Accurate sequence-based prediction of catalytic residues

Tuo Zhang1,2, Hua Zhang1,2, Ke Chen2, Shiyi Shen1,3, Jishou Ruan1,3 and Lukasz Kurgan2,∗

1College of Mathematical Science and LPMC, Nankai University, Tianjin, PRC, 2Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB, Canada and 3Chern Institute of Mathematics, Tianjin, PRC

Received on May 5, 2008; revised on August 6, 2008; accepted on August 14, 2008

Advance Access publication August 18, 2008

Associate Editor: Anna Tramontano

ABSTRACT

Motivation: Prediction of catalytic residues provides useful information for the research on function of enzymes. Most of the existing prediction methods are based on structural information, which limits their use. We propose a sequence-based catalytic residue predictor that provides predictions with quality comparable to modern structure-based methods and that exceeds quality of state-of-the-art sequence-based methods.

Results: Our method (CRpred) uses sequence-based features and the sequence-derived PSI-BLAST profile. We used feature selection to reduce the dimensionality of the input (and explain the input) to support vector machine (SVM) classifier that provides predictions. Tests on eight datasets and side-by-side comparison with six modern structure- and sequence-based predictors show that CRpred provides predictions with quality comparable to current structure-based methods and better than sequence-based methods. The proposed method obtains 15–19% precision and 48–58% TP (true positive) rate, depending on the dataset used. CRpred also provides confidence values that allow selecting a subset of predictions with higher precision. The improved quality is due to newly designed features and careful parameterization of the SVM. The features incorporate amino acids characterized by the highest and the lowest propensities to constitute catalytic residues, Gly that provides flexibility for catalytic sites and sequence motifs characteristic to certain catalytic reactions. Our features indicate that catalytic residues are on average more conserved when compared with the general population of residues and that highly conserved amino acids characterized by high catalytic propensity are likely to form catalytic sites. We also show that local (with respect to the sequence) hydrophobicity contributes towards the prediction.

Availability: http://biomine.ece.ualberta.ca/CRpred/CRpred.htm

Contact: ikurgan@ece.ualberta.ca

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

Enzymes are biomolecules (virtually all of them are proteins) that catalyze chemical reactions. Enzymes bind temporarily to reactants and by doing so they lower the amount of activation energy needed which results in speeding up the reaction. Many cellular processes require enzymes in order to occur at significant rates. Only a small number of amino acids that compose the enzyme, which are known as catalytic residues, are directly involved in the reaction and thus are fundamental to the enzyme’s biological functions. Compared with the rapidly increasing volume of protein sequence and structure information, experimental methods used to infer catalytic residues are lagging behind. This motivates development of high-throughput in silico methods for identifying catalytic residues.

To date, several approaches for prediction of protein functional residues that are involved in catalytic reactions have been proposed. One group of such methods is based solely on protein sequence and they use evolutionary information in the form of multiple sequence alignments (Capra and Singh, 2007; Fischer et al., 2008; La et al., 2005; Pande et al., 2007; Sterner et al., 2007). La et al. (2005) demonstrated that phylogenetic motifs, which are sequence alignment fragments that approximate the overall familial phylogeny, are useful for predictions of regions surrounding enzyme active sites. In Pande et al. (2007), the authors show that neural networks can accurately predict catalytic residues using only residue identity and sequence conservation. In the most recent attempt, Fischer and colleagues (2008) proposed a method that improves on the prediction quality when compared with existing sequence-based methods by integrating the information concerning sequence conservation, predicted secondary structure and relative solvent accessibility, and amino acid frequencies. Another group of methods is based on protein structure (Chea et al., 2007; Gutteridge et al., 2003; Ota et al., 2003; Petrova and Wu, 2006; Sacquin-Mora et al., 2007; Torrance et al., 2005; Youn et al., 2007). Structure-based methods usually obtain better prediction accuracy due to the combination of both sequence and structure information, which were shown to be complementary (Gutteridge et al., 2003; Petrova and Wu, 2006; Youn et al., 2007). We emphasize that their application is limited only to proteins with known structure, which constitute only a small fraction of all known proteins. One of the cornerstone methods for prediction of catalytic residues was developed by Thornton’s group (Gutteridge et al., 2003). This group also built the Catalytic Site Atlas (CSA) database which provides catalytic residue annotations for enzymes in Protein Data Bank (Porter et al., 2004). Two most recent structure-based methods include catalytic residue predictor based on closeness centrality measure (Chea et al., 2007) and support vector machine (SVM)-based method that utilizes...
2 MATERIALS AND METHODS

2.1 Datasets
We prepared nine datasets to design and test the proposed method, including six datasets that were used in previous studies. They include SCOP fold dataset (EF fold), SCOP superfamily dataset (EF superfamily) and SCOP family dataset (EF family) from Youn et al. (2007), SCOP superfamily dataset (HA superfamily) from Chea et al. (2007), dataset (PC) from Petrova and Wu (2006) and non-homologous dataset (NN) from Gutteridge et al. (2003). These datasets include sequences at various levels of homology, i.e. one sequence per fold, family and superfamily, and allow for an unbiased comparison with the competing methods. The remaining three datasets include two test datasets (T-124 and T-37) and one design dataset (ST-1109). We use CSA v. 2.2.5 (Porter et al., 2004) to annotate catalytic residues. Table 2 gives an overview of the datasets used for testing.

