Amino Acids (2010) 39:1353–1361
DOI 10.1007/s00726-010-0587-2

ORIGINAL ARTICLE

# Prediction of catalytic residues based on an overlapping amino acid classification

**Yongchao Dou · Xiaoqi Zheng · Jialiang Yang · Jun Wang**

Received: 18 October 2009 / Accepted: 27 March 2010 / Published online: 10 April 2010


**Abstract** Protein sequence conservation is a powerful and widely used indicator for predicting catalytic residues from enzyme sequences. In order to incorporate amino acid similarity into conservation measures, one attempt is to group amino acids into disjoint sets. In this paper, based on the overlapping amino acids classification proposed by Taylor, we define the relative entropy of Venn diagram (RVD) and RVD2. In large-scale testing, we demonstrate that RVD and RVD2 perform better than many existing conservation measures in identifying catalytic residues, especially than the commonly used relative entropy (RE) and Jensen–Shannon divergence (JSD). To further improve RVD and RVD2, two new conservation measures are obtained by combining them with the classical JSD. Experimental results suggest that these combination measures have excellent performances in identifying catalytic residues.

**Keywords** Catalytic sites prediction · Sequence conservation · Stereochemical properties · Overlapping sets · Combination measure

Y. Dou
School of Mathematical Science, Dalian University of Technology, Dalian 116024, People's Republic of China
URL: http://cast.dlut.edu.cn/Materials/ycdou/index.html

Y. Dou
College of Advanced Science and Technology, Dalian University of Technology, Dalian 116024, People's Republic of China

J. Wang
Scientific Computing Key Laboratory of Shanghai Universities, Shanghai 200234, People's Republic of China

J. Yang
MPI-Institute of Computational Biology, CAS, Shanghai 200031, People's Republic of China

X. Zheng · J. Wang (✉)
Department of Mathematics, Shanghai Normal University, Shanghai 200234, People's Republic of China
e-mail: jwang@shnu.edu.cn

## Introduction

Catalytic residues are essential for maintaining enzyme activity, so they are relatively more conserved than other residues in evolutionary history (Valdar 2002). Thus, the degree of conservation is an important indicator and widely used tool to predict catalytic residues from enzyme sequences. Actually, the conservation information may be used straightforward (Capra and Singh 2007; Donald and Shakhnovich 2005; Mirny and Shakhnovich 1999; Wang and Samudrala 2006), or in conjunction with phylogenetic motif, phylogenetic tree or protein structure information in predicting catalytic sites (Ye et al. 2008; Mihalek et al. 2004; David et al. 2005; Innis et al. 2004; Dukka and Dennis 2008; Panchenko et al. 2003). Additionally, it is also a key feature for machine-learning based methods, such as the Support Vector Machine (SVM) and the Neural Network (NN) methods (Petrova and Wu 2006; Zhang et al. 2008a; Tang et al. 2008; Pande et al. 2007; Gutteridge et al. 2003; Youn et al. 2007).

Usually, the first step to calculate the conservation index for a target chain is to extract amino acid position-specific distributions (PSDs). In most cases, these distributions are extracted from the multiple sequence alignment (MSA) of the protein and its homologous sequences (Capra and Singh 2007). And thus, the quality of these PSDs rely highly on that of the alignment. In the paper, PSDs are also extracted

from the weighted observed percentages (WOP) in the corresponding position-specific scoring matrix (PSSM). After the PSDs are extracted, many computational methods can be used to estimate protein sequence conservation (Pei and Grishin 2001; Caffery et al. 2004; Wang and Samudrala 2006; Sterner et al. 2007; del sol mesa et al. 2003; Reva et al. 2007; Mihalek et al. 2007; Merkl and Zwick 2008; Zhang et al. 2008b; Liu and Guo 2008; Dou et al. 2009a, b). A group of them consider amino acids as symbols in a uniformly diverse alphabet and are often referred to as residue-centric methods. For example, one of the most popular residue-centric methods is the Shannon entropy (SE) measure, which estimates conservation by calculating the Shannon entropy scores of amino acids distribution for each column (Shenkin and Erman 1991). To improve SE, RE takes amino acid background distribution into account, so it could identify residues which are highly deviated from the background distribution (Wang and Samudrala 2006). These sites indicate strong evolutionary constraint and are thought to be important for maintaining the structure or function of the protein. Similar procedure is also used in JSD (Capra and Singh 2007) and the variance measure (VM) method (Pei and Grishin 2001). Besides, several residue-centric methods try to incorporate phylogenetic information in measuring conservation. For example, the two entropy analysis-objective (TEA-O) method traces evolutionary pressure from the root to the branches of a phylogenetic tree (Ye et al. 2008) and the Rate4site infers the rate of evolution at each site (Mayrose et al. 2004). A low rate of evolution means high conservation at each site. However, a deficiency of the above residue-centric methods is that they fail to account for the stereochemical properties of amino acids. To solve the problem, another group of methods attempt to take amino acid similarity into account. For example, the Shannon entropy of residue property method incorporates amino acid similarity by grouping them into six or nine disjoint sets (Mirny and Shakhnovich 1999; Taylor 1986). Williamson (1995) further incorporated the background distribution of these properties to improve the method. Innis et al. (2004) predicted catalytic sites using the conserved functional group analysis method, which depends on the protein structure information. Compared with residue-centric methods, methods in this category are tolerant to mutations between amino acids in the same set.

