Chinese Science Bulletin 2004 Vol. 49 No. 3 273—276

Statistical analysis on protein-protein interface in crystals: Specific and non-specific interfaces are differentially distributed

FENG Dan & ZENG Zonghao

Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101, China
Correspondence should be addressed to Zeng Zonghao (e-mail: zzh@sun5.ibp.ac.cn)

Abstract **The distribution of contact areas, or fractions of contacting, of protein-protein interfaces in crystals of pure polypeptides contains two components: a major exponential distribution and a minor flatter distribution. Suppose the two components belong to specific and non-specific contacts, respectively, then the probability of a contact with a given area, or fraction of contacting, can be estimated. By dividing the whole database into two sub-databases, one of them is known to contain more specific contacts than the other, this hypothesis is confirmed and it is also proved that the fraction of contacting is more effective than the contact area on discriminating specific and non-specific contacts in protein crystals.**

Keywords: crystal structure, protein-protein interface, statistical analysis, specific interface, protein oligomer.

DOI: 10.1360/03wc0280

Protein-protein interactions are essential for their biological functions. A large amount of crystal structures provide abundant information of protein-protein interactions. Many researchers have analyzed the interfaces in protein crystals[1—15], and attempted to discover biologically relevant contacts according to the features of physical chemistry, stereochemistry and conservation in evolution. But it turns out not to be easy to discriminate crystal packing contacts and biologically relevant contacts, because they adopt the same microscopic mechanism to produce binding forces.

The extent of a contact can be evaluated by contact areas. It is expected that the binding force produced by a contact is directly proportional to contact area, after statistical average. The relation between contact areas and binding forces is probabilistic. In 1998, the standard to judge whether an interface could lead to the formation of an oligomer was proposed to be 4 nm^2 [16], then it was reset to 8 nm^2 [13] in 2000. Obviously, with a higher standard an oligomer protein may be wrongly judged as a monomer protein, on the other hand, with a lower standard a true monomer protein may be judged as an oligomer. The "best" standard would change with the database used. Because of the probabilistic feature of the question, instead of searching for the "best" standard, it is more meaningful to promote and answer the following question: "What is the probability for an interface with a given contact area to be specific?"

This question can be answered by analyzing properly prepared representative database. Janin et al.[12] found that interface areas are exponentially distributed by a statistical analysis on monomer protein crystals. As these crystals are supposed not to contain specific interfaces, it is not reasonable to estimate from these data the probability of an interface with a given area to be specific, although the author did do so in an unscientific way. To carry out studies on this direction forward, we analyzed interfaces in a much wider extent, including all the protein crystals of pure peptides. The restriction to "pure peptides" is to make the chemical components on the interfaces as simple as possible, with the aim of strengthening the statistical expectation on the proportionality between the binding forces and contacting areas. The possibility of using "fraction of contacting" for replacing contact area was also discussed.

1 Materials and methods

1378 crystal structures were selected out from the Protein Data Bank (PDB)[17] at the end of 2001. All these structures have resolutions superior to 0.25 nm, and contain no non-polypeptide components, such as nucleic acids, organic small molecules, metal ions, etc. When the same protein crystallized into isomorphous crystals, the crystal structure with the best resolution was selected.

Polypeptide chains bound through covalent bonds, e.g. disulfide bonds, are regarded as in one monomer. Molecular surface areas were calculated by the method as in the SURFACE program[18]. Contact area was calculated as the total surface area buried in the contact, i.e. the difference of the sum of surface areas of the two partners when isolated and the surface area of the contacting pair. Therefore, a contact area contains contributions from both contacting molecules. Only symmetry un-relevant contacts are considered. Finally, the quantity "fraction of contacting" is calculated as the quotient of the contacting area divided by the surface area of the smaller partner.

2 Results

Altogether 12604 non-symmetry-related interfaces are generated. Statistical distributions of contact areas and fractions of contacting are shown in Fig. 1 (left bottom). The exponential distribution is expressed as

$$p(s) = b\exp(-as) \quad (1)$$

with s as the contact area, or fraction of contacting, and a and b as parameters to be fitted to the experimental data. The corresponding fitting curves and experimental data are shown with lines and dots, respectively. Parameter values obtained from fitting are listed in Table 1. Since

