Journal of Biomolecular Structure & Dynamics ISSN 0739-1102
Volume 28, Issue Number 4, February 2011

Comment

Short-Range Contact Preferences and Long-Range Indifference: Is Protein Folding Stoichiometry Driven?

http://www.jbsdonline.com

Hue Sun Chan

Departments of Biochemistry, Molecular Genetics, and of Physics, University of Toronto, Toronto, Ontario M5S 1A8, Canada

Mittal *et al*. (1) recently advanced an unconventional view on protein folding. By analyzing the spatial neighborhoods of amino acid residues in an extensive set of structures in the Protein Data Bank (PDB), the authors concluded that preferential interactions between amino acid residues do not drive protein folding. In this connection, it should be noted that preferential interactions between amino acids are the basis for introducing knowledge-based potentials, which in turn provide the underpinning for present day three-dimensional protein structure prediction by modeling and simulation (2-5 and references therein). Instead of these preferential interactions, Mittal *et al.* indicate that "protein folding is a direct consequence of a narrow band of stoichiometric occurrences of amino-acids in the primary sequences" (1). According to the authors, this observation is akin to Chargaff's discovery that the molar ratios of adenine and thymine and that of guanine and cytosine in DNA were not far from unity (6).

This assertion runs counter to prevalent views, most notably the decades-old consensus that hydrophobic interactions is a major driving force for folding (7, 8). The view of Mittal *et al*. is counterintuitive because folded proteins do have a "hydrophobic inside, polar outside" organization; the average buried area (not exposed to solvent) of an amino acid residue in folded proteins correlates with its hydrophobicity (9). The authors' conclusion is all the more puzzling in light of established statistical potentials derived from the PDB that clearly demonstrate preferences in contacts among amino acids (10–12). A major contribution to those preferences is none other than the hydrophobic effect (13).

The conclusion of Mittal *et al*. was based on enumerating the spatial distribution of pairs of Cα positions among PDB structures. For each of the 20 × 20 pairs of the twenty types of amino acids, they obtained the number of residue pairs (termed "contacts") within a variable distance from each other (the residues were referred to as "neighbors" regardless of distance), and fitted the distance dependence of the number of such contacts to a particular sigmoidal-shaped function. They found that the fitted sigmoidal trends were similar for all 20 × 20 types of neighbors, and that asymptotically (at large distances) the number of contacts of an amino acid type is proportional to its overall composition in the PDB structures considered. They interpreted the results of this "neighborhood analysis" of theirs (1) as implying a lack of preferential interactions. Mittal *et al*. did not address the inconsistency of their conclusion with established statistical potentials. But this contradiction is significant because it should not have arisen. After all, the authors' results and the statistical potentials were both derived from the PDB.

Corresponding Author:
Hue Sun Chan
Phone: (416)978-2697
Fax: (416)978-8548
E-mail: chan@arrhenius.med.toronto.edu

Is Mittal *et al.'s* assertion warranted by the analysis they presented? To answer this question, it is instructive to perform a neighborhood analysis on the hydrophobic-polar (HP) model (Figure 1). Folded structures of short HP sequences configured on the two-dimensional square lattice have ratios of inside and outside residues similar to those of real proteins (14). The only favorable interaction energy in the HP model is that between a pair of H residues that are not next to each other along the chain sequence but are spatial nearest neighbors on the lattice. Although this simple potential does not provide a full account of protein energetics (15), it captures important features of the sequence to structure mapping of real proteins (16), and thus is a valuable tool for studying molecular evolution (17). The HP model has preferential interactions by construction. Regardless of the model's ability, or lack thereof, to rationalize real protein properties, we may use it to evaluate the interpretive logic of Mittal *et al*. by asking whether the folded structures in the model exhibit neighborhood properties similar to those obtained by the authors. If the answer is affirmative, it would indicate that the results presented by Mittal *et al*. do not necessarily imply that preferential interactions do not drive folding of real proteins.

In Figure 1, the behavior exhibited in (A) for a single HP sequence with $n = 25$ residues (18) are similar to that in (C) for more than six thousand $n = 18$ HP sequences (17). Both

