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


***Editorial***

# A Conversation on Protein Folding

**Ramaswamy H. Sarma**

Department of Chemistry, State University of New York at Albany, Albany NY 12222, USA

http://www.jbsdonline.com

When the paper on stoichiometry-driven protein folding by Mittal *et al.* (1) was published, I knew that it would become a controversial publication, bordering on revolutionary. The ideas, such as (i) no preferences are related to side-chain chemistry and (ii) notwithstanding size and 3D fold, the probability that two amino acids will be close together depends mainly on their percent populations, go against what is being currently practised and taught in biochemistry and molecular biology. Preferential interaction between amino acids is the basis of the development of knowledge-based potentials, which in turn form the underpinning of protein structure prediction by modeling and simulation, now routinely performed in many laboratories across the globe (2-5 and references therein).

It was very obvious that I should invite protein structural chemists to comment on the claims of Mittal *et al.* In my letter of invitation, I made it explicitly clear that their comments would not be subject to the normal peer review so that they could express their personal points, views, and evaluation of Mittal *et al.* without interference from the referees. This is perfectly fine because the comments themselves are not research articles which require mandatory peer evaluations, but just pure comments, and referees inserting moderation and balance into the comments is not right. These comments are brief items without abstracts, contain only a short list of references, and do not contain original research data. They are essentially open referee reports on Mittal *et al*. As Editor-in-Chief of the *Journal*, I thoroughly read each one of the comments; it was necessary to make certain minor editorial changes in a few of them so that all the comments were in resonance with the *Journal* format and mission. The comments are published without dates received and without the name of the Communicating Editor because they are not regular research articles. Finally this *Journal* is publishing the section consisting of this editorial, the 29 comments and the author response as *Open Access*. This is because this *Journal* strongly believes that doctoral students in biochemistry and molecular biology will benefit a great deal from a study of these comments; and *Open Access* publication enables this.

I have received comments from 29 laboratories across the globe. I thank all of them for reading Mittal *et al.* and expressing their opinion. I am particularly grateful to the senior and highly respected investigators in the discipline, Harold Scheraga, Cornell Univ., Brian Matthews, Univ. of Oregon, Alexei Finkelstein, RAS, Pushchino, Russia and Herman Berendesen, Univ. of Groningen, The Netherlands for participating in this Conversation and providing their comments within the deadline.

Mittal *et al*. have written a single collective response to the 29 comments so as to avoid redundancy and to provide the response in a clear well structured setting. There is very little one could do about the repetition of a few items from comment

Corresponding Author:
Ramaswamy H. Sarma
Phone: 518-456-9362
Fax: 518-452-4955
E-mail: rhs07@albany.edu

to comment. Nevertheless, the collection of comments and the author response form an engaging Conversation. Many of the commentators in their presentation have visited the literature on how protein folding evolved over the last 70 years. There have been interesting incidents in which a commentator answers certain questions raised by another commentator, even though commentators themselves had no previous notion of who was writing what. While several people were very critical of the Mittal *et al*. thesis of protein folding, to others it appeared like a fresh breeze with a novel perspective; even there were young investigators excited about potential development of biologically meaningful *ab initio* modeling, based on universal principle of stoichiometry. There were discussions of the evolution of the amino acids, a few evolving earlier than others, and how this might have influenced their distribution in the proteins. There was also the question of whether there was any relationship between the proportions of amino acids and their proprotions in the codons. There are also people who believe that the *organization* and the *hierarchical structure* of the living system makes the living system unique, and that the molecular biology at the lowest level does not dictate the terms and conditions for the higher level.

***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).

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


***Comment***

# Stoichiometry *versus* Hydrophobicity in Protein Folding

http://www.jbsdonline.com

**Brian W. Matthews**

Institute of Molecular Biology and
Department of Physics,
1229 University of Oregon,
Eugene, OR 97403-1229

In their recent report in this journal, A. Mittal, B. Jayaram and coworkers (1) suggest that protein folding is driven by stoichiometric occurrences of amino acids ("Chargaff's Rules"), and not by preferred interactions (*e.g*., hydrophobic interactions) between specific amino acids. The proposal is a radical departure from conventional wisdom and is shown to be without merit.

The analysis of Mittal *et al.* (1) is based on the counting of $C_\alpha$-$C_\alpha$ distances in 3718 structures taken from the Protein Data Bank. Beginning with a single protein one takes an amino acid of interest (*e.g*., leucine). Around the $C_\alpha$ atom of the first leucine in the protein structure one draws a series of concentric spheres of increasing radius, *X*. The number of $C_\alpha$ atoms of any other residues that happen to be located within the sphere is counted separately for the 20 possible types of amino acid and binned as a function of the sphere radius. In this way one determines both how many and what sort of neighbors (or "contacts") there are for this leucine residue within a given distance.

The procedure is then repeated in turn for all other leucine residues in the protein and the results summed. The same analysis is carried out for each of the proteins from the PDB and these results are also included in the summation. The overall totals give the leucine-specific neighborhood distribution for all leucines in all the proteins analyzed.

Starting from the beginning, the same procedure is used to find an alanine-specific neighborhood distribution, and likewise for all 20 types of amino acids. The results can be seen in Figure 2 of Reference 1.

*Total $C_\alpha$-$C_\alpha$ Contacts:* In the first part of their analysis of the neighborhood distribution curves Mittal *et al*. (1) focus on the total number of contacts made by each of the 20 different types of amino acids. They hypothesize that "in case amino-acids prefer certain neighborhoods due to preferential interactions (*e.g.,* hydrophobic, hydrogen bonding, electrostatics), one would not be able to predict a direct relationship between the total number of contacts of a given amino-acid with its frequency of occurrence in folded proteins".

To test this hypothesis they estimate, for each type of amino acid, the total number of contacts generated by all $C_\alpha$ atoms in all proteins and plot this against the percentage occurrence of that amino acid in the overall sample of proteins. The points lie on a straight line, leading the authors to state "To our surprise, the total number of contacts made by an amino acid were correlated excellently with the average occurrence of that amino acid (stoichiometry) in folded proteins as

Corresponding Author:
Brian W. Matthews
Phone: (541)346-2572
Fax: (541)346-5870
E-mail: brian@uoregon.edu

shown by Figure 2E. This strongly supported our hypothesis and directly implied an 'absence' of any preferential interactions between amino acids". Unfortunately, as shown by the simple example below, this plot has no predictive value because the calculation always gives a straight line, whether the $C_\alpha$ atoms correspond to a real protein or to some other arbitrary arrangement.

Consider, for example, a protein consisting of one methionine (Met), two serines (Ser) and seven alanines (Ala) arranged as in Figure 1. It can be assumed that the structure is fully extended although this will not enter into the analysis.

Met-Ser-Ser-Ala-Ala-Ala-Ala-Ala-Ala-Ala

**Figure 1.** Hypothetical protein of 10 amino acids.

We now wish to plot the equivalent of Figure 2E of Reference 1 for this hypothetical protein. For each type of amino acid we need to count the total number of "contacts" that can be made with all other $C_\alpha$ atoms in the chain. For the single methionine there are nine such contacts. Because there are two serines and each can make nine contacts there are 18 contacts. Similarly, the seven alanines make a total of 7 x 9 = 63 contacts. As summarized in Table 1, these total contact values are directly proportional to the abundance of each amino acid. It will also be apparent from the example that the same straight-line dependence seen in Table 1 would be obtained no matter what sequence or structure or arbitrary arrangement of $C_\alpha$ positions was assumed for the protein. For the same reason, the straight-line dependence seen in Figure 2E of Reference 1 cannot be taken as support for the suggestion that preferential interactions between amino acids in proteins are unimportant. Because Figure 2E will be a straight line regardless of the assumed structure or amino acid composition, it does not support the idea that the percentage occurrence of the 20 amino acids needs to be within specified limits in order for a protein to fold (Table 1 of Reference 1).

