Protein and Peptide Letters, Vol. 11, No. 4, pp. 361-366, 2004
Bentham Science Publishers Ltd.
0929-8665/04 $ 45.00 + .00

SPECIFIC AND NON-SPECIFIC CONTACTS IN PROTEIN CRYSTALS

Feng Dan* and Zeng Zong-Hao

Center of Molecular Biology, Institute of Biophysics, Chinese Academy of Sciences
15 Datun Road, Chaoyang District, Beijing 100101, China; fengdan_bj@yahoo.com

Abstract: Statistical analysis of protein-protein interfaces in a database of pure peptide crystals shows that the distribution of the contact area contains two components: a major exponential distribution and a minor flat distribution. Analysis of two sub-databases provides evidence that the two components represent specific and non-specific contacts, respectively. The probability of an interface with a given area being specific can be estimated. A scaled quantity (contact ratio) is introduced that is more useful than contact area for discriminating specific and non-specific contacts in protein crystals.

Keywords: crystal structure, interface, statistical analysis, specificity of interface, oligomer protein.

INTRODUCTION

Protein-protein interaction is essential for subsequent biological function. The large number of crystal structures in Brookhaven Protein Data Bank (PDB) [17] provides abundant information regarding molecular interaction. Many researchers have studied interfaces in protein crystals [1-15], and much effort has been expended to developing new methods to discriminate between crystal packing and a functional protein-protein interaction. This is difficult because packing contacts make use of the same forces that govern specific recognition in protein-protein complexes and oligomeric proteins.

Contact area has been commonly used to help crystallographers to classify the found protein-protein interactions into either crystal packing or a likely quaternary biological assembly. A contact area cut-off of 4 nm^2 [16] was proposed in 1998 and later improved to 8 nm^2 in 2000 [13]. Obviously, the calculated “best” standard would change with the database selected, and also it was just an empirical value. Instead of searching for the “best” standard, it is more meaningful to propose and answer following question: “What is the probability for a given area interface to be specific?”

After analysing 152 monomeric protein crystals, Janin [12] found that contact areas reveal an exponential distribution. He also gave an equation to estimate the probability of a given area interface being **non**-specific. In this work, we analyzed 1,817 pure peptide crystals and found that there was deviation of the experimental distribution from an exponential character. The deviation was not pronounced but was quite persistent. In addition, the feasibility of assessing specificity *via* “contact ratio”, instead of “contact area”, was also discussed.

MATERIALS AND METHODS

Non-redundant 1,817 crystal structure data were selected from Protein Data Bank (PDB) [17] in August 2003. All these data met the following criteria: each was determined by X-ray diffraction method with a resolution better than 2.5Å; did not containing non-polypeptide compositions such as nucleic acid, small organic molecule, metal ion, etc; and did not contain polypeptide chains with homology exceeding 70 and with length less than 20 residues. The data set with the best resolution was selected as representative of existing isomorphous crystals.

Molecule surface area was calculated by SURFACE program [18]. Contact area of two molecules is the area of the protein surface that becomes buried in contacts between molecules. Only non-symmetrical relevant contacts were considered. We introduced another quantity, “contact ratio”, to measure the extent of a contact. Contact ratio is evaluated as the contact area divided by the smaller surface area of the two partners.

RESULTS

Altogether 16,455 non-symmetrical relevant interfaces were generated. Statistical distributions of contact area and contact ratio are shown with dotted curves in the left bottom of Figure **1**. Experimental data is fitted by exponential function. The expression equation of exponential function is as follows:

$$P(s)=b\ \exp(´as),$$

where s is contact area (or contact ratio), parameters a and b are needed to determine. The corresponding fitting curves are showed with thin lines. Parameter values of fitting lines are listed in Table **1**. From the figure, one can find that the experimental distribution of contact area fits with the exponential curve much better than that found in a much smaller database [12], especially in lower contact area range. The exponential distribution of contact area does reflect the apparent randomness in the process of crystal packing. But, if checked in detail, mismatches are obvious with the increasing of contact area. The deviation is more readily found in the case of contact ratio. We suggest that there exists another minor and flat distribution in addition to the major exponential one.

Table 1. Parameters value of three databases

	Database	*a*	*b*	*b*/*a*	1-*b*/*a*
Contact Area	Whole	0.1201	0.1086	0.9042	0.0958
	A	0.1430	0.1217	0.8507	0.1493
	B	0.1244	0.1143	0.9192	0.0807
Contact Ratio	Whole	0.1229	0.1101	0.8964	0.1036
	A	0.1431	0.1190	0.8317	0.1683
	B	0.1327	0.1209	0.9108	0.0892

If the exponential distribution represents the crystal contacts, then the minor and flat distribution represent the biologically meaningful specific contacts. The relative abundance of the two component distributions in the composite distribution can be estimated by *b*/*a* and 1-*b/a* respectively (Table **1**).