It is beneficial to group amino acids into disjoint groups. However, a deficiency of the previous classifiers is that residues of different types are treated equally different. To overcome the weakness, Taylor's ten overlapping sets of amino acids are incorporated in measuring conservation (Dou et al. 2009a). However, the method fails to account for background frequencies of these types and depends on a free parameter $\lambda$. In order to factor in the background distribution and remove the free parameter, we propose two new methods, the relative entropy of Venn diagram (RVD) and RVD2. They build Taylor's ten overlapping classes into the relative entropy model by the following process. First, the fractional frequencies of each type are calculated for each site. Then, the background frequencies of each type are extracted based on the amino acid BLOSUM62 background distribution. Finally, the conservation index is calculated from these PSDs and the background distribution using the relative entropy model. To further improve the performance, RVD and RVD2 are combined with JSD, and the resulting two new measures are referred to as VJSD and V2JSD, respectively.

To compare new methods with existing ones, we apply them to identify catalytic residues on our data set and a data set constructed by Capra and Singh (2007). The performances are evaluated by two criteria, receiver operator characteristic curve (ROC) analysis (Gribskov and Robinson 1996) and 'top-$n$' ranks (Wang and Samudrala 2006). The following are our main findings: (1) RVD and RVD2 perform better than most of the previous methods. (2) Taylor's overlapping classification is superior to disjoint classifications in predicting catalytic residues. (3) Combination measures perform excellently in predicting catalytic residues. (4) WOP can be served as an alternative source of PSDs.

## Data and methods

### Benchmark sets

We first construct our benchmark data set based on the Catalytic Site Atlas (CSA) database (Porter et al. 2003). For each literature-based entry in the CSA version 2.2.11, we obtain protein chains from the Protein Data Bank (PDB) (Berman et al. 2000). These sequences are clustered at 95% sequence identity and redundant sequences are removed. For each remaining sequence, we download its alignment from the HSSP database (Dodge et al. 1998). Then, sequences in each alignment are clustered at 95% sequence identity and redundant sequences are removed. Next, within each alignment, sequences whose lengths are less than half of the target sequence length are removed. After that, alignments with fewer than five sequences are also removed. In this paper, catalytic sites are treated as positive sites and all other sites are negative. After filtering, 818 alignments with an average of 357 sequences and 2559 positive sites remain. On the other hand, PSSMs are obtained for each remained target chain. These PSSMs are extracted using PSI-BLAST against the NCBI non-redundant (NR) database with parameters -j 3 -e 0.001 (Altschul et al. 1997). The above data set is referenced as New-CSA

data set in the paper. For comparison purposes, we also test all methods on a large data set recently published by Capra and Singh (2007). After filtering, 652 alignments and 1984 positive sites remain. PSSMs for each target sequence are also extracted by the above method. This benchmark data set is referenced as the Capra-CSA data set.

In order to compare with an SVM-based method, our new methods are evaluated on the EF-fold data set created by Zhang et al. (2008a). Protein chains and catalytic position data are extracted from the data set. Then PSSMs are obtained for each chain using the above method. There are 189 chains and 606 catalytic sites within the data set.

## Conservation scores

For any target chain $K$, let $P_C$ be the PSD on site $C$ and $p_C(i)$ be the frequency of the $i$th residue. PSDs for each site are extracted from two sources: MSA and WOP. As usual, Gaps are ignored when extracting PSDs from MSAs, and the BLOSUM62 background distribution is used as the amino acid background distribution in the paper. Let $P_0$ be the background distribution and $p_0(i)$ be the background frequency of the $i$th residue. Similarly, let $P_C^s$ be the PSDs of each type in site $C$ and $P_0^s$ be the background distribution of these types. In addition, conservation indices calculated by method $M$ based on MSAs are denoted by $M-m$, and those based on WOPs are denoted by $M-w$.

### *Previous methods*

In this section, we introduce eight well-tested protein sequence conservation measures, namely SE, RE, JSD, $FR_{\text{basic}}$,VM, TEA-O, the property relative entropy (PRE) and stereochemical property divergence (SPD). The first six methods are residue-centric methods, whereas amino acid similarity is taken into account in the last two methods.

*The Shannon entropy of residues (SE) method* Shannon entropy score is one of the most commonly used sequence conservation measures. The SE score of a column $C$ is defined as

$$\text{SE}_C = \sum_{i=1}^{20} p_C(i) \log p_C(i).$$

SE score is the smallest for a column with complete conservation (Cover and Thomas 1991).

*The relative entropy (RE) method* Wang and Samudrala (2006) applied RE to find relatively conserved residues. In contrast to SE, RE incorporates amino acid background distribution. The RE score of a column $C$ is calculated as

$$\text{RE}_C = \sum_{i=1}^{20} p_C(i) \log \frac{p_C(i)}{p_0(i)}.$$

*The Jensen–Shannon Divergence (JSD) method* The Jensen–Shannon divergence is widely used to quantify the similarity between two probability distributions (Lin 1991). Capra and Singh (2007) used it to measure the conservation at each alignment position. The JSD score of a column $C$ is computed as

$$\text{JSD}_C = \frac{1}{2}\sum_{i=1}^{20} p_C(i) \log \frac{p_C(i)}{\frac{1}{2}p_0(i) + \frac{1}{2}p_C(i)} + \frac{1}{2}\sum_{i=1}^{20} p_0(i) \log \frac{p_0(i)}{\frac{1}{2}p_C(i) + \frac{1}{2}p_0(i)}.$$

*The variance measure (VM) method* Pei and Grishin (2001) introduced VM as a protein sequence conservation measure. Unlike entropy-based methods, it is straightforward and intuitive. The VM score of a column $C$ is given by

$$\text{VM}_C = \sqrt{\sum_{i=1}^{20} (p_C(i) - p_0(i))^2}.$$

It has the advantage over RE of being symmetric and less extreme in scoring deviations in frequencies of rare residues. Actually, this measure is the Euclidean distance between a position-specific distribution and the background distribution.

