

Topology Improves Phylogenetic Motif Functional Site Predictions

Dukka B. KC and Dennis R. Livesay

Abstract—Prediction of protein functional sites from sequence-derived data remains an open bioinformatics problem. We have developed a phylogenetic motif (PM) functional site prediction approach that identifies functional sites from alignment fragments that parallel the evolutionary patterns of the family. In our approach, PMs are identified by comparing tree topologies of each alignment fragment to that of the complete phylogeny. Herein, we bypass the phylogenetic reconstruction step and identify PMs directly from distance matrix comparisons. In order to optimize the new algorithm, we consider three different distance matrices and 13 different matrix similarity scores. We assess the performance of the various approaches on a structurally nonredundant data set that includes three types of functional site definitions. Without exception, the predictive power of the original approach outperforms the distance matrix variants. While the distance matrix methods fail to improve upon the original approach, our results are important because they clearly demonstrate that the improved predictive power is based on the topological comparisons. Meaning that phylogenetic trees are a straightforward, yet powerful way to improve functional site prediction accuracy. While complementary studies have shown that topology improves predictions of protein-protein interactions, this report represents the first demonstration that trees improve functional site predictions as well.

Index Terms—Phylogenetic motif, functional site prediction, phylogenetic tree, distance matrix.

1 INTRODUCTION

PREDICTING functionally relevant information from a newly discovered protein is one of the most challenging jobs in this postgenomic era. Currently, there are two major paradigms within this realm. The first is related to classification of a protein into its broad functional class (e.g., gene ontology [1] or enzyme classification number [2]). While this information is clearly important, it provides little mechanistic insight. As such, the second paradigm uses computational strategies to predict *functional sites* that define the function and/or regulation of the protein. Not only do functional site descriptions allow for improved descriptions of *how* a given protein functions, they can also be used to identify functionally deleterious mutations. Consequently, there has been a steady output of new computational approaches for protein functional site prediction. Current approaches can be classified as either sequence-based (actually, alignment), structure-based, or a combination thereof [3], [4].

The most common methods are based on conservation analysis across a multiple sequence alignment, which has been used to detect functional sites in myriad settings [5], [6], [7]. Several more sophisticated approaches that attempt to detect some additional sort of “feature” conservation have also been developed [8]. Most of these approaches attempt to identify positions whose variability is dictated by the functional evolution of the family [9], [10]. For example, evolutionary trace (ET) [9] and its variants [11], [12], [13] attempt to identify *trace residues*, which are individual

alignment positions that are conserved within functionally distinct subfamilies. Once the trace residues have been identified, the most common usage of ET is to map these positions to structure, and structural clusters of ET+ conserved mutations are put forth as functional site predictions.

Previously, we have demonstrated that alignment fragments taken from a priori identified functional sites tend to conserve the overall familial phylogeny [14]. Subsequently, we reversed this scenario in order to predict protein functional sites by screening all possible alignment fragments for this phylogenetic feature [10]. Our method uses a sliding alignment window algorithm to scan all possible fragments. A tree is generated for each fragment, which is compared to the overall familial tree using a partition metric algorithm. Alignment fragments that most closely reflect the overall phylogeny are put forth as functional site predictions. The resultant phylogenetic motifs (PMs) are very likely to correspond to protein functional sites [10], [15], [16], [17], [18]. The predictive power of the PM approach was most recently demonstrated in our report that used PM information to make statistically significant improvements in prediction accuracy over raw conservation score [19]. Our PM identification algorithm has been implemented in an easy to use Web server called MINER [20]. (Note that the MINER has been moved to <http://coit-apple01.uncc.edu/MINER>.)

While the predictive value of the MINER approach has been shown repeatedly, one potential source of criticism results from building trees on such small alignment fragments. This concern is partially mitigated by our results that show (via bootstrap analysis) PM tree stability is acceptable [10]. Nevertheless, if methods based on comparison of the underlying distance matrices (versus trees) can be shown to perform equally well, this would represent an attractive alternative. Recently, Manning et al. [21] introduced the SMERFS algorithm that uses distance matrices as a source of evolutionary information. In a manner analogous to MINER,

• The authors are with the Department of Bioinformatics and Genomics, University of North Carolina at Charlotte, 9201 University City Blvd., Charlotte, NC 28223. E-mail: {dbahadur, drlivesa}@uncc.edu.

Manuscript received 12 Mar. 2009; revised 9 June 2009; accepted 16 June 2009; published online 9 July 2009.

For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBB-2009-03-0034. Digital Object Identifier no. 10.1109/TCBB.2009.60.

SMERFS compares distance matrices constructed from alignment fragments and the global distance matrix, which is constructed from the whole alignment. Pairwise matrix similarity is calculated using Pearson's correlation coefficient. In a head-to-head comparison, SMERFS was shown to outperform MINER in the prediction of domain-domain interactions; however, MINER did a better job of predicting small molecule binding sites. The latter observation is entirely consistent with our results demonstrating that MINER does a much better job of predicting localized functional sites (catalytic sites, active sites, and ligand-binding sites).

