Improved Phylogenetic Motif Detection Using Parsimony

Usman Roshan
New Jersey Institute of Technology
usman@cs.njit.edu

Dennis R. Livesay
California State Polytechnic University, Pomona
drlivesay@csupomona.edu

David La
California State Polytechnic University, Pomona
dla@csupomona.edu

Abstract

*We have recently demonstrated (La et al, Proteins, **58**:2005) that sequence fragments approximating the overall familial phylogeny, called phylogenetic motifs (PMs), represent a promising protein functional site prediction strategy. Previous results across a structurally and functionally diverse dataset indicate that phylogenetic motifs correspond to a wide variety of known functional characteristics. Phylogenetic motifs are detected using a sliding window algorithm that compares neighbor joining trees on the complete alignment to those on the sequence fragments. In this investigation we identify PMs using heuristic maximum parsimony trees. We show that when using parsimony the functional site prediction accuracy of PMs improves substantially, particularly on divergent datasets. We also show that the new PMs found using parsimony are not necessarily conserved in sequence, and, therefore, would not be detected by traditional motif (information content-based) approaches.*

1. Introduction

The identification of protein functional sites is an especially important post-genomic problem. For example, knowing the location of functional sites is the first step in: understanding enzyme catalysis, assessing the result of nonsynonymous single nucleotide polymorphisms, and the identification of potential drug targets. Many protein functional site strategies have been presented in the literature (see [1] for an excellent review). Due to the richness of the data, many prediction strategies rely on protein structure information. However, despite many promising advances in high-throughput x-ray crystallography, the number of solved structures is still less than 5% of the known sequence space. As a consequence, it is imperative that accurate strategies for predicting protein functional sites from sequence be developed. Only after such prediction methods have matured will the promised biomedical benefits of large-scale sequencing efforts be more forthcoming. We have recently [2-4] demonstrated that sequence-based phylogenetic motifs (PMs) represent a promising functional site prediction strategy.

PMs [3] are short sequence alignment fragments that approximate the overall familial phylogeny. Across a structurally and functionally diverse protein dataset, we have demonstrated that PMs consistently correspond to a wide variety of known functional features [2], including catalytic sites, substrate binding epitopes, and protein-protein interfaces. Similarity between traditional and phylogenetic motifs is generally observed. However, there are instances when PMs are not (overall) well conserved in sequence. This point is enticing because it suggests that PMs are able to functionally annotate regions where traditional motifs fail. The PM approach is similar in spirit to the evolutionary trace (ET) [5-7] and similar [8-10] methods. As expected, PM results parallel those from ET investigations. Ostensibly, PMs correspond to sequence clusters of ET residues, which has the general effect of improving their functional site prediction accuracy [3]. Whereas the common use of the ET method is to map the tree-determinant positions to structure [11], no structural information is used in PM identification. Furthermore, PMs can be used in many of the same ways as traditional motifs [3;4] because they are sequence profiles with width, versus a non-contiguous collection of single alignment positions.

In this report we use the maximum parsimony (MP) optimization criterion [12] for constructing phylogenies in our phylogenetic motif detection algorithm. Maximum parsimony is an NP-hard problem [13] and therefore we resort to hill-climbing heuristics in our study. We show that with heuristic maximum parsimony trees (obtained using hill-climbing searches) our algorithm does a better job in accurately predicting protein functional sites, especially on divergent datasets. We also describe the newest implementation of our algorithm which accepts trees in the commonly used Newick format, thus making PM comparisons vis-à-vis phylogenetic reconstruction methods now possible. We implement a

modified bipartition metric based upon the TREEDIST program of the PHYLIP [14] suite phylogeny programs. Our previous calculation was specific for neighbor-joining [15] trees generated by CLUSTALW [16].

