Structure, Vol. 12, 1405–1412, August, 2004, ©2004 Elsevier Ltd. All rights reserved. DOI 10.1016/j.str.2004.05.012

Finding Functional Sites in Structural Genomics Proteins

Alexander Stark, Alexander Shkumatov, and Robert B. Russell*
EMBL
Meyerhofstrasse 1
69117 Heidelberg
Germany

Summary

Assigning function to structures is an important aspect of structural genomics projects, since they frequently provide structures for uncharacterized proteins. Similarities uncovered by structure alignment can suggest a similar function, even in the absence of sequence similarity. For proteins adopting novel folds or those with many functions, this strategy can fail, but functional clues can still come from comparison of local functional sites involving a few key residues. Here we assess the general applicability of functional site comparison through the study of 157 proteins solved by structural genomics initiatives. For 17, the method bolsters confidence in predictions made based on overall fold similarity. For another 12 with new folds, it suggests functions, including a putative phosphotyrosine binding site in the Archaeal protein Mth1187 and an active site for a ribose isomerase. The approach is applied weekly to all new structures, providing a resource for those interested in using structure to infer function.

Introduction

Structural genomics projects and the generally increased pace of structure determination mean that structures are often available for proteins of unknown function (e.g., Burley, 2000). Functions for these proteins can often be predicted by comparison to other structures. Common comparison methods like DALI (Holm and Sander, 1993b, 1995), VAST (Gibrat et al., 1996), SSAP (Orengo and Taylor, 1996), or STAMP (Russell and Barton, 1992) detect similarities through structural alignment and thus identify proteins with a similar fold (see Novotny et al. [2004] for a recent comparison of fold comparison servers). These approaches are very often successful at detecting ancient relationships between proteins that are not found by sequence comparison. Indeed, their availability leads to numerous discoveries not seen when structures were first determined (e.g., Artymiuk et al., 1995, 1994b, 1997; Holm and Sander, 1993a; Russell and Barton, 1993; Swindells et al., 1993; for a recent review, see Zhang and Kim, 2003). These similarities, though remote, are often associated with a similar function that can be highly revealing for a new structure.

However, since the focus of many structural genomics efforts is to deliberately target proteins most likely not to resemble any others, structural alignment methods will fail for many new structures. Moreover, finding a similarity between a new structure and a highly populated fold, like the $\beta\alpha$-(TIM)-barrels, can sometimes produce ambiguous results, since the best functional match does not always have the best overall structural similarity. For example, protein Mpt51 of *Mycobacterium tuberculosis* is not catalytically active despite 40% sequence identity and extensive fold similarity to α/β hydrolases (Wilson et al., 2004). Fold similarities thus often require additional investigation of key residues before functions can be inferred (Babbitt, 2003).

To assign functions more explicitly, several methods have been developed to look for similarities between functional sites, regardless of whether or not there is a similarity in overall fold (e.g., Artymiuk et al., 1994a; Kleywegt, 1999; Russell, 1998; Wallace et al., 1997), and we have recently presented a statistical model for such searches (Stark et al., 2003). Application of these methods has uncovered many interesting examples (e.g., Lorentzen et al., 2003; Sanishvili et al., 2003), though there is still no general picture of their applicability based on an analysis of a large test set.

There are an increasing number of protein structures of unknown function: between 2001 and 2003, the number of structural genomics protein structures has risen from only 3 to 228, and in the first four months of this year, already 164 structures became available (Berman et al., 2000). This makes automated approaches to assign function on a large scale highly relevant in structural biology. Here we apply a method to probe for functional site similarities in 254 currently available structures from structural genomics projects. Examples of functional similarities despite no sequence or overall fold similarity demonstrate the complementarity of this approach to those based on structural alignment. We also discuss a weekly service where all new structures are probed in the same way that will serve as a useful tool for structural biologists seeking functions.

Results and Discussion

Overall Performance of Functional Site Comparison

Removing sequences with clear homology to already known structures (see Experimental Procedures) left 157 of the original 254 structures for further analysis. Dali finds matches with $Z \geq 10$ for 61 (39%), and PINTS reports matches with $E \leq 10^{-3}$ for 29 (18%). For 17 (11%), both methods find significant matches; 44 (28%) were only found by Dali and 12 (8%) only by PINTS. The proportions are similar when structures labeled as "unknown function" are used instead (Dali: 41%; PINTS: 21%; overlap: 12%; Dali-only: 29%; PINTS-only: 8%). The numbers change if different Z score or E value thresholds are chosen (Table 1); however, the complementarity of the two approaches remains, even with the most extreme thresholds: with a strict one for Dali ($Z \geq$

*Correspondence: russell@embl-heidelberg.de

Table 1. Overall Performance of Functional Site Comparison

Cutoffs		Number of Matches				
Dali Z Score	PINTS E Value	Dali	PINTS	Both	Only Dali	Only PINTS
12	10^{-4}	50 (32)	23 (15)	12 (8)	38 (24)	11 (7)
12	10^{-3}	50 (32)	29 (18)	14 (9)	36 (23)	15 (10)
12	10^{-2}	50 (32)	52 (33)	25 (16)	25 (16)	27 (17)
9	10^{-4}	69 (44)	23 (15)	15 (10)	54 (34)	8 (5)
9	10^{-3}	69 (44)	29 (18)	18 (11)	51 (32)	11 (7)
9	10^{-2}	69 (44)	52 (33)	31 (20)	38 (24)	21 (13)
6	10^{-4}	105 (67)	23 (15)	19 (12)	86 (55)	4 (3)
6	10^{-3}	105 (67)	29 (18)	23 (15)	82 (52)	6 (4)
6	10^{-2}	105 (67)	52 (33)	42 (27)	63 (40)	10 (6)

Number (percentage) of proteins for which Dali or PINTS reported matches above the indicated thresholds. Note that in an earlier study, 9% of protein pairs with Dali $Z \leq 12$ were not functionally related (13% for $Z \leq 9$ and 28% for $Z \leq 6$ [Holm and Sander, 1997]).

