Published in final edited form as:

J Mol Biol. 2010 March 12; 396(5): 1451–1473. doi:10.1016/j.jmb.2009.12.037.

Evolutionary Trace Annotation of Protein Function in the Structural Proteome

Serkan Erdin^{1,2,†}, R. Matthew Ward^{1,2,3,†}, Eric Venner³, and Olivier Lichtarge^{1,2,3}

Serkan Erdin: serdin@bcm.tmc.edu; R. Matthew Ward: ; Eric Venner: ; Olivier Lichtarge: lichtarge@bcm.tmc.edu

¹Departments of Molecular and Human Genetics

²W.M. Keck Center for Interdisciplinary Bioscience Training, Houston, TX 77005

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

Abstract

By design, structural genomics (SG) solves many structures that cannot be assigned function based on homology to known proteins. Alternative function annotation methods are therefore needed and this study focuses on function prediction with three-dimensional (3D) templates: small structural motifs built of just a few functionally critical residues. Although experimentally proven functional residues are scarce, we show here that Evolutionary Trace (ET) rankings of residue importance are sufficient to build 3D templates, match them, and then assign Gene Ontology (GO) functions in enzymes and non-enzymes alike. In a high specificity mode, this Evolutionary Trace Annotation (ETA) method covered half (53%) of the 2384 annotated SG protein controls. Three-quarters (76%) of predictions were both correct and complete. The positive predictive value for all GO depths (all-depth PPV) was 84%, and it rose to 94% over GO depths 1–3 (depth 3 PPV). In a high sensitivity mode coverage rose significantly (84%) while accuracy fell moderately: 68% of predictions were both correct and complete, all-depth PPV was 75%, and depth 3 PPV was 86%. These data concur with prior mutational experiments showing that ET rank information identifies key functional determinants in proteins. In practice, ETA predicted functions in 42% of 3461 un-annotated SG proteins. In 529 cases—including 280 non-enzymes and 21 for metal ion ligands—the expected accuracy is 84% at any GO depth and 94% down to GO depth 3, while for the remaining 931 the expected accuracies are 60% and 71%, respectively. Thus local structural comparisons of evolutionarily important residues can help decipher protein functions to known reliability levels and without prior assumption on functional mechanisms. ETA is available at <http://mammoth.bcm.tmc.edu/eta>.

Keywords

Function prediction; proteome annotation; structural motif; structural genomics; evolution

© 2009 Elsevier Ltd. All rights reserved.

Correspondence to: Olivier Lichtarge, lichtarge@bcm.tmc.edu.

[†]These authors contributed equally to this work

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Introduction

Most proteins lack known function¹. This is also the case for proteins with known structure, since, as of November 2008, 1784 of the 3433 structures solved by the Protein Structure Initiative (PSI)^{2; 3; 4; 5}, or over half, were labeled “hypothetical” or “unknown function” in the Protein Data Bank (PDB)⁶. One cause for this lack of knowledge is that detailed functional characterization of individual protein function are lacking, given their demands on time and resources. One possibility is to use high-throughput experimental strategies, but large-scale functional screens are only now being developed⁷. Moreover, protein function can be context dependent^{8; 9}, or non specific, and a battery of positive *in vitro* assays may overestimate the actual *in vivo* function. For now, fewer than five percent of annotations are experimental^{10; 11}; our own analysis of the Uniprot database¹² suggests that number may be as low as 2.4%. A second reason is that the most widely used annotation method, which transfers annotation among sequence homologs identified by BLAST¹³ or PSI-BLAST¹⁴, is increasingly error prone as evolutionary distances grow¹⁵, or as variations impinge nearer to functional sites¹⁶. Thus, homology-based function transfer of the four Enzyme Commission (EC) digits that describe enzymatic reactions¹⁷ are only 90% accurate down to 70% sequence identity^{10; 18; 19; 20}. For three EC digits annotation transfers, the standard that we use to define the correctness of enzyme function predictions, they are reliable only down to 50%–60% sequence identity^{18; 19; 20}. Yet many annotations are based on 30% sequence identity homologs²¹. This has raised repeated concerns that annotation errors occur and then propagate^{22; 23; 24; 25; 26; 27}, and justifies the search for new approaches^{2; 28}.

Among these, 3D templates are of special interest because they probe directly the molecular basis of function through small structural motifs of just a few key amino acid that, ideally, identify functional determinants and their functionally relevant matches in other proteins^{29; 30; 31; 32; 33; 34; 35; 36; 37; 38; 39; 40; 41; 42; 43; 44; 45; 46; 47}. Preliminary studies hint that 3D templates may be accurate when sequence identity is below 40%⁴⁸, and thus prove complementary to sequence comparison methods. Two limitations decrease the potential impact of 3D templates, however. First, precise and experimentally validated groups of functional residues, such as the well-known catalytic triad⁴⁹, that together define a structure-function motif are few⁵⁰. Second, when templates are built from residues chosen through heuristics, simple geometric thresholds of root mean square deviation (RMSD) between a template and its matches can be non-specific, failing to reliably discriminate between relevant and random matches^{51; 52}.

The Evolutionary Trace Annotation (ETA) pipeline was developed to address these issues. It exploits the rankings of evolutionary importance produced by the Evolutionary Trace (ET)⁵³. These rankings are obtained when ET correlates residue variation patterns observed during evolution with sequence divergences. Both retrospective computational studies and numerous experimental validation studies show that top-ranked ET residues on protein surfaces have desirable properties: they cluster structurally; these clusters overlap and predict functional sites^{54; 55}; and mutations targeted to top-ranked residues reliably block and separate functions^{56; 57; 58}, or transfer function among homologs upon exchange of cognate residues⁵⁹. These data buttress the basic hypothesis that surface clusters of top-ranked residues may indicate essential determinants of function and specificity. Since this is precisely the type of information that 3D templates seek to embody, ETA analyzes these ET clusters to create 3D templates. Importantly, in so doing, it bypasses the need for any prior knowledge of functional mechanisms. By contrast, the ProFunc metasever’s Enzyme Active Site method⁶⁰ for example, relies on experimental knowledge of functional sites taken from the Catalytic Site Atlas⁵⁰—although another of its component methods does use *de novo* templates, but was improved upon by ETA⁴⁵. A notable strength of the ETA approach is that it stringently filters the matches between a 3D template, *s*, extracted from protein *S*, and a protein structure

T, so as to enhance specificity. First, the matched site in T must be evolutionarily important in its own right⁵¹; second, a template *t* extracted from T must reciprocally match the structure of protein S61; and finally, T's function must achieve a plurality over all matches⁶². ETA thus reached 92% 3 digit EC positive predictive value (PPV) in large, retrospective controls of enzyme function prediction using the Enzyme Classification (EC) system. These encouraging results fell short, however, since (a) most proteins are not enzymes; and (b) the EC scheme does not carry over to non-enzymes. Thus, for ETA to be tested realistically it must be (a) controlled over enzymes and non-enzymes alike; and (b) adapted to a universal function classification scheme.

To these ends, this work generalized ETA to the GO classification, and then evaluated its annotation performance in enzymes and in non-enzymes—first on small control sets, and then on all previously annotated proteins solved by structural genomics projects. Finally, the possibility of predicting ligand type was assessed for the basic case of proteins that bind metal ions.

Gene Ontology (GO) was a natural choice because it provides a controlled vocabulary that hierarchically describes the biological processes, cellular components, and molecular functions that any protein takes part in⁶³. Mapping a protein to its correct GO term, or terms, is not entirely straightforward, however. Often multiple overlapping GO annotations are relevant—either due to multiple functions, or because GO typically splits functions into several component terms, each at different levels of specificity. ETA was thus extended to predict multiple, hierarchical functions. To control that ETA annotations reflected in part the choice of templates residues, templates built from evolutionarily unimportant residues were also tried, and performed worse. Surprisingly, and for a number of possible reasons, so did templates that took Catalytic Site Atlas information as the primary basis for templates. The control studies focused on the positive predictive value down to depth three of the GO hierarchy (depth 3 PPV), which for enzymes corresponds to the second of the four hierarchical Enzyme Classification numbers. But PPV for any GO depth is reported as well as the fraction of predictions that are both accurate at all GO depths and entirely complete. The results are that when specificity is maximized (depth 3 PPV 94%), the prediction availability, or coverage, was moderate, at 53%. But this coverage could be increased to 84% if one was willing to decrease accuracy (86%, depth 3 PPV). Finally, ETA was then applied to structural genomics proteins without known annotations. This produced 529 predictions at an expected GO depth 3 PPV of 94%, including 280 predicted non-enzymes and 21 ion-binding proteins, and 931 additional predictions with a lower depth 3 PPV of 71%, yielding a total of 1460 new annotations.

Results

Overview of ETA

The ETA pipeline proceeds in a series of steps best illustrated in an example. To annotate *Mycobacterium tuberculosis* v1626 (PDB 1sd5, chain A), shown in Figure 1, ET ranked its residues by evolutionary importance. The first cluster of 10 or more top-ranked residues on the protein's surface appeared at the sixth percentile rank. From these ETA picked a template: the C_α atom coordinates for the six top-ranked amino acid positions, each one labeled by its allowed side chain type(s), given the variations seen frequently in homologs (in this case, there are 14 additional amino acid combinations). A hierarchical search algorithm for C_α atoms with near identical distances between pairs, the PDM algorithm, identified a match. This was to a structure with 22% sequence identity, namely the *Tolypothrix* species PCC 7601 phytochrome response regulator rcpb (PDB ID 1k66, chain A). Specifically, six amino acids in 1k66 had side chains and C_α inter-atom distances that matched those of their template counterparts to within ± 2.5 Å. The match was then evaluated by a support vector machine (SVM) trained to

identify functionally relevant matches on the basis of geometric similarity to the template (0.79 Å RMSD), and of evolutionary importance similarity between the template and the matched residues (the average difference was less than five percentile ranks). The same search repeated one-to-many (OTM) over the entire PDB 90 identified another eight other matches to various structures that were all also accepted by the SVM.

Conversely, ETA applied to 1k66 created a template that matched back to 1sd5. This match was of high quality both geometrically and evolutionarily (1.05 Å RMSD, ± 2 percentile ranks), and therefore also accepted by the SVM. The two templates were not identical, however, (they shared only three of six residues) so that the two reciprocal matches between 1sd5 and 1k66 were essentially independent. This enhanced the likelihood that they were functionally relevant rather than chance events. Similarly, two of the other eight matches searched many-to-one (MTO) against 1sd5 reciprocally matched it (1i3c and 1jbe; 23% and 32% sequence identity, respectively). Finally, all three reciprocally matching proteins were annotated with GO: 0000156: two-component response regulator activity. This consistent annotation, which arose independently from distinct sources, was then predicted by ETA to be 1sd5's function—which was correct.