2. Methods

2.1. Phylogenetic motif identification

PMs are identified using the sliding sequence window algorithm described in [3]. Starting with a multiple sequence alignment, the algorithm parses the alignment into all possible windows of some fixed width. We find that small (five alignment positions) windows result in the most accurate functional site predictions, whereas larger windows should be more appropriate for alternate uses (e.g. assigning function to ORFans). Using standard approaches (described below) a phylogenetic tree is constructed for the complete alignment and each fragment window. The similarity (distance actually) of each window tree versus the complete familial tree is computed using a modified partition metric (see section 2.2). The partition metric scores are recast as Phylogenetic Similarity Z-scores (PSZs), which are simply the number of standard deviations away from the mean. The phylogenetic similarity spectrum, which plots PSZs vs. window number, of triosephosphate isomerase (TIM) is shown in Figure 1. All overlapping windows scoring below an adjustable PSZ threshold are grouped into a single PM. Lower partition metrics and thus, lower PSZs, indicate increased tree similarity. In the TIM example shown in Figure 1, seven PMs are identified with the smallest equal to 1 window and the longest equal to ten windows. In our early investigations [3;4], PSZ thresholds were manually adjusted to maximize functional site prediction accuracy. However, we have recently [2] implemented a clustering-based algorithm for a fully automated threshold determination.

2.2. Phylogenetic reconstruction methods

In this investigation, two types of phylogenetic reconstruction methods (MP and NJ) are used by the PM identification algorithm. In all cases, alignments are generated using CLUSTALW. CLUSTALW NJ results are qualitatively the same as the results from PRODIST NJ trees. PRODIST is also part of the PHYLIP suite of programs. NJ trees are calculated using CLUSTALW. MP trees are computed using a TBR-based hill climbing heuristic for maximum parsimony implemented in the TNT software package [17]. TBR (Tree Bisection and Reconnection) [12] is a common heuristic used to find acceptable solutions to the MP problem. The method begins with a starting tree and then modifies it to find better ones. In TBR, an edge is removed from the starting tree; the new tree arises from the reconnection of any two edges within the partitions. After trying all possible edge reconnections, the best "new" tree is selected. TNT has been demonstrated to quickly find accurate solutions to the MP problem [18].

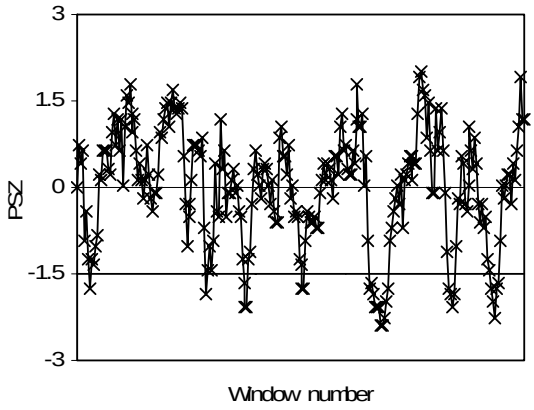

Figure 1: The phylogenetic similarity spectrum, which plots the phylogenetic similarity z-score (PSZ) vs. window number, of triosephosphate isomerase is presented as a typical case. All overlapping windows scoring past the PSZ threshold are grouped into a single PM. In this example, the threshold is -1.5, which results in the identification of seven distinct PMs.

2.3. Calculating tree distances

TREEDIST calculates the distance between a pair of trees (A and B), each with *n* leaves. The distance is calculated using the Symmetric Distance [19], which essentially enumerates the number of topological differences between the pair. The algorithm splits (partitions) each tree into two sub-trees by systematically deleting branches. After a pair of branches have been removed (one from A and one from B), the partition metric is iterated by +1 if a pair of partitions is not conserved across the A and B sub-trees.

As can be seen in Figure 2, PM window trees contain large numbers of zero-length edges. However, they also have multiple (generally two or three) ET positions that cluster in the same way as the complete alignment. It is these positions that are conserving the phylogenetic information that leads to them being identified as PMs. However, if a strict partition metric is used, windows like this are unlikely to score well because they are highly unresolved. Tree pairs that score well using the strict partition metric tend to *look*

like trees, meaning they have very few unresolved nodes. Despite a poor score using the strict partition metric, the example shown in Figure 2 clusters in the same way as the complete alignment (i.e. into two large clades that represent the two major TIM subfamilies). Conversely, leaves from the two major subfamilies of windows that do score well with the strict partition metric tend to be completely interspersed. Consequently, a modified partition metric is required.