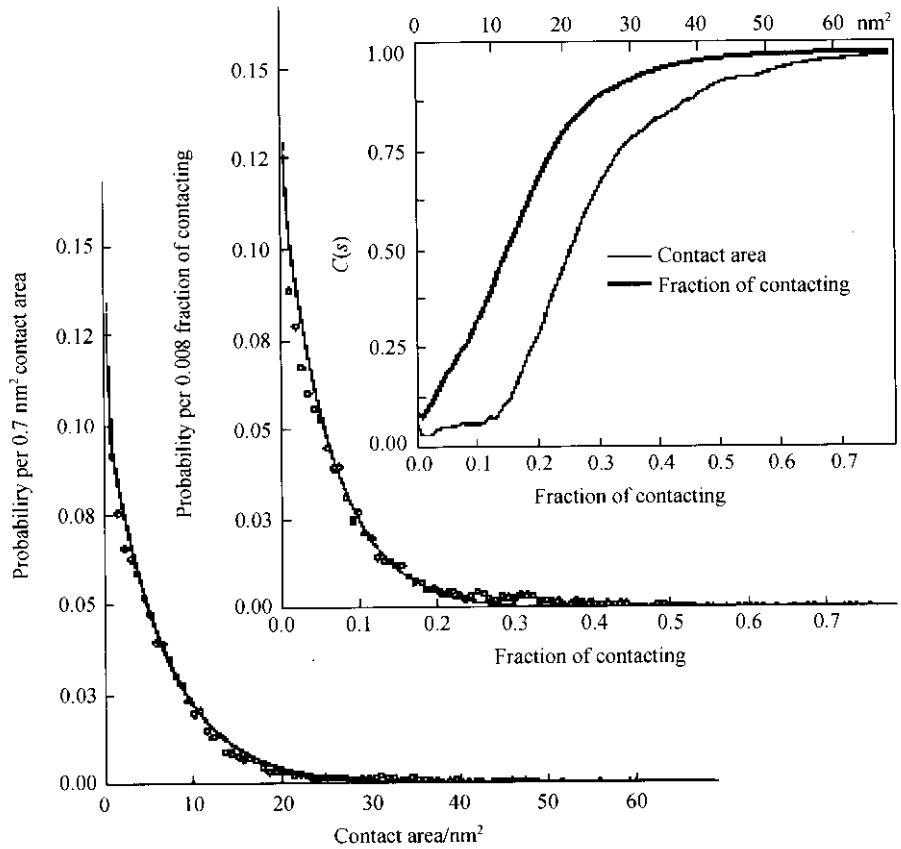

Fig. 1. Probability distribution of contact areas and fractions of contacting in protein crystals. Thin lines are the exponential curves fitting the experimental data (circles). The probability of a contact with an area (or fraction of contacting) more than s and being specific, $C(s)$, is shown at top-right.

there are much more data in the present analysis than before[12], the quality of fitting is greatly improved.

Careful examination of the experiment distribution reveals that there exists a long and flatter tail in both distributions of contact areas and fractions of contacting, respectively. This suggests the existence of another minor and flatter distribution in addition to the major exponential one. If $p_e(s)$ represents for experimental distribution, then the postulated minor distribution can be expressed as $p_e(s) - p(s)$. If the exponential distribution is considered to represent the crystal contacts and the minor and flatter distribution represent the biologically meaningful specific contacts, then

$$C(s) = \int_s^\infty [p_e(s') - p(s')]ds' / \int_s^\infty p_e(s')ds', \qquad (2)$$

that is, the probability of a contact is specific when its contact area, or fraction of contacting, is larger than s. Values of $C(s)$ are shown in Fig. 1 (top-right). Judged by the slope of $C(s)$ curves, the minor distribution peaks are at about 20 nm^2 for contact area (or 0.15 for fraction of contacting).

Relative importance of the two component distributions in the composite distribution can be estimated by b/a and $1 - b/a$, respectively (Table 1). To check these estimations and ascertain the suggestion that the major exponential distribution represent the crystal contacts and the minor flatter distribution represent the specific contacts, the whole dataset is divided into two sub-databases: sub-database A and sub-database B, each of them contains crystals with peptides having different lengths of amino acid sequences, more than or less than 5, respectively. If the argument can be affirmed that, when two different proteins can co-crystallize into one single crystal, specific contacts should exist between them, then there should be more specific contacts in sub-database A and less specific contacts in sub-database B compared with that in the whole database. Statistical analysis on the two sub-databases (Table 1) supports this conclusion. Irrespective of judging with contact areas or fractions of contacting, the

Table 1 Parameter values of three databases

	Database	a	b	b/a	$1-b/a$
Fraction of contacting	whole	0.1345	0.1240	0.9216	0.0784
	B	0.1327	0.1238	0.9325	0.0675
	A	0.1431	0.1249	0.8726	0.1274
Contact area	whole	0.1275	0.1200	0.9415	0.0585
	B	0.1244	0.1182	0.9494	0.0506
	A	0.1430	0.1296	0.9065	0.0935

components of specific contacts (measured by 1- *b/a*) in the three databases have the following order: database A > whole database > database B. In turn, this analysis supports the hypothesis that the exponential distribution comes from crystal contacts and the flatter distribution comes from specific contacts.

