doi:10.1006/jmbi.2001.5327 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (2002) 316, 139–154

JMB

Structural Clusters of Evolutionary Trace Residues are Statistically Significant and Common in Proteins

Srinivasan Madabushi[1,2], Hui Yao[1,2], Mike Marsh[1,2]
David M. Kristensen[1,2], Anne Philippi[2], Mathew E. Sowa[1,3]
and Olivier Lichtarge[1,2]*

[1]*Structural and Computational Biology and Molecular Biophysics Program*

[2]*Department of Molecular and Human Genetics*

[3]*Department of Biochemistry and Molecular Biology, Baylor College of Medicine, One Baylor Plaza, Houston TX 77030, USA*

**Corresponding author*

Given the massive increase in the number of new sequences and structures, a critical problem is how to integrate these raw data into meaningful biological information. One approach, the Evolutionary Trace, or ET, uses phylogenetic information to rank the residues in a protein sequence by evolutionary importance and then maps those ranked at the top onto a representative structure. If these residues form structural clusters, they can identify functional surfaces such as those involved in molecular recognition. Now that a number of examples have shown that ET can identify binding sites and focus mutational studies on their relevant functional determinants, we ask whether the method can be improved so as to be applicable on a large scale. To address this question, we introduce a new treatment of gaps resulting from insertions and deletions, which streamlines the selection of sequences used as input. We also introduce objective statistics to assess the significance of the total number of clusters and of the size of the largest one. As a result of the novel treatment of gaps, ET performance improves measurably. We find evolutionarily privileged clusters that are significant at the 5 % level in 45 out of 46 (98 %) proteins drawn from a variety of structural classes and biological functions. In 37 of the 38 proteins for which a protein-ligand complex is available, the dominant cluster contacts the ligand. We conclude that 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 suggest that ET can be applied on a large scale to identify functional sites in a significant fraction of the structures in the protein databank (PDB). This approach to combining raw sequences and structure to obtain detailed insights into the molecular basis of function should prove valuable in the context of the Structural Genomics Initiative.

Keywords: bioinformatics; structural genomics; protein interaction; active site evolution; ligand binding

Introduction

As the production of putative genes from whole genome shotgun sequencing continues to grow exponentially, the Structural Genomics Initiative (SGI) aims to produce an equally daunting number of new protein structures, thus raising an important question as to how this mass of raw data will be transformed into biologically meaningful information. One approach, the Evolutionary Trace (ET),[1,2] combines both sequence and structure information to identify the location and specificity determinants of functional sites in proteins. This can be immediately useful because functional surfaces mediate all protein-ligand interactions. Knowledge of their location and specificity determinants helps target mutagenesis to the relevant residues of a protein in order to understand the molecular basis of function,[3,4] and it is also useful for drug design and for engineering new, desirable properties into proteins.[5]

Abbreviations used: ET, evolutionary trace; PDB, Protein Data Bank; SGI, Structural Genomics Initiative.

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

0022-2836/02/010139–16 $35.00/0

Determining functional sites on the surface of a protein is not trivial, however. The best approach to date is mutational analysis whereby individual residues are altered one at a time and the mutant protein's various functions are subsequently assayed. This can be slow and costly and, critically, it also requires that accurate assays of each of the protein's multi-faceted functions be available. This requirement can be problematic, since the functions of many proteins are either not known or are only partially known. Nevertheless, elegant studies have shown that this strategy is highly successful at delineating functional hot spots, and reveal that these are often a smaller subset of the entire set of interfacial contact residues seen in crystal complexes.[6]

As an alternative to the protein-specific nature of mutational analysis, a number of laboratories have sought general computational approaches to characterize functional sites. Analyses that seek motifs, either based on complementarity of shape[7] or of charge,[8] on empirical energy functions,[9] including the energetic links between proteins and ligands[10] or the energetic effect of substitutions,[11] are all best done in the setting of an already identified binding partner. A second set of methods seeks to predict functional surfaces *de novo*. Casari *et al*. developed an algebraic method termed principal component analysis that treats proteins as vectors in a sequence space.[12] Henikoff & Henikoff seek block-like motifs and then measure the chance distribution of their matches against the SWISS-PROT database.[13] Jones & Thornton use physico-chemical descriptors of surface residues to score their probability of interacting in protein-protein interactions[14] based on a database of interfaces. Shatsky and colleagues align hinge regions and flexible regions of proteins by detecting maximal congruent rigid fragments so as to obtain their optimal arrangement and to identify functional interfaces.[15] At their core, most of these methods focus on measuring global aspects of residue conservation as a marker of importance.

In contrast, ET seeks to identify local patterns of conservation and global patterns of variation that intrinsically indicate functional or structural importance. This method uses a phylogenetic tree derived from a multiple sequence alignment to approximate the functional clustering of family members. By partitioning the tree into distinct branches (deemed equivalent to functional classes), consensus sequences can be generated for each one and then compared. Residue positions that are invariant within each branch but variable among them are termed trace or class-specific residues. By construction, these class-specific residues are closely coupled to evolutionary divergence and hence, presumably, to functional importance. The minimum number of branches into which the tree has to be divided in order for a residue to become class-specific is termed the rank of that residue. Top ranks (1, 2, 3,...) indicate residues that have become fixed within each of the most ancient evolutionary clades of a family suggesting a fundamental link to function, whereas low ranks $(\ldots, n-2, n-1, n;$ where n is the maximum number of sequences in the family) indicate residues that vary even among the most closely related of proteins suggesting they have little impact on function.

The ET was tested extensively in SH2 and SH3 modular signaling domains,[2] type-II zinc fingers from nuclear hormone receptors,[16] heterotrimeric G proteins,[1] RGS proteins,[3] and G protein-coupled receptors.[17] Significantly, the G protein and RGS traces are *bona fide* predictions that were only later confirmed by mutational and crystallographic analysis.[4,18,19] Other laboratories have successfully used ET in BPTI,[20] heregulin,[21] TGFβ and related growth factors,[22] and in PHD zinc fingers[23] occasionally proposing variations on the basic scheme of the trace.[21,24,25]

Since ET is demonstrably useful in these examples, can it now be applied to proteins on a large scale? There are two major bottlenecks in the way. First, the selection of homologues that make up the input to ET is complicated by the need to minimize gaps in their multiple sequence alignment. This is because ET eliminates all positions with gaps from further analysis (the rationale being that a position cannot be functionally critical to all the proteins in the alignment if it can be removed altogether from some). Since insertions and deletions are common, this creates a dilemma, as ET traces more sequences it covers a diminishing percentage of the protein under study. The compromise is to selectively remove sequences that introduced the most gaps in the alignment. One drawback is that this step is subjective, another is that it also deprives ET of the evolutionary information in those sequences. Here, we introduce a novel, gap-tolerant trace, by adopting the convention that gaps are a virtual 21st amino acid type.

A second bottleneck is that ET requires visual interpretation of its results. The user must recognize, by eye, clusters of top-ranked residues in 3D space and visually estimate their significance based on the level of scattered signal throughout the protein. A few large clusters would be interpreted as true signal, while many, small clusters scattered homogeneously about the protein would indicate noise. Although this evaluation is fairly straightforward, it too is subjective, especially near the signal-to-noise threshold. To replace human assessment we use two statistics, the overall number of clusters and the size of the largest cluster. Random sampling of residues in several structures allows us to estimate the distributions of these statistics, and to measure the significance of the actual ET-generated values.

The novel gap-tolerant trace reaches statistical significance by at least one of the two clustering statistics in 45 of the 46 proteins we tested. Using only the number of cluster statistics, 44 protein traces are significant, and if we use the largest cluster statistics only, 42 are significant. By comparison using the older gap-intolerant method, only 33 traces and 36 traces reach significance using the cluster number or the cluster size criteria, respectively. Moreover, in the subset of 38 proteins for which a ligand is present in the crystallographic structure, a significant cluster was identified at the ligand-binding site for 37 proteins. These results establish this novel ET as a general method that is applicable on a large scale and that can identify evolutionarily privileged sites on protein structures that are both statistically significant and functionally relevant.

