

# Evolutionary predictions of binding surfaces and interactions

Olivier Lichtarge\* and Mathew E Sowa†

Rapid progress in structural biology and whole-genome sequencing technology means that, for many protein families, structural and evolutionary information are readily available. Recent developments demonstrate how this information can be integrated to identify canonical determinants of protein structure and function. Among these determinants, those residues that are on protein surfaces are especially likely to form binding sites and are the logical choice for further mutational analysis and drug targeting.

## Addresses

\*Department of Molecular and Human Genetics, 1 Baylor Plaza, Baylor College of Medicine, Houston, Texas 77030, USA; e-mail: lichtarge@bcm.tmc.edu

†Department of Biochemistry and Molecular Biology, Program in Structural and Computational Biology and Molecular Biophysics, 1 Baylor Plaza, Baylor College of Medicine, Houston, Texas 77030, USA; e-mail: ms691994@bcm.tmc.edu

Current Opinion in Structural Biology 2002, 12:21–27

0959-440X/02/\$ – see front matter

© 2002 Elsevier Science Ltd. All rights reserved.

## Abbreviations

ET	evolutionary trace
G $\alpha$	G protein $\alpha$ subunit
HMM	hidden Markov model
IR	intracellular hormone receptor
MSA	multiple sequence alignment
PDB	Protein Data Bank
RGS	regulator of G protein signaling
SH	Src homology

## Introduction

Protein functional sites play a central role in biology. They include catalytic centers, interfaces with ligands and conformational switches, and can be defined as clusters of amino acids whose structural, dynamic and physicochemical properties directly impact ligand interaction and transformation. Many important applications follow from their identification, including drug design, engineering protein mimetics and elucidating molecular pathways through site-directed mutagenesis [1,2<sup>••</sup>]. Furthermore, the proper identification of binding surfaces can lead to protein–protein association models, for example, by understanding whether protein oligomers are biologically relevant or simply crystallization artifacts [3,4], and also by guiding protein docking [5<sup>••</sup>]. Finally, convergent evolution and molecular mimicry in bacterial pathogens [6] and in autoimmune diseases suggest that functional sites can be viewed as irreducible modules of biological activity. Their precise topological description [7] could therefore be useful for functional annotation. This review discusses sequence- and structure-based approaches to recognizing functional sites, and then focuses on emerging methods that explicitly incorporate evolutionary information.

## Sequence variations and function

It has long been recognized that functional sites undergo fewer mutations during evolution than other parts of a protein [8,9]. Amino acid differences at functional sites can be described in terms of sequence conservation patterns using matrices of position-specific variations, called profiles [10], or through statistical means, such as hidden Markov models (HMMs) [11,12]. Both methods can sensitively detect local sequence motifs, which can be associated with specific functions and compiled, as, for example, in the INTERPRO database [13]. Global sequence homologies can also be identified by these methods or by the related and even more sensitive PSI-BLAST search [14,15], thereby suggesting putative functions and functional sites, by analogy.

These sequence-based methods are fundamental to the functional annotation of proteins, but they can be limited for two reasons. First, many functions involve large interfacial areas, rather than short local sequence motifs. Second, functional analogies can be specious, especially when sequence identity falls below 40%. Under that threshold, functional differences are increasingly common [16] and homologs may have unrelated functions, distinct chemistries and different functional sites [17<sup>••</sup>,18]. Circumventing these limitations requires careful examination of a protein structure's physical features.

## Physical and compositional heterogeneity of functional sites

Protein–ligand interactions are so diverse that no single attribute guarantees or precludes a site from being part of an interface. Geometric and electrostatic complementarity to ligands are universal features of functional surfaces [19,20] and, in general, small ligands bind in cavities [21] and in areas of increased surface roughness [22]. On the other hand, protein–protein interactions typically involve large, accessible and mostly planar sites where the solvation potential, interface propensities and protrusion of residues cannot be easily distinguished *a priori* from the rest of the protein's surface [23].

This heterogeneity of functional sites is consistent with their plasticity and with the context dependence of the functional contributions of their amino acid components. A review of 2325 alanine mutants in 22 protein–protein complexes [24] reveals that only a fraction of the residues that contact a ligand actually contribute to the binding free energy [25]. Furthermore, the efficacy of some active sites is allosterically regulated [2<sup>••</sup>,26] and the site can adapt to mutations through extensive structural rearrangements of as much as 3 Å by both mainchain and sidechain atoms located up to 15 Å away [27]. Thus, the effective functional site not only includes part of the ligand contact site, but

also extends beyond it through allosteric interactions, and the effect of specific mutations depends largely on their surrounding environment.

