research articles Journal of proteome research

Protein−Protein Interfaces: Properties, Preferences, and Projections

Jeffrey J. Headd, Y. E. Andrew Ban, Paul Brown, Herbert Edelsbrunner, Madhuwanti Vaidya, and Johannes Rudolph*

Department of Biochemistry, Duke University Medical Center, Durham, North Carolina 27710

Received January 12, 2007

Herein, we study the interfaces of a set of 146 transient protein−protein interfaces in order to better understand the principles of their interactions. We define and generate the protein interface using tools from computational geometry and topology and then apply statistical analysis to its residue composition. In addition to counting individual occurrences, we evaluate pairing preferences, both across and as neighbors on one side of an interface. Likelihood correction emphasizes novel and unexpected pairs, such as the His−Cys pair found in most complexes of serine proteases with their diverse inhibitors and the Met−Met neighbor pair found in unrelated protein interfaces. We also present a visualization of the protein interface that allows for facile identification of residue−residue contacts and other biochemical properties.

Keywords: protein−protein interactions • protein interfaces • geometric topology • visualization

Introduction

Protein−protein interactions play a significant role in the majority of intracellular processes, and understanding how proteins transiently form complexes is essential for grasping the nuances of biological systems. While we have yet to derive universal rules that allow us to identify interaction sites a priori or to reliably predict protein docking, formation of protein complexes and their subsequent stability have been linked to a few specific interfacial residues. These residues, commonly called hotspots, contribute the bulk of the binding energy between proteins, whereas a majority of other residues apparently serve as tolerant bystanders. Thus, the composition of surface residues involved in transient interactions is important to their function. With the ever increasing number of available high-resolution structures of protein complexes, studying the residue composition and pairing preferences of known protein−protein interfaces allows a better understanding of the fundamentals of protein−protein association.

Many prior studies have examined the composition of protein−protein recognition sites from a number of different perspectives.[1−21] The work of Chakrabarti and Janin[1] has shown these sites have an amino acid composition similar to the overall protein surface. They break the sites down to core and rim regions, with the core having a composition different from the rest of the protein surface, suggesting that protein interfaces possess some unique characteristics. Glaser et al.[2] studied the pairing preferences for a much larger number of protein−protein complexes by including homo-dimeric complexes in addition to transient hetero-complexes. Though statistically more rigorous, their work does not take into account the inherent interfacial differences between transient hetero-complexes and obligate homo-dimers, with the latter more resembling protein interiors. Ofran and Burkhard[3] have accounted for these differences by classifying protein−protein interfaces into six categories based on interaction type and studying them individually. Mintz et al.[20] have developed a method for clustering similar interfaces based on shape and location of chemical functional groups from the entire PDB wherein the overall data set is again biased toward homo-dimers. Ma et al.[15] have focused on elucidating the structurally conserved residues at protein−protein interfaces. Although these prior statistical studies have given us a foundation for thinking about protein−protein interactions, with a few exceptions,[19,22] they have not found significant applications to the prediction of protein interaction sites, docking of proteins, and the identification of hotspot residues.

This lack of progress in employing the results of previous statistical studies can be attributed to two key drawbacks. First, the limited number of structurally characterized transient protein−protein complexes introduces significant bias, and the resulting statistics have potentially limited applications.[1,8] As noted above, larger data sets contain disproportionately more homo-dimeric complexes[2,3,20] or are assembled through automated processes that may include nonphysiological protein−protein pairs such as crystal contacts. Second, the lack of a rigorous definition of what constitutes a protein interface makes reliable automation for broader scale analysis difficult. Frequently, studies have relied on either a distance cutoff (e.g., at most 6 Å between atoms across the interface)[2,3] and/or an area cutoff (e.g., at least 0.1 Å^2 increase in solvent accessible area upon separation).[1,2,5−14] The choice of a cutoff, which does not arise from an intrinsic property of protein−protein complexes, can greatly influence the size and composition of the protein interface and frequently gives rise to artificial holes,

* To whom correspondence should be addressed: Johannes Rudolph, Department of Biochemistry, Mailstop 3813 LSRC, C125 Duke University Medical Center, Durham, NC 27710. Tel, (919) 668-6188; fax, (919) 613-8642; e-mail, johannes@alum.mit.edu.

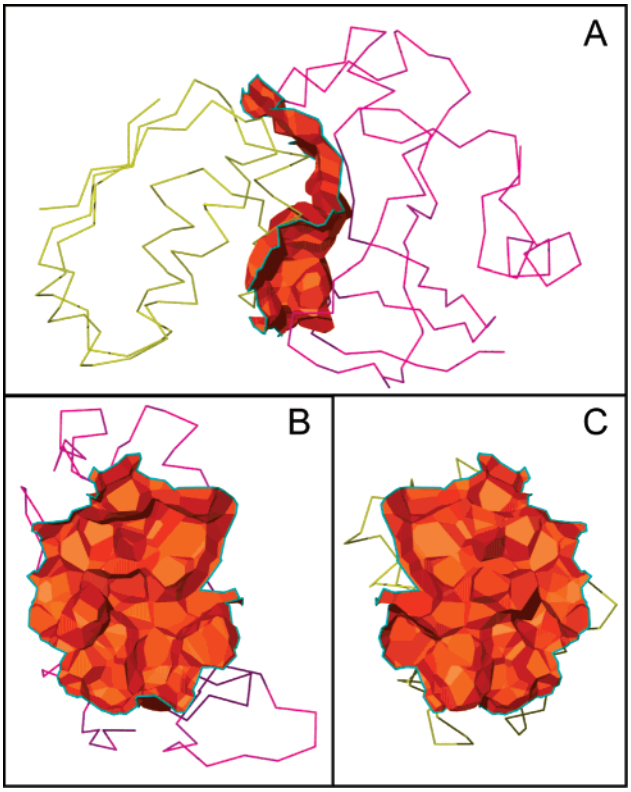

Figure 1. Three-dimensional interface. (A) Interface surface for barnase/barstar (1BRS) complex, with barnase colored purple and barstar colored yellow; (B) image from panel A rotated to show the barstar side of the interface against the barnase protein; (C) image from panel A rotated to show the barnase side of the interface against the barstar protein.

overlaps, or extensions. Thus, we and others have developed alternative definitions of the interface between two proteins based on geometric topology,[21] "halfway points",[16] or molecular contacts.[17,18]

To address the limitations of previous methods and handle a wider range of protein−protein complex types, we base our work on a concrete and unambiguous definition of the interface. Our interface surface is a subset of the Voronoi diagram of the entire complex. As seen in Figure 1, the surface separates two chains in a protein complex and resembles a wrinkled sheet of paper. Its polygonal assembly captures features of the contacts between two chains in more detail than previously described methods, better representing the complexity and complementarity of protein−protein interactions. Its hierarchical construction allows for a retraction of the interface surface to a core region that is highly enriched in hotspot residues.[21]

In this study, we analyze the amino acid composition of the interface of a set of 135 manually selected high-resolution protein−protein complexes yielding 146 protein interfaces. Our database of transient protein interfaces is more than twice as large as previously assembled, hand-culled data sets, and is significantly more diverse. Using our definition of protein interfaces, we not only derive a more rigorous statistical analysis, but are also able to directly measure pairing preferences across and on one side of the surface. Our analysis reveals a number of intriguing results, including a higher than expected contribution of backbone atoms. Finally, we provide a novel flattened view of the interface surface, allowing for facile visualization of interfacial residue composition, residue contacts, and comparison of homologous structures.

Experimental Section

Interface Construction. We define a protein−protein interface surface as described previously.[21] Briefly, the surface formed by two or more proteins is a subset of the Voronoi diagram of the set of spheres making up the space-filling diagrams of the involved proteins. We center a space-filling ball at each atom, and grow the balls simultaneously such that the Voronoi cells do not change. As a ball grows, we clip it to within its Voronoi cell, and each time two or more such clipped balls intersect, we add the convex hull of their centers (generically a simplex) to K, which is the dual complex of the Voronoi diagram. This process creates a filtration of dual complexes, adding simplices to K until K equals the Delaunay triangulation (the dual of the Voronoi diagram). Simplices that are formed at the same time enter the complex at the same time t, which defines the ordering of the filtration. This ordering gives each simplex a rank value, which is assigned chronologically. Participating atoms at the interface are those that share Voronoi polygons in K with an atom on a complementary protein chain. The collection of Voronoi polygons that belong to a single residue are denoted a *tile*. Tiles are defined on both sides of the surface, defining two tilings, one for each of the two proteins separated by the surface. Overlapping tiles correspond to interchain pairs and adjacent tiles correspond to intrachain pairs.