3 Discussions and conclusions

The introduction of "fraction of contacting" is provoked by observations on how proteins aggregate. Not all aggregations are formed through monomer-monomer interfaces. Larger aggregations often result from interfaces between smaller aggregations. Larger aggregations may have larger opportunities to form larger interfaces with each other just by accident. Therefore it is not enough by using only contact areas to judge whether smaller aggregations form larger aggregations. A scaled quantity independent of the size of an aggregation is needed. The statistical behavior of this newly introduced quantity, i.e. fractions of contacting, is roughly the same as that of contacting area (Fig. 1), except that the component of specific contacts is a little more when judged by this new quantity than by contact areas (Table 1). Compared to other physicochemical quantities about an interface, contact areas and fraction of contacting are the same type of quantities.

More careful analysis shows that "fraction of contacting" is more closely related with the specificity of a contact. Generally, individual amino acids or short peptides consisting of a few residues added in crystallization solutions are difficult to be found in electronic density maps. But, once they are found out, they must bind with proteins specifically. Because of their small sizes, they can only produce small interfaces with proteins (less than 20 nm^2). When contact areas are used, this kind of binding would be excluded from specific contacts due to their small areas. On the contrary, they all have "fractions of contacting" higher than 0.5. As such, it is easy to understand that "fractions of contacting" are more directly related with specificity in the case of smaller peptides binding with larger peptides. But when both partner proteins are larger, would "fractions of contacting" be advantageous over contact areas on judging contact specificity? To answer this question, we examined the crystal packing of 30 crystals, each crystal has polypeptide of more than 400 residues while the largest "fraction of contacting" is less than 0.1. The largest contact area in each of these crystals is in the range of 8—25 nm^2. A slapdash conclusion from a glance at these area values is that all these interfaces are specific. But, as a matter of fact, they lead to infinite assemblies in 25 out of the 30 crystals. Because proteins in these crystals have no functions like that of microtubules, these infinite assemblies should have no biological meaning. Therefore, in the 30 crystals only a very small number of contacts might be specific. The conclusion is that "fractions of contacting" are more reliable than "contact areas" even for larger proteins.

To make a systematic study, a statistics on the number of crystals, in each of them the largest contact areas (or fractions of contacting) fell into given ranges (Table 2), was carried out. If contact areas are used, the peaks of the crystal number distributions for all the three databases are all at 10—20 nm^2; while fractions of contacting are used, the peaks for the whole database and sub-database B are both at 0.1—0.2, the peak for sub-database A is in the range of 0.2—0.3 and extends to 0.3—0.4. This result correctly reflects the fact that more specific contacts exist in sub-database A. In a word, based on all these facts we are in more confidence to say that "fractions of contacting" have more power than "contact areas" on discriminating specific and non-specific contacts.

Although the author of ref. [12] used the term "non-specificity" instead of "random" which was used in his early works, the exponential distribution of contact areas (or fractions of contacting) does reveal the apparent randomness in the process of crystal packing. This is the feature that obviously resulted from the varieties of shapes and surfaces of proteins, and can only be observed when statistics is performed over the whole database. In any given crystal, molecule packing is highly ordered. The phenomenon that almost all parts of a protein surface can be used in some crystal, like that happened for pancreases ribonuclease[3], cannot be observed for all kinds of proteins. In the most general sense, randomness and specificity are mutually excluding and overlapping properties and a phenomenon can be both random and specific in different senses. For the problem discussed in this work, it is still appropriate to use them to refer to the two differently distributed contacts respectively. The difference of the two

Table 2 Number of crystals, in each of them the largest contact area (or fractions of contacting) is in a given range (area unit: nm^2)

Contact area	0—10	10—20	20—30	30—40	40—50	> 50	Total
Whole	347	837	292	194	70	77	1817
A	7	91	35	61	15	8	217
B	340	746	257	133	55	69	1600
Fraction of contacting	0—0.1	0.1—0.2	0.2—0.3	0.3—0.4	0.4—0.5	> 0.5	Total
Whole	315	691	385	207	115	104	1817
A	19	17	57	54	34	36	217
B	296	674	328	153	81	68	1600

kinds of contacts on distribution provides us the possibility to differentiate specific and non-specific contacts. For a given contact area (or fraction of contacting) s, $C(s)$ is the probability that an interface with contact area (or fraction of contacting) larger than s is specific.