A complicating factor is that GO terms often lie in parallel or at different hierarchical levels, which required a more complex hierarchical voting scheme than previously devised for the simple branching hierarchy of the EC annotation of enzymes⁶¹. The new scheme, in Figure 2, starts at the top of the molecular function hierarchy and chooses the most frequently represented plurality function as the most likely. If there is a tie, the tied functions are all considered correct, since matches support them equally, and the search splits and continues in parallel, as shown in Figure 2a. For each winning term, child functions are likewise examined for plurality, and so forth, until there are either no votes or the bottom of the hierarchy is reached. Ultimately, this produces one or more annotations, each one proceeding from general to specific terms. In some cases, such as Figure 2b, the final term predicted may have fewer votes than its parent terms, because some matches are not annotated all the way to a leaf node. However, since there are still matches supporting this term, it is given as the prediction in order to provide the most specific annotation supported by matches. Finally, Figure 2c shows a case where terms that were eliminated earlier in voting can still be predicted when they represent an alternate path to a predicted term, since GO specifies that annotation by a term implies annotation by its parent terms as well.

Assessing annotation performance for GO term predictions is difficult because of the multi-functional and hierarchical nature of GO. Pal et al.⁶⁴ assessed performance at each depth in the GO hierarchy, and considered a prediction correct if it identified at least one known GO term, and false if no prediction overlapped the known annotation. Positive predictive value is then the number of correct predictions divided by the total number of predictions.

In addition to the depth-dependent measurements, we wish to describe overall performance, independent of GO depth, and to indicate when predictions are useful but less-than-ideal. We therefore define four categories of predictions. *Correct and complete* predictions identify all currently known GO terms down to the deepest levels of the hierarchy, including, if applicable, multiple GO terms, and also allowing additional predicted functions that may point to incomplete reference annotations. *Incomplete* predictions identify some, but not all GO terms at their lowest depths. These two categories correspond to correct predictions according to the definition in Pal et al.⁶⁴. *Partially correct* predictions agree down to GO depth three, but they diverge starting at depth four or below. We select a depth of three for this threshold because of its consistent meaning for enzymes (2 EC digits) and because it represents a compromise between the deep enzyme hierarchies and shallow non-enzyme hierarchies. This standard is more stringent than that used to evaluate some methods⁶⁵; 66. Partially correct predictions also

include predictions that are otherwise incomplete but also suggest additional functions that cannot be verified. By the definition of Pal et al.⁶⁴ these extra predictions will be correct at some depths and incorrect at others. Finally, *incorrect* predictions disagree with known annotations at depths one, two or three, and they would be defined as incorrect according to Pal et al.⁶⁴ at and below those depths. It is possible that some of these incorrect predictions may actually be true and simply lack support, but here they are assumed for simplicity's sake to be false. Accordingly, to calculate depth 3 PPV we consider the “correct and complete”, “incomplete”, and “partially correct” predictions as positive and the “incorrect” predictions as negative. Depth 3 PPV gives a general description of performance, regardless of GO depth, supplemented as needed by all-depth PPV, and by the fraction of predictions that is fully correct and complete. In practice, as we shall see below, ETA annotations are largely correct at any GO depth.

Gene Ontology Predictions for Enzymes

This GO voting scheme was tested first on all 1889 annotated SG enzymes (395 with multiple GO terms) to verify that its performance was comparable to ETA's EC predictions. ETA predictions covered 1031 proteins (or 55%, Figure 3a,) such that 763 of the GO annotations (74%) were both fully correct and complete. In 213 of these cases, ETA predicted additional or more specific functions. For example, for Shikimate5-Dehydrogenase from *Thermus thermophilus* (PDB 2d5c, chain A), in addition to the known shikimate5-dehydrogenase activity (GO:0004764), ETA suggested NADP binding (GO:0050661) on the strength of two reciprocal matches—PDB 1p77, chain A, from *Haemophilus influenzae*, and PDB 1nyt, chain A, from *Escherichia coli*. Although the 2d5c structure did not have NADP bound, two alternate structures of the same protein (2ev9 and 2cy0) were complexed with NADP, suggesting that ETA identified a missing annotation. In another case, beyond the prediction of hexokinase activity (GO:0004396) for human hexokinase ii (PDB 2nzt, chain A), ETA also suggested ATP binding (GO:0005524) based on a match to human glucokinase (PDB 1v4s, chain A). This was again an expected function since hexokinases phosphorylate hexose by reducing ATP to ADP. These examples show that current GO annotations may be incomplete even for well-characterized proteins, and that ETA can recover such missing annotations.

Another 80 (of the 1031) cases had fully correct but incomplete annotations, yielding an all-depth PPV of 82%. Finally, 124 cases were partially correct: 20 were fully correct at all depths but incomplete and included additional predictions that could not be verified, 84 were correct until the lowest GO depth, and the remaining 20 were accurate until the second lowest GO depth. For example, diaminopimelate decarboxylase from *Mycobacterium tuberculosis* (PDB 2o0t, chain A) is annotated with EC 4.1.1.20, which maps to GO:0008836, whereas ETA predicted ornithine decarboxylase activity (GO:0004586 or EC 4.1.1.17) for this protein. As is typical of difference in the fourth EC digit, the error is in the prediction of the substrate, not the mechanism. Both GO terms share a parent, GO:0016831 (carboxy-lyase activity), equivalent to the correct three EC digit annotation of 4.1.1, our standard for EC annotation in prior studies⁶¹. Predictions were incorrect (at GO depth one, two, or three) for only 64 proteins. Thus, ETA predictions of enzyme GO annotations achieved 94% GO depth 3 PPV (967/1031), comparable to the previous benchmark of 92% 3 EC digit PPV⁶¹.

This high success rate raises the possibility that some of these annotations arose from matches to structures of homologs with high sequence identity. These would be trivial, since sequence searches would identify similar annotations without difficulty. As a control, we therefore monitored ETA's performance on all test proteins when target structures were systematically eliminated from consideration as a function of sequence identity (Figure 3a). Although prediction coverage decreased marginally with sequence identity, it remained at about 45% down to the 50% cutoff. By contrast depth 3 PPV remained high, staying above 90% until the

30% cutoff, at which point it fell slightly to 87% and then to 84% at the 20% cutoff. Likewise, neither the all-depth PPV nor the fraction of fully complete and accurate annotations depended on trivial similarities: both are robust at least down to a 40% sequence identity cutoff.

One caveat is that this performance may reflect an overly favorable choice of GO depth, three, as the standard for a positive prediction. To obtain a more thorough picture of how ETA's performance varies with the specificity of its predictions, we therefore evaluated predictions at every GO depth, similar to Pal et al.⁶⁴ (Figure 4a). Keeping in mind that a protein may have multiple annotations at the same GO depth, ETA predictions for a single protein and for each depth may be: correct and complete, correct but incomplete, partially correct (correct but incomplete predictions, but also extra functions), or completely incorrect. The first case defines a "lower bound PPV", the first two define an "incomplete PPV", and the first three define a "best case PPV". The latter was nearly uniformly high: over 90% at all depths except for depth five, where it fell to 88%. The incomplete PPV is essentially identical, and the lower bound PPV is slightly lower initially, but becomes identical past depth 6. This suggests that even highly specific (deep) ETA predictions are likely true. Coverage was above 40% at all depths except eight. Thus ETA's PPV does not seem to vary consistently with ontology depth, and the depth 3 PPV of 94% is a reasonably accurate snapshot of the method's performance over most depth depths.

Another possible concern is that the sensitivity, or coverage, is on the order of 50%. This is a methodological bias taken in order to gather fewer but more reliable predictions—an especially desirable feature at lower sequence identity, when sequence methods become more error-prone. Nevertheless, if sensitivity were paramount, we note that simply including in the tally non-reciprocal ETA matches—a high sensitivity mode we call *all-match ETA*—can increase it. Under such less stringent constraints, an additional 634 predictions were made. Of these, respectively, 345, 32, 83, and 174 proved to be either correct and complete, incomplete, partially correct, or false. Altogether, reciprocal ETA and all-match ETA thus yielded 88% overall coverage such that two thirds (67%) of predictions were both fully correct and complete; nearly three quarters (73%) were correct, if incomplete, and 86% were accurate down to GO depth 3. The true cost was that fully 238 (14%) cases were now incorrect, compared to only 64 (6%) with standard, reciprocal ETA annotations; i.e. the penalty for increasing coverage to nearly 9 proteins in every 10 is a three-fold increase in possibly false annotations.

Finally, to put ETA's performance in perspective, it was compared to two competing methods. First, it was compared to JAJA, a sequence-based metaserver⁶⁷, for 50 of the enzymes (Figure 5). JAJA made 40 predictions (80% availability) as opposed to ETA's 50 (100% availability). The depth 3 PPV was 78% (vs ETA's 100%), the all-depth PPV was 68% vs ETA's 88%, and only 20 of these were complete (50% vs ETA's 82%). JAJA failed to annotate udp-n-acetylmuramate-alanine ligase murc (tm0231) from *Thermotoga maritima* (PDB 1j6u; chain A), while ETA correctly predicts ATP binding (GO:0005524) and UDP-N-acetylmuramate-L-alanine ligase activity (GO:0008763) based on a reciprocal match with udp-n-acaetylmuramic acid:l-alanine ligase (murc) from *Haemophilus influenzae* (PDB 1p3d, chain B) with 28% sequence identity. Time demands on the JAJA server precluded larger scale comparisons.

Second, to compare ETA's performance to another template-based method, the ProFunc metaserver's Reverse Templates (RT)⁴⁵ method was assessed for 120 of the enzymes (again, time and computational considerations limited this number) (Figure 5). For these proteins, RT had 54 correct and complete annotations (ETA had 79), 4 were incomplete (ETA had 13), 16 partially correct (ETA had 10), and 17 incorrect (ETA had 1). Overall, RT had 81% depth 3 PPV and 76% coverage (ETA had 99% depth 3 PPV and 86% coverage for the same proteins). Thus ETA improved on the RT method in both metrics.

Gene Ontology Predictions for Non-enzymes

Having ascertained in enzymes that ETA's GO annotation scheme was reliable, proof of concept was sought next in non-enzymes by predicting functions in 50 non-SG, including 14 proteins that had multiple GO annotations. In this small test set, ETA yielded 36 annotations (72% prediction coverage), of which 32 (89%) were correct and complete (Figure 3b). In one of these instances, the ETA annotation was more specific than the available annotation. We treat this as two separate predictions. First, ETA identified the annotated function fully, and we consider this portion of the prediction correct. Second, it suggested a more specific annotation than is currently available, which we report neutrally since it cannot be verified or refuted, but mention it to note that many current GO annotations are incomplete, and ETA can provide evidence for extending them. A putative transcriptional repressor from *Salmonella typhimurium* lt2 (PDB 1t33, chain A) was annotated with transcription factor activity (GO: 0003700) and regulator activity (GO:0030528). However, beyond the GO:0003700 prediction, ETA more specifically suggested transcriptional repressor activity (GO:0016566) based on a match with a transcription regulator from *Bacillus subtilis* (PDB 1vi0, chain A) with 20% sequence identity.

Two predictions were incomplete. For example, the outer membrane transporter FecA from *Escherichia coli* (PDB 1kmp, chain A) was annotated with receptor activity (GO:0004872), iron ion binding (GO:0005506), and siderophore-iron transmembrane activity (GO:0015343). ETA predicted receptor activity and siderophore-iron transmembrane activity but not iron binding (although this was strongly suggested by ETA's other predictions). Hence, ETA identified two thirds of the annotated functions (67%). In the other case, ETA covered 75% of annotated functions. Two predictions were partially correct, identifying one or more functions down to the second-most-specific depth of the hierarchy, but did not identify the other known functions, and suggested extra functions that could not be verified. No predictions were incorrect, giving 100% depth 3 PPV. As before, sensitivity could be increased with all-match ETA, which annotated six more proteins: four correctly and completely, but the other two incorrectly (Figure 3b). Thus overall prediction coverage could be raised from 72% to 84% (42/50), with the depth 3 PPV dropping from 100% to 95% (40/42).