*The property relative entropy (PRE) method* In contrast to the above methods, biochemical similarity among amino acids is taken into account in this method. We use the following grouping: aliphatic [AVLIMC], aromatic [FWYH], polar [STNQ], positive [KR], negative [DE], and special conformations [GP] (Mirny and Shakhnovich 1999). The PRE score of a column $C$ is given by

$$\text{PRE}_C = \sum_{i=1}^{6} p_C^s(i) \log \frac{p_C^s(i)}{p_0^s(i)},$$

where $p_C^s(i)$ is the fractional frequency of class $i$ in the column, and $p_0^s(i)$ is the background frequency of the class $i$.

*The stereochemical property divergence (SPD) method* Unlike PRE methods, SPD incorporates ten overlapping groups of amino acids (Dou et al. 2009a). The method is based on the assumption that a column in a multiple sequence alignment is evolved from a column with identical entries in the evolutionary history. It measures conservation in two steps. First, divergence between a real column and a putative column with identical entries is quantified by the cosine function. Second, the above divergence is combined with conservation scores proposed

by other methods. We consider the SPR scores and chose $\lambda = 0.05$.

*The two entropy analysis-objective (TEA-O) method* Ye et al. (2008) proposed this method to trace the evolutionary pressures from the root to branches of a phylogenetic tree. Given an alignment and a related phylogenetic tree, TEA-O calculates two types of SE scores at each position. Then, the conservation score is calculated from these two SE scores for each site. We use the freely available TEA-O version for Linux and phylogenetic trees are generated by the Clustalw for Linux (Thompson et al. 1994). Since the method depends on a phylogenetic tree, its use is limited to MSA based data.

*The $FR_{basic}$ method* Fischer et al. (2008) introduced a functional residues prediction method using a probability density estimation method. The $FR_{\text{basic}}$ conservation score of a column $C$ is defined as

$$\text{FR}_{\text{basic}} = (\log \sum_{i=1}^{i=20} p_C^2(i)/p_0(i))/(\log \sum_{i=1}^{i=20} p_C(i)/p_0(i)).$$

*New methods*

Our new methods take amino acid similarity into account by using Taylor's ten overlapping classes of amino acids (Taylor 1986). These classes are Polar [NQSDECTKRHYW], Positive [KHR], Negative [DE], Charged [KHRDE], Hydrophobic [AGCTIVLKHFYWM], Aliphatic [IVL], Aromatic [FYWH], Small [PNDTCAGSV], Tiny [ASGC] and Proline [P]. Given a column $C$, let $P^s{}_C = (p^s{}_C(1), p^s{}_C(2), ..., p^s{}_C(10))$, where $p^s{}_C(i)$ is the fractional frequency of class $i$ in the column; similarly, let

$$\overline{P_C^s} = (\overline{p_C^s(1)}, \overline{p_C^s(2)}, \ldots, \overline{p_C^s(10)}),$$

where $\overline{p_C^s(i)} = 1 - p_C^s(i); P^s{}_0 = (p^s{}_0(1), p^s{}_0(2), ..., p^s{}_0(10))$, where $p^s{}_0(i)$ is the fractional frequency of class $i$ based on the BLOSUM62 amino acid background distribution; and $\overline{P_0^s} = (\overline{p_0^s(1)}, \overline{p_0^s(2)}, \ldots, \overline{p_0^s(10)})$, where $\overline{p_0^s(i)} = 1 - p_0^s(i)$.

Then the conservation score of the column $C$ is defined as

$$\text{RVD}_C = \sum_{i=1}^{10} p_C^s(i) \ln \frac{p_C^s(i)}{p_0^s(i)}.$$

This method is referred to as the relative entropy of Venn diagram (RVD). According to the definition, the position-specific distribution $P^s{}_C$ satisfies that $1 \le \sum_{i=1}^{10} p_C^s(i) \le 5$. In specific, it equals to 1 if and only if the column consists of M or Q, because M and Q belong to only one of these classes. Similarly, it equals to 5 if and only if the column consists of H, for H belongs to five different classes. To achieve some important mathematical properties of distance measures, we define an alternative conservation score for this column as

$$\text{RVD2}_C = \sum_{i=1}^{10} RE_C(i),$$

where $RE_C(i) = p_C^s(i) \ln \frac{p_C^s(i)}{p_0^s(i)} + \overline{p_C^s(i)} \ln \frac{\overline{p_C^s(i)}}{\overline{p_0^s(i)}}$, which is the classical relative entropy measure. Mathematically, $RE_C(i)$ is non-negative, convex, additive, and non-symmetric (Lin 1991; Johnson 1979), so RVD2 also has these properties.

However, there is an inherent weakness for RVD and RVD2, as Taylor's classifier cannot distinguish several residues. For example, it cannot distinguish a column constructed by only Y and that constructed by Y and W. So do the residues G and A, I, and L. But the methods based on residue frequencies can distinguish them naturally. So, we combine RVD and RVD2 with a residue-centric method JSD to overcome the weakness. The combination process consists of two steps. First, they are normalized to 0 and 1, and the normalized conservation measure is denoted by '*n*Measure'. Second, the square root of the sum of their squares is treated as a new measure. Hence, we get two combination measures for column $C$:

$$\text{VJSD}_C = \sqrt{nRVD_C^2 + nJSD_C^2},$$

$$\text{V2JSD}_C = \sqrt{nRVD2_C^2 + nJSD_C^2}.$$

Evaluation methods

To validate the proposed methods, we apply them to predict catalytic sites on the New-CSA and Capra-CSA data sets. Their performances are evaluated by two criteria: receiver operator characteristic curve (ROC) analysis (Gribskov and Robinson 1996) and 'top-*n* hits' (Wang and Samudrala 2006).

