doi:10.1016/S0022-2836(02)01336-0 *J. Mol. Biol.* (2003) **326**, 255–261

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

An Accurate, Sensitive, and Scalable Method to Identify Functional Sites in Protein Structures

Hui Yao[1,2]†, David M. Kristensen[1,2,3]†, Ivana Mihalek[1] Mathew E. Sowa[2,4], Chad Shaw[1], Marek Kimmel[3,5], Lydia Kavraki[3,6,7] and Olivier Lichtarge[1,2,3,8,9]*

[1]*Department of Molecular and Human Genetics Baylor College of Medicine One Baylor Plaza T921 Houston, TX 77030, USA*

[2]*Program in Structural and Computational Biology and Molecular Biophysics Baylor College of Medicine One Baylor Plaza T921 Houston, TX 77030, USA*

[3]*W.M. Keck Center for Computational Biology Houston, TX 77030, USA*

[4]*Verna and Marrs McLean Department of Biochemistry and Molecular Biology Baylor College of Medicine One Baylor Plaza T921 Houston, TX 77030, USA*

[5]*Department of Statistics Rice University, Houston TX 77030, USA*

[6]*Department of Computer Science, Rice University Houston, TX 77030, USA*

[7]*Department of Bioengineering Rice University, Houston TX 77030, USA*

[8]*Program in Developmental Biology, Baylor College of Medicine, One Baylor Plaza T921, Houston TX 77030, USA*

[9]*Human Genome Sequencing Center, Baylor College of Medicine, One Baylor Plaza T921, Houston TX 77030, USA*

**Corresponding author*

Functional sites determine the activity and interactions of proteins and as such constitute the targets of most drugs. However, the exponential growth of sequence and structure data far exceeds the ability of experimental techniques to identify their locations and key amino acids. To fill this gap we developed a computational Evolutionary Trace method that ranks the evolutionary importance of amino acids in protein sequences. Studies show that the best-ranked residues form fewer and larger structural clusters than expected by chance and overlap with functional sites, but until now the significance of this overlap has remained qualitative. Here, we use 86 diverse protein structures, including 20 determined by the structural genomics initiative, to show that this overlap is a recurrent and statistically significant feature. An automated ET correctly identifies seven of ten functional sites by the least favorable statistical measure, and nine of ten by the most favorable one. These results quantitatively demonstrate that a large fraction of functional sites in the proteome may be accurately identified from sequence and structure. This should help focus structure–function studies, rational drug design, protein engineering, and functional annotation to the relevant regions of a protein.

Keywords: molecular recognition; protein evolution; protein interactions; structural motif; drug target

† These authors contributed equally to this work.

Abbreviations used: SGI, structural genomics initiative; ET, Evolutionary Trace.

E-mail address of the corresponding author: lichtarge@bcm.tmc.edu

Introduction

The structural genomics initiative (SGI) aims to solve 10,000 protein structures in this decade. The biological functions of many of these proteins will not be known but one approach to uncover their roles is to identify their functional sites. This can focus experiments on biologically relevant regions, help define the molecular basis of function,[1] and reveal instances of molecular mimicry among functional sites.[2–4] On the other hand, once function is known, functional site identification is also essential for drug design.[5–8] Possible methods to identify these sites were reviewed recently[9] and include the use of block-like motifs in protein sequences;[10] physicochemical descriptors of surface residues;[7] variable rates of divergence among phylogenetic branches;[11] and energetic pathways.[12]

The approach taken by the Evolutionary Trace[13] (ET) ranks the evolutionary importance of residues in a protein family by correlating their variations with evolutionary divergences. Many studies show consistently that top-ranked residues occur in diverse protein families and that they cluster at functional sites.[13–17] This approach[18–21] and variations upon it[22–24] have now been used by others to identify many protein functional sites and to study protein–protein docking.[25] Targeted mutagenesis also recently demonstrated ET's accurate prediction of functional specificity determinants in the regulator of G protein signaling (RGS) family. Top-ranked trace residues at a predicted functional site of the RGS domain were swapped between two family members, thereby conferring one's activity to the other.[1,17]

In light of these detailed studies on specific protein families, it would be useful to understand whether ET can accurately identify functional sites in protein structures on a large scale. We already showed that in a wide variety of protein families the most evolutionarily important trace residues form structural clusters (referred to as trace clusters) that are significantly different than if residues were picked randomly,[26] but the extent and significance of their overlap with functional sites remains unclear. To address this central question, we show here that the overlap is statistically significant and frequently covers half, or more, of the residues in contact with a ligand. To minimize bias, the test set of proteins was only partially chosen by us (and includes the initial structures solved by the SGI), and our results reflect automated trace results as well as those that were manually optimized. We conclude that trace clusters identify functional sites in a sensitive and accurate manner that is scalable for structural proteomics.

