Predicting Protein Functional Sites with Phylogenetic Motifs: Past, Present and Beyond

Dennis R. Livesay, Dukka Bahadur KC, and David La

Abstract More than sequence or structure, function imposes very tight constraints on the evolutionary variability within a protein family. As such, numerous functional site prediction methods are based on algorithms to uncover conserved regions that lead to conserved function. Nevertheless, evolution does allow for some systematic variability within functional regions. Based on this tenet, we have introduced the MINER algorithm to predict functional regions from phylogenetic motifs. Specifically, our approach identifies alignment fragments that parallel the overall phylogeny of the family, which are more likely to be functional due to increased evolutionary signature. In this chapter, we provide an overview of the method, summarize recent developments, and comment on future work.

Introduction

Due to the rapid increased in the number of solved sequences from next-generation sequencing technologies, accurate prediction of protein function and functional sites from sequence-derived data is now more important than ever. There are many different functional site prediction algorithms in the literature [1], most of which attempt to identify some sort of evolutionary feature within the input alignment. Meaning, they are primarily based on the simple and common dogma that conservation of function is the ultimate evolutionary driving force.

The evolutionary constraints imposed by function severely limits sequence variability at certain sites, which has led to myriad algorithms to predict function from conservation [2–4]. However, the constraints imposed by function need not completely limit variability within a given site. Rather, functional sites frequently vary somewhat dependent upon exact functional criteria (i.e., substrate specificity, catalytic efficiency, etc.) and the context of the rest of the protein, thus yielding systematic variations between subfamilies [5–6]. Unfortunately, prediction algorithms

D.R. Livesay (✉)
Department of Bioinformatics and Genomics, University of North Carolina at Charlotte,
9201 University City Blvd., Charlotte, NC 28223, USA
e-mail: drlivesa@uncc.edu

D. Kihara (ed.), *Protein Function Prediction for Omics Era*,
DOI 10.1007/978-94-007-0881-5_5, © Springer Science+Business Media B.V. 2011

based solely on these "evolutionary trace" positions result in an unsatisfactory number of false positives [7–9]. MINER is based on a similar notion; however, it attempts to identify phylogenetic motifs (PMs), which are contiguous alignment fragments, not alignment positions, that have co-evolved to satisfy the functional evolutionary constraints. Along with some judicious algorithmic implementation details discussed below, it is this distinction that leads to improved prediction accuracy of our approach.

The Past

> Based on work published between 2005 and 2007, this section describes the original MINER algorithm. In addition, we present a summary of application of the approach to the NSS protein family, which highlights MINER's utility and limitations.

The MINER Algorithm

The MINER algorithm, originally introduced in La et al. [7], is inspired by our earlier observation that motifs taken from regions known to be functionally important a priori conserve the overall phylogeny of the family [10]. Meaning, MINER reverses this scenario to look for regions that reproduce the phylogenetic clustering, and then presents them as putative functional sites. The algorithm, which is summarized in Fig. 1, begins with a sliding sequence window that generates all possible alignment fragments of fixed width from an input alignment. Subsequently, a tree is constructed on each fragment using standard phylogenetic reconstruction algorithms, which is compared to the phylogenetic tree of the whole family using a bipartition metric algorithm that counts topological differences between the pair [11]. In the original implementation of the algorithm, all overlapping fragments that score pass some threshold are grouped into a single PM. Based on the competition between site specificity and evolutionary signal, we have determined that a window width of five is ideal in most situations [7].

Tree similarity is quantified using the ubiquitous bipartition metric algorithm [13], which is also commonly referred to as the symmetric difference or the Robinson-Foulds distance. The bipartition metric simply counts the number of partitions, defined by tree branch points, varying across the pair. To improve prediction accuracy, the bipartition metric employed by MINER is actually a slightly modified algorithm, but the details of the modification and its rationale are beyond the scope of this discussion (cf. Roshan et al. [11] for a full discussion).