The ST-1109 dataset is based on all chains which have catalytic residue annotations in CSA v. 2.2.5 that were filtered at 40% sequence identity with CD-HIT (Li and Godzik, 2006) with parameters –c 0.4 –n 2. These 2152 chains were then filtered at 40% sequence identity against the sequences in the above six datasets from Youn et al. (2007), Chea et al. (2007), Petrova and Wu (2006) and Gutteridge et al. (2003) with CD-HIT-2D (Li and Godzik, 2006) (parameters: –c 0.4 –n 2 –s2 0.3) to avoid overlap with the sequence used for testing. The remaining 1109 chains constitute the design dataset.

The remaining two datasets were used to compare with competing methods and to present results using tests on an independent dataset, rather than based on cross-validation which was applied on EF fold, EF superfamily, EF family, HA superfamily, PC and NN datasets. The T-124 dataset was built by selecting chains in the HA superfamily dataset (large dataset independent of the contribution that introduced the EF fold dataset) that have sequence identity <30% when compared with the chains in both EF fold and ST-1109; the latter two datasets were used to design the method and to train the classification model. To compare with the method by Fischer et al. (2008), we further filtered the T-124 dataset to remove chains that have sequence identity ≥30% with respect to the chains in CSA-cat dataset used in Fischer et al. (2008). Among the selected 40 chains, three were removed since they produced errors when executed on the server that implements the method by Fischer et al. (2008); the remaining 37 chains constitute T-37 dataset.

We use the whole protein sequence as an input, including residues which have no coordinate information. The residues with the missing information are considered as non-catalytic residues. While this allows the users to enter the sequence without the necessity to verify whether all its residues have coordinates in the PDB file (some competing methods predict/ were tested only for residues with the coordinate information), it also increases the difficulty of the prediction. This is since the number of actual catalytic residues does not change, while we may risk making additional false positive predictions due to the increased number of non-catalytic residues. The predictions are evaluated based on true positive rate and precision computed for catalytic residues, which are larger when more true positives and fewer false positives are predicted and which are independent of true negative predictions. As shown in Table 2, the datasets are heavily imbalanced, with the ratio of catalytic to non-catalytic residues varying between 1 : 80 and 1 : 128. Using the original data result in a strong bias toward prediction of all residues as non-catalytic; thus we undersample the non-catalytic residues to create the training data and keep the original ratio for the test data. We use 1 : 6 ratio by randomly selecting six non-catalytic residues for one catalytic residue in each training sequence, which is consistent with the sampling performed in Gutteridge et al. (2003) and Youn et al. (2007).

2.2 Overview of the prediction system
The proposed system extracts five sets of features based on (1) position-specific scoring matrix (PSSM) values, (2) entropy computed from weighted observed percentages (EntWOP), both of which are generated with PSI-BLAST (Altschul et al., 1997), (3) residue type (ResType), (4) average cumulative hydrophobicity (AveCH) and (5) catalytic residue pairs (CRPair) that were developed using the ST-1109 dataset. The latter three sets are generated directly from a protein sequence. Total of 210 features, which were selected among 544 considered features, are fed into the SVM classifier to predict catalytic residues, see Figure 1.
Table 2. Summary of the datasets used to validate the proposed method

<table>
<thead>
<tr>
<th>Reference</th>
<th>EF fold</th>
<th>EF superfamily</th>
<th>EF family</th>
<th>HA superfamily</th>
<th>NN</th>
<th>PC</th>
<th>T-124</th>
<th>T-37</th>
</tr>
</thead>
<tbody>
<tr>
<td>Catalytica</td>
<td>606</td>
<td>712</td>
<td>1183</td>
<td>933</td>
<td>599</td>
<td>283</td>
<td>379</td>
<td>130</td>
</tr>
<tr>
<td>Non-catalybc</td>
<td>48362</td>
<td>62236</td>
<td>107284</td>
<td>98130</td>
<td>56309</td>
<td>24499</td>
<td>48634</td>
<td>14501</td>
</tr>
</tbody>
</table>

aNumber of catalytic residues in the dataset.
bNumber of non-catalytic residues in the dataset.
cRatio between catalytic and non-catalytic residues.

As a result, we obtain 21 * 20 = 420 PSSM features and 21 EntWOP features for each predicted residue. In case of residues at the sequence termini, we use 0’s to fill in blanks in both PSSM and WOP vectors.

2.4 Feature-based sequence representation

2.4.1 ResType The 20 amino acids are characterized by different propensities to constitute catalytic residues (Bartlett et al., 2002). We use binary encoding to represent the type of a predicted residue, i.e. the residue corresponding amino acid is set to 1 and the remaining positions equal 0.

2.4.2 PSSM and EntWOP Several studies have shown that sequence conservation is important for catalytic residue prediction (Capra and Singh, 2007; Fischer et al., 2008; Petrova and Wu, 2006; Youn et al., 2007). Therefore, we designed two types of features based on PSI-BLAST profiles to incorporate sequence conservation information. We extracted the PSSM and weighted observed percentages (WOP) using PSI-BLAST with parameter \( \text{–j } 3 \), i.e. 20-dimensional PSSM and 20-dimensional WOP vectors are obtained for each residue.