Under the first criterion, we plot ROC curves with false-positive rate (FPR) as abscissa and true positive rate (TPR) as ordinate. TPR and FPR are often called sensitivity and 1-specificity, respectively. Then we calculate the area under each ROC curve (AUC) at a range of false-positive rates: 0.05 ($\text{AUC}_{0.05}$), 0.1 ($\text{AUC}_{0.1}$), and 0.15 ($\text{AUC}_{0.15}$). After that, the $\text{AUC}_\alpha s$ over all alignments within the same set are averaged with $\alpha = 0.1$, 0.5 and 1, respectively. Then, we plot $\text{ROC}_n$ curves as described by Panchenko et al. (2003) and choose $n = 100$. Under the second criterion, we count the number of catalytic residues across all alignment in the 'top-*n*' highest scoring residues and set $n$ to be 10, 20, and 30, respectively. The top-10, 20, and 30 ranks are summed across all alignments within the same set, respectively. The three summations are used to evaluate a method. The higher the $\text{AUC}_\alpha s$ and the top-*n* numbers, the better the method in predicting functional residues.

We use the Friedman test to judge whether the performance statistics (e.g. top-10 numbers) for these methods

are significantly different. We also apply a Bonferroni correction for choosing a *P*-value for our analysis. The difference between the performances of two methods is called statistically significant if the *P*-value estimated by Friedman test with a Bonferroni correction is less than 0.05 (Capra and Singh 2007).

## Results

The performance statistics, top-*n* numbers and averaged $AUC_\alpha s$, are calculated to evaluate the performances of these methods. The performance statistics of 'top-*n* hits' method are shown in Tables 1 and 2. The averaged $AUC_\alpha s$ are shown in Tables 3 and 4. Table 5 shows the *P*-values estimated by the Friedman test for seven methods in terms of top-10 numbers on the Capra-CSA data set. Table 6 shows the performances of several methods on the EF-fold data set. High-confidence regions of the ROC curves for RE, JSD, RVD, and VJSD on the New-CSA data set are shown in Fig. 1. Figure 2 shows the high-confidence regions of the ROC curves for RE, JSD, RVD, and VJSD on the Capra-CSA data set. In the following, we describe our main findings in detail.

### RVD and RVD2 perform better than most of the previous methods

For the New-CSA data set, Tables 1 and 3 show that RVD and RVD2 perform better than all the previous methods using both criteria. Although JSD-m has similar performance to RVD2-m, JSD-w is outperformed by RVD2-w on the data set. Similarly, Table 3 shows that RVD and RVD2 outperform other methods judged by both $AUC_{0.05}$ and $AUC_{0.1}$. Figure 1 shows the ROC curves of several methods. This figure illustrates that RVD performs better than RE and JSD on the New-CSA data set in FPR (0, 0.1). For the Capra-CSA data set, Tables 2 and 4 also show that RVD and RVD2 outperform other methods. For example, RVD-m performs better than all the previous methods judged by top-10 and top-20 numbers based on MSAs. In detail, the top-10 number of RVD-m is more than 920, while those of all the previous methods are less than 820 based on MSAs. Furthermore, Table 5 shows that the differences between RVD-m and the previous methods are all statistically significant in terms of top-10 numbers. We find in Table 4 that RVD and RVD2 outperform the previous methods in terms of $AUC_{0.05}$ and $AUC_{0.1}$. Figure 2

**Table 1** Evaluation of the performances of all methods on the New-CSA data set by the top-*n* numbers

| Method | MSA | | | WOP | | |
|---|---|---|---|---|---|---|
| | top-10 | top-20 | top-30 | top-10 | top-20 | top-30 |
| $FR_{basic}$ | 649 | 1,171 | 1,457 | 947 | 1,424 | 1,681 |
| SE | 770 | 1,243 | 1,518 | 898 | 1,363 | 1,597 |
| RE | 976 | 1,509 | 1,809 | 916 | 1,389 | 1,689 |
| TEA-O | 983 | 1,515 | 1,771 | – | – | – |
| SPR | 989 | 1,509 | 1,787 | 935 | 1,388 | 1,624 |
| PRE | 1,092 | 1,468 | 1,702 | 960 | 1,339 | 1,550 |
| JSD | 1,136 | 1,643 | 1,893 | 996 | 1,437 | 1,682 |
| VM | 1,151 | 1,629 | 1,853 | 1,003 | 1,435 | 1,652 |
| RVD | 1,180 | 1,576 | 1,748 | 1,118 | 1,491 | 1,686 |
| RVD2 | 1,147 | 1,516 | 1,692 | 1,048 | 1,377 | 1,574 |
| VJSD | 1,299 | 1,737 | 1,935 | 1,147 | 1,568 | 1,778 |
| V2JSD | 1,221 | 1,689 | 1,889 | 1,055 | 1,456 | 1,710 |

**Table 2** Evaluation of the performances of all methods on the Capra-CSA data set by the top-*n* numbers

| Method | MSA | | | WOP | | |
|---|---|---|---|---|---|---|
| | top-10 | top-20 | top-30 | top-10 | top-20 | top-30 |
| $FR_{basic}$ | 705 | 1,104 | 1,330 | 779 | 1,141 | 1,380 |
| SE | 696 | 1,083 | 1,312 | 732 | 1,107 | 1,309 |
| RE | 673 | 1,133 | 1,376 | 767 | 1,141 | 1,366 |
| TEA-O | 695 | 1,079 | 1,307 | – | – | – |
| SPR | 704 | 1,153 | 1,380 | 777 | 1,146 | 1,339 |
| PRE | 812 | 1,115 | 1,307 | 812 | 1,108 | 1,278 |
| JSD | 797 | 1,213 | 1,435 | 845 | 1,170 | 1,387 |
| VM | 811 | 1,205 | 1,409 | 837 | 1,165 | 1,366 |
| RVD | 925 | 1,250 | 1,408 | 925 | 1,222 | 1,376 |
| RVD2 | 894 | 1,188 | 1,356 | 868 | 1,132 | 1,287 |
| VJSD | 983 | 1,342 | 1,507 | 960 | 1,280 | 1,449 |
| V2JSD | 911 | 1,278 | 1,474 | 882 | 1,203 | 1,450 |

