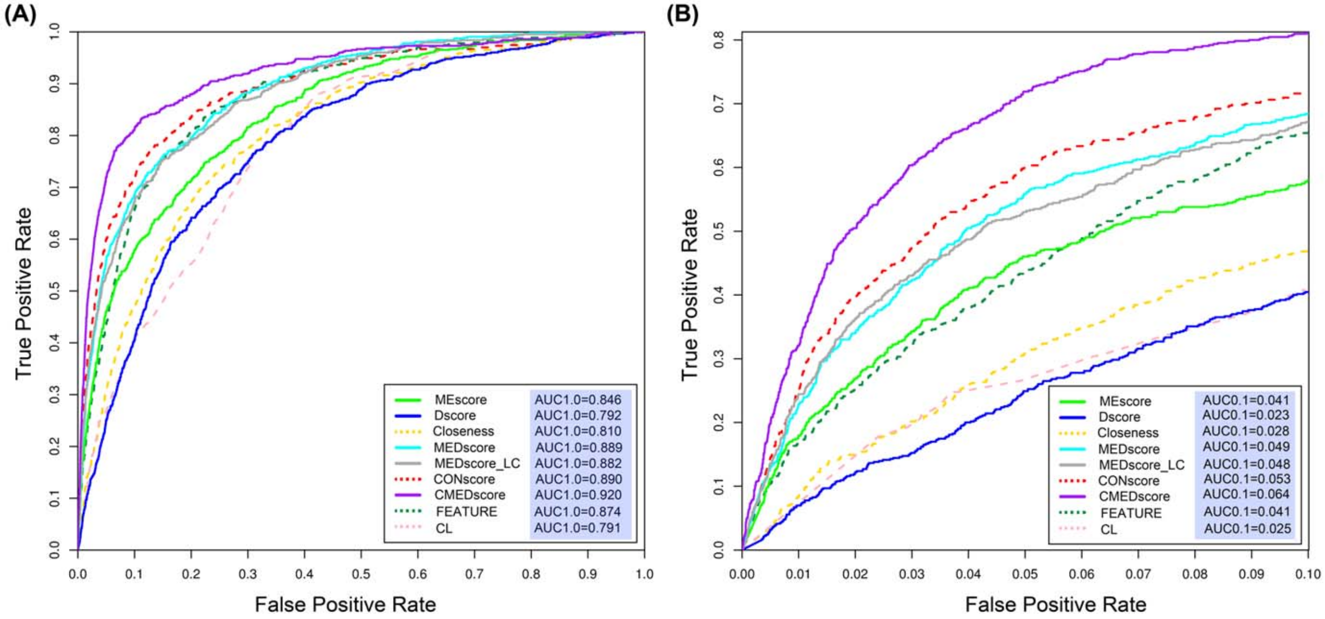

Figure 4. The ROC curves of different features/predictors. Panel A gives the ROC curves at each possible control of false positive rate, while panel B only plots ROC curves at a false positive rate ≤10%.
doi:10.1371/journal.pone.0041370.g004

enzyme dataset. In the context of residue interaction networks, the Closeness measure compared favorably with other features in predicting catalytic residues [27,37–39,43,47]. Ben-Shimon and Eisenstein found that there was a strong correlation between Dscore and Closeness based on four known enzyme structures [24]. Here, we confirm the presence of this high correlation in the current enzyme dataset [Figure 5; Pearson's correlation coefficient (PCC) = 0.946]. Therefore, Dscore and Closeness do contain similar protein structural information, despite the fact that they are extracted based on different protein structure representations. Since Closeness also describes a residue's geometrical distance to the center of protein structure, it is less likely that a minor conformation change of protein structure will considerably affect the performance of Closeness [39]. Considering that catalytic residues prefer to be located in the center of protein structures, it is more favorable for catalytic residues to cooperate with other neighboring residues during the course of catalysis [54] and have more evolutionary constraints [58]. More importantly, this may be a universal phenomenon for functional residues; for example, other functional important residues also tend to be located in the center of protein structures. Therefore, Closeness has been also used as a feature for predicting single amino acid polymorphisms (SAP) [59] and other functional sites in protein structures [37,60].

Apparently, MEscore captures different structural information in comparison to Dscore and Closeness; it is necessary to benchmark their performance on the same enzyme dataset. As depicted in Figure 4, at the R_{cutoff} of 9 Å, the AUC1.0 value of MEscore is 0.846 in comparison to 0.810 for Closeness and 0.792 for Dscore. The performance of MEscore is significantly better than Closeness (DeLong's test [61,62], p-value = 0.000152) and Dscore (DeLong's test, p-value = 2.023e-07). In terms of AUC0.1 (i.e. the area under ≤10% false positive rate control), MEscore achieved an AUC0.1 of 0.041, which is also significantly better than that of Dscore (AUC0.1 = 0.023; Bootstrap test [62], p-value = 1.200e-20) and Closeness (AUC0.1 = 0.028; Bootstrap test, p-value = 2.148e-11).

In order to analyze the overlapping predictions by MEscore, Dscore and Closeness, we drew a Venn diagram based on their prediction results at the ≤10% false positive rate (Figure 6A). The Venn diagram further confirms that Dscore provided a similar prediction capacity as Closeness. For instance, 221 catalytic residues were consistently predicted by both Dscore and Closeness, accounting for 86.7% of the Dscore and 74.7% of Closeness predictions, respectively. The high number of overlapping predictions indicates again that these two features describe similar structural properties of the protein. On the other hand, only 48.2% and 57.0% of the catalytic residues predicted by MEscore were consistently predicted by Dscore and Closeness, respectively (Figure 6A). Moreover, there is a weak correlation between MEscore and the other two features (PCC = 0.112 for Dscore and PCC = 0.133 for Closeness, respectively). These results suggest that MEscore is strongly complementary to Dscore and Closeness and an integration of MEscore with Dscore or Closeness may result in an even more powerful structural feature. Considering that the mathematical implementation of Dscore is much easier than that of Closeness, only the integration of MEscore and Dscore was carried out in this work.

Integrating MEscore and Dscore into a novel structural feature

We developed a novel nonlinear integrative structural feature by combining MEscore and Dscore, termed as MEDscore. MEDscore is a modified MEscore with the positional information (i.e. Dscore) taken into account. More details are given in the 'Materials and Methods' Section. As illustrated in Figure 4A, MEDscore achieved an AUC1.0 value of 0.889, which is a significantly improved performance than the individual MEscore (DeLong's test, p-value = 3.236e-34) or Dscore (DeLong's test, p-value = 1.219e-34). At the false positive rate control of 10%, the AUC0.1 value of MEDscore is 0.049, which is 113.0% and 19.5% higher than Dscore (Bootstrap test, p-value = 3.385e-60) and MEscore (Bootstrap test, p-value = 5.099e-19), respectively.

