Extent and Nature of Contacts Between Protein Molecules in Crystal Lattices and Between Subunits of Protein Oligomers

Swagata Dasgupta,[1] Ganesh H. Iyer,[1] Stephen H. Bryant,[2] Charles E. Lawrence,[3] and Jeffrey A. Bell[1]*
[1]*Department of Chemistry and Center for Biophysics, Rensselaer Polytechnic Institute, Troy, New York 12180*
[2]*National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894*
[3]*Laboratory of Biometrics, Wadsworth Center for Laboratories and Research, New York State Department of Health, Albany, New York 12237*

ABSTRACT **A survey was compiled of several characteristics of the intersubunit contacts in 58 oligomeric proteins, and of the intermolecular contacts in the lattice for 223 protein crystal structures. The total number of atoms in contact and the secondary structure elements involved are similar in the two types of interfaces. Crystal contact patches are frequently smaller than patches involved in oligomer interfaces. Crystal contacts result from more numerous interactions by polar residues, compared with a tendency toward nonpolar amino acids at oligomer interfaces. Arginine is the only amino acid prominent in both types of interfaces. Potentials of mean force for residue–residue contacts at both crystal and oligomer interfaces were derived from comparison of the number of observed residue–residue interactions with the number expected by mass action. They show that hydrophobic interactions at oligomer interfaces favor aromatic amino acids and methionine over aliphatic amino acids; and that crystal contacts form in such a way as to avoid inclusion of hydrophobic interactions. They also suggest that complex salt bridges with certain amino acid compositions might be important in oligomer formation. For a protein that is recalcitrant to crystallization, substitution of lysine residues with arginine or glutamine is a recommended strategy. Proteins 28:494–514, 1997.**
© 1997 Wiley-Liss, Inc.

Key words: potential of mean force; molecular recognition; protein interfaces; salt bridges; hydrophobic interaction; protein crystallization; contact patches

INTRODUCTION

The Protein Data Bank[1,2] contains extensive information regarding protein–protein interactions at crystal contacts, which has not been systematically exploited. Several authors have compared the lattice contacts found in multiple crystal forms of individual proteins and have made interesting observations regarding the way in which these specific proteins interact in the crystal.[3–7] Takahashi and colleagues[8] have shown for two protein crystals that electrostatic interactions are complementary, and may be essential. Relatively few efforts have been made to draw conclusions regarding crystal lattice contacts from a large sample of crystal structures.[9–11] Recently, a thorough examination of buried surface area at crystal contacts has been published with interesting conclusions regarding the specificity and symmetry of crystal contacts.[12]

No doubt reflecting the biological importance of subunit interactions in oligomeric proteins, the common characteristics of this type of interface have been surveyed by several authors.[13–19] The present survey of protein–protein contacts is focused more on crystal lattice contacts than on subunit interactions in protein oligomers. Nevertheless, intersubunit contacts are included in this study because they provide an interesting and informative comparison.

Oligomer interactions form with high affinity[17,20] and strict specificity. The exchange of subunits between two unrelated oligomeric proteins has apparently not been observed. Protein crystal contacts are tenuous and generally form only under conditions designed to limit protein solubility and maximize protein–protein interactions. However, they do form with a kind of specificity, at least in the sense that only certain contacts are repeated within each unit cell of any particular crystal packing arrangement.

For oligomeric proteins, amino acids at the subunit interfaces have been selected in order to main-

Contract grant sponsor: National Institutes of Health, Contract grant number: GM50817.

Dr. Dasgupta's present address is Department of Chemistry, Indian Institute of Technology, Kharagpur, 721 302 India.
*Correspondence to: Jeffrey A. Bell, Department of Chemistry, Rensselaer Polytechnic Institute, Troy, NY 12180.
E-mail: bell@xray.chem.rpi.edu

Received 25 June 1996; Accepted 27 February 1997

© 1997 WILEY-LISS, INC.

tain affinity and specificity properties. By contrast, protein crystal contacts form by happenstance and have not been optimized or selected for by evolution. At least for some proteins that crystallize in multiple forms, any residue on the protein surface may be involved in lattice contacts in one crystal form or another.[5]

Several characteristics of oligomer and crystal interfaces are compared in this work, with the goal of understanding in more detail how oligomer contacts are optimized by evolution and how crystal contacts might be improved.

MATERIALS AND METHODS

Selection of Protein Coordinates

The January 1993 release of the Protein Data Bank[1,2] contains 944 protein entries. These entries, however, evince considerable redundancy, since some proteins are represented by multiple entries differing only in the identity of bound small molecules (e.g., enzyme inhibitors) or in one or a few changes in amino acid sequence (e.g., site-directed mutants). Isomorphous structures must be eliminated to prevent similar contacts from being overrepresented in the data. Algorithms[21–23] have been developed that allow extraction of representative datasets from the Protein Data Bank. The selection of proteins for this study was not based on any one of the previous algorithms but was devised for this particular analysis. Proteins were excluded from this study according to the following criteria:

1. Resolution worse than 3.5 Å and R factor greater than 32%. Rather liberal criteria for structure quality were permitted in order to obtain the largest dataset possible. Errors in structure determination, especially at lattice contacts, appear in this analysis as increased noise, which may be tolerated in exchange for having the broadest coverage of interfaces possible. The numbers of structures with low resolution or high R factor are few (Fig. 1).
2. More than 10% of the residues in the protein missing. Many protein structures in the Protein Data Bank lack coordinates for several residues at the N terminus, C terminus, or, occasionally, in the middle of the protein. This criterion allows exclusion of proteins in which large parts of the lattice contact might be missing, without excluding the many structures that lack a few disordered residues.
3. Multiple structures with sequences at least 80% identical, all in the same space group, and with

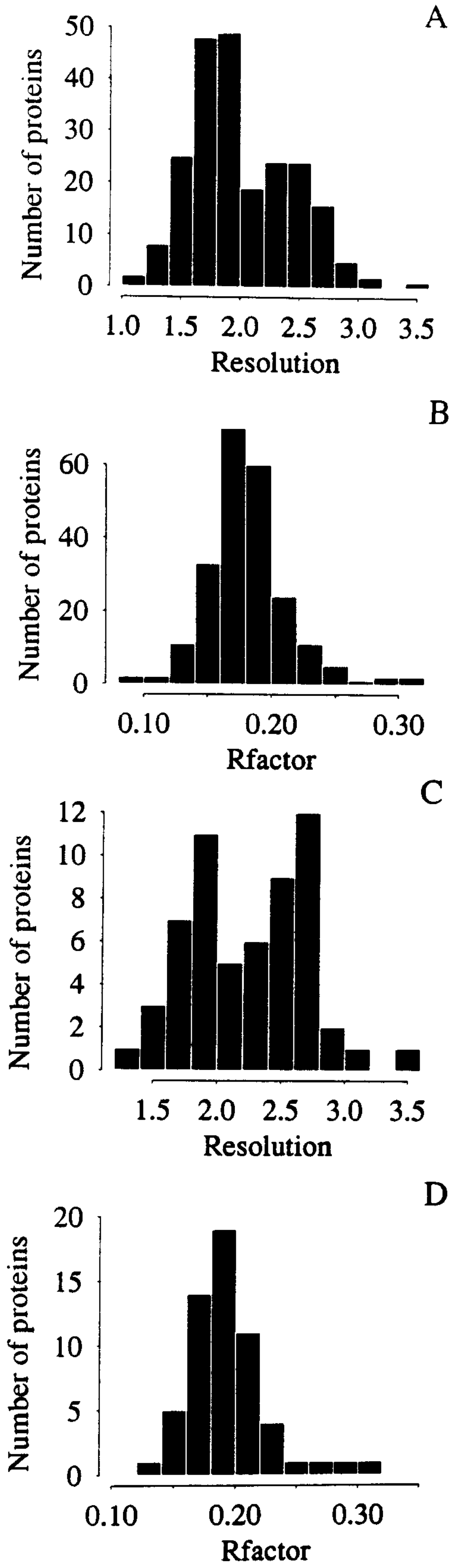

Fig. 1. Histograms showing the distribution of resolution **(A)** and R factor **(B)** in the crystal lattice contacts dataset, and the distribution of resolution **(C)** and R factor **(D)** in the oligomer subunit contacts dataset.

TABLE I. Protein Data Bank Entries Employed*

1AAP, 1AK3, **1ALC, 1APH, 1BBP, 1BP2,** 1C2R, **1CA2,** 1CC5, **1CCP, 1CCR, 1CDP,** 1CDT, 1CHO, **1CRN,** 1CSE, 1CTS, 1DHF, **1DRF, 1DTX,** 1FBX, 1FDL, **1FKF, 1FNR,** 1FXA, 1FXI, **1GCR,** 1HHO, 1HHP, **1HIP,** 1HNE, **1HOE,** 1HSA, **1IFB,** 1IGF, **1LPE,** 1LYM, **1LZ1, 1MBC, 1MBW, 1MCP, 1MCW,** 1MEE, **1OMD,** 1OVA, **1PAL, 1PAZ,** 1PBX, **1PCY,** 1PFC, 1PFK, **1PGX, 1PP2, 1Q21, 1R69,** 1RBB, **1RDG, 1RMS, 1RNH, 1S01,** 1SAR, **1SDH, 1SGT, 1SNC, 1STP, 1TGN,** 1TGS, **1TNF, 1TON,** 1TPA, 1TPK, 1TRM, **1UBQ,** 1ULA, 1UTG, 1WRP, **1YCC,** 256B, 2ABX, **2ACT, 2ALP, 2ATC,** 2AZA, **2BJL, 2CCY, 2CDV,** 2CGA, 2CPK, **2CPP,** 2CTS, **2F19, 2FBJ, 2FCR, 2FXB, 2GBP, 2GCH,** 2GCR, 2GN5, **2HAD, 2HBG,** 2HFL, 2HIP, **2HPR, 2I1B,** 2INS, **2LBP,** 2LHB, **2LIV, 2LTN, 2LYM, 2LZ2, 2LZT, 2MCG, 2MCM, 2MHR,** 2MLT, **2MM1,** 2PHH, 2PKA, **2PRK,** 2RHE, **2SDH, 2SGA,** 2SNI, 2SOD, **2ST1,** 2TEC, 2TGP, 2TRX, **2TSC, 2UTG, 2YPI, 2ZTA,** 3AAT, **3BLM, 3C2C, 3CHY,** 3CLA, **3CLN, 3DFR, 3EBX, 3EST, 3FGF, 3FXC, 3GAP, 3GBP,** 3GRS, 3HFM, **3ICB,** 3IL8, 3INS, **3LZM, 3MCG, 3P2P, 3PEP,** 3PHV, **3PSG, 3RN3,** 3RP2, 3SGB, **3TLN,** 3TMS, **451C, 4APE,** 4BLM, **4BP2, 4CLN, 4CMS, 4CNA, 4CTS,** 4DFR, 4ENL, **4FAB,** 4FBP, **4FD1, 4FXN, 4GPD, 4HHB,** 4ICD, **4MDH, 4P2P, 4PEP, 4PTP, 4RXN,** 4TMS, **4TNC,** 4XIA, **5ABP, 5ACN,** 5CHA, **5CSC,** 5CTS, **5CYT, 5DFR,** 5FBX, 5HVP, **5MBA, 5P21, 5P2P, 5PEP, 5TIM, 6CPA, 6DFR, 6FAB,** 6Q21, **6RXN,** 7API, **7DFR, 7PCY, 7RXN, 8ADH,** 8API, **8DFR, 8I1B, 8PTI,** 8RSA, 9HVP, **9INS, 9PAP, 9PTI, 9RNT, 9RUB, 9WGA**

*All protein crystal structures listed were employed for crystal contact surveys. The underscored proteins were employed in oligomer interface surveys. Bold face proteins have exactly one molecule per asymmetric unit and were used in calculating the results in Table II.

unit cell parameters varying less than 5%. This category includes the point mutations and various enzyme-inhibitor complexes. One representative variant was chosen from each set of multiple structures on the basis of resolution and R factor; neither an apoprotein nor a wild-type protein sequence was necessarily chosen over a holoprotein structure or over a mutant sequence.

4. Nonstandard representation of a space group and virus structures. The individual attention required by these proteins to interpret the crystal symmetry exceeds the value of the data lost by their elimination.

Application of the above criteria produced a set of 223 protein structures for the crystal lattice contact analyses (Table I). The same protein may appear twice in this list if the space group or unit cell parameters in the two structures are clearly different. The 58 oligomeric proteins which were employed in the examination of subunit interactions are underlined. They were chosen from the crystal dataset with the additional constraint that no two members of the oligomer dataset have over 80% sequence identity, to avoid overrepresentation of a particular type of oligomer interface. The proteins that provided data for Table II are shown in boldface. These proteins have exactly one molecule per asymmetric unit.

Where more than one conformation for a protein residue is specified in the coordinate file, the one assigned higher occupancy was used for this analysis. Where alternate conformations with the same occupancy occur, the first one listed in the coordinate set was used.

Computational Methods

All routines for the analyses have been written in 'S', an interactive programming environment for data analysis with high-level functions for graphics and statistical calculations.[24] A program suite designed to perform protein structure computations, called Protein Knowledge Base,[25] is an application of 'S.' The Protein Knowledge Base combines a relational database of three-dimensional protein structures derived from the Protein Data Bank with algorithms for pattern recognition, data analysis, and graphics. The ADXP routine from the Protein Knowledge Base was used to generate the list of atom pairs in contact with distances between two molecules or subunits.

To determine surface patches, a cluster analysis was performed on the atoms involved in contacts for the crystal and oligomer datasets. The "connected" method for cluster analysis in the 'S' programming language[24] considers the minimum distance between two clusters to be the criterion for joining clusters. Initially, each point, in this case each atom, is an isolated cluster, which then joins other points or clusters separated by the shortest distance. This algorithm identifies continuous groups of atoms when applied to surface residues because if any path exists between any atom or group of atoms that involves a gap of no more than 5.0 Å, then the two clusters are combined into a single one.

