A Comparison of SCOP and CATH with Respect to Domain-Domain Interactions

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

Running title: SCOP and CATH Domain-Domain Interactions

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

*Corresponding Author

Geoffrey J. Barton

School of Life Sciences

University of Dundee

Dow St.

Dundee

DD1 5EH

Scotland

UK

email: geoff@compbio.dundee.ac.uk

Telephone: +44 1382 385860

Fax: +44 1382 345764

Summary

The analysis and prediction of protein-protein interaction sites from structural data, is restricted by 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 paper, 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 non redundant 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%. 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 where as only 22.5% of SCOP interfaces were in this category.

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, 2, 3], prediction of protein

interaction sites [4, 5, 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 non-redundant interacting pairs in the current PDB is approximately 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 unit 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 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 paper, 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 so no direct mapping between the two domain classification systems was attempted here. Instead, this paper 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 paper, the term 'pairwise superfamily' is used to describe, 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.ssemblies from the PQS Database [22] were employed rather than the asymmetric units seen in PDB files since this increases the number of non-redundant 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 [26]) +0.5 Ångstroms. 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, 28, 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 ≥ 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 ≥ 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 coverage of interface sites of SCOP and CATH are compared rather than direct comparison of the single domains.

In the example shown in Figure 1 the reference interface (Figure 1a) consists of 50 interacting residues and is shown in red (domain labelled Ref1) and orange (domain labelled Ref2). The equivalent interface shown in Figure 1b is composed of 3 different alternative interfaces: the first is between domains coloured purple (Alt1) and yellow (Alt2) and has an interface size of 20 interacting residue pairs; the second is between domains coloured purple (Alt1) and green (Alt3) and has 17 interacting residues and the third is between domains coloured cyan (Alt4) and pink (Alt5) and has 3 interacting residue pairs. In this example, the alternative interface between domains coloured cyan (Alt4) coloured 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 number 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 above, there were three possible alternative interfaces, but only 2 of these (Alt1-Alt2 and Alt1-Alt3) had an interface of ≥ 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 comprises 50 residues this gave the percentage overlap of 74%.

Results and Discussion

Several properties of CATH and SCOP domain-domain interactions where 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 SCOP interfaces (mean of 39.8 residues) where generally smaller than CATH interfaces (mean of 45.7 residues) which can be explained by the difference in single domain sizes when comparing SCOP and CATH i.e. SCOP single domains are generally smaller than CATH single domains. The promiscuity of domains was found to be similar for both CATH and SCOP. The frequencies with which domain-domain interactions are observed in different pairwise superfamilies was 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).

Table I summarises 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 these results were not biased by the redundancy of the data, the overlap of the interfaces was also analysed in a non-redundant 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 non-redundant set compared to the redundant set. The analysis of interfaces generated by a non-redundant 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.

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 analysing 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 considered 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 non-redundant 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 analysing 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 approximately 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 do-

main 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'. 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 non-redundant set was determined by taking a random sample of 10 interactions from each pairwise superfamily where 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 number 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 was no CATH equivalent interfaces for SCOP reference domain-domain interactions were split more evenly across the 4 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 2 equivalent alternative interfaces. Table IV shows the number of equivalent

alternative interfaces for each reference interface. The non-redundant set was determined by taking a random sample of 10 interactions from each pairwise superfamily. The size of the non-redundant set is therefore an order or magnitude greater than is seen in Tables I and II. 70.9% (83.4% redundant set) of the SCOP interfaces had 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 which 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 4 equivalent SCOP interfaces.

The SCOP interactions which had 8 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] was employed for this study shows the biological unit to be a dimer. CATH divides the single domain assigned by SCOP into 3 separate domains. Therefore, in the PQS structure one interaction was observed between the two SCOP domains but when the CATH domains were analysed the 6 CATH domains were seen to form 8 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 5 or more CATH equivalents were found to be classified into 4 different SCOP family level interactions, all of which were homo-fam-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 above).

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:

- 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%.

- 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.

Acknowledgements

References

[1] S. Jones and J.M. Thornton. Principles of protein-protein interactions. *Proc Natl Acad Sci USA*, 93(1):13–20, 1996.

[2] L. Lo Conte, C. Chothia, and J. Janin. The atomic structure of protein-protein recognition sites. *Journal Molecular Biology*, 285(5):2177–2198, 1999.

