*Journal of Structural and Functional Genomics* **4:** 159–166, 2003.



# Accurate and scalable identification of functional sites by evolutionary tracing

Olivier Lichtarge[1,2,*], Hui Yao[2], David M. Kristensen[2], Srinivasan Madabushi[2], Ivana Mihalek[1]

*[1]Department of Molecular and Human Genetics, [2]Program in Structural and Computational Biology and Molecular Biophysics, Baylor College of Medicine, Houston, TX 77030, USA; * Author for Correspondence e-mail: lichtarge@bcm.tmc.edu*

Received 29 January 2003; Accepted in revised form 30 April 2003

*Key words:* Evolutionary Trace, binding site, structural genomics, functional surface, bioinformatics, molecular recognition

## Abstract

A common difficulty in post genomics biology is that large-scale techniques of data collection often strip away information on the biological context of these data. The result is a massive number of disconnected observations on sequence, structure, and function from which underlying patterns and biological meaning are obscured. One solution is to build computational filters that pick out sufficiently few facts, relevant to a query, that their relationship is immediately apparent and experimentally testable. Typically, these filters rely on mathematics and statistics, and on first principles from physics and chemistry. We show here that evolution itself can be used to filter sequence and structure data in order to identify evolutionarily important amino acids. A general property of these residues is that they form clusters in native protein structures and point to regions where mutations have the greatest biological impact. The result is an accurate method of functional site annotation that is scalable for structural proteomics.

## Introduction

Protein functional sites are intimately connected with biological activity and their characterization bears on drug design (Kuntz, 1992) and protein engineering (Hellinga, 1998). To the extent that functional sites can be described as structural units, their characterization may also lead to 3-dimensional motifs that are suitable for function annotation in novel protein structures (Nussinov and Wolfson, 1991, Wallace et al., 1996, de Rinaldis et al., 1998). Unfortunately, protein functional sites cannot be predicted from protein structure alone. For example, when complexes reveal physical interaction sites between proteins, not all interfacial residues contribute equally to binding affinity and specificity (Pearce et al., 1996). The best approach to link a residue with its function remains mutational analysis (Pearce et al., 1996, Fersht, 1987), but this is expensive and necessarily dependent on assays that are protein, cell, and species specific. Given the rapid growth of solved protein structures, a key problem is to develop a method of functional site identification that would be as accurate as mutational analysis but also scalable and cost-efficient.

The Evolutionary Trace model (ET) aims to mimic mutational analysis in the laboratory by using the experiments that already occurred during evolution (Lichtarge et al., 1996b). This model is depicted in Figure 1, and it states simply that (i) sequence variations are equivalent to mutations and that (ii) evolutionary divergence is equivalent to a functional change as measured by an assay. Taken literally, these two hypotheses imply that functionally important residues are those where sequence variations are linked to evolutionary divergences. In a multiple sequence alignment of $n$ related proteins with an associated family tree $\Im$, we therefore define *trace residues*, at rank $k$, as those that are invariant *within* each of the

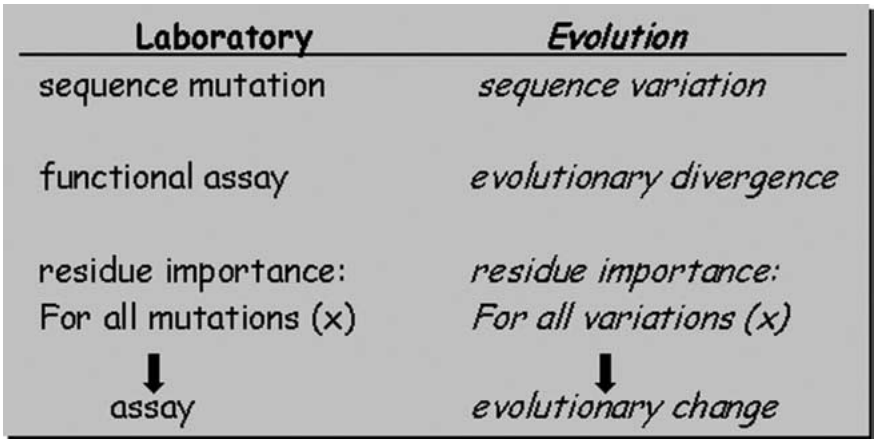


*Figure 1.* The Evolutionary Trace model mimics mutational analysis.

first $k$ branches of $\Im$ but that do vary in at least one of the first $k-1$ branches. At one extreme, trace residues with a rank of 1 are invariant and, presumably, the most important to function and structure. At the other extreme, residues that vary as far as between adjacent terminal leaves of tree $\Im$ will have ranks approaching $n$ and, presumably, be among the least important (Lichtarge et al., 1996b). This process is illustrated in Figure 2, where $k = 4$.

