Comprehensive statistical analysis of residues interaction specificity at protein-protein interfaces

ARTICLE *in* PROTEINS STRUCTURE FUNCTION AND BIOINFORMATICS · JUNE 2007

Impact Factor: 2.63 · DOI: 10.1002/prot.21363 · Source: PubMed

Anastasia Anashkina
Engelhardt Institute of Molecular Biology (EI…

Comprehensive Statistical Analysis of Residues Interaction Specificity at Protein–Protein Interfaces

Anastasya Anashkina,[1*] **Eugene Kuznetsov,**[2] **Natalia Esipova,**[1] **and Vladimir Tumanyan**[1]

[1]*Laboratory of bioinformatics and system biology, Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow, Russia*

[2]*Laboratory of data analysis, Institute of Control Sciences, Russian Academy of Sciences, Moscow, Russia*

ABSTRACT **We calculated interchain contacts on the atomic level for nonredundant set of 4602 protein-protein interfaces using an unbiased Voronoi-Delaune tessellation method, and made 20×20 residue contact matrixes both for homodimers and heterocomplexes. The area of contacts and the distance distribution for these contacts were calculated on both the residue and the atomic levels. We analyzed residue area distribution and showed the existence of two types of interresidue contacts: stochastic and specific. We also derived formulas describing the distribution of contact area for stochastic and specific interactions in parametric form. Maximum pairing preference index was found for Cys-Cys contacts and for oppositely charged interactions. A significant difference in residue contacts was observed between homodimers and heterocomplexes. Interfaces in homodimers were enriched with contacts between residues of the same type due to the effects of structure symmetry. Proteins 2007;67:1060–1077.** **© 2007 Wiley-Liss, Inc.**

Key words: protein–protein interaction; protein interface; Voronoi tessellation; residue contact preferences

INTRODUCTION

Protein–protein interactions are crucial in most *in vivo* processes: cellular regulation, metabolic pathways, signal transduction, DNA replication initiation, transcription and translation, oligomer and multimolecular complex formation, virus packing, and immune response. Given their central role in most biological processes, protein interactions have been addressed by many research groups.[1–7] Understanding the nature of protein–protein interactions is not straightforward, since such understanding is related to another practically unsolved problem—the problem of protein folding. In both protein–protein interactions and protein-folding there is a delicate balance between the relative contributions of hydrophobic effects and electrostatic interactions, and so a wide variety of motifs are observed at the interface region. We try to formulate and solve the problem of residue contacts statistics in protein–protein recognition and binding sites. Here, we describe an automatic approach for extraction and analysis of such information.

PDB[8] contains information about various types of protein–protein complexes, which can be used to study the nature of interactions in a protein–protein complex. Amino acid residues that take part in protein binding and recognition are called the protein–protein interface. One can find at least three types of contacts in PDB[1]: contacts between different chains in protein–protein complexes[4,9,2]; contacts between monomers in homodimers[10,11,3]; contacts between monomers because of crystallographic symmetry.[3,12] Specific interactions of biological significance occur in all cases, especially in Cases 1 and 2. Protein crystallization also takes place in natural conditions and plays a part in biological processes, so we cannot assert that crystallographic symmetry contacts are nonspecific.

It is generally accepted that interacting proteins have a high degree of complementarity of contacting surfaces.[13] High density atomic packing was observed for structural elements both inside and between interacting proteins.[14,15] It is often assumed that such tight packing provide optimal van der Waals interactions and solution exclusion, maximizing hydrophobic stabilization of the molecule. The role of hydrophobic residues in protein recognition and interaction has been well established many years ago.[16–21]

In some cases domain–domain interfaces may be taken into account instead of protein–protein interfaces, implying that a protein domain is a basic unit of protein structure, function, and evolution.[22] Domain units can fold independently, and correspond to specific biological functions, being a part of large complexes. Analysis of protein–protein and domain-domain interfaces[15] on different levels of their organization has been performed by numerous researchers.[2,3,6,9,20,21,23–27] Using a different

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

Grant sponsor: Molecular and Cellular Biology RAS; Grant sponsor: Russian Ministry of Education and Science; Grant number: 02.434.11.1008; Grant sponsor: Russian Foundation for Basic Research; Grant number: 05-04-49625.

*Correspondence to: Anastasya Anashkina, Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Laboratory of bioinformatics and system biology, 32 Vavilova St., Moscow, Russian Federation 119991. E-mail: anastasya.anashkina@gmail.com

Received 22 May 2006; Revised 15 September 2006; Accepted 29 November 2006

Published online 13 March 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21363

dataset of interfaces from PDB, investigators studied amino acid pairing preferences and conservation of sequence and structure of interfaces.[28–35] As a result, a substantial progress was achieved with respect to our understanding of the physics and the evolution of protein interactions. Here, we try to make some improvements in this area[1]: for our analysis we used the largest interface dataset to date[2]; at first, we detected all interactions on the atomic level, and on the basis of these interactions we determined the contacts between the residues[3]; instead of a distance cut-off, we used the Voronoi–Delaunay method for identifying the contact neighbors[4]; we computed the contact area from Voronoi polyhedra instead of accessible surface area[5]; we introduced rigorous statistical criteria to validate our results.

For the first time, in the context of protein structure, the Voronoi tessellation was used by Richards in 1974[36] to evaluate the volumes of the atoms in a globular protein, defined as the volume of their Voronoi polyhedra. There are some other methods for spatial tessellation of three-dimensional structures reviewed by Poupon.[37] These methods also allow to split the space into zones of influence for each atom.

The Voronoi–Delaunay tessellation has been used successfully in chemistry and structural biology for a long time. By means of this method the packing density of atoms in proteins was analyzed.[38] Average volumes of residues, number of faces per cell, and edges per face were defined in various papers.[38,39] Atom volumes at protein interfaces were also examined using the Voronoi tessellation.[40] Mean volumes of atoms in crystal structure of a large number of inorganic compounds were defined by Christensen and Thomas.[41] Similar findings were described for proteins.[42,43]

METHODS

Voronoi-Delaunay Tessellation

Software has been developed, which unambiguously defines the contact pairs of amino acid residues and the area of the contact between them. It is based on the Voronoi–Delaunay tessellation method. For an arbitrary center of the system of centers $\{A\}$ one can define a region of this space where all points are close to it, rather than another center. This region is called the Voronoi region. In three-dimensional space, the Voronoi region of any center i of a system $\{A\}$ is a convex polyhedron. Voronoi polyhedra for each center from a system $\{A\}$ form a patchwork of polyhedra, which is called the Voronoi tessellation.

The Delaunay method of empty spheres as well as Voronoi faces reveals the same system of points. Each simplex (polyhedron of Delaunay mosaic) corresponds to some vertex of Voronoi mosaic, and vice versa, each vertex of Voronoi polyhedron corresponds with the appropriate Delaunay simplex. These constructions are called dual, and they are topologically equivalent. This duality of Voronoi and Delaunay tessellations is used in the analysis algorithm of the program. Its use greatly improves the algorithm working time. As a rule, main neighbors correspond to the largest faces of polyhedra. Minor neighbors are usually more distant, their faces are small cuts of vertexes or edges on Voronoi polyhedron. Detailed and thorough analysis of the Voronoi–Delaunay tessellation, related algorithms, and its applications can be found elsewhere.[44]

Our program treats protein atoms as a set of mathematical points. To specify the boundary conditions, the proteins are surrounded by virtual water. For each atom, the neighbors are determined by the Delaunay method. Using the Delaunay tessellation, the Voronoi polyhedra are calculated. Note that in this work the neighboring atoms are determined by the Voronoi-Delaunay method, that is, the actual neighbors are determined independently of the interatomic distance. Faces of a Voronoi polyhedron shared by the atoms belonging to different protein chains are classified as a protein–protein interface on the atomic level. According to this mathematical model, two residues are considered to be in contact, and thus to be neighbors if the Voronoi polyhedra of any two atoms share a face of nonzero area. The areas of shared faces are added up to determine the contact area between these two residues. Based on spatial data, we calculated the number C_{ij} of contacts between residues of each type.

Statistical Interpretation

Statistical model of randomly contacting residues

We need a statistical model of residue contacts for evaluation and interpretation of deviations from the stochastic process. The simplest model is a sample with replacement, when amino acid residues are approximated by colored spheres of equal size.

However, this model is rather simple and does not take into account the fact that one residue can form more then one contact. Therefore, we have used the number of contacts n_i instead of the number of residues.

In that case, the probability of contact between them is proportional to their quantity:

$$P_{ij,i\neq j} = \frac{n_i n_j}{N(N-1)}, \tag{1}$$

$$P_{ii} = \frac{n_i(n_i-1)}{N(N-1)}, \tag{2}$$

where $N = \sum_{i=1}^{20} n_i$ is the sum of all contacts formed by these residues. In this case the expected number of casual contacts is calculated as a product NP_{ij}, where N is the total number of contacts.

Two-overlapping-circles model

The purpose of the model is to estimate the distribution of the areas of contact between residues at a protein–protein interface.

Casual contacts. The problem of contact formation can be reduced to the problem of two identical circles of radius r intersecting by chance in a square region with

R being the length of its side. The area of intersection for two circles of the same radius r will be

$$S(L) = 2\left(r^2 \arcsin\left(\sqrt{1-\frac{L^2}{4r^2}}\right) - \frac{Lr\sqrt{1-\frac{L^2}{4r^2}}}{2}\right), \qquad (3)$$

where L is the distance between the centers of these circles.

The coordinates of centers could take values from the interval [r, $R - r$], if the coordinates of square vertexes are (0,0); (R,R); (0,R); and (R,0). The following formula

$$P = \frac{\pi L^2}{2(R-2r)^2} \qquad (4)$$

gives the probability that the distance between the centers of the circles does not exceed $2r$ ($L \in [0, 2r]$), and, therefore, the intersection area is greater than zero.

To find a dependence of probability density $\frac{dP}{dS}$ on intersection area S, we can write the corresponding equation in parametric form, because we cannot express L as a function of S in an explicit form:

$$\frac{dP}{dS}(L) = \frac{dP}{dL}\frac{dL}{dS} = \left(-\frac{\pi L}{(R-2r)^2}\right)\left(-\frac{1}{2\sqrt{r^2 - \frac{L^2}{4}}}\right) = \frac{\pi L}{r(R-2r)^2\sqrt{1-\frac{L^2}{4r^2}}}, \qquad (5)$$

where $S(L)$ is expressed by Eq. (3).

In other words, the distribution of casual contacts contains a lot of small-area contacts (Fig. 1, curve A). As the contact area increases, the number of contacts decreases rapidly. Consequently, the average of the distribution is close to zero.