In our analysis, we use the ordered filtration to unambiguously delineate between boundary and core regions of the interface surface through a retraction process that isolates those polygons whose rank value is less than or equal to the median rank value for a given surface. We define this subset as the *core* of an interface.

Since our surface defines atom pairs across the interface, residue pairs easily follow. In addition, the total number of atom pairs per residue pair can be determined, clarifying the extent of the interaction. We can also identify residue neighbor pairs on the same side of the interface, defined by atoms whose Voronoi polygons share an adjacent edge in the interface surface. Finally, the concrete definition allows for normalization by area and perimeter contributions of residues.

Interface Flattening. For the purpose of visualization, we fully triangulate the interface surface and then map the triangulation into the plane, effectively flattening the surface. Almost all surfaces defined by only two proteins are simply connected and can therefore be flattened to a round disk in the plane. While the disk does not necessarily represent the general shape of the surface, it is easy to view, and it lends itself to comparing different interfaces. Importantly, the flattening process preserves all connectivity information, both across the surface and between neighboring tiles. If a complex consists of three or more proteins, we flatten the sheets defined by the pairs separately. Flattening is only performed once per interface, and retracted regions are removed from display, leaving the remnants of the original disk to represent the remaining interface surface. Finally, in the uncommon case in which an interface surface has nonzero genus, we have to cut the surface to remove the genus before flattening. For an example of an interface with nonzero genus, see the neurotoxic vipoxin complex from Western Sand Viper, PDB code 1JLT, as depicted in Figure 8 in Ban et al.[21]

The algorithm we use for flattening is based on a theorem by Tutte[23] which states that a convex mapping of a simple,

3-connected, planar graph is a valid straight-line embedding of the graph. Specifically, if we draw the boundary of the surface as a convex curve in the plane and we express the image of each interior vertex as a convex combination of the images of its neighbor vertices, then Tutte's theorem guarantees that we indeed have a flattening of the triangulation. Let N be the total number of vertices and $n < N$ the number of interior vertices. The most straightforward implementation of Tutte's theorem maps the $N - n$ boundary vertices to equally spaced points in sequence along a unit circle, and it solves a system of linear equations to compute the image of the interior vertices. Letting v_1 to v_N be the vertices of the surface triangulation and μ the map to the plane, the equation for the ith vertex is

$$\mu(v_i) = \sum_{j=1}^{N} \lambda_{i,j} \cdot \mu(v_j)$$

where $\lambda_{i,j} = 0$ if (v_i,v_j) is not an edge in the triangulation, $\lambda_{i,j} = 1/d_i$ if (v_i,v_j) is an edge, d_i is the number of neighbor vertices of v_i, and v_i is assumed to be an interior vertex. We have n linear equations in n unknowns and therefore a unique solution. Although the system of equations can be big, it is necessarily sparse and therefore permits efficient computation.

We refer to the above implementation of Tutte's theorem as the *uniform method* because it treats the neighbors of each vertex the same way. While it produces topologically accurate flattened images, it introduces a significant amount of distortion. To reduce the distortion, we use a more sophisticated implementation of Tutte's theorem referred to as the *mean value coordinates method* as described by Floater.[24] The boundary vertices are again mapped in sequence to points on a unit circle, but the spacing along the circle is chosen according to the lengths of the boundary edges. The second difference to the uniform method is in the choice of the $\lambda_{i,j}$. To describe the new weights, we assume v_j is a neighbor of v_i and let $\alpha_{i,j}$ and $\beta_{i,j}$ be the angles at v_i of the two triangles that share the edge (v_i,v_j). Setting

$$w_{i,j} = \frac{\tan(\alpha_{i,j}/2) + \tan(\beta_{i,j}/2)}{||v_i - v_j||}$$

the corresponding weight used in the linear systems is

$$\lambda_{i,j} = \frac{w_{i,j}}{\sum_{l=1}^{N} w_{i,l}}$$

assuming $w_{i,j} = 0$ if v_i and v_l are not neighbors. For a further description of how the weights $w_{i,j}$ are calculated and affect the flattening, see Floater.[24] It is easy to see that the $\lambda_{i,j}$ are non-negative and add up to 1, if summed over all j. In words, the weights express each interior vertex as a convex combination of its neighbors; hence, Tutte's theorem applies.

To assess the difference between the two implementations, we measure area distortion as a weighted sum of square ratios of normalized areas of triangles. Specifically, we normalize such that the (one-sided) area of the interface surface is 1 and the area of its image (the disk in the plane) is 1. Denoting a triangle in the surface by t and its image by $\mu(t)$, we define

$$D = \sum_{\text{triangles } t} \left[\frac{\text{area}(\mu(t))}{\text{area}(t)}\right]^2 \text{area}(t)$$

Table 1. The Data Set Used for All Statistical Analysis, Consisting of 135 Complexes Hand-Culled from the PDB

A0O	1BKD	1DN2	1FSS	1JHL	1MLC	1SG1	1WEJ	2KAI
1A0R	1BQL	1DQJ	1FVC	1JQJ	1NCA	1SPB	1Y8R	2MTA
1A22	1BRC	1DVF	1G3N	1JRH	1NFD	1STF	1YCQ	2PCC
1A2K	1BRS	1DX5	1GC1	1JTG	1NMB	1TAB	1YCS	2PTC
1A4Y	1BTH	1EER	1GG2	1KB5	1NSN	1TBQ	1YDR	2SIC
1ACB	1BUH	1EFN	1GLA	1KF6	1NVU	1TCO	1YYM	2SNI
1AGR	1BVK	1EFU	1GOT	1KKL	1OMW	1TGS	1Z92	2TEC
1AHW	1BXI	1F47	1GUA	1KXV	1OSP	1TOC	1ZJD	2TRC
1AIP	1C4Z	1FBI	1HIA	1L0Y	1P22	1TT5	2ASS	3HFL
1AK4	1CBW	1FC2	1HWG	1LDK	1PPE	1UDI	2B4J	3HFM
1AO7	1CHO	1FDL	1I1R	1LM8	1PPF	1UGH	2B4S	3HHR
1ATN	1CSE	1FIN	1IAI	1MCT	1PVH	1US7	2B5I	3SGB
1AVW	1DAN	1FLE	1IGC	1MDA	1QFU	1USU	2BTF	3TPI
1AVZ	1DFJ	1FQ1	1JDH	1MEL	1RZK	1UUG	2C2V	4CPA
1BI8	1DHK	1FS1	1JEL	1MKW	1SEB	1VFB	2JEL	4HTC

It is not difficult to see that $D \geq 1.0$ and $D = 1.0$ if, and only if, the area of each triangle is the same as the area of its image. More generally, the smaller D is, the closer μ is to an equiareal map. We have measured the distortion of flattenings as computed by the uniform and the mean value coordinates methods. As shown in Supplemental Table 1 on Supporting Information, the latter method has $D \leq 2.0$ for almost all cases and thus introduces significantly less area distortion than the uniform method with values of D between 5.0 and 64.0. Because of this difference, the mean value coordinates method is used throughout this paper. A visual comparison between the 1BRS interface surface flattened using both methods is shown in Supplemental Figure 1 in Supporting Information.

Data Set. Our data set (Table 1) is a significant augmentation of the set of PDB complexes used by Chakrabarti and Janin.[1] It consists of 135 high-resolution hetero-complexes (ranging from 1.2 to 3.0 Å resolution) containing 146 protein−protein interfaces and 10 296 interfacial amino acids. The distribution of number of interface residues and side-chains versus area of each complex is depicted in Figure 2.

Single Residue Statistics. Letting $\#(a)$ be the number of type a residues that contribute tiles to interface surfaces and $\#_{\text{total}} = \sum_a \#(a)$ the total number of contributing residues (of any type), we define the *relative frequency of residue type a* equal to

$$\text{Prob}[a] = \frac{\#(a)}{\#_{\text{total}}}$$

which is the probability of a randomly chosen contributing residue to be of type a. Similarly, letting $area(a)$ be the total area contributed by type a residues and $area_{\text{total}} = \sum_a area(a)$ the total (two-sided) area of the interface surfaces, we define the *area-weighted relative frequency of residue type a* equal to

$$\text{Prob}_{\text{area}}[a] = \frac{area(a)}{area_{\text{total}}}$$

which is the probability that a randomly chosen point and side of an interface surface belongs to a tile of a type a residue. Both relative and area-weighted relative frequencies will be used to calculate the statistics for single residue occurrences (see Results).