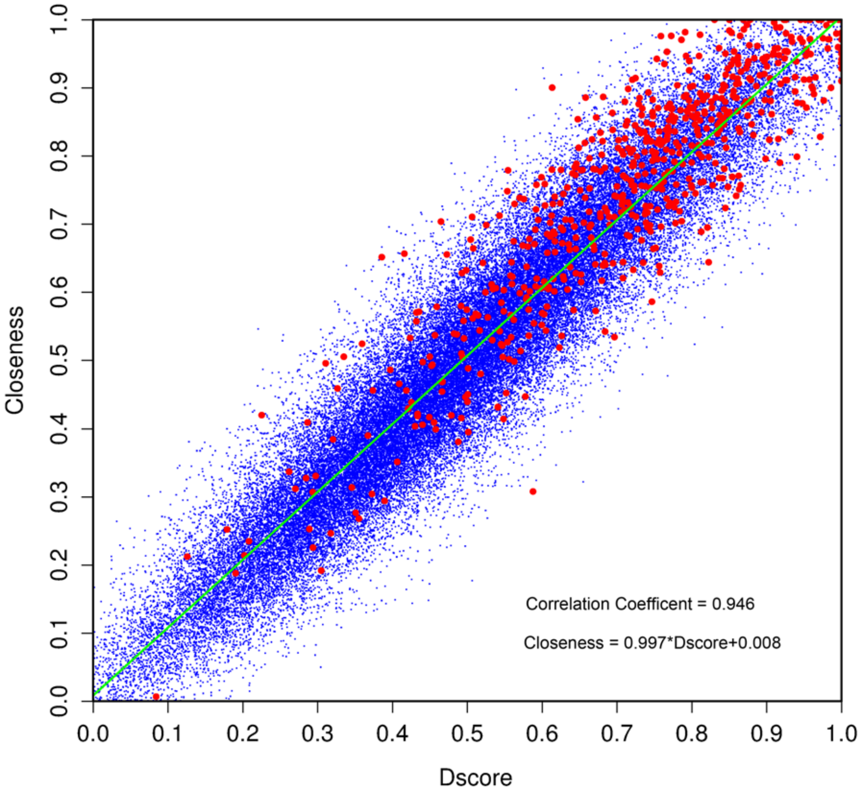

Figure 5. The relationship between Dscore and Closeness. The red points denote catalytic residues, while the blue points represent the non-catalytic residues. The correlation coefficient between Dscore and Closeness is 0.946 as derived by regression equation (i.e. the green line).
doi:10.1371/journal.pone.0041370.g005

As an alternative integration, we also combined MEscore and Dscore in a linear way to form a feature called MEDscore_LC, which is defined as a weighted sum of MEscore and Dscore. MEDscore_LC also achieved a significantly better performance than MEscore and Dscore (Figure 4). For instance, the AUC1.0 value of MEDscore_LC is 0.882, slightly lower than that of MEDscore (DeLong's test, p-value = 0.014). At a false positive rate control of 10%, MEDscore_LC also revealed a slightly lower performance in comparison to MEDscore (Figure 4B). The improved performance of MEDscore and MEDscore_LC further confirms that MEscore and Dscore are complementary with each other. The integrative strategies of MEDscore and MEDscore_LC

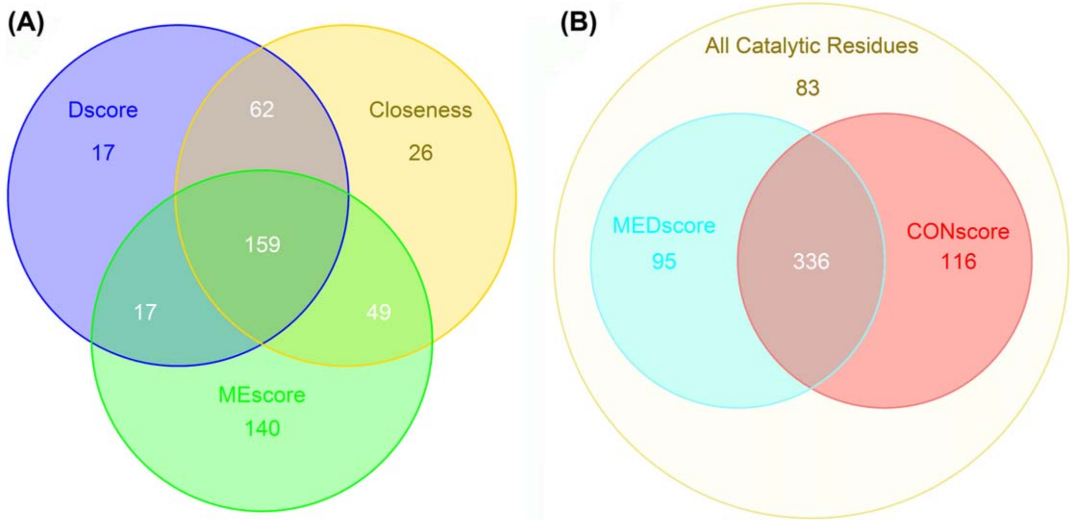

Figure 6. Venn diagrams showing the numbers of catalytic residues identified at a false positive rate ≤10%. Panel A shows the overlapping predictions by three different features (Dscore, Closeness and MEscore), and panel B summarizes the prediction results by MEDscore and CONscore.
doi:10.1371/journal.pone.0041370.g006

present several common advantages. For instance, both strategies are simple and easy to implement, and the two new features have clear and interpretable physicochemical meanings. Nevertheless, the integrative strategy of MEDscore is superior to the linear combination used by MEDscore_LC in terms of the prediction performance and whether or not there is a requirement for parameter optimization. Moreover, the construction of MEDscore has offered a novel and important way to integrate more relevant features to improve the prediction performance of catalytic residues as well as other functionally important residues.

Performance comparison between MEDscore and FEATURE

As a well-established protein functional site predictor, FEATURE extracts a set of features from the structural ME of a query residue and conducts the prediction through a Bayesian classifier [63]. A key idea that MEDscore and FEATURE share is the concept of ME representation. It is of great interest here to benchmark MEDscore against FEATURE. As can be seen from Figure 4A, MEDscore outperformed FEATURE slightly (AUC1.0 = 0.889 vs. 0.874). At the false positive rates of less than 10%, MEDscore was significantly better than FEATURE (AUC0.1 = 0.049 vs. 0.041; DeLong's test, p-value = 2.518e-7) (Figure 4B). Note that another measure MEscore also showed a better performance than FEATURE at the false positive rate control of 5% (0.0146 vs. 0.0135, Figure 4B), yet the AUC1.0 of MEscore was lower than that of FEATURE. However, different to FEATURE that employs high dimensional feature vectors and a machine learning-based algorithm to conduct the prediction, our proposed MEDscore has a clearly defined physicochemical meaning.