Specific contacts. It is reasonable to assume that specific contacts have a well-defined nonzero contact area, so, some average specific contact area exists that originates from some specific (physicochemical) interactions of amino acid residues. Let us consider such specific interaction as a tendency to form the maximal contact area. In this case, the centers of circles have a tendency to coincide, and the problem is equivalent to a problem of shooting at a target in the statistical sense. The distribution of distances between points and the center of the target in this case is a normal one. Thus, in the model of specific contacts, the distance between centers of circles follows the formula for normal distribution:

$$f(L) = \frac{1}{\sigma\sqrt{2\pi}}\ \exp^{\frac{(\mathrm{L-a})^2}{2\sigma^2}}. \qquad (6)$$

In a general case, formulas (5) and (6) enter in a sum function with some coefficients that reflect the proportion between casual and specific contacts (Fig. 1). The area under the sum curve is equal to 1. Thus, the contact area distribution can be represented as a composite one, one part of which is formed by casual contacts and the other by specific contacts (responsible for protein recognition and binding). Casual contact distribution reflects the fact that there is a large number of extremely small (nearly zero) contacts, and the number of contacts rapidly decreases with the increase of the contact area. The distribution of specific contacts, on the other hand, has some average nonzero contact area, and it is dome-shaped.

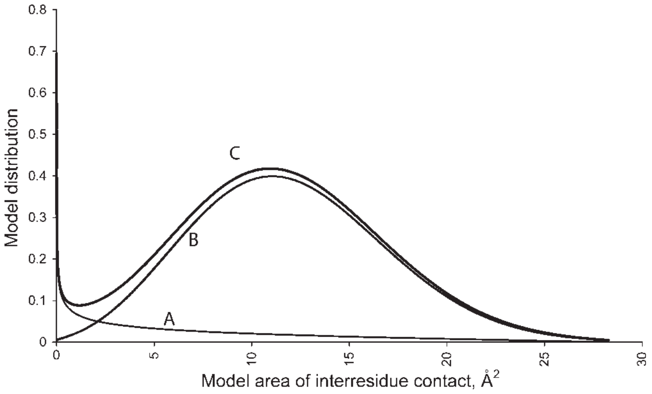

Fig. 1. Plots of Eqs. (3), (5), and (6) modeling the distribution of contact area for casual and specific (responsible for protein recognition and binding) contacts. **A.** Plot of Eqs. (3) and (5) in parametric form. This plot reflects the distribution of areas of stochastic contacts. **B.** Plot of Eqs. (3) and (6) in parametric form. This plot reflects the distribution of areas of specific contacts. **C.** Plot of the sum of Eqs. (3) and (5) and Eqs. (3) and (6). This plot reflects the sum of specific and stochastic distributions. Calculations were performed for parameters from model Eq. (3), (5), and (6) r = 3 Å, R = 20 Å, a = 3, and σ = 1. The distribution of stochastic contacts reflects the fact that there are a lot of minor contacts (practically with zero area). The number of these contacts becomes negligibly small with the enlargement of the contact area. In contrast, the distribution of specific contacts is characterized by significantly nonzero mean contact area, and it is dome-shaped. Thus, the distribution of residue contact areas can be presented as a composite one. Stochastic contacts are responsible for the first part of the distribution, and the other part is formed by specific contacts involved in protein recognition and binding.

Correlation coefficient

To evaluate the significance of deviation in amino acid composition, we chose the coefficient of linear correlation, indicating the degree of data linear interconnection:

$$R = \frac{\mathrm{cov}(x,y)}{\sigma_x \sigma_y}, \qquad (7)$$

here, $\mathrm{cov}(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ and

$$\sigma_x = \sqrt{\tfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad \sigma_y = \sqrt{\tfrac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}.$$

Pairing preference index

For pairing preference index G_{ij} we used the ratio of contact values C_{ij} (where contacts are calculated according to the Voronoi mathematical model, and two residues are in contact if the Voronoi polyhedra of any two atoms share a face of nonzero area) and theoretical values (1), (2) $P_{ij} * N$, where P_{ij} is the probability of contact between residues of i and j types, and N is the total number of contacts:

$$G_{ij} = \frac{C_{ij}}{P_{ij} * N} \tag{8}$$

This value reflects a correlation of the observed number of contacts with the expected one. It has the following properties: if the expected number of contacts is close to the observed number, G_{ij} takes a value near 1. G_{ij} is less than 1 if the expected number of contacts exceeds the observed number of contacts, and is greater than 1 if the expected number of contacts is less than the observed number of contacts.

MATERIALS

Protein–protein interface dataset was taken from a previously published paper.[45] The authors[45] extracted all two-chain complexes from PDB with resolution lower than 3.5 Å that have at least 10 interacting residues in each chain. The first filter rules out interfaces that have no biological significance, when interface occurs because of crystallographic symmetry (space group and unit cell). The second filtering procedure removes structurally redundant complexes according to SCOP classes. In the final step, authors extract only representatives of the groups, resulting in 4602 interfaces. So, these interfaces are combined into 3422 PDB files. Selected dataset of protein–protein interfaces includes 3067 interfaces formed by two identical chains. Identical chains were determined by the FASTA[46] alignment program. Chains are treated as identical if their coefficient of identity was no less than 94%. So, the number of homodimer interfaces forms 67% from the total number of interfaces. This dataset is the largest to date.

The list of all contacting atoms with distance and contact area between them was compiled by the Voronoi–Delaunay tessellation method for each of the 4602 interfaces. After this, we calculated the total interaction area for each interacting residue. Common faces of Voronoi polyhedra form spatial boundaries between interacting protein chains.

RESULTS

The overall set of interacting units contains 4602 interfaces, 421,956 interresidue contacts, and 2,057,304 interatomic contacts. A histogram of residue contact area is shown in Figure 2. The maximum of the distribution is attained around null contact area. As the area of the contact rises, the number of the contacts drops significantly. There is a local minimum at 3 Å^2 region. Additionally, the distribution is dome-shaped with the mean area about 8 Å^2. Obviously, the envelope curve for this histogram can be interpreted as a density of the distribution composed of two parts, namely, casual and specific contacts (Fig. 1), as we proposed in the *Statistical Interpretation* subsection of the Methods section. It is possible to assign appropriate values to parameters in Eqs. (3), (5), and (6), and their corresponding shares in the summary curve to approximate the experimental data (Fig. 2). We suggested that these specific contacts are responsible for protein recognition and binding.

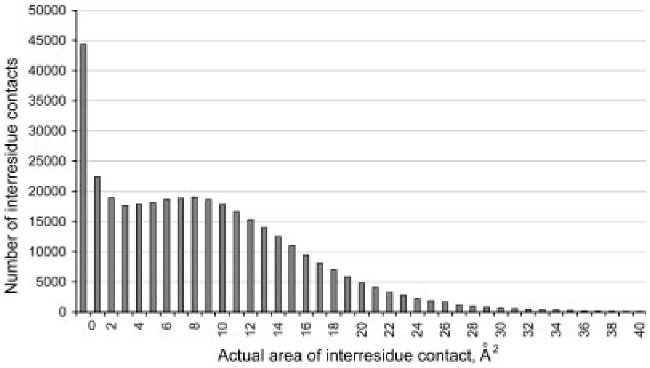

Fig. 2. Histogram of residue contact areas at protein–protein interfaces. Data were obtained by the Voronoi-Delaunay tessellation method for all 4602 protein–protein interfaces.[45] The maximum of this distribution is located around the null contact area. As the area of the contact increases, the number of the contacts drops significantly. There is a local minimum about 3 Å^2. Additionally, the distribution is dome-shaped with a mean about 8 Å^2. The envelope curve for this histogram can be interpreted as a density of the distribution composed of two parts, representing stochastic and specific contacts (Fig. 1). It is possible to assign appropriate values of parameters in Eqs. (3), (5), and (6) and their corresponding shares in the summary curve to fulfill the approximation (Fig. 1).

Amino Acid Composition

All previous studies of protein–protein interfaces operated with a similar definition of interresidue contact. A contact was defined as existing, if two residues (atoms) belonged to different chains and had a mutual distance of less than some cutoff value. In this work we defined the contact using the mathematical Voronoi–Delaunay model. Two residues are in contact if the Voronoi polyhedra of any two atoms share a face of nonzero area.

Our program provides information about the neighbors of each atom, the contact area between Voronoi polyhedra for neighboring atoms, the volume of each Voronoi polyhedron, the contact area between amino acid residues, and the volume of amino acid residues. We can find all spatial neighbors for each atom, volumes of each polyhedron and contact areas between neighbor atoms. As a result, we can define precisely all residues involved in protein–protein interface formation. The calculated residue composition statistics of protein–protein interfaces and proteins as a whole are shown in Figure 3 and Table I.

The amino acid composition in proteins is in a good agreement with the previously published data.[47] We detected some insignificant deviations in residue composition in proteins and in combining sites. Ala, Arg, Gly, Tyr, and Val show the most prominent variations. Correlation coefficient for composition in proteins and in homodimer interfaces is 0.94.

We compared the share of the residues, W_j, at protein interfaces observed in this study with that obtained by Chakrabarti and Janin,[4] Bahadur et al.,[10] Keskin et al.,[29] Ansari and Helms,[28] Glaser et al.,[33] and Saha et al.[48] There is evident similarity of residue compositions for homodimers in this study and in Bahadur et al.[10] The correlation coefficient for residue composition at homodimer interfaces is 0.98 in this case. However, there is almost no similarity between our data and those of Chakrabarti and Janin[4] on heterocomplexes (correlation coefficient is 0.57).

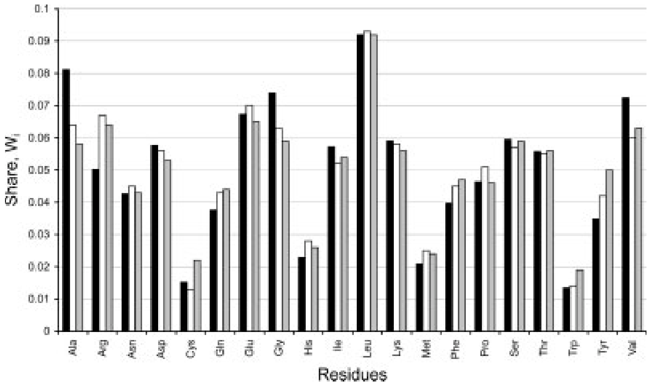

Fig. 3. Amino acid residue composition, W_i, in proteins and protein–protein interfaces. The shares are shown for whole proteins (black bars), for protein–protein interfaces of homodimers (white bars), and for protein–protein interfaces of heterocomplexes (grey bars). These data were obtained for all 4602 interfaces,[45] 3067 of which are homodimers. Contacting residues of protein–protein interfaces of homodimers and heterocomplexes were obtained by the Voronoi–Delaunay tessellation method. The differences in residue composition at protein–protein interfaces in homodimers and heterocomplexes are not significant. We detected some significant deviations in residue composition in proteins and at interfaces. Ala, Arg, Gly, Tyr, and Val show the most pronounced variations. The shares of amino acids in proteins are in good agreement with the previously published data.[46]

It has been shown[4,10] that homodimer interfaces contain more hydrophobic residues than heterocomplex interfaces. However, according to our findings, the compositional difference is not so significant. The correlation coefficient is 0.98 for amino acid interface compositions of homodimers and heterocomplexes.