The PSSM vector for a given residue represents the log-likelihood of the substitution of 20 amino acids at that sequence position (Jones, 1999). Each PSSM value \( x \) in the vector indicates the degree of conservation of a given amino acid type for that residue and is normalized by \( 1/(1+e^{-x}) \). We use a sliding window of size 21 as suggested in Youn et al. (2007) to extract the PSSM features.

Since the WOP vector for a given residue represents a frequency distribution of 20 amino acids at that sequence position (Altschul et al., 1997), we introduce EntWOP by computing Shannon entropy:

\[
\text{EntWOP} = \sum_i -p_i \log(p_i)
\]

where \( p_i = n_i / \Sigma n_i \), \( i = 1, 2, \ldots, 20 \), and \( (n_1, n_2, \ldots, n_{20}) \) is the WOP vector. EntWOP ranges between 0 (the most conserved; only one amino acid type has non-zero value at the corresponding position in the WOP vector) and 2.996 (the least conserved; all 20 amino acids have the same non-zero value in the WOP vector).

As a result, we obtain 21 * 20 = 420 PSSM features and 21 EntWOP features for each predicted residue. In case of residues at the sequence termini, we use 0’s to fill in blanks in both PSSM and WOP vectors.

2.4.3 Average cumulative hydrophobicity Bartlett et al. (2002) indicates that most of catalytic residues have limited exposure to solvent. Since hydrophobicity index can be used to quantify the intrinsic propensity of an amino acid to be exposed to solvent and since it was shown to provide useful input for prediction of DNA/RNA binding sites (Wang and Brown, 2006), we introduce AveCH features by computing the average of the cumulative hydrophobicity index over a sliding window, varying window size between 3 and 21. We applied Eisenberg’s hydrophobicity index (Sweet and Eisenberg, 1983) due to its superior performance when compared with other indices (Juretic and Lucin, 1998; Kurgan et al., 2007). We used 0’s to fill in blanks for residues at the sequence termini. Total of 10 AveCH features are computed.

2.4.4 Catalytic residue pairs We also built several features based on sequence motifs associated with catalytic residues. A CRPair is defined as \( \text{CRPair}_n = [A, D] \) where \( n = 2, 3, \ldots \), \( A = \{A_1, A_2, \ldots, A_n\} \) denotes \( n \)-adjacent catalytic residues in a sequence with direction from N-terminus to C-terminus, and \( D = \{d_1, d_2, \ldots, d_n-1\} \) denotes the distances between each two adjacent catalytic residues. We computed the occurrence of each \( \text{CRPair}_n \) using the ST-1109 dataset and we selected 73 \( \text{CRPair}_n \) (including 47 \( \text{CRPair}_{12} \), 21 \( \text{CRPair}_{13} \) and 5 \( \text{CRPair}_{14} \)) that occur more than eight times in this dataset. For each predicted residue, we computed 73 binary features that correspond to the selected 73 \( \text{CRPair}_n \) i.e. value of 0 represents the fact that the a given residue is associated with a certain \( \text{CRPair}_n \), which means that this residue is in set A and the remaining \( n-1 \) residues in set A can be found in sequence according to the distance values in \( D \), and value of 1 is used otherwise.

Total of 544 features (20 ResType, 420 PSSM, 21 EntWOP, 10 AveCH and 73 CRPair) were generated for each predicted residue.

2.5 Design

The design of the proposed predictor concerns selection of a subset of the proposed features and parameterization of the SVM classifier. Feature selection reduces the dimensionality, which decreases the computational time and complexity of the prediction model, and allows finding factors (encoded as features) associated with the prediction of catalytic residues.

2.5.1 Feature ranking In the first step, we ranked the 544 features based on their contribution to the prediction of catalytic residues. We use \( \chi^2 \)-statistic to perform ranking (Liu and Setiono, 1995). The \( \chi^2 \)-values were computed with 10-fold cross-validation on EF-fold dataset, i.e. average over the 10-fold was computed to avoid overfitting, and features that give higher average \( \chi^2 \)-value receive higher rank. Usage of this method was motivated by recent research that shows that it performs well if the goal of the subsequent classification is to improve precision (Forman, 2003), which is the case for
The selected SVM parameters \((C, \gamma)\) for each \(n\) are shown near the ACI point. We show ACI values for \(n \geq 90\), since values for smaller \(n\) were low and would distort the plot.

2.5.2 Parameter selection Parameterization of the SVM aims at improving the prediction quality. Values of \(C\) and \(\gamma\) were optimized based on the top \(n\) selected features. After several initial tests that allowed establishing an approximate range of well-performing \(C\) and \(\gamma\) values, we executed a grid search over a Cartesian product of these ranges. The values of parameter \(C\) were chosen from \([1.0, 1.5, 2.0]\) and for parameter \(\gamma\) from \([0.01, 0.02, \ldots, 0.1]\). The grid search involved 30 10-fold cross-validations on EF fold dataset each time with a different set of \((C, \gamma)\) values. We measured two quality indices

\[
\text{TP rate} = \frac{TP}{(TP + FN)} \quad \text{and} \quad \text{precision} = \frac{TP}{(TP + FP)}
\]

where TP, FP and FN denote true positives, false positives and false negatives, respectively. These indices are commonly used to evaluate catalytic residue prediction systems (Chea et al., 2007; Fischer et al., 2008; Gutteridge et al., 2003; Youn et al., 2007). We selected the parameters for which precision equals at least 17% and which provide the highest TP rate. The minimal precision equals the precision obtained in Youn et al. (2007) and Chea et al. (2007). We note that \(C\) values above 2.0/\(\gamma\) below 0.1 would lead to low precision/poor TP rate.

