BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 15 2007, pages 1875–1882
doi:10.1093/bioinformatics/btm270

Sequence analysis

Predicting functionally important residues from sequence conservation

John A. Capra and Mona Singh*

Department of Computer Science and Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ 08540, USA

Received on March 26, 2007; revised on May 4, 2007; accepted on May 10, 2007

Advance Access publication May 22, 2007

Associate Editor: Keith Crandall

ABSTRACT

Motivation: All residues in a protein are not equally important. Some are essential for the proper structure and function of the protein, whereas others can be readily replaced. Conservation analysis is one of the most widely used methods for predicting these functionally important residues in protein sequences.

Results: We introduce an information-theoretic approach for estimating sequence conservation based on Jensen–Shannon divergence. We also develop a general heuristic that considers the estimated conservation of sequentially neighboring sites. In large-scale testing, we demonstrate that our combined approach outperforms previous conservation-based measures in identifying functionally important residues; in particular, it is significantly better than the commonly used Shannon entropy measure. We find that considering conservation at sequential neighbors improves the performance of all methods tested. Our analysis also reveals that many existing methods that attempt to incorporate the relationships between amino acids do not lead to better identification of functionally important sites. Finally, we find that while conservation is highly predictive in identifying catalytic sites and residues near bound ligands, it is much less effective in identifying residues in protein–protein interfaces.

Availability: Data sets and code for all conservation measures evaluated are available at http://compbio.cs.princeton.edu/conservation/

Contact: mona@cs.princeton.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

One of the most important and widely studied problems in protein sequence analysis is identifying which residues in a protein are responsible for its function. Knowledge of a protein's functionally important sites has immediate relevance for predicting function, guiding experimental analysis, analyzing molecular mechanisms and understanding protein interactions.

Many computational methods have been developed to predict functionally important residues given a protein sequence. In this article, we focus on one of the most common approaches: the analysis of a multiple sequence alignment (MSA) of the protein and homologous sequences in order to find columns that are preferentially conserved. These sites are presumed to be functionally or structurally important because they have accepted fewer mutations relative to the rest of the alignment.

Conservation analysis has proven to be a powerful indicator of functional importance and has been used to detect residues involved in ligand binding (Liang *et al.*, 2006; Magliery and Regan, 2005), in protein–protein interaction interfaces (Caffrey *et al.*, 2004; Guharoy and Chakrabarti, 2005; Mintseris and Weng, 2005), in maintaining structure (Karlin and Brocchieri, 1996; Schueler-Furman and Baker, 2003; Valdar and Thornton, 2001), and in determining protein functional specificity (Hannenhalli and Russell, 2000; Kalinina, *et al.*, 2003; Lichtarge *et al.*, 1996). Conservation analysis has also been used in conjunction with structural information in many of these applications (Landau *et al.*, 2005; Panchenko *et al.*, 2004).

Computational methods for identifying functional residues that do not use conservation exist, but they typically require structural information and are usually employed in the unusual case where there is an absence or paucity of sequence homologs. Such structural approaches (review, Jones and Thornton, 2004) work by either identifying local shared structural patterns (Fetrow and Skolnick, 1998; Stark and Russell, 2003; Wallace *et al.*, 1997) or by identifying residues in the protein structure with unusual electrostatic and ionization properties (Elcock, 2001; Ondrechen *et al.*, 2001). Many recent methods have used conservation along with other predictors of functional importance (e.g. solvent accessibility, secondary structure, catalytic propensities of amino acids, etc.) in a statistical learning framework (Bordner and Abagyan, 2005; Chung *et al.*, 2006; Gutteridge *et al.*, 2003). It has been found that conservation is the single most powerful attribute in predicting functional importance in these settings (Petrova and Wu, 2006).

While analysis of conservation is a very common approach with an intuitive basis (Valdar, 2002), there is no universally agreed upon technique. Here, we introduce and evaluate a new information-theoretic measure for estimating sequence conservation that is motivated by the notion that conserved

*To whom correspondence should be addressed.

positions are under significant evolutionary pressure, and that positions under pressure are expected to have amino acid distributions very different from those of columns under no pressure (Wang and Samudrala, 2006). We quantify this difference using the Jensen-Shannon divergence (JSD) and an appropriate background 'no pressure' distribution. We also give a window-based extension of our algorithm that incorporates the estimated conservation of sequentially adjacent residues into the score for each column; this window approach can be applied to any conservation scoring method that gives columnwise scores.

To compare the JSD conservation measure to previously proposed methods, we create three data sets that correspond to different types of functional residues—catalytic residues, residues close to ligands and residues in protein–protein interfaces (PPIs)—and give the first large-scale evaluation of several popular conservation measures in identifying functional sites.

We consider six previously introduced methods for estimating the conservation of a column within an MSA. The first and most commonly used method estimates conservation by calculating the Shannon entropy (SE) of the amino acid distribution of each column (Durbin *et al*., 1998). The second attempts to take amino acid similarity into account by partitioning the amino acids into stereochemically similar groups and then calculating the SE in terms of this partition (Mirny and Shakhnovich, 1999; Williamson, 1995). The third incorporates the similarities between amino acids by adapting the von Neumann entropy (VNE) to operate on a substitution matrix (Caffrey *et al*., 2004). The fourth calculates the relative entropy (RE) (Cover and Thomas, 1991) between a column distribution and a background distribution (Wang and Samudrala, 2006); it is similar to our measure in that it attempts to identify sites that have amino acid distributions very different from those of columns under no evolutionary pressure. The fifth takes all pairs of amino acids in a column and sums their pairwise similarity according to a similarity matrix (Karlin and Brocchieri, 1996). The sixth, Rate4Site (R4S), is a sophisticated, computationally intensive approach that builds a phylogenetic tree for a family of protein sequences and infers the rate of evolution at each site (Mayrose *et al*., 2004).

