Sequence and structure continuity of evolutionary importance improves protein functional site discovery and annotation

A. D. Wilkins,[1,2] R. Lua,[1] S. Erdin,[1,2] R. M. Ward,[1,2,3] and O. Lichtarge[1,2,3]*

[1]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas, 77030

[2]W. M. Keck Center for Interdisciplinary Bioscience Training, Houston, Texas, 77005

[3]Graduate Program in Structural and Computational Biology and Molecular Biophysics, Baylor College of Medicine, Houston, Texas, 77030

Received 10 February 2010; Revised 18 April 2010; Accepted 19 April 2010
DOI: 10.1002/pro.406
Published online 27 April 2010 proteinscience.org

Abstract: Protein functional sites control most biological processes and are important targets for drug design and protein engineering. To characterize them, the evolutionary trace (ET) ranks the relative importance of residues according to their evolutionary variations. Generally, top-ranked residues cluster spatially to define evolutionary hotspots that predict functional sites in structures. Here, various functions that measure the physical continuity of ET ranks among neighboring residues in the structure, or in the sequence, are shown to inform sequence selection and to improve functional site resolution. This is shown first, in 110 proteins, for which the overlap between top-ranked residues and actual functional sites rose by 8% in significance. Then, on a structural proteomic scale, optimized ET led to better 3D structure-function motifs (3D templates) and, in turn, to enzyme function prediction by the Evolutionary Trace Annotation (ETA) method with better sensitivity of (40% to 53%) and positive predictive value (93% to 94%). This suggests that the similarity of evolutionary importance among neighboring residues in the sequence and in the structure is a universal feature of protein evolution. In practice, this yields a tool for optimizing sequence selections for comparative analysis and, via ET, for better predictions of functional site and function. This should prove useful for the efficient mutational redesign of protein function and for pharmaceutical targeting.

Keywords: binding site prediction; sequence selection; functional annotation; evolutionary trace

Introduction

The knowledge of which amino acids mediate protein function is necessary to unravel molecular mechanisms,[1,2] to redesign function rationally,[3,4] and to target drugs.[5] The gold standard to identify these residues remains systematic mutational analysis,[6–8] but this approach has some high throughput limitations. Inadequate choice and availability of assays reduce sensitivity while the promiscuity of binding[9] or catalysis,[10] as well as poor reproducibility of the relevant cellular context,[11] reduce specificity.

This prompts complementary computational methods to discover functional sites, their residues and their biological roles. Approaches based on preexisting structure may be grouped broadly into those using energetics,[12–15] and others using structural and geometric analysis.[16–19] Here, we focus on comparative, or evolutionary approaches,[20–26] and specifically on the evolutionary trace (ET).[27,28]

ET maps functional hotspots on protein structures: areas of the protein where amino acids that

Additional Supporting Information may be found in the online version of this article.

Grant sponsor: NIH; Grant number: GM079656; Grant sponsor: NSF DBI; Grant number: 0547695; Grant sponsor: NIH; Grant number: GM066099; Grant sponsor: CCF; Grant number: 0905536; Grant sponsor: NLM; Grant number: 5T15LM07093.

*Correspondence to: O. Lichtarge, Department of Molecular and Human Genetics Baylor College of Medicine One Baylor Plaza, Room T921 Mail stop BCM225 Houston, TX 77030. E-mail: lichtarge@bcm.edu

Table 1. *Summary of Quality Measures*

Q_m	Formula	Description
z-score based measures		
$Q_{structure,1}$	$f(i,j) = (j - i)$	Structural clustering
$Q_{structure,2}$	$f(i,j) = \sqrt{j - i}$	Structural clustering
$Q_{structure,3}$	$f(i,j) = 1$	Structural clustering
$Q_{surface}$	$f(i,j) = s_i s_j, s_i = \begin{cases} 1 & \text{if surface} \\ 0 & \text{otherwise} \end{cases}$	Surface clustering
$Q_{sequence}$	$f(i,j) = \delta_{j-i,1}$	Neighbors in sequence
non *z*-score based measures		
$Q_{contrast}$	$\frac{\sum_{j>i}^{L} \sum_{a_i,a_j} A(a_i,a_j)\lvert(c_j - c_i)\rvert}{n}$	Contact rank difference
Q_{RI}	TI × RE	Information content of ranks

There are five *z*-score based Q_m's that measure the statistical significance of the clustering among ET top-ranked residues within structure and sequence. The measures are a function of *z*-scores, $z_c = \frac{w - \langle w \rangle}{\sigma}$ where $w = \sum_{j>i}^{L} S(i)S(j)A(i,j)f(i,j)$. The quantity w measures the top-ranked residues in contact spatially. The difference lies in the weighting term $f(i,j)$, which weighs the contribution of residues i and j differently based on their relative position in the structure and sequence. The last two measures ($Q_{contrast}$ and Q_{RI}) are unique in formula. $Q_{contrast}$ is a measure of the rank gradient over the structure. Q_{RI} is structureless measure of the information content of the rank distribution. Further detail of the Q_m's can be found in "Materials and Methods" section.

impact function concentrate. In large-scale analyses, ET ranked amino acids by evolutionary importance[8,29] such that the top-ranked ones formed structural clusters[30–32] that overlapped and predicted functional sites.[33,34] Case studies further showed that bona fide ET-guided mutants could then block, separate and even swap functions *in vitro* and *in vivo*.[35–38] ET thus predicts key functional determinants and enables their rational perturbation.

With the goal to optimize accuracy for high throughput automated ET, this study now aims to increase the functional consistency among ET's input sequences. On the one hand, this is not trivial. Simply relying on BLAST[39,40] to pool homologous sequence often leads to functionally heterogeneous sequence selections,[41–43] and, in turn, since ET identifies the functional sites that are common to all the proteins which it analyzes, such functional heterogeneity reduces accuracy. On the other hand, it is possible to optimize the selection of sequences. When some of these homologous sequences are pruned away so as to improve either the structural clustering of ET ranks,[44,45] or their information content[46] in the sequence, then the overlap between top-ranked residues and the functional site increases. Building on these results, the hypothesis of this work is that basic features of the ET rank distribution can be found that inform the selection of sequences and improve ET accuracy.

This article presents evidence that continuity of ET ranks across adjacent residues is one such fundamental characteristic of evolutionary forces. One type of test for successful ET improvement will be whether top-ranked ET residues overlap better with known functional sites. As this involves just a small fraction of the known structural proteome, however, a second, high throughput test will be whether these improvements carry over to protein function predictions via ET Annotation (ETA).[47,48] In this approach, without any prior knowledge of functional or catalytic sites, ET guides the selection, in a query protein of unknown function, of a structural motif of (six) top-ranked, neighboring, surface residues that together define a 3D template. ETA then matches the 3D templates to previously annotated proteins across the PDB. Matches identify analog substructures with similar geometric and evolutionary features that may therefore mediate identical functions, and this is the basis for function predictions with ETA, or with related approaches.[49] Predictably, an improved assignment of ET ranks yields better templates for functional annotation. For example, replacing 3D template residues by presumed catalytic site residues lowered prediction accuracy from 96% to 81%.[50] Thus, ETA annotations depend sensitively on an all-against-all comparisons of ET ranks, and their improvement will confirm broad gains in accuracy across the structural proteome.

Results

Quantitative features of the rank distribution

To probe features of the ET rank distributions that are universal and that correlate with accuracy, we define in the "Materials and Methods" section seven quality measures, Q_m, where the subscript m reflects the choice in measure, see Table 1. We show below that all of them fulfill three conditions that are necessary and sufficient to guide the selection of input sequences for ET: (1) they are computable without reference to prior known functional sites; (2) they correlate with the overlap between high ranked residues and the known functional site, $A_{overlap}$; (3) and sequence selections that increase their value also improve $A_{overlap}$. Accordingly, each Q_m can guide

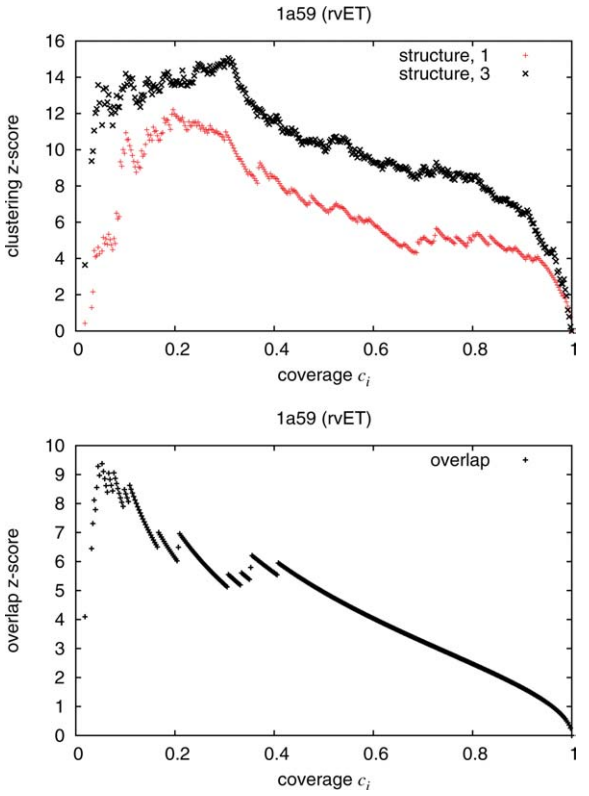

Figure 1. (a) The clustering *z*-score measures the nonrandomness of the clustering of top-ranked residues in space. The *z*-scores are a direct result of the ranking of the residues in a protein structure. This diagram shows an example of the clustering *z*-scores as a function of c_i using the rvET method for a cold-active citrate synthase [*Antarctic bacterium*, PDB 1a59]. The high clustering *z*-scores would indicate similarly ranked residues proximate in the structure and would be considered a positive result. Quality measures $Q_{\text{structure},1}$ and $Q_{\text{structure},3}$ are variants of the clustering *z*-scores. (b) To represent a method's ability to predict a known site, the overlap *z*-score is also calculated using a simple hypergeometric distribution. An example of the overlap *z*-scores as a function of c_i can be seen in bottom figure. The overlap measure A_{overlap} is derived from the these *z*-scores.

the selection of sequences in order to improve functional site detection without prior knowledge of that site. Notably, some of these different Q_{m} depend on the structure and others on the sequence, but most focus on the similarity of physically neighboring ET ranks, that is, their continuity in the sequence or in the structure (Fig. 1). This common theme suggests that many other derivative quality measures also related to continuity could be devised easily.