**Table I**
Stoichiometry calculation for the hypothetical protein shown in Figure 1.

| Type of amino acid | Total number of $C_\alpha$- $C_\alpha$ "contacts" possible for the whole protein | Percentage occurrence of amino acid in protein |
|---|---|---|
| Met | 9 | 10% |
| Ser | 18 | 20% |
| Ala | 63 | 70% |

*Short-range $C_\alpha$-$C_\alpha$ Contacts:* In the second part of their analysis Mittal *et al.* (1) focus on shorter-range $C_\alpha$- $C_\alpha$ contacts. Their analysis is based on the following relationship (Equation 1) which they use to describe how the number of contacts depends on distance.

$$Y = Y_{max}(1 - e^{-kX})^n \quad [1]$$

$Y$ is the number of contacts within a sphere of radius $X$, and $Y_{max}$ is the number of contacts for the whole sample (*i.e.,* when the sphere radius becomes large enough to encompass the whole protein). The distributions are sigmoidal in nature and $k$ and $n$ are arbitrary variables which are determined by fitting the equation to the experimental data. In practice, Mittal *et al*. found for each of the 20 amino acids that $k$ has a value of about 0.07 and $n$ a value around 4.5.

To look for possible short-range differences between one type of amino acid and another, Mittal *et al.* focus especially on the value of $n$ because the "lift-off" points of the sigmoid distributions are strongly dependent on $n$. In their words "if there were any preferential neighborhoods (of amino acids) they would be reflected particularly in $n$".

Returning to Equation 1, if $(-kX)$ is small then

$$e^{-kX} \approx 1 - kX \quad [2]$$

which leads to

$$Y \approx Y_{max}(kX)^n \quad [3]$$

Since $k$ equals about 0.07, Equation 3 should be reasonably accurate up to radii $X$ of about 5 Å.

Equation 3 predicts that the number of $C_\alpha$ contacts increases in proportion to the $n^{th}$ power of the sphere radius $X$. For a hypothetical one-dimensional protein as in Figure 1, it would be expected that the number of $C_\alpha$ contacts would be approximately proportional to the sphere radius $X$, *i.e*., $n$ would equal about 1. For a hypothetical two-dimensional protein, with $C_\alpha$ positions in a plane, the number of $C_\alpha$ contacts would increase in proportion to the sphere radius squared, *i.e*., $n$ would be approximately 2. For three-dimensional proteins it would expected that $n \sim 3$, *i.e*., the volume of a sphere of radius $X$ is proportional to $X^3$.

How can it be that the actual value determined by Mittal *et al.* (1) is $n \sim 4.5$? As shown below, it arises from limitations on $C_\alpha$-$C_\alpha$ distances in proteins. For consecutive amino acids in a polypeptide chain the $C_\alpha$-$C_\alpha$ distance is 3.8 Å (ignoring *cis* peptide bonds which are extremely rare). For non-consecutive amino acids the closest $C_\alpha$-$C_\alpha$ approach is limited to about 3.2 Å by van der Waals contact. The consequences of these restrictions on the $C_\alpha$-$C_\alpha$ distance can be seen in Figures 3(A-D) of Reference 1. For "neighborhood distances" of 1 Å, 2 Å and 3 Å there are no $C_\alpha$-$C_\alpha$ contacts whatsoever and even at 4 Å the number of contacts remains very small. This type of distribution, with zero or small values from $X = 0$ to 4 Å, and rapidly increasing thereafter, is exactly what one would expect for a distribution of the form $X^n$ with $n \sim 4.5$.

Mittal *et al*. find that the value of *n* is essentially constant, regardless of the amino acid, and infer that this demonstrates the absence of preferred interactions between amino acids. In contrast, the analysis given above shows that the value of *n* arises from the inherent stereochemistry of protein backbones. Because *n* is determined primarily by backbone geometry it is expected to be largely independent of the identity of the side-chain, consistent with observation.

In summary, the conclusions reached by Mittal *et al.* based on their analysis of both longer and shorter $C_\alpha$-$C_\alpha$ distances in known protein structures are not justified.

Some years ago, Rose *et al*. (2) also carried out an analysis of known protein structures, but based on the accessibility to solvent of individual amino acids in each protein. Their analysis showed that hydrophobic amino acids like leucine, isoleucine, methionine, *etc.,* are often fully buried within the core of the protein, whereas this happens infrequently for the polar amino acids. This analysis strongly indicated that the non-polar residues do tend to cluster together within the cores of proteins and this provides hydrophobic driving energy for the folding process.

In this context it might be worth noting that the purpose of the analysis of Mittal *et al.* (1) was to look for preferential interactions between specific pairs of amino acids in proteins. Such interactions will be between side-chains, but the analysis was based on $C_\alpha$-$C_\alpha$ distances, which may be more diagnostic of the backbone. As a suggestion, it might be instructive to repeat the analysis of Mittal *et al*. using $C_\beta$-$C_\beta$ rather than $C_\alpha$-$C_\alpha$ distances since the former may be more representative of sidechain-sidechain interaction. Also, rather than attempting an interpretation based on Equation 1, a simpler calculation would suffice. If, for example, a $C_\beta$-$C_\beta$ analysis was made of amino acids in the neighborhood of leucine, the key question would be whether the hydrophobic amino acids in the immediate vicinity of the hydrophobic leucine occur more frequently than the polar ones. Such a calculation would have to be appropriately normalized to take into account the abundance of all of the amino acids involved.

***References***

1. A. Mittal, B. Jayaram, S. Shenoy, and T. S. Bawa. *J Biomol Struct Dyn 28*, 133-142 (2010).
2. G. D. Rose, A. R. Geselowitz, G. J. Lesser, R. H. Lee, and M. H. Zehfus. *Science 229*, 834-838 (1985).

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


***Comment***

# On the Information Content of Protein Sequences

http://www.jbsdonline.com

**S. Rackovsky**[†]
**H. A. Scheraga**[*]

[†]Dept. of Pharmacology and Systems Therapeutics
Mount Sinai School of Medicine of NYU
One Gustave L. Levy Place
New York, NY 10029
[*]Dept. of Chemistry and Chemical Biology
Baker Laboratory
Cornell University
Ithaca, NY 14853

In recent work, Mittal *et al.* have investigated the spatial organization of amino acids in proteins (1). They considered residue-specific radial distributions of the 20 amino acids, and concluded, on the basis of an observed universality in the form of the distribution function, that the folding of proteins is "a direct consequence of a narrow band of stoichiometric occurrences of amino acids in primary sequences" They reject the role of specific amino acid interactions as a driving force in protein folding, and suggest that this result represents a protein version of the Chargaff rules which have long been known to govern nucleic acid folding.

In the present work, we make several points relevant to the conclusions of Mittal *et al.*

1. It has been long known (2, 3) that neither α-carbon nor side-chain radial distribution functions alone are reliable indicators of the placement of residues in proteins, and are therefore not reliable indicators of inter-residue interactions in proteins. Rather, it was shown that one must consider the angular dispositions of side chains in order to obtain a reliable indication of interactive properties. This result was analyzed with respect to its significance for hydrophobic interactions, but is even more certainly correct for the atom-specific interactions which are responsible for Van der Waals, hydrogen-bonding and electrostatic interactions.

   It should be noted more generally that "contacts", defined by the presence of given atoms within a specified distance from a point of interest in the structure, do *not* automatically carry information about stabilizing interactions. A contact map is a geometric tool, not an energetic measure. It is entirely possible that atoms which fall within the specified distance, particularly at the longer distances (up to ~90 Å) considered by Mittal *et al.* do not interact with the central atom in a way which meaningfully stabilizes the structure.

   For these reasons, conclusions cannot be drawn about the global stabilization of a molecule from radial distribution functions alone.