This model leaves open the exact type of tree to be used. This is because any number of hierarchical functional classification schemes can be built by a user, and ET probes for each one the hypothesis that there are residues commonly important to these sequences when they are classified functionally according to this tree. This model can also be modified to tolerate conservative substitutions within branches (Landgraf et al., 1999, Armon et al., 2001), thereby increasing sensitivity at the expense of specificity (Lichtarge et al., 1996b). In practice, however, we find that sequence identity trees are good default approximations of evolutionary and functional divergences in individual protein families. Together with strict invariance between branches, they are sufficiently robust and sensitive to trace sequences, assign ranks to residues, and identify structural clusters of top-ranked trace residues in structures that predict functional sites and guide mutational analysis. We review these data and then discuss progress on the generalization of ET towards functional site characterization on a proteomic scale.

**Functional site predictions in G$\alpha$ and in RGS**

Control studies in modular signaling or DNA binding domains validated ET (Lichtarge et al., 1996b, Lichtarge et al., 1997), but the first *bona fide* Evolutionary Trace prediction was in the $\alpha$-subunit of G proteins. A site including the C-terminal, the distal helix $\alpha$5, and strand $\beta$6 was predicted to mediate G protein activation through direct contacts with G Protein Coupled Receptors (GPCR) (Lichtarge et al., 1996a). Subsequent and independent alanine-scanning mutagenesis of 106 residues then identified a GPCR interface composed of most of the same residues in the same secondary structure elements (Onrust et al., 1997). A recent mutational study further supports the role of helix $\alpha$5 in GPCR-mediated activation (Marin et al., 2002).

This was followed by a second blind prediction, this time in Regulators of G protein Signaling (RGS). RGS proteins normally bind onto activated G$\alpha$ (complexed to GTP) and enhance its intrinsic rate of GTP hydrolysis, whereby G$\alpha$ turns itself back off (complexed to GDP) (He and Wensel, 2002). In the visual pathway, an additional interaction of the G$\alpha$-RGS complex with PDE$\gamma$ (the visual effector protein) further boosts the GTP-ase accelerating property of RGS 9 but slows down that of RGS7. In order to understand the molecular basis of PDE$\gamma$'s differential effect on GTP hydrolysis by G$\alpha$, we performed an ET analysis of the RGS family. This identified a novel site on the surface of RGS, called $S_2$, at which residues differ markedly in RGS7 and RGS9. A direct interaction $S_2$-PDE$\gamma$ was also suggested by the proximity of $S_2$ in the G$\alpha$-RGS complex to a site in G$\alpha$ shown to