Results

Distribution of the number of clusters

The significance of an ET analysis can be obtained by comparing clusters formed by trace residues to clusters created when an equal number of residues are randomly chosen. For example, a trace of pyruvate decarboxylase (1pvd) reveals two large clusters confined predominantly to one face of the protein, as shown in Figure 1, where (a) to (d) are rotated by increments of 90 ° about a vertical axis. In contrast, a random sampling of the same number of residues yields many smaller clusters distributed uniformly over the surface, as shown in Figure 1(e)-(h). From here onwards, a cluster is defined as a collection of residues such that at least one atom (not counting hydrogen atoms) of each member of the cluster is within 4 Å of another member of the cluster. Also, in order to compare statistics across proteins, we find it useful to define coverage as a fraction which is the number of trace residues at and above a given rank, divided by the total number of trace residues at the maximum possible rank.

Random distributions were established in 12 proteins ranging from 80 to 537 residues in length. For each, a number of residues representing 5 %, 10 %, 15 %, 20 %, 25 %, and 30 % coverage were picked randomly and repeatedly 5000 times. This allowed us to estimate the random distributions of the number of clusters expected by chance (see Methods). In turn, we could then test whether an observed number of clusters was consistent with the null hypothesis that they arose by chance (H_0) and, if not, determine the specific level of confidence at which we could reject H_0. As shown in pyruvate decarboxylase (1pvd) (Figure 2), when 81 of its 537 residues are picked randomly over 5000 trials (for a coverage of nearly ~15 %), the number of clusters formed each time forms a histogram distributed unevenly about 38, with extremes ranging from 25 to 55. In 99.7 % of the trials, the number of generated clusters is greater than 26, so that if a trace yields fewer than 26 clusters, we could reject H_0 at a significance level (p-value) of 0.3 %. Similarly, the number of cluster thresholds to reach significance levels of 1 %, 5 % and 30 % are 27, 30, and 35, respectively. The trace of pyruvate decarboxylase identifies a total of ten clusters and the probability that this reflects random chance is much less than 0.3 %.

The significance thresholds depend linearly on protein size

Remarkably, for a given coverage, the threshold of the number of clusters needed to reach a given level of significance is nearly a linear function of protein size. This is demonstrated in the inset of Figure 2, which shows the number of clusters corresponding to the 1 % significance threshold for each of the 12 proteins as a function of their size, at 15 % coverage. This relationship holds for all of the defined significance levels at coverages ranging from 5 % to 30 % of the protein and even beyond (data not shown). In addition, the high quality-of-fit values, R^2 (ranging from 0.89-0.99), suggest that the specific choice of these 12 proteins is unlikely to bias the results. Thus, instead of building individual clustering distributions for each protein under study, many computational cycles can be saved by using a generic lookup table† with the information from the linear fit in order to determine the significance level for an observed number of clusters identified by the trace. For example, using the inset of Figure 2, if the top 55 trace residues in a 370 amino acid residue protein (~15% coverage) formed 17 clusters or fewer, they would achieve a significance at the 1% level.

Distribution of the size of the dominant cluster

To further understand the significance of individual trace clusters, we also compared the size of the largest, dominant cluster to the size expected due to chance. For this purpose, we built distributions of the largest cluster size using the method already described above, and shown in Figure 4 for protein pyruvate decarboxylase (1pvd) at 15 % coverage. Typically the largest cluster contained eight residues, with sizes ranging from four to 34 residues over 5000 trials. In order to achieve a significance level of 30 %, 5 %, 1 %, or 0.3 %, the largest trace cluster would have to comprise at least 11, 19, 26, or 30 residues, respectively. In fact, the largest cluster traced in Figure 1 includes 74 residues, and thus achieves a significance much better than 0.3 %.

Here again, for a given coverage and as shown in the inset of Figure 4, the threshold for the size of the largest cluster needed to reach a given level of significance is nearly a linear function of protein size. This relationship holds for all levels of significance up to 40 %, and the high quality-of-fit R^2 values (ranging from 0.77-0.97) allowing us to apply this linear relationship to other proteins. For example, using the inset of Figure 4 showing the linear fit thresholds for 20 % coverage at the 0.3 % level of significance, a 250 amino acid residue

† http://imgen.bcm.tmc.edu/molgen/labs/lichtarge/ET-stats

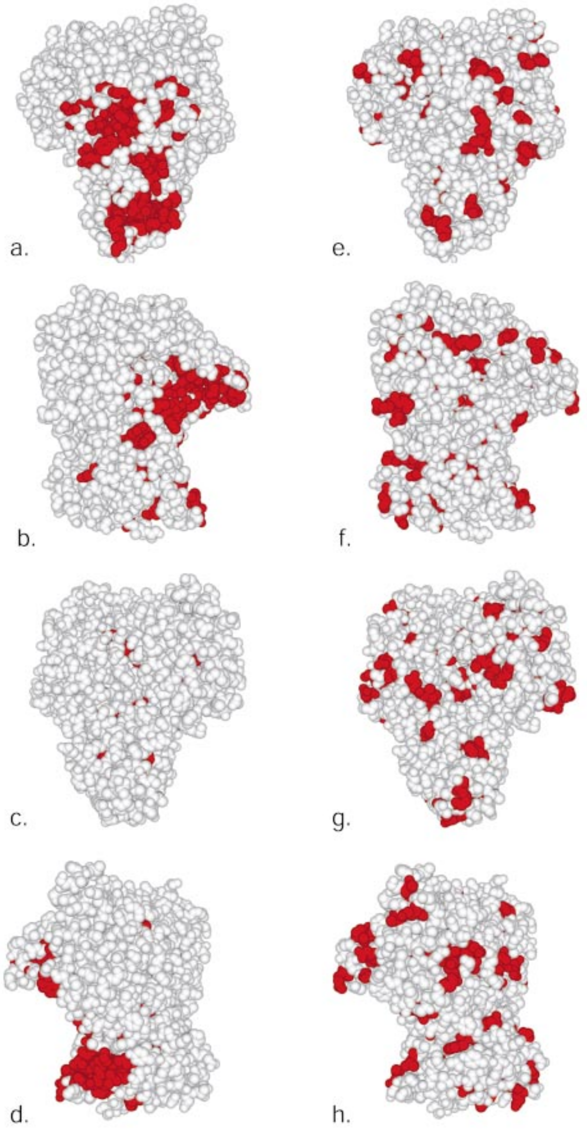

Figure 1. Trace residues cluster non-randomly. In pyruvate decarboxylase (shown here), and in general, trace residues tend to form a small number of large clusters ((a)-(d)); (a)-(d) are rotated with respect to each other by 90 ° about the *y*-axis), while an equivalent number of randomly selected residues form many small clusters scattered homogeneously throughout the protein ((e)-(h)); (e)-(h) are rotated in the same manner as (a)-(d). The trace residues shown here correspond to those identified at rank 10, or 20 % coverage of the protein (PDB identifier: 1pvd), where 90 residues are predicted to be important by ET.