Secondary structure was determined by using routines that are part of the Protein Knowledge Base,[25] based on criteria first employed by Liebman and colleagues.[26]

Characterization of Contacts

For the purposes of this study, a protein composed of more than one polypeptide chain in a complex that is stable under physiological conditions (i.e., the biologically relevant entity) is considered to be one molecule. Thus, a hemoglobin tetramer is spoken of as a single molecule, even though by analysis of the covalent bonding each subunit might be considered to be an individual molecule in another context.

For the analysis of patches, unless otherwise specified, two atoms were considered to be in contact if their center-to-center distance was less than 5.0 Å.

TABLE II. Number of Proteins With Different Combinations of Space Group and Coordination Number in the Crystal Lattice*

Space group	Total in sample	Coordination number										
		4	5	6	7	8	9	10	11	12	13	14
P $2_12_12_1$	45	1		5		14		12		11		2
P 2_1	21					3		11		6		1
P 2_12_12	9				1	2		2	2	2		
C 2	9			1	1	2	2	2	1			
P 3_121	8		1		5	2						
P 1	7							2		3		2
P 3_221	6		1	1	1	1				1		1
P 4_32_12	5	1			1		1	1	1			
C 222_1	4			1			2					1
P 4_3	3	2						1				
P 4_12_12	3			2					1			
P 4_22_12	3			2				1				
I 222	2			1		1						
R 3	2			1				1				
P 6_1	2	1				1						
P 6_3	2			1		1						
P 6_122	2				1	1						
P 6_522	2		1	1								
I 2_13	2			2								
P 4_1	1			1								
P 4_2	1			1								
I 4	1							1				
I 4_122	1	1										
R 32	1				1							
P 6	1			1								
P 6_5	1			1								

*All contacts generated from a single symmetry operator form a single coordinating interaction.

Distances much larger than 5.0 Å might allow a water molecule to interpose between two atoms that are putatively in contact.

Oligomer contacts were classified as such, whether they happened to be part of an interface between asymmetric units in the crystal or occurred within the asymmetric unit. Oligomer interfaces were usually identified from information contained in the Protein Data Bank entry or, occasionally, from references listed within each entry. All other intermolecular interfaces were included as crystal contacts.

For contacts between asymmetric units, the Protein Data Bank SCALE transformation in each entry was used to calculate fractional crystallographic coordinates from the orthogonal coordinates in the Protein Data Bank files. From the fractional coordinates, all asymmetric units were generated by applying the set of symmetry operations associated with the space group of that protein entry. The 728 adjacent unit cells, ranging over the set from −4 to +4 along each principal axis, were then generated by translation. This large number was essential to avoid missing contacts because Protein Data Bank asymmetric units are not necessarily at the origin.

Some of the translated asymmetric units were much too distant from the original to contribute any contacts. The center of the original asymmetric unit was computed as the mean of the coordinates, and the distance from the center to the farthest atom in the structure was calculated. If twice that distance plus 6.0 Å was less than the distance between the centers of the original and any symmetry-related asymmetric unit, then that symmetry-related asymmetric unit was eliminated from consideration.

Bound water molecules were not employed explicitly in this analysis because criteria for identification of electron density peaks as solvent molecules varies widely among protein crystallographers. Instead, the analysis of contacts was carried out in a distance-dependent fashion to a maximum interatomic distance of 6 Å, so that information about interactions bridged by solvent or ions would be implicit in the pairwise interaction potentials.

The quality of X-ray structures in the Protein Data Bank is not uniform,[27] and not all experimenters exercise the same care to avoid steric clashes at crystal contacts. The count of atom pairs in contact will be artificially inflated where amino acids are misplaced in this manner. Of all of the contacts observed, 0.16% were at an unreasonably short distance when judged against frequently used short contact distances for unified atom models.[28] For individual structures, this statistic varied from a high of 2.04% to a low of 0%. The number of short contacts does not seem high enough to be a major source of error.

The observed contacts for crystal (oligomer) interfaces were tabulated as a four-dimensional array with 20 × 20 × 4 × 223(58) elements, where each cell in the array contained the number of atom–atom contacts for that residue pair, for that distance interval, for that protein, N_{rsdp}. Observed interactions were counted and tabulated by considering interactions in both directions across each interface. In this way, the natural symmetry of the contact matrix could be preserved, under the reasonable assumption that a particular contact type between residue *i* on one molecule (subunit) and residue *j* on another molecule (subunit) is equivalent to a contact between residue *j* on the first entity and residue *i* on the second.

Reference State

The significance of each pairwise contact at a particular distance can be evaluated from the comparison of the number of contacts observed with the number of contacts that might be expected by mass action (i.e. by random pairing).

In accordance with the assumption that solvent exposure of a residue is directly related to its probability of forming random contacts, accessible surface area[29] might be used as the basis of a reference state to compute the number of random contacts expected. However, because the observed contacts are categorized by distance, a distance-dependent reference state is preferred. Shrake and Rupley[30] first suggested a method with "near" and "long" test atoms to determine the exposure of an atom. For this work, the opportunity for residues of type *r* on a specific protein *p* to form random contacts within a distance interval *d* is assumed to be directly proportional to C^{g}_{rdp}, the number of contacts that that residues type makes with a grid of points external to the protein. For computation of C^{g}_{rdp} a three-dimensional grid is constructed about the protein such that the extremities are 6 Å beyond any atom in any direction. Grid points are placed 1.3 Å apart along the standard orthogonal axes used for Protein Data Bank entries. Contacts between the constructed grid and the original protein molecule are computed, and grid points within 2.6 Å of a protein atom are eliminated, resulting in a "hole" in the grid. The crystal contact reference state is obtained by computing the contacts between the remaining grid points (the grid with a hole) and each atom in the protein molecule. In Figure 2, the accessible surface area is compared with the results of the grid method for each residue in eight proteins, including two oligomers. This figure shows that grid results can be closely related to accessible surface area, given appropriate grid parameters.

Equation (1) allows the calculation of the number of expected contacts, N^{0}_{rsdp}, from grid contacts, where summations are over all residue types *i* and *j*.

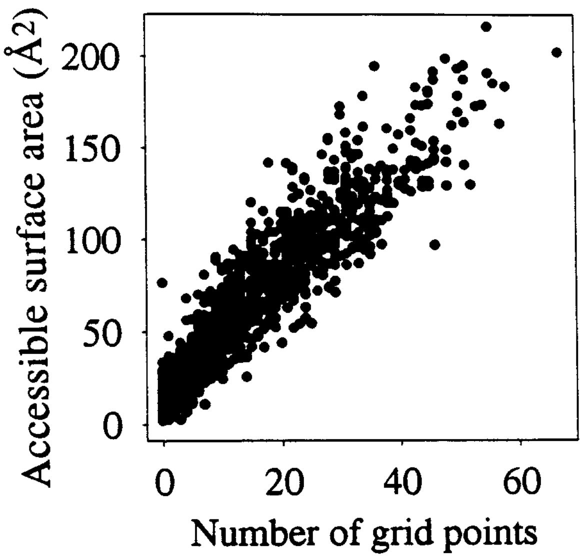

Fig 2. The accessible surface area versus the number of grid points in contact for each amino acid in 8 proteins: 1CRN, 1SNC, 2FCR, 2TSC, 2ZTA, 3ICB, 5ACN, and 5MBA. Correlation = 0.95. Accessible surface area was calculated with the program ACCESS, of Lee and Richards.[29] Grid parameters were chosen to correspond to the probe size employed in the accessible surface area calculation: minimum grid distance = 2.6 Å, maximum grid distance = 3.8 Å, distance between grid points along axes = 2.1 Å.

$$N^{0}_{rsdp} = \frac{C^{g}_{rdp} \times C^{g}_{sdp}}{\sum_{i}\sum_{j} C^{g}_{idp} \times C^{g}_{jdp}} \times \sum_{i}\sum_{j} N_{ijdp} \qquad (1)$$

The product of each value of C^{g}_{rdp} with every value of $C^{g}{}_{sdp}$ provides a number proportional to the number of expected random pairings of residues *r* with *s* for that protein and distance interval. In order to obtain the probability of forming a contact, this product is divided by the sum over all such products, which serves to normalize the sum of all probabilities to unity. Finally, in order to obtain the expected number of contacts between a given pair of residues within a given distance interval, this probability is multiplied by the total number of contacts observed for that protein in that distance interval.

When interactions between two identical subunits or molecules involve a symmetry operator (such as a twofold axis), the distribution of contacts is not strictly independent but is conditionally dependent on the distribution of residues on the surface of that entity. However, since the location and the orientation of the symmetry axes are variable under the null, the independent distribution used here [equation (1)] is a good approximation of the mass action reference state.

For crystal contacts, the reference state algorithm operates on the coordinates of the whole protein molecule, whether it is less than, equal to, or greater than the contents of the asymmetric unit. The reference state for oligomer contacts is computed in a very similar manner except that grid point contacts are computed with each individual chain as if it were an isolated molecule. Since contacts are observed in both directions between two molecules (subunits), reference state contacts were treated in the same way.

Log-Linear Models

A four-dimensional contingency table is constructed with margins of 20 residue types × 20 residue types × 4 distance intervals × 223 proteins for the crystal contact data (58 proteins for the oligomer contact data). Each cell of the tables is C_{rsdp}. This ratio [equation (2)] is the deviation from mass action behavior of contact formation between residues.

$$C_{\mathrm{rsdp}} = \frac{N_{\mathrm{rsdp}}}{N^{0}_{\mathrm{rsdp}}} \tag{2}$$

In order to explain any statistical dependence between the categorical variables and the deviation from mass action in the table, log-linear models were tested. Details of the analysis have been described.[31,32] In brief, assuming a Boltzmann-like relationship, relative chemical potentials can be defined by:

$$\mu_{\mathrm{rsdp}} = -\ln C_{\mathrm{rsdp}} \tag{3}$$

The negative natural logarithm of the contents of each cell is estimated as the sum of *ln* odds that correspond to some or all of the terms of equation (4).

$$\mu_{\mathrm{rsd}} = \mu'_{\mathrm{r}} + \mu'_{\mathrm{s}} + \mu'_{\mathrm{rd}} + \mu'_{\mathrm{sd}} + \mu'_{\mathrm{rs}} + \mu'_{\mathrm{rsd}} \tag{4}$$

Each term gives increments in *ln* odds associated with each categorical variable. The simplest model attempts to explain contact frequencies in each cell from only the type of residue involved in contacts, but not accounting for any distance or pairwise effects of contact formation. This model corresponds to using only the first two terms of equation (4). Other terms of equation (4) add distance dependence to the residue variable, pairwise specificity of interactions, and distance-dependent pairwise interactions. Note that, because of equation (3), a negative potential corresponds to an interaction that is found more frequently than predicted by mass action, and that might be judged to be energetically favorable.

One method of distributing the terms of equation (4) that has proven useful[32] is given in equation (5). In this case, departures from mass action are considered as either distance-dependent hydrophobic (*H*) or pairwise (*P*) interactions. Although the term "hydrophobic interaction" has been employed,[32] "nonpairwise interactions" is preferred terminology in this work because those residues that score high in "hydrophobic" or nonpairwise potential in crystal formation are usually quite polar. Note also that, because the data from which these parameters are derived are symmetric with respect to pairwise contacts, the first two lines in equation (5) are essentially interchangeable.

$$\begin{aligned} \mu^{\mathrm{H}}_{\mathrm{rd}} &= \mu'_{\mathrm{r}} + \mu'_{\mathrm{rd}} \\ \mu^{\mathrm{H}}_{\mathrm{sd}} &= \mu'_{\mathrm{s}} + \mu'_{\mathrm{sd}} \\ \mu^{\mathrm{P}}_{\mathrm{rsd}} &= \mu'_{\mathrm{rs}} + \mu'_{\mathrm{rsd}} \end{aligned} \tag{5}$$

The method of iterative proportional fitting was used to find the maximum likelihood estimates for parameters.[33] The most elaborate model considered, which includes distance-dependent pairwise interactions, contains 2200 parameters. For the crystal data, 1,310,366 pairwise contacts were observed, resulting in an average of 596 observations per parameter. For the oligomer dataset 363,192 pairwise contacts were observed, resulting in an average of 165 observations per parameter.

A cell in the contingency table may contain no contacts, although a contact is possible according to random pairing. The zero in such a cell contains information that contributes to the calculation of parameters. However, if the numerator and denominator of equation (2) are both zero because residue *r* or *s* does not appear on the surface of that protein, then that cell contains no information about the likelihood that those two residues will be involved in a contact, and the corresponding cell is therefore assigned a "structural zero." These structural zeros do not contribute to the parameter estimation. If the number of such cells over all proteins is large, then the sample will be too small to estimate some of the parameters involving that cell. In the crystal contact data, the minimum number of cells used for estimation of any parameter was 155. In the worst case within the oligomer data, 16 of the 58 proteins contained a structural zero, leaving 42 values to be evaluated in the fitting of parameters. These numbers of cells should be adequate for estimation of the parameters required by the various models.

RESULTS

Coordination Numbers for Proteins in Crystals

The number of other protein molecules in contact with the protein molecule in an asymmetric unit can be considered as a coordination number. Of the 223 proteins in the crystal contact dataset only 144 were appropriate to use for this analysis. Proteins with more or less than one molecule in the asymmetric unit were excluded since no single coordination number can be clearly defined for these situations.