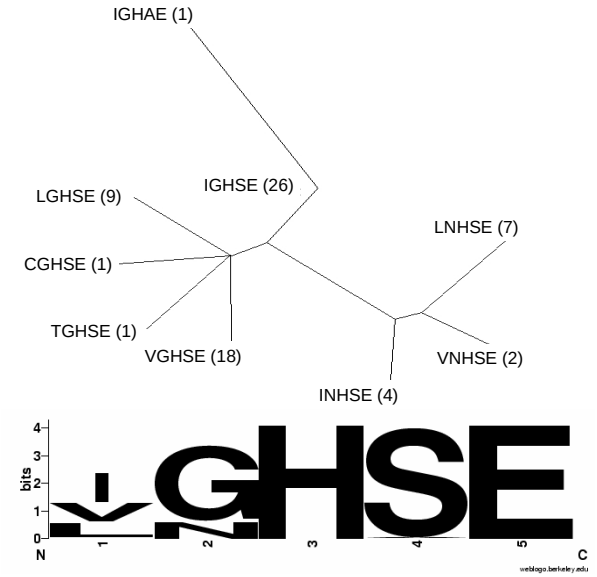

Figure 2: An example PM window tree (top) taken from the triosephosphate isomerase example. The leaves are labeled with the corresponding sequence fragment and the number of zero-length edges (in parentheses). A sequence logo [20] for the window is also provided (bottom). In this example, the differences within only two evolutionary trace-like positions (with one exception in position 4) results in leaf discrimination.

Our modified bipartition metric first contracts all internal edges of length at most and including zero. We then use the sum of the false positives and false negatives as a measure of tree similarity. This measure also captures some motifs that are also conserved in sequence just because highly conserved windows will naturally lead to unresolved trees. The low resolution of the trees in turn produces low false positives. Note, however, that our modified metric does capture motifs which are not conserved in sequence, but are conserved in phylogeny (see Section 3.3 for examples).

2.4. Determining prediction accuracy

Determining what constitutes a functional site is an exceptionally difficult problem. Several automated approaches that rely on structural proximity to known functional sites have been developed. However, automated approaches inevitably miss known functionality. For example, we have shown [2] that an automated assessment scheme can incorrectly scores one of the seven PMs in Figure 1 as a false positive (FP) [21]. The incorrect FP actually corresponds to an evolutionarily conserved dimer interface epitope that includes several stabilizing monomer-monomer interactions [22]. Despite being far removed from the active site, binding of a small molecule at the dimer interface can inactivate the enzyme [23]. As a consequence, it can be argued that this PM is indeed *functional*. This short discussion encapsulates the ambiguity involved in functional site definitions and the difficulty in assessing their predictions.

In large-scale analyses, these types of incorrect assignments must be tolerated in order to automate the process. However, in smaller datasets, more thoughtful analyses can be performed. In this report, we investigate a structurally and functionally diverse dataset of twelve proteins that we are quite familiar with. Ten of the twelve are taken from our original PM report [3], the eleventh is from [24], and the twelfth is a previously unstudied family. Functional site prediction accuracy is gauged based on the relative number of FPs and true positives (TPs) from a wide variety of functional features, including: active sites, deleterious mutation sites, co-factor binding sites, protein-protein interfaces, etc. These structural assessments (FPs and TPs) should not to be confused with the tree bipartition similarity discussed at the end of section 2.3.

3. Results and discussion

3.1. Comparison to the true partition metric

As described above, it does not make sense to consider zero-length edges when comparing the window and complete trees. This point is exemplified in Figure 3, which compares the TIM phylogenetic similarity spectrums using the modified and strict partition metrics. No correlation exists between the two plots. However, an anti-correlation between the strongest singles is qualitatively observed. This result indicates that the best scoring windows using the modified partition metric correspond to the poorest scoring windows using the true partition metric.

The explanation for this initially surprising result follows directly from the discussion in section 2.2. The four windows scoring the poorest using the strict partition metric are highly conserved (e.g. like the sequence window shown in Figure 2). Because there are so many zero-length edges, the strict partition

metric score is actually reflecting how unresolved the tree is. Conversely, after contracting all zero-length edges, the modified partition metric highlights the fact that the PM window subfamily classification closely parallels that of the complete familial tree [3]. Similar results (not shown) are observed on the other datasets. Hence forth, tree similarity is solely calculated using the modified partition metric.