Figure 1 includes a typical phylogenetic similarity spectrum for the glycolytic enzyme triosephosphate isomerase, which plots tree similarity (recast as statistical z-scores) versus fragment number. We call the MINER output values phylogenetic similarity z-scores (PSZs). Because the bipartition metric provides a distance,

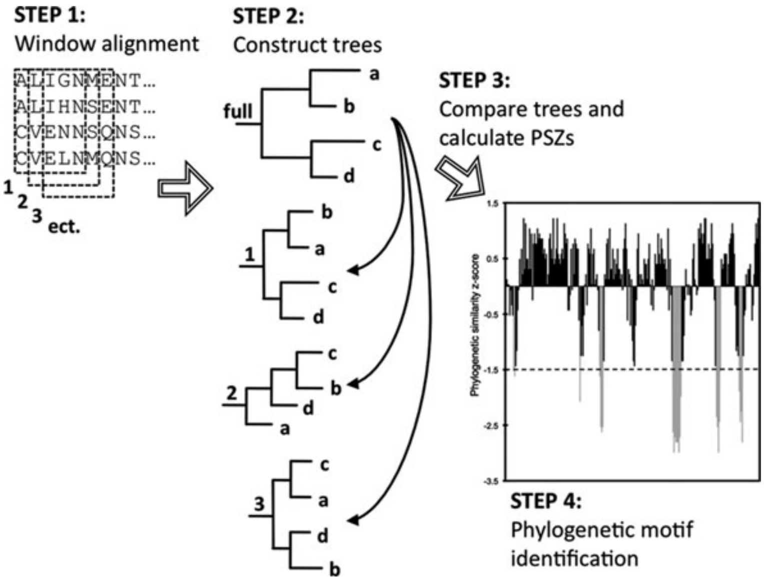

Fig. 1 A cartoon describing the original MINER algorithm. First, MINER starts with a multiple sequence alignment, from which all possible $L - W + 1$ windows are generated, where L is the alignment length and W is the window width. Second, using standard phylogenetic reconstruction techniques, a tree is generated for each window and the complete alignment. Third, the topological similarity of each window tree is compared to the overall phylogeny using a bipartition metric algorithm. The raw bipartition metric scores are then converted to statistical z-scores, called PSZs for phylogenetic similarity z-scores. Note that more negative values indicate greater tree similarity since the raw bipartition metric values represent distances. Finally, all overlapping windows scoring past some threshold are grouped into a phylogenetic motif (PM). As applied to the *triosephosphate isomerase* enzyme family, six PMs are identified. The smallest PM is composed of only a single window, whereas the largest is composed of eight contiguous windows. Note that the largest PM overlaps the PROSITE definition of the family [12]. The algorithm is more fully described in La et al. [7]

smaller values indicate greater phylogenetic similarity. Using a PSZ threshold of 1.5, Fig. 2 clearly demonstrates that the predicted functional sites map to the enzyme's active site region. In fact, with the exception of the one PM in the upper-right corner, all identified PMs (dark grey) clearly cluster around the enzyme's active site (a co-crystalized substrate analog is shown in light grey). However, this PM is actually interacting with the substrate at the active site of its homodimer partner, meaning all six PMs overlap the enzyme's active site. In a later follow-up study [14], we demonstrated the functional roles of PM residues are commonly explained in a rational way by sophisticated continuum electrostatics calculations. Therein, the biophysical calculations demonstrated that the PM residues were interacting with the strictly conserved catalytic residues to fine-tune their chemical properties. This result highlights the power of synergistically combining empirical and first principles viewpoints to understand protein function. However, biophysical calculations