We evaluate how well these seven conservation measures perform in identifying functional sites using ROC curves and by analyzing how often functional sites are within the top ranking sites. Our main findings are: (1) JSD and R4S perform similarly, and are not significantly outperformed by any other method on any data set. However, JSD is several orders of magnitude faster, suggesting its use in many applications, such as genome-scale analyses. (2) The performance of JSD improves when using our approach for incorporating the conservation of neighboring positions in the protein sequence. Incorporating the signal from neighboring residues also improves the performance of the other six methods. While considering the conservation of positions neighboring in 3D has been previously shown to improve predictions (Panchenko *et al*., 2004), structural information is often unavailable. As a result, this finding has immediate relevance in conservation analysis. (3) Many of the conservation methods that explicitly incorporate amino acid similarity fail to consistently improve upon the simple SE method. (4) As compared to identifying catalytic sites and residues near ligands, all the conservation methods tested are only weakly predictive in identifying residues in PPIs. This confirms previous analysis (Caffrey *et al*., 2004; Mintseris and Weng, 2005) and suggests that conservation should only be used as one component in ensemble methods for predicting interaction interfaces (Bordner and Abagyan, 2005; Chung *et al*., 2006).

Overall, our testing demonstrates that JSD, used with our window method to incorporate information from sequential amino acids, provides a fast, state-of-the art method for identifying functionally important residues via conservation analysis. This combined approach performs significantly better than SE, which is likely the most commonly used method in conservation analysis. Moreover, we find that our simple heuristic for incorporating the conservation scores from sequentially neighboring amino acids results in improved performance for all methods tested; this suggests that further development of conservation analysis methods should focus on better exploiting the signal from neighboring residues. Finally, our data sets and testing methodology provide a comprehensive framework for gaining an empirical understanding of the real-world performance of using conservation scores to identify functionally important sites, and analysis similar to the one performed here should be useful in evaluating new proposed conservation measures.

2 METHODS AND ALGORITHMS

2.1 Conservation scores

Brief descriptions of all methods, previous and new, are given subsequently. For most methods, there are a number of parameters to optimize. We have explored the space of reasonable settings and report the best found parameter settings in the following text. See Valdar (2002) for a more complete discussion of similar methods and their evolution.

2.1.1 Preliminaries Each method takes as input a MSA M of length L over N sequences. Let M_C denote the Cth column of the alignment, and M_{Ci} denote the symbol in column C of sequence i. $M_{Ci} \in AA$, where AA is the 21 element set of amino acids plus the gap symbol.

Gaps: Any column that is more than 30% gaps is ignored in the analysis presented here, because a column with many gaps is unlikely to be functionally important. Additionally, a simple gap penalty was applied to all methods except R4S, which handles gaps in its software. In particular, each raw column score is multiplied by the fraction of non-gapped positions in the column (Valdar, 2002). If sequence weighting is used, the gap penalty is weighted as well. We also performed the analyses ignoring all columns with gaps, and our overall conclusions did not change (data not shown).

Sequence weighting: An alignment will often contain sequences at a range of evolutionary distances. If an alignment consists of several very similar sequences, all columns may look conserved, and it will be difficult to discriminate positions under evolutionary pressure from those that are not. We implemented the sequence weighting method proposed in Henikoff and Henikoff (1994) that rewards sequences that are 'surprising'. Sequence weighting is used with all methods and results given subsequently, except for R4S, which builds an evolutionary tree as the first part of its analysis.

Estimating probabilities: Let $\boldsymbol{p_C}$ be the distribution of the set AA in column C; $\boldsymbol{p_C}$ is computed subsequently using the observed (weighted) frequency of each symbol of AA in the column, with a pseudocount of 10^{-6}.

2.1.2 Previous methods We first describe the six previous methods.

Shannon entropy of residues: SE (Cover and Thomas, 1991) is one of the simplest and most common measures of conservation at a site (Sander and Schneider, 1991; Shenkin *et al.*, 1991). It is defined for a column C as:

$$\mathrm{SE}_C = -\sum_{\alpha \in AA} p_C(\alpha) \log p_C(\alpha). \quad (1)$$

The SE is smallest for a column with complete conservation.

Shannon entropy of residue properties: The previous method does not take into account biochemical similarity between amino acids. Instead of treating the amino acids as distinct symbols in the entropy calculation, several groups (Mirny and Shakhnovich, 1999; Williamson, 1995) have proposed partitioning the amino acids into stereochemically defined sets, and then computing the entropy of the column with respect to these sets. We refer to this conservation scoring method as property entropy (PE). We use the following grouping (Mirny and Shakhnovich, 1999): aliphatic [AVLIMC] aromatic [FWYH], polar [STNQ], positive [KR], negative [DE] and [P].

von Neumann entropy: Caffrey *et al.* (2004) introduced the use of VNE (Nielsen and Chuang, 2000), a concept from quantum mechanics, as an information-theoretic measure of conservation that incorporates the physicochemical similarity between amino acids. The VNE of a column C is computed as:

$$VNE_C = -Tr(\rho \log \rho) \quad (2)$$

where ρ is the density matrix of column C normalized so that $Tr(\rho) = 1$. The density matrix of a column is computed by creating a matrix where the diagonal elements are the relative frequencies of amino acids in each column (ignoring gaps and without a pseudocount), and all other entries are zero, and multiplying this by target frequencies for an amino acid similarity matrix. We use the BLOSUM62 matrix as suggested in Caffrey *et al.* (2004).

