WILEY InterScience® DISCOVER SOMETHING GREAT

PROTEINS STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

# A comparison of SCOP and CATH with respect to domain–domain interactions

**Emily R. Jefferson, Thomas P. Walsh, and Geoffrey J. Barton***

University of Dundee, School of Life Sciences, Dow Street, Dundee, DD1 5EH Scotland, United Kingdom

*ABSTRACT*

***The analysis and prediction of protein–protein interaction sites from structural data are restricted by the limited availability of structural complexes that represent the complete protein–protein interaction space. The domain classification schemes CATH and SCOP are normally used independently in the analysis and prediction of protein domain–domain interactions. In this article, the effect of different domain classification schemes on the number and type of domain–domain interactions observed in structural data is systematically evaluated for the SCOP and CATH hierarchies. Although there is a large overlap in domain assignments between SCOP and CATH, 23.6% of CATH interfaces had no SCOP equivalent and 37.3% of SCOP interfaces had no CATH equivalent in a nonredundant set. Therefore, combining both classifications gives an increase of between 23.6 and 37.3% in domain–domain interfaces. It is suggested that if possible, both domain classification schemes should be used together, but if only one is selected, SCOP provides better coverage than CATH. Employing both SCOP and CATH reduces the false negative rate of predictive methods, which employ homology matching to structural data to predict protein–protein interaction by an estimated 6.5%.***

Proteins 2008; 70:54–62.


***Key words: SCOP; CATH; protein–protein interactions; domain–domain interactions.***

## INTRODUCTION

The determination of protein–protein interactions is fundamental to the understanding of many biological processes. High resolution structural data have been widely used in prediction and investigation of protein–protein interactions since they provide information about the interfaces at the atomic level. Such investigations and predictive methods which use structural data include analysis of the properties of protein–protein interaction interfaces,[1–3] prediction of protein interaction sites,[4–6] protein–protein interaction prediction based on homologues to known structural complexes,[7,8,9] and docking methods.[10]

However, many types of protein–protein interactions are not represented in the available structures found in the PDB.[11] Studies of protein interaction data from different sources have produced estimates that there are about 10,000 distinct types of protein–protein interactions but the number of nonredundant interacting pairs in the current PDB is ~2000.[12] It has also been predicted that at the current rate of structure determination it will be 20 years before a full complement of interactions is elucidated.[12] Given this paucity of data, it is imperative that analytical procedures make the best use of all available information.

Analysis of protein–protein interactions is usually performed at the level of domains rather than complete protein chains since domains are considered to be the fundamental functional and structural units of proteins. Investigation and prediction of domain–domain interactions from structural data have typically employed systematic structural domain classifications, such as SCOP,[13,14,15] which classifies domains based on structural, evolutionary and functional similarity, and CATH,[16,17,18] which classifies domains primarily on the basis of structural features. The majority of studies and predictive methods employ only one of these structural domain classification systems.

Analysis at the level of domain–domain interactions is clearly dependent upon the nature of the domain classification system. Hadley and Jones[19] and Veretnik *et al.*[20] examined the differences between SCOP and CATH domain assignments and found that, while there is a high degree of correspondence between CATH and SCOP, there are instances

The Supplementary Material referred to in this article can be found online at http://www.interscience.wiley.com/jpages/0887-3585/suppmat/.

Grant sponsors: BBSRC, TEMBLOR; Grant number: QLRI-CT-2001-00015; Grant sponsor: Scottish Funding Council, Scottish Bioinformatics Research Network; Grant number: 7030-355-064105-35OF; Grant sponsor: ENFIN, a Network of Excellence funded by the European Commission; Grant number: LSHG-CT-2005-518254.

*Correspondence to: Geoffrey J. Barton, School of Life Sciences, University of Dundee, Dow Street, Dundee, DD1 5EH, Scotland, UK. E-mail: geoff@compbio.dundee.ac.uk

Received 27 July 2006; Revised 25 February 2007; Accepted 7 March 2007
Published online 16 July 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21496

where a domain found in one classification has no equivalent in the other. Given these observations, it may also be hypothesised that the set of domain–domain interactions that can be derived using SCOP will contain interactions not found in the set derived using CATH and vice versa. If this were so, it would be advantageous to employ the union of both sets of domain classifications. Accordingly, in this work, the validity of this hypothesis is investigated in addition to the differences in the properties of domain–domain interactions observed using CATH and SCOP.

## MATERIALS AND METHODS

Hadley and Jones[19] and Veretnik *et al.*[20] have found that there are many differences in the organisation of the CATH and SCOP hierarchies and so no direct mapping between the two domain classification systems was attempted here. Instead, this work focuses on comparison of interfaces implied by the two classification systems.