Probably, the differences observed are associated with the dataset size. Statistical methods allowed estimating the minimal size of the set necessary for the estimated parameter to fall into a confidence interval of the given length at the given confidence level.[49] If we want to estimate the amino acid composition of interfaces with an accuracy of 2% at 95% confidence, then, by our estimates, the minimal set must contain about 460 interfaces.

$$n = \hat{p}\hat{q} * \left(\frac{z_{\alpha/2}}{E}\right)^2 = 0.05 * 0.95 * \left(\frac{1.96}{0.02}\right)^2 = 456.19,$$

where $E = 0.02$ is the accuracy of estimation, $\hat{p} = 0.05$ is the known mean share of residue composition, $\hat{q} = 1 - \hat{p}$, and for 95% confidence $z_{\alpha/2}$ is 1.96.

Our data are in good agreement with those of Keskin et al.,[29] with the exception of Leu, which is twice as frequent in our data for all proteins, and at interfaces than

TABLE I. Amino Acid Residue Composition in Proteins and at Protein-Protein Interfaces

Amino acid residues	Residues at the interfaces of homodimers		Residues at the interfaces of heterocomplexes		Amino acid residues in the proteins	
	Value	Share	Value	Share	Value	Share
ALA	12,993	0.064	5,496	0.058	131,027	0.081
ARG	13,659	0.067	6,029	0.064	80,990	0.072
ASN	9,085	0.045	4,062	0.043	68,845	0.092
ASP	11,314	0.056	5,025	0.053	93,016	0.057
CYS	2,570	0.013	2,091	0.022	24,635	0.021
GLN	8,825	0.043	4,147	0.044	60,719	0.015
GLU	14,201	0.070	6,094	0.065	108,574	0.074
GLY	12,770	0.063	5,524	0.059	119,302	0.060
HIS	5,611	0.028	2,449	0.026	37,004	0.056
ILE	10,597	0.052	5,090	0.054	92,294	0.058
LEU	18,828	0.093	8,646	0.092	148,527	0.067
LYS	11,793	0.058	5,283	0.056	95,364	0.043
MET	4,982	0.025	2,295	0.024	33,766	0.038
PHE	9,057	0.045	4,417	0.047	64,094	0.050
PRO	10,354	0.051	4,314	0.046	74,825	0.059
SER	11,509	0.057	5,604	0.059	96,021	0.046
THR	11,255	0.055	5,295	0.056	90,042	0.040
TRP	2,939	0.014	1,755	0.019	21,685	0.035
TYR	8,623	0.042	4,668	0.05	56,240	0.013
VAL	12,193	0.060	5,954	0.063	116,826	0.023

Share is the ratio of the number of amino acids of definite type to the overall number of amino acids.

the frequencies observed by Keskin et al. The dataset of Keskin et al.[29] contains 156 protein–protein interfaces. They define interface contact if two residues belong to different chains and if the distance between the backbone atoms was less than 6.4 Å. However, we can not calculate a numerical estimation for comparison because there are no numerical data on amino acid composition in the work[29] and other works, mentioned below.

We also have observed a good qualitative correlation with the Ansari and Helms[28] data. Although numerical correspondence was not so good, the same tendency was observed for 16 out of 20 types of residues. Frequencies of occurrence of Leu, Cys, and Glu were the same in proteins and in interfaces as in our data. At the same time, Ansari and Helms[28] found that interfaces were enriched by Leu and Glu in comparison to whole proteins, and Cys was abundant in proteins, not in interfaces. According to the Ansari and Helms[28] data, Gly had the same frequency of occurrence in all proteins and at interfaces, but in our data Gly was rare at interfaces. Some discrepancy between the data is apparently a consequence of the method used for contact definition and the dataset choice. Ansari and Helms[28] used a dataset of 170 interfaces, and the cutoff distance of 5 and 3.5 Å for side chain and backbone atoms, respectively.

Only 6 residues out of 20 (Tyr, Gln, His, Met, Cys, and Trp) were in a good agreement in our data and that of Glaser et al.[33] The observation of Glaser et al.[33] was that Gly had a maximum frequency of occurrence, which does not correlate with our results. This tendency can be related to the cutoff distance of 6 Å between C_β atoms is used to define the residue-residue contacts by Glaser et al.[33] The dataset of Glaser et al.[33] contained 621 interfaces, 404 from them were homodimers and 217 were heterodimers.

Saha et al.[48] used a cutoff distance of 4.5 Å between atoms. The dataset of Saha et al.[48] contained only 70 interfaces. Their data have no overall correlations with ours, but individual observations do correlate, for example, the Leu prevalence at protein interfaces.

Note, that the discrepancy in frequency of occurrence of residues at protein–protein interfaces can be explained by either the choice of the dataset or the choice of the method used for contact definition. Various authors set a different cutoff value. However, there is no universal choice method for a cutoff as a matter of fact because at low value cutoffs some contacts may be missed whereas some contacts can be erroneously defined when using large cutoff values. The difficulties in using the distance-dependent approach are sometimes apparent, for example, the ones when erroneous glycine clusters are identified at the interface surface. Additionally, the poor correspondence between our data and the results of other researchers can be partially explained by the small size of most other datasets analyzed to date.

Note that the distance between neighbor interacting atoms does not exceed 6 Å. In our scheme there always are neighbor atoms, including artificial water environment.

Residue–Residue Contacts

Homodimers

We calculated the number of contacts C_{ij} between residue types i and j for homodimers (Table II) and heterocomplexes (Table III). There is an obvious tendency (Tables II) for Leu, Val, and Arg to form contacts (Leu-Leu 6235, Leu-Val 4223, Leu-Ile 4199, Arg-Glu 4499). Probably, these residues form contacts owing to their large size and their high share in the amino acid composition (Table I).

Heterocomplexes

There is also a tendency to contact between oppositely charged residues (Glu-Arg 2122, Glu-Lys 1704, Asp-Arg 1684, Asp-Lys 1331) and between hydrophobic residues (Leu-Leu 2237, Leu-Val 2478, Leu-Ile 2291, Leu-Phe 1984, Leu-Ala 1675) (Table III).

However, we cannot compare data for homodimers and heterocomplexes directly because of the different size of these sets. To account for differences in set sizes, we used a preference index G_{ij} (8).

Residue-Residue Contact Preferences

On the basis of C_{ij} values, we estimated the prevalence of contacts between each pair of residues i and j using the pairwise preference indices G_{ij} (8) as explained in the Methods. The calculated values of G_{ij} are given as the 20 × 20 symmetric matrices in Tables IV and V. The residues in Tables II, III, IV, V, VI, and VII are arranged in alphabetical order. From the formula for G_{ij} (8) it follows that index value 1.0 corresponds to a stochastic process. If the index value is greater than 1.0, there are more frequent contact than in a stochastic process. Finally, for the index values less than 1.0, the contacts are less frequent than in the casual model.

Homodimers

The largest preference index values are observed for contacts Cys-Cys (4.85). Large values are also characteristic of interactions between residues with opposite charges: Arg-Asp (1.96), Arg-Glu (1.72), Lys-Asp (1.87), and Lys-Glu (1.92). Rather small values are observed for the contacts Gly-Pro (1.68) and Met-Met (1.64). In the range 1.43–1.31 lie hydrophobic contacts between large hydrophobic residues, such as Leu, Ile, and Phe. The lowest G_{ij} values are found in the case of Asp contacts with Leu (0.6), Ile (0.63), Phe, and Cys (0.64).

Heterocomplexes

As with homodimers, the largest value of preference index is observed for contacts Cys-Cys (3.48). Likewise, there are interactions between oppositely charged residues (2.14–1.89) and contacts between large hydrophobic residues (1.56–1.53). However, in comparison with homodimers, we see a higher preference index for contacts between cysteines and glycines (1.55), cysteines and his-

TABLE II. Residue-Residue Contacts in Homodimers, C_{ij}

	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
ALA	1,798	2,092	1,210	1,223	366	1,252	1,529	1,557	835	1,777	3,370	1,328	1,018	1,811	1,390	1,510	1,582	563	1,499	2,022
ARG	2,092	2,297	1,875	3,869	407	1,811	4,499	2,351	1,093	1,759	3,140	1,820	989	1,828	1,971	2,121	2,056	738	2,078	2,082
ASN	1,210	1,875	1,270	1,226	254	1,172	1,505	1,310	687	1,023	1,862	1,341	530	1,117	1,140	1,377	1,438	410	1,153	1,099
ASP	1,223	3,869	1,226	999	189	1,265	1,392	1,303	982	910	1,545	2,609	544	904	1,073	1,513	1,368	454	1,405	1,084
CYS	366	407	254	189	334	218	251	360	175	338	580	260	199	332	338	318	333	121	305	393
GLN	1,252	1,811	1,172	1,265	218	1,126	1,465	1,263	650	1,103	2,211	1,352	684	1,212	1,217	1,302	1,349	416	1,149	1,127
GLU	1,529	4,499	1,505	1,392	251	1,465	1,618	1,590	1,257	1,386	2,905	3,552	840	1,426	1,519	1,811	1,680	542	1,777	1,655
GLY	1,557	2,351	1,310	1,303	360	1,263	1,590	1,480	785	1,322	2,104	1,538	771	1,334	2,321	1,414	1,545	554	1,668	1,373
HIS	835	1,093	687	982	175	650	1,257	785	592	741	1,423	711	396	803	711	809	773	295	940	769
ILE	1,777	1,759	1,023	910	338	1,103	1,386	1,322	741	2,200	4,199	1,220	1,108	2,135	1,239	1,261	1,575	601	1,507	2,283
LEU	3,370	3,140	1,862	1,545	580	2,211	2,905	2,104	1,423	4,199	6,235	2,342	1,849	3,692	2,290	2,289	2,703	1,061	2,790	4,223
LYS	1,328	1,820	1,341	2,609	260	1,352	3,552	1,538	711	1,220	2,342	1,278	657	1,224	1,188	1,509	1,576	462	1,521	1,395
MET	1,018	989	530	544	199	684	840	771	396	1,108	1,849	657	760	954	737	684	778	335	811	1,103
PHE	1,811	1,828	1,117	904	332	1,212	1,426	1,334	803	2,135	3,692	1,224	954	1,857	1,435	1,287	1,419	628	1,733	2,206
PRO	1,390	1,971	1,140	1,073	338	1,217	1,519	2,321	711	1,239	2,290	1,188	737	1,435	1,452	1,278	1,306	597	1,682	1,491
SER	1,510	2,121	1,377	1,513	318	1,302	1,811	1,414	809	1,261	2,289	1,509	684	1,287	1,278	1,374	1,531	513	1,410	1,452
THR	1,582	2,056	1,438	1,368	333	1,349	1,680	1,545	773	1,575	2,703	1,576	778	1,419	1,306	1,531	1,571	524	1,298	1,798
TRP	563	738	410	454	121	416	542	554	295	601	1,061	462	335	628	597	513	524	294	534	596
TYR	1,499	2,078	1,153	1,405	305	1,149	1,777	1,668	940	1,507	2,790	1,521	811	1,733	1,682	1,410	1,298	534	1,200	1,666
VAL	2,022	2,082	1,099	1,084	393	1,127	1,655	1,373	769	2,283	4,223	1,395	1,103	2,206	1,491	1,452	1,798	596	1,666	2,208

Contact values C_{ij} were calculated in accordance with the Voronoi mathematical model: two residues are in contact if the Voronoi polyhedra of any two atoms share a face of nonzero area. We computed all contacts between residues of definite types from the set of all homodimer interfaces (3067 interfaces out of 4602[45]). For example, in all homodimer interfaces of the set we observed 1499 contacts between Ala and Tyr residues.