12) and lenient one for PINTS ($E \leq 10^{-2}$) there are still 16% found only by Dali; in the reverse situation (Dali $Z \geq 6$, PINTS $E \leq 10^{-4}$), there are still 3% found only by PINTS (Table 1). However, it is important to note that Dali is not explicitly designed to find functional similarities: correct, structurally similar yet functionally different matches are often found with high Z scores. Even with a Z score threshold of 12, it is expected that 9% of matches will be functionally unrelated, increasing to 28% for $Z \geq 6$ (Holm and Sander, 1997).

There are several reasons why similarities are found by Dali and not by PINTS. For example, active sites can sometimes be distorted by binding to other molecules, or indeed be incorrectly modeled owing to poor X-ray or NMR data, and cannot be detected with statistical significance. This effect is most pronounced for similarities involving a small number of side chains. For example, our best match for Tm1158 (1o1y) is to three residues from the active site of a glutamine amidotransferase domain (1a9x). Although the E value $E = 0.035$ is above the threshold used here, the match is from the same family as the best Dali match (1qdl, $Z = 20.4$). Other missed similarities include those lacking common small-ligand binding sites, such as scaffolding proteins (e.g., 1oyz/1b3u, Dali $Z = 15$) or DNA/RNA binding proteins (1jyh/1d5y, $Z = 14$; 1ljo/1d3b, $Z = 12$). Some Dali matches are to other proteins that are also of unknown function, where no functional pattern is present in any database (e.g., 1o13/1p90, $Z = 11.5$), or involve fold matches without a similarity in function (e.g., helical bundles [1n1q/1bcf, Dali $Z = 18$] or a periplasmic divalent cation tolerance protein with fold similarity to anthranilate isomerase [1p1l/2pii, $Z = 10$]).

The 12 structures matched only by PINTS are usually novel folds where a functional similarity is found between proteins with different overall folds. Of these, five are metal binding sites, two are ligand binding sites, three are anion binding sites, and two are short linear motifs with similar conformations probably due to their secondary structure context but lacking an apparent functional role.

Using a large number of structural genomics targets without sequence similarity to known structures, we can find functional centers within an overall similar fold for 11% and detect functional similarities across folds that cannot be detected by structural alignment methods for an additional 6% of all structures. Specific examples of how functional site similarity can aid structure-based annotation of function are discussed in the sections that follow.

Confirmation of Superfamily, or Resolution of Ambiguity

Overall sequence or fold similarity does not always reveal the correct function. For example, the archaeal fructose-1,6 bisphosphate aldolase (FBPA) (Lorentzen et al., 2003) shows the highest fold similarity to a triosephosphate isomerase (1hg3, Dali $Z = 17.7$), high above the FBPAs from eukaryotes (Dali $Z = 7.4$ for 1fbp; though note in this case that VAST [Gibrat et al., 1996] identifies the correct function match). Functional site comparison methods have already shown some promise in resolving these situations (e.g., β/α-barrels [Lorentzen et al., 2003] or α/β hydrolases [Wilson et al., 2004; Sanishvili et al., 2003]; see Babbitt [2003] for a general discussion).

There are several structures for which we could support functional similarities suggested by a Dali match through the identification of a functional center. These include the similarities between cephalosporin c deacetylase and α/β hydrolases (PINTS $E = 1 \times 10^{-8}$, Dali $Z = 20$; Figure 1A), between Mj0882 and methyltransferases ($E = 3 \times 10^{-5}$, $Z = 13.2$; Figure 1B), between Hi0754 and glucosamine 6-phosphate synthase ($E = 6 \times 10^{-09}$, $Z = 14.2$; Figure 1C), or between Tm1643 and lactate dehydrogenase ($E = 3 \times 10^{-4}$, Dali $Z = 9.3$; Figure 1D).

For Yjee (Teplyakov et al., 2002) (Figure 1E), the best Dali match is marginal (RecA, $Z = 6.3$), not readily allowing any functional conclusions. However, the functional site found here is highly significant, involving five residues from the GDP binding sites of Ran ($E = 2 \times 10^{-5}$) or other P loop nucleotide hydrolases from the same superfamily. The subsequently determined ADP-bound form of Yjee shows that the two nucleotides superimpose perfectly (Figure 1E, right).

Sites Found by Similarities between Different Folds

Functional sites found across different folds are both intriguing and useful: they can suggest aspects of convergent evolution or can suggest functional details for proteins adopting folds not seen before. Those detected here fall into broad classes that we discuss below.