2. It has been known for some time, thanks to the work of numerous investigators, that the amino acid composition of proteins contains information about the fold of the molecule (4-15). Their results showed that amino acid composition can give reasonably accurate classifications of structure class and fold family. Classification schemes, however, do not give information about the quantitative differences between structural classes. In very recent work (16), one of us (S.R.) addressed this question, and demonstrated that, when properly represented, amino acid composition encodes the *quantitative* organization of protein structure space.

Corresponding Author:
H. A. Scheraga
S. R. Rackovsky
E-mail: has5@cornell.edu
shalom.rackovsky@mssm.edu

In order to extract this information it is necessary to describe protein sequences numerically in a statistically significant manner, and then construct a metric function which acts on that numerical sequence information to give quantitative distances between pairs of sequences or pairs of groups of sequences. The numerical description of sequences was carried out using 10 physical property factors derived by Kidera *et al.* (17, 18) with a factor analysis. The 10 factors were shown to carry a large fraction of the variance for all the known physical property sets described for the 20 naturally occurring amino acids. These factors are orthonormal by construction, eliminating problems which arise in other numerical representations due to incompleteness and statistical correlation. (The values of the 20 factors are given in Table I, and their definitions in Table II.)

The intersequence metric function was shown (16) to be a natural result of the application of a Euclidean distance measure to the 10-dimensional space which arises when sequence compositions are represented by the Kidera factors. An excellent three-dimensional picture of the resulting space was obtained (16) by using a Principal Component Analysis of the 10-dimensional space, and shows clearly that the organization of protein architectural groups, using the sequence metric, corresponds well to that obtained in earlier work (19) on the organization of protein space using a structure-based metric.

**Table I**
Values of The Kidera Property Factors.[a]

| | f1 | f2 | f3 | f4 | f5 | f6 | f7 | f8 | f9 | f10 |
|---|---|---|---|---|---|---|---|---|---|---|
| ALA | –1.56 | –1.67 | –0.97 | –0.27 | –0.93 | –0.78 | –0.20 | –0.08 | 0.21 | –0.48 |
| ASP | 0.58 | –0.22 | –1.58 | 0.81 | –0.92 | 0.15 | –1.52 | 0.47 | 0.76 | 0.70 |
| CYS | 0.12 | –0.89 | 0.45 | –1.05 | –0.71 | 2.41 | 1.52 | –0.69 | 1.13 | 1.10 |
| GLU | –1.45 | 0.19 | –1.61 | 1.17 | –1.31 | 0.40 | 0.04 | 0.38 | –0.35 | –0.12 |
| PHE | –0.21 | 0.98 | –0.36 | –1.43 | 0.22 | –0.81 | 0.67 | 1.10 | 1.71 | –0.44 |
| GLY | 1.46 | –1.96 | –0.23 | –0.16 | 0.10 | –0.11 | 1.32 | 2.36 | –1.66 | 0.46 |
| HIS | –0.41 | 0.52 | –0.28 | 0.28 | 1.61 | 1.01 | –1.85 | 0.47 | 1.13 | 1.63 |
| ILE | –0.73 | –0.16 | 1.79 | –0.77 | –0.54 | 0.03 | –0.83 | 0.51 | 0.66 | –1.78 |
| LYS | –0.34 | 0.82 | –0.23 | 1.70 | 1.54 | –1.62 | 1.15 | –0.08 | –0.48 | 0.60 |
| LEU | –1.04 | 0.00 | –0.24 | –1.10 | –0.55 | –2.05 | 0.96 | –0.76 | 0.45 | 0.93 |
| MET | –1.40 | 0.18 | –0.42 | –0.73 | 2.00 | 1.52 | 0.26 | 0.11 | –1.27 | 0.27 |
| ASN | 1.14 | –0.07 | –0.12 | 0.81 | 0.18 | 0.37 | –0.09 | 1.23 | 1.10 | –1.73 |
| PRO | 2.06 | –0.33 | –1.15 | –0.75 | 0.88 | –0.45 | 0.30 | –2.30 | 0.74 | –0.28 |
| GLN | –0.47 | 0.24 | 0.07 | 1.10 | 1.10 | 0.59 | 0.84 | –0.71 | –0.03 | –2.33 |
| ARG | 0.22 | 1.27 | 1.37 | 1.87 | –1.70 | 0.46 | 0.92 | –0.39 | 0.23 | 0.93 |
| SER | 0.81 | –1.08 | 0.16 | 0.42 | –0.21 | –0.43 | –1.89 | –1.15 | –0.97 | –0.23 |
| THR | 0.26 | –0.70 | 1.21 | 0.63 | –0.10 | 0.21 | 0.24 | –1.15 | –0.56 | 0.19 |
| VAL | –0.74 | –0.71 | 2.04 | –0.40 | 0.50 | –0.81 | –1.07 | 0.06 | –0.46 | 0.65 |
| TRP | 0.30 | 2.10 | –0.72 | –1.57 | –1.16 | 0.57 | –0.48 | –0.40 | –2.30 | –0.60 |
| TYR | 1.38 | 1.48 | 0.80 | –0.56 | 0.00 | –0.68 | –0.31 | 1.03 | –0.05 | 0.53 |

[a]The property factors are dimensionless by construction. The reader is referred to references 17 and 18 for details of their derivation.

**Table II**
Definitions of The Kidera Property Factors.[a]

| | |
|---|---|
| 1. Helix/bend preference | 2. Side-chain size |
| 3. Extended structure preference | 4. Hydrophobicity |
| 5. Double-bend preference | 6. Partial specific volume |
| 7. Flat extended preference | 8. Occurrence in alpha region |
| 9. pK-C | 10. Surrounding hydrophobicity |

[a]The first four factors are essentially pure physical properties; the remaining six factors are superpositions of several physical properties, and are labelled for convenience by the name of the most heavily weighted component.

3. The result presented by Mittal *et al.* does not, in fact, support the existence of Chargaff-like rules for proteins. On the contrary, their result suggests that the pairwise interaction of amino acids is purely random, governed by the relative proportions of the 20 amino acids in the sequence. The Chargaff rules, on the other hand, arise from pairwise-specific interactions of nucleotides in DNA molecules, and do not reflect random association in any way.

These observations suggest that one must exercise considerable care in interpreting the results of a study of simple radial distributions. It would seem to be somewhat premature to discount the role of specific pairwise amino acid interactions in the folding of proteins.

***References***


1. A. Mittal, B. Jayaram, S. Shenoy, and T. S. Bawa. *J Biomol Struct & Dyn 28*, 133-142 (2010).
2. S. Rackovsky and H. A. Scheraga. *Proc Nat Acad Sci USA 74*, 5248-5251 (1977).
3. H. Meirovitch, S. Rackovsky, and H. A. Scheraga. *Macromolecules 13*, 1398-1405 (1980).
4. H. Nakashima, K. Nishikawa, and T. Ooi. *J Biochem 99*, 153-162 (1986).
5. M. van Heel. *J Mol Biol 220*, 877-897 (1991).
6. M. Reczko and H. Bohr. *Nucleic Acids Res 22*, 3616-3619 (1994).
7. I. Dubchak, I Muchnik, S. R. Holbrook, and S.-H. Kim. *Proc Nat Acad Sci USA 92*, 8700-8794 (1995).
8. W. Hobohm and C. Sander. *J Mol Biol 255*, 390-399 (1995).
9. Z.-X. Wang and Z. Yuan. *Proteins: Structure, Function and Genetics 38*, 165-175 (2000).
10. C. H. Q. Ding and I. Dubchak. *Bioinformatics 17*, 349-358 (2001).
11. L. Edler, J. Grassmann, and J. Suhai. *Math Comput Model 33*, 1401-1417 (2001).
12. Q.-S. Du, Z.-Q. Jiang, W.-Z. He, D.-P. Li, and Q.-C. Chou. *J Biomol Struct Dyn 23*, 634-640 (2006).
13. Y. Ofran and H. Margalit. *Proteins: Structure, Function and Bioinformatics 64*, 275-279 (2006).
14. H.-B. Shen and K.-C. Chou. *Bioinformatics 22*, 1717-1722 (2006).
15. Y.-H. Taguchi and M.M. Gromiha. *BMC Bioinformatics 8*:404 (2007).
16. S. Rackovsky. *Proc Nat Acad Sci USA 34*, 14345-14348 (2009).
17. A. Kidera, Y. Konishi, M. Oka, T. Ooi, and H. A. Scheraga. *J Prot Chem 4*, 23-55 (1985).
18. A. Kidera, Y. Konishi, T. Ooi, and H. A. Scheraga. *J Prot Chem 4*, 265-297 (1985).
19. S. Rackovsky. *Proteins: Structure, Function and Genetics 7*, 378-402 (1990).