Nevertheless, useful generalizations can be made. For stable protein–protein interactions, binding surfaces are found to have a lack of charged groups and an excess of hydrophobic residues, leading to an amino acid composition intermediate to that of the protein interior and surface [28]. For transient protein–protein interactions, such as those found among proteins in signaling pathways, the binding interface does not differ greatly in composition from the rest of the protein surface. On average, a ‘standard interface’ buries approximately  $1600 \pm 400 \text{ \AA}^2$  and contains either  $10 \pm 5$  hydrogen bonds or 2 salt bridges per interface, but large deviations are also observed [29].

In the case of protein–DNA interfaces, lysine and arginine residues interact electrostatically with the phosphate DNA backbone, with an average of 1 hydrogen bond/ $125 \text{ \AA}^2$ , and there are an equal number of hydrogen bonds involving water molecules [30,31]. A recent study using 129 protein–DNA complexes revealed that two out of three of the direct interactions between amino acids and nucleotides involve van der Waals contacts, one out of six involves hydrogen bonds and the remaining one out of six involves water-mediated bonds [32]. In the case of protein–RNA interfaces, a study of 32 protein–RNA complexes by Jones *et al.* [33] noted the expected importance of arginines, but found that single aromatic residues (phenylalanine and tyrosine) are also key factors in binding. Finally, in a study of protein–carbohydrate sites (19 protein–sugar complexes), arginine, aspartate and glutamate are found to play prominent roles, in addition to aromatic residues [34]. However, the lack of a single recognition template for these interactions remains a problem in their identification.

Thus, functional sites are sufficiently unique that their detailed characterization requires direct mutational analysis, for example, alanine scanning mutagenesis, which directly links sequence variation with a functional consequence. Systematic and exhaustive mutations, however, are resource-intensive, protein-specific and critically dependent on assays being available to test functional differences before and after mutation. An alternative, embodied by the evolutionary trace (ET) method [35], is to emulate mutational analysis using the mutations and functional assays that have already occurred during evolution, rather than those made in the laboratory.

## Evolutionary tracing

### Theory

ET analysis is predicated on two hypotheses. The first hypothesis is that functional sites evolve through variations of a conserved architecture. If so, whereas architecture-defining residues might be mostly invariant, the residues responsible for functional specificity could undergo many

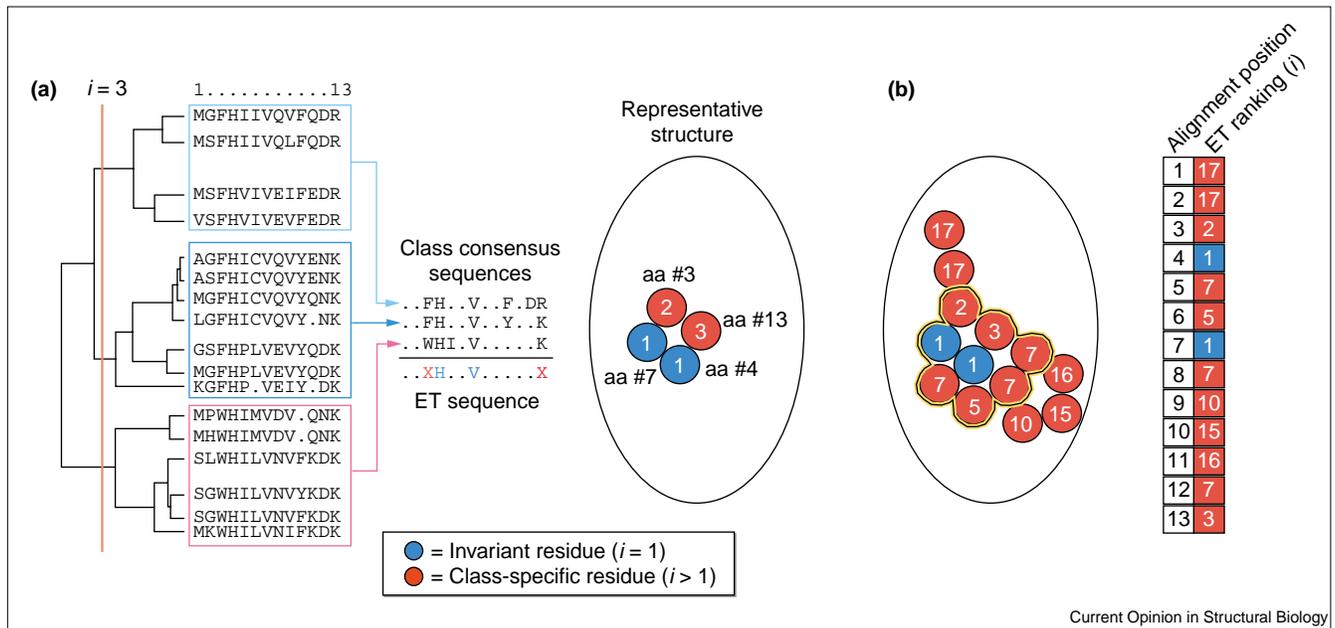
substitutions, each associated with a functional variation and, hence, divergence. A problem in identifying residues whose variations always correlate with changes in function is that the functional differences among dozens or even hundreds of proteins within a family must be known. ET employs a second hypothesis to address this problem by postulating that sequence identity trees approximate functional classifications. This approach is plausible because proteins with greater sequence identity have diverged more recently than sequences with lesser similarity, and therefore have had less time to functionally diverge [35].