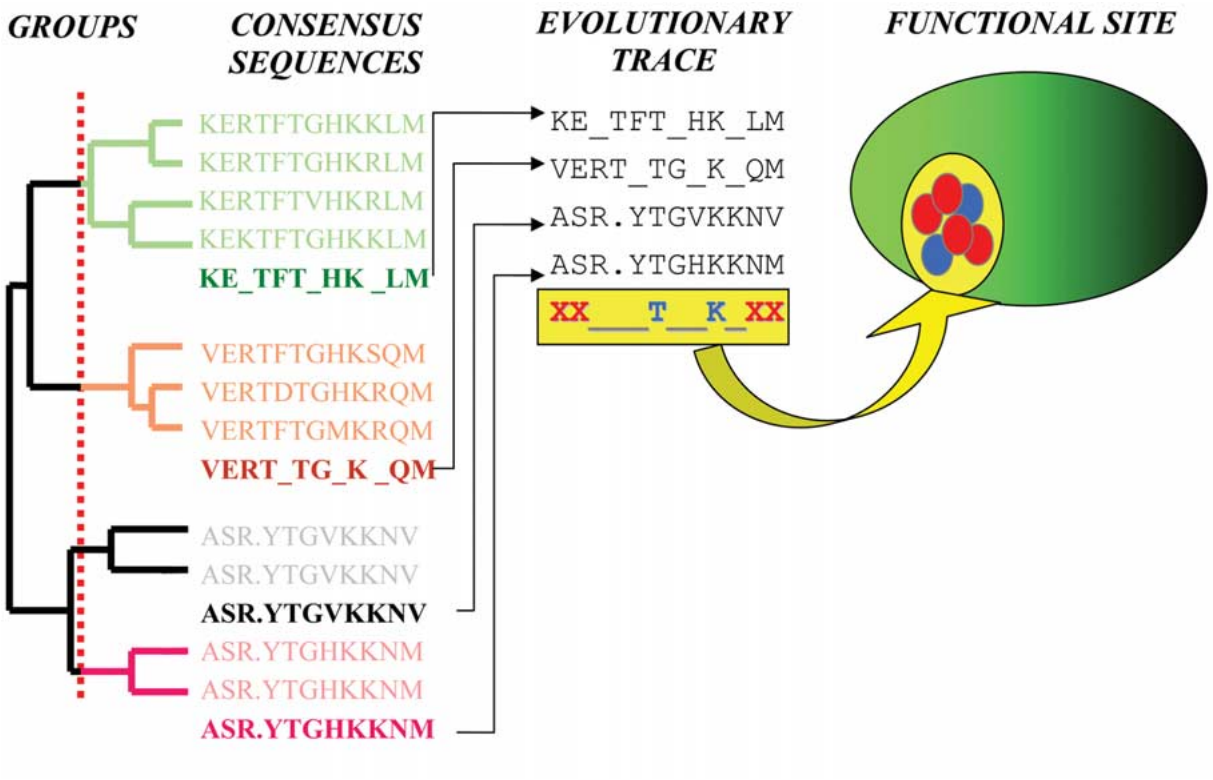


*Figure 2.* The Evolutionary Trace mechanism. The proteins in the multiple sequence alignment are divided into groups by phylogeny. The consensus sequence of invariant residues is generated for each group, and those that are invariant in each one but possibly variable between groups are called trace residues (red). These are then mapped onto a structure to visualize a functional site. Completely invariant residues with rank 1 are shown in blue and group-specific residues are shown in red.

interact with PDE$\gamma$ in peptide studies (Skiba et al., 1996). Together these data suggested a low resolution heterotrimer model where PDE$\gamma$ straddles the RGS-G$\alpha$ complex and interacts directly with $S_2$, which, in turn, mediates PDE$\gamma$'s effects on the GTP-ase accelerating property of RGS proteins (Sowa et al., 2000).

This model predicts that a swap at $S_2$ of cognate residues from RGS9 into RGS7 should confer to the RGS7 chimera many of the properties of RGS9 with respect to PDE$\gamma$. Out of 65 different residues between RGS7 and RGS9, we therefore focused only on the 7 in $S_2$, and after 5 mutation experiments identified 3 key residues shown in Figure 3. Mutations E353L and P360R, diminished the GTP-ase activating property of RGS7 to a level comparable to wild type RGS7 in the presence of PDE$\gamma$, and adding PDE$\gamma$ made no difference. S367G then conferred onto the RGS7 mutant an activity level near that of RGS9, and adding PDE$\gamma$ fully restored activity to that of RGS9 in the presence of PDE$\gamma$. These experiments therefore confirmed that $S_2$ plays a key in the PDE$\gamma$ interaction. Moreover the low-resolution quaternary structure model of PDE$\gamma$ straddling the RGS-G$\alpha$ complex and binding to $S_2$ was independently verified crystallographically (Slep et al., 2001), as shown in Figure 3.

These protein-specific studies uncover some of the general mechanisms that trigger and regulate G protein signaling. But more generally, they suggest that evolutionary analysis can link raw sequence and structure data to the molecular basis of function with sufficient resolution to target mutational experiments to the relevant regions of a protein and to engineer novel functional specificity into them. Other protein-specific studies by us and by others have been reviewed and support the validity of this Evolutionary Trace model (Lichtarge and Sowa, 2002, Innis et al., 2000, Landgraf et al., 2001, Aloy et al., 2001, Pritchard and Dufton, 1999, Armon et al., 2001).

