Supplementary Information

Hamann Distance Function

Bit vectors were compared using the Hamann distance measure, $\text{dist}_{\text{hamann}}$, a rescaled and reversed version of the traditional Hamann similarity coefficient, $\text{sim}_{\text{hamann}}$, developed for use in plant systematics (Hamann, 1961). To compare bit vectors X and Y of length m, the numbers of matches and mismatches for on and off states are first counted. The Hamann distance function is then computed as detailed below (Eq 2).

	$X_i = 1$	$X_i = 0$
$Y_i = 1$	n_a	n_b
$Y_i = 0$	n_c	n_d

$$\text{sim}_{\text{hamann}} = \frac{(n_a + n_d) - (n_b + n_c)}{m} = [-1, 1] \qquad (1)$$

$$\text{dist}_{\text{hamann}} = 1 - \frac{(\text{sim}_{\text{hamann}} + 1)}{2} = [0, 1] \qquad (2)$$

Topological Fingerprints

Complexes

A *domain connectivity graph* is generated for each structure, describing the pattern of domain-domain contacts. In this graph, a domain is a node, and a binary domain interface is an edge (Figure 2(b)). This graph captures the arrangement of the individual domains in the complex. A crude linear representation of this graph is then computed. This string is generated by labeling the nodes with their degree (*ie*, the number of nodes they are connected to). The edges are then listed using the labels of their composite nodes. The edge list is sorted lexically and the resulting string is used as a crude topological fingerprint to group the structures into topological classes (Figure 2(d)). While degeneracy exists in this representation (*ie*, two distinct topologies may have the same sorted edge list), it is useful as both a query and crude clustering term.

Binding Sites and Interfaces

Similar to the domain connectivity graphs, a crude linear representation of the secondary structure topology graph was generated for use as a topological fingerprint. This string is generated by labeling each node on the secondary structure topology graph with their degree and their secondary structure type. The edges are then listed using the labels of their composite nodes. The edge list is sorted lexically and the resulting string is used as a crude fingerprint to group the structures into *topological classes* (Figure 2(c), 2(d)).

Kd-trees algorithm

Inter-atomic distances were computed using an in-house ANSI C implementation, called kd-contacts, of the median kd-trees algorithm (Friedman *et al.*, 1977; Berg *et al.*, 1998). The kd-trees algorithm, a commonly used computational geometry algorithm, performs nearest neighbor queries by first building a tree in $O(n \log n)$ time, and then querying it in $O(n^{1-(1/d)} + k)$ time, where n is the number of data points in the d-dimensional space, and k is the number of reported points. This approach is much faster than the naive approach of all *vs* all distance calculation ($O(n^2)$). The logarithmic scaling allows rapid calculation of contact maps even for large structures with tens of thosuands of atoms, such as PQS entries of virus capsids. Briefly, the algorithm begins by building a binary tree that recursively decomposes the d-dimensional (here, $d = 3$) input space on alternate axes along the median partition. At each branch the splitting value is stored, which allows a unique bounding box of points to be associated with each node in the tree. The result is a non-uniform hypercube binning of d-dimensional space, which is adapted to the actual distribution of data points. A nearest neighbor query, given a query point and radius, is then performed by traversing only those branches that may possibly contain a nearest neighbor according to their bounding box definitions. This procedure allows rapid elimination of entire branches of the tree, leaving only those bins that may contain a nearest neighbor. Distance calculations are required only for the points in these candidate bins.

References

Adai, A. T., Date, S. V., Wieland, S. and Marcotte, E. M. (2004) Lgl: creating a map of protein function with an algorithm for visualizing very large biological networks. *J Mol Biol*, **340**, 179–190.

Berg, M. D., Kreveld, M. V., Overmars, M. and Schwarzkopf, O. (1998) *Computational Geometry: Algorithms and Applications*. Springer Verlag, Berlin, 2nd edition.

Friedman, J. H., Bentley, J. L. and Finkel, R. A. (1977) An algorithm for finding best matches in logarithmic expected time. *ACM Trans Math Software*, **3**, 209–226.