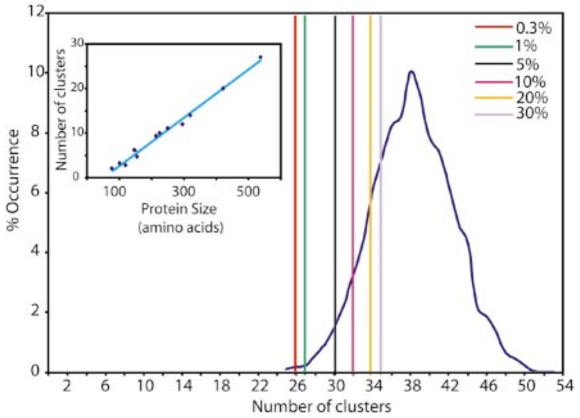

Figure 2. The random distribution of the expected number of clusters can be used to establish significance thresholds: 15 % of the residues of pyruvate decarboxylase (1pvd) were randomly selected and the number of residue clusters thus generated was counted. This was repeated 5000 times and the resulting histogram is shown here. Note that we have used a best fitting smoothed envelope to model discrete data points. The significance thresholds are shown as vertical lines on the distribution and are colored accordingly (see Methods). Inset: The linear relationship between protein size and the number of clusters predicted by random simulations. Each point represents 5000 random simulations performed on a different protein (12 proteins in all) at 15 % coverage with a significance threshold of 1 %.

protein would achieve significance at the 0.3 % level if its largest cluster contained at least 33 residues.

Treatment of gaps

We treated gaps as if they were a 21st amino acid type. Here the convention that a gap can be interpreted in the same way that Ala, Val, or any of the other 20 amino acid positions is not meant to carry biophysical meaning. It is simply a computational device, which is reasonable because gaps often occur in blocks in a multiple sequence alignment. These blocks indicate that a deletion or insertion took place that was then conserved in all protein descendants, suggesting some functional importance at the location of those gaps. In practice, the ability to rank gapped positions eliminates ''holes'' from ET analyses: now at maximum coverage all the residues in the structure are ranked. To evaluate this new approach, we compared it to the original, gap-intolerant method, where gapped positions are excluded from the trace and remain unranked.

Statistical significance of evolutionary traces

We tested the novel ET method in 46 proteins and found that in all but two the observed number of clusters reaches a significance level of 5 % or better for at least one level of coverage (96 % success rate). In fact, almost two-thirds (30 out of 46, or 65 %) of the traces reach a significance level of 0.3 % or better, as shown in Table 1. The first protein in which the trace failed reach the 5 % significance level is the ligand-binding domain of the LDL receptor. This trace involved essentially no pruning of its sequence family, so that 184 homologues were traced together, and the protein itself is

Table 1. The trace reaches statistical significance in 45 of 46 proteins

	Number of clusters				Size of largest cluster			
	With gaps		Without gaps		With gaps		Without gaps	
Significance level (%)	Number of proteins	Fraction (%)	Number of proteins[a]	Fraction (%)	Number of proteins	Fraction (%)	Number of proteins[a]	Fraction (%)
<0.3	30	65	26	57	34	74	29	63
0.3-1	6	13	3	7	5	11	6	13
1-5	8	17	4	9	3	7	1	2
5-10	0	0	6	13	3	7	2	4
10-20	0	0	2	4	1	2	1	2
20-30	1	2	1	2	0	0	4	9
>30	1	2	3	7	0	0	2	4

For each significance level, the number of proteins achieving that significance at least once at some level of coverage is shown for both the number of clusters and the size of largest cluster statistical method. For each method, results are given both for the case where gaps in the multiple sequence alignment caused residue positions to be ignored (without gaps) and for the case where those gaps were treated as a "virtual" amino acid type (with gaps). The number of proteins is expressed as a fraction of the total number of proteins in the test set.

[a] Of the 46 proteins in the test set, one protein, growth hormone receptor, did not reach any significance level in the case where gaps were ignored due to the fact that none of the rank values for this protein were directly convertible to any of the standard coverage levels.

very small (37 residues) which favors the natural clustering of randomly drawn residues. Interestingly the cluster size statistic discussed below shows that this trace is nevertheless significant. The other trace that does not reach significance is the biotinyl domain involved in fatty acid synthesis. In comparison, the original gap-intolerant method yields statistically significant traces in fewer cases. A total of 33 out of 46 traces (or 72 %, down from 96 %) reach a significance level of 5 % or better at least once (Table 1) at some coverage between 5 and 30 % (Table 2), and 26 achieve a significance level of 0.3 % or better (or 57 % down from 65 %).

These results can be further broken down for specific coverages. Thus, among the 22 proteins that have a rank corresponding to nearly 5 % coverage, 45 % are significant (the other 24 proteins have ranks that correspond to coverages other than 5 %). For the 33 proteins that have a rank near 10 % coverage, 79 % achieve significance. At coverages of 15, 20, 25, and 30 %, the number of proteins with matching ranks are 32, 32, 35, and 35, respectively, and the fractions that achieve the 5 % significance level are 75, 91, 74 and 49 %, respectively (Table 2).

Nearly similar results are obtained by considering the size of the dominant cluster. Using the novel trace approach, 42 of 46 proteins (91 %) reached the 5 % significance level, and 34 (74 %) reached the 0.3 % significance level, as shown in Figure 5 and in Table 1). As before, the statistics worsen somewhat when gaps are not counted in the trace. Nevertheless 36 (78 %) of the 46 proteins reach the 5 % significance level, and 29 (63 %) are at the 0.3 % significance level.

As above, these results can be further broken down in terms of the coverage as shown in Table 3. The number of proteins with matching ranks at 5, 10, 15, 20, 25 and 30 % coverages are 22, 33, 32, 32, 35 and 35, respectively. Specifically, a significance level of 5 % was achieved at 5 % coverage in 50 % of the proteins with matching ranks, 76 % at 10 % coverage, 72 % at 15 % coverage, 75 % at 20 % coverage, 57 % at 25 % coverage, and 71 % at 30 % coverage.

It is important to stress that the proteins listed in Table 1 and Table 4 are the totality of those examined during this study. The alignments and the corresponding phylogenetic trees were not adjusted or optimized in any way (see Methods) after the respective proteins were chosen to be included in the test set. In fact, the sets of homologous sequences that make up the input to the trace were minimally optimized: obviously fragmented sequences were removed, as were single outlying sequences that diverged from the main protein of interest near the root of the family tree. Our analyses therefore should be representative of the results expected from a novice rather than expert ET.

Signal-to-noise threshold

The ET signal-to-noise threshold varies among proteins. Initially, at top ranks, relatively few residues are class-specific, therefore, coverage is low and the trace residues may be too sparse to make direct contacts (within 4 Å), and thus they may not cluster significantly. Thus, as shown in Table 2, relatively few proteins achieve significance at 5 % coverage. As we lower the rank threshold, more residues become class-specific and these tend to fill the gaps between top-ranked residues traced earlier, thereby coalescing many small clusters into fewer, larger ones. This reflects the tendency of ET clusters to expand outward from small cores of critically important residues. In keeping with this scenario, most traces reach significance between 15 and 25 % coverage. Eventually, when coverage reaches 30 to 35 %, so many residues are being considered that they will cluster even when picked randomly, and the significance of trace clusters diminishes as seen in Figure 3 and in Table 2. The threshold at which the trace residue clusters cease to be significant varies, most often in the 20 to 35 % coverage range.

Ligand contacts

Among the 38 structures in the test set that had bound ligands, 36 (95 %) traces reached significance with the new method, compared to 30 (79 %) which reached significance with the gap-intolerant approach. In one more protein, HIV reverse transcriptase, ET identified a few residues contacting the ligand, but the clusters were not significant. However, with additional pruning of the original 278 sequences, that trace became significant as well. In all but one of these cases, the known ligand(s) directly contacted a trace cluster as shown for two representative traces in Figure 6. ET identifies some but not necessarily all of the residues contacting the ligand, consistent with the fact that not all interfacial residues are important for ligand binding.[6] This is also a function of the rank, or coverage, since top ranked residues should represent the most critical determinants of molecular recognition, while somewhat lesser ranked residues may play a role in specificity. The overall results for the 38 proteins having bound ligands are shown in Table 4. Despite the lack of optimization of these traces, it is noteworthy that ET provides biologically relevant information in 37 out of 38 cases and in 36 of those cases, that information is also statistically significant.