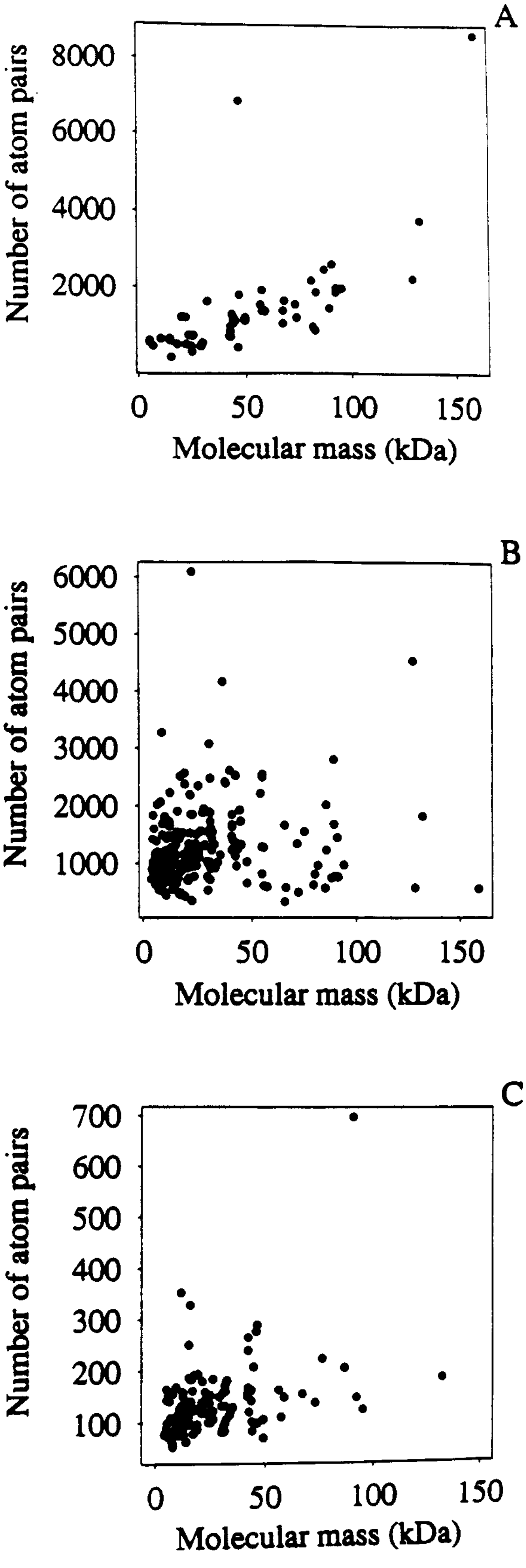

Fig. 3. The number of atom–atom contacts plotted versus protein molecular mass. **A:** Oligomer subunit interfaces. **B:** Crystal lattice interfaces. **C:** The number of atom–atom contacts at crystal interfaces divided by the number of neighboring molecules (coordination number, Table II) is plotted versus the protein molecular mass.

Table II shows the distribution of coordination numbers as a function of space group. The distribution of coordination numbers is broad, with a range from 4 to 14. Apparently in some protein crystals the molecules form a complex network of interactions with multiple neighbors, while in others contacts occur with only a few adjacent neighbors. Even though space group P $2_12_12_1$ contains 31% of the entries, no odd coordination number was found for these proteins. Odd-valued coordination numbers are only present in space groups with twofold rotational symmetry intrinsic to the space group, as has been observed previously by Janin and Rodier[12] for a slightly larger set of proteins.

Atoms in Contact

Figure 3 shows the total number of atom–atom contacts within 5.0 Å at oligomer interfaces and at crystal interfaces as a function of molecular mass. Larger proteins do tend to form a more extensive number of pairwise contacts in oligomers. However, at crystal interfaces, no close relationship exists between the size of a protein and the number of atom–atom contacts.

The number of pairwise contacts may be divided by the coordination number (Fig. 3C) to obtain the average number of contacts per intermolecular interface. The number of contacts between pairs of molecules at individual crystal interfaces is of the order of 0.1 of the number of contacts that occur between subunits at oligomer interfaces.

Contact Patches

When two proteins (or subunits) interact, one could conceive of two extreme cases. At one extreme, all contacts could form a single continuous surface at the interface. At the other extreme, the interactions could occur as so many isolated contacts. A patch is defined as a collection of adjacent atoms on the surface of a protein, all of which are involved in contacts, and all of which are connected by some pathway that contains no gap greater than 5 Å.

Figure 4A,B shows the number of patches for the two types of interfaces as a function of the mass of the molecule. The numbers of patches for oligomer contacts are generally low, whereas the numbers of patches for crystal contacts range significantly higher. However, in evaluating this observation, one should consider that crystal contacts between proteins must form a three-dimensional network. By contrast, for oligomer contacts, the geometry required for formation of common multimers is more consistent with patches confined to the same region of the protein surface.

The large coordination numbers for proteins in crystals (Table II) might be construed to imply that the patch number should be 4 to 14 times higher for crystal contacts than for subunit contacts, if each interface between neighbors was made up of at least one unique patch. Figure 4C shows the crystal interface patch information corrected for coordination number. Note, however, that interfaces between several neighboring molecules in the crystal sometimes form a single patch, so that dividing the patch number by the coordination number overcorrects for the geometric constraint.

The number of atoms in the largest patch as a percentage of the total number of atoms involved in contacts for that protein are shown as two histograms in Figure 5. Oligomer interfaces (Fig. 5A) usually have a single patch containing greater than 90% of the atoms in contact. On the other hand, considering all crystal interfaces that a reference molecule makes (Fig. 5B), the largest patch will often contain only a small portion of the total number of contacts. Again, the geometric constraint of forming a three-dimensional network may be a cause of this difference.

Many of the proteins that contribute to the rightmost bar in Figure 5B have molecular mass <20,000 Da, the situation where having only one or two crystal contact patches is fairly common (Fig. 4B). In other words, for small proteins, all crystal contacts across all interfaces to all neighbors can be one big patch. For larger proteins, this situation rarely occurs.

Ultimately, a clearer difference in the spatial organization of interfaces can be shown. Figure 6 contains histograms showing the log of the number of atoms in each patch in the datasets. Small patches occur in both oligomer and crystal interfaces. Medium-sized patches (containing 10 to 100) atoms are characteristic of crystal interfaces. Large patches (containing 100 to 1000 atoms), although very common in oligomer interfaces, are not frequently found at crystal interfaces.

The distribution of the average size of patches for each protein in the oligomer and crystal contact datasets is shown in Figure 7. The average patch size is much smaller for crystal interfaces than for oligomer interfaces, consistent with the presence of few very large patches at crystal interfaces.

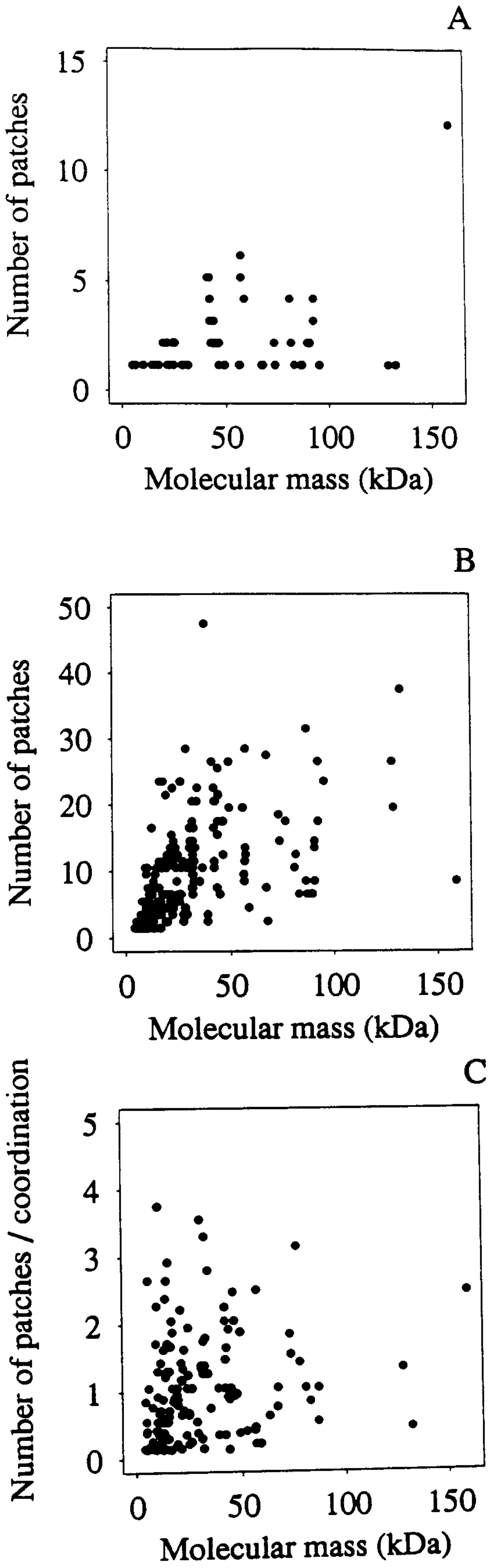

Fig. 4. The number of patches of contacts involved in contacts plotted versus protein molecular mass. **A:** Oligomer subunit interfaces. **B:** Crystal lattice interfaces. **C:** The number of patches divided by the coordination number in Table II is plotted versus molecular mass.

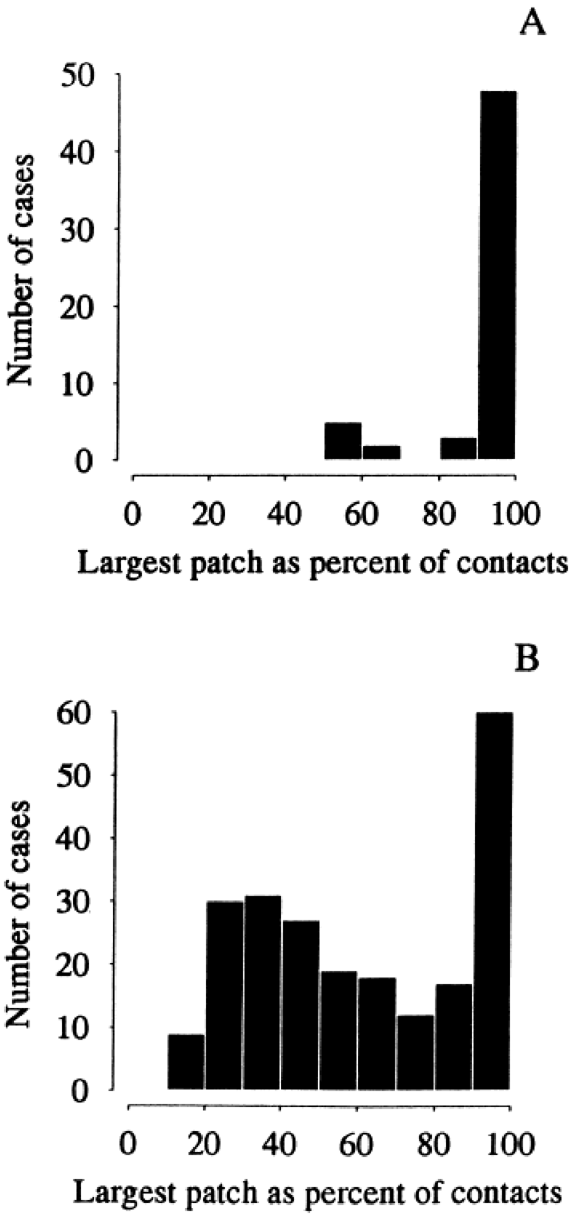

Fig. 5. Histogram of the number of atoms in the largest patch as a percentage of the total number of atoms involved in interactions in that protein. **A:** Oligomer subunit interfaces. **B:** Crystal lattice interfaces.

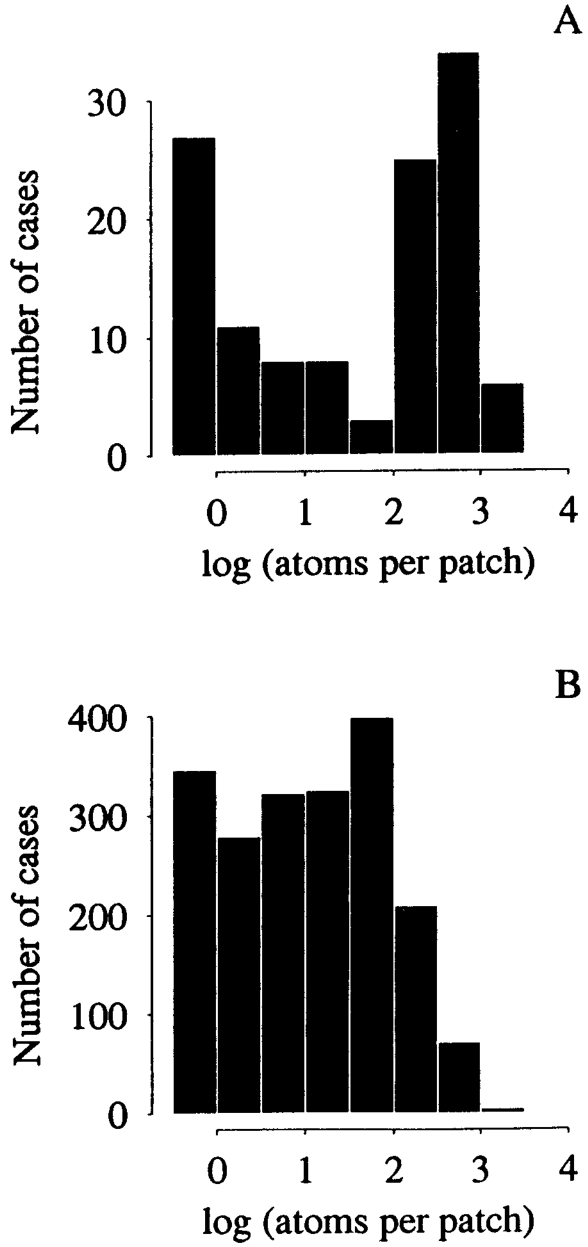

Fig. 6. Histogram of the occurrence of various patch sizes, plotted on a logarithmic scale. **A:** Oligomer subunit interfaces. **B:** Crystal lattice interfaces.

Secondary Structures of Contact Residues

Table III lists the percentage distribution of the secondary structural elements (helix, strand, or coil) at the two types of interface. The differences in the percentages are small.

Residues in Contact

To contrast the nature of the contacts involved in the association of subunits in oligomers and protein molecules in crystals, the numbers of contacts involving atoms from each amino acid type are plotted in Figure 8 as a percentage of the total contacts. The residues most frequently found involved in contacts at oligomer interfaces are R, F, L, Y, and T, in order of decreasing number of contacts. The top five residues at crystal contacts are K, E, R, S and D.