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


***Comment***

# Cunning Simplicity of a Stoichiometry Driven Protein Folding Thesis

http://www.jbsdonline.com

**Oxana V. Galzitskaya**
**Michael Yu. Lobanov**
**Alexey V. Finkelstein***

Institute of Protein Research, Russian Academy of Sciences, Institutskaya str., 4 Pushchino, Moscow Region, 142290, Russia

Mittal *et al*. present two statements, which, as they say, 'reveal a surprisingly simple unifying principle of backbone organization in protein folding' (1). Namely, they assert (i) 'that protein folding is a direct consequence of a narrow band of stoichiometric occurrences of amino-acids in primary sequences', and (ii) 'that "preferential interactions" between amino-acids do not drive protein folding, contrary to all prevalent views'.

However, the first of these statements, presented quite ambiguously by Mittal *et al.* is not novel, while the second one is wrong (and moreover, to prove this statement, Mittal *et al.* used a method that is hardly sensitive to interactions between the chain links). Thus, the paper by Mittal *et al*. is useful in one respect only: it clearly shows that consideration of interactions in protein molecules at the level of Cα - Cα distances has no sense.

Mittal *et al.* write that there is 'a narrow band of stoichiometric occurrences of amino-acids in primary sequences' for the "folded" (*i.e.,* having 'crystal structures in the Protein Data Bank' (2)), while 'stoichiometry of "unstructured" proteins' deviates 'from the frequencies of occurrence of amino-acids in folded proteins'.

This needs to be commented.

(a). Actually, the reported numbers in Table I of Mittal *et al.* clearly show that the "band" *is not* that narrow: the content of Ala in "folded" proteins is 7.8 ± 3.4%, the content of Cys is 1.8 ± 1.5%, and so on. Thus, individual proteins can have significant deviations from a "typical" (for folded proteins) amino-acid composition. Do Mittal *et al*. really mean that *any* (even with a random sequence) polypeptide chain folds when and only when its amino-acid composition is in the "band" outlined by them? This would contradict to the known (3-5) large fold-destabilizing effect of quite a few point mutations. If not, what do they have in mind, really?

(b). The typical amino-acid compositions of proteins that one can find in any textbook (*e.g.,* 6, 7) are rather similar to those reported in Mittal *et al.* for folded proteins. The composition of a folded globular protein (especially of eukaryotes and their viruses) is known to be close, as a rule, to that yielded by random usage of codons of the genetic code (8). However, amino acid compositions of some well folded proteins (*e.g.,* collagen, a fibrous protein) can be entirely different from a 'typical' amino acid composition of the folded globular proteins (6, 7).

*Corresponding Author:
Alexey V. Finkelstein
Phone: +7-495-514-0218
Fax: +7-495-514-0218
E-mail: afinkel@vega.protres.ru

(c). A modest difference between typical amino-acid compositions of "folded" and "unstructured" (usually called "natively unfolded") proteins, as observed by Mittal *et al.* is well known; moreover, it is widely used to predict the state of a protein (*e.g.,* 9-14). Amazingly, Mittal *et al.* ignore all these works completely!

(d). It is also odd that Mittal *et al*. ignore the famous work by H. Fisher (15), who many years ago related the size and crude shape of protein molecules to their amino-acid composition.

(e). Although Mittal *et al*. promise to present 'rules for protein folding', they, actually, talk only on the *ability* of amino-acid sequences to form *some* folded structure, and do not say a word on why and/or how a sequence chooses its native fold among zillions of alternatives. It may be noted that just the latter question is normally called "the protein folding problem"!

Thus, although amino-acid composition is undoubtedly important for protein folding, this is not a novel finding and it cannot solve the protein folding problem without the help of geometry-specific interactions of amino-acid residues.

Now, let us consider the method used by Mittal *et al*. to prove that "preferential interactions" between amino-acids do not drive protein folding. In this connection, it should be noted that that preferential interactions between amino acids are the basis for introducing knowledge-based potentials, which in turn provide the underpinning for present day 3D protein structure prediction by modeling and simulation (16-19 and references therein).

To analyze this method, we will consider DNA double helices rather than proteins (we have the right to do this, because Mittal *et al*. also refer to Chargaff's rules established for DNA). The "preferential", complementary interactions of nucleotides, underlying the structure of double helices, are well known (20). Let us see, if it is possible to detect these interactions by the method used by Mittal *et al.* in search for the "preferential interactions" in protein chains.

The method of Mittal *et al.* briefly, consists in calculation of the number of various "contacts" of residues (Ala→Ala, Ala→Gly, ..., Asp→Lys,..., *etc.*), which are not adjacent in the chain at distances of ≤3 Å; ≤4 Å; ≤5 Å; ...; ≤60 Å *between their backbone Cα atoms* (1). Mittal *et al.* assume that 'if two amino-acids were to interact with each other ... their respective Cα- atoms would be expected to occur in fixed neighborhoods relative to each other'.

The same, literally, analysis we applied to DNA double helices, the only difference being that in DNA, we have to use C1' atoms, which play here the same side-chain binding role as Cα atoms in polypeptides. The structures of complementary DNA double helices, of at least 10 base pairs each, are taken from the Protein Data Bank (2); we chose the structures with resolution of 2.5 Å or better from the files, which contain only DNA having all four types of bases (+ water, and no other atoms). The Bank contains 18 such double helices of 10–12 nucleotide pairs. Figure 1 shows

