Prediction and experimental validation of enzyme substrate specificity in protein structures

Shivas R. Amin,a,b,1 Serkan Erdin,a,c,1 R. Matthew Ward,a Rhonald C. Lua,a and Olivier Lichtargea,c,2

aDepartment of Molecular and Human Genetics and cComputational and Integrative Biomedical Research Center, Baylor College of Medicine, Houston, TX 77030; and bBiology Department, University of Saint Thomas, Houston, TX 77006

Edited by Barry Honig, Howard Hughes Medical Institute, Columbia University, New York, NY, and approved September 24, 2013 (received for review March 17, 2013)

Structural Genomics aims to elucidate protein structures to identify their functions. Unfortunately, the variation of just a few residues can be enough to alter activity or binding specificity and limit the functional resolution of annotations based on sequence and structure; in enzymes, substrates are especially difficult to predict. Here, large-scale controls and direct experiments show that the local similarity of five or six residues selected because they are evolutionarily important and on the protein surface can suffice to identify an enzyme activity and substrate. A motif of five residues predicted that a previously uncharacterized Siliicibacter sp. protein was a carboxylesterase for short fatty acyl chains, similar to hormone-sensitive-lipase-like proteins that share less than 20% sequence identity. Assays and directed mutations confirmed this activity and showed that the motif was essential for catalysis and substrate specificity. We conclude that evolutionary and structural information may be combined on a Structural Genomics scale to create motifs of mixed catalytic and noncatalytic residues that identify enzyme activity and substrate specificity.

A

As the list of known genes grows exponentially, the elucidation of their function remains a major bottleneck and lags behind the production of sequences (1–5). The best approach remains to search computationally for functionally characterized sequence homologs, ideally with greater than 50% sequence identity (6). Binding specificity, however, is sensitive to subtle amino acid differences, and the transfer of substrate between related enzymes is prone to errors when sequence identity is below 65–80% (7–9). These thresholds vary from case to case: Some orthologs will maintain identical functions down to 25% sequence identity (9), whereas paralogs can take on highly diverse activities (10). Other difficulties that plague annotation transfer between homologs are that individual small molecules may bind to multiple and distinct molecular pockets (11), that different residues can support similar chemistries (12), and that activity can vary even when catalytic residues are conserved (13–18). To raise annotation accuracy, Structural Genomics (19) made structural information widely available and spurred the development of annotation methods dependent on local chemical and physical environments (20), sequence and structural comparisons (21), or 3D templates (22). In the case of the latter, these methods search between proteins for local structural similarities over a few signature residues that represent the telltale parts of a functional site, so-called “3D templates” (3, 14, 18, 22–24). The residue composition of 3D templates is critical, however, and derived from experiments (25) or from analyses of functional sites and determinants (14, 15, 26). The sensitivity and specificity of template-based annotations still needs to be established experimentally (27, 28), but retrospective controls suggest they often predict enzyme catalytic activity (14, 16, 17, 29, 30).

Here, to extend the functional resolution of 3D template annotations to substrates, we exploit Evolutionary Tracing (ET) (31, 32). ET ranks sequence positions by the tendency of their evolutionary variations to correlate with major or with minor divergences. Top-ranked ET sequence positions are the most evolutionarily and, presumably, functionally important, and indeed they map out functional sites and specificity determinants (33) accurately enough to efficiently design mutations that block or swap functions among homologs in vitro (34–36) or in vivo (37, 38).

Accordingly, given a query protein of unknown function, the ET Annotation pipeline (ETA) builds a 3D template from five or six top-ranked ET residues that also cluster together on surface regions of protein structures (31, 32). ETA then searches already annotated protein structures, the targets, for those that match the query 3D template (Fig. 1 and Movie S1). False positive matches are common but can be recognized because they typically (i) involve unimportant residues in the target (39), (ii) are not reciprocated back to the query (40), and (iii) point to multiple proteins that each bear unrelated functions. With appropriate specificity filters to eliminate these false positives, ETA identified enzyme activity down to the first three Enzyme Commission (EC) levels with 92% accuracy (40), as well as in nonenzymes (41) in large-scale Structural Genomics retrospective controls. The prediction of substrate specificity remains an open question and further requires accurate identification of the fourth and last EC level (42) presumably by adding a more discriminating use of 3D template residues than is sufficient to specify a general chemical process (43). Some sequence methods (29, 30) and other structure methods (14, 44) have aimed to predict all four EC levels, but to our knowledge they have

Significance

Many proteins solved by Structural Genomics have low sequence identity to other proteins and cannot be assigned functions. To address this problem, we present a computational approach that creates structural motifs of a few evolutionarily important residues, and these motifs probe local geometric and evolutionary similarities in other protein structures to detect functional similarities. This approach does not require prior knowledge of functional mechanisms and is highly accurate in computational benchmarks when annotations rely on homologs with low sequence identity. We further demonstrate the accuracy of this approach using biochemical and mutagenesis studies to validate two predictions of unannotated proteins.