Relative Entropy: RE, or the Kullback–Leibler divergence, is often used to compare probability distributions (Cover and Thomas, 1991). The RE conservation score for a column is defined as:

$$RE_{\boldsymbol{p_C},\boldsymbol{q}} = \sum_{\alpha \in AA} p_C(\alpha) \log \frac{p_C(\alpha)}{q(\alpha)}. \quad (3)$$

If the background distribution, $\boldsymbol{q}$, lacks gaps (as it does here), then $\boldsymbol{p_C}$ will ignore gaps as well. Magliery and Regan (2005) have applied RE in order to identify unconserved hypervariable positions, and Wang and Samudrala (2006) have applied RE to the problem of finding conserved positions. Unless otherwise stated, we use the overall amino acid distribution in the BLOSUM62 alignments as the background distribution.

Sum-of-pairs measure: The SP method scores the conservation of a column using a similarity matrix S, where $S(x, y)$ is the similarity score between amino acids x and y. Typically S is a matrix such as one from the BLOSUM series (Henikoff and Henikoff, 1992). The SP method encapsulates the overall pairwise similarity between amino acids in a column. The SP measure for a column C is given by:

$$SP_C = \frac{1}{\sum_i^N \sum_{j>i}^N w_i \cdot w_j} \cdot \sum_i^N \sum_{j>i}^N w_i \cdot w_j \cdot S(C_i, C_j), \quad (4)$$

where w_i and w_j are the sequence weights for the ith and jth sequences respectively. While transformations have been proposed to make all diagonal elements equal to one (Karlin and Brocchieri, 1996), or to give immutable amino acids greater self-similarity than mutable ones (Valdar, 2002), we have found that untransformed matrices yielded the best performance. All results presented subsequently use the untransformed BLOSUM62 matrix, unless otherwise indicated.

Rate4Site. In contrast to the methods described earlier, the R4S algorithm (Mayrose *et al.*, 2004) uses a statistical model of evolution to estimate the rate of evolution, and thus the conservation, at each site. Briefly, a phylogenetic tree is constructed for the input alignment. The rates of evolution are assumed to follow a Gamma distribution, and this distribution is used as the prior in a Bayesian inference scheme. A low rate of evolution means high conservation at a position. We use the freely available source code with the default parameters.

2.1.3 New methods We describe a new conservation scoring method and an extension that can be applied to any of the methods.

Jensen-Shannon divergence score: The JSD (Lin, 1991) quantifies the similarity between probability distributions. As compared to RE, it has the advantages of being symmetric and bounded with a range of zero to one. A 'background' amino acid distribution $\boldsymbol{q}$, estimated from a large sequence set, can be used to approximate the distribution of amino acid sites subject to no evolutionary pressure. Then, positions in an alignment that are found to have amino acid distributions very different from this background distribution are proposed to be functionally important or constrained by evolution. JSD is defined for a column C as:

$$D_C^{JS} = \lambda RE_{\boldsymbol{p_C},\boldsymbol{r}} + (1-\lambda) RE_{(\boldsymbol{q},\boldsymbol{r})} \quad (5)$$

where: $\boldsymbol{r} = \lambda \boldsymbol{p_C} + (1-\lambda)\boldsymbol{q}$, $\boldsymbol{p_C}$ is the column amino acid distribution, $\boldsymbol{q}$ is a background distribution and λ is a prior weight. We use $\lambda = 1/2$ and have found that it performs better than other options. Unless otherwise stated, we use the overall amino acid distribution in the BLOSUM62 alignments as the background distribution. Using alignment specific backgrounds can provide a slight improvement, but we have found it is not great enough to justify the added complexity.

While to the best of our knowledge, this is the first use of JSD to assess sequence conservation, it has been previously used in the context of comparing sequence profiles (Yona and Levitt, 2002).

Incorporating sequential residues: Positions near in space and sequence to functionally important residues are known to be more conserved than average (Bartlett *et al.*, 2002). The conservation of spatial neighbors can be exploited to improve prediction of functionally important residues (Panchenko *et al.*, 2004). The conservation of spatial neighbors is stronger than that of positions near in sequence, but 3D structures are often unavailable. Thus we developed the following heuristic method to incorporate the conservation of positions near in sequence into the score for a column:

$$WindowScore_C = \lambda S_C + (1-\lambda) \frac{\sum_{i \in window} S_i}{|window|} \quad (6)$$

where S_i is the raw score of column i and $window$ is a set containing the indices of all columns in the window around column C. We find $\lambda = 1/2$ and a window size of three residues on either side of C works well. This window technique can be applied to any conservation scoring method that gives columnwise scores. When discussing the windowed version of a method, we will append '+W' to the name of the method. Additionally, we call the non-windowed version of a method 'basic.'

2.2 Data sets

We have created three data sets that reflect varying contexts in which conservation-based analysis is commonly applied. The data sets are by nature imperfect, as we rarely know all the functionally important residues in a protein. Indeed it is often not clear how to define 'functionally important'. Moreover, it is difficult to determine whether

a position that appears to be conserved, but is not known to be functionally important, is constrained or simply has not had enough time to diverge. To account for this uncertainty, we construct the data sets to include different types of functional sites, with the hope that the shortcomings of one will be less prevalent in another. We will look for consistent results across these data sets, using various performance metrics, to judge the performance of conservation measures.

2.2.1 Catalytic site data set We created the first data set using known catalytic sites obtained from the Catalytic Site Atlas (CSA) (Porter *et al*., 2003), a literature derived database of enzyme active sites and catalytic residues. For each literature based entry in the CSA as of June 8, 2006, we obtained the 3D structure of the protein chain from the Protein Data Bank (PDB) (Berman *et al*., 2000). The structures' sequences were then clustered at 95% sequence identity and redundant structures were removed. Sequence alignments for each remaining structure were then obtained from HSSP (http://swift.cmbi.kun.nl/gv/hssp/) (Dodge *et al*., 1998). These alignments were filtered to improve alignment quality by removing sequences with more than 95% sequence similarity to the original CSA sequence or whose length was more than two SD away from it. Any alignment with fewer than five sequences was removed. After filtering, 645 alignments with an average of ~79 sequences per alignment and ~1900 catalytic sites remained. The annotated catalytic sites for each protein serve as positives (i.e. functionally important residues) and all other residues are negatives.