[3] S.A. Teichmann. Principles of protein-protein interactions. *Bioinformatics*, 18(Suppl 2):S249, 2002.

[4] Piero Fariselli, Florencio Pazos, Alfonso Valencia, and Rita Casadio. Prediction of protein–protein interaction sites in heterocomplexes with neural networks. *Eur J Biochem*, 269(5):1356–1361, Mar 2002.

[5] Jemima Hoskins, Simon Lovell, and Tom L Blundell. An algorithm for predicting protein-protein interaction sites: Abnormally exposed amino acid residues and secondary structure elements. *Protein Sci*, 15(5):1017–1029, May 2006.

[6] Yoichi Murakami and Susan Jones. SHARP2: protein-protein interaction predictions using patch analysis. *Bioinformatics*, May 2006.

[7] P. Aloy and R.B. Russell. Interrogating protein domain-domain interaction networks through structural biology. *Proc Natl Acad Sci USA*, 99(9):5896–5901, 2002.

[8] U. Pieper, N. Eswar, H. Braberg, M.S. Madhusudhan, F.P. Davis, A.C. Stuart, N. Mirkovic, A. Rossi, M.A. Marti-Renom, and A. Fiser. MODBASE, a database of annotated comparative protein structure models, and associated resources. *Nucleic Acids Research*, 32(Database issue)(D217-222), 2004.

[9] L. Lu and J. Lu, H. Skolnick. MULTIPROSPECTOR: an algorithm for the prediction of protein-protein domain-domain interactions by multimeric threading. *Proteins*, 49(3):350–364, 2002.

[10] J. Janin, K. Henrick, J. Moult, L. T. Eyck, M. J. Sternberg, S. Vajda, I. Vakser, and S. J. Wodak. Capri: a critical assessment of predicted interactions. *Proteins*, 52(1):2–9, 2003.

[11] J. Westbrook, Z. Feng, S. Jain, T.N. Bhat, N. Thanki, V. Ravichandran, G.L. Gilliland, W. Bluhm, H. Weissig, and D.S. Greer. The protein data bank: unifying the archive. *Nucleic Acids Research*, 30(1):245–248, 2002.

[12] P. Aloy and R.B. Russell. Ten thousand domain-domain interactions for the molecular biologist. *Nature Biotechnology*, 22(10):1317, 1321 2004.

[13] A.G. Murzin, S.E. Brenner, T. Hubbard, and C. Chothia. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal Molecular Biology*, 247(4):536–540, 1995.

[14] L. Lo Conte, S.E. Brenner, T.J.P. Hubbard, C. Chothia, and A. Murzin. SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Research*, 30(1):264–267, 2002.

[15] A. Andreeva, D. Howorth, S.E. Brenner, T.J.P. Hubbard, C. Chothia, and A.G. Murzin. SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Research*, 32:D226–D229, 2004.

[16] C.A. Orengo, A.D. Michie, S. Jones, D.T. Jones, M.B. Swindells, and J.M. Thornton. CATH - a hierarchic classification of protein domain structures. *Structure*, 5(8):1093–1108, 1997.

[17] F.M.G. Pearl, D. Lee, J.E. Bray, I. Sillitoe, A.E. Todd, A.P. Harrison, J.M. Thornton, and C.A. Orengo. Assigning genomic sequences to CATH. *Nucleic Acids Research*, 28(1):277–282, 2000.

[18] F. Pearl, A. Todd, I. Sillitoe, M. Dibley, O. Redfern, T. Lewis, C. Bennett, R. Marsden, A. Grant, D. Lee, A. Akpor, M. Maibaum, A. Harrison, T. Dallman, G. Reeves, I. Diboun, S. Addou, S. Lise, C. Johnston, A. Sillero, J. Thornton, and C. Orengo. The CATH domain structure database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis. *Nucleic Acids Research*, 33:D247–D251, 2005.

[19] C. Hadley and D.T. Jones. A systematic comparison of protein structure classifications: Scop, cath and fssp. *Structure Fold Des*, 7(9):1099–1112, 1999.

[20] S. Veretnik, P.E. Bourne, N.N. Alexandrov, and I.N. Shindyalov. Toward consistent assignment of structural domains in proteins. *Journal Molecular Biology*, 339(3):647–678, 2004.

[21] E.R. Jefferson, T.P. Walsh, and G.J. Barton. SNAPPI-DB: A database and API of structures, interfaces and alignments for protein-protein interactions. *(Submitted)*, 2006.