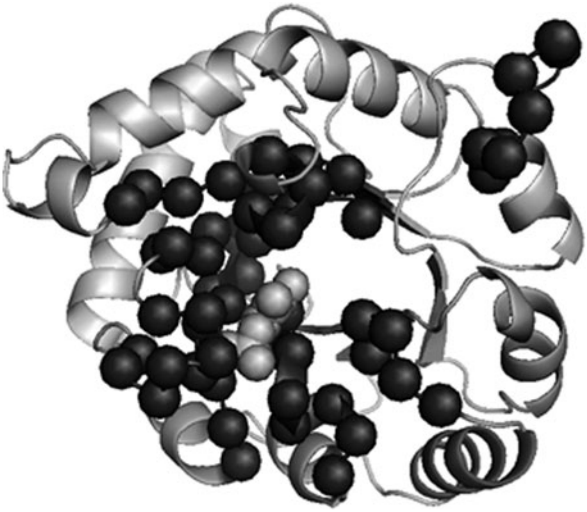

Fig. 2 Triosephosphate isomerase. The PMs identified in Fig. 1 are mapped to the structure of an example structure. *Dark grey colored spheres* represent α-carbon atoms of the predicted sites. The enzyme's substrate analog is colored *light grey* and shown in *spacefill*

are generally expensive and require structural input, thus limiting their utility for high-throughput investigations.

The PSZ threshold used can be predefined by the user or automatically determined. Threshold values of $\sim 1.5 \pm 0.5$ standard deviations are generally ideal; however, large prediction differences can occur within this range. While there are myriad signal-to-noise methods, we have developed the EXTREME algorithm to be specifically appropriate to the problem at hand [15]. The approach is based on three primary features. The first is that we pre-process the MINER output to highlight the evolutionary signal. Specifically, because they are associated with a single PM, contiguous stretches of scores within the above range are represented by a single data point, which we call sharpening. Second, the sharpened scores are then clustered into $k = 2$ groups using partition around medoids clustering. We use k-medoids clustering because it is less sensitive to outliers compared to the more common k-means clustering. The threshold is defined as the largest score within the second (more negative) cluster. Finally, there are number of algorithmic overrides that have been developed to ensure that the resultant threshold has the desired properties, such as not predicting too many PMs. A quantitative assessment of prediction accuracy on a small dataset of 32 protein families demonstrates that EXTEME leads to 69% correct predictions and 23% useful predictions. Only 11% were deemed wrong. As previously done with evolutionary trace [16], the assessment of *correct*, *useful*, and *incorrect* is determined from whether the predicted sites are, respectively, *within*, *overlapping*, or *distinct* from the known functional site. However, as we have discussed previously, this assessment is overly strict because it completely ignores functional roles outside the active site.

Prediction of Functional Sites Within the NSS Protein Family

The accuracy and utility of the MINER functional site predictions has been born out many times. As an example, we focus here on our application of the method to the neurotransmitter/sodium symport (NSS) family, which is a large and functionally diverse family of transporter proteins. In the NSS family, free energy provided by the flux of sodium and chloride ions with their electrochemical gradients across a membrane barrier is used to move chemical substrates against theirs. The chemical substrates recognized by members of the family are extremely chemically diverse, and include amino acids, biogenic amines and osmolytes. Application of MINER, along with a number of other common functional site prediction methods, identified a large number of putative functional sites, which were compared to residues identified as important from the leucine transporter transporter solved by Yamashita et al. [17] and an exhaustive survey of the experimental mutagenesis data.

MINER had the best prediction coverage of the six methods considered, predicting an impressive 62% of the benchmark sites. Moreover, MINER's overall performance was among the best considered. Interestingly, the others with similar performance were primarily conservation measures. Yet, MINER performed much better than evolutionary trace and SDPpred [18], which is another common prediction technique based on subfamily differences. To provide a balanced description of coverage and accuracy, overall performance is calculated as the Cartesian distance between (coverage, accuracy) of each method to a hypothetical perfect method (coverage = 1.00, accuracy = 1.00). The distances are normalized such that a method with 0.00 coverage and accuracy would have a value of unity. The reason that MINER's overall performance is slightly below the conservation measures is that it tends to over-predict sites. This is simply due to each prediction within MINER actually corresponding to five residues. As such, we also evaluated a relative accuracy, which is normalized by the number of predicted windows (not residues). The relative accuracy of MINER is very good, but should not necessarily be compared to the site specific methods since they are fundamentally different quantities. These results are presented in Table 1.