Correlations with deleterious sequence perturbations

To test condition 2, Q_{m} perturbations were introduced in ET's input and a correlation was assessed between spatial clusters of top-ranked residues and known functional sites, that is, between the quality measure Q_{m} and the overlap A_{overlap}. Different ranking methods were used each time to control for method-specific bias: the integer value ET[27] (ivET), Shannon Entropy[51] and the current standard real value ET[28] (rvET), which is resilient to errors like entropy, but exploits the phylogenetic information like ivET. Each method is reviewed in "Materials and Methods" section.

The first type of perturbation was to add more sequences to the input to ET, starting from just the query and two homologs. The sequences were taken at random from an initial BLAST[52] search. A representative example shows that Q_{contrast} and A_{overlap} were well correlated (>0.9) (see Fig. 2) and this generally irrespective of the ranking method (Fig. 3). There was one sole outlier $Q_{\text{structure},3}$, that had a poor correlation for ivET. Also note that one method, ivET, had more proteins with little or no correlation. This is consistent with the high sensitivity of ivET to errors, gaps, misalignments or polymorphisms that break the perfect match between sequence variations and phylogenetic divergences that is the hallmark of ivET rankings. Once such a sequence was added to the input, it decreased the overlap to a

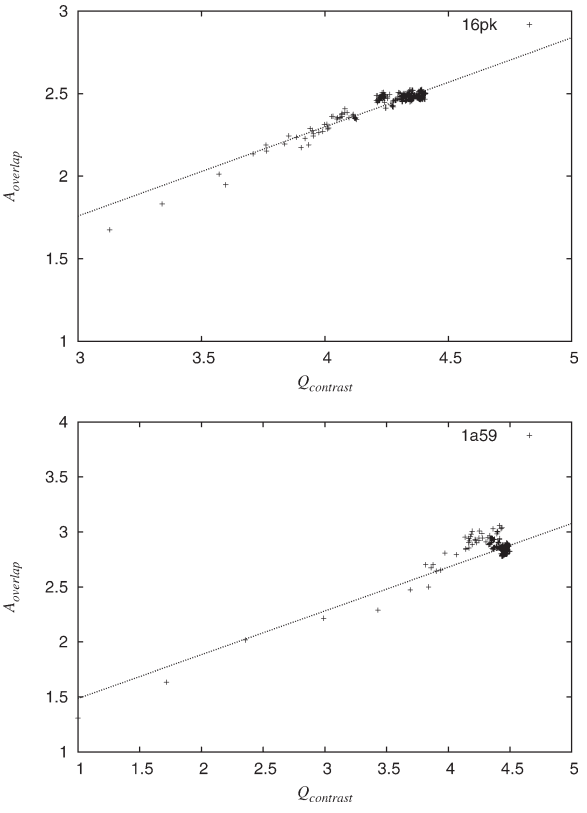

Figure 2. A correlation between quality measures and overlap of known site was found when variations were considered in alignment. The quality measures are a result of the ranking of the sequences in an alignment. These diagrams show examples of the values of quality measure Q_{contrast} and overlap measure A_{overlap} as sequences are added into the analysis randomly. The values for the first 30 sequences added to the analysis were used to calculate correlation.

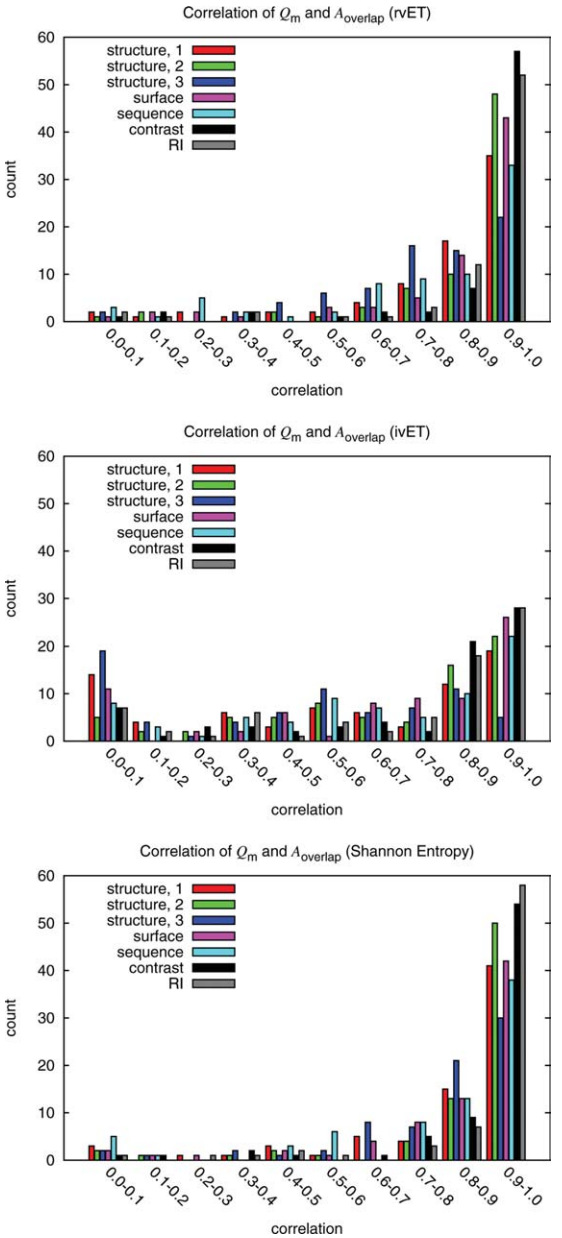

Figure 3. Distribution of Pearson correlations between quality measure variations and overlap measure variations in 74 proteins when sequences are added randomly added to an alignment. The purpose of the study was to test the methods and quality measures as a function of sequence selection. The histograms show the correlations of the possible quality measures and functional site measure $A_{overlap}$ for the rvET, ivET, and Shannon Entropy method when 30 sequences are randomly added to the ranking analysis. The $Q_{contrast}$ (labeled EC), Q_{RI} and $Q_{structure,2}$ had the highest correlations amongst the quality measures for the ranking methods though all measures where found to have some correlation. Note that one method, ivET, had more proteins with little or no correlation. This is consistent with the high sensitivity of ivET to errors, gaps, misalignments or polymorphisms that break a perfect match between sequence variations and phylogenetic divergences. Once such a sequence was added to the input, it decreased the overlap to a known site irretrievably, yielding traces with lower quality and lower correlation.

known site irretrievably, yielding traces with lower quality and lower correlations overall.

A second type of perturbation (Fig. 4) was introduced to further test these correlations. To corrupt the alignments, an increasing number of mutations were introduced to simulate errors (in steps of 0.25% and up to a 5% error). Each time, a sequence and residue location was randomly picked and replaced with another residue or a gap, also picked randomly (each had an equal chance of occurring). The procedure was repeated 10 times to find how the average quality measure Q_m correlated with the average functional site overlap, $A_{overlap}$, as a function of errors. Again, Q_m's and $A_{overlap}$ were both strongly correlated for most proteins and ranking methods (Fig. 5). This time, $Q_{structure,3}$ was comparable to $Q_{structure,2}$, and even outperformed it with integer value ET. These observations suggest that Q_m's are adequate surrogate markers of the impact of input sequence perturbations on the accuracy of ET hotspots.

Test set: Individual quality measures

Next, we tested whether sequence selections that maximized Q_m also improved ET predictions. An

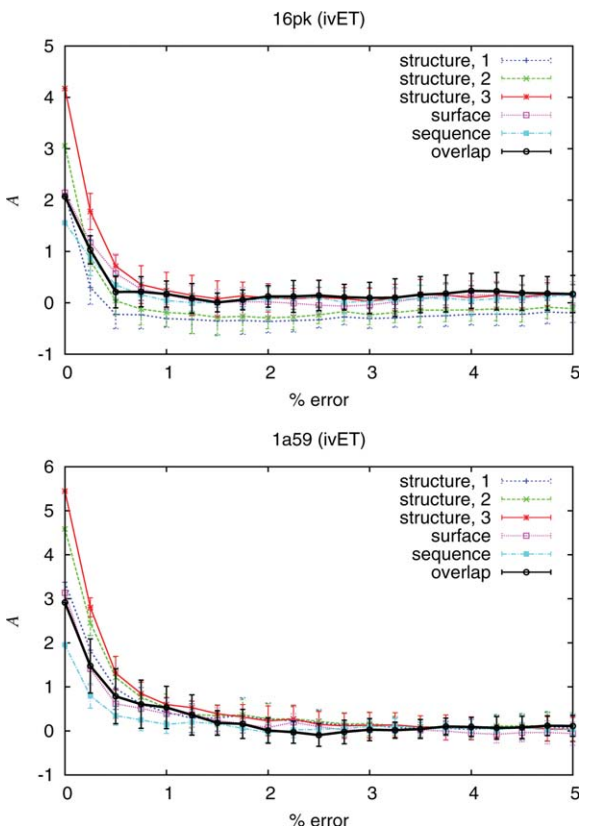

Figure 4. Analysis was performed to study the performance of the quality measures and the ranking methods as errors were introduced. The deterioration of the quality measures and overlap measure $A_{overlap}$ as a function of random mutations in the analysis is observed in protein 16pk and 1a59. Correlation was determined from the values of the quality measures and overlap measure $A_{overlap}$.