We note that many positions in protein cores are conserved for structural reasons. We do not want to penalize methods for giving these likely non-catalytic positions high scores. However, many catalytic sites have low relative solvent accessibility (RSA); for example 5% of catalytic sites have 0% RSA (Bartlett *et al*., 2002). To resolve this tension between leaving out known positives and excluding positions that are likely important but unannotated, we performed the analysis both with and without residues that have RSA less than 1%. There was little change in the relative performance of the measures on all data sets (see Supplementary Material). The results presented here include all columns and catalytic sites. Most sequences do not have known structures, thus this represents the more common scenario in which conservation analysis is applied.

2.2.2 Ligand Distance The second data set is based on a less restrictive definition of functionally important. The increased conservation found in the binding sites of enzyme ligands (Bartlett *et al*., 2002) is used to compare methods without making many assumptions about the type of functional site sought.

The Enzyme Commission (EC) (Bairoch, 2000; Webb, 1992) provides a classification of known enzymes into functional groups. For each EC class, we retrieve all structures present in the PDB. For each structure with resolution better than 2.5 Å, we check to see if it contains bound ligands similar to the substrates required for the reaction catalyzed by the enzyme, using a similarity cutoff of 50% as defined by PDBSum (Laskowski *et al*., 2005). The structures' sequences are then clustered at the level of 95% sequence identity within each EC class, and a non-redundant set is kept for analysis. For each structure that remains in each EC class, we download the alignment from HSSP and filter it as for the catalytic site data set. For each structure, we put all residues within 4 Å of any ligand atom into the set of putative positive residues. This may include some positions that are not functionally important, but the area around the active site contains a strong enough conservation signal that we are able to distinguish between methods by the number of highly conserved positions each predicts near ligands. All remaining residues comprise our set of negatives. We are left with 828 alignments with an average of ~92 sequences per alignment. The alignments span 495 EC classes and provide an average of ~1.6 alignments per class. We also performed the analysis excluding all residues that have less than 1% solvent accessibility and obtained similar results (see Supplementary Material).

2.2.3 Protein-protein interfaces We use the data set of Caffrey *et al*. (2004) consisting of 64 PPIs: 42 homodimers, 12 heterodimers and 10 transient complexes. For each interface, they provide an alignment of close homologs and an alignment of diverse homologs. We present results for the close homolog alignments; performance is similar on the diverse alignments (Supplementary Material). Interface residues, comprising the set of positives, are defined as those losing more than 1% RSA on complex formation; as suggested by Caffrey *et al*. (2004), we compute RSA using NACCESS (Hubbard and Thornton, 1993; Lee and Richards, 1971) with a probe size of 1.4 Å. All other residues are the negatives. We also evaluated the methods by removing all positions that have less than 1% solvent accessibility in the monomer; results were similar on this modified data set (see Supplementary Material).

2.3 Evaluation methods

Conservation scoring methods are compared on a data set by considering how well they rank the positive set of functionally important residues, as well as by computing receiver operator characteristic (ROC) curves.

For ROC analysis, a ROC curve is constructed for each method on each alignment, and all the ROC curves for a method are averaged across all alignments to obtain its overall curve. For the ligand distance data set, ROC curves are first averaged over each EC class and then averaged across classes; this is done because some EC classes have more alignments than others. We report the area under the ROC curve (AUC) at a range of false positive rates: 0.1 ($AUC_{0.1}$), 0.5 ($AUC_{0.5}$) and 1.0 (AUC_1). The higher the AUC, the better the method has done at identifying functional residues.

In the rank analysis, for each alignment we compute the conservation scores for all columns and note the rank of the known functionally important columns. We report the fraction of the top 30 ranked columns that are functionally important (Wang and Samudrala, 2006); however, since the number of positives may be less than 30, we normalize the statistic so that perfect performance (i.e. all possible positives in the top 30 predictions) gets a score of one. These top-30 statistics are averaged over all alignments, and in the case of the ligand distance data set, are first averaged over EC class and then averaged over the classes.

We use the Friedman test, as implemented in Matlab, to judge whether the performance statistics (e.g. AUC_1) for the methods are significantly different. For the CSA and PPI data sets, when judging each statistic, comparisons of its value on each alignment are considered; for the ligand distance data, comparisons are made between its averaged value for each EC class. Since for each statistic, the values for all pairwise combinations of methods are compared, we further apply a Bonferroni correction to judge significance. The difference in performance of two methods using a particular statistic is called statistically significant if the *P*-value computed using the Friedman test with a Bonferroni correction is less than 0.05.

3 RESULTS

The seven methods are evaluated in their ability to identify functional sites in the three data sets. The performance statistics—averaged AUCs and top-30—for all basic methods are summarized for the catalytic site (CSA) data set in Table 1, the ligand distance data set in Table 2 and the close homolog PPI data set in Table 3. The relative performance of these methods using the AUC_1 performance statistic is also

Table 1. Performance statistics for all methods on the catalytic site data set

Method	$AUC_{0.1}$	$AUC_{0.5}$	AUC_1	Top-30
Shannon Entropy	0.0524	0.4248	0.9235	0.6783
Property Entropy	0.0338	0.3780	0.8749	0.4328
von Neumann Entropy	0.0499	0.4211	0.9166	0.6462
Sum-of-pairs measure	0.0528	0.4291	0.9271	0.6374
Relative Entropy	0.0599	0.4436	0.9428	0.7120
Rate4Site	0.0615	0.4451	0.9412	0.7240
Jensen-Shannon divergence	**0.0623**	**0.4464**	**0.9440**	**0.7338**

Area under the ROC curve is given for the 0.1 ($AUC_{0.1}$), 0.5 ($AUC_{0.5}$) and 1.0 (AUC_1) false positive rates. Top-30 is the normalized fraction of the top 30 scoring sites that are functionally important (see text). The best scores are in bold. JSD and R4S are significantly better than all other methods at the AUC_1 level.