### Method

The input to ET consists of a protein family with divergently related sequences in a multiple sequence alignment (MSA) and a related sequence identity tree, which can be generated using programs such as Accelrys’s PILEUP (found in the Genetics Computer Group [GCG®] Wisconsin Package™), CLUSTALW and PHYLIP. ET then iteratively partitions the protein family into an increasing number of subgroups delineated by branch points in the tree (Figure 1), beginning with one group containing all of the sequences in the family and ending with each protein being its own subgroup. For example, the  $i^{\text{th}}$  trace is done with the family divided into the first  $i$  branches, as shown in Figure 1 for  $i = 3$ . Positions in the MSA that have invariant residues within each of the subgroups, but for which residue identity varies among the subgroups are called trace residues (also known as class-specific residues). The evolutionary rank of a residue is the minimum number of branches into which the tree must be divided for it to become a trace residue. Accordingly, a residue with rank  $k$  will be invariant in each of the first  $k$  branches, but variable in at least one of the first  $k-1$  branches. Importantly, because nodes near the root of the tree reflect the most basic evolutionary divergences, residues ranked high (1, 2, 3,...) correlate with the most fundamental features of the protein. As the rank increases, class specificity is linked with increasingly minor evolutionary divergences, until it eventually loses significance after a given signal-to-noise threshold. The tree can therefore define a hierarchy of functional importance among residues in the protein. The final step in ET is to map the top-ranked residues onto the structure and then to assess whether they are spatially clustered. Trace residues that form clusters in the three-dimensional structure of the protein are likely to define active sites, as changes in the amino acid composition of these regions are linked with evolutionary divergence and, hence, functional specificity [35]. Recent variations on this basic scheme add quantitative descriptions of residue variability (weighted evolutionary tracing [36]), use amino acid exchange matrices to better tolerate variations within branches, utilize better methods to build trees (ConSurf [37]) and account for gaps in sequence alignments [38\*].

Deeper concerns that a tree inadequately accounts for convergent evolution and for different evolutionary rates at

Figure 1



The ET method. (a) All of the sequences in a protein family are aligned and a tree is generated to illustrate the relatedness of individual family members. The tree can then be delineated into groups ( $i$ ) approximating functional classes (in this case, three classes). For each class, a consensus sequence is created and these are then compared to form the ET sequence. Residue positions that are invariant within each class, but that vary among them are called class-specific or trace residues (labeled X in the ET sequence, colored red) and those that are class-specific at rank  $i = 1$  are denoted by amino acid single-letter

code in the ET sequence and colored blue. The number of classes into which the tree has to be divided for a residue to become class-specific is called the rank of that residue. Finally, trace residues are mapped onto the three-dimensional structure of a family member, with clusters of trace residues indicating a functional site [yellow line in panel (b)]. (b) The process described in (a) can be repeated from rank 1 to  $N$  ( $N$  = total number of sequences), so that each residue position is assigned a rank. Residues with lower numbered ranks are considered to be more important than those with higher numbered ranks.

different functional sites and in different branches have spawned related approaches that seek to eliminate it altogether. Thus, Hannenhalli and Russell [39<sup>\*</sup>] used experimentally predefined functional subtypes, rather than a tree, to group sequences and then identified class-specific residues using HMMs and relative entropy. Landgraf *et al.* [40<sup>\*</sup>] further added a variability window around each residue, defined structurally as the spatially neighboring amino acids, and then identified important residues as those whose variations deviate significantly from the average. Nevertheless, the tree has important advantages. It conveys simply and transparently the complex interplay between sequence and function variations. Thus, once top-ranked residues are identified, their variations can be tracked along the tree to identify which variations are the most likely to impart functional specificity to each different functional group during evolution [2<sup>\*\*</sup>,41,42]. It can also be easily modified to test whether a given site is conserved in any number of subgroups within a protein family [5<sup>\*\*</sup>,41] (Figure 2).