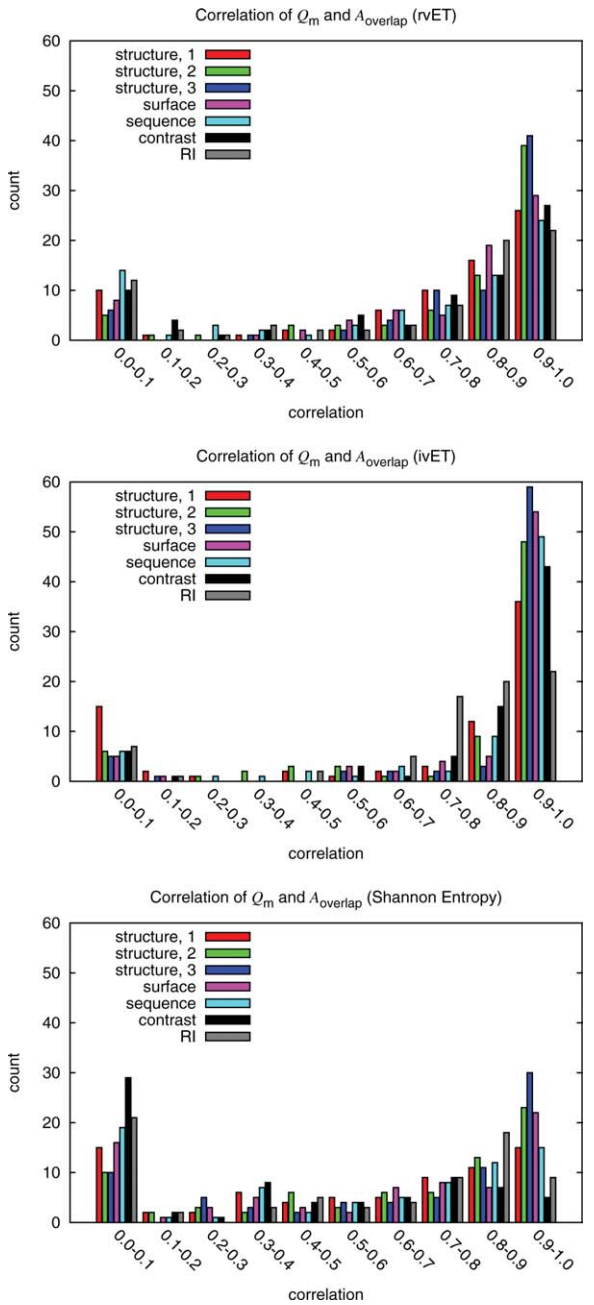

Figure 5. To test ranking methods and quality measures, random mutations were inserted into the alignment. These histograms show the correlations of the possible quality measures and functional site measure $A_{overlap}$ for the rvET, ivET, and Shannon Entropy method. The $Q_{structure,2}$ and $Q_{structure,3}$ measures consistently have the best correlations in all three methods for the majority of the proteins. All measures were shown to have some correlation. The Shannon Entropy and the rvET methods had a significant number of proteins with low correlation when compared to the ivET method. This is because ivET is very sensitive to errors while the other methods are more resilient. Thus, as errors were added, ivET rapidly lost accuracy and showed better correlations than the two other, more robust methods for which the overlap with the known site would not change dramatically up until the alignment had 20% error. Though this decreased correlation may impair optimization, it is desirable for good initial functional site prediction.

optimization algorithm described in Materials and Methods added, or removed, sequences or whole evolutionary tree branches from ET's input depending on whether Q_m values rose, or fell. It was then applied to 74 diverse proteins (that variously bind substrates, cofactors, DNA or proteins). In each case, an initial BLAST[52] search gathered sequences; those with obvious gaps and fragments were removed (a step referred to as coarse heuristic pruning); and the remaining sequence selections were refined with the optimization algorithm.

Consistent with the hypothesis, most sequence selections could be manipulated to increase Q_m and, in turn, lead to better overlap with the known functional sites (see Table 2). Optimized ET improved on coarse heuristic pruning, which itself had improved on the naive ET result taken over the initial, raw set of sequence homologs. This held for every Q_m, independent of the ranking method. Specifically, the robust rvET method yielded the best final overlap between top-ranked residues and known functional sites, with z-scores rising as much as 9% (see Table 2). The ivET method, which is sensitive to sequence perturbations, gained the most (up to 15% z-score increases) but still lagged behind rvET. Strikingly, similarity among sequence neighbors alone, measured via the $Q_{sequence}$ measure, was sufficient to improve overlap of known site ($\langle z_o \rangle$ increased 7%). Thus, ET rank similarities among neighbors computed without knowledge of the 3D structure are significant in their own right. Overall, the rank

Table 2. *Training Set Optimized to Find a Better Sequence Selection Using the rvET, ivET, and Shannon Entropy Methods for the Individual Quality Measures*

	$\langle z_o \rangle$		
Q	rvET	ivET	Shannon Entropy
No pruning	3.14	1.08	3.28
Pruning only	3.71	2.98	3.61
Cluster, 1	3.75 (+1.1%)	3.39 (+13.8%)	3.65 (+1.1%)
Cluster, 2	3.89 (+4.9%)	3.38 (+13.4%)	3.69 (+2.2%)
Cluster, 3	4.06 (+9.4%)	3.45 (+15.8%)	3.70 (+2.5%)
Surface	4.05 (+9.2%)	3.45 (+15.8%)	3.64 (+0.8%)
Sequence	3.96 (+6.7%)	3.35 (+12.4%)	3.76 (+4.2%)
Contrast	4.07 (+9.7%)	3.45 (+15.8%)	3.71 (+2.8%)
RI	3.68 (−0.8%)	3.08 (+3.4%)	3.58 (−0.8%)
Combined measure	4.16 (+12.1%)	—	—

All quality measures were shown to improve the overlap except the Q_{RI} which decreased the overlap measure $\langle z_o \rangle$ in the case of the rvET and Shannon Entropy methods. The small decrease of $\langle z_o \rangle$ due to Q_{RI} optimized may be because the value of the measure is already near maximized. The optimization with the ivET method had a larger improvement due to a new sequence selection but did not give the equivalent results of the rvET method before optimization. The $\langle z_o \rangle$ for the pruned set is considered the original/starting value for the alignments described in "Materials and Methods" section.

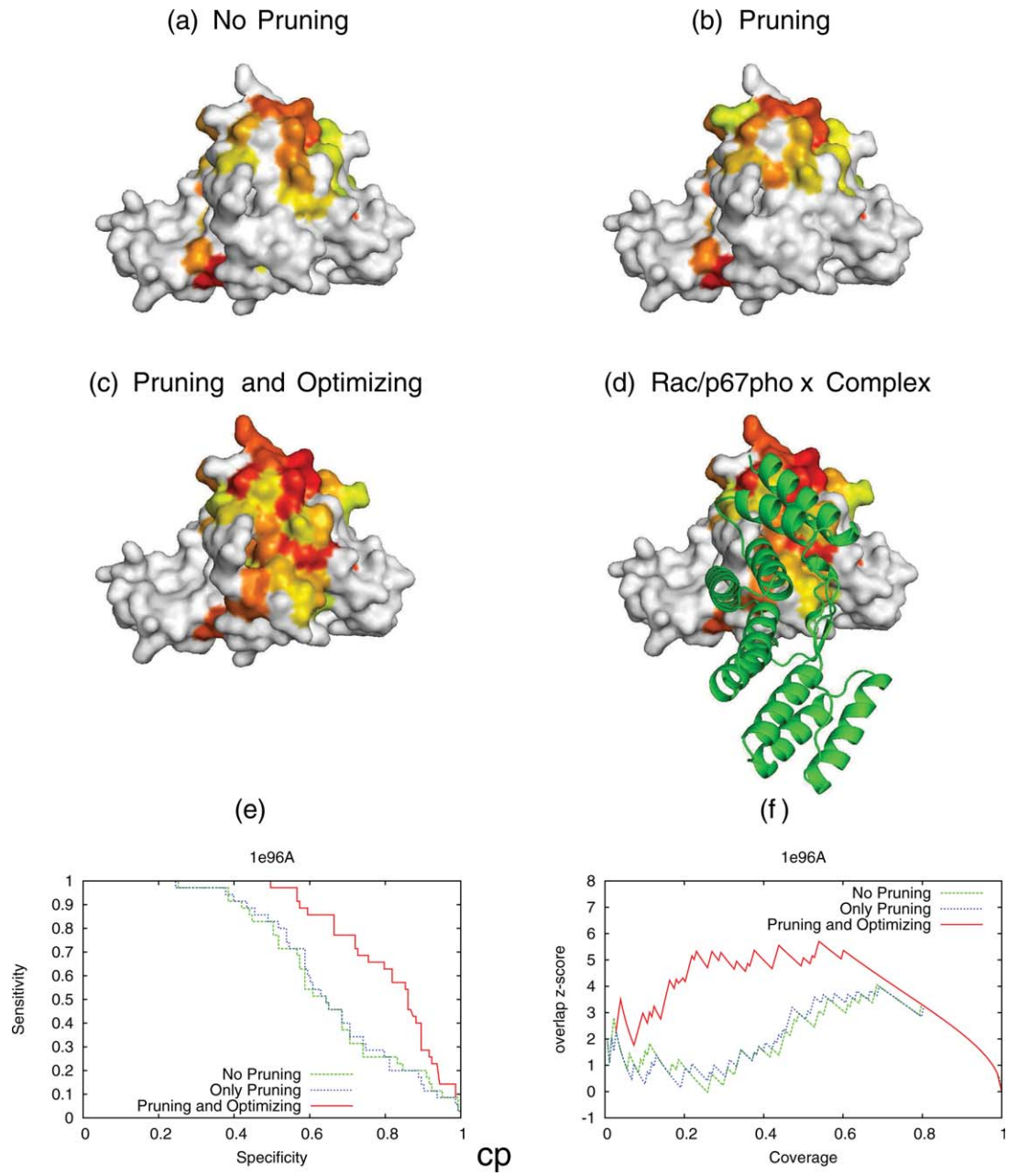

Figure 6. The sequence selection was optimized with quality measure Q_{contrast} for human Rac/p67phox complex [PDB 1e96]. The top 25% ranked residues before and after the optimization are shown here. The individual rankings with no pruning (a), only pruning (b) and after optimization (c) are shown. (d) shows the actual protein–protein interface. The bound protein p67phox is shown in green. Before optimization the average overlap *z*-score $\langle z_o \rangle$ after pruning is 0.96 while the optimization improves $\langle z_o \rangle$ to 2.76. The new alignment predicts more residues proximate to the known protein-protein interface. The optimization of the sequence selection dramatically improves the ability to predict the interfaces. An interactive view is available in the electronic version of the article.

distribution features measured by Q_{m} are sufficiently correlated with ET accuracy to inform sequence selection and to optimize ET results.

In practice, the human Rac/p67phox complex [PDB 1e96][53] illustrates these gains (Fig. 6). The GTPase Rac and p67phox assemble to form an active enzyme complex, the NADPH oxidase, which generates superoxide in the phagosome of neutrophils as part of their attempts to kill bacteria during infection. After collecting BLAST hits [Fig. 6(a)], culling sequences with blatant problems [Fig. 6(b)] and further Q_{contrast} optimization [Fig. 6(c)], the top 25% ranked residues are shown in rainbow coloring (Red is most important and yellow is 25th percentile rank). The bound protein p67phox is shown in green [Fig. 6(d)]. Optimization specifically improved the resolution of the protein-protein interface, with the additional recovery of 5 interfacial residues (I21, T24, T25, F28, D29). Likewise, the receiver-operator curve (ROC) of sensitivity versus specificity [Fig.