The types of atom–atom pairing observed in the two classes of interfaces is shown in Figure 9. Contacts between carbon atoms are consistent with hydrophobic interactions. Oligomer interfaces show a higher percentage of carbon–carbon contacts. Heteroatom–heteroatom contacts, which are consistent with salt bridges and hydrogen bonds, are less common in oligomer interfaces. Both types of inter-

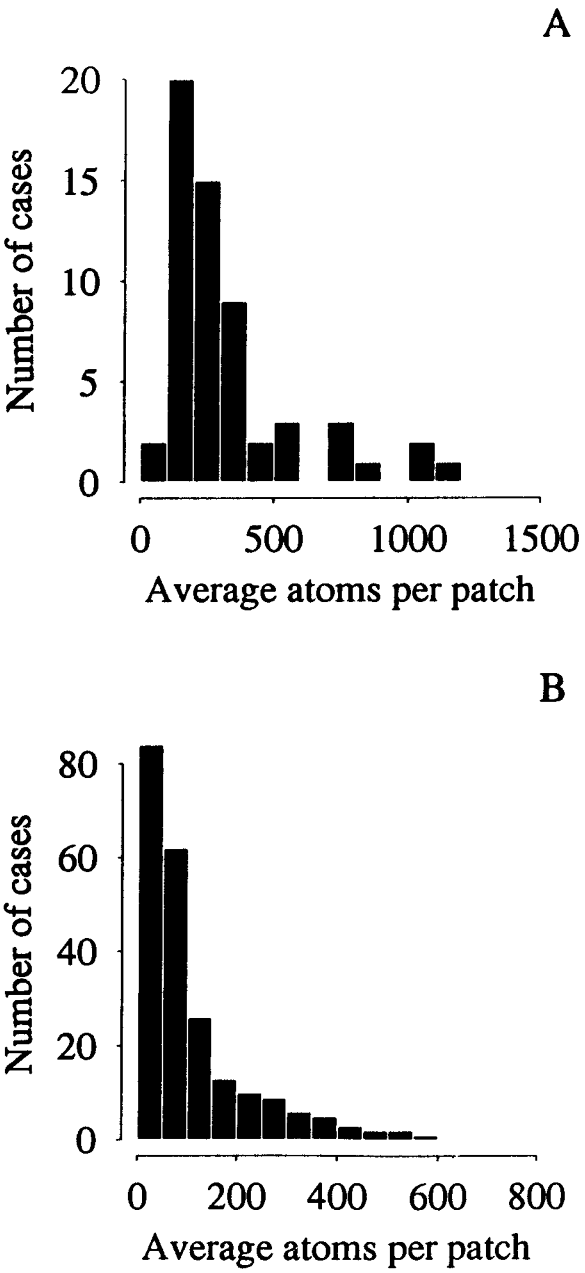

Fig. 7. Histogram of the average patch size among the proteins in each dataset. **A:** Oligomer subunit interfaces. **B:** Crystal lattice interfaces.

faces have a high percentage of "mismatched" contacts between heteroatoms and carbon atoms. The difference in distribution of contact types between oligomer and crystal interfaces is moderate.

Models for Contact Preferences

Beyond the number of contacts observed by residue (Fig. 8), information regarding preferential interactions is available if the number of contacts are divided by the number expected by mass action. As detailed under the Methods section, various models with a range of complexity might be invoked to explain the dependence of that ratio on the identity of the variables along the margin of the contact contingency tables.

TABLE III. Secondary Structure Type at Interfaces

	Oligomer interface	Crystal interface
This work		
Helix	40.4	36.3
Sheet	36.7	37.7
Coil	22.9	26.1
Argos[14]		
Helix	23.7	—
Sheet	16.7	—
Coil	30.3	—
Turn	29.3	—
Jones and Thornton[16]		
Helix	41	—
Sheet	12	—
Coil	34	—
Turn	12	—

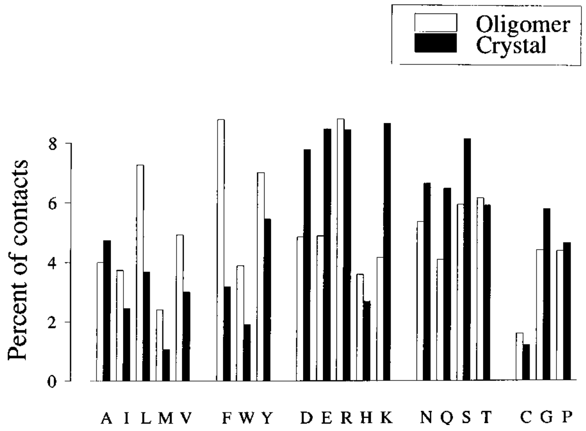

Fig. 8. Percentage of atom–atom contacts which are due to each amino acid type at oligomer subunit and crystal lattice interfaces.

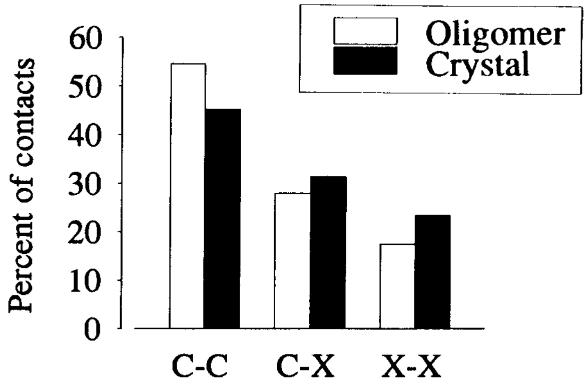

Fig. 9. Percentage of atom–atom contacts that involve only carbon atoms (C–C), only heteroatoms (X–X), or mixed carbon–heteroatom types (C–X), for both oligomer subunit and crystal lattice interfaces.

TABLE IV. Statistics for Stepwise Log-Linear Modeling of Relative Contact Frequencies

Model*	Component	$g^2 \times 10^{-5}$ †	$\Delta g^2 \times 10^{-4}$	$\%\Delta g^2$ ‡	*df*	Δ*df*	$\chi^2 \times 10^{-6}$
Oligomer contacts							
None	constant	5.399			92799		
1	μ'_r, μ'_s	4.914	4.85	51.2	92761	38	1.891
2	$+\mu'_{rd}$, μ'_{sd}	4.906	0.08	0.8	92644	117	1.823
3	$+\mu'_{rs}$	4.502	4.04	42.7	92283	361	1.106
4	$+\mu'_{rsd}$	4.452	0.50	5.3	91200	1083	1.021
Total			9.47	100			
Crystal contacts							
None	constant	13.978			356799		
1	μ'_r, μ'_s	13.666	3.12	27.8	356761	38	35.953
2	$+\mu'_{rd}$, μ'_{sd}	13.530	1.36	12.1	356644	117	16.600
3	$+\mu'_{rs}$	12.986	5.44	48.6	356283	361	7.009
4	$+\mu'_{rsd}$	12.858	1.28	11.4	355200	1083	5.998
Total			11.20	100			

*Model 1 considers preferences in amino acid incorporation at interfaces, independent of what other amino acid is forming the contact across the interface. Model 2 adds distance dependence. Model 3 adds pairwise contact preferences without distance dependence. Model 4 adds distance dependence to the pairwise contact model.

†The probability of finding the observed reductions in g^2 by chance, given the change in the number of degrees of freedom with addition of each component, $P < 0.001$, in all cases.

‡$\%\Delta g^2$ is the proportion of the overall goodness-of-fit that may be attributed to the components added to the overall model one at a time, as derived from the change in the g^2 value between successive models. The corresponding percentages for the stepwise addition of models 1 to 4 using internal contact data[32] were 62.8, 5.5, 13.9, and 17.9.

The usefulness of increasingly complex models must be justified. In addition, statistics identifying which components account for the most "fitting power" provide information regarding the relative importance of various interaction terms [equation (4)]. Table IV shows the statistics for the stepwise addition of components to the models employed to explain the observed contact preferences for both oligomer and crystal interfaces. All of the components contribute significantly to the fit of the models to the two datasets. However, in comparing oligomer and crystal contacts, differences are observed in which components contributed most to the overall reduction of g^2, the likelihood ratio statistic. For oligomer contacts, just over 50% of the fit is due to nonpairwise interactions, for which the distant-dependent component contributes only a small amount. For crystal contacts, only 40% of the fit is due to nonpairwise interactions, and those interactions are more distance dependent than they are in oligomer contacts. Conversely, pairwise interactions play a larger part in crystal contacts than in oligomer contacts. For internal protein contacts, to make another relevant comparison, over two-thirds of the fit is due to nonpairwise interactions (Table IV, footnote). In summary, these data indicate that nonpairwise interactions play a dominant role in internal contacts, an important role in oligomer contacts, and a smaller role in crystal contacts.

Pairing-Independent Preferences

Table V shows the nonpairwise component of the contact potentials for oligomer, crystal, and internal contacts. Oligomer contact potentials and internal contact potentials are correlated, but at least one systematic difference is observed. While internal contacts show an approximately equally favorable potential for any nonpolar amino acid, hydrophobic interactions at oligomer interfaces form with a preference for methionine and aromatic residues.

Crystal contacts have a very different pattern of pairing-independent preferences: so much so, that the term "hydrophobic potentials" does not seem to be an apt description.[32] Of the nonpolar amino acids, only tyrosine is preferred in a nonpairwise manner, perhaps because its side chain can associate both through polar and nonpolar interactions. Although, in general, polar and charged amino acids are preferred at crystal interfaces, lysines are selected against. The most favorable amino acids to have at such interfaces, independent of any pairwise preferences, are glutamine and arginine. This situation is very different from either oligomer or internal contacts where preferential association through nonpolar residues in a nonpairwise manner suggest a true hydrophobic component to the binding free energy. In crystal interface formation, hydrophobic contact formation is actually avoided.

Pairwise Preferences

Tables VI and VII give the pairwise potentials (equation 5) for oligomer and crystal contacts, respectively. These potentials give information on chemical complementarity between pairs of residues, or the lack thereof, independent of nonpairwise considerations such as the hydrophobic effect.

Such a large mass of data is difficult to assimilate. Therefore, in order to represent part of it in a

TABLE V. Nonpairwise Contact Potentials for Oligomer, Crystal, and Internal Contacts

	Oligomer				Crystal				Internal†	
Distance	0 to 4.5	>4.5 to 5.1	>5.1 to 5.6	>5.6 to 6.0	0 to 4.5	>4.5 to 5.1	>5.1 to 5.6	>5.6 to 6.0	0 to 5.0	>5.0 to 6.0
A	0.24	0.23	0.18	0.13	−0.01	0.03	0.01	0.01	−0.24	0.08
I	−0.08	−0.09	−0.15	−0.10	0.30	0.19	0.13	0.20	−0.29	−0.44
L	−0.09	−0.23	−0.21	−0.23	0.12	−0.05	−0.08	−0.08	−0.13	−0.35
M	−0.35	−0.34	−0.31	−0.31	0.15	−0.06	−0.09	−0.26	−0.20	−0.14
V	−0.08	−0.14	−0.17	−0.14	0.21	0.08	0.12	0.11	−0.29	−0.28
F	−0.87	−0.92	−0.83	−0.84	0.03	0.04	0.08	0.06	−0.26	−0.23
W	−0.50	−0.56	−0.49	−0.45	0.18	0.18	0.17	0.22	−0.18	−0.16
Y	−0.40	−0.29	−0.27	−0.36	−0.19	−0.14	−0.10	−0.07	−0.22	−0.23
D	0.77	0.85	0.74	0.81	0.02	0.04	0.04	−0.04	0.38	0.44
E	0.39	0.46	0.44	0.46	0.12	0.13	0.13	0.12	0.46	0.39
R	−0.03	0.04	0.06	0.12	−0.30	−0.21	−0.19	−0.18	0.37	0.20
H	−0.09	0.00	−0.05	−0.08	−0.26	−0.09	−0.08	−0.10	−0.04	0.03
K	0.88	0.75	0.76	0.68	0.31	0.24	0.23	0.22	0.50	0.48
N	−0.17	−0.04	0.04	0.06	−0.12	−0.03	−0.06	−0.04	0.32	0.22
Q	−0.03	0.10	0.09	0.11	−0.27	−0.18	−0.11	−0.09	0.29	0.21
S	0.06	0.09	0.05	0.14	−0.14	−0.10	−0.10	−0.07	0.10	0.15
T	0.05	0.06	0.00	0.07	−0.10	−0.06	−0.03	0.00	0.06	0.08
C	−0.07	−0.27	−0.33	−0.44	0.30	0.29	0.29	0.21	−0.33	−0.38
G	0.12	0.15	0.22	0.21	−0.24	−0.21	−0.18	−0.14	−0.01	0.18
P	0.12	−0.01	0.08	0.00	−0.05	−0.05	−0.12	−0.08	−0.02	0.22

†Nonpairwise potentials provided from Bryant and Lawrence[32] for qualitative comparison only. Distance intervals were measured from the side chain centroid for these potentials rather than from atom center to atom center. Also, a different reference state was employed. Both of these differences in methodology may affect the absolute magnitude of these potentials but the ordering of contact preferences should be little affected.

manner more easily comprehended, the pairwise potentials for formation of oligomer and crystal contacts are compared in Table VIII. Each pairwise interaction is categorized according to its value in the first distance interval as being very unfavorable, unfavorable, slightly unfavorable, slightly favorable, favorable, or very favorable. The majority of pairwise interactions lie near the center of the figure, indicative of weak pairwise preferences. Interactions listed along the diagonal from upper left to lower right have similar contact potentials in both oligomer and crystal interfaces. Potentials shown in italics, away from the diagonal, are those that differ by two or more categories between the two types of interfaces.

Even this simplification leaves a large amount of information to be considered. A few of the pairwise interactions in Table VIII have characteristics that seem worthy of being highlighted. For example, the C—C pair is favored strongly in both oligomer contacts and crystal contacts. If two cysteines occur on two adjacent molecules, a very strong tendency is observed that they be found close together, presumably in the formation of a disulfide linkage. Although this interaction is very favorable, cysteine residues are rare at crystal or oligomer interfaces (Fig. 8), and thus this interaction could not be abundant.