Performance comparison between MEDscore and the residue conservation score CONscore

As a widely used feature, the residue conservation score has proved to be the most effective singular feature in catalytic residue prediction [8–12]. In this section, we benchmarked MEDscore against one of the most advanced residue conservation measures, i.e. CONscore. The CONscore was extracted from the Consurf_DB database [64], which used the Rate4Site algorithm [16] to measure the conservation score for each residue in the protein. As shown in Figure 4, MEDscore showed a comparable performance with CONscore (AUC1.0 = 0.890; DeLong's test, p-value = 0.400). However, the performance of MEDscore was slightly lower than that of CONscore at the false positive rate control of ≤10% (AUC0.1 = 0.049 vs. 0.053; Bootstrap test, p-value = 0.018).

In addition we further benchmarked MEDscore against two common conservation scoring methods (i.e. the Shannon entropy [14] and Shannon entropy with residue properties [65]).As shown in Figure S6, the performance of MEDscore is better than these two conservation scores, indicating that MEDscore is competitive with different residue conservation measures.

Since MEDscore and CONscore target and capture different properties in proteins, their inter-correlation should be generally low. Indeed, for all the catalytic residues in the dataset, the PCC between MEDscore and CONscore was only 0.192, suggesting that there may be a large complementarity between the two features. Furthermore, we also generated the Venn diagram based on their prediction results at the ≤10% false positive rate control (Figure 6B). The Venn diagram further suggests that CONscore and MEDscore are complementary with each other to some extent. 53.4% of the catalytic residues that were not identified by CONscore, were correctly predicted by MEDscore. On the other hand, 58.3% of the catalytic residues that were not identified by MEDscore, could be correctly predicted by CONscore (Figure 6B). Therefore, the integration of these two features may lead to a more accurate and comprehensive predictor of catalytic residues.

To examine the feasibility and advantage of combining MEDscore and CONscore, we further integrated CONscore and MEDscore to form a new feature termed as CMEDscore using the weighted sum of CONscore and MEDscore. As shown in Figure 4A, CMEDscore achieved the highest AUC1.0 value of 0.920, outperforming MEDscore and CONscore with an increase in AUC1.0 of 3.48% (DeLong's test, p-value = 6.627e-25) and 3.37% (DeLong's test, p-value = 2.202e-06), respectively. At the false positive rate control of 10%, CMEDscore correctly recognized 81.1% of the catalytic residues and its corresponding AUC0.1 was 0.064. This is also remarkably higher than MEDscore (Bootstrap test, p-value = 2.681e-37) and CONscore (Bootstrap test, p-value = 1.132e-15). Taken together, these results demonstrate that MEDscore has a competitive performance and an excellent complementarity to CONscore.

Case studies

We performed two case studies to illustrate the prediction performance of all the features we have developed. In the first case study, we predicted the catalytic residues of a tryptophan biosynthesis related enzyme. Tryptophan is an important substrate for protein biosynthesis in microorganisms and plants [66]. The first step in synthesizing tryptophan is the biosynthesis of anthranilate from chorismate, catalyzed by anthranilate synthase (AnthS, PDB entry: 1QDL [67]). The small domain of AnthS (TrpG, SCOP family index: c.23.16.1) is a glutamine amidotransferase (EC 4.1.3.27) that hydrolyzes glutamine and transfers the ammonia group to a substrate to form a new carbon-nitrogen group [66,68]. The function of TrpG is carried out through a catalytic triad of the active site (C84, H175 and E177) [67]. In Figure 7A, the cartoon representation of TrpG and the prediction performance at a false positive rate control of 3% are illustrated. At such a low false positive rate control, Dscore failed to recognize any catalytic residue and MEscore correctly identified only one catalytic residue (H175). Remarkably, MEDscore correctly identified all three catalytic residues, suggesting the robustness of this nonlinear combination between MEscore and Dscore. In fact, MEDscore in this case also performed better than CONscore. We also found that CMEDscore achieved an even better performance, with all three catalytic residues ranked as the top hits according to the CMEDscore values.

The second case study concerns the catalytic residue prediction of diaminopimelate (DAP) epimerase (EC 5.1.1.7), a typical member of pyridoxal phosphate (PLP)-independent amino acid racemase involved in lysine biosynthesis [69]. The structure of DAP epimerase was characterized from *Haemophilus influenzae* (PDB entry: 1BWZ), consisting of two homology domains [70]. In the C-terminal domain (residues 118–262, SCOP family index: d.21.1.1), three catalytic residues (H159, E208 and C217) directly contribute to the catalytic process. Figure 7B presents the performance of different features at the false positive rate control of 3%. At this rigorous control, the performance of both of MEDscore and CONscore was weak. MEDscore only correctly detected one of the three catalytic residues while CONscore identified none. In comparison, CMEDscore was able to correctly predict all three catalytic residues. These results indicate that there indeed exists a complementarity between CONscore and MEDscore.