### Terminology

SCOP and CATH are hierarchical domain classification systems. Domains within the same SCOP superfamily have structural and functional similarities that suggest a common evolutionary origin but may not share a detectable sequence similarity. The equivalent CATH category is the "Homologous Superfamily" (H-level). Here, the term "superfamily" is used for both the SCOP level of superfamily and the CATH level of Homologous Superfamily. Similarly, the term "family" is adopted to mean both the SCOP family level and the CATH level of Sequence families (S-level), which groups domains with detectable sequence similarity.

SCOP and CATH domains were classified at both the superfamily and family level of similarity for their respective domain classification systems. Domain–domain interactions were classified by "pairwise superfamily" or "pairwise family." In this article, the term "pairwise superfamily" is used to describe the classification of a domain–domain interaction based on the superfamily classification of each of the interacting domains. Similarly, the term "pairwise family" is used to describe the classification of a domain–domain interaction based on the family classification of each of the interacting domains.

The region where the residues from one domain interact with the residues from another domain is termed the *interface*. The domain classification system employed as the base is referred to as the "reference" classification system and the classification system to which the reference is compared is referred to as the "alternative" classification system. For example, when SCOP interfaces are compared to CATH interfaces the reference classification system is SCOP and the alternative is CATH.

### The data

This study employed SNAPPI-DB (Structures, iNterfaces, and Alignments of Protein–Protein Interactions DataBase),[21] a high performance object-oriented database. Assemblies from the PQS Database[22] were employed rather than the asymmetric units seen in PDB files since this increases the number of nonredundant domain–domain interfaces by 34.5% over that seen in the PDB and removes 2981 probable crystal packing artefacts.[23] Only structures that had both CATH and SCOP definitions assigned were considered in this study.

SCOP (release 1.65) and CATH (version 2.5.1) domains were assigned to positions on chains for each protein structure in the database using the mapping provided by the Macromolecular Structure Database.[24,25] Interfaces between domains were determined based on distance. Atoms were considered to interact if the distance between them was less than the sum of their van der Waals radii (using the radii determined in Ref. 26) +0.5 Å. Two domains were considered to interact if there were 10 or more interacting residue pairs between the domains. The threshold of 10 residue–residue interactions was chosen based on visual inspection of domain–domain interactions sites and published evidence.[27–29]

### Determination of equivalence in domain–domain interaction interface

A method was devised to compare coverage at interfaces for CATH and SCOP interactions, which involved no direct mapping between the two domain classification systems. Comparison of interfaces between SCOP and CATH is complicated since interface residues in a SCOP domain pair may be split across more than two domain pairs in CATH and vice versa. For this reason, a flexible definition of equivalence was developed, which allows a reference interface to be composed of several equivalent alternative interfaces. Figure 1 illustrates how the flexible definition of equivalence was implemented. A valid alternative interface is defined as an interface generated from the alternative domain definition that contains $\geq$5 of the pairwise residue–residue interactions found in the reference interface. Since a reference domain may correspond to more than one alternative domain, the reference interface may have several disjoint alternative interfaces that each contain $\geq$5 residues from the reference interface; in this case, the alternative interfaces are merged and treated as a single alternative interface. This definition of equivalence does not penalise situations where the alternative definition splits a single reference domain into two or more alternative domains. The definition is appropriate since only the coverages of interface sites by SCOP and CATH are compared rather than direct comparison of the single domains.

In the example shown in Figure 1, the reference interface [Fig. 1(a)] consists of 50 interacting residues and is