Other pairwise contacts strongly favored at both oligomer and crystal contacts are N–N, M–M, C–R, W–P, K–E, V–W, H–H, W–K, and F–M. The reason for some of these favorable pairings might be easily postulated, such as the charge complementarity in the K–E pairing. The favorable H–H interaction suggests a transition metal ion-mediated interaction. In other cases, the chemical interaction that leads to favorable potentials are obscure. Some of the properties of the side chains that might play a part in these preferences are shape complementarity, length, polarity, polarizability, and binding of solvent molecules.

Strongly disfavored pairs at both interface types are Y–C, H–A, S–M, F–D, and V–E. The same properties involved in favorable interactions also probably play a role here. For example, a negatively charged aspartic acid near an electron-rich, nonpolar phenylalanine is apparently a strongly disfavored juxtaposition of residues. Perhaps the relatively short side chain of aspartic acid also plays a role if it rarely extends far enough from the surface of a protein to form an interaction with a phenylalanine. Indeed, size seems to play a role, since E–F interactions are only slightly disfavored in crystal and oligomer interfaces.

Table IX compares pairwise potentials for oligomer and internal contacts. Again, most pairwise preferences are weak, and occur near the center of the diagram. Strongly favored pairings in both oligomer and internal contacts are K–E, W–K, W–I, F–M, V–L, W–I and C–C. Strongly disfavored pairs are P–I and Y–T.

Considering all three types of interfaces, only three pairwise interactions, K–E, W–K, and C–C, are consistently strongly favored. Of these three, the

TABLE VI. Pairwise Potentials for Oligomer Contacts*

	A–A	R–A	R–R	N–A	N–R	N–N	D–A	D–R	D–N	D–D	C–A	C–R	C–N
0.0–4.5	−0.23	0.07	−0.94	−0.07	0.32	−0.86	0.52	−1.12	0.05	−1.02	2.40	−0.71	−0.43
4.5–5.1	−0.22	0.17	−0.90	−0.55	0.33	−0.67	0.60	−0.97	0.32	−1.01	0.32	−0.18	−1.13
5.1–5.6	−0.02	0.23	−0.90	0.21	0.40	−0.83	−0.30	−0.85	−0.01	−0.98	0.26	−0.49	−0.49
5.6–6.0	0.21	0.38	−0.63	−0.23	0.48	−0.73	−0.26	−0.79	−0.04	−0.95	0.32	−0.10	−0.62
	C–D	C–C	Q–A	Q–R	Q–N	Q–D	Q–C	Q–Q	E–A	E–R	E–N	E–D	E–C
0.0–4.5	−0.67	−3.38	0.59	−0.44	0.19	−0.35	−0.73	−0.37	−0.07	−0.10	0.35	1.31	−2.10
4.5–5.1	−0.49	−2.87	0.70	−0.56	0.10	−0.25	−1.33	−0.51	0.30	−0.07	0.32	0.77	−1.90
5.1–5.6	−0.65	−2.13	0.54	−0.60	0.22	0.21	−1.59	−0.03	0.19	−0.02	0.20	0.33	−1.80
5.6–6.0	−0.26	−1.77	0.72	−0.59	0.45	0.11	−1.63	−0.35	0.13	−0.24	0.15	0.23	−1.50
	E–Q	E–E	G–A	G–R	G–N	G–D	G–C	G–Q	G–E	G–G	H–A	H–R	H–N
0.0–4.5	−0.25	−0.48	−0.78	0.23	−0.22	−0.05	−1.26	−0.20	0.78	−0.02	0.70	0.80	−0.35
4.5–5.1	−0.01	−0.78	−0.54	0.08	−0.19	−0.19	−1.52	−0.09	0.73	−0.15	0.54	0.98	−0.79
5.1–5.6	−0.01	−0.72	−0.49	−0.01	−0.05	0.00	−1.26	−0.23	1.03	−0.11	0.74	0.99	−0.48
5.6–6.0	0.26	−0.56	−0.35	−0.25	−0.01	0.36	−0.18	−0.24	1.02	0.11	0.50	1.30	−0.76
	H–D	H–C	H–Q	H–E	H–G	H–H	I–A	I–R	I–N	I–D	I–C	I–Q	I–E
0.0–4.5	−0.23	−0.21	0.28	−0.39	0.04	−0.90	−0.39	0.57	−0.24	0.09	0.29	−0.48	0.42
4.5–5.1	−0.33	−0.26	0.38	0.09	0.63	−0.72	0.05	0.25	0.25	0.00	0.73	−0.24	0.06
5.1–5.6	−0.07	−0.86	0.23	−0.37	0.04	−0.60	0.04	0.31	0.14	0.47	0.71	−0.25	0.23
5.6–6.0	−0.09	−0.90	0.11	−0.38	0.22	−0.52	−0.28	−0.03	0.00	0.06	0.86	−0.08	0.23
	I–G	I–H	I–I	L–A	L–R	L–N	L–D	L–C	L–Q	L–E	L–G	L–H	L–I
0.0–4.5	−0.73	0.55	0.63	−0.13	0.39	0.22	1.52	−0.16	0.46	0.47	−0.30	−0.11	−0.37
4.5–5.1	−0.70	0.28	0.70	−0.45	0.30	0.51	0.91	−0.08	0.27	0.59	0.04	−0.22	−0.14
5.1–5.6	−0.50	0.11	0.15	−0.11	0.14	0.16	0.99	0.41	0.21	0.59	−0.14	0.21	−0.46
5.6–6.0	−0.71	0.65	0.30	−0.28	0.25	0.14	0.76	0.05	0.11	0.58	−0.06	−0.16	−0.34
	L–L	K–A	K–R	K–N	K–D	K–C	K–Q	K–E	K–G	K–H	K–I	K–L	K–K
0.0–4.5	−0.36	0.10	0.53	0.64	−0.14	0.03	0.83	−1.06	−0.02	−0.58	−0.05	0.46	−0.40
4.5–5.1	−0.15	−0.04	0.40	0.61	−0.02	0.38	0.59	−0.76	−0.19	−0.66	0.11	0.45	−0.23
5.1–5.6	−0.37	−0.10	0.30	0.38	0.07	0.49	0.60	−0.43	0.22	−0.97	0.26	0.31	−0.35
5.6–6.0	−0.20	−0.10	0.15	0.47	−0.13	0.82	0.48	−0.36	−0.41	−1.13	0.28	0.43	−0.13
	M–A	M–R	M–N	M–D	M–C	M–Q	M–E	M–G	M–H	M–I	M–L	M–K	M–M
0.0–4.5	−0.65	−0.07	0.08	0.61	0.38	1.43	−0.26	−0.19	0.94	−0.93	−0.44	0.20	−0.78
4.5–5.1	−0.41	−0.04	0.52	0.97	0.88	1.88	−0.59	−0.05	0.36	−0.43	−0.31	0.27	−0.90
5.1–5.6	−0.19	−0.09	0.19	0.51	0.94	1.40	−0.41	0.49	0.23	0.31	−0.80	0.10	−0.99
5.6–6.0	0.07	−0.12	0.31	0.88	0.75	0.60	−0.41	0.08	0.76	0.23	−0.21	0.05	−1.06
	F–A	F–R	F–N	F–D	F–C	F–Q	F–E	F–G	F–H	F–I	F–L	F–K	F–M
0.0–4.5	−0.82	0.86	0.64	0.89	−0.28	0.06	0.44	0.58	−0.02	−0.30	−0.48	0.19	−1.27
4.5–5.1	−0.81	0.93	0.62	0.60	−0.36	−0.07	0.26	0.44	−0.02	−0.26	−0.53	0.16	−0.52
5.1–5.6	−0.67	0.89	0.62	0.74	−0.23	0.03	0.22	0.25	0.06	−0.19	−0.43	0.11	−0.73
5.6–6.0	−0.53	0.84	0.59	0.62	−0.16	0.12	0.23	0.14	−0.04	−0.30	−0.31	0.16	−0.70
	F–F	P–A	P–R	P–N	P–D	P–C	P–Q	P–E	P–G	P–H	P–I	P–L	P–K
0.0–4.5	−0.29	−0.81	0.13	−0.20	0.17	0.21	0.65	0.29	0.34	0.32	1.33	−0.07	0.08
4.5–5.1	−0.28	−0.44	0.34	0.08	0.22	0.60	0.79	−0.17	0.59	0.52	0.39	0.05	0.32
5.1–5.6	−0.35	−0.50	0.30	−0.32	0.45	0.94	0.53	0.01	0.17	0.83	0.49	−0.08	0.21
5.6–6.0	−0.31	−0.19	0.24	−0.24	0.48	0.52	0.54	−0.31	−0.17	0.87	0.36	0.11	0.53
	P–M	P–F	P–P	S–A	S–R	S–N	S–D	S–C	S–Q	S–E	S–G	S–H	S–I
0.0–4.5	−0.07	−0.30	−1.29	0.35	0.37	−0.06	0.29	−1.19	−0.47	−0.76	0.54	−0.34	0.18
4.5–5.1	−2.14	0.11	−1.24	0.48	0.13	−0.09	0.21	−0.59	−0.55	−0.25	0.40	−0.09	0.28
5.1–5.6	−1.31	−0.02	−0.97	0.29	0.23	−0.22	0.30	−0.50	−0.73	−0.27	0.44	−0.08	−0.09
5.6–6.0	−1.82	0.06	−1.21	0.12	0.20	−0.23	0.23	−0.31	−0.36	−0.28	0.36	0.04	0.16
	S–L	S–K	S–M	S–F	S–P	S–S	T–A	T–R	T–N	T–D	T–C	T–Q	T–E
0.0–4.5	−0.05	0.43	0.71	−0.30	0.32	−0.09	−1.09	0.08	−0.29	−0.35	1.21	−0.08	−0.29
4.5–5.1	−0.01	0.30	0.10	−0.29	0.32	0.00	−0.57	0.11	−0.41	−0.34	0.96	−0.05	−0.13
5.1–5.6	−0.02	0.29	0.06	−0.14	0.09	0.24	−0.42	0.08	−0.33	−0.34	0.49	0.05	−0.09
5.6–6.0	0.02	0.08	0.31	−0.25	0.17	−0.12	−0.64	−0.19	−0.36	−0.27	0.51	−0.06	0.02
	T–G	T–H	T–I	T–L	T–K	T–M	T–F	T–P	T–S	T–T	W–A	W–R	W–N
0.0–4.5	0.31	−0.69	0.05	−0.20	−0.21	0.59	0.31	0.40	−0.44	0.06	0.53	−0.64	0.53
4.5–5.1	−0.03	−0.45	−0.29	−0.25	−0.28	0.88	0.33	0.14	−0.30	0.18	1.14	−0.71	0.51
5.1–5.6	0.00	−0.39	−0.52	0.02	−0.39	0.72	0.20	0.22	−0.15	0.20	0.59	−0.43	0.28
5.6–6.0	0.09	−0.48	−0.47	−0.12	−0.13	0.80	0.34	0.28	0.17	0.45	0.27	−0.60	0.45

TABLE VI. (Continued)

	W–D	W–C	W–Q	W–E	W–G	W–H	W–I	W–L	W–K	W–M	W–F	W–P	W–S
0.0–4.5	−0.73	2.36	−0.08	0.74	0.32	1.25	−0.90	0.02	−0.66	−0.49	0.38	−1.07	−0.05
4.5–5.1	−0.64	1.50	−0.16	0.72	0.47	0.74	−0.80	−0.01	−0.78	−0.72	0.15	−0.74	−0.19
5.1–5.6	−0.52	1.55	0.00	0.50	0.08	0.60	−0.91	0.01	−0.53	−0.43	0.03	−0.77	−0.04
5.6–6.0	−0.53	1.23	0.03	0.53	0.26	0.64	−0.34	0.01	−0.46	−0.74	−0.16	−0.47	−0.16
	W–T	W–W	Y–A	Y–R	Y–N	Y–D	Y–C	Y–Q	Y–E	Y–G	Y–H	Y–I	Y–L
0.0–4.5	0.17	0.29	−0.02	−0.55	−0.69	−0.81	4.92	−0.68	−0.08	0.12	−1.00	−0.08	−0.09
4.5–5.1	0.38	0.61	−0.05	−0.66	−0.64	−0.72	5.86	−0.68	−0.02	0.21	−0.57	−0.32	−0.36
5.1–5.6	0.49	0.59	−0.23	−0.47	−0.53	−0.71	4.88	−0.56	0.08	0.17	−0.36	−0.30	−0.12
5.6–6.0	0.20	0.47	0.01	−0.36	−0.16	−0.67	3.01	−0.35	0.11	0.11	−0.16	−0.20	−0.13
	Y–K	Y–M	Y–F	Y–P	Y–S	Y–T	Y–W	Y–Y	V–A	V–R	V–N	V–D	V–C
0.0–4.5	−0.73	0.18	−0.22	0.16	0.09	0.75	−1.15	0.02	−0.12	0.12	0.41	0.01	−0.28
4.5–5.1	−0.70	0.10	−0.34	0.03	−0.10	0.47	−1.26	−0.27	−0.18	0.05	0.38	0.38	−0.07
5.1–5.6	−0.49	0.20	−0.32	−0.10	0.12	0.32	−0.95	−0.34	−0.08	−0.02	0.47	0.36	−0.22
5.6–6.0	−0.57	0.50	−0.28	−0.02	−0.22	0.16	−1.05	0.01	0.13	0.09	0.38	0.33	−0.25
	V–Q	V–E	V–G	V–H	V–I	V–L	V–K	V–M	V–F	V–P	V–S	V–T	V–W
0.0–4.5	−0.21	0.97	0.38	0.23	0.09	−0.83	0.26	−0.08	−0.27	−0.70	0.47	−0.08	−0.66
4.5–5.1	−0.01	0.70	−0.05	−0.28	−0.04	−0.66	0.00	0.10	−0.19	−0.20	0.18	−0.10	−0.22
5.1–5.6	0.11	0.60	−0.11	0.22	0.06	−0.61	−0.15	−0.16	−0.19	−0.38	0.13	0.04	−0.11
5.6–6.0	0.11	0.42	−0.32	−0.27	−0.26	−0.68	−0.05	−0.28	−0.18	−0.02	0.09	−0.09	0.40
	V–Y	V–V											
0.0–4.5	0.11	0.07											
4.5–5.1	0.35	−0.13											
5.1–5.6	−0.01	0.01											
5.6–6.0	0.46	−0.03											

*Distance intervals are indicated at the left edge of this table, and correspond to the ranges given in Table V. Mean = 0.00, standard deviation = 0.65.