In this report, we test an exhaustive array of distance matrix methods and compare their predictive power to the MINER. The considered distance matrix variants are derived from the MINER algorithm; however, phylogenetic similarity is determined by a direct matrix-to-matrix comparison that bypasses topology. Toward that end, we consider three different distance matrices and evaluate 13 matrix similarity metrics that span several different strategies, including distance-based methods (e.g., euclidean distance, etc.), correlation-based metrics (e.g., Pearson's correlation coefficient, etc.), and information-theoretic methods (e.g., mutual information, etc.). None of the considered distance matrix permutations comes close to approximating the predictive power of the original MINER algorithm. Even though the considered distance matrix methods fail to improve upon the original approach, they do provide key insight into the predictive power of MINER. Based on the similarity of the implementation details between MINER and the distance matrix methods considered here, we conclude that the additional evolutionary information provided by tree topologies is the basis of the improved predictive power. While not completely unexpected, these results conclusively demonstrate for the first time that the added evolutionary descriptions within phylogenetic trees can be easily translated into improved predictions of protein functional sites.

2 PREDICTING PMS WITH MINER

MINER begins by windowing over a "masked" input alignment. Masking indicates that highly gapped positions (> 50 percent gaps) are ignored. From the generated set of $(N_{mask} - W + 1)$ alignment fragments (where N_{mask} is the length of the masked alignment and W is the window width), windows whose evolutionary descriptions match that of the overall alignment are predicted to be functional. To evaluate evolutionary similarity, MINER compares the topology of a tree reconstructed for each alignment window to that of the overall phylogeny using a modified bipartition metric [18] that counts the topological differences. These raw differences are subsequently converted into phylogenetic similarity z-scores (PSZs), where $PSZs < 0$ indicate similarities better than the mean. Subsequently, the PSZs are used to rank the various windows based on their putative prediction accuracy (meaning that lower PSZs are expected to be better predictions than larger values).

3 PREDICTING PMS WITHOUT TREES

For the first time, we compare the predictive power of the original MINER algorithm to analogous methods that bypass phylogenetic reconstruction. Instead, the methods presented

here compare *distance matrices* (not trees) generated from alignment fragments to that of the whole alignment. Meaning that the considered variants correspond to the SMERFS method. However, to optimize their performance, we consider many different algorithmic options. In all cases, the generic algorithm used by the distance matrix methods is exactly same as the original approach except that, of course, the similarity score is obtained by comparing the distance matrices instead of phylogenetic trees. The key steps of the basic algorithm are:

1. Given a masked multiple sequence alignment, an overall distance matrix is created using one of three different distance matrix methods (as discussed below).
2. For each window distance matrix, the corresponding distance matrix is calculated using the same approach as in step 1.
3. Using one of the matrix comparison measures (also discussed below), the similarity between the overall distance matrix and each window distance matrix is calculated.
4. The raw matrix similarity scores are converted to z-scores.

While our previous results have found a window width of five to be ideal for MINER [10], it is not clear what the ideal window width for the distance matrix methods should be. As such, we consider a series of window sizes ranging in width from three to nine for each variant.

3.1 Generation of Distance Matrices

We consider three different programs for calculation of the underlying distance matrices: ProtDist (PD) [22], ClustalDist (CD) [23], and TREE-PUZZLE (TP) [24]. All three construct a symmetrical $N \times N$ matrix based on $N(N-1)/2$ pairwise sequence comparisons; hence, in each case, we only utilize one-half of the matrix for our comparisons. PD is a module in the PHYLIP phylogenetic reconstruction suite of programs developed at Washington University. PD computes a distance measure for protein sequences using a variety of maximum likelihood estimates. Herein, we use the likelihood table by Jones et al. [25], which is the default setting. Similarly, CD and TP are also used to calculate their respective distance matrices using default parameters.

3.2 Distance-Based Matrix Comparisons

In order to verify the predictive power of distance matrix comparisons, we consider a variety of methods to compare the window distance matrices to the distance matrix constructed from the whole alignment. Each of the 13 different methods is described below and grouped into one of four categories:

1. distance-based methods,
2. correlation-based metrics,
3. information-theoretic methods, and
4. the Tanimoto coefficient.

In this section, we begin by describing the distance-based approaches.

Let $P = (p_1, p_2, \dots, p_n)$ and $Q = (q_1, q_2, \dots, q_n)$ be the list of matrix elements arranged with the same corresponding order of the elements taken from window distance matrix P

and overall distance matrix Q . The *euclidean distance* between the two distance matrix P and Q is defined by

$$Dist_{PQ}^{Eucl} = \left[\sum_{i=1}^n (p_i - q_i)^2 \right]^{1/2}. \quad (1)$$

The Marcotte distance is very similar to the euclidean distance, but it includes an overall (meaning outside the sum) normalization factor that is based on the size of the matrix:

$$Dist_{PQ}^{Mar} = \left\{ 2 \cdot Dist_{PQ}^{Eucl} / [n(n-1)] \right\}^{1/2}. \quad (2)$$

The Delta distance is also similar to the euclidean distance, but it includes a normalization inside the sum that is defined by the sum of the two constituent matrix values:

$$Dist_{PQ}^{Del\&ChiSq} = \sum_{i=1}^n (p_i - q_i)^2 / D_{norm}, \quad (3)$$

where $D_{norm} = (p_i + q_i)$. Juxtaposed to the Delta distance, the chi-square distance is simply normalizing by one of the constituent matrix values ($D_{norm} = q_i$).