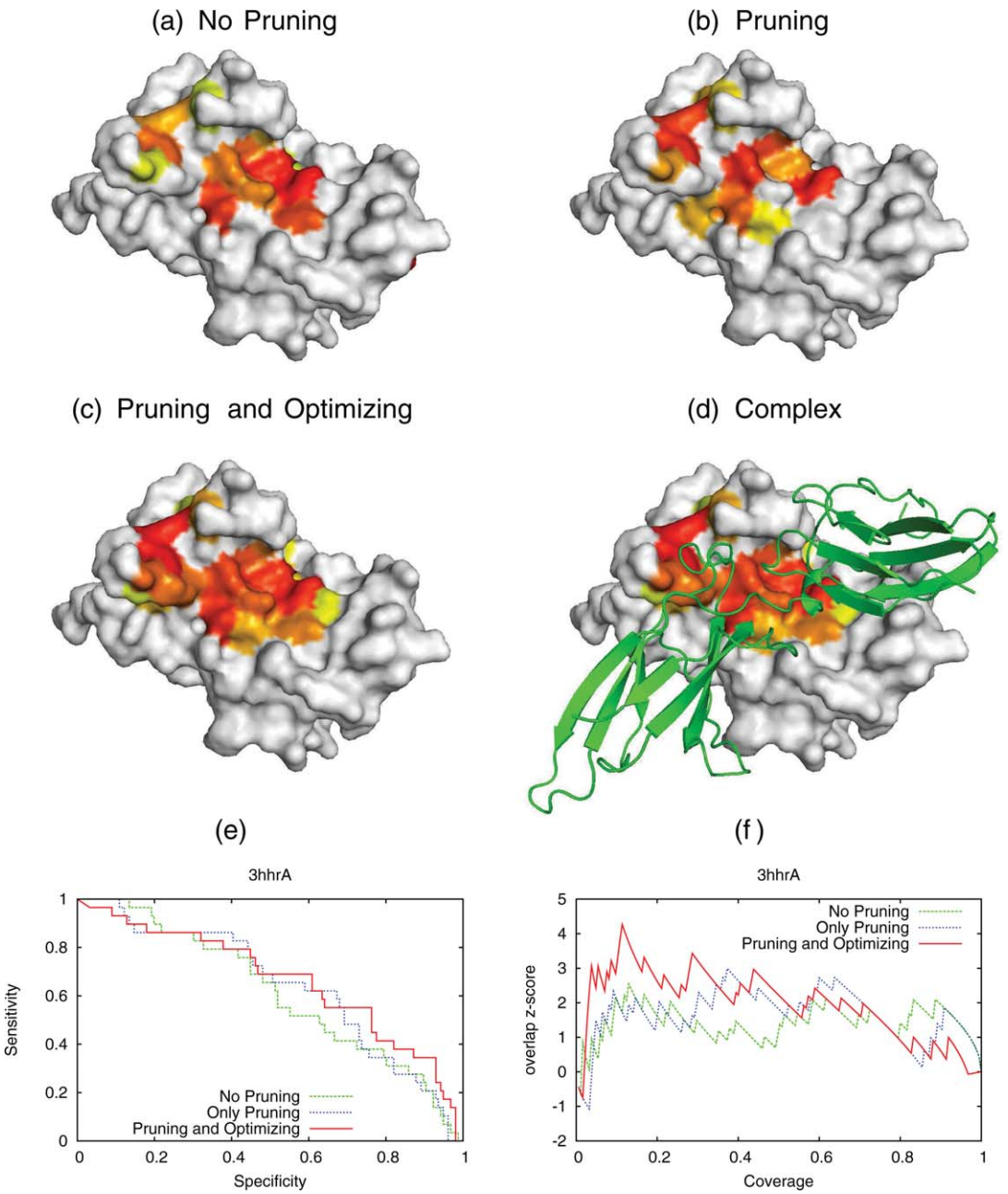

Figure 7. The optimization was performed with the $Q_{surface}$ quality measure for the human growth hormone and receptor complex [PDB 3hhr]. The individual rankings with no pruning (a), only pruning (b) and after optimization (c) are shown (Red is most important and yellow is 25th percentile rank). The new selection of sequences enables the ranking method to recover the protein–protein interface with the receptor (shown in green). The average overlap *z*-scores starts $\langle z_o \rangle = 1.30$ (no pruning), after pruning $\langle z_o \rangle$ is 1.48 and after quality measure optimization the $\langle z_o \rangle = 3.14$. The new sequence selection improves the ability to the predict the protein interface.

6(e)] and the overlap *z*-scores [Fig. 6(f)] improved with the optimized ranks.

Similar observations held in the human growth hormone and receptor complex (Fig. 7, PDB 3hhr).[54] This complex comprises the growth hormone (Chain A) bound asymmetrically to two receptor molecules (Chain B & C) and it is essential for growth and development and a potential drug target.[55,56] This time, the new selection of sequences illustrated in Figure 7 was guided by $Q_{surface}$ and it enabled the ET recovery of the protein–protein interface with the receptor (Chain B). The color-coding is as before.

Test set: Combined quality measures

Next, we asked whether the Q_m's might be combined together. This is plausibly useful since each Q_m responds slightly differently to different perturbations and, in turn, optimizes different ranking methods to different extents. After trial and error, a single composite score $Q_{composite}$ for the ET ranks

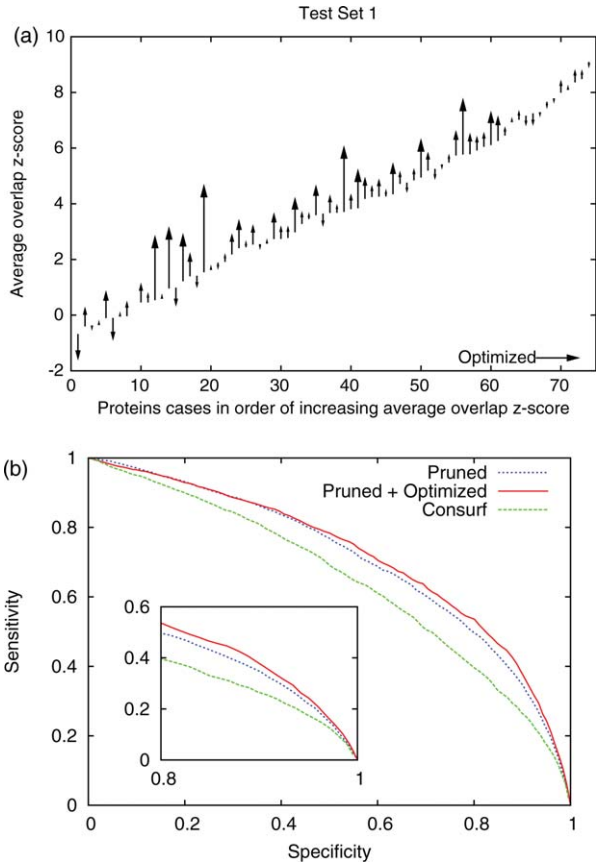

Figure 8. Optimization of the sequence selection using the combined quality measure further improved functional site prediction. Best results were obtained by first pruning the alignment and then followed by quality measure optimization with a combination of the standard score of the quality measures, $Q_{surface}$, $Q_{structure,2}$, $Q_{sequence}$, and $Q_{contrast}$. (a) The diagram shows the functional site measure $\langle z_o \rangle$ before and after the optimization of the pruned alignments is compared for the 74 individual proteins. The average overlap *z*-scores increased by 12% when rankings depend the optimized alignments compared to the pruned only. (b) The differences in methods can also be seen in receiver-operator curve. The pruned traces and pruned/optimized out performed the Consurf[20] results.

emerged. It combined the standard scores of $Q_{surface}$, $Q_{structure,2}$, $Q_{sequence}$, and $Q_{contrast}$ and it improved the average *z*-score $\langle z_o \rangle$ of the 74 test proteins (on which it was trained) from 3.71 to 4.16 (+12%) (shown in Fig. 8). This suggested that, independent of the ranking method, functional site prediction improved the most when sequence selection led to ET ranks that are the most evenly smoothed out and concentrated over the whole structure, its surface, or the sequence.

To confirm these results are free of circular bias, ET optimization guided by this composite score $Q_{composite}$ was next tested in 110 unrelated proteins taken from the literature.[57,58] Their known ligands defined the gold standard functional sites from PDBsite.[59] The optimized ET overlap *z*-score $\langle z_o \rangle$ improved was 3.75, and 8% improvement on the standard ET server (3.46) [Fig. 9(a)], at percentile ranks within 20%. For reference, another functional site detection method, Consurf,[20] yielded overlaps with average *z*-scores $\langle z_o \rangle$ of 2.17. Receiver-operator curve further illustrate the gain in sensitivity and specificity after ET optimization [Fig. 9(b)]. The equivalent results (standard ET, optimized ET and Consurf) for each proteins are in Supporting Information. The average overlap *z*-scores decreased in a few cases, those proteins typically had multiple ion-binding sites. The prediction would then improve for one site but lose overlap with respect to the secondary sites

Bovine ribonuclease A [PDB 7rsa][60] illustrates the gain accuracy. The enzyme has four catalytic residues (H12, K41, H119, and F120). Figure 10 shows the catalytic residues (spheres) and the ET top-ranked residues (colored by rank, red is most important evolutionarily and yellow is at the 20% ET rank). To recover all four catalytic residues the standard ET needed a coverage reaching as far as a percentile rank of 52%. By contrast, the optimized

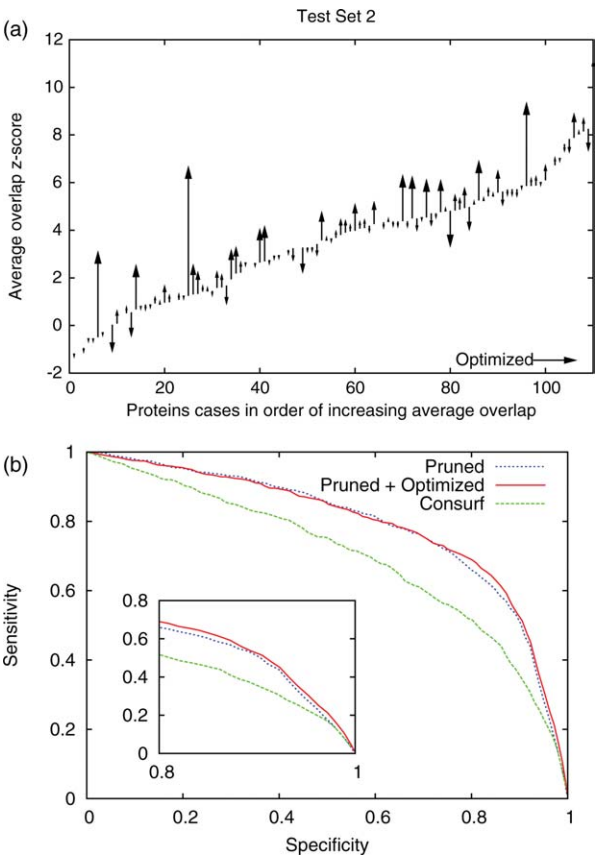

Figure 9. To test quality measure optimization method a second set was optimized for improvement in site prediction. The average *z*-score before and after the optimization for the 110 proteins was compared. (a) We found that after optimized sequence selection the dataset improved site prediction (average *z*-score improved from 3.46 to 3.75, an 8% increase). (b) The pruned traces and pruned/optimized out performed the Consurf results.