Residue Pair Statistics. We distinguish between *interchain* pairs of residues that are separated by at least one shared Voronoi polygon in the interface surface, and *intrachain* or *neighbor* pairs of residues whose tiles belong to the same side and share at least one Voronoi edge in their boundaries. We note that residue pairs from the same chain are only counted

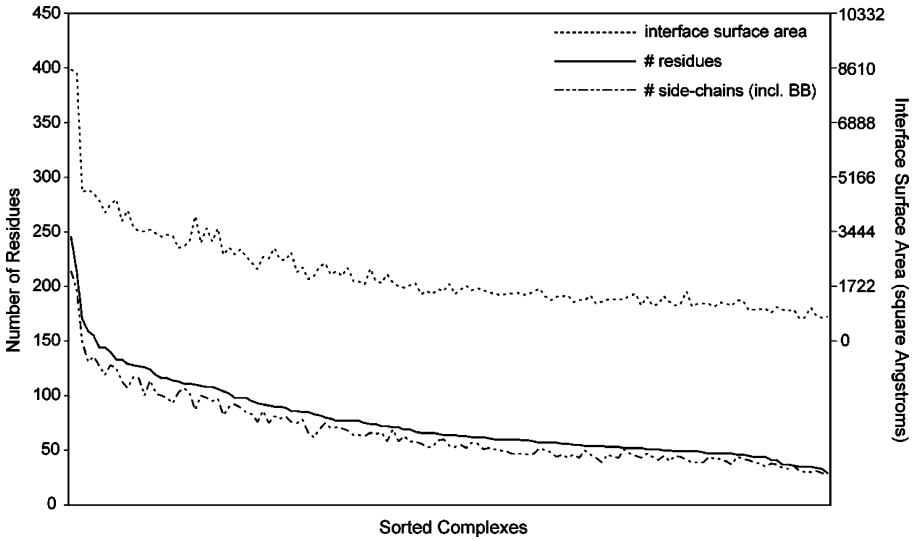

Figure 2. Graphical summary of data sets. From top to bottom: the surface area, the number of residues, and the number of side chains (including backbone as one category) of the complexes sorted by number of residues from left to right. Numbers of residues and side chains are marked on the left, and areas (in Å^2) are marked on the right.

if they are adjacent to each other at the interface surface, not if they are adjacent elsewhere within the protein or they are contiguous along the backbone. It will be convenient to assume an arbitrary but fixed ordering among the residue types so we can write $a \leq b$ by which we mean that a is equal to b or a precedes b in this ordering.

Letting $\#C(a,b)$ be the number of interchain pairs in which one residue is of type a and the other of type b and $C_{\text{total}} = \sum_{a\leq b} \#C(a,b)$ the total number of interchain pairs, we define the *relative frequency of the pair a, b* to be equal to

$$\text{Prob}[a,b] = \frac{\#C(a,b)}{\#C_{\text{total}}}$$

which is the probability that a randomly chosen interchain pair consists of a type a and a type b residue. We note that the pairs are unordered, so $\text{Prob}[a,b] = \text{Prob}[b,a]$. To get area weighted formulas, we observe that $area_{\text{total}} = \sum_{a\leq b} area(a,b)$, where $area(a,b)$ is the total (two-sided) area contributed by pairs in which one residue is of type a and the other of type b. We define the *area-weighted relative frequency of the pair a, b* to be equal to

$$\text{Prob}_{\text{area}}[a,b] = \frac{area(a,b)}{area_{\text{total}}}$$

which is the probability that a randomly chosen point on an interface surface belongs to the tile of a type a residue on one side and to the tile of a type b residue on the other side of the interface surface.

Letting $\#N(a,b)$ be the number of neighbor pairs in which one residue is of type a and other of type b and $\#N_{\text{total}} = \sum_{a\leq b} \#N(a,b)$ the total number of neighbor pairs, we define the *relative frequency of the neighbor pair a, b* to be equal to

$$\text{Prob}[a,b] = \frac{\#N(a,b)}{N_{\text{total}}}$$

which is the probability that a randomly chosen neighbor pair consists of a type a and a type b residue. We note again that the pairs are unordered, so $\text{Prob}[a,b] = \text{Prob}[b,a]$. If two tiles share a common portion of their boundary, we can measure the length or perimeter of that portion and use it as a weight in our statistics. Letting $perim(a,b)$ be the total shared perimeter of tiles contributed by pairs of type a and b and $perim_{\text{total}} = \sum_{a\leq b} perim(a,b)$, we define the *perimeter-weighted relative frequency of the neighbor pair a, b* to be equal to

$$\text{Prob}_{\text{perim}}[a,b] = \frac{perim(a,b)}{perim_{\text{total}}}$$

which is the probability that a randomly chosen point in the interior of an interface surface that does not belong to the interior of a tile belongs to the shared boundary of tiles contributed by a type a and type b residue.

We define a triplet as an intrachain pair whose members both form interchain pairs with a third residue on the other chain. Propensities for triplets are identified by visual inspection of a representative sample (20−50% of observed cases) for each type.

Likelihood Correction. The probabilities for inter- and intrachain pairs depend on the probabilities of single residues. For example, in the absence of any bias, the relative frequency of the pair $a \neq b$ is $\text{Prob}[a,b] = 2\text{Prob}[a]\text{Prob}[b]$, where the factor of 2 accounts for the fact that the pair is unordered. On the other hand, $\text{Prob}[a,a] = \text{Prob}[a]^2$, without the factor of 2. It thus makes sense to consider the ratio of the left over the right side, which in the absence of any bias is one. Taking the logarithm, we obtain positive numbers for pairs that are more likely than warranted by the probabilities of its constituents and negative numbers for pairs that are less likely. The formulas are

$$\text{LogOdds}(a,b) = \log_2 \frac{\text{Prob}[a,b]}{2\text{Prob}[a]\text{Prob}[b]}, \quad \text{where } a \neq b$$

$$\text{LogOdds}(a,a) = \log_2 \frac{\text{Prob}[a,a]}{\text{Prob}[a]\text{Prob}[a]}$$

We use similar area- and perimeter-weighted formulas to study the bias for or against forming inter- and intrachain pairs. Previous studies[2,3] have used the $\text{LogOdds}(a,a)$ formula to calculate all log odds values, regardless of pairing partners, but

we treat each pair uniformly as unordered pairs, giving a consistent statistical result.

Results and Discussion

Larger Data Set of Transient Hetero-Complexes. A major drawback of using the characteristics of known complexes to study protein−protein interactions is the limited available data set. Although the PDB holds more than 34 000 structures, only a small fraction of these represents complexes of proteins that can exist independently in a folded native state. The data set is sufficiently limited that statistical filtering for redundancy is often not applied. For example, the largest previous analysis of hand-selected protein−protein recognition sites contains 75 transient complexes,[8] seven of which are complexes of a serine protease bound to a protein inhibitor. Thus, we have attempted to increase the number and diversity of readily available protein−protein complexes for study. Our current list, built on the foundation of Chakrabarti and Janin,[1] adds complexes from the protein docking benchmark data set,[4] and others culled manually from the literature. New complexes were selected to capture as much interface diversity as possible, taking advantage of recent additions to the PDB consisting of multi-protein systems (e.g., ubiquitination). Our data set includes 135 hetero-complexes, some with multiple interfaces (Table 1).

Applying our interface definition to the 135 complexes in our database (Table 1), we generate 146 protein−protein interfaces. There are $\#_{total} = 10\ 292$ residues in total, $P_{total} = 25\ 875$ interchain pairs, and $N_{total} = 23\ 545$ neighbor pairs in our data set. This yields an average of $2P_{total}/\#_{total} = 5.03$ pairs and $2N_{total}/\#_{total} = 4.58$ neighbors per residue. The most common residue is present #(Glu) = 705 times and the least common residue is found #(Met) = 193 times, demonstrating the significant size of our interface database. The interfaces range in size from 709.4 to 8,544.8 Å^2, with an average ± standard deviation of 2094.2 ± 1263.3 Å^2, counting both sides so as to compare with accessible surface area methods that have previously shown an average of 1906 ± 759 Å^2.[1] The distribution of area for our data set is depicted in Figure 2. As a point of reference, there are 64,640 residues in the 135 complexes, yielding 54,348 non-interface residues, such that the interface regions account for almost 16% of the residues in the data set.