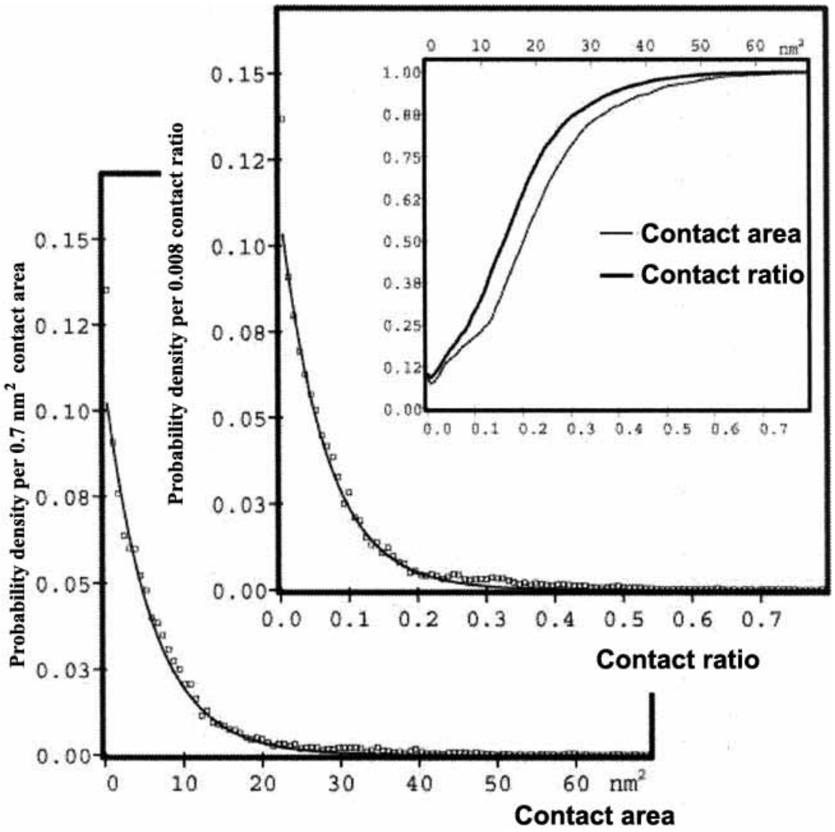

Figure 1. Histogram of contact area (and contact ratio) found in protein crystals. Thin lines are the fitted exponential curves. The probability of a contact with an area (or contact ratio) larger than s, *C*(*s*), are shown at top-right.

To check these estimations and ascertain the suggestion that the major exponential distribution represent the crystal contacts and the minor flat distribution represent the specific contacts, the whole database was divided into two sub-databases: sub-database A and sub-database B. Those crystals containing heterologous polypeptide chains are included in sub-database A. All other crystals are in sub-database B. It may be argued that when two heterologous polypeptide chains can co-crystallize into one single crystal, specific contacts should exist between them. Consequently, there should be more specific contacts in sub-database A and less specific contacts in sub-database B compared with those in the whole database. Statistical analysis of the two sub-databases supports this conclusion. Regardless of whether judged by contact area or contact ratio, the components of specific contacts (measured by 1-*b/a*) in the three databases have the following order: Database A > whole database > database B (Table **1**). This analysis supports the hypothesis that the exponential distribution comes from crystal contacts and the flat distribution comes from specific contacts.

Because of the relatively small amount of specific contacts, it is difficult to fit the flat distribution alone with some theoretical curve. But the postulated minor distribution can be expressed as $P_e(s)$-$P(s)$, where $P_e(s)$ represents for experimental distribution. Then

$$C(s) = \int [P_e(s') - P(s')]\, ds' / \int P_e(s')\, ds', \qquad (1)$$

is the probability of finding an interface with contact area (or contact ratio) larger than *s* being specific. Values of $C(s)$ are showed in Figure1 top-right. Judging by the slope of $C(s)$ curves, the minor distribution peaks at about 18 nm^2 for contact area (or 0.16 for contact ratio).

DISCUSSION AND CONCLUSIONS

There are three types of specific interaction in crystals. The first two types are well known and easily understood. They exist in specific complexes and oligomeric proteins respectively. The third type of specific interaction refers to those between subunits of “crystal oligomer”. “Crystal oligomers” differ from true oligomeric proteins. They can only exist in the special environment of crystals. But in recent years, increasing experimental evidence supports the view that “crystal oligomers” can occur in pre-crystallization solution and are likely intermediates in crystallizations. This suggests that there is a degree of specificity in that particular contact.

A previous study by Janin [12] used a crystal database that consisted of only monomeric proteins. Specific interfaces in that dataset are those between subunits of “crystal oligomer”. So there were only very few examples of specific interfaces. Compared to the monomeric proteins database, the pure peptide database in this work includes all types of specific interfaces. The greater amount of specific interfaces (about 10% of whole interfaces) provided the opportunity to find out the deviation of the experimental distribution from an exponential character. This suggests that specific interactions do not fit to exponential distribution. In the other hand, it is more reasonable to judge the probability of an interface to be specific with this dataset. For a given contact area (or contact ratio) s, $C(s)$ is the probability that an interface with contact area (or contact ratio) larger than s and be specific. For an interface larger than 18nm^2 or 0.16 contact ratio, the chance to be specific is larger than 50%.