Hamann, U. (1961) Merkmalsbestand und verwandtschaftsbeziehungen der farinosae. ein beitrag zum system der monokotyledonen. *Willdenowia*, **2**, 639–768.

Figure Legends

1. Distribution of buried solvent accessible surface area in interacting SCOP domain pairs. The distribution is calculated from the non-redundant set of interacting domain pairs (Methods). The largest interface is 7225 Å^2 (not shown). The dashed vertical lines shows the cutoff on the solvent accessible surface area used to obtain interfaces for subsequent annotation and analysis.

2. Topology properties calculated for each structure. Color represents the SCOP classification of the domains (blue, g.3.9.1; red, g.3.11.1; black, c.10.2.5). (a) Ribbon and surface representation of the complex of human epidermal growth factor and receptor extracellular domains (PQS entry 1ivo_1.mmol). The interfaces are colored in green. (b) Graph representation of the domain connectivity. Solid lines represent intra-chain interfaces; Dashed lines represent inter-chain interfaces. (c) Graph representation of the binding site and interface topology for interface 4, as defined in (b). The interface topology is defined by the subgraph containing only dashed edges. The topologies of the two interacting binding sites are defined by the two disconnected subgraphs containing only solid edges. Node shapes represent secondary structure type (triangle is β sheet, circle is α helix, filled box is loop, and open box is unassigned). (d) The topological fingerprints are listed for the overall complex, interface number 4, as defined in (b), and one of its corresponding binding sites. The characters in the interface and binding site fingerprints represent secondary structure type (B is β sheet, H is α helix, T is Loop, and _ = unassigned). Structure visualization by PyMOL (http://pymol.sourceforge.net); Graph

layout by LGL (Adai *et al.*, 2004).

3. PIBASE Interface Property Distributions. All the plots are calculated from the non-redundant set of interfaces. (a) Number of secondary structure elements at the interface. Maximum 253 elements (not shown). (b) Number of continuous sequence segments at the interface. Maximum 148 segments (not shown). (c) Number of structural patches at the interface. Maximum 43 structural patches (not shown). (c) Number of sequence segments *vs* buried surface area ($r^2 = 0.71$).

[Table 1 about here.]

[Figure 1 about here.]

[Figure 2 about here.]

[Figure 3 about here.]

List of Figures

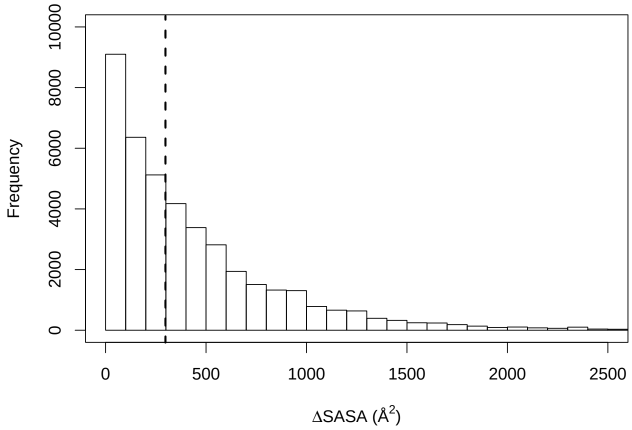
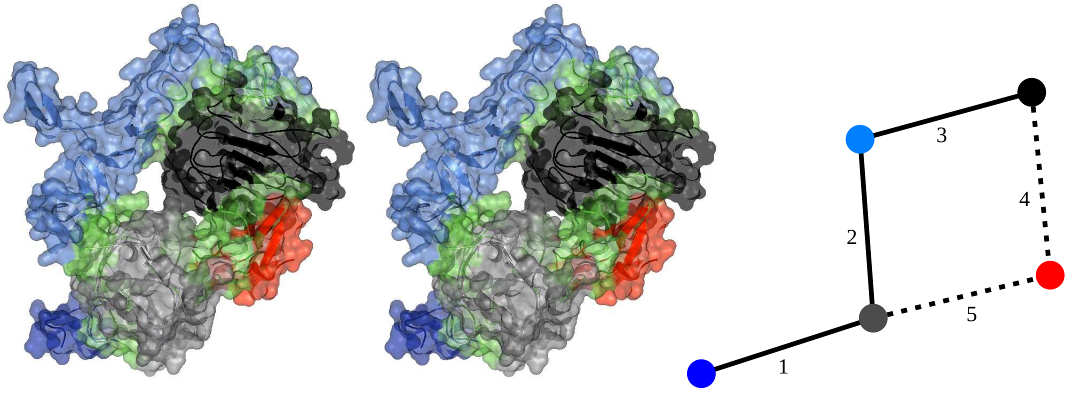
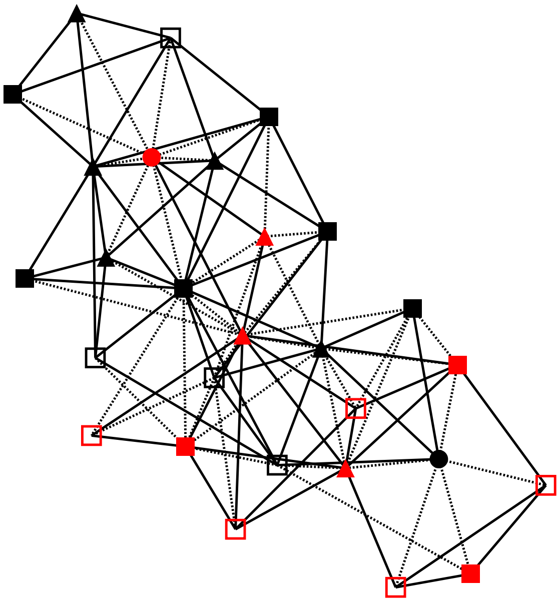