**Table 3** Evaluation of the performances of all methods on the New-CSA data set by averaged $AUC_\alpha s$ (all values are $\times 10^{-2}$)

| Method | MSA | | | WOP | | |
|---|---|---|---|---|---|---|
| | $AUC_{0.05}$ | $AUC_{0.10}$ | $AUC_{0.15}$ | $AUC_{0.05}$ | $AUC_{0.10}$ | $AUC_{0.15}$ |
| $FR_{basic}$ | 1.134 | 3.742 | 6.959 | 1.627 | 4.804 | 8.574 |
| SE | 1.358 | 4.087 | 7.399 | 1.537 | 4.546 | 8.153 |
| RE | 1.697 | 5.059 | 9.009 | 1.600 | 4.749 | 8.541 |
| TEA-O | 1.742 | 5.045 | 8.855 | – | – | – |
| SPR | 1.733 | 5.067 | 8.879 | 1.613 | 4.676 | 8.265 |
| PRE | 1.955 | 5.186 | 8.926 | 1.743 | 4.754 | 8.319 |
| JSD | 2.001 | 5.551 | 9.567 | 1.714 | 4.876 | 8.588 |
| VM | 2.042 | 5.514 | 9.391 | 1.761 | 4.904 | 8.621 |
| RVD | 2.182 | 5.599 | 9.381 | 2.114 | 5.423 | 9.111 |
| RVD2 | 2.138 | 5.430 | 9.059 | 1.949 | 5.015 | 8.498 |
| VJSD | 2.363 | 6.076 | 10.166 | 2.173 | 5.574 | 9.423 |
| V2JSD | 2.238 | 5.867 | 9.908 | 1.964 | 5.218 | 8.942 |

**Table 4** Evaluation of the performances of all methods on the Capra-CSA data set by averaged $AUC_{\alpha}s$ (all values are $\times 10^{-2}$)

| Method | MSA | | | WOP | | |
|---|---|---|---|---|---|---|
| | $AUC_{0.05}$ | $AUC_{0.10}$ | $AUC_{0.15}$ | $AUC_{0.05}$ | $AUC_{0.10}$ | $AUC_{0.15}$ |
| $FR_{basic}$ | 1.562 | 4.672 | 8.464 | 1.728 | 5.020 | 8.915 |
| SE | 1.546 | 4.613 | 8.338 | 1.615 | 4.744 | 8.512 |
| RE | 1.513 | 4.777 | 8.714 | 1.706 | 4.977 | 8.915 |
| TEA-O | 1.556 | 4.613 | 8.306 | – | – | – |
| SPR | 1.572 | 4.874 | 8.773 | 1.736 | 4.946 | 8.695 |
| PRE | 1.888 | 5.074 | 8.812 | 1.846 | 4.972 | 8.671 |
| JSD | 1.787 | 5.228 | 9.249 | 1.837 | 5.160 | 9.028 |
| VM | 1.832 | 5.209 | 9.282 | 1.866 | 5.156 | 8.996 |
| RVD | 2.162 | 5.640 | 9.536 | 2.193 | 5.628 | 9.451 |
| RVD2 | 2.079 | 5.450 | 9.219 | 2.052 | 5.246 | 8.822 |
| VJSD | 2.303 | 5.982 | 10.112 | 2.277 | 5.799 | 9.795 |
| V2JSD | 2.096 | 5.715 | 9.786 | 2.066 | 5.457 | 9.351 |

presents the ROC curves of several methods based on both MSAs and WOPs. This figure illustrates that RVD performs better than RE and JSD on the Capra-CSA data set in FPR (0, 0.1). Although it is found that methods incorporating amino acid similarity fail to provide significant improvement over methods which ignore the underlying chemistry (Capra and Singh 2007), RVD and RVD2 are counterexamples of this fact. To sum up, RVD and RVD2 perform better than most of the previous methods in FPR (0, 0.1).

### Taylor's overlapping classification is superior to disjoint classification in predicting catalytic residues

RVD and PRE are both measures in the relative entropy model and take amino acid similarity into account. The difference between them is that PRE considers disjoint partition of amino acids, whereas RVD considers overlapping sets. However, RVD performs significantly better than PRE by both criteria on both data sets. For example, Table 1 shows that the 'top-*n*' numbers of PRE-m on the New-CSA data set are 1,092, 1,468, and 1,702, respectively; in contrast, those of RVD-m are 1,180, 1,576, and 1,748, respectively. Table 2 shows that the 'top-*n*' numbers of PRE-m on the Capra-CSA data set are 812, 1115, and 1307, and those of RVD-m are 925, 1,250, and 1,408, respectively. Tables 1 and 2 also show that RVD-w performs better than PRE-w on both data sets. Table 5 shows that the difference between RVD-m and PRE-m is statistically significant in terms of the top-10 numbers. On the other hand, PRE is compared with RVD2 on both data sets by both criteria and the results are similar. For example, the top-10 numbers of RVD2 on the Capra-CSA data set are 894 and 868, whereas those of PRE are only 812, and 812, respectively. These results suggest that some biological information is lost during the process of grouping amino acids into disjoint classes. So, Taylor's overlapping classes of amino acids help to predict catalytic residues from protein sequence.