2.5.3 Feature selection Feature selection was performed using 10-fold cross-validation on the EF fold, EF superfamily, EF family, HA superfamily and NN datasets with the top \(n\) features and the selected SVM parameters. We compute average cumulative improvement (ACI) by summing up the improvements, i.e. the difference between the results of the proposed method and the corresponding results reported for a given dataset in Gutteridge et al. (2003), Youn et al. (2007) and Chea et al. (2007) of the two quality indices and next averaging the sum by the number of the datasets. Figure 2 shows the ACI values with respect to different values of \(n\). We select \(n = 210\) that corresponds to the largest value of ACI. We observe two large dips in ACI values for \(n = 220\) and \(n = 270\); the reason is that the SVM parameters (relatively low \(\gamma\)-value) selected on the EF-fold dataset did not translate well into the other four datasets.

2.6 Conservation score

We compute a baseline prediction for the eight benchmarking datasets (EF fold, EF superfamily, EF family, HA superfamily, NN, PC, T-124 and T-37), based on three simple conservation score-based approaches, SP score (Valdar, 2002), Rate4site (Mayrose et al., 2004) and information per position (IPP) (Karypis, 2006; Youn et al., 2007). These three scores are based on multiple sequence alignment (MSA). IPP comes directly from the output of PSI-BLAST, while both SP score and Rate4site use MSAs, which are extracted using PSI-BLAST with search parameters \(-j 3-h 0.001-e 0.001\) against the NCBI non-redundant (NR) database (Martin et al., 2006), as the input.

3 RESULTS AND DISCUSSION

3.1 Comparison with competing structure- and sequence-based methods

The proposed method (CRpred) is compared with four structure-based methods and one sequence-based method using 10-fold cross-validation on six datasets; see Table 3. To facilitate comparison using two quality indices, we report the TP rate and precision obtained by the competing methods and by CRpred (top four rows), together with TP rate values obtained by CRpred at precision equal to the precision of a given competing method (fifth row), CRpred’s precision obtained at equal TP rate (sixth row) and precision of a baseline prediction with Rate4site at TP rate equal to the TP rate of CRpred (seventh row). The values in the last three rows are computed by adjusting the threshold with respect to the probability of predicting a given residue as a catalytic residue provided by the proposed SVM, and with respect to the Rate4site’s conservation score. The Rate4site is used as the baseline due to its favorable quality when compared with the other two baseline conservation scores (Fig. 3). The TP rate of Rate4site at equal precision is not provided, since the baseline method cannot obtain precision as high as that reported for CRpred.

When compared with the method by Youn et al. (2007), CRpred’s TP rates at equal precision for the EF fold and EF superfamily datasets are about 2–3% lower and for the EF family dataset are 1% higher, and precision values at equal TP rates are 1% lower for the EF fold and EF superfamily datasets and 1% higher for the EF family dataset. When compared with the method by Chea et al. (2007), CRpred obtains 20.4% better TP rate and 8.2% better precision at equal precision and equal TP rate, respectively. The structure-based method without spatial clustering by Gutteridge et al. (2003) is characterized by lower TP rate (by 9.9%) and lower precision (by 4.0%) for measurements at equal precision and TP rate, respectively. The structure-based method with spatial clustering (Gutteridge et al., 2003) has 7.2% higher TP rate and 3.5% higher precision, when compared at equal precision and TP rate with CRpred, respectively. The method by Petrova and Wu (2006) obtains 5.5% better TP rate at equal precision and 1.4% higher precision at equal TP rate. Finally, the TP rate and precision of the sequence-based method by Gutteridge et al. (2003) are lower by 16.9% and 6.3% at the same precision and TP rate, respectively. CRpred shows a consistent improvement roughly doubling the baseline precision at equal TP rate for all six datasets. Overall, the results indicate that CRpred provides predictions of quality comparable to predictions of the considered structure-based methods (only the method that applies spatial clustering provides a clear improvement) and which are better than predictions of the included sequence-based method.

