Appeared in Protein Science 15(6):1530–1536, 2006

(Manuscript accepted to special issue of Protein Science relating to function annotation work shown at the 2005 ISMB conference)

Recurrent Use of Evolutionary Importance for Functional Annotation of Proteins Based on Local Structural Similarity

David M. Kristensen[1,2], Brian Y. Chen[3], Viacheslav Y. Fofanov[4], R. Matthew Ward[1,2], Andreas Martin Lisewski[1], Marek Kimmel[4], Lydia E. Kavraki[2,3,5], and Olivier Lichtarge[1,2*]

[1]Department of Molecular and Human Genetics, [2]Graduate Program in Structural and Computational Biology and Molecular Biophysics, Baylor College of Medicine, Houston, TX 77030, [3]Department of Computer Science, [4]Department of Statistics, [5]Department of Bioengineering, Rice University, Houston, Texas 77030

[*]Corresponding author: Olivier Lichtarge, M.D., Ph.D., Department of Human and Molecular Genetics, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, Phone: (713) 798-5646, Fax: (713) 798-5386, e-mail: lichtarge@bcm.tmc.edu

Abstract

The annotation of protein function has not kept pace with the exponential growth of raw sequence and structure data. An emerging solution to this problem is to identify 3D motifs or templates in protein structures that are necessary and sufficient determinants of function. Here, we demonstrate the recurrent use of Evolutionary Trace information to construct such 3D templates for enzymes, search for them in other structures, and distinguish true from spurious matches. Serine protease templates built from evolutionarily important residues distinguish between proteases and other proteins nearly as well as the classic Ser-His-Asp catalytic triad. In 53 enzymes spanning 33 distinct functions, an automated pipeline identifies functionally related proteins with an average positive predictive power of 62%, including correct matches to proteins with the same function but with low sequence identity (the average identity for some templates is only 17%). Although these template building, searching, and match classification strategies are not yet optimized, their sequential implementation demonstrates a functional annotation pipeline which does not require experimental information, but only local molecular mimicry among a small number of evolutionarily important residues.

Keywords: function prediction; proteome annotation; evolution; structural genomics; structural motif

Introduction

By August 2005, the NCBI Entrez Genome Project contained 273 fully sequenced genomes yielding almost 1.7 million putative protein sequences in NCBI's RefSeq database. However, up to 40% of these genes still lacked any annotation of biological function (Pruitt et al. 2005), thus illustrating the importance of reliable methods to identify protein function.

To address this problem, broad categories of computational methods for functional annotation have emerged that rely on either sequence or structure, considered whole or through motifs. Whole sequence

methods can fail when homologs develop unrelated functions, distinct chemistries, or different functional sites as sequence identity falls below 40% (Olmea and Valencia 1997; Russell et al. 1998; Todd et al. 2001). Local sequence motifs, however, cannot adequately capture functions distributed over non-adjacent stretches of primary structure. These limitations motivated the extension of the concept of functional motifs from sequence to structure (Wallace et al. 1996; Wallace et al. 1997; Russell 1998; Kleywegt 1999; Bartlett et al. 2002; Barker and Thornton 2003; Stark et al. 2003; Ivanisenko et al. 2004; Porter et al. 2004; Shulman-Peleg et al. 2004; Ausiello et al. 2005; Ivanisenko et al. 2005; Torrance et al. 2005).

The rationale for structural motifs ("3D templates") is that typically, just a few key residues directly mediate catalysis or binding. These same residues in the same conformation should therefore be reasonably expected to carry out the same function even in a different fold unless long-range effects impact their biophysical behavior.

Many methods aim to derive 3D templates and match them to protein structures. Some map sequence motifs onto structures (Kasuya and Thornton 1999; Liang et al. 2003); others compare enzymes with known functional sites against a structural database (Fischer et al. 1994) or against each other (Wallace et al. 1997) (de Rinaldis et al. 1998) (Torrance et al. 2005) (Laskowski et al. 2005). However, fundamental difficulties remain. First, 3D templates that rely on experimental data are limited by the availability of such information. Second, while the search for 3D matches to small templates (3-4 residues) is not computationally expensive, this quickly changes for larger motifs that include amino acid substitutions when searched against the full Protein Data Bank (PDB) (Berman et al. 2000). Third, although sequence methods such as BLAST (Altschul et al. 1990) and PSI-BLAST (Altschul et al. 1997) can confidently claim to find sequence homologs and suggest—but do not prove—functional similarity between proteins, it is not yet clear what degree of functional similarity can be inferred from a structural match.

With these issues in mind, we present an evolution-directed series of algorithms, which in the absence of experimental data, aim to identify relevant 3D templates, to guide an efficient search for molecular mimicry in other protein structures, and finally to isolate from among all matches a subset that is highly enriched in proteins that perform the same function. Together these represent the first steps towards an automated functional annotation pipeline for proteins that can complement experimentally-driven annotation efforts.