### Combination measures have excellent performances in predicting catalytic residues

In order to further improve RVD and RVD2, we propose two combination methods, VJSD, and V2JSD. For example, Tables 1 and 3 show that they outperform all other methods by two criteria on the New-CSA data set. In detail, the three 'top-*n*' numbers of VJSD-m are 1299, 1737, and 1935, whereas, those of all other methods are less than 1180, 1643, and 1893 based on MSAs. Figure 1 also illustrates that VJSD performs better than RVD, JSD, and RE. Tables 2 and 4 show that combination methods outperform other methods by both criteria on the Capra-CSA data set. For example, the three 'top-*n*' numbers of VJSD-m are 983, 1342, and 1507. In contrast, the three 'top-*n*' numbers of JSD-m are only 797, 1213, and 1435. Table 5 shows that the differences between VJSD-m and other methods are statistically significant in terms of the top-10 numbers. Table 4 shows that VJSD also outperforms other

**Table 5** A table to show the *P*-values on the Capra-CSA data set

| Method | $FR_{basic}$-m | RE-m | PRE-m | VM-m | JSD-m | RVD-m | VJSD-m |
|---|---|---|---|---|---|---|---|
| $FR_{basic}$-m | 1 | $4.65 \times 10^{-1}$ | $2.66 \times 10^{-4}$ | $2.27 \times 10^{-7}$ | $2.52 \times 10^{-5}$ | $1.67 \times 10^{-15}$ | 0 |
| RE-m | 0 | 1 | $9.82 \times 10^{-4}$ | $7.51 \times 10^{-11}$ | $2.10 \times 10^{-14}$ | 0 | 0 |
| PRE-m | 0 | 0 | 1 | $2.74 \times 10^{-1}$ | $6.03 \times 10^{-1}$ | $3.76 \times 10^{-7}$ | $8.88 \times 10^{-16}$ |
| VM-m | 0 | 0 | 0 | 1 | $4.14 \times 10^{-1}$ | $9.04 \times 10^{-5}$ | $6.53 \times 10^{-13}$ |
| JSD-m | 0 | 0 | 0 | 0 | 1 | $2.75 \times 10^{-7}$ | $2.22 \times 10^{-16}$ |
| RVD-m | 0 | 0 | 0 | 0 | 0 | 1 | $2.50 \times 10^{-4}$ |
| VJSD-m | 0 | 0 | 0 | 0 | 0 | 0 | 1 |

The difference between the performance of two methods is called statistically significant if the *P*-value is less than 0.05. Because there are 12 methods in the paper, the threshold of statistically significant is adjusted to $4.17 \times 10^{-3}$ according to the Bonferroni correction

**Table 6** Top-10 hits of 1uas-A judged by RE-m, JSD-m, RVD-m, and VJSD-m on the Capra-CSA data set

| RE-m | | JSD-m | | RVD-m | | VJSD-m | |
|---|---|---|---|---|---|---|---|
| order | residue | order | residue | order | residue | order | residue |
| 237 | W | 237 | W | 84 | H | 84 | H |
| 180 | W | 180 | W | **130** | **D** | **130** | **D** |
| 54 | W | 101 | C | **185** | **D** | **185** | **D** |
| 16 | W | 271 | Q | 214 | D | 214 | D |
| 122 | W | 93 | Y | 216 | D | 216 | D |
| 164 | W | 16 | W | 125 | D | 125 | D |
| 101 | C | 54 | W | 272 | D | 272 | D |
| 132 | C | 122 | W | 115 | D | 237 | W |
| 84 | H | 84 | H | 233 | H | 128 | K |
| 13 | W | 73 | P | 128 | K | 51 | D |

Bold font indicates catalytic residues

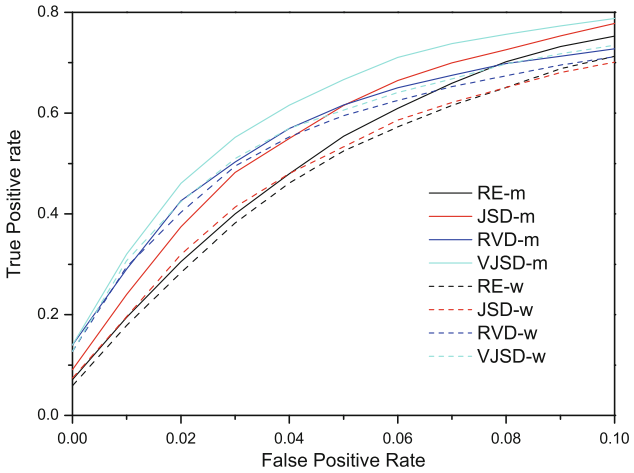


**Fig. 1** High confidence region of the ROC curves of RE, JSD, RVD, and VJSD on the New-CSA data set. These methods are evaluated based on the PSDs extracted form both MSAs and WOPs. This figure illustrates that RVD performs better than RE and JSD in false positive rate (0, 0.1). In addition, results based on MSAs are slightly better than those based on WOPs on the data set

methods judged by all $AUC_{0.05}$, $AUC_{0.1}$ and $AUC_{0.15}$ levels. Figure 2 illustrates that VJSD outperforms other methods in predicting catalytic residues in FPR (0, 0.1). In brief, combination methods improve RVD and RVD2 successfully and have excellent performances in predicting catalytic residues.