TABLE III. Residue-Residue Contacts in Heterocomplexes, C_{ij}

	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
ALA	461	889	524	461	178	514	659	673	354	868	1,675	537	457	874	549	687	694	339	837	949
ARG	889	673	918	1,684	378	943	2,122	1,092	462	903	1,415	821	487	840	899	1,021	935	468	1,161	956
ASN	524	918	273	539	183	580	634	690	295	497	707	639	265	670	514	614	593	212	665	496
ASP	461	1,684	539	240	114	625	599	538	423	477	736	1,331	261	482	513	749	565	262	735	550
CYS	178	378	183	114	402	250	179	404	228	214	845	213	87	401	188	232	226	98	282	359
GLN	514	943	580	625	250	342	748	636	303	607	1,005	602	291	570	565	657	643	274	714	711
GLU	659	2,122	634	599	179	748	362	583	529	668	1,116	1,704	352	602	605	928	772	283	910	826
GLY	673	1,092	690	538	404	636	583	410	354	642	935	657	321	660	621	652	691	307	750	730
HIS	354	462	295	423	228	303	529	354	116	382	568	320	222	346	270	498	394	185	444	351
ILE	868	903	497	477	214	607	668	642	382	733	2,291	582	512	1,101	555	612	790	403	896	1,255
LEU	1,675	1,415	707	736	845	1,005	1,116	935	568	2,291	2,237	1,074	950	1,984	946	1,049	1,177	682	1,506	2,478
LYS	537	821	639	1,331	213	602	1,704	657	320	582	1,074	341	337	524	501	762	739	321	843	683
MET	457	487	265	261	87	291	352	321	222	512	950	337	171	503	300	340	351	201	380	504
PHE	874	840	670	482	401	570	602	660	346	1,101	1,984	524	503	563	640	779	758	448	1,113	1,063
PRO	549	899	514	513	188	565	605	621	270	555	946	501	300	640	341	631	589	398	859	714
SER	687	1,021	614	749	232	657	928	652	498	612	1,049	762	340	779	631	435	773	311	774	776
THR	694	935	593	565	226	643	772	691	394	790	1,177	739	351	758	589	773	406	354	807	876
TRP	339	468	212	262	98	274	283	307	185	403	682	321	201	448	398	311	354	123	379	440
TYR	837	1,161	665	735	282	714	910	750	444	896	1,506	843	380	1,113	859	774	807	379	422	1,029
VAL	949	956	496	550	359	711	826	730	351	1,255	2,478	683	504	1,063	714	776	876	440	1,029	776

Contact values C_{ij} were calculated in accordance with the Voronoi mathematical model: two residues are in contact if the Voronoi polyhedra of any two atoms share a face of nonzero area. We computed all contacts between residues of definite type for all heterocomplex interfaces (1535 interfaces out of 4602[45]). For example, in all heterocomplexes of the set we observed 1984 contacts between Leu and Phe residues.

tidines (1.53), cysteines and leucines (1.58), and we do not see a high preference index for contacts Met-Met. The lowest G_{ij} values were detected in the case of contacts of residues with similar charges (0.41–0.48). Pairs of hydrophobic–hydrophilic residues are also characterized by low G_{ij} (e.g., Leu-Asp).

Both for homodimers and heterocomplexes maximal preference index values exist for contacts between two cysteine residues. Contacts between pairs of cysteine residues produce the highest G_{ij} value 4.85 for homodimers (Table IV) and 3.48 for heterocomplexes (Table V), indicating that such pairing is very likely. Second largest value indices G_{ij} correspond to the interaction of residues with opposite charges. As expected, the charged residues show a tendency to interact with residues of opposite charge. This tendency is characteristic both for interfaces in homodimers and in heterocomplexes. The hydrophobic residues show intermediate tendency to pair with each other.

Neither for homodimers nor for heterocomplexes there was preference for small residues such as Gly and Ala to get paired as reported by Glaser et al.[33] Apparently, this is connected with the differences in methods used to compute the contacts. The distance cutoff method allowed revealing contacts with small residues lying close to a protein surface, including buried ones. The Voronoi–Delaunay method does not define such kind of contact, registering only direct neighbors.

Differences in Preference Index Tables

Comparing the preference indices G_{ij} for homodimers (Table IV) and heterocomplexes (Table V) we can see more figures greater than 1.2 in central diagonal cells in case of homodimers than for heterocomplexes. One can see only Cys-Cys preference index larger than 1.2 for heterocomplexes and also Asn-Asn, His-His, Ile-Ile, Leu-Leu, Met-Met, and Trp-Trp for homodimers. Preference index for contact residues of the same type is greater for interactions of identical protein chains than for heterocomplexes. The same effect was reported by Dasgupta et al.[12] They revealed the strongest preference, in descending order, for Cys-Cys, Glu-Cys, Trp-Lys, Val-Trp, Trp-Pro, and Met-Met in contacts between oligomer subunits. Also, pairs with the same charge His-His, Arg-Arg, Asp-Asp have large interaction preference according to Dasgupta et al.,[12] which is contrary to our data.

The difference between indices G_{ij} for homodimers and for heterocomplexes calculated by subtraction of Table V from Table IV is given in Table VI. The greatest distinctions are observed for contacts Cys-Cys (1.37), Met-Met (0.81), Trp-Trp (0.74), and His-His (0.73), that is, the corresponding preference indices are higher in homodimers than in heterocomplexes. Note that this is true for all central diagonal elements (contacts between residues of the same type). For example, for contacts of Pro-Pro type and Trp-Trp type the prevalence of homodimers is 0.42 and 0.74, respectively (Table VI).

Although pairs with the same type have a greater preference at homodimer interfaces than in heterocomplexes, we have not found any preference for residues with similar charges. Furthermore, we note an increase in preference index for Pro-Gly on 0.52 and a decrease for Cys-Leu on 0.61 in the case of homodimers. The properties of the interaction site formed by identical protein chains differ from those in heterocomplexes. In particular, the former contain contacts of symmetric pairs. It is reasonable to assume that homodimer interfaces are enriched in contacts between amino acids of the same type owing to the symmetry relations including twofold symmetry.[50] So, for example, if a contact exists between the 70th residue of the first chain and the 345th residue of second, one may expect, that there is also a contact between the 345th residue of the first chain and the 70th of the second chain. Thus, in the central zone of such an interface the spatial approach of residues with identical numbers in the polypeptide chain, that is, identical residues, is possible. We have found 6600 such contacts among 167,000 residue–residue contacts at homodimer interfaces.

The comparison with Glaser et al.[33] is of particular interest in view of the prevalence of homodimers in their data (404 out of 621 interfaces), even though this comparison was not straightforward due to the fact that heterocomplexes and homodimers were not separated in the their work. There is a reasonably good agreement with respect to amino acid pairs preferences, such as Cys-Cys, Met-Met, His-His, and different results for Gly-Gly, Leu-Leu, Phe-Trp, Arg-Trp, and Asp-Ser.

The importance of contacts between the identical amino acids because of the 2-fold symmetry was mentioned by Saha et al.[48] and in some other works. Unfortunately, Saha et al.[48] have no reported quantitative results to support this argument, but they present only qualitative characteristics. Saha et al.[48] noted the striking feature of the dimeric interfaces, the existence of contacts between the same types of residues, however, they observed few Cys-Cys, Ile-Ile, Tyr-Tyr, Trp-Trp, His-His, and Ser-Ser contacts. We show analogous data only for Tyr-Tyr, Trp-Trp, and Ser-Ser pairs. For Ile-Ile and His-His pairs the frequency does not deviate from the expected, and Cys-Cys interactions are more frequent than it could be expected by chance.

Cys-Cys Contacts

Using the Cys-Cys pairs we illustrate the ability of the Voronoi–Delaunay tessellation method to solve the protein–protein contact problem at the atomic level. The histogram of distances between sulfur atoms of the cysteine residues is presented in Figure 4. These data are calculated on the basis of the full set of 4602 protein–protein interfaces, without division into homodimers and heterocomplexes. The histogram demonstrates the strong specificity of S—S interactions (Figure 4). We have checked up all Cys-Cys contacts by DSSP.[51] All S—S contacts with distances less than 2.85 Å are disulfide bonds.

TABLE IV. Pairing Preference Index for Homodimers, G_{ij}

	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
ALA	1.09	0.92	0.95	0.85	1.08	0.96	0.80	1.00	0.97	1.08	1.15	0.83	1.16	1.11	0.95	1.01	1.01	0.99	0.96	1.14
ARG	0.92	0.74	1.07	1.96	0.88	1.01	1.72	1.10	0.93	0.77	0.78	0.82	0.82	0.81	0.98	1.04	0.95	0.94	0.97	0.85
ASN	0.95	1.07	1.28	1.10	0.97	1.17	1.02	1.09	1.04	0.80	0.82	1.08	0.78	0.89	1.00	1.20	1.19	0.93	0.95	0.80
ASP	0.85	1.96	1.10	0.80	0.64	1.12	0.84	0.96	1.32	0.63	0.60	1.87	0.71	0.64	0.84	1.17	1.00	0.92	1.03	0.70
CYS	1.08	0.88	0.97	0.64	4.85	0.82	0.65	1.13	1.00	1.00	0.97	0.79	1.11	1.00	1.13	1.05	1.04	1.04	0.95	1.08
GLN	0.96	1.01	1.17	1.12	0.82	1.10	0.98	1.04	0.96	0.85	0.96	1.07	0.99	0.95	1.06	1.11	1.10	0.93	0.94	0.81
GLU	0.80	1.72	1.02	0.84	0.65	0.98	0.74	0.89	1.27	0.73	0.86	1.92	0.83	0.76	0.90	1.06	0.93	0.83	0.99	0.81
GLY	1.00	1.10	1.09	0.96	1.13	1.04	0.89	1.01	0.97	0.85	0.76	1.02	0.94	0.87	1.68	1.01	1.05	1.04	1.13	0.82
HIS	0.97	0.93	1.04	1.32	1.00	0.96	1.27	0.97	1.33	0.87	0.93	0.85	0.87	0.95	0.93	1.05	0.95	1.00	1.16	0.83
ILE	1.08	0.77	0.80	0.63	1.00	0.85	0.73	0.85	0.87	1.33	1.43	0.76	1.27	1.31	0.85	0.85	1.01	1.06	0.96	1.28
LEU	1.15	0.78	0.82	0.60	0.97	0.96	0.86	0.76	0.93	1.43	1.20	0.82	1.19	1.27	0.88	0.87	0.97	1.05	1.00	1.33
LYS	0.83	0.82	1.08	1.87	0.79	1.07	1.92	1.02	0.85	0.76	0.82	0.82	0.77	0.77	0.83	1.04	1.03	0.84	1.00	0.81
MET	1.16	0.82	0.78	0.71	1.11	0.99	0.83	0.94	0.87	1.27	1.19	0.77	1.64	1.10	0.95	0.87	0.94	1.11	0.98	1.17
PHE	1.11	0.81	0.89	0.64	1.00	0.95	0.76	0.87	0.95	1.31	1.27	0.77	1.10	1.15	0.99	0.88	0.92	1.12	1.12	1.26
PRO	0.95	0.98	1.00	0.84	1.13	1.06	0.90	1.68	0.93	0.85	0.88	0.83	0.95	0.99	1.12	0.97	0.94	1.18	1.21	0.94
SER	1.01	1.04	1.20	1.17	1.05	1.11	1.06	1.01	1.05	0.85	0.87	1.04	0.87	0.88	0.97	1.03	1.08	1.00	1.00	0.91
THR	1.01	0.95	1.19	1.00	1.04	1.10	0.93	1.05	0.95	1.01	0.97	1.03	0.94	0.92	0.94	1.08	1.06	0.97	0.87	1.06
TRP	0.99	0.94	0.93	0.92	1.04	0.93	0.83	1.04	1.00	1.06	1.05	0.84	1.11	1.12	1.18	1.00	0.97	1.50	0.99	0.97
TYR	0.96	0.97	0.95	1.03	0.95	0.94	0.99	1.13	1.16	0.96	1.00	1.00	0.98	1.12	1.21	1.00	0.87	0.99	0.81	0.99
VAL	1.14	0.85	0.80	0.70	1.08	0.81	0.81	0.82	0.83	1.28	1.33	0.81	1.17	1.26	0.94	0.91	1.06	0.97	0.99	1.15