Results and Discussion

Our first finding is that manually optimized traces identify structural clusters of trace residues that are statistically significant by both the number of clusters and size of largest cluster statistics in 79 out of the 86 test proteins (92%). Among these 79 proteins, 32 of 37, 29 of 29, and 18 of 20 are from the protein–ligand, enzyme, and SGI datasets, respectively. This proportion of proteins with significant structural clusters is consistent with the 98% previously reported in 46 proteins, where significance was only required in any one of the two statistics.[26]

In order to determine whether trace residues that form significant structural clusters identify the location of functional sites accurately, we focused next on these 79 proteins and sought to measure the statistical significance of the overlap between such trace clusters and functional sites defined by ligand proximity (protein–ligand and SGI sets) or experiment (enzyme set). The first statistic, total connected residues (Figure 1(a)), counts as a success any trace residue in any cluster that overlaps with the functional site. This may be overly generous but it is consistent with functional sites that extend beyond a ligand's immediate vicinity.[9,12] By this measure, there is always at least one rank at which the trace clusters overlap significantly with the functional sites for all 79 proteins (Figure 2(a), red bars). The second statistic, the largest cluster overlap statistic (Figure 1(b)), counts as a success only those residues that are both in the largest cluster and in the functional site. Despite the stringency of this definition, the overlap is significant for at least one rank in 88%, 97%, and 100% of the protein–ligand, enzyme, and SGI data sets, respectively. The third statistic, the average overlap statistic (Figure 1(c)), counts as a success all trace residues that overlap with the functional site, but it penalizes predictions with multiple overlaps due to distinct clusters. Again significance is high: 91%, 86%, and 94% in the same sets. Finally, the hypergeometric distribution statistic (Figure 1(d)) counts as a success only trace residues directly part of the functional site. By this measure, other top-ranked residues that may extend this site or be important for structural and dynamical properties will count as failures, regardless of the cluster they belong to. For this reason, this statistic is likely to systematically underestimate the true significance of overlap and to provide a lower bound for the cluster-based ones. Even so, it indicates significant overlaps for 91%, 79%, and 89% of the protein–ligand, enzyme, and SGI sets, respectively.

A limitation of these results lies in the repeated sampling problem. Namely, since we compute clusters at multiple ranks, we test the null hypothesis multiple times. This inevitable difficulty is mitigated by two facts. First, these tests are not independent. Trace clusters incorporate all trace residues determined at prior ranks, and they are therefore nested. Second, nearly each one of these nested trace clusters overlaps with the functional site far better than expected by chance. For example, the total connected residues statistic shows that the overlap is significant (p-value $\leq 5\%$) for approximately 90% of the ranks that