Two distance-based metrics based on probability distributions are also considered. Let $X = (x_1, x_2, \dots, x_n)$ and $Y = (y_1, y_2, \dots, y_n)$ be the two corresponding lists of matrix elements after normalizing P and Q (from above) by the total sum of the elements in each list, respectively. Meaning that X and Y can be thought of as discrete probability distributions. From which, the Hellinger distance between the two distance matrices P and Q is defined as

$$Dist_{PQ}^{Hel} = \sum_{i=1}^n [(x_i)^{1/2} - (y_i)^{1/2}]^2. \quad (4)$$

In a similar way, the Bhattacharya distance between two distance matrices P and Q is defined as

$$Dist_{PQ}^{Bha} = \sum_{i=1}^n (x_i \cdot y_i)^{1/2}. \quad (5)$$

3.3 Correlation-Based Comparisons

Three correlation coefficients are also considered. The first two, Pearson's and Spearman's rank correlation coefficients, need not be introduced due to their ubiquity. We also consider the Kendall tau rank correlation, which is given by

$$Corr_{PQ}^{KT} = 4D / [n(n-1)] - 1, \quad (6)$$

where D is the total number of concordant pairs, which are defined as the pairs for which $\text{sign}(p_j - p_i) = \text{sign}(q_j - q_i)$. The denominator can be thought of as the total number of pairs; thus, a high value of D means that most pairs are concordant implying that two rankings are consistent. Note that, as discussed above, SMERFS also compares distance matrices using Pearson's correlation coefficient. As expected, our results indicate that the predictive power of SMERFS is consistent with our Pearson's correlation distance matrix variant (unpublished results).

3.4 Information-Theoretic Comparisons

From the discrete probability distributions X and Y defined above, the mutual information can be calculated from:

$MI_{XY} = H(X) + H(Y) - H(X, Y)$, where $H(R)$ represents the Shannon entropy of R and $H(X, Y)$ is the joint entropy. While mutual information measures the mutual dependence on the two variables, the Kullback-Leibler divergence is used to directly measure the difference between the two probability distributions. Specifically, the Kullback-Leibler divergence is defined as

$$Div_{PQ}^{KL} = \sum_{i=1}^n X_i \log(X_i/Y_i). \quad (7)$$

Similarly, the Jensen-Shannon divergence is also used to directly measure the similarity between the two distributions; in fact, it is based on the Kullback-Leibler divergence:

$$Div_{PQ}^{JSD} = 0.5 Div_{XM}^{KL} + 0.5 Div_{YM}^{KL}, \quad (8)$$

where Div_{XM}^{KL} is the KL divergence between distributions X and M and Div_{YM}^{KL} is the KL divergence between distributions Y and M . M is the distribution whose i th element is given by $m_i = 0.5(x_i + y_i)$.

3.5 The Tanimoto Coefficient

The Tanimoto coefficient, which does not fit into any of the above categories, is a measure of similarity between two vectors P and Q . The Tanimoto coefficient is calculated by

$$T_{PQ} = P \cdot Q / (\|P\|^2 + \|Q\|^2 - P \cdot Q), \quad (9)$$

where $P \cdot Q$ is the dot product of two vectors P and Q and $\|R\|$ is the magnitude of vector R .

4 FUNCTIONAL SITE BENCHMARKS

We assess the performance comparison of each measure on our recently developed functional site benchmark composed of 163 structurally nonredundant enzyme families [19]. This data set, which is based on the Catalytic Site Atlas [26], includes annotations for each entry in terms of three different functional site definitions: 1) catalytic residues; 2) active sites; and 3) ligand-binding sites. The *catalytic residues* are defined by the Catalytic Site Atlas, which is a manual curation of residues directly involved in the reaction coordinate of the enzyme-catalyzed reaction. The *active sites* are defined by the union of the catalytic residues and those residues that are directly interacting with them (as defined by HBPLUS [27]). Finally, the *ligand-binding sites* are defined by all enzyme-substrate interactions (again, as identified by HBPLUS). As discussed in KC and Livesay [19], each definition is treated as independent of the others, which results in a lower bound on the assessed predictive power. All three benchmark data sets can be freely downloaded at <http://cs.uncc.edu/~drlivesa/dataset.html>.

In the results below, we demonstrate that MINER systematically outperforms all of the corresponding distance matrix comparison methods, most in a statistically significant way. However, this is in conflict with results published by Manning et al. [21] where they demonstrated that MINER outperformed SMERFS, but not in a statistically significant way. In fact, as applied to ligand-binding sites, we predict MINER's predictive power to be twice of what Manning et al. reported. While distinct, it is unlikely that this result is simply due to benchmark composition since both benchmarks are of