Ultimately, methods that predict functional sites should be judged along three different lines: do they efficiently guide mutational, protein engineering and other studies to the relevant sites of a protein structure; are they statistically significant; and can they be applied to the proteome at

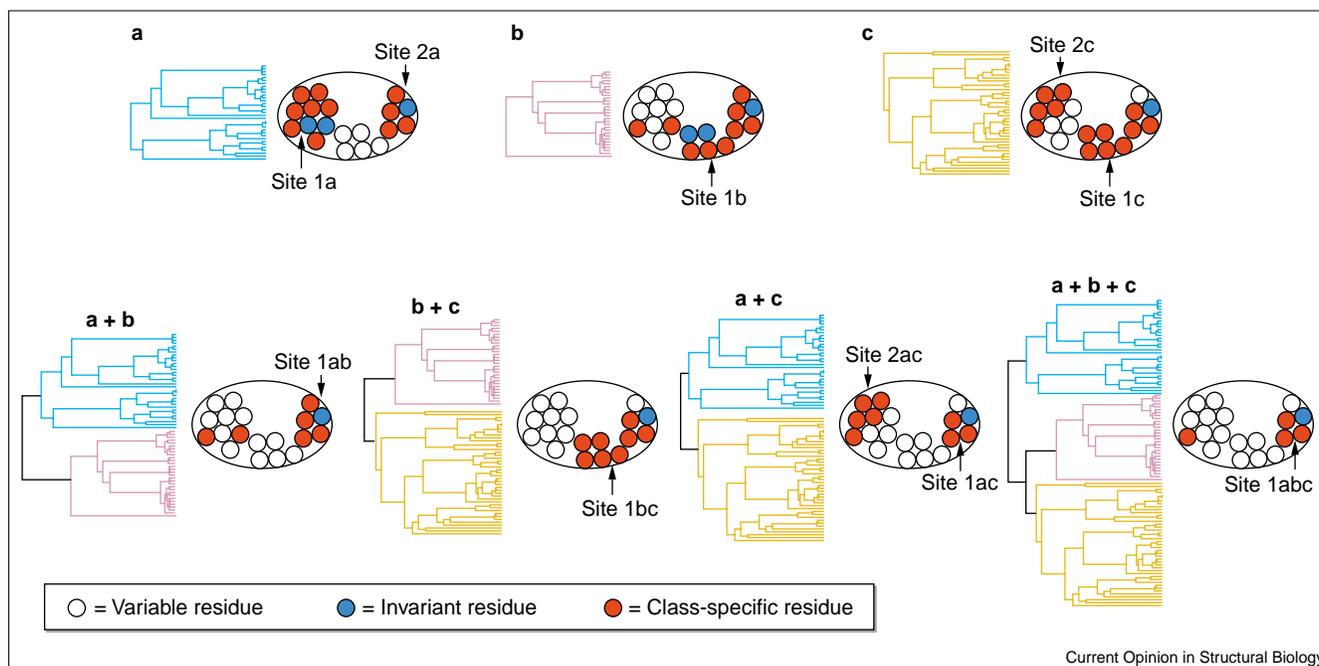
large? Increasing evidence suggests that ET-based methods fulfill these criteria.

## Guiding experiments to biologically relevant sites

### Controls

Control studies in SH2, SH3 [35] and nuclear hormone receptor zinc finger domains [41] show that residues of increasing evolutionary rank form enlarging structural clusters that match precisely the structurally determined ligand interfaces. Then, as the rank increases further, residues begin to scatter across the protein surface, defining the signal-to-noise threshold. Comparisons with the literature show that mutations of the best-ranked residues eradicate activity, mutations at lesser-ranked residues modulate it and mutations at the worst-ranked residues do not change activity, even when they are part of the structural epitope. Finally, in a striking verification of the fundamental hypothesis of ET, the top-ranked trace residues from the zinc finger domains form two groups. The first group consists of the highest ranked residues. They are nearly invariant, contact only invariant bases in the DNA response element and thus appear to define the essential characteristics of DNA binding. The second group includes slightly lower ranked residues. These undergo frequent highly nonconservative substitutions and they

Figure 2



ET can be used as a functional filter: 'subgroup analysis'. For protein families that are large and divergent, ET analysis of the entire family may miss functional sites that are specific to individual subgroups. In this example, one large family can be divided into three main subgroups (a–c). Separate traces of these subgroups reveal some overlapping and some subgroup-specific functional surfaces (labeled as site 1a, etc.). Trees from subgroups can then be added together to determine whether subgroup-specific functional sites are maintained, altered or

abolished, indicating that the groups share very similar functions, slightly similar functions or no functions, respectively. For example, subgroups a and b share one functional site (site 2a/part of site 1b), indicating that this location plays a common role in both. However, site 1a is not shared by subgroup b, but is shared by subgroup c (a+c), indicating that, at this location, a and c are more similar than a and b, or b and c. This process of subgroup analysis is an efficient way to take advantage of the tree in order to delineate functional sites in a large protein family.

contact variable bases, consistent with a key role in binding specificity.