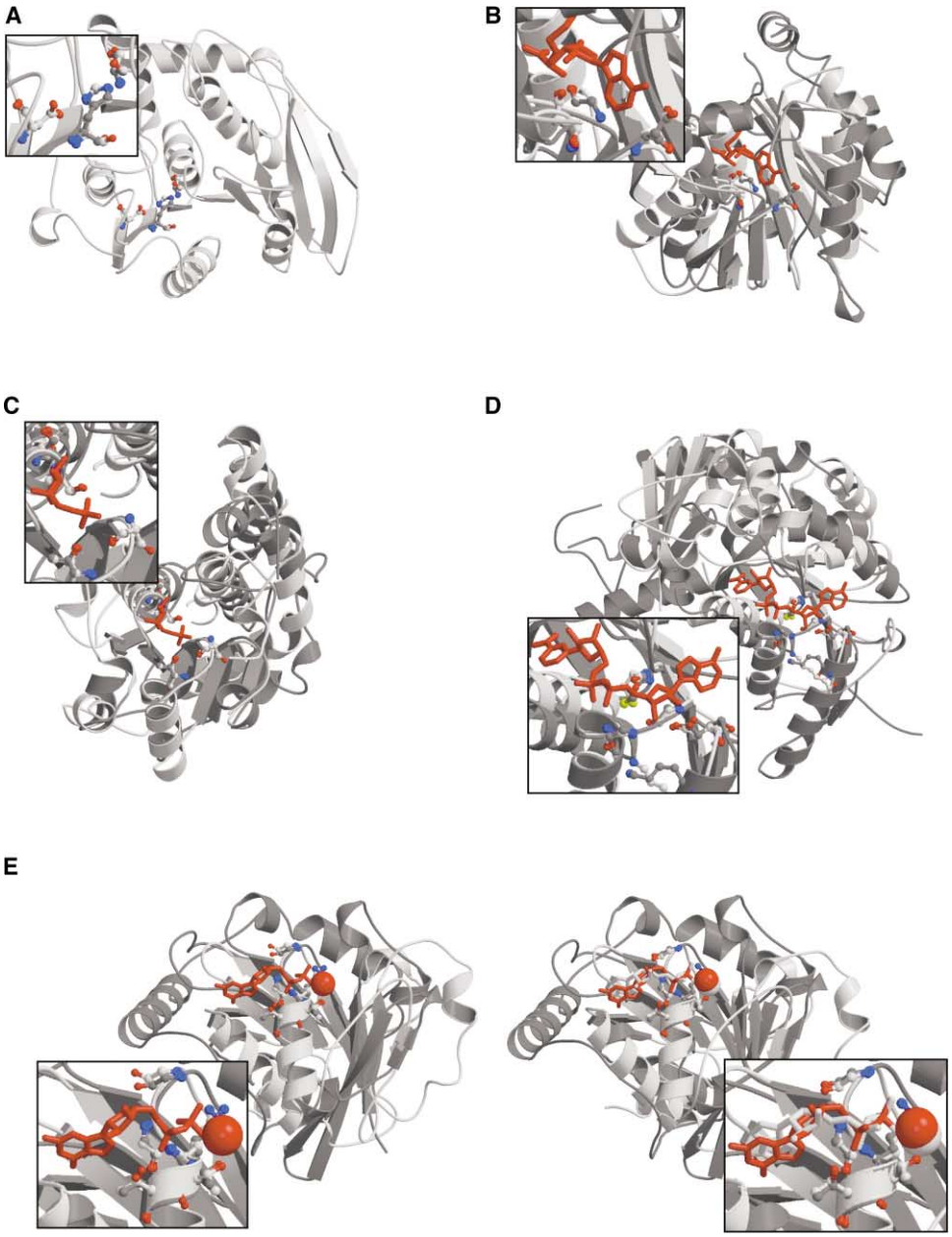

Figure 1. Functional Site Conservation within Superfamily or Fold

Molscript (Kraulis, 1991)/Raster3D (Merritt and Murphy, 1994) figures showing matches in similar folds. Structural genomics proteins (query) are shown in light gray, and the match is shown in dark gray cartoons. Matched residues are shown in ball-and-stick, with the ligand of the database structure in red (magnified as insets).

(A) Similarity between cephalosporin c deacetylase (1l7a) and the catalytic triad of prolyl-oligopeptidase (1h2x, $E = 1 \times 10^{-8}$).

(B) Residues Gly63, Asp84, and Asp113 from the hypothetical protein (HP) Mj0882 (1dus) matched to the S-adenosylmethionine binding site from isoflavone o-methyltransferase (1fpx, $E = 3 \times 10^{-5}$, Dali $Z = 13.2$). Here Dali's first match (1nv8, $Z = 18.2$) ranks 2nd in PINTS ($E = 9 \times 10^{-3}$).

(C) Residues Thr78, Ser79, Ser147, and Thr150 from the HP Hi0754 (1nri) match to the glucosamine 6-phosphate binding site in the isomerase domain of glucosamine 6-phosphate synthase (1moq, $E = 6 \times 10^{-9}$, Dali $Z = 12.6$). Dali's best match (1jeo, $Z = 14.2$) belongs to the same superfamily (c.80.1).

(D) Residues Gly7, Gly9, Gly12, Asp28, Lys32, and Cys55 from HP Tm1643 (1j5p) match the NAD binding site in lactate dehydrogenase (2ldb, $E = 3 \times 10^{-4}$, $Z = 9.3$).