G_{ij} is the ratio of contact numbers C_{ij} to the theoretical values (1), (2) $P_{ij}{*}N$, where N is the total number of contacts. The contacts were calculated in accordance with the Voronoi mathematical model: two residues are in contact if the Voronoi polyhedra of any two atoms share a face of nonzero area. For example, we obtained $G_{ij} = 0.96$ for the interaction between Ala and Tyr residues for a subset of all homodimer interfaces. Although pairs with one and the same type of residue have a greater preference at homodimer interfaces than in heterocomplexes, we have not found any preference for residues with similar charges.

TABLE V. Pair Formation Preference Index for Heterocomplexes, G_{ij}

	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
ALA	0.69	0.91	0.98	0.76	0.64	0.87	0.85	1.07	0.99	1.14	1.29	0.78	1.23	1.15	0.96	1.01	1.04	1.02	1.06	1.13
ARG	0.91	0.48	1.18	1.92	0.94	1.10	1.89	1.20	0.89	0.82	0.76	0.82	0.91	0.76	1.09	1.04	0.97	0.98	1.02	0.78
ASN	0.98	1.18	0.64	1.12	0.82	1.23	1.03	1.37	1.03	0.82	0.69	1.16	0.89	1.10	1.13	1.14	1.11	0.80	1.05	0.74
ASP	0.76	1.92	1.12	0.44	0.45	1.17	0.86	0.95	1.31	0.69	0.63	2.14	0.78	0.70	1.00	1.23	0.94	0.88	1.03	0.72
CYS	0.64	0.94	0.82	0.45	3.48	1.02	0.56	1.55	1.53	0.68	1.58	0.75	0.56	1.27	0.79	0.83	0.81	0.71	0.86	1.03
GLN	0.87	1.10	1.23	1.17	1.02	0.66	1.10	1.15	0.96	0.90	0.88	0.99	0.89	0.85	1.13	1.10	1.09	0.94	1.03	0.96
GLU	0.85	1.89	1.03	0.86	0.56	1.10	0.41	0.80	1.28	0.76	0.75	2.14	0.82	0.69	0.92	1.19	1.00	0.74	1.00	0.85
GLY	1.07	1.20	1.37	0.95	1.55	1.15	0.80	0.70	1.05	0.90	0.77	1.02	0.92	0.93	1.16	1.03	1.10	0.99	1.01	0.93
HIS	0.99	0.89	1.03	1.31	1.53	0.96	1.28	1.05	0.60	0.94	0.82	0.87	1.12	0.85	0.88	1.38	1.10	1.05	1.05	0.78
ILE	1.14	0.82	0.82	0.69	0.68	0.90	0.76	0.90	0.94	0.84	1.56	0.74	1.21	1.27	0.85	0.79	1.04	1.07	1.00	1.31
LEU	1.29	0.76	0.69	0.63	1.58	0.88	0.75	0.77	0.82	1.56	0.90	0.81	1.33	1.35	0.86	0.80	0.91	1.07	0.99	1.53
LYS	0.78	0.82	1.16	2.14	0.75	0.99	2.14	1.02	0.87	0.74	0.81	0.48	0.88	0.67	0.85	1.10	1.07	0.95	1.04	0.79
MET	1.23	0.91	0.89	0.78	0.56	0.89	0.82	0.92	1.12	1.21	1.33	0.88	0.83	1.19	0.95	0.91	0.95	1.10	0.87	1.08
PHE	1.15	0.76	1.10	0.70	1.27	0.85	0.69	0.93	0.85	1.27	1.35	0.67	1.19	0.65	0.99	1.02	1.00	1.20	1.24	1.11
PRO	0.96	1.09	1.13	1.00	0.79	1.13	0.92	1.16	0.88	0.85	0.86	0.85	0.95	0.99	0.70	1.10	1.04	1.42	1.28	1.00
SER	1.01	1.04	1.14	1.23	0.83	1.10	1.19	1.03	1.38	0.79	0.80	1.10	0.91	1.02	1.10	0.64	1.15	0.93	0.97	0.91
THR	1.04	0.97	1.11	0.94	0.81	1.09	1.00	1.10	1.10	1.04	0.91	1.07	0.95	1.00	1.04	1.15	0.61	1.07	1.02	1.04
TRP	1.02	0.98	0.80	0.88	0.71	0.94	0.74	0.99	1.05	1.07	1.07	0.95	1.10	1.20	1.42	0.93	1.07	0.76	0.97	1.06
TYR	1.06	1.02	1.05	1.03	0.86	1.03	1.00	1.01	1.05	1.00	0.99	1.04	0.87	1.24	1.28	0.97	1.02	0.97	0.45	1.04
VAL	1.13	0.78	0.74	0.72	1.03	0.96	0.85	0.93	0.78	1.31	1.53	0.79	1.08	1.11	1.00	0.91	1.04	1.06	1.04	0.73

G_{ij} is the ratio of contact numbers C_{ij} to the theoretical values (1), (2) $P_{ij}*N$, where N is the total number of contacts. The contacts have been calculated accordance with the Voronoi mathematical model: two residues are in contact if the Voronoi polyhedra of any two atoms share a face of nonzero area. For example, we $G_{ij} = 0.99$ for the interaction between Leu and Phe residues for a subset of all heterocomplex interfaces.

TABLE VI. The Difference in Formation Preference Index for Homodimers and Heterocomplexes

	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
ALA	**0.40**	0.01	–0.03	0.09	**0.44**	0.09	–0.05	–0.07	–0.02	–0.06	–0.14	0.05	–0.07	–0.04	–0.01	0.00	–0.03	–0.03	–0.10	0.01
ARG	0.01	0.26	–0.11	0.04	–0.06	–0.09	–0.17	–0.10	0.04	–0.05	0.02	0.00	–0.09	0.05	–0.11	0.00	–0.02	–0.04	–0.05	0.07
ASN	–0.03	–0.11	**0.64**	–0.02	0.15	–0.06	–0.01	–0.28	0.01	–0.02	0.13	–0.08	–0.11	–0.21	–0.13	0.06	0.08	0.13	–0.10	0.06
ASP	0.09	0.04	–0.02	**0.36**	0.19	–0.05	–0.02	0.01	0.01	–0.06	–0.03	–0.27	–0.07	–0.06	–0.16	–0.06	0.06	0.04	0.00	–0.02
CYS	**0.44**	–0.06	0.15	0.19	**1.37**	–0.20	0.09	–**0.42**	–**0.53**	**0.32**	–**0.61**	0.04	**0.55**	–0.27	**0.34**	0.22	0.23	**0.33**	0.09	0.05
GLN	0.09	–0.09	–0.06	–0.05	–0.20	**0.44**	–0.12	–0.11	0.00	–0.05	0.08	0.08	0.10	0.10	–0.07	0.01	0.01	–0.01	–0.09	–0.15
GLU	–0.05	–0.17	–0.01	–0.02	0.09	–0.12	**0.33**	0.09	–0.01	–0.03	0.11	–0.22	0.01	0.07	–0.02	–0.13	–0.07	0.09	–0.01	–0.04
GLY	–0.07	–0.10	–0.28	0.01	–**0.42**	–0.11	0.09	**0.31**	–0.08	–0.05	–0.01	0.00	0.02	–0.06	**0.52**	–0.02	–0.05	0.05	0.12	–0.11
HIS	–0.02	0.04	0.01	0.01	–**0.53**	0.00	–0.01	–0.08	**0.73**	–0.07	0.11	–0.02	–0.25	0.10	0.05	–**0.33**	–0.15	–0.05	0.11	0.05
ILE	–0.06	–0.05	–0.02	–0.06	**0.32**	–0.05	–0.03	–0.05	–0.07	**0.49**	–0.13	0.02	0.06	0.04	0.00	0.06	–0.03	–0.01	–0.04	–0.03
LEU	–0.14	0.02	0.13	–0.03	–**0.61**	0.08	0.11	–0.01	0.11	–0.13	0.30	0.01	–0.14	–0.08	0.02	0.07	0.06	–0.02	0.01	–0.20
LYS	0.05	0.00	–0.08	–0.27	0.04	0.08	–0.22	0.00	–0.02	0.02	0.01	**0.34**	–0.11	0.10	–0.02	–0.06	–0.04	–0.11	–0.04	0.02
MET	–0.07	–0.09	–0.11	–0.07	**0.55**	0.10	0.01	0.02	–0.25	0.06	–0.14	–0.11	**0.81**	–0.09	0.00	–0.04	–0.01	0.01	0.11	0.09
PHE	–0.04	0.05	–0.21	–0.06	–0.27	0.10	0.07	–0.06	0.10	0.04	–0.08	0.10	–0.09	**0.50**	0.00	–0.14	–0.08	–0.08	–0.12	0.15
PRO	–0.01	–0.11	–0.13	–0.16	**0.34**	–0.07	–0.02	**0.52**	0.05	0.00	0.02	–0.02	0.00	0.00	**0.42**	–0.13	–0.10	–0.24	–0.07	–0.06
SER	0.00	0.00	0.06	–0.06	0.22	0.01	–0.13	–0.02	–**0.33**	0.06	0.07	–0.06	–0.04	–0.14	–0.13	**0.39**	–0.07	0.07	0.03	0.00
THR	–0.03	–0.02	0.08	0.06	0.23	0.01	–0.07	–0.05	–0.15	–0.03	0.06	–0.04	–0.01	–0.08	–0.10	–0.07	**0.45**	–0.10	–0.15	0.02
TRP	–0.03	–0.04	0.13	0.04	**0.33**	–0.01	0.09	0.05	–0.05	–0.01	–0.02	–0.11	0.01	–0.08	–0.24	0.07	–0.10	**0.74**	0.02	–0.09
TYR	–0.10	–0.05	–0.10	0.00	0.09	–0.09	–0.01	0.12	0.11	–0.04	0.01	–0.04	0.11	–0.12	–0.07	0.03	–0.15	0.02	**0.36**	–0.05
VAL	0.01	0.07	0.06	–0.02	0.05	–0.15	–0.04	–0.11	0.05	–0.03	–0.20	0.02	0.09	0.15	–0.06	0.00	0.02	–0.09	–0.05	**0.42**

Each element in this table presents the difference between indices G_{ij} for homodimers and heterocomplexes, i.e. between the values from Table IV and Table V. The elements inside the (–0.3, 0.3) range are marked in bold. The most prominent differences between G_{ij} for homodimers and heterocomplexes are observed for contacts Cys-Cys (1.37) and Leu-Cys (–0.61). For all diagonal elements (i.e. contacts between residues of the same type) the preference index is markedly higher in homodimers than in heterocomplexes. The same effect was reported by Dasgupta et al.[12]. For example, for contacts of Pro-Pro type the difference is 0.42, and for contacts of Trp-Trp type it is 0.74. The larger preference index for pairs of the same type is due to the dyad symmetry of homodimers. Consequently, residues with the same numbers are close to each other in the central zone of the interface.

