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

Author Response

The Newest View on Protein Folding: Stoichiometric and Spatial Unity in Structural and Functional Diversity

http://www.jbsdonline.com

Aditya Mittal[1#*]
B. Jayaram[1,2#*]

[1]School of Biological Sciences, Indian Institute of Technology Delhi, New Delhi, India
[2]Department of Chemistry and Supercomputing Facility for Bioinformatics & Computational Biology, Indian Institute of Technology Delhi, New Delhi, India

[#]Equal contribution.

Recently we proposed (1):

1. There is a universal spatial distribution for the backbones of folded proteins, regardless of their size, shape and sequence.
2. This universality appears to primarily arise out of stoichiometric (relative frequencies of occurrences of amino acids) margins of life that dictate the neighborhoods of individual amino acids in folded proteins.
3. These neighborhoods defy the conventional views on "preferential interactions" stabilizing folded protein structures.
4. The apparent "preferential interactions" that have formed the current view on protein folding are *post-facto* inferences rather than drivers of protein folding.

Several investigators have carefully and critically examined the above findings (2-30), especially in terms of a very large body of literature on calculated propensities of different amino acids for different environments. It is very encouraging that none of the investigators disagree with our results. In our opinion, an objective, weighed, and neutral articulation of our work is most elegantly put forward by Berendsen (5).

However, there is a clear polarization of opinion on our proposals, with skepticism from the critics regarding the applicability of our methodology to understanding of protein folding. On one hand, several comments agree, to varying degrees, with our findings on the stoichiometric margins of life (4, 8, 9, 10, 11, 12, 13, 14, 17, 18, 21, 23, 24, 25, 30), depending on (a) whether C_α-C_α neighborhoods can be considered informative, and, (b) even if stoichiometric margins of life are considered necessary for protein folding, they are not sufficient. Some investigators agree with our findings in general, including the fascinating universal spatial distribution discovered by us (5, 20, 22, 23, 29), that may provide a lead into the "sufficient" condition(s) required for protein folding. On the other hand several comments on our work are either quite skeptical or critical (2, 3, 4, 6, 7, 15, 16, 19, 21, 25, 26, 27, 28), based on the large body of literature that has evolved sophisticated formalisms using knowledge-based potentials towards establishing a mechanistic view on protein folding. Interestingly, in the apparent debate on "for" (the former group) and "against" (the latter group) our proposals, several arguments provided by "for" comments answer questions raised by the "against" comments sufficiently. For example, it is remarkable that the major issues raised by Matthews (2) are directly answered experimentally by Song *et al*. (29).

[*]Corresponding Authors:
Aditya Mittal
B. Jayaram
Phone: +91-11-26591052
+91-11-26591505
Fax: 091-11-26582037
E-mail: amittal@bioschool.iitd.ac.in
bjayaram@chemistry.iitd.ac.in

In this report, we address several issues regarding our proposals to enable a clearer and objective emergence of the "newest" view on protein folding. For doing so we first address some points regarding our methodology:

1. The dataset of 3718 crystal structures of proteins was randomly collected with the following constraints – (a) 2.5 Å or better structural resolution, (b) Structural data of only A-chains was considered to understand folding of single polypeptide chains, and (c) only soluble proteins were considered.
2. We specifically exclude immediate neighbors along the sequence.
3. In considering C_α-C_α neighborhoods, neither do we consider that the backbone carbon atoms interact with each other, nor do we suggest any such possibility. Our premise is that spatial organization of C_α-pairs of amino acids whose side chains interact would be distinct from the spatial organization of C_α-pairs of amino acids whose side chains do not interact. Number of C_α-pairs are termed as number of contacts within a defined distance.
4. Instead of arriving at stoichiometric margins of life by simply compiling statistics of protein sequences, we arrive at these margins through a surprising route of discovering the universal spatial distributions. Thus, while the end result appears to be "trivial", the path taken towards the discovery is certainly not. This in fact, resembles several classical examples in mathematics, physics and chemistry where an apparently intriguing path has yielded rather simple solutions to problems.