We observe that although the HA superfamily and EF superfamily datasets are characterized by the same level of homology, CRpred’s predictions show differences in precision. Although the reason for the 2% decrease for the HA superfamily dataset is unknown (we hypothesize that this could be due to more imbalanced nature of this
Table 3. Comparison with competing methods that include four structure-based methods and one sequence-based method on six datasets

<table>
<thead>
<tr>
<th>Method</th>
<th>Reported index</th>
<th>Structure-based competing methods</th>
<th>Sequence-based prediction of catalytic residues</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>EF fold</td>
<td>EF super family</td>
</tr>
<tr>
<td>Competing methods</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CRpred</td>
<td>TP rate</td>
<td>51.1a</td>
<td>53.9a</td>
</tr>
<tr>
<td></td>
<td>Precision</td>
<td>17.1a</td>
<td>16.9a</td>
</tr>
<tr>
<td>Equal precision</td>
<td>TP rate</td>
<td>48.2</td>
<td>52.1</td>
</tr>
<tr>
<td></td>
<td>Precision</td>
<td>17.0</td>
<td>17.0</td>
</tr>
<tr>
<td>Equal to TP</td>
<td>Precision</td>
<td>16.1</td>
<td>15.9</td>
</tr>
<tr>
<td>Rate4site</td>
<td>Precision</td>
<td>9.3</td>
<td>9.4</td>
</tr>
</tbody>
</table>

*Table 3. Comparison with competing methods that include four structure-based methods and one sequence-based method on six datasets*

<table>
<thead>
<tr>
<th>Method</th>
<th>TP rate</th>
<th>FN</th>
<th>FP</th>
<th>TN</th>
<th>TP rate</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRpred</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CRpred</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Table 4. Summary of predictions on T-124 dataset*

3.2 Test on T-124 dataset

Since the performed feature selection includes use of the five datasets from Section 3.1, we also provide tests on an independent dataset. We generate the prediction model by training on the entire EF fold dataset (dataset with the lowest homology) using the selected 210 features and SVM parameters, and next we test this model on the T-124 dataset. The latter dataset is characterized by low pairwise sequence identity (<30%) with respect to the EF fold and ST-1109 datasets that were used for training the CRpred model. We compare these results with the structure-based HA method by Chea et al. (2007) to establish a point of reference. The HA method was computed based on two different filters, residue identity and combination. Since the competing method works only for residues that have coordinate information, we compute the CRpred’s quality indices based on (1) all residues and (2) based only on the residues that have the coordinate information.

As shown in Table 4, CRpred’s results computed for all residues have the same TP rate and slightly lower precision when compared with the results for residues that have complete coordinate information, which agrees with the discussion in Section 2.1. The results (TP rate = 50.1% and precision = 14.4%) are consistent with the results (TP rate = 54.0% and precision = 14.9%) shown in Table 3. The small decrease is likely due to the smaller size of the T-124 dataset; we emphasize that a similar decrease is observed for the HA method (TP rate = 27.7% and precision = 16.1% for the T-124 dataset, and TP rate = 29.3% and precision = 16.5% for the HA superfamily dataset). When compared with the HA method, the proposed method provides a large improvement of the TP rate as a tradeoff for a small decrease of precision. However, HA method is based on a relatively simple model, which in contrast to other considered methods does not use alignment conservation.

3.3 Comparison with FRPred method

The comparison with the most recent sequence-based method by Fischer et al. (2008) is based on the T-37 dataset, which has low pairwise sequence identity (<30%) with the two training datasets of CRpred (EF fold and ST-1109) and with the training dataset used.

![Fig. 3. TP rate versus FP rate for CRpred, FRpred and three baseline predictions (IPP, SP score and Rate4site) on two datasets: (A) EF fold and (B) T-37. Based on Fischer et al. (2008), FP rate was constrained to (0, 0.05).](https://academic.oup.com/bioinformatics/article-abstract/24/20/2329/259116/fig3.png)
by Fischer et al. (2008). The chains in the T-37 were predicted with CRpred built in Section 3.2, web server (FRPred) provided by Fischer et al. (2008) and using the HA method by Chea et al. (2007). CRpred obtained TP rate = 52.3% and precision = 18.8% and the HA method obtained TP rate = 25.4% and precision = 17.2%. Since the results of the FRPred can be adjusted using a threshold, we choose two configurations, one that provides the same TP rate as obtained by the CRpred and the other with the same precision. FRPred obtains precision = 18.8% with TP rate = 26.2%, and TP rate = 52.3% with precision = 10.7%. We observe that in both cases CRpred provides improvements over the FRPred method.

Since the T-37 dataset is relatively small, we also use the EF fold dataset. We compare predictions provided by the FRPred server (10 sequences in EF fold generated errors when executed on the FRPred and thus were excluded) with results of CRpred based on 10-fold cross-validation on this dataset. At 17.0% precision, CRpred’s TP rate = 48.2%, which is about 8% higher than the TP rate of FRPred (40.5%). At TP rate = 48.2%, CRpred obtains precision = 17.0%, which is 2% higher than the precision of FRPred (15.0%). We again observe that the proposed method provides more accurate predictions than FRPred, even though in case of FRPred the results could be overestimated due to an overlap between the EF fold dataset and the training dataset of FRPred. Using CD-HIT (parameters: -c 0.4 -n 2), we found 60 chains (about one-third of sequences in the EF fold) with sequence identity > 95% when compared with the sequences in the training dataset of FRPred, and 77 chains with sequence identity > 40%. Although the results suggest that CRpred is characterized by improved prediction quality, we note that FRPred is a more generic method that can also provide ligand-binding residue prediction, while CRpred is specific to prediction of catalytic residues.