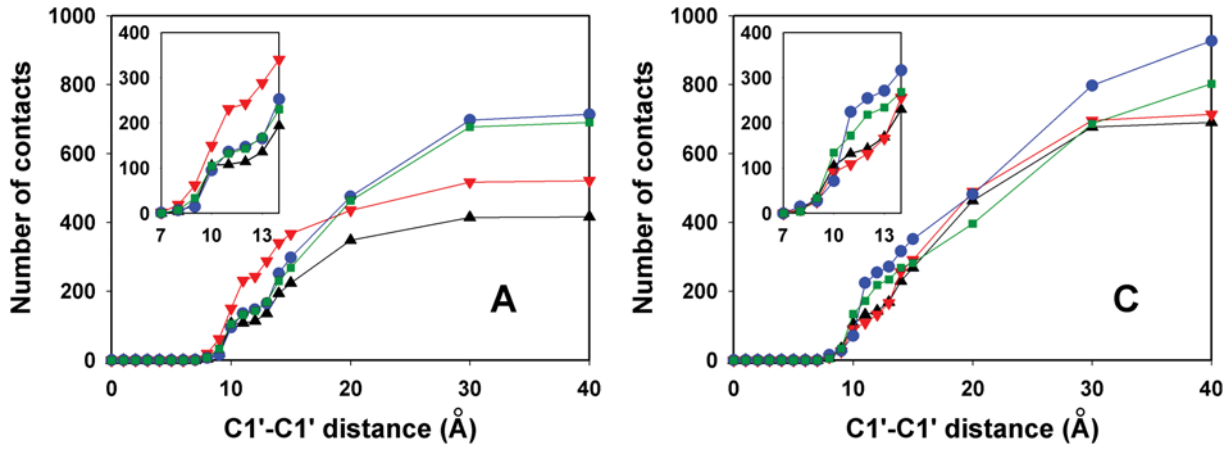


**Figure 1:** Analysis of C1'-C1' distances in crystal structures of 18 DNA double helices, including, in total, 85 A-T and 113 C-G nucleotide pairs — (**Left**) The number of nucleotide pairs A→A (black triangles), A→C (green squares), A→G (blue circles), A→T (red triangles) within a given C1'-C1' distance limit. The upper limit is 40Å (as in (1), we took the largest possible distance for our molecules), although no interaction of nucleotides is expected at C1'-C1' distances above 15Å. A physically meaningful C1'-C1' distance range is given in the Insert. The points correspond to the limits 0-1Å, 0-2Å, ..., 0-14Å, 0-15Å, 0-20Å, 0-30Å, 0-40Å. (**Right**) The number of nucleotide pairs C→A (black triangles), C→C (green squares), C→G (blue circles), C→T (red triangles) within a given C1'-C1' distance limit. One can see that both panels show a small difference in occurrence of complementary (A-T, C-G) and non-complementary nucleotide pairs: this difference is comparable to differences in the numbers of contacts between various non-complementary bases.

that the numbers obtained demonstrate a *small* (maximum twofold, but often much less) difference in occurrence of complementary and non-complementary nucleotide pairs even at short C1'-C1' distances, where nucleotides can be in a complementary contact (Figure 1 insets). This means that the approach used by Mittal *et al*. is a poor tool for singling out even the well known preferential interactions that stabilize the structures of double helices. Therefore, it is no wonder that this approach, applied to proteins (1), is hardly capable of singling out the "preferential interactions" in them.

The reason is that the distances between Cα atoms (in proteins) or between C1' atoms (in DNA) are not that sensitive to "preferential interactions" of chain links: these atoms are far from the place where a specific side chain-side chain interaction occurs. It is no mere chance that Miyazawa and Jernigan used distances between the side chain centers (*not* Cα atoms!) to obtain their famous potentials (21, 22), and Galzitskaya *et al*. used "contact distances" determined as distances between the closest (in space) heavy atoms of two amino-acid residues (23).

When we used the nucleotide contact distances (instead of the C1'-C1' distances) to analyze the crystal structures of DNA double helices, the "preferential interactions" of complementary nucleotide pairs in DNA became quite evident (Figure 2) at the shortest contact distances (Figure 2 inset), — while larger distances, which have nothing to do with physical interactions of the chain links, were again, of course, influenced by nucleotide content only as in Mittal *et al*. By the way, certain "preferential interactions" of residues at short distances are also seen in protein statistics (21, 22), but here they are quite smoothed-out and less specific than in DNA.

Some traces of these interactions can be found even in Figures 3A-D in Mittal *et al. i.e*., the Figures that concern short-range (≤10 Å) Cα-Cα distances (they are especially evident in Figures. 3B, 3D at the distances of 6–7 Å: here, they are comparable to deviations in occurrences of complementary and non-complementary nucleotide pairs at the C1'-C1' distances of 10–13 Å). However, Mittal *et al*. prefer to neglect these small deviations and concentrate on larger (up to 60 Å!) distances (where interactions must be negligible and therefore the deviations are absent) — and, of course, they miss all "preferential interactions" and obtain only some more or less trivial stoichiometric relationships.

Odd enough, Mittal *et al.* treat these stoichiometric relationships as an analogue of Chargaff's rules. The Chargaff's rules read: %A = %T and %G = %C in a double-stranded DNA (24), while Mittal *et al*. give no equalities of this kind: they end up with an already known amino-acid contents of "folded" and "unstructured" proteins, and seem to think that these contents by themselves can elucidate the intricate spatial organization of proteins!

The authors thank E. V. Serebrova for assistance in manuscript preparation. This work was supported by the grants from the HHMI (#55005607), MCB (#01200957492 and #01200959110) and "Leading Sci. Schools" (#NSh-2791.2008.4) programs, RFBR (#10-04-00162-a and #08-04-00561-a) and FASI (#02.740.11.0295).

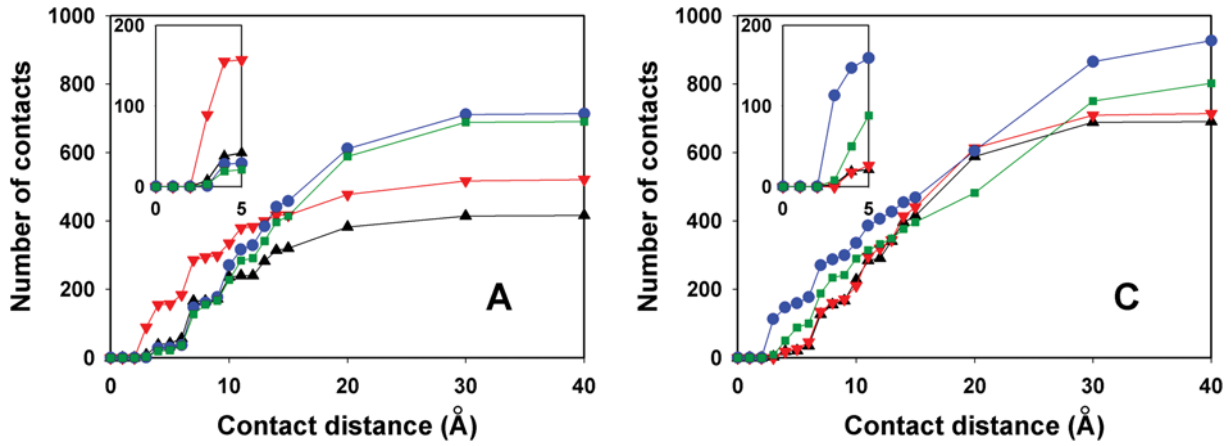


**Figure 2:** Analysis of contact distances (determined as distances between the closest in space heavy atoms of two nucleotides) in crystal structures of 18 DNA double helices — (**Left**) The number of nucleotide pairs A→A (black triangles), A→C (green squares), A→G (blue circles), A→T (red triangles) within a given contact distance limit. (**Right**) The number of nucleotide pairs C→A (black triangles), C→C (green squares), C→G (blue circles), C→T (red triangles) within a given contact distance limit. *Now* both panels show a definite (~tenfold) predominance in occurrence of complementary (A-T, C-G) nucleotide pairs at short (3 - 4Å) contact distances (while the larger distances, as in Figure 1, reflect mainly the nucleotide content of all the remote DNA pieces and, of course, do not show any "predominant interactions").