To measure the evolutionary importance of each protein residue, we use the Evolutionary Trace (ET) method (Lichtarge et al. 1996b). ET ranks residue importance by correlating amino acid variations in a multiple sequence alignment with evolutionary divergences in a phylogenetic tree. The quality of the analysis is measured by the extent to which top-ranked (trace) residues cluster in the structure (Madabushi et al. 2002); (Mihalek et al. 2003). Remarkably, these clusters match functional sites (Madabushi et al. 2002; Yao et al. 2003) precisely enough to guide rational protein engineering (Lichtarge et al. 1996a; Lichtarge et al. 1997; Landgraf et al. 1999; Pritchard and Dufton 1999; Innis et al. 2000; Pascual et al. 2000; Sowa et al. 2000; Sowa et al. 2001; Imanishi et al. 2002; Lichtarge et al. 2003; Madabushi et al. 2004; Raviscioni et al. 2005). These data suggest that top-ranked trace residues represent the key determinants of protein function. It is

therefore logical to use ET ranks to design 3D templates, to prioritize matching of residues by their importance, and then again to interpret their matches.

Results

The key steps of the pipeline (see Methods) are: the identification of evolutionarily important residues in the protein of interest (the query); the selection of some of these residues to construct a 3D template; the search in other structures (targets) for matches based on residue type and geometry (Chen et al. 2005; Chen et al. 2006); the assessment of the significance of a match based on its least-root-mean-square deviation (LRMSD) from the template and finally, a selection of the most biologically relevant matches based on the evolutionary importance of the matched target residues.

An underlying hypothesis is that templates built using ET rank information can be useful in cases where the functional site of a protein has not been determined by experimental methods. Accordingly, we start with a comparison in serine proteases of a 3D template (positive control) composed of the well-known Ser195-His57-Asp102 "catalytic triad" (Wallace et al. 1996)—the gold standard for proteolytic activity—with two neighboring but non-overlapping templates: one composed of highly ranked residues (the test template), and the other of poorly ranked residues (the negative control). Figure 1 shows the distribution of matches of these templates against the PDB, with vertical lines marking the points at which p-value=1% (solid green) and 5% (dashed purple). Matches to the catalytic triad template shown in Figure 1A (geometric positions for the template are obtained from bovine chemotrypsin, PDB code 1acb), exhibit a bimodal distribution in which the left LRMSD peak is smaller but rich in proteases (312 true positives shown as red bars) and the right LRMSD peak is larger but contains mostly functionally unrelated proteins (blue bars). The separation between the two modes shows that LRMSD from the template acts as a good discriminator of function (Wallace et al. 1996).

Remarkably, Figure 1B shows that a template of non-catalytic but highly ranked neighboring residues separates these two peaks nearly as well. These residues were chosen because they are near the catalytic triad and are ranked within the top 5%, i.e., among the 12 most important residues in this 245-residue protein. Unlike the triad, however, this "non-catalytic quartet" contains fold-specific residues: Cys42 and Cys58 form a disulfide bond, Asp194 is involved in a salt bridge, and Ser214 is implicated in ligand binding. As a result, it does not find matches to proteases with different folds, although it is able to find matches to proteins with less than 30% sequence identity. In comparison, the negative control template of poorly ranked neighboring residues shown in Figure 1C cannot distinguish at all between proteases and other proteins. This suggests that structural templates built from evolutionarily important residues will be useful, particularly when experimental data on functional residues is not available.

To confirm this hypothesis, we systematically selected high-ranking residues in a test set of 53 enzymes spanning 36 folds and 33 distinct functions and searched the PDB for matches to each template for which the LRMSD has a p-value ≤ 1%. We examine only enzymes because the Enzyme Nomenclature provides an easy and reliable way to define each protein's exact function (see Methods). Figure 2A displays the distribution of these 1% significant matches over all 53 proteins as a function of LRMSD. As with the match

distribution of the "catalytic triad" and the "non-catalytic quartet" above, true positive matches (red) generally have lower LRMSDs, suggesting that as before, ET rank allows us to build discriminating templates. Also as before, matches frequently occur to proteins with the same function but low sequence identity to the source (the average identity for some templates is only 17%). However, unlike the previous examples, there are many false positives (blue) even at the 1% p-value threshold. Furthermore, the true and false hits are not well separated in the LRMSD dimension, which suggests that there is no universal LRMSD threshold to separate true from false geometric matches.

