

Predicting Protein Functional Sites With Phylogenetic Motifs

David La,¹ Brian Sutch,¹ and Dennis R. Livesay^{2*}

¹Department of Biological Sciences, California State Polytechnic University, Pomona, California

²Department of Chemistry, California State Polytechnic University, Pomona, California

ABSTRACT In this report, we demonstrate that phylogenetic motifs, sequence regions conserving the overall familial phylogeny, represent a promising approach to protein functional site prediction. Across our structurally and functionally heterogeneous data set, phylogenetic motifs consistently correspond to functional sites defined by both surface loops and active site clefts. Additionally, the partially buried prosthetic group regions of cytochrome P450 and succinate dehydrogenase are identified as phylogenetic motifs. In nearly all instances, phylogenetic motifs are structurally clustered, despite little overall sequence proximity, around key functional site features. Based on calculated false-positive expectations and standard motif identification methods, we show that phylogenetic motifs are generally conserved in sequence. This result implies that they can be considered motifs in the traditional sense as well. However, there are instances where phylogenetic motifs are not (overall) well conserved in sequence. This point is enticing, because it implies that phylogenetic motifs are able to identify key sequence regions that traditional motif-based approaches would not. Further, phylogenetic motif results are also shown to be consistent with evolutionary trace results, and bootstrapping is used to demonstrate tree significance. *Proteins* 2005;58:309–320.

© 2004 Wiley-Liss, Inc.

Key words: phylogenetic motif; phylogenetic tree; phylogenomics; functional site prediction; sequence–function relationships

INTRODUCTION

The identification of regions responsible for protein stability and function is an especially important postgenomic problem.¹ There have been several recent attempts to predict functional sites from sequence alone. Sequence motifs have been used with some success in this endeavor^{2,3}; however, motif-based approaches result in too many false positives to be useful in large-scale analyses. Phylogenomic techniques, which use evolutionary information to improve functional classification accuracy, are particularly useful in large-scale analyses.⁴ Similarly, methods that search for evolutionary trace (ET) residues,⁵ or positions that define subfamily classification within a phylogenetic tree, frequently correspond to sites critical to function. ET methods, and others like it,^{6–9} identify key

features structurally clustered around substrates and dimer interfaces.^{10–16} It has also been shown that ET positions form statistically significant clusters.¹⁷ Using structural clusters, the utility of the ET method in large-scale postgenomic endeavors has been established.¹⁸

We have demonstrated that motifs taken from regions known to be functionally important a priori conserve the overall phylogeny of the family. Using the ubiquitous enzyme copper, zinc superoxide dismutase (CuZnSOD), we have shown that a contiguous subsequence taken from 3 functionally annotated motifs, each 5–10 residues long and representing less than 10% of the overall alignment, conserve the overall phylogeny of the family.¹⁹ Functional annotation was determined by the experimental mutagenesis results of Bordo et al.²⁰ Using the same strategy, we have shown that functional motifs taken from the enolase superfamily also conserve the overall phylogeny. Due to a lack of sufficient functional annotation, enolase superfamily functional regions were identified from calculated pK_a shifts.²¹ In both instances, randomly generated test cases of the similar lengths fail to reproduce the overall phylogeny.

In this report, we demonstrate the usefulness of using motifs conserving the overall familial phylogeny [termed phylogenetic motifs (PMs)] as functional site predictions. Reversing the previous scenario, we find that PMs are frequently associated with key functionality. The ability of our approach to identify functional sites is conserved across a structurally and functionally heterogeneous data set. Our exemplar protein data set (Table I) includes single- and multidomain proteins, monomer and protein complexes, and is representative of the 4 main Structural Classification of Proteins (SCOP) classes.²² Inasmuch as is possible, structure is used to determine the accuracy of the PM predictions. PM functional site predictions are mapped

The Supplementary Materials referred to in this article can be found at <http://www.interscience.wiley.com/jpages/0887-3585/suppmat/index.html>

Grant sponsor: American Chemical Society Petroleum Research Fund; Grant number: 36848-GB4. Grant sponsor: NIH; Grant number: S06 GM53933-07. Grant sponsor: National Center for Supercomputing Applications; Grant number: MCB00018N.

*Correspondence to: Dennis Livesay, Department of Chemistry, California State Polytechnic University, Pomona, 3801 W. Temple Avenue, Pomona, CA 91768. E-mail: drivesay@csupomona.edu

Received 27 March 2004; Accepted 28 July 2004

Published online 30 November 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20321

TABLE I. Summation of Results on Our Protein Data Set^a

Protein family	Number sequences	Average Shannon S	Z-score threshold	Number ph. motifs	SCOP class	Example PDB ID
Acetylglucosamine-6-phosphate deacetylase	42	2.34	-2.2	6	α/β	1O12
Alcohol dehydrogenase	82	2.48	-2.2	6	α/β	1JVB
CuZnSOD	124	1.48	-1.5	5	β	1SPD
Cytochrome P450	71	3.02	-2.2	6	α	1N4G
Enolase	73	1.63	-2.0	5	α/β	2ONE
Glycerolaldehyde-3-phosphate dehydrogenase	94	2.02	-1.8	4	$\alpha+\beta$	1DC4
Glycerol kinase	53	1.79	-2.0	8	α/β	1BO5
Glutamate dehydrogenase	67	2.21	-1.8	8	α/β	1HWZ
Inorganic pyrophosphatase	60	2.19	-2.2	3	β	1I6T
Myoglobin	102	1.43	-1.5	4	α	1MBA
Succinate dehydrogenase—FAD	84	2.50	-2.0	7	α/β	1NEK
Succinate dehydrogenase—Fe/S	64	2.19	-2.2	4	$\alpha+\beta$	1NEK
Succinate dehydrogenase—heme-binding/lipid bilayer anchor	27	2.32	-2.2	2	α	1NEK
TATA box-binding protein	25	1.73	-1.5	5	$\alpha+\beta$	1TBP
Triosephosphate isomerase	70	2.32	-1.5	6	α/β	7TIM

^aA sequence window width of 5 is used in each of the above examples.

onto closely related protein structures to establish prediction accuracy.

The PM approach is similar *in spirit* to the ET methods. In essence, PMs identify sequence clusters of ET positions. A thorough comparison of the predictions from the 2 methods is provided below. There is significant overlap between the results. However, some subtle differences occur. PMs tend to be more structurally clustered around key functionality than trace residues, especially substrate epitopes. Consistent with previous trace residue observations,²³ PMs are also generally conserved in sequence; thus, they can be considered motifs in the traditional sense. Therefore, we conclude that PMs often represent a subset of motif space. This result is notable because PMs are identified by tree topology only. We also critically examine the sensitivity of our unique approach to adjusting the key window width and phylogenetic similarity threshold parameters.

METHODS

Phylogenetic Motif Identification

We employ a sliding sequence window algorithm to comprehensively evaluate the phylogenetic similarity of each window compared to that of the complete alignment. An input alignment is parsed into a series of windows of predetermined length. A phylogenetic tree is used to cluster each window based on sequence. The similarity between each window and the overall tree is computed from their topological similarity using the partition metric algorithm.²⁴ The partition metric simply counts the number of outgroups found in one tree or the other, but not both, meaning it counts the number of topological differences. Therefore, the smaller the partition metric is, the greater the tree similarity. PMs represent overlapping windows displaying significant tree similarity. Empirically, we find that short sequence windows, length 5–10, are the most sensitive [Fig. 1(A)]. Larger window sizes are less able to identify statistically significant regions conserv-

ing phylogeny. Highly gapped windows (with more than 50% of the alignment positions that are more than 50% gapped) are purged. The number of gapped motifs is very low (less than 5%) for any portion of the protein other than the terminal ends. In this work, myoglobin and CuZnSOD sequences are taken from SWISS-PROT,²⁵ with all fossil and remote homolog sequences purged. The only sequences that have been purged are sequences with different functions. All other sequences are from the recently updated version of the Clusters of Orthologous Groups (COG) database.^{26,27} All orthologs are included in the familial alignments—none are purged. Multiple sequence alignments are constructed using CLUSTALW.²⁸ CLUSTALW is not always the best alignment method, especially in cases with appreciable sequence divergence. However, this is not an issue owing to the high sequence similarity in the data sets. The average column Shannon entropy score (in “bits”) for each masked alignment is given in Table I (lower scores indicate higher conservation); alignment masking involves deleting positions with more than 50% gaps. Additionally, the computational speed of CLUSTALW, compared to T-Coffee,²⁹ for example, makes it attractive for large-scale analyses.⁴ Phylogenetic trees are calculated using the distance-based algorithms within CLUSTALW and PHYLIP.³⁰ Distance-based approaches are used to ensure computational efficiency as well. Additionally, as Kuhner and Felsenstein³¹ point out, distance-based approaches sometimes outperform maximum likelihood methods on short sequences. Phylogenetic similarity is quantified using z scores calculated from the partition metric distribution. (Lower partition metric values, and thus lower z values, indicate higher phylogenetic similarity.) Plotting the phylogenetic similarity z scores (PSZs) against window number facilitates sequence comparison [Fig. 1(A)].