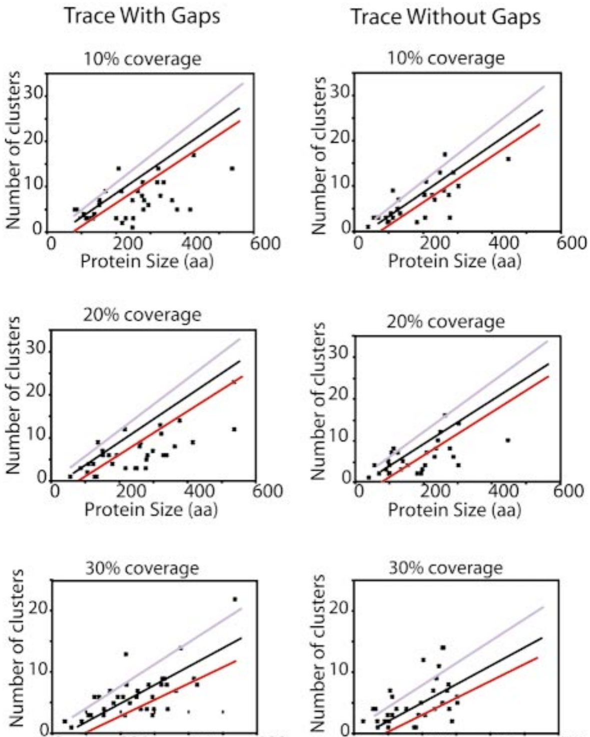

Figure 3. Significance of ET predictions using the number of clusters statistics. For 10, 20, and 30 % coverages, the number of clusters identified by ET was plotted against protein size for each of the 46 proteins with a rank directly convertible to a coverage level. Trace With Gaps refers to ET data generated when considering gaps in the alignment and Trace Without Gaps refers to the ET data generated without this information. The significance thresholds are shown as: 0.3 %, red line; 5 %, black line; 30 %, blue line and were generated using the linear fits from the number of clusters simulations. Most ET clusters have a significance better than 5 % (that is, there is a less than 5 % chance of observing such clusters randomly).

Discussion

In order to assess the significance of functional sites predicted by ET analysis, we carried out a statistical analysis of residue clusters in proteins. One aim was to detect and measure the significance of 3D trace clusters objectively. Another aim was to assess the feasibility of annotating protein functional surfaces on a large scale using ET. To fulfil these goals, we developed an approach based on the fact that random residues will be scattered homogeneously throughout a structure and thus will tend to give rise to multiple clusters of small size rather than a few clusters of large size. Thus, the significance of clusters as predicted by ET can be determined by comparing the number of clusters and the size of the largest cluster with the results expected by random chance. The user can then select the rank with the greatest statistical significance and thus likely to have greater residue clustering. We also introduce a novel approach to insertions and deletions that is more tolerant of gaps in the multiple sequence alignments. This simplifies the preparation of input data and allows us to exploit the evolutionary information in homologous sequences more fully. We found that over all, ET identifies active sites that are significant at a level of 5 % or better in 45 of the 46 (98 %) proteins tested using either number of clusters or largest cluster size criteria. To interpret these results, it is important to note that the proteins tested represent a diverse cross-section of the proteome as well as families that are well populated in sequence space.

Implications for large-scale use of the ET

The widespread identification of evolutionarily privileged clusters of residues that are non-random and that overlap with ligand interaction sites (Figure 6) in such a general and randomly selected test set suggests that most proteins will be amenable to trace analysis provided enough sequences are available in their respective family. The proteins analyzed here have a broad cross-section of structural, functional, and evolutionary characteristics, summarized in Table 4. They participate in metabolic, signaling, transcriptional, and many other pathways where they perform catalysis,

Table 2. Distribution of the significance levels for each level of coverage using the number of clusters statistical method and including gap positions

	Coverage											
	5%		10%		15%		20%		25%		30%	
Significance level (%)	No. of proteins	Fraction (%)	No. of proteins	Fraction (%)	No. of proteins	Fraction (%)	No. of proteins	Fraction (%)	No. of proteins	Fraction (%)	No. of proteins	Fraction (%)
<0.3	6	27	20	61	19	59	19	59	16	46	7	20
0.3-1	1	5	3	9	2	6	4	13	5	14	3	9
1-5	3	14	3	9	3	9	6	19	5	14	7	20
5-10	2	9	2	6	3	9	1	3	2	6	10	29
10-20	1	5	1	3	2	6	0	0	2	6	0	0
20-30	3	14	0	0	1	3	1	3	3	9	2	6
>30	6	27	4	12	2	6	1	3	2	6	6	17
Total	22	100	33	100	32	100	32	100	35	100	35	100

The number of proteins indicates how many had ranks that matched the indicated percentage coverage, for a given significance level. This nomenclature holds true for Tables 2 and 3.

Table 3. Distribution of the significance levels for each level of coverage using the size of largest cluster statistical method and including gap positions

	Coverage											
	5%		10%		15%		20%		25%		30%	
Significance level (%)	No. of proteins	Fraction (%)	No. of proteins	Fraction (%)	No. of proteins	Fraction (%)	No. of proteins	Fraction (%)	No. of proteins	Fraction (%)	No. of proteins	Fraction (%)
<0.3	0	0	17	52	16	50	18	56	19	54	18	51
0.3-1	9	41	4	12	1	3	1	3	0	0	1	3
1-5	2	9	4	12	6	19	5	16	1	3	6	17
5-10	2	9	3	9	2	6	4	13	2	6	2	6
10-20	1	5	4	12	5	16	2	6	3	9	1	3
20-30	0	0	0	0	0	0	2	6	4	11	2	6
>30	8	36	1	3	2	6	0	0	6	17	5	14
Total	22	100	33	100	32	100	32	100	35	100	35	100

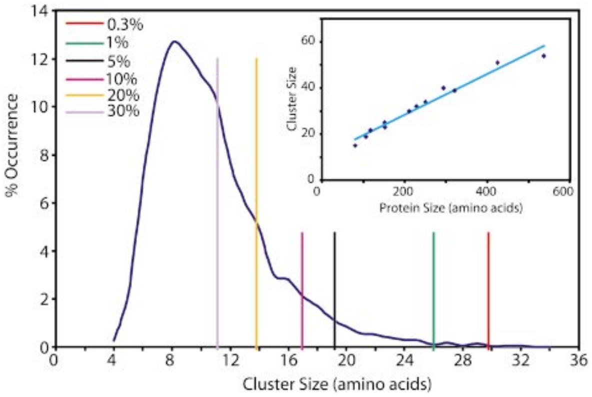

Figure 4. The size of the largest cluster is another useful statistical measure: 15 % of the residues of pyruvate decarboxylase (1pvd) were randomly selected and the size of the largest cluster was recorded. This process was repeated 5000 times and the resulting distribution is shown. Note that we have used a best fitting smoothed envelope to model discrete data points. The significance thresholds are shown as vertical lines on the distribution and are colored accordingly. Similar to the number of clusters study, the linear relationship between protein size and the size of the largest cluster predicted by random simulations is shown in the inset. Each point represents 5000 random simulations performed on a different protein (12 proteins in all) at 20 % coverage with a significance threshold of 0.3 %.