### Predictions and confirmations

*Bona fide* predictions, subsequently followed by experimental verification, were made for G protein  $\alpha$  subunit ( $G_\alpha$ ) proteins and for the regulator of G protein signaling (RGS) proteins that act to increase  $G_\alpha$  GTP hydrolysis rates. ET analysis of the  $G_\alpha$  family identified a functional surface, A1, from which the C-terminal tail of  $G_\alpha$  extends [43]. As this tail had already been linked to receptor specificity, A1 was predicted to be part of the  $G_\alpha$ -receptor interface. 108 alanine mutants in this region were constructed and assayed, showing an overall agreement with the ET prediction of 68% (sensitivity = 75%, specificity = 65%) [44]. However, the most complete demonstration that ET can anticipate both mutational and crystallographic analyses was the RGS study. ET analysis of 42 members of the RGS family revealed a novel functional surface, R2, located next to, but distinct from, the interface between the RGS and  $G_\alpha$ . Based on both the pattern of amino acid variation in R2 among RGS proteins that are enhanced and inhibited by the G protein effector subunit PDE $\gamma$ , and the proximity of R2 to a cluster of class-specific residues in  $G_\alpha$  that contains residues involved in the PDE $\gamma$  interaction, it was predicted

that the effector would bind the RGS- $G_\alpha$  complex by straddling both  $G_\alpha$  and R2 [42]. Mutagenesis of the RGS catalytic core domain, based on the ET prediction, revealed that three residues out of the six selected for mutagenesis had profound effects on the regulation of activity by PDE $\gamma$ . Two of the three residues do not directly interact with  $G_\alpha$ , whereas the third residue contacts the G protein, indicating that there is some form of allosteric communication among these residues [2\*\*]. After the ET-based mutagenesis was completed, the crystal structure of RGS9- $G_{i/t\alpha}$ -PDE $\gamma$  was solved by Slep *et al.* [45] and confirmed the predicted position of the PDE $\gamma$  interaction site on the RGS domain.

### Statistics and large-scale development

These studies suggest that ET can be an efficient and powerful tool for understanding protein function if it can be applied to a significant fraction of the proteome. Madabushi *et al.* [38\*] addressed this issue by streamlining input preparation, developing formal statistics to assess the significance of trace clusters and testing performance on proteins with diverse folds, structures and evolutionary history. Input was simplified by treating gaps as an artificial amino acid type and, as a result, nearly all sequences can be tolerated (sequence fragments still need to be excluded).

Random distributions of both the number of clusters and the size of the largest cluster expected if trace residues arose by chance were generated. These statistics were then applied to traces of 46 proteins. According to these criteria, 95% of the traces were found to be significant, with the largest cluster being significant in 92% of the proteins ( $p$  value  $\leq 0.05$ , each time). Moreover, the largest cluster of trace residues directly contacted the ligand in 39 of the 40 proteins for which the co-crystal structure was available. This demonstrates that the spatial clustering of evolutionarily important residues is a general phenomenon, consistent with the cooperative nature of residues that determine structure and function.

In practice, these results also suggest that ET will identify functional sites in nearly all protein structures. Similar findings were found in control studies of methods closely related to ET. Hannehalli and Russell [39•] used a HMM and amino acid exchange matrices to correctly assign subtypes to 91.2% of 2593 sequences from the PFAM protein family database at a 20% similarity threshold and to 94% of these proteins at a 30% similarity threshold. Landgraf *et al.* [40•] used a three-dimensional residue clustering analysis, which relies on a structure and a multiple sequence alignment, to identify binding interfaces in 35 protein families with four main types of interactions (protein–DNA/RNA, protein–small ligand and both stable and transient protein–protein). For sequences with an  $E$ -score (i.e. the number of sequences expected to be homologous to the query sequence by chance) threshold of  $10^{-20}$ , 67% of the interfacial residues were identified, assuming that 10% of the high scores might occur randomly. Aloy *et al.* [5••] used a modified version of ET, focusing only on invariant polar residues, to locate functional sites in a test set of 86 proteins with  $\leq 30\%$  identity. Using this method, an identified site with more than 50% overlap with the known active site was found in 79% of the proteins, with 15% of the sites overlapping less than 50% with the known site and 6% having no overlap. Other trace studies and predictions further attest to the generality of these methods ([36,46–48]; see also Update).

## Future directions

### Functional annotation

The ability to combine sequence and structure information to identify functional sites on a large scale is likely to be useful for functional annotation. Currently, less than 1% of annotated sequences in SWISS-PROT reflect experimental observations. The remaining annotated sequences (about 4%) rely on homology-based annotation [49]. This is a concern because homology does not necessarily imply functional similarity [50] and because incorrectly assigned functions can easily propagate to other homologs [15].