A very interesting result from this work was that the set of predictions from each of the six methods are generally orthogonal to each other. As such, we demonstrated that predictions based on simple intersections of the various methods significantly improve prediction accuracy. Meaning, only positions that are simultaneously predicted by multiple methods are put forth as a prediction. Impressively, prediction by any three methods (except for SDPpred that was excluded due to poor overall performance), the coverage and accuracy reached 0.56 and 0.44, respectively, which is much higher than any of the individual methods. Another interesting result from this work is based on consensus predictions. We demonstrated that predictions with better support, meaning they are predicted by multiple methods, are more likely to cluster around the leucine-binding site and the proposed transport route (cf. Fig. 3). Taken together, these two sets of results highlight the synergy and complementarity across various functional site prediction methods.

Table 1 Coverage and accuracy of the various functional site prediction schemes across all the NSS functional site benchmark

Method[a]	Coverage (%)	Accuracy[b] (%)	Overall performance[c]
Phylogenetic motif	62	24 (55%)	0.40
Motif conservation	53	35 (90%)	0.43
Position conservation	59	35	0.45
Rate4Site	50	37	0.43
Evolutionary trace	44	27	0.34
SDPpred	12	27	0.19
Intersect 2[d]	71	29	0.46
Intersect 3	56	44	0.50
Intersect 4	32	50	0.40
Intersect 5	18	67	0.37

[a] These results are reproduced from Livesay et al. [8], which provides details of the methods employed.

[b] Accuracies are reported as the ratio of correct to total alignment positions predicted. For methods that are based on alignment fragments, the relative accuracy that describes the ratio of correct predictions to the total number of alignment windows is provided in parentheses.

[c] Overall performance is calculated as the Cartesian distance between (coverage, accuracy) of each method and that of a perfect method (coverage = 1.00, accuracy = 1.00). The distance is normalized such that a method with 0.00 coverage and accuracy would have a value of unity.

[d] The Intersect predictions describe a hybrid approach composed of the unique prediction strategies. Whenever the number of predictions for a particular site are greater than the intersect value, that site is put forth as a prediction

The Present

Based on work published between 2008 and 2010, this section describes our recent attempts to improve the MINER algorithm and to explain its predictive power. Specifically, we demonstrate that the accuracy of MINER is improved by translating it into a site-specific model. Moreover, development of more rigorous hybrid methods that combine PMs and conservation provide very good predictions. Finally, we have also demonstrated, not unexpectedly, that the bulk of the predictive power of MINER comes from its topological description of evolutionary variability.

Residue Specific Predictions

As discussed above, MINER does a very good job of identifying known functional sites; however, its accuracy is somewhat tempered by its window-centric view. Moreover, the NSS family results above are just a single example, which may or may not be representative of average performance. To determine how well MINER performs relative to conservation measures, we have constructed a large well-curated and nonredundant benchmark dataset based on the catalytic site atlas

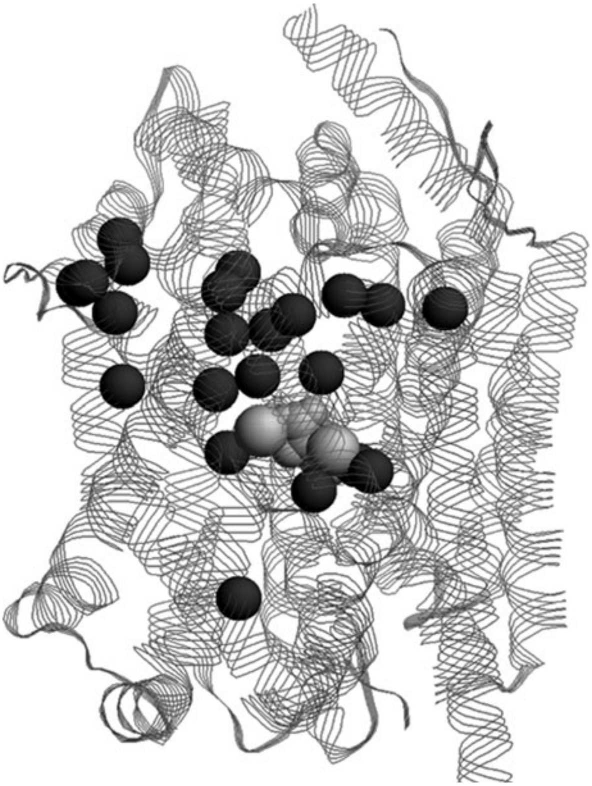

Fig. 3 The NSS family. Structural superposition of all functional site predictions within the neutrostransmitter/sodium symporter family that are predicted by at least four methods are highlighted in *dark grey* within the leucine transporter structure, which do a good job of covering the extracellular/periplasmic gate and ligand-binding site residues. The leucine, sodium ions and chloride ion are also shown in spacefill (*light grey*) at the center of the structure. Reproduced from Livesay et al. [8]