(E) Residues Gly43 and 45–48 from Yjee (1fl9, unliganded, left) (Teplyakov et al., 2002) match the Ran GDP binding site (1a2k, $E = 2 \times 10^{-5}$). Superposition of the ADP-bound form of Yjee (1htw, right) showing the similar position of the nucleotide.

Metal or Phosphate Binding Sites

Nature frequently reinvents similar metal binding sites (Russell, 1998), and unsurprisingly, several similarities observed across folds involve metals. For example, Tm1083 has a highly significant similarity with the calcium binding site of staphylococcal nuclease ($E = 5 \times$

10^{-5}), despite an obvious difference in fold (Figure 2A). Aq1354 contains a site similar to the Zn containing active site of carbonic anhydrase ($E = 8 \times 10^{-3}$) and other Zn binding sites (Figure 2B). Although no metal is present in the structure, the conservation of the histidine residues suggests that the site is real, Zn being absent from the structure owing to EDTA in the purification protocol (Oganesyan et al., 2003). The Dali server, in contrast, reported only a marginal match to glucuronidase (1mqp, $Z = 3.3$) that didn't allow reliable functional inferences.

Phosphate site similarities also arise convergently. For example, the survival protein Eα (SurEα) has a site similar to the active site of a phosphate binding periplasmic protein ($E = 2 \times 10^{-4}$; Figure 2C). Although SurEα is not liganded itself, a homolog (SurE, 1j9l) contains a VO^{4-} ion at the corresponding site. Conserved residues lining this surface lead to the protein being identified as a putative phosphatase site with a preferred specificity toward purine nucleotides (Mura et al., 2003).

Active Site in Ribose-5-Phostphase Isomerase

The alternate ribose-5-phosphate isomerase (RpiB/AlsB) catalyzes the conversion of ribose-5-phosphate to ribulose-5-phosphate. Structure comparison confirmed the similarity to Rossman fold proteins but did not reveal any insights into the reaction, though the authors used their knowledge about preferred binding sites, residue conservation, and surface curvature (i.e., surface cavity) to locate a putative active site pocket (Zhang et al., 2003). The best match for this protein in our study is the similarity between three residues that line one side of this pocket and the substrate-bound active site of phosphoglycerate mutase ($E = 3 \times 10^{-2}$; Figure 2D). Although the similarity does not comprise the full active site pocket, it is useful for determining the location and specificity for phosphorylated ligands.

A Predicted Phosphotyrosine Binding Site in an Archaeal Protein

Structure comparison revealed that Mth1187 adopted a ferrodoxin-like fold, and the authors speculated that it might be a protein-protein interaction module (Tao et al., 2003). They noted, however, that the residues lining the binding site of a sulfate ion showed enhanced conservation indicative of a functional site or a binding site for an unknown ligand (Tao et al., 2003). We found a highly significant similarity to the phosphotyrosine binding sites in SH2 domains (1fyr, $E = 3 \times 10^{-4}$; Figure 2E). The similarity includes residues contacting the sulfate ion in addition to others that contact the tyrosine ring. Indeed, a reverse search of the Mth1187 binding site against all phosphate/sulfate binding sites or against a representative set of complete structures (Hobohm and Sander, 1994) finds no other significant similarities. Phosphotyrosine is thus an excellent candidate for the natural ligand. This is particularly intriguing, as tyrosine-specific protein kinases and phosphatases have only recently been recognized to play important roles in prokaryotic organisms (Bakal and Davies, 2000; Kennelly, 2002, 2003; Kim et al., 2003; Shi et al., 1998).

Comparison of PINTS with PROCAT and Rigor

We wanted to compare the performance of PINTS to other services or programs that perform searches for small functional patterns. PROCAT allows proteins to be compared to a database of 36 partially redundant 3D enzyme active site templates from 5 enzyme classes (Wallace et al., 1997). When we compared the 254 structural genomics structures to the PROCAT templates using standard settings, 127 of them had no matches and 24 gave errors due to an unknown problem, including the catalytic triad containing cephalosporin c deacetylase (1l7a; C. Porter, personal communication). For the remaining 103 structures, PROCAT reported 288 matches to 16 templates including 5 lysozyme templates (82 matches), 2 glucosidase templates (48 matches), and a β-amylase template (31 matches). The large number of matches to these 2 residue templates confirms their low specificity noted earlier by the authors (Wallace et al., 1997) and suggests that most are likely to be false positives. We suspect this because for some of the proteins PROCAT suggests a function that is different from that actually known or because there is no overall structural similarity between the proteins as would have been expected from the cases where Dali and PINTS agree (see above). In addition, visual inspection of the structures for the top ten matches ranked by root-mean-square deviation (rmsd) and for all matches to the only 3 residue template (Asp-His-Asp catalytic triad) confirmed that they are probably nonfunctional. The overall performance was largely as expected, since most of the proteins do not belong to the enzyme classes covered by PROCAT.