Now, one of the major issues in several comments on our work is the understanding of the sigmoidal universal spatial distributions, barring a few investigators (5, 20, 22, 23, 29). While we and others (20) strongly agree that the single sigmoid provides solid computational insight into the "protein folding space" and its constraints, the stoichiometric margins of life appear to be more understandable and appreciated in general. Therefore, we examined simply the raw data of the number of contacts as a function of the percentage occurrences, at different distances (the complete dataset is provided as supplementary material). By doing so, we directly observe the presence or absence of correlations between the total number of contacts at specified distances, rather than in terms of "n" and "k" as done previously (1). Figure 1 clearly shows that regardless of the defined distance, number of contacts made by leucines with individual amino acids is well correlated to frequency of occurrences of the respective amino acids. Overall, including the insets, Figure 1 shows:

$$\text{Number of Contacts} = m \times \text{Number of pairs}$$

This is also a distance independent relationship, where "Number of pairs" is directly proportional to (Percentage Occurrence)^2. These results establish our proposals directly, in a model independent manner. Here, it is extremely important to appreciate that development of a variety of knowledge based potentials, applications of none of which have more than 75% success in explaining folded proteins and require customized corrections to achieve high resolution structural predictions, has originated from analyzing the apparent deviations of points from the straight line shown in Figure 1A. In this regard, we wish to state the following explicitly:

1. It is quite incongruous to analyze these apparent deviations in sub-10 Å regimes and ignore them for 20 Å or higher distance regimes based on the assumption that only the former matter and the latter do not. At the same time, it would be equally incongruous to propose that weak "preferential interactions" do occur at distance scales of 20 Å or higher. Thus, over-analyses of amino acid pairs limited to sub-10 Å distances has resulted in somewhat misleading knowledge based potentials.
2. A clear and conclusive stoichiometric dependence of number contacts that increases in correlation with increasing distances points out a uniform distribution constrained only by the sampling size. Lower the sample size - more is the observed variation from the expected. For example, if percentage occurrence of an amino acid is 7% in a 100 residue protein, then every set of 10 residues of the protein would be expected to have either 1 or 0 of this amino-acid, on an average. Thus, the "closer" we look into subsets of 10 or lesser residues, the more noise we would see in terms of the average occurrence of this residue. Thus, the deviation seen in Figure 1A is simply noise.
3. Over-analyses of the above noise (Figure 1A) has led to sophisticated formalisms and development of numerous knowledge based potentials, none of which are universally applicable to known protein crystal structures.
4. Percentage occurrence statistics of the 20 amino acids have now been collected for 131855 protein sequences (confirmed by annotation and experimentally and having 50 or more residues) from the ExPASy Proteomics Server (http://www.expasy.ch/sprot/) and are shown in Table 1. The stoichiometric margins of life found by us for 3718 proteins correlate extremely well with those for 131855 sequences in the Swiss-Prot server (Table 1). In fact, the very minor deviations between the margins of life (1) and Table 1 here are probably due to presence of a (small) number of unstructured proteins also.

Having established our findings in a model independent manner, we emphasize below some particularly remarkable