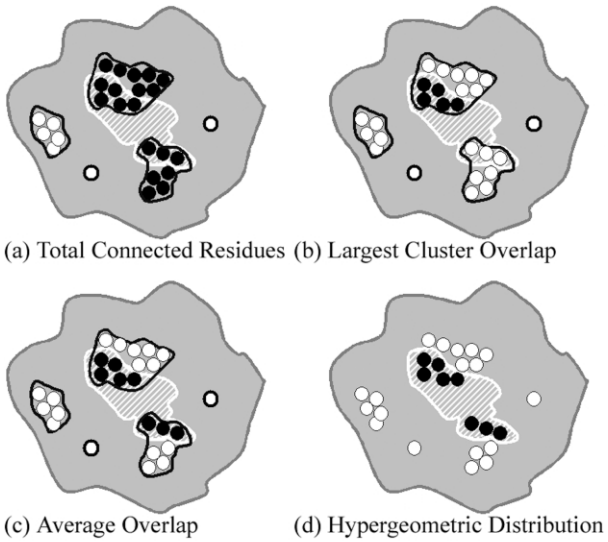

Figure 1. Overlap statistics. The protein is gray and its functional site is delineated by the striped white region. Trace residues are the small circles and they form trace clusters, outlined in black lines. Trace residues that meet the criteria set by the illustrated statistic used to measure the overlap between trace clusters and the functional site are filled in black, or left white otherwise. (a) Total connected residues counts as positive all residues connected to the functional site (in this case, 19) and as negative the rest. (b) Largest cluster overlap only counts as positive the residues shared by the largest cluster and by the functional site (5). (c) Average overlap averages all overlapping trace residues by the number of overlapping trace clusters (8/2). (d) Hypergeometric distribution counts as positive any trace residue in the functional site, regardless of their clustering properties (8).

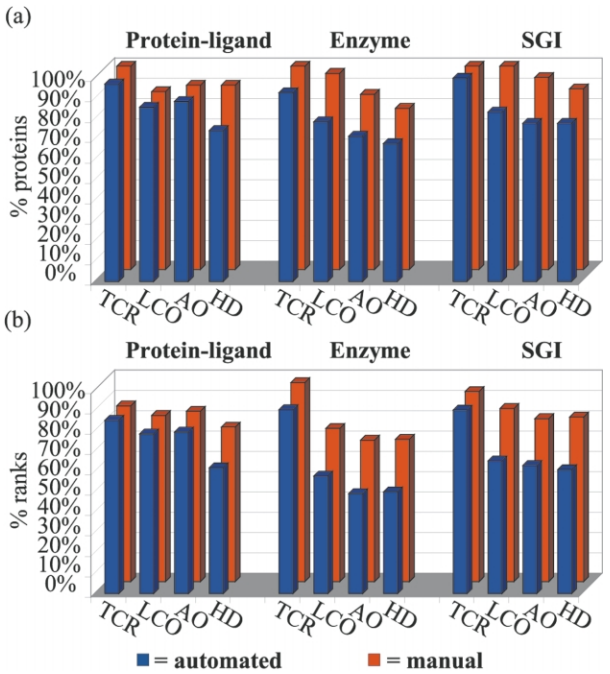

Figure 2. Trace clusters overlap significantly with functional sites. (a) The fraction of manually optimized traces (red) or automated traces (blue) that significantly overlap with functional sites for at least one rank, is shown for each statistic: total connected residues (TCR), largest cluster overlap (LCO), average overlap (AO), and hypergeometric distribution (HD). (b) The fraction of trace ranks with significant clusters that also significantly overlap the functional site. This is averaged for each dataset.

identify non-random trace clusters (86%, 98%, 93%, respectively, for the protein–ligand, enzyme, and SGI sets, Figure 2(b)). This fraction is lower but always remains greater than 70% for the largest cluster overlap statistic (82%, 75%, 85%, for the three sets, respectively); the average overlap statistic (84%, 70%, 80%); and the hypergeometric distribution (76%, 70%, 81%).

These highly non-random overlaps between trace clusters and known functional sites indicate that ET can accurately identify the location of functional sites in many proteins, but its use on a proteomic scale still requires automation. Therefore we tested ET's performance without manual pruning of sequence fragments and of evolutionary outliers from the input. This automated trace identifies statistically significant structural clusters of evolutionarily important residues in 81/86 (94%) of the proteins (35/37, 28/29, and 18/20 in the protein–ligand, enzyme, and SGI datasets, respectively). Of these 81 proteins, the fraction that achieve significant overlap for at least one rank is 97%, 93%, and 100% for the protein–ligand, enzyme, and SGI sets, as measured by the total connected residues statistic (Figure 2(a), blue bars). Thus, the automated trace correctly identifies functional sites in 78 of the 81 proteins (96%) by the most favorable statistic. This success rate is 83% by the largest cluster overlap, 80% by the average overlap, and 73% using the lower-bound estimate of the hypergeometric distribution. Again, these findings hold for the majority of trace ranks with non-random trace clusters. For example, the percentage of significant ranks that have significant overlap, averaged by dataset, is: protein–ligand 85%, enzyme 90%, SGI 90%, using total connected residues; Figure 2(b), blue bars). We expect that these results will grow closer to those obtained manually as better heuristics are devised for sequence selection. Nevertheless among the 86 proteins tested here, the automated ET identifies functional sites in 69% according to the least favorable statistic and in 91% according to the most favorable one, suggesting that ET can be made applicable to the proteome at large.