Rigor (Kleywegt, 1999) also allows the comparison of a protein to a database of patterns similar to those used here (ftp://xray.bmc.uu.se/pub/gerard/spasm/rigor_sep01.lib.gz). We searched the 254 structural genomics proteins against the 10,588 (non-"ENGINEERABLE") patterns in the database using standard settings. The database contains both collections of residues making contact with ligands and others probably more relevant to broader descriptions of structures (e.g., clusters of negative, positive, or hydrophobic residues and consecutive residues), which we ignored. As Rigor does not provide a measure for the significance of matches, we chose to compare the matches directly, i.e., how often PINTS or Rigor report the correct SCOP superfamily among the top matches without E value or rmsd thresholds. For 44 structures, PINTS ranks a functional match to the correct superfamily on top—Rigor in only 5 cases. This changes to 47 versus 12 structures when the top three matches are examined for both programs. Rigor makes predictions not found by PINTS for two structures; however, one of these appears not to be functionally correct, as visual inspection revealed that the correct superfamily is found, but not the correct binding site (shikimate 5-dehydrogenase-like protein HI0607 [1npy] matches to the NAD binding site in alcohol dehydrogenase [1b16]). In addition, we examined the structures for which PINTS finds significant matches in different folds (see above). Rigor finds the correct or a chemically similar ligand (i.e., sulfate versus phosphate, pyrophosphate versus ATP or ADP, etc.) for 6 out of 12 structures as best match, and for 3 additional structures among the top 10 matches. The calcium binding site in Tm1083 (1j3v) and the putative phospho-tyrosine binding site in Mth1187 (1lxn) are missed.

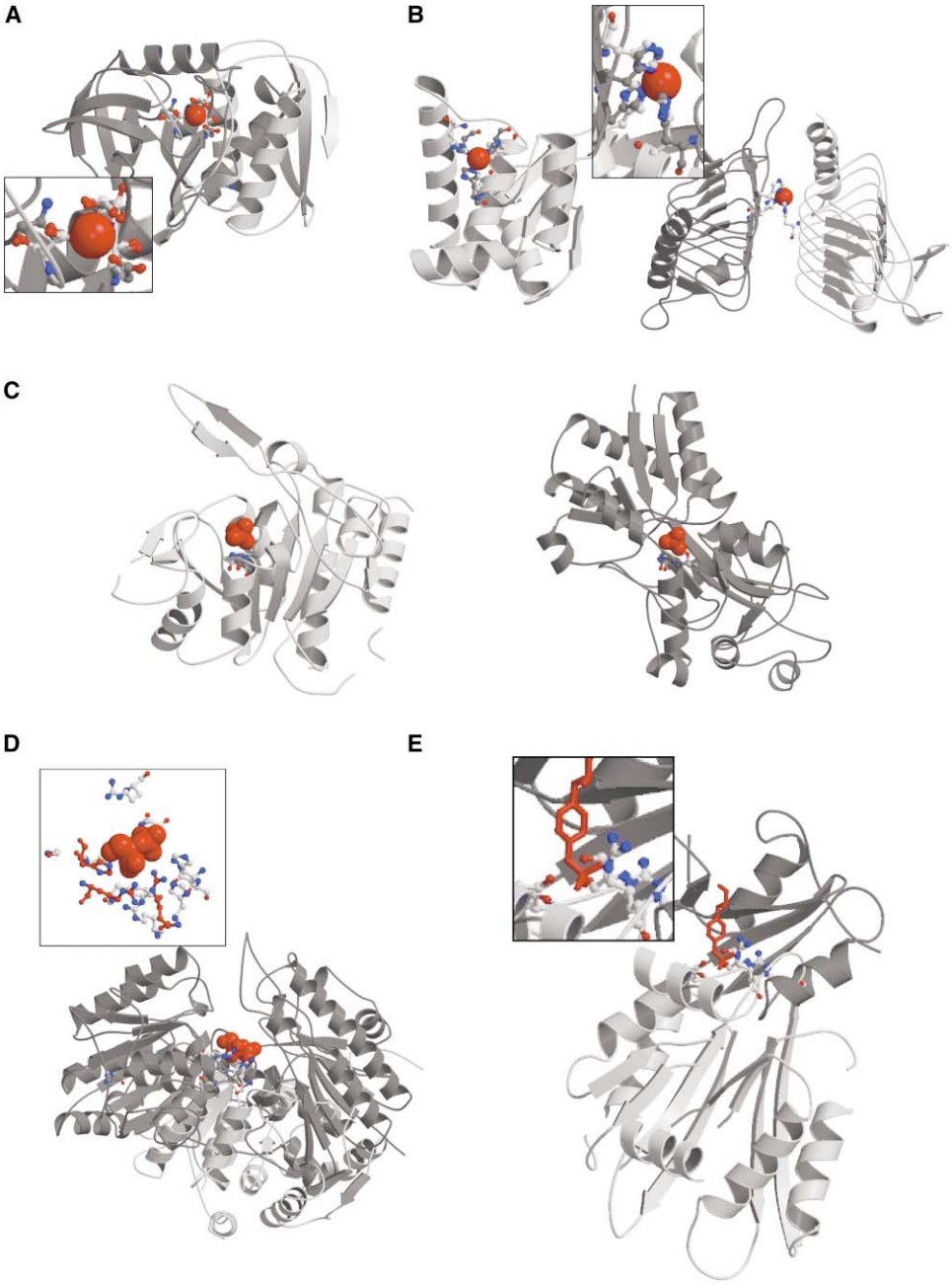

Figure 2. Functional Similarities between Different Folds

Molscript (Kraulis, 1991)/Raster3D (Merritt and Murphy, 1994) figures showing matches between different folds (details are as in Figure 1).