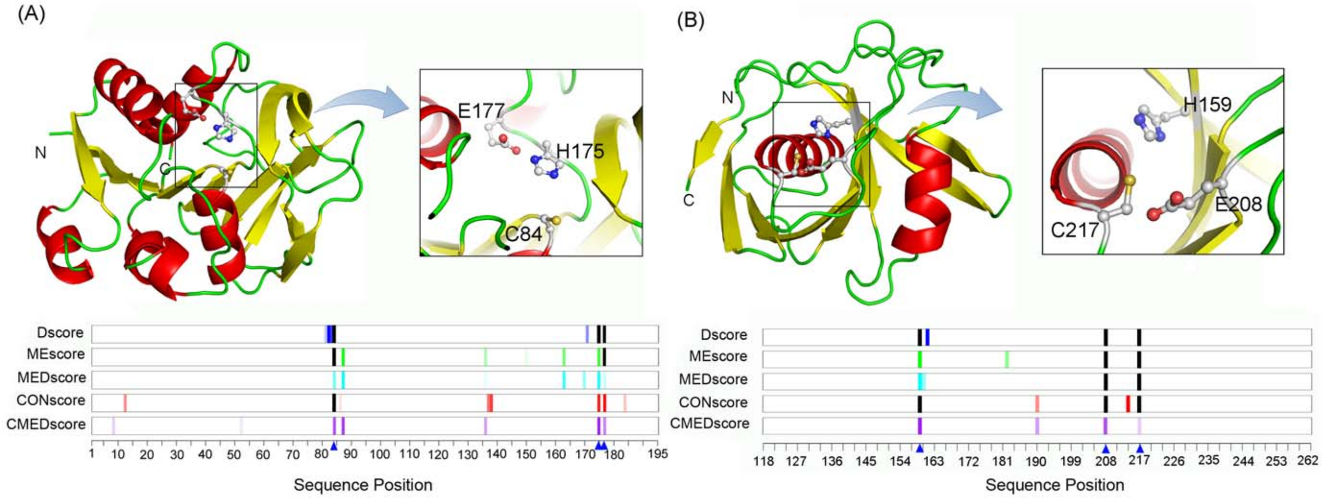

Figure 7. Two case studies illustrating the prediction performance of different features at a false positive rate control of 3%. Panel A shows the predicted catalytic residues of TrpG (the small domain of anthranilate synthase; PDB entry: 1QDL), and panel B gives the predictions of diaminopimelate (DAP) epimerase (PDB entry: 1BWZ). Top parts: Protein structures are represented by cartoon ribbons and the corresponding catalytic residues are highlighted by ball-and-stick-models, as seen in the insets. Lower parts: The blue triangles represent the sequence positions of the catalytic residues. With respect to the prediction results of each feature, the sequence positions of the predicted catalytic residues are marked using colored bars, with a higher score corresponding to a more saturated color. The black bars denote catalytic residues which a corresponding feature failed to predict.

doi:10.1371/journal.pone.0041370.g007

It is noteworthy that both of the two query proteins (TrpG and DAP epimerase) share less than 40% sequence identity with any other domains in the enzyme dataset used in this work. The findings from the above case studies provide supportive evidence that MEDscore and its integration with CONscore (i.e. CMEDscore) can favourably identify a large portion of catalytic residues from the given protein structure with better accuracy. This suggests that these new features can be efficiently employed for practical applications.

Conclusion

A number of structural features have been developed to identify catalytic residues from enzyme structures. However, the performance of most features is not comparable to the most powerful sequence-based feature, namely, the residue conservation score. In this work we develop a novel promising structural feature termed as MEDscore for the identification of catalytic residues. The superior performance of MEDscore can be ascribed to its capability of capturing the intrinsic ME and geometrical location properties of the residues. In particular, it allows the ME of a residue to be converted into a series of spatially neighboring residue pairs such that the likelihood of belonging to the catalytic ME could be deduced from a pre-existing enzyme dataset. To the best of our knowledge, this research represents the first endeavor to characterize ME of a residue based on this strategy. From the practical perspective, we find that the proposed MEDscore performs better in catalytic residue prediction when being integrated with other features, such as CONscore. Moreover, it should be noted that MEDscore may be the first structural feature that shows a competitive performance compared to the residue conservation score. We anticipate that this novel structural feature can be applied to reliably identify catalytic residues, facilitate the functional annotation of structural genomics targets and improve our understanding of the complex sequence-structure-function relationships of enzymes.

Materials and Methods

Benchmark enzyme dataset

The benchmark enzyme dataset used in this study was extracted from the Catalytic Site Atlas (CSA) database (version 2.2.12) [71]. Total of 7,124 entries with catalytic residues annotated directly in the literature were extracted. These entries were mapped onto the SCOP database (version 1.75) [72] and the corresponding PDB files were downloaded from the ASTRAL database (http://astral.berkeley.edu/pdbstyle-1.75.html) [73]. These enzymes were further filtered based on the following criteria: a) the sequence identity between any two sequences should be less than 30%; b) the sequence length of any enzyme should be larger than 100; c) the PDB structures with 10 consecutive missing residues were excluded; d) only the PDB structures belonging to four SCOP structural classes (i.e. all-α, all-β, $\alpha+\beta$ and α/β) were included; e) if an enzyme had two or more NMR structure models in our dataset, only the first model was retained; and f) some enzymes were discarded because that the number of homologous sequences of the enzymes was insufficient to permit an accurate calculation of residue conservation scores. Based on the above criteria, 223 enzyme catalytic domains were retained in our final dataset, covering six top levels of the EC classifications. These 223 enzymes cover 112 folds, 139 superfamilies and 185 families in terms of the SCOP classification. In this non-redundant benchmark enzyme dataset, 630 residues are defined as catalytic residues according to the CSA annotation, while the remaining 60,658 residues are regarded as non-catalytic residues. The details about the enzyme dataset are listed in Supporting Information Text S1.

Definition and Calculation of MEscore

The definition of ME in the context of residue interaction networks. Given that a protein structure can be represented as a residue interaction network, residues are viewed as nodes and an edge can be established if the distance between the two residues is less than a distance cutoff (R_{cutoff}). The residue interaction network can be represented as an adjacency matrix D as follows

$$D_{ij}=\begin{cases}1 & r_{ij}\le R_{cutoff}\ i\neq j\\ 0 & otherwise\end{cases} \tag{1}$$

where r_{ij} is the spatial distance between residue i and j. In this work, r_{ij} is defined as the shortest distance between any pairs of the heavy atoms (C, N, O, S) in residue i and j.

In the context of residue interaction networks, the direct neighboring residues constitute the ME of a given residue. In other words, the ME of a residue can be defined using its direct interacting neighbors in the residue interaction network. Here, a 400-dimensional residue pair frequency vector called $\mathbf{F_{ME}}$ is used to represent the ME information of a residue, which is defined as

$$\mathbf{F_{ME}}=\left(N_{AA}^{ME},N_{AC}^{ME},...,N_{A_mA_n}^{ME},\ \ldots,N_{YW}^{ME},N_{YY}^{ME}\right)_{400} \tag{2}$$

Note that the residue pair representation in $\mathbf{F_{ME}}$ is orientation-dependent. More specifically, the residue pair A_mA_n has the orientation from the query residue A_m to its spatially neighboring residue A_n. It should be noted that the residue pairs between any two neighboring residues of A_m are not taken into account. The value of each feature, such as $N_{A_mA_n}^{ME}$, denotes the number of the corresponding residue pair involved in the ME, which can be readily extracted from the pre-computed adjacency matrix D of the protein structure. Although $\mathbf{F_{ME}}$ of a residue is defined as a 400-dimensional vector, it is highly sparse in nature and only contains 20 parameters with potential non-zero values.

The residue pair weight coefficient vector. Our hypothesis is that spatially neighboring residue pairs of the catalytic residues should have a specific frequency distribution and such specificity can be determined and used to identify catalytic residues in a given protein structure. To explore the frequency distribution, we introduce a residue pair weight coefficient vector for each residue in a protein structure, which is expressed as

$$\mathbf{W_{ME_i}}=\left(W_{AA}^{ME},W_{AC}^{ME},\ \ldots,\ W_{A_mA_n}^{ME},\ \ldots,\ W_{YW}^{ME},W_{YY}^{ME}\right)_{400} \tag{3}$$

Here $W_{A_mA_n}^{ME}$ is calculated as

$$W_{A_mA_n}^{ME}=N_{A_mA_n}^{ME}/(N_{A_m}^{ME}\times N_{A_n}^{ME}) \tag{4}$$

where N_{A_m} and N_{A_n} are the numbers of residues A_m and A_n in the ME of the query residue, respectively. Note that the query residue itself is also included when counting N_{A_m}.