To illustrate these results, Figure 3(a)–(f) shows manually optimized traces with significant overlap, whereas Figure 3(g)–(l) shows cases that do not reach significance by at least one statistical measure†. In Galectin-3 (Figure 3(a) and (b),

† All results are located at http://imgen.bcm.tmc.edu/molgen/labs/lichtarge/LSET

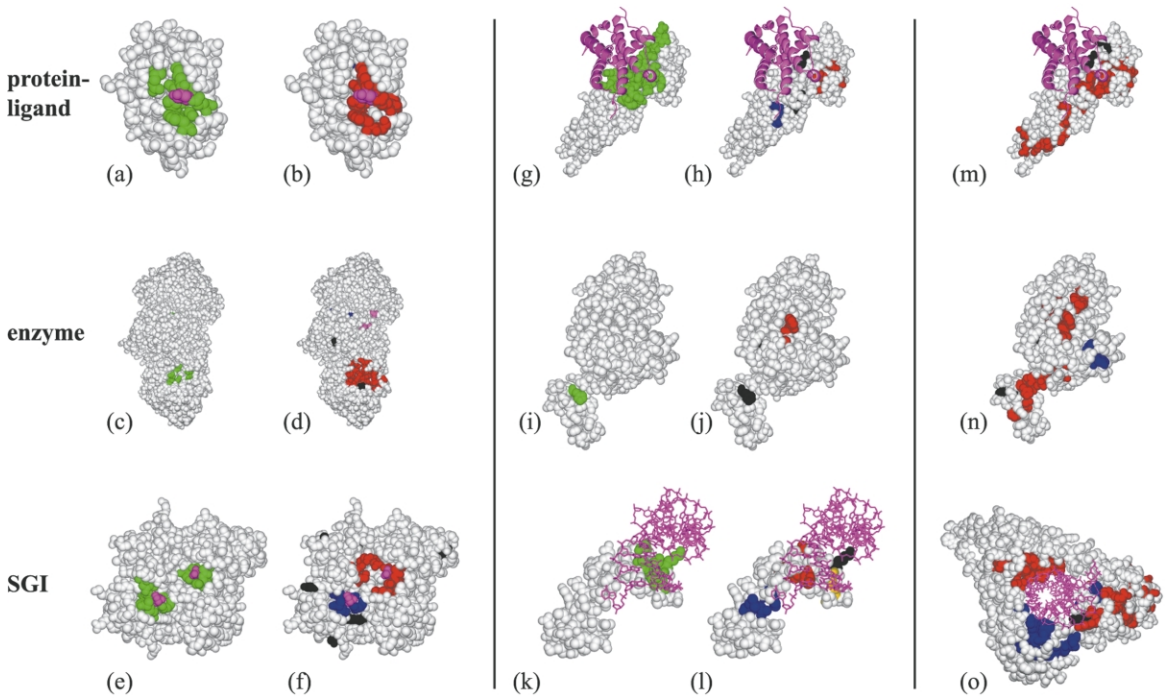

Figure 3. Examples of functional site identification. (a)–(f) Some of the most significant overlaps between trace clusters and functional sites; (g)–(l) others that failed to be significant by at least one statistic; (m)–(o) alternate traces that often establish significance even in these cases (see the text). Small ligands are shown in yellow, large ligands (protein partners and oligonucleotides) in purple, functional sites in green ((a), (c), (e), (g), (i) and (k)), and trace clusters are individually colored ((b), (d), (f), (h), (j), (l), (m), (n) and (o)). The structures are Galectin-3 (PDB code:1A3K) ((a) and (b)) and growth hormone receptor (1A22) ((g), (h) and (m)) from the ligand set; pyruvate phosphate dikinase (1DIK) ((c) and (d)) and xenobiotic acetyltransferase (1XAT) ((i), (j) and (n)) from the enzyme set; and phosphoribosylaminoimidazole-succinocaboxamide synthase (1A48) ((e) and (f)) and transcriptional regulator Nc2 alpha chain (1JFI) ((k), (l) and (o)) from the SGI set (the latter including Nc2 beta chain and a TATA-Box-Binding Protein).