In order to better separate true and false geometric matches we focus next on the evolutionary importance of the matched residues. The hypothesis is that a spurious match is less likely to occur at evolutionarily important residues than a true match. Indeed, Figure 2B demonstrates the strikingly different evolutionary importance of matched sites (the average ET rank of the matched residues) between functionally related (red) and unrelated (blue) proteins. Thus, consideration of the evolutionary importance in the target protein should let us separate true from random matches and thereby improve the positive predictive power of our templates.

We formalized and tested this observation with a support vector machine (SVM) trained to classify matches as true or false using the average ET rank of the target residues and/or LRMSD. Table 1 shows that an SVM based only on ET rank identifies 554 of 570 true positives (97% sensitivity), and 4,959 out of 5,450 true negatives (91% specificity). Its overall accuracy is 92% and its positive and negative predictive powers are 53% and 99%, respectively. In contrast, an SVM that uses only the LRMSD feature has a reduced accuracy of 85% and positive predictive power of 37%. An SVM that uses both ET rank and LRMSD yields the best performance, with 94% accuracy and 61% positive predictive power. Thus, most of the discriminatory power of this classifier comes from ET rank, with some complementary information arising from LRMSD.

Since we ultimately wish to predict protein function, we must test the classifier on proteins it has not been trained on. This was done through leave-one-out cross-validation experiments. For each of the 33 enzyme classes in the test set, we trained an SVM on the other 32 classes, and then tested performance on the left-out class. Table 1 shows that, overall, performance changes by less than 1%, comparing All vs. average Cross-validation results for each of the metrics used. While the standard deviation for most of the metrics is on the order of 1-10%, it reaches as high as 39% for positive predictive power. These results indicate that, while template performance is not uniform across all enzyme classes, this classifier is not highly dependent on the proteins in this dataset, and therefore should work for other proteins as well.

We can now revisit the serine protease example using the annotation pipeline from beginning to end. The 5-residue template chosen by this automated method partially overlaps the catalytic triad (Ser195) and the test template (Cys42, Cys58, and Ser214), since these residues are highly ranked by ET. The distribution of matches, as before, is shown in Figure 1D. Compared to the catalytic triad, the positive and negative predictive powers both decrease by only 1%, yielding a positive and negative predictive power of 93%. The similarity of these numbers is remarkable because no experimental data about the active site was used to build the new template except for

that inferred from evolution and structure. This example suggests that this approach will be useful in enzymes (and non-enzymes) whose functional mechanisms are unclear.

Discussion

We linked algorithms that exploit evolutionary information in different ways towards the creation of an automated functional annotation pipeline. First, 3D templates are built from residues that are top-ranked by ET, cluster in the protein structure, and are solvent accessible. This choice follows from past studies that consistently show that top-ranked trace residues are key functional determinants. Indeed, in serine proteases, several templates built from top-ranked residues can distinguish serine proteases from other proteins nearly as well as the catalytic triad itself.

Second, the structural matching algorithm exploits ET rank to prioritize its search. Rather than perform a geometric search for the entire template in a single step (which would be very computationally expensive due to the amino acid labels), MA performs a fast search for the three most important trace residues (the "seed") and then iteratively expands matches to template residues of lesser rank. This method is fast enough to search the entire PDB and generate a non-parametric estimate of each p-value for any LRMSD.

At that point in the pipeline, however, many false matches are found. Even at a 1% significance LRMSD threshold the average positive predictive power of the 53 3D templates is only 14%, and indeed Figures 2A and 2B display more false matches than true ones. This may reflect template limitations such as the choice of residues; the choice of points for geometric representation of those residues (C-alpha atoms); the choice of size (five residues); and, unlike most other template search algorithms, the allowance for amino acid substitutions as they occur in the query's multiple sequence alignment. We note that this finding of many functionally unrelated geometric matches is in keeping with other studies (Laskowski et al. 2005; Torrance et al. 2005).

For this reason, additional separation of the biologically relevant matches is imperative. This is accomplished, again, through evolutionary importance—but this time in the matches themselves. Used in this novel way, ET rank proves a powerful and robust discriminator that separates true from false geometric matches with 92% accuracy alone, and 94% with LRMSD added. The average positive predictive power of all templates after using this classifier is 63% – a 4.5-fold improvement from the 14% seen without its use.

Future improvements may arise from better template design, from the inclusion of biophysical features in the SVM classifier, and from larger scale studies with broader scope (including non-enzymes). For now, these results show that the recurrent use of evolutionary information in the form of ET rank is a novel and useful approach for the functional annotation of protein structures based on local molecular mimicry among a small number of evolutionarily important residues.

Methods

Test set

The test set consists of 53 proteins with 36 folds and 33 unique functions. These proteins were chosen from the PDB-SELECT-25 (Hobohm and Sander 1994) and thus each has less than 25% sequence

identity to all the others, including those with the same function. A complete description is available in supplementary materials.