After all phylogenetic comparisons are made, the PSZ threshold can be adjusted to alter what constitutes a “hit.” The threshold can be adjusted to be more stringent or more

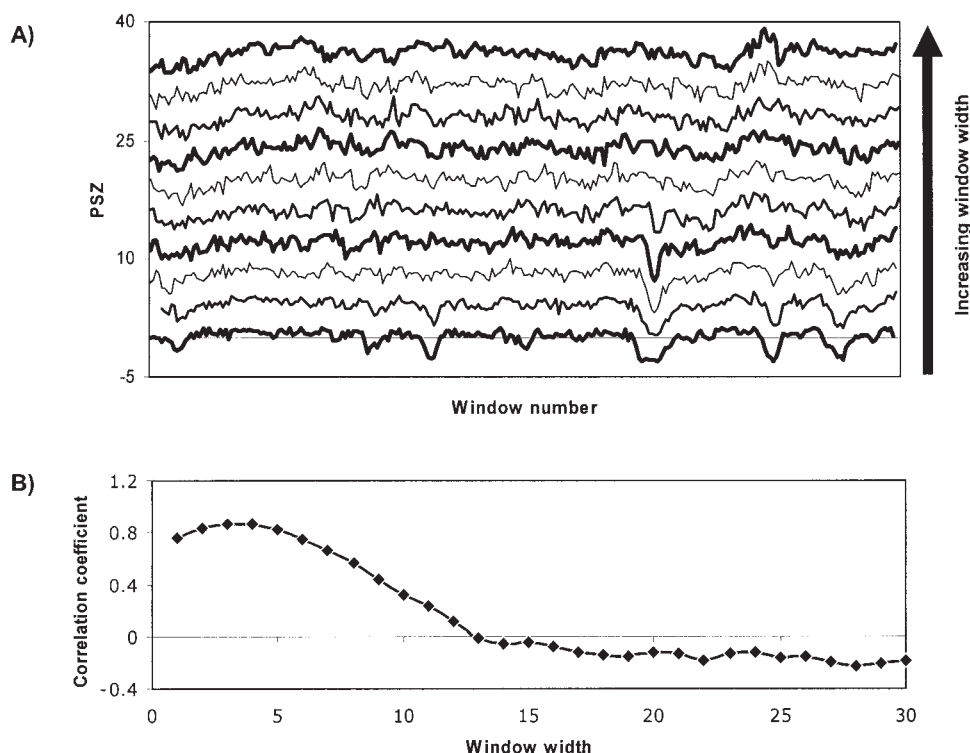


Fig. 1. (A) Empirical results indicate that smaller window widths give much better signal to noise. All TIM window PSZs are plotted against window number. From bottom to top, the series represent window widths of 5, 7, 9, 11, 13, 15, 18, 21, 24, and 27. Each subsequent series has been shifted upwards by 4 to facilitate comparisons. Some series have been shifted left or right to overlap corresponding regions. Window widths greater than ~11 do not give any signal appreciably greater than noise. In the text, all discussed results use a window width of 5. (B) There is a strong linear correlation between phylogenetic and traditional motif scores. This correlation falls off sharply with increasing window size.

accommodating. Relaxing the z -score threshold identifies regions with no obvious functional relevance. On the other hand, setting the threshold too stringently will result in a large number of false negatives. We have empirically examined the effect of adjusting the z -score threshold. Comprehensive testing of window widths between 2 and 30 on our structurally heterogeneous data set indicates that a window size of 5 and a PSZ threshold between -1.5 and -2.0 are generally best at identifying regions structurally clustered around the active site. PMs are defined as all overlapping windows scoring past the PSZ threshold. Our results also indicate that the likelihood of PMs to correspond to functional sites is largely independent of the number of familial sequences within the original alignment. Sequence data sets range from 25 (TATA box-binding protein) to several hundred (see Table I). In each example, annotating the identified PMs (using the above parameters) as functional is consistent with structural and biochemical data.

All identified PMs are compared to ET predictions. ET predictions are made using the Evolutionary Trace Server (<http://www-cryst.bioc.cam.ac.uk/~jiye/evoltrace/evoltrace.html>), a web implementation of the approach.^{10,11} Starting with the same alignments as above, the Evolutionary Trace Server uses PHYLIP³⁰ to build a phylogenetic tree. Ten tree partitions (cut levels) are examined in each

example. Results are plotted against the sequence of the structural examples to facilitate comparisons.

Bootstrap Analysis

Tree stability is gauged using bootstrap analysis. The SEQBOOT feature in PHYLIP is used to generate 100 resampled bootstrapped data sets. Bootstrapping is generally employed to test the significance of internal portions of a given tree. For example, the CONSENSE feature in PHYLIP computes a consensus tree from the resamples and scores the stability of each outgroup. Ranking global tree significance is more problematic, especially when not considering the underlying sequence data.³² The problem arises from the fact that even a single pair of adjacent leaves conserved between two trees is usually enough for probability-based approaches to ensure that the trees are significantly closer than random. To test global tree stability, we again rely on topological similarity. The topology of each resampled tree is compared to the original. Tree stability (which ranges from 0 to 100) is the count of resampled trees with normalized partition metric scores below some threshold (as before, lower partition metric scores indicate higher tree similarity). The partition metric indicates the number of outgroups that vary between the two trees in question. These scores are normalized by dividing by the theoretical maximum number of possible