Figure 1: Distribution of buried solvent accessible surface area in interacting SCOP domain pairs.



(a) Structure (PQS 1ivo_1.mmol).

(b) Complex topology.

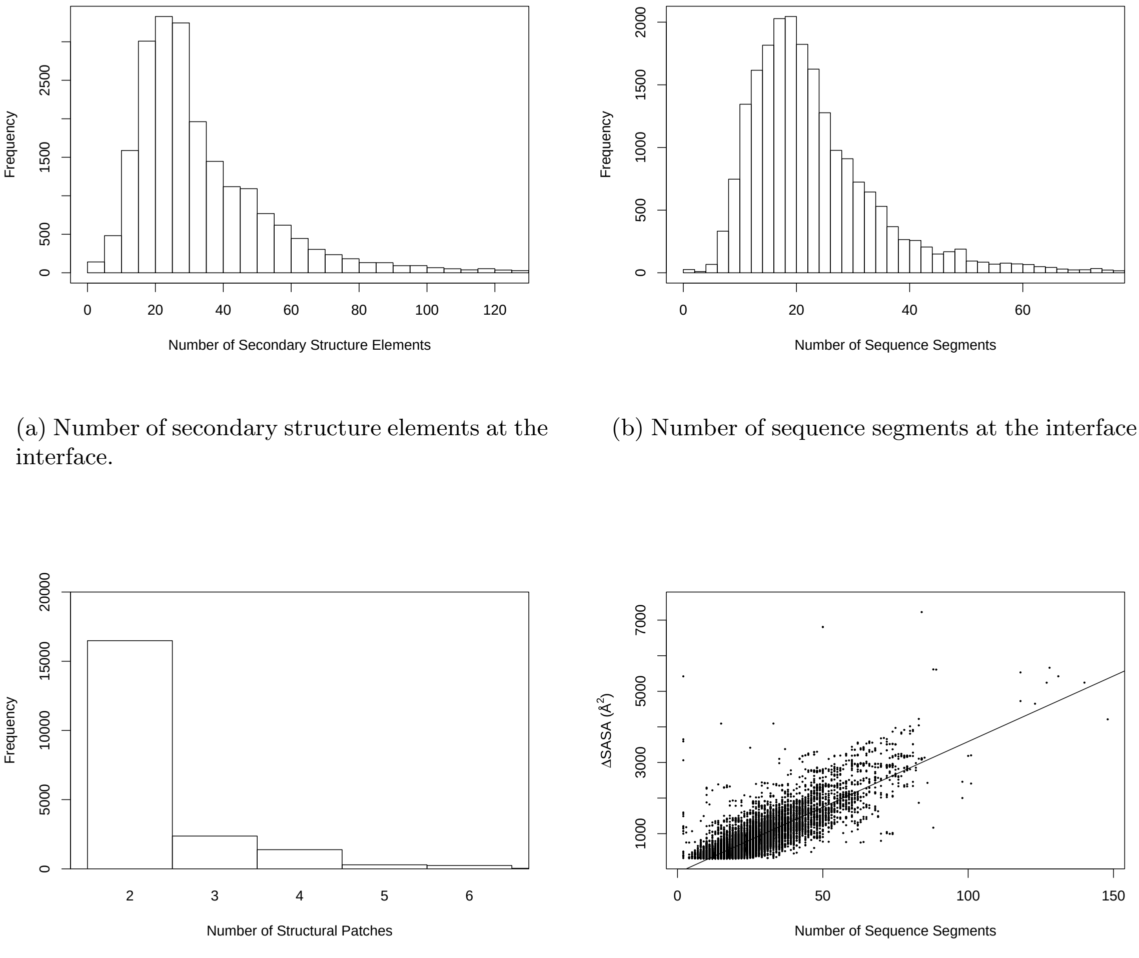


(c) Interface and binding site topologies for interface number 4, as defined in (b).

complex	3-2.3-2.3-1.2-2.2-2
interface 4	3_-2T.4B-3_.4T-2_.4T-3T.4T-4B.5B-2T.5B-2T.5H-1_.5H-1_.5H-2T.5H-3T.5H-4B.5T-1_.5T-2_.5T-3_.5T-5B.5T-5T.5_-1_.5_-2_.5_-5B.5_-5T.6B-2_.6B-3T.6B-4B.6B-5B.6B-5T.7B-1T.7B-2B.7B-2T.7B-4T.7B-5T.7B-5_.7B-6B.8H-1B.8H-1B.8H-1B.8H-1T.8H-1_.8H-2B.8H-2T.8H-5T
red binding site (interface 4)	2H-2B.3_-2T.3_-2T.3_-3_.4T-2_.4T-3_.4_-4T.4_-4T.4_-4_.6B-3_.6B-4T.6B-4T.6B-4_.6B-4_.8B-2B.8B-2H.8B-2_.8B-4T.8B-4T.8B-4_.8B-4_.8B-6B

(d) Topological Fingerprints.

Figure 2: Topology properties calculated for each structure.



(a) Number of secondary structure elements at the interface.

(b) Number of sequence segments at the interface.

(c) Number of structural patches at the interface.

(d) Number of Sequence Segments *vs* Buried Surface Area.

Figure 3: PIBASE Interface Property Distributions.

List of Tables

	(a) *Complexes*
1	Domain connectivity graph

	(b) *Domains*
