

Sequence analysis

Improving position-specific predictions of protein functional sites using phylogenetic motifs

Dukka Bahadur K. C. and Dennis R. Livesay*

Department of Computer Science and Bioinformatics Research Center, University of North Carolina at Charlotte, Charlotte, NC 28223, USA

Received on May 20, 2008; revised on August 14, 2008; accepted on August 20, 2008

Advance Access publication August 21, 2008

Associate Editor: Burkhard Rost

ABSTRACT

Motivation: Accurate computational prediction of protein functional sites is critical to maximizing the utility of recent high-throughput sequencing efforts. Among the available approaches, position-specific conservation scores remain among the most popular due to their accuracy and ease of computation. Unfortunately, high false positive rates remain a limiting factor. Using phylogenetic motifs (PMs), we have developed two combined (conservation + PMs) prediction schemes that significantly improve prediction accuracy.

Results: Our first approach, called position-specific MINER (psMINER), rank orders alignment columns by conservation. Subsequently, positions that are also not identified as PMs are excluded from the prediction set. This approach improves prediction accuracy, in a statistically significant way, compared to the underlying conservation scores. Increased accuracy is a general result, meaning improvement is observed over several different conservation scores that span a continuum of complexity. In addition, a hybrid MINER (hMINER) that quantitatively considers both scoring regimes provides further improvement. More importantly, it provides critical insight into the relative importance of phylogeny versus alignment conservation. Both methods outperform other common prediction algorithms that also utilize phylogenetic concepts. Finally, we demonstrate that the presented results are critically sensitive to functional site definition, thus highlighting the need for more complete benchmarks within the prediction community.

Availability: Our benchmark datasets are available for download at <http://www.cs.uncc.edu/~drlivesa/dataset.html>.

Contact: drlivesa@uncc.edu

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

1 INTRODUCTION

Accurate prediction of protein functional sites from sequence- and structure-derived data remains an open problem (Pazos and Bang, 2006; Watson *et al.*, 2005). The most common solutions to this problem are based on relatively simple conservation scores that are used to rank the importance of alignment positions (Capra and Singh, 2007; Mayrose *et al.*, 2004; Valdar, 2002). The argument supporting such an approach is that evolution has

drastically suppressed variation at the conserved positions due to their functional importance. However, conservation scores are frequently plagued by an unsatisfactory number of false positives. In response, Cheng *et al.* (2005) developed an elegant hybrid conservation/protein design algorithm to tease out functional from structural conserved positions. The method uses all-atom protein design (Alm *et al.*, 2002; Kuhlman and Baker, 2000) to identify structurally important sites that were subsequently removed from the list of possible functional sites. This approach is quite elegant, but is limited by computational expense and the need, of course, for a structural template. Similar studies have used a variety of biophysical calculations in the same manner ((Chelliah *et al.*, 2004; Pei *et al.*, 2003). Below, we present an analogous method to filter out putative false positives, but do so in a way that drastically reduces computational complexity and eliminates the dependence upon solved structures.

Myriad protein functional site prediction methods that go beyond alignment conservation have been described in the literature; unfortunately, most are based on structurally derived features. For example, a number of network centrality based approaches, where protein structures are described as topological graphs, have been demonstrated to predict enzyme catalytic residues with good success (Amitai *et al.*, 2004; Chea and Livesay, 2007; Thibert *et al.*, 2005). In addition, Poisson–Boltzmann continuum theory can also be used to predict catalytic residues from atypical titration profiles (Ondrechen *et al.*, 2001) and from electrostatic strain energies (Elcock, 2001). Additionally, there are a number of functional site prediction methods based on machine learning approaches, primarily support vector machines (Bradford and Westhead, 2005; Cai *et al.*, 2004; Petrova and Wu, 2006) and artificial neural networks (Blom, N. *et al.*, 1999; Gutteridge *et al.*, 2003; Pande *et al.*, 2007). These latter approaches are attractive due to their unparalleled ability to combine a wide variety of inputs (i.e. conservation, residue type, secondary structure, solvent accessibility, etc.). However, their appeal is reduced by their ‘black box’ nature, meaning it is quite difficult to glean knowledge from their outputs.

Arguably, the most common functional site prediction methods (outside of strict conservation scores) are based on phylogenetic information. For example, evolutionary trace (ET) uses a phylogenetic tree to cluster sequences within a multiple alignment (Lichtarge *et al.*, 1996). Subsequently, trace residues, which are alignment positions that are conserved within the phylogenetic clusters, are identified. Finally, structural clusters of conserved

*To whom correspondence should be addressed.

positions and the trace residues are used to identify functional regions (Aloy *et al.*, 2001; Madabushi *et al.*, 2002, 2004; Yao *et al.*, 2003). Related methods that utilize phylogenetic information include SMERFS (Manning *et al.*, 2008) and the mutational behavior method (MB-method) (del Sol *et al.*, 2003). These approaches provide key insight into subfamily differentiation (Madabushi *et al.*, 2004; Pritchard and Dufton, 1999), and in the case of ET, have been used to identify experimentally verified allosteric pathways (Sowa *et al.*, 2001). Nevertheless, in spite of the importance of these methods, we demonstrate below that when considering alignment information alone, they are unable to predict functional sites as well as the simpler conservation scores.

Similar in spirit to the above approaches, we have developed a protein functional site prediction scheme based on phylogenetic motifs (PMs) (La and Livesay, 2005; La *et al.*, 2005). PMs, which are identified using MINER (La and Livesay, 2005), are alignment fragments that best recapitulate the overall phylogeny. We have previously demonstrated that PMs accurately predict protein functional sites, specifically active and ligand-binding sites, from alignment information alone (Livesay and La, 2005; Livesay *et al.*, 2007; Roshan *et al.*, 2005). Ostensibly, PMs identify sequence clusters of ET positions, and, as expected, the results are somewhat similar. However, compared to raw ET predictions without structural filtering, PMs are more likely to coincide with functional regions (La *et al.*, 2005; Livesay *et al.*, 2007).

The ability of PMs to highlight active site regions from alignment alone is very promising; however, the fragment-based approach limits their specificity. As such, we have developed a position-specific MINER (psMINER) algorithm that attempts to leverage the most attractive prediction characteristics of PMs (ability to highlight active site regions) and conservation scores (position-specific descriptions of importance). The psMINER method represents the first incorporation of conservation information into the PM framework. Unlike the other phylogeny-based methods considered here, the simple approach results in a statistically significant improvement over conservation scores alone no matter which conservation metric is considered. Moreover, the performance improvement comes at minimal computational cost. In addition, a hybrid MINER (hMINER) that quantitatively considers both scoring regimes is also presented. While hMINER results in further performance improvements, its most important consequence is that the method highlights how ideal relative weightings of the two regimes is entirely context dependent. Finally, an additional important result from this work is that the extent of improvement is critically dependent upon the functional site definition. Herein, we assess the developed algorithms against benchmarks composed of enzyme catalytic residues, active sites and ligand-binding sites.