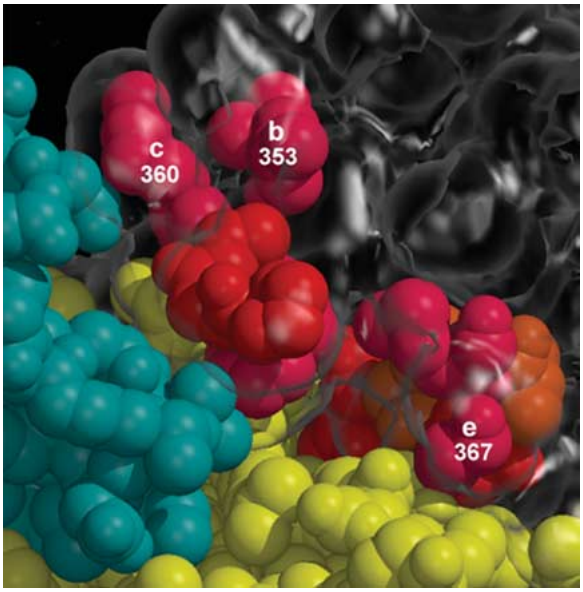


*Figure 3.* This functional site was predicted by the Evolutionary Trace and later confirmed by mutation experiments and crystallographically. Here the complex between RGS9 (clear), Gα (yellow), and PDEγ (cyan) shows direct contact between the residues predicted to be important in RGS9 ($S_2$, red) and the interface of Gα and PDEγ. Three of the residues (labeled 353,360,367) in $S_2$ were identified as key residues by mutation experiments.

### Large-scale identification of functional sites

Two recent studies test the scalability, accuracy, and automation of the evolutionary trace model. They show that statistically significant clusters of trace residues are a universal finding in proteins [18], and that their overlap with known functional sites is also significant [32]. Trace clusters are defined as collections of trace residues within 4Å of at least one other trace residue in the same cluster. For a given protein, let $T_k$ be the number of residues ranked $\leqslant k$, $N_k$ be the number of clusters that they define, and $S_k$ the size of the largest cluster. In the first study, the structural significance of the trace clusters was then measured from distributions $D_{cluster}$ (Number of Clusters) and $D_{size}$ (Size of the Largest) expected if the $T_k$ residues were picked randomly. Each distribution was approximated by 5000 repeated random samplings, the significance thresholds used were 5%, 1%, 0.03% and they were linear functions of protein size in each case.

The test set consisted of 46 proteins selected from the PDB so as to represent a variety of structural classes and biological functions. About half a dozen were from prior analyses, the others were selected blindly. Over the entire test set, 24 were uniquely of eukaryotic origin, 18 were both eukaryotic and prokaryotic, 2 were prokaryotic only, and 2 others were viral. Sequence identity was typically between 30 to 50% in the various protein families, and their structures represented a diversity of folds as well, including 19 with $\alpha/\beta$ folds, 15 were all $\alpha$, 7 all $\beta$, 1 multidomain, and one was a membrane protein. Sequence alignments were taken as is, but obvious sequence fragments were discarded. The web site http://imgen.bcm.tmc.edu/molgen/labs/lichtarge/trace_of_the_week/traces.html describes these proteins.

We find that at a *p-value* of 0.05, 95% and 92% of test proteins have statistically significant clusters of trace residues by the Number of Cluster and Size of the Largest cluster statistics, respectively. This represents 45 of 46 true positives by one measure and 44 of 46 by the other. These fractions remain high (75% for $D_{cluster}$ and 85% for $D_{size}$) and (65% for $D_{cluster}$ and 74% for $D_{size}$) at *p-values* of 0.01, and 0.003 respectively. All seven protein families with 30 or fewer homologs reached a p-value of 0.003, including the smallest family that only had 19 sequences. This is consistent with our experience that families with 15 or more homologs are traceable. Moreover, side-by-side comparisons of the largest predicted, significant cluster with the physical binding site, defined by proximity to a ligand when available, show that trace clusters match the functional sites (see the above URL). False negatives are not easily quantified since clusters that do not match known functional sites can simply represent internal residues that are important to folding, structural stability, allostery, or they may represent an as yet unrecognized functional site.

In order to quantify how well trace clusters match experimentally determined functional sites, and thus ET's predictive accuracy, a follow-up study compared their overlap to that expected by chance (Figure 4). To further reduce a possible selection bias, the test set was enlarged to 86 proteins, those above that had ligands, 29 enzymes whose active sites had been biochemically characterized in the literature (Todd et al., 2001), and 20 from a first set of Structural Genomics Initiative (SGI, the worldwide effort to solve the structure of every single natural protein fold) structures solved in complex with a ligand. Given many possible ways to measure the overlap of trace clusters and true sites, we used three statistics that take clustering into account. The Total Connected Residues (TCR) statistic, which counts any residues in