ETA was then tested on a larger scale on all 311 annotated SG non-enzymes (Figure 3c). It made 132 predictions (42% coverage). Strikingly, none were false down to GO depth 3 (100% depth 3 PPV), and only one was false at any GO depth (99% all-depth PPV). Finally, 116 (88%) were fully correct and complete. Moreover, ETA further proposed additional functions in six cases. Since these were consistent with the annotated function, but could not be verified, they were considered neutral, plausible predictions that suggest that ETA can extend current annotations. In one of these six cases, the first c2 domain of synaptotagmin iv from human fetal brain (PDB 1ugk, chain A) had reciprocal matches to the calcium binding domain c2b from *Rattus norvegicus* (PDB 1uow, chain A) and the c2 domain from protein kinase C (alpha) (PDB 1dsy, chain A), thus recovering the reference annotation of transporter activity (GO: 0005215). However, the protein kinase C match also added the novel annotation of phosphotransferase activity with an alcohol group as an acceptor (GO:0016773). In another case where ETA extends existing annotations, a Zn binding protein (PDB 1uqw, chain A) had transporter activity (a term at depth one) as its reference annotation. A reciprocal match with the periplasmic nickel transporter from *Escherichia coli* (PDB 1zlq, chain B) predicted nickel transporting ATPase activity (GO:0015413, a term at depth seven). Although this zinc-binding protein is not likely to bind nickel, transporter activity is a parent term and thus this is annotation is considered correct and complete with respect to the reference annotation. Further, its immediate parent, cation-transporting ATPase activity (GO:0019829, depth six), represents a novel extension of the reference annotation. Predictions for 15 proteins were incomplete but correct and spanned one to two thirds of the annotated functions. One prediction was partially

correct, identifying one function correctly down to the second-most-specific term, but not identifying some functions, and suggesting some functions that could not be verified. Again, all-match ETA could extend coverage—this time to from 42% to 69% (214/311) with a decrease from 100% to 90% depth 3 PPV—by predicting functions for 82 additional cases. Fifty-six were correct and complete, five were incomplete and 21 were incorrect.

As for enzymes, in order to ascertain that the PPV was not due simply to trivial matches, structural matches to proteins were progressively removed from consideration as a function of their sequence identity. In the small test set depth 3 PPV, all-depth PPV and the fraction of fully correct and complete predictions remained nearly steady regardless of sequence identity stringency (Figure 3b). Critically, these observations held true for larger-scale predictions: except for a dip at the 30% threshold that corrected at 20%, the depth 3 PPV remained nearly 100% at all thresholds. The all-depth PPV was steady around 97%, and the fully correct and complete fraction was about 80%. Similarly, prediction coverage was down only 10% (to 32%) at the 50% cutoff (Figure 3c). Although coverage can be increased with non-reciprocal matches, it is clear that, as-is, ETA's very high PPV at all sequence identity thresholds on a large scale makes it a useful tool to annotate non-enzymatic functions reliably.

Strikingly, the level by level analysis showed all of the ETA predictions overlapped the reference annotations, yielding 100% best-case and incomplete PPVs at all GO depths (Figure 4b). Completeness varied however, with a low of 71% at depth five, but was otherwise better than 94%. Moreover, since ETA lacked predictions mostly for proteins that were mostly annotated to shallower GO depths, the coverage actually rose at deeper ontology depths. Thus performance for non-enzymes seems to be high at all ontology depths in this exhaustive test set of previously annotated non-enzymes with structures solved by SG.

In order to gauge the significance of these results, we compared Jafa vs ETA among the 311 SG proteins (Figure 5). The PPVs suggest a specificity advantage for ETA: 100% vs 98% and 99%, vs 97% for the depth 3 PPV and all-depth PPV, respectively, although the Jafa coverage was better (93%). Thus when fewer sequence homologs are available ETA, which uses structural details, yields more specific annotations, and so may be complementary to sequence-based methods⁶¹.

ETA's non-enzyme performance was also compared to Reverse Templates in 224 of the 311 proteins (a limitation due to server time constraints). While ETA had much smaller coverage, it was more specific (Figure 5). Of note, ETA correctly and completely annotated eight of the 37 incorrect RT predictions, and did not make predictions for the other 29. Also, ETA correctly and completely annotated two of the six partially correct RT predictions and made no prediction for the other four. Finally, ETA made correct and complete predictions for two of the three proteins not annotated by RT. This suggests that for non-enzymes, ETA is complementary to reverse templates. RT's much higher prediction coverage is consistent with its use of multiple templates per protein, and is also consistent with lower PPV. Taken together with the enzyme results, this suggests that ETA is generally better suited to enzymatic function prediction than RT, and that it is also better for non-enzymes when specificity is paramount, but that RT is more sensitive than ETA for the annotation of non-enzymes.

Metal-ion Binding Predictions

In order to assess the prediction of binding to simple ligands, ETA was next applied to 184 SG proteins with known metal binding specificity and diverse functions, including catalysis⁶⁸, transport, and signaling⁶⁹. ETA extracted templates from these proteins and searched for matches in all PDB structures that had ligands. It identified 89 matches that suggested specific and functionally relevant ligands (48% prediction coverage). Of these, 84 cases predicted metal ion binding (94% depth 3 PPV, since the GO term for metal binding (GO:0043167) is at depth

three). Seventy-three had exactly the correct type of metal ion (82% correct and complete), while 11 had the correct electronic charge but the wrong chemical element. Just five cases erroneously suggested that a small molecule ligand, such as nicotinamide-adenine-dinucleotide or mupirocin. Given that proteins may bind several types of ions depending on the local environment⁷⁰, the identification of a specific metal ligand is neither trivial⁸, nor always complete. As before, the prediction accuracy is largely independent of ET sequence identity cutoffs (Figure 3d). All-match ETA adds 39 new predictions and raises coverage from 48% to 70%, but metal binding accuracy (depth 3 PPV) falls from 94% to 80%, and the perfect identification of the metal (all-depth PPV) falls from 82% to 70%.

A confounding factor is that many of these ion-binding proteins were enzymes, so that ion prediction performance could just be a corollary to the quality of enzymatic predictions. But that was not the case: only 29 predictions were enzymes. The other proteins mediated signaling (8 Ca²⁺), response regulation (3 Mg²⁺ and 1 Mn²⁺), gene regulation (20 Zn²⁺), protein-protein interaction (20 Zn²⁺) and metal transport (1 Zn²⁺) (Table 1). Eight proteins did not have a known biological function. Thus the ETA templates could broadly capture essential metal ion binding characteristics and often predict the correct ion.

ETA also matched or improved on Jafa for metal ion-binding predictions (Figure 5), with better depth 3 PPV and a higher fraction correct and complete, but slightly lower coverage.

Structural Genomics Predictions

Taken together, these controls consistently suggested that ETA annotations were specific for enzymes and non-enzymes alike—reaching between 90–100% depth 3 PPV and between 74–88% fully correct and complete predictions. Moreover, such performance was neither dependant on trivial matches among structures with high sequence similarity, nor the annotation of non-specific GO terms. Logically, it follows that ETA predictions should be generally and equally reliable to identify the function of heretofore un-annotated structural genomics proteins. We therefore applied it to 2652 un-annotated SG proteins. In its high specificity mode (94% PPV and 76% correct and complete for the controls), reciprocal ETA matches predicted 268 enzymatic and 414 non-enzymatic functions in 514 distinct proteins (19% coverage, see Table 2). Of these, 167 proteins had two or more functions predicted, and 280 had entirely non-enzymatic functions predicted. In its high sensitivity mode (86% PPV and 68% correct and complete for the controls), all-match ETA added predictions for another 888 proteins, including 257 predicted non-enzymes to raise coverage to 53%.

These predictions are non-obvious, as seen from the percent sequence identity between each un-annotated SG protein and its most closely related match (Figure 6). Most of the annotations are based on matches with structures that have only 21% to 50% sequence identity. Some even occur in the 11%–20% range, where traditional homology-based methods are least accurate. Conversely, very few predictions fall in the 60%–90% range, suggesting that most of these SG proteins have already been annotated. Surprisingly, occasional predictions were in the 91%–100% region, representing trivial annotations that eluded the PDB and GO databases. Thus, most ETA predictions involve proteins with remote homology and appear truly novel. A notable failure is that there were no annotations from matches to proteins with different folds. Apparently, successful predictions require divergently related, pre-annotated proteins. Indeed, catalytic triad⁴⁹ aside, we know of no examples of 3D templates capable of annotating proteins with different folds. This in part explains why coverage is not higher.

We examine two predictions in detail. For the human bromodomain containing protein (PDB 2nxb, chain A), ProFunc⁷¹ predicts transferase activity and Jafa returned no hits. ETA finds a reciprocal match to human histone acetyltransferase bromodomain (PDB 1n72, chain A),

with 35% sequence identity, and predicts histone acetyltransferase activity (GO:0004402), which is more detailed than the ProFunc prediction.

Another example is a protein structure from locus ef_2437 in *Enterococcus Faecalis* (PDB 2p0o, chain A). ETA predicts peptidyl-prolyl cis-trans isomerase activity (GO:0003755) based on a single reciprocal match to a *Bacillus Cereus* protein with 26% sequence identity. Jafa did not return any hits, while ProFunc predicted only binding, with a low score.

Likewise, since ETA can identify proteins that bind metal ions when those ions are not already present in the structure, we applied ETA to 2249 SG proteins whose structures do not contain any bound ions or small molecules. In the end, ETA predicted metal-ion binding in 21 cases (see Table 3). All-match ETA adds 43 predictions.

Discussion

This study extends automated function prediction to any type of protein structure regardless of whether its function is enzymatic or not. ETA first identifies structural motifs of key functional residues, then it tallies local similarities of these residues among all other already annotated structures, and finally it transfers GO annotation between matches, using independent arguments to filter those that best reflect functional rather than random similarities. Prior studies suggested that ETA was highly specific but limited to enzymes and their specialized EC classification system¹⁷. Here, ETA was extended to the much more general GO classification. When used in a high specificity mode ETA can be expected to achieve 94% GO depth 3 PPV, 84% accuracy at any GO depth, and to produce perfect annotation both fully correct and complete in 76% of cases. ETA can also be used in a high sensitivity mode; if so, the GO depth 3 PPV decreases to 86%, but coverage jumps from about 53% to nearly 84%.

One concern is that these performance benchmarks, which were established in already annotated proteins, may not reflect future ETA performance in the SG proteins that lack annotation, which most likely have fewer homologs. Indeed, coverage was much higher among the controls (53%) than in the un-annotated SG proteins (23%). This suggests compositional differences between the two sets, and is consistent with the reduced ETA coverage seen below 30% sequence identity, where it is about 26% for enzymes and 14% for non-enzymes. A key point, however, is that for this threshold, ETA's PPV for GO depth 3 is 87% (423/489) for enzymes and 98% (43/44) for non-enzymes. This suggests an overall 87% GO depth 3 PPV (466/533) below the 30% sequence identity threshold, and 63% PPV over all depths. Thus performance in the yet un-annotated proteins may be less than in the controls but not dramatically so.