The overall $\mathbf{W_{ME}}$ vector for catalytic residues is measured by averaging all the weighted vectors of the related catalytic residues:

$$\mathbf{W_{ME}}=\left(\frac{\sum_{n=1}^{CN_{AA}}W_{AA}^{ME}}{\ln CN_A},\frac{\sum_{n=1}^{CN_{AC}}W_{AC}^{ME}}{\ln CN_A},\ \ldots,\ \frac{\sum_{n=1}^{CN_{A_mA_n}}W_{A_mA_n}^{ME}}{\ln CN_{A_m}},\right.$$
$$\left.\ldots,\frac{\sum_{n=1}^{CN_{YW}}W_{YW}^{ME}}{\ln CN_Y},\frac{\sum_{n=1}^{CN_{YY}}W_{YY}^{ME}}{\ln CN_Y}\right)_{400} \tag{5}$$

where $CN_{A_mA_n}$ is the total number of residue pairs of A_mA_n in all the MEs of catalytic residues, while CN_{A_m} represents the total number of catalytic residues of type A_m.

MEscore

MEscore is proposed to measure the likelihood of a query residue being catalytic, which is derived using the following equation:

$$\text{MEscore}=\mathbf{F_{ME}}\times\mathbf{W_{ME}^T} \tag{6}$$

Here $\mathbf{W_{ME}^T}$ is the transposed matrix of $\mathbf{W_{ME}}$. The MEscores of all residues in a protein structure are further normalized using the following equation:

$$norm\_score=\frac{score-\min(score)}{\max(score)-\min(score)} \tag{7}$$

In this study, six different R_{cutoff} values (from 3 to 13 Å at an interval of 2Å) were examined in order to obtain the optimal one. As a result, MEscore almost achieved the maximal performance when R_{cutoff} was set as 9 Å (Figure S7). Therefore, the optimal value of R_{cutoff} was set as 9.0 Å.

Definition and calculation of Dscore

Since catalytic residues tend to be located in the center of protein structures, we further develop a feature termed as Dscore to characterize this property. Briefly, each residue in a protein structure is represented by its C_α atom and the atomic coordinates of the geometrical center of this protein are calculated as follows:

$$(c_x,c_y,c_z)=\left(\frac{\sum_{i=1}^{N}x_i}{N},\frac{\sum_{i=1}^{N}y_i}{N},\frac{\sum_{i=1}^{N}z_i}{N}\right) \tag{8}$$

where c_x, c_y and c_z are the coordinates of the center of the protein structure; x_i, y_i and z_i are the trajectory of the C_α atom in residue i; while N is the total number of residues in the protein. The distance between a residue i and the center of the structure (i.e. Dscore_i) is then calculated as

$$Dscore_i=\sqrt{(c_x-x_i)^2+(c_y-y_i)^2+(c_z-z_i)^2} \tag{9}$$

The Dscore values of each protein are normalized using the above Eq. (7). After normalization, its values in a protein vary in the range from 0 to 1. To transform Dscore from a dissimilarity to similarity measure, each Dscore is subtracted by 1. Thus, a residue near the center of the protein structure presumably has a relatively high Dscore.

Definition and calculation of MEDscore and MEDscore_LC

Dscore is a global feature that describes the positional information of a residue in protein, while MEscore is a local feature that describes the local environment surrounding the residue of interest. As these two types of features capture different and complementary information of the catalytic residue, we integrate them to constitute a novel feature in a nonlinear manner. The rationale behind this nonlinear integration is to construct a refined MEscore, termed as MEDscore, where the positional information (Dscore) of each residue involved in the ME representation is also considered. To achieve this, $\mathbf{F_{ME}}$ is firstly modified to $\mathbf{F_{MED}}$:

$$\mathbf{F_{MED}} = \left(N_{AA}^{MED},\, N_{AC}^{MED},\, ...,\, N_{AmAn}^{MED},\, \ldots,\, N_{YW}^{MED},\, N_{YY}^{MED}\right)_{400} \quad (10)$$

where N_{AmAn}^{MED} is calculated as

$$N_{AmAn}^{MED} = \sum^{All\ AmAn\, pairs} Dscore_{Am} \times Dscore_{An} \quad (11)$$

where $Dscore_{Am}$ and $Dscore_{An}$ are the Dscore of the query residue A_m and its neighboring residue A_n, respectively. Note that the definition of A_mA_n is the same as described in $\mathbf{F_{ME}}$. For each occurring residue pair A_mA_n, the two corresponding Dscore values are multiplied to obtain a coefficient. N_{AmAn}^{MED}corresponds to the summation of the corresponding coefficients for all the observed A_mA_n. Then, the residue pair weight coefficient vector for each residue in a protein structure is modified to

$$\mathbf{W_{MED_i}} = \left(W_{AA}^{MED},\, W_{AC}^{MED},\, ...,\, W_{AmAn}^{MED},\, ...,\, W_{YW}^{MED},\, W_{YY}^{MED}\right)_{400} \quad (12)$$

Here W_{AmAn}^{MED} is calculated as

$$W_{AmAn}^{MED} = N_{AmAn}^{MED} / (N_{Am}^{MED} \times N_{An}^{MED}) \quad (13)$$

where N_{Am}^{MED} stands for the summation of the Dscores for all the occurring A_m in the ME and N_{An}^{MED} represents the summation of the Dscores for all the occurring A_n in the ME, respectively. Similar to Eq.(4), the query residue itself is also considered when counting N_{Am}^{MED}.

The overall weighted vector ($\mathbf{W_{MED}}$) for catalytic residues is measured by averaging all the weighted vectors of the related catalytic residues:

$$\mathbf{W_{MED}} = \left(\frac{\sum_{n=1}^{CN_{AA}} W_{AA}^{MED}}{\ln CN_A}, \frac{\sum_{n=1}^{CN_{AC}} W_{AC}^{MED}}{\ln CN_A}, \ldots, \frac{\sum_{n=1}^{CN_{AmAn}} W_{AmAn}^{MED}}{\ln CN_{Am}}, \ldots, \frac{\sum_{n=1}^{CN_{YW}} W_{YW}^{MED}}{\ln CN_Y}, \frac{\sum_{n=1}^{CN_{YY}} W_{YY}^{MED}}{\ln CN_Y} \right)_{400} \quad (14)$$

In the above equation, CN_{AmAn} and CN_{Am} have the same meanings as in Eq. (5).

Finally, a nonlinear combination between MEscore and Dscore is obtained using the following equation:

$$\mathrm{MEDscore} = \mathbf{F_{MED}} \times \mathbf{W_{MED}^T} \quad (15)$$

To ensure that the MEDscores of all residues in a protein structure range from 0 to 1, the original MEDscores are further normalized by Eq. (7).