[22] K. Henrick and J.M. Thornton. PQS: a protein quaternary structure file server. *Trends Biochem Sci*, 23(9):358–361, 1998.

[23] E.R. Jefferson, T.P. Walsh, and G.J. Barton. Biological units and their effect upon the properties and prediction of protein-protein interactions. *(Submitted)*, 2006.

[24] A. Golovin, T.J. Oldfield, J.G. Tate, S. Velankar, G.J. Barton, H. Boutselakis, D. Dimitropoulos, J. Fillon, A. Hussain, and J.M. Ionides. E-MSD: an integrated data resource for bioinformatics. *Nucleic Acids Research*, 32 (Database issue):D211–216, 2004.

[25] S. Velankar, P. McNeil, V. Mittard-Runte, A. Suarez, D. Barrell, R. Apweiler, and K. Henrick. E-MSD: an integrated data resource for bioinformatics. *Nucleic Acids Research*, 33(Database issue):D262–265, 2005.

[26] C. Chothia. The nature of the accessible and buried surfaces in proteins. *Journal Molecular Biology*, 105(1):1–12, 1976.

[27] J. Park, M. Lappe, and S.A. Teichmann. Mapping protein family domain-domain interactions: intramolecular and intermolecular protein family domain-domain interaction repertoires in the pdb and yeast. *Journal Molecular Biology*, 307(3):929–938, 2001.

[28] P. Aloy, H. Ceulemans, A. Stark, and R.B. Russell. The relationship between sequence and domain-domain interaction divergence in proteins. *Journal Molecular Biology*, 332(5):989–998, 2003.

[29] J. Wojcik and V. Schachter. Protein-protein interaction map inference using interacting domain profile pairs. *Bioinformatics*, 17(Suppl 1):S296–305, 2001.

[30] C. V. Iancu, T. Borza, H. J. Fromm, and R. B. Honzatko. Imp, gtp, and 6-phosphoryl-imp complexes of recombinant mouse muscle adenylosuccinate synthetase. *J Biol Chem*, 277(30):26779–26787, 2002.

[31] C. V. Iancu, T. Borza, J. Y. Choe, H. J. Fromm, and R. B. Honzatko. Recombinant mouse muscle adenylosuccinate synthetase: overexpression, kinetics, and crystal structure. *J Biol Chem*, 276(45):42146–42152, Nov 2001.

Figures

Reference domain-
domain interface

Alternative domain-domain interface
involving 2 equivalent interfaces

ref1 ref2

alt1 alt3 alt2 alt4 alt5

(a) (b)

Small interaction site of less than 5
residues so not considered to be
an equivalent interaction

Two domain-domain interactions which
are considered to be large enough to
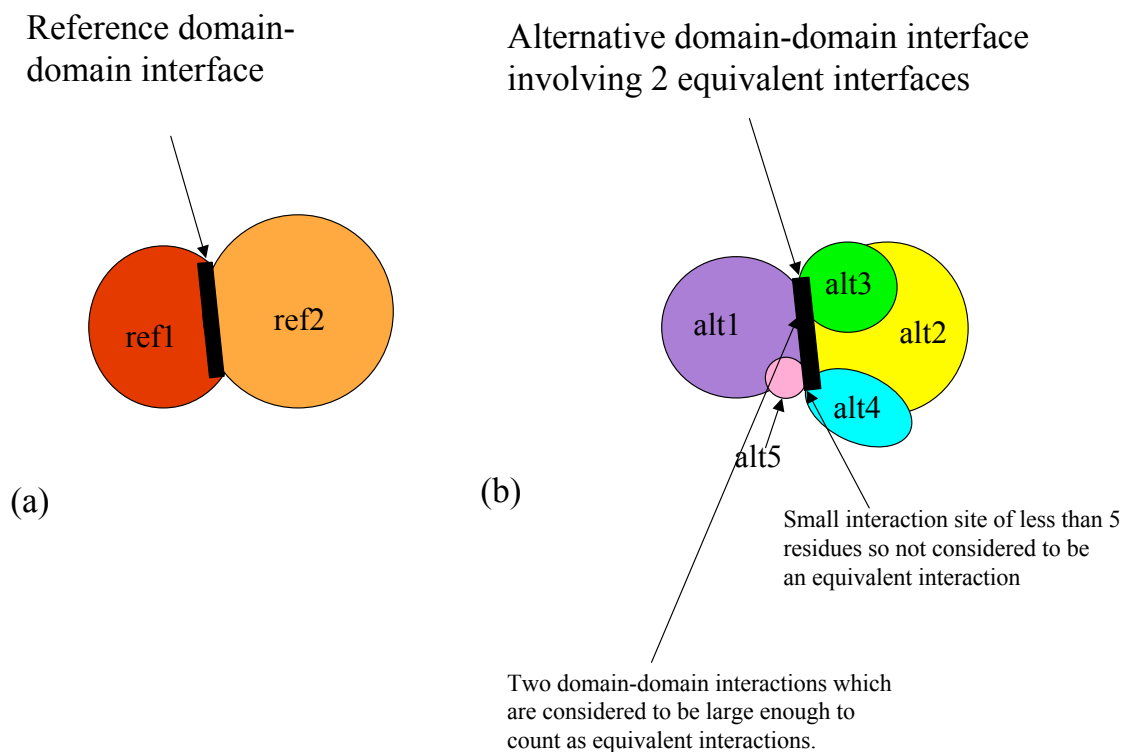count as equivalent interactions.