[19]. Specifically, we defined *active sites* from the catalytic residues plus all residues interacting with them [9]. To make MINER position-specific, a given alignment position is simply assigned the phylogenetic similarity score of the window centered on it.

As such, there are two key differences between this approach and what we have done prior. First, of course, we have removed MINER's inherent window-centric view. Second, we have also removed the threshold needed to group windows into a PM. Rather, like other site-specific measures, we now just have a list of scores rank-ordered from best to worst predictions. And like all methods along these lines, the appropriate cut-off to balance sensitivity and specificity is a degree of freedom to be optimized. To eliminate the arbitrariness of defining such a threshold, we apply receiver operator characteristic (ROC) analysis to quantify the balance between the two over a systematic range of cut-offs. Table 2 provides the area under curve (AUC) at a false positive rate of 0.1, which is a standard measure of the predictive power

Table 2 Receiver operator characteristic analysis for position specific predictions of active site residues across a large nonredundant dataset[a]

Method	$\text{AUC}^{\text{b}}_{0.10}$
MINER (based on phylogenetic similarity of window centered on target position)	2.13
SCORECONS[c] (which is a sum of pairs conservation score)	1.95
psMINER (based on SCORECONS)	2.38
hMINER (based on SCORECONS and $\alpha = 0.6$)	2.48

[a] These results are reproduced from KC and Livesay [9]. While not provided here, statistical significance of the improvements is discussed in the cited paper.
[b] All reported values are $\times\ 10^{-2}$.
[c] The citation for SCORECONS is Valdar [2]

of functional site prediction algorithms. (Note that AUCs at larger false positive rates are generally not considered because they would produce too many spurious predictions to be of practical usefulness as a guide for experimental studies.) Our results demonstrate that the modified PM approach is very powerful. Specifically, the PSZs result in a 9% improvement over the common sum-of-pairs conservation metric. Similar results are observed for other conservation scores.

Integrating Conservation and Evolutionary Viewpoints

Based on the complementarity between variability and conservation viewpoints discussed above, we have recently developed more rigorous algorithms to integrate both approaches. Specifically, we have developed two hybrid site-specific versions of MINER [9]. Both incorporate conservation information, but do so in distinct ways. The first approach, called psMINER for position-specific MINER, starts by rank ordering each alignment position with respect to a calculated conservation score. Next, each position is interrogated about whether or not it is found within a PM. If so, then its ranking is unaffected. However, if not, the position is shuffled to the bottom of the list and is never considered to be a possible functional site. Again, based on ROC AUC values, Table 2 demonstrates that psMINER leads to large improvements over MINER itself and the underlying conservation scores in the prediction of active site residues. In fact, the improvement over sum-of-pairs is 22% and the improvement over the PSZs is nearly 12%.

One drawback of the psMINER algorithm is that it only uses PM information as a binary, meaning a residue is either part of a PM or not. To incorporate a quantitative aspect, we have also developed a hybrid MINER (hMINER) approach that averages the phylogenetic similarity and conservation scores using optimized statistical weight α. The hMINER score for alignment position i is given by: $H_i = \alpha M_i + (1 - \alpha)C_i$, where M_i is the MINER similarity score and C_i is the conservation score. Again, Table 2 demonstrates that hMINER does an excellent job of improving predictive power. The improvement over sum-of-pairs is 27% and the improvement over the PSZs is 16%. Nevertheless, an equally interesting aspect of the hMINER approach is that the value of α dissects the relative importance of the evolutionary variability and conservation aspects of the hybrid approach. For example, depending upon the conservation score used, typical values range from $\alpha \sim 0.5 - 0.7$, indicating that descriptions of the phylogenetic variability are generally slightly more important than conservation.