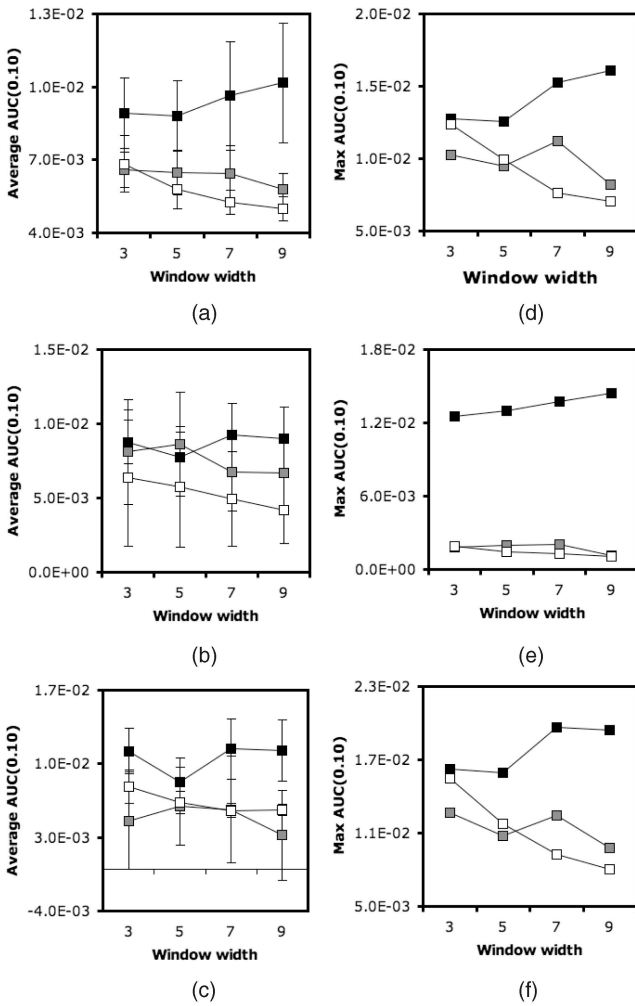


Fig. 1. The average AUC0.10 values (over the 13 different matrix similarity metrics) versus window width are plotted for each distance matrix (black squares = ClustalDist, gray squares = TREE – PUZZLE, and white squares = ProtDist) and functional site benchmark: (a) active site, (b) ligand-binding site, and (c) catalytic residue. The error bars = 1 standard deviation. (d), (e), and (f) The best of the 13 matrix similarity metrics are plotted (using the same coloring scheme and respective functional site benchmark order) in panels. Similar results are observed when considering AUC0.05 and AUC0.15 values.

similar size. (Yet, we point out while our benchmark is structurally nonredundant, whereas this is not the case within the Manning et al. report.) We believe that this dichotomy can be resolved by the fact that we have biased our data set toward families with greater numbers of sequences, which is a simple way to improve MINER performance. Moreover, it appears that Manning et al. simply used the PFAM alignments, which are notoriously poor. As such, we expect that realignment of the Manning et al. benchmark would be a straightforward way to improve MINER’s performance against that data set.

5 PREDICTION ASSESSMENT

The receiver operating characteristic (ROC) is a good statistic to assess predictive power because the area under the curve (AUC) combines descriptions of sensitivity and specificity. However, the full ROC curve is generally not appropriate to assess functional site prediction power

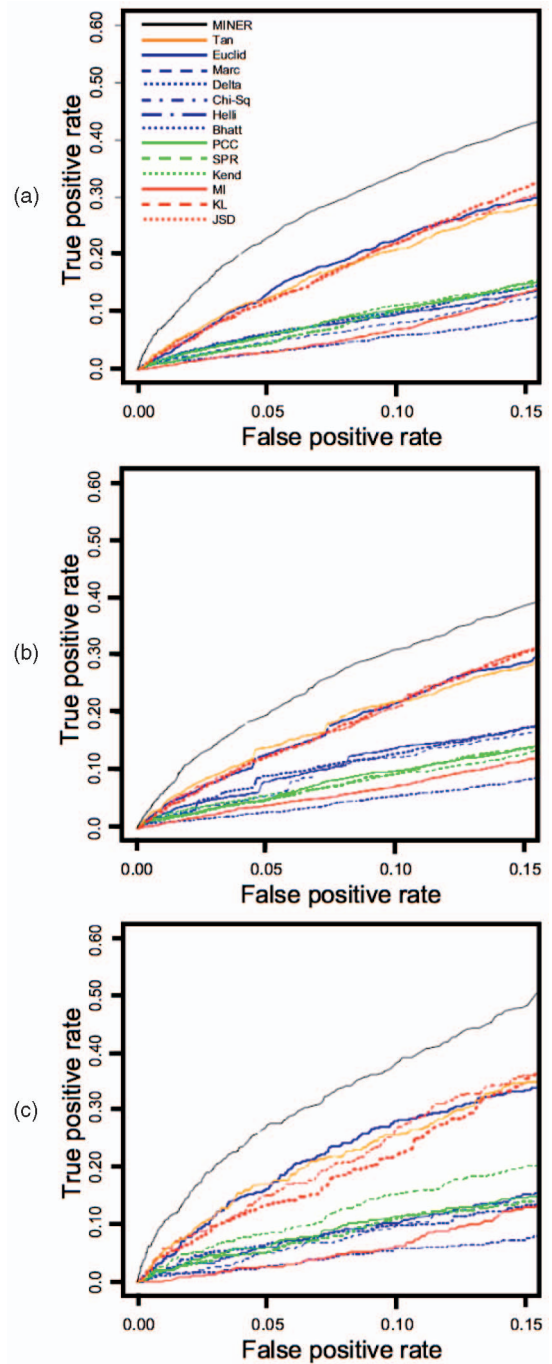


Fig. 2. Receiver operating characteristic curves for MINER and each of the 13 ClustalDist variants (window width = 5) against the (a) active site, (b) ligand-binding site, and (c) catalytic residue benchmarks. The matrix similarity metrics are color-coded based on class. Note that in all cases, MINER substantially outperforms the distance matrix variants. While the performance of the CD variants does improve at window widths of 9, they never approach the predictive power of the original MINER algorithm using a width of 5.

because it is highly biased by false positive rates that are outside the realm of experimental utility. As such, we assess the methods considered herein against all three benchmarks up to false positive rates of 0.05, 0.10, and 0.15 (AUC0.05, AUC0.10, and AUC0.15, respectively). Unfortunately, there is no metric to assess the statistical significance of differences within AUCX < 1.00 values. As an acceptable alternative, we employ the modified McNemar’s test [21] to