The authors declare no conflict of interest.

This article is a PNAS Direct Submission. Freely available online through the PNAS open access option.

5S.R.A. and S.E. contributed equally to this work.
6To whom correspondence should be addressed. E-mail:lichtarge@bcm.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1305162110/-/DCSupplemental.

www.pnas.org/cgi/doi/10.1073/pnas.1305162110
not been directly tested on de novo predictions of substrate specificity.

In this study, we improve the functional resolution of the ETA pipeline to identify relevant functional homology down to very low sequence identity and add substrate specificity to its large-scale predictions. We then experimentally validate the predictions and show that both catalytic and noncatalytic residues are essential for 3D templates to pinpoint activity and substrate specificity.

Results

Substrate-Level Predictions from 3D Templates. To predict substrates, a new confidence score was created that empirically favored multiple mutually consistent matches between a query and other protein structures if they had identical fourth-level EC numbers (42) (Materials and Methods and Fig. 1B). In a retrospective control, 605 enzymes from all six major classes were stripped of their annotations and then matched to a set of 3,082 annotated target structures (Fig. S1A). Almost all of the ETA annotations that scored above 1 on this new substrate confidence score (Materials and Methods) were correct over all four EC levels (99%), and only two-thirds of those with scores below 1 were (63%, Fig. S1B). Strikingly, high scoring annotations remained nearly perfect even when the sequence identity between a query and its matches fell below 30% (Fig. 1 C and D). By contrast, the template-based matching method COFACTOR (45) performed worse than ETA (96% vs. 92%, Table S1) despite having a larger target set and matching to proteins with greater than 30% sequence identity (see Materials and Methods for further detail). Similarly, annotations based on overall structural matches became increasingly inaccurate below 45% sequence identity (Fig. 1D). These data show that when a plurality of ETA matches agree on all four EC levels, function predictions based on just a few evolutionarily important and structurally clustered residues yield accurate predictions of function and substrates that improve on current protein structure comparison methods.

A potential limitation of using plurality to determine confidence may occur when only a single structure of a functional class is solved that yields a maximum confidence score equal to 1. However, ETA remains accurate when the confidence score is equal to 1, with correct predictions in 92% of cases (Fig. S1A). Furthermore, in the 24 cases that were based on the only representative of a particular functional class, the ETA annotation was correct 100% of the time (n = 24). These data suggest that even in the absence of plurality ETA accurately identifies substrate specificity.

To probe the basis of ETA accuracy, we compared these ET-derived 3D template residues to those in the Mechanism, Annotation and Classification in Enzymes (MACIE) database of catalytic sites (Fig. 2A) (46). Histidine, aspartic acid, and arginine are preponderant in both, consistent with catalytic roles and hence with being structurally invariant in functionally similar enzymes (47). Noncatalytic residues, however, are mostly absent from MACIE but frequent in 3D templates. For example, glycines and prolines are in 69% and 27% of 3D templates, respectively. The low all-atom root-mean-square deviation (rmsd) between ETA templates and their cognate match sites, over a wide range of sequence identities (Fig. 2B), shows that just like catalytic residues these noncatalytic template residues are also structurally invariant among enzymes that catalyze the same reaction, and thus do not hinder recognition by ETA templates. This is distinct from the rest of the structure, as the rmsd between a query protein and its ETA matches, over the entire structures or between clusters of evolutionarily unimportant residues, increases as sequence identity decreases (Fig. 2B). This structural invariance of ETA templates does not compromise recognition of similarities among structures with and without
Fig. 2. Noncatalytic residues are prevalent in structurally invariant ETA templates. (A) Comparison of log propensities of ETA six-residue templates and known catalytic residues from MACiE database 3.0. ETA templates use glycine and proline residues at higher propensity than they appear in catalytic sites (Pearson coefficient = 0.58 considering all residues; 0.91 when ignoring G and P). (B) The rmsd for structural alignments for ETA matches, binned according to sequence identity. Alignments were generated using ETA templates and the entire structures using all atoms (lovoalign) and only the alpha carbons (TM-Align) for all matches. Negative control templates were also made using clusters of evolutionarily unimportant residues and aligned (alpha carbon only).