Template creation

ET analyses were performed using an automated (Yao et al. 2003), real-valued (Mihalek et al. 2004) version of the ET algorithm (Lichtarge et al. 1996b). For each protein, template residues were chosen as the 5 top-ranked residues for which the largest trace cluster contained at least 10 surface residues, defined by DSSP solvent-accessibility values ≥ 2 (Kabsch and Sander 1983). The 5 top-ranked surface residues in that cluster were chosen to make the 3D template, representing each by the geometric coordinates of its C-alpha atom and labeled by ET rank and allowed amino acid substitutions (those appearing at least twice in the corresponding column of the multiple sequence alignment used for ET).

Matches

The Match Augmentation (MA) algorithm is described elsewhere (Chen et al. 2005; Chen et al. 2006). In brief, MA matches a query template to a target structure in two-stages: Seed Matching identifies several low LRMSD matches for the template's three best-ranked residues; Augmentation then iteratively adds template residues in order of their ET rank. The output is the lowest LRMSD match, or none if all LRMSDs exceed 4Å. MA can match a typical template to the entire PDB in ~40 min on a single processor. We then compute the statistical significance (p-value) of a match using a nonparametric density estimate of the distribution of match LRMSDs to all protein chains in the PDB (Chen et al. 2005; Chen et al. 2006).

For this study, matches were searched against 13,600 chains from the PDB. This representative subset is redundant at the protein level, but includes only a single chain in cases where multiple structures are available due to crystallographic symmetry. Mutants, ionically perturbed structures, and small peptide fragments were manually removed, although structures bound to inhibitors were retained.

Evaluation of Matches

Throughout the paper, Enzyme Nomenclature (EC) (NC-IUBMB 1992) annotations are those reported in the PDB. A true match means exact agreement of all 4 digits of the hierarchical EC number. In all, 5,200 proteins (38%) have full, unambiguous EC annotation while 7,900 (58%) have none, although this number may include some unannotated enzymes. Only 500 proteins (<4%) have partial or ambiguous EC annotation (such as large proteins performing multiple functions). As there were only 248 matches to these proteins (<2%), these were discarded.

Machine learning

We traced every matched protein and averaged the percentile ET rank of its matched residues. This average ET rank, the LRMSD of a match, or both were used to train either a 1- or 2-dimensional support vector machine (SVM) using the Spider package for MATLAB (see http://www.kyb.tuebingen.mpg.de/bs/people/spider). Default parameters were used with a linear kernel and a balanced ridge calculated as the difference between the proportions of the two classes.

Acknowledgements

We wish to gratefully acknowledge support from the NSF DBI-0318415, MRI-0420984 (OL & LEK). Work on this paper by OL has also been supported in part by the March of Dimes FY03-93 and the NIH GM066099. Work on this paper by LEK has also been supported in part by a Sloan Fellowship and the Brown School of Engineering at Rice University. This work was also supported by training fellowships from the Keck Center for Interdisciplinary Bioscience Training from the W.M. Keck Foundation (AML) and NLM Grant No. 5T15LM07093 (RMW, DMK, BYC), and from the VIGRE Training in Bioinformatics Grant NSF DMS 0240058 (VYF).

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tool. *J Mol Biol* **215:** 403-410.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25:** 3389-3402.

Ausiello, G., Zanzoni, A., Peluso, D., Via, A., and Helmer-Citterich, M. 2005. pdbFun: mass selection and fast comparison of annotated PDB residues. *Nucleic Acids Res* **33:** W133-137.

Barker, J.A., and Thornton, J.M. 2003. An algorithm for constraint-based structural template matching: application to 3D templates with statistical analysis. *Bioinformatics* **19:** 1644-1649.

Bartlett, G.J., Porter, C.T., Borkakoti, N., and Thornton, J.M. 2002. Analysis of catalytic residues in enzyme active sites. *J Mol Biol* **324:** 105-121.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. 2000. The Protein Data Bank. *Nucleic Acids Res* **28:** 235-242.

Chen, B.Y., Fofanov, V.Y., Bryant, D.H., Dodson, B.D., Kristensen, D.M., Lisewski, A.M., Kimmel, M., Lichtarge, O., and Kavraki, L.E. 2006. Geometric Sieving: Automated Distributed Optimization of 3D Motifs for Protein Function Prediction. In *Proceedings of The Tenth Annual International Conference on Computational Molecular Biology (RECOMB 2006)*, Venice, Italy.

Chen, B.Y., Fofanov, V.Y., Kristensen, D.M., Kimmel, M., Lichtarge, O., and Kavraki, L.E. 2005. Algorithms for structural comparison and statistical analysis of 3D protein motifs. *Pac Symp Biocomput***:** 334-345.

de Rinaldis, M., Ausiello, G., Cesareni, G., and Helmer-Citterich, M. 1998. Three-dimensional profiles: a new tool to identify protein surface similarities. *J Mol Biol* **284:** 1211-1221.