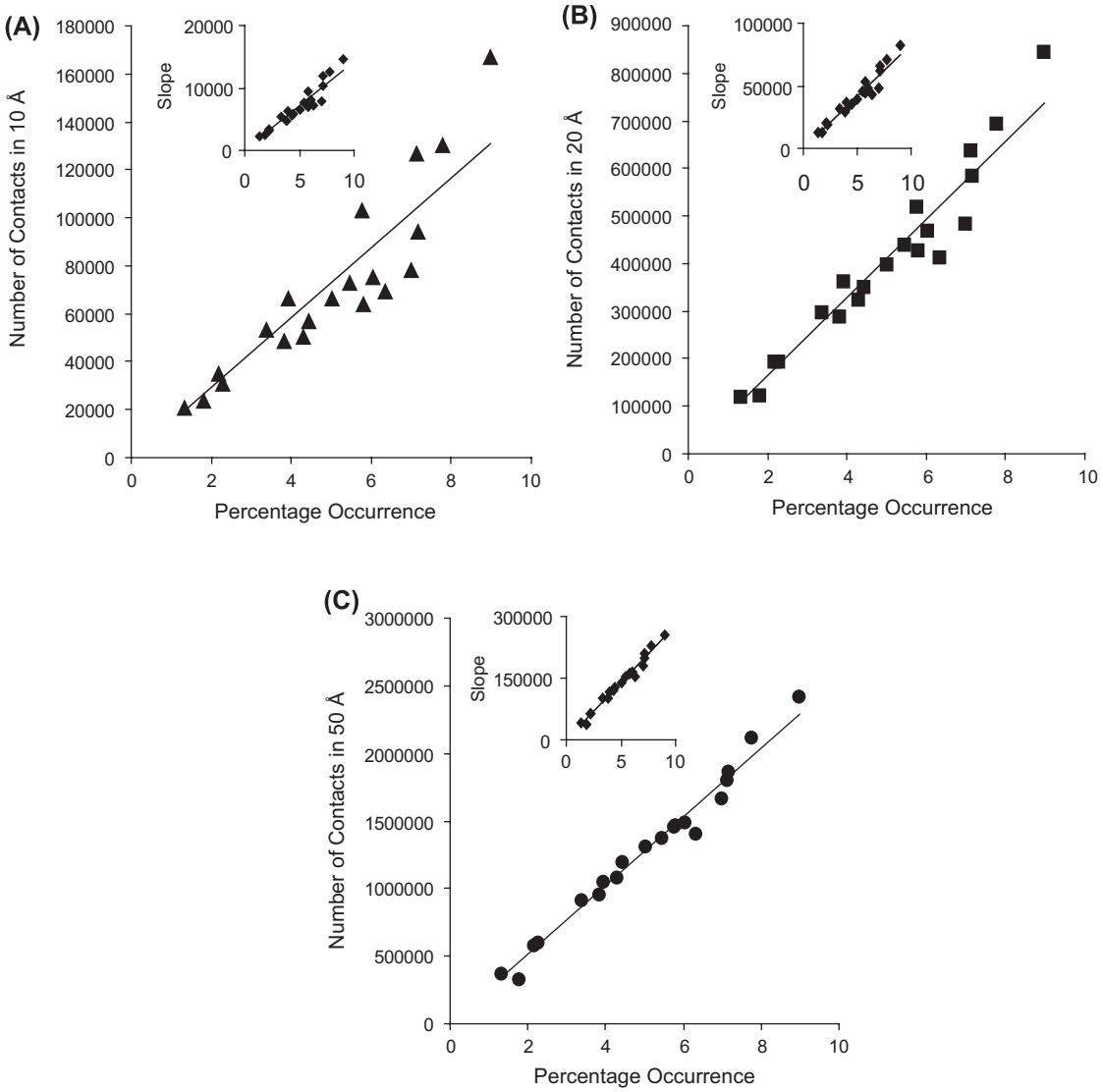

Figure 1: Neighborhoods of amino-acids in folded proteins are determined by simply their stoichiometry in primary sequences, regardless of the definition of neighborhood distance – (**A**) Neighbors of leucine within a 10 Å neighborhood are correlated well with their frequency of occurrence in folded proteins, regardless of the size of the protein. The relationship is of the form "Number of Contacts = Slope x Percentage Occurrence". Inset shows that "Slopes" from such relationships for all the 20 amino-acids are also excellently correlated with the frequency of occurrence of the respective amino-acids. (**B**) Neighbors of leucine within a 20 Å neighborhood are correlated well with their frequency of occurrence in folded proteins, regardless of the size of the protein. The relationship is of the form "Number of Contacts = Slope x Percentage Occurrence". Inset shows that "Slopes" from such relationships for all the 20 amino-acids are also excellently correlated with the frequency of occurrence of the respective amino-acids. (**C**) Neighbors of leucine within a 50 Å neighborhood are correlated well with their frequency of occurrence in folded proteins, regardless of the size of the protein. The relationship is of the form "Number of Contacts = Slope x Percentage Occurrence". Inset shows that "Slopes" from such relationships for all the 20 amino-acids are also excellently correlated with the frequency of occurrence of the respective amino-acids. From the insets of (**A**), (**B**) and (**C**) we find "Slope = m x Percentage Occurrence". Therefore, regardless of the definition of neighborhood distance, we get Number of Contacts = m x Number of pairs, where Number of pairs is directly proportional to (Percentage Occurrence)^2 for every amino-acid in a folded protein, regardless of the size of the protein.