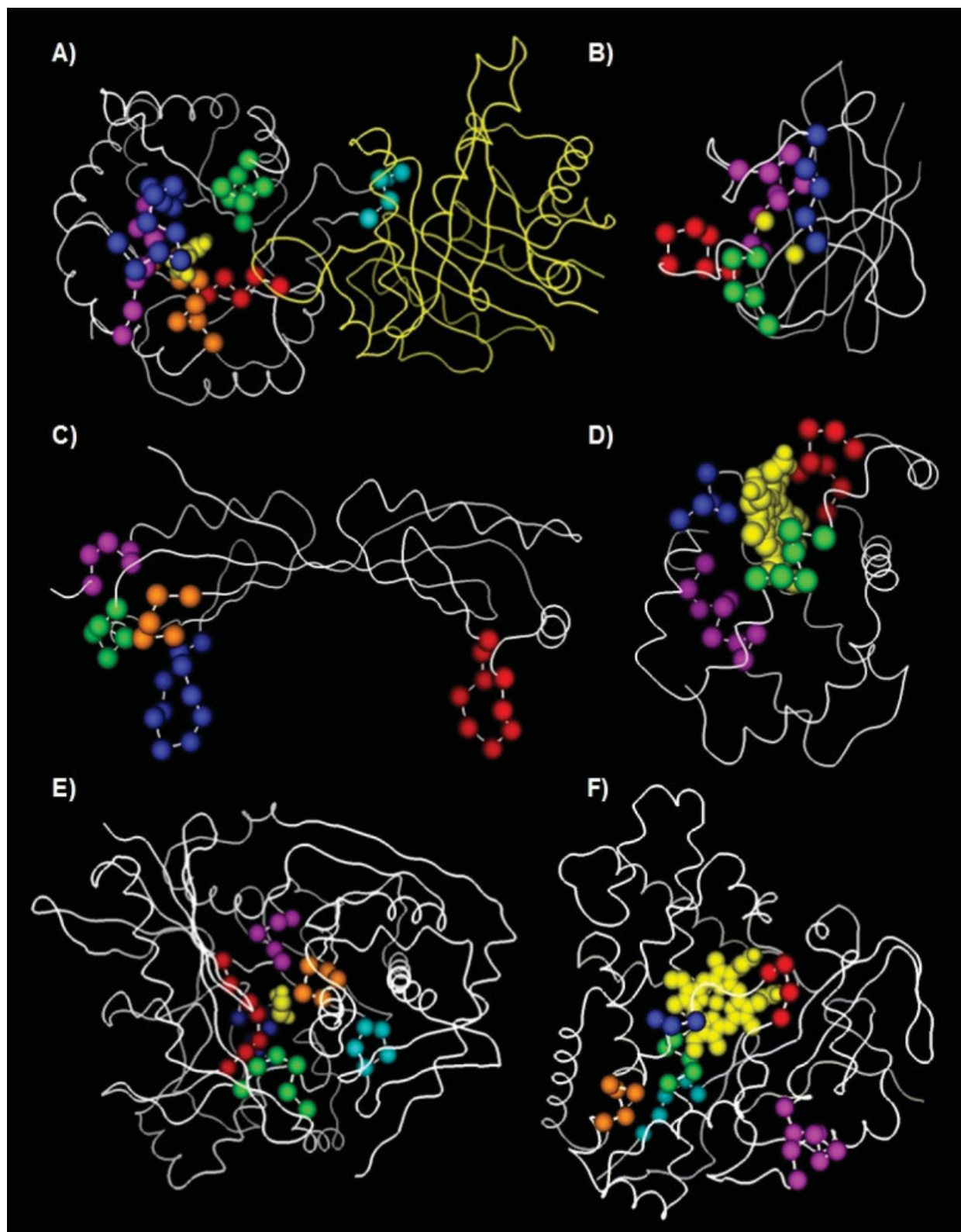


Fig. 2. PMs consistently correspond to key functional sites. This figure shows a sampling of the 15 structurally diverse examples studied here. Colored spheres represent PM α -carbons; varying colors correspond to each identified PM. Identified PMs are structurally clustered and correspond to functional sites defined by the plastic loop regions of (A) TIM, (B) CuZnSOD, and (C) TATA box-binding protein; the active site clefts of (D) myoglobin and (E) glycerol kinase; and the partially buried heme binding region of (F) cytochrome P450. Substrates are colored yellow and correspond to (A) phosphoglycolohydroxamic acid, (B) copper and zinc cations, (D and F) heme, and (E) glycerol. For reference, the second chain of the TIM homodimer is also shown in yellow; the second chain of the CuZnSOD is not shown.

differences. We employ a threshold equal to 20% of the normalized value. For example, the theoretical maximum number of differences in the triosephosphate isomerase (TIM) trees is 138. Therefore, the tree stability metric counts the number of resampled trees with less than $0.2 \times 138 = 28$ outgroup differences between it and the original.

False-Positive Expectation Calculation

Motif false-positive expectation (FPE) is calculated in a manner reminiscent of eMOTIF.³³ A regular expression is generated for each window, from which an FPE is calculated. This is the likelihood of encountering each motif randomly. Unlike with eMOTIF, substitution groups are not used in generating the regular expression. Background probabilities of each residue observed in a position are summed, resulting in the likelihood of randomly encountering a residue included in that position. The overall likelihood of randomly encountering a given sequence is calculated by multiplying the probabilities calculated for each position. For example, the FPE of the regular expression $A[V,I,L]T[K,R]P$ is calculated by the equation:

$$p(\text{motif}) = p(A) \cdot [p(V) + p(I) + p(L)] \cdot p(T) \cdot [p(K) + p(R)] \cdot p(P).$$

Background probabilities are calculated from the updated version of the COG database. Although not theoretically rigorous, gaps are treated as a "21st amino acid." Residue and gap probabilities are determined from the alignments of each orthologous group with the COG database. To eliminate overbiasing gaps, alignment positions with more than 50% gaps are not tabulated. The resulting gap probability is slightly less than alanine, the most probable residue within our data set. Future work will attempt to more rigorously account for gap probabilities. While simplistic, the approach is corroborated through comparison to a more sophisticated (traditional) motif identification method (MEME).

MEME-Identified Motifs

To confirm the generally conserved nature of the PMs, MEME³⁴ is also used to identify motifs in each example. MEME uses expectation maximization to identify conserved regions in a set of ungapped DNA or protein sequences. We have demonstrated that using MEME with a single set of robust parameters is suitable for heterogeneous data sets.³⁵ Here, custom settings include a minimum and maximum motif width of 10 and 20, a motif model biased toward 0 or 1 motif occurrence per sequence, a maximum motif search number of 5, and an *E* value threshold of 0.01.

RESULTS AND DISCUSSION

Molecular Examples

Triosephosphate isomerase (TIM) exemplifies the connection between PMs and protein functional sites. The TIM-barrel fold is one of the most ubiquitous in nature.³⁶ The active site is defined by loop regions at the top of the barrel that connect the β - α - β segments. This modular design allows large amounts of sequence plasticity, which is necessary to catalyze such a wide variety of chemical

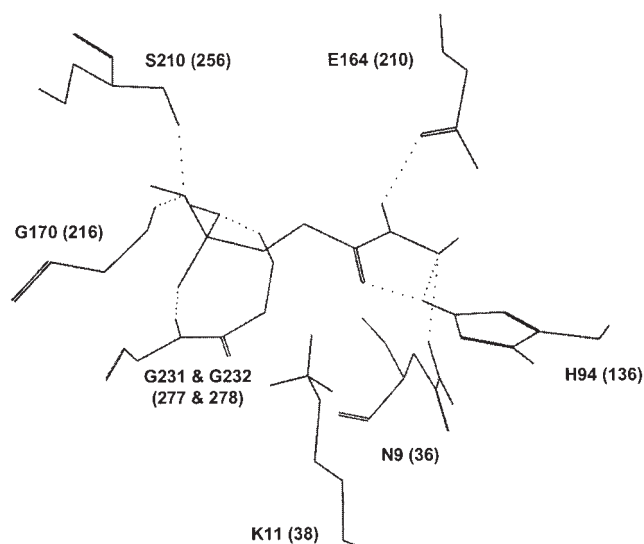


Fig. 3. Mapping the TIM PMs to structure highlights their functional relevance [see Fig. 2(A)]. In the case of TIM, all 8 of the LIGPLOT-identified enzyme-substrate contacts are PM residues. This image is centered on the substrate analog phosphoglycolohydroxamic acid. Seven PM residues are hydrogen-bonded to, and one forms an ion pair with, the substrate analog. Residues are labeled corresponding to both the PDB and alignment (in parentheses) numbering schemes.

reactions. In this report, we focus much of our discussion on TIM to highlight our combined results. Using a window width of 5, TIM windows conserving the overall familial phylogeny consistently correspond to key functionality within the protein. As acknowledged in del Sol et al.,⁸ defining what constitutes a "functional site" is not trivial. In this work, we solely judge prediction accuracy by literature results and qualitative and quantitative comparisons to structure. PMs structurally clustered around known functionality (i.e., catalytic residue, binding sites, etc.) are assumed to be functional.

Figure 2(A) highlights the 6 best-scoring TIM PM regions. Five of the 6 regions are directly interacting [H-bond, salt bridge, and van der Waals (VDW) interactions] with the substrate (Fig. 3). The sixth window is involved in quaternary structure interactions at the dimer interface. TIM PMs cover all 8 LIGPLOT³⁷ identified electrostatic interactions (excluding VDW interactions) between the enzyme and substrate. Four of these positions, Lys11, His94, Glu164, and Gly170 [using *Saccharomyces cerevisiae* ortholog numbering from the Protein Data Bank (PDB ID: 7TIM)], are strictly conserved throughout the family. In fact, many positions within identified PMs are strictly (or nearly so) conserved. More than 30% of the PMs positions are conserved better than 90% (Fig. 3). This result illustrates the conserved nature of the PMs. (Using calculated FPEs and MEME, we examine this result more thoroughly in the next section.) In general, the remaining positions are globally variable, yet locally conserved across subfamilies, resulting in conservation of the overall familial phylogeny.