W–K interaction is the only one for which the chemical origin is not easily postulated, and therefore it is possibly the most interesting.

DISCUSSION

Space Group and Coordination Numbers

A relatively few space groups make up the majority of protein crystals. Of the 65 enantiomorphic space groups, Padmaja and colleagues[34] found 29 different space groups in a dataset of 209 protein crystals. Only 26 space groups were found in the present study with a dataset of 144 proteins.

The distribution of coordination numbers found in Table II may be compared to that found by Janin and Rodier (see Fig. 2 of reference 12). Their results were based on buried surface area, as opposed to the approach of counting atom–atom interactions. Nevertheless, the results obtained were similar.

Islam and Weaver[10] found in a dataset of 58 proteins that the number of neighbors varied from 6 to 16. The distance between proteins at which they considered two proteins to be neighbors was not stated. Use of a somewhat longer distance than 5.0 Å might explain the trend to higher number of neighbors, compared to Table II and to the results of Janin and Rodier.[12] Islam and Weaver[10] also found 10 proteins in the 58 crystal structures (17%) that had an odd number of neighbors, a similar fraction to that found in the present survey (16%).

A number of questions remain unanswered as to why certain space groups are preferred in protein crystals. A correspondence between molecular symmetry and space group symmetry elements has been postulated.[35] Wukovitz and Yeates[36] have made a persuasive case for the idea that the more degrees of freedom offered by a particular space group, the more favored that space group is for protein crystallization. P $2_12_12_1$ is the most common space group, and the one that offers the most degrees of freedom. Janin and Rodier[12] have shown that crystal interfaces that incorporate a twofold symmetry on average produce larger, and probably more stable, interfaces than those that do not. Since twofold rotations are not intrinsic to P $2_12_12_1$, other space groups consistent with twofold symmetry might be favored for proteins that form this kind of stable symmetric interaction, even if they allow for fewer degrees of freedom. These two factors partially explain the distribution of space groups observed. Further insight may come from future analysis of contacts at crystal interfaces.

Microscopic and Macroscopic Properties of Crystal and Oligomer Interfaces

We have examined six different properties of oligomer and crystal interfaces: the number of atom–atom contacts, the patch size distribution, the secondary structure involved, the types of residues or atoms

TABLE VII. Pairwise Potentials for Crystal Contacts*

	A–A	R–A	R–R	N–A	N–R	N–N	D–A	D–R	D–N	D–D	C–A	C–R	C–N
0.0–4.5	−0.01	−0.24	0.18	0.02	0.21	−0.48	−0.20	−0.34	−0.16	0.06	−0.25	−0.39	0.03
4.5–5.1	−0.03	−0.14	0.19	0.03	−0.03	−0.29	−0.22	−0.17	−0.14	0.05	0.07	−0.13	0.57
5.1–5.6	−0.22	−0.09	0.14	−0.02	−0.04	−0.32	−0.13	−0.15	−0.01	−0.05	0.10	−0.16	0.34
5.6–6.0	−0.16	−0.14	0.23	0.03	0.15	−0.24	−0.06	−0.11	0.00	−0.06	0.14	−0.67	0.08

	C–D	C–C	Q–A	Q–R	Q–N	Q–D	Q–C	Q–Q	E–A	E–R	E–N	E–D	E–C
0.0–4.5	0.50	−1.47	0.17	−0.12	0.08	−0.21	0.58	0.05	0.32	−0.31	−0.31	−0.14	−0.06
4.5–5.1	−0.17	−0.66	0.18	−0.20	−0.22	−0.08	0.46	0.01	0.29	−0.09	−0.18	0.01	−0.25
5.1–5.6	−0.26	−0.62	0.19	−0.11	−0.14	−0.18	0.40	−0.14	0.10	−0.09	−0.19	−0.09	−0.02
5.6–6.0	−0.32	−0.40	0.29	−0.05	−0.07	−0.15	0.09	−0.09	0.10	−0.16	−0.11	−0.01	−0.19

	E–Q	E–E	G–A	G–R	G–N	G–D	G–C	G–Q	G–E	G–G	H–A	H–R	H–N
0.0–4.5	0.17	−0.26	0.23	−0.01	−0.23	−0.02	−0.05	−0.01	0.16	−0.03	0.42	−0.30	0.21
4.5–5.1	0.01	−0.17	0.10	−0.02	−0.07	0.01	0.13	0.02	0.17	−0.15	0.48	−0.23	0.14
5.1–5.6	0.05	−0.17	0.20	−0.04	−0.15	0.00	0.01	0.00	0.14	−0.08	0.38	−0.26	0.19
5.6–6.0	0.02	−0.21	0.08	0.09	−0.01	0.24	−0.11	0.02	−0.01	−0.17	0.36	−0.12	0.16

	H–D	H–C	H–Q	H–E	H–G	H–H	I–A	I–R	I–N	I–D	I–C	I–Q	I–E
0.0–4.5	0.12	−0.57	0.16	−0.37	0.37	−0.81	0.05	0.13	0.02	0.18	−0.28	−0.47	0.64
4.5–5.1	0.09	−0.70	0.29	−0.42	0.09	−1.04	−0.10	−0.13	−0.02	0.28	−0.11	−0.45	0.53
5.1–5.6	0.10	−0.79	0.22	−0.42	0.05	−0.49	0.12	−0.01	0.15	0.23	−0.81	−0.19	0.45
5.6–6.0	0.19	−0.66	0.21	−0.32	0.21	−0.67	−0.11	0.10	−0.09	0.36	−0.54	−0.25	0.37

	I–G	I–H	I–I	L–A	L–R	L–N	L–D	L–C	L–Q	L–E	L–G	L–H	L–I
0.0–4.5	−0.17	−0.28	−0.43	0.08	0.68	0.04	−0.15	−0.04	0.02	0.37	−0.04	0.37	−0.37
4.5–5.1	−0.11	−0.01	−0.54	0.20	0.52	0.11	0.14	−0.25	0.14	0.34	0.08	0.39	−0.30
5.1–5.6	−0.02	−0.09	−0.44	0.16	0.40	0.28	−0.05	−0.33	0.13	0.56	−0.01	0.21	−0.31
5.6–6.0	−0.12	−0.22	−0.37	0.20	0.42	0.37	0.10	−0.09	0.21	0.59	0.08	0.29	−0.21

	L–L	K–A	K–R	K–N	K–D	K–C	K–Q	K–E	K–G	K–H	K–I	K–L	K–K
0.0–4.5	−0.38	−0.05	0.30	−0.38	−0.64	0.96	−0.12	−0.55	−0.10	0.36	0.38	0.49	−0.26
4.5–5.1	−0.43	0.01	0.14	−0.25	−0.42	0.61	0.00	−0.34	−0.15	0.07	0.06	0.39	−0.27
5.1–5.6	−0.18	−0.06	0.08	−0.27	−0.37	0.77	0.03	−0.33	−0.11	0.15	0.08	0.30	−0.28
5.6–6.0	−0.12	0.04	−0.03	−0.20	−0.35	0.51	0.07	−0.28	−0.20	−0.06	0.10	0.36	−0.22

	M–A	M–R	M–N	M–D	M–C	M–Q	M–E	M–G	M–H	M–I	M–L	M–K	M–M
0.0–4.5	−0.06	0.80	0.88	−0.49	−0.19	0.15	0.11	−0.08	−0.45	−0.12	−0.95	0.57	−0.67
4.5–5.1	0.00	0.66	0.59	−1.13	0.15	0.25	0.06	−0.23	−0.24	0.24	−1.45	0.39	−0.15
5.1–5.6	−0.23	0.34	0.00	−0.89	0.75	0.14	0.12	0.18	−0.15	0.03	−1.20	0.40	−0.24
5.6–6.0	0.03	0.43	0.24	−1.87	0.70	0.25	0.45	0.22	0.13	0.42	−3.10	0.37	−0.16

	F–A	F–R	F–N	F–D	F–C	F–Q	F–E	F–G	F–H	F–I	F–L	F–K	F–M
0.0–4.5	0.02	0.07	0.46	0.59	0.36	0.02	0.33	0.23	0.00	0.35	−0.61	0.26	−0.74
4.5–5.1	0.03	0.07	0.30	0.57	−0.17	−0.05	0.17	0.22	0.02	0.28	−0.55	0.41	−0.35
5.1–5.6	−0.10	0.24	0.43	0.50	0.25	−0.01	0.15	0.17	−0.05	0.05	−0.46	0.26	−0.23
5.6–6.0	−0.10	0.21	0.24	0.54	−0.25	0.13	0.18	0.16	0.02	0.11	−0.34	0.28	0.14

	F–F	P–A	P–R	P–N	P–D	P–C	P–Q	P–E	P–G	P–H	P–I	P–L	P–K
0.0–4.5	−1.29	−0.16	−0.31	0.09	0.62	0.85	−0.28	0.15	−0.30	0.18	0.26	0.26	0.04
4.5–5.1	−0.94	−0.16	−0.32	0.07	0.48	0.43	−0.32	0.03	−0.26	0.09	0.05	0.35	0.19
5.1–5.6	−1.05	−0.23	−0.24	0.04	0.49	0.05	−0.27	0.05	−0.08	0.04	0.06	0.18	0.14
5.6–6.0	−1.15	−0.11	−0.24	0.01	0.36	0.56	−0.16	0.02	−0.27	−0.11	−0.06	0.31	0.08

	P–M	P–F	P–P	S–A	S–R	S–N	S–D	S–C	S–Q	S–E	S–G	S–H	S–I
0.0–4.5	−0.25	−0.17	−0.55	−0.19	0.11	−0.08	−0.39	0.03	0.12	−0.10	0.00	0.12	0.01
4.5–5.1	0.14	−0.26	−0.29	−0.17	0.12	−0.08	−0.23	0.00	0.10	−0.04	0.06	0.07	−0.02
5.1–5.6	0.30	−0.28	−0.22	0.08	0.11	−0.06	−0.22	0.02	−0.07	0.11	0.05	0.04	0.10
5.6–6.0	0.53	−0.29	−0.25	0.00	0.08	−0.14	−0.15	0.11	−0.06	0.02	0.18	−0.08	0.13

	S–L	S–K	S–M	S–F	S–P	S–S	T–A	T–R	T–N	T–D	T–C	T–Q	T–E
0.0–4.5	−0.06	−0.19	0.40	0.02	−0.12	−0.47	−0.15	−0.08	−0.30	−0.26	−0.22	−0.05	−0.13
4.5–5.1	0.06	−0.14	0.20	0.00	−0.07	−0.46	−0.45	0.05	−0.26	−0.11	−0.06	0.04	−0.07
5.1–5.6	−0.02	−0.09	0.27	0.05	−0.01	−0.45	−0.34	−0.03	−0.14	−0.19	0.23	0.15	−0.03
5.6–6.0	0.02	−0.03	0.27	0.07	−0.12	−0.46	−0.36	−0.04	−0.20	0.03	0.22	−0.07	−0.05

	T–G	T–H	T–I	T–L	T–K	T–M	T–F	T–P	T–S	T–T	W–A	W–R	W–N
0.0–4.5	0.18	0.39	−0.01	0.28	−0.23	−0.53	0.07	0.23	0.27	−0.28	0.41	−0.80	−0.28
4.5–5.1	0.10	0.45	−0.11	0.22	−0.06	−0.31	0.06	0.19	0.12	−0.13	0.59	−0.80	−0.38
5.1–5.6	−0.06	0.39	0.20	0.15	−0.23	−0.14	−0.06	0.22	0.04	−0.22	0.49	−0.54	−0.34
5.6–6.0	−0.13	0.44	0.03	0.11	−0.15	0.20	0.13	0.01	−0.02	−0.18	0.27	−0.55	−0.31

TABLE VII. (Continued)

	W–D	W–C	W–Q	W–E	W–G	W–H	W–I	W–L	W–K	W–M	W–F	W–P	W–S
0.0–4.5	1.02	−0.35	0.21	−0.49	−0.05	−0.13	0.17	−0.12	−0.80	1.99	−0.10	−0.36	0.13
4.5–5.1	0.73	−0.43	0.20	−0.49	−0.03	0.21	0.37	−0.13	−0.71	1.66	0.07	−0.45	0.29
5.1–5.6	0.97	−0.59	0.44	−0.69	−0.16	0.44	0.16	−0.03	−0.42	1.08	0.01	−0.35	−0.07
5.6–6.0	0.69	−0.14	0.01	−0.52	−0.27	0.47	0.11	0.16	−0.34	1.01	−0.06	−0.19	0.14
	W–T	W–W	Y–A	Y–R	Y–N	Y–D	Y–C	Y–Q	Y–E	Y–G	Y–H	Y–I	Y–L
0.0–4.5	0.63	−0.21	−0.17	0.23	0.18	0.00	0.53	−0.18	−0.14	−0.18	0.03	0.25	−0.04
4.5–5.1	0.29	−0.09	−0.23	0.23	0.02	−0.02	0.45	−0.14	−0.03	−0.04	0.26	0.47	−0.08
5.1–5.6	0.13	−0.15	−0.21	0.20	0.06	−0.04	0.45	−0.20	−0.06	−0.10	0.04	0.36	0.06
5.6–6.0	0.22	0.05	−0.27	0.21	−0.03	0.01	0.61	−0.10	−0.11	−0.10	0.00	0.31	0.20
	Y–K	Y–M	Y–F	Y–P	Y–S	Y–T	Y–W	Y–Y	V–A	V–R	V–N	V–D	V–C
0.0–4.5	0.19	−0.05	−0.27	−0.36	0.06	0.31	−0.08	−0.51	−0.20	0.16	0.10	0.07	−0.05
4.5–5.1	0.23	−0.37	−0.06	−0.15	0.00	0.23	−0.21	−0.89	−0.37	0.25	0.19	0.28	−0.02
5.1–5.6	0.09	−0.16	−0.04	−0.05	−0.07	0.26	−0.02	−0.85	−0.11	0.22	0.21	0.36	0.21
5.6–6.0	0.13	0.15	−0.09	−0.07	−0.11	0.18	−0.23	−0.88	−0.17	0.10	0.16	0.42	0.36
	V–Q	V–E	V–G	V–H	V–I	V–L	V–K	V–M	V–F	V–P	V–S	V–T	V–W
0.0–4.5	−0.18	0.52	0.09	0.06	−0.33	0.06	−0.10	−0.13	0.37	0.23	0.41	−0.13	−0.67
4.5–5.1	−0.12	0.37	0.07	−0.10	−0.29	0.04	−0.12	0.13	0.13	0.25	0.26	−0.18	−0.62
5.1–5.6	−0.29	0.27	0.02	−0.06	−0.17	0.00	−0.03	−0.12	0.20	0.17	0.22	−0.29	−0.33
5.6–6.0	−0.25	0.19	0.09	−0.23	−0.08	0.10	−0.02	−0.17	0.09	0.09	0.21	−0.26	−0.44
	V–Y	V–V											
0.0–4.5	0.26	−0.64											
4.5–5.1	0.36	−0.60											
5.1–5.6	0.34	−0.86											
5.6–6.0	0.26	−0.51											

*Distance intervals are indicated at the left edge of this table, and correspond to the ranges given in Table V. Mean = 0.00, standard deviation = 0.35.

making contact, nonpairwise contact potentials, and pairwise contact potentials. What can be said regarding the relevance of each property to the specificity and energetics of interface formation?