bound to galactose shown in yellow) and phosphoribosylaminoimidazole-succinocarboxamide synthase (Figure 3(e) and (f), bound to sulfate ions shown in yellow), the functional sites are identified by trace clusters (Figure 3(b) and (f)) which overlap significantly with the known ligand-binding sites (Figure 3(a) and (e); green) at all significant ranks by all four statistics. Similarly for pyruvate phosphate dikinase, (Figure 3(c) and (d)) the functional site is identified by a large trace cluster (Figure 3(d); red) which overlaps significantly with the known active site (Figure 3(c); green) at all significant ranks by all four statistics. By contrast, the overlap is much poorer in the other three cases. It is notable, however, that a simple explanation exists for each one. For the trace of growth hormone receptor (Figure 3(g) and (h)) in the protein–ligand set, the overlap does become significant, but only after the distantly related prolactin subfamily is pruned from the family tree (Figure 3(m)). In xenobiotic acetyltransferase (Figure 3(i) and (j)) from the enzyme set, it is the second largest cluster rather then the largest one that overlaps with the active site, causing the largest cluster overlap, average overlap, and hypergeometric distribution statistics to fail to reach significance, although the total connected residues statistic does reach significance due to the trace residues identified surrounding the active site. In fact, at higher ranks these two clusters merge together to form one large, statistically significant cluster that extends through the interior of the protein (Figure 3(n)), which suggests that the function performed by each site is not independent of the other.[12] Finally, in the transcription regulator Nc2 alpha chain (Figure 3(k) and (l)) from the SGI set, trace residues do not cluster significantly, although they tend to be near the active site. This protein is a member of a trimeric protein complex that also includes transcription regulator Nc2 beta chain and a TATA-Box-Binding Protein. Strikingly, these partners have significant overlap with the shared ligand (a 19 base-pair TATA-containing oligonucleotide; Figure 3(o), where the complex is viewed from "above" in relation to Figure 3(k) and (l), and the Nc2 alpha chain extends to the right), suggesting perhaps that these members of the complex provide the key interactions with the ligand rather than the Nc2 alpha chain.

A complementary approach to measure the accuracy of functional site identification is to determine how much of the functional site is identified by the largest trace cluster when the trace reaches its signal-to-noise rank threshold. For manually

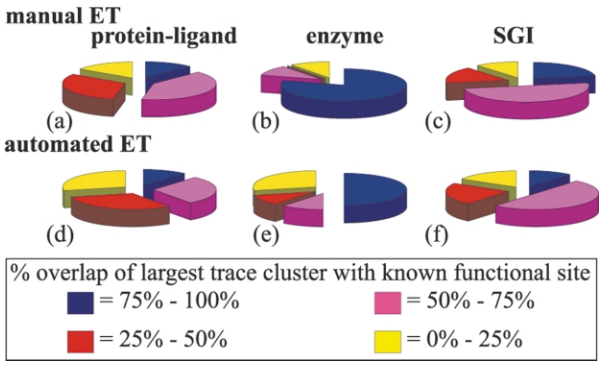

Figure 4. The largest significant trace cluster overlaps most of the functional site. This is shown for both manually refined traces in (a)–(c) and automated traces in (d)–(f). The overlap is especially extensive in the enzyme set where the sites are small, but it is also extensive in the sites defined by the approximate criterion of ligand proximity, often covering more than 50% of the site, even with the automated traces.

optimized traces and among the 79 proteins with significant clusters, more than 50% of the site is traced in 53% of the protein–ligand set, in 90% of the enzyme set, and in 72% of the SGI set (Figure 4(a)–(c), blue and purple). These variations arise from basic differences between functional sites defined by ligand proximity (ligand and SGI sets) and those defined experimentally (enzyme set). In the latter, active sites are mostly limited to key catalytic residues. This is only four residues on average but it can be as few as one, such as in chloramphenicol acetyltransferase (PDB code 1xat, EC number 2.3.1.28) where His79 is the lone putative general base for hydroxyl deprotonation.[27] Trace clusters will overlap most of these residues, but because they are so few, the overlap will tend to be less statistically significant. On the other hand, ligand proximity is only a rough measure of the true functional site since only a fraction of the residues near a ligand contribute to binding.[28] In that case, the overlap in the protein–ligand and SGI data sets tends to be less extensive but more significant. At worst, the largest cluster covers less than 25% of the functional site in 16% of the protein–ligand set, 10% of the enzyme set, and in 11% of the SGI set (Figure 4(a)–(c), yellow). When the automated ET is used, the percentage of the 81 proteins with significant clusters that have greater than 50% overlap with the functional site is 40% in the protein–ligand set, 61% in the enzyme set, and 61% in the SGI set (Figure 4(d)–(f), blue and purple).