(A) Residues Asp16, Asp129, and Thr130 from the hypothetical protein (HP) Tm1083 (1j5u) match to a Ca binding site in staphylococcal nuclease (1sty, $E = 5 \times 10^{-5}$).

(B) His115, His125, and His119 of HP Aq1354 (1oz9) (Oganesyan et al., 2003) match to the zinc binding active site of carbonic anhydrase (1thj, $E = 9 \times 10^{-3}$, only the two subunits contributing to the zinc binding site are shown for clarity).

(C) Ser106, Gly107, and Thr108 of the survival protein E (SurE) homolog (1l5x) (Mura et al., 2003) match to the phosphate binding site of a phosphate binding periplasmic protein (1a40, $E = 2 \times 10^{-4}$).

(D) Residues His10, Arg133, and Arg40 of alternate ribose-5-phosphate isomerase Rpib/Alsb (1nn4) (Zhang et al., 2003) match to the active site of phosphoglycerate mutase (1o98 and 1o99; 2-phosphoglycerate bound; $E = 3 \times 10^{-2}$ [Rigden et al., 2003]). The inset shows the residues that line the putative active site pocket, with the matching residues in red.

(E) Residues Ser17, Ser19, Arg1081, Arg1086, and Ser1094 of the HP Mth1187 (1lxn) (Tao et al., 2003) match to the phosphotyrosine binding site of an SH2-domain (1fyr, $E = 3 \times 10^{-4}$).

Generally, matches were missed by Rigor because the patterns were missing from its database (e.g., the catalytic triad) or because the method requires complete matches to the patterns. We suspect that many more matches would be found if such partial matches were permitted and the database was enlarged to cover func-

tional residues not in contact with a bound ligand. In addition, Rigor finds many more matches than PINTS and often does not rank correct matches first. Both are due to small patterns that are easily matched within the threshold. Inspection suggests that the majority of these are not functionally meaningful and are the result of over-prediction.

The comparison to PROCAT and Rigor is informative, as it reveals important facts about how to best detect functionally meaningful patterns. Simply for detection, partial matches to larger patterns are clearly essential to allow for the type of variation seen even in homologous proteins. However, the most important point is that a reliable statistic is necessary to rank the matches correctly and to distinguish real functional matches from noise. Using a measure like rmsd, methods will tend to over-predict, particularly for small patterns, and produce a vast number of incorrect predictions. For a single query, with a single protein, it is usually possible to assess the results by visual inspection. However, to be applicable to a wide variety of structures and databases, reliable statistics are a must.

Conclusions

We have tested the applicability of functional site comparison on a large dataset of new proteins with unknown function. Many structures show similarities between functional residues to those solved previously, which can lead to functional hypotheses to be tested further. The examples show a variety of situations, ranging from confirmation of a similarity inferred by overall structural similarity to detecting a convergently evolved mode of ligand binding. For specific examples, we have uncovered intriguing similarities that give suggestions for experiments.

Overall, the results demonstrate how searches for similar functional sites complement those for similar folds. A combined strategy where both types of searches are used for structure-based functional annotation can help overcome problems inherent to each when applied separately. Even when structural alignment searches reveal fold similarities, active site comparison can highlight the presence (Sanishvili et al., 2003) or absence (Wilson et al., 2004) of an active site and can sometimes resolve functional ambiguities (Lorentzen et al., 2003). It can also help to identify "migrating" catalytically equivalent residues that are located on different parts of homologous structures (e.g., Todd et al., 2001). Moreover, newly determined active sites can be sought in previously existing structures regardless of any similarity in overall fold.

The complementarity can also work in reverse: a similarity in fold as revealed by structural alignment can boost confidence in a marginally significant functional site match. This is particularly relevant for matches involving only a few residues that require too narrow geometrical constraints (i.e., small rmsd) to be distinguishable from noise (Stark et al., 2003), or those involving residues distorted by bound ligands. Functional site matches involving proteins of the same fold can be more believable even when the matches themselves are marginal.

Both approaches will benefit from the increasing number of functionally annotated protein structures. There are also recent efforts to catalog active sites in structures based on studies of their function (Bartlett et al., 2002; Porter et al., 2004). These will increase the coverage, sensitivity, and specificity of methods like that described here. Investigating both types of similarities discussed here while the number structures and known functional sites grows will also complete the picture of how Nature evolves or reinvents proteins to perform different functions with a diverse array of ligands.

Experimental Procedures

Analysis of Structural Genomics Proteins

We considered 254 structures labeled as "Structural Genomics" with release dates up to October 2003 and compared these to the sequences in the protein databank (PDB) (Berman et al., 2000) using BLAST (Altschul et al., 1997). Proteins that matched any other sequence in the database with an expectation E value $\leq 10^{-10}$ were not considered further. Although this degree of sequence similarity is not always associated with a similarity in function (Rost, 2002; Tian and Skolnick, 2003; Todd et al., 2001), our threshold ensures that the analysis excludes all cases where functional similarities are obvious from sequence comparison and is thus reliable in assessing the added value of structure comparison. Altering this threshold does not, however, greatly affect the overall findings.