## WOP is an alternative source of PSDs

In most previous conservation measures, PSDs are extracted from MSAs. So the process relies on the quality of the alignment and is limited to only a subset of potential queries. In this paper, we try to extract PSDs from PSSMs using PSI-BLAST. Then, all methods are tested on distributions extracted from both MSAs and WOPs. In detail,

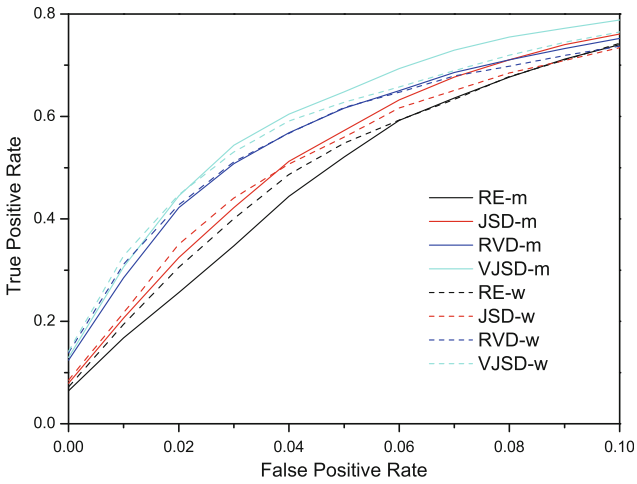


**Fig. 2** High confidence region of the ROC curves of RE, JSD, RVD and VJSD on the Capra-CSA data set. These methods are evaluated based on the PSDs extracted form both MSAs and WOPs. The figure illustrates that RVD performs better than RE and JSD in false positive rate (0, 0.1). In addition, results based on WOPs are comparative to results based on MSAs on the data set

Tables 1 and 3 show that the results based on MSAs are better than those based on WOPs on the New-CSA data set, except for SE and $FR_{basic}$. Figure 1 also illustrates that the results based on MSAs are slightly better than those based on WOPs, respectively. Tables 2 and 4 show that the results based on MSAs are comparable to those based on WOPs on the Capra-CSA data set. Figure 2 also illustrates that the results based on MSAs are comparable to those based on WOPs. Experimental results suggest that although PSDs extracted from MSAs are slightly better than those extracted from WOPs, extracting PSDs from WOPs is more convenient and can be used for any protein chain. Therefore, extracting PSDs from WOPs can be served as an alternative method to protein sequence conservation measures.

## Case study

To further illustrate our methods, we select an alignment (PDB ID 1uas-A) from the Capra-CSA data set. The selected protein chain contains two catalytic residues (D and D). They are the 130th and 185th residues in the sequence. In our analysis, RVD-m and VJSD-m identify catalytic sites in their top-10 hits, but RE-m and JSD-m fail to do so. The top-10 hits of these methods are shown in Table 6. Table 6 shows that there is only one charged residue in the top-10 hits of RE-m and JSD-m, but there are ten charged residues in those of RVD-m and nine charged residues in those of VJSD-m. Because most catalytic sites are charged (Bartlett et al. 2002), RVD-m and VJSD-m have a higher probability than RE-m and JSD-m to catch catalytic residues in their top-10 hits.

## Conclusion and discussion

Given the PSDs of a protein chain, a number of methods have been proposed to extracted conservation information in recent years. Several of them attempt to incorporate amino acid similarity by grouping residues into disjoint sets. In contrast, RVD and RVD2 consider ten overlapping classes of amino acids. Experimental results suggest that they perform better than many previous methods, such as commonly used RE, PRE, and JSD. Thus, our results suggest that Taylor's overlapping classification of amino acids is superior to disjoint classes in predicting catalytic residues. In addition, similar to residue-centric methods, we find that methods based on the classification can be improved by incorporating background distribution of overlapping sets.

Although RVD and RVD2 perform better than many previous methods, there is an inherent weakness: that they cannot distinguish some residues. While residue-centric methods can distinguish them naturally. To solve the problem, we propose a formula by combining RVD and RVD2 with residue-centric methods. As a result, we get two combination methods and the experimental results suggest that they have excellent performances in finding catalytic residues. On the other hand, the results imply that it helps to predict catalytic residues from sequence by incorporating different aspects of biological information. So, further attention should be focused on incorporating different aspect of biologic information in scoring protein sequence conservation.

In most previous methods, PSDs are often extracted from MSAs. So these distributions rely highly on the quality of the alignments, and are limited to only a subset of potential queries. In this paper, PSDs are also extracted from WOPs using PSI-BLAST. Then all methods are tested on PSDs extracted from both PSSMs and MSAs. The experimental results suggest that although PSDs based on MSAs are slightly better than those based on WOPs, it is more convenient to extract distributions from WOPs, and the method can be used for any protein chain. So WOP is an alternative source of PSDs for protein sequence conservation measures.

In this paper, we also compare our methods with the SVM approach proposed by Zhang et al. (2008a) on the EF-fold data set. The experimental results are shown in Table 7, which show that although VJSD-w is based only on conservation information, it achieves similar performance with SVM. Moreover, our method is much simpler than SVM. On the other hand, several previous studies suggest that conservation information is a key feature for machine-learning based methods. So our methods may be used to present new features for these methods.

**Table 7** Evaluation of SVM and several conservation based methods on the EF-fold data set (all AUC values are $\times 10^{-2}$)

| Method | top-*n* method | | | AUC method | | |
|---|---|---|---|---|---|---|
| | top-10 | top-20 | top-30 | $AUC_{0.05}$ | $AUC_{0.10}$ | $AUC_{0.15}$ |
| RE-w | 238 | 346 | 414 | 1.630 | 4.684 | 8.389 |
| JSD-w | 261 | 357 | 418 | 1.728 | 4.832 | 8.470 |
| RVD2-w | 243 | 346 | 388 | 1.806 | 4.690 | 8.090 |
| V2JSD-w | 261 | 366 | 415 | 1.870 | 5.001 | 8.638 |
| RVD-w | 285 | 366 | 404 | 1.945 | 5.099 | 8.630 |
| VJSD-w | 290 | 379 | 430 | 2.060 | 5.328 | 9.074 |
| SVM | 301 | 407 | 455 | 2.101 | 5.502 | 9.456 |

**Acknowledgments** This work was partially supported by the Natural Science Foundation of China (No.10731040), Shanghai Leading Academic Discipline Project (No. S30405) and Innovation Program of Shanghai Municipal Education Commission (No. 09zz134).