Of the conserved TIM-substrate contacts, Glu164 is the catalytic residue, Lys11 forms a stabilizing ion pair with

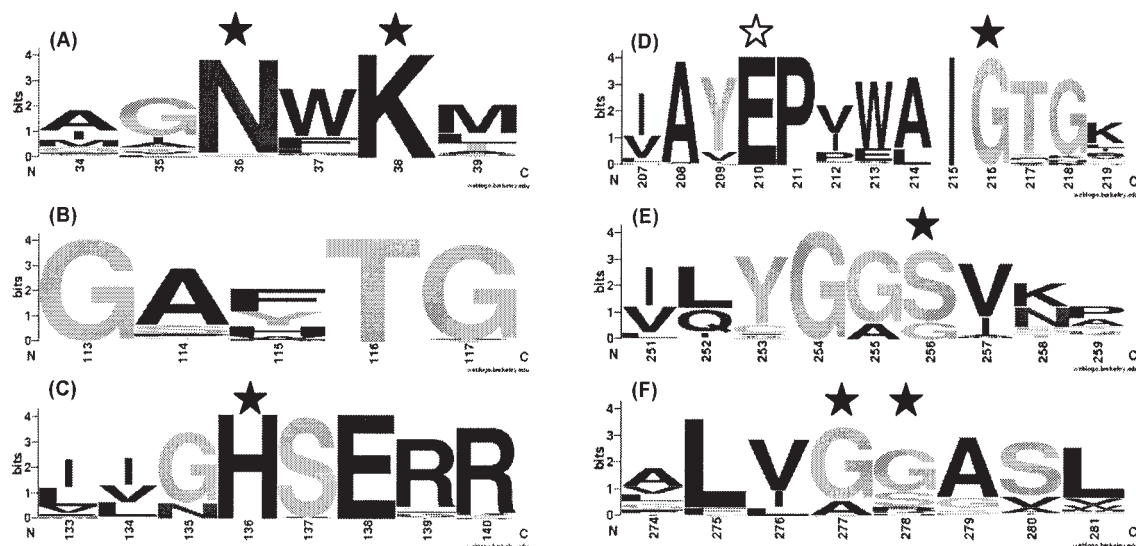


Fig. 4. Sequence logos of the 6 TIM PMs (A–F) visually highlight their conserved nature. PMs are defined as all overlapping windows scoring past the PSZ threshold. Residues are labeled corresponding to the alignment numbering scheme. Stars indicate positions highlighted in Figure 3 (E210 is the catalytic residue). Sequence logos are generated using the UC Berkeley server at <http://weblogo.berkeley.edu>.

the phosphate of the substrate, and His94 is hydrogen-bonded to both the substrate carbonyl and alcohol oxygens. Because of their strictly conserved nature, these positions do not affect the overall phylogeny. However, these (and the other) conserved residues are identified within PMs, because several proximal (in sequence) positions do. This point can be exemplified by analyzing the 4 variable positions in contact with the substrate. Figure 4 shows that position 256 (now using alignment numbering) is a conserved serine 80% of the time; the remaining 20% is a subfamily-conserved glycine. Gly256 is also conserved 80% of the time, whereas (all but one of) the remaining 20% is a subfamily-conserved alanine. The subfamily defined by Gly256 is the same as that defined by Ala277. Position 278 is also a conserved glycine (67% conservation). Most of the differences at position 278 correspond to the same subfamily defined by Gly256 and Ala 277. However, 4 of the differences at position 278 (a subfamily-conserved arginine) correspond to a small subfamily that include other positions in the alignment. Taken together, this brief and incomplete description encapsulates how conserved-subfamily variations can reproduce the overall tree. More comprehensive analysis of all PM positions reveals further subfamily-conserved residues and leads to locally conserved groups of (functionally conserved) mutations that reproduce the overall familial phylogeny.

In addition to TIM, we also apply our method to 14 other protein examples, representing a structurally and functionally heterogeneous data set (for a sampling see Fig. 2; representations of the remainder of the data set are provided in the Supplementary Material). As expected, functional sites defined by surface loops are consistently identified as PMs. In the enolase example, PMs correspond to active site loops contributed by both the C-terminal (TIM-barrel) and N-terminal domains.³⁸ CuZnSOD PMs correspond to loop regions mediating the shape of the

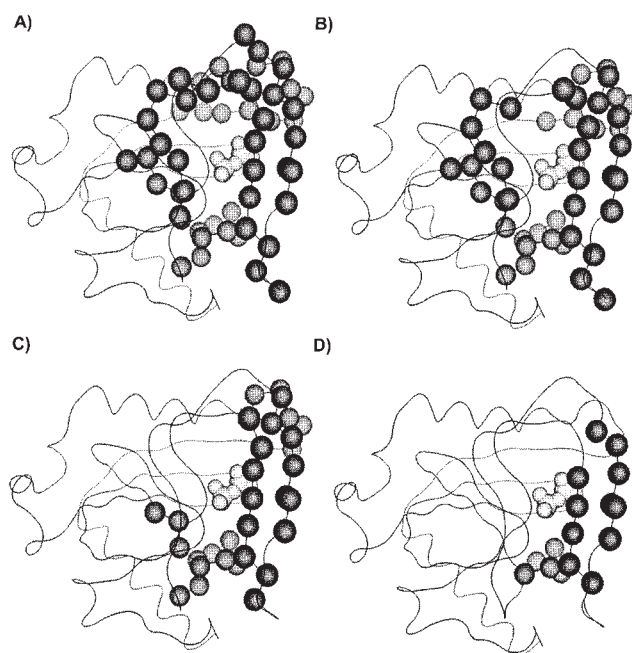


Fig. 5. In general, increasingly stringent PSZ thresholds focus the identified PMs toward the active site. This point is generally conserved on the entire data set and is visually demonstrated here on inorganic pyrophosphatase. Dark gray spheres indicate α -carbons of PMs observed past the (A) -1.2 , (B) -1.5 , (C) -1.8 , and (D) -2.2 thresholds. The pyrophosphate substrate is colored light gray.

cationic funnel that conserves the high enzyme–superoxide anion encounter rate.²⁰ PMs are identified in the TATA box–binding protein family, two of which correspond to the DNA sequence specifying “stirrup” regions³⁹; 2 of the remaining 3 correspond to the TFIIB binding site.⁴⁰ Additionally, PMs identified in the inorganic pyrophosphatase⁴¹ and *N*-acetylglucosamine-6-phosphate deacetylase⁴² fami-