Flattening. Visualization is another significant hurdle faced when attempting to understand protein−protein interfaces. A common representation shows the separate surfaces of the two contributing proteins, often as a GRASP view.[25] From these disjointed views, it is difficult to see the relative area contributions of residues or residue interactions across the interface. Gabdoulline and Wade[26] have previously described a method that attempts to remedy this dilemma by projecting their analytically defined interface onto a flat surface with approximate conservation of area. We introduce here a flattened view and associated web tool that takes advantage of our interface definition to yield a voidless map independent of embedding and without overlapping points. As described in Experimental Section, we flatten the potentially complicated surface embedded in three-dimensional space to a disk. Critically, the flattening procedure retains all neighbor connectivity, both across as well as on each side of the interface. The flattened interface can be colored by selectable attributes. For example, the default view in our MAPS web tool divides and colors the flattened interface by contributing residues (Figure 3A,B). Thick black lines separate the tiles (residue contributions) that are colored by type according to the indicated palette. Thin black lines separate the atom contributions within each tile, and both atom name and residue identification are obtained using a mouse-over tool. Other selectable attributes for viewing include atom type, electrostatics, backbone versus side-chain, and distance between atoms across the interface. For example, the large contribution of backbone atoms is immediately apparent in Figure 3F, whereas hydrogen bonds across the interface are seen as the regions of closest distance in Figure 3E. Of course, any of the selected attributes can originate from either of the two contributing proteins, as seen for residue type in Figure 3A,B.

To directly compare the attributes from two complexed proteins, we have created a merged view (Figure 3C,D) in which the bottom side remains unchanged while the top side is reduced to frames that outline its tiles. These frames have a black edge to aid in viewing and are colored to indicate the selected attribute (e.g., residue type, atom type, distance). Thus, it is easy to see which residue/attribute from one protein sits across from which residue/attribute from the other protein.

Our MAPS web tool (http://biogeometry.cs.duke.edu/research/docking/index.html) contains interfaces for all 135 complexes from our data set (Table 1) and displays both the 3D interface between the generating protein chains as well as the flattened view with functionality as described above.

Comparison of Related Protein Interfaces. A particularly powerful application of flattened interfaces is the direct comparison of similar protein complexes to reveal key similarities and differences. As an example, we consider the pig RNase inhibitor (Figure 4A) bound to bovine RNase (1DFJ) (Figure 4C) compared to the human RNase inhibitor (Figure 4B) bound to angiogenin (1A4Y) (Figure 4D), two well-characterized protein−protein complexes. The pig and human RNase inhibitors are highly homologous (about 77% identity), whereas bovine RNase and angiogenin, despite their similar protein folds, are significantly different in sequence (only about 36% identity) and function (ribonuclease activity vs inducer of angiogenesis). The similarity of the binding interfaces is readily seen from the side of the inhibitor (Figure 4A,B). Clearly, the RNase inhibitors use the corresponding binding sites, showing the same interfacial residues, including the hotspot residues found in both complexes (Tyr434 and Asp435). Some differences include the presence of Trp375 in 1A4Y but not 1DFJ and the more subtle change of the backbone-mediated interaction of residue 436, which is an Ile in 1A4Y and a Thr in 1DFJ. Despite a few similarities in the region of the hotspot interaction (His119, Lys41, Gln11 in 1DFJ vs His114, Lys40, Gln117 in 1A4Y), the protein interfaces are dramatically different from the other side. The direct visual comparison of two or more protein interfaces is expected to facilitate experimental investigations aimed toward understanding how protein−protein interaction can attain both specificity and flexibility.

Single Amino Acid Statistics. Understanding the composition of binding sites is essential to understanding how transient protein−protein complexes form. Despite differences in the definition of the interface and the database of protein complexes, our amino acid composition for protein interfaces is comparable to previous studies.[1] For example, four of the five most (Ser, Glu, Gly, and Asp) and the five least (Met, Trp, Cys, His, and Phe) represented amino acids are identical in the two data sets. Also, our retraction process toward the protected core of about 50% area (see Experimental Section) shows a trend in enrichment of amino acids similar to the selection of the

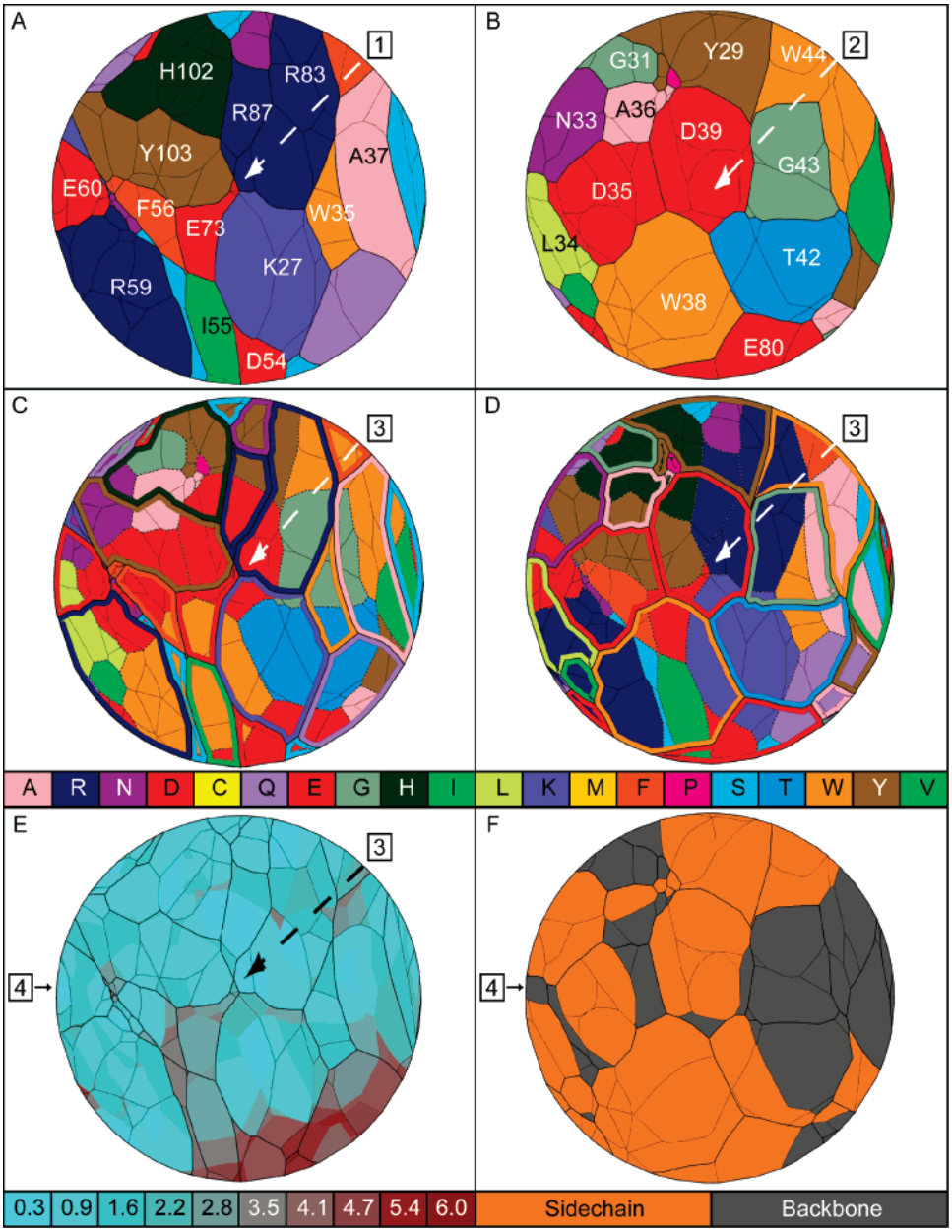

Figure 3. Flattened views of the interface of barnase/barstar (1BRS). (A) Interface surface of barnase (chain A), colored by residue type. The $N_{\eta 2}$ atom of Arg 83 is indicated by 1. (B) Interface surface of barstar (chain D), colored by residue type. The $O_{\delta 1}$ atom of Asp 39 is indicated by 2. (C and D) Merged view of the interface between barnase and barstar, colored by residue type, as viewed from barnase in C, and from barstar in D. The salt-bridge between $N_{\eta 2}$ Arg 83 (barnase) and $O_{\delta 1}$ Asp 39 (barstar) is indicated by 3. (E) Interface surface of barnase colored by distance gradient. The Arg 83−Asp 39 salt-bridge is indicated by 3, and a hydrogen bond between a backbone N from Leu 34 of barstar to $O_{\epsilon 2}$ of Glu 60 of barnase is indicated by 4. (F) Interface surface of barstar colored by backbone/side chain. The Leu 34−Glu 60 hydrogen bond is indicated by 4.

53% core residues that contribute 72% of the interface area in Chakrabarti and Janin.[1] In particular, both methods and data sets reveal a substantial enrichment in aromatic and some hydrophobic residues, especially Leu and Ile, and a decrease in all charged residues, especially Asp and Glu (Table 2).