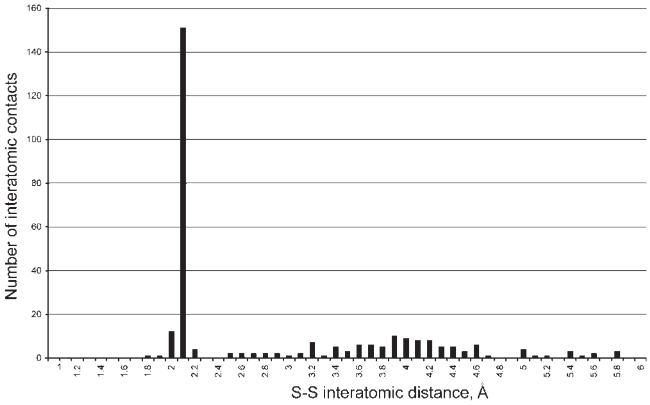

Fig. 4. Histogram of distances between two sulfur atoms of cysteine residues from different protein chains. All interatomic contacts of S—S type were selected from the full set of interatomic contacts.[45] "Contact" here means that any two atoms share a common face of Voronoi polyhedra with nonzero area. Actually, distances between sulfur atoms lie in the range from 1.8 to 5.8 Å. There is a sharp peak in the 2 Å^2 area, which coincides with the disulfide bond length. According to DSSP[51] all S—S contacts with distances less than 2.85 Å are interchain disulfide bonds.

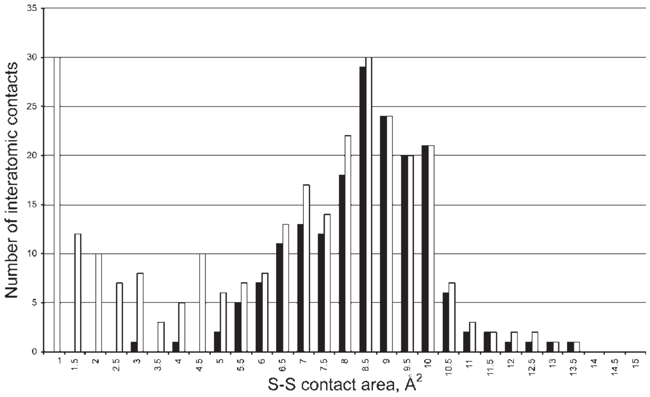

Fig. 5. Atomic contact area distribution for the S–S contacts. S—S contacts were determined by the Voronoi–Delaunay tessellation method for the full set of protein–protein interfaces.[45] All S—S contacts are shown in white bars, and disulfide bonds (according to DSSP[49] with distance less 2.85 Å, see Fig. 4) are shown in black bars. It is reasonable to assume that all S—S contacts comprise stochastic contacts and specific contacts (disulfide bonds in this case) as it was suggested in the *Statistical Interpretation* in the Methods. Specific interatomic S—S contacts that arise from disulfide bonding have an average contact area of about 8.23 Å^2.

In Figure 5, the distribution of the area of interatomic contacts is shown for all S—S contacts, and separately for the disulfide bonds annotated in DSSP. This plot is an illustrative example of the abstract model in the Statistical Interpretation section, being the sum of specific and casual interactions. Specific S—S contacts that arise from disulfide bonding have an average contact area of 8.23 Å^2.

Homodimers

The preference index G_{ij} for this type of contact reaches 4.85 (Table IV). The lowest values are observed for Cys contacts with Asp (0.64) and Glu residues (0.65). The values for all other contacts of cysteines are within 0.88–1.13.

Heterocomplexes

The maximal preference index G_{ij} for this type of contact in heterocomplexes is lower than in homodimers, and takes the value of 3.48 (Table V). The lowest values are also observed for Cys contacts with Asp (0.45) and Glu (0.56) residues.

Note that both for homodimers (Table IV) and for heterocomplexes (Table V) the Cys-Cys contacts have the highest preference index. In all interfaces, Cys-Cys interactions have very high specificity, which is emphasized by the extremely low percentage of Cys residues in the amino acid composition of the interfaces (1.3–2.2%).

Oppositely Charged Residues

Homodimers

The residue pairing indices of Table IV indicate that oppositely charged residues tend to be in contact with each other. The preference index G_{ij} for Glu-Lys contacts is rather high, comprising 1.92, for Asp-Lys 1.87, for Asp-Arg 1.96, and 1.72 for Glu-Arg contacts. The preference index values for His residues are not so high: only 1.27 for Glu-His and 1.32 for Asp-His contacts. The minimal values of preference index are for contacts of residues with similar charges and also for contacts of charged residues (Asp, Glu, Lys, and Arg) with large hydrophobic residues (Leu, Phe, Val, and Ala).

Heterocomplexes

The preference index G_{ij} for charged residues in case of heterocomplexes is also high: it is 2.14 for Glu-Lys and Asp-Lys, 1.92 for Asp-Arg, and 1.89 for Glu-Arg contacts. As in case of homodimers, the minimal values of preference index are for contacts of residues with similar charges and for contacts of charged residues (Asp, Glu, Lys, and Arg) with large hydrophobic residues (Leu, Phe, Val, and Ala).

It is necessary to note that charged residues (Arg, Glu, Asp, Lys) make ∼25% of the overall number of residues at protein–protein interfaces. The prevalence of interfacial pairs of residues with opposite charges confirms the importance of electrostatic interactions, such as salt bridges and hydrogen bonding.

Hydrophobic Interactions

Homodimers

Large hydrophobic residues (Leu, Ile, Val, and Phe) form mutual contacts with preference index values within 1.15–1.43 (Table IV). Their specific values are: 1.43 (Leu-Ile), 1.33 (Leu-Val), 1.27 (Leu-Phe), 1.28 (Ile-Val), and 1.31 (Ile-Phe). One should note a high preference index for Pro-Gly contacts (1.68).

Heterocomplexes

As in homodimers, the higher preference index values are observed for contacts between bulky hydrophobic residues: Leu-Ile (1.56), Leu-Val (1.53), Leu-Phe (1.35), Ile-Val (1.31), and Ile-Phe (1.27) (Table V).

Hydrophobic residues (Leu, Ile, Val, and Phe) as well as charged residues make ~25% of the overall number of residues at protein–protein interfaces and have high preference index values G_{ij} for hydrophobic-hydrophobic pairs Leu-Ile, Leu-Val, Leu-Phe, Ile-Val, and Ile-Phe both for homodimers and heterocomplexes. We cannot describe the role of the bulky aromatic residues because of the low preference index G_{ij} for Trp interactions in spite of others authors' findings, but we report the high preference index values for Trp-Pro contacts (1.18 for homodimers and 1.42 for heterocomplexes).

Hydrophobic-Charged Pairs

Extremely low preference index values are observed for hydrophobic-charged pairs (Tables IV and V). In homodimers for various interactions between (Arg, Lys, Glu, Asp) and (Ile, Leu, Phe, Val) the G_{ij} range from 0.6 (Asp-Leu) to 0.85 (Arg-Val), and in heterocomplexes they range from 0.63 (Asp-Leu) to 0.78 (Arg-Val). We do not find any singularity in Trp-Arg pair reported by Glasier et al.[33] The preference index values for this pair suggest a random nature of the Trp-Arg contacts (Tables IV and V).

Residues With Similar Charges

The pairwise index values for similarly charged residues indicate their extremely low propensity to form contacts in homodimers or heterocomplexes (Tables IV and V). For negatively charged residues, the preference index G_{ij} takes on values 0.74–0.84 for homodimers and 0.41–0.86 for heterocomplexes. The positively charged residues, Arg and Lys, behave similarly.

Error Estimation

We used the Jackknife procedure for G_{ij} error estimation. The procedure is based on the construction of randomly chosen half-subsets from the whole sets including 3067 homodimeric interfaces and 1535 interfaces in heterocomplexes. For each random subset, a matrix of pairing preference index G_{ij} was made. We performed calculations for 100 such random subsets, separately for homodimer interfaces and heterocomplex interfaces. The formula (9) gives the error.

$$E = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad (9)$$

Here, $\bar{x}$ is the pairing preference index for the whole set, x_i for random subset i, $n = 100$. Based on this formula, Tables VII and VIII were made. The error ranges from 0.03 to 0.49 with a mean of 0.06 for homodimers (Table VII) and from 0.02 to 0.18 with a mean of 0.04 for heterocomplexes (Table VIII). The percentage mean error is 6.33% for homodimers and 4.57% for heterocomplexes. Maximal errors take place for rare residues: Cys, His, Met, and Trp. Thus for contacts of cysteine the mean error is 0.14 (12.47%) for homodimers and 0.07 (7.32%) for heterocomplexes. The maximal error in homodimers is 22.78% for contacts Cys-Lys (0.18), whereas the minimal error is 2.80% for contacts Asn-Arg (0.03). In heterocomplexes the maximal error is 16.07% for contacts Cys-Met (0.16), and the minimal error is 1.59% for contacts Arg-Glu (0.03). For prevalent residues, such as Leu, Glu, Arg, and Ala, the mean error does not exceed 6% in homodimers and 4.5% in heterocomplexes.

CONCLUSIONS

It is common knowledge that interresidue contacts contribute to primary, secondary, tertiary, and ternary protein structures. A serious study of favorable amino acid interactions is necessary to attain progress in the problem of protein folding. Contacts in protein–protein interfaces are of special interest as the simplest model of the event. The data collected in this case can be extended to more complex systems. The biological importance of protein–protein complexes is obvious. Large molecular machines, for example, spliceosomes,[52] are formed by complexes of many proteins. Systems biology mainly addresses networks of protein–protein interrelations.[34] Indeed, combination of protein–protein interactions provides a huge variety of complexes, enabling diverse biological functions.