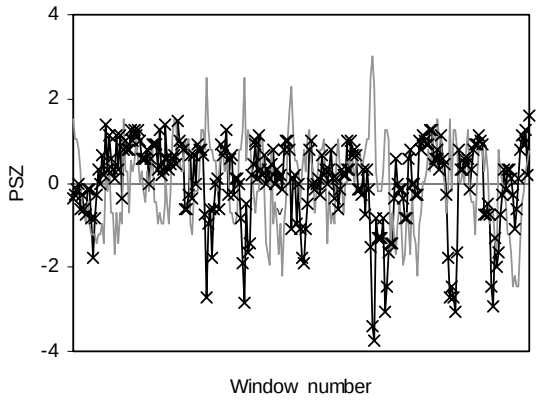

Figure 3: PSZs of triosephosphate isomerase calculated using the modified (solid line with data points) and true (grey line) partition metrics. No global correlation (R=-0.15) exists between the two plots.

3.2. Comparison of MP vs. NJ results

3.2.1. General observations. Table 1 summarizes the comparisons between MP and NJ results. Across the structurally and functionally diverse dataset, PMs calculated using MP trees generally result in improved functional site predictions, especially in more divergent datasets. Using MP trees increases the number of TPs and decreases FPs. PM predictions using MP trees are clearly superior in five of the twelve examples investigated, whereas predictions using NJ trees are superior only once. The remaining examples are determined to be equally good. Table 1 also demonstrates that the accuracy of the MP predictions improves (relative to the NJ trees) as the families become more divergent.

When considering the equally good results, the phylogenetic similarity spectrums of the MP and NJ results are virtually superimposable. As a consequence, the identified PMs are remarkably conserved. The sole exception to this trend is with the TATA-box binding protein (TBP). TBP is the smallest family investigated, which appears to be the critical point affecting the results. (We have previously argued that 25 is the minimum number of sequences necessary for accurate PM predictions [2].) Despite the TBP PM differences, the relative accuracy of the predictions is similar, meaning in some cases, MP makes good predictions that NJ does not, and in other cases the reverse occurs.

A critical examination of three examples follows. We have previously discussed the general success of PM functional site predictions in all three examples. Therefore, the following discussion is solely focused on the prediction differences using the two different phylogenetic reconstruction techniques.

3.2.2. Myoglobin. We begin our molecular-level discussions with myoglobin (Mb). Mb is the primary mode of oxygen transport in the muscles. Like its structural cousin hemoglobin, oxygen is bound to the protein via an iron-containing heme group at its active site. We have previously demonstrated that PMs are structurally clustered around the active site [3], and make several structural contacts to the heme.

The Mb family is interesting for several reasons. First, the family has the lowest average Shannon entropy of all twelve examples in our dataset, meaning it's the most conserved family. Furthermore, it is the only example within the dataset whose PM functional site predictions are more accurate using NJ trees. This is not entirely unexpected as distance-based methods generally perform well in more conserved instances. Three stark differences between the MP and NJ results are highlighted in Figure 4. The structural locations of the observed differences are also highlighted. It can be clearly observed from their structural superposition, the three additional NJ PMs identified are all structurally clustered around the heme, and thus are clearly expected to be functional. No other clear differences are observed between the two sets of predictions.

3.2.3. Triosephosphate isomerase. We have previously demonstrated that TIM PMs correspond to several important functional sites [2-4]. In fact, the functional role of TIM PMs was the primary focus of discussion in [3] and [4]. PMs correspond to all electrostatic interactions (H-bonds and salt bridges) between the enzyme and substrate and to a well-conserved monomer-monomer interface region. Furthermore, the catalytically important "flexible lid" has also been demonstrated to be identified as a PM.

Figure 4 clearly demonstrates that the results from the two methods superimpose remarkably well. All the peaks in one series have corresponding peaks in the other. However, there is a small difference between the two phylogenetic reconstruction methods. The difference (highlighted in Figure 4) results in a correctly identified functional site (a TP) using MP trees, but not with NJ. While there is a peak at this location in the NJ phylogenetic similarity spectrum, the peak is not strong enough to be unequivocally distinct from noise, whereas it is in the MP results.