***References***


1. A. Mittal, B. Jayaram, S. Shenoy, and T. S. Bawa. *J Biomol Struct Dyn 28*, 133-142 (2010).
2. H. Berman, K. Henrick, H. Nakamura, and J. L. Markley. *Nucleic Acids Res 35*, D301-D303 (2007).
3. K. Yutani, K. Ogasahara, and Y. Sugino. *Adv Biophys 20*, 13-29 (1985).
4. T. Alber, D. P. Sun, J. A. Nye, D. C. Muchmore, and B. W. Matthews. *Biochemistry 26*, 3754-3758 (1987).
5. W. A. Baase, L. Liu, D. E. Tronrud, and B. W. Matthews. *Protein Sci 19*, 631-641 (2010).
6. G. E. Schulz and R. H. Schirmer. *Principles of protein structure*. New York - Heidelberg – Berlin, Springer-Verlag (1979).
7. A. V. Finkelstein and O. B. Ptitsyn. *Protein Physics*. Amsterdam – Boston – London – New York – Oxford – Paris – San Diego – San Francisco – Singapore – Sydney – Tokyo, Academic Press, An Imprint of Elsevier Science (2002).
8. N. S. Bogatyreva, A. V. Finkelstein, and O. V. Galzitskaya. *J Bioinform Comput Biol 4*, 597-608 (2006).
9. V. N. Uversky, J. R. Gillespie, and A. L. Fink. *Proteins 41*, 415-427 (2000).
10. V. N. Uversky. *Protein Sci* 11, 739-756 (2002).
11. S. O. Garbuzynskiy, M. Yu. Lobanov, and O. V. Galzitskaya. *Protein Sci 13*, 2871-2877 (2004).
12. A. L. Fink. *Curr Opin Struct Biol 15*, 35-41 (2005).
13. O. V. Galzitskaya, S. O. Garbuzynskiy, and M. Yu. Lobanov. *PLoS Comput Biol 2*, e177 (2006).
14. M. Yu. Lobanov, S. O. Garbuzynskiy, and O. V. Galzitskaya. *Biokhimiya (Russia)* 75(2), 192-200 (2010).
15. H. Fisher. *Proc Natl Acad Sci USA 51*, 1285-1291 (1964).
16. P. Sklenovský and M. Otyepka. *J Biomol Struct Dyn 27*, 521-539 (2010).
17. M. J. Aman, H. Karauzum, M. G. Bowden, and T. L. Nguyen. *J Biomol Struct Dyn 28*, 1-12 (2010).
18. Y. Tao, Z. H. Rao, and S. Q. Liu. *J Biomol Struct Dyn* 28, 143-157 (2010).
19. C. Koshy, M. Parthiban, and R. Sowdhamini. *J Biomol Struct Dyn* 28, 71-83 (2010).
20. J. D. Watson and F. H. Crick. *Nature 171*, 737-738 (1953).
21. S. Miyazawa and R. L. Jernigan. *Macromolecules 18*, 534-552 (1985).
22. S. Miyazawa and R. L. Jernigan. *Proteins 34*, 49-68 (1999).
23. S. O. Garbuzynskiy, B. S. Melnik, M. Yu. Lobanov, A. V. Finkelstein, and O. V. Galzitskaya. *Proteins 60*, 139-147 (2005).
24. E. Chargaff. *Experientia 6,* 201-209 (1950).

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


***Comment***

# The Relevance of Distance Statistics for Protein Folding

http://www.jbsdonline.com

**Herman J. C. Berendsen**

Molecular Dynamics Group, Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, Nijenborgh 4, 9747 AG Groningen, The Netherlands

This paper is a comment on a recent article by Mittal *et al.* (1), written at the request of the Editor of the Journal of Biomolecular Structure and Dynamics. The authors of this article investigate the statistics of inter-aminoacid distances (measured as Cα distances) in over 3700 folded proteins with known crystal structures. They report a remarkable insensitivity of this distance statistics to the nature of the amino-acids and conclude that the total number of neighbors within a given distance (of any amino-acid) depends only on the overall occurrence of that amino-acid. Thus it appears that there are neither short- nor long-ranged interactions between particular amino-acids that could be a determining factor in protein folding. The authors suggest that protein folding is determined by such packing of residues (excluding water) that the surface-to-volume ratio is minimized; they do not support the conventional view that classification into polar and nonpolar residues plays a role.

In this comment I shall first take a closer look at the meaning of the observations reported in the article. My conclusion will be that the way the results are presented emphasizes the stoichiometric aspects and hides possible relevant details. Some observations (relating to total long-range counts) that are reported as surprising and indicating the absence of long-range interactions, appear to be a simple consequence of the stoichiometry alone without sensitivity to possible interactions. The more relevant observations at shorter ranges do indeed seem to indicate absence of specific short-range interactions; I discuss the consequences for effective potential energy models used in protein folding simulations. Finally I suggest how the presentation of the results could be made more relevant.

*A few definitions:* In order to be able to discuss the results, a few definitions are given here.

With "aa" we mean "amino-acid" and with "set" we mean the set of all 3718 proteins.

$i, j = 1 \ldots 3718$ enumerates the proteins,

$k, l = 1 \ldots 20$ enumerates the type of aa,

$n_{ik}$ is the number of aa's of type $k$ in protein $i$,

$L_i = \sum_k n_{ik}$ is the length (number of aa's) in protein $i$,

$N_k = \sum_i n_{ik}$ is the total number of aa's of type $k$ in the set

Corresponding Author:
Herman J. C. Berendsen
E-mail: H.J.C.Berendsen@rug.nl

$N_{tot} = \sum_k N_k = \sum_i L_i$ is the total number of aa's in the set,

$f_k = N_k / N_{tot}$ is the fractional occurrence of aa of type $k$

***Long-range Counts***

The authors count the number of neighbors within a given range $X$ of each aa of type $k$ (excluding itself and its immediate sequential neighbors) over the whole set and sort the neighbors according to aa type $l$. Thus they obtain a 20 x 20 matrix $Y_{kl}$ for each chosen $X$. They find that each of the elements of this matrix follows a functional relationship with the distance $X$:

$$Y_{kl} = Y_{kl}^{max} [1\text{-}\exp(-\kappa X)]^n, \quad [1]$$

where $\kappa$ and $n$ are parameters that appear to be independent of the type of aa ($\kappa \approx 0.075$ and $n \approx 4.5$). Note that $k$ in the article has been replaced by $\kappa$ in order to avoid confusion with the index $k$ in this comment.

The authors then plot (in figure 2E) the sum of all 20 values of $Y_{kl}^{max}$, *i.e.*, $\sum_l Y_{kl}^{max}$, versus the percentage of occurrence of the central aa of type $k$, *i.e.*, 100 $f_k$, and find a perfect proportionality between the two variables. From figure 4C, which replots the same data, we can read the proportionality constant:

$$\sum_l Y_{kl}^{max} = 3.25 \text{ x} 10^8 f_k . \quad [2]$$

Now consider the fact that all proteins have a limited size and the asymptotic large range from any aa in any protein will include *all* aa's present in the protein. If summed over all types, the number of neighbors at a limiting large range will equal the total number of aa's in the protein minus the excluded three (two if the central aa is a terminal), *i.e.*, $L_i - 3$. For aa of type $k$ there are $n_{ik}$ occurrences in protein $i$, so that the total summation over all proteins yields:

$$\sum_l Y_{kl}^{max} = \sum_i n_{ik} (L_i - 3). \quad [3]$$

Under the assumption that the aa distribution is homogeneous over proteins of all sizes, implying that $n_{ik} = f_k L_i$, eq. [3] reduces to

$$\sum_l Y_{kl}^{max} = f_k \sum_i L_i (L_i - 3) \approx f_k \sum_i L_i^2.$$

We have recovered the empirical relation, eq. [2], without any reference to the strength of "long-range interactions." In fact such interactions are completely irrelevant for this result. The fact that the points in figure 2E are not exactly on a straight line relate to possible deviations of the homogeneity assumption, which could be validated separately.

From the considerations above I conclude that the presentation of the long-range data is such that the results are completely determined by stoichiometry without any possible influence of long-range interactions. Is this also true for the shorter-range data?

***Shorter-range Counts***

The validity of eq. [1] for all ranges $X$ with parameters $\kappa$ and $n$ independent of aa type would indicate that there are no specific interactions between amino-acids also at shorter range. The shape of the functions is then a result of the spatial distribution of each protein and the length distribution of the set of proteins. The authors claim (figure 3) that also at shorter distances (5-10 Å) the distributions follow the general form. However, closer inspection of figure 3 shows appreciable deviations of the data points from the fits to eq. [1]. Further analysis on the basis of the article cannot be given as the data points are not specified. It would be worthwhile for the authors to analyze the *deviations* from eq. [1] in terms of the amino-acids involved.