DeLano, W.L. 2002. *The PyMOL Molecular Graphics System*. DeLano Scientific, San Carlos, CA, USA.

Fischer, D., Wolfson, H., Lin, S.L., and Nussinov, R. 1994. Three-dimensional, sequence order-independent structural comparison of a serine protease against the crystallographic database reveals active site similarities: potential implications to evolution and to protein folding. *Protein Sci* **3:** 769-778.

Hobohm, U., and Sander, C. 1994. Enlarged representative set of protein structures. *Protein Sci* **3:** 522-524.

Imanishi, Y., Li, N., Sokal, I., Sowa, M.E., Lichtarge, O., Wensel, T.G., Saperstein, D.A., Baehr, W., and Palczewski, K. 2002. Characterization of retinal guanylate cyclase-activating protein 3 (GCAP3) from zebrafish to man. *Eur J Neurosci* **15:** 63-78.

Innis, C.A., Shi, J., and Blundell, T.L. 2000. Evolutionary trace analysis of TGF-beta and related growth factors: implications for site-directed mutagenesis. *Protein Eng* **13:** 839-847.

Ivanisenko, V.A., Pintus, S.S., Grigorovich, D.A., and Kolchanov, N.A. 2004. PDBSiteScan: a program for searching for active, binding and posttranslational modification sites in the 3D structures of proteins. *Nucleic Acids Res* **32:** W549-554.

Ivanisenko, V.A., Pintus, S.S., Grigorovich, D.A., and Kolchanov, N.A. 2005. PDBSite: a database of the 3D structure of protein functional sites. *Nucleic Acids Res* **33:** D183-187.

Kabsch, W., and Sander, C. 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22:** 2577-2637.

Kasuya, A., and Thornton, J.M. 1999. Three-dimensional structure analysis of PROSITE patterns. *J Mol Biol* **286:** 1673-1691.

Kleywegt, G.J. 1999. Recognition of spatial motifs in protein structures. *J Mol Biol* **285:** 1887-1897.

Landgraf, R., Fischer, D., and Eisenberg, D. 1999. Analysis of heregulin symmetry by weighted evolutionary tracing. *Protein Eng* **12:** 943-951.

Laskowski, R.A., Watson, J.D., and Thornton, J.M. 2005. Protein function prediction using local 3D templates. *J Mol Biol* **351:** 614-626.

Liang, M.P., Brutlag, D.L., and Altman, R.B. 2003. Automated construction of structural motifs for predicting functional sites on protein structures. *Pac Symp Biocomput***:** 204-215.

Lichtarge, O., Bourne, H.R., and Cohen, F.E. 1996a. Evolutionarily conserved Galphabetagamma binding surfaces support a model of the G protein-receptor complex. *Proc Natl Acad Sci U S A* **93:** 7507-7511.

Lichtarge, O., Bourne, H.R., and Cohen, F.E. 1996b. An evolutionary trace method defines binding surfaces common to protein families. *J Mol Biol* **257:** 342-358.

Lichtarge, O., Yamamoto, K.R., and Cohen, F.E. 1997. Identification of functional surfaces of the zinc binding domains of intracellular receptors. *J Mol Biol* **274:** 325-337.

Lichtarge, O., Yao, H., Kristensen, D.M., Madabushi, S., and Mihalek, I. 2003. Accurate and scalable identification of functional sites by evolutionary tracing. *J Struct Funct Genomics* **4:** 159-166.

Madabushi, S., Gross, A.K., Philippi, A., Meng, E.C., Wensel, T.G., and Lichtarge, O. 2004. Evolutionary trace of G protein-coupled receptors reveals clusters of residues that determine global and class-specific functions. *J Biol Chem* **279:** 8126-8132.

Madabushi, S., Yao, H., Marsh, M., Kristensen, D.M., Philippi, A., Sowa, M.E., and Lichtarge, O. 2002. Structural clusters of evolutionary trace residues are statistically significant and common in proteins. *J Mol Biol* **316:** 139-154.

Mihalek, I., Res, I., and Lichtarge, O. 2004. A family of evolution-entropy hybrid methods for ranking protein residues by importance. *J Mol Biol* **336:** 1265-1282.

Mihalek, I., Res, I., Yao, H., and Lichtarge, O. 2003. Combining inference from evolution and geometric probability in protein structure evaluation. *J Mol Biol* **331:** 263-279.

NC-IUBMB. 1992. *Enzyme Nomenclature*. Academic Press, San Diego.

Olmea, O., and Valencia, A. 1997. Improving contact predictions by the combination of correlated mutations and other sources of sequence information. *Fold Des* **2:** S25-32.