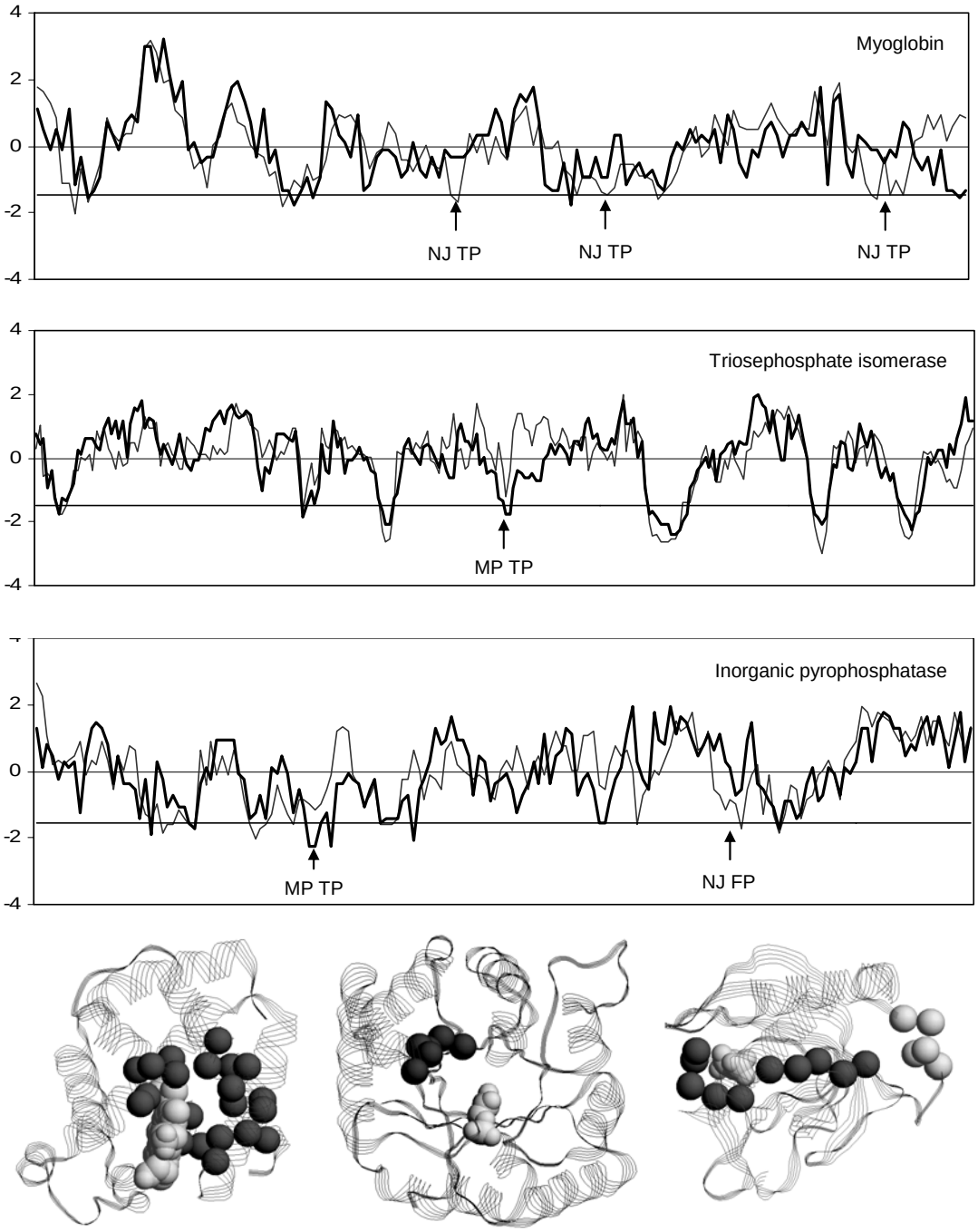

Figure 4: The phylogenetic similarity spectrum of three examples is presented. The most significant differences within the PM predictions using MP (bold lines) and NJ (thin lines) are indicated. Highlighting the alpha-carbons of the PM differences allows us to structurally assess their relative accuracy. True positives are colored grey, whereas false positives are colored white. Protein substrates are displayed in spacefill and are also colored white. The three additional myoglobin (left) PMs (grey spheres) identified using NJ trees are structurally clustered around the heme. In the center, the one additional triosephosphate isomerase PM identified using MP trees is structurally proximal to the substrate analog. In inorganic pyrophosphatase (right), the true positives identified using MP trees and the false positive identified using NJ trees are both highlighted.