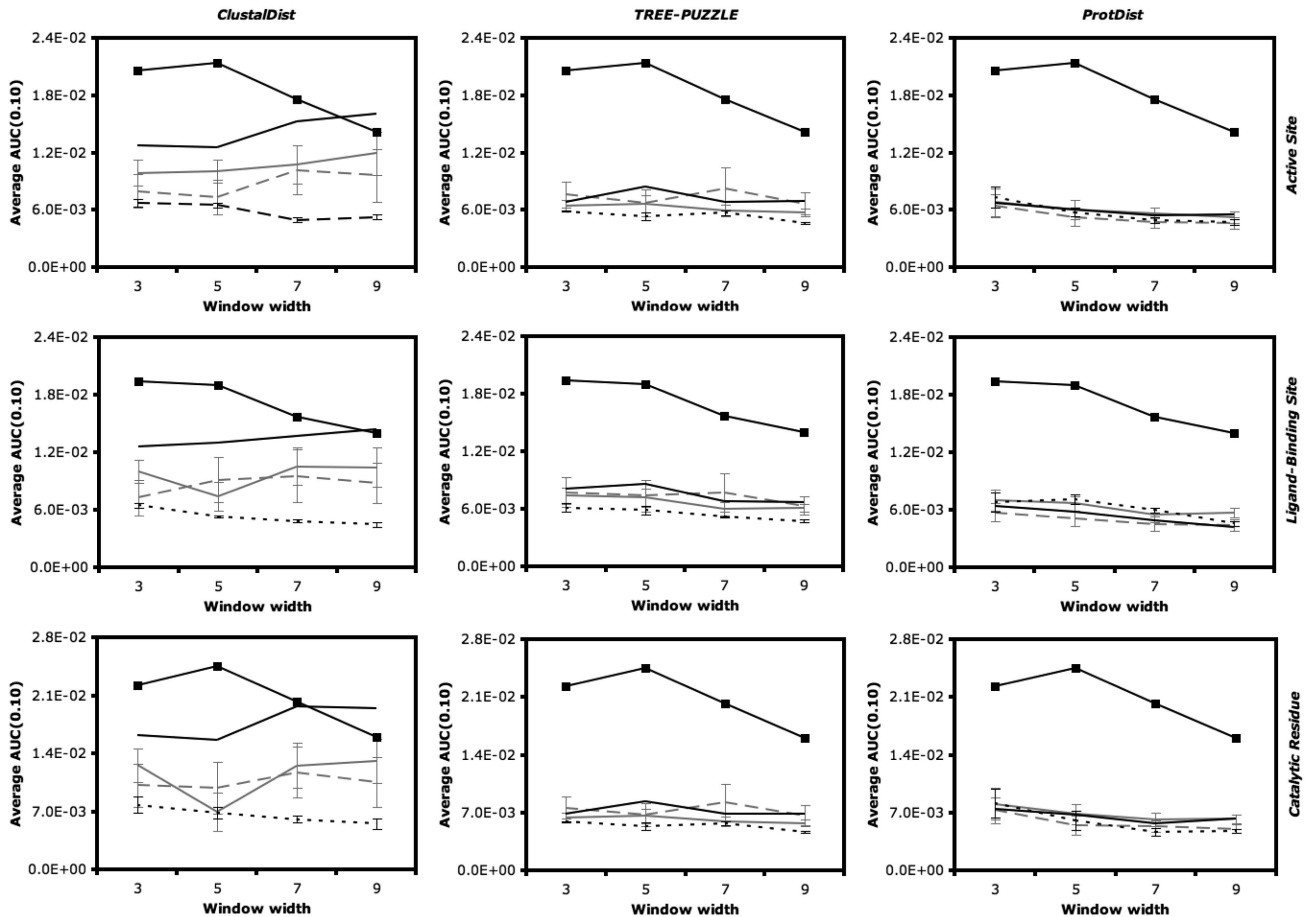


Fig. 3. The average AUC0.10 values are plotted for each matrix similarity metric class (black squares = MINER, solid black line = Tanimoto coefficient, solid gray line = distance – based, dashed gray line = information theory, and dotted black line = correlation coefficient). The error bars correspond to one standard deviation. There are no error bars for MINER and the Tanimoto coefficient because each is only a single metric. Each row corresponds, respectively, to the active site, ligand-binding site, and catalytic residue benchmarks, whereas each column corresponds, respectively, to ClustalDist, TREE-PUZZLE, and ProtDist.

assign statistical significance between TP rates at a given FP rate using a chi-square statistic. Note that while we do not compare AUC1.00 values herein due to space considerations, it should be pointed out that our primary conclusion (namely, the original MINER algorithm significantly outperforms the considered distance matrix methods) is also observed across the full ROC curve.

6 RESULTS AND DISCUSSION

6.1 Distance Matrix Methods

We begin by comparing the performance of the distance matrix comparisons vis-à-vis which underlying distance matrix is considered. Figs. 1a, 1b, and 1c plot the average AUC0.10 over the 13 distance matrix comparison metrics versus window width. The figure clearly indicates that CD generally outperforms the other methods; however, there is considerable variability across the performance within the CD metrics. The error bar in Fig. 1 is equal to one standard deviation. Example ROC curves for each of the 13 metrics using CD are provided in Fig. 2. While the variation across the TP and PD values is much less than CD, the superior predictive power of CD is again demonstrated when considering the best performing metric per scoring matrix; see Figs. 1d, 1e, and 1f. Similar conclusions are drawn when considering AUC0.05 and AUC0.15.