examples of simulations by some investigators who while attempting to refute our conclusions, actually support our proposals extremely well:

1. Galzitskaya *et al*. (4) very clearly demonstrate (involuntarily) that in case of well established preferred interactions, such as A-T and C-G in DNA, application of our approach yields very clearly the following:
   (a) Spatial organizations of complementary base pairs clearly do not follow the same behavior as that observed for non-complementary base pairs. The curves obtained in Figure 1 in (4) cannot be fit by a single equation, with the complementary base pairs showing unique/different forms.
   (b) The preferred interactions are extracted, although to varying degrees.

Table 1
The average percentage occurrence of each amino-acid from the ExPASy Server.

Amino Acid	Protein sequences confirmed by annotation and experiments (mean ± std, n = 131855)
A	7.2 ± 3.0
V	6.3 ± 2.1
I	5.1 ± 2.2
L	9.6 ± 2.9
Y	3.0 ± 1.5
F	3.9 ± 1.8
W	1.2 ± 0.9
P	5.4 ± 2.6
M	2.2 ± 1.3
C	1.9 ± 2.3
T	5.5 ± 1.8
S	7.9 ± 2.8
Q	4.3 ± 2.0
N	4.2 ± 1.9
D	5.2 ± 1.9
E	6.8 ± 2.8
H	2.4 ± 1.3
R	5.3 ± 2.9
K	6.0 ± 2.9
G	6.6 ± 2.8

Thus, application of our methodology to DNA sequences clearly extracts the complementary base pairs. Therefore, the corollary stands that in absence of extraction of individual amino acid paired interactions, folded proteins do not have the conventionally assumed preferential interactions.

2. Chan (6) shows that even in presence of presumed preferential interactions in a lattice model, reproduction of our results is seen. A significant aspect of this work is that if one was to simply reverse the positioning of red and blue beads in the simulation, while keeping the numbers the same as original, similar results would be obtained. In other words, the results are essentially dependent on the numbers of the red and blue beads only. Thus, H-H contacts or P-P contacts are not required to be defined in this simulation. Simply keeping total number of beads as the same and keeping H/P ratio also the same in the example would yield the same results. Therefore, the conclusion must be that H-H or P-H/H-P or P-P contacts in this simulation are simply a *post-facto* inference resulting from the number of H and P beads considered in the simulation system. Interestingly enough, before applying our methodology, Chan states "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." Thus, stoichiometric margins have already been fixed to obtain the results by Chan. The corollary is one would expect to obtain similar spatial distributions that result from fixed stoichiometries. This is exactly our conclusion.
3. Mitternacht and Berezovsky (7) state "The distributions seen in the paper are an effect of general protein geometry and the natural frequencies of the different amino acids". We could not agree more. We are the first ones to demonstrate this intuitive statement. Further, in their simulations, the authors appear to consciously avoid the use of simple frequency of occurrence on their data set and utilize somewhat complex formalisms. It is apparent from their Figure 1 that simple division by frequency of occurrence for the amino acids would yield indistinguishable data sets for neighbors of leucines.
4. Wang *et al.* (15) show occurrence probabilities of the twenty amino acids in different structural classes of proteins. Interestingly, they specifically investigate a limited number of sequences based on "similar folds" to four selected proteins representing four "different structures". Neither do the authors consider single polypeptide chains (as we have done), nor do they consider the possibility of exceptions to general biology. After all, even all of DNA is not double helical. Further, the authors apparently avoid a figure with all of their data pooled as one. One needs to appreciate that the margin of life is in form of distributions and not absolutes. Moreover, the stoichiometric margins of life found by us for 3718 proteins correlate extremely well with those for 131855 sequences in the Swiss-Prot server.
5. Matthews incorrectly calculates the total number of contacts in a hypothetical protein by including immediate sequence neighbors (2). Now, let us carefully consider the example provided by Matthews and compare 3 sequences composed of only 3 amino acids (Met, Ser, Ala) but with varying stoichiometries: (i) Met-Ser-Ser-Ala-Ala-Ala-Ala-Ala-Ala-Ala (ii) Met-Ala-Ala-Ser-Ser-Ser-Ser-Ser-Ser-Ser (iii) Met-Ser-Ser-Met-Ala-Met-Met-Met-Met-Met. It is straightforward to apply our methodology and find that in sequence (i) Met, Ser and Ala have stoichiometric percentages of 10, 20 and 70 resp., and, a total of 8, 14 and 50 contacts resp. Thus, for all the 3 hypothetical proteins, Met, Ser and Ala have stoichiometric percentages of 30.00 ± 34.64, 36.67 ± 28.87%, 33.33 ± 32.15% resp. and a total of 67, 78, 71 contacts resp. Firstly, the stoichiometric standard deviations in this example are clearly very high compared to the margins of life. Secondly, the regression between percentage occurrence and total contacts is already lower than that found by us for 3718 proteins. We are also enthused to observe the final sentence by Matthews, while suggesting analysis of side chain contacts – "Such a calculation would have to be appropriately normalized to