Table 1: Summary of results

Protein family[1]	Average Shannon S	Number sequences	MP vs. NJ[2]	Comment[3]
Myoglobin	1.43	102	NJ	NJ predicts +3 TP
Ammonia channel (cog0004)	1.46	58	push	
Enolase (cog0148)	1.63	72	push	
TATA-box binding protein (cog2101)	1.73	25	push	
Glycerol kinase (cog0554)	1.79	53	push	
Permeases of the major facilitator superfamily (cog0477)	2.18	382	MP	see below[4]
Inorganic pyrophosphatase (cog0221)	2.19	60	MP	MP predicts +1 TP and NJ predicts +1 FP
Glutamate dehydrogenase (cog0334)	2.21	67	push	
Triosephosphate isomerase (cog0149)	2.32	70	MP	MP predicts +1 TP
Acetylglucosaminephosphate deacetylase (cog1820)	2.34	42	MP	MP predicts +2 TPs
Alcohol dehydrogenase (cog1064)	2.48	82	push	
Cytochrome P450 (cog2124)	3.02	71	MP	MP predicts +1 TP

[1] Except for the myoglobin family, all sequences are taken from the most recent update of the COG database [25]. Myoglobin sequences are taken from the Swissprot database [26]. [2] MP = Maximum parsimony; NJ = Neighbor-joining. [3] FP = False Positive; TP = True Positive. Numbers presented are net differences. For example, in the case Cytochrome P450, MP predicts two additional true positives, beyond a conserved set of predictions, whereas NJ only predicts one additional true positive, thus making a net +1 true positive for MP. [4] PMs correctly identifies one of two PROSITE [27] definitions of the family, whereas none are identified using NJ trees. Furthermore, the signal to noise ratio is much stronger within the phylogenetic similarity spectrum when using the MP trees.

3.2.4. Inorganic pyrophosphatase. Within several metabolic reactions, pyrophosphate is frequently hydrolyzed to two inorganic phosphates by the enzyme inorganic pyrophosphatase (IP). The energetically favorable hydrolysis of pyrophosphate is frequently used to "pull" anabolic reactions (e.g. protein or nucleic acid biosynthesis) to completion. Like the two previous examples, we have previously investigated IP using PMs, which are generally structurally clustered around the active site of the enzyme.

Two main differences arise in the PM results when comparing MP and NJ trees. First, the D-X-D-X-X-D PROSITE [27] definition of the family is correctly identified when using MP trees, but not NJ. The three conserved aspartate residues bind divalent metal ions, which are directly involved in catalysis. The second difference arises from a prediction the only occurs when using NJ trees. The site occurs on the polar opposite end as the active site and has no known functional significance. As a consequence, the prediction is determined to be a FP.

3.3. PMs vs. information content

Going back to our original PMs paper [3], we have demonstrated that PMs are frequently motifs in the traditional sense, meaning they have low information content. As a consequence, PMs seem to bridge the two most common techniques for predicting functional sites, namely motif-based and ET methods [2]. However, we have encountered numerous instances when PMs are not overall conserved in sequence, i.e. cytochrome P450 (CytP). These unconserved PMs occur because of the large number of subfamilies within these families.

Comparing the overall MP and NJ results, we observe that PMs from both techniques generally correspond to low information content regions in the well conserved families. Conversely, in the more divergent datasets, PMs using the two methods may or may not correspond to one another. In these more difficult datasets, well conserved (low information content) PMs are generally identified by both methods. However, differences arise within the less conserved sequence regions. It is the ability of MP to do a better job on these difficult windows that leads to its overall improvement in the functional site predictions.

For example, sequence logos of high entropy PMs from MP and NJ are compared in Figure 5. The two exemplar PMs are taken from CytP, which is the most divergent dataset investigated. Both windows have PSZs ~ -2.0. However, the fact that no single position is conserved within the NJ window makes it highly suspect – the standard dogma of molecular evolution requires that something be conserved within a *functional site*. Granted, the MP window is not as conserved as the one shown in Figure 2, but evolutionary conservation in the last four positions is evident. Furthermore, conserved subfamily discrimination leads to it being identified as a PM.