But let us assume that such further analysis will not show specificity: the conclusion must then be that folding does not imply enhanced or reduced contacts between specific aa pairs. This indeed seems contrary to the commonly held view that a folded protein contains internal clusters of hydrophobic aa's, with hydrophilic aa's situated mostly on the surface in contact with water. If such views *are* valid, the conclusion must be that Cα distances are irrelevant indicators of aa clustering.

***Consequences for Computational Folding***

Computational protein folding using detailed atomic models with explicit solvent was (2) and still is a formidable problem, mainly because the huge space that must be sampled. The sampling problem can be eased by using effective models in a reduced coordinate space. The simplest models place amino-acids on a grid with effective pair interactions. With the results of Mittal *et al.* we can now say that such approaches must fail because the pair interactions between amino-acids are non-descriptors for the interactions that determine folding. Such models are mere toys for playing with sampling methods.

More sophisticated coarse-grained models must involve several degrees of freedom per aa. At least an appropriate interaction site at – or in the direction of - the Cα position must be included in order to provide proper local densities. In the formulation of an effective Hamiltonian, it now seems that pair interactions between Cα atoms are irrelevant and can be left out (this does not mean that angular and dihedral terms for successive Cα atoms are also irrelevant). Thus the article

reviewed here may help to construct better effective Hamiltonians for coarse-grained protein models.

### *Conclusions*

It is concluded that the long-range neighbor counts do not reflect any interactions between aa's, but follow exclusively from the stoichiometry. For shorter ranges this is not true, but the presentation of the data emphasizes the non-specific distribution. From the presented data it is not clear whether any short-range specificities exist and it is recommended to re-analyze the statistics on the basis of deviations from the non-specific (average) distribution. If there are indeed no short-range specificities, one may conclude that Cα pair interactions are irrelevant descriptors for effective coarse-grained interaction models used for computational protein folding.

### *References*

1. A. Mittal, B. Jayaram, S. Shenoy, and T. S. Bawa. *J Biomol Struct Dyn 28*, 133-142 (2010).
2. H. J. C. Berendsen. *Science 282*, 642-643 (1998).

*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

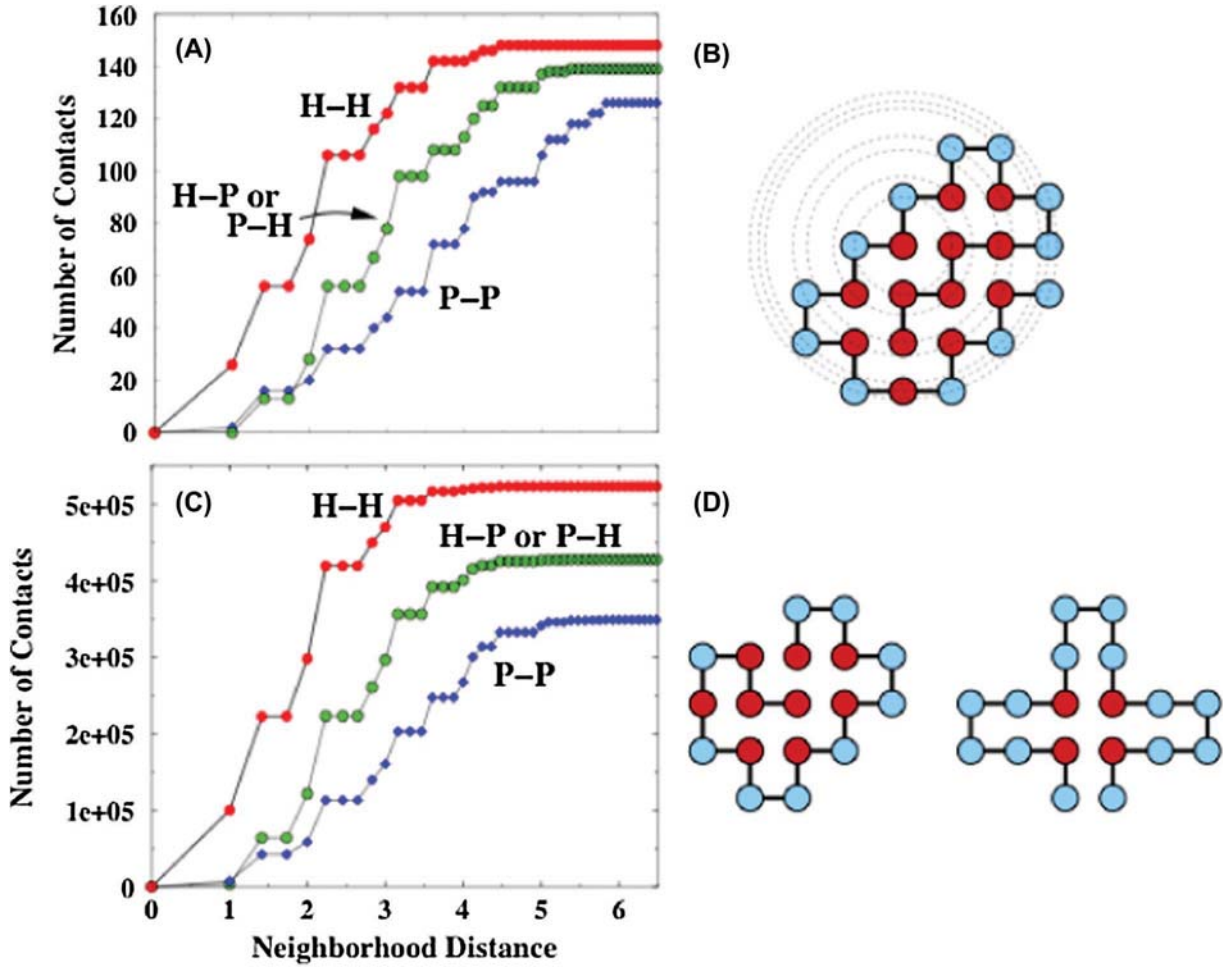


**Figure 1:** Neighborhood analysis in the two-dimensional HP model. (**A**) Following the terminology of Mittal *et al*. (1), the "number of contacts" (vertical axis) is the number of residue positions within a given distance (horizontal axis, in unit of lattice bond length) from a residue of a given 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).

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


***Comment***

# On the Importance of Amino Acid Sequence and Spatial Proximity of Interacting Residues for Protein Folding

http://www.jbsdonline.com

**Simon Mitternacht[1,2]**
**Igor N. Berezovsky[2*]**

[1]Department of Informatics,
[2]Computational Biology Unit, Bergen Center for Computational Science, University of Bergen, Norway

We discuss here the paper by Mittal *et al.* (1), which claims that stoichiometry plays a key role in protein folding. Though the very use of the term stoichiometry seems to be unjustified in this context, we will follow the authors' definition of stoichiometry in order to discuss the results, the way they were obtained, and the suggested conclusions on protein folding. The main issue we have with how the analysis is done is the range of interaction distances studied, which is mostly irrelevant to protein structure. The authors measure the cumulative distribution of the number of $C_\alpha$-$C_\alpha$ interactions over the range 0–60 Å for specific residue types and fit this to simple sigmoids. The fits show no clear residue dependence at the range of interaction distances typically known to be relevant to protein structure and folding (0–10 Å). This is not surprising since the fits are dominated by the data at long distances and would look very similar for scrambled sequences, implying that folding is sequence-independent. Most physical interactions between amino acids, however, are mediated by direct contact between residues or perhaps one or two water molecules or hydration shells. Specifically, van der Waals interactions between different atom types have most of their potential well in the range 2.5–5.0 Å, with vanishingly weak interactions at distances longer than 5.0–7.0 Å (2). The limit for hydrogen bonding is around 4.2 Å (3). Coulomb interactions are screened due to the high dielectric constant of the solvent but may still have a somewhat longer range (up to 10–15 Å). Translated to $C_\alpha$-$C_\alpha$ distances, anything beyond 15–20 Å should thus in most cases be ignored when calculating statistics over residue-residue interactions.