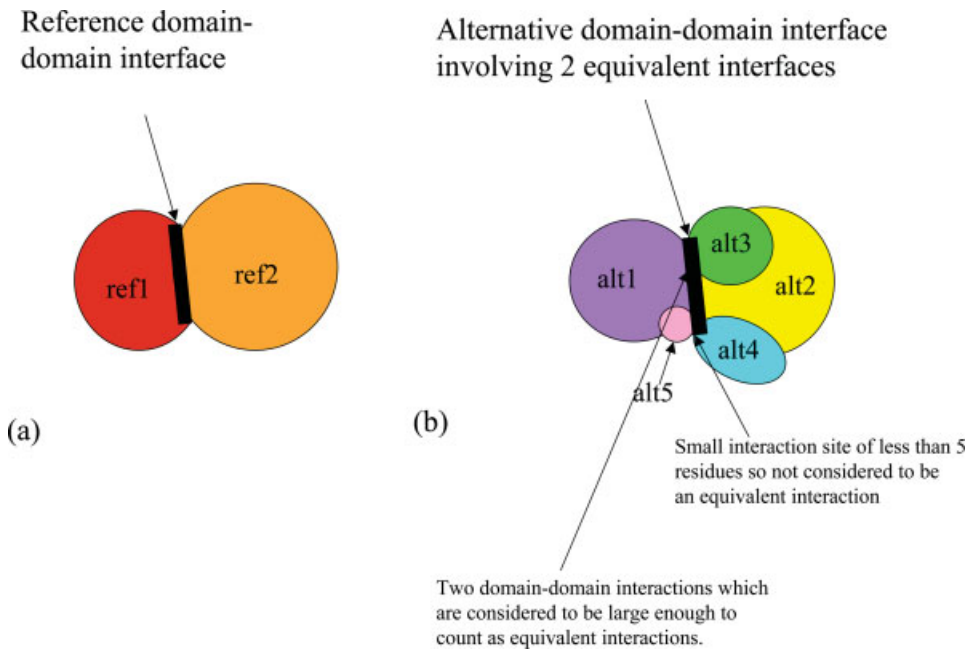


**Figure 1**

*The determination of equivalent domains. The reference interface is between the domains labelled Refs. 1 and 2. There are three interfaces using the alternative domain definition, which covers the same interaction site as the reference interface. The interface between the domains labelled Alt4 and Alt5 is <5 residues in size and so is not considered to be an equivalent interaction. The other two interfaces are ≥5 residues in interaction site size and so are considered to be equivalent to the reference interface.*

shown in red (domain labelled Ref. 1) and orange (domain labelled Ref. 2). The equivalent interface shown in Figure 1(b) is composed of three different alternative interfaces: the first is between domains colored purple (Alt1) and yellow (Alt2) and has an interface size of 20 interacting residue pairs; the second is between domains colored purple (Alt1) and green (Alt3) and has 17 interacting residues and the third is between domains colored cyan (Alt4), and pink (Alt5) and has three interacting residue pairs. In this example, the alternative interface between domains colored cyan (Alt4) colored pink (Alt5) has $<5$ interacting residues and so is considered to be too small to be counted as an equivalent interface. The purple domain (Alt1) is the interaction partner for both the green (Alt3) and yellow (Alt2) domains. This can occur when one of the reference domains is considered to be two domains by the alternative domain definition. To determine the percentage of overlap at the interface site the numbers of common interacting residues for each of the valid alternative interfaces were summed and the total divided by the number of interacting residues seen in the reference interface. For the example discussed earlier, there were three possible alternative interfaces, but only two of these (Alt1-Alt2 and Alt1-Alt3) had an interface of $\geq 5$ residues in size. Accordingly, the number of equivalent interacting residues for the two equivalent interfaces in the example was 37 (17 + 20) and since the reference interface site comprised 50 residues, this gave the percentage overlap of 74%.

## RESULTS AND DISCUSSION

Several properties of CATH and SCOP domain–domain interactions were investigated including the promiscuity of domains, the classification into pairwise superfamilies and families and the sizes of the domains and interaction interfaces. These results are all available in the supplementary material. In summary, it was found that CATH interfaces (mean of 39.8 residues) were generally smaller than SCOP interfaces (mean of 45.7 residues), which can be explained by the difference in single domain sizes when comparing SCOP and CATH, i.e., CATH domains tend to be smaller than SCOP domains and vary less in size. The promiscuity of domains was found to be similar for both CATH and SCOP. The frequencies with which domain–domain interaction were observed in different pairwise superfamilies were similar for the two domain classification systems; however, since CATH uses a more stringent sequence similarity threshold to define families than SCOP (5406), there are more CATH pairwise families than SCOP pairwise families (2932).


DOI 10.1002/prot

**Table I**
*The Percentage of Interfaces, Which Have Equivalent Alternative Interactions at a Range of Overlaps*

| Percentage overlap at interface | Percentage of domain–domain interactions | | | |
|---|---|---|---|---|
| | CATH referenced to SCOP alternative | | SCOP referenced to CATH alternative | |
| | Redundant set | Nonredundant set | Redundant set | Nonredundant set |
| 0 | 13.1 (12089) | 12.4 (243) | 8.3 (6616) | 17.3 (299) |
| 1 to 10 | 0.0 (15) | 0.9 (17) | 0.1 (55) | 0.1 (19) |
| 11 to 20 | 0.1 (102) | 0.8 (15) | 0.6 (470) | 1.2 (20) |
| 21 to 30 | 0.1 (113) | 1.0 (20) | 0.9 (722) | 2.1 (37) |
| 31 to 40 | 0.4 (390) | 1.4 (28) | 1.1 (870) | 2.0 (35) |
| 41 to 50 | 0.9 (817) | 1.4 (28) | 1.4 (1130) | 3.1 (55) |
| 51 to 60 | 0.9 (861) | 1.7 (33) | 1.2 (978) | 3.4 (59) |
| 61 to 70 | 1.1 (1061) | 3.0 (58) | 1.9 (1529) | 4.7 (81) |
| 71 to 80 | 1.5 (1344) | 3.8 (72) | 3.3 (2644) | 6.9 (120) |
| 81 to 90 | 2.7 (2471) | 4.8 (95) | 5.1 (4023) | 9.0 (155) |
| 91 to 99 | 2.4 (2220) | 7.6 (150) | 7.0 (5612) | 9.5 (165) |
| 100 | 76.8 (70940) | 61.3 (1202) | 69.1 (54995) | 39.6 (684) |
| $<25$ | 13.2 (12227) | 14.1 (277) | 9.3 (7404) | 20.0 (345) |
| $\geq 25$ | 86.8 (80196) | 85.9 (1684) | 90.7 (72240) | 80.0 (1384) |
| $<75$ | 17.1 (15776) | 23.6 (463) | 16.5 (13131) | 37.3 (645) |
| $\geq 75$ | 82.9 (76647) | 76.4 (1498) | 83.5 (66513) | 62.7 (1084) |