## References

Altschul SF et al (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 15:3398–3402

Berman H et al (2000) The protein data bank. Nucleic Acids Res 28:235–242

Bartlett G, Porter C, Borkakoti N, Thornton J (2002) Analysis of catalytic residues in enzyme active sites. J Mol Biol 324:105–121

Capra J, Singh S (2007) Predicting functionally important residues from sequence conservation. Bioinformatics 23:1875–1882

Caffery D, somaroo S, Hughes J, Mintseris J, huang E (2004) Are protein–protein interfaces more conserved in sequence than the rest of the protein surface? Protein Sci 13:190–202

Cover T, Thomas J (1991) Elements of information theory. Wiley, New York

David L, Sutch B, Livesay DR (2005) Predicting protein functional sites with phylogenetic motifs. Proteins 58:309–320

Donald JS, Shakhnovich EI (2005) Determining functional specificity from protein sequence. Bioinformatics 21:2629–2635

Dukka B, Dennis R (2008) Improving position-specific predictions of protein functional sites using phylogenetic motifs. Bioinformatics 24:2308–2316

del Sol Mesa A, Pazos F, Valencia A (2003) Automatic methods for predicting functionally important residues. J Mol Biol 326:1289–1302

Dou YC, Zheng XQ, Wang J (2009a) Several appropriate background distributions for entropy-based protein sequence conservation measures. J Theor Biol 262(2):317–322

Dou YC, Zheng XQ, Wang J (2009b) Prediction of catalytic residues using the variation of stereochemical properties. Protein J 28:29–33

Dodge C, Schneider R, Sander C (1998) The hssp database of protein structure-sequence alignments and family profiles. Nucleic Acids Res 26:313–315

Fischer JD, Mayer CE, Söding J (2008) Prediction of protein functional residues from sequence by probability density estimation. Bioinformatics 24:613–620

Gribskov M, Robinson N (1996) Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching. Comput Chem 20:25–33

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

Innis CA, Anand AP, Sowdhamini R, Brocchieri L (2004) Prediction of functional sites in proteins using conserved functional group analysis. J Mol Biol 337:1053–1068

Johnson RW (1979) Axiomatic characterization of the directed divergances and their linear combinations. IEEE Trans Inf Theory 6:709–716

Liu XS, Guo WL (2008) Robustness of the residue conservation score reflecting both frequencies and physicochemistries. Amino acids 34:643–652

Lin J (1991) Divergence measure based on the Shannon entropy. IEEE Trans Inf Theory 37:145–151

Mihalek I, Reš I, Lichtarge O (2007) Background frequencies for residue variability estimates: BLOSUM revisited. BMC Bioinformatics 8:488

Mihalek I, Reš I, Lichtarge O (2004) A family of evolution–entropy hybrid methods for ranking residues by importance. J Mol Biol 336:1265–1282

Merkl R, Zwick M (2008) H2r: Identification of evolutionary important residues by means of an entropy based analysis of multiple sequence alignments. BMC Bioinformatics 9:151

Mirny L, Shakhnovich E (1999) Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. J Mol Biol 291:177–196

Mayrose I, Graur D, Ben-Tal N, Pupko T (2004) Comparison of sitespecific rate-inference methods for protein sequences: empirical bayesian methods are superior. Mol Biol and Evol 21:1781–1791

Pande S, Raheja A, Liversay DR (2007) Prediction of enzyme catalytic sites from sequence using neural networks. IEEE symp CIBCB 07:247–253

Panchenko A, Kondrashov F, Bryant S (2003) Prediction of functional sites by analysis of sequence and structure conservation. Protein Sci 13:884–892

Petrova N, Wu C (2006) Prediction of catalytic residues using support vector machines with selected protein sequence and structural properties. BMC Bioinformatics 7:312

Pei J, Grishin N (2001) AL2CO: calculation of positional conservation in a protein sequence alignment. Bioinformatics 17:700–712

Porter C, Bartlett G, Thornton J (2003) The catalytic site atlas: a resource of catalytic sites and residues identified in enzymes using structural data. Nucleic Acids Res 32:D129–D133

Reva B, Antipin Y, Sander C (2007) Determinants of protein function revealed by combinatorial entropy optimization. Genome Biol 8:R232

Sterner B, Singh R, Berger B (2007) Predicting and annotating catalytic residues: an information theoretic approach. J Comput Biol 14:1058–1073

Shenkin P, Erman BLM (1991) Information-theoretical entropy as a measure of sequence variability. Proteins 11:297–313

Taylor W (1986) The classification of amino acid conservation. J Theor Biol 119:205–218

Tang Y, Sheng Z, Chen Y, Zhang Z (2008) An improved prediction of catalytic residues in enzyme structures. Protein Eng Des Sel 21:295–302

Thompson J, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680

Valdar W (2002) Scoring residue conservation. Proteins 48:227–241

Wang K, Samudrala R (2006) Incorporating background frequency improves entropy-based residue conservation measures. BMC Bioinformatics 7:385

Williamson R (1995) Information theory analysis of the relationship between primary sequence structure and ligand recognition among a class of facilitated transporters. J Theor Biol 174:179–188

Ye K, Vriend G, IJzerman A (2008) Tracing evolutionary pressure. Bioinformatics 24:908–915

Youn E, Peters B, Radivojac P, Mooney SD (2007) Evaluation of features for catalytic residue prediction in novel folds. Protein Sci 16:216–226

Zhang T, Zhang H, Chen K, Shen SY, Ruan J, Kurgan L (2008a) Accutate sequence-based prediction of catalytic residues. Bioinformatics 24:2329–2338

Zhang SW, Zhang YL, Pan Q, Cheng YW, Chou KC (2008b) Estimating residue evolutionary conservation by introducing von Neumann entropy and a novel gap-treating approach. Amino acids 35:495–501