Figure 3 shows the ROC curves for CRpred, FRPred and the three baseline predictions (IPR, SP score and Rate4site) on the EF fold and T-37 datasets. The FP rate is constrained to (0.05) range; the same assumption was made in Fischer et al. (2008). The corresponding global ROC curves (that cover entire range of FP rate values) and precision versus TP rate curves, which were introduced and used in Fischer et al. (2008) for all eight benchmarking datasets, are provided in the Supplementary Material. The proposed method is shown to perform better than Rate4site, which is the best performing among the baseline predictors, on both datasets.

Test on the EF fold dataset shows that CRpred and FRPred provide favorable predictions when compared with the three baselines, except for FRPred that performs slightly worse than Rate4site on the T-37 dataset. When compared with FRPred, CRpred provides higher TP rate when FP rate ≤ 0.07, while FRPred obtains higher TP rate when FP rate > 0.07. However, since the datasets are highly unbalanced, i.e. catalytic residues (positive samples) constitute < 1.5% of residues (Table 2), FP rates of above 0.07 result in relatively low precision. For instance, CRpred’s precision of 17%, which equals precision obtained in Youn et al. (2007) and Chea et al. (2007), corresponds to FP rate of 0.06 for the EF-fold dataset. Therefore, we conclude that CRpred constitutes an improvement over FRPred in the case when relatively high precision is required. We note that results of FRPred on the EF fold dataset could be overestimated since training set of this method overlaps with this test set, while CRpred’s results are based on 10-fold cross-validation. Test on the T-37 dataset, which has low similarity with the training sets of both CRpred and FRPred, shows that our predictor provides favorable quality of predictions when considering the entire range of precision.

3.4 Confidence values

The proposed CRpred method supplements the prediction of catalytic/non-catalytic residues with an estimate of a probability of predicting a given residue as a catalytic residue, which is derived from the SVM model (Fan et al., 2005). This probability estimate constitutes a confidence value, i.e. the higher the confidence value, the more credible is the corresponding prediction. Removal of the positive (catalytic residue) predictions with lower confidence values should lead to reduction of the false positive predictions, which in turn should provide better precision.

We use 10-fold cross-validation test results on the EF fold dataset and results of the test on the T-124 dataset to validate whether precision can be improved with the help of the confidence values. We rank all predicted positives (residues predicted as catalytic residues) by their confidence values. Next, we select a given percentage of the top ranked predictions as the positive predictions, while the remaining predictions are regarded as non-catalytic residues. We vary the threshold between 5% and 100% with a step of 5%, see Figure 4. Selection of the top 5% predicted positives results in the highest precision = 33%, however, only 10% of the true positives were included among the predicted positives (TP rate = 4.8%). Increase of the threshold results in lower precision value and better TP rate. We observe that the slope of TP rate curve is higher for lower values of the threshold and it gradually decreases as the threshold increases. At the same time, the slope of the precision is approximately constant, which indicates that the confidence values help in obtaining better precision as a tradeoff for reduced TP rate. For instance, when selecting the top 75% of predicted positives, precision = 20.1% is obtained with TP rate = 42.9%, compared with precision = 17.0% and TP rate = 48.2% when all predictions are considered. This shows that 3.1% more true positives that in fact constitute catalytic residues are obtained when predictions cover 5.3% less of the actual catalytic residues. We also plot inter_TP rate, which is defined as the number of true positives divided by the number of all true positives in predicted positives that follows similar trend as the TP rate. We note that similar relations are obtained for both EF fold and T-124 datasets.
An example application of the confidence values is summarized in Table 5. We rank all residues in the T-124 dataset by their confidence values and select three proteins which contain the top three ranked residues. When selecting the predicted positives (including true and false positives) that have confidence value >0.6 (which corresponds to the top 90% predicted positives), three false positives are removed while the number of true positives is not affected.

We observe that some false positives may have high confidence values, some of which are higher than the confidence of true positives. However, a false positive could correspond to a catalytic residue which is not yet annotated in CSA database, or to a residue that is related to the function of an enzyme and which may be a (missing) part of an annotated catalytic site. In order to analyze these false (unverified) positives, we compute the distance between them and the actual positives (including true positives and false negatives) in each protein by finding the minimal distance between atoms on the two residues. Some of the false positives are in fact close to the actual positives. For instance, the false (unverified) positive 100Gly in 1A7UA, 291Tyr and 21His, 1NVMG, could be in the catalytic site formed by 291Tyr and 21His.

3.5 Analysis of SVM parameterization and selected features

We compare a linear kernel with the RBF kernel that was used in EntWOP. When testing these two kernels using all 544 features and default parameters (default RBF kernel has $C=1.0$ and $\gamma=0.01$; default linear kernel has $C=1.0$) on the six benchmark datasets (EF fold, EF superfamily, EF family, HA superfamily, NN and PC), we find that the results based on linear kernel have a higher TP rate (on average by 12%) but much (twice) lower precision. Similar results are found while using the selected 210 features. This indicates that RBF kernel provides higher precision and still relatively high TP rate, while linear kernel gives higher TP rate as a tradeoff for below-standard precision. The comparison between the CRpred and the results where all 544 features and the default RBF kernel are used show that CRpred obtains on average 9.6% improvement of TP rate (on average by 12%) but much (twice) lower precision. Similar results are found while using the selected 210 features. This indicates that RBF kernel provides higher precision and still relatively high TP rate, while linear kernel gives higher TP rate as a tradeoff for below-standard precision. The comparison between the CRpred and the results where all 544 features and the default RBF kernel are used show that CRpred obtains on average 9.6% improvement of TP rate.