The Importance of Topology

In prior work, we demonstrated that improving phylogenetic descriptions is another straightforward way of improving the predictive power of MINER [11]. Specifically, we demonstrated that we could improve prediction accuracy by focusing on phylogenetic trees reconstructed using parsimony, rather than neighbor-joining methods. However, an interesting report recently demonstrated that an algorithm similar to MINER, but instead focused on distance matrix comparisons rather than phylogenetic trees, could also provide acceptable prediction accuracies [20]. To test their assertions more rigorously, we performed an exhaustive assessment of 39 different variants of their approach over a range of window widths [21]. We considered three different types of distance matrices [22–24] and thirteen different matrix-to-matrix comparison metrics. Figure 4 summarizes our results. Specifically, it plots the ROC

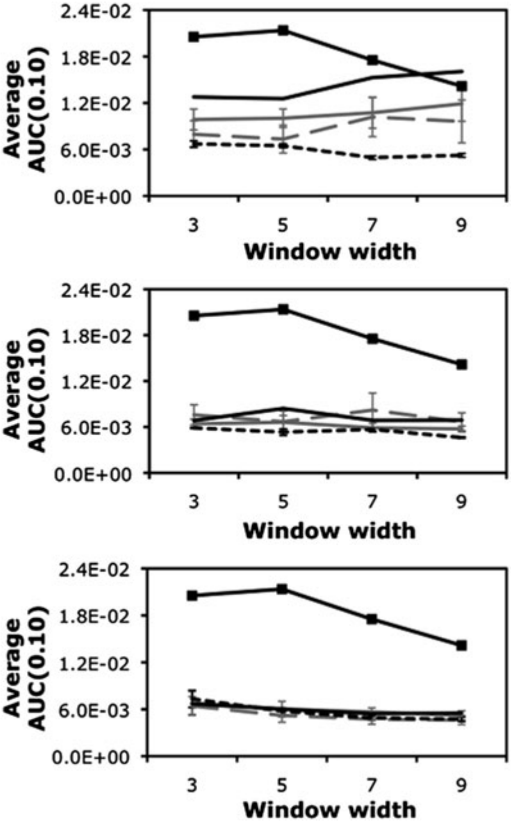

Fig. 4 The importance of topology. Average $\mathrm{AUC}_{0.10}$ values for active site prediction are plotted for each matrix similarity metric class (*black squares* = MINER, *solid black line* = Tanimoto coefficient, *solid grey line* = distance-based, *long dashed grey line* = information theory, and *short dashed 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 to, respectively, ClustalDist, TREE-PUZZLE, and ProtDist distance matrices

$AUC_{0.10}$ values for the prediction of active sites for the distance-matrix variants and the original MINER approach. The results clearly demonstrate the superior predictive power of MINER. The three panels correspond to the three different distance matrices and each curve corresponds to the four matrix comparison metric classes + MINER over a range of window widths. Taken as a whole, these results establish that the improved predictive power arises from the added evolutionary insight provided by phylogenetic trees. Meaning, tree topologies represent a simple, yet powerful way to improve the accuracy of PM functional site predictions.

The Future

The sum of our work to date in the realm of protein functional site prediction clearly indicates that strategies based on strict conservation scores and alternate strategies based on evolutionary variability both have merit. Moreover, we have clearly demonstrated that integrating viewpoints is a convenient way to improve predictive power. However, while not specifically discussed, another general conclusion from our work to date is that *all* current functional site prediction algorithms (MINER included) lack prediction specificity. While all published methods produce better than random predictions of which positions within an alignment are important, that is all they are able to do. The methods indiscriminately identify evolutionarily important sites and/or regions, but provide little additional insight. Meaning they fail to explain *how* or *why* these positions are important. In order to provide such mechanistic descriptions, we continually attempt to layer the results from biophysical calculations on representative structures from the family onto the predicted sites to assist interpretation and provided added value. Alternately, to provide the same sorts of mechanistic detail from sequence-derived data alone, the bioinformatics community needs to identify new ways to improve functional site prediction specificity by development of algorithms that are able to distinguish between various functional roles (i.e., catalytic residues, allosteric/regulatory sites, ligand-binding sites, trafficking signals, etc.). *Second generation* functional site prediction algorithms must provide this sort of specificity if we are ever going to fully extract biochemical insight from the massive amounts of sequence information that is currently being produced. To us, development of algorithms that include such mechanistic specificity is the next grand challenge for the functional site prediction community.