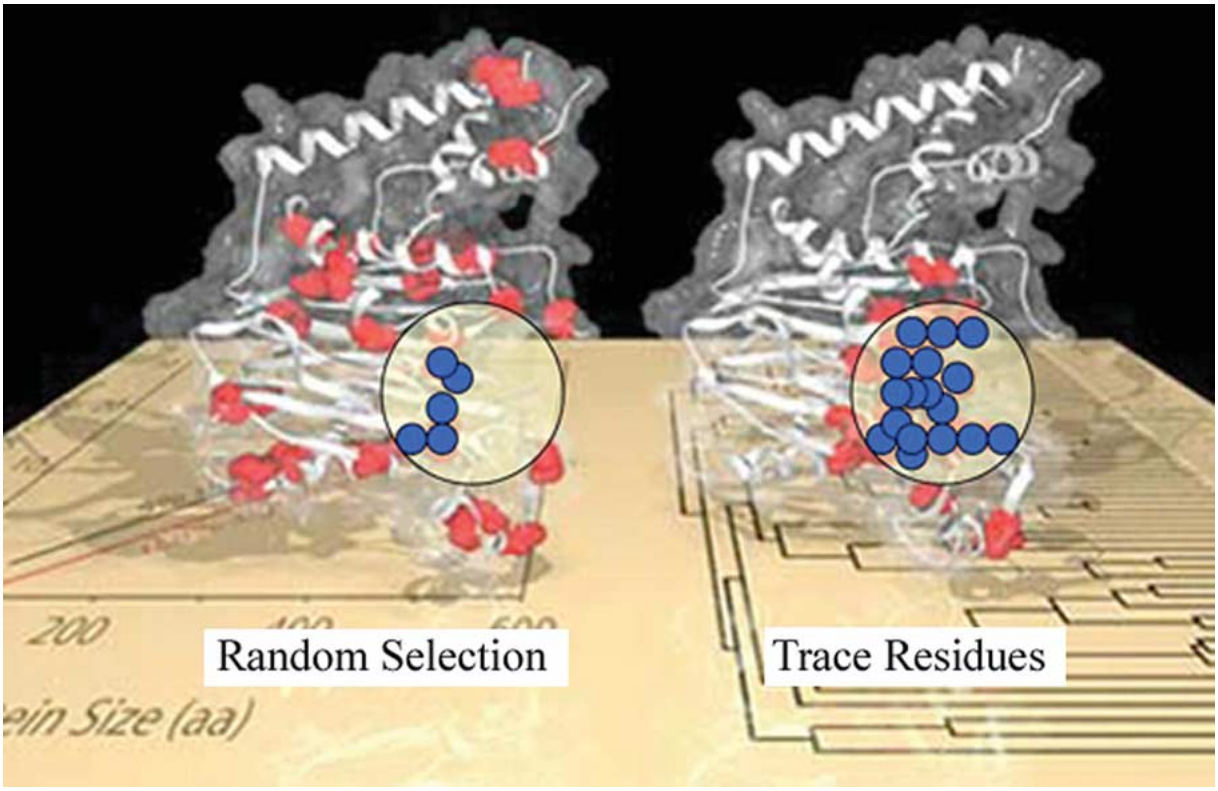


*Figure 4.* Trace residues (right) in pyruvate decarboxylase (PDB code: 1pvd) form significantly larger and fewer clusters than the same number of randomly selected residues (left), and they overlap the functional site far more.

any cluster that intersect with the true site. The Largest Cluster Overlap (LCO), which counts only residues common to the largest trace cluster and the true site. The Average Overlap (AO), which is the number of trace residues in the true site divided by the number of trace clusters that overlap that site. Finally a fourth statistic, which is the hypergeometric (HG) distribution, simply counts the number of trace residues in the functional site without regard to clustering.

We find that by the most favorable statistic, TCR, there is always one rank for which trace clusters significantly overlap true functional sites, as shown in Figure 5 for a *p-value* of 0.05. Even by the least favorable non-cluster-based HG statistic, this remains true in 86% of proteins. These numbers must be interpreted with care because they do represent multiple samplings: one for each rank. However, these ranks are not independent of one another, and in fact the overlap with the true site tends to be significant at most ranks where clustering is significant. With the TCR statistic this true in ˜90% of ranks, and in ˜75% of ranks by HG. Even though these numbers are averaged over all proteins, they do not vary markedly among the enzymes, the SGI proteins, or proteins from our original data set.