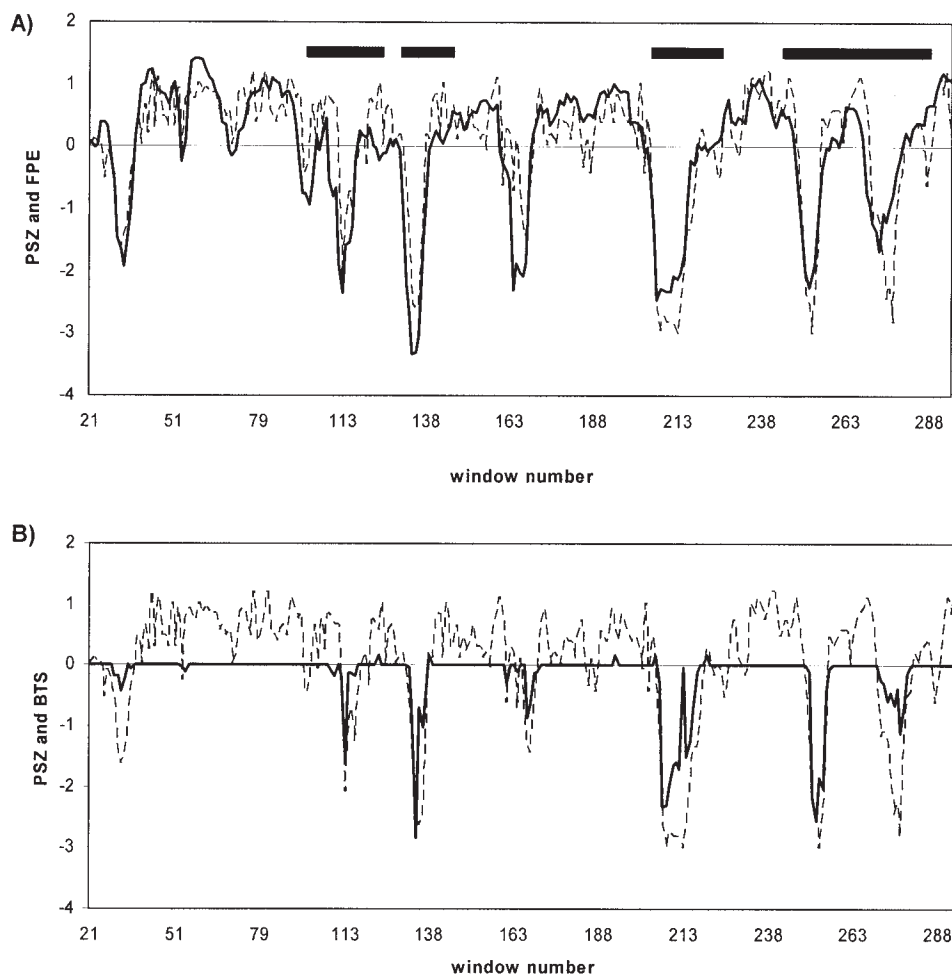


Fig. 6. (A) TIM PSZ (dashed line) and FPE (solid line) also recast as z scores versus window number. This plot demonstrates the sequence correspondence of phylogenetic and traditional motifs. MEME identified motifs are labeled in bold (at the top) and correspond with low FPEs. Similar plots are observed for the remainder of our diverse data set. (B) TIM PSZs (dashed line) and bootstrap tree stability (BTS; solid line) versus window number. This plot demonstrates that the PM window trees are more stable than non-PM windows. Tree stability is calculated as described in the Methods section. Tree stability ranges from 0 to 100. For comparison ease, the baseline of the tree stability values has been shifted to correspond to PSZ = 0; further, tree stability values have been inverted and normalized against the most extreme PSZ.

lies correspond to active site loop regions. In each of the above examples, PMs are structurally clustered, despite little overall proximity in sequence.

As a general rule, making the z-score threshold more stringent focuses the PMs toward the catalytic region. Figure 5 clearly demonstrates this point in the inorganic pyrophosphatase example. Z-score thresholds of -1.5 , -1.8 , and -2.2 result in 5, 4, and 3 identified PMs, respectively. The same number of PMs is identified for the thresholds of -1.2 and -1.5 . However, motif size is reduced as the threshold is made stricter (average reduction is 1.8 residues). The more relaxed thresholds result in the identification of several regions not likely to be functional. Predictions made at the -1.8 threshold are consistent with literature results. All 3 residues salt-bridged (Lys29, Arg43, and Lys142) to the pyrophosphate substrate (in the *Escherichia coli* structure, PDB ID: 1I6T⁴¹) correspond to PMs. All 3 have been shown to be critical to

function.⁴³ Each salt bridge partner residue corresponds to a different PM. The fourth PM corresponds to the conserved D-X-D-X-X-D sequence pattern⁴⁴; the 3 Asp residues bind catalytically requisite divalent metal ions.⁴¹ Making the threshold more stringent eliminates the PM containing Lys142, suggesting that -1.8 represents a good balance.

Encouragingly, application of our approach to proteins whose active sites are not defined by plastic loop regions also predicts key functionality. PMs predict regions corresponding to active site clefts, partially buried functional sites, sites from multiple domains, and functionally linked sites within a multiprotein complex. PMs identified in the myoglobin,⁴⁵ glycerol kinase,⁴⁶ and glutamate dehydrogenase⁴⁷ families all correspond to their respective active site clefts. In larger protein structures, PMs are identified in each domain. In the enolase, alcohol dehydrogenase⁴⁸ and glyceraldehyde-3-phosphate dehydrogenase⁴⁹ ex-

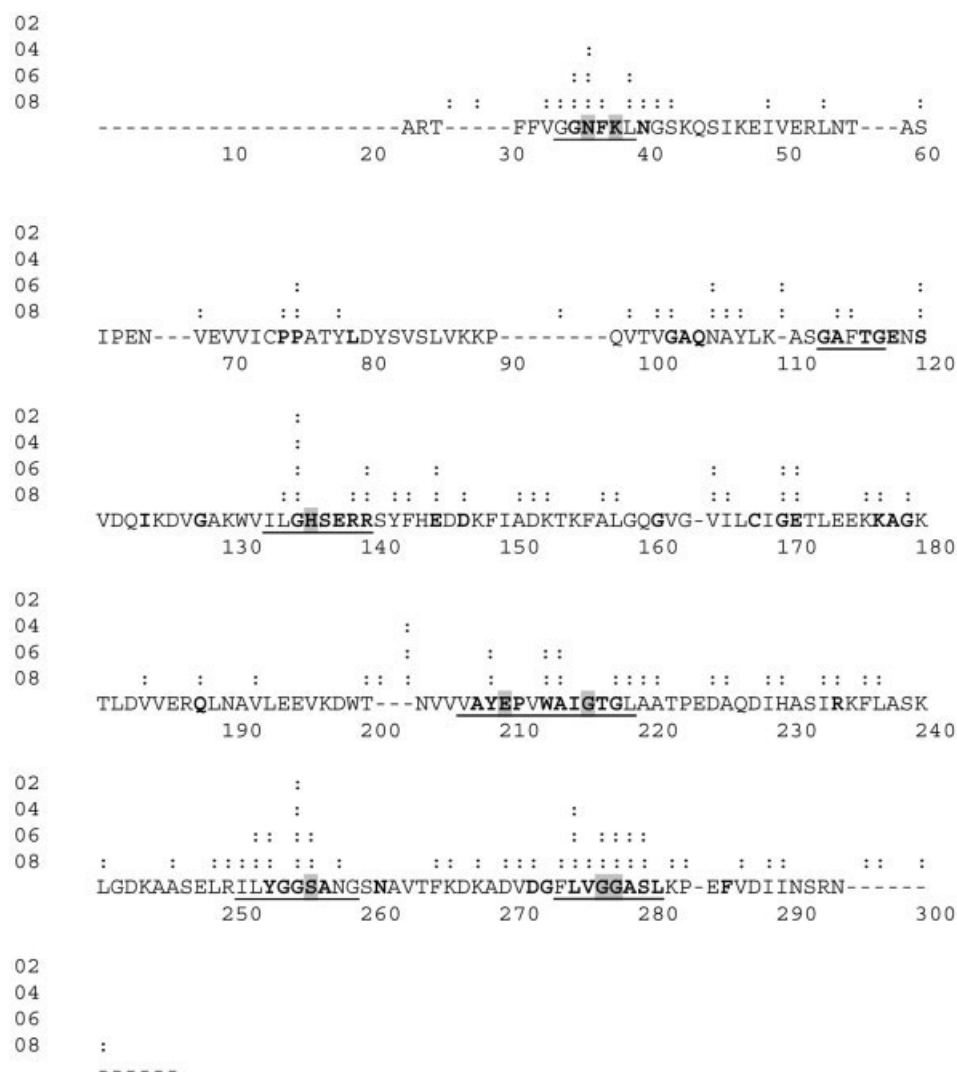


Fig. 7. Comparison of ET predictions versus PMs for TIM. PMs are defined as all overlapping windows scoring past the PSZ threshold. ET positions at particular tree cut levels are indicated by colons above each position. PMs are underlined; positions directly interacting (H-bond or salt bridge) with the substrate are shaded in gray; alignment positions conserved more than 60% of the time are highlighted in bold. The two methods have similar results, but (as highlighted in Fig. 8) the PM predictions are more structurally clustered around known functional features.

amples, active site regions from both domains are identified as PMs. PMs also successfully predict partially buried regions mediating prosthetic group specificity. They are structurally clustered around the heme of cytochrome P450⁵⁰ and approximate the perceived electron transport path within succinate dehydrogenase (Complex II of electron transport), from heme to Fe/S ($\times 3$) to flavin adenine dinucleotide (FAD).⁵¹ Succinate dehydrogenase is composed of 3 nonhomologous proteins (the FAD-binding protein, the Fe/S cluster-binding protein, and the heme-binding/lipid bilayer anchor). Because function is dependent upon each member of the complex, it is not unexpected that PMs from the 3 unique families are structurally linked. Taken together, these results encourage us that PMs represent a robust functional site prediction scheme.