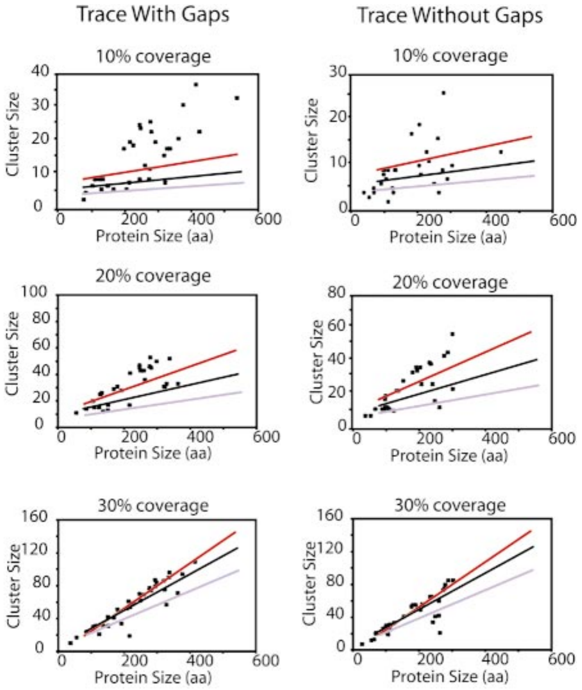

Figure 5. Significance of ET predictions using the size of largest cluster statistics. For 10, 20, and 30 % coverages, the size of the largest cluster predicted by ET is plotted against protein size for each of the 46 proteins with a rank directly convertible to a coverage level. Trace With Gaps refers to ET data generated when considering gaps in the alignment and Trace Without Gaps refers to the ET data generated without this information. The significance thresholds are shown as: 0.3 %, red line; 5 %, black line; 30 %, blue line and were generated using the linear fits from the size of the largest cluster simulations. Similar to the number of clusters study, most of the ET clusters have a significance better than 5 % (that is, there is less than 5 % chance of observing such a cluster size randomly).

proteolysis, phosphorylation, and many other biochemical activities. Their structures also vary widely with representatives from all α-helix, all β-sheet, α-helix and β-sheet-containing proteins, and one integral membrane protein (visual rhodopsin). In addition, these proteins come from a range of species, including eukaryotic (mammals, plants, fungi, and others), prokaryotic (Eubacteria and Archaebacteria), and viral representatives. Eukaryotic examples include HSP-90 and growth hormone receptor; prokaryotic examples include β-lactamase and citrate synthase; and viral examples include HIV reverse transcriptase and F-MuLV. Another aspect of this study is that trace input was not optimized beyond the elimination of obvious fragmented sequences and of evolutionary outliers from the tree. Even such minimal pruning was omitted in some cases (such as in HIV reverse transcriptase), and what pruning was performed was not done so uniformly or systematically. In addition, none of the alignments was adjusted manually or underwent multiple rounds of refinement. Thus our results represent those that might be readily obtained by an occasional user.

Of particular note is the fact that all seven proteins with fewer than 30 homologues reached a significance level of 0.3 %. Most, including the smallest family with only 19 homologues, did so using both the number of clusters and the size of largest cluster statistics. This is consistent with previous experience suggesting that proteins with more than 15-20 homologues will achieve good signal-to-noise levels. This assumes that these proteins offer a good sample of evolutionary diversity, as opposed, for example, to being uniquely mammalian.

Cluster number and size: two complementary statistics

For most proteins (30 out of 46), the number of clusters and size of largest cluster statistics yield the same level of confidence. Occasionally, however, these statistics are complementary rather than redundant. For instance, a trace may identify one large dominant cluster plus several isolated residues that each count as a cluster. If there are enough of these singleton clusters, the overall number of clusters may be too great to be significant. Yet, the size of the dominant cluster would still be significant and suggest that this trace is biologically relevant. In a number of instances, pro-

Table 4. Summary of the 46 proteins in the test set

								Best significance: no. of clusters			Best significance: cluster size		
Name	PDB code	Function	Evolutionary breadth	SCOP class	Protein size	No. of seq. In alignment	Min % identity	Significance level (%)	Percent coverage	Numerical value	Significance level (%)	Percent coverage	Numerical value
Ligand binding domain of LDL receptor	1ajj	LDL receptor	E	Small proteins	37	184	46	30	25	2	0.3	30	8
c-Src tyrosine kinase; SH3	1nlo	Tyrosine kinase	E	β	56	71	37	5	20	1	0.3	30	11
Biotinyl domain	1bdo	Carboxylase	E+P	β	80	37	45	>30			20	15	8
Acyl CoA binding protein	1aca	Binding protein	E	α	86	38	46	5	20	3	0.3	30	24
c-Src tyrosine kinase; SH2	1a09	Tyrosine kinase	E	α & β	106	137	34	1	20	2	0.3	30	30
Bikunin	1bik	Kunitz type inhibitor	E	Small proteins	110	36	43	5	20	4	10	20	15
Mannose binding protein	2msb	Binds mannose	E	α & β	113	71	34	5	10	3	0.3	30	31
Trp1 domain of Hop	1elw	Chaperone	E	α	117	42	41	1	10	3	1	10	9
Pseudoazurin	1bqk	Electron transport	E+P	β	124	29	37	1	15	4	10	20	15
Tpr2a domain of Hop	1elr	Chaperone	E+P	α	128	41	30	0.3	25	2	1	5	5
Regulator of G-protein signaling	1fqi	Regulator of G-protein signaling	E	α	133	43	43	0.3	25	1	0.3	25	33
Galectin-3 CRD	1a3k	Galectin carbohydrate recognition domain	E	β	137	70	32	1	10	4	1	10	9
Myoglobin	1a6m	Oxygen transport	E	α	151	171	35	5	30	3	0.3	30	42
Thermosome	1ass	Chaperonin	E+P	α & β	152	84	36	5	25	5	10	20	17
Poly(A)binding protein	1cvj	Gene regulation	E	α & β	169	73	26	0.3	25	3	0.3	25	38
Growth hormone	1a22-A	Growth hormone	E	α	180	67	36	0.3	20	4	0.3	30	51
Growth hormone receptor	1a22-B	Growth hormone receptor	E	α	192	21	30	1	20	6	0.3	20	28
Astacin	1ast	Metalloproteinase (hydrolase)	E	α & β	200	38	44	0.3	25	3	0.3	25	47
von Willebrand factor	1auq	Blood coagulation	E	α & β	208	44	34	5	5	5	1	5	6
HSP-90	1am1	Chaperone	E+P	α & β	213	78	55	0.3	30	3	0.3	30	61
Glutathione *S*-transferase, type-III	1aw9	Transferase	E+P	α	216	86	30	5	10	9	5	10	8
Adenylate kinase	1aky	Phosphotransferase	E+P	α & β	218	42	45	0.3	25	5	0.3	25	47
F-MuLV	1aol	Viral glycoprotein	V	β	228	21	39	0.3	25	5	0.3	25	54
Estrogen receptor	3ert	Nuclear receptor	E	α	247	93	44	0.3	25	2	0.3	25	60
Indole-3-glycerophosphate synthase	1a53	Synthase	E+P	α & β	247	19	30	0.3	20	3	0.3	30	70
Triosephosphate isomerase	1amk	Gluconeogenesis	E+P	α & β	250	73	47	0.3	25	6	0.3	25	52
Cyclins	1fin-B	Transferase	E	α	260	23	34	0.3	30	4	0.3	30	69
Beta-lactamase	1btl	Hydrolase	P	Multi-domain proteins	263	50	45	0.3	20	9	0.3	25	54
Deacetoxycephalosporin C	1rxg	Oxidoreductase	E+P	β	275	24	25	0.3	15	8	1	15	19
2,5-Diketo-D-gluconic acid reductase A	1a80	Oxidoreductase	E+P	α & β	277	83	46	0.3	30	4	0.3	30	77
Endonuclease IV	1qum	Endonuclease	E+P	α & β	279	27	39	0.3	20	5	0.3	20	53
Dihydropteroate synthase	1aj2	Synthase	P	α & β	282	42	37	0.3	20	6	0.3	25	60
Protein phosphatase-1	1fjm	Hydrolase	E	α & β	294	68	65	0.3	30	3	0.3	30	87
Signal sequence recognition protein	1ng1		E+P	α	294	73	45	0.3	15	6	0.3	30	79
Cyclins	1fin-A	Transferase	E	α	298	37	64	0.3	30	4	0.3	30	84
Thioredoxin reductase	1f6m	Reductase	E+P	α & β	320	44	56	0.3	25	8	0.3	30	86