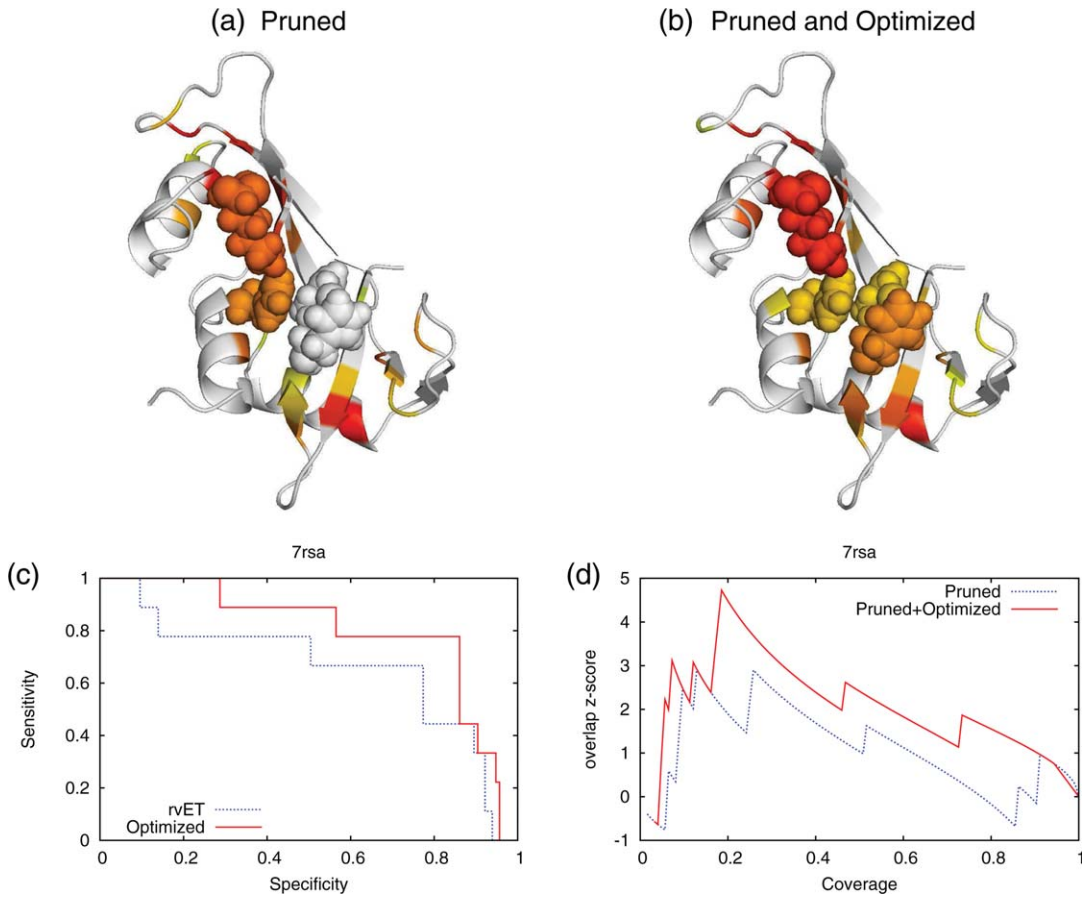

Figure 10. The example of the optimized sequence selection for phosphate-free bovine ribonuclease [PDB 7rsa] known to have an active site with catalytic residues. The top 20% ranked residues before (a) and after the optimization (b) are shown in both diagram. Residues marked red are most important and yellow are the 20th percentile rank. The overlap *z*-scores (c) and sensitivity/specifity (d) had significant improvement with a new selection of sequences based on quality measures.

ET identified all of them with a scant coverage of only the 20th percentile rank. Thus, maximizing $Q_{\text{composite}}$ significantly improved the resolution of the functional sites.

Application: Annotation set

To assess more generally the meaning of an of 8% increase in the *z*-score of functional site overlap, we asked next whether it improved function prediction via 3D-templates. This is a stringent test for two reasons: it requires large-scale comparisons of traces over a representative subset of the PDB; and it focuses specifically on molecular determinants of function as defined by a few of the very best ranked residues, so accurate ET rank order is paramount.

In more details, 3D templates are small structural motifs made up of just a few (six) of the most functionally important neighboring surface residues in a (query) structure. Ideally, they embody the key residues that are necessary and sufficient to determine function, and a classic example is the catalytic triad of serine proteases. When such templates can be matched in other (target) structures in terms of geometry and evolutionary importance, repeatedly and reversibly,[47,48,50] then such matches are likely to be relevant, rather than random, and to indicate that the query and the target have the same enzymatic activity and hence the same Enzyme Commission (EC) number.

The challenge is that for most proteins, the functional sites are not known a priori. Hence there are no obvious templates. The Evolutionary Trace Annotation (ETA) server[61] obviates the need for any prior knowledge of function, functional site location, and functional site composition by building templates solely on the basis of residue rank order and distribution: it picks a six residue template from clusters of top-ranked surface residues. Then it searches and finds matches across the PDB as described above to suggest likely functions. Recently, when ETA was controlled with the standard rvET analysis on 1217 structural genomics enzymes that were already annotated with EC numbers, the positive predictive value (PPV) was 93%, but the sensitivity was much lower, 43%.[47] A prediction is correct if the first three digits of the EC annotation are correct. Typically, this defines the chemical reaction, although not its substrate, which would require the fourth digit.

Table 3. *Protein Set Annotated by the ETA Method Using Default Alignment and the Optimized Sequence Selection Alignment*

	Pruned	Optimized
Number of proteins	1217	1217
Number of predictions	522/1217 (43%)	690/1217 (57%)
Correct predictions	483/522 (93%)	648/690 (94%)

The new selection of sequence made a dramatic improvement in the number of prediction without compromising accuracy. The sensitivity increased to 53% from 40%. The quality measure optimization will contribute considerably to prediction for functional annotation.

Here, ETA was run again on the same set, but this time using optimized traces to create templates for both the 1217 control queries and the target set of all annotated 2006 PDB90 protein structures. Overall, for 1217 proteins, the three digit EC PPV rose to 94% from 93% (Table 3). More strikingly, sensitivity rose to 53% from 40%, where PPV = correct predictions/(correct predictions + incorrect predictions) and sensitivity = correct predictions/number of proteins in the test set.

This trend was robust, even when trivial matches to proteins of high sequence identity are progressively removed from consideration. For example, at the 40% threshold (meaning that all annotations are based on matches to other structures with less than 40% identity), the three digit EC PPV rose to 92% from 89%, and sensitivity rose to 30% from 25% (Fig. 11). ETA with optimized ET added 243 predictions, where 227 were correct and 16 were not. Conversely, optimized ETA missed prediction for 75 cases that unoptimized ETA analysis alone had recovered. Thus, overall, optimized traces improve the rank distribution sufficiently to raise the quality of picked templates, the relevance of their matches and overall net predictions of enzyme function.

As an example, standard unoptimized ETA found no matches for the template it extracted from Dephospho-CoA kinase (EC 2.7.1.24) (Dephosphocoenzyme A kinase) (tm1387) from *Thermotoga maritima* at 2.60 *Å* resolution [PDB 2grj; chain A]. That template consisted of residues: 12G, 13K, 113G, 142L, 134R, 139D and 142L. The optimized ETA, however, created a different template (see Fig. 12) in which four of six residues were different: 6T (old ET percentile rank 10.3% → new percentile rank 2.9%), 84H (7.4% → 5.1%), 85P (10.9% → 4.0%), 107A (8.0% → 3.4%) while 12G (1.7% → 2.9%) and 13K (1.7% → 2.9%) were unchanged. The average percentile rank of the optimized template improved from 6.7% to 3.5%, and ETA was able to match a dephospho-coenzyme A kinase from *Haemophilus influenzae* [PDB 1jjv; chain A] of 29% sequence identity with 2grj (chain A), leading to a correct prediction of EC 2.7.1.24.

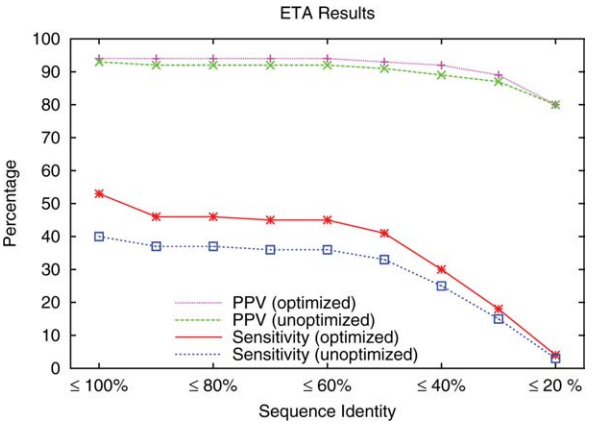

Figure 11. ETAs performance for 1217 enzymes with optimized and unoptimized ET. Positive predictive value (PPV) and sensitivity are calculated removing matches above a sequence identity threshold.

Discussion

This study is part of a long-term effort to identify evolutionary hotspots[27] in proteins in order to design functional variants[62] or peptidomimetics[63] that

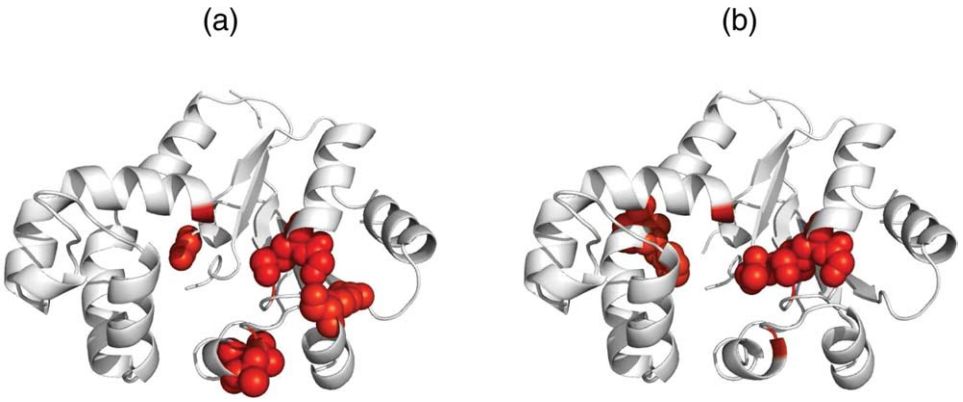

Figure 12. Pictures show the ETA templates as spheres on the PDB 2grj (chain A) structure. Both templates are taken at 5.14% ET percentile rank. Left structure (a) shows the template from unoptimized ET while the right (b) is the template from quality measure optimized ET.

selectively perturb pathways involved in signaling,[38,63,64] transcription,[65,66] or genomic stability.[34] The approach relies on the Evolutionary Trace, a method that integrates sequence, structure and function analyses into a single framework to characterize structural sites and functional residues. Some recurrent features of top-ranked ET ranks residues[27] are that: these top-ranked residues (in the 10th, 20th, 30th top-percentile rank) cluster non-randomly in protein structures[30]; and these clusters overlap significantly with, and therefore reveal, functional sites.[31,67] These observations are highly reliable and can efficiently guide experiments, for example, to separate functions,[8,34] rewire specificity,[29] design peptide inhibitors,[63] or reveal the conformational trigger of an allosteric pathway and recode it to respond to a different ligand.[68] Beyond these varied experimental case studies, ETA function prediction further validated the basic premise that clusters of top-ranked ET residues point to functionally essential residues, but this time on a large scale.