Because evolutionary selection of amino acids occurs via side-chain variation on an invariant backbone, we chose to analyze backbone contributions to protein interfaces separately. This was accomplished by designating backbone its own category, which consists of all backbone atoms of the residues making up a protein chain. The side chains are the residues without their backbone atoms. As a result, Gly is absorbed into the backbone because it does not have a sidechain. In the case where both side chain and backbone atoms from one residue appear at the interface, both are counted in the appropriate categories. As previously noted by Lo Conte et al.,[8] backbone atoms comprise a significant portion of the interface. By frequency, about 32% of all interfacial pairs are either backbone−backbone or backbone−side-chain contacts, with backbone atoms accounting for about 23% of total interface area (e.g., Figure 3F). Though prior studies have shown that backbone carbonyl O atoms are commonly involved in hydrogen bonding at protein−protein interfaces,[8,20] we find that all backbone atoms make a significant contribution. While not significantly

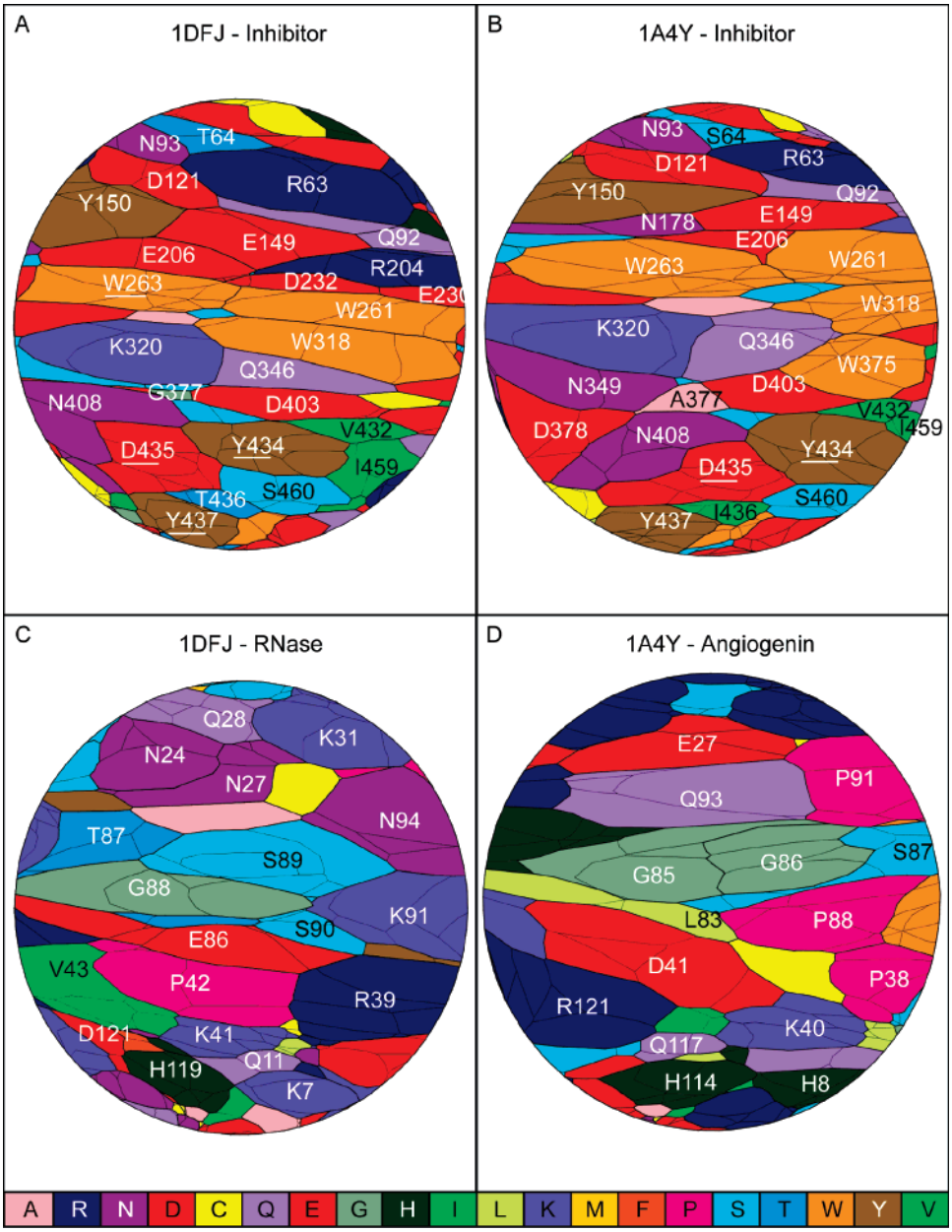

Figure 4. Comparison between the flattened views of 1DFJ and 1A4Y: (A) 1DFJ inhibitor chain, (B) 1A4Y inhibitor chain, (C) 1DFJ RNase chain, (D) 1A4Y angiogenin chain. Residues of interest are labeled in all four pictures, and the hotspot residues are underlined for the two inhibitors.

changing the rank order for amino acid prevalences at the interface, our classification serves as a noise filter, allowing the five most frequent side chains (Glu, Ser, Asp, Lys, and Arg) to each have more than 7% representation and the four least represented side chains (Met, Cys, Trp, and His) to each have less than 3.5% representation. In the halfway retracted interface, there are some changes in the ranked frequencies of the larger side chains. In particular, Tyr and Leu join Ser as the three most prevalent side chains. Additionally, our reclassification decreases the number from $\#_{total} = 10\,292$ residues at the interface to $\#_{total} = 8934$ side chains (including the backbone category), with the number of pairs dropping from $P_{total} = 25\,875$ to $P_{total} = 16\,603$ pairs. Accordingly, the average number of pairs drops from $2P_{total}/\#_{total} = 5.03$ per residue to $2P_{total}/\#_{total} = 3.72$ per side chain. This change is fairly uniform across the complexes (Figure 2). The large drop indicates that many residues contribute only backbone atoms to the interface, often forming cross-interface interactions with other backbone-only contributors. The change in average pairs indicates that residues tend to interact with backbone atoms from multiple other residues, while not necessarily interacting with the corresponding side chains. Our observation suggests that some interfacial residues are selected for their contribution to folding and/or stability of the substituent proteins rather than for their contribution to transient complex formation. The high prevalence of backbone at the interface also implies that a complete energetic characterization of interfacial contacts by experimentation remain elusive given the challenges in substitutions of backbone atoms.

To emphasize that most of the binding energy of protein–protein interactions is thought to be contributed by van der Waals interactions, we weigh amino acid occurrences by

Table 2. Amino Acid Occurrences by Interfacial Surface Area

	standard 20 AA[a]		side chain/backbone[b]			
	frequency		frequency		area	
	full	core	full	core	full	core
ALA	4.09	4.29	3.68	4.04	1.84	1.79
ASP	7.01	5.96	7.53	6.26	4.76	3.26
ARG	6.66	6.06	7.30	6.47	8.25	6.48
ASN	5.66	5.14	6.00	5.20	4.26	3.27
CYS	2.74	3.54	2.46	3.03	0.99	1.05
GLU	7.18	5.24	7.89	5.20	5.62	3.23
GLN	4.05	4.06	4.32	4.15	4.13	3.08
GLY	7.12	6.82	0.00	0.00	0.00	0.00
HIS	3.11	3.50	3.37	3.64	3.09	3.35
ILE	3.75	4.43	3.83	4.72	3.51	4.04
LYS	6.89	5.03	7.38	4.68	6.18	5.15
LEU	5.59	6.58	5.75	6.76	4.98	5.23
MET	2.09	2.63	2.16	2.63	2.34	2.51
PHE	3.50	4.47	3.68	4.68	3.87	5.00
PRO	4.20	3.84	4.05	3.51	2.86	2.49
SER	7.78	7.53	7.76	6.87	3.76	3.15
THR	6.08	5.16	6.12	4.78	3.59	2.96
TRP	2.60	3.58	2.75	3.88	3.64	4.97
TYR	5.75	7.34	6.09	7.70	6.36	7.80
VAL	4.16	4.81	4.16	4.74	2.98	3.21
BBA	na	na	3.70	7.05	22.98	27.97

[a] Standard 20 amino acid definitions are used to calculate the frequency of each side chain type or backbone for the full and halfway retracted interface surface. [b] Frequency and area contribution of each side chain or backbone (see Results) for the full interface and the core.