Annexin III	1axn	Calcium/phospholipid binding protein	E	α	323	70	40	0.3	20	11	5	20	31
Transferrin	1a8e	Iron transport	E	α & β	329	52	46	0.3	20	6	0.3	15	26
Peroxidase	1aru	Peroxidase	E	α	336	29	55	0.3	15	13	0.3	30	90
Rhodopsin	1f88	Signaling protein	E	Membrane and cell surface protein	338	59	33	0.3	30	4	0.3	30	96
Serine/threonine phosphatase	1a6q	Hydrolase	E	α & β	363	58	38	0.3	20	8	0.3	15	31
Citrate synthase	1a59	Synthase	E+P	α	377	63	32	0.3	20	14	0.3	25	74
Phosphoglycerate kinase	16pk	Kinase	E+P	α & β	415	95	41	0.3	25	11	0.3	25	93
Alpha amylase	1bag	Alpha-amylase	E+P	β	425	55	24	0.3	30	8	0.3	30	116
HIV reverse transcriptase	1c1b	Reverse transcriptase	V	α & β	536	278	61	1	20	23	5	20	43
Pyruvate decarboxylase	1pvd	Carbon-carbon lyase	E+P	α & β	537	43	37	0.3	25	11	0.3	25	114

The 46 proteins in the test set are listed. In addition to the full name of the protein, listed for each are the PDB identifier, class, size (amino acid residues), known function, number of sequences in the multiple sequence alignment, minimum percentage identity between sequences in the alignment and the selected protein structure, the best significance level the protein achieves using the number of clusters statistical method, the best significance level the protein achieves using the size of largest cluster method when gaps are considered as informative, and the evolutionary breadth of the protein family tree (E, eukaryotic; P, prokaryotic; V, viral). As can be seen, the vast majority of the proteins achieve significance at or better than the 5 % level at least once.

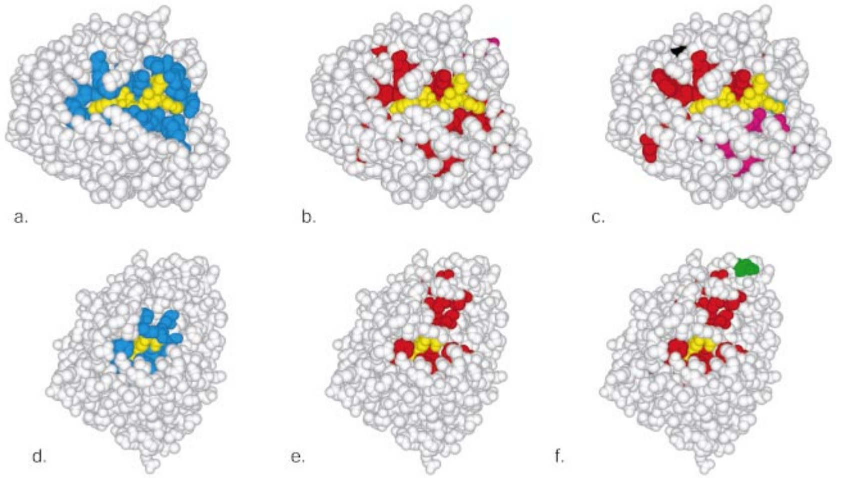

Figure 6. ET clusters overlap with known ligand binding domains. In the representative cases of 2,5-diketo-D-gluconic acid reductase A (1a80, (a)-(c)) and dihydropteroate synthase (1aj2, (d)-(f)), the structural epitopes, defined as all the residues within 5 Å of the ligand (shown in yellow), are shown in blue ((a) and (d)). ET-identified residues surround and include residues from the structural epitopes for both proteins when gaps are excluded ((b) at rank 66, (e) at rank 13) and included ((c) at rank 55, (f) at rank 23) from the ET analysis. Individual ET-identified residue clusters are shown in red, purple, and black ((b) and (c)) and in red and green ((e) and (f)). In the case of 1a80, the clustering pattern is noticeably different when gaps are included or excluded from the analysis. This is due to the fact that when gaps are excluded, the rank at which the structural epitope is identified is greater than when gaps are included (compare (c) to (b)). As rank increases, separate small clusters tend to coalesce into larger clusters (compare (c) to (b)). These ET-identified residues do not overlap completely with the structural epitope, consistent with the fact that not all of the residues in the binding site contribute to ligand interaction.

teins did achieve greater significance using the largest cluster size method and include the ligand binding domain of the LDL receptor (1ajj), SH3 domain (1nlo), Acyl CoA binding protein (1aca), SH2 domain (1a09), mannose binding protein (2msb), myoglobin (1a6 m), growth hormone (1a22-B), and von Willebrand factor (1auq). Conversely, a trace of a protein with multiple ligands may identify a few clusters of nearly equal size. None may be large enough to be significant, but their low overall number would be significant. Instances of proteins that achieved greater significance with the number of clusters method include bikunin (1bik), pseudoazurin (1bqk), Tpr2a-domain of Hop, an adaptor protein mediating association of Hsp70 and Hsp90 (1elr), thermosome (1ass), deacetoxycephalosporin C (1rxg), annexin III (1axn), and HIV reverse transcriptase (1c1b). In practice, a trace that fails to reach significance with either statistics is unlikely to lead to biologically relevant information.

Statistical limitations

A potential liability of our approach is that we assume a linear relationship between the significance thresholds and protein size. Deviations from this norm are likely to occur, however, in proteins that have a marked increase in the surface-to-volume ratio of a protein. In such a case, the average residue is more often on the surface and thus makes fewer contacts to other residues. Accordingly, a given fraction of randomly picked residues will form more, smaller clusters than if they were picked from a typical globular protein, such as were included in the set of 12 proteins for which random simulations were performed. In other words, our significance thresholds are too stringent for proteins with irregular (non-globular) shapes. This bias is acceptable, however, since it will underestimate trace significance.

Another potential shortcoming is that statistical significance, for example at the 5 % level, is established after repeated sampling, say at 5, 15, 20, 25 % coverage. Thus to report that a 5 % level of significance has been reached at a given coverage is not formally correct, especially if that significance level was not achieved for any of the other coverages. While the repeated sampling pitfall is inevitable if one examines multiple coverages, significance is often reached over a range of coverage levels, and each one of these incorporates the information in the previous levels, so they are not truly independent trials. In practice, if trace residues arose randomly and thus had no privileged structural relationship with one another, the null hypothesis H_0 could not be confidently rejected. If in fact we find that H_0 can be rejected over a certain range of coverage, then this suggests that these traces uncover biologically meaningful information. The fact that ET clusters overlap with ligand binding surfaces in 37 of 38 proteins supports this interpretation.

Accounting for insertions and deletions

The tolerance of gaps in the input multiple sequence alignment significantly improves trace performance. With the number of clusters statistics, 44 proteins (96 %) reached the 5 % significance level or better; and 30 (65 %) reached the 0.3 % significance level. But when gaps are excluded these numbers are 72 % and 57 %, respectively. With the size of largest cluster statistics, 42 (91 %) reached the 5 % significance level and 34 (74 %) reached the 0.3 % significance level with gaps *versus* 78 % and 63 %, respectively, without.