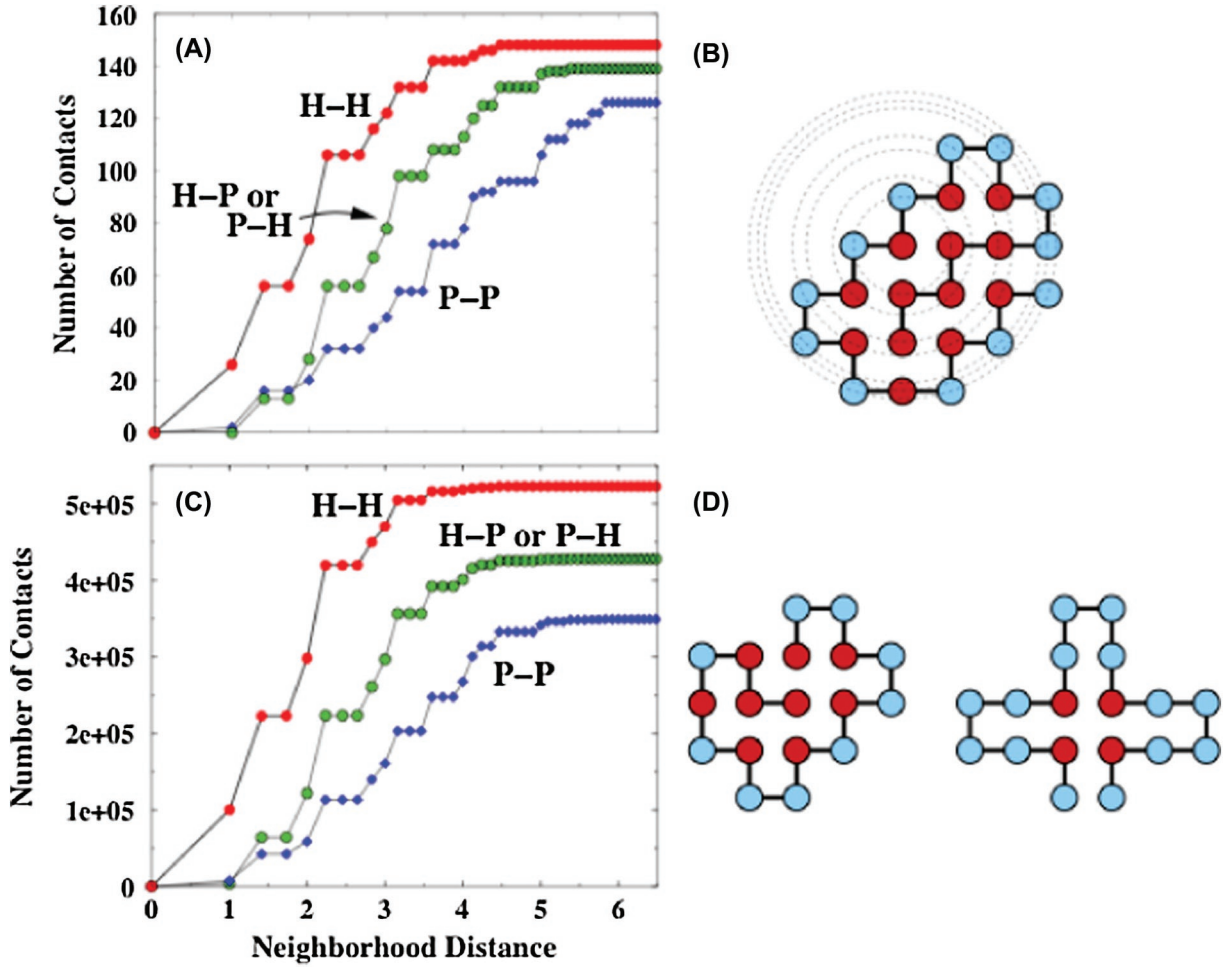

Figure 1: Neighborhood analysis in the two-dimensional HP model. (**A**) Following the terminology of Mittal *et al*. (1), the "number of contacts" (vertical axis) is the number of residue positions within a given distance (horizontal axis, in unit of lattice bond length) from a residue of a given type (H or P). Results in (**A**) are for the HP sequence and structure in (**B**). The curve labeled "H–H" (red circles) shows the sum of numbers of H residues in the neighborhood of each H residue; the curve labeled "H–P or P–H" (green circles) shows the sum of numbers of P residues in the neighborhood of each H residue, and vice versa; similarly, the curve labeled "P–P" (blue diamonds) shows the sum of numbers of P residues in the neighborhood of each P residue. (**B**) The HP sequence studied in (**A**) is one of 325 HP sequences determined by Irbäck and Troein to encode uniquely for the structure shown (18). H and P residues are drawn as red and blue beads, respectively. The concentric dotted circles illustrate the neighborhoods of a residue. Results in (**C**) are for 6,349 18-residue HP sequences that encode uniquely, with each sequence contributing equally to the data plotted. A total of 1,475 different native structures are encoded by these sequences (17). For this set of 6,349 sequences, the overall P/H ratio of fractional occurrence is equal to 51,602/62,680 = 0.8233. The corresponding ratio for the total number of contacts with P versus that with H is equal to 776,644/950,284 = 0.8173. (**D**) Two examples among the 6,349 sequences studied in (**C**) are depicted in their respective native structures.