take into account the abundance of all of the amino acids involved." This confirms the acceptance of our proposals regarding the need for a straightforward accounting (we urge this; for example if C_β contacts were analyzed, proper and simple normalization for number of glycines would be required since it lacks one!) for compositional stoichiometries in natural proteins. We are convinced that if simple stoichiometric rules of physical chemistry are applied, the field of protein folding will certainly benefit substantially from our newest view.

6. Rackovsky and Scheraga (3) emphasize the importance of the four weak interactions in protein folding while correctly pointing out that contact maps are geometric tools and not energetic measures. Surprisingly, these authors ignore that a contact map resulting from energetics of presumed interactions should certainly show those interactions. While completely mis-stating the Chargaff's rules that primarily indicated stoichiometric equivalences of A with T and G with C respectively (the pairings/preferential interactions were inferred much later by Watson and Crick), these authors do correctly summarize our findings in terms of the importance of relative proportions of amino acids in protein sequences. In a nutshell the arguments presented by these authors are extremely well captured in support of our proposals by the elegant views of Gruebele (9) - "In a compact random heteropolymer, this result is what one would expect. But there is another possible explanation. If folding is governed by myriad weak interactions (van der Waals contacts, entropically driven solvent exclusion or hydrophobicity, hydrogen bonds, salt bridges, *etc.,*), the free energy terms summed up to produce a given pair-distribution will act as random variables, and the Central Limit Theorem applies approximately. A universal sigmoid should then provide a fairly good fit to the data. Thus the result of Mittal reinforces the notion that no single 'magic bullet' interaction holds proteins together in a compact state."

Having established our findings in a model independent manner, we are now aware that the next challenge is to be able to provide a mechanism towards solving of "Chargaff's rules" of protein folding in terms of stoichiometric margins of life, analogous to the hydrogen-bonded pairing of complementary bases proposed by Watson and Crick. The first solid step has already been taken in this direction (31), in which we have discovered the existence amino acid side chain and location independent invariant neighborhoods in backbones of folded proteins. We hope that utilization of these invariant neighborhoods for developing knowledge based potentials could be a strong step towards completely solving the protein folding problem.

It is important to mention here that in (31) we also find that out of the possible 400 pairs of amino-acids, Cys-Cys pairs show a distinct spatial organization compared to the remaining 399. These results are a direct "test for validity" of our methodology suggested by Agutter (21) (these results had been informally shared with Prof. Ramaswamy Sarma along with the formal submission of (1)). Finally, in the spirit of our work and comments received on it, we present a quote attributed to Alfred E. Newman (the fictional mascot of Mad magazine): "*We are living in a world today where lemonade is made from artificial flavors and furniture polish is made from real lemons*."