Conclusions

The central feature of ET analysis is the hierarchical classification of functional characteristics, approximated through evolutionary trees. Since such trees are intrinsic to any protein family, ET should in principle be applicable to any protein structure with enough sequence homologs to sample the natural history of evolutionary variations and selections. This study provides the first large-scale, quantitative demonstration that functional sites can be identified accurately and generally in this manner and in diverse proteins selected either by us, by others, or by the SGI. Furthermore, we have shown that ET is scalable since an automated version identifies seven of ten functional sites by the least favorable statistical measure, and nine of ten by the most favorable one. The automated method also identifies significant trace clusters in 35 (88%) of the 40 SGI proteins available for this study. Hence, besides the 18 that had bound ligands, we have identified likely functional sites in 17 others.

As a caveat, it is important to note that current SGI target selection favors larger protein families for which a diverse sample of the families' evolutionary history is available. The extent to which ET can accurately identify functional sites in future SGI proteins will reflect a balance between the rate at which proteins from smaller families are targeted for crystallization, and the rate these same families are populated by sequencing efforts. In general it is reasonable to expect that as sequence and structure data continue their relentless growth, evolutionary tree-based functional classification methods will be increasingly useful to identify functionally important residues and to focus experiments, drug design, and functional annotation to the most biologically relevant sites of a protein.

Materials and Methods

Data sets

The overlap of trace clusters and functional sites were examined in 86 proteins drawn from three data sets. For lack of direct experimental evidence, the functional site was defined in many cases as all residues with a non-hydrogen atom within five angstroms of a ligand. Thus the "protein–ligand" set consists of the 37 protein–ligand complexes out of a larger set of 46 previously defined by Madabushi *et al.*[26] Such structurally determined functional sites are large, with an average of 28(±7) residues representing 13(±3)% of the protein. By contrast, in the "enzyme" set of representatives from 29 superfamilies defined by Todd *et al.*[27] individual active site residues are known experimentally from the literature. These sites are small with only four residues on average, or about 1% of the protein. As a further objective test of ET, we also defined the "SGI" set, which consists of the 22 protein–ligand complexes out of the 42 structures solved in the context of the SGI and readily accessible in the PDB[29] as of the end of 2001. Again, for lack of direct biochemical evidence, the functional sites were defined by proximity to a ligand. This set was further reduced to 20 proteins after removal of two outliers: the subunits in the Cyanate Lyase complexes (PDB codes 1dwk and 1dw9) whose decameric arrangement yields a functional site covering 72% of the protein, far more than the average 10(±3)%

(21(±8) residues). Detailed information regarding each protein in the three datasets is available on-line†.

Evolutionary Trace

The automated ET identifies homologs to the query structure with BLAST (using NCBI's non-redundant protein sequence database, the blosum62 substitution matrix, and default parameters).[30] It then retrieves from NCBI's Protein database (using NCBI's Entrez search engine)‡ the complete sequence of the top 100 homologs with an *e*-value better than 0.05 and covering at least 50% of the query sequence. Next, CLUSTALW (using the full alignment option and default parameters: gap open penalty = 10, gap extension penalty = 0.05)[31] is used to generate a multiple sequence alignment of the regions overlapping with the query structure and this alignment is the input for a gapped ET.[26] Manual refinement then involves re-aligning and tracing the sequences after pruning evolutionary outliers and sequence fragments. An ET server is under construction at our web site§. As previously described, the trace rank of a residue is the minimum number of branches into which the protein family tree must be partitioned for that residue to be invariant within each branch. Thus a residue of rank i is invariant within each of the first i branches of the tree (starting from the root), but variable within one of the first $(i-1)$ branches.[13] After identifying all residues ranked $\leq i$, as i varies from 1 to the maximum number of sequences, we cluster together those with non-hydrogen atoms that are within 4 Å of one another, and count the number of clusters and the size of the largest one at each rank. These values can then be compared to the number and size expected if trace residues were drawn randomly and thus assigned a statistical significance against the null hypothesis of randomly chosen residues.[26] Here, we consider trace clusters significant if both the number of clusters and size of largest cluster occur in less than 5% of the random simulations. Our random distributions were built at each rank, and for each protein, from (500 × (the number of amino acids in the protein)) random draws, up until more than 30% of the residues were class-specific, a threshold after which clusters typically are not significant.[26]