2 METHODS

2.1 Quantifying alignment conservation

Alignment conservation is calculated using four different conservation scores: the sum-of-pairs (SP) score, the Williamson property entropy (WPE), the Jensen–Shannon divergence (JSD) and the Rate4Site algorithm. Each is introduced below.

2.1.1 sum-of-pair score The SP score is one of the most commonly used conservation measures. The SP score is generally a much more powerful method than simpler metrics that fail to consider the likelihood

of residue-to-residue substitution patterns (e.g. percent conservation and the Shannon entropy). Rather, SP scores use scoring matrix substitution values to determine the amount of conservation within an alignment column. The SP score of column m_i , designated $S(m_i)$, is calculated as:

$$S_i^{\text{SP}} = \sum_{i=1}^K \sum_{j>i} s(m_i, m_j) \quad (1)$$

where $s(m_i, m_j)$ is the scoring matrix substitution value. The sum is enumerated over all possible pairs within a single alignment column. Since larger $s(m_i, m_j)$ values indicate greater similarity, larger SP scores indicate greater conservation. Herein, we use the program Scorecons (Valdar, 2002) to calculate SP scores.

2.1.2 Williamson property entropy WPE is a relative entropy measure that explicitly attempts to introduce physicochemical considerations into the measure. WPE is calculated by:

$$S_i^{\text{WPE}} = \sum_{i=1}^K p_i \ln \left(\frac{p_i}{\langle p_i \rangle} \right) \quad (2)$$

where $K=9$ are different physicochemical classes, p_i is the probability of physicochemical class i in that column and $\langle p_i \rangle$ is the average class probability over all alignment columns. The WPE is a significant improvement over standard methods because its reduced alphabet is more forgiving of chemically conservative mutations. In fact, a recent study by Manning *et al.* (2008) concluded that the WPE was the best conservation measure (of 16 that were considered) for predicting domain interfacial residues and ligand-binding sites. Capra and Singh (2007) have reported that incorporating information from sequential residues can substantially improve functional site prediction; their results indicate that inclusion of three positions on either side of i works well. As such, we compute S_i^{WPE} over a window of seven alignment positions, where i is at the center.

2.1.3 Jensen–Shannon divergence The third conservation score considered is an information-theoretic approach based on the JSD (Capra and Singh, 2007). The JSD is a common method used to compare probability distributions. In this application, the first distribution is based on the amino acid distribution in column i , and the second distribution is based on a background distribution (calculated from the BLOSUM62 alignments). The JSD is calculated by:

$$S_i^{\text{JSD}} = \frac{1}{2} \text{RE}_{p,r} + \frac{1}{2} \text{RE}_{q,r} \quad (3)$$

where $\text{RE}_{p,r}$ and $\text{RE}_{q,r}$ are relative entropies of the same form as Equation (2), comparing the column and background distributions (p and q , respectively) to $r=(p+q)/2$. Again, sequential residues are included to improve predictive power. As with WPE, a window width of seven, which is used herein, has been determined to be ideal (Capra and Singh, 2007). We compute JSD and WPE scores using software provided by the Singh lab.

2.1.4 Rate4site The last conservation score considered is Rate4Site (R4S) (Pupko *et al.*, 2002). R4S constructs a rather compute-intensive description of the underlying phylogeny in order to improve determining the rate of evolution at each site. The rate of evolution at each site is then estimated using the maximum likelihood principle, which considers both phylogenetic tree branch lengths and the stochastic nature of evolution. Subsequently, R4S normalizes the calculated evolutionary rates such that the average is 0 and the SD is 1. The Rate4Site software is freely provided by the Ben-Tal lab, and is used with default parameters. Lower values indicate greater conservation.

2.2 The psMINER algorithm

As deftly argued by Cheng *et al.* (2005), the chief problem with conservation-based predictions of important protein sites is that they fail to discriminate between structural and functional sites. To circumvent this problem, they used all-atom protein design to identify putative structural sites that were

subsequently subtracted from the list of highly conserved sites. Conversely, we use PMs to confirm functional importance; meaning, non-PM positions are simply subtracted from the list of highly conserved sites.

MINER (La and Livesay, 2005) compares the topology of each alignment window (width=5) tree to that of the complete phylogeny using a modified bipartition metric (Roshan *et al.*, 2005) that counts topological differences. These raw differences are subsequently converted into phylogenetic similarity z -scores (PSZs), where PSZs <0 indicate similarities better than the mean. All overlapping windows that score past a signal-to-noise threshold are collapsed into a single PM. We have developed a MINER-Extreme algorithm that automatically determines good estimates of the signal-to-noise threshold (La and Livesay, 2005). The algorithm attempts to determine, based on PSZs, if a window is likely functional or not. Ideal PSZ thresholds, which are dataset dependent, generally occur between -1.0 and -2.0 . Windows with PSZs >-1.0 are not expected to be functional, whereas windows with PSZs <-2.0 are. The MINER-Extreme algorithm uses partition around medoids clustering (PAMC) of PSZs to automatically identify good estimates of threshold values. PAMC is a clustering algorithm analogous to partition around means (which is sometimes called k -means clustering). However, each cluster is represented by its median position instead of its average, making it less biased by outliers. In the MINER-Extreme algorithm, PSZs between -1.0 and -2.0 are clustered into $k=2$ groups, one corresponding to noise and the other corresponding to signal. In most circumstances, the ideal threshold is defined to be the largest value in the signal cluster. When there are too few or too many data points between -1.0 and -2.0 , there is an algorithmic override that sets the threshold to be the first (rank ordered) PSZ below -2.0 .

It should be pointed out that the MINER approach is conceptually similar to the MB-method (del Sol *et al.*, 2003) and SMERFS (Manning *et al.*, 2008). The MB-method compares the mutational behavior of each alignment position, represented by a residue pair similarity matrix, to that of the whole protein, which is represented by a similarity matrix for all pairs in the protein family. The two matrices are compared using a rank-correlation criterion. Alignment columns that best parallel the familial mutation behavior are predicted as functional sites. Similarly, SMERFS compares window-based distance matrices to that of the whole alignment. While these methods are conceptually similar to MINER, there is a key difference. That being, MINER evaluates phylogenetic similarity using a topological comparison versus the matrix comparisons used by SMERFS and the MB-method. We (D.B.K.C. and D.R.L.) have recently evaluated a large number of matrix comparison methods in MINER (ranging from correlation coefficients to information theory metrics); however, none have the predictive power of our current topological comparison (unpublished data). Consequently, it is our hypothesis that MINER's predictive power is derived from topological information.