The introduction of “contact ratio” is based on observations on how proteins aggregate. Not all aggregates are formed through monomer-monomer interfaces. Larger aggregates are often formed through interfaces between smaller aggregates. For large aggregates, there is more chance to form large interfaces between them just by randomness. Therefore it is not enough using only contact areas to judge whether smaller aggregates can assemble into large aggregates. A scaled quantity independent of the size of an aggregate is needed. The statistical behaviors of the newly introduce quantity is roughly the same with that of contact area (Figure **1**), except that the component of specific contacts is a little more when judged by this new quantity than that judged by contact area (Table **1**).

Contact ratio is a more reasonable quantity because it 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 found, they must bind with proteins specifically. Because of their small sizes, they can only produce small interfaces with proteins (less than 4 nm^2). This kind of binding is unlikely to be specific if using “contact area” as measurement. But they all have “contact ratio” higher than 0.5. As such, it is easy to understand that “contact ratio” is more directly related with specificity in the case of small peptides binding with large peptides.

In another case, when both partner proteins are large, “contact ratio” still has an advantage over contact areas for judging contact specificity. We examined the crystal packing of 30 crystals; in each of these crystals all polypeptides are longer than 400 residues. The largest contact area in each of these crystals is in the range of 8~25 nm^2, while the largest “contact ratio” is less than 0.1. A slapdash conclusion from a glance at these contact area values is that all these interfaces are very likely to be 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 “contact ratio” is more reliable than “contact area” even for large proteins.

In the region where the two distributions overlap, it is difficult to judge whether an interface is specific or not only by contact area or contact ratio. Moreover, to a large extent, based on our present knowledge it is still a subjective matter to assess whether an interface is biologically relevant. For instance, a crystal of pancreases ribonuclease contains a dimer-like interface with contact area of 18 nm^2. The dimer-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, as our biological chemistry knowledge on cells increases, to find that the dimer-like interface regulates the enzyme activity: when the ribonucleases are over-expressed, they cover their catalytic site to decrease activity. After all, evolution does not always tend to increase binding force. Cells require proteins to aggregate or disaggregate at the right time. In such a case, it is necessary to consider more factors for reference, such as symmetry rules governing interfaces in crystals, physical chemistry properties of interface, frequencies of an interface happening in different crystals, evolutional conservation of interface, etc. Further work will clarify the distribution of interfaces in crystals.

REFERENCES

[1] Janin, J., Miller, S. and Chothia, C. J. (**1988**) *Mol. Biol., 204*, 155-164.
[2] Janin, J. and Chothia, C. (**1990**) *J. Biol. Chem., 265*, 16027-16030.
[3] Crosio, M.P., Janin, J. and Jullien, M. (**1992**) *J. Mol. Biol., 228*, 243-251.
[4] Janin, J. and Rodier, F. (**1995**) *Proteins, 23*, 580-587.
[5] Jones, S. and Thornton, J. M. (**1995**) *Prog. Biophys. Mol. Biol., 63*, 31-65.
[6] Jones, S. and Thornton, J. M. (**1996**) *Proc. Natl. Acad. Sci. USA, 93*, 13-20.
[7] Jones, S. and Thornton, J. M. (**1997**) *J. Mol. Biol., 272*, 121-132.
[8] Jones, S. and Thornton, J. M. (**1997**) *J. Mol. Biol., 272*, 133-143.
[9] Lijnzaad, P. and Argos, P. (**1997**) *Proteins, 28*, 333-343.
[10] Tsai, C. J., Lin, S. L., Wolfson, H. J. and Nussinov, R. (**1997**) *Protein Sci., 6*, 63-64.
[11] Carugo, O. and Argos, P. (**1997**) *Protein Sci., 6*, 2261-2263.

[12] Janin, J. (**1997**) *Nature Struct. Biol.*, *4*, 973-974.
[13] Ponstingl, H., Henrick, K. and Thornton, J. M. (**2000**) *Proteins*, *41*, 47-57.
[14] Jones, S. Marin, A. and Thornton, J. M. (**2000**) *Protein Eng.*, *13*, 77-82.
[15] Elcock, A. H. and McCammon, J.A. (**2001**) *Proc. Natl. Acad. Sci. USA*, *98*, 2990-2994.
[16] Henrick, K and Thornton, J. M. (**1998**) *Trends Biochem. Sci.*, *23*, 358-361
[17] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. and Bourne, P. E. (**2000**) *Nucleic Acids Res.*, *28*, 235-242.
[18] Collaborative Computational Project, Number 4. (**1994**) *Acta Cryst.*, *D50*, 760-763.

Received on December 16, 2003, accepted on February 25, 2004.