One may address this problem by tracing homologous protein families separately, as well as together, and then examining whether they share identical functional surfaces [5••]. This is a special case of a general strategy to test

whether any two (or  $i$ ) evolutionary branches of a family share common functional sites by tracing a subtree of the entire family that only contains these two (or  $i$ ) branches, as shown in Figure 2. For example, in DNA-binding domains from intracellular hormone receptors (IRs), an ET restricted to steroid receptors reveals a dimerization interface that is typical of that subgroup and that is not observed when all IRs are traced jointly. This interface vanishes whenever a nonsteroid branch is added back to the steroid family, except in the case of all-*trans* retinoic acid IR (RAR), which must therefore use the steroid dimerization interface for some aspect of its function. A similar strategy has been used to identify similarities and differences between G-protein-coupled receptors from class A and class B [1]. A large-scale implementation of this approach was recently shown to be 94% accurate when tested in 88 proteins and confirms that this is a viable strategy to identify which homologs are functionally analogous [5••]. Importantly, although a shared functional site does not imply a common function, it is easy to further compare the sidechains of top-ranked residues to assess whether they are consistent with a conserved function.

### Protein–protein interactions

Quaternary structure modeling is another area in which the identification of functional interfaces is useful. For example, the identification of a probable interface led to a model for the association of the G protein and its receptor [43] that is consistent with mutational analysis [44]. Additionally, a model for the RGS–G $\alpha$ –PDE $\gamma$  heterotrimer predicted by a trace of the RGS family [42] was consistent with the actual crystallographic structure [2••,45]. Aloy *et al.* [5••] have now demonstrated on a large scale that docking solutions are achieved faster and more accurately when the binding partner is targeted towards a site of evolutionary importance. Yet another viable possible application is to determine whether an observed crystallographic interface is functionally relevant or merely the result of crystallization [3,4].

## Conclusions

Functional sites are key targets for manipulating protein activity and cellular behavior, yet they remain difficult to identify because of their compositional and structural heterogeneity. This reflects a multiplicity of possible ligands, the context dependence of sidechain contributions to activity and the allosteric regulation of active sites. Thus, functional sites are sufficiently unique that they cannot be characterized statistically as accurately as with protein-specific mutational analysis. One goal of ET, therefore, is to combine the precision of mutational analysis with the scalability of a computational approach.

ET correlates sequence variations with functional divergence, using an evolutionary tree to approximate functional groupings. In this view, every pairwise variation between sequences is a virtual mutation and each node in the tree is a virtual assay that functionally distinguishes between proteins in each of the daughter branches. Thus,

it becomes clear that ET is a computational filter that sifts through thousands of virtual mutations (all of which yield biologically active proteins) using dozens of virtual assays to identify sequence positions whose mutation leads to a nondeleterious change in function. Although similar to mutational analysis, ET benefits from more mutations and far more assays than are typically available in the laboratory. The sheer number of these evolutionary mutations and assays makes up for the approximate character of the evolutionary tree and explains ET's many applications: targeting experiments to the relevant sites of a protein; modeling quaternary structure; and functional annotation. On the other hand, a limitation is that, in multifunctional proteins, it remains a challenge to deconvolute which function a given residue contributes to most, although subgroup analysis may be helpful in this regard.

Already, ET has been shown to identify statistically significant and functionally relevant regions in more than 80% of the proteins tested. As ever more sequences and structures are solved through large-scale sequencing and structural genomics efforts, ET-based approaches should prove increasingly useful to integrate sequence, structure and function information for a large fraction of the proteome.

## Update

Dean *et al.* [51\*] have recently employed the ET method to analyze over 700 G-protein-coupled receptors (GPCRs), using Monte Carlo techniques to assess the significance of ET-identified residue clusters. For each of the eight GPCR families traced, trace residues cluster in a nonrandom manner and were generally found in helices 5 and 6 (proposed to be involved with receptor dimerization), and in helices 2 and 3 (a novel functional site).

## Acknowledgements

OL gratefully acknowledges the support of the American Heart Association, March of Dimes, NSF and the National Human Genome Research Institute (NHGRI). MES is supported by the WM Keck Center for Computational Biology and National Institutes of Health grant R01-EY11900.

## References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- of special interest
  - of outstanding interest
1. Lichtarge O, Sowa ME, Philippi A: **Evolutionary traces of functional surfaces along the G protein signaling pathway.** *Methods Enzymol* 2001, **344**:536-556.
  2. Sowa ME, He W, Slep KC, Kercher MA, Lichtarge O, Wensel TG:
    - **Prediction and confirmation of a site critical for effector regulation of RGS domain activity.** *Nat Struct Biol* 2001, **8**:234-237.
 The best experimental demonstration to date that the ET technique can anticipate and guide mutational analysis, as well as predict protein-protein interactions. This paper follows-up on the prior predictions described in [42].
  3. Thornton JM: **From genome to function.** *Science* 2001, **292**:2095-2097.
  4. McCammon JA: **Theory of biomolecular recognition.** *Curr Opin Struct Biol* 1998, **8**:245-249.
  5. Aloy P, Querol E, Aviles FX, Sternberg MJ:
    - **Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from**

**homology in genome annotation and to protein docking.** *J Mol Biol* 2001, **311**:395-408.

This paper provides an excellent validation of an automated ET-based method. Its application to 88 pairs of homologs tests whether they are likely to share a common functional site and demonstrates that ET can help guide docking to the relevant interfaces.