We then used these structures to search for similarities in a database of side chain patterns using the PINTS (Patterns In Non-homologous Tertiary Structures) method (Stark and Russell, 2003; Stark et al., 2003). These patterns consist of residues that are either close to bound ligands (*ligand binding sites*) or labeled by the authors of the structure (*SITE* records) (Stark and Russell, 2003). When comparing a new structure to these databases, the method identifies patterns of residues with similar spatial arrangements that need not be close in sequence nor in the same relative sequence order. It optimally superimposes the matching patterns and assigns a statistical significance E value to the rmsd (Stark et al., 2003). For comparison, we also compared the structure to PDB representatives in the FSSP database using Dali (Holm and Sander, 1993b) with default options. Dali scores similarities by a Z score that assesses the nonrandomness of the matches. For a normal distribution, Z scores can be easily interpreted, such that, for example, less than 0.14% of random matches would achieve $Z = 3$ or better. However, in practice the distributions are usually not normal but are instead heavily skewed, making the interpretation of Z scores difficult.

For NMR structures, we considered only the first model of an ensemble. We do not anticipate that the results would change significantly if different models were chosen, since we generally observe a greater degree of similarity across models in functionally relevant parts of the protein.

Structural Similarity Thresholds

PINTS usually detects binding site similarities for chemically similar ligands with E values between 10^{-4} and 10^{-2}, whereas negative matches generally have $E \geq 0.1$ (Stark et al., 2003). We thus inspected the number of matches for E value cutoffs 10^{-4}, 10^{-3}, and 10^{-2} (see Table 1) and found reliable functional clues to come from matches with $E \leq 10^{-3}$. However, we also inspected the best matches for each structure manually.

As Dali Z scores are defined around structural similarity rather than functional similarity, the accuracy of inferring functional relationship for a given Z score is specific family (Dietmann and Holm, 2001). For example, TIM-barrels can have different functions at comparatively high values (e.g., $Z = 18$) (Lorentzen et al., 2003), while Rossman-type NAD binding domains are reliably detected with values as low as 6. However, earlier observations showed that fewer than 10% of structure pairs with values above 12 are functionally unrelated (see legend to Table 1; and Holm and Sander, 1997). We thus decided to use this as the threshold for our study.

We report and discuss only the best matches for PINTS or Dali

and only consider one representative for groups of structures sharing 90% sequence identity.

Weekly Updates

We modified the scheme above slightly to apply it to new structures on a weekly basis. Updates to the PDB often contain structures that are either slight variants (e.g., different bound small molecules, mutants, etc.) or close homologs of proteins already present in the database. We thus first compare the sequences of new structures to the current database using BLAST and remove those that match any structure with an E value $\leq 10^{-20}$, a sequence identity $\geq 80\%$, and a length difference of $\leq 10\%$ or ≤ 50 residues. We compare the remaining structures to the pattern database discussed above and flag those labeled as "Structural Genomics" or "Unknown Function." This service is available at www.russell.embl.de/pints-weekly.

Received: February 5, 2004
Revised: May 13, 2004
Accepted: May 14, 2004
Published: August 10, 2004

References

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.

Artymiuk, P.J., Poirrette, A.R., Grindley, H.M., Rice, D.W., and Willett, P. (1994a). A graph-theoretic approach to the identification of three-dimensional patterns of amino acid side-chains in protein structures. J. Mol. Biol. *243*, 327–344.

Artymiuk, P.J., Rice, D.W., Poirrette, A.R., and Willet, P. (1994b). A tale of two synthetases. Nat. Struct. Biol. *1*, 758–760.

Artymiuk, P.J., Rice, D.W., Poirrette, A.R., and Willett, P. (1995). β-glucosyltransferase and phosphorylase reveal their common theme. Nat. Struct. Biol. *2*, 117–120.

Artymiuk, P.J., Poirrette, A.R., Rice, D.W., and Willett, P. (1997). A polymerase I palm in adenylyl cyclase? Nature *388*, 33–34.

Babbitt, P.C. (2003). Definitions of enzyme function for the structural genomics era. Curr. Opin. Chem. Biol. *7*, 230–237.

Bakal, C.J., and Davies, J.E. (2000). No longer an exclusive club: eukaryotic signalling domains in bacteria. Trends Cell Biol. *10*, 32–38.

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.

Burley, S.K. (2000). An overview of structural genomics. Nat. Struct. Biol. Suppl. *7*, 932–934.

Dietmann, S., and Holm, L. (2001). Identification of homology in protein structure classification. Nat. Struct. Biol. *8*, 953–957

Gibrat, J.F., Madej, T., and Bryant, S.H. (1996). Surprising similarities in structure comparison. Curr. Opin. Struct. Biol. *6*, 377–385.

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

Holm, L., and Sander, C. (1993a). Globin fold in a bacterial toxin. Nature *361*, 309.

Holm, L., and Sander, C. (1993b). Protein structure comparison by alignment of distance matrices. J. Mol. Biol. *233*, 123–138.

Holm, L., and Sander, C. (1995). Dali: a network tool for protein structure comparison. Trends Biochem. Sci. *20*, 478–480.

Holm, L., and Sander, C. (1997). Decision support system for the evolutionary classification of protein structures. Proc. Int. Conf. Intell. Syst. Mol. Biol. *5*, 140–146.

Kennelly, P.J. (2002). Protein kinases and protein phosphatases in prokaryotes: a genomic perspective. FEMS Microbiol. Lett. *206*, 1–8.