In practice, the human zeta-crystallin [Protein Data Bank (PDB) ID code 1yb5, chain A] illustrates how both catalytic and noncatalytic residues contribute to predictions. Out of the 3,082 targets, a 3D template from this protein correctly matched a quinone oxidoreductase from *Escherichia coli* (PDB ID code 1qor, chain B, EC 1.6.5.5) (52) (Fig. 3A). The matched residues had near identical geometry and evolutionary importance to those from the zeta-crystallin template, including four glycines previously described as contributing to stability due to their structural and values closer to 100 are unimportant. The yellow highlights represent residues where the zeta-crystallin/alcohol dehydrogenase match site leads to different substrate specificity despite the similarity in global topology. The cluster alignments are also shown in sequence form in A next to the target structures, where one dot signifies aligned residues and two dots signify an rmsd of less than 0.5 Å. (C and D) ETA alpha-carbon templates for *S. solfataricus* alcohol dehydrogenase (EC 1.1.1.1). Alcohol dehydrogenases have the same Rossmann fold as quinone reductases but, unlike them, require a zinc ion for proper function (53, 54). Therefore, in alcohol dehydrogenases, it is the residues near that zinc ion that are most evolutionarily important, and these are not matched by the zeta-crystallin query template (Fig. 3B). Indeed, such an erroneous
match to alcohol dehydrogenases would include a residue as far as 4 Å off from its position in the template (Fig. 3D). These data show the extent to which both evolutionarily important catalytic and noncatalytic template residues capture local features that specifically determine function.

**Experimental Validation Studies.** We sought next to directly validate ETA with experimental controls. For this, we selected predictions for which there was no prior knowledge of activity and substrate and that were based on matches to proteins with less than 30% sequence identity, as accuracy is especially challenging in these distant homologs.

The first case is dhaf_2064 from *Desulfutobacterium hafniense* (PDB ID code 3db2, chain A; Fig. S2 and Fig. 4A), which is a member of the functionally diverse family of Rossman fold proteins. In keeping with the prior discussion, the dhaf_2064 template, (G11, G13, E95, K96, P97, H184), included two glycines and one proline that are noncatalytic (Fig. S2). This template matched three proteins that were evolutionarily distant with sequence identity of 21%, 17%, and 18%, namely *Lactobacillus plantarum* (PDB ID code 3ee3), *Corynebacterium glutamicum* (PDB ID code 3euw), and *Salmonella typhimurium* (PDB ID code 3sec), respectively. The mean rmsd of the matches was 0.8 Å, and the mean evolutionary importance rank of the matches was 1.5%. Strikingly, all three proteins carried the same oxidoreductase catalytic function, and thus dhaf_2064 was predicted to have that function as well: EC 1.1.1.18, namely, to convert myo-inositol into scyllo-inosose via the reduction of NAD+.

To confirm this prediction, we expressed dhaf_2064 in *E. coli* and tested crude extracts in vitro for myo-inositol activity. Extracts containing dhaf_2064 had significantly more activity toward the predicted substrates (4 U/mg) than negative control extracts that lacked the protein (0 U/mg) (Fig. 4A). As a further control, and because template residue E95 is thought to mediate a key binding event to cofactor NAD+ (55), we showed that an extract from an E95A mutant abolished activity (0 U/mg). These data confirm three aspects of this ETA prediction: the enzyme activity, the substrate specificity, and the critical role of at least one template residue.

A second case focused on tm1040_2492 from *Silicibacter sp.* (PDB ID code 2pbl, chain C), hereafter protein x for short, selected because it belongs to the highly populated α/β hydrolase fold and may therefore match a wide spectrum of possible functions. ETA matched protein x to three proteins: EstE1 (PDB ID code 2c7b, chain B; 18% identity) from the metagenome of thermophilic organisms, Est2 (PDB ID code 2hm7, chain A; 18% identity) from *Alcyclobacillus acidocaldarius*, and AFEST (PDB ID code 1jji, chain D; 16% identity; Fig. 4B) from the archaeon *Archaeoglobus fulgidus*. Strikingly, all were carboxylesterases from the hormone-sensitive lipase (HSL)–like family (Fig. 4B, EC 3.1.1.1), suggesting that, like them, protein x catalyzed the hydrolysis of an ester bond into an alcohol and a carboxylic acid.