Atom–atom contacts

The total numbers of atom–atom contacts are not very different between the oligomer and crystal interfaces, and this property does not distinguish the two types of association, although the pairwise interactions between any two molecules in a crystal are usually less numerous than the interactions at a given oligomer interface (Fig. 3).

A similar conclusion might be made from accessible surface areas. For a sample of 23 oligomeric proteins, the interface area per subunit was found to range from 670 to 5540 Å^2, except for one unusual protein, catalase, with 10,570 Å^2 of buried surface on each subunit.[13] For 152 crystal structures,[12] the total area buried per monomer protein within the lattice was found to range from 1100 to 4400 Å^2. Except for catalase, given only the surface area buried by either oligomerization or crystallization, one could not identify with certainty the type of association being measured.

The simplest model for evaluating the energetics of these two types of interfaces might be simply the number of atom–atom contacts or total buried surface area. However, if every contact contributed equally to the interaction energy, then placing a protein in a crystal lattice would be roughly the same as placing a subunit in association with another subunit, and docking two protein subunits would be purely a matter of finding complementary-shaped surfaces. Clearly, this is not so, since lattice forces are much smaller than those involved in oligomer formation[20,37,38] and docking protein molecules is not that simple (discussion below).

Patches

A hypothesis worth considering is that groups of interactions may be more favorable than isolated contacts. This hypothesis could be elaborated to say that dehydration and association of protein surfaces are cooperative processes for adjacent atoms in a patch, in analogy to the arguments for the contribution of cooperativity in protein folding.[39] Thus, patch size may be more than a rough measurement of surface shape complementarity: it may delimit the size of a putative cooperative unit.

Large patches might also allow for a particular spatial organization of residues, such as the arrangement of polar and charged residues surrounding a core of hydrophobic contacts that has been previously suggested.[16] Such an arrangement is another possible factor contributing to stable protein–protein recognition and binding.

TABLE VIII. Comparison of the Pairwise Potentials for Formation of Oligomer and Crystal Contacts*

	Oligomer contacts											
Crystal contacts	Very disfavored	Disfavored		Slightly disfavored			Slightly favored			Favored		Very favored
Very disfavored				*K–C*	*M–N*		*M–R*	*W–M*		*W–D*		
Disfavored	Y–C	H–A F–D	S–M V–E	I–E L–R L–E K–L	M–K P–D P–C F–N	W–A W–T V–S	*L–H* *K-H*	*K–I* *F–C*	*F–I* *V–F*	*C–D* *Q–C*	*T–H*	
Slightly disfavored	*M–Q* *P–I*	G–E F–R	Y–T	N–R Q–A Q–N H–Q H–G I–R I–D L–N L–Q K–R F–E	F–G F–K F–Q P–E P–H P–K S–R S–G S–I T–G	T–I T–F T–P Y–S V–R V–N V–D V–G V–H V–Y	N–A C–N Q–Q E–A E–Q H–N H–D I–A	I–N L–A M–E P–N P–L S–Q S–H S–F	T–L T–S W–Q W–S Y–R Y–I V–P	*R–R* *D–D* *G–A* *F–A* *S–C* *W–I*	*Y–N* *Y–D* *Y–H* *Y–K* *V–L*	
Slightly favored	*C–A* *E–D* *L–D*	*H–R* *K–Q*	*T–C* *W–H*	R–A D–A D–N E–N G–R I–C I–H K–A M–C	P–Q P–G S–A S–K S–P T–R W–N W–G	W–L W–F W–W Y–A Y–G Y–M Y–P V–K	A–A Q–R Q–D E–R E–E G–N G–D G–Q G–G L–C L–G K–G	K–K M–A M–G F–H P–M P–F S–N S–L T–N T–D T–Q	T–E T–K T–T Y–E Y–L Y–F V–A V–C V–Q V–M V–T	D–R M–I P–A S–E T–A	Y–W I–G Y–Q G–C	*E–C*
Favored	*W–C*	*M–H*	*W–E*	*I–I* *P–R* *S–D*	*K–N* *M–D* *T–M*	*Y–Y* *V–I* *V–V*	H–C H–E I–Q	L–I L–L K–D	F–L P–P S–S	N–N C–R K–E	M–M W–P V–W	
Very favored							*M–L*	*F–F*	*W–R*	H–H F–M	W–K	C–C

*Pairwise interactions in the first distance interval are categorized as slightly disfavored, disfavored or very disfavored by whether they are 0–1, >1–2 or >2–3 standard deviations greater than the mean, respectively; and similarly for favorable interactions having negative potentials. Values in italics are two or more categories apart in the two types of contacts compared.

Small patches of atoms in contact occur in oligomers approximately as frequently as in crystal associations (Fig. 6) . Whether these isolated contacts make a more or less significant contribution to the energetics of the interface than contacts in larger patches would be interesting to know, and is approachable by site-directed mutagenesis.

Two measures of shape complementarity related to patch size have been devised: a parameter that directly correlates surface shapes of molecules at an interface,[40] and measurements of gap size at interfaces.[16,41] The former method has shown that oligomer interfaces are equally complementary as protease–inhibitor interfaces, and significantly more complementary than are antibody–antigen interfaces. Oligomer interfaces scored between 0.70 and 0.74 where a score of 1.0 is perfectly complementary, and 0.0 indicates two surfaces whose shapes are uncorrelated. The latter methods have shown that, on average, 10.7% of subunit interface surface area is composed of cavities that interrupt surface complementarity[41] and place the three types of interfaces in the same order[16] as did the former method.[40] The conclusion to be drawn from this information is that surface complementarity is high but not perfect.

Although simple shape complementarity contains almost enough information to correctly dock proteins in the case of proteins that form complexes,[42,43] most successful protein docking algorithms incorporate information on the chemistry of some or all types of nonbonded interactions,[42,43–51] suggesting that shape complementarity is required but is insufficient without chemical complementarity.

Janin and Rodier[12] have shown that the buried surface areas that protein–protein interfaces produce have a similar size distribution as that achieved

TABLE IX. Comparison of the Pairwise Potentials for Formation of Oligomer and Internal Contacts*

	Oligomer contacts											
Internal contacts	Very disfavored	Disfavored		Slightly disfavored			Slightly favored			Favored		Very favored
Very disfavored				*K–R*	*K–C*	*W–W*	*E–E*			*W–D* *W–K*	*Y–D*	
Disfavored	P–I	Y–T		I–E L–R L–N L–Q	L–E M–N M–K F–G	W–N Y–S V–D	*I–N* *K–K* *M–R*	*F–N* *S–L* *S–F*	*W–Q*	*D–D* *Q–C*	*H–H* *V–P*	
Slightly disfavored	*C–A* *L–D* *M–Q* *W–C* *Y–C*	H–R M–H F–D	S–M T–C V–E	R–A D–A H–Q H–G I–D K–A K–N F–Q F–E F–K M–D P–D	P–G P–H P–K S–A S–I S–P T–R T–I T–M T–F T–P W–A	W–G W–T Y–G Y–P Y–Y V–N V–G V–H V–K V–S V–Y	A–A N–A C–N Q–R E–A H–E I–Q	L–H K–G M–E M–G P–N P–P S–Q	S–H T–Q T–L W–S Y–R V–T	*C–R* *G–A* *I–G* *F–A*	*P–A* *T–A* *T–H*	*E–C*
Slightly favored	*E–D*	*G–E* *H–A* *K–Q*	*F–R* *W–E* *W–H*	N–R Q–A Q–N E–N G–R I–H I–I I–R	K–L M–C P–R P–C P–E S–R S–G	S–K T–G W–F Y–A Y–M V–R V–I	Q–D G–N G–Q G–G H–N H–D H–C L–A L–C L–G L–I K–H	K–I M–L F–C F–H F–F L–L P–L P–M P–F S–S T–N T–E	T–K T–T W–R Y–E Y–I Y–L V–C V–Q V–M V–F	R–R N–N C–D G–C M–I M–M S–C	S–E W–P Y–Q Y–H Y–K Y–W V–W	
Favored				*D–N* *I–C*	*P–Q* *S–D*	*W–L* *V–V*	Q–Q G–D I–A M–A	F–I F–L S–N T–D	T–S Y–F V–A	D–R F–M W–I	Y–N V–L	C–C
Very favored							*E–R* *E–Q*	*K–D*	*W–M*	K–E	W–K	

*Pairwise interactions in the first distance interval are categorized as slightly disfavored, disfavored or very disfavored by whether they are 0–1, >1–2, or >2–3 standard deviations greater than the mean, respectively; and similarly for favorable interactions having negative potentials. Values in italics are two or more categories apart in the two types of contacts compared. Internal contact information is from Bryant and Lawrence.[32]

by random pairing of two proteins. One might argue that crystal interfaces should form to maximize buried surface area if that would result in an assembly with lower free energy. However, their results indicate little or no such tendency.[12] The small average patch size found for protein crystal contacts is consistent with this view of crystal interface formation. From this information, one must conclude either that these small patches are the most stable way of organizing contacts formed between two random (nonoptimized) surfaces, that only small patches *can* form given the shape of globular proteins, or that some reason exists that extensive interfaces should hinder the growth of protein crystals.

Secondary structure

The type of secondary structure at the two interfaces does not seem to be very different (Table III), and probably is not closely related to the association properties of molecules in the two datasets. Previous results on secondary structure at oligomer interfaces are also presented in Table III. Some variability occurs among these numbers for oligomer interfaces. Three factors probably play a role in this variation. In the two previous studies cited, each residue was counted once if it appeared at an oligomer interface in their datasets. First, in this study, the number of atom–atom contacts was counted, so that the relative importance of each secondary structure is weighted by the number of contacts contributed at the interface. Second, Miller[15] has shown that one can classify oligomer interfaces into four motifs, two of which contain a contribution from sheets. The distribution of these four motifs might be different in the three datasets. Finally, the other two studies identified secondary structure by the method of Kabsch and Sander,[52] which involves analysis of backbone conformation and hydrogen bonding patterns. The present study employed only local backbone conformation, using the criteria of Liebman and colleagues.[26]

Secondary structural elements involved in lattice contacts have not previously been documented for a large set of proteins. A study of four different forms of trypanosomal triosephosphate isomerase revealed that 77% of the residues that were involved in crystal contacts were found in "loop" regions,[6] which is probably equivalent to combined coil and turn regions by the definition of Kabsch and Sander.[52] This percentage is clearly much higher than average (Table III).

Amino acid types

Crystal contacts are characterized by the involvement of amino acids with polar side chains, and oligomer contacts are characterized by more frequent contacts between amino acids with nonpolar side chains. This result is as might be expected from comparisons of subunit interfaces with the overall protein surface.[13,17,19]

Table VI of Janin and colleagues[13] shows the percent composition of subunit interfaces, calculated from residue surface area at 23 oligomer interfaces. These results are similar, but far from identical, to those in Figure 8, which are based on number of atom–atom contacts (correlation = 0.75, not shown). That these two metrics measure related properties of the oligomer interface might be expected, since any decrease in accessible surface area is ultimately attributable to atom–atom contacts. The difference that does occur is difficult to account for.

Remarkably, arginine was found in the top three residues in forming contacts at both types of interfaces, suggesting that this residue might be especially well suited for formation of intermolecular interactions. Arginine is especially adaptable to the formation of multiple salt bridge interactions, which may explain its structural role at oligomer and crystal interfaces.[53]

It is worth noting that, while on average oligomer interfaces are more hydrophobic than protein surfaces and less hydrophobic than protein interiors,[13,14,17,19] examples exist that span this whole range of hydrophobicity. Korn and Burnett[19] observed that some oligomer interfaces are quite hydrophilic and that such interfaces may be found more frequently at flexible interfaces that play a role in allosteric interactions. Average hydrophobicity or amino acid content could be misleading if a subunit interface may be constructed equally well with predominantly polar or nonpolar amino acids.

Nonpairwise preferences

Figure 8 and Table V contain information that may be related to interface properties. In regard to crystal contacts, of the five most important residues in terms of number of contacts (K, E, R, S and D), only R and S are selected for at crystal interfaces (in absence of a particularly favorable pairwise interaction). In regard to oligomer contacts, of the five most important residues in terms of number of contacts (F, R, L, Y and T); F, L and Y are favored in nonpairwise manner, while R and T are near neutral with regard to nonpairwise contact potentials. More numerous favored contacts at oligomer interfaces is likely related to their high affinity. Specificity of oligomer contacts must somehow reside in the pairwise preferences.

Hydrophobic interactions are very important for oligomer interfaces,[13,14,17,19] but the types of nonpolar amino acids favored at these interfaces are restricted to the aromatic amino acids and methionine (Table V). This is very unlike the case with internal contacts, where aliphatic side chains are equally favored with aromatic side chains. What advantage might be obtained through the use of these particular residues is unclear. It must be a substantial advantage since this preference for a subset of hydrophobic residues is strong compared to all of the other nonpairwise interactions. Also related to this observation is the fact that phenylalanine is involved in more contacts at oligomer interfaces than any other amino acid (Fig. 8).