One of the most interesting results from the distance matrix variants is that the best matrix similarity metric is strongly dependent upon which underlying distance matrix is used. For example, over the 117 permutations using CD (3 AUC values \times 3 benchmarks \times 13 metrics = 117), the Tanimoto coefficient is the best performing metric 95.8 percent of the time. However, when considering TP, the Kullback-Leibler divergence is the best performing metric 97.5 percent of the time, and when considering PD, Hellinger distance is the best performing metric 95.8 percent of the time. Curiously, none of these three best performing metrics are even in the same class of matrix comparison methods. After collapsing the various metrics into their four classes, the differences between TP and PD are shown not to be statistically significant; see Fig. 3, which plots the average AUC0.10 values for each matrix similarity metric class using each distance matrix (error bars = 1 standard deviation). While there is variability within the CD over the various metrics, there are no strong significant differences between classes when using TP or PD. Interestingly, the differences observed with CD are generally conserved over all three functional site benchmarks.

It is surprising to note that there is very little sensitivity to window width when using TP and PD. While the best performing methods occur at width = 3, as indicated in Figs. 1d, 1e, and 1f, there are no consistent trends when

TABLE 1
MINER versus Best Distance Matrix Variant for a Given Distance Matrix, Width, and Functional Site Definition

		3	5	7	9
		<i>ClustalDist</i>			
<i>Active site</i>	Best variant	1.3e-2	1.3e-2	1.5e-2	1.6e-2
	MINER	2.0e-2	2.1e-2	1.8e-2	1.4e-2
	p-value	2.2e-16	2.2e-16	1.7e-1	n/a
<i>Ligand-binding site</i>	Best variant	1.3e-2	1.3e-2	1.4e-2	1.4e-2
	MINER	1.9e-2	1.9e-2	1.6e-2	1.4e-2
	p-value	8.4e-8	1.3e-11	4.5e-1	n/a
<i>Catalytic residue</i>	Best variant	1.6e-2	1.6e-2	2.0e-2	1.9e-2
	MINER	2.2e-2	2.5e-2	2.0e-2	1.6e-2
	p-value	3.4e-3	5.2e-3	9.8e-1	n/a
		<i>TREE-PUZZLE</i>			
<i>Active site</i>	Best variant	1.0e-2	9.5e-3	1.1e-2	8.2e-3
	MINER	2.0e-2	2.1e-2	1.8e-2	1.4e-2
	p-value	2.2e-16	2.2e-16	2.8e-12	3.4e-14
<i>Ligand-binding site</i>	Best variant	1.1e-2	1.0e-2	1.0e-2	7.4e-3
	MINER	1.9e-2	1.9e-2	1.6e-2	1.4e-2
	p-value	2.2e-16	2.2e-16	1.1e-8	2.2e-16
<i>Catalytic residue</i>	Best variant	1.3e-2	1.1e-2	1.2e-2	9.8e-3
	MINER	2.2e-2	2.5e-2	2.0e-2	1.6e-2
	p-value	2.2e-6	1.1e-8	1.8e-4	2.2e-3
		<i>ProDist</i>			
<i>Active site</i>	Best variant	1.2e-2	9.9e-3	7.7e-3	7.1e-3
	MINER	2.0e-2	2.1e-2	1.8e-2	1.4e-2
	p-value	2.2e-16	2.2e-16	2.2e-16	2.2e-16
<i>Ligand-binding site</i>	Best variant	1.1e-2	9.3e-3	7.2e-3	7.1e-3
	MINER	1.9e-2	1.9e-2	1.6e-2	1.4e-2
	p-value	2.2e-16	2.2e-16	2.2e-16	2.2e-16
<i>Catalytic residue</i>	Best variant	1.6e-2	1.2e-2	9.3e-3	8.0e-3
	MINER	2.2e-2	2.5e-2	2.0e-2	1.6e-2
	p-value	2.6e-4	9.1e-7	9.9e-4	1.0e-1

AUC0.10 is used to assess predictive power. Similar results are observed with AUC0.05 and AUC0.15. The p-values, calculated using the modified McNemar's test, assess the statistical significance of the improvement of MINER over the corresponding variant. Statistically significant ($p < 0.05$) improvements by MINER are highlighted in boldface. Instances where MINER fails to do better than the corresponding variant are indicated by "n/a."

considering the matrix comparison classes in Fig. 3. In contrast, CD coupled with the Tanimoto coefficient does improve with increasing window widths, whereas the other matrix comparison classes are either unaffected or slightly by worsened by increased window widths. These general conclusions regarding window width are all robust to functional site definition.

6.2 MINER

The performance of the original MINER algorithm is also plotted in Fig. 2. While there is much diversity within the myriad permutations of the overall distance matrix comparisons, the superior performance of the original MINER approach is nearly universal. This result is especially true at smaller window widths. As we have discussed previously [10], the performance of MINER peaks at window widths of five, and then, falls off as larger windows are considered. Even though the performance of MINER diminishes at larger widths, whereas the performance of TP and PD is roughly constant, MINER still soundly outperforms both over all considered widths. Conversely, there is a crossing over point with CD at window width = 9 where a small

TABLE 2
MINER versus Optimized Distance Matrix Variant

	Active site	Ligand-binding site	Catalytic residue
MINER	2.1e-2 (W=5)	1.9e-2 (W=3)	2.5e-2 (W=5)
Best of all distance matrix variants	1.6e-2 (CD, W=9)	1.4e-2 (CD, W=9)	2.0e-2 (CD, W=7)
p-value	1.3e-4	5.6e-5	1.5e-1

AUC0.10 is used to assess predictive power. Similar results are observed with AUC0.05 and AUC0.15. The p-values, calculated using the modified McNemar's test, assess the statistical significance of the improvement of MINER over the corresponding variant. Statistically significant ($p < 0.05$) improvements by MINER are highlighted in boldface.

number of the CD variants outperform MINER. However, none of these CD permutations approach the optimized predictive power of MINER at width = 5.