Numbers in parentheses are the actual numbers of such interfaces found in each case. The bottom four rows show cumulative percentages of interfaces for $<25\%$, $\geq 25\%$, $<75\%$, and $\geq 75\%$ overlap. For example, when a percentage overlap of 75% or more was taken to be sufficient to consider two interfaces to be equivalent, 17.1% (15776 interfaces) of the CATH interfaces had no equivalent SCOP interface and 16.5% (13131 interfaces) SCOP interfaces had no equivalent CATH interface.

Table I summarizes the degree of overlap between reference interfaces and their alternatives. Analysis of the full redundant set of domain pair interfaces found that 77.4% of the reference SCOP interfaces had either full 100% overlap with an equivalent CATH interface (69.1% of interfaces) or 0% overlap (8.3% of interfaces). When CATH was taken as the reference interface, 89.9% of the CATH interfaces either had 0% (13.1% of interfaces) or 100% (76.8%) coverage by SCOP equivalents. Between the extremes of 0 and 100% overlap, the percentage of interfaces with a given coverage increases steadily for both domain definitions. Comparison of the fully redundant sets of interfaces showed no significant difference between CATH and SCOP. When a percentage overlap of 75% or more was taken to be sufficient to consider two interfaces to be equivalent, 17.1% (15776 interfaces) of the CATH interfaces had no equivalent SCOP interface and 16.5% (13131 interfaces) SCOP interfaces had no equivalent CATH interface.

To ensure that these results were not biased by the redundancy of the data, the overlap of the interfaces was also analyzed in a nonredundant set. For each pairwise superfamily, 10 interactions were chosen at random to represent the superfamily interaction. The overlap was determined for each of the 10 different random selections and the mean overlap taken to represent the pairwise superfamily. Table I shows the overlap for the nonredundant set compared to the redundant set. The analysis of interfaces generated by a nonredundant set suggested that the coverage of CATH interfaces by SCOP was greater than the coverage of SCOP interfaces by CATH. If a percentage overlap of 75% or more is again taken to be sufficient to consider two interfaces to be equivalent, 23.6% of the CATH interfaces had no equivalent SCOP interface and 37.3% of SCOP interfaces had no equivalent CATH interface. Only 39.6% of the SCOP interfaces had 100% overlap by equivalent CATH interfaces whereas 61.3% of the CATH interfaces were covered 100% by SCOP interfaces. The pairwise superfamily level of similarity was used to generate the nonredundant data set since SCOP and CATH definitions of what constitutes a superfamily were more similar to that of the SCOP and CATH family-level definitions. However, employing the pairwise family level of similarity to generate the nonredundant data set was found not to alter these results significantly (nor any of the proceeding results) (data not shown).

Although one reference interface may have no equivalent alternative domain–domain interaction in the structure where it was observed, there may be a valid alternative to the domain–domain interaction in a homologous structure. Therefore, the overlap at the interface was also investigated by analyzing equivalences for all available interfaces in a given pairwise superfamily. The reference interfaces were classified at the pairwise superfamily level and if any of the reference interfaces within a pairwise superfamily classification had equivalent alternative interfaces then the entire superfamily of interfaces was consid-

**Table II**
*The Percentage of Pairwise Superfamilies Where at Least One Member of the Pairwise Superfamily Has an Equivalent Alternative Interaction for Different Thresholds of Percentage Overlap*

| | Percentage of nonredundant interfaces | | | |
|---|---|---|---|---|
| | CATH referenced to SCOP alternative | | SCOP referenced to CATH alternative | |
| Percentage overlap | Superfamily pairs with equivalent | Superfamily pairs without equivalent | Superfamily pairs with equivalent | Superfamily pairs without equivalent |
| 25 | 87.8 (1721) | 12.2 (240) | 82.5 (1427) | 17.5 (302) |
| 50 | 86.6 (1699) | 13.4 (262) | 81.1 (1402) | 18.9 (327) |
| 75 | 84.7 (1660) | 15.3 (301) | 77.8 (1346) | 22.2 (383) |
| 100 | 79.4 (1558) | 20.6 (403) | 67.3 (1164) | 32.7 (565) |

For example, if the overlap was ≥75%, then 77.8% of the SCOP pairwise superfamilies had an equivalent CATH superfamily level interaction. Numbers in parentheses are the actual numbers of such interfaces found in each case.

ered to have an equivalent interface. So, for a reference interface to be considered to have an equivalent at 75% overlap, at least one of the interactions belonging to the same pairwise superfamily classification needed to have an equivalent interface, which covered at least 75% of the interface area.