In the region where the two distributions overlap, it is difficult to judge only by contact areas, or fractions of contacting, whether an interface is specific or not. Moreover, to a large extent, based on our present knowledge it is still a subjective matter to judge whether an interface is biologically relevant. For instance, crystals of pancreases ribonuclease contain a dimer-like interface with contact area of 18 nm^2. The dimmer-like interface was considered as an artifact[3] of crystal packing because it happened to cover the catalytic site of the enzyme. But it is possible that, as our biological-chemistry knowledge on cells increases, it may be found that the dimer-like interface regulates the enzyme activity: when the ribonucleases are overexpressed, they cover their catalytic site to decrease the activity. Furthermore, evolution does not always tend to increase binding force. Cells require proteins to aggregate or dissociate at the right time. In addition, different oligomer states of a protein may all be biologically relevant, just different oligamer states playing different functions. In such a case, it is necessary to consider more factors for reference, such as the symmetry rules governing interfaces in crystals, the physico-chemistry properties of interfaces, the frequencies of an interface happening in different crystals, the conservation of an interface in evolution, etc. The clarification of the distribution of interfaces in crystals promised the further studies.

References

1. Janin, J., Miller, S., Chothia, C., Surface, subunit interfaces and interior of oligomeric proteins, J. Mol. Biol., 1988, 204: 155—164.
2. Janin, J., Chothia, C., The structure of protein-protein recognition sites, J. Biol. Chem., 1990, 265: 16027—16030.
3. Crosio, M. P., Janin, J., Jullien, M., Crystal packing in six crystal forms of pancreatic ribonuclease, J. Mol. Biol., 1992, 228: 243—251.
4. Janin, J., Rodier, F., Protein-protein interaction at crystal contacts, Proteins, 1995, 23: 580—587.
5. Jones, S., Thornton, J. M., Protein-protein interactions: a review of protein dimmer structures, Prog. Biophys. Biol., 1995, 63: 31—65.
6. Jones, S., Thornton, J. M., Principles of protein-protein interactions, Proc. Natl. Acad. Sci. USA, 1996, 93: 13—20.
7. Jones, S., Thornton, J. M., Analysis of protein-protein interaction sites using surface patches, J. Mol. Biol., 1997, 272: 121—132.
8. Jones, S., Thornton, J. M., Prediction of protein-protein interaction sites using surface patches, J. Mol. Biol., 1997, 272: 133—143.
9. Lijnzaad, P., Argos , P., Hydrophobic patches on the surfaces of protein structures, Proteins, 1997, 28: 333—343.
10. Tsai, C. J., Lin, S. L., Wolfson, H. J. et al., Studies of protein-protein interfaces: A statistical analysis of the hydrophobic effect, Protein Sci., 1997, 6: 63—64.
11. Carugo, O., Argos, P., Protein-protein crystal-packing contacts, Protein Sci., 1997, 6: 2261—2263.
12. Janin, J., Specific versus non-specific contacts in protein crystals, Nat. struct. Biol., 1997, 4: 973—974.
13. Ponstingl, H., Henrick, K., Thornton, J. M., Discriminating between homodimeric and monomeric proteins in the crystalline state, Proteins, 2000, 41: 47—57.
14. Jones, S., Marin, A., Thornton, J. M., Protein domain interfaces: characterization and comparison with oligomeric protein interfaces, Protein Eng., 2000, 41: 77—82.
15. Elcock, A. H., McCammon, J. A., Identification of protein oligomerization states by analysis of interface conservation, Proc. Natl. Acad. Sci. USA, 2001, 98: 2990—2994.
16. Henrick, K., Thornton, J. M., PQS: a protein quaternary structure file server, Trends Biochem. Sci., 1998, 23: 358—361.
17. Berman, H. M., Westbrook, J., Feng, Z. et al., The protein databank, Nucleic Acids Res., 2000, 28: 235—242.
18. Collaborative Computational Project, Number 4. The CCP4 Suite: Programs for protein crystallography, Acta Cryst., 1994, D50: 760—763.

(Received June 4, 2003; accepted December 11, 2003)