We analyzed the currently largest dataset of 4602 protein–protein interfaces.[45] This set of interfaces contains the maximal number of nonhomologous structures. This dataset includes additional interfaces with amino acid substitutions inserted by site-directed mutagenesis. Inclusion of these "non-natural" interfaces allows enlarging the database, whereas rare point uncertainties cannot distort the overall statistical conclusions.

The Voronoi–Delaunay method demonstrated its validity in this study. The method, being nonparametric, deduces the contacts only from atomic coordinates by an unambiguous algorithm. At the same time, this method appears to be of ample mathematical accuracy. We tested our mathematical procedure against small inaccuracies in coordinate values and detected that its results were stable. The Voronoi–Delaunay method displays resistance to errors even for rather large deviations in atom positions after omitting small contacts (area of contacts <3 Å). These contacts may be treated as casual ones (data not shown).

We suggested an a priori model of residue interactions. In this scheme, the resulting curve of contact area distribution comprises two curves, one for casual contacts and the other for specific contacts. Thus, one can observes a coexistence of casual and specific contacts at protein–protein interfaces. The distribution of casual contacts

TABLE VII. Standard Errors of Pair Formation Preference Index G_{ij} in Homodimers

	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
ALA	0.08	0.03	0.05	0.05	0.13	0.05	0.04	0.05	0.07	0.05	0.04	0.04	0.06	0.06	0.06	0.05	0.05	0.06	0.06	0.05
ARG	0.03	0.03	0.03	0.06	0.08	0.06	0.05	0.04	0.05	0.03	0.03	0.04	0.04	0.03	0.05	0.04	0.05	0.07	0.04	0.03
ASN	0.05	0.03	0.08	0.05	0.15	0.06	0.04	0.06	0.06	0.05	0.04	0.05	0.06	0.04	0.06	0.05	0.08	0.10	0.04	0.04
ASP	0.05	0.06	0.05	0.05	0.07	0.07	0.06	0.06	0.08	0.03	0.03	0.07	0.06	0.04	0.04	0.06	0.04	0.09	0.07	0.03
CYS	0.13	0.08	0.15	0.07	0.49	0.12	0.07	0.12	0.14	0.11	0.07	0.18	0.16	0.09	0.14	0.11	0.15	0.20	0.10	0.11
GLN	0.05	0.06	0.06	0.07	0.12	0.05	0.06	0.06	0.08	0.05	0.04	0.05	0.07	0.05	0.06	0.06	0.05	0.08	0.05	0.06
GLU	0.04	0.05	0.04	0.06	0.07	0.06	0.03	0.06	0.06	0.04	0.03	0.05	0.05	0.03	0.05	0.05	0.04	0.06	0.04	0.05
GLY	0.05	0.04	0.06	0.06	0.12	0.06	0.06	0.06	0.06	0.04	0.03	0.06	0.05	0.04	0.20	0.05	0.07	0.10	0.05	0.04
HIS	0.07	0.05	0.06	0.08	0.14	0.08	0.06	0.06	0.10	0.06	0.04	0.05	0.08	0.06	0.07	0.07	0.08	0.08	0.07	0.07
ILE	0.05	0.03	0.05	0.03	0.11	0.05	0.04	0.04	0.06	0.05	0.06	0.04	0.06	0.05	0.05	0.08	0.06	0.08	0.04	0.04
LEU	0.04	0.03	0.04	0.03	0.07	0.04	0.03	0.03	0.04	0.06	0.05	0.04	0.04	0.05	0.03	0.04	0.05	0.06	0.03	0.06
LYS	0.04	0.04	0.05	0.07	0.18	0.05	0.05	0.06	0.05	0.04	0.04	0.03	0.05	0.04	0.04	0.04	0.05	0.06	0.05	0.04
MET	0.06	0.04	0.06	0.06	0.16	0.07	0.05	0.05	0.08	0.06	0.04	0.05	0.09	0.08	0.06	0.07	0.08	0.11	0.06	0.07
PHE	0.06	0.03	0.04	0.04	0.09	0.05	0.03	0.04	0.06	0.05	0.05	0.04	0.08	0.04	0.06	0.04	0.07	0.08	0.07	0.05
PRO	0.06	0.05	0.06	0.04	0.14	0.06	0.05	0.20	0.07	0.05	0.03	0.04	0.06	0.06	0.12	0.05	0.07	0.11	0.05	0.08
SER	0.05	0.04	0.05	0.06	0.11	0.06	0.05	0.05	0.07	0.08	0.04	0.04	0.07	0.04	0.05	0.08	0.05	0.10	0.05	0.04
THR	0.05	0.05	0.08	0.04	0.15	0.05	0.04	0.07	0.08	0.06	0.05	0.05	0.08	0.07	0.07	0.05	0.07	0.08	0.05	0.05
TRP	0.06	0.07	0.10	0.09	0.20	0.08	0.06	0.10	0.08	0.08	0.06	0.06	0.11	0.08	0.11	0.10	0.08	0.18	0.08	0.07
TYR	0.06	0.04	0.04	0.07	0.10	0.05	0.04	0.05	0.07	0.04	0.03	0.05	0.06	0.07	0.05	0.05	0.05	0.08	0.04	0.04
VAL	0.05	0.03	0.04	0.03	0.11	0.06	0.05	0.04	0.07	0.04	0.06	0.04	0.07	0.05	0.08	0.04	0.05	0.07	0.04	0.04

The Jackknife procedure was used to estimate the standard error for G_{ij}. The procedure was based on 100 randomly chosen half-size subsets of the whole set of 3067 homodimeric interfaces. The G_{ij} tables were calculated for each random subset. The error estimation follows the formula (9).

TABLE VIII. Standard Errors of Pair Formation Preference Index G_{ij} in Heterocomplexes

	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
ALA	0.04	0.03	0.04	0.03	0.06	0.04	0.03	0.04	0.05	0.04	0.03	0.03	0.06	0.04	0.04	0.05	0.04	0.06	0.03	0.04
ARG	0.03	0.02	0.04	0.05	0.05	0.04	0.03	0.04	0.04	0.03	0.02	0.03	0.04	0.03	0.04	0.03	0.04	0.04	0.02	0.03
ASN	0.04	0.04	0.03	0.05	0.06	0.05	0.04	0.04	0.06	0.03	0.03	0.05	0.06	0.05	0.05	0.05	0.04	0.07	0.04	0.03
ASP	0.03	0.05	0.05	0.03	0.04	0.05	0.03	0.04	0.07	0.03	0.02	0.06	0.06	0.03	0.04	0.04	0.04	0.05	0.03	0.03
CYS	0.06	0.05	0.06	0.04	0.18	0.06	0.04	0.07	0.11	0.05	0.06	0.06	0.09	0.06	0.07	0.07	0.06	0.08	0.05	0.04
GLN	0.04	0.04	0.05	0.05	0.06	0.04	0.04	0.04	0.05	0.03	0.03	0.04	0.06	0.03	0.05	0.04	0.05	0.06	0.04	0.04
GLU	0.03	0.03	0.04	0.03	0.04	0.04	0.02	0.04	0.06	0.03	0.02	0.05	0.04	0.03	0.04	0.04	0.03	0.04	0.03	0.03
GLY	0.04	0.04	0.04	0.04	0.07	0.04	0.04	0.03	0.06	0.04	0.02	0.04	0.06	0.04	0.05	0.04	0.04	0.06	0.04	0.03
HIS	0.05	0.04	0.06	0.07	0.11	0.05	0.06	0.06	0.06	0.05	0.03	0.05	0.08	0.05	0.06	0.06	0.05	0.08	0.05	0.05
ILE	0.04	0.03	0.03	0.03	0.05	0.03	0.03	0.04	0.05	0.04	0.03	0.03	0.05	0.04	0.04	0.03	0.04	0.06	0.03	0.04
LEU	0.03	0.02	0.03	0.02	0.06	0.03	0.02	0.02	0.03	0.03	0.02	0.03	0.05	0.03	0.03	0.02	0.03	0.04	0.02	0.03
LYS	0.03	0.03	0.05	0.06	0.06	0.04	0.05	0.04	0.05	0.03	0.03	0.03	0.05	0.03	0.04	0.04	0.04	0.06	0.04	0.03
MET	0.06	0.04	0.06	0.06	0.09	0.06	0.04	0.06	0.08	0.05	0.05	0.05	0.08	0.05	0.06	0.05	0.05	0.10	0.05	0.05
PHE	0.04	0.03	0.05	0.03	0.06	0.03	0.03	0.04	0.05	0.04	0.03	0.03	0.05	0.03	0.04	0.04	0.04	0.05	0.04	0.04
PRO	0.04	0.04	0.05	0.04	0.07	0.05	0.04	0.05	0.06	0.04	0.03	0.04	0.06	0.04	0.04	0.04	0.04	0.07	0.05	0.03
SER	0.05	0.03	0.05	0.04	0.07	0.04	0.04	0.04	0.06	0.03	0.02	0.04	0.05	0.04	0.04	0.03	0.04	0.05	0.04	0.03
THR	0.04	0.04	0.04	0.04	0.06	0.05	0.03	0.04	0.05	0.04	0.03	0.04	0.05	0.04	0.04	0.04	0.03	0.06	0.04	0.03
TRP	0.06	0.04	0.07	0.05	0.08	0.06	0.04	0.06	0.08	0.06	0.04	0.06	0.10	0.05	0.07	0.05	0.06	0.07	0.04	0.05
TYR	0.03	0.02	0.04	0.03	0.05	0.04	0.03	0.04	0.05	0.03	0.02	0.04	0.05	0.04	0.05	0.04	0.04	0.04	0.02	0.03
VAL	0.04	0.03	0.03	0.03	0.04	0.04	0.03	0.03	0.05	0.04	0.03	0.03	0.05	0.04	0.03	0.03	0.03	0.05	0.03	0.03

The Jackknife procedure was used to estimate the standard error for G_{ij}. The procedure was based on 100 randomly chosen half-size subsets of the whole set of 1535 heterocomplex interfaces. The algorithm of error estimation for heterocomplexes was the same as for homodimers (see Table VII).

reflects the prevalence of small (nearly of null area) contacts. The number of casual contacts falls rapidly with the increase of contact area. In contrast, the distribution of specific contacts has a dome-like form, and it is characterized by a certain nonzero mean contact area.

Residue composition of protein components of complexes and interfaces need to be ascertained before performing the analysis of amino acid preference in contacts. We showed the similarity of residue composition between heterocomplex interfaces and homodimer interfaces, including hydrophobic residues, the correlation coefficient being 0.98. However, it was shown in Refs. 4 and10 that homodimer interfaces are enriched in hydrophobic amino acids as compared with heterocomplex interfaces. Probably, this difference is a consequence of different sets of interfaces. As for the difference between amino acid compositions of whole proteins and their interfaces, we observed a slightly lower correlation coefficient of 0.94. The main differences in this case concern Ala, Arg, Gly, Tyr, and Val (Table I, Fig. 3).