The psMINER algorithm employed here is extremely simple (Fig. 1). The method begins by rank ordering all alignment positions based on whichever conservation score is being considered. Subsequently, PMs are identified (within the same alignment) using MINER in the traditional manner. Finally, alignment positions that were not identified as PMs have their conservation score reset to the lowest possible value. Over the section of the receiver operating characteristic (ROC) curve in which we evaluate our method (see below), this process prevents non-PM positions from ever being considered as a functional site prediction. When it was originally developed, we purposely designed the MINER-Extreme algorithm to be overly strict (meaning it favors false negatives over false positives). However, in the context of the psMINER algorithm we find it to be overly strict (data not shown). After considering a range of PSZ thresholds around the one identified by MINER-Extreme, we find that simply relaxing it by 0.5 SDs gives the best results.

2.3 The hMINER algorithm

Hybrid approaches that quantitatively combine conservation scores with other metrics have proven quite powerful (Mihalek *et al.*, 2004). In this

(a) The psMINER algorithm:

1. For each alignment column, compute raw conservation score.
2. For each alignment window (width = 5):
 - a. Compute PSZ score,
 - b. Compute PSZ threshold using MINER Extreme,
 - c. If PSZ $<$ PSZ threshold, then the window is a PM.
3. If alignment column does not overlap a PM, reset conservation score to zero.

(b) The hMINER algorithm:

1. For each alignment column, compute raw conservation score.
2. For each alignment column \bullet {1, 2, n-1, or n}, compute Z_i^{cent} .
3. Calculate hMINER score using Equation (4).

Fig. 1. Descriptions of the (a) psMINER and (b) hMINER algorithms.

regard, besides simply filtering out non-PM alignment columns, it is possible that a hybrid method that considers both metrics could further improve prediction accuracy. However, it is not straightforward to quantitatively combine the two regimes since one is window-based (MINER), whereas the other is position-specific. In order to resolve this problem, we devise two new types of position-specific PSZs. The first, Z_i^{cent} , assigns column i the PSZ score for the window that is centered on i . The second, Z_i^{max} , assigns column i the maximal PSZ score of the five windows that it is found within. (Note that i will be found in less than five windows at the ends of the alignment.) In each case, the score is normalized from 0 to 1.

After normalization of the metrics, a simple hybrid MINER (hMINER) approach is obtained by performing linear combinations of the two:

$$H_i = \alpha Z_i^X + (1 - \alpha) S_i^Y \quad (4)$$

where α is a weight coefficient, Z_i^X is the modified PSZ score and S_i^Y is one of four conservation scores. We examine a series of values of α ranging (0.0, 1.0) and compared the performance of the hybrid method for each type of functional site.

2.4 Functional site benchmarks

As discussed within Dessailly *et al.* (2007), one of the biggest stumbling blocks to development of new and improved functional site prediction methods is the development of complete functional site benchmarks. Too many investigations of functional site prediction methods continue to assess their methods on a single definition of importance. The most common benchmark is based on enzymatic catalytic sites. However, evolution selects for a wide array of sites and it remains unclear how varying functional site definitions affects the assessment. As such, we benchmark psMINER and hMINER against three different functional site definitions: active sites, ligand-binding sites and catalytic residues. In order to avoid introducing sampling bias, we have designed our dataset so that each member will include at least one site from each of the three classes. Our dataset is based on a structurally non-redundant subsection of the Catalytic Site Atlas (CSA) (Porter *et al.*, 2004), thus every member of the benchmark is assured to have at least one catalytic residue (as defined by the CSA). Catalytic residues are extremely well conserved, whereas other positions within the active site region can be more variable. To test our ability to predict which positions define active site structure, we have identified all residues contacting the catalytic residues using HBPLUS (McDonald and Thornton, 1994). The union of these 'secondary catalytic sites' and the CSA catalytic residues define the active site benchmark. Finally, ligand-binding sites are defined by identifying (also with HBPLUS) all enzyme-ligand interactions.

We have explicitly chosen to consider each benchmark independently, meaning a positive for one benchmark may be a negative for another. As such, this approach provides a lower bound on predictive power. The only general exception to this is the active site benchmark, which, by definition, must include the catalytic residues. Finally, it is possible for a residue to occur in both the ligand-binding site and either the catalytic residue or active site benchmarks.

Construction of our dataset begins with 581 manually annotated CSA entries. Any structure without a substrate is removed in order to assure each functional class is present within each benchmark class. Homologous sequences for the remaining proteins are obtained from five rounds of PSI-BLAST (Altschul *et al.*, 1997) search against the SwissProt/TrEMBL database (Boeckmann *et al.*, 2003) using an *E*-value cut-off of 0.01 on a local machine. Once all homologs are collected, an all-to-all pairwise sequence comparison is performed in order to remove highly similar (>85% pairwise identity) and dissimilar (<25% pairwise identity) sequences. As dictated by the requirements of MINER, all families with less than 25 sequences are removed from the dataset. Finally, as we have done previously (Chea and Livesay, 2007), the remaining entries are filtered based on SCOP family in order to assure structural uniqueness. When two or more entries belong to the same family, the one with greatest number of sequences is retained. Our final dataset is composed of 163 structurally non-redundant enzyme families. The average number (and SD) of residues within each benchmark are: active site = 22.2 (9.4), ligand-binding site = 18.0 (15.5) and catalytic residue = 3.1 (1.5). Complete statistics describing the dataset are provided in Supplementary Table 1.

2.5 Prediction assessment and statistical significance

ROC analysis is used to assess prediction performance. ROC curves compare true and false positive rates over a continuum of prediction thresholds, which herein correspond to varying conservation scores. Thus, the area under the curve (AUC) provides a global assessment of predictive power. However, as we have discussed before (Chea and Livesay, 2007), methods that perform well at low false positive rates (versus the complete ROC curve) are of more practical benefit. It follows, based on the small number of residues within the benchmark, that it might make sense to consider predictive power against the catalytic residue benchmark at even lower false positive rates. However, defining a definitive false positive rate threshold remains an open problem. As such, we assess the methods considered herein against all three benchmarks at false positive rates of 0.05, 0.10, 0.15 ($AUC_{0.05}$, $AUC_{0.10}$ and $AUC_{0.15}$, respectively), which is in line with other functional site prediction studies (Capra and Singh, 2007; Cheng *et al.*, 2005; Manning *et al.*, 2008).

While, it is possible to assign statistical significance to differences within the full ROC curve (Vergara *et al.*, 2008), there is, unfortunately, no test to assign significance to $AUC_{<1.00}$ differences. Nevertheless, differences within the number of true positives at a single false positive rate can be tested for statistical significance. To do so, we employ the modified McNemar's test used by Manning *et al.* (2008). McNemar's test (McNemar, 1947) is based on χ^2 -statistic, from which *P*-values can be computed. However, it should be pointed out that differences within AUCs and number of false positives have fundamentally different meanings. AUC_X evaluates performance over the continuum of false positive rates from 0 to *X*, whereas McNemar's test only considers performance at *X*.