Table II shows that at 75% overlap 77.8% of the SCOP pairwise superfamily interfaces have an equivalent CATH interface and 84.7% of the CATH pairwise superfamily interfaces have an equivalent SCOP interface. The percentage of nonredundant interfaces with equivalent interfaces decreases steadily as the degree of overlap at the interface increases. At 25% overlap, 82.5% of SCOP interfaces and 87.8% of the CATH interfaces had equivalents decreasing to 67.3% and 79.4%, respectively, for 100% overlap.

The difference in the overlap again shows that the interfaces derived using the SCOP classification were less well covered by CATH equivalents than CATH interfaces were covered by equivalent SCOP interfaces. This can partly be explained by the larger size of SCOP domains and the SCOP interfaces thereby providing a greater coverage of the equivalent CATH interfaces than CATH interfaces provide of equivalent SCOP interfaces. These results suggest that, as the coverage of CATH interfaces by SCOP was greater than the coverage of SCOP interfaces by CATH, SCOP should be a better choice of domain definition for analyzing protein–protein interactions if interface coverage is the principal criterion. However, using both SCOP and CATH in conjunction leads to a significant increase in the number of interfaces observed. As the shortage of domain–domain interfaces is a limiting factor in the use of structural data in protein–protein interaction investigation and prediction and the majority of work in this area is performed using only one domain classification system this result represents a significant finding.

To illustrate this point, if CATH is used as the domain classification system for prediction of protein–protein interactions through homology matching of sequences to domains observed interacting in structural data then there would be 1961 pairwise superfamiles with which to match. If the domain–domain interactions observed in SCOP that had no CATH equivalent were added there would be 2606 pairwise superfamilies with which to match (1961 + 645). Aloy *et al.* predicted that the number of types of protein–protein interaction is 10,000.[7] If a pairwise superfamily is considered to be a type of protein–protein interaction then just using CATH would mean that 19.6% (1961/10,000) of protein–protein interaction types are observed in structural data. Including the SCOP pairwise superfamily interactions where there is no equivalent CATH interaction would increase this value to 26.1% (2606/10,000). This could reduce the false negative rate of predictive methods, which employ homology matching to structural data (such as[7, 8, 9]) by ~6.5%.

## Reasons for no equivalent interface

The interfaces of domain–domain interactions where there was no equivalent interface observed were classified according to whether one or both of the interacting domains had no corresponding domain in the other classification system. Figure 2 shows the different classifications of interaction site state. If there was no equivalent interface because there was no alternative domain present for one of the interface partners then this is described as "one absent." If there was no alternative domain present for both of the interface partners then this is described as "both absent." If there were alternative domains present for both interface partners but these domains were not present at the same area of the interface then this is described as "both present but not overlapping at the interface." This can happen when the interaction site is large and where part of one half of an interaction site is covered by an alternative domain and part of the other half the interaction is also cov-