These prior results suggest that ET ranks highlight fundamental, general and useful patterns linking the distribution of evolutionary importance in sequence residues to their structural location and to their biological roles. The question posed here, is whether other quantifiable features can be defined to improve the resolution of this evolutionary relationship, and to lead to more accurate ranks, more accurate functional sites, and more accurate function predictions.

All seven of the quality measures proposed here do so, as does the 8th composite one. They guided sequence selections that improved the match between top-ranked residues and functional sites, independent of the precise ranking algorithm. The rise in statistical significance, the *z*-score of overlap, with the composite quality measure was 8% in 110 proteins unrelated to the training set. This gain translated into better resolution of the functional sites in both individual case studies and on a proteome-wide scale for function prediction. ETA sensitivity rose sharply (a 13% gain from 40% to 53%, which is a 33% relative increase) with no loss of specificity (positive predictive value rose 1% from 93% to 94%). The results reflect the large impact of even subtle improvements in ET ranks.

These data also confirm the hypothesis: for most, if not all proteins, quantifiable features of ET rank distributions can be optimized towards more accurate views of the sequence-structure-function evolutionary relationship. But the multiplicity of Q_m, their complementarity, and the better performance of the ad hoc, composite quality measure $Q_{composite}$, suggest that a deeper, more general and more basic feature of the distribution of evolutionary importance in proteins is at play.

It is therefore noteworthy that a common theme among the Q_m is to focus on neighboring residues and rate whether they have similar evolutionary importance. The more this is so, the more top-ranked residues will cluster,[30] and ET accuracy measures will increase. This suggests two related broader reformulation of our results: First, that ET ranks should distribute continuously, that is, such that from one neighbor to the next the change in evolutionary importance is as small, or smooth as possible. Second, that ET ranks should be most ordered, that is, their distribution entropy is least. Both statements hold for sequence neighbors, or spatial neighbors in the structure, and suggest, most simply that over time, molecular evolution operates in sequence and structure in an orderly and continuous fashion.

In this light, mutations are still random events, but the physical and function constraints that lead to natural selection lead to the apparent order and continuity of evolutionary importance. The different quality measures reflect different ideas of what organization of evolutionary ranks best reflects nature: how to measure the continuity or the entropy of a distribution of evolutionary importance in the discrete context of a sequence or structure? In practice, the suitability of aligned sequences for site prediction analysis can be measured and optimized to improve the statistical significance of functional site predictions. These gains are scalable to the proteome and carry over to the prediction of functional determinants since these translate to improved function prediction by 3D templates.

In summary, the results suggests that a finer definition of the "clustering" property that ties top-ranked residues with function is the continuity and order of ET ranks distributions in sequence and in space. The generality of this statement is supported by all the correlations between Q_m and ET accuracy, which is so reliable and so general that it guides sequence selections that optimize ET, and ETA, on a proteomic scale. The maximal rank continuity suggests a more succinct formulation than the phenomenological (ad hoc) nature of the quality measures themselves. It remains to be tested in the future whether other, and more general means to improve rank continuity can further improve ET, and in so doing point to a more definitive ET rank order quality than $Q_{composite}$. For now, this study provides significant improvements to the automated, large-scale functional site identification and the annotation of their key residues and functions.

Materials and Method

Quality measures

The first group focuses on the notion of "clusters" of top-ranked residues. These are residues that are in

contact and are evolutionarily important, given ET rank threshold. The more such residues are in contact at a given threshold, the greater Q_{cluster} will be. Q_{cluster} is an accumulative value derived from the clustering z-scores at each unique rank,

$$Q_{\text{cluster}} = \frac{1}{L}\sum_{i}^{L} z_{\text{c}}^{(i)}(1 - c_i) \tag{1}$$

where $z_{\text{c}}^{(i)}$ is the clustering z-score of the residues within a threshold based on the rank of residue i and c_i is the fraction of the residues falling within this threshold. The term $1 - c_i$ weighs more heavily z-scores arising from top-ranked residues. L is the residue length of the protein structure. The clustering z-score z_{c} is the distance of w, the actual clustering of top-ranked residues, from its average expected value $\langle w \rangle$, measured in units of standard deviation σ and is expressed

$$z_{\text{c}} = \frac{w - \langle w \rangle}{\sigma}. \tag{2}$$

Finally, the quantity w is defined by the top-ranked residues in contact and can be expressed

$$w = \sum_{j>i}^{L} S(i)S(j)A(i,j)f(i,j). \tag{3}$$

Here, the adjacency matrix $A(i,j)$ assigns 1 to any pair of residues i and j if they are defined as neighbors (within 4 Å), and is 0 otherwise. S_i is the selection threshold which assigns the value of 1 if the residue i falls into a given c_i. Detailed explanations of the methods used to calculate $\langle w \rangle$ and σ can be found in Mihalek *et al.*[32]

Of note, the function $f(i,j)$ weighs the contribution of residues that are near in the structure but not in the sequence. Until now contacts between residues that were further apart in the sequence were weighed ($f(i,j) = j - i$, where i and j is the residue numbering) more heavily.[44,45] But other choices (referred to as $Q_{\text{structure},1}$) are possible including no special weight $Q_{\text{structure},3}$, or a drop-off such as the square root is taken as the weight $Q_{\text{structure},2}$. This gives rise to three quality measures.

Moreover, clustering among surface residues may be more relevant to identify functional sites for protein ligand interactions. This is the purpose of Q_{surface}, constructed as shown in Eqs. (1) to (3) where now $f(i,j)$ is equal 1 if both i and j are on the surface residues (solvent accessibility > 5 Å^2 according to DSSP[69]) and equal $f(i,j) = 0$ otherwise. This yields a two-dimensional measure of quality.

One may further reduce dimensionality and consider sequence clustering only. This yields Q_{sequence}, akin to the previous Q_{cluster} measures but with the adjacency matrix $A(i,j)$ set at 1 only for residues that are next to each other in sequence, and set to 0 otherwise. This quality measure is structure independent.

We also consider the previously defined Rank Information (Q_{RI})[46] which does not explicitly rely on clustering, and which is also structure independent. (Q_{RI}) is a product of two expressions related to the information content of the ET rank distribution

$$Q_{\text{RI}} = \text{TI} \times \text{RE} \tag{4}$$

where TI is the trace integral and RE is the rank entropy. The trace integral sums ET ranks over all possible positions and is written

$$\text{TI} = \sum_{r=1}^{N} f_{\text{r}} \frac{(N + 1 - r)}{N}. \tag{5}$$

where f_{r} is the frequency a rank appears in the analysis and N is the number of sequences. The value r is an integer position based on the score from the ranking method, where it is the integer value modified by leaving gaps before the sets of equally ranked items. For example, if four residues have the evolutionary scores 1.0, 1.1, 1.1, and 1.3 then r is equal to 1, 3, 3, and 4, respectively. This transformation was necessary to compare methods. The rank entropy measures the diversity of the rank values over the possible positions as shown

$$\text{RE} = -\sum_{r=1}^{N} f_r \log f_r. \tag{6}$$

The origin of Q_{RI} is the following. First, the fewer data are corrupted or inconsistent, the better the ranks of functional residues should be. So TI should be as large as possible. However, in the extreme, this process leads to sequences that are identical to each other so that every residue is top-ranked. RE balances this process by requiring that the rank distribution remains as diverse as possible.

Last, a partially related view of clustering among top-ranked residues is that the relative rank difference between contact pairs of the individual atoms in the residues should be minimized. This observation introduces the notion of smoothness within the distribution of ranked residues and it is quantified by Q_{contrast}. Q_{contrast} measures the rank difference between residues within the structure. Q_{contrast} is calculated as follows

$$\frac{1}{Q_{\text{contrast}}} = \frac{\sum_{j>i}^{L}\sum_{a_i,a_j} A(a_i,a_j)|(c_j - c_i)|}{n} \tag{7}$$

The adjacency matrix $A(a_i,a_j) = 1$ if the atoms a_i and a_j are within a minimum distance of 4 Å (otherwise $A(a_i,a_j) = 0$). n is the number of atoms within the structure. The value of c_i contains the

evolutionary rank information as a fraction of residues with the evolutionary rank of residue i or better. The reciprocal is taken since we want the function to be maximized similar to the other quality measures. In this study, Q_{contrast} will only be considered across the whole structure, but it can be narrowed to surface residues only.

Finally, to assess the recovery of known sites, we use A_{overlap} which is the measure of overlap with the "gold standard" functional site. This function is derived from the overlap z-scores z_o, which are based on the hypergeometric distribution describing the overlap between the "gold standard" residues and the ranked residue (Fig. 1). The measure A_{overlap} similar to Q_{cluster} is defined as

$$A_{\text{overlap}} = \frac{1}{L}\sum_{i}^{L} z_o^{(i)}(1 - c_i) \tag{8}$$

where $z_o^{(i)}$ is referred to as the overlap z-score corresponding to the residues within c_i.

Note that in order to normalize the gains in functional site overlap among many different proteins, they are expressed in terms of $\langle z_o \rangle$, the average statistical z-score of overlap between the functional site and the trace residues at the percentile ranks within the top 20%. We focus on the residues in the top 20% because the performance of the method at top ranks is more relevant when guiding experiments.

Residue ranking methods

The focus of our studies will be integer value ET, Shannon Entropy and a hybrid method (real value ET). Mihalek, *et al*. discuss a comparative study of the ability of these methods to predict important residues.[28]

The integer value ET rank[27] (ivET) for the residue position i in the query protein can be expressed

$$r_i = 1 + \sum_{n=1}^{N-1} \delta \tag{9}$$

The summation considers all the nodes N (branches) in the evolutionary tree. The value $\delta = 0$ if residue position i is conserved within the sequences that make up the node in the evolutionary tree, and $\delta = 1$ otherwise. The ranking method ignores the groups that are not completely conserved at position i. Assigning a rank r_i to each of the residues leads to a relative ranking scheme: given any two residues, the one with smaller rank r_i is considered more important. The magnitude of r_i for residue position i is not important except relative to the ranks at other residue positions in the query protein. Each method we consider shares this idea.