Table 5. Summary of the prediction results for three proteins that include predictions with the three highest confidence values in the T-124 dataset

<table>
<thead>
<tr>
<th>PDB id</th>
<th>Position 1</th>
<th>Type</th>
<th>Confidence</th>
<th>Position 2</th>
<th>Type</th>
<th>Confidence</th>
<th>Position 3</th>
<th>Type</th>
<th>Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1A7UA</td>
<td>98</td>
<td>Ser</td>
<td>0.995</td>
<td>57</td>
<td>Asp</td>
<td>0.946</td>
<td>99</td>
<td>Met</td>
<td>0.471</td>
</tr>
<tr>
<td></td>
<td>257</td>
<td>His</td>
<td>0.959</td>
<td>30</td>
<td>His</td>
<td>0.877</td>
<td>32</td>
<td>Phe</td>
<td>0.381</td>
</tr>
<tr>
<td></td>
<td>228</td>
<td>Asp</td>
<td>0.604</td>
<td>100</td>
<td>Gly</td>
<td>0.672</td>
<td>140</td>
<td>Ala</td>
<td>0.049</td>
</tr>
<tr>
<td></td>
<td>97</td>
<td>Phe</td>
<td>0.587</td>
<td>141</td>
<td>Pro</td>
<td>0.037</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2PLC</td>
<td>45</td>
<td>His</td>
<td>0.993</td>
<td>79</td>
<td>Arg</td>
<td>0.832</td>
<td>84</td>
<td>Arg</td>
<td>0.408</td>
</tr>
<tr>
<td></td>
<td>46</td>
<td>Asp</td>
<td>0.842</td>
<td>119</td>
<td>Glu</td>
<td>0.78</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>278</td>
<td>Asp</td>
<td>0.678</td>
<td>66</td>
<td>Gln</td>
<td>0.596</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>93</td>
<td>His</td>
<td>0.646</td>
<td>236</td>
<td>His</td>
<td>0.555</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1NVMG</td>
<td>21</td>
<td>His</td>
<td>0.774</td>
<td>202</td>
<td>His</td>
<td>0.992</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>291</td>
<td>Tyr</td>
<td>0.648</td>
<td>17</td>
<td>Arg</td>
<td>0.973</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>200</td>
<td>His</td>
<td>0.957</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>18</td>
<td>Asp</td>
<td>0.904</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>203</td>
<td>His</td>
<td>0.861</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>170</td>
<td>Asp</td>
<td>0.784</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>204</td>
<td>Asn</td>
<td>0.721</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>13</td>
<td>Asp</td>
<td>0.691</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>171</td>
<td>Ser</td>
<td>0.671</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>236</td>
<td>Asn</td>
<td>0.635</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

aFalse positives (some of these predictions could correspond to not yet annotated catalytic residues).
bResidue number in the PDB file.
cResidue type.
dConfidence value.

Fig. 5. Distribution of the number of residues constituting catalytic sites; x-axis shows the number of residues in a catalytic site; y-axis shows the count of the corresponding sites (the numbers above the bars are the actual count).

Table 6. Analysis of contributions of new features based on 10-fold cross-validation on 6 datasets

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>TP Rate</th>
<th>Precision</th>
<th>New Features Removed</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRpred</td>
<td>48.2</td>
<td>16.8</td>
<td></td>
</tr>
<tr>
<td>CRpred</td>
<td>48.2</td>
<td>16.8</td>
<td></td>
</tr>
<tr>
<td>CRpred</td>
<td>48.2</td>
<td>16.8</td>
<td></td>
</tr>
<tr>
<td>CRpred</td>
<td>48.2</td>
<td>16.8</td>
<td></td>
</tr>
<tr>
<td>CRpred</td>
<td>48.2</td>
<td>16.8</td>
<td></td>
</tr>
<tr>
<td>CRpred</td>
<td>48.2</td>
<td>16.8</td>
<td></td>
</tr>
</tbody>
</table>

aCRpred model using 210 features.
bPredictions when EntWOP, AveCH and CRPair features were removed from the 210 features.
3.5.1 Relation between residue conservation and catalytic residues

Among the 21 EntWOP features, nine that represent the entropy of the central and four immediately adjacent residues on both sides are selected. The feature corresponding to the central residue has the largest score, and the scores decrease with the increasing distance. The distribution of an example selected EntWOP feature, EntWOP-0, is shown in Figure 6. We observe that catalytic residues tend to have lower entropy values, i.e. they are more conserved, while the non-catalytic residues are skewed toward higher values. Hence, EntWOP features help to distinguish between catalytic and non-catalytic residues based on residue conservation. This is measured irrespective of the residue type, which contrasts these features from the PSSM features.