Table 1 demonstrates that the improvements by MINER over each corresponding variant are statistically equivalent at widths = 3 and 5. At larger widths, the differences between MINER and the TP and PD variants are significant in all cases but one. As discussed above, the CD variants improve with increasing window width. At width = 7, the improvement by MINER is no longer significant, and at width = 9, CD actually outperforms MINER. Nevertheless, Table 2 clearly demonstrates that MINER with optimized parameters outperforms the best distance matrix method in a statistically significant way on the active site and ligand-binding site benchmarks. The optimized MINER also outperforms the best MINER variant against the catalytic residue benchmark, but the difference is not deemed significant. This result is exactly consistent with our previous results that indicate MINER does a better job of predicting active and ligand-binding sites, relative to other approaches, than it does at predicting catalytic residues [19].

The entirety of these results highlights the importance of tree topology information within the PM functional site prediction paradigm. The improved performance of MINER over the distance matrix variants must come about because of the added evolutionary information that topology provides. No other explanation is likely because we have, inasmuch as possible, controlled all the variables. We have controlled for length dependence by systematically scanning over four different window widths. We have controlled for how gaps are dealt and other implementation details by building the variants within the original MINER software. In fact, because MINER uses ClustalW to construct the trees, there are precisely two differences between the original MINER and the CD distance matrix variants, specifically: 1) in MINER, ClustalW is used to construct trees, whereas in the variants, it is used to construct distance matrices; and 2) in MINER, as dictated by the use of trees, a bipartition metric is used to compare the window clustering to the complete alignment, whereas in the variants, one of the 13 different matrix comparison metrics is used. As such, it is straightforward to conclude that the improved predictive power of MINER is due to the added evolutionary insight provided by tree topology. Note that this result is not without precedent. Similar conclusions vis-à-vis the improved predictive power

of phylogenetic trees have been drawn in regards to protein-protein interaction prediction [28], [29]. However, the results presented herein retain importance because they represent the first definitive demonstration that the same is true within the protein functional site prediction problem.

7 CONCLUSION

PMs predict functional sites based on the notion that conservation of function is the ultimate evolutionary driving force. As such, PMs are identified from regions of the protein whose evolution mostly closely resembles that of the complete protein. In this report, we identify PMs using three different distance-based matrices for the calculation of phylogenetic motifs and 13 different matrix similarity measures. The performance of each permutation is evaluated and compared to the original MINER algorithm that instead relies on topological similarity. Based on the assessment of the results on a comprehensive data set of 163 proteins consisting of active site, ligand-binding site, and catalytic residue benchmarks, we demonstrate that, without exception, our traditional MINER algorithm outperforms the distance matrix variants. While the considered distance matrix variants fail to improve upon the original MINER algorithm, the results are important as they clearly demonstrate that the improved predictive power arises from the added evolutionary insight provided by phylogenetic trees. As such, they represent a simple, yet powerful way to improve the accuracy of PM functional site predictions.

ACKNOWLEDGMENTS

The authors thank members of Dr. Geoff J. Barton's Lab (University of Dundee) for assisting them with SMERFS. Luis Carlos Gonzalez and Deeptak Verma (both of the University of North Carolina at Charlotte) are thanked for helpful discussions regarding the matrix similarity scores.