1	Solvent accessible surface area
2	Secondary structure assignment
3	Domain classification code

	(c) *Interfaces and Binding Sites*		
		Interface	Binding Site
1	Buried solvent accessible surface area (ΔSASA)	✓	
2	Buried polar solvent accessible surface area ($\Delta\text{SASA}_{polar}$)	✓	
3	Number of residues	✓	✓
4	Residue types present	✓	✓
5	Number of secondary structure elements	✓	✓
6	Secondary structure types present	✓	✓
7	Number of structural patches	✓	✓
8	Number of sequence segments	✓	✓
9	Number of residue contacts	✓	✓
10	Residue contact types	✓	✓
11	Inter-atomic contacts (distance binned) ($\leq$ 4.5, 5, 5.5, 6.05 Å)	✓	✓
12	Number of secondary structure element contacts	✓	✓
13	Secondary structure type contacts	✓	✓
14	Secondary structure topology	✓	✓
15	Number of H bonds	✓	
16	Number of salt bridges	✓	
17	Number of disulfide bridges	✓	

Table 1: PIBASE properties. The computed properties characterize each complex at four levels: the overall complex, its constituent domains, interfaces, and binding sites. A subset of the properties are later used to remove interface redundancy and cluster the complexes, interfaces, and binding sites into topological classes.