Pascual, J., Martinez-Yamout, M., Dyson, H.J., and Wright, P.E. 2000. Structure of the PHD zinc finger from human Williams-Beuren syndrome transcription factor. *J Mol Biol* **304:** 723-729.

Porter, C.T., Bartlett, G.J., and Thornton, J.M. 2004. The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res* **32:** D129-133.

Pritchard, L., and Dufton, M.J. 1999. Evolutionary trace analysis of the Kunitz/BPTI family of proteins: functional divergence may have been based on conformational adjustment. *J Mol Biol* **285:** 1589-1607.

Pruitt, K.D., Tatusova, T., and Maglott, D.R. 2005. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* **33:** D501-504.

Raviscioni, M., Gu, P., Sattar, M., Cooney, A.J., and Lichtarge, O. 2005. Correlated evolutionary pressure at interacting transcription factors and DNA response elements can guide the rational engineering of DNA binding specificity. *J Mol Biol* **350:** 402-415.

Russell, R.B. 1998. Detection of protein three-dimensional side-chain patterns: new examples of convergent evolution. *J Mol Biol* **279:** 1211-1227.

Russell, R.B., Sasieni, P.D., and Sternberg, M.J. 1998. Supersites within superfolds. Binding site similarity in the absence of homology. *J Mol Biol* **282:** 903-918.

Shulman-Peleg, A., Nussinov, R., and Wolfson, H.J. 2004. Recognition of functional sites in protein structures. *J Mol Biol* **339:** 607-633.

Sowa, M.E., He, W., Slep, K.C., Kercher, M.A., Lichtarge, O., and Wensel, T.G. 2001. Prediction and confirmation of a site critical for effector regulation of RGS domain activity. *Nat Struct Biol* **8:** 234-237.

Sowa, M.E., He, W., Wensel, T.G., and Lichtarge, O. 2000. A regulator of G protein signaling interaction surface linked to effector specificity. *Proc Natl Acad Sci U S A* **97:** 1483-1488.

Stark, A., Sunyaev, S., and Russell, R.B. 2003. A model for statistical significance of local

similarities in structure. *J Mol Biol* **326:** 1307-1316.

Todd, A.E., Orengo, C.A., and Thornton, J.M. 2001. Evolution of function in protein superfamilies, from a structural perspective. *J Mol Biol* **307:** 1113-1143.

Torrance, J.W., Bartlett, G.J., Porter, C.T., and Thornton, J.M. 2005. Using a library of structural templates to recognise catalytic sites and explore their evolution in homologous families. *J Mol Biol* **347:** 565-581.

Wallace, A.C., Borkakoti, N., and Thornton, J.M. 1997. TESS: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites. *Protein Sci* **6:** 2308-2323.

Wallace, A.C., Laskowski, R.A., and Thornton, J.M. 1996. Derivation of 3D coordinate templates for searching structural databases: application to Ser-His-Asp catalytic triads in the serine proteinases and lipases. *Protein Sci* **5:** 1001-1013.

Yao, H., Kristensen, D.M., Mihalek, I., Sowa, M.E., Shaw, C., Kimmel, M., Kavraki, L., and Lichtarge, O. 2003. An accurate, sensitive, and scalable method to identify functional sites in protein structures. *J Mol Biol* **326:** 255-261.