To test this prediction in vitro, we monitored the degradation of a carboxylesterase substrate, 4-nitrophenyl acetate, by a crude extract of *E. coli* containing recombinant protein x. This extract had significantly more activity (13 U/mg) than a negative control extract that lacked protein x (1.5 U/mg) (Fig. 4A). Although these data show that as predicted protein x hydrolyzes some ester bonds, it is important to recognize that 4-nitrophenyl acetate is not representative of the entire spectrum of substrates with ester bonds. HSL-like carboxylesterases (EC 3.1.1.1) target ester bonds from short fatty acid chains, and their activity falls dramatically in chains longer than 8 or 10 carbons (56, 57). By contrast, lipases (EC 3.1.1.3) hydrolyze ester bonds from fatty acid carbon chains with more than 10 carbons (58, 59).

To confirm the substrate selectivity of protein x, we purified it (Fig. S3) and tested its activity in vitro against fatty acid chains of increasing length: acetate (2C), butyrate (4C), octanoate (8C), decanoate (10C), and palmitate (16C, lipase substrate) (Fig. 4E).

Protein x activity decreased from 100% against 4-nitrophenyl acetate (2C) to 15%, 3%, 2.5%, and 0.5% activity toward 4-nitrophenyl butyrate, octanoate, decanoate, and palmitate, respectively. These data show that like other HSL-like carboxylesterases, protein x has extremely low activity toward fatty acid carbon chain lengths of 8 or 10, and no activity toward the lipase
Interestingly, two of the three proteins that matched protein sensitivity from 62% to 70% in our benchmarks (Fig. S4). In contrast, mutations of nontemplate residues plateau residues, whether catalytic or not, always abolish enzymatic activity. To probe the functional role of the 3D template residues, \{G71, W73, P104, S136, H241\}, we structurally aligned protein x with the HSL-like carboxylesterases. The mean rmsd of these residues with their targets is 0.6 Å, so they are structurally invariant in HSL-like family members (Fig. 4C).

The template contains two of the three residues from the Ser–His–Glu catalytic triad, serine 136, and histidine 241. The glutamic acid at position 214 was not included in the template as it is relatively buried and ETA preferentially picks surface-exposed residues. The template also contains glycine 71, which forms the critical oxyanion hole in HSL-like carboxylesterases (60, 61). The role of the tryptophan at position 73, and of the proline at position 104, remains unknown, but their spatial invariance suggests they may stabilize and promote hydrophobic interactions in the active site cleft consistent with a role in substrate specificity and reaction mechanisms (60).

This last possibility is further supported by a second structural alignment of protein x with the lipase from Candida rugosa (Fig. 4D). In that case, the catalytic serine and histidine are structurally invariant between the two proteins, as is the glycine in the oxyanion hole and the tryptophan that lines the active site. The proline, however, cannot be aligned with any residue in the C. rugosa lipase. Lipases, unlike the HSL-like carboxylesterases, undergo conformational changes to accommodate large lipid substrates in their active sites, and in this region lipases are rich with flexible random coils. This suggests that, unlike HSL-like carboxylesterases, lipases might not tolerate a rigid proline residue at the cognate location. These data show that a single noncatalytic template residue, proline 104, is critical to distinguish substrate specificity, whereas the catalytic residues cannot do so by themselves, as these are common to both the lipases and the HSL-like carboxylesterases enzyme families.

To confirm that the template residues were essential to function, we individually mutated each one to alanine, purified the modified proteins, and assayed their activity toward the most reactive substrate (Fig. 4F). Mutation of any of the template residues, whether catalytic or not, always abolishes enzymatic activity. In contrast, mutations of nontemplate residues near the catalytic triad do not necessarily inactivate the protein. For instance, the control mutation, E105A, preserved activity even though this residue is within 2 Å of the active site. Likewise, Y12A, which is within 10 Å of the active site, retains significantly greater activity than any of the template mutations (Fig. 4F). These control mutations were picked based on their ET ranks of evolutionary importance, which put E105 in the bottom 5% of importance and Y12 in the bottom 20%. These data show that the template residues are necessary for activity, unlike nearby residues with poor ET rank.