Shannon Entropy is a measure of variability at a given position in a set of aligned sequences.[51] The rank s_i for residue position i is defined as

$$s_i = -\sum_{a=1}^{20} f_{ia} ln f_{ia} \tag{10}$$

where f_{ia} is the frequency that amino acid a appears in the column containing residue position i.

Real value ET (rvET) method[28] ranks the evolutionary importance of residues in a protein family, which is based on the column variation in multiple sequence alignments and evolutionary information extracted from the phylogenetic tree. Shannon Entropy is calculated for the entire alignment, and then recalculated for all the subgroups of the alignment selected by the phylogenetic tree. The rank ρ_i of residue i is calculated as follows:

$$\rho_i = 1 + \sum_{n=1}^{N-1} \frac{1}{n} \sum_{g=1}^{n} \left\{ -\sum_{a=1}^{20} f_{ia}^{g} ln f_{ia}^{g} \right\} \tag{11}$$

where f_{ia}^{g} is the frequency of the amino acid of type a within the sub-alignment of group g. The number of possible nodes in the evolutionary tree is $N - 1$ where N is the number of sequences in the alignment. The nodes in the phylogenetic tree are numbered in the order of increasing distance from the root. Note that the node $n = 1$ term would be Shannon Entropy.

Training set

The dataset used to determine best set of quality measures consists of 74 proteins with protein–ligand and/or protein–protein interactions. The set was chosen to be diverse in function and to include proteins with more then one functional site. The protein data bank IDs are: 16pk, 1a09, 1a0oE, 1a22A, 1a22B, 1a2kA, 1a2kD, 1a3k, 1a48, 1a4mA, 1a53, 1a59, 1a6m, 1a6q, 1a80, 1aca, 1ad3A, 1ai2, 1aj2, 1aj8A, 1aky, 1am1, 1amk, 1aonF, 1ars, 1aru, 1ast, 1axn, 1b54, 1bag, 1bqk, 1bto, 1c1bA, 1cg0, 1cio, 1cvjA, 1cxzA, 1dam, 1dig, 1dqr, 1dqx, 1e96A, 1e96B, 1ee9, 1efaB, 1eg2, 1eje, 1elrA, 1elwA, 1f6mA, 1f88A, 1finA 1finB, 1fjmA, 1fqjB, 1gnjA, 1jfiB, 1k7vA, 1ng1, 1nzcA, 1pvdA, 1qumA, 1qupA, 1rrpA, 1rrpB, 1vh4A, 1w1uA, 1ycsA, 1ycsB, 2bif, 2mjpA, 2msbA, 3hhrA, 6gst. The "gold standard" functional sites of the protein–ligand interactions are defined by the database PDBsum.[70] The protein–protein functional sites are defined as the residues within five angstroms distance of the residues in the second protein.

Pruning algorithm

For each query protein, a set of sequences were obtained with the default settings of the ETviewer.[71] The set was retrieved with BLAST[52] (using NCBI's non-redundant protein sequence database, the

blosum62 substitution matrix, and default parameters). The top 500 homologs with an *e*-value better than 0.05 were retrieved from NCBI's Protein database. CLUSTALW[72] (using the default parameters) was used to generate a multiple sequence alignment for the query structure. The set was pruned to remove evolutionary outliers and sequence fragments (referred to as Pruned set): sequences were removed if sequence identity was less then 26% with the query and less then 70% of the query sequence length. The sequences were then re-aligned. The set of alignments was used to test performance of the quality measures and ranking methods.

Feedback optimization algorithm

The sequence selection algorithm aims to eliminate problem sequences rather than to pinpoint a single best set. The reason is that many similarly good selections differ only by any number of combinations of close homologs, which would have no impact on the distribution of top-ranked residues. Starting from the sequences collected as described in previous section, the algorithm identifies a reasonable initial selection in the first two steps, and then adds new sequences in the third step, guided each time by the quality score measure Q_{m}. Specifically:

1. The evolutionary tree nodes that contain the query protein as a leaf are used to define nested sequence selection and each one is traced. The node with the best Q_{m} value defines the starting selection, thus initially removing outlying homologs that may conflict with closer homologs to the query.
2. Each of the smaller nodes of this new tree is then removed, in turn and one at a time, and the remaining sequences are retraced. If the value of the quality measure increases then the node is removed from the analysis. Thus as the tree, and sequence selection shrinks, Q_{m} increases further.
3. Finally, all individual sequences are then added/removed from analysis and the Q_{m} is evaluated based on the new rankings due to the change in sequence selection. Thus, any one of the sequences removed in earlier steps may be incorporated back into the tree. The sequences are added/removed in the order of the value of the quality measures Q_{m}, each time retesting the sequence against the new evolutionary tree. The sequences not selected are repeatedly tested until the subset of sequences is left unchanged. The algorithm allows for five iterations to insure convergence but in a majority of the cases the selection converges after no more then three iterations.

Composite quality measure

$Q_{\mathrm{composite}}$ is a single score made up of the standard scores of a subset of the quality measures. $Q_{\mathrm{composite}}$ is formulated

$$Q_{\mathrm{composite}} = \sum_{\mathrm{m}} \frac{Q_{\mathrm{m}} - \langle Q_{\mathrm{m}} \rangle}{\sigma_{Q_{\mathrm{m}}}}. \qquad (12)$$

where the sum is over subset of quality measures chosen. The expected average of a quality measures $\langle Q_{\mathrm{m}} \rangle$ and standard deviation $\sigma_{Q_{\mathrm{m}}}$ are evaluated from the values of the quality measures obtained during the steps of the Feedback Optimization Algorithm.

Sensitivity and specifity

The receiver-operator curve (ROC) was calculated on the test sets as follows: sensitivity is found as $\frac{\mathrm{TP}}{\mathrm{TP+FN}}$, where a true positive (TP) is the number of residues defined to be part of the "gold standard" functional site and predicted by the ranking method, while a false negative (FN) is the "gold standard" residues that the method misses. The specificity is equal to $\frac{\mathrm{TN}}{\mathrm{TN+FP}}$, where the true negative (TN) is neither "gold standard" nor predicted by the ranking method, while the false positive (FP) is the residues not listed as part of the "gold standard" site but still predicted by the ranking method. The ROC curve was calculated with the total TP, FN, TN, and FP found as the rank coverage increased in the test sets.

Comparison to Consurf

The amino acid conservation scores were taken from the pre-calculated results obtained from the Consurf website[20] We were unable to obtain a pre-calculated result for PDB ID 1cxzA. For the 73 proteins we found Consurf results, we found the average *z*-scores for predictions within the top 20% ranked for Consurf ($\langle z_o \rangle = 2.75$), were lower then our standard ET server ($\langle z_o \rangle = 3.89$), and the optimized ET ($\langle z_o \rangle = 4.20$). We were also unable to obtain a pre-calculated result for PDB 1iyu. After adjusting for the missing protein, the standard ET server traces ($\langle z_o \rangle = 3.49$) and optimized ET traces ($\langle z_o \rangle = 3.84$) out-performed the Consurf results ($\langle z_o \rangle = 2.17$). A complete comparison of the individual proteins making up the test set can be found in Supporting Information.

Acknowledgment

The authors thank Panos Katsonis and and Dan Morgan for helpful discussions contributing to the article. A.D.W., S.E. and R.M.W. were also supported by training fellowships from the National Library of Medicine to the Keck Center for Interdisciplinary Bioscience Training of the Gulf Coast Consortia.

References

1. Lee D, Redfern O, Orengo C (2007) Predicting protein function from sequence and structure. Nat Rev Mol Cell Biol 8:995–1005.
2. Laskowski RA, Thornton JM (2008) Understanding the molecular machinery of genetics through 3D structures. Nat Rev Genet 9:141–145.