As an alternative combination, we linearly combine MEscore and Dscore into another feature called MEDscore_LC, which is defined as

$$\mathrm{MEDscore\_LC} = \alpha \times \mathrm{MEscore} + (1-\alpha) \times \mathrm{Dscore} \quad (16)$$

To determine the optimal value of α, we benchmarked the performance of MEDscore_LC using different α values, ranging from 0.0 to 1.0 at an interval of 0.05. The optimal α corresponded to the maximal AUC1.0. In this work, the optimal value of α was assigned to 0.55.

Other existing features and predictors

Catalytic likelihood of each residue (CL). The CL value of each residue can be inferred from the benchmark enzyme dataset, defined as

$$CL_{Am} = \frac{CN_{Am}/CN}{TN_{Am}/TN} \quad (17)$$

where CN_{Am} denotes the number of residue A_m as catalytic, CN is the total number of catalytic residues, TN_{Am} is the number of residue A_m, and TN is the total number of residues, respectively. For a given residue, the corresponding value of CL can be used to predict whether a residue is catalytic or not.

Closeness. Derived from residue interaction networks, Closeness has been previously shown to be a powerful feature in predicting catalytic residues [27,37,39,43,47]. In this study, Closeness of a given residue i in the residue interaction network of a protein is calculated as

$$Closeness_i = \frac{N-1}{\sum_{i \neq j} SD_{ij}} \quad (18)$$

where N is the number of residues in the protein structure and SD_{ij} is the shortest path between residues i and j. To construct the residue interaction network, the value of R_{cutoff} was set as 9.0 Å. Note that the Closeness values also need to be further normalized using Eq. (7).

FEATURE. FEATURE employs a Bayesian classifier to predict protein functional sites [63]. The input feature vector of FEATURE is constructed from the ME surrounding the query residue, including atom properties, residue properties, partial charge, solvent accessibility etc. We downloaded the stand-alone version of FEATURE (version 3.0) from https://simtk.org/home/feature, and evaluated its performance based on the benchmark enzyme dataset in this study. For performance comparison, as reported in [46], the ratio of catalytic to non-catalytic residues in the training stage was set as 1:6.

Residue conservation score. The residue conservation scores of the enzymes were directly obtained from the ConSurf-DB database [64], which consists of pre-calculated conservation scores for each protein structure. In order to detect functionally important residues in protein structures, ConSurf-DB employs the Rate4Site algorithm [16] to compute the conservation score of each residue based on the generated MSA. The criteria to generate the MSA are detailed in [64]. The number of sequences in the MSAs of the 223 enzyme domains ranges from 8 to 199 and the average number is 124. Approximately 90% of the enzymes have more than 50 aligned sequences in their respective MSAs, while only two enzymes have less than 10 aligned sequences in the corresponding MSAs. In addition to the use of an empirical Bayesian inference, Rate4Site also takes into account the phylogenic relations within proteins. Likewise, the pre-calculated conservation score (CONscore) of each residue is further normalized using Eq. (7).

Linear combination of CONscore and MEDscore

Similar to the construction of MEDscore_LC, we further integrate CONscore and MEDscore into a new feature, defined as CMEDscore:

$$\mathrm{CMEDscore} = \beta \times \mathrm{CONscore} + (1-\beta) \times \mathrm{MEDscore} \qquad (19)$$

Similar to the determination of the optimal α in Eq. (16), the value of β was optimized to attain the maximal AUC1.0 of CMEDscore. In the current enzyme dataset, the optimal value of β was set as 0.70.

Performance assessment of different features or predictors

In this work, we used five-fold cross validation tests to evaluate the performance of the six features/predictors, i.e. MEscore, MEDscore, MEDscore_LC, CL, FEATURE and CMEDscore. In particular, the benchmark enzyme dataset was randomly divided into five subsets and each subset contained roughly equal number of protein domains (the SCOP entries of these five subsets are available in Text S1). In each cross-validation evaluation step, four subsets were merged into a training dataset to infer the residue pair weight vectors ($\mathbf{W_{ME}}$ and $\mathbf{W_{MED}}$) and the remaining were utilized as a test dataset to assess the performance of each feature. The final performance of this feature was averaged over all of the five subsets. We also randomly repeated the subset partition three times and similar results were obtained for each feature/predictor. Since Dscore, Closeness and CONscore do not require a training phase, the performance of these three features was evaluated directly on the whole dataset rather than the five-fold cross validation tests.

We used the ROC curve and two corresponding parameters (AUC1.0 and AUC0.1) to assess the overall performance of each feature. In this work, the ROC curve was prepared on per subset basis. Briefly, we generated a ROC curve in each subset and the overall ROC curve on the whole dataset was averaged over the generated five ROC curves. Note that the ROC curve can also be generated on per enzyme basis. That is to say, we can plot a ROC curve in each enzyme domain and obtain the average ROC curve on the 223 enzyme domains. Since normalization at the domain level was conducted for each feature, the ROC curves based on the above two strategies should generate close results. For comparison, the ROC curves on per enzyme basis are also shown in Figure S2.

Analysis and Visualization

All computational and analytic scripts were written in Perl/R languages. The implemented R packages included ROCR [74] and pROC [62] for ROC analysis and visualization, as well as igraph [75] for network parameter calculation. All the figures were prepared using either R (http://www.r-project.org/) or PyMol (http://www.pymol.org/).

Supporting Information

Figure S1 The MEscore distribution across the five subsets in 5-fold cross-validation tests.
(PDF)

Figure S2 The ROC curves of different features/ predictors on per enzyme basis.
(PDF)

Figure S3 The ROC curves of MEscore and MEDscore based on two different datasets.
(PDF)

Figure S4 Performance of MEscore in predicting the buried and the exposed catalytic residues.
(PDF)

Figure S5 The performance variation of different features/predictors in different structural folds.
(PDF)

Figure S6 The ROC curves of MEDscore and two simple conservation scoring methods (i.e. the Shannon entropy and Shannon entropy with residue properties).
(PDF)

Figure S7 The performance of MEscore and MEDscore at different R_{cutoff} values.
(PDF)

Text S1 This file contains the FASTA sequences and the experimentally verified catalytic residues of 223 enzymes used in this work. The SCOP entries of the five subsets used in the five-fold cross validation tests are also given.
(DAT)

Table S1 This file contains the calculated weight coefficient vector $\mathbf{W_{ME}}$, which is represented by a 20×20 amino acid matrix. Note that the listed $\mathbf{W_{ME}}$ was derived from the whole enzyme dataset with 223 enzymes.
(DOC)

Table S2 The performance of MEscore in five subsets.
(DOC)

Acknowledgments

We thank Prof. Russ B. Altman (Stanford University) for sharing the FEATURE program and Prof. Tal Pupko (Tel Aviv University) for providing the Rate4Site program. The authors also thank Dr. Zhi-Ping Feng at the Walter and Eliza Hall Institute of Medical Research (WEHI), Australia for helpful discussions. MSL thanks China Agricultural University for hosting his visits and acknowledges the support from CSIRO-Advanced Materials TCP and Computational and Simulation Sciences-TCP.