3 RESULTS

3.1 Raw conservation scores

Before evaluating the performance of the psMINER and hMINER algorithms, we use this opportunity to reconsider the performance of the various conservation scores. This assessment is important because of the contradictory results that have recently been presented in the literature. For example, the recent report by Manning *et al.* (2008) indicates that WPE is superior to other conservation measures. However, they failed to consider the two best performing methods (namely, R4S and JSD) of the Capra and Singh (2007) investigation. (Note that Manning *et al.* did briefly mention JSD within a note added to the proof.)

The ROC curves for each prediction method up to false positive rates of 0.15 are shown in Figure 2a–c. At a false positive rate of 0.15,

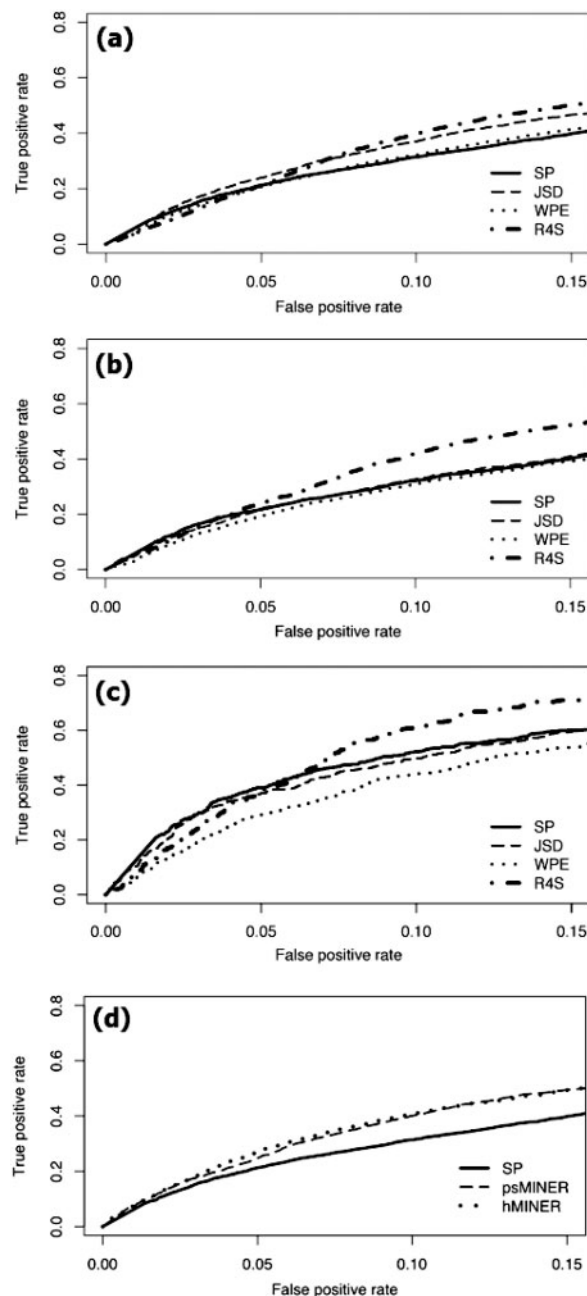


Fig. 2. ROC curves of the four conservation scores on the (a) active site, (b) ligand-binding site and (c) catalytic residue benchmarks. In (d), psMINER and hMINER active site predictive power are compared to that of the SP score. Similar trends are observed within the other conservation scores. See Supplementary Figures 1–4 for a complete set of comparisons over each functional site definition and all four considered conservation scores.

Figure 2 clearly indicates that R4S outperforms the other methods regardless of the benchmark used. The performance gap between R4S and the other methods is most stark in the ligand-binding site and catalytic residue, whereas all methods are much closer in their ability to predict active sites. In fact, JSD performs nearly as well as R4S ($AUC_{0.15} = 4.35 \times 10^{-2}$ versus 4.37×10^{-2}).

As discussed above, there are currently no methods for assigning statistical significance to differences within AUC_X values. As such, we assign significance to differences within the number of true positives at a *fixed* false positive rate using McNemar's test (Supplementary Tables 2–4). The improvement by R4S at $X=0.15$ over the other methods is always significant within the ligand-binding site benchmark; however, interestingly, only one (R4S versus WPE) of the six pairwise differences is significant within the catalytic residue benchmark. In the active site benchmark, both R4S and JSD have significant improvements over WPE; the improvement by R4S over SP is also significant.

The results are much different at low false positive rates. In fact, there is little consensus over the three benchmarks at a false positive rate of 0.05; however, somewhat surprisingly, SP scores have the best average rank ordered AUC, followed by JSD, R4S and WPE, respectively. Additionally, it should be pointed out that R4S is never the best prediction method at a false positive rate of 0.05. In fact, it is the worst at predicting active site residues over this range. When considering significance at $X=0.05$, SP, JSD and R4S are all significantly better than WPE on the ligand-binding site. There is only one difference each in the remaining two benchmarks that is considered significant. Taken together, these results highlight the precariousness of drawing firm conclusions regarding predictive power from a single AUC_X value.

It is also interesting to compare relative prediction accuracies across the three functional site benchmarks. The conservation scores are best at, not surprisingly, predicting catalytic sites, whereas the methods generally do worst at predicting ligand-binding sites. Active site prediction accuracy falls in between. While these trends are generally upheld across the four conservation scores, the exact details of the performance gap separating each of the benchmarks varies significantly. Based on these results, it is expected that prediction of active sites and ligand-binding sites is most likely to be improved by including additional information.

To summarize this section, our results suggest that a unique assessment of the relative predictive power of various protein functional site prediction algorithms is dubious. Rather, these results clearly indicate that the observed trends depend critically upon the underlying functional site benchmark and how deep into the false positive regime one is willing to consider. Nevertheless, a few general trends have emerged. First, R4S is generally the best performing prediction method at $X=0.15$. Second, the performance gap between R4S and the others is the greatest within the ligand-binding site and catalytic residue benchmarks. However, McNemar's test reveals only the differences within the ligand-binding site benchmark to be statistically significant. Finally, either JSD or SP scores is the best of the remaining two methods; however, the power of the three methods is actually quite similar. This is especially true in the ligand-binding site benchmark; in fact, the $AUC_{0.15}$ of JSD, WPE and SP score varies by <3%, none of which are considered to be statistically significant (for statistics comparing the four methods over the full ROC curve see Supplementary Tables 5–6).