Nonpairwise crystal contact potentials clearly disfavor or are neutral toward the nonpolar side chains, except perhaps for tyrosine. So, even when hydrophobic residues have some surface exposure, crystallization does not favor their inclusion at crystal interfaces.

The apparent selection against nonpolar residues at crystal interfaces may reflect a selection of contact types during crystallization. In this view, because association due to hydrophobic interactions are not as directional as hydrogen bonds and other polar interactions, hydrophobic interactions lead to disordered precipitates, and not to crystals. Proteins which associate through this contact type would not show up very frequently in the sample of crystal contacts observed. This explanation is consistent with simulations of protein crystallization and aggregation that suggest that very strong interactions between proteins are unfavorable toward crystallization and lead to less ordered structures.[54,55]

Clearly, both the balance of forces involved (hydrophobic versus polar interactions) and the complementarity of the interfaces (large vs. small patches) appear distinct when comparing oligomer to crystal interfaces. A connection between these two observations can be suggested and provides an alternative explanation for the nonpairwise preferences at crystal interfaces. In this view, the lack of large continuous surface patches at crystal interfaces may make dehydration and hydrophobic association less cooperative and therefore less energetically favorable than might otherwise be. In simple terms, formation of many small "oil droplets" through protein–protein association may not be as favorable as formation of one large one using only adjacent residues. If so, polar interactions might then be competitive for formation of crystal interfaces.

Pairwise preferences

Pairwise preferences are difficult to discuss because they are so numerous. Most pairwise preferences are similar between crystal and oligomer contacts (Table VIII).

At oligomer interfaces, favored interactions between pairs of charged residues include K–E and D–R. Surprisingly, R–R and D–D pairs are also favored at oligomer interfaces. Meanwhile, E–D pairs are very disfavored. This result seems inconsistent, since the difference between an E–D pair and a D–D pair would seem to be small. One additional piece of information that may be relevant is that E–R and E–E contacts are slightly favored. This result may be related to the importance of complex salt bridges at oligomer interfaces.[53] Complex salt bridges are those that involve more than two residues. Simple two-residue salt bridges predominate within intramolecular interactions. Complex salt bridges are more common at oligomer interfaces than simple salt bridges.[53] Perhaps the statistics obtained herein are indicative of the types of combinations of residues that may occur in complex salt bridges. One possible hypothesis is that two glutamic acids or two aspartic acids from one subunit frequently interact with an arginine from another subunit. However, a mixed pair of acid residues, because of their difference in length, might not be able to adopt a geometry that produces a complex salt bridge interaction. If nothing else, this result suggests that further study of charge–charge interactions at oligomer interfaces could be informative.

At crystal interfaces, certain pairings of nonpolar residues, M–M, M–L, L–L, L–I, F–L, F–F, V–W and W–P are favored (Tables VII and VIII). Looking at that list, one might be tempted to regard these contacts as evidence of a hydrophobic interaction. Yet Table V clearly indicates that generalized nonpairwise interactions between nonpolar residues (i.e., hydrophobic interactions) are disfavored. Also, a longer list of contacts between nonpolar residues, Y–C, Y–M, V–Y, Y–A, Y–P, Y–Y, P–C, P–I, I–I, W–A, W–L, W–F, W–W, V–I, and V–V are disfavored. These favored pairings must be attributed to specific van der Waals interactions that make these pairs especially complementary because of their shape, size, or polarizability.

Other Conclusions From Contact Potentials

In the situation where a particular protein cannot be crystallized, the results of Table V suggest one possible strategy for modification of the protein to favor crystallization. Lysine residues, the least favored amino acid at crystal interfaces, could be replaced with either arginine or glutamine, the two most favored residues at crystal interfaces. In absence of any information about pairwise interactions (which otherwise might be obtained from a low resolution structure using poorly diffracting crystals),[56] this strategy seems to be the most likely to succeed.

An overall contact potential [equation (4)] can be obtained for any pairwise contact across a crystal or oligomer interface as the sum of the nonpairwise contact potentials for the two residues (Table V) plus the pairwise potential (Table VI or VII). These overall contact potentials have a variety of possible uses, in docking simulations between proteins, and in the design of crystal or oligomer interfaces. They should have some predictive power regarding the effect of point mutations at interfaces. At very least, they provide a measure of whether a particular contact between residues at an interface is typical of such interactions. They should also provide an estimation of the relative energetics of such contacts.

ACKNOWLEDGMENTS

We thank Todd M. Przybycien for helpful discussion. This work was made possible by a grant from the National Institutes of Health (GM50817) to J.A.B.

REFERENCES

1. Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer Jr., E.F., Brice, M.D., Rogers, J.R., Kennard, O., Shimanouchi, T., Tasumi, M. The Protein Data Bank: A computer-based archival file for macromolecular structures. J. Mol. Biol. 112:535–542, 1977.
2. Abola, E., Bernstein, F.C., Bryant, S.H., Koetzle, T.F., Weng, J. Protein Data Bank. In: Allen, F.H., Bergerhoff, G., Sievers, R. "Crystallographic Databases: Information Content, Software Systems, Scientific Applications." Bonn/Cambridge/Chester: Data Commission of the International Union of Crystallography, 1987:107–132.
3. Crosio, M.P., Rodier, F., Jullien, M. Packing forces in ribonuclease crystals. FEBS Lett. 271:152–156, 1990.
4. Svensson, L.A., Dill, J., Sjölin, L., Wlodawer, A., Toner, M., Bacon, D., Moult, J., Veerapandian, B., Gilliland, G.L. The crystal packing interactions of two different crystal forms of bovine ribonuclease A. J. Crystal Growth 110:119–130, 1991.
5. Crosio, M.P., Janin, J., Jullien, M. Crystal packing in six crystal forms of pancreatic ribonuclease. J. Mol. Biol. 228:243–251, 1992.
6. Radha Kishan, K.V., Zeelen, J.P., Noble, M.E.M., Borchert, T.V., Wierenga, R.K. Comparison of the structures and the crystal contacts of trypanosomal triosephosphate isomerase in four different crystal forms. Protein Sci. 3:779–787, 1994.

7. Zhang, X.J., Wozniak, J.A., Matthews, B.W. Protein flexibility and adaptability seen in 25 crystal forms of T4 lysozyme. J. Mol. Biol. 250:527–552, 1995.
8. Takahashi, T., Endo, S., Nagayama, K. Stabilization of protein crystals by electrostatic interactions as revealed by a numerical approach. J. Mol. Biol. 234:421–432, 1993.
9. Matthews, B.W. Solvent content of protein crystals. J. Mol. Biol. 33:491–497, 1968.
10. Islam, S.A., Weaver, D.L. Molecular interactions in protein crystals: Solvent accessible surface and stability. Proteins 8:1–5, 1990.
11. Frey, M., Genovesio-Taverne, J.C., Fontecilla-Camps, J.C. Molecular packing and morphology of protein crystals. J. Phys. D. Appl. Phys. 24:105–110, 1991.
12. Janin, J., Rodier, F. Protein–protein interactions at crystal contacts. Proteins 23:580–587, 1995.
13. Janin, J., Miller, S., Chothia, C. Surface, subunit interfaces and interior of oligomeric proteins. J. Mol. Biol. 204:155–164, 1988.
14. Argos, P. An investigation of protein subunit and domain interfaces. Protein Eng. 2:101–113, 1988.
15. Miller, S. The structure of interfaces between subunits of dimeric and tetrameric proteins. Protein Eng. 3:77–83, 1989.
16. Jones, S., Thornton, J.M. Protein–protein interactions: A review of protein dimer structures. Prog. Biophys. Mol. Biol. 63:31–65, 1995.
17. Miller, S., Lesk, A.M., Janin, J., Chothia, C. The accessible surface area and stability of oligomeric proteins. Nature 328:834–836, 1987.
18. 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. 6:53–64, 1997.
19. Korn, A.P., Burnett, R.M. Distribution and complementarity of hydropathy in multisubunit proteins. Proteins 9:37–55, 1991.
20. Neet, K.E., Timm, D.E. Conformational stability of dimeric proteins: Quantitative studies by equilibrium denaturation. Protein Sci. 3:2167–2174, 1994.
21. Boberg, J., Salakoski, T., Vihinen, M. Selection of a representative set of structures from the Brookhaven Protein Data Bank. Proteins 14:265–276, 1992.
22. Hobohm, U., Scarf, M., Schneider, R., Sander, C. Selection of representative protein data sets. Protein Sci. 1:409–417, 1992.
23. Hobohm, U., Sander, C. Enlarged representative set of protein structures. Protein Sci. 3:522–524, 1994.
24. Becker, R.A., Chambers, J.M., Wilks, A.R. "The New 'S' Language: A Programming Environment for Data Analysis and Graphics." Pacific Grove, CA: Wadsworth, 1988.
25. Bryant, S.H. PKB: A program system and data base for analysis of protein structure. Proteins 5:233–247, 1989.
26. Liebman, M.N., Venanzi, C.A., Weinstein, H. Structural analysis of carboxypeptidase A and its complexes with inhibitors as a basis for modeling enzyme recognition and specificity. Biopolymers 24:1721–1758, 1985.
27. Morris, A.L., Hutchinson, E.G., MacArthur, M.W., Thornton, J.M. Stereochemical quality of protein structure coordinates. Proteins 12:345–364, 1992.
28. Tronrud, D.E., Ten Eyck, L.F., Matthews, B.W. An efficient general-purpose least-squares refinement program for macromolecular structures. Acta Crystallogr. A 43:489–503, 1987.
29. Lee, B., Richards, F.M. The interpretation of protein structure: Estimation of static accessibility. J. Mol. Biol. 55:379–400, 1971.
30. Shrake, A., Rupley, J.A. Environment and exposure to solvent of protein atoms: Lysozyme and insulin. J. Mol. Biol. 79:351–371, 1973.
31. Bryant, S.H., Lawrence, C.E. The frequency of ion-pair substructures in proteins is quantitatively related to electrostatic potential: A statistical model for non-bonded interactions. Proteins 9:108–119, 1991.
32. Bryant, S.H., Lawrence, C.E. An empirical energy function for threading protein sequence through the folding motif. Proteins 16:92–112, 1993.
33. Fienberg, S.E. "The Analysis of Cross-Classified Categorical Data." Cambridge, MA: MIT Press, 1980.
34. Padmaja, N., Ramakumar, S., Viswamitra, M.A. Space-group frequencies of proteins and of organic compounds with more than one formula unit in the asymmetric unit. Acta Crystallogr. A46:725–730, 1990.
35. Brock, C.P., Dunitz, J.D. Towards a grammar of crystal packing. Chem. Mater. 6:1118–1127, 1994.
36. Wukovitz, S.W., Yeates, T.O. Why protein crystals favour some space-groups over others. Nature Struct. Biol. 2:1062–1067, 1995.
37. McPherson, A. "The Preparation and Analysis of Protein Crystals." Malabar, FL: Robert E. Kreieger Publishing, 1989:174–179.
38. McPherson, A. A brief history of protein crystal growth. J. Crystal Growth 110:1–10, 1991.
39. Creighton, T.E. An empirical approach to protein conformation stability and flexibility. Biopolymers 22:49–58, 1983.
40. Lawrence, M.C., Colman, P.M. Shape complementarity at protein/protein interfaces. J. Mol. Biol. 234:946–950, 1993.
41. Hubbard, S.J., Argos, P. Cavities and packing at protein interfaces. Protein Sci. 3:2194–2206, 1994.
42. Cherfils, J., Duquerroy, S., Janin, J. Protein–protein recognition analyzed by docking simulation. Proteins 11:271–280, 1991.
43. Helmer-Citterich, M., Tramontano, A. PUZZLE: A new method for automated protein docking based on surface shape complementarity. J. Mol. Biol. 235:1021–1031, 1994.
44. Zielenkiewicz, P., Rabczenko, A. Protein–protein recognition: Method for finding complementary surfaces of interacting proteins. J. Theor. Biol. 111:17–30, 1984.
45. Jiang, F., Kim, S.H. "Soft docking": Matching of molecular surface cubes. J. Mol. Biol. 219:79–102, 1991.
46. Bacon, D.J., Moult, J. Docking by least-squares fitting of molecular surface patterns. J. Mol. Biol. 225:849–858, 1992.
47. Meng, E.C., Shoichet, B.K., Kuntz, I.D. Automated docking with grid-based energy evaluation. J. Comput. Chem. 13:505–524, 1992.
48. Shoichet, B.K., Kuntz, I.D. Protein docking and complementarity. J. Mol. Biol. 221:327–346, 1991.
49. Shoichet, B.K., Kuntz, I.D. Matching chemistry and shape in molecular docking. Protein Eng. 6:723–732, 1993.
50. Walls, P.H., Sternberg, M.J.E. New algorithm to model protein recognition based on surface complementarity: Applications to antibody antigen docking. J. Mol. Biol. 228:277–297, 1992.
51. Cherfils, J., Janin, J. Protein docking algorithms: Simulating molecular recognition. Curr. Opin. Struct. Biol. 3:265–269, 1993.
52. Kabsch, W., Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22:2577–2637, 1983.
53. Musafia, B., Buchner, V., Arad, D. Complex salt bridges in proteins: Statistical analysis of structure and function. J. Mol. Biol. 254:761–770, 1995.
54. Patro, S.Y., Przybycien, T.M. Simulations of kinetically irreversible protein aggregate structure. Biophys. J. 66:1–16, 1994.
55. Patro, S.Y., Przybycien, T.M. Simulations of reversible protein aggregate and crystal structure. Biophys. J. 70: 2888–2902, 1996.
56. Lawson, D.M., Artymiuk, P.J., Yewdall, S.J., Smith, J.M.A., Livingstone, J.C., Treffry, A., Luzzago, A., Levi, S., Arosio, P., Cesareni, G., Thomas, C.D., Shaw, W.V., Harrison, P.M. Solving the structure of human H ferritin by genetically engineering intermolecular crystal contacts. Nature 349: 541–544, 1991.