Table 8 summarizes 166 PSSM features selected from the 420 values from the PSSM. The count and the average score (higher score corresponds to higher ranked feature) of selected features for each position in the window show that most of the top ranked selected PSSM values are close to the center of the window. The selected features are asymmetrical with respect to the central position in the window. They have a preference towards N-terminus positions, as most of them occupy positions between +4 and −7. At the moment, we have no explanation for this skewed distribution. The count and the average score for the residue types show consistency with respect to their catalytic propensity defined in Bartlett et al. (2002). The residue types with higher average scores have either high or low catalytic propensities. More specifically, the top 12 highest scored residue types include six out of the seven residue types with the highest catalytic propensities and five out of the six residue types with the lowest propensities. Nine out of the top 10 selected features are based on PSSM values of two amino acids (Cys and His) with the highest catalytic propensity and three amino acids (Val, Ile, and Ala) with the lowest catalytic propensity (Bartlett et al., 2002). Figure 7 shows catalytic propensity of residues with a given range of PSSM-0-V and PSSM-0-C values (these two features have the highest scores among the residues with the lowest and the highest propensities). We observe that residues with larger PSSM-0-C values, i.e. conserved with respect to Cys, have higher propensity to be catalytic, while residues with larger PSSM-0-V values, i.e. conserved with respect to Val, have relatively low propensity. This shows that the highly conserved amino acids characterized by high catalytic propensity are likely to form catalytic sites.

3.5.2 Relation between residue type and catalytic residues

Different residue types have different propensity towards formation of catalytic sites (Bartlett et al., 2002). Thirteen out of the 20 features that represent different residue types are selected. Six of them (His, Cys, Asp, Arg, Glu and Tyr) concern amino acids characterized by the highest propensities to constitute catalytic residues, another six (Val, Ala, Ile, Pro, Leu, and Met) by the lowest propensities (Bartlett et al., 2002), while the remaining Gly was suggested to provide flexibility for enzyme active sites (Yan and Sun, 1997). This way...
This motif is shown to be essential for catalysis of redox reactions (Bartlett et al., 2002), i.e. percentage of catalytic residues with PSSM values in a given range among all residues divided by the percentage of all residues for the same PSSM range among all residues in the EF fold dataset.

CRPair2 is a special case of the former one, and the corresponding CysXXCys motif constitutes a strong pattern associated with catalytic residues. This motif is shown to be essential for catalysis of redox reactions (Chivers et al., 1997). In another example, selected feature \{[Asp,Lys,Asn],1,2\} serves as the catalytic loop of kinase which is important to the catalysis of the phosphate transfer to the substrate (Stegert, 2005).

### 4 CONCLUSIONS

We propose an accurate sequence-based method (CRpred) for the prediction of catalytic residues. CRpred uses SVM classifier that takes five types of interpretable features, including (1) residue type, (2) PSSM values, (3) entropy computed over WOP vector that are extracted with PSI-BLAST, (4) average cumulative hydrophobicity and (5) several sequence motifs, as the input. We perform feature selection which reduces the dimensionality of the input and allows for investigation into the relations between the input features and the prediction of catalytic residue. The most important factor that contributes towards accurate predictions is the residue conservation. Catalytic residues, irrespective of their types, tend to be more conserved when compared with the general population of residues. We show that highly conserved amino acids characterized by high catalytic propensity are likely to form catalytic sites. We also show that amino acids characterized by the highest (His, Cys, Asp, Arg, Glu and Tyr) and the lowest (Val, Ala, Ile, Pro, Leu and Met) propensities to constitute catalytic residues, Gly that is known to provide flexibility for catalytic sites, and certain sequence motifs that are associated with catalytic reactions contribute to the prediction. Our results suggest that although catalytic residues prefer a relatively more hydrophobic neighborhood, they are likely to be locally (with respect to the sequence) surrounded by hydrophilic residues. We also introduce confidence values that allow selection of a subset of predictions with increased precision as a tradeoff for reduced TP rate.

Based on comprehensive tests that include eight datasets, CRpred shows comparable quality when contrasted with the modern structure-based methods and provides improved quality with respect to the state-of-the-art sequence-based methods. CRpred is characterized by 15–19% precision and 48–58% TP rate depending on the dataset used. When compared with several recent structure-based methods, CRpred obtains similar result to methods by Youn et al. (2007) and Petrova and Wu (2006), and it improves both TP rate and precision with respect to the method by Chea et al. (2007). Comparison with methods introduced in Gutteridge et al. (2003) shows that although their method based on spatial clustering provides better quality, CRpred shows improvement when contrasted with the method that does not utilize the clustering. Comparison with two sequence-based methods (Fischet et al., 2008; Gutteridge et al., 2003) reveals that the proposed method obtains better results with respect to both TP rate and precision.

### ACKNOWLEDGEMENTS

We thank Dennis Livesay, Eunseog Youn and Alex Gutteridge for providing their datasets, which were supplemented with helpful explanations. We are particularly grateful to Dr Livesay for clarifications concerning his prediction method.

Funding: National Education Committee of China (to T.Z. and H.Z.); NSFC (grant 10671100 to S.S. and J.R.); Liuhui Center for applied mathematics (to S.S. and J.R.); the joint program of Tianjin and Nankai Universities (to S.S. and J.R.); NSERC (to L.K.); Alberta Ingenuity (to K.C.).
Conflict of Interest: none declared.

REFERENCES