From the tables of the interresidue contact preference indices (Tables IV and V) follows an important role of cysteines, oppositely charged residues, and hydrophobic residues in protein–protein interactions.

Cysteines

Disulfide bonds between adjacent cysteine residues are common in the native conformations of proteins. The cysteine residues play an important and substantial role in cofactor binding, intersubunit interactions, DNA binding inhibition, membrane binding, and subcellular localization.[53] The results of our computations demonstrate that interaction between cysteines has the maximal preference indices both for homodimers and heterocomplexes. This observation is especially interesting since the cysteine content in the interfaces is very low (0.3%).

Charged Amino Acids

Electrostatic interactions are known to play an important role in protein–protein recognition and binding. It is believed that these forces are responsible for mutual fitting of complex components at long distances. We obtained maximal preference indices G_{ij} for oppositely charged residues in both homodimers and heterocomplexes. Note that the charged residues (Arg, Glu, Asp, Lys) constitute ~25% of the overall number of residues at protein–protein interfaces. Thus, the data collected confirm the importance of electrostatic interactions, including salt bridges and hydrogen bonds between charged groups.

Hydrophobic Interactions

Hydrophobic residues (Leu, Ile, Val, and Phe) also make ~25% of the overall number of residues at protein–protein interfaces. They have slightly lower preference index values G_{ij} than oppositely charged residues. Surface hydrophobicity at protein–protein interfaces was a special subject of study in Ref. 17.

Among other interesting examples we found high preference index values for such pairs as Trp-Pro in heterocomplexes (1.42) and Pro-Gly in homodimers (1.68). Naturally, low preference index values are characteristic of contacts between residues with similar charges.

An important difference exists between homodimers and heterocomplexes concerning the occurrence of contacts between equivalent residues. Homodimer interfaces were enriched in such contacts owing to symmetry relations, including the twofold axis symmetry.[50] As a result, residues with the same numbers (and, as a consequence, of the same type) are brought closer together in the central zone of interface. An increase in the number of contacts of identical residues is highlighted in the central diagonal of the Table IV (for homodimers). This effect is especially obvious in comparison with the respective diagonal for heterodimers (Table V). Glaser et al.[33] and Saha et al.[48] also pointed out the large number of contacts between residues of the same type. They also found an increased number of contacts for residues with similar charges. In our work, preference indices for such contacts are very low. Evidently, these contacts are energy-poor.

In this work we undertook efforts to obtain accurate and reliable data on contacts between amino acid residues on the surface of protein–protein interfaces both in homodimers and in heterocomplexes. Pairs of residues responsible for specific recognition were identified and residues avoiding one another were detected. The basic differences between homodimers and heterocomplexes were characterized in respect of contact preference. The material collected should contribute to improving of programs for prediction of protein interaction sites.

ACKNOWLEDGMENTS

The authors thank F. Kondrashov and A. Galkin for critical reading of the manuscript, and for their valuable comments.

REFERENCES

1. Chothia C, Janin J. Principles of protein-protein recognition. Nature 1975;256:705–708.
2. Jones S, Thornton JM. Principles of protein-protein interactions. Proc Natl Acad Sci USA 1996;93:13–20.
3. Bahadur RP, Chakrabarti P, Rodier F, Janin J. A dissection of specific and non-specific protein-protein interfaces. J Mol Biol 2004;336:943–955.
4. Chakrabarti P, Janin J. Dissecting protein-protein recognition sites. Proteins 2002;47:334–343.
5. Tsai CJ, Lin SL, Wolfson HJ, Nussinov R. Protein-protein interfaces: architectures and interactions in protein-protein interfaces and in protein cores. Their similarities and differences. Crit Rev Biochem Mol Biol 1996;31:127–152.
6. Valdar WS, Thornton JM. Protein-protein interfaces: analysis of amino acid conservation in homodimers. Proteins 2001;42:108–124.
7. Xu D, Lin SL, Nussinov R. Protein binding versus protein folding: the role of hydrophilic bridges in protein associations. J Mol Biol 1997;265:68–84.

8. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. Nucleic Acids Res 2000;28:235–242.
9. Lo Conte L, Chothia C, Janin J. The atomic structure of protein-protein recognition sites. J Mol Biol 1999;285:2177–2198.
10. Bahadur RP, Chakrabarti P, Rodier F, Janin J. Dissecting subunit interfaces in homodimeric proteins. Proteins 2003;53: 708–719.
11. Ponstingl H, Henrick K, Thornton JM. Discriminating between homodimeric and monomeric proteins in the crystalline state. Proteins 2000;41:47–57.
12. Dasgupta S, Iyer GH, Bryant SH, Lawrence CE, Bell JA. Extent and nature of contacts between protein molecules in crystal lattices and between subunits of protein oligomers. Proteins 1997; 28:494–514.
13. Sternberg MJ, Gabb HA, Jackson RM. Predictive docking of protein-protein and protein-DNA complexes. Curr Opin Struct Biol 1998;8:250–256.
14. Ponder JW, Richards FM. Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes. J Mol Biol 1987;193:775–791.
15. Kim WK, Ison JC. Survey of the geometric association of domain-domain interfaces. Proteins 2005;61:1075–1088.
16. Korn AP, Burnett RM. Distribution and complementarity of hydropathy in multisubunit proteins. Proteins 1991;9:37–55.
17. Young L, Jernigan RL, Covell DG. A role for surface hydrophobicity in protein-protein recognition. Protein Sci 1994;3:717–729.
18. Lijnzaad P, Argos P. Hydrophobic patches on protein subunit interfaces: characteristics and prediction. Proteins 1997;28:333–343.
19. Tsai J, Gerstein M, Levitt M. Simulating the minimum core for hydrophobic collapse in globular proteins. Protein Sci 1997;6: 2606–2616.
20. Hubbard SJ, Argos P. Cavities and packing at protein interfaces. Protein Sci 1994;3:2194–2206.
21. Argos P. An investigation of protein subunit and domain interfaces. Protein Eng 1988;2:101–113.
22. Ponting CP, Russell RR. The natural history of protein domains. Annu Rev Biophys Biomol Struct 2002;31:45–71.
23. Janin J, Miller S, Chothia C. Surface, subunit interfaces and interior of oligomeric proteins. J Mol Biol 1988;204:155–164.
24. Park J, Lappe M, Teichmann SA. Mapping protein family interactions: intramolecular and intermolecular protein family interaction repertoires in the PDB and yeast. J Mol Biol 2001;307: 929–938.
25. Aloy P, Ceulemans H, Stark A, Russell RB. The relationship between sequence and interaction divergence in proteins. J Mol Biol 2003;332:989–998.
26. Keskin O, Tsai CJ, Wolfson H, Nussinov R. A new, structurally nonredundant, diverse data set of protein-protein interfaces and its implications. Protein Sci 2004;13:1043–1055.
27. Tsai CJ, Lin SL, Wolfson HJ, Nussinov R. A dataset of protein-protein interfaces generated with a sequence-order-independent comparison technique. J Mol Biol 1996;260:604–620.
28. Ansari S, Helms V. Statistical analysis of predominantly transient protein-protein interfaces. Proteins 2005;61:344–355.
29. Keskin O, Bahar I, Badretdinov AY, Ptitsyn OB, Jernigan RL. Empirical solvent-mediated potentials hold for both intra-molecular and inter-molecular inter-residue interactions. Protein Sci 1998;7:2578–2586.
30. Bahar I, Jernigan RL. Inter-residue potentials in globular proteins and the dominance of highly specific hydrophilic interactions at close separation. J Mol Biol 1997;266:195–214.
31. Hendlich M, Lackner P, Weitckus S, Floeckner H, Froschauer R, Gottsbacher K, Casari G, Sippl MJ. Identification of native protein folds amongst a large number of incorrect models. The calculation of low energy conformations from potentials of mean force. J Mol Biol 1990;216:167–180.
32. Prlic A, Domingues FS, Sippl MJ. Structure-derived substitution matrices for alignment of distantly related sequences. Protein Eng 2000;13:545–550.
33. Glaser F, Steinberg DM, Vakser IA, Ben-Tal N. Residue frequencies and pairing preferences at protein-protein interfaces. Proteins 2001;43:89–102.
34. Aloy P, Russell RB. Interrogating protein interaction networks through structural biology. Proc Natl Acad Sci USA 2002;99: 5896–5901.
35. Robert CH, Ho PS. Significance of bound water to local chain conformations in protein crystals. Proc Natl Acad Sci USA 1995;92:7600–7604.
36. Richards FM. The interpretation of protein structures: total volume, group volume distributions and packing density. J Mol Biol 1974;82:1–14.
37. Poupon A. Voronoi and Voronoi-related tessellations in studies of protein structure and interaction. Curr Opin Struct Biol 2004; 14:233–241.
38. Soyer A, Chomilier J, Mornon JP, Jullien R, Sadoc JF. Voronoi tessellation reveals the condensed matter character of folded proteins. Phys Rev Lett 2000;85:3532–3535.
39. Angelov B, Sadoc JF, Jullien R, Soyer A, Mornon JP, Chomilier J. Nonatomic solvent-driven Voronoi tessellation of proteins: an open tool to analyze protein folds. Proteins 2002;49:446–456.
40. Gerstein M, Tsai J, Levitt M. The volume of atoms on the protein surface: calculated from simulation, using Voronoi polyhedra. J Mol Biol 1995;249:955–966.
41. Christensen SW, Thomas NW. Structure characterization and predictability by Voronoi analysis. Acta Crystallogr A 1999; 55(Part 5):811–820.
42. Quillin ML, Matthews BW Accurate calculation of the density of proteins. Acta Crystallogr D Biol Crystallogr 2000;56(Part 7): 791–794.
43. Tsai J, Voss N, Gerstein M. Determining the minimum number of types necessary to represent the sizes of protein atoms. Bioinformatics 2001;17:949–956.
44. Medvedev N. Voronoi-Delaune method in noncrystal systems investigations. Novosibirsk: 2000 Siberian Branch of Russian Academy of Sciences (in Russian).
45. Mintz S, Shulman-Peleg A, Wolfson HJ, 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.
46. Pearson WR. Rapid and sensitive sequence comparison with FASTP and FASTA. Methods Enzymol 1990;183:63–98.
47. Jones DT, Taylor WR, Thornton JM. The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci 1992;8:275–282.
48. Saha RP, Bahadur RP, Chakrabarti P. Interresidue contacts in proteins and protein-protein interfaces and their use in characterizing the homodimeric interface. J Proteome Res 2005;4: 1600–1609.
49. Kendall M, Stuart A. Kendall's advanced theory of statistics. Arnold, London Volume 2A Classical Inference and the linear Model. 6 th edition. 1998.
50. Brown JH. Breaking symmetry in protein dimers: designs and functions. Protein Sci 2006;15:1–13.
51. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.
52. Neubauer G. The analysis of multiprotein complexes: the yeast and the human spliceosome as case studies. Methods Enzymol 2005;405:236–263.
53. Swaisgood HE. The importance of disulfide bridging. Biotechnol Adv 2005;23:71–73.