3. Jiang L, Althoff EA, Clemente FR, Doyle L, Röthlisberger D, Zanghellini A, Gallaher JL, Betker JL, Tanaka F, Barbas CF, Hilvert D, Houk KN, Stoddard BL, Baker D (2008) De novo computational design of retro-aldol enzymes. Science 319:1387–1391.
4. Thyme SB, Jarjour J, Takeuchi R, Havranek JJ, Ashworth J, Scharenberg AM, Stoddard BL, Baker D (2009) Exploitation of binding energy for catalysis and design. Nature 461:1300–1304.
5. Hardy JA, Wells J (2004) Searching for new allosteric sites in enzymes. Curr Opin Struct Biol 14:706–715.
6. Matsumura M, Matthews B (1989) Control of enzyme activity by an engineered disulfide bond. Science 243: 792–794.
7. Clackson T, Wells JA (1995) A hot spot of binding energy in a hormone-receptor interface. Science 267: 383–386.
8. Onrust R, Herzmark P, Chi P, Garcia PD, Lichtarge O, Kingsley C, Bourne HR (1997) Receptor and beta-gamma binding sites in the alpha subunit of the retinal G protein transducin. Science 275:381–384.
9. Nobeli I, Favia AD, Thornton JM (2009) Protein promiscuity and its implications for biotechnology. Nat Biotechnol 27:157–167.
10. Ota M, Kinoshita K, Nishikawa K (2003) Prediction of catalytic residues in enzymes based on known tertiary structure, stability profile, and sequence conservation. J Mol Biol 327:1053–1064.
11. Marklund U, Lightfoot K, Cantrell D (2003) Intracellular Location and Cell Context-Dependent Function of Protein Kinase D. Immunity 19:491–501.
12. Elcock A (2001) Prediction of functionally important residues based solely on the computed energetics of protein structure. J Mol Biol 312:885–896.
13. Ondrechen MJ, Clifton JG, Ringe D (2001) THEMATICS: A simple computational predictor of enzyme function from structure. Proc Natl Acad Sci USA 98: 12473–12478.
14. Jones S, Shanahan H, Berman H, Thornton J (2003) Using electrostatic potentials to predict DNA-binding sites on DNA-binding proteins. Nucleic Acids Res 31: 7189–7198.
15. Keskin O, Ma B, Nussinov R (2005) Hot regions in protein–protein interactions: the organization and contribution of structurally conserved hot spot residues. J Mol Biol 345:1281–1294.
16. Gutteridge A, Bartlett G, Thornton J (2003) Using a neural network and spatial clustering to predict the location of active sites in enzymes. J Mol Biol 330: 719–734.
17. delSol A, O'Meara P (2005) Small-world network approach to identify key residues in protein-protein interaction. Proteins 58:672–682.
18. Lisewski AM, Lichtarge O (2006) Rapid detection of similarity in protein structure and function through contact metric distances. Nucleic Acids Res 34:e152.
19. Glaser F, Morris RJ, Najmanovich RJ, Laskowski RA, Thornton JM (2006) A method for localizing ligand binding pockets in protein structures. Proteins 62: 479–488.
20. Glaser F, Pupko T, Paz I, Bell R, Bechor-Shental D, Martz E, Ben-Tal N (2003) ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. Bioinformatics 19:163–164. Available at: http://consurftauacil/.
21. Pupko T, Bell RE, Mayrose I, Glaser F, Ben-Tal N (2002) Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. Bioinformatics 18:S71–S77.
22. Valdar W (2002) Scoring Residue Conservation. Proteins 43:227–241. Available at: http://wwwebiacuk/thornton-srv/databases/cgi-bin/valdar/scorecons_serverpl.
23. Innis CA (2007) siteFiNDER—3D: a web-based tool for predicting the location of functional sites in proteins. Nucleic Acids Res 35:W489–W494.
24. Sankararaman S, Sjölander K (2008) INTREPID–INformation-theoretic TREe traversal for Protein functional site IDentification. Bioinformatics 24:2445–2452.
25. Engelen S, Trojan LA, Sacquin-Mora S, Lavery R, Carbone A (2009) Joint evolutionary trees: a large-scale method to predict protein interfaces based on sequence sampling. PLoS Comput Biol 5:e1000267.
26. Halabi N, Rivoire O, Leibler S, Ranganathan R (2009) Protein sectors: evolutionary units of three-dimensional structure. Cell 138:774–786.
27. Lichtarge O, Bourne H, Cohen F (1996) An evolutionary trace method defines binding surfaces common to protein families. J Mol Biol 257:342–358.
28. Mihalek I, Reš I, Lichtarge O (2004) A family of evolution-entropy hybrid methods for ranking protein residues by importance. J Mol Biol 336:1265–1282.
29. Sowa M, He W, Slep K, Kercher A, Lichtarge O, Wensel T (2001) Prediction and confirmation of a site critical for effector regulation of RGS domain activity. Nat Struct Biol 8:234–237.
30. Madabushi S, Yao H, Marsh M, Kristensen D, Philippi A, Sowa M, Lichtarge O (2002) Structural clusters of evolutionary trace residues are statistically significant and common in proteins. J Mol Biol 316:139–154.
31. Yao H, Kristensen D, Mihalek I, Sowa M, Shaw C, Kavraki L, Kimmel M, Lichtarge O (2003) An accurate, sensitive, and scalable method to identify functional sites in protein structures. J Mol Biol 326:255–261.
32. Mihalek I, Reš I, Yao H, Lichtarge O (2003) Combining inference from evolution and geometric probability in protein structure evaluation. J Mol Biol 331:263–279.
33. Madabushi S, Gross A, Philippi A, Meng E, Wensel T, Lichtarge O (2004) Evolutionary trace of G protein-coupled receptors reveals clusters of residues that determine global and class-specific functions. J Biol Chem 279:8126–8132.
34. Ribes-Zamora A, Mihalek I, Lichtarge O, Bertuch AA (2007) Distinct faces of the ku heterodimer mediate DNA repair versus telomeric functions. Nat Struct Mol Biol 14:301–307.
35. Sowa ME, He W, Wensel T, Lichtarge O (2000) A regulator of G protein signaling interaction surface linked to effector specificity. Proc Natl Acad Sci USA 97: 1483–1488.
36. Shenoy S, Drake M, Nelson C, Houtz D, Xiao K, Madabushi S, Reiter E, Premont R, Lichtarge O, Lefkowitz R (2006) Beta-arrestin-dependent, G protein-independent ERK1/2 activation by the beta2 adrenergic receptor. J Biol Chem 281:261–273.
37. Rajagopalan L, Patel N, Madabushi S, Goddard JA, Anjan V, Lin F, Shope C, Farrell B, Lichtarge O, Davidson AL, Brownell WE, Pereira FA (2006) Essential helix interactions in the anion transporter domain of prestin revealed by evolutionary trace analysis. J Neurosci 26:12727–12734.
38. Kobayashi H, Ogawa K, Yao R, Lichtarge O, Bouvier M (2009) Functional rescue of beta-adrenoceptor dimerization and trafficking by pharmacological chaperones. Traffic 10:1019–1033.

39. Rost B (1999) Twilight zone of protein sequence alignment. Protein Eng 12:85–94.
40. Wilson C, Kreychman J, Gerstein M (2000) Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. J Mol Biol 297:233–249.
41. Rost B (2002) Enzyme function less conserved than anticipated. J Mol Biol 318:595–608.
42. Tian W, Skolnick J (2003) How well is enzyme function conserved as a function of pairwise sequence identity. J Mol Biol 333:863–882.
43. He Y, Chen Y, Bryan PN, Orban J (2008) NMR structures of two designed proteins with high sequence identity but different fold and function. Proc Natl Acad Sci USA 105:14412–14417.
44. Mihalek I, Reš I, Lichtarge O (2006) Evolutionary and structural feedback on selection of sequences for comparative analysis of proteins. Proteins 63:87–99.
45. Mihalek I, Reš I, Lichtarge O (2006) A structure and evolution guided Monte Carlo sequence selection strategy for multiple alignment-based analysis of proteins. Bioinformatics 22:149–156.
46. Yao H, Mihalek I, Lichtarge O (2006) Rank information: a structure-independent measure of evolutionary trace quality that improves identification of protein functional sites. Proteins 65:111–123.
47. Ward RM, Erdin S, Tran TA, Kristensen DM, Lisewski AM, Erdin S, Lichtarge O (2008) De-orphaning the structural proteome through reciprocal comparison of evolutionarily important structural features. PLoS ONE 3:e2136.
48. Kristensen DM, Ward RM, Lisewski AM, Erdin S, Chen BY, Fofanov VY, Kimmel M, Kavraki LE, Lichtarge O (2008) Prediction of enzyme function based on 3D templates of evolutionarily important amino acids. BCM Bioinformatics 9:17.
49. Polacco BJ, Babbitt PC (2006) Automated discovery of 3D motifs for protein function annotation. Bioinformatics 22:723–730.
50. Erdin S, Ward RM, Venner E, Lichtarge O (2010) Evolutionary trace annotation of protein function in the structural proteome. J Mol Biol 396:1451–1473.
51. Shenkin P, Erman B, Mastrandrea L (1991) Information-theoretical entropy as a measure of sequence variability. Proteins 11:297–313.
52. Altschul S, Madden T, Schaffer A, Zhang J, Zhang Z, Miller W, Lipman D (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402.
53. Lapouge K, Smith SJ, Walker PA, Gamblin SJ, Smerdon SJ, Rittinger K (2000) Structure of the TPR domain of p67phox in complex with RacGTP. Mol Cell 6: 899–907.
54. de Vos AM, Ultsch M, Kossiakoff AA (1992) Human growth hormone and extracellular domain of its receptor: crystal structure of the complex. Science 255: 306–312.
55. Benedini S, Terruzzi I, Lazzarin A, Luzi L (2008) Recombinant human growth hormone: rationale for use in the treatment of HIV-associated lipodystrophy. BioDrugs 22:101–112.
56. Tallet E, Rouet V, Jomain JB, Kelly PA, Bernichtein S, Goffin V (2008) Rational design of competitive prolactin/growth hormone receptor antagonists. J Mammary Gland Biol Neoplasia 13:105–117.
57. delSol A, Pazos F, Valencia A (2003) Automatic Methods for Predicting Functionally Important Residues. J Mol Biol 326:1289–1302.
58. Nimrod G, Schushan M, Steinberg DM, Ben-Tal N (2008) Detection of functionally important regions in "hypothetical proteins" of known structure. Structure 16:1755–1763.
59. Ivanisenko VA, Grigorovich DA, Kolchanov NA (2000) PDBSite: a database on biologically active sites and their spatial surroundings in proteins with known tertiary structure. In: The Second International Conference on Bioinformatics of Genome Regulation and Structure (BGRS'2000), Novosibirsk, Russia, August 7–11, 2000, Vol. 2, pp 171–174.
60. Wlodawer A, Svensson LA, Sjolin L, Gilliland GL (1988) Structure of phosphate-free ribonuclease A refined at 126 A. Biochemistry 27:2705–2717.
61. Ward RM, Venner E, Daines B, Murray S, Erdin S, Kristensen DM, Lichtarge O (2009) Evolutionary trace annotation server: automated enzyme function prediction in protein structures using 3D templates. Bioinformatics 25:1426–1427.
62. Sheikh S, Zvyaga T, Lichtarge O, Sakmar T, Bourne H (1996) Rhodopsin activation blocked by metal-ion-binding sites linking transmembrane helices C and F. Nature 383:347–350.
63. Baameur F, Morgan D, Yao H, Tran T, Hammitt R, Sabui S, McMurray J, Lichtarge O, Clark R (2010) Role for the RH Domain of GRK5 and 6 in beta2-adrenergic receptor and rhodopsin phosphorylation. Mol Pharmacol 77:405–415.
64. Sheikh SP, Vilardarga J, Baranski T, Lichtarge O, Iiri T, Meng E, Nissenson R, Bourne H (1999) Similar structures and shared switch mechanisms of the beta2-adrenoceptor and the parathyroid hormone receptor. zn(II) bridges between helices III and VI block activation. J Biol Chem 274:17033–17041.
65. Quan X, Denayer T, Yan J, Jafar-Nejad H, Philippi A, Lichtarge O, Vleminckx K, Hassan B (2004) Evolution of neural precursor selection: functional divergence of proneural proteins. Development 131:1679–1689.
66. Raviscioni M, Gu P, Sattar M, Cooney A, Lichtarge O (2005) Correlated evolutionary pressure at interacting transcription factors and DNA response elements can guide the rational engineering of DNA binding specificity. J Mol Biol 350:402–415.
67. Lichtarge O, Bourne H, Cohen F (1996) Evolutionarily conserved Galphabetagamma binding surfaces support a model of the G protein-receptor complex. Proc Natl Acad Sci USA 93:1483–1488.
68. Rodriguez GJ, Yao R, Lichtarge O, Wensel T (2010) Evolution-guided discovery and recoding of allosteric pathway specificity determinants in psychoactive bioamine receptors. Proc Natl Acad Sci USA. Apr 27; 107(17):7787–92.
69. Kabsch W, Sander C (1993) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22: 2577–2637.
70. Laskowski R, Chistyakov V, Thornton J (2005) PDBSum more: new summaries and analyses of the known 3D structures of proteins and nucleic acids. Nucleic Acids Res 33:D266–D268.
71. Morgan DH, Kristensen DM, Mittleman D, Lichtarge O (2006) ET viewer: an application for predicting and visualizing functional sites in protein structures. Bioinformatics. 2006 Aug 15;22(16):2049–50.
72. Thompson J, Higgins D, Gibson T (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680.