A significant additional concern is that most annotations used for both control proteins and target proteins have the GO evidence code "IEA", meaning that they were inferred from electronic annotations. This raises the possibility that our own annotation used already adulterated functional input, and regardless of the methodological accuracy simply furthers the propagation of errors. Conversely, since ETA's "incorrect" predictions usually only differ with established annotations at the second or third lowest depth of the hierarchy, some of these cases may represent annotation errors in the databases instead of errors on ETA's part. Ideally, ETA would be benchmarked using only proteins annotated with experimental data, and this is an important future direction. Currently, however, a search of the Uniprot12 database indicates that, only 324 have GO Molecular Function annotations with direct experimental evidence (i.e. a GO evidence code of EXP, IDA, IPI, IMP, IGI, or IEP).. So experimentally annotated structures may still be too rare to provide both control and target data sets. ETA was previously benchmarked⁶¹ against 13 proteins that were newly annotated to three EC digits by a high throughput experimental annotation pipeline¹¹ and achieved 86% 3 EC digit PPV (6/7

proteins). However, these annotations still relied on a target set consisting primarily of electronically annotated proteins.

The question remains as to why there are not more ETA hits at very low sequence identity, below 30%, 20%, or even lower in the twilight zone?

It is important to first note that although most of the control ETA annotations involve structures with significant sequence identity to another annotated structure (above 30%), ETA also accurately annotates structures with poorer sequence identity, below 30%. Specifically, there are 489 enzymes annotated, 44 non-enzymes annotated, and 51 ion-binding proteins annotated below 30% sequence identity among controls. Second, the data show that they are nearly as accurate as their counterparts with more easily recognizable homology: ETA's depth 3 PPV declines only from 94% to 88%; likewise at the 20% threshold the decline is to 85%. Admittedly, all-depth PPV declines a bit more, but much less so for non-enzymes. In any event, at GO depth 3, ETA continues to give high specificity results. Third, a plurality of ETA's predictions for un-annotated proteins (150) fall in the 20–30% range.

So, the one aspect of ETA performance that suffers considerably at lower sequence identity thresholds is coverage. But, even this is subject to debate. The reference annotation for these control structures must have necessarily arisen from low sequence identity homologs, which are sensitive but not so specific. Thus it may very well be that many of the reference (gold standard) annotations in that low sequence identity region are not accurate. In such cases, ETA, which has greater specificity, will accurately not "cover" these proteins. Here, we report such disagreements in our disfavor, as lower coverage, but this may be better thought of as a lower bound on coverage.

Finally, at lower sequence identity functions diverge significantly, or at least the biological context certainly does. Hence a protein may well change EC number or GO function, or maintain them through important changes to the 3D template. A reasonable future direction would be to test whether templates that are smaller, or that tolerate more residue variability (such as to allow for substitutions seen in remote sequence alignments) would extend coverage. More definitively, actual experimental assays to test the veracity of both the predictions and the presumed gold standard controls would help resolve these issues.

To what extent is ETA's diminishing number of annotations as sequence identity falls a limit on the method for structural genomics proteins? Perhaps not so much since after all, as 1259 of the previously un-annotated proteins have annotated homologs in the PDB90 with a BLAST e-value of 0.05 or better. Hence, this explains why ETA was able to suggest novel annotations for 514 proteins in high specificity mode, and an additional 888 in its high sensitivity mode. Thus while some SG proteins may be beyond ETA's scope, many are not.

Molecular Analysis of Templates

The results on control proteins are consistent with the fundamental ETA hypothesis: that proper ranking of evolutionary importance, combined with a knowledge of the protein structure, leads to 3D templates of key amino acid that capture enough determinants of function to define useful *structure-function motifs*: hence a different protein with the same residues will be likely to perform the same function.

One way that this hypothesis may be further ascertained would be to verify whether 3D templates overlap known functional sites residues. In enzymes, this means the amino acids listed in the Catalytic Site Atlas (CSA)⁵⁰. In non-enzymes, this means, for lack of better functional data, the residues in structural proximity to the ligand of interest (5 Å). Among 846 SG enzymes the template directly overlapped at least one, and often more, catalytic residue(s)

in 67% of cases. If not, it was immediately adjacent (within 10 Å) in another 28%, as shown in Figure 7. Thus ETA's templates typically fell on or near the catalytic sites in most cases (95%), as expected.

We also examined several non-enzyme templates from the non-SG small control in detail to illustrate their functional relevance with well-known examples. For the human follistatin-like and EF-hand calcium-binding domain (PDB 1bmo, chain A), ETA predicts calcium ion binding based on a reciprocal match to the extracellular Ca^{2+} -binding module in BM-40 (PDB 1sra). The 1bmo structure shows that three template residues (E234, D261 and E268; Figure 8a) directly coordinate calcium. Conversely, the two E residues are also part of the 1sra reciprocal template that matches the query protein.

For the human oxy T state hemoglobin (PDB 1gzx, chain B), ETA correctly predicts heme- (GO:0020037), oxygen- (GO:0020037), and iron ion-binding (GO:0005506)—based on six matches. One of those is to the deoxy form of hemoglobin from *Dasyatis akajei* (PDB 1cg5, chain A). Both the templates for 1gzx and the matching 1cg5 include four heme-binding residues (F185, H235, L231 and F246; Figure 8b).

Finally, the annotated function for human growth hormone (PDB 1a22, chain A) is hormone activity (GO:0005179), and ETA predicts this based on a reciprocal match with human affinity-matured growth hormone (PDB 1huw). The templates for both contain two residues (P61 and K168; Figure 8c) that are within 5 Å of the receptor interface. Thus ETA templates appear to be functionally relevant in non-enzyme functions as diverse as ion-binding, small molecule binding, and protein-protein interaction.

More broadly, overlap is even better for non-enzymes and ion-binding sites than for enzymes. Binding sites were defined as residues that had non-hydrogen atoms within 5 Å of a non-hydrogen atom in the binding partner. Among 63 non-enzymes examined, 89% of the 3D templates overlapped binding sites by an average of nearly four residues. This is highly significant since the templates themselves have just six residues. Six of the seven remaining templates fell within 10 Å of the site. Thus overall, 99% of templates had residues at or near the functional site. Likewise, in 184 metal ion-binding proteins, 83% overlap the binding the metal ion-binding site, by an average of four residues. Of the remaining 30 proteins, 23 had templates within 10 Å of the binding site.

These comparisons confirm that ET-based templates are structurally closely related to the known determinants of functions. Intriguingly, however, they also show that ETA templates are not perfect reproductions of what is typically considered to be the “functional site”. Their overlap is rarely complete, and occasionally there is no overlap at all. This suggests some degree of delocalization of the functional site, in the sense that it may extend to regions and to residues that are not necessarily in direct contact with a ligand.

Effect of Residue Selection on Template Performance

Having shown that templates are generally both structurally and functionally relevant still does not itself establish that their specific selection is responsible for the quality of ETA predictions. Moreover, because as we just saw the templates rarely include all known functional residues, it is uncertain whether better overlap yields better templates. Previous experiments⁵¹ showed that a template constructed from highly-ranked residues could identify serine proteases almost as well as the catalytic triad⁷², but this was not reproduced on a larger scale. In order to properly establish the causal link between evolutionary ranking and function-associated templates, central to the ETA hypothesis, and to explore the significance of incomplete functional site overlap, two additional controls were therefore designed—one positive, one negative—and applied to both enzymes and non-enzymes.

The design of the controls is as follows. Fifty-one annotated SG enzymes from the CSA were randomly picked. Best case, positive control templates were built by including all available CSA residues, and then, since there are rarely six residues in a CSA active site, completing the template with the nearest, best-ranked ET amino acid positions. By contrast, worst-case, negative control templates were built from six poorly ranked (residues with ET coverage >50%) but clustered amino acids that were not within 15 Å of the CSA residues (see Methods for more detail). Similar templates for non-enzymes were constructed using binding sites instead of CSA residues. The expectation was that the standard 3D templates picked automatically by ETA would yield better annotations than the negative controls, since the latter contained unimportant amino acids, but that they would not reach the quality of the positive control templates, since these embody the best available catalytic or functional information directly from the CSA.

The striking surprise was that in enzymes and non-enzymes alike the ETA templates outperformed not only the negative but also the positive controls, as shown in Figures 9a and 9b, respectively. Depth 3 PPV for enzymes was 96%, 81% and 55% for one-to-many ETA, the positive and the negative controls, respectively, and ETA also had more fully and partially complete predictions. Even more obviously, the fractions of fully correct and complete predictions were 70%, 52%, and 30%, respectively. Thus templates that contain the known catalytic or functional site residues perform no better, and sometimes worse than templates with fewer catalytic residues but that contain key evolutionary determinants, at least in this limited test. One explanation is that a catalytic site has important features besides its core reaction mechanism, such as its specificity, affinity, and kinetics, which are necessarily regulated by more than just the few residues in the CSA. Presumably, such residues are evolutionarily important and present in ETA templates, so that even without the specific CSA amino acids, they carry sufficient information to identify functionally relevant matches. Thus, a useful 3D template need not contain all, or even any of the CSA residues, to be relevant, confirming our earlier findings about the catalytic triad.

Why, however, should CSA residues hinder annotation performance? One possibility is that by rigidly dictating a precise amino acid catalytic arrangement, some flexible substitutions of residues, such as in the enolase and other enzyme families, cannot be readily accommodated⁷³; if so missing these residues from an ETA template would then be an advantage. A related possibility is that CSA based 3D templates are not adapted to the ETA filters which demand matches with similar evolutionary importance. Finally, it may also be that strict definition of CSA residues based on experimental evidence does not always capture all key functional features. On the other hand, non-enzyme positive control templates may over-estimate functional residues as interfacial proximity across a protein-ligand interface is no guarantee of binding importance⁷⁴. This observation suggests that reliable functional information is not entirely equivalent to known catalytic residues or structural views from crystallographic analysis, and that the former is paramount for function prediction as much as for rational mutation design.

Intriguingly, when the negative control was reciprocally filtered, it produced fewer annotations, but none that were false. Thus, even though 18 one-to-many predictions were incorrect (45%), ETA still produced 100% depth 3 PPV for the enzyme negative control, and likewise for non-enzymes. In large part this is because the reciprocal many-to-one search used standard ETA templates rather than negative control templates, so that the reciprocal matches are high quality. Nevertheless, it also provides dramatic evidence for the power of reciprocal matching to filter out false positives.

Template Composition

Since ETA templates dependably convey functionally relevant information, they may also yield insights into the composition of functional sites. One may thus compare the amino acid

frequencies found in the Catalytic Site Atlas⁵⁰ to the frequency of residues in the templates from the 1889 SG enzymes and the 311 SG non-enzymes, as shown in Figure 10a. Consistent with past observations, these frequencies show that charged and polar residues occur more frequently in enzyme templates⁶⁰ than in non-enzyme templates. Interestingly, although glycine is a somewhat common CSA residue (5%), it is by far the most prevalent residue in ETA enzyme templates (22%), likely due to its structural importance. Since glycine is the smallest and most flexible residue, it might benefit the flexibility of the site that is linked to the function. Hydrophobic residues such as F, L, and P seemed to be the major elements of non-enzyme templates⁷⁵.