6. Stebbins CE, Galan JE: **Structural mimicry in bacterial virulence.** *Nature* 2001, **412**:701-705.
  7. de Rinaldis M, Ausiello G, Cesareni G, Helmer-Citterich M: **Three-dimensional profiles: a new tool to identify protein surface similarities.** *J Mol Biol* 1998, **284**:1211-1221.
  8. Livingstone CD, Barton GJ: **Protein sequence alignments — a strategy for the hierarchical analysis of residue conservation.** *Comput Appl Biosci* 1993, **9**:745-756.
  9. Zvelebil MJ, Barton GJ, Taylor WR, Sternberg MJ: **Prediction of protein secondary structure and active sites using the alignment of homologous sequences.** *J Mol Biol* 1987, **195**:957-961.
  10. Gribskov M, Luthy R, Eisenberg D: **Profile analysis.** *Methods Enzymol* 1990, **183**:146-159.
  11. Krogh A, Brown M, Mian IS, Sjolander K, Haussler D: **Hidden Markov models in computational biology. Applications to protein modeling.** *J Mol Biol* 1994, **235**:1501-1531.
  12. Eddy SR: **Profile hidden Markov models.** *Bioinformatics* 1998, **14**:755-763.
  13. Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, Biswas M, Bucher P, Cerutti L, Corpet F, Croning MD *et al.*: **The InterPro database, an integrated documentation resource for protein families, domains and functional sites.** *Nucleic Acids Res* 2001, **29**:37-40.
  14. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25**:3389-3402.
  15. Brenner SE: **Errors in genome annotation.** *Trends Genet* 1999, **15**:132-133.
  16. Olmea O, Rost B, Valencia A: **Effective use of sequence correlation and conservation in fold recognition.** *J Mol Biol* 1999, **293**:1221-1239.
  17. Todd AE, Orengo CA, Thornton JM: **Evolution of function in protein superfamilies, from a structural perspective.** *J Mol Biol* 2001, **307**:1113-1143.
- This paper provides an extremely detailed study of the evolution of functional sites, specifically focusing on 31 functionally distinct superfamilies selected from the PDB. The authors combined sequence and structural data to reveal that, in general, enzymes within a superfamily have a variety of specificities, while maintaining a conserved reaction mechanism.
18. Russell RB, Sasieni PD, Sternberg MJ: **Supersites within superfolds. Binding site similarity in the absence of homology.** *J Mol Biol* 1998, **282**:903-918.
  19. Elcock AH, Sept D, McCammon JA: **Computer simulation of protein-protein interactions.** *J Phys Chem B* 2001, **105**:1504-1518.
  20. Sheinerman FB, Norel R, Honig B: **Electrostatic aspects of protein-protein interactions.** *Curr Opin Struct Biol* 2000, **10**:153-159.
  21. Kuntz ID, Blaney JM, Oatley SJ, Langridge R, Ferrin TE: **A geometric approach to macromolecule-ligand interactions.** *J Mol Biol* 1982, **161**:269-288.
  22. Pettit FK, Bowie JU: **Protein surface roughness and small molecular binding sites.** *J Mol Biol* 1999, **285**:1377-1382.
  23. Jones S, Thornton JM: **Analysis of protein-protein interaction sites using surface patches.** *J Mol Biol* 1997, **272**:121-132.
  24. Bogan AA, Thorn KS: **Anatomy of hot spots in protein interfaces.** *J Mol Biol* 1998, **280**:1-9.
  25. Clackson T, Wells JA: **A hot spot of binding energy in a hormone-receptor interface.** *Science* 1995, **267**:383-386.
  26. Lockless SW, Ranganathan R: **Evolutionarily conserved pathways of energetic connectivity in protein families.** *Science* 1999, **286**:295-299.