These data clarify nearly opposite interpretations of gapped positions in a multiple sequence alignment. The view followed in the original trace method was that a gapped position cannot be functionally important in all the proteins in the multiple sequence alignment, since in at least one instance, in one species it is eliminated altogether. Another view, however, is that insertions may clearly introduce novel and important functional elements in a protein, perhaps simply by blocking a ligand from binding to its normal site. By the same token, a deletion might reveal a new functional surface or directly remove a functional component. Either way, gapped positions may play important roles and should be considered. Functional information in gapped regions can still be obtained with the original trace, but only if the, sequences with the gaps are removed from the alignment. This however, deprives the ET of much evolutionary information and diminishes its significance. On the other hand, the simple convention that gaps are a new amino acid type allows us to trace gapped regions and to avoid any loss of sequence information.

In addition, there are two structural advantages to including residue positions in regions containing gaps in the multiple sequence alignment. First, an active site of any given size will appear less significant in a smaller protein using our statistics, since it will occupy a greater fraction of the total size, making it more difficult to distinguish a significant result from random clustering by an equivalent fraction of residues randomly scattered onto the same protein. Thus, by including more residues in the protein, an active site of the same size can be more easily distinguished from random chance. Second, removing residues from consideration tends to generate a pseudo-protein structure with an increased surface-to-volume ratio, which as seen above, increases the stringency of our statistical methods and underestimates the significance of the trace cluster.

Guidelines for selecting input sequences and performing traces

The techniques introduced suggest a few simple guidelines for performing a trace. The initial selection of sequences should strive to eliminate sequence fragments and evolutionary outliers. The former introduces isolated gaps in the alignment that are not aligned with gaps in other sequences and thus that do not form blocks of gaps. The latter are also easily recognizable as offshoot branches rooted deeply in the phylogenetic tree. Otherwise, gap tolerance allows us to keep sequences which bear deletions/insertions, thereby reducing considerably the pruning of input sequences. Overall sequence identity in an alignment is another factor that affects the quality of the trace output. An input data set consisting of closely related proteins (say >70 % identity) causes many trace residues to be top-ranked. A large percentage coverage will therefore be reached very quickly (say >30 % coverage), at which point most clusters lose statistical significance. On the other hand, if identity among input sequences is too low, some of the proteins may have diverged sufficiently so that they no longer perform similar functions, and they may even use distinct functional surfaces. In that case the patterns of conservation and variation will be different in those proteins leading to few top-ranked residues, poor clustering and low significance. In fact, this feature is useful to identify which homologues perform identical functions and which do not.[16,26] These guidelines are necessarily approximate and may not apply to specific protein families. In our data, useful traces were generated with as little as 25 % identity in an enzyme, deacetoxycephalosporin C, and as much as 65 % identity (protein phosphatase 1). In G protein-coupled receptors, for example, it is clearly possible to trace jointly sequences with as low as 12 % sequence identity and extract biologically meaningful functional site information.[27]

HIV reverse transcriptase is an example of trace input optimization. This is one of the 38 proteins tested in this study having a bound ligand, and its initial trace included all 278 sequences retrieved by a simple search, including many fragments. The trace on this input set still yielded a significant dominant cluster (5 %) but did not identify the ligand-binding site. However, a simple pruning of all protein fragments from the multiple sequence alignment, bringing the number of remaining homologues to 70, enlarges this cluster so that it matches the ligand-binding site and increases its significance to 0.3 %.

An evolutionary link between sequence-structure-function

Functional sites are the key mediators of protein activity, and the ability to predict their location on the structure and the relative importance of their constituent residues has important applications. Prior studies established proof of concept that the ET can predict the location of functional surfaces. First, known binding sites were accurately retro-predicted in control proteins.[2] Then, ET correctly anticipated later mutational studies and quaternary structures revealed by X-ray crystallography.[1,3,4,18,19] Now, this study further extends our understanding of the method's range of application by showing that: (1) clusters of trace residues can be defined objectively; (2) they are repeatedly identified in diverse types of proteins; and (3) they match the known ligand binding sites whether they are on the surface, or buried in the protein core.

The statistics introduced here formalize the clustering characteristic of class-specific residues. It is interesting to stress that the rank of a residue depends only on its variational history and how this correlates with evolutionary divergence. Thus the trace operates on sequences, with no *a priori* knowledge of the spatial distribution. Why then do highly ranked residues cluster, and why do they match functional surfaces? Both observations suggest that this is because, as anticipated, top-ranked residues are evolutionarily important. Trace residue variation, by definition, always correlates with evolutionary divergence. These residues must, therefore, contribute significantly to one, or many, functions (such as folding, cellular location targeting, protein dynamics, ligand interactions, biochemistry, degradation, and others), the most likely of which is molecular recognition when they are found on the surface. Lastly, it is difficult to imagine how a residue important to any of these functions could be entirely surrounded by residues that lacked importance themselves. It would seem that the latter would insulate their important neighbors and likely diminish the influence of its variation. This argument hinges on the cooperative nature of residues within the polypeptide chain, and suggests that the key determinants of structure and function should be organized into networks of contacts. Our results show that such networks can indeed be identified.

While the practical purpose of the trace is to identify functional surfaces, we note that most top-ranked residues clusters internally where they are likely to play key roles in folding, protein dynamics, and allosteric phenomena. As more traces become available it will be interesting to study this evolutionary core in its own right to understand its components, variations, and substructure as they relate to the rest of the protein. It will also be useful to probe the relationship between trace residues and those defined by the related concept of ''conservatism-of-conservatism'',[28] or through conservation in common protein substructures.[29]

In conclusion, this study shows that the ET should be applicable to most proteins in the protein structure database and those that are yet to emerge from the Structural Genomics Initiative.[30] As ever more genomes are sequenced, the minimum number of homologues needed for ET to achieve significance will be increasingly met. Even when relatively few sequences are available, the statistics presented here still allow users to analyze a structure and assign an objective significance to the ET-identified clusters. The experimentalist can then judge whether the significance justifies targeted experimentation. Often, in an experimental context, a significance level of only 30 % may in fact warrant targeted mutagenesis as a better alternative to random mutagenesis. The ability to objectively assess cluster significance, and the diminished requirement for removing gapped sequences beyond those that are obviously fragments should also allow for streamlining and automating ET. It will thus provide a general and natural mechanism to extract from the raw data in sequence and structure databases the answers to at least two critical biological questions: where are the functional sites, and what are their key residues?

Methods

Protein test set

A total of 12 proteins (1bdo (80 residues), 1a09 (106 residues), 1elw (117 residues), 1a6m (151 residues), 1ass (152 residues), 1am1 (213 residues), 1aol (228 residues), 1amk (250 residues), pp1 (294 residues), 1axn (323 residues), 1bag (425 residues), 1pvd (537 residues); full names are given in Table 4) were included in the test set for random cluster simulations based on the criteria that their sizes should adequately sample the region between a small protein (~80 residues) and a large protein (~500 residues) and that their shapes are mostly globular. For each protein, the fraction of residues chosen randomly was determined as a percentage of the total number of residues present in that protein, beginning with 5 % and increasing in increments of 5 % to 95 % (although only coverages up to 30 % are shown herein). At each coverage level, individual residues were selected randomly and both the total number of clusters and the size of each cluster were recorded. This process was repeated 5000 times (a compromise between statistical significance and computational time) for each protein at each coverage level to generate the complete data set for further analysis.

Clustering analysis

The randomly selected residues were defined as a cluster if any atom in one residue was within 4 Å of any other atom in another residue (hydrogen atoms excluded). The typical distribution of the number of clusters followed a long-tailed distribution as shown in Figure 2. We calculated the number of clusters at threshold values of 0.3, 1, 5, 10, 20 and 30 % significance: 0.3 % significance, for example, implies that the probability of randomly observing the corresponding number of clusters is 3 in 1000.