Figure 1: The determination of equivalent domains. The reference interface is between the domains labelled Ref1 and Ref2. There are 3 interfaces using the alternative domain definition which cover 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.

Reference domain-domain interaction

Reasons for no equivalent
alternative domain-domain
interaction

One absent

Both present but
not overlapping
at the interface

Both absent

One alternative
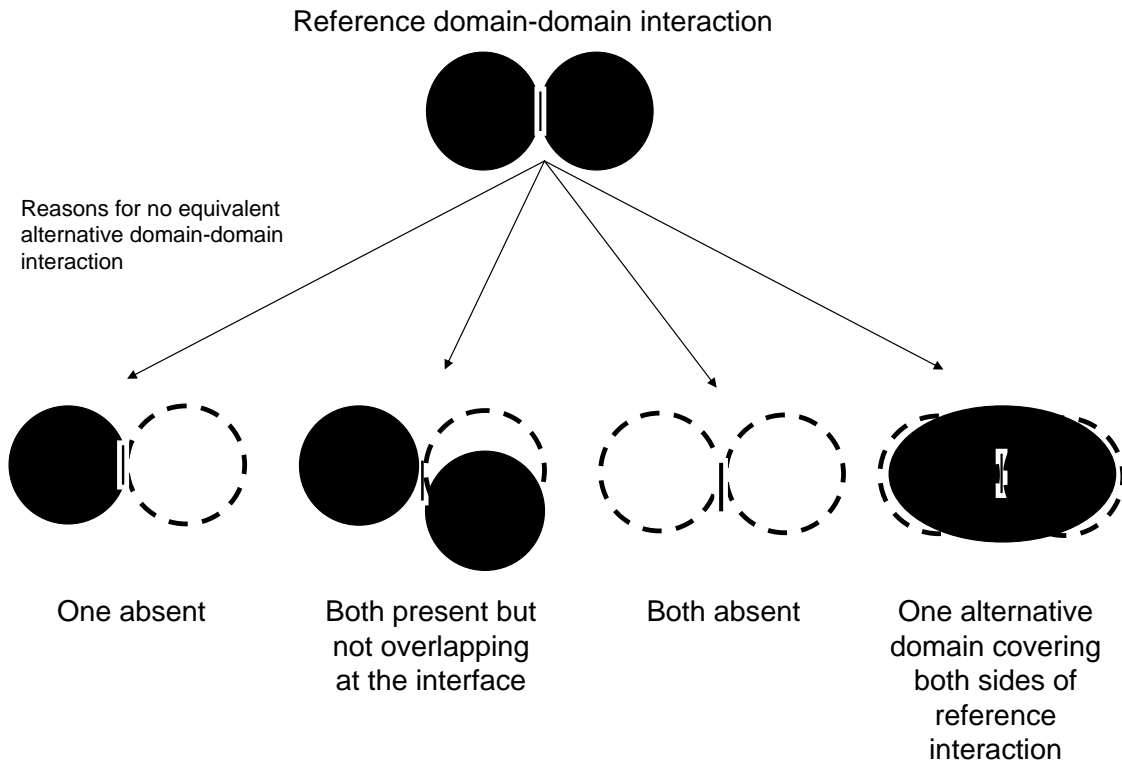domain covering
both sides of
reference
interaction

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.