Kennelly, P.J. (2003). Archaeal protein kinases and protein phosphatases: insights from genomics and biochemistry. Biochem. J. *370*, 373–389.

Kim, S.H., Shin, D.H., Choi, I.G., Schulze-Gahmen, U., Chen, S., and Kim, R. (2003). Structure-based functional inference in structural genomics. J. Struct. Funct. Genomics *4*, 129–135.

Kleywegt, G.J. (1999). Recognition of spatial motifs in protein structures. J. Mol. Biol. *285*, 1887–1897.

Kraulis, P.J. (1991). MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. J. Appl. Crystallogr. *24*, 946–950.

Lorentzen, E., Pohl, E., Zwart, P., Stark, A., Russell, R.B., Knura, T., Hensel, R., and Siebers, B. (2003). Crystal structure of an archaeal class I aldolase and the evolution of $(\beta\alpha)_8$ barrel proteins. J. Biol. Chem. *278*, 47253–47260.

Merritt, E.A., and Murphy, M.E.P. (1994). Raster3D version 2.0. A program for photorealistic molecular graphics. Acta Crystallogr D Biol. Crystallogr. *50*, 869–873.

Mura, C., Katz, J.E., Clarke, S.G., and Eisenberg, D. (2003). Structure and function of an archaeal homolog of survival protein E (SurEα): an acid phosphatase with purine nucleotide specificity. J. Mol. Biol. *326*, 1559–1575.

Novotny, M., Madsen, D., and Kleywegt, G.J. (2004). Evaluation of protein fold comparison servers. Proteins *54*, 260–270.

Oganesyan, V., Busso, D., Brandsen, J., Chen, S., Jancarik, J., Kim, R., and Kim, S.H. (2003). Structure of the hypothetical protein AQ-1354 from Aquifex aeolicus. Acta Crystallogr. D Biol. Crystallogr. *59*, 1219–1223.

Orengo, C.A., and Taylor, W.R. (1996). SSAP: sequential structure alignment program for protein structure comparison. Methods Enzymol. *266*, 617–635.

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–D133.

Rigden, D.J., Lamani, E., Mello, L.V., Littlejohn, J.E., and Jedrzejas, M.J. (2003). Insights into the catalytic mechanism of cofactor-independent phosphoglycerate mutase from X-ray crystallography, simulated dynamics and molecular modeling. J. Mol. Biol. *328*, 909–920.

Rost, B. (2002). Enzyme function less conserved than anticipated. J. Mol. Biol. *318*, 595–608.

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., and Barton, G.J. (1992). Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. Proteins *14*, 309–323.

Russell, R.B., and Barton, G.J. (1993). An SH2-SH3 domain hybrid. Nature *364*, 765.

Sanishvili, R., Yakunin, A.F., Laskowski, R.A., Skarina, T., Evdokimova, E., Doherty-Kirby, A., Lajoie, G.A., Thornton, J.M., Arrowsmith, C.H., Savchenko, A., et al. (2003). Integrating structure, bioinformatics, and enzymology to discover function: BioH, a new carboxylesterase from Escherichia coli. J. Biol. Chem. *278*, 26039–26045.

Shi, L., Potts, M., and Kennelly, P.J. (1998). The serine, threonine, and/or tyrosine-specific protein kinases and protein phosphatases of prokaryotic organisms: a family portrait. FEMS Microbiol. Rev. *22*, 229–253.

Stark, A., and Russell, R.B. (2003). Annotation in three dimensions. PINTS: Patterns in Non-homologous Tertiary Structures. Nucleic Acids Res. *31*, 3341–3344.

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.

Swindells, M.B., Orengo, C.A., Jones, D.T., Pearl, L.H., and Thornton, J.M. (1993). Recurrence of a binding motif? Nature *362*, 299.

Tao, X., Khayat, R., Christendat, D., Savchenko, A., Xu, X., Goldsmith-Fischman, S., Honig, B., Edwards, A., Arrowsmith, C.H., and Tong, L. (2003). Crystal structures of MTH1187 and its yeast ortholog YBL001c. Proteins *52*, 478–480.

Teplyakov, A., Obmolova, G., Tordova, M., Thanki, N., Bonander, N., Eisenstein, E., Howard, A.J., and Gilliland, G.L. (2002). Crystal structure of the YjeE protein from Haemophilus influenzae: a putative Atpase involved in cell wall synthesis. Proteins *48*, 220–226.

Tian, W., and Skolnick, J. (2003). How well is enzyme function conserved as a function of pairwise sequence identity? J. Mol. Biol. *333*, 863–882.

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.

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.

Wilson, R.A., Maughan, W.N., Kremer, L., Besra, G.S., and Futterer, K. (2004). The structure of Mycobacterium tuberculosis MPT51 (FbpC1) defines a new family of non-catalytic alpha/beta hydrolases. J. Mol. Biol. *335*, 519–530.

Zhang, C., and Kim, S.H. (2003). Overview of structural genomics: from structure to function. Curr. Opin. Chem. Biol. *7*, 28–32.

Zhang, R.G., Andersson, C.E., Skarina, T., Evdokimova, E., Edwards, A.M., Joachimiak, A., Savchenko, A., and Mowbray, S.L. (2003). The 2.2 A resolution structure of RpiB/AlsB from Escherichia coli illustrates a new approach to the ribose-5-phosphate isomerase reaction. J. Mol. Biol. *332*, 1083–1094.