To further test scalability, we automated trace analysis by using the standard BLAST Evalue < 0.05 as the sole criterion for sequence selection. The performance degraded only marginally (Figure 5, blue bars) since the overlap of trace cluster remains significant well-above 90% of proteins according to TCR, and in nearly 75% by HG. We expect that the gap with manual trace results will narrow as the sequence selection heuristics improve.

The accuracy of trace clusters can also be ascertained by measuring how much of the true site is overlapped by the largest significant trace cluster. Figure 6 shows that this cluster covers more than 50% of the site in 70% of proteins, and less than 25% in fewer than 15% of proteins. Since the manual traces were more effective at removing sequence fragments and evolutionary outliers, their results are consistently better than those of automated traces. Fractional overlap is especially large in enzymes, for which the ‘true

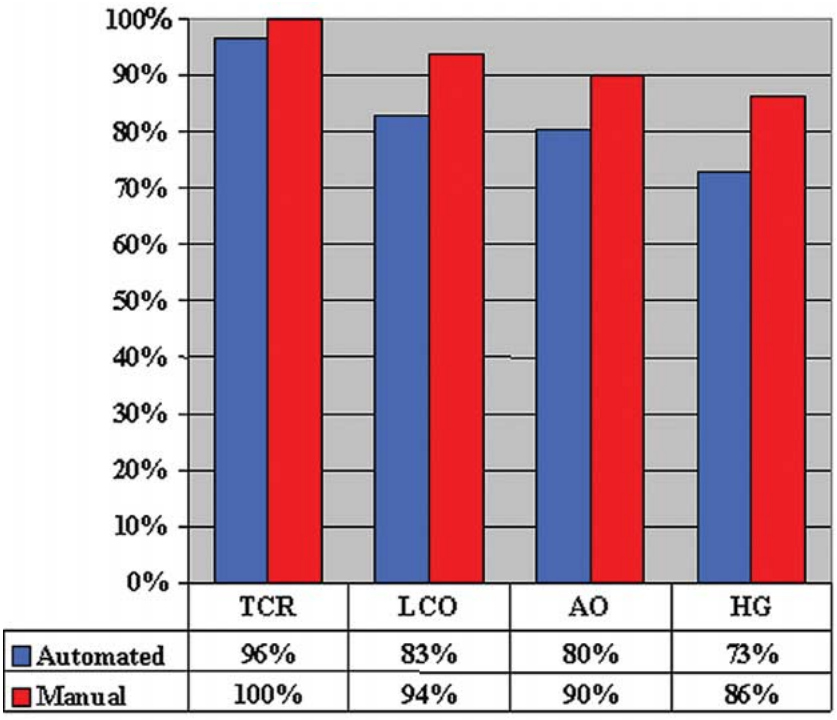


*Figure 5.* In four statistical measures, a large fraction of automated traces (blue) or manual traces (red) overlap significantly with known functional sites for at least one rank.

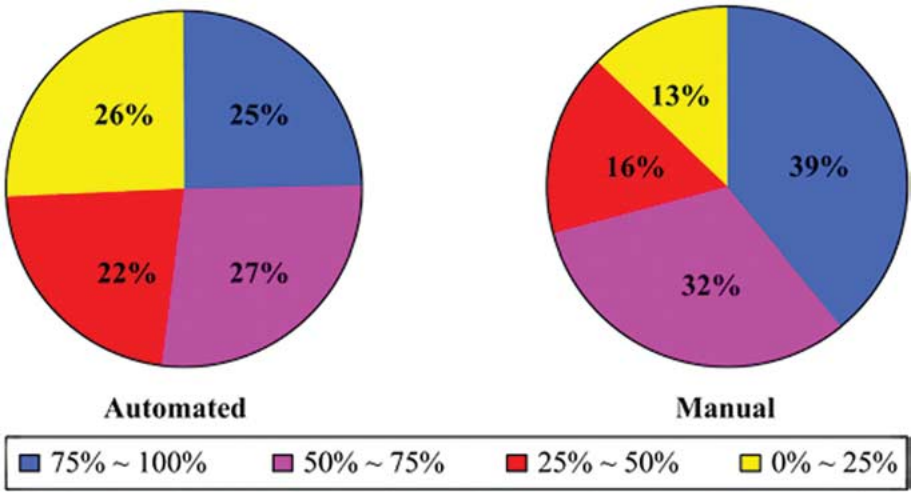


*Figure 6.* The largest significant trace cluster covers most of the known functional site in both the automated traces (left) and manual traces (right). The values in each pie chart reflect the fraction of proteins that cover the proportion of the known functional site indicated in the range shown at bottom.

site' was determined mutationally and therefore tends to be small. Conversely, fractional overlap is smaller in proteins where the 'true site' is a large interface determined, for lack of a better method, as all the residues that are within 5 Å of the ligand. This overestimates the importance of many residues, and it is likely that the small trace cluster appropriately reflects the small number of key interfacial residues.