Author Contributions

Conceived and designed the experiments: LH MSL ZZ. Performed the experiments: LH. Analyzed the data: LH YJZ JS MSL ZZ. Wrote the paper: LH JS MSL ZZ.

References

1. Benkovic SJ, Hammes-Schiffe S (2003) A perspective on enzyme catalysis. Science 301: 1196–1202.
2. Burley SK (2000) An overview of structural genomics. Nat Struct Mol Biol 7: 932–934.
3. Baker D (2001) Protein structure prediction and structural genomics. Science 294: 93–96.
4. Laskowski RA, Watson JD, Thornton JM (2003) From protein structure to biochemical function? Journal of Structural and Functional Genomics 4: 167–177.
5. Noble MEM, Endicott JA, Johnson LN (2004) Protein kinase inhibitors: insights into drug design from structure. Science 303: 1800–1805.
6. Andrianantoandro E, Basu S, Karig DK, Weiss R (2006) Synthetic biology: new engineering rules for an emerging discipline. Mol Syst Biol 2: 2006 0028.
7. Bartlett GJ, Porter CT, Borkakoti N, Thornton JM (2002) Analysis of catalytic residues in enzyme active sites. J Mol Biol 324: 105–121.
8. Petrova N, Wu C (2006) Prediction of catalytic residues using support vector machine with selected protein sequence and structural properties. BMC Bioinformatics 7: 312.