3.2 Improved prediction accuracy via psMINER

Figure 2d compares ROC curves of the raw and PM filtered (by psMINER) conservation scores up to false positive rates of 0.15. Note that exact AUC values at three different false positive

Table 1. Performance of psMINER versus the raw conservation scores (all values are $\times 10^{-2}$)

	$AUC_{0.05}$		$AUC_{0.10}$		$AUC_{0.15}$	
	Raw	psMINER	Raw	psMINER	Raw	psMINER
Active site						
SP	0.62	0.72	1.95	2.38	3.73	4.65
WPE	0.57	0.67	1.91	2.32	3.76	4.59
JSD	0.68	0.76	2.24	2.48	4.35	4.76
R4S	0.53	0.83	2.10	2.70	4.37	5.05
Avg Δ , SD	0.15	0.10	0.42	0.15	0.71	0.22
Ligand-binding site						
SP	0.66	0.71	2.01	2.24	3.82	4.30
WPE	0.52	0.62	1.79	2.09	3.55	4.08
JSD	0.60	0.66	1.97	2.17	3.81	4.17
R4S	0.63	0.70	2.01	2.24	4.68	4.79
Avg Δ , SD	0.07	0.02	0.24	0.04	0.38	0.20
Catalytic residue						
SP	1.23	1.21	3.55	3.33	6.37	5.86
WPE	0.80	0.88	2.64	2.83	5.11	5.29
JSD	1.15	1.10	3.34	3.19	6.08	5.71
R4S	0.99	1.24	3.53	3.63	76.88	6.31
Avg Δ , SD	0.07	0.14	-0.02	0.20	-0.32	0.34

The values in each cell represent the AUCs for raw conservation score (left) and the psMINER algorithm (right). The average difference (Avg Δ) and SD between the psMINER and raw conservation score AUC is also presented. Boldface indicates the best method within each subgroup, which is defined by functional site type at each AUC_X .

cutoffs (0.05, 0.10, and 0.15) are provided in Table 1. At a given false positive rate, psMINER returns more true positives than the raw conservation score. The most exciting aspect of these results is that psMINER improves prediction accuracy regardless of the underlying conservation score, indicating that incorporation of PM information universally improves predictive power. Analogous plots for each of the conservation scores are provided in Supplementary Figures 1–4.

Table 2 lists the results of McNemar's test applied to each psMINER permutation. When applied to the active site benchmark, the difference within the number of true positives output by psMINER versus the raw conservation score is always statistically significant at false positive rates of 0.05 and 0.10. At false positive rates of 0.15, the difference is statistically significant when applied to the SP and WPE scores, and nearly significant when applied to JSD. However, the difference is clearly not significant when applied to R4S in spite of the large improvement within $AUC_{0.10}$. Again, this apparently contradictory result is due to the fundamentally different meanings of AUCs and McNemar's test. While, the difference within the area under the ROC curve up to a false positive rate of 0.15 is substantial for R4S (in fact, $AUC_{0.15}^{R4S}$ is 50% larger than $AUC_{0.15}^{JSD}$), the difference within the number of true positives at that *specific* false positive rate is not statistically significant. When considering the ligand-binding site benchmark, the improvement is mostly significant for the SP and WPE scores, but not JSD and R4S.

The psMINER algorithm generally fails to improve predictive power against the catalytic site benchmark. By removing alignment positions, as dictated by the PM-based filter, the number of true positives is reduced. The rationale behind psMINER is

Table 2. Statistical significance of psMINER improvements

	0.05	0.10	0.15
Active site			
SP	9.92E-4	2.81E-9	4.85E-10
WPE	2.48E-4	1.55E-8	1.68E-7
JSD	1.40E-3	5.26E-3	5.71E-2
R4S	3.57E-11	3.69E-2	1.00E0
Ligand-binding site			
SP	2.10E-1	5.30E-3	6.50E-3
WPE	3.49E-3	1.16E-2	2.51E-3
JSD	7.64E-2	5.67E-2	6.08E-2
R4S	9.02E-07	4.23E-1	NA
Catalytic residue			
SP	NA	NA	NA
WPE	6.44E-1	8.30E-1	8.96E-1
JSD	NA	NA	NA
R4S	2.30E-1	NA	NA

Results of McNemar’s test (P -values) comparing the difference within the number of true positives from psMINER versus the corresponding raw conservation score at the given false positive rates. Boldface indicates differences that are statistically significant (P -value < 0.05).

that it will concurrently remove a greater number of false positives, thus improving performance. However, the ratio of true to false positives removed is unfavorably skewed within the catalytic residue benchmark. Consequently, the relative number of true positives removed from consideration is disproportionately increased. Interestingly, the trends across the catalytic residue benchmark are not uniform. For example, psMINER reduces JSD and SP-score predictive power at all false positive rate cutoffs; however, it slightly improves WPE over the same cutoffs.

3.3 Elucidating relative importance via hMINER

Before comparing hMINER to the underlying conservation scores, we must first determine which PSZ scoring alternative, Z_i^{cent} versus Z_i^{max} , is best suited to the hMINER approach. While the results are similar, Z_i^{cent} universally outperforms Z_i^{max} (data not shown), which is consistent with the results reported by (Manning *et al.*, 2008). As such, it is used throughout the remainder of this report. The next task is to determine which value of α is the best. To do this, we simply linearly adjust α from 0.0 to 1.0, which corresponds to 100% Z_i^{cent} to 100% conservation. The parameter α is a free parameter to be optimized. Consequently, we perform a 5-fold cross-validation analysis to ensure that we do not overfit α . Figure 3 plots an exemplar case where the average $\text{AUC}_{0.10}$ values for each value of α on the five 80% cross-sections of the dataset (the error bars = 1 SD). It should be noted that this plot effectively represents a free parameter optimized within a test set that is composed of the complete validation set. Results showing the optimized value of α applied to the 20% cross-section held back are reported in Supplementary Tables 7–9.

After optimizing α , the optimized values are applied to the complete dataset. Table 3 provides the coefficient values at the maximal $\text{AUC}_{0.10}$ values and the results of McNemar’s test. Interestingly, there is little *overall* conservation within α^{max} (the value of α at the maximal $\text{AUC}_{0.10}$ value) across the three benchmarks. Nevertheless, some general trends do emerge. For

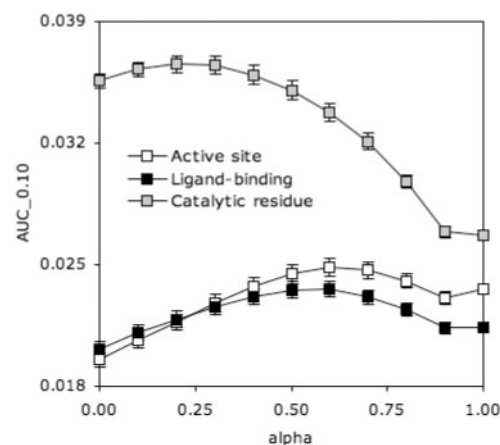


Fig. 3. Average hMINER $\text{AUC}_{0.10}$ values from the five dataset cross-sections. Error bars represent one SD. The left-hand side of the figure coincides to the SP score only, whereas the right-hand side coincides with Z_i^{cent} only. Similar figures are observed for the other conservation scores.