27. Atwell S, Ultsch M, De Vos AM, Wells JA: **Structural plasticity in a remodeled protein-protein interface.** *Science* 1997, **278**:1125-1128.
28. Valdar WS, Thornton JM: **Protein-protein interfaces: analysis of amino acid conservation in homodimers.** *Proteins* 2001, **42**:108-124.
29. Conte LL, Chothia C, Janin J: **The atomic structure of protein-protein recognition sites.** *J Mol Biol* 1999, **285**:2177-2198.
30. Nadassy K, Wodak SJ, Janin J: **Structural features of protein-nucleic acid recognition sites.** *Biochemistry* 1999, **38**:1999-2017.
31. Jones S, van Heyningen P, Berman HM, Thornton JM: **Protein-DNA interactions: a structural analysis.** *J Mol Biol* 1999, **287**:877-896.
32. Luscombe NM, Laskowski RA, Thornton JM: **Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level.** *Nucleic Acids Res* 2001, **29**:2860-2874.
33. Jones S, Daley DT, Luscombe NM, Berman HM, Thornton JM: **Protein-RNA interactions: a structural analysis.** *Nucleic Acids Res* 2001, **29**:943-954.
34. Taroni C, Jones S, Thornton JM: **Analysis and prediction of carbohydrate binding sites.** *Protein Eng* 2000, **13**:89-98.
35. Lichtarge O, Bourne HR, Cohen FE: **An evolutionary trace method defines binding surfaces common to protein families.** *J Mol Biol* 1996, **257**:342-358.
36. Landgraf R, Fischer D, Eisenberg D: **Analysis of heregulin symmetry by weighted evolutionary tracing.** *Protein Eng* 1999, **12**:943-951.
37. Armon A, Graur D, Ben-Tal N: **ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information.** *J Mol Biol* 2001, **307**:447-463.
38. Madabushi S, Yao H, Marsh M, Kristensen DM, Philippi A, Lichtarge O: **Structural clusters of evolutionary trace residues are statistically significant and common in proteins.** *J Mol Biol* 2002, in press.
- In an effort towards large-scale application, the authors streamlined the preparation of ET input by introducing tolerance to gaps in multiple sequence alignments. They also defined statistics to objectively assess the significance of clusters of trace residues. Significant trace residue clusters are found in 45 of the 46 proteins tested.
39. Hannehalli SS, Russell RB: **Analysis and prediction of functional sub-types from protein sequence alignments.** *J Mol Biol* 2000, **303**:61-76.
- The authors have developed a method for identifying the determinants of already known functional domains in proteins. By grouping proteins in a multiple sequence alignment using various criteria, the pattern of residue variation can be examined and then linked to functional specificity using HMMs.
40. 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.
- This paper aims to extend ET by focusing directly on the structure and eliminating the tree from consideration. The authors can identify clusters of residues whose evolutionary variations deviate significantly from the protein average. This is sufficient to identify two out of three interfacial residues with moderate background noise.
41. 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.
42. Sowa ME, He W, Wensel TG, Lichtarge O: **A regulator of G protein signaling interaction surface linked to effector specificity.** *Proc Natl Acad Sci USA* 2000, **97**:1483-1488.
43. Lichtarge O, Bourne HR, Cohen FE: **Evolutionarily conserved G $\alpha$  $\beta\gamma$  binding surfaces support a model of the G protein-receptor complex.** *Proc Natl Acad Sci USA* 1996, **93**:7507-7511.
44. Onrust R, Herzmark P, Chi P, Garcia PD, Lichtarge O, Kingsley C, Bourne HR: **Receptor and  $\beta\gamma$  binding sites in the  $\alpha$  subunit of the retinal G protein transducin.** *Science* 1997, **275**:381-384.
45. Slep KC, Kercher MA, He W, Cowan CW, Wensel TG, Sigler PB: **Structural determinants for regulation of phosphodiesterase by a G protein at 2.0 Å.** *Nature* 2001, **409**:1071-1077.
46. Pritchard L, Dufton MJ: **Evolutionary trace analysis of the Kunitz/BPTI family of proteins: functional divergence may have been based on conformational adjustment.** *J Mol Biol* 1999, **285**:1589-1607.
47. Pascual J, Martinez-Yamout M, Dyson HJ, Wright PE: **Structure of the PHD zinc finger from human Williams-Beuren syndrome transcription factor.** *J Mol Biol* 2000, **304**:723-729.
48. 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.
49. Karp PD, Paley S, Zhu J: **Database verification studies of SWISS-PROT and GenBank.** *Bioinformatics* 2001, **17**:526-532; 533-534.
50. Benner SA, Gaucher EA: **Evolution, language and analogy in functional genomics.** *Trends Genet* 2001, **17**:414-418.
51. Dean MK, Higgs C, Smith RE, Bywater RP, Snell CR, Scott PD, Upton GJG, Howe TJ, Reynolds CA: **Dimerization of G-protein coupled receptors.** *J Med Chem* 2001, **44**:4595-4614.
- This paper extends the conventional ET analysis by introducing Monte Carlo techniques to assess class-specific residue clustering. The authors used ET to identify a possible receptor dimerization interface that would have important ramifications for drug design. Furthermore, this paper contains an excellent review of GPCR mutational studies.