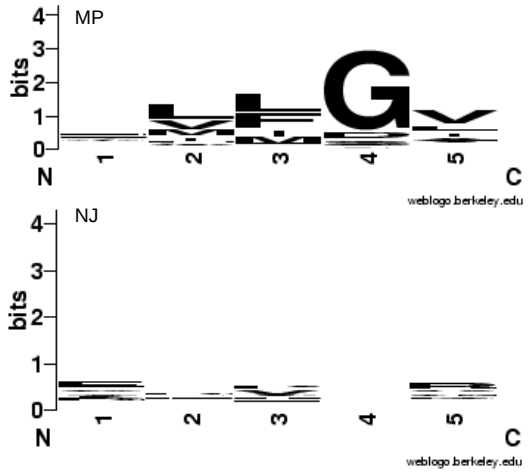

Figure 5: Sequence logos [20] comparing typical high sequence entropy PMs from MP and NJ results. The lack of any conservation within the NJ window makes it highly suspect, whereas the subfamily distinctions of the complete alignment (so-called ET positions) are maintained in the last four alignment positions of the MP window.

Globally, our results indicate that MP is doing a much better job reconstructing the phylogeny on the more difficult cases (less conserved regions). This global improvement is evident in Figure 6, which plots the PSZ vs. false positive expectation (FPE). FPE gives the probability of randomly encountering a given sequence window; lower FPEs indicate greater conservation. (The technical details of the computation are provided in [3].) The results in Figure 6 demonstrate that fewer suspect PMs (windows with FPEs < 0.2) are identified using MP. Moreover, the PSZ magnitudes from the best scoring PMs using MP actually decrease. This decrease occurs because MP is doing a much better job on the more difficult windows. The raw partition metrics of the easier windows (i.e. best scoring PMs) are generally conserved between NJ and MP. However, because the NJ distribution is more spread out, the easy PMs are farther from the mean. Conversely, the PSZs of MP PMs are decreased in magnitude for precisely the opposite reason, namely MP identifies more good PMs. With the exception of Mb, these trends are conserved across the dataset.

4. Conclusions

PMs represent a promising approach for predicting protein functional sites. Previously, phylogenetic trees have been determined using distance-based approaches. In this paper, we demonstrate that PMs identified using MP are generally superior to those identified using NJ trees. Not surprisingly, this is especially true in more divergent datasets. Future work will also compare maximum likelihood trees to the previous two as well as examine the effect of different alignments on the accuracy of phylogenetic motif detection using our algorithm.

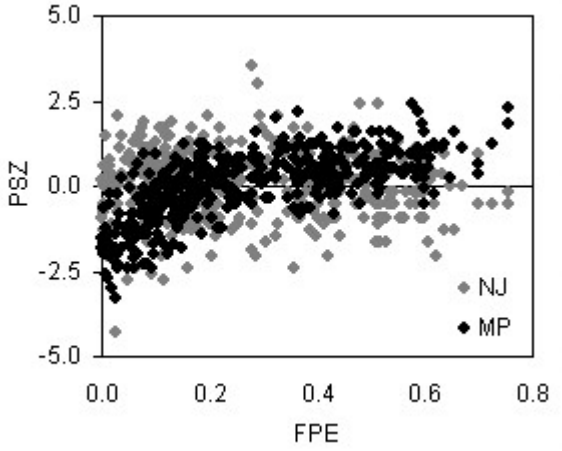

Figure 6: PSZ vs. FPE for both sets of cytochrome P450 results. The overall quality of the MP PMs is substantially improved relative to the NJ results. This result occurs because MP does a much better job reconstructing trees on the more "difficult" windows. Similar results are generally obtained in the other datasets.

5. References

[1] Jones S and Thornton JM. Searching for functional sites in protein structures. *Curr Opin Chem Biol* 8 (2004) pp:3-7.

[2] La D and Livesay DR. Predicting functional sites with an automated algorithm suitable for heterogeneous datasets. *BMC Bioinformatics* 6 (2005) pp:116.

[3] La D, Sutch B, and Livesay DR. Predicting protein functional sites with phylogenetic motifs. *Proteins* 58 (2005) pp:309-320.