Figures

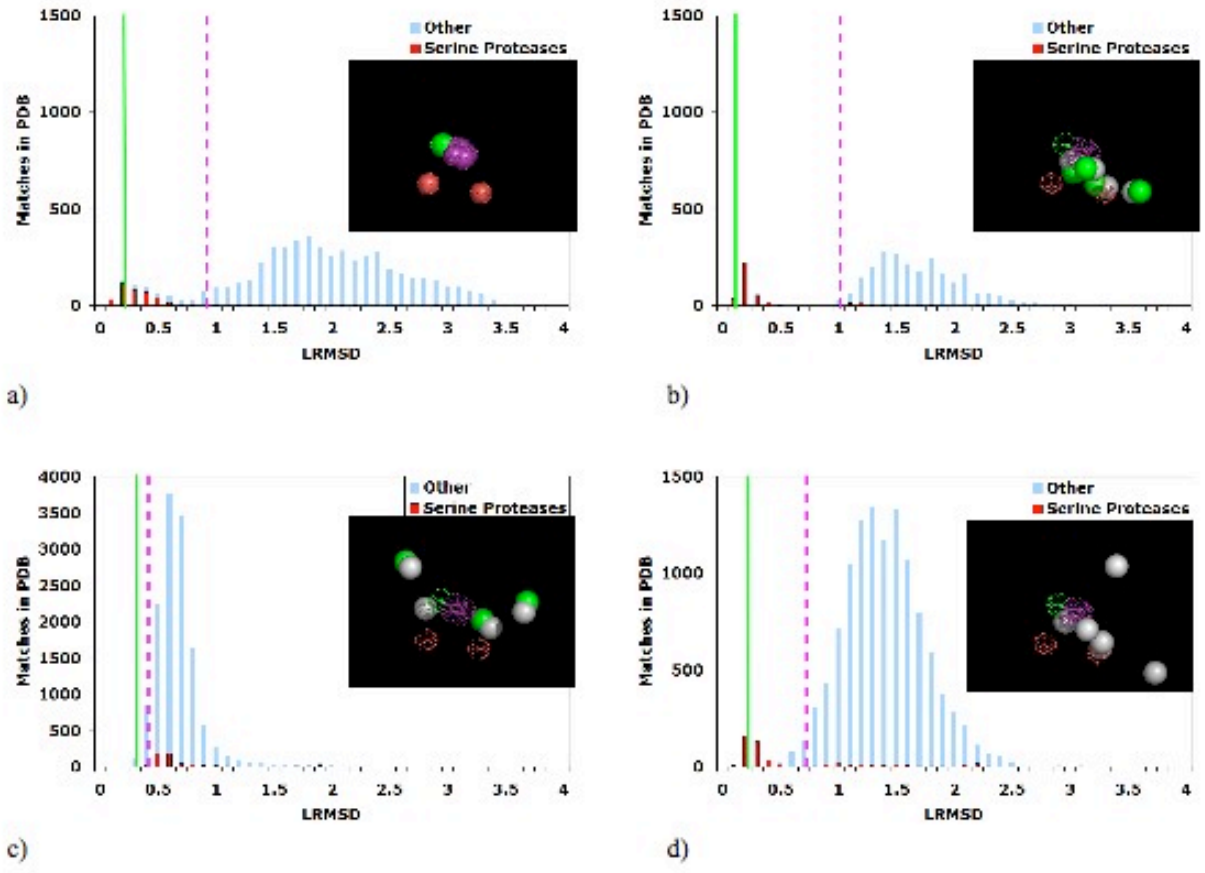
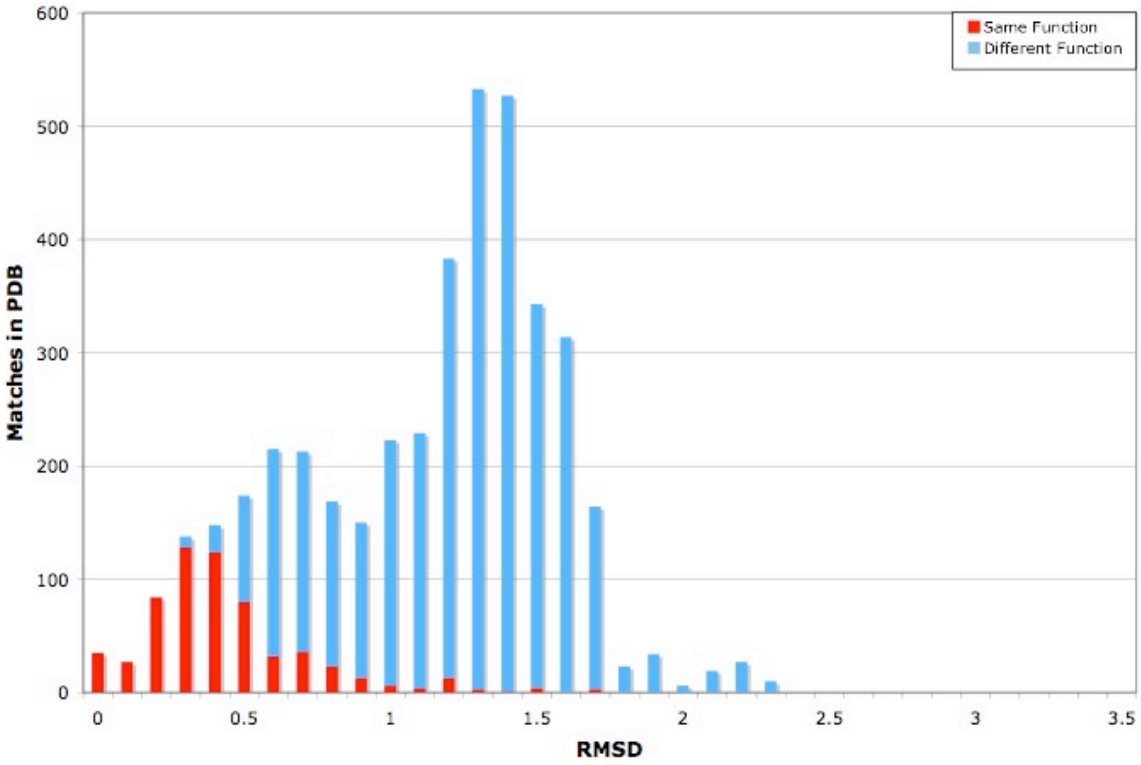
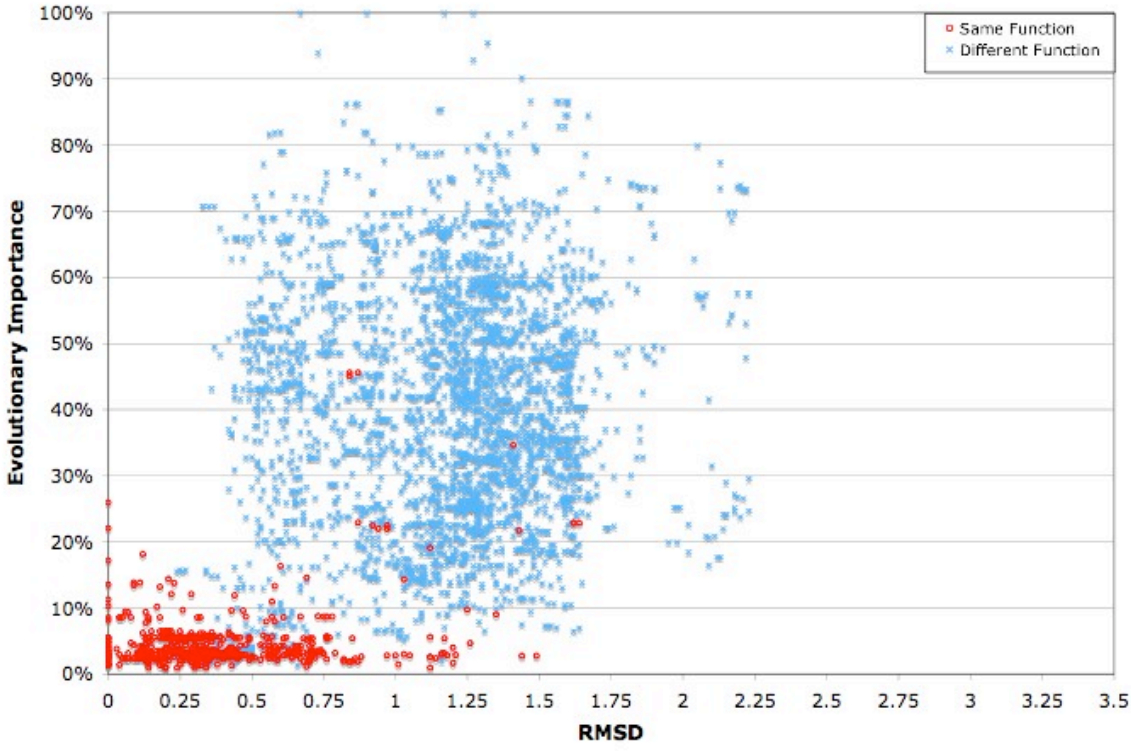