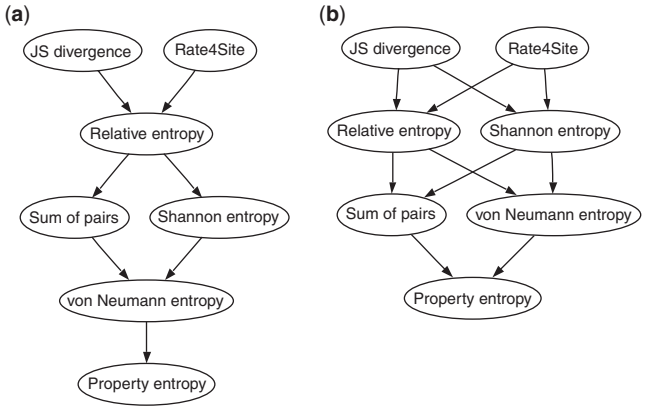

Fig. 1. Significance relationships at AUC_1 level for the (**a**) catalytic site and (**b**) ligand distance data sets. An edge from method X to method Y means that method X performs significantly better than method Y. A path between two nodes implies a significant difference as well.

Table 2. Performance statistics for all methods on the ligand distance data set

Method	$AUC_{0.1}$	$AUC_{0.5}$	AUC_1	Top-30
Shannon Entropy	0.0093	0.3238	0.8036	0.3960
Property Entropy	0.0049	0.2813	0.7590	0.2822
von Neumann Entropy	0.0089	0.3138	0.7934	0.3816
Sum-of-pairs measure	0.0086	0.3141	0.7898	0.3759
Relative Entropy	0.0098	0.3311	0.8119	0.4076
Rate4Site	**0.0109**	**0.3394**	**0.8238**	**0.4312**
Jensen-Shannon divergence	0.0107	0.3345	0.8153	0.4220

See Table 1 for a description of the statistics.

Table 3. Performance statistics for all methods on the close homolog protein interface data set

Method	$AUC_{0.1}$	$AUC_{0.5}$	AUC_1	Top-30
Shannon Entropy	0.0060	0.1352	0.5203	0.1692
Property Entropy	0.0037	0.1160	0.4968	0.1225
von Neumann Entropy	0.0059	0.1367	0.5265	0.1670
Sum-of-pairs measure	0.0069	0.1380	0.5217	0.1806
Relative Entropy	**0.0079**	**0.1529**	**0.5468**	0.1948
Rate4Site	0.0075	0.1466	0.5433	0.1772
Jensen-Shannon divergence	**0.0079**	0.1516	0.5437	**0.1960**

See Table 1 for a description of the statistics.

depicted graphically for the CSA and ligand distance data sets in Figure 1. These relationships are not shown for the PPI data set, as the only significant differences on it between methods at the AUC_1 level involve comparisons with the worst performing method. The top-30 improvement provided by using the window heuristic on each method on each data set is given in Table 4. In the following text we describe our main findings in detail.

Table 4. Improvement in top-30 performance provided by the window heuristic for all methods

Method	CSA	Ligand	PPI
SE	0.6783	0.3960	0.1692
SE+W	**0.7077**	**0.4583**	**0.2000**
PE	0.4328	0.2822	0.1225
PE+W	**0.5928**	**0.3840**	**0.1772**
VNE	0.6462	0.3816	0.1670
VNE+W	**0.6979**	**0.4422**	**0.1960**
SP	0.6374	0.3759	0.1806
SP+W	**0.6995**	**0.4264**	**0.2034**
RE	0.7120	0.4076	0.1948
RE+W	**0.7507**	**0.4546**	**0.2205**
R4S	**0.7240**	0.4312	0.1772
R4S+W	0.7197	**0.4795**	**0.2205**
JSD	0.7338	0.4220	0.1960
JSD+W	**0.7539**	**0.4703**	**0.2205**

The better score between the basic method and its windowed version is in bold. Full statistics are provided in the supplement.

3.1 JSD is not significantly outperformed by any other basic method

Tables 1, 2 and 3 show that JSD, RE and R4S perform better than the other four basic methods when considering any of the performance statistics on any of the data sets. The significance chart in Figure 1 illustrates that JSD and R4S perform significantly better at the AUC_1 level on both the CSA and ligand distance data sets than all the other methods, including RE. For the CSA data set, JSD outperforms R4S on all criteria, whereas on the ligand distance data set, R4S outperforms JSD using all criteria (Tables 1 and 2). The differences between these methods on the PPI data set are not significant, but Table 3 shows that JSD and RE are the best performing methods on

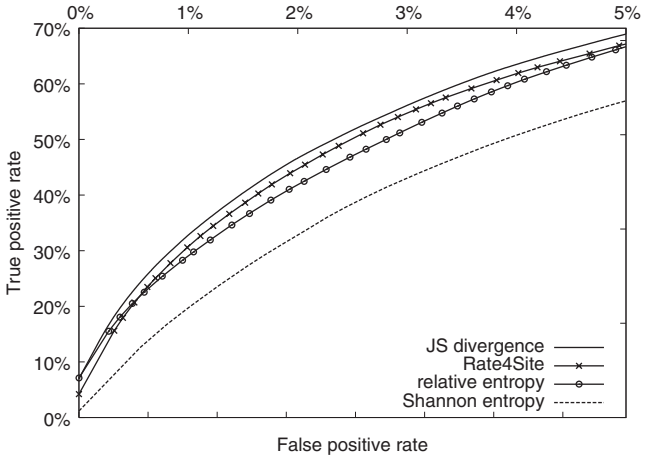

Fig. 2. High confidence region of the ROC curves for SE, JSD, R4S, and RE on the catalytic site data set. The three methods all perform significantly better than SE.