9. Chien T-Y, Chang DT-H, Chen C-Y, Weng Y-Z, Hsu C-M (2008) E1DS: catalytic site prediction based on 1D signatures of concurrent conservation. Nucl Acids Res 36: W291–W296.
10. Gutteridge A, Bartlett GJ, Thornton JM (2003) Using a neural network and spatial clustering to predict the location of active sites in enzymes. J Mol Biol 330: 719–734.
11. Youn E, Peters B, Radivojac P, Mooney SD (2007) Evaluation of features for catalytic residue prediction in novel folds. Protein Science 16: 216–226.
12. La D, Sutch B, Livesay DR (2005) Predicting protein functional sites with phylogenetic motifs. Proteins 58: 309–320.
13. Dukka Bahadur KC, Livesay DR (2008) Improving position-specific predictions of protein functional sites using phylogenetic motifs. Bioinformatics 24: 2308–2316.
14. Shenkin PS, Erman B, Mastrandrea LD (1991) Information-theoretical entropy as a measure of sequence variability. Proteins 11: 297–313.
15. Capra JA, Singh M (2007) Predicting functionally important residues from sequence conservation. Bioinformatics 23: 1875–1882.
16. Mayrose I, Graur D, Ben-Tal N, Pupko T (2004) Comparison of site-specific rate-inference methods for protein sequences: empirical bayesian methods are superior. Molecular Biology and Evolution 21: 1781–1791.
17. Caffrey DR, Somaroo S, Hughes JD, Mintseris J, Huang ES (2004) Are protein–protein interfaces more conserved in sequence than the rest of the protein surface? Protein Science 13: 190–202.
18. Yao H, Kristensen DM, Mihalek I, Sowa ME, Shaw C, et al. (2003) An accurate, sensitive, and scalable method to identify functional sites in protein structures. J Mol Biol 326: 255–261.
19. del Sol Mesa A, Pazos F, Valencia A (2003) Automatic methods for predicting functionally important residues. J Mol Biol 326: 1289–1302.
20. Dou Y, Zheng X, Yang J, Wang J (2010) Prediction of catalytic residues based on an overlapping amino acid classification. Amino Acids 39: 1353–1361.
21. Lee B-C, Park K, Kim D (2008) Analysis of the residue-residue coevolution network and the functionally important residues in proteins. Proteins 72: 863–872.
22. Lengauer T, Kowarsch A, Fuchs A, Frishman D, Pagel P (2010) Correlated mutations: a hallmark of phenotypic amino acid substitutions. PLoS Comput Biol 6: e1000923.
23. Marino Buslje C, Teppa E, Di Doménico T, Delfino JM, Nielsen M (2010) Networks of high mutual information define the structural proximity of catalytic sites: implications for catalytic residue identification. PLoS Comput Biol 6: e1000978.
24. Ben-Shimon A, Eisenstein M (2005) 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 351: 309–326.
25. Ota M, Kinoshita K, Nishikawa K (2003) Prediction of catalytic residues in enzymes based on known tertiary structure, stability profile, and sequence conservation. J Mol Biol 327: 1053–1064.
26. Sonavane S, Chakrabarti P (2010) Prediction of active site cleft using support vector machines. J Chem Inf Model 50: 2266–2273.
27. Tang YR, Sheng ZY, Chen YZ, Zhang Z (2008) An improved prediction of catalytic residues in enzyme structures. Protein Eng Des Sel 21: 295–302.
28. Malabanan MM, Amyes TL, Richard JP (2010) A role for flexible loops in enzyme catalysis. Current Opinion in Structural Biology 20: 702–710.
29. Yuan Z, Zhao J, Wang ZX (2003) Flexibility analysis of enzyme active sites by crystallographic temperature factors. Protein Eng Des Sel 16: 109–114.
30. Wang K, Horst JA, Cheng G, Nickle DC, Samudrala R (2008) Protein meta-functional signatures from combining sequence, structure, evolution, and amino acid property information. PLoS Comput Biol 4: e1000181.
31. Ondrechen MJ (2001) THEMATICS: A simple computational predictor of enzyme function from structure. Proc Natl Acad Sci USA 98: 12473–12478.
32. Ko J, Murga LF, André P, Yang H, Ondrechen MJ, et al. (2005) Statistical criteria for the identification of protein active sites using theoretical microscopic titration curves. Proteins 59: 183–195.
33. Elcock AH (2001) Prediction of functionally important residues based solely on the computed energetics of protein structure. J Mol Biol 312: 885–896.
34. Bryliński M, Prymula K, Jurkowski W, Kochańczyk M, Stawowczyk E, et al. (2007) Prediction of functional sites based on the fuzzy oil drop model. PLoS Comput Biol 3: e94.
35. Sacquin-Mora S, Laforet É, Lavery R (2007) Locating the active sites of enzymes using mechanical properties. Proteins 67: 350–359.
36. Atilgan AR, Akan P, Baysal C (2004) Small-world communication of residues and significance for protein dynamics. Biophysical Journal 86: 85–91.
37. Amitai G, Shemesh A, Sitbon E, Shklar M, Netanely D, et al. (2004) Network analysis of protein structures identifies functional residues. J Mol Biol 344: 1135–1146.
38. del Sol A, Fujihashi H, Amoros D, Nussinov R (2006) Residue centrality, functionally important residues, and active site shape: analysis of enzyme and non-enzyme families. Protein Science 15: 2120–2128.
39. Chea E, Livesay DR (2007) How accurate and statistically robust are catalytic site predictions based on closeness centrality? BMC Bioinformatics 8: 153.
40. Bagley SC, Altman RB (1995) Characterizing the microenvironment surrounding protein sites. Protein Science 4: 622–635.
41. Zvelebil MJ, Sternberg MJ (1988) Analysis and prediction of the location of catalytic residues in enzymes. Protein Engineering 2: 127–138.
42. Liang MP (2003) WebFEATURE: an interactive web tool for identifying and visualizing functional sites on macromolecular structures. Nucl Acids Res 31: 3324–3327.
43. Li Y, Li G, Wen Z, Yin H, Hu M, et al. (2011) Novel feature for catalytic protein residues reflecting interactions with other residues. PLoS ONE 6: e16932.
44. Cilia E, Passerini A (2010) Automatic prediction of catalytic residues by modeling residue structural neighborhood. BMC Bioinformatics 11: 115.
45. Li G-H, Huang J-F (2010) CMASA: an accurate algorithm for detecting local protein structural similarity and its application to enzyme catalytic site annotation. BMC Bioinformatics 11: 439.
46. Xin F, Myers S, Li YF, Cooper DN, Mooney SD, et al. (2010) Structure-based kernels for the prediction of catalytic residues and their involvement in human inherited disease. Bioinformatics 26: 1975–1982.
47. Sankararaman S, Sha F, Kirsch JF, Jordan MI, Sjolander K (2010) Active site prediction using evolutionary and structural information. Bioinformatics 26: 617–624.
48. Tong W, Wei Y, Murga LF, Ondrechen MJ, Williams RJ (2009) Partial order optimum likelihood (POOL): maximum likelihood prediction of protein active site residues using 3D Structure and sequence properties. PLoS Comput Biol 5: e1000266.
49. Zhang T, Zhang H, Chen K, Shen S, Ruan J, et al. (2008) Accurate sequence-based prediction of catalytic residues. Bioinformatics 24: 2329–2338.
50. Yahalom R, Reshef D, Wiener A, Frankel S, Kalisman N, et al. (2011) Structure-based identification of catalytic residues. Proteins 79: 1952–1963.
51. Pande S, Raheja A, Livesay DR (2007) Prediction of enzyme catalytic sites from sequence using neural networks. IEEE Symp CIBCB 7: 247–253.
52. Zhang Z, Tang Y-R, Sheng Z-Y, Zhao D (2009) An overview of the de novo prediction of enzyme catalytic residues. Current Bioinformatics 4: 197–206.
53. Xin F, Radivojac P (2011) Computational methods for identification of functional residues in protein structures. Curr Protein Pept Sci 12: 456–469.
54. Yang L-W, Bahar I (2005) Coupling between catalytic site and collective dynamics: a requirement for mechanochemical activity of enzymes. Structure 13: 893–904.
55. Gutteridge A, Thornton JM (2005) Understanding nature's catalytic toolkit. Trends in Biochemical Sciences 30: 622–629.
56. Holliday GL, Mitchell JBO, Thornton JM (2009) Understanding the functional roles of amino acid residues in enzyme catalysis. J Mol Biol 390: 560–577.
57. Hubbard SJ, Thornton JM (1993) NACCESS Version 2.1.1, Computer Program, Department of Biochemistry and Molecular Biology, University College London.
58. Schueler-Furman O, Baker D (2003) Conserved residue clustering and protein structure prediction. Proteins 52: 225–235.
59. Li Y, Wen Z, Xiao J, Yin H, Yu L, et al. (2011) Predicting disease-associated substitution of a single amino acid by analyzing residue interactions. BMC Bioinformatics 12: 14.
60. David-Eden H, Mandel-Gutfreund Y (2008) Revealing unique properties of the ribosome using a network based analysis. Nucl Acids Res 36: 4641–4652.
61. DeLong ER, DeLong DM, Clarke-Pearson DL (1988) Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 44: 837–845.
62. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, et al. (2011) pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 12: 77.
63. Wu S, Liang M, Altman R (2008) The SeqFEATURE library of 3D functional site models: comparison to existing methods and applications to protein function annotation. Genome Biology 9: R8.
64. Goldenberg O, Erez E, Nimrod G, Ben-Tal N (2009) The ConSurf-DB: pre-calculated evolutionary conservation profiles of protein structures. Nucl Acids Res 37: D323–D327.
65. Mirny LA, Shakhnovich EI (1999) Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. J Mol Biol 291: 177–196.
66. Romero RM, Roberts MF, Phillipson JD (1995) Anthranilate synthase in microorganisms and plants. Phytochemistry 39: 263–276.
67. Knöchel T, Ivens A, Hester G, Gonzalez A, Bauerle R, et al. (1999) The crystal structure of anthranilate synthase from sulfolobus solfataricus: functional implications. Proc Natl Acad Sci USA 96: 9479–9484.
68. Morollo AA, Eck MJ (2001) Structure of the cooperative allosteric anthranilate synthase from salmonella typhimurium. Nat Struct Mol Biol 8: 243–247.
69. Koo CW, Sutherland A, Vederas JC, Blanchard JS (2000) Identification of active site cysteine residues that function as general bases: diaminopimelate epimerase. J Am Chem Soc 122: 6122–6123.
70. Cirilli M, Zheng R, Scapin G, Blanchard JS (1998) Structural symmetry: the three-dimensional dtructure of haemophilus Influenzae diaminopimelate epimerase. Biochemistry 37: 16452–16458.
71. Porter CT, Bartlett GJ, Thornton JM (2004) The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. Nucl Acids Res 32: D129–D133.
72. Andreeva A, Howorth D, Chandonia JM, Brenner SE, Hubbard TJ, et al. (2008) Data growth and its impact on the SCOP database: new developments. Nucl Acids Res 36: D419–D425.
73. Chandonia JM, Hon G, Walker NS, Lo Conte L, Koehl P, et al. (2004) The ASTRAL compendium in 2004. Nucl Acids Res 32: D189–D192.
74. Sing T, Sander O, Beerenwinkel N, Lengauer T (2005) ROCR: visualizing classifier performance in R. Bioinformatics 21: 3940–3941.
75. Csardi G, Nepusz T (2006) The igraph software package for complex network research. InterJournal, Complex Systems: pp 1695.