show a sigmoidal trend similar to that observed by Mittal *et al*. This behavior is not surprising because the number of contacts of any residue must saturate for large neighborhood distances (denote as r below) if the sizes of the folded structures are finite (as in the model and for real proteins). For smaller r values, the number of contacts should be roughly proportional to the available volume of the neighborhood, meaning that it should increase approximately as $(r - r_{ex})^2$ in two dimensions and $(r - r_{ex})^3$ in three dimensions, where r_{ex} is a threshold r value below which contacts are impossible because of excluded volume. (In Mtital *et al*. (1), the number of contacts ~ r^4 for small r; a larger exponent of ≈ 4 instead of 3 is apparently needed in their formulation to compensate for the effect of r_{ex}.) For a structure with chain length n, the total number of contacts as defined by Mittal *et al*. is $n - 3$ for every residue not at the chain ends and $n - 2$ for the two terminal residues. Thus, aside from a small chain-end correction, the total number of contacts so defined for a residue type is necessarily proportional to its fractional occurrence. This is illustrated by the structure in Figure 1B. It has a P/H residue ratio of 12/13 = 0.9231, which is almost identical to the corresponding ratio of 265/287 = 0.9233 for the total number of contacts. In general, for a collection of structures (labeled by i) with chain lengths n_i and compositions $\phi^a{}_i$ for any amino acid type a, the total fractional occurrence of the amino acid type a is, by definition, $\sum_i n_i \phi^a{}_i / \sum_i n_i$ and the total number of contacts of a is essentially $\sum_i n_i (n_i - 3) \phi^a{}_i$. It follows that an approximate proportionality relationship between the overall fractional occurrence and the total number of contacts of an amino acid type as observed by Mittal *et al*. is expected if n_i is a constant (as in Figure 1C), or if $\phi^a{}_i$ varies little with n_i — which is apparently the case for real proteins.

In Figure 1C, the overall sigmoidal shapes for the H and P contacts are similar. Yet a P residue is on average 3.65 times more exposed than an H residue in this set of structures. Therefore, similarity of the overall sigmoidal fits for different residues does not necessarily imply a lack of preferential interactions. Because the HP model interactions have a short spatial range, the differences between H and P contacts are apparent for small neighborhood distances but the differences are less conspicuous if one takes a panoramic view and assigns equal significance to "contacts" at all neighborhood distances when fitting the data to sigmoidal functions. Mittal *et al*. noted deviations from their overall fits at small neighborhood distances but dismissed the deviations as "only noise in the data" (1). However, if most interactions among real amino acids have short spatial ranges, behaviors at small neighborhood distances should be regarded as key signals for the underlying physics, not noise in the data. Mittal *et al*. did not identify the amino acid types of the data points for small neighborhood distances in their Figure 3A–D. If provided, this information would help resolve the contradiction between the authors' conclusion and the preferential interactions underscored by statistical potentials (10–12).

Although the conclusion of Mittal *et al*. is not supported by the evidence presented thus far, the authors' suggestion of a near-universal amino acid composition among globular proteins is thought-provoking and deserves further investigation. If validated, it would be extremely interesting to relate this organization principle to the study of evolution of the genetic code (19) as well as theoretical perspectives that emphasize interaction heterogeneity (13, 20) as a critical requirement for efficiency (20) and cooperativity (15) of protein folding. In this as in any scientific endeavor, it is prudent to heed Chargaff's timeless advice: "generalizations in science are both necessary and hazardous; they carry a semblance of finality which conceals their essentially provisional character" (6).

References

1. A. Mittal, B. Jayaram, S. Shenoy, and T. S. Bawa. *J Biomol Struct Dyn 28*, 133-142 (2010).
2. P. Sklenovský and M. Otyepka. *J Biomol Struct Dyn 27*, 521-539 (2010).
3. M. J. Aman, H. Karauzum, M. G. Bowden, and T. L. Nguyen. *J Biomol Struct Dyn 28*, 1-12 (2010).
4. C. Koshy, M. Parthiban, and R. Sowdhamini. *J Biomol Struct Dyn 28*, 71-83 (2010).
5. Y. Tao, Z. H. Rao, and S. Q. Liu. *J Biomol Struct Dyn 28*, 143-157 (2010).
6. E. Chargaff. *Experientia 6*, 201-209 (1950).
7. W. Kauzmann. *Adv Protein Chem 14*, 1-63 (1959).
8. C. Tanford. *Adv Protein Chem 23*, 121-282 (1968).
9. G. D. Rose, A. R. Geselowitz, G. J. Lesser, R. H. Lee, and M. H. Zehfus. *Science 229*, 834-838 (1985).
10. S. Tanaka and H. A. Scheraga. *Macromolecules 9*, 954-950 (1976).
11. S. Miyazawa and R. L. Jernigan. *Macromolecules 18*, 534-552 (1985).
12. M.-Y. Shen and A. Sali. *Protein Sci 15*, 2507-2524 (2006).
13. H. S. Chan. *Nature Struct Biol 6*, 994-996 (1999).
14. K. A. Dill, S. Bromberg, K. Yue, K. M. Fiebig, D. P. Yee, P. D. Thomas, and H. S. Chan. *Protein Sci 4*, 561-602 (1995).
15. H. S. Chan. *Proteins Struct Funct Genet 40*, 543-571 (2000).
16. A. Irbäck and E. Sandelin. *Biophys J 79*, 2252-2258 (2000).
17. Y. Cui, W. H. Wong, E. Bornberg-Bauer, and H. S. Chan. *Proc Natl Acad Sci USA 99*, 809-814 (2002).
18. A. Irbäck and C. Troein. *J Biol Phys 28*, 1-15 (2002).
19. J. T. F. Wong. *Proc Natl Acad Sci USA 72*, 1909-1912 (1975).
20. P. G. Wolynes. *Nature Struct Biol 4*, 871-874 (1997).