Significance of overlap between trace clusters and functional sites

To test the ability of trace clusters (trace residues that form significant structural clusters as described in the previous paragraph) to accurately predict functional sites, we developed three cluster-based overlap statistics and determined their *p*-values by comparison to the randomly chosen residue distributions at each rank. The total connected residues statistic (Figure 1(a)) is the total number of trace residues in the union of all clusters that overlap the functional site. The largest cluster overlap statistic (Figure 1(b)) is the number of residues in the intersection of the functional site present and its largest overlapping trace cluster. The average overlap statistic (Figure 1(c)) is the average number of residues in overlaps between trace clusters and the functional site. Finally, we also used the hypergeometric distribution (Figure 1(d)) as a non-cluster-based measure of the likelihood that t out of k trace residues will overlap by chance a functional site of R residues in a protein with N residues. The p-value of t is $1 - \Pr(X \leq t-1)$, where the probability mass function $\Pr(X = x)$ is $[C(R,x) \times C(N-R, k-x)/C(N,k)]$, and where $C(x,y)$ denotes the binomial coefficient (the number of combinations of x objects chosen y at a time). To perform our statistical testing against the null hypothesis of randomly chosen residues, we define for all of our tests a significance threshold of 0.05, the traditional benchmark in the biosciences.

† http://imgen.bcm.tmc.edu/molgen/labs/lichtarge/LSET/

‡ http://www.ncbi.nlm.nih.gov/Entrez/

§ http://imgen.bcm.tmc.edu/molgen/labs/lichtarge/

Acknowledgements

We gratefully acknowledge support from the American Heart Association, the March of Dimes, the NSF (DBI-0114796), NHGRI (O.L.), and the Keck Center for Computational Biology (D.M.K.).

References

1. Sowa, M. E., He, W., Slep, K. C., Kercher, M. A., Lichtarge, O. & Wensel, T. G. (2001). Prediction and confirmation of a site critical for effector regulation of RGS domain activity. *Nature Struct. Biol.* **8**, 234–237.
2. Strubbins, C. & Galan, J. (2001). Structural mimicry in bacterial virulence. *Nature*, **412**, 701–705.
3. de Rinaldis, M., Ausiello, G., Cesareni, G. & Helmer-Citterich, M. (1998). Three-dimensional profiles: a new tool to identify protein surface similarities. *J. Mol. Biol.* **284**, 1211–1221.
4. Wallace, A. C., Borkakoti, N. & Thornton, J. M. (1997). TESS: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites. *Protein Sci.* **6**, 2308–2323.
5. Kuntz, I. D., Blaney, J. M., Oatley, S. J., Langridge, R. & Ferrin, T. E. (1982). A geometric approach to macromolecule–ligand interactions. *J. Mol. Biol.* **161**, 269–288.
6. Sheinerman, F. B., Norel, R. & Honig, B. (2000). Electrostatic aspects of protein–protein interactions. *Curr. Opin. Struct. Biol.* **10**, 153–159.
7. Jones, S. & Thornton, J. M. (1997). Prediction of protein–protein interaction sites using patch analysis. *J. Mol. Biol.* **272**, 133–143.
8. Pettit, F. K. & Bowie, J. U. (1999). Protein surface roughness and small molecular binding sites. *J. Mol. Biol.* **285**, 1377–1382.
9. Lichtarge, O. & Sowa, M. E. (2002). Evolutionary predictions of binding surfaces and interactions. *Curr. Opin. Struct. Biol.* **12**, 21–27.
10. Henikoff, S. & Henikoff, J. G. (1991). Automated assembly of protein blocks for database searching. *Nucl. Acids Res.* **19**, 6565–6572.
11. Gaucher, E. A., Miyamoto, M. M. & Benner, S. A. (2001). Function–structure analysis of proteins using covarion-based evolutionary approaches: elongation factors. *Proc. Natl Acad. Sci.* **98**, 548–552.