REFERENCES

- [1] M. Ashburner et al., "Gene Ontology: Tool for the Unification of Biology. The Gene Ontology Consortium," *Nature Genetics*, vol. 25, pp. 25-29, 2000.
- [2] "Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB), Enzyme Supplement 5 (1999)," *European J. Biochemistry*, vol. 264, pp. 610-650, 1999.
- [3] F. Pazos and J. Bang, "Computational Prediction of Functionally Important Regions in Proteins," *Current Bioinformatics*, vol. 1, pp. 15-23, 2006.
- [4] J.D. Watson, R.A. Laskowski, and J.M. Thornton, "Predicting Protein Function from Sequence and Structural Data," *Current Opinion in Structural Biology*, vol. 15, pp. 275-284, 2005.
- [5] W.S. Valdar, "Scoring Residue Conservation," *Proteins*, vol. 48, pp. 227-241, 2002.
- [6] J.A. Capra and M. Singh, "Predicting Functionally Important Residues from Sequence Conservation," *Bioinformatics*, vol. 23, pp. 1875-1882, 2007.
- [7] T. Pupko, R.E. Bell, I. Mayrose, F. Glaser, and N. Ben-Tal, "Rate4Site: An Algorithmic Tool for the Identification of Functional Regions in Proteins by Surface Mapping of Evolutionary Determinants within Their Homologues," *Bioinformatics*, vol. 18, pp. S71-77, 2002.
- [8] S. Jones and J.M. Thornton, "Searching for Functional Sites in Protein Structures," *Current Opinion in Structural Biology*, vol. 8, pp. 3-7, 2004.
- [9] O. Lichtarge, H.R. Bourne, and F.E. Cohen, "An Evolutionary Trace Method Defines Binding Surfaces Common to Protein Families," *J. Molecular Biology*, vol. 257, pp. 342-358, 1996.
- [10] D. La, B. Sutch, and D.R. Livesay, "Predicting Protein Functional Sites with Phylogenetic Motifs," *Proteins*, vol. 58, pp. 309-320, 2005.
- [11] P. Aloy, E. Querol, F.X. Aviles, and M.J. Sternberg, "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. Molecular Biology*, vol. 311, pp. 395-408, 2001.
- [12] A.D.S. Mesa, F. Pazos, and A. Valencia, "Automatic Methods for Predicting Functionally Important Residues," *J. Molecular Biology*, vol. 326, pp. 1289-1302, 2003.
- [13] F. Pazos, A. Rausell, and A. Valencia, "Phylogeny-Independent Detection of Functional Residues," *Bioinformatics*, vol. 22, pp. 1440-1448, 2006.
- [14] D.R. Livesay, P. Jambeck, A. Rojnuckarin, and S. Subramaniam, "Conservation of Electrostatic Properties within Enzyme Families and Superfamilies," *Biochemistry*, vol. 42, pp. 3464-3473, 2003.
- [15] D. La and D.R. Livesay, "Predicting Functional Sites with an Automated Algorithm Suitable for Heterogeneous Datasets," *BMC Bioinformatics*, vol. 6, no. 116, 2005.
- [16] D.R. Livesay, P.D. Kidd, S. Eskandari, and U. Roshan, "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*, vol. 8, no. 397, 2007.
- [17] D.R. Livesay and D. La, "The Evolutionary Origins and Catalytic Importance of Conserved Electrostatic Networks within TIM-Barrel Proteins," *Protein Science*, vol. 14, pp. 1158-1170, 2005.
- [18] U. Roshan, D.R. Livesay, and D. La, "Improved Phylogenetic Motif Detection Using Parsimony," *Proc. Fifth IEEE Int'l Symp. Bioinformatic and Bioeng.*, pp. 19-26, 2005.
- [19] D.B. KC and D.R. Livesay, "Improving Position-Specific Predictions of Protein Functional Sites Using Phylogenetic Motifs," *Bioinformatics*, vol. 24, pp. 2308-2316, 2008.
- [20] D. La and D.R. Livesay, "MINER: Software for Phylogenetic Motif Identification," *Nucleic Acids Research*, vol. 33, pp. W267-270, 2005.
- [21] J.R. Manning, E.R. Jefferson, and G.J. Barton, "The Contrasting Properties of Conservation and Correlated Phylogeny in Protein Functional Residue Prediction," *BMC Bioinformatics*, vol. 9, no. 51, 2008.
- [22] J. Felsenstein, "PHYLIP (Phylogeny Inference Package) Version 3.6," Distributed by the Author, Dept. of Genome Sciences, Univ. of Washington, 2004.
- [23] J.D. Thompson, T.J. Gibson, and D.G. Higgins, "Chapter 2: Multiple Sequence Alignment Using ClustalW and ClustalX," *Current Protocols in Bioinformatics*, pp. pp 2.3.1-2.3.22, Wiley, 2003.
- [24] H.A. Schmidt, K. Strimmer, M. Vingron, and A. von Haeseler, "TREE-PUZZLE: Maximum Likelihood Phylogenetic Analysis Using Quartets and Parallel Computing," *Bioinformatics*, vol. 18, pp. 502-504, 2002.
- [25] D.T. Jones, W.R. Taylor, and J.M. Thornton, "The Rapid Generation of Mutation Data Matrices from Protein Sequences," *Computer Applications in the Biosciences*, vol. 8, pp. 275-282, 1992.
- [26] C.T. Porter, G.J. Bartlett, and J.M. Thornton, "The Catalytic Site Atlas: A Resource of Catalytic Sites and Residues Identified in Enzymes Using Structural Data," *Nucleic Acids Research*, vol. 32, pp. D129-133, 2004.
- [27] I.K. McDonald and J.M. Thornton, "Satisfying Hydrogen Bonding Potential in Proteins," *J. Molecular Biology*, vol. 238, pp. 777-793, 1994.
- [28] R.A. Craig and L. Liao, "Improving Protein Protein Interaction Prediction Based on Phylogenetic Information Using a Least-Squares Support Vector Machine," *Annals of the New York Academy of Sciences*, vol. 1115, pp. 154-167, 2007.
- [29] R.A. Craig and L. Liao, "Phylogenetic Tree Information Aids Supervised Learning for Predicting Protein-Protein Interaction Based on Distance Matrices," *BMC Bioinformatics*, vol. 8, no. 6, 2007.



Dukka B. KC received the BS degree in computer science and the PhD degree in information science (bioinformatics) from Kyoto University in 2001 and 2006, respectively. From 2006 to 2007, he was a postdoctoral fellow at Georgia Institute of Technology. Since 2007, he has been worked as a postdoctoral fellow at the University of North Carolina at Charlotte. His research interests

are in the development of algorithms for prediction of protein structure and function.



Dennis R. Livesay received the BS degree in chemistry from Ball State University in 1996, and the PhD degree in physical chemistry from the University of Illinois at Urbana-Champaign in 2000. From 2000 to 2006, he was an assistant, then associate professor in the Department of Chemistry at California State Polytechnic University at Pomona. From 2006 to 2008, he was an associate professor in the Department of Computer Science at the University of North

Carolina at Charlotte, where he is currently an associate professor in the Department of Bioinformatics and Genomics. His research interests are deciphering sequence/structure/function relationships in protein families and superfamilies using a synergistic combination of bioinformatics and computational biology techniques.

▷ **For more information on this or any other computing topic, please visit our Digital Library at www.computer.org/publications/dlib.**