ETA exploits an important feature of top-ranked ET positions—namely, that although they are important, they are often not invariant. Instead the residues at these positions vary, in direct correlation with evolutionary branches. This fact is revealed by allowing 3D template residues to “wobble,” meaning that they can match to alternative side chains if they appear recurrently in the multiple sequence alignment. This wobble is therefore different from case to case, and in practice, it increases sensitivity from 62% to 70% in our benchmarks (Fig. S4). Interestingly, two of the three proteins that matched protein x had such wobbles. Whereas the native protein x template contains W73, the matched cognate structural position was a phenylalanine in both EstE1 and AFEST. Because all of these proteins have similar substrate specificity, we hypothesize that regardless of this tryptophan to phenylalanine wobble, these positions fulfill identical mechanistic roles.

To confirm that the wobble does not fundamentally alter function, we introduced a W73F mutation into protein x. The specific activity of the mutant was reduced (229 U/mg), but still remained larger than the activity of the Y12A control mutation (198 U/mg), and it was much more active than a control W73A mutation (1.9 U/mg) (Fig. 4F). These data show that position 73 is critical to function and yet tolerates a phenylalanine substitution remarkably well despite dramatic sequence differences between protein x, EstE1, and AFEST.

Discussion

The key findings of this work are that ETA can now reliably predict substrate-level specificity and the experimental demonstration of this fact on two previously unannotated proteins. The approach relies on narrowly identifying molecular similarity between proteins in the immediate vicinity of their most evolutionarily important site, which can potentially include both catalytic and noncatalytic residues. The role of catalytic residues is well understood and has motivated several other template-based annotations methods (16, 62, 63) and spurred many approaches to annotate structural genomics structures based on local structural similarities (14–16, 23, 24, 62–67). However, because the fundamental catalytic mechanism may be invariant between proteins that operate on varied substrates (68), the challenge has been to add noncatalytic residues that constrain the catalytic mechanism to a more specific substrate (24, 69). The selection of which noncatalytic residues in the vicinity of an active site to include in a template is not trivial, however because not all residues near a catalytic site are themselves important.

To solve this problem, ETA relies on evolution and on structure to pick the most evolutionarily important residues that also cluster tightly at or near the protein surface. It has been previously shown that evolutionarily important residues are necessary for function (31–33, 70, 71), map functional sites (71, 72), guide a variety of targeted mutational experiments that efficiently block function (73, 74), separate functions in multifunctional proteins (38, 75), or recode function by swapping amino acids between functionally distinct homologs (34, 36). For these reasons, these residues are good candidates from which to pick templates.

As a result, the ETA annotation pipeline reliably predicted both the activity and the substrate of an enzyme structure. In retrospective controls, ETA accuracy was consistently high (96%) even when the annotations were based on matches to uncharacterized proteins (14, 25, 26, 46) to validate unique ETA predictions of function. Additionally, mutagenesis studies verified the essential role played by the noncatalytic residues that were selected to be in the 3D template. Thus, all template residues, catalytic or noncatalytic, are necessary for complete enzyme activity toward its substrate. Our data also show that evolutionary wobble substitutions enhance computational specificity.

Taken together, this work suggests that ETA 3D templates capture the essential local elements of the enzyme structure, dynamics, and chemical activity that combine to define substrate-specific mechanisms. A central finding is the essential role that noncatalytic 3D template residues play in specifying substrates. As a result, this study validates a high-throughput method that combines evolution and structure to identify the activity and
substrate specificity of unique enzymes, with applications to enzyme annotations on a structural genomics scale.

An ETA webserver is available that makes 4 E.C. digit predictions using five or six residue templates: http://mammoth.bcm.tmc.edu/AminErdinetalPNAS/eta. This webserver also integrates the newly described confidence scoring. We are actively maintaining this site, and several updates to enhance prediction power and protein visualization are currently in production.