Table 5. Running time on a set of 25 alignments randomly selected from the catalytic site data set

Method	Min Time	Max Time	Average	Total
Shannon Entropy	0.18 s	1.18 s	0.45 s	11.18 s
Jensen-Shannon divergence	0.20 s	1.21 s	0.47 s	11.81 s
Sum-of-pairs measure	0.20 s	23.44 s	3.87 s	96.75 s
Rate4Site	18.66 s	1976.38 s	382.53 s	9563.22 s

The JSD takes several orders of magnitude less time than R4S and provides competitive performance. All information-theoretic methods have similar running times.

this data set. Overall, the results are similar if we consider the top-30 statistic; JSD, RE and R4S are all significantly better than the other methods on the catalytic site and ligand distance data sets. Note that while SE is probably the most commonly used method for identifying functionally important residues via conservation analysis, our evaluation shows that in all settings tested it is outperformed by other methods (Fig. 2).

3.2 JSD performs similarly to R4S, but is much faster

Overall, JSD and R4S perform similarly; none of the differences in performance observed between them on any of the data sets using any of the statistics are significant (Fig. 1). However, JSD and the other information-theoretic methods have a significant advantage over R4S when considering run time. Table 5 gives (processor) running time statistics for several methods on a benchmark set of 25 randomly chosen alignments from the CSA data set. R4S took over 2.5 h (9563.22 s) to score these 25 alignments while JSD required only 11.81 s; JSD finishes scoring all 25 in less time than R4S needs to score the smallest alignment. RE's running time is similar to JSD's. In light of this and the performance results, JSD is the best method for the estimation of conservation in contexts where speed is an issue, such as large, genome-scale analysis.

3.3 Incorporating the conservation of sequentially adjacent positions improves performance

Our heuristic for exploiting conservation scores within a sequence window around the residue of interest can be applied to any scoring method that produces independent column scores. Using a window of size seven (three residues on either side of the current residue), all methods improve on each of the three data sets (Table 4), as judged by top-30. The results are similar for the other statistics (see Supplementary Material for all performance statistics for windowed approaches). Note that the window method improves predictions for all types of functional sites, not just those with low conservation. In fact, as Table 4 shows, the improvement is greater for sites with high conservation (catalytic residues and residues near ligands) than for sites in protein interfaces.

Figure 3 shows the improvement on the CSA data set for SE and JSD when our window approach is used. The figure depicts the high-confidence region of the ROC curve. The difference between JSD+W and SE illustrates the improvement provided by methods introduced in this article; at a false positive rate of 2%, JSD+W identifies over 50% of the true positives while SE finds only ~30%. Note that when SE is extended to incorporate the conservation of sequentially adjacent positions, it performs nearly as well as the basic JSD method. This highlights the power of simply using the window approach with existing scoring methods. The consistent improvement provided by the window heuristic suggests that it can improve predictions in a range of settings.

3.4 Incorporating relationships between amino acids is not always helpful

Three of the methods considered, VNE, SP and PE, attempt to incorporate information about the similarity of amino acids. One would expect that, since pairs of amino acids have differing physicochemical similarities, incorporation of such information would improve upon other methods that do not. Our evaluation framework allows us to assess this claim by characterizing the performance of VNE, SP and PE relative to the commonly used SE.

Figure 1 shows that using the AUC_1 criterion, SE is significantly better than VNE, PE and SP on the ligand distance data set, and significantly better than VNE and PE on the CSA data set. While SP performs better than SE on the CSA data set using the AUC_1 criterion, the difference is not statistically significant. The differences between these methods on the PPI data set are not significant; however, SE performs best of the four as judged by top-30, SP performs best as judged by $AUC_{0.1}$ and $AUC_{0.5}$ and VNE performs best as judged by AUC_1.

We also evaluated the effect of using different BLOSUM matrices with VNE and SP. We found that the choice of matrix does not change our overall results. For the SP method, we additionally experimented with alignment-specific matrices. In particular, for each alignment considered, we computed the average pairwise sequence identity and selected the nearest of BLOSUM45, BLOSUM62 and BLOSUM80. This scheme also did not improve overall results (see Supplementary Table 8).

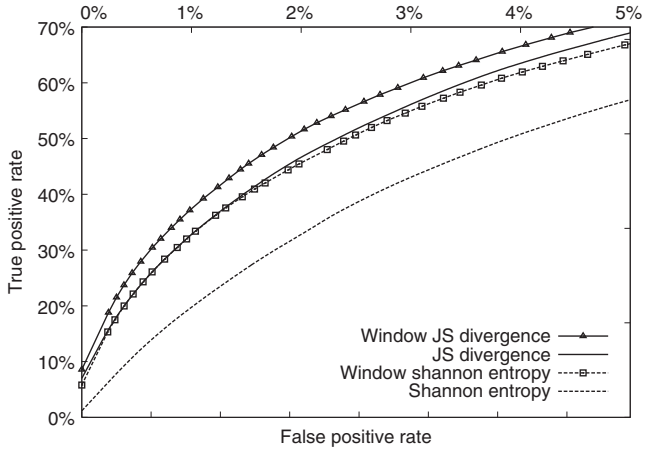

Fig. 3. High confidence ROC curves demonstrating the improvement for SE and JS divergence when used with the window method on the CSA set. The difference between SE (dashes) and Window JSD (triangles) is the improvement provided by methods introduced in this article. Similar improvement is seen across methods and data sets.