Accessibility and miniMINER

MINER is accessible in three ways. First, we have developed a web-based implementation called webMINER [25]. The implementation contains full functionality, including the EXTREME algorithm. It uses ClustalW [22] to construct phylogenetic trees, and uses our own partition metric implementation to compare them. Input is either a multiple alignment or a set of unaligned sequences that we will

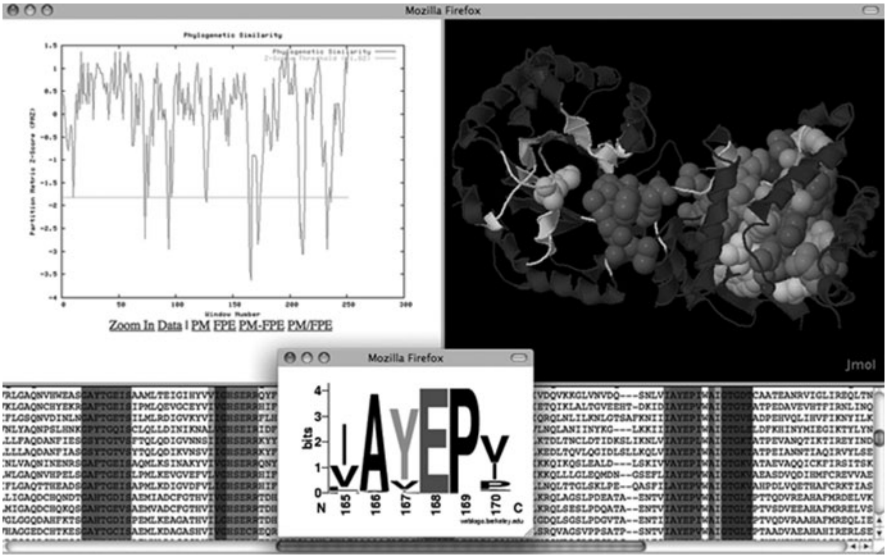

Fig. 5 Screenshot of the webMINER output. In addition to PM identification, webMINER includes a variety of additional functionality, including the option to map the PM predictions to a representative structure (as shown here). The multiple sequence alignment, which highlights the identified PMs, is hyperlinked to the structure viewer such that structural context of one or all of the PMs can be interactively analyzed. Additionally, webMINER provides sequence logo descriptions of the PMs so that the user can quickly evaluate the evolutionary variability within the identified region. Finally, all of the raw data is available for download so that user can port the data to other analysis programs

align for you. The basic ouput is a phylogenetic similarity z-score for each window, but depending upon user options, a number of additional analysis tools are also provided. The webMINER is currently accessible at http://coit-apple01.uncc.edu/MINER/. A screenshot from a typical output is provided in Fig. 5.

The second option is that, upon request, we will provide a standalone PERL program that integrates all the relevant software used by the webMINER. Meaning, it includes all of the visualization options, which can be either used or not. Unfortunately, the standalone version is rather difficult to compile and integrate. As such, if you have only a few families to analyze, we recommend that you use our web-version. Conversely, if you want to apply MINER in a large-scale way, we have recently developed a third option.

Our most recent work has focused on improving the utility of MINER by providing a streamlined version of MINER that has no dependencies upon other installed software (but Java). Specifically, this miniMINER has been programmed to ease the high-throughput use of MINER. The program simply outputs the PSZs for a given input alignment. To ease installation, we have re-implemented all of the underlying phylogenetic reconstruction and tree similarity functionalities within a self-contained Java jar file that should work seamlessly on any computer with Java installed. This miniMINER is available upon request, and a paper describing these results is currently being prepared.