## Future directions

In theory, these studies depend on adequate alignment and tree building methods, as well as on error free sequences. Certainly it is possible to make alternative uses of phylogenetic trees and modify the definition of branch specific invariance (Armon et al., 2001), and to exploit other aspects of structural clustering (Landgraf et al., 2001). But, in practice, we used here standard CLUSTALW and distance-based UPGMA trees. The Evolutionary Trace model proves sufficient, as described above, to extract from sequence and structure databases detailed information on the functional importance of residues and of their structural neighbors. Protein specific applications have been, thus far, to rank the relative evolutionary importance of residues and identify functional sites (Lichtarge et al., 1996b, Pritchard and Dufton, 1999, Innis et al., 2000); to derive functional specificity determinants and target them for mutations in order to engineer novel functions (Sowa et al., 2000, Sowa et al., 2001); to build low resolution models of protein-protein quaternary structure (Lichtarge et al., 1996a); to infer functions shared by distant homologs (Lichtarge et al., 1997); to align and study GPCR (Sheikh et al., 1996, Sheikh et al., 1999, Baranski et al., 1999, Dean et al., 2001) and to study protein-protein interaction and docking (Aloy et al., 2001). Finally, the large-scale studies demonstrate that the ET model is general and scalable so these small scale studies should be replicable in any other protein in the proteome (Madabushi et al., 2002, Yao et al., 2003).

In the context of the Structural Genomics Initiative, there are many ways in which Evolutionary Trace identification of functional sites may be increasingly useful to understand protein function. If a protein's biological activity is known, its functional sites and key determinants of specificity can be identified and then targeted for modification or drug design. When biological activity is only partially known, functional annotation of protein structures and homologous genes can focus precisely on the presence or absence of key trace residues that confer activity rather than on overall sequence identity (Lichtarge et al., 1997, Aloy et al., 2001). Finally in the absence of any functional information at all, functional convergence may be inferred by recognizing that some region in a novel protein bears local structural similarities to the Evolutionary Trace site of a functionally characterized protein, thereby extending to 3-dimensions the typical functional motif searches in sequence space (Fischer et al., 1993, Stark et al., 2003). Even more generally, it will also be useful to understand the mechanism by which the ET model extracts information from sequences and structures. Simply put, ET mimics mutational analysis but with important differences. Since it views sequence variations as mutations, ET's 'in vivo' mutations are always active and adequate for survival. Since it views tree nodes as assays, ET's functional assays are also far more numerous then typically available in the wet lab. Thus for any proteins with enough sequence homologs in the database, ET benefits from large number of evolutionary mutation and assay experiments.

A more formal view is that the tree embodies a hierarchical classification scheme that filters, at each branchpoint, the residues that are most functionally and structurally relevant. A novel but basic principle of protein structure, function, and interaction, thus appears to be that these trace residues form structural clusters and networks that link key residues and functional sites to one another. As the SGI produces ever more structures, it will be interesting to study each one from an Evolutionary Trace perspective and understand how these trace clusters and networks inform studies of protein folding, function, and interaction.

## Acknowledgements

O.L. gratefully acknowledges support from the American Heart Association, the March of Dimes, the NSF (DBI-0114796), and NHGRI (HG02345).

## References

1. Aloy, P., Querol, E., Aviles, F. X. and Sternberg, M. J. (2001) *J Mol Biol,* **311,** 395–408.
2. Armon, A., Graur, D. and Ben-Tal, N. (2001) *J Mol Biol,* **307,** 447–63.