Figure 1: Distribution of matches for serine protease motifs: (a) catalytic triad, (b) non-catalytic quartet, (c) negative control, (d) surface trace cluster. The vertical lines represent the points at which p-value=1% (solid green) and 5% (dashed purple). Structural template representations created using PyMOL (DeLano 2002).



a)



b)

Figure 2: Distribution of matches for 53 enzymes in (a) a single dimension, LRMSD, and (b) two-dimensions, LRMSD and evolutionary importance.

Table 1

Classification (53 proteins)							
	3D Matches	**ET+LRMSD**		**LRMSD**		**ET**	
		Positive	**Negative**	**Positive**	**Negative**	**Positive**	**Negative**
Same EC	570	553	17	515	55	554	16
Different EC	5450	350	5100	874	4576	491	4959
Total	6020	903	5117	1389	4631	1045	4975

SVM Performance					
		ET+LRMSD		**LRMSD**	**ET**
Performance Metric		**All**	**Cross-validated**	**All**	**All**
Accuracy		0.94	0.94 ± 0.09	0.85	0.92
Sensitivity		0.97	0.96 ± 0.10	0.90	0.97
Specificity		0.94	0.94 ± 0.09	0.84	0.91
Positive Predictive Power		0.61	0.62 ± 0.39	0.37	0.53
Negative Predictive Power		1.00	1.00 ± 0.01	0.99	1.00

Table 1: Performance of the SVM classifier in distinguishing between true and false matches for the attributes: ET+LRMSD, LRMSD alone, and ET alone. Leave-one-out cross-validation is done for each enzyme class in the dataset for ET+LRMSD.