Table 3. Relative importance of phylogeny versus alignment conservation within the hMINER results ($\text{AUC}_{0.10}$ values are $\times 10^{-2}$)

	α	Raw $\text{AUC}_{0.10}$	hMINER $\text{AUC}_{0.10}$	P -value
Active site				
SP	0.6	1.90	2.48	5.12E-7
WPE	0.7	2.25	2.39	4.89E-8
JSD	0.5	2.23	2.51	3.00E-2
R4S	0.2	2.09	2.79	3.00E-2
Ligand-binding site				
SP	0.6	2.01	2.36	2.16E-5
WPE	0.6	1.97	2.17	3.90E-4
JSD	0.5	1.96	2.21	1.10E-2
R4S	0.1	2.28	2.90	1.30E-1
Catalytic residue				
SP	0.2	3.55	3.65	7.48E-1
WPE	0.5	3.38	2.86	NA
JSD	0.1	3.33	3.37	7.2E-1
R4S	0.1	3.53	4.42	4.14E-1

The values in the second column represent the average value of α over the five cross-sections of the dataset at maximal hMINER $\text{AUC}_{0.10}$, which are then used to compute the values in the third column. The values in the last column represent the result of McNemar’s test at false positive rates of 0.10. Boldface indicates hMINER improvement is statistically significant (P -value < 0.05) over the raw conservation score. Complete cross-validation results are provided in Supplementary Tables 7–9.

example, the ideal relative weights are strongly biased toward the raw conservation scores within the catalytic residue benchmark. Nevertheless, inclusion of PM information, sometimes as little as 10% weighting, can appreciably improve predictive power in most cases.

As one might expect based on the psMINER results, there is a general decrease in catalytic residue prediction accuracy as PSZ scores are up-weighted, whereas inclusion of PM information improves prediction accuracy against the other two benchmarks.

Within the latter two classes, the ideal relative weights are much closer to even weighting. These trends are mostly conserved across three of the four conservation scores; R4S is the clear exception. The ideal coefficients when using R4S within hMINER are strongly biased towards conservation. This is true across all three benchmarks. Nevertheless, slight inclusion of PM information does improve predictive power. This result is likely explained by the fact that, as we have already ascertained, R4S is generally more powerful than the other conservation scores. Moreover, unlike the other methods, R4S already includes phylogenetic information within its algorithm, which may further reduce its ability to be improved by PMs.

To summarize, α^{max} is generally conserved between 0.5 and 0.7 across three of the conservation scores (SP, WPE and JSD) when applied to the active and ligand-binding site benchmarks. The value of α^{max} is shifted down when R4S is used, which may be due to the fact that it already includes evolutionary information. When considering the catalytic residue benchmark, α^{max} is consistently downshifted ($\alpha^{max} \sim 0.1-0.2$) due to the extreme evolutionary constraints placed on catalytic residues. The sole exception to this latter trend is WPE, where $\alpha^{max} = 0.5$. However, it should be noted that the WPE $AUC_{0.10}$ versus α curve is very flat and that the $AUC_{0.10}$ value at $\alpha^{max} = 0.2$ is almost the same as at 0.5. (Details of how psMINER and hMINER perform over the full ROC curve are provided in Supplementary Tables 5 and 10.)

3.4 Comparison to related methods

As stated above, there are several related functional site prediction methods that use phylogenetic information, including ET, the MB-method and SMERFS. (We do not consider R4S a phylogeny-based method because, while it does use phylogenetic concepts to correct for sampling bias, it is fundamentally a conservation score.) As a benchmark to assess how well our approach is doing, we apply each of these methods to our dataset. We utilize web-server implementations of ET (Innis *et al.*, 2000) and SMERFS (Manning *et al.*, 2008). In the ET predictions, the tree is partitioned at 20 cut-levels, and each alignment position is scored based on the deepest tree partition (0–20) that identifies that position as a trace residue. As dictated by Manning *et al.* (2008), we apply SMERFS with the following parameters: window size = 9, scoring scheme = max and gap threshold = 1.00. We also apply Xdet (del Sol *et al.*, 2003; Pazos *et al.*, 2006), which is a stand-alone implementation of the MB-method, with default parameters.

The results from each of these three methods are compared to psMINER and hMINER in Supplementary Figure 5 (the AUC values are provided in Supplementary Table 11). In all cases, hMINER results in the largest AUC values. Moreover, in all remaining cases but one, psMINER has the second largest AUC value. Statistical significance between the differences between psMINER and each of the additional methods are provided in Supplementary Table 12. In all cases except for the MB-method applied to the catalytic residue benchmark, the improvement by psMINER versus the other methods is always statistically significant. As a general rule, ET and SMERFS perform the worst, whereas the MB-method is between these two and the two MINER variants. While, Supplementary Figure 5 and Supplementary Table 11 only provide MINER-variant results based on the SP score, similar results are observed for the other three conservation scores (as demonstrated elsewhere in this report).

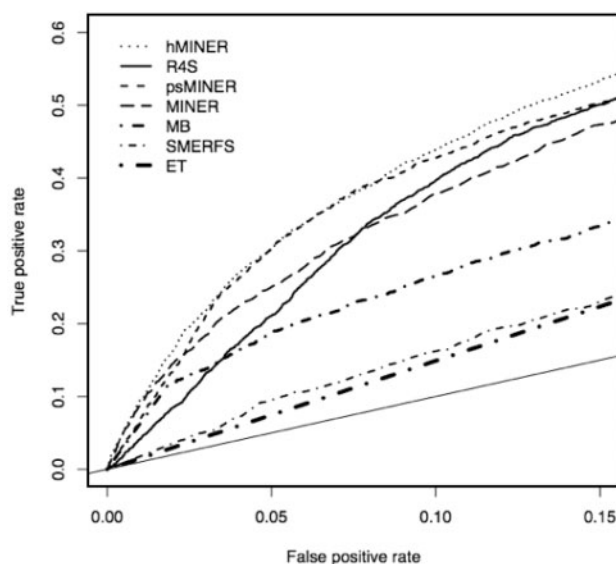


Fig. 4. ROC curves of both MINER variants, the two constituent metrics that make them up, and the three other phylogeny-based methods. Similar relationships to MINER (Z_i^{cent}) are observed for the other considered conservation scores. Note: to facilitate analysis, the ordering of the legend is approximately consistent with the rank ordering at false positive rates of 0.15. The solid gray line at the bottom is the diagonal of the full ROC curve, which indicates a random predictive method.