Acknowledgements

BJ acknowledges the funding support from the Department of Biotechnology, and Department of Information Technology, Govt. of India. We are extremely grateful to the authors of the comments on our work, who were gracious in giving their valuable time to, and, even more important valuable opinions on, our findings. We are thankful beyond words to Prof. Ramaswamy Sarma for taking such energetic interest in our work and giving us an opportunity to discuss our results in such depth at such a global scale.

Supplementary Material

The supplementary material, in form of MS-Excel file, is freely available at the following address: http://www.scfbio-iitd.res.in/publication/PrimaryContactData.xls.

References

1. A. Mittal, B. Jayaram, S. R. Shenoy, and T. S. Bawa. *J Biomol Struct Dyn* 28, 133-142 (2010).
2. B. W. Matthews. *J Biomol Struct Dyn* 28, 589-591 (2011).
3. S. Rackovsky and H. A. Schraga. *J Biomol Struct Dyn* 28, 593-594 (2011).
4. O. V. Galzitskaya, M. Yu. Lobanov, and A. V. Finkelstein. *J Biomol Struct Dyn* 28, 595-598 (2011).
5. H. J. C. Berendsen. *J Biomol Struct Dyn* 28, 599-602 (2011).
6. H. S. Chan. *J Biomol Struct Dyn* 28, 603-606 (2011).
7. S. Mitternacht and I. N. Berezovsky. *J Biomol Struct Dyn* 28, 607-610 (2011).
8. S. Akella and C. K Mitra. *J Biomol Struct Dyn* 28, 611-614 (2011).
9. M. Gurebele. *J Biomol Struct Dyn* 28, 615-616 (2011).
10. R. I. Dima. *J Biomol Struct Dyn* 28, 617-618 (2011).
11. B-G Ma and H-Y Zhang. *J Biomol Struct Dyn* 28, 619-620 (2011).
12. X-L Ji and S-Q Liu. *J Biomol Struct Dyn* 28, 621-624 (2011).
13. M. Mezei. *J Biomol Struct Dyn* 28, 625-626 (2011).
14. I. Ghosh. *J Biomol Struct Dyn* 28, 627-628 (2011).
15. J. Wang. Z. Cao, and J. Yu. *J Biomol Struct Dyn* 28, 629-632 (2011).
16. K. Berka and M. Otyepka. *J Biomol Struct Dyn* 28, 633-634 (2011).
17. C. H. T. P. Silva and C. A. Taft. *J Biomol Struct Dyn* 28, 635-636 (2011).
18. B. P. Mukhopadhyay and H. R. Bairagya. *J Biomol Struct Dyn* 28, 637-638 (2011).

19. R. Nagaraj. *J Biomol Struct Dyn* 28, 639-640 (2011).
20. J. C. Burnett and T. L. Nguyen. *J Biomol Struct Dyn* 28, 641-642 (2011).
21. P. S. Agutter. *J Biomol Struct Dyn* 28, 643-644 (2011).
22. T. C. Ramalho and E. F. F. da Cunha. *J Biomol Struct Dyn* 28, 645-646 (2011).
23. R. A. Bryce. *J Biomol Struct Dyn* 28, 647-648 (2011).
24. S. Mishra. *J Biomol Struct Dyn* 28, 649-652 (2011).
25. A. Bagchi and T. C. Ghosh. *J Biomol Struct Dyn* 28, 653-654 (2011).
26. S. Ventura. *J Biomol Struct Dyn* 28, 655-656 (2011).
27. J. J. Perez. *J Biomol Struct Dyn* 28, 657-659 (2011).
28. R. Ramanathan and A. Verma. *J Biomol Struct Dyn* 28, 661-662 (2011).
29. Y. Song, Y. Song, and X. Chen. *J Biomol Struct Dyn* 28, 663-665 (2011).
30. R. Joshi. *J Biomol Struct Dyn* 28, 667-668 (2011).
31. A. Mittal and B. Jayaram. *J Biomol Struct Dyn* 28, 443-454 (2011).