Second, a similar analysis of secondary structure elements found within templates may also be done, based on DSSP data⁷⁶, as in Figure 10b. This shows a notable difference between non-enzyme and enzyme templates in alpha helix frequency. A little more than 30% of non-enzyme template residues reside on alpha helix, while only about 22% of enzyme residues are alpha helical. Such differences in amino acid frequencies and secondary-structure content might be useful in identifying enzymatic and non-enzymatic functions. This follows the observation by Dobson and colleagues that an SVM could distinguish enzymes from non-enzymes based on such differences⁷⁷.

Conclusions

ETA transfers functional annotations among protein based on local structural and evolutionary similarities of their functional sites. It differs from related applications in two important ways. First, it relies on evolutionary analysis by ET to pick templates. This circumvents any need for prior information on protein function or mechanisms, instead letting evolution dictate where functionally relevant motifs are likely to lie in the structure. Second, it is biased to make specific predictions with low false positive rates. This is achieved through three stringent constraints: that matches fall on residues that are themselves important, as measured by ET; that multiple matches sufficiently corroborate each other so as to reach annotation plurality; and that matches be reciprocated between a queries and their targets. When combined, these specificity filters insure that annotations due to chance are much reduced. Third, sensitivity can be increased, if desirable, and at a known cost to the positive predictive value.

The key finding of this work is that automated ETA annotation of protein functions are reliable and can achieve on the order of 94% accuracy at GO depth 3, and 84% at any GO depth for many types of protein, whether enzymes or not. The GO molecular function terms enable ETA to blindly tackle any protein, and let a voting mechanism trickle down the hierarchical and multifunctional structure of GO. Strikingly, this generalized ETA makes only 6% false annotations among enzymes and none in non-enzymes. Moreover, it attains full and correct annotations in 75% and 88% of these cases, respectively, with the balance made up of correct but incomplete or partially correct predictions.

A limitation of the method at this level of specificity, compared to sequence-based methods such as JAFa, is lower sensitivity. However, it is important to emphasize that this is a design choice made to optimize PPV and minimize the possibility of ETA-generated functional annotation errors. For in-depth studies of a particular protein, however, possibly less accurate predictions may still be useful in guiding experimental studies. In such cases, non-reciprocal matches may be added back, and this dramatically increases prediction coverage. Over all of our controls, all-match ETA would, for example, increase coverage from 53% to 84%, while decreasing GO depth 3 PPV from 94% to 86%, or all-depth PPV from 84% to 75%.

Many controls highlight the contribution that ETA may provide to protein function annotation. While sequence-based methods, such as in the JAFa metaserver, may have greater sensitivity,

consistent with often plentiful sequence data compared to the limited body of structural data, evolutionary and structural comparisons together still can improve the annotation specificity. This was observed especially for enzymes and metal-ion binding proteins. Moreover, ETA not only predicts function, but it also offers insight about possible functional sites to experimentalists via 3D templates, useful not just for annotation but also for targeting mutations. The comparison of templates with known functional residues strengthens this point. Extensive controls suggest that the majority of the 500 new ETA annotations should be reliable and accurate, and should lead to efficient, systematic, and targeted assays for experimental verification. The fact that ETA templates perform as well or better than templates that include structurally defined catalytic residues further suggests that subtle differences between functional and structural insights have consequences for function prediction.

Other studies demonstrated that ET rankings of residue importance are sufficiently accurate to predict binding sites^{53; 55}, and then to efficiently guide mutations to top-ranked residues in order to rationally swap functions⁵⁹, constitutively activate or knock out functions, separate protein functions^{56; 57; 58}, and even recently construct mimetic peptides⁷⁸. Here, these same ET rankings prove sufficient to build structure-function motifs of a few key residues that embody biological information that is necessary and sufficient to identify protein function. The method has a low likelihood of producing false annotations—a major concern in light of the difficulty of performing functional assays on a proteomic scale and validating computational annotations. In the future it will be important to benchmark ETA's performance for the recognition of more complex ligands, to further improve its coverage of structure space, without compromising specificity, and to blend ETA with complementary annotation methods such as SIFTER²⁸, ProFunc⁷¹, JAFa⁶⁷, and Proknow⁶⁴. An ETA server for enzyme predictions has now been expanded to include GO annotation for enzymes and non-enzymes alike, and is available at <http://mammoth.bcm.tmc.edu>, Source code is available upon request.

Addendum

After the original submission, another template method called FLORA was released⁷⁹ and showed marked improvement in predicting 3 digit EC¹⁷ functions compared to Reverse Templates⁴⁵, CATHEDRAL⁸⁰ and CE⁸¹. We benchmarked ETA against FLORA on the exact same data set of 821 protein chains, using the same leave-one-out procedure to match each protein against the other 820. ETA predictions of 3 digit EC functions were made as described previously⁶¹, and the target set consisted of 6117 annotated proteins in the 2008 PDB SELECT 90⁸². Although this is not a stringent comparison on independent sets of proteins, the data in Supplementary Figure S1 suggests that ETA may be more specific in either the high sensitivity or high specificity mode.

Materials and Methods

Function Definition

Protein functions were defined for controls as the GO molecular function terms found on the Gene Ontology website, (<http://www.geneontology.org>) and the EC annotations from the PDB, mapped to GO⁶³. For metal ion-binding predictions, the type of ion bound in the structure was taken as the function.

Data Sets

A training set of 53 enzymes⁵¹ was used to train the SVM (see below) that selects functionally relevant matches, and to initially select parameter values for the template search (also below).

The “Non-enzyme Test Set” is composed of 50 protein non-enzyme structures with Gene Ontology annotations, chosen by hand from the PDB. Functions include nucleic acid binding,

metal ion binding, protein binding, lipid binding, oxygen transport, growth hormone activity, a structural role, and ion channel activity.

The “Structural Genomics Set” consists of proteins with the keywords “structural genomics” and “unknown function” in the PDB⁸³. For each protein, unique chains were considered as unique structures. There were 5463 structures in this set, 4852 of which also had ET results. These 4852 proteins were divided into enzymes, non-enzymes, and un-annotated proteins based on their EC numbers, GO molecular function terms, or names when these suggested annotations. Some proteins in the enzyme set also have non-enzymatic functions. There are 1889 proteins in the “Annotated Structural Genomics Enzymes” subset; 50 were chosen at random for comparison to JAFa, and an additional 70 (120 total) were chosen at random for comparison to ProFunc Reverse Templates. There are 311 proteins in the “Annotated Structural Genomics Non-enzymes” subset; 224 of these were chosen at random for comparison to the Reverse Template method. There remain 2652 proteins in the “Un-annotated Structural Genomics” subset.

The “Metal Ion-binding Set” was composed of proteins from the “Structural Genomics Set” that bind a single type of ion. This set consists 184 proteins binding Ca^{2+} , Cu^+ , Co^{2+} , Fe^{2+} , Mg^{2+} , Mn^{2+} , Ni^{2+} and Zn^{2+} . The “GO Target Set” was the subset of the 2006 PDB-SELECT-90⁸⁴ with ET results (9001 proteins) and GO or EC annotations (5971 proteins). The “Metal Ion Target Set” was the subset of PDB structures with ET results and a known ion or ligand bound to the structure (3730 structures, 470 with bound ions and 3260 with bound small molecules). For the sequence identity controls, target structures above a sequence identity threshold were removed from consideration, and predictions made without any matches to these proteins.

Template Creation

Templates were created as described elsewhere⁸⁵. Proteins were traced using automated⁸⁶, real-valued⁸⁷ ET⁵³ with default parameters (available at <http://mammoth.bcm.tmc.edu/traceview/index.html>) to determine their residues’ relative evolutionary importance. Residues were added in order of importance to form a structural cluster of at least 10 surface residues (solvent accessibility of at least $2 \text{ \AA}^2$ calculated by DSSP⁷⁶), and the six most important (numerically lowest ET rank) were chosen and represented geometrically by the coordinates of their C_α . To determine allowed residue types, the multiple sequence alignment used by ET was examined for unique combinations of residue types for the six positions comprising the template. Any combination that appeared more than once was allowable as a match.

Template Searching

Template searching is performed using Paired Distance Matching, described in detail elsewhere⁶¹. Briefly, the Paired Distance Matching algorithm builds matches residue-by-residue, adding one residue in each step and checking the geometry of this partial match against the template. The algorithm can be conceptualized as a breadth-first search of a tree, with each internal node representing a partial match, and each leaf node representing a full match. The depth of each node (its depth in the tree) corresponds to how many residues have been matched. The tree is only searched below nodes whose geometry match the template, which prunes the search tree to limit its breadth and thus decrease search time.

More specifically, starting with residue r_I in a template $\mathbf{R}=\{r_i\}$, PDM identifies all residues of type t_I in the target protein. For the first iteration, each of these is a possible match m_i to the template, and each is stored in the set $\mathbf{M}=\{m_i\}$.

For residue r_2 , all residues of type t_2 are identified. Each new residue is added combinatorically to each of the possible matches m_i in $\mathbf{M}$, expanding $\mathbf{M}$. Each m_i is then checked against distance constraints and retained or discarded. The distance between the new residue r_2 and the old residue r_1 is computed; in this case distance $d(r_1, r_2)$. For each m_i , the corresponding distances between the new residue r_2' and the residues in the current m_i are computed and compared; in this case the distance of the corresponding matched residues $d(r_1', r_2')$ is compared to $d(r_1, r_2)$. The match is removed if $|d(r_1, r_2) - d(r_1', r_2')| \geq \epsilon$; where ϵ represents a tolerance value (2.5 Å in this study); otherwise m_i remains in $\mathbf{M}$.

These steps are repeated for r_3 , with each residue of type t_3 in the target added to each m_i , distances $d(r_2, r_3)$, and $d(r_1, r_3)$ computed and compared to their counterparts in m_i , and each m_i with all distances within ϵ of the template distances retained in $\mathbf{M}$. This process continues for each remaining template residue r_i , halting when $\mathbf{M}$ becomes empty or all residues in the template have been examined. The result is a set of matches whose distances between residues match those of the original template plus or minus ϵ . If the distances match, the residues in m_i are likely in a similar geometry to those in R , so the residue numbers of each m_i are reported with their RMSD.

For example, consider a template from the structure of Invt, chain A, composed of T71, P73, K75, N96, D111, and Q259. The first depth of the tree is composed of nodes representing each T in the target protein. At the second depth, each node represents a pair of T's and P's in the target, with all possible pairs represented. For each node, the geometry of the residues is checked against that of the template by comparing the distance between the C_α coordinates of each residue pair to the analogous distance in the template. These distances must be within a certain ϵ (2.5 Å) of each other; that is, residues r_1 and r_2 match r_1' and r_2' only if $|d(r_1, r_2) - d(r_1', r_2')| \leq \epsilon$, where $d(a, b)$ is the distance between the C_α atoms of residues a and b.