3.5 Dissecting the performance improvement

While all five of the phylogeny-based methods do *implicitly* provide some sort of conservation information (e.g. a trace residue occurring at a deep ET-cut must be highly conserved), of the methods considered here, only psMINER and hMINER *explicitly* utilize conservation scores. As such, this leads to the question of whether or not the performance of the MINER variants is simply a consequence of their explicit usage of conservation scores. In Figure 4, we plot the ROC curves of both MINER variants, an exemplar conservation score, PMs (the Z_i^{cent} score), and the three other phylogeny-based methods. From this figure, it is clear that PMs, in themselves, contain much higher predictive power than the other three other phylogeny-based methods. The exact relative performance between the conservation scores and Z_i^{cent} varies—at times the raw conservation scores perform better (as shown in Fig. 4), whereas at other times Z_i^{cent} is a better predictor (e.g. Z_i^{cent} does better than the SP score). Nevertheless, overall the conservation scores and Z_i^{cent} values are in the same ballpark relative to the other three methods. Consequently, it is straightforward to conclude that the improved predictive power of the MINER variants is due to both underlying constituent metrics. Future work will use machine learning techniques to identify optimized ways of combining the two scoring schemes.

4 DISCUSSION

4.1 Prediction accuracies

By themselves, the predictive power of the raw conservation scores is quite impressive. In fact, they dramatically outperform three common methods that utilize phylogeny information. Nevertheless,

the conservation scores still result in an undesirable number of false positives. As discussed above, this is due to the various, sometimes competing, evolutionary constraints acting upon enzymes. As a consequence, the choice of functional site benchmark drastically affects the observed accuracies of the conservation score methods. Across our dataset, the conservation scores do best at predicting catalytic residues, followed by active site and ligand-binding sites, respectively. Nevertheless, the difference between the latter two is small. There is little consensus when comparing the four conservation scores at $X=0.05$. However, our results indicate that R4S is generally best at higher false positive rates. However, it should again be stressed that R4S is substantially more compute intensive than the other methods. Based on computational ease, the predictive power of JSD and SP is quite impressive.

Whether using the psMINER or the hMINER approach, marrying conservation scores and phylogenetic descriptions improves predictive power against the active site and ligand-binding site benchmarks. The improvement is most stark when considering active sites, which is consistent with our previous results indicating that PMs do a good job of identifying them (La and Livesay, 2005; La *et al.*, 2005; Livesay and La, 2005). The improvement (relative to the underlying conservation scores) when applied to the ligand-binding sites is mostly statistically significant, but is less than the improvement observed within the active sites. Nevertheless, the observed improvements against both benchmarks strongly indicate that the method is useful and should be considered further. Moreover, the ubiquity of the improvement over all four conservation scores suggests the method is general and is expected to be applicable to any conservation score. The only caveat here is that the hMINER results indicate the appropriate balance between PMs and conservation within R4S is much more biased towards the underlying conservation score. This result is likely due to the innate predictive power of R4S, which may occur because phylogenetic information is already incorporated into its algorithm.

While the two introduced MINER variants generally outperform their underlying conservation scores against these two benchmarks, their power is clearly revealed when comparing to the three other functional site prediction methods that utilize phylogenetic information. When considering the active site and ligand-binding site benchmarks, the two MINER variants substantially outperform the other three methods. In fact, none of the other methods are even able to approach the predictive power of the SP score, which conceptually is the simplest of the considered conservation scores. This result indicates that incorporation of phylogenetic information fails to always improve accuracy. Nevertheless, both psMINER and hMINER do generally improve over their constituent conservation scores. Meaning that from a prediction point of view, only psMINER and hMINER are worth the added complexity of including phylogenetic information. While the improvement within psMINER and hMINER is moderate (yet generally statistically significant), these methods represent a promising class of phylogeny + conservation algorithms that are able to outperform conservation scores. Future work will seek test additional variants that incorporate PM information in the same vein in order to further improve the approach.

Conversely, incorporating phylogenetic information into the conservation scores fails to uniformly improve predictive power. This result occurs because of the tight constraints that evolution puts on catalytic residues; catalytic residues are among the most

conserved groups within the enzyme. While some evolutionary plasticity is allowed at active site and ligand-binding residues, mutation at catalytic residues will nearly always result in loss of function. Consequently, incorporating phylogenetic information, which is based on *diversity*, fails to improve the ability of the underlying conservation scores to predict catalytic residues. This dichotomy between the active site + ligand-binding site benchmarks and the catalytic residue benchmark is, in fact, quite interesting and noteworthy because it reveals much about the different natures of the functional sites considered herein.

It is noteworthy that the performance of the MB-method (relative to the other methods) is much improved against the catalytic residue benchmark. This is especially true at very low false positive rates (<0.05). Nevertheless, at false positive rates ≥ 0.10 , both psMINER and hMINER outperform the other three methods. This result is somewhat surprising because psMINER and hMINER perform poorest relative to the underlying conservation scores, when applied to the catalytic residue benchmark. Nevertheless, the mostly statistically significant improvements by psMINER (and thus hMINER) over the three methods reinforce the notion that the PMs represent a better method to include phylogenetic information into functional site prediction algorithms.

4.2 The importance of comprehensive benchmarks

Recently, there has been an increasing interest in computational prediction of protein functional sites using both sequence- and/or structure-based methods. However, the benchmarks that the prediction community uses to assess these methods have not kept up with prediction methodological improvements. In fact, most studies continue to assess prediction accuracy using a single functional site definition, generally either based on catalytic importance (from the CSA) or ligand binding. As a consequence, we have constructed benchmarks based on active sites, ligand-binding sites and catalytic residues. Using such a comprehensive benchmarking strategy allows us to more completely assess our algorithms. For example, had we limited ourselves to the most common approach of prediction assessment (catalytic residues), then we would have determined our approach to be unsuccessful. However, by also benchmarking against active sites and ligand-binding sites, the merit of our methods is clearly revealed. In future work, we hope to expand our assessment protocols to include an even broader array of functional definitions, including allosteric sites, structural sites and trafficking signals.

ACKNOWLEDGEMENTS

We thank Dr Mona Singh for providing her research group's software to calculate the JSD and WPE scores, Dr Nir Ben-Tal for providing R4S, Dr William S. J. Valdar for providing Scorecons, Dr Florencio Pazos for providing Xdet and Dr Jonathan R. Manning and Dr Geoff J. Barton for assisting us with SMERFS. We thank Dr Hiroto Saigo for helpful discussions and the anonymous reviewers for a number of valuable suggestions.

Funding: Bioinformatics Research Center at UNC-Charlotte (to D.R.L.).

Conflict of Interest: none declared.