These results highlight the need for large-scale evaluation. While it might be expected that PE, VNE and SP would improve on SE, none provide any significant gain. In fact, in several settings, some perform significantly worse than SE. VNE was introduced for the prediction of PPI residues (Caffrey *et al*., 2004), but is not significantly better on the PPI data set as judged by any of the four statistics tested. PE was introduced for the analysis of ligand recognition in transport proteins, and SP for the analysis of DNA-binding proteins. It is possible that these methods could achieve better performance in other specific settings, but the three contexts investigated here are quite common and similar to those in which they were introduced.

3.5 Identifying residues in the PPI from conservation alone is difficult

Recently, conservation analysis has been employed to predict and analyze protein-protein interaction sites (Border and Abagyan, 2005; Caffrey *et al*., 2004 and Guharoy and Chakrabarti, 2005). Several of these groups have found that it is difficult to predict the interface using various measures of conservation. Here, we find that none of the seven conservation measures studied perform particularly well in identifying residues in protein interfaces (Table 3). We report statistics from the close homolog alignments, but results are similar for the diverse homolog set (see Supplementary Material). The AUC_1 values for all methods are approximately 0.55, as compared to much better performance in identifying catalytic site residues ($AUC_1 \approx .95$) and the sites near ligands ($AUC_1 \approx .80$). However, while the conservation signal on the interface is weak, it is still detectable; all methods other than PE performed significantly better than random guessing. This suggests that conservation alone should not be used to predict residues in PPI. However, it is an important component of ensemble-based approaches (Chung *et al*., 2006).

4 DISCUSSION AND CONCLUSION

Despite the prevalence of conservation-based analysis, there are few agreed upon best practices. This article establishes an empirical understanding of the relationships between approaches. We describe several methods for quantifying conservation and introduce a method based on the JSD as well as a heuristic for incorporating the conservation signal from sequentially neighboring residues. We then quantitatively compare the performance of all methods in three realistic settings: the identification of catalytic sites, residues near ligands and residues comprising PPIs.

Our evaluation demonstrates that methods such as JSD and RE that incorporate a background amino acid distribution are preferable to SE (Fig. 2). R4S also provides similar improvement over SE, but is quite slow in comparison to the information theoretic methods (Table 5). The speed of JSD would allow researchers to modify alignments and re-predict functional sites on the fly. It also makes large-scale analysis faster and more appealing. While JSD and RE are similar measures, overall RE does not perform quite as well as JSD. RE is unbounded; events that are unlikely according to the background or column distributions tend to contribute more to the RE score than to the JSD score, and this likely causes the difference in performance between the two methods.

We also demonstrate that our window heuristic provides a way to boost the conservation signal, and thus performance, even in the absence of structural information. This improvement is seen across methods and data sets. The approach is fast, flexible and can be applied to any method that produces column scores.

Perhaps most surprisingly, we find that several methods that intend to improve conservation estimation by incorporating amino acid similarity fail to provide any significant improvement over methods that ignore the underlying chemistry. In fact, some perform significantly worse than SE. While it may be the case that incorporating amino acid similarity is not critical for identifying functional sites, it is more likely that the existing set of methods are not adequate, and other as yet undeveloped methods may be able to exploit better the similarities between amino acids. Additionally, it is possible that the data sets of known functional sites are biased towards absolutely conserved residues, and thus incorporating relationships between amino acids is not essential for good performance on them.

The poor performance of all methods on the PPI data set demonstrates that the difficulties encountered in previous attempts (Bordner and Abagyan, 2005; Caffrey *et al*., 2004) exist across a range of conservation methods. It is likely that the results could be improved by dividing the data set into transient and obligate interactions (Mintseris and Weng, 2005) and further dividing the interface into central and peripheral residues. Nevertheless, it is clear that conservation alone is insufficient to predict all residues in PPIs.

When interpreting the predictions made by a conservation-based method, it is natural to ask whether a site is important for maintaining structure, for catalysis, or for binding ligands, other proteins or DNA. Conservation alone cannot distinguish among these possibilities; however, features such as amino acid composition, electrostatic potential and known or predicted

structural properties (e.g. secondary structure and solvent accessibility), used along with conservation, can be used within machine-learning methods to identify particular types of functional residues (Bordner and Abagyan, 2005; Petrova and Wu, 2006).

Overall our results highlight the necessity for rigorous evaluation of conservation methods. Conservation analysis is beginning to be applied in settings where the signal is not strong (e.g. the prediction of protein interaction sites). Thus, comprehensive analyses such as the one performed here are increasingly important in order to develop an empirical understanding of the strengths and weaknesses of various methods; this understanding can then be used to guide development of more powerful techniques for estimating sequence conservation in diverse biological settings.

ACKNOWLEDGEMENTS

J. A. C. has been supported by the Quantitative and Computational Biology Program NIH grant T32 HG003284. M. S. thanks the NSF for grants IIS-0612231 and PECASE MCB-0093399, and the NIH for grant GM076275. This research has also been supported by the NIH Center of Excellence grant P50 GM071508. The authors would like to thank Prof. Tom Funkhouser and members of the Singh group for helpful discussions. Funding to pay the Open Access publication charges was provided by NSF grant IIS-0612231.

Conflict of Interest: none declared.

REFERENCES

Bairoch,A. (2000) The enyzme database in 2000. *Nucleic Acids Res.*, **28**, 304–305.

Bartlett,G. *et al.* (2002) Analysis of catalytic residues in enzyme active sites. *J. Mol. Biol.*, **324**, 105–121.