Table 3. Most Common Interchain Pairs, Both in Area-Weighted Ranking and with Likelihood Correction for the Full as Well as the Halfway Retracted Interface Surface

	full				core			
	%			Log Odds	%			Log Odds
1	Glu−Arg	1.72	Ala−Cys	1.34	Glu−Arg	1.22	Cys−His	2.34
2	Asp−Arg	1.37	Asp−Lys	1.21	Asp−Lys	1.20	Asp−Lys	1.84
3	Asp−Lys	1.36	Cys−Leu	1.16	Arg−Trp	1.15	Arg−Glu	1.54
4	Glu−Lys	1.29	Cys−His	1.11	Arg−Tyr	1.06	Ala−Cys	1.44
5	Arg−Tyr	1.12	Ala−Ala	1.01	Leu−Phe	0.96	Phe−Phe	1.35
6	Asn−Tyr	0.72	Arg−Glu	0.89	Lys−Tyr	0.92	Met−Met	1.31
7	Lys−Tyr	0.71	Glu−Lys	0.89	Asn−Tyr	0.84	Met−Pro	1.08
8	Glu−Tyr	0.70	Phe−Phe	0.89	Ile−Trp	0.81	Met−Val	1.07
9	Arg−Trp	0.70	Ile−Leu	0.89	Asp−Arg	0.75	Ile−Trp	1.02
10	Arg−Asn	0.70	Leu−Met	0.88	Gln−Tyr	0.75	Leu−Val	1.01

interfacial surface area throughout the rest of our analysis (Table 2). Not surprisingly, the biggest gainers from this procedure are the large amino acids Arg, Tyr, Trp, and Phe, whereas the biggest losers are Ser, Asp, Thr, and Ala. Despite some reordering, however, many of the same amino acids remain similarly ranked in this area-weighted analysis.

Interchain Pair Statistics. We next mine our data set for residue pairings across the interface surface. Pairwise statistics prove to be more informative than single residue or side-chain statistics because they more directly extract information about what interactions drive protein−protein association and/or prevent their disassociation. Interchain pairs are weighted by area to emphasize significant residue contacts across the interface. In addition, we use the log odds function to determine how different the probability of a pair is from uniformly random given the probabilities of its constituents (Table 3, Supplemental Table 2a−d in Supporting information). As for the single side chains, we are primarily interested in interactions that can be readily selected by evolution or tested by site-directed mutagenesis and thus ignore pairings that involve the backbone. Area-weighted pairwise statistics yield a number of interesting results:

Not unexpectedly, the four highest scoring area-weighted pairs for the full interface are the four salt-bridge pairs (Glu−Arg, Asp−Arg, Asp−Lys, and Glu−Lys) (Table 3). Of these, the Glu−Arg pair is by far the most common, with 1.72% compared to 1.37%, 1.36%, and 1.29% for the Asp−Arg, Asp−Lys, and Glu−Lys pairs, respectively. Although some of these pairs arise from van der Waals contacts along the uncharged part of the side chain, many of them form salt-bridges. The prevalence of salt-bridges at the interface of transient protein−protein complexes has been previously noted by Ofran and Rost[3] and emphasizes the importance of charge complementarity at protein interfaces, which tend to be protected from solvent.

The next most prevalent pairs are Tyr with the sidec hains of Arg, Asn, Lys, and Glu (Table 3). Of these, the Arg−Tyr pair shows the most interesting configurations, often as a hydrogen bond between the hydroxyl of Tyr and one of the three nitrogens (usually $N_{\eta 1}$ or $N_{\eta 2}$) of Arg (about 40%). Almost as often, a classical cation−π interaction is observed (about 40%).[27] Of these, about two-thirds orient the amino group over the center of the ring, while about one-third orient the C_δ or $C\gamma$ atoms over the ring. Asn−Tyr pairs are seen most often as a hydrogen bond between the hydroxyl of Tyr and the $O_{\delta 1}$ or $N_{\delta 2}$ atom of the Asn residue (about 55%). Occasionally, Asn−Tyr pairs display an orientation similar to cation−π packing despite not being positively charged (about 15%). Lys−Tyr pairs are seen most often as hydrogen bonds between the hydroxyl of the Tyr and the N_ζ atom of the Lys (about 55%). About one-third of these pack Lys carbon atoms against the Tyr ring, often with a hydrogen bond between the N_ζ of Lys and the carbonyl of the Tyr backbone. Glu−Tyr pairs are seen most often as hydrogen bonds between the hydroxyl of the Tyr and either $O_{\epsilon 1}$ or $O_{\epsilon 2}$ of the Glu residue (about 65%).

Arg is part of the ninth and tenth ranked pairs. For the Arg−Trp pair, the cation−π interaction is most prevalent (about 75%). The core region is significantly enriched in these Arg−Trp pairs, which is consistent with their increased prevalence as hotspot residues. For the tenth ranked Arg−Asn pair, hydrogen bonds between the functional groups are by far the most common mode of interaction (about 90%).

Similar results are observed for the frequency statistics not weighted by area (data not shown).

Likelihood correction based on probabilities from individual occurrences serves to highlight two types of pairs. First, pairing preferences for residues that are rarely present at the interface (i.e., Cys, Met, His) are revealed and can be identified as recurring motifs. For example, the Cys−His pair arises from the proximity of a cysteine disulfide bridge that packs against the active site His of the catalytic triad in serine protease inhibitor complexes. This motif is particularly interesting because of the diversity observed in serine protease inhibitors. Although there is high homology among the serine-proteases, the inhibitors themselves are quite different with the exception of the cysteine disulfide positioned about 3.5 Å from the His. Although this disulfide is known to contribute significantly to the stability of these protease inhibitors,[28] our observations suggest that there may be other roles for these highly conserved Cys pairs, such as this interaction with the His of the serine protease. It is interesting to note that substitutions of this Cys−Cys pair have been performed that yield a protein of similar stability to wild-type yet retain a similar potency as trypsin

inhibitors only if the substitutions are approximately size-neutral (e.g., Gly−Leu but not Leu−Val).[29]

The second class of pairs emphasized by likelihood correction includes those that remain high on the list despite their intrinsic prevalence at the interface. These unusually prevalent pairs fall into two categories. Leu−Val and Ile−Leu suggest a hydrophobic component to transient protein complexes, as noted previously,[1] whereas the charge pairs Asp−Lys, Glu−Arg, Glu−Lys, and Asp−Arg again emphasize the importance of charge complementarity at protein interfaces. Similar results with only a slight reordering were observed for the core residues following retraction (Table 3).

The 10 least common pairs (without likelihood correction) almost all involve Cys (data not shown), in accord with its infrequency at protein interfaces. Following likelihood correction, the self-pairs His−His, Lys−Lys, Arg−Arg, Tyr−Tyr, and Asp−Asp are especially rare, with log odd ratios ranging from −1.03 to −0.38, suggesting that the high steric cost of packing like charges against each other is selected against by evolution.

These interchain pair statistics differ significantly from those reported by Glaser et al.[2] In their study, Cys−Cys, Pro−Trp, Asp−His and Arg−Trp are the most prevalent unweighted pairs, and Arg−Trp, Pro−Trp, and Cys−Cys are the most prevalent pairs as weighted by volume of the contributing amino acid. The basis for the discrepancy between our results herein and these prior studies is twofold. First, their data set of 621 interfaces is dominated by 404 homo-dimers, whose interfaces more resemble protein interiors.[3] Second, Glaser et al. use overall residue volume, not area contribution to the interface, to weight their amino acids, thus biasing their results toward larger residues even if they contribute only a single atom to the protein interface.

Although not our primary focus, we also examined the extremely prevalent backbone-atom−backbone-atom (BBA−BBA) pairs and BBA pairing to selected amino acids in more detail. As about 70% of BBA−BBA contacts are at distance between 3 and 6 Å, these appear to constitute van der Waals packing or hydrogen bonding. In a number of complexes, we observe series of hydrogen bonds between backbone O and N atoms that mimic the hydrogen-bonding pattern of β-sheets. These interactions, both parallel and antiparallel, occur in 2−3 residue stretches per side, such as the parallel β-sheet formation between residues Gly 42, Val 43, and Met 44 of actin with residues Tyr 65, Val 66, and Val 67 of DNase I in the complex 1ATN as previously noted.[30] Additionally, these β-sheet-like motifs are seen in a number of proteinase/inhibitor complexes (1BTH, 1CBW, 1FLE, 1HIA, 2KAI, 2PTC, and 3TPI).[31,32]

Intrachain Pairs or Neighbor Statistics. Unique to our definition of the interface, neighbor information on each side of a given protein complex is also captured. Neighbor pairing preferences reveal the composition of common patches on a protein that may be responsible for initial docking or subsequent stabilization of a transient interaction. Analogous to weighting by area applied to the interchain pairs, we here accumulate statistics in which each intrachain pair is weighted by the length of the shared boundary between the contributed regions (Supplemental Table 3a,b in Supporting Information). We again exclude the prevalent BBA-pairings from our tabulated analysis (Table 4).