 DOI 10.1002/prot

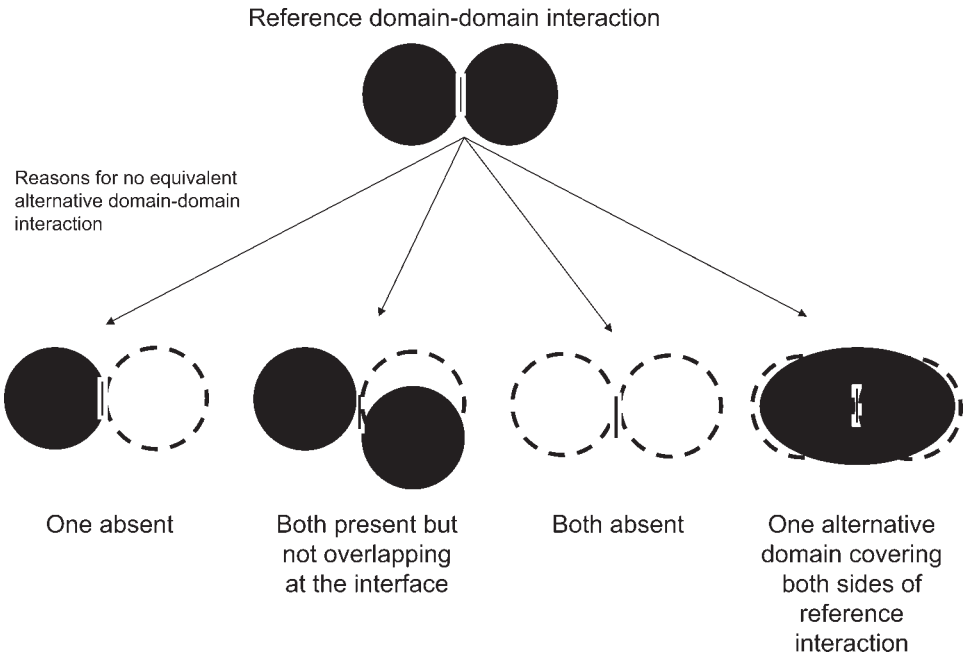


**Figure 2**

*Examples of situations where there is no equivalent alternative interface for a reference interface. "One absent" is where there is no equivalent interface because one of the reference domains has no equivalent in the alternative classification. "Both absent" is where there are no equivalent domains for both of the domains in the reference pair. "Both present but not overlapping at the interface" is where there are equivalents for both domains in the reference interface but the coverage of the interface by the alternatives is too small. "One alternative domain covering both sides of reference interaction" is where one alternative domain covers both of the reference domains observed interacting.*

ered by an alternative domain; however, these two parts are not actually covering the same region in the interaction site. If there was one alternative domain covering both of the reference domains then this is described as "one alternative domain covering both sides of reference interaction." A nonredundant set was determined by taking a random sample of 10 interactions from each pairwise superfamily where a non-equivalent interface had been found.

Table III shows the properties of the interfaces, which had less than 75% interface site overlap. As only interactions where there is no equivalent interface are considered the numbers of interactions are considerably smaller than in Tables I and II. When there was no equivalent SCOP interface for a CATH interface, 86.4% of the interfaces were in the "one alternative domain covering both sides of reference interaction" category (93.4% for a redundant data set). As the number of known proteins increases this problem may be alleviated by an increase in the number of SCOP domains seen as a separate domain in isolation or in more than one context in different multi-domain proteins. CATH defining the region as a single domain but SCOP classifying the region as two separate domains was not as important a reason for why there was no CATH equivalent interface for SCOP reference domain–domain interactions with only 22.5% (40.0% for the redundant data set) of the interfaces with no equivalent being in this category. The reasons why there were no CATH equivalent interfaces for SCOP reference domain–domain interactions were split more evenly across the four different scenarios.

## Number of equivalent alternative interfaces per reference interface

Protein chains can be split into several domains by one classification and considered to be only one domain by the other classification.[19, 20] This situation may cause one reference interface to comprise several equivalent alternative interfaces. For example, in Figure 1, the reference interface comprises two equivalent alternative interfaces. Table IV shows the number of equivalent alternative interfaces for each reference interface. The nonredundant set was determined by taking a random sample of 10 interactions from each pairwise superfamily. The size of the nonredundant set is therefore an order of magnitude greater than is seen in Tables I and II. About 70.9% (83.4% redundant set) of the SCOP interfaces had

**Table III**
*Percentage of Interfaces That Had No Equivalent Alternative Domain–Domain at the 75% Overlap Threshold*

| | Percentage of domain–domain interactions | | | |
|---|---|---|---|---|
| | CATH referenced to SCOP alternative | | SCOP referenced to CATH alternative | |
| | Redundant set | Nonredundant set | Redundant set | Nonredundant set |
| Both present but not overlapping at the interface | 5.4 (853) | 12.0 (73) | 28.8 (3777) | 28.2 (2461) |
| One alternative domain covering both sides of reference interaction | 93.4 (14729) | 86.5 (5238) | 40.0 (4459) | 22.5 (2026) |
| One absent | 1.2 (194) | 1.5 (91) | 25.2 (3317) | 33.5 (3030) |
| Both absent | 0 | 0 | 12.0 (1578) | 17.0 (1533) |

Numbers in parentheses are the actual numbers of such interfaces found in each case. The assignments of "both present but not overlapping at the interface," "one alternative domain covering both sides of reference interaction," "one absent," and "both absent" are explained in Figure 2.

only a single equivalent CATH interface and 83.2 (85.8% redundant set) of the CATH interfaces had only a single SCOP equivalent. For the fully redundant set, there were 2 SCOP interfaces, where each had 8 equivalent CATH interfaces, 18 SCOP interfaces that each had 6 equivalent CATH interfaces, and 45 SCOP interfaces that each had 5 equivalent CATH interfaces. There were no CATH interfaces, which had more than four equivalent SCOP interfaces.