Materials and Methods

Evolutionary Trace Annotation Pipeline. ETA is made of five modules that separately operate to suggest a predicted function for a given query protein. In the first module, ET algorithm (31, 32) assigns evolutionary importance ranks to the query protein’s residues based on correlation of branches in the phylogenetic tree, with the variations in the multiple sequence alignment generated by ClustalX (77) for homologs identified for the sequence of query protein by BLAST (78). The second module is the Template Picker algorithm (26), which selects five or six best-ranked residue positions near the center of a cluster of at least 11 evolutionarily important residues on the protein surface where solvent accessibility of the residues is greater than 2 Å2, as calculated by DSSP algorithm (79). Each position is geometrically represented by the 3D Cartesian coordinates of the selected residue’s alpha carbon atoms. ETA uses both native and similarity-based (query structure) templates in the query space (native templates) and a combination of variations that were observed at least twice in the multiple sequence alignment (variations) for the identified positions (40). Further, the paired-distance-algorithm (40) searches the query template against a “target” library of proteins with known functions for geometric similarity and, in doing so, identifies geometric matches in which residues are matched with those in the query template and surface criterion that each pair of residues in the template and the matched region are within a distance of 2.5 Å. In the fourth module, found matches are passed to the support vector machine (SVM), which identifies significant matches based on geometric and evolutionary similarities. ETA uses a support SVM trained for a set of 53 enzymes based on six-residue templates (40) using SPIDER package (www.kyb.tuebingen.mpg.de/bjpeople/spider) for MATLAB with a radial basis function kernel and α = 0.5. SVM feature space is seven-dimensional, with one dimension representing rmsd between the query template residues and found matched site in the target structure, whereas six other dimensions are for ET percentile rank differences of the query six-residue templates and the residues they are linked with in the target structures. In the case of the five-residue ETA, we use the same six-residue SVM by constructing a virtual sixth residue as the average of the other five positions for the five-residue data. ETA repeats all these steps reciprocally, generating templates from target structures and searching for significant matches on the “query” protein. In the last module, ETA suggests the most seen function as a prediction among a set of identified reciprocal significant matches.

Confidence Score. In developing a confidence score, we considered the following two criteria: First, confidence should be high if the number of matches that ETA identifies for a given enzyme set is large, and second, confidence should be low if the number of matches with differing EC numbers is large. A simple model that holds these criteria is defined to be $n_{op}/n_{tm}$, where $n_{op}$ is the number of matches with winning function, $n_{n}$ is the number of identified distinct function, and $n_{tm}$ is the number of matches.

Negative Templates. To generate negative templates, we started with the surface residue (solvent accessible area greater than 2 Å2 as calculated by the DSSP algorithm) (79) with the poorest EC percentile rank. Next we selected the neighbor surface residue with the poorest EC percentile rank that lies between 70% and 100% ET percentile to the first selected one within a distance of 8 Å. We further selected the surface residues iteratively with the poorest percentile rank that lies between 70% and 100% ET percentile within the distance of 8 Å with the center of mass of chosen residues in the previous iteration.

Comparison of Template, Negative Template, and Global Structure Alignments. Alignment data are retrieved from 1,157 pairs of matches in which query and target protein functions are in full agreement for 451 proteins. We used TM-align (80) and lovoalign (81) structural alignment methods for alpha-carbon rmsd and all-atom rmsd, respectively. To calculate template alignment rmsd, we aligned query templates with their corresponding match sites that are identified by ETA through TM-align and lovoalign. Again, TM-align and lovoalign were used to align the whole structures of query and target proteins in reciprocal match pairs. We applied the negative templates generated as described above from the query proteins to the previously found reciprocal match partners. Among those, ETA’s paired-distance matching algorithm identified a matched site for 445 proteins, with 1,127 matches in total. In generation of Fig. 2B, these pairs of matches were used for comparison of alignments at levels of templates, clusters, and global structures.

ETA and TM-Align Comparison for Annotation Performance. We compared performance of ETA and TM-align (80) on a set of 430 protein structures where ETA made correct or incorrect predictions with six-residue templates. TM-align annotation was made by identifying the lowest rmsd hit upon aligning query proteins with the protein structures in the target set. Next, we grouped ETA and TM-align predictions according to the maximum sequence identity of reciprocal matches found by ETA for 430 proteins.

ETA and COFACTOR Comparison for Annotation Performance. We specifically selected test proteins that were matched with high (>$1$ or medium ($<$)= 0.9) confidence to proteins with less than 30% sequence identity for the comparison ($n$ = 137). All proteins were submitted to the COFACTOR webserver (http://zhanglab.ccb.med.umich.edu/COFACTOR/) between 8/5/13 and 8/14/13. Due to the fact that the COFACTOR webserver has a larger target enzyme set than the one used in this publication, many of the query proteins were matched to homologs of >30% sequence identity. Therefore, both the top hit (highest Cscore) and the top hit with less than 30% sequence identity (defined as IDEN*Cov. < 30%) were recorded. For the library size comparison, only the four-digit PDB codes were used; chains were not considered. For example, 1JID and 1JIC would be considered the same protein.