3. Baranski, T. J., Herzmark, P., Lichtarge, O., Gerber, B. O., Trueheart, J., Meng, E. C., Iiri, T., Sheikh, S. P. and Bourne, H. R. (1999) *J Biol Chem,* **274,** 15757–65.
4. de Rinaldis, M., Ausiello, G., Cesareni, G. and Helmer-Citterich, M. (1998) *J Mol Biol,* **284,** 1211–21.
5. Dean, M. K., Higgs, C., Smith, R. E., Bywater, R. P., Snell, C. R., Scott, P. D., Upton, G. J., Howe, T. J. and Reynolds, C. A. (2001) *J Med Chem,* **44,** 4595–614.
6. Fersht, A. R. (1987) *Biochemistry,* **26,** 8031–7.
7. Fischer, D., Norel, R., Wolfson, H. and Nussinov, R. (1993) *Proteins,* **16,** 278–92.
8. He, W. and Wensel, T. G. (2002) *Methods Enzymol,* **344,** 724–40.
9. Hellinga, H. W. (1998) *Nat Struct Biol,* **5,** 525–7.
10. Innis, C. A., Shi, J. and Blundell, T. L. (2000) *Protein Eng,* **13,** 839–47.
11. Kuntz, I. D. (1992) *Science,* **257,** 1078–82.
12. Landgraf, R., Fischer, D. and Eisenberg, D. (1999) *Protein Eng,* **12,** 943–51.
13. Landgraf, R., Xenarios, I. and Eisenberg, D. (2001) *J Mol Biol,* **307,** 1487- 502.
14. Lichtarge, O., Bourne, H. R. and Cohen, F. E. (1996a) *Proc Natl Acad Sci U S A,* **93,** 7507–11.
15. Lichtarge, O., Bourne, H. R. and Cohen, F. E. (1996b) *J Mol Biol,* **257,** 342- 58.
16. Lichtarge, O. and Sowa, M. E. (2002) *Curr Opin Struct Biol,* **12,** 21–7.
17. Lichtarge, O., Yamamoto, K. R. and Cohen, F. E. (1997) *J Mol Biol,* **274,** 325–37.
18. Madabushi, S., Yao, H., Marsh, M., Kristensen, D. M., Philippi, A., Sowa, M. E. and Lichtarge, O. (2002) *J Mol Biol,* **316,** 139–54.
19. Marin, E. P., Krishna, A. G. and Sakmar, T. P. (2002) *Biochemistry,* **41,** 6988–94.
20. Nussinov, R. and Wolfson, H. J. (1991) *Proc Natl Acad Sci U S A,* **88,** 10495–9.
21. Onrust, R., Herzmark, P., Chi, P., Garcia, P. D., Lichtarge, O., Kingsley, C. and Bourne, H. R. (1997) *Science,* **275,** 381–4.
22. Pearce, K. H., Jr., Ultsch, M. H., Kelley, R. F., de Vos, A. M. and Wells, J. A. (1996) *Biochemistry,* **35,** 10300–7.
23. Pritchard, L. and Dufton, M. J. (1999) *J Mol Biol,* **285,** 1589–607.
24. Sheikh, S. P., Vilardarga, J. P., Baranski, T. J., Lichtarge, O., Iiri, T., Meng, E. C., Nissenson, R. A. and Bourne, H. R. (1999) *J Biol Chem,* **274,** 17033–41.
25. Sheikh, S. P., Zvyaga, T. A., Lichtarge, O., Sakmar, T. P. and Bourne, H. R. (1996) *Nature,* **383,** 347–50.
26. Skiba, N. P., Bae, H. and Hamm, H. E. (1996) *J Biol Chem,* **271,** 413–24.
27. Slep, K. C., Kercher, M. A., He, W., Cowan, C. W., Wensel, T. G. and Sigler, P. B. (2001) *Nature,* **409,** 1071–7.
28. Sowa, M. E., He, W., Slep, K. C., Kercher, M. A., Lichtarge, O. and Wensel, T. G. (2001) *Nat Struct Biol,* **8,** 234–7.
29. Sowa, M. E., He, W., Wensel, T. G. and Lichtarge, O. (2000) *Proc Natl Acad Sci U S A,* **97,** 1483–8.
30. Stark, A., Sunyaev, S. and Russell, R. B. (2003) *J Mol Biol,* **326,** 1307–16.
31. Todd, A. E., Orengo, C. A. and Thornton, J. M. (2001) *J Mol Biol,* **307,** 1113–43.
32. Wallace, A. C., Laskowski, R. A. and Thornton, J. M. (1996) *Protein Sci,* **5,** 1001–13.
33. Yao, H., Kristensen, D. M., Mihalek, I., Sowa, M. E., Shaw, C., Kimmel, M., Kavraki, L. and Lichtarge, O. (2003) *J Mol Biol,* **326,** 255–61.