The Conserved Nature of Phylogenetic Motifs

As highlighted above, PMs appear to be generally conserved in sequence. This qualitative observation is apparent from a cursory examination of sequence logos (Fig. 3). In order to quantify this observation, we calculate FPEs to determine the probability of randomly encountering each motif. Comparison of calculated FPEs with phylogenetic similarity indicates that PMs are generally motifs in the traditional sense. Windows with lower FPEs are less likely to be encountered by chance and are more conserved. In relation to windows of the same size, none of the observed TIM PMs are significantly likely to occur randomly. Figures 1(B) and 6(A) clearly indicate that traditionally and PMs correspond to each other consistently (correlation coefficient = 0.80). The observed correlation between the 2

methods falls off drastically with increasing window width. The molecular explanation for this result is (at least partially) related to the importance of these positions in defining substrate specificity. Figures 3 and 4 indicate that all positions directly interacting with the substrate through side-chain interactions are strictly (or nearly so) conserved (e.g., the catalytic Glu164). Further, residues interacting with the substrate through backbone H-bonds are conserved a majority of the time. Clearly, conservation of these positions is critical to substrate specificity. Presumably, the remaining conserved PM positions maintain tertiary and/or active site structure. Several plastic positions are interspersed between conserved residues. However, plasticity is limited as evolution conserves residue identity within subfamilies. Very few PM positions are totally variable.

Based on our combined results, we generally conclude that PMs are a subset of motif space. There are instances that deviate from this trend [i.e., cytochrome P450, succinate dehydrogenase FAD-binding protein, and (to a lesser extent) succinate dehydrogenase Fe/S-binding protein (see Supplementary Material)]. Some PMs from these examples are not significantly conserved in sequence, yet are identified as PMs because mutations between subfamilies are still conserved. As with the average PM, these atypical observations are still structurally clustered around key functional sites. This fortuitous result suggests that PMs might be able to functionally annotate sequence regions that traditional motif-based methods would not. However, we do note a general tendency for the more specific PMs to better correlate to the prosthetic group binding regions, which is consistent with our earlier findings.

Comparing differences between phylogenetic and traditional motifs is quite telling. In the case of TIM, 6 of 7 traditional motifs are also identified as PMs. In one case, the corresponding pair structurally maps to the dimer interface. All of the remaining motifs (phylogenetic and traditional) are near the active site region. Each of the 5 PMs corresponding to traditional motifs contributes to substrate specificity through hydrogen bonding and/or ionic interactions. Whereas the 5 active site PMs at least partially correspond to surface loop segments, the remaining traditional motif is more buried, corresponding to a barrel β -strand. Furthermore, in spite of being near the active site, the remaining traditional motif never directly interacts with the substrate. Across our exemplar data set, there are some instances without significant differences between the 2 methods (e.g., glyceraldehydes-3-phosphate dehydrogenase and glycerol kinase), and there are a few cases with dramatic differences (i.e., cytochrome P450 and succinate dehydrogenase FAD-binding domain). In most cases, however, differences are subtle, yet significant. For example, PMs not identified as traditional motifs predict one of the DNA sequence specifying stirrups of TATA box-binding protein, a buried region of myoglobin directly interacting with the heme, and several enolase-substrate contacts. A complete case-by-case structural comparison of the motif approaches is provided in the Supplementary Material.

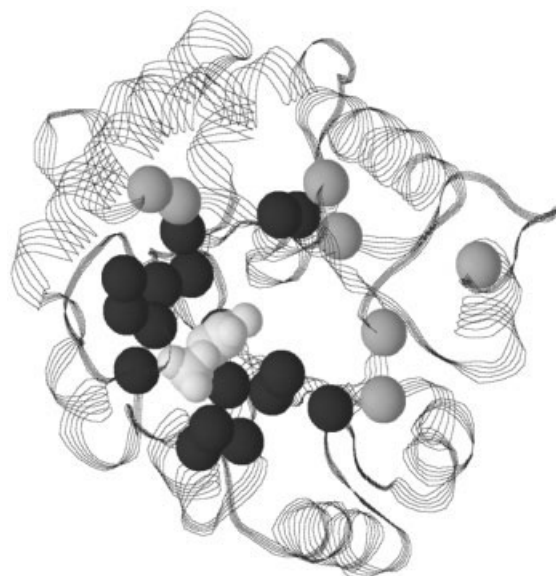


Fig. 8. The α -carbons of TIM ET positions scoring past the cut level = 6 are highlighted. ET positions also identified as PMs are colored dark gray, whereas the remaining positions are colored light gray. The substrate analog is also shown (at the top) in light gray. The orientation is the same as in Figure 2(A).

Because of the relative simplicity of our false-positive calculation, MEME³⁴ is used to confirm the conclusions drawn. MEME motifs reinforce our earlier results, confirming their conserved nature. As described above, 7 conserved regions are identified in the TIM family [Fig. 6(A)]. MEME-identified TIM motifs are annotated in Figure 6(A). In each case, the 5 identified MEME motifs (which is the search limit) directly correspond to the lowest FPE values. (The 2 remaining conserved regions would likely be identified by MEME had the search limit not been set to 5.) Across the complete data set, MEME results consistently correspond to low FPEs.

Comparison to Evolutionary Trace Predictions

The ET technique has proven to be a particularly powerful protein functional site prediction technique.^{5,11–18} The method begins with a sequence alignment and an accompanying phylogenetic tree. The ET method ranks each alignment position based on the minimum number of tree branches required to keep that position invariant within each outgroup. Frequently, structural clusters of ET positions, which are statistically significant,¹⁷ are used to fine-tune functional site predictions. Here, PM predictions are compared to ET results using the same underlying alignment. Figure 7 compares the TIM results from the two techniques. As expected, PM regions are significantly populated by trace residues. All ET positions (except one) scoring at or better than cut level = 4 correspond to PM regions. (The one exception is a position that is mostly gaps.) Several trace residues scoring at more relaxed cut levels occur outside PM regions. In total, 26 ET positions scoring at or better than cut level = 6 are identified; 17 (65%) of these correspond to PM positions.