Propensity Calculation. Propensity of a particular residue is defined to be the ratio of its frequency among all template residues for the protein structures in the experiment to its frequency within all residues for the protein structures (82). To get better insight, we used log propensity, which is defined to be the logarithm of the calculated ratio. Log propensity becomes positive for the values with a ratio greater than 1, whereas it becomes negative when the ratio is less than 1. MACIE’s log propensities were retrieved from MACIE database version 3.0 (www.ebi.ac.uk/thornton-srv/databases/MACIE/) (46):

$$\text{Pearson correlation coefficient: } r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where $n$ is the total number of measurements, $x_i$ is the $i$th measurement, and $\bar{x} = \frac{\sum_{i=1}^{n}x_i}{n}$.

Performance Measures. Positive predictive value (PPV) = TP/(TP + FP), where TP and FP stand for the numbers of True Positives and False Positives, respectively. True Positive is defined for cases where the prediction fully agrees with the known function at the four-digit EC level. False Positive is the case in which the predicted EC number for a protein does not agree fully with its known EC number. Sensitivity = TP/(TP + FN), where FN denotes False Negatives, which is defined for the cases for which ETA does not make predictions.

Datasets. In selecting proteins for the benchmark test set, we started with 837 Structural Genomics enzymes with full EC annotations in the SwissProt database (83). The structures were retrieved from PDB (84) as of October 2010, and any pair of enzymes has at most 90% sequence identity with one another. One distinct feature of those proteins was that they had a varying truncation ratio, which was defined to be the ratio of number of amino acids in the protein structure to the actual sequence length. This ratio quantifies how much the structure represents the whole protein sequence. Therefore, a ratio of less than 1 means that the whole sequence is not represented in the structure. The cases with such ratios might pose a potential problem in automated computational annotation efforts, as these structures might lack the regions essential for function. This motivated us to assess how ETA performed on those cases. To this end, we calculated the prediction coverage of query proteins in a given bin of truncation ratios (Fig. S1). The prediction coverage reaches its maximum values (>71%) when the ratio is >0.9, whereas the PPV does not show a clear pattern as the prediction coverage below this threshold. Therefore, we used 605 protein structures with a truncation ratio greater than 0.9 as a benchmark test set.

The target set contains 3,082 protein structures with a truncation ratio greater than 0.95 and full EC annotation from 2008PDB90, among which any pair has at most 90% sequence identity with one another. EC annotations of these proteins are retrieved from PDB (317), SwissProt (2454), and Trembl (311) databases, where the numbers inside the parentheses denote the number of protein structures with the associated annotation source.
Determination of Enzymatic Activity. Protein concentration of the supernatant was then determined using the buster Mastermix (Novagen). Lysates were kept shaking at room temperature overnight at 18°C. Tseng YY, Li WH (2012) PSC: Protein surface classification. J Biomol Struct Dyn 29(6):738–748.

Cloning and Expression. Each uncharacterized gene was PCR amplified from vectors received from a structural genomics center using the primers identified in Table S1. The resulting PCR products contained NotI and SalI restriction sites, which were used to insert the genes into the pet28a vector, resulting in an N-terminal His tag. The vectors were electroentrapped into BL21D cells for protein expression. Protein expression was carried out as previously described (87).

Initial enzymatic assays were carried out using crude lysate. Crude lysates were produced by spinning down 15 mL of induced culture, freezing the pellets overnight at −80°C, and resuspending the pellets in 1 mL of bug-buster Mastermix (Novagen). Lysates were kept shaking at room temperature for 20 min, followed by centrifugation to remove the insoluble debris. Protein concentration of the supernatant was then determined using the MicroBCA kit (Thermo).

Determination of Enzymatic Activity. Activity of myo-inositol dehydrogenase was determined by adding crude lysate to a mixture of 25 mM myo-inositol, 10 mM sodium pyrophosphate (pH 9.0), and 0.5 mM j-nicotinamide adenine dinucleotide and monitoring absorbance at 340 nm (88). Activity of the putative aspartate aminotransferase was determined by adding crude lysate to a mixture of 50 mM L-aspartate, 5 mM alpha-ketoglutarate, 100 mM Tris-HCl (pH 8.0), 0.1 mM NADH, and two units of malate dehydrogenase (Sigma) and monitoring absorbance at 340 nm. All reactions were initiated following the addition of the substrate (myo-inositol or L-aspartate) and mixed vigorously. For both reactions, one unit of activity is defined as the conversion of 1.0 μmol of NAD/NADH per minute using 6.22 as the millimolar extinction coefficient (89). P values were obtained using Student t-test.