12. Lockless, S. W. & Ranganathan, R. (1999). Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*, **286**, 295–299.
13. Lichtarge, O., Bourne, H. R. & Cohen, F. E. (1996). An Evolutionary Trace method defines binding surfaces common to protein families. *J. Mol. Biol.* **257**, 342–358.
14. Lichtarge, O., Bourne, H. R. & Cohen, F. E. (1996). Evolutionarily conserved Gabg binding surfaces support a model of the G protein–receptor complex. *Proc. Natl Acad. Sci. USA*, **93**, 7507–7511.
15. Lichtarge, O., Yamamoto, K. R. & Cohen, F. E. (1997). Identification of functional surfaces of the zinc binding domains of intracellular receptors. *J. Mol. Biol.* **274**, 325–337.
16. Onrust, R., Herzmark, P., Chi, P., Garcia, P. D., Lichtarge, O., Kingsley, C. & Bourne, H. R. (1997). Receptor and beta gamma binding sites in the alpha subunit of the retinal G protein transducin. *Science*, **275**, 381–384.
17. Sowa, M. E., He, W., Wensel, T. G. & Lichtarge, O. (2000). A regulator of G protein signaling interaction surface linked to effector specificity. *Proc. Natl Acad. Sci. USA*, **97**, 1483–1488.
18. Pritchard, L. & Dufton, M. J. (1999). Evolutionary trace analysis of the Kunitz/BPTI family of proteins: functional divergence may have been based on conformational adjustment. *J. Mol. Biol.* **285**, 1589–1607.
19. Landgraf, R., Fischer, D. & Eisenberg, D. (1999). Analysis of heregulin symmetry by weighted evolutionary tracing. *Protein Eng.* **12**, 943–951.
20. Innis, C. A., Shi, J. & Blundell, T. L. (2000). Evolutionary trace analysis of TGF-beta and related growth factors: implications for site-directed mutagenesis. *Protein Eng.* **13**, 839–847.
21. Pascual, J., Martinez-Yamout, M., Dyson, H. J. & Wright, P. E. (2000). Structure of the PHD zinc finger from human Williams-Beuren syndrome transcription factor. *J. Mol. Biol.* **304**, 723–729.
22. Landgraf, R., Xenarios, I. & Eisenberg, D. (2001). Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. *J. Mol. Biol.* **307**, 1487–1502.
23. Armon, A., Graur, D. & Ben-Tal, N. (2001). ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *J. Mol. Biol.* **307**, 447–463.
24. Hannenhalli, S. S. & Russell, R. B. (2000). Analysis and prediction of functional sub-types from protein sequence alignments. *J. Mol. Biol.* **303**, 61–76.
25. Aloy, P., Querol, E., Aviles, F. X. & Sternberg, M. J. (2001). Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. *J. Mol. Biol.* **311**, 395–408.
26. Madabushi, S., Yao, H., Marsh, M., Kristensen, D., Philippi, A., Sowa, M. E. & Lichtarge, O. (2002). Structural clusters of Evolutionary Trace residues are statistically significant and common in proteins. *J. Mol. Biol.* **316**, 139–154.
27. Todd, A. E., Orengo, C. A. & Thornton, J. M. (2001). Evolution of function in protein superfamilies, from a structural perspective. *J. Mol. Biol.* **307**, 1113–1143.
28. Cunningham, B. C. & Wells, J. A. (1993). Comparison of a structural and a functional epitope. *J. Mol. Biol.* **234**, 554–563. published erratum appears in J. Mol. Biol. 1994 Apr 8;237(4):513.
29. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H. *et al.* (2000). The Protein Data Bank. *Nucl. Acids Res.* **28**, 235–242.
30. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.
31. Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucl. Acids Res.* **22**, 4673–4680.

Edited by F. E. Cohen

(Received 10 September 2002; received in revised form 7 November 2002; accepted 14 November 2002)