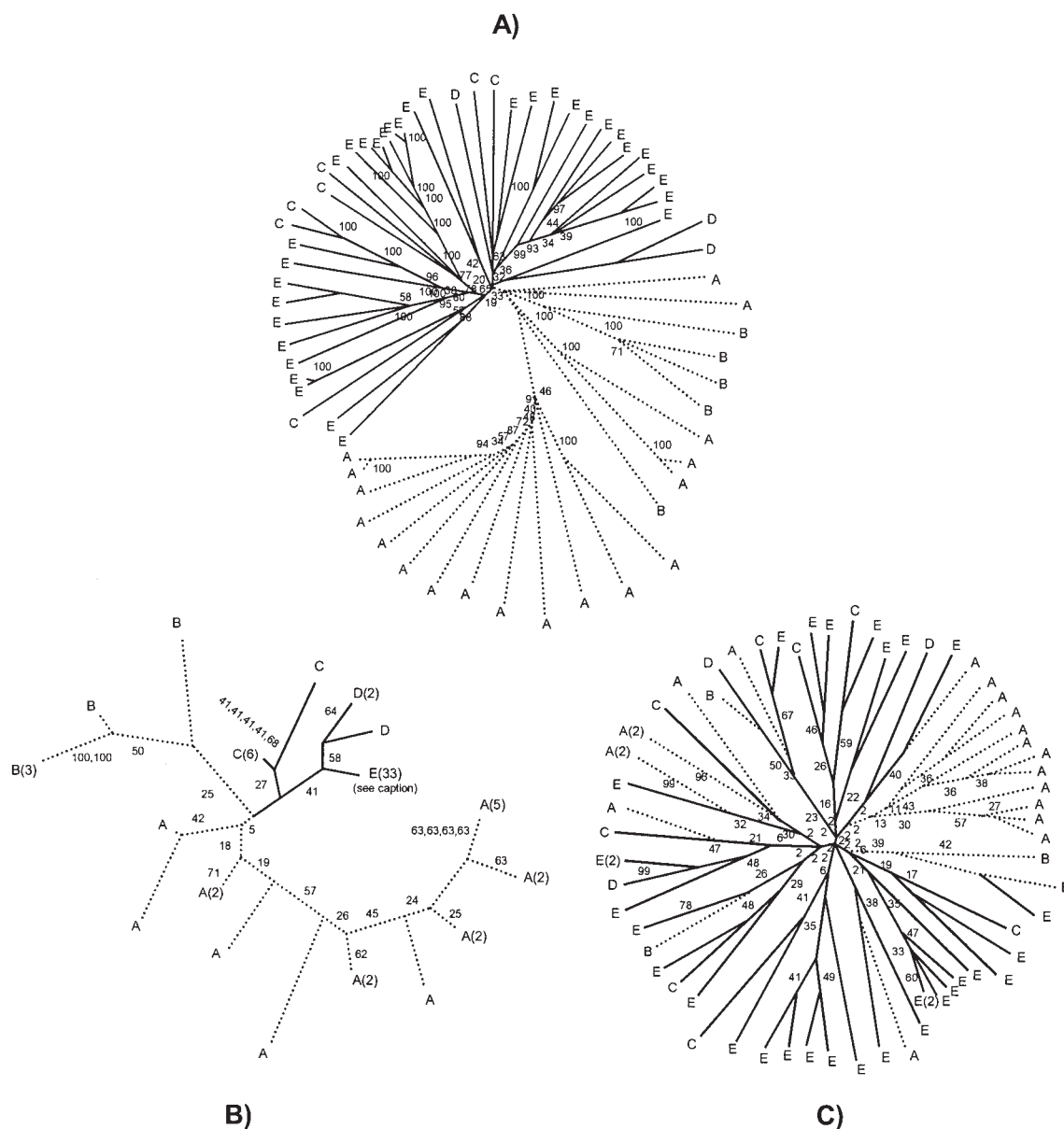


Fig. 9. Complete TIM (A) sequence window 276 and (B) sequence window 61 phylogenetic trees. (B) is representative of trees with high phylogenetic similarity, whereas (C) is representative of low similarity. Due to reduced information content, the window trees generally have shorter branch lengths and are less overall branched. In the window trees, branch label numbers in parentheses indicate the number of identical sequences represented. (B) is subdivided into 5 subfamilies, which are largely conserved in the whole family tree. In fact, (A) and (B) can be divided into two identical subfamilies (indicated by solid and dashed lines). Subfamily conservation between (A) and (C) is greatly reduced. Bootstrap scores indicate the number of times each outgroup is conserved in the 100 resamples. The PM window tree is more stable than its counterpart. Bootstrap scores for the 31 identical sequences in subfamily E (B) are 87, 79, 63, 54, 48, 47, 47, 46, 44, 44, 44, 44, 44, 43, 43, 42, 42, 41, 40, 40, 37, 25, 25, 24, 24, 23, 23, and 23.

Most identified trace residues are expected to be critically related to protein function. Nevertheless, we observe a clear tendency for ET positions also implicated as PMs to cluster more tightly around known catalytic regions (Fig. 8). For example, the average distance between TIM α -carbons and the substrate geometrical center for trace residues occurring within PMs is 8.7 Å, whereas the average distance for non-PM trace residues is 16.0 Å. (Note: gaps in the 7TIM sequence occur for 2 non-PM trace residues, meaning that those distances could not be computed.) A complete set of PM versus ET predictions is provided in the Supplementary

Material. The general conclusions described here are conserved across the remainder of the data set, suggesting that PM predictions are generally more structurally clustered around known functionality, especially substrate-binding regions. In essence, PMs are focusing trace residues toward structural epitopes. We note that the use of ET structural clusters¹⁷ should have the same effect. However, we are purposely avoiding all structural considerations in our predictions here. More stringent cut levels do focus the trace residues in the same way; however, focusing comes at the expense of several key interaction predictions.

Tree Comparisons and Bootstrap Analysis

Figure 9 compares the complete TIM tree to two window trees. One is taken from a PM window [alignment positions 276–280, see also Fig. 3(F)] and the other has little overall phylogenetic similarity (positions 61–65). Because of reduced information content, the window trees have shorter branch lengths and are less branched than their whole alignment counterpart. Despite these global differences, most PM outgroups are conserved within the whole alignment tree. In fact, they can be divided into 2 identical subfamilies. Outgroup conservation between the second window and whole family tree is greatly reduced.

Based on the limited information content, calculating trees on small sequence windows is potentially a concern. To gauge tree stability, bootstrap analysis is employed. A cursory examination of the window tree bootstrap results indicates that many of the interior branches are suspect, yet most of the outer subfamilies are significant. Additionally, comparison of PM tree against the non-PM tree suggests that the PM tree is much more stable. Figure 7(B) quantifies this result and shows that it is conserved throughout all TIM windows (results for the remaining protein examples are provided in the Supplementary Material). In fact, the best global bootstrap stability scores consistently correspond to PM windows. This result indicates that PM tree topologies are less likely to change than non-PM windows. These results satisfy us that, at the very least, the phylogenetic trees are meaningful. Future work will seek to develop methods to increase the statistical significance of the window trees.

CONCLUSIONS

We report that PMs, sequence regions conserving the overall familial phylogeny, frequently correspond to key protein functional sites, including regions defined by plastic surface loops, active site clefts, and partially buried regions. In general, PMs are also conserved in sequence, and can be considered motifs in the traditional sense. However, there are instances where PMs are not overall conserved in sequence (i.e., cytochrome P450, succinate dehydrogenase Fe/S-binding protein, and succinate dehydrogenase FAD-binding protein). PMs from these fortuitous examples still correspond to key functional sites (they are all structurally clustered around key prosthetic groups), which suggests that PMs can functionally annotate regions that traditional motif-based methods would not. PM results are systematically compared to those from the ET technique. The results of the 2 methods are consistent, but PM predictions tend to be more structurally clustered around known functional sites. Structural clusters of trace residues could be used to filter ET predictions, resulting in the same effect. However, structure is purposely avoided here, because high-throughput structural initiatives will be unable to keep pace with genome sequencing for some time. All together, these initial results encourage us that PMs should supplement current sequence–function annotation strategies. Future testing on larger data sets will determine the robustness of these initial conclusions.

ACKNOWLEDGMENTS

We thank Dr. Shankar Subramaniam (University of California, San Diego) for several key suggestions and review of the manuscript. Dr. Don Jacobs (California State University, Northridge) and Dr. Patrick Mobley (California State Polytechnic University, Pomona) are also acknowledged for proofreading an early version of the manuscript.