REFERENCES

- Alm,E. *et al.* (2002) Simple physical models connect theory and experiment in protein folding kinetics. *J. Mol. Biol.*, **322**, 463–476.
- Aloy,P. *et al.* (2001) Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. *J. Mol. Biol.*, **311**, 395–408.
- Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
- Amitai,G. *et al.* (2004) Network analysis of protein structures identifies functional residues. *J. Mol. Biol.*, **344**, 1135–1146.
- Blom,N. *et al.* (1999) Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *J. Mol. Biol.*, **294**, 1351–1362.
- Boeckmann,B. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
- Bradford,J.R. and Westhead,D.R. (2005) Improved prediction of protein-protein binding sites using a support vector machines approach. *Bioinformatics*, **21**, 1487–1494.
- Cai,Y.D. *et al.* (2004) Identify catalytic triads of serine hydrolases by support vector machines. *J. Theor. Biol.*, **228**, 551–557.
- Capra,J.A. and Singh,M. (2007) Predicting functionally important residues from sequence conservation. *Bioinformatics*, **23**, 1875–1882.
- Chea,E. and Livesay,D.R. (2007) How accurate and statistically robust are catalytic site predictions based on closeness centrality? *BMC Bioinformatics*, **8**, 153.
- Chelliah,V. *et al.* (2004) Distinguishing structural and functional restraints in evolution in order to identify interaction sites. *J. Mol. Biol.*, **342**, 1487–1504.
- Cheng,G. *et al.* (2005) Improvement in protein functional site prediction by distinguishing structural and functional constraints on protein family evolution using computational design. *Nucleic Acids Res.*, **33**, 5861–5867.
- del Sol,A. *et al.* (2003) Automatic methods for predicting functionally important residues. *J. Mol. Biol.*, **326**, 1289–1302.
- Dessailly,B.H. *et al.* (2007) Relating destabilizing regions to known functional sites in proteins. *BMC Bioinformatics*, **8**, 141.
- Elcock,A.H. (2001) Prediction of functionally important residues based solely on the computed energetics of protein structure. *J. Mol. Biol.*, **312**, 885–896.
- 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.
- Innis,C.A. *et al.* (2000) Evolutionary trace analysis of TGF- β and related growth factors: implications for site-directed mutagenesis. *Prot. Eng.*, **13**, 839–847.
- Kuhlman,B. and Baker,D. (2000) Native protein sequences are close to optimal for their structures. *Proc. Natl Acad. Sci. USA*, **97**, 10383–10388.
- La,D. and Livesay,D.R. (2005) MINER: software for phylogenetic motif identification. *Nucleic Acids Res.*, **33**, W267–W270.
- La,D. and Livesay,D.R. (2005) Predicting functional sites with an automated algorithm suitable for heterogeneous datasets. *BMC Bioinformatics*, **6**, 116.
- La,D. *et al.* (2005) Predicting protein functional sites with phylogenetic motifs. *Proteins*, **58**, 309–320.
- Lichtarge,O. *et al.* (1996) An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.*, **257**, 342–358.
- Livesay,D.R. and La,D. (2005) The evolutionary origins and catalytic importance of conserved electrostatic networks within TIM-barrel proteins. *Protein Sci.*, **14**, 1158–1170.
- Livesay,D.R. *et al.* (2007) Assessing the ability of sequence-based methods to provide functional insight within membrane integral proteins: a case study analyzing the neurotransmitter/Na⁺ symporter family. *BMC Bioinformatics*, **8**, 397.
- Madabushi,S. *et al.* (2002) Structural clusters of evolutionary trace residues are statistically significant and common in proteins. *J. Mol. Biol.*, **316**, 139–154.
- Madabushi,S. *et al.* (2004) Evolutionary trace of G protein-coupled receptors reveals clusters of residues that determine global and class-specific functions. *J. Biol. Chem.*, **279**, 8126–8132.
- Manning,J.R. *et al.* (2008) The contrasting properties of conservation and correlated phylogeny in protein functional residue prediction. *BMC Bioinformatics*, **9**, 51.
- Mayrose,I. *et al.* (2004) Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior. *Mol. Biol. Evol.*, **21**, 1781–1791.
- McNemar,Q. (1947) Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, **12**, 153–157.
- Mihalek,I. *et al.* (2004) A family of evolution-entropy hybrid methods for ranking protein residues by importance. *J. Mol. Biol.*, **336**, 1265–1282.
- McDonald,I.K. and Thornton,J.M. (1994) Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol.*, **238**, 777–793.
- Ondrechen,M.J. *et al.* (2001) THEMATICS: a simple computational predictor of enzyme function from structure. *Proc. Natl Acad. Sci. USA*, **98**, 12473–12478.
- Pande,S. *et al.* (2007) Prediction of enzyme catalytic sites from sequence using neural networks. In *Proceedings of the IEEE Symposium on Computational Intelligence and Bioinformatics and Computational Biology*, IEEE Press, Honolulu, USA, pp. 247–253.
- Pazos,F. and Bang,J. (2006) Computational prediction of functionally important regions in proteins. *Curr. Bioinformatics*, **1**, 15–23.
- Pazos,F. *et al.* (2006) Phylogeny-independent detection of functional residues. *Bioinformatics*, **22**, 1440–1448.
- Pei,J. *et al.* (2003) Using protein design for homology detection and active site searches. *Proc. Natl Acad. Sci. USA*, **100**, 11361–11366.
- Petrova,N.V. and Wu,C.H. (2006) Prediction of catalytic residues using Support vector machine with selected protein sequence and structural properties. *BMC Bioinformatics*, **7**, 312.
- Porter,C.T. *et al.* (2004) The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res.*, **32**, D129–D133.
- Pritchard,L. and Dufton,M.J. (1999) Evolutionary trace analysis of the Kunitz/BPTI family of proteins: functional divergence may have been based on conformational adjustment. *J. Mol. Biol.*, **285**, 1589–1607.
- Pupko,T. *et al.* (2002) Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics*, **18**, S71–S77.
- Roshan,U. *et al.* (2005) Improved phylogenetic motif detection using parsimony. In *Proceedings of the Fifth IEEE Symposium on Bioinformatics and Bioengineering*, IEEE Press, Minneapolis, USA, pp. 19–26.
- Sowa,M.E. *et al.* (2001) Prediction and confirmation of a site critical for effector regulation of RGS domain activity. *Nat. Struct. Biol.*, **8**, 234–237.
- Thibert,B. *et al.* (2005) Improved prediction of critical residues for protein function based on network and phylogenetic analyses. *BMC Bioinformatics*, **6**, 213.
- Valdar,W.S. (2002) Scoring residue conservation. *Proteins*, **48**, 227–241.
- Vergara,I.A. *et al.* (2008) StAR: a simple tool for the statistical comparison of ROC curves. *BMC Bioinformatics*, **9**, 265.
- Watson,J.D. *et al.* (2005) Predicting protein function from sequence and structural data. *Curr. Opin. Struct. Biol.*, **15**, 275–284.
- Yao,H. *et al.* (2003) An accurate, sensitive, and scalable method to identify functional sites in protein structures. *J. Mol. Biol.*, **326**, 255–261.