[4] Livesay DR and La D. The evolutionary origins and catalytic importance of conserved electrostatic networks within TIM-barrel proteins. *Protein Sci* 14 (2005) pp:1158-1170.

[5] Lichtarge O, Bourne HR, and Cohen FE. An evolutionary trace method defines binding surfaces common to protein families. *J Mol Biol* 257 (1996) pp:342-358.

[6] Lichtarge O, Sowa ME, and Philippi A. Evolutionary traces of functional surfaces along G protein signaling pathway. *Methods Enzymol* 344 (2002) pp:536-556.

[7] Lichtarge O, Yao H, Kristensen DM, Madabushi S, and Mihalek I. Accurate and scalable identification of functional sites by evolutionary tracing. *J Struct Funct Genomics* 4 (2003) pp:159-166.

[8] Armon A, Graur D, and Ben Tal N. ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *J Mol Biol* 307 (2001) pp:447-463.

[9] Bickel PJ, Kechris KJ, Spector PC, Wedemayer GJ, and Glazer AN. Finding important sites in protein sequences. *Proc Natl Acad Sci U S A* 99 (2002) pp:14764-14771.

[10] del Sol MA, Pazos F, and Valencia A. Automatic methods for predicting functionally important residues. *J Mol Biol* 326 (2003) pp:1289-1302.

[11] Yao H, Kristensen DM, Mihalek I, Sowa ME, Shaw C, Kimmel M, Kavraki L, and Lichtarge O. An accurate, sensitive, and scalable method to identify functional sites in protein structures. *J Mol Biol* 326 (2003) pp:255-261.

[12] Swofford DJ and Olsen GJ. *Phylogeny reconstruction, Molecular Systematics*. Hills D, Moritz C, Marble BK, eds Sinauer Ass. Inc. (1996) pp:407-514.

[13] Foulds LR and Graham RL. The Steiner problem in phylogeny is NP-complete. *Advances in Appled Math* 3 (1982) pp:43-49.

[14] Felsenstein J. PHYLIP -- Phylogeny Inference Package [Version 3.2]. *Cladistics* 5 (1989) pp:164-166.

[15] Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* 4 (1987) pp:406-425.

[16] Thompson JD, Higgins DG, and Gibson TJ. 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 (1994) pp:4673-4680.

[17] Goloboff P. Analyzing large data sets in reasonable times: solution for composite optima. *Cladistics* 15 (1999) pp:415-428.

[18] Roshan U, Moret BME, Warnow T, and Williams TL. Rec-I-DCM3: a fast algorithmic technique for reconstructing large phylogenetic trees. *Proceedings of the IEEE Computational Systems Bioinformatics conference [CSB]*. Stanford, California (2004).

[19] Robinson DF and Foulds LR. Comparison of phylogenetic trees. *Math BioSci* 53 (1981) pp:131-147.

[20] Crooks GE, Hon G, Chandonia JM, and Brenner SE. WebLogo: a sequence logo generator. *Genome Res* 14 (2004) pp:1188-1190.

[21] Aloy P, Querol E, Aviles FX, and Sternberg MJ. 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 (2001) pp:395-408.

[22] Kursula I and Wierenga RK. Crystal structure of triosephosphate isomerase complexed with 2-phosphoglycolate at 0.83-A resolution. *J Biol Chem* 278 (2003) pp:9544-9551.

[23] Tellez-Valencia A, Olivares-Illana V, Hernandez-Santoyo A, Perez-Montfort R, Costas M, Rodriguez-Romero A, Lopez-Calahorra F, Tuena DG-P, and Gomez-Puyou A. Inactivation of triosephosphate isomerase from Trypanosoma cruzi by an agent that perturbs its dimer interface. *J Mol Biol* 341 (2004) pp:1355-1365.

[24] La D and Livesay DR. MINER: Software for phylogenetic motif identification. *Nucleic Acids Res* (2005) in press.

[25] Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, and Natale DA. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 4 (2003) pp:41.

[26] Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, and Schneider M. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res* 31 (2003) pp:365-370.

[27] Hulo N, Sigrist CJ, Le S, V, Langendijk-Genevaux PS, Bordoli L, Gattiker A, De Castro E, Bucher P, and Bairoch A. Recent improvements to the PROSITE database. *Nucleic Acids Res* 32 Database issue (2004) pp:D134-D137.