The SCOP interactions, which had eight equivalent CATH interfaces were both formed between two copies of Nitrogenase iron protein-like domain (SCOP family c.37.1.10), in PDB entries 1loo[30] and 1j4b.[31] Although these are seen as single domains in the PDB entries, PQS[22] which was employed for this study, shows the biological unit to be a dimer. CATH divides the single domain assigned by SCOP into three separate domains. Therefore, in the PQS structure, one interaction was observed between the two SCOP domains but when the CATH domains were analyzed the six CATH domains were seen to form eight different interfaces along the interaction surface of the SCOP interaction. This example shows the advantages of using PQS as the structures of 1loo[30] and 1j4b[31] are those proposed as the biological units by the authors of the original structures.

The SCOP interfaces which had five or more CATH equivalents were found to be classified into four different SCOP family level interactions, all of which were homofam-pairs (both interaction partners have the same family classification). These domains included the Reverse transcriptase domain (e.8.1.2), the L-aspartase/fumarase domain (a.127.1.1), the Cystathionine synthase-like domain (c.67.1.3), and the Nitrogenase iron protein-like domain (discussed earlier).

**Table IV**
*Percentage of SCOP and CATH Interfaces That Had a Given Number of Equivalent Alternative Interfaces*

| Number of equivalent alternative interactions | Percentage of domain–domain interactions | | | |
|---|---|---|---|---|
| | CATH referenced to SCOP alternative | | SCOP referenced to CATH alternative | |
| | Redundant set | Nonredundant set | Redundant set | Nonredundant set |
| 0 | 13.1 (12089) | 15.7 (3076) | 8.3 (6616) | 23.5 (4064) |
| 1 | 85.8 (79334) | 83.2 (16311) | 83.4 (66440) | 70.9 (12253) |
| 2 | 0.9 (865) | 1.1 (214) | 6.5 (5184) | 4.3 (750) |
| 3 | 0.1 (131) | 0.0 (8) | 1.3 (1070) | 1.0 (173) |
| 4 | 0.0 (4) | 0.0 (1) | 0.3 (266) | 0.2 (40) |
| 5 | 0 | 0 | 0.0 (45) | 0.1 (10) |
| 6 | 0 | 0 | 0.0 (18) | 0 |
| 7 | 0 | 0 | 0.0 (0) | 0 |
| 8 | 0 | 0 | 0.0 (2) | 0 |

The numbers shown in brackets are the actual number of interactions.


DOI 10.1002/prot

## CONCLUSIONS

The effect of including data from both CATH and SCOP domain assignments on observed domain–domain interactions has been systematically investigated. The general conclusions are:

- About 23.6% of CATH interfaces had no SCOP equivalent and 37.3% of SCOP interfaces had no CATH equivalent. Therefore, using both SCOP and CATH domain classification systems significantly increases the number of domain–domain interaction interfaces observed in structural data. As the shortage of domain–domain interfaces is a limiting factor in the use of structural data in protein–protein interaction investigation and prediction this result represents a significant finding.
- If only one domain classification is to be used, SCOP should be a better choice of domain definition for analysing protein–protein interactions since the coverage of CATH interfaces by SCOP was greater than the coverage of SCOP interfaces by CATH.
- Employing both SCOP and CATH reduces the false negative rate of predictive methods, which employ homology matching to structural data to predict protein–protein interaction, by an estimated 6.5%.
- About 86.5% of the structures in which CATH interfaces had no equivalent SCOP interface can be attributed to SCOP assigning a single domain to the region while CATH assigns two separate domains.
- Only 22.5% of the structures in which SCOP interfaces had no equivalent CATH interface can be attributed to CATH assigning a single domain to the region while SCOP assigns two separate domains.

## ACKNOWLEDGMENTS

We thank the MSD group at EBI for discussions and information and Dr Jonathan Monk and Mr Eduardo Damato for network and systems support.