As described by Jones and Thornton,[7] surface patches that correlate with protein docking sites in hetero-complexes show a propensity for hydrophobic residues, particularly Ile, Leu, Met, Phe, and Val, as well as Arg and the polar aromatic residues Trp, Tyr, and His. Our observation of neighbor pair preferences agree with these findings and complement them by identifying specific neighbor contacts as well as their interaction partners across the interface (selected triplets).

Table 4. Most Common Intrachain or Neighbor Pairs, Both in Perimeter-Weighted Ranking and with Likelihood Correction for the Full as Well as the Halfway Retracted Interface Surface

	full				core			
		%		Log Odds		%		Log Odds
1	Asp−Arg	1.00	Cys−Cys	3.85	Ile−Leu	0.78	Cys−Cys	3.68
2	Glu−Lys	0.92	Met−Met	1.80	Trp−Tyr	0.77	Met−Met	2.05
3	Glu−Arg	0.87	Trp−Trp	1.31	Asp−Tyr	0.70	Trp−Trp	1.37
4	Asp−Lys	0.60	Gln−His	1.09	Asp−Arg	0.70	Gln−His	1.33
5	Asp−Tyr	0.59	Ile−Leu	1.07	Leu−Tyr	0.70	Pro−Pro	1.33
6	Ile−Leu	0.57	Glu−Lys	1.06	Tyr−Tyr	0.69	Thr−Thr	1.22
7	Arg−Tyr	0.55	Asp−Arg	1.05	Ser−Tyr	0.68	His−Ser	1.21
8	Arg−Lys	0.54	Ser−Ser	0.93	Leu−Phe	0.66	Leu−Leu	1.20
9	Asn−Tyr	0.52	Met−Pro	0.92	Leu−Val	0.62	Met−Pro	1.19
10	Leu−Tyr	0.52	Ala−Ile	0.90	Ile−Tyr	0.58	Met−Val	1.17

As with the interchain pairing preferences, the four oppositely charged pairs (Asp−Arg, Glu−Lys, Glu−Arg, and Asp−Lys) are the most prevalent neighbor pairs (Table 4). In contrast to their interactions across the interface, however, these pairs do not typically form salt-bridges. Instead, they appear to form small dipoles, mostly on the periphery of interfaces. These dipoles do not necessarily form salt-bridges across the interface as they are paired with other charged residues (about 31%), polar residues (about 30%), backbone atoms (about 24%), as well as with hydrophobic residues (about 15%). How these dipoles facilitate transient protein interactions remains to be studied in more detail using electrostatic potentials. However, these peripheral dipoles are reminiscent of the concept of electrostatic steering pioneered by Fersht and Schreiber[33] with an added element of local directionality. Asp−Arg dipoles are enriched at protein interfaces versus noninterface in a ratio of about 2:1, whereas Glu−Lys, Glu−Arg, and Asp−Arg dipoles are about equally common at and outside interfaces.

The fifth most common neighbor pair is Asp−Tyr, which has a notable preference for tripling with Arg and Lys (see for example the Tyr 434−Asp 435 pair with Lys 40 in 1A4Y in Figure 4). This generates a salt-bridge across the interface flanked by a Tyr residue. This configuration is consistent with the high prevalence of Arg−Tyr and Lys−Tyr pairs across the interface surface (Table 3).

The sixth most prevalent neighbor pair is the hydrophobic Ile−Leu, which is enriched to be the most common neighbor pair following halfway retraction. Most of the time, these Ile−Leu pairs form triplets with other hydrophobic residues (about 50%), but we also see triplets with polar residues (about 26%), backbone atoms (about 18%), and the occasional charged amino acid (about 6%). The reason this hydrophobic pair is more common than any of the others remains unclear, though it is important to note that Ile and Leu are often found packed near each other, primarily in hydrophobic protein interiors. In fact, many hydrophobic pairs are seen in the core of the interface (Table 4), suggesting that a pre-existing hydrophobic patch can serve as a docking site for protein interactions. As for Ile−Leu, these hydrophobic pairs are not necessarily across from other hydrophobic residues. For example, Leu−Tyr forms triplets with other hydrophobic residues (about 37%) and with polar residues (about 30%), and Trp−Tyr pairs preferably form triplets with polar (about 30%), backbone (about 30%), and

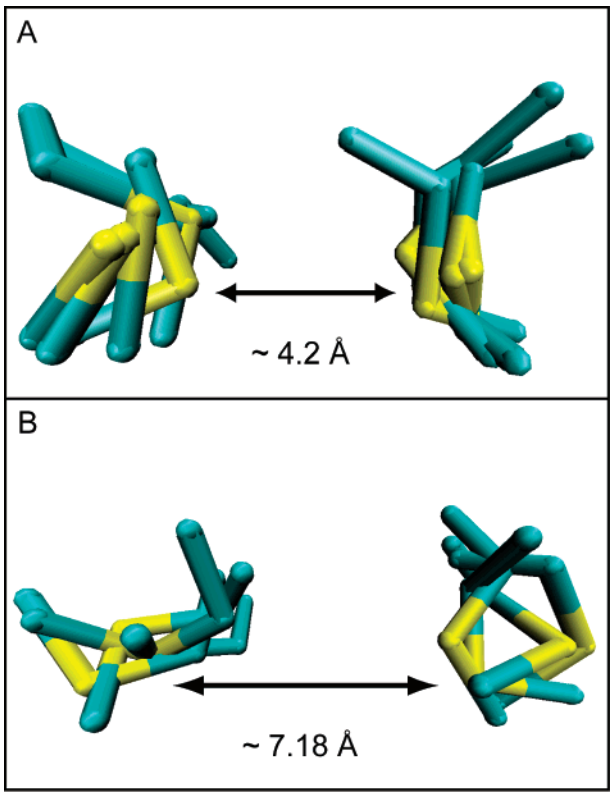

Figure 5. (A) Near Met−Met neighbor pairs from 1LDK, 1MDA, 1DN2, 1YCS, and 1AIP overlaid using rmsd alignment for side-chain atoms, excluding C_β. (B) Far Met−Met neighbor pairs from 1ATN, 1MDA, and 1AIP(2) overlaid using rmsd alignment for side-chain atoms, excluding C_β.

charged (about 24%) residues and relatively infrequently with hydrophobic residues (about 16%).

Arg−Tyr and Lys−Tyr neighbor pairs are also often seen at the interface (7th and 12th most frequently, respectively), which is consistent with their propensity to form the cation−π motif. Collectively, 7.6% of the observed Arg−Tyr and Lys−Tyr neighbor pairs occur between consecutive residues in a protein chain, which is consistent with the previously noted 7.3% for all cation−π motifs, suggesting that many occur on α-helices.[27]

Likelihood correction highlights a number of interesting neighbor pairs (Table 4, Supplemental Table 3c,d in Supporting Information). The Cys−Cys disulfide of trypsin protease inhibitors noted above is detected again, which not surprisingly forms triplets with His, as described above. Met−Met neighbor pairs are the second most prevalent following likelihood correction. We observe two similar yet different motifs, which we refer to as *near* and *far* (Figure 5). The near Met−Met motif contains S−S distances ranging from 3.5 to 4.7 Å, with an average ± standard deviation of 4.2 ± 0.4 Å, and contains only nonconsecutive pairs (i.e., not consecutive along the protein chain). The far Met−Met motif contains S−S distances ranging from 5.54 to 8.81 Å, with an average ± standard deviation of 7.18 ± 1.2 Å, and contains both consecutive Met−Met residue pairs and nonconsecutive pairs. Both Met−Met motifs pack against hydrophobic residues or hydrophobic regions of charged residues, moving close to a C_β and C_γ atom in all but one observed case.

The Gln−His pair, which is the fourth most prevalent neighbor pair in the perimeter normalized statistics, is seen most frequently forming hydrogen bonds, either as Gln $O_{\epsilon1}$/$N_{\epsilon2}$ with His $N_{\delta1}$/$N_{\epsilon2}$ (about 65%) or Gln $O_{\epsilon1}$/$N_{\epsilon2}$ with His backbone O/N (about 20%). Spurious contacts (about 15%) include Gln−His neighbors that occur consecutively on the protein chain, where steric requirements preclude H-bond formation.