Carboxylesterase Purification, Activity Measurement, and Mutation. Purification and measurement of activity was conducted as previously defined (87). All 4-nitrophenyl substrates were dissolved in 100% DMSO at a stock solution of 100 mM. The final reaction volumes consisted of 50 mM 2-(N-Morpholino)ethanesulfonic acid, pH 6.0, 3% (vol/vol) DMSO, and 1 mM 4-nitrophenyl substrate and enzyme, either as crude lysate or purified. The reaction was initiated upon addition of the substrate and monitored at 405 nm using an Amersham Ultraspec 3100pro spectrophotometer. All mutations were carried out using the Quickchange II Site-Directed Mutagenesis Kit (Agilent) using primers found in Table S1.

ACKNOWLEDGMENTS. We thank the following University of Saint Thomas students for bioinformatics Tracit support: Anthony Pham, Anthony Su, and Medina Batimairo. We gratefully acknowledge grant support from the National Institutes of Health (GM079565, GM066999, and K12 GM084987) and from the National Science Foundation (CCF-0905336 and DBI-1062455).


**Supporting Information**

Amin et al. 10.1073/pnas.1305162110

---

**Fig. S1.** ET Annotation pipeline (ETA) accuracy was benchmarked with a diverse dataset. (A) Distribution of protein structures in query and target sets according to their first digit Enzyme Commission (EC) classes. (B) Benchmark results of ETA with six-residue and five-residue templates according to confidence scores of predictions. Black bars correspond to correct predictions, gray bars are incorrect predictions, and the triangles represent the positive predictive values. (C) Comparison of ETA results using 248 enzymes in both Apo and Holo conformations. Matching was conducted using six-residue templates from both conformations. The majority of cases resulted in the selection of nonidentical templates that still give a reciprocal match between the Apo and Holo structures.
Fig. S2. Template residues of 3db2 and their Evolutionary Tracing (ET) percentile rank scores. At the bottom, reciprocal matches with their sequence identity with 3db2 and EC number are given.

![Evolutionary Tracing (ET) percentile rank scores](image)

**ETA SUGGESTS** INOSITOL 2-DEHYDROGENASE ACTIVITY (EC 1.1.1.18) BASED ON MATCHES

- 3CEA (EC 1.1.1.18) 21% sequence identity
- 3CE7 (EC 1.1.1.18) 18% sequence identity
- 3EUW (EC 1.1.1.18) 17% sequence identity

Fig. S3. A total of 10 μg of purified protein from wild-type, template mutants, and control mutants for tm1040_2492. The second band in the mutant preparations is a cleavage product, determined by mass spectrometry.

![Purification Bands](image)

Fig. S4. Reciprocal ETA performance based on default ETA templates (all), native templates (using residue labels in the query structure), and templates using variation in the multiple sequence alignment (variation).

![ETA Performance Chart](image)

**Fig. S2.** Template residues of 3db2 and their Evolutionary Tracing (ET) percentile rank scores. At the bottom, reciprocal matches with their sequence identity with 3db2 and EC number are given.

**Fig. S3.** A total of 10 μg of purified protein from wild-type, template mutants, and control mutants for tm1040_2492. The second band in the mutant preparations is a cleavage product, determined by mass spectrometry.

**Fig. S4.** Reciprocal ETA performance based on default ETA templates (all), native templates (using residue labels in the query structure), and templates using variation in the multiple sequence alignment (variation).
### Table S1. Comparison of ETA and cofactor for 137 test proteins matched by ETA to targets of <30% sequence identity

<table>
<thead>
<tr>
<th>Method</th>
<th>Total proteins</th>
<th>Unique to dataset</th>
<th>Found in both datasets</th>
<th>Accuracy at ETA high confidence predictions, $n = 47$</th>
<th>Accuracy at ETA medium confidence predictions, $n = 90$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cofactor Webserver, Enzyme Library</td>
<td>7,902</td>
<td>6,166</td>
<td>1,736</td>
<td>91.49%</td>
<td>76.60%</td>
</tr>
<tr>
<td>ETA target set in this publication</td>
<td>2,967</td>
<td>1,231</td>
<td>1,736</td>
<td>—</td>
<td>95.74%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>70.00%</td>
</tr>
</tbody>
</table>

**Movie S1.** An example of how ETA uses structure and sequence data to identify protein function.