REFERENCES

1. Attwood TK. The quest to deduce protein function from sequence: the role of pattern databases. *Int J Biochem Cell Biol* 2000;32:139–155.
2. Liu AH, Zhang X, Stolovitzky GA, Califano A, Firestein SJ. Motif-based construction of a functional map for mammalian olfactory receptors. *Genomics* 2003;81:443–456.
3. Puntervoll P, Linding R, Gemund C, Chabanis-Davidson S, Matningsdal M, Cameron S, Martin DM, Ausiello G, Brannetti B, Costantini A, Ferre F, Maselli V, Via A, Cesareni G, Diella F, Superti-Furga G, Wyrwicz L, Ramu C, McGuigan C, Gudavalli R, Letunic I, Bork P, Rychlewski L, Kuster B, Helmer-Citterich M, Hunter WN, Aasland R, Gibson TJ. ELM server: A new resource for investigating short functional sites in modular eukaryotic proteins. *Nucleic Acids Res* 2003;31:3625–3630.
4. Sjolander K. Phylogenomic inference of protein molecular function: advances and challenges. *Bioinformatics* 2004;20:170–179.
5. Lichtarge O, Bourne HR, Cohen FE. An evolutionary trace method defines binding surfaces common to protein families. *J Mol Biol* 1996;257:342–358.
6. Bickel PJ, Kechris KJ, Spector PC, Wedemayer GJ, Glazer AN. Finding important sites in protein sequences. *Proc Natl Acad Sci USA* 2002;99:14764–14771.
7. Casari G, Sander C, Valencia A. A method to predict functional residues in proteins. *Nat Struct Biol* 1995;2:171–178.
8. del Sol MA, Pazos F, Valencia A. Automatic methods for predicting functionally important residues. *J Mol Biol* 2003;326:1289–1302.
9. Landgraf R, Xenarios I, Eisenberg D. Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. *J Mol Biol* 2001;307:1487–1502.
10. Innis CA, Shi J, Blundell TL. Evolutionary trace analysis of TGF-beta and related growth factors: implications for site-directed mutagenesis. *Protein Eng* 2000;13:839–847.
11. Lichtarge O, Yamamoto KR, Cohen FE. Identification of functional surfaces of the zinc binding domains of intracellular receptors. *J Mol Biol* 1997;274:325–337.
12. Lichtarge O, Sowa ME. Evolutionary predictions of binding surfaces and interactions. *Curr Opin Struct Biol* 2002;12:21–27.
13. Lichtarge O, Sowa ME, Philippi A. Evolutionary traces of functional surfaces along G protein signaling pathway. *Methods Enzymol* 2002;344:536–556.
14. Madabushi S, Gross AK, Philippi A, Meng EC, Wensel TG, Lichtarge O. Evolutionary trace of G protein-coupled receptors reveals clusters of residues that determine global and class-specific functions. *J Biol Chem* 2004;279:8126–8132.
15. Mihalek I, Res I, Yao H, Lichtarge O. Combining inference from evolution and geometric probability in protein structure evaluation. *J Mol Biol* 2003;331:263–279.
16. Yao H, Kristensen DM, Mihalek I, Sowa ME, Shaw C, Kimmel M, Kaviraki L, Lichtarge O. An accurate, sensitive, and scalable method to identify functional sites in protein structures. *J Mol Biol* 2003;326:255–261.
17. Madabushi S, Yao H, Marsh M, Kristensen DM, Philippi A, Sowa ME, Lichtarge O. Structural clusters of evolutionary trace residues are statistically significant and common in proteins. *J Mol Biol* 2002;316:139–154.
18. Lichtarge O, Yao H, Kristensen DM, Madabushi S, Mihalek I. Accurate and scalable identification of functional sites by evolutionary tracing. *J Struct Funct Genomics* 2003;4:159–166.
19. Livesay D, Jambeck P, Rojnuckarin A, Subramaniam S. Conservation of electrostatic properties within enzyme families and superfamilies. *Biochemistry* 2003;42:3464–3473.
20. Bordo D, Djinovic K, Bolognesi M. Conserved patterns in the

- Cu,Zn superoxide dismutase family. *J Mol Biol* 1994;238:366–386.
21. Elcock AH. Prediction of functionally important residues based solely on the computed energetics of protein structure. *J Mol Biol* 2001;312:885–896.
 22. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* 1995;247:536–540.
 23. Mihalek I, Res I, Lichtarge O. A family of evolution–entropy hybrid methods for ranking protein residues by importance. *J Mol Biol* 2004;336:1265–1282.
 24. Penny D, Hendy M. The use of tree comparison metrics. *Systematic Zoology* 1985;34:75–82.
 25. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res* 2003;31:365–370.
 26. Tatusov RL, Galperin MY, Natale DA, Koonin EV. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res* 2000;28:33–36.
 27. Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND, Koonin EV. The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res* 2001;29:22–28.
 28. Thompson JD, Higgins DG, 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* 1994;22:4673–4680.
 29. Notredame C, Higgins DG, Heringa J. T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J Mol Biol* 2000;302:205–217.
 30. Felsenstein J. PHYLIP—Phylogeny Inference Package (version 3.2). *Cladistics* 1989;5:164–166.
 31. Kuhner MK, Felsenstein J. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol Biol Evol* 1994;11:459–468.
 32. Felsenstein J. *Inferring phylogenies*. 1st ed. Sunderland, MA: Sinauer Associates; 2004.
 33. Nevill-Manning CG, Wu TD, Brutlag DL. Highly specific protein sequence motifs for genome analysis. *Proc Natl Acad Sci USA* 1998;95:5865–5871.
 34. Bailey TL, Gribskov M. Methods and statistics for combining motif match scores. *J Comput Biol* 1998;5:211–221.
 35. La D, Silver M, Edgar RC, Livesay DR. Using motif-based methods in multiple genome analyses: a case study comparing orthologous mesophilic and thermophilic proteins. *Biochemistry* 2003;42:8988–8998.
 36. Wierenga RK. The TIM-barrel fold: a versatile framework for efficient enzymes. *FEBS Lett* 2001;492:193–198.
 37. Wallace AC, Laskowski RA, Thornton JM. LIGPLOT: a program to generate schematic diagrams of protein–ligand interactions. *Protein Eng* 1995;8:127–134.
 38. Babbitt PC, Hasson MS, Wedekind JE, Palmer DR, Barrett WC, Reed GH, Rayment I, Ringe D, Kenyon GL, Gerlt JA. The enolase superfamily: a general strategy for enzyme-catalyzed abstraction of the alpha-protons of carboxylic acids. *Biochemistry* 1996;35:16489–16501.
 39. DeDecker BS, O'Brien R, Fleming PJ, Geiger JH, Jackson SP, Sigler PB. The crystal structure of a hyperthermophilic archaeal TATA-box binding protein. *J Mol Biol* 1996;264:1072–1084.
 40. Juo ZS, Kassavetis GA, Wang J, Geiduschek EP, Sigler PB. Crystal structure of a transcription factor IIIB core interface ternary complex. *Nature* 2003;422:534–539.
 41. Samygina VR, Popov AN, Rodina EV, Vorobyeva NN, Lamzin VS, Polyakov KM, Kurilova SA, Nazarova TI, Ayaeva SM. The structures of *Escherichia coli* inorganic pyrophosphatase complexed with Ca(2+) or CaPP(i) at atomic resolution and their mechanistic implications. *J Mol Biol* 2001;314:633–645.
 42. Ferreira FM, Mendoza-Hernandez G, Calcagno ML, Minauro F, Delboni LF, Oliva G. Crystallization and preliminary crystallographic analysis of *N*-acetylglucosamine 6-phosphate deacetylase from *Escherichia coli*. *Acta Crystallogr D Biol Crystallogr* 2000;56:670–672.
 43. Sivula T, Salminen A, Parfenyev AN, Pohjanjoki P, Goldman A, Cooperman BS, Baykov AA, Lahti R. Evolutionary aspects of inorganic pyrophosphatase. *FEBS Lett* 1999;454:75–80.
 44. Cooperman BS, Baykov AA, Lahti R. Evolutionary conservation of the active site of soluble inorganic pyrophosphatase. *Trends Biochem Sci* 1992;17:262–266.
 45. Bolognesi M, Onesti S, Gatti G, Coda A, Ascenzi P, Brunori M. *Aplysia limacina* myoglobin: crystallographic analysis at 1.6 Å resolution. *J Mol Biol* 1989;205:529–544.
 46. Ormo M, Bystrom CE, Remington SJ. Crystal structure of a complex of *Escherichia coli* glycerol kinase and an allosteric effector fructose 1,6-bisphosphate. *Biochemistry* 1998;37:16565–16572.
 47. Smith TJ, Peterson PE, Schmidt T, Fang J, Stanley CA. Structures of bovine glutamate dehydrogenase complexes elucidate the mechanism of purine regulation. *J Mol Biol* 2001;307:707–720.
 48. Korkhin Y, Kalb G, Peretz M, Bogin O, Burstein Y, Frolov F. NADP-dependent bacterial alcohol dehydrogenases: crystal structure, cofactor-binding and cofactor specificity of the ADHs of *Clostridium beijerinckii* and *Thermoanaerobacter brockii*. *J Mol Biol* 1998;278:967–981.
 49. Duee E, Olivier-Deyris L, Fanchon E, Corbier C, Branlant G, Dideberg O. Comparison of the structures of wild-type and a N313T mutant of *Escherichia coli* glyceraldehyde 3-phosphate dehydrogenases: implication for NAD binding and cooperativity. *J Mol Biol* 1996;257:814–838.
 50. Mowat CG, Leys D, McLean KJ, Rivers SL, Richmond A, Munro AW, Ortiz LM, Alzari PM, Reid GA, Chapman SK, Walkinshaw MD. Crystallization and preliminary crystallographic analysis of a novel cytochrome P450 from *Mycobacterium tuberculosis*. *Acta Crystallogr D Biol Crystallogr* 2002;58:704–705.
 51. Yankovskaya V, Horsefield R, Tornroth S, Luna-Chavez C, Miyoshi H, Leger C, Byrne B, Cecchini G, Iwata S. Architecture of succinate dehydrogenase and reactive oxygen species generation. *Science* 2003;299:700–704.