Acknowledgements The development and application of MINER is based on the dedicated work of a number of people. Specifically, we wish to thank Dr. Usman Roshan, Ehsan Tabari, Brian Sutch, Patrick Kidd, and Dr. Sepehr Eskandari for contributions to the results discussed herein.

References

1. K.C., D., Livesay, D.R. A spectrum of phylogenetic-based approaches for predicting protein functional sites. Bioinformatics for systems biology Krawetz, S. (ed.) Humana Press, New York, NY: pp. 315–337 (2009).
2. Valdar, W.S. Scoring residue conservation. Proteins **48**: 227–241 (2002).
3. Pupko, T. 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 (2002).
4. Capra, J.A., Singh, M. Predicting functionally important residues from sequence conservation. Bioinformatics **23**: 1875–1882 (2007).
5. Lichtarge, O., et al. An evolutionary trace method defines binding surfaces common to protein families. J. Mol. Biol. **257**: 342–358 (1996).
6. del Sol, A., et al. Automatic methods for predicting functionally important residues. Ann. Mat. Pura. Appl. **326**: 1289–1302 (2003).
7. La, D., et al. Predicting protein functional sites with phylogenetic motifs. Proteins **58**: 309–320 (2005).
8. Livesay, D.R., et al. 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 Bioinform. **8**: 397 (2007).
9. K.C., D., Livesay, D.R. Improving position-specific predictions of protein functional sites using phylogenetic motifs. Bioinformatics **24**: 2308–2316 (2008).
10. Livesay, D.R., et al. Conservation of electrostatic properties within enzyme families and superfamilies. Biochemistry **42**: 3464–3473 (2003).
11. Roshan, U., et al. Improved phylogenetic motif identification using parsimony. Proc. IEEE Syms. Bioinform. Bioeng. **BIBE05**: 19–26 (2005).
12. Hulo, N., et al. Recent improvements to the PROSITE database. Nucleic Acids Res. **32**: D134–D137 (2004).
13. Penny, D., Hendy, M. The use of tree comparison metrics. Sys. Zoo. **34**: 75–82 (1985).
14. Livesay, D.R., La, D. Probing the evolutionary origins and catalytic importance of conserved electrostatic networks in TIM-barrel proteins. Protein Sci. **14**: 1158–1170 (2005).
15. La, D., Livesay, D.R. Accurate functional site prediction using an automated algorithm suitable for heterogeneous datasets. BMC Bioinform. **6**: 116 (2005).
16. Aloy, P., Querol, E., Aviles, F.X., Sternberg, M.J.E. Automated structure-based prediction of functional sites in proteins. J. Mol. Biol. **311**: 395–408 (2001).
17. Yamashita, A. Crystal structure of a bacterial homologue of Na+/Cl− dependent neurotransmitter transporters. Nature **437**: 215–223 (2005).
18. Kalinina, O.V. SDPpred: a tool for prediction of amino acid residues that determine differences in functional specificity of homologous proteins. Nucleic Acids Res. **32**: W424–W428 (2004).
19. Porter, C.T. The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. Nucleic Acids Res. **32**: D129–D133 (2004).
20. Manning, J.R., et al. The contrasting properties of conservation and correlated phylogeny in protein functional residue prediction. BMC Bioinform. **9**: 51 (2008).

21. K.C., D., Livesay, D.R. Topology improves phylogenetic motif functional site predictions. Trans. Comp. Biol. Bioinf. (2010) In press.
22. Thompson, J.D., et al. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22**: 4673–4680 (1994).
23. Felsenstein, J. PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the Author, Department of Genome Sciences, University of Washington (2004).
24. Schmidt, H. A., et al. TREE-PUZZLE: Maximum Likelihood Phylogenetic Analysis Using Quartets and Parallel Computing. Bioinformatics **18**: 502–504 (2002).
25. La, D., Livesay, D.R. MINER: software for phylogenetic motif identification. Nucleic Acids Res. **33**: W267–W270 (2005).