## REFERENCES

1. Jones S, Thornton JM. Principles of protein–protein interactions. Proc Natl Acad Sci USA 1996;93:13–20.
2. Lo Conte L, Chothia C, Janin J. The atomic structure of protein–protein recognition sites. J Mol Biol 1999;285:2177–2198.
3. Teichmann SA. Principles of protein–protein interactions. Bioinformatics 2002;18 (Suppl 2):S249.
4. Fariselli P, Pazos F, Valencia A, Casadio R. Prediction of protein–protein interaction sites in heterocomplexes with neural networks. Eur J Biochem 2002;269:1356–1361.
5. Hoskins J, Lovell S, Blundell TL. An algorithm for predicting protein–protein interaction sites: abnormally exposed amino acid residues and secondary structure elements. Protèin Sci 2006;15:1017–1029.
6. Murakami Y, Jones S. SHARP2: protein–protein interaction predictions using patch analysis. Bioinformatics 2006;22:1794–1795.
7. Aloy P, Russell RB. Interrogating protein domain–domain interaction networks through structural biology. Proc Natl Acad Sci USA 2002;99:5896–5901.
8. Pieper U, Eswar N, Braberg H, Madhusudhan MS, Davis FP, Stuart AC, Mirkovic N, Rossi A, Marti-Renom MA, Fiser A. MODBASE, a database of annotated comparative protein structure models, and associated resources. Nucleic Acids Res 2004;32(Database issue):D217–D222.
9. Lu L, Lu J, Skolnick H. MULTIPROSPECTOR: an algorithm for the prediction of protein–protein domain–domain interactions by multimeric threading. Proteins 2002;49:350–364.
10. Janin J, Henrick K, Moult J, Eyck LT, Sternberg MJ, Vajda S, Vakser I, Wodak SJ. Capri: a critical assessment of predicted interactions. Proteins 2003;52:2–9.
11. Westbrook J, Feng Z, Jain S, Bhat TN, Thanki N, Ravichandran V, Gilliland GL, Bluhm W, Weissig H, Greer DS. The protein data bank: unifying the archive. Nucleic Acids Res 2002;30:245–248.
12. Aloy P, Russell RB. Ten thousand domain–domain interactions for the molecular biologist. Nat Biotechnol 2004;22:1317, 1321.
13. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
14. Lo Conte L, Brenner SE, Hubbard TJP, Chothia C, Murzin A. SCOP database in 2002: refinements accommodate structural genomics. Nucleic Acids Res 2002;30:264–267.
15. Andreeva A, Howorth D, Brenner SE, Hubbard TJP, Chothia C, Murzin AG. SCOP database in 2004: refinements integrate structure and sequence family data. Nucleic Acids Res 2004;32:D226–D229.
16. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH—a hierarchic classification of protein domain structures. Structure 1997;5:1093–1108.
17. Pearl FMG, Lee D, Bray JE, Sillitoe I, Todd AE, Harrison AP, Thornton JM, Orengo CA. Assigning genomic sequences to CATH. Nucleic Acids Res 2000;28:277–282.
18. Pearl F, Todd A, Sillitoe I, Dibley M, Redfern O, Lewis T, Bennett C, Marsden R, Grant A, Lee D, Akpor A, Maibaum M, Harrison A, Dallman T, Reeves G, Diboun I, Addou S, Lise S, Johnston C, Sillero A, Thornton J, Orengo C. The CATH domain structure database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis. Nucleic Acids Res 2005;33:D247–D251.
19. Hadley C, Jones DT. A systematic comparison of protein structure classifications: SCOP, CATH and fssp. Struct Fold Des 1999;7:1099–1112.
20. Veretnik S, Bourne PE, Alexandrov NN, Shindyalov IN. Toward consistent assignment of structural domains in proteins. J Mol Biol 2004;339:647–678.
21. Jefferson ER, Walsh TP, Barton GJ. SNAPPI-DB: a database and API of structures, interfaces and alignments for protein–protein interactions. Nucleic Acids Res 2007;35(Database issue):D580–D589.
22. Henrick K, Thornton JM. PQS: a protein quaternary structure file server. Trends Biochem Sci 1998;23:358–361.
23. Jefferson ER, Walsh TP, Barton GJ. Biological units and their effect upon the properties and prediction of protein–protein interactions. J Mol Biol 2006;364:1118–1129.
24. Golovin A, Oldfield TJ, Tate JG, Velankar S, Barton GJ, Boutselakis H, Dimitropoulos D, Fillon J, Hussain A, Ionides JM. E-MSD: an integrated data resource for bioinformatics. Nucleic Acids Res 2004;32(Database issue):D211–D216.
25. Velankar S, McNeil P, Mittard-Runte V, Suarez A, Barrell D, Apweiler R, Henrick K. E-MSD: an integrated data resource for bioinformatics. Nucleic Acids Res 2005;33(Database issue):D262–D265.

26. Chothia C. The nature of the accessible and buried surfaces in proteins. J Mol Biol 1976;105:1–12.
27. Park J, Lappe M, Teichmann SA. Mapping protein family domain–domain interactions: intramolecular and intermolecular protein family domain–domain interaction repertoires in the pdb and yeast. J Mol Biol 2001;307:929–938.
28. Aloy P, Ceulemans H, Stark A, Russell RB. The relationship between sequence and domain–domain interaction divergence in proteins. J Mol Biol 2003;332:989–998.
29. Wojcik J, Schachter V. Protein–protein interaction map inference using interacting domain profile pairs. Bioinformatics 2001;17 (Suppl 1):S296–305.
30. Iancu CV, Borza T, Fromm HJ, Honzatko RB. Imp, gtp, and 6-phosphoryl-imp complexes of recombinant mouse muscle adenylosuccinate synthetase. J Biol Chem 2002;277:26779–26787.
31. Iancu CV, Borza T, Choe JY, Fromm HJ, Honzatko RB. Recombinant mouse muscle adenylosuccinate synthetase: overexpression, kinetics, and crystal structure. J Biol Chem 2001;276:42146–42152.


DOI 10.1002/prot