Having computed the distance from T to P for each node and compared it to that from T71 to P73 in the template, PDM retains only those nodes that meet the distance constraints. Thus at the third depth of the hierarchy, all K's in the target are combinatorically added only to those depth two nodes. The K–T and K–P distance constraints are then checked, and the process continues at the fourth depth. The search ends when it reaches the leaf nodes (sixth depth), and reports those for which all pairs of residues fit the distance constraints as matches, or when all nodes on the current depth have been eliminated. RMSD is then calculated for all matches using a freely available C implementation⁸⁸.

To make the search more sensitive, it is repeated with different amino acid labels taken from homologous proteins. For example S71, P73, K75, N96, D111, and Q259 is a combination of labels observed 69 times in the ET multiple sequence alignment of InvtA and its homologs, compared to 397 occurrences of the set of labels seen natively in InvtA. Here, templates include all such combinations observed more than once in the alignment.

Match Filtering

Three filters removed likely false matches. First, matches with an RMSD greater than 2 Å were eliminated, as, based on prior observations⁶¹, true matches are very rare above 2 Å, and removing these obvious false matches improves the SVM.

Next, an SVM filters additional matches based on geometric and evolutionary similarity. The SVM feature vector is seven-dimensional, containing match RMSD (1 dimension)—which quantifies geometric similarity—and the sorted absolute values of the difference between the percentile ET ranks (coverage) of each pair of matched residues (6 dimensions)—which quantifies evolutionary similarity. The SVM was created with the Spider package for MATLAB⁸⁹. Default parameters were used, with the exception of the use of an RBF kernel

with $\sigma=0.5$, and the balanced ridge set to the difference of the proportion of true and false matches in the training data. Training data consisted of matches to the 53 enzyme “Training Set”⁸⁵, using 4 EC digits to determine whether matches were correct, and was not changed to accommodate non-enzymes, as preliminary studies⁶¹ showed that reciprocal ETA identifies functionally relevant matches in both enzymes and non-enzymes. For further implementation details, see⁶¹.

Finally, reciprocal ETA removes non-reciprocal matches, taking only those in the intersection of the set of matches found by both matching methods.

Function Prediction

Given a list of significant template matches, ETA uses a hierarchical voting procedure to assign specific GO functions to a protein. For each match, ETA considers all annotated GO terms and their parent terms. Starting at the first depth of the GO molecular function hierarchy, each term at that depth is considered and gets one vote for each match annotated with that term. The one with the most votes is chosen as the predicted function. For the next depth of the hierarchy, each child term is considered, and again the one with a plurality is chosen as the function at that depth. If there is a tie, all tied functions are considered correct, and for each one, its children are considered as above (Figure 2a), resulting in multiple predicted functions. Voting continues until a term has no more children (a leaf node) or there are no votes for any of that term’s children. These most specific terms are the predicted function of the protein, as are their parent terms, in some cases including terms rejected earlier in the voting process (Figure 2b), or terms that have fewer votes than their parent terms (Figure 2c).

For ion-binding predictions, the more traditional plurality scheme is used: each filtered match represents one vote for a bound ion or ligand. The ion or ligand with a plurality of votes in this is the predicted ion or ligand.

When reciprocal ETA fails to annotate a protein, non-reciprocal matches may be included in a second attempt to annotate them, called all-match ETA.

Functional Sites

Catalytic Site Atlas⁵⁰ residues were used as functional sites for comparison to ETA templates. Identifying known functional sites in non-enzymes is more difficult, so for proteins with known binding partners (nucleic acids, small molecules, ions, and other proteins) residues with non-hydrogen atoms within 5 Å of non-hydrogen atoms of the binding partner (according to the PDB structure) were used as the functional site.

Positive and Negative Template Controls

Positive control templates were constructed for enzymes and non-enzymes by beginning with clustered residues from the functional sites, as defined above. Since most enzymes have only one to four CSA residues templates were supplemented by identifying residues that clustered with these functional sites, then choosing the best ranked among them. For non-enzymes, the six best-ranked residues around the ligand of interest were chosen. Negative control templates were selected by first identifying residues with ET percentile rank greater than 50% that were at least 15 Å from the known functional residues. One of these was then chosen at random, and if it clustered with five others, these six were chosen as the negative template. Otherwise, another residue was chosen, and so forth until a cluster that meets the distance and percentile rank constraints was found.

Performance Measurements

For overall performance, annotations were divided into four categories. *Correct and complete* predictions identify all known GO terms to the greatest available level of detail, and, optionally may provide additional or more specific terms. *Incomplete* predictions fully identify some GO terms down to the greatest available level of detail, but exclude others, and do not predict any additional functions. *Partially correct* predictions, match one or both of the following criteria: at least one predicted function matches known predictions to a minimum depth of three, but differ somewhere below that; or some terms are identified down to the lowest available ontology depth, but some terms are excluded, and extra terms are suggested that cannot be verified. *Incorrect*; predictions do not match known functions at any ontology depth below two. This threshold was chosen because an ontology depth of three maps to two digit EC numbers, but the depth of three digit EC numbers is inconsistent, and as a compromise between enzymatic functions, which tend to have deep hierarchies, and non-enzymatic functions, which tend to have shallower hierarchies.

For example, YibK from *Haemophilus Influenzae* protein (PDB ID 1j85, chain A) has two known functions: GO:0003723 (RNA binding) at depth three and GO:0008173 (RNA methyltransferase activity) at depth five. Suppose that ETA predicts GO:0008175 (tRNA methyltransferase activity) at depth six, GO:0003723 and GO:0005524 (ATP binding) for this protein. The GO:0008175 has an ontology depth of six and GO:0008173 is a parent of this function. ETA therefore predicts a function that includes the known function but is more specific. Since it cannot be verified as part of the existing annotation, it is disregarded when assessing performance. The prediction of GO:0003723 is fully correct, and the GO:0005524 is a new prediction. Since the protein is not annotated with GO:0005524, this is considered an extra prediction, and is disregarded since the ETA's predictions fully identify all annotated functions. This example is therefore correct and complete.

Alternately, suppose ETA predicts only GO:0008173. This term matches the lowest known level of one of the existing annotations, but ETA fails to identify GO:0003723, so this is considered incomplete. As a third example, suppose ETA predicts GO:0008173 and GO:0005524. Although the GO:0008173 is accurate to the lowest available depth, one prediction is missing (GO:0003723), and there is an extra term (GO:0005524), so this prediction is considered partially correct. In a fourth case, suppose ETA predicts GO:0016741 (transferase activity, transferring one-carbon groups) and GO:0003723. Although the latter completely identifies one known function, the former only agrees to the a depth of three, so this predictions is also considered partially correct. Finally, suppose ETA predicts the function as GO:0030170 (pyridoxal phosphate binding), so this prediction only agrees with the existing prediction at the first ontology depth, with GO:0005488 (binding), a parent of GO:0003723. Therefore, this is considered an incorrect prediction.

For depth-specific performance, ETA made predictions only down to the depth being considered, and was evaluated against all GO terms at that depth, including parent terms of annotations deeper in the hierarchy. Here, *correct and complete* predictions identify all terms at that depth, but no other terms. *Incomplete* predictions identify some functions at that depth, but exclude others, and do not predict additional functions. *Partially correct* predictions identify some know functions, but also identify additional functions. *Incorrect* predictions do not identify any known functions. If there are multiple paths to the most specific term, all are evaluated.

Depth 3 Positive Predictive Value (PPV) for general assessments and best-case PPV for depth-specific assessments were defined as the proportion of proteins for which ETA gives useful predictions—that is, correct and complete, incomplete, or partially correct—out of the total proteins with predictions. In the case of depth-specific predictions, this number will be

comparable to that used in Pal et al.⁶⁴, but we provide additional data breaking down correct predictions into *correct and complete*, *incomplete*, and *partially correct* predictions to provide information about predictions that are useful but may not be optimally specific. All-depth PPV for the general assessments and incomplete PPV for the depth-specific assessments are the proportion of predictions that are either correct and complete or incomplete. Finally, the fraction correct and complete (general) and lower bound PPV (depth-specific) are the proportion of predictions that are correct and complete only. Prediction coverage in all cases was defined as the proportion of proteins having some ETA annotation.

JAFAs⁶⁷ annotations were those given directly by the server. ProFunc Reverse Template⁴⁵ predictions were taken as the GO annotations provided by the server for the top-ranked annotated match.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

OL wishes to gratefully acknowledge partial support from NSF DBI-0547695 and CCF-0905536, and NIH-GM079656 and GM066099. Work by SE and RMW was also supported by training fellowships from the National Library of Medicine to the Keck Center for Interdisciplinary Bioscience Training of the Gulf Coast Consortia (NLM Grant No. 5T15LM07093).

References

1. Lee D, Redfern O, Orengo C. Predicting protein function from sequence and structure. *Nat Rev Mol Cell Biol* 2007;8:995–1005. [PubMed: 18037900]
2. Rentzsch R, Orengo CA. Protein function prediction - the power of multiplicity. *Trends Biotechnol.* 2009
3. Chandonia JM, Brenner SE. The impact of structural genomics: expectations and outcomes. *Science* 2006;311:347–351. [PubMed: 16424331]
4. Burley SK. An overview of structural genomics. *Nat. Struct. Biol* 2000;(7 Suppl):932–934. [PubMed: 11103991]
5. Brenner SE. A tour of structural genomics. *Nat Rev Genet* 2001;2:801–809. [PubMed: 11584296]
6. Xie L, Bourne PE. Functional coverage of the human genome by existing structures, structural genomics targets, and homology models. *PLoS Comput Biol* 2005;1:e31. [PubMed: 16118666]
7. Mercier KA, Baran M, Ramanathan V, Revesz P, Xiao R, Montelione GT, Powers R. FAST-NMR: functional annotation screening technology using NMR spectroscopy. *J Am Chem Soc* 2006;128:15292–15299. [PubMed: 17117882]
8. Waldron KJ, Robinson NJ. How do bacterial cells ensure that metalloproteins get the correct metal? *Nat Rev Microbiol* 2009;7:25–35. [PubMed: 19079350]
9. Friedberg I. Automated protein function prediction--the genomic challenge. *Brief Bioinform* 2006;7:225–242. [PubMed: 16772267]
10. Valencia A. Automatic annotation of protein function. *Curr Opin Struct Biol* 2005;15:267–274. [PubMed: 15922590]
11. Kuznetsova E, Proudfoot M, Sanders SA, Reinking J, Savchenko A, Arrowsmith CH, Edwards AM, Yakunin AF. Enzyme genomics: Application of general enzymatic screens to discover new enzymes. *FEMS Microbiol Rev* 2005;29:263–279. [PubMed: 15808744]
12. The Universal Protein Resource (UniProt) 2009. *Nucleic Acids Res* 2009;37:D169–D174. [PubMed: 18836194]
13. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* 1990;215:403–410. [PubMed: 2231712]

14. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;25:3389–3402. [PubMed: 9254694]
15. Gerlt JA, Babbitt PC. Can sequence determine function? *Genome Biol* 2000;1 REVIEWS0005.
16. Zhang B, Rychlewski L, Pawlowski K, Fetrow JS, Skolnick J, Godzik A. From fold predictions to function predictions: automation of functional site conservation analysis for functional genome predictions. *Protein Sci* 1999;8:1104–1115. [PubMed: 10338021]
17. Webb, EC. International Union of Biochemistry and Molecular Biology. Nomenclature Committee. Enzyme nomenclature 1992 : recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the nomenclature and classification of enzymes. San Diego: Academic Press; 1992.
18. Todd AE, Orengo CA, Thornton JM. Evolution of function in protein superfamilies, from a structural perspective. *J. Mol Biol* 2001;307:1113–1143. [PubMed: 11286560]
19. Tian W, Skolnick J. How well is enzyme function conserved as a function of pairwise sequence identity? *J Mol Biol* 2003;333:863–882. [PubMed: 14568541]
20. Addou S, Rentzsch R, Lee D, Orengo CA. Domain-based and family-specific sequence identity thresholds increase the levels of reliable protein function transfer. *J Mol Biol* 2009;387:416–430. [PubMed: 19135455]
21. Devos D, Valencia A. Intrinsic errors in genome annotation. *Trends Genet* 2001;17:429–431. [PubMed: 11485799]
22. Brenner SE. Errors in genome annotation. *Trends Genet* 1999;15:132–133. [PubMed: 10203816]
23. Galperin MY, Koonin EV. Sources of systematic error in functional annotation of genomes: domain rearrangement, non-orthologous gene displacement and operon disruption. *In Silico Biol* 1998;1:55–67. [PubMed: 11471243]
24. Karp PD. What we do not know about sequence analysis and sequence databases. *Bioinformatics* 1998;14:753–754. [PubMed: 10366280]
25. Kyrpides NC, Ouzounis CA. Whole-genome sequence annotation: 'Going wrong with confidence'. *Mol Microbiol* 1999;32:886–887. [PubMed: 10361291]
26. Ouzounis CA, Karp PD. The past, present and future of genome-wide re-annotation. *Genome Biol* 2002;3 COMMENT2001.
27. Pallen M, Wren B, Parkhill J. 'Going wrong with confidence': misleading sequence analyses of CiaB and clpX. *Mol Microbiol* 1999;34:195. [PubMed: 10540297]
28. Engelhardt BE, Jordan MI, Muratore KE, Brenner SE. Protein molecular function prediction by Bayesian phylogenomics. *PLoS Comput Biol* 2005;1:e45. [PubMed: 16217548]
29. Wallace AC, Borkakoti N, Thornton JM. TESS: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites. *Protein Sci* 1997;6:2308–2323. [PubMed: 9385633]
30. Barker JA, Thornton JM. An algorithm for constraint-based structural template matching: application to 3D templates with statistical analysis. *Bioinformatics* 2003;19:1644–1649. [PubMed: 12967960]
31. Kleywegt GJ. Recognition of spatial motifs in protein structures. *J Mol Biol* 1999;285:1887–1897. [PubMed: 9917419]
32. Stark A, Russell RB. Annotation in three dimensions. PINTS: Patterns in Non-homologous Tertiary Structures. *Nucleic Acids Res* 2003;31:3341–3344. [PubMed: 12824322]
33. Artymiuk PJ, Poirrette AR, Grindley HM, Rice DW, Willett P. A graph-theoretic approach to the identification of three-dimensional patterns of amino acid side-chains in protein structures. *J Mol Biol* 1994;243:327–344. [PubMed: 7932758]
34. de Rinaldis M, Ausiello G, Cesareni G, Helmer-Citterich M. Three-dimensional profiles: a new tool to identify protein surface similarities. *J Mol Biol* 1998;284:1211–1221. [PubMed: 9837739]
35. Laskowski RA. SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J Mol Graph* 1995;13:323–330. 307–8. [PubMed: 8603061]
36. Kleywegt GJ, Jones TA. Detection, delineation, measurement and display of cavities in macromolecular structures. *Acta Crystallogr D Biol Crystallogr* 1994;50:178–185. [PubMed: 15299456]

37. Binkowski TA, Naghibzadeh S, Liang J. CASTp: Computed Atlas of Surface Topography of proteins. *Nucleic Acids Res* 2003;31:3352–3355. [PubMed: 12824325]
38. Shulman-Peleg A, Nussinov R, Wolfson HJ. Recognition of functional sites in protein structures. *J Mol Biol* 2004;339:607–633. [PubMed: 15147845]
39. Binkowski TA, Freeman P, Liang J. pvSOAR: detecting similar surface patterns of pocket and void surfaces of amino acid residues on proteins. *Nucleic Acids Res* 2004;32:W555–W558. [PubMed: 15215448]
40. Glaser F, Morris RJ, Najmanovich RJ, Laskowski RA, Thornton JM. A method for localizing ligand binding pockets in protein structures. *Proteins* 2006;62:479–488. [PubMed: 16304646]
41. Kinoshita K, Furui J, Nakamura H. Identification of protein functions from a molecular surface database, eF-site. *J Struct Funct Genomics* 2002;2:9–22. [PubMed: 12836670]
42. Schmitt S, Kuhn D, Klebe G. A new method to detect related function among proteins independent of sequence and fold homology. *J Mol Biol* 2002;323:387–406. [PubMed: 12381328]
43. Ivanisenko VA, Pintus SS, Grigorovich DA, Kolchanov NA. PDBSiteScan: a program for searching for active, binding and posttranslational modification sites in the 3D structures of proteins. *Nucleic Acids Res* 2004;32:W549–W554. [PubMed: 15215447]
44. Ivanisenko VA, Pintus SS, Grigorovich DA, Kolchanov NA. PDBSite: a database of the 3D structure of protein functional sites. *Nucleic Acids Res* 2005;33:D183–D187. [PubMed: 15608173]
45. Laskowski RA, Watson JD, Thornton JM. Protein function prediction using local 3D templates. *J Mol Biol* 2005;351:614–626. [PubMed: 16019027]
46. Shulman-Peleg A, Nussinov R, Wolfson HJ. SiteEngines: recognition and comparison of binding sites and protein-protein interfaces. *Nucleic Acids Res* 2005;33:W337–W341. [PubMed: 15980484]
47. Redfern OC, Dessailly B, Orengo CA. Exploring the structure and function paradigm. *Curr Opin Struct Biol* 2008;18:394–402. [PubMed: 18554899]
48. Watson JD, Sanderson S, Ezersky A, Savchenko A, Edwards A, Orengo C, Joachimiak A, Laskowski RA, Thornton JM. Towards fully automated structure-based function prediction in structural genomics: a case study. *J Mol Biol* 2007;367:1511–1522. [PubMed: 17316683]
49. Wallace AC, Laskowski RA, Thornton JM. Derivation of 3D coordinate templates for searching structural databases: application to Ser-His-Asp catalytic triads in the serine proteinases and lipases. *Protein Sci* 1996;5:1001–1013. [PubMed: 8762132]
50. Porter CT, Bartlett GJ, Thornton JM. The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res* 2004;32:D129–D133. [PubMed: 14681376]
51. Kristensen DM, Chen BY, Fofanov VY, Ward RM, Lisewski AM, Kimmel M, Kaviraki LE, Lichtarge O. Recurrent use of evolutionary importance for functional annotation of proteins based on local structural similarity. *Protein Sci* 2006;15:1530–1536. [PubMed: 16672239]
52. Gherardini PF, Helmer-Citterich M. Structure-based function prediction: approaches and applications. *Brief Funct Genomic Proteomic* 2008;7:291–302. [PubMed: 18599513]
53. Lichtarge O, Bourne HR, Cohen FE. An evolutionary trace method defines binding surfaces common to protein families. *J Mol Biol* 1996;257:342–358. [PubMed: 8609628]
54. Madabushi S, Yao H, Marsh M, Kristensen DM, Philippi A, Sowa ME, Lichtarge O. Structural clusters of evolutionary trace residues are statistically significant and common in proteins. *J Mol Biol* 2002;316:139–154. [PubMed: 11829509]
55. Yao H, Kristensen DM, Mihalek I, Sowa ME, Shaw C, Kimmel M, Kaviraki L, Lichtarge O. An accurate, sensitive, and scalable method to identify functional sites in protein structures. *J Mol Biol* 2003;326:255–261. [PubMed: 12547207]
56. Madabushi S, Gross AK, Philippi A, Meng EC, Wensel TG, Lichtarge O. Evolutionary trace of G protein-coupled receptors reveals clusters of residues that determine global and class-specific functions. *J Biol Chem* 2004;279:8126–8132. [PubMed: 14660595]
57. Onrust R, Herzmark P, Chi P, Garcia PD, Lichtarge O, Kingsley C, Bourne HR. Receptor and betagamma binding sites in the alpha subunit of the retinal G protein transducin. *Science* 1997;275:381–384. [PubMed: 8994033]
58. Ribes-Zamora A, Mihalek I, Lichtarge O, Bertuch AA. Distinct faces of the Ku heterodimer mediate DNA repair and telomeric functions. *Nat Struct Mol Biol* 2007;14:301–307. [PubMed: 17351632]

59. Sowa ME, He W, Slep KC, Kercher MA, Lichtarge O, Wensel TG. Prediction and confirmation of a site critical for effector regulation of RGS domain activity. *Nat Struct Biol* 2001;8:234–237. [PubMed: 11224568]
60. Bartlett GJ, Porter CT, Borkakoti N, Thornton JM. Analysis of catalytic residues in enzyme active sites. *J Mol Biol* 2002;324:105–121. [PubMed: 12421562]
61. Ward RM, Erdin S, Tran TA, Kristensen DM, Lisewski AM, Lichtarge O. De-orphaning the structural proteome through reciprocal comparison of evolutionarily important structural features. *PLoS ONE* 2008;3:e2136. [PubMed: 18461181]
62. Kristensen DM, Ward RM, Lisewski AM, Erdin S, Chen BY, Fofanov VY, Kimmel M, Kavraki LE, Lichtarge O. Prediction of enzyme function based on 3D templates of evolutionarily important amino acids. *BMC Bioinformatics* 2008;9:17. [PubMed: 18190718]
63. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000;25:25–29. [PubMed: 10802651]
64. Pal D, Eisenberg D. Inference of protein function from protein structure. *Structure* 2005;13:121–130. [PubMed: 15642267]
65. Jensen LJ, Gupta R, Staerfeldt HH, Brunak S. Prediction of human protein function according to Gene Ontology categories. *Bioinformatics* 2003;19:635–642. [PubMed: 12651722]
66. Shen HB, Chou KC. EzyPred: a top-down approach for predicting enzyme functional classes and subclasses. *Biochem Biophys Res Commun* 2007;364:53–59. [PubMed: 17931599]
67. Friedberg I, Harder T, Godzik A. JAFa: a protein function annotation meta-server. *Nucleic Acids Res* 2006;34:W379–W381. [PubMed: 16845030]
68. Andreini C, Bertini I, Cavallaro G, Holliday GL, Thornton JM. Metal ions in biological catalysis: from enzyme databases to general principles. *J Biol Inorg Chem* 2008;13:1205–1218. [PubMed: 18604568]
69. Shi W, Chance MR. Metallomics and metalloproteomics. *Cell Mol Life Sci* 2008;65:3040–3048. [PubMed: 18560755]
70. Tottey S, Waldron KJ, Firbank SJ, Reale B, Bessant C, Sato K, Cheek TR, Gray J, Banfield MJ, Dennison C, Robinson NJ. Protein-folding location can regulate manganese-binding versus copper- or zinc-binding. *Nature* 2008;455:1138–1142. [PubMed: 18948958]
71. Laskowski RA, Watson JD, Thornton JM. ProFunc: a server for predicting protein function from 3D structure. *Nucleic Acids Res* 2005;33:W89–W93. [PubMed: 15980588]
72. Henschel A, Kim WK, Schroeder M. Equivalent binding sites reveal convergently evolved interaction motifs. *Bioinformatics* 2006;22:550–555. [PubMed: 16287935]
73. Rakus JF, Fedorov AA, Fedorov EV, Glasner ME, Hubbard BK, Delli JD, Babbitt PC, Almo SC, Gerlt JA. Evolution of enzymatic activities in the enolase superfamily: L-rhamnonate dehydratase. *Biochemistry* 2008;47:9944–9954. [PubMed: 18754693]
74. Clackson T, Wells JA. A hot spot of binding energy in a hormone-receptor interface. *Science* 1995;267:383–386. [PubMed: 7529940]
75. Bate P, Warwicker J. Enzyme/non-enzyme discrimination and prediction of enzyme active site location using charge-based methods. *J Mol Biol* 2004;340:263–276. [PubMed: 15201051]
76. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 1983;22:2577–2637. [PubMed: 6667333]
77. Dobson PD, Doig AJ. Distinguishing enzyme structures from non-enzymes without alignments. *J Mol Biol* 2003;330:771–783. [PubMed: 12850146]
78. Baameur F, Morgan D, Yao G, Tran T, Sabui S, McMurray J, Lichtarge O, Clark R. Evolutionary Trace Residues in the RGS Homology Domain Define a Site Important for GRK5 and 6 Phosphorylation of GPCRs. 2009 unpublished.
79. Redfern OC, Dessailly BH, Dallman TJ, Sillitoe I, Orengo CA. FLORA: a novel method to predict protein function from structure in diverse superfamilies. *PLoS Comput Biol* 2009;5:e1000485. [PubMed: 19714201]