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

Bordner,A. and Abagyan,R. (2005) Statistical analysis and prediction of protein-protein interfaces. *Proteins*, **60**, 353–366.

Caffrey,D. *et al.* (2004) Are protein-protein interfaces more conserved in sequence than the rest of the protein surface? *Protein Sci.*, **13**, 190–202.

Chung,J. *et al.* (2006) Exploiting sequence and structure homologs to identify protein-protein binding sites. *Proteins*, **62**, 630–640.

Cover,T. and Thomas,J. (1991) *Elements of Information Theory*, John Wiley and Sons, New York.

Dodge,C. *et al.* (1998) The hssp database of protein structure-sequence alignments and family profiles. *Nucleic Acids Res.*, **26**, 313–315.

Durbin,R. *et al.* (1998) *Biological Sequence Analysis: probabilistic models of proteins and nucleic acids*. Cambridge University Press, Cambridge, UK.

Elcock,A. (2001) Prediction of funtionally important residues based solely on the computed energetics of protein structure. *J. Mol. Biol.*, **312**, 885–896.

Fetrow,J. and Skolnick,J. (1998) Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T1 ribonucleases. *J. Mol. Biol.*, **281**, 949–968.

Guharoy,M. and Chakrabarti,P. (2005) Conservation and relative importance of residues across protein-protein interfaces. *Proc. Natl Acad. Sci. USA*, **102**, 15447–15452.

Gutteridge,A. *et al.* (2003) Using a neural network and spatial clustering to predict the location of active sites in enzymes. *J. Mol. Biol.*, **330,** 719–734.

Hannenhalli,S. and Russell,R. (2000) Analaysis and prediction of functional subtypes from protein sequence alignments. *J. Mol. Biol.*, **303**, 61–76.

Henikoff,S. and Henikoff,J. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.

Henikoff,S. and Henikoff,J. (1994) Position-based sequence weights. *J. Mol. Biol.*, **243**, 574–578.

Hubbard,S. and Thornton,J. (1993) Naccess. Computer Program.

Jones,S. and Thornton,J. (2004) Searching for functional sites in protein structures. *Curr. Opin. Chem. Biol.*, **8**, 3–7.

Kalinina,O. *et al.* (2003) Automated selection of positions determining functional specificity of proteins by comparative analysis of orthologous gropus in protein families. *Protein Sci.*, **13**, 443–456.

Karlin,S. and Brocchieri,L. (1996) Evolutionary conservation of reca genes in relation to protein structure and function. *J. Bacteriol.*, **178**, 1881–1894.

Landau,M. *et al.* (2005) Consurf 2005: the projection of evolutionary conservation scores of residues on protein structures. *Nucleic Acids Res.*, **33**, W299–W302.

Laskowski,R. *et al.* (2005) Pdbsum more: new summaries and analyses of the known 3d structures of proteins and nucleic acids. *Nucleic Acids Res.*, **33**, D266–D268.

Lee,B. and Richards,F. (1971) The interpretation of protein structures: Estimation of static accessibility. *J. Mol. Biol.*, **55**, 379–400.

Liang,S. *et al.* (2006) Protein binding site prediction using and empirical scoring function. *Nucleic Acids Res.*, **34**, 3698–3707.

Lichtarge,O. *et al.* (1996) An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.*, **257**, 342–358.

Lin,J. (1991) Divergence measures based on the shannon entropy. *IEEE Trans. Inf. Theory*, **37**, 145–151.

Magliery,T. and Regan,L. (2005) Sequence variation in ligand binding sites in proteins. *BMC Bioinformatics*, **6**, 240.

Mayrose,I. *et al.* (2004) Comparison of sitespecific rate-inference methods for protein sequences: Empirical bayesian methods are superior. *Mol. Biol. and Evol.*, **21**, 1781–1791.

Mintseris,J. and Weng,Z. (2005) Structure, function, and evolution of transient and obligate protein-protein interactions. *Proc. Natl Acad. Sci. USA*, **102**, 10930–10935.

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

Nielsen,M. and Chuang,I. (2000) *Quantum Computation and Quantum Information*. Cambridge University Press, Cambridge, UK.

Ondrechen,M. *et al.* (2001) Thematics: a simple computational predictor of enzyme function from structure. *Proc. Natl Acad. Sci. USA*, **98**, 12473–12478.

Panchenko,A. *et al.* (2004) Prediction of functional sites by analysis of sequence and structure conservation. *Protein Sci.*, **13**, 884–892.

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

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

Sander,S. and Schneider,R. (1991) Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins*, **9**, 56–68.

Schueler-Furman,O. and Baker,D. (2003) Conserved residue clustering and protein structure prediction. *Proteins*, **52**, 225–235.

Shenkin,P. and Erman,B.L.M. (1991) Information-theoretical entropy as a measure of sequence variability. *Proteins*, **11**, 297–313.

Stark,A. and Russell,R. (2003) Annotation in three dimensions. PINTS: patterns in non-homologous tertiary structures. *Nucleic Acids Res.*, **31**, 3314–3344.

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

Valdar,W. and Thornton,J. (2001) Conservation helps to identify biologically relevant crystal contacts. *J. Mol. Biol.*, **313**, 399–416.

Wallace,A. *et al.* (1997) Tess: a geometric hasing algorithm for deriving 3d coordinate templates for searching structural databases. *Protein Sci.*, **6**, 2308–2323.

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

Webb,E. (1992) *Enzyme Nomenclature. Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology*. Academic Press, New York.

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.

Yona,G. and Levitt,M. (2002) Within the twilight zone: a sensitive profile-profile comparison tool based on information theory. *J. Mol. Biol.*, **315**, 1257–1275.