Conclusions

We have presented here a statistical analysis of protein−protein interfaces from a large and diverse data set using a reliable and consistent definition of the interface. Our analysis serves as a foundation for prediction problems in protein−protein docking. For example, our pairing and neighbor preferences can be used as weights in scoring functions to distinguish between true and false predictions. Previously generated lists of 2000−10 000 possible docking configurations containing one or more correct answers[34] and data sets of native and decoy docking configurations[4,35] will serve as useful test sets for such implementations. Additionally, residue frequencies and neighbor preferences can be used to predict probable binding sites for proteins whose 3D coordinates are available but whose interaction sites remain unclassified. Identification of these binding sites will allow the potential identification of novel protein−protein pairs, leading to a greater understanding of the networks of interactions in the proteome. We have also provided a novel visualization that facilitates the analysis of the intrinsic complexity of protein−protein interfaces. Our simplified view allows easier recognition of interfacial residue contacts and other biochemical characteristics. Insights derived from such visual inspections will aid in the design of experiments toward elucidating the specificity of protein−protein association. Also, comparative studies of related interfaces are made easier by having a single independent and simplified entity. Combined, our statistical analysis and visualization serve as a novel toolset for biochemists interested in the fundamentals of protein−protein interactions.

Supporting Information Available: Tables listing the area distortion measures for a sampling of complexes in the data set; the complete area-weighted statistics and complete likelihood corrected statistics for interchain pairs for the full interface, complete area-weighted statistics and complete likelihood corrected statistics for interchain pairs for the core of the interface; complete perimeter-weighted statistics and complete likelihood corrected perimeter-weighted statistics for the intrachain neighbors for the full interface and complete perimeter-weighted statistics and complete likelihood corrected perimeter-weighted statistics for the intrachain neighbors for the core of the interface; and figure of the comparison between Uniform and Mean Value Coordinate flattening methods for 1BRS. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Chakrabarti, P.; Janin, J. Dissecting protein-protein recognition sites. *Proteins* **2002**, *47*, 334−343.
(2) Glaser, F.; Steinberg, D. M.; Vakser, I. A.; Ben-Tal, N. Residue frequencies and pairing preferences at protein-protein interfaces. *Proteins* **2001**, *43*, 89−102.
(3) Ofran, Y.; Rost, B. Analyzing six types of protein-protein interfaces. *J. Mol. Biol.* **2003**, *325*, 377−387.
(4) Chen, R.; Mintseris, J.; Janin, J.; Weng, Z. A protein-protein docking benchmark. *Proteins* **2003**, *52*, 88−91.
(5) Janin, J.; Chothia, C. The structure of protein-protein recognition sites. *J. Biol. Chem.* **1990**, *265*, 16027−16030.

(6) Hubbard, S. J.; Argos, P. Cavities and packing at protein interfaces. *Protein Sci.* **1994**, *3*, 2194−2206.
(7) Jones, S.; Thornton, J. M. Analysis of protein-protein interaction sites using surface patches. *J. Mol. Biol.* **1997**, *272*, 121−132.
(8) Lo Conte, L.; Chothia, C.; Janin, J. The atomic structure of protein-protein recognition sites. *J. Mol. Biol.* **1999**, *285*, 2177−2198.
(9) Bahadur, R. P.; Chakrabarti, P.; Rodier, F.; Janin, J. A dissection of specific and non-specific protein-protein interfaces. *J. Mol. Biol.* **2004**, *336*, 943−955.
(10) Nooren, I. M. A.; Thornton, J. M. Structural characterization and functional significance of transient protein-protein interactions. *J. Mol. Biol.* **2003**, *325*, 981−1018.
(11) Jones, S.; Thornton, J. M. Principles of protein-protein interactions. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 13−20.
(12) Tsai, C.-J.; Lin, S. L.; Wolfson, H. J.; Nussinov, R. Studies of protein-protein interfaces: a statistical analysis of the hydrophobic effect. *Protein Sci.* **1997**, *6*, 53−64.
(13) Xu, D.; Tsai, C. J.; Nussinov, R. Hydrogen bonds and salt bridges across protein-protein interfaces. *Protein Eng.* **1997**, *10*, 999−1012.
(14) Rodier, F.; Bahadur, R. P.; Chakrabarti, P.; Janin, J. Hydration of protein-protein interfaces. *Proteins* **2005**, *60*, 36−45.
(15) Ma, B.; Elkayam, T. W. H.; Nussinov, R. Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 5772−5777.
(16) Duncan, B. S.; Olson, A. J. Texture mapping parametric molecular surfaces. *Mol. Graphics* **1995**, *13*, 258−264.
(17) Varshney, A. F. P.; Brooks, J.; Richardson, D. C.; Wright, W. V.; Manocha, D. Defining, computing, and visualizing molecular interfaces. *Proc. IEEE Visualization* **1995**, 36−43.
(18) Word, J. M.; Lovell, S. C.; LaBean, T. H.; Taylor, H. C.; Zalias, M. E.; Presley, B. K.; Richardson, J. S.; Richardson, D. C. Visualizing and quantifying molecular goodness-of-fit: small-probe contact dots with explicit hydrogen atoms. *J. Mol. Biol.* **1999**, *285*, 1711−1733.
(19) Neuvirth, H.; Raz, R.; Schreiber, G. ProMate: A structure based prediction program to identify the location of protein-protein binding sites. *J. Mol. Biol.* **2004**, *338*, 181−199.
(20) Mintz, S.; Shulman-Peleg, A.; Wolfson, H. J.; Nussinov, R. Generation and analysis of a protein-protein Interface data set with similar chemical and spatial patterns of interactions. *Proteins* **2005**, *61*, 6−20.
(21) Ban, Y. E.; Edelsbrunner, H.; Rudolph, J. Interface surfaces for protein-protein complexes. *J. Assoc. Comput. Mach.* **2006**, *53* (3), 361−378.
(22) Gottschalk, K.-E.; Neuvirth, H.; Schreiber, G. A novel method for scoring of docked protein complexes using predicted protein-protein binding sites. *Protein Eng. Des. Sel.* **2004**, *17* (2), 183−189.
(23) Tutte, W. How to draw a graph. *Proc. London Math. Soc.* **1963**, 743−768.
(24) Floater, M. Mean value coordinates. *Comput. Aided Geometric Des.* **2003**, *20*, 19−27.
(25) Nicholls, A. J. GRASP: graphical representation and analysis of surface properties. *Biophys. J.* **1993**, *64*, A116.
(26) Gabdoulline, R. R.; Wade, R. C. Analytically defined surfaces to analyze molecular interaction properties. *J. Mol. Graphics* **1996**, *14*, 341−353.
(27) Gallivan, J. P.; Dougherty, D. A. Cation-pi interactions in structural biology. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 9459−9464.
(28) Schwarz, H.; Hinz, H.; Mehlich, A.; Tschesche, H.; Wenzel, H. Stability studies of derivatives of the bovine pancreatic trypsin inhibitor. *Biochemistry* **1987**, *26* (12), 3544−3551.
(29) Hagihara, Y.; Shiraki, K.; Nakamura, T.; Uegaki, K.; Takagi, M.; Imanaka, T.; Yumoto, N. Screening for stable mutants with amino acid pairs substituted for the disulfide between residues 14 and 38 of bovine pancreatic trypsin inhibitor (BPTI). *J. Biol. Chem.* **2002**, *277* (52), 51043−51048.
(30) Kabsch, W.; Mannherz, H. G.; Suck, D.; Pai, E. F.; Holmes, K. C. Atomic structure of the actin:DNase I complex. *Nature* **1990**, *347*, 37−44.
(31) Bode, W.; Huber, R. Natural protein proteinase inhibitors and their interaction with porteinases. *Eur. J. Biochem.* **1992**, *204* (2), 433−451.
(32) van de Locht, A.; Bode, W.; Huber, R.; Le Bonniec, B. F.; Stone, S. R.; Esmon, C. T.; Stubbs, M. T. The thrombin E192Q-BPTI complex reveals gross structural rearrangements: implications for the interaction with antithrombin and thombomodulin. *EMBO J.* **1997**, *16* (11), 2977−2984.
(33) Schreiber, G.; Fersht, A. R. Rapid, electrostatically assisted association of proteins. *Nat. Struct. Biol.* **1996**, *3* (5), 427−431.
(34) Wang, Y.; Agarwal, P. K.; Brown, P.; Edelsbrunner, H.; Rudolph, J. Coarse and reliable geometric alignment for protein docking. *Proc. Pac. Symp. Biocomput.* **2005**, 64−75.
(35) Gray, J. J.; Moughon, S.; Wang, C.; Schueler-Furman, O.; Kuhlman, B.; Rohl, C. A.; Baker, D. Protein-protein docking with simultaneous optimization of Rigid-body Displacement and Side-chain conformations. *J. Mol. Biol.* **2003**, *331*, 281−299.

PR070018+