80. Redfern OC, Harrison A, Dallman T, Pearl FM, Orengo CA. CATHEDRAL: a fast and effective algorithm to predict folds and domain boundaries from multidomain protein structures. *PLoS Comput Biol* 2007;3:e232. [PubMed: 18052539]
81. Shindyalov IN, Bourne PE. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng* 1998;11:739–747. [PubMed: 9796821]
82. Berman H, Henrick K, Nakamura H, Markley JL. The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res* 2007;35:D301–D303. [PubMed: 17142228]
83. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. *Nucleic Acids Res* 2000;28:235–242. [PubMed: 10592235]
84. Hobohm U, Scharf M, Schneider R, Sander C. Selection of representative protein data sets. *Protein Sci* 1992;1:409–417. [PubMed: 1304348]
85. Kristensen DM, Ward RM, Lisewski AM, Chen BY, Fofanov VY, Kimmel M, Kavraki L, Lichtarge O. Prediction of Enzyme Function Based on 3D Templates of Evolutionarily Important Amino Acids. *BMC Bioinformatics*. 2007 Submitted.
86. Morgan DH, Kristensen DM, Mittelman D, Lichtarge O. ET viewer: an application for predicting and visualizing functional sites in protein structures. *Bioinformatics* 2006;22:2049–2050. [PubMed: 16809388]
87. Mihalek I, Res I, Lichtarge O. A family of evolution-entropy hybrid methods for ranking protein residues by importance. *J Mol Biol* 2004;336:1265–1282. [PubMed: 15037084]
88. Ho, BK. 2005. <http://boscoh.com/code/rmsd.c>
89. Weston, J.; Elisseeff, A.; BakIr, G.; Sinz, F. Spider; 2006.

Figure 1.

Reciprocal match between *Mycobacterium tuberculosis* v1626 (PDB 1sd5, chain A; green cartoon) to *Tolypothrix* species PCC 7601 phytochrome response regulator rcpb (PDB 1k66, chain A; orange cartoon). ET analysis of 1sd5A identified a 10-residue functional site (yellow spheres), from which the template picker chose six residues (D21, P59, D65, A93, Y111, K114, red spheres). Their C_{α} coordinates and amino acid types (with some variations allowed) matched 1k66A (D14, P63, D69, T100, Y118, K121, red spheres). A trace of 1k66A identified a 10-residue functional site, and six residues were chosen for a template (E13, D14, D69, P73, K121, P122, blue spheres), which reciprocally matched 1sd5 (E20, D21, D65, P69, K114, P115, blue spheres). Three residues (D21, D65, K114 from 1sd5A; D14, D69, K121 from 1k66A; purple spheres) were in both templates.

Figure 2a

Figure 2b

Figure 2c

Figure 2. Illustration and examples of plurality voting procedure with GO molecular function terms. Each box is a GO term. Green boxes represent terms accepted by the voting procedure; red boxes were rejected. Colored dots next to the box represent a match with that function, and correspond to the matches shown at the side of each figure. 2a Annotation of 1nhz, chain A, illustrating the assignment of multiple functions to a protein when there are ties; 2b Annotation of 1q45, chain A, illustrating the prediction of the most specific term available; 2c Annotation of 2p68A, illustrating a case where an initially rejected term (Transferase, GO:0016740) is included if one of its children is selected.

Figure 3.

ETA performance. The performance of reciprocal ETA performance as matches above a sequence identity cutoff were removed is shown (the test sets remain the same size in all cases), as are additional predictions made by all-match ETA. Proteins with correct and complete predictions are shown in red; incomplete, orange; partially correct, yellow; incorrect, gray; no predictions, white. Coverage (orange circles, depth 3 PPV (red squares), all-depth PPV (red circles), and fraction correct and complete (red triangles) are plotted against the right axis. 3a Performance for 1889 SG enzymes. 3b Proof-of-concept performance for 50 non-SG non-enzymes. 3c Performance for 311 SG non-enzymes. 3d Performance for 184 SG ion-binding proteins.

Figure 4.

Reciprocal ETA performance at varying GO depths. Performance is reported only with respect to predictions at that depth, using the color scheme from Figure 3 (substituting best-case PPV for depth 3 PPV, incomplete PPV for all-depth PPV, and lower bound PPV for the fraction correct and complete). 4a Performance for 1889 SG enzymes. 4b Performance for 311 SG non-enzymes.

Figure 5.

Comparisons of ETA performance for SG proteins to other methods. ETA is compared, using the color scheme in Figure 3, to JAJA for 50 enzymes, 311 non-enzymes, and 184 ion-binding proteins; and is compared to ProFunc's Reverse Templates method (RT) for 120 enzymes and 224 non-enzymes.

Figure 6. Distribution of match sequence identity for un-annotated proteins. Histogram showing the percentage sequence identity for ETA-annotated SG proteins with their highest sequence identity match.

Figure 7.

Overlap between template residues and known functional sites for 846 enzymes, 63 non-enzymes and 184 metal ion-binding SG proteins. The number of overlapping residues is shown in the legend; when no residues overlapped, templates were divided into those that were within 10 Å of any non-hydrogen atom and those farther away from the functional site.

Figure 8.
Examples of non-enzyme templates. Green cartoon, query protein; purple, reciprocal template residues; red, one-to-many residues; blue, bound ions, ligands or protein-protein interface residues. 7a 1bmo, chain A, with calcium ion; 7b 1gzs, chain B, with a heme molecule; 7c Human growth hormone 1a22, chain A, and human growth hormone receptor 1a22, chain B, (orange cartoon) with the hormone receptor's interface residues R271, W304, I305 and P306.

Figure 9.

Annotation performance for ETA's template picker and two control template pickers. Performance is shown for both one-to-many and reciprocal ETA (the many-to-one portion of the reciprocal search used the standard ETA template picker). ETA templates were constructed as described elsewhere. Positive controls are constructed from known functional sites as described below. Negative control templates are constructed from poorly ranked ET residues that are not near the known functional site. 8a Performance for 51 SG enzymes. CSA+ETA (positive control) templates start with CSA residues and then supplement these with nearby highly ranked ET residues. 8b Performance for 41 SG non-enzymes. "Binding+ETA" (positive control) templates start with residues from a ligand or ion-binding site and supplement these with nearby highly ranked ET residues.

Figure 10a

Figure 10b

Figure 10.

Composition of ETA enzyme templates and CSA residues (846 SG proteins); and of non-enzyme templates and non-enzyme binding sites (63 SG proteins). 9a Amino acid composition. 9b Secondary structure composition.

Table 1
Functions of metal ion-binding proteins
Summary of functions of predicted ion-binding proteins. Predictions are listed as correct predictions/total predictions.

Function	Ca ²⁺	Co ²⁺	Cu ⁺	Fe ²⁺	Mg ²⁺	Mn ²⁺	Zn ²⁺	Total
Catalytic activity	0 / 1	0 / 2	2 / 2	1 / 1	6 / 10		8 / 13	17 / 29
Signal Transduction	8 / 8							8 / 8
Response regulator					0 / 1	1 / 2		1 / 3
Gene regulation							20 / 20	20 / 20
Protein-Protein interaction							20 / 20	20 / 20
Metal transport							1 / 1	1 / 1
Hypothetical or unknown	0 / 1				3 / 4		3 / 3	6 / 8
Total	8 / 10	0 / 2	2 / 2	1 / 1	9 / 15	1 / 2	52 / 57	73 / 89

Table 2
ETA predictions for 2642 Structural Genomics proteins

Summary of depth one GO terms predicted by ETA predictions for un-annotated structural genomics proteins.

	Reciprocal	All Match
Total Predicted Enzymes	234	631
Oxidoreductase activity (GO:0016491)	58	168
Transferase activity (GO:0016740)	90	225
Hydrolase activity (GO:0016787)	67	227
Lyase activity (GO:0016829)	15	53
Isomerase activity (GO:0016853)	24	51
Ligase activity (GO:0016874)	7	26
Catalytic activity (GO:0003824)	7	15
Total Predicted Non-enzymes	280	257
Binding (GO:0005488)	323	375
Transcription regulator activity (GO:0030528)	43	42
Molecular transducer activity (GO:0060089)	24	24
Structural molecule activity (GO:0005198)	19	10
Enzyme regulator activity (GO:0030234)	10	9
Transporter activity (GO:0005215)	6	24
Motor activity (GO:0003774)	3	0
Translation regulator activity (GO:0045182)	2	7
Antioxidant activity (GO:0016209)	1	7
No Predictions	2138	1250
Total Proteins	2642	2138

Table 3
ETA predictions for SG metal ion-binding proteins

Summary of metal ions predicted to bind structural genomics proteins.

	Reciprocal	All Match
Total Predicted Metal Ion-binding	21	43
Ca ²⁺	2	15
Co ²⁺	0	0
Cu ⁺	3	5
Fe ²⁺	1	0
Mg ²⁺	5	8
Mn ²⁺	3	2
Ni ²⁺	1	0
Zn ²⁺	6	13
No Predictions	2228	2185
Total Proteins	2249	2228