For each of the 5000 iterations, at each coverage level, the size of the largest cluster was recorded and their distribution was plotted for each coverage level (Figure 4). The resulting distributions closely resembled those observed for the number of clusters analysis and therefore a similar approach of determining threshold values was used. However, the tail of the distribution corresponding to a larger value (of cluster size) was used rather than a smaller one as before, although for the same reason (signal is defined as a small number of large clusters).

Linear fitting and look-up tables

Both the threshold values for the number of clusters and the size of the largest cluster analyses followed a linear relationship with respect to protein size. In order to extrapolate our data to proteins of all sizes (within the range we tested), we plotted each threshold value against protein size to allow comparison of an observed threshold value with the value expected randomly (and the significance of such a comparison) for a given coverage (Figures 3 and 5). Since a linear relationship was found irrespective of protein size, significance, or coverage, we obtained linear fits for each coverage and significance level and used the resulting data to generate look-up tables whereby the fits at a given coverage and significance level could be used to extrapolate a threshold value for a protein of any size within the range test. The statistics used to construct Figures 3 and 5 are discrete and not continuous as may be implied by the smooth envelope of the histogram. All the thresholds obtained by linear extrapolation have been rounded off to the nearest integer.

ET analysis

ET analysis was performed as previously described.[1,2,16] A total of 46 proteins were selected from the PDB so that they represented a range of protein sizes (the smallest one 37 residues and the largest one 537 residues in length), a wide range of protein folds, and a diversity of biological function (Table 4). The multiple sequence alignments were generated using pileup (of GCG package) or CLUSTALW using their default variables and the trees obtained are rooted and unbalanced. Two different types of ET analyses were performed on each of these 46 proteins. The first method of analysis discounted any residue position containing a gap; while the second method of ET analysis involved includes such residue positions by treating gaps in the sequence alignment as a 21st amino acid. Considering a gap as an artificial amino acid type allows every residue in the structure to be assigned a rank (as opposed to the first method where gap positions are completely ignored), leading to 100 % coverage of the protein structure at the maximum rank, and thus enables direct comparison between ET predictions and the random clustering. The first method, however, requires an adjustment to be made in the coverage values so as to account for the fact that at maximum rank, 100 % coverage of the protein may not have been achieved. In both the cases we used an input data set that was not optimized (i.e. fragmented sequences, incomplete sequences, distant homologues, mutants, and sequences that vary in length from the sequence of the structure were not pruned from the input to the multiple sequence alignment). These sequences were not removed in order to mimic an input that could be fed into the program by a scientist not trained to use the ET. An ET server is under construction† and more Figures of traces performed on various proteins discussed here or otherwise can be found‡.

ET mapping

Ranks were converted into their corresponding coverage levels by dividing the number of class-specific residues at a given rank by the total number of residues able to be assigned a rank. When residue positions containing gaps are excluded from trace analyses the total number of residues becomes the number of class-specific residues found at the maximum rank. When gaps are treated as artificial amino acids, the total number of residues is simply the total number of residues in the protein structure. The significance of any rank can be determined by examining where the observed number of clusters (at that rank's coverage) falls with respect to the significance thresholds established from the linear fitting of the random cluster data. In the case where gaps are excluded from the analyses, the significance of the ranks is determined in the same manner as when gaps are included, but the protein size is considered to be the maximum number of trace residues and not the total number of residues in the protein. Visualization of trace predictions was done using Rasmol molecule viewer.

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

‡ http://imgen.bcm.tmc.edu/molgen/labs/lichtarge/trace_of_the_week/traces.html

Acknowledgments

We thank Chad Shaw for his help. We gratefully acknowledge support from the American Heart Association, the March of Dimes, NHGRI HG0251-03 and NSF DBI-0114796 (O.L.), the NLM and the W.M. Keck Center for Computational Biology (M.E.S.), as well as from the Rockwell Funds Inc. (S.M.).

References

1. 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.
2. 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.
3. 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.
4. 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.
5. Ma, B. W. H & Nussinov, R. (2001). Protein functional epitopes: hot spots, dynamics and combinatorial libraries. *Curr. Opin. Struct. Biol.* **11**, 364-369.
6. Pearce, K. H., Jr, Ultsch, M. H., Kelley, R. F., de Vos, A. M. & Wells, J. A. (1996). Structural and mutational analysis of affinity-inert contact residues at the growth hormone-receptor interface. *Biochemistry*, **35**, 10300-10307.
7. 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.
8. Honig, B. & Nicholls, A. (1995). Classical electrostatics in biology and chemistry. *Science*, **268**, 1144-1149.
9. Miranker, A. & Karplus, M. (1991). Functionality maps of binding sites: a multiple copy simultaneous search method. *Proteins: Struct. Funct. Genet.* **11**, 29-34.
10. Lamb, M. L. & Jorgensen, W. L. (1997). Computational approaches to molecular recognition. *Curr. Opin. Chem. Biol.* **1**, 449-457.
11. Reyes, C. M. & Kollman, P. A. (2000). Investigating the binding specificity of U1A-RNA by computational mutagenesis. *J. Mol. Biol.* **295**, 1-6.
12. Casari, G., Sander, C. & Valencia, A. (1995). A method to predict functional residues in proteins. *Nature Struct. Biol.* **2**, 171-178.

13. Henikoff, S. & Henikoff, J. G. (1991). Automated assembly of protein blocks for database searching. *Nucl. Acids Res.* **19**, 6565-6572.
14. Jones, S. & Thornton, J. M. (1997). Prediction of protein-protein interaction sites using patch analysis. *J. Mol. Biol.* **272**, 133-143.
15. Shatsky, M. F. Z. Y., Nussinov, R. & Wolfson, H. J. (2000). Alignment of flexible protein structures. *Intell. Syst. Mol. Biol.* **8**, 329-343.
16. 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.
17. Baranski, T. J., Herzmark, P., Lichtarge, O., Greber, B. O., Trueheart, J., Meng, E. C., Iiri, T. *et al.* (1999). C5a receptor activation. Genetic identification of critical residues in four transmembrane helices. *J. Biol. Chem.* **274**, 15757-15765.
18. Onrust, R., Herzmark, P., Chi, P., Garcia P., D., Lichtarge, O., Kingsley, C. & Bourne, H. R. (1997). Receptor and betagamma binding sites in the alpha subunit of the retinal G protein transducin. *Science,* **275**, 381-384.
19. Slep, K. C., Kercher, M. A., He, W., Cowan, C. W., Wensel, T. G. & Sigler, P. B. (2001). Structural determinants for regulation of phosphodiesterase by a G protein at 2.0 Å. *Nature,* **409**, 1071-1077.
20. Pritchard, A. L. D. 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.
21. Landgraf, R., Fischer, D. & Eisenberg, D. (1999). Analysis of heregulin symmetry by weighted evolutionary tracing. *Protein Eng.* **12**, 943-951.
22. 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.* **12**, 839-847.
23. 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.
24. 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.
25. 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.
26. 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.
27. Lichtarge, O., Sowa, M. E. & Philippi, A. (2001). Evolutionary traces of functional surfaces along the G protein signaling pathway. *Methods Enzymol.* **344**, 537-556.
28. Mirny, L. A. & Shakhnovich, E. I. (1999). Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. *J. Mol. Biol.* **291**, 177-196.
29. Reddy, B. V., Li, W. W., Shindyalov, I. N. & Bourne, P. E. (2001). Conserved key amino acid positions (CKAAPs) derived from the analysis of common substructures in proteins. *Proteins: Struct. Funct. Genet.* **42**, 148-163.
30. Brenner, S. (2001). A tour of structural genomics. *Nature Rev. Genet.* **2**, 801-809.

Edited by F. E. Cohen

(Received 14 August 2001; received in revised form 20 November 2001; accepted 26 November 2001)