The distributions seen in the paper are an effect of general protein geometry and the natural frequencies of the different amino acids. The number of residues within a given sphere should scale as the volume of that sphere. At some point the sphere will be larger than the protein and no matter how big you make it no more residues will be counted. Hence the sigmoid shape of the distribution. These are the gross features of the distribution, and they would be the same regardless of structure and also for scrambled sequences. However, as mentioned, any effects due to specific interactions are expected in the range 0–20 Å. Mittal *et al.* acknowledge this and zoom in to the region 0–10 Å and find that the fitted curves are very similar for all amino acids. The clear deviations of the data from the fitted curves are however not discussed. To illustrate the significance of the deviations we have plotted the probability density function (instead of the distribution function) for Leucine contacts, normalized by a factor $1/r^2$ to remove the effect of the number of possible contacts growing as the distance increases, in figure 1A. From this figure it is clear that there are strong deviations from the behavior described in the paper. First of all there are irregularities in the form of peaks

[*]Corresponding Author:
Igor N. Berezovsky
Phone: +47-55584712
Fax: +47-55584295
E-mail: Igor.Berezovsky@uni.no

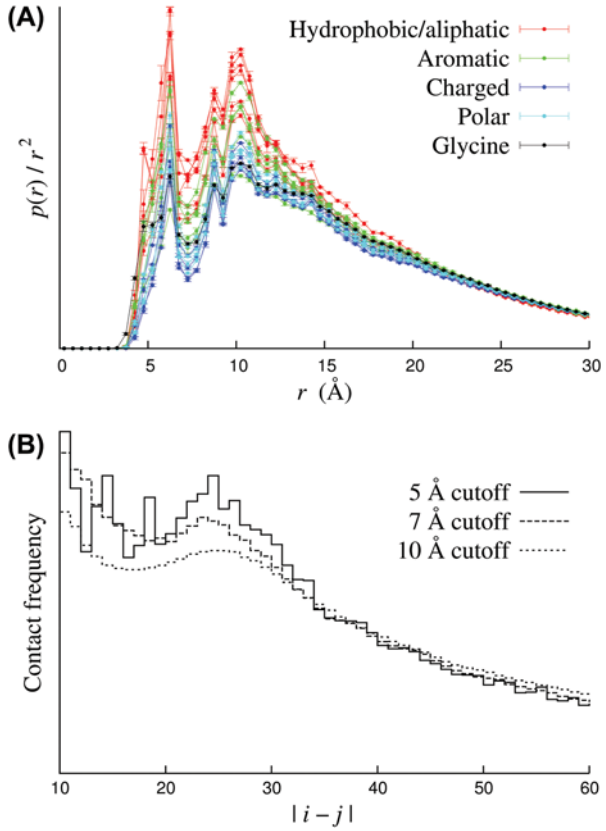


**Figure 1:** (**A**) $C_\alpha$-$C_\alpha$ contacts between Leucine and other residues. Residues are classified as follows, hydrophobic/aliphatic: Ala, Val, Ile, Leu, Met and Cys; aromatic: Phe, Tyr, Pro, Trp and His; charged: Arg, Lys, Asp and Glu; polar: Asn, Gln, Thr, Ser. Error bars are calculated by dividing the set of proteins into three and calculating the standard deviation between the distributions in each bin. (**B**) $C_\alpha$-$C_\alpha$ contacts between residues of different distance along the chain. The plots are calculated from a set of 4135 proteins with 30 % max sequence identity, 1.8 Å or better resolution and a R-factor cutoff 0.25. The list of PDBs was downloaded from PISCES September 3 2010 (26).

corresponding to contacts in the secondary structures. Second, and more importantly in this context, there is a strong preference for Leucine to make contacts with hydrophobic residues rather than polar or charged ones. Above 20 Å the different curves are identical, *i.e.,* when there is no physical interaction there is no preference for any certain residue to be in a contact with a given amino acid. This result holds for contacts between any amino acid types, yielding favorable interactions between hydrophobic amino acids as well as between polar/charged ones and unfavorable interactions between hydrophobic and polar/charged residues (data not shown). This is a well-known fact and the contact preferences between different amino acid types were discussed in detail by for example Miyazawa and Jernigan (4). At relevant spatial distances between amino acids (up to 10–15 Å), the identities of the interacting residues do matter, and folding thus depends on the order of the amino acids in the sequence, contrary to what is implied by the conclusions of Mittal *et al*. In this connection it may be noted that the preferential interactions between amino acids served as the basis for the development of knowledge-based potentials, which in turn provided a foundation for modern day protein structure prediction by modeling and simulation (5, 6 and references therein).

The polymer nature of proteins, which prompted the authors to their exercise, is indeed one of the major determinants of protein structure, evolution and folding. Figure 1B contains the distribution of closed loop sizes, *i.e.,* the distance along the sequence given that two $C_\alpha$-atoms are within 5, 7, and 10 Å of each other, respectively. The distribution yields a preferential size for closed loops (7), or chain returns, at around 25–30 amino acid residues. It has been shown exhaustively that closed loops are a universal basic unit of folds/domains of soluble proteins (8), and any protein globule can be represented as a combination of closed loops, which tightly follow one another forming a loop-n-lock structure (9). It was also hypothesized that closed loops are units of co-translational protein folding, which consists of the sequential loop closure and formation of the hydrophobic core (10, 11). The importance of closed loops for the structure and stability of soluble proteins stems from their evolutionary history (12). Specifically, closed loops are apparently descendants of primordial ring-like peptides, the first proteins with primitive functions. These peptides were brought together by gene fusion into different combinations, thus forming complex enzymatic functions (12-14). High-throughput analysis of proteomic sequences allows a reconstruction of closed loop prototypes (13), and thus an alphabet for modern proteins using these prototypes (14). Specific signatures of elementary chemical functions can also be determined for prototypes, turning them into prototypes of elementary functional loops (EFLs). An exhaustive description of a protein fold and its enzymatic function as a combination of EFLs illuminates how this fold emerged in pre-domain evolution by fusion of prototype genes and opens an opportunity to (re) design folds with desired functions by combining EFLs. A rigorous computational procedure for deriving prototypes of EFLs has recently been developed (15).

We believe that our demonstration of some basic results and the discussion thereof reveals the nature of the misconceptions in the paper by Mittal *et al.* and brings one back to the concepts and ideas dominating the protein folding "endeavor" for almost half a century. These studies started from Anfinsen's experimental observations and thermodynamic hypothesis (16, 17), formalized and rigorously described in theory (18), and are still full of problems to be resolved. The theory of protein folding has reached a mature state where the general mechanisms and driving forces are understood to a large extent (18). Due to a fine balance between energy and entropy the stability of most proteins is only marginal, which means

that the process is extremely sensitive to the details of the different interactions involved. Hence formulating a realistic unbiased energy function that can be used to describe protein folding is a difficult and unsolved problem (19), although some progress has been made (20-25). With improved energy functions and simulation techniques, and modern state-of-the-art experimental approaches, one can address new exciting problems in protein folding, such as co-translational folding, folding in crowded cellular environments, chaperone assisted folding and refolding, folding of large multidomain proteins and membrane proteins, and folding upon binding for intrinsically disordered proteins.

This work was supported by FUGE II (Functional Genomics) Program allocated through Norwegian Research Council.

***References***

1. Mittal, B. Jayaram, S. Shenoy, and T. S. Bawa. *J Biomol Struct Dyn 28*, 133-142 (2010).
2. G. Nemethy, M. S. Pottle, and H. A. Sheraga. *J Phys Chem 87*, 1883-1887 (1983).
3. D. F. Stickle, L. G. Presta, K. A. Dill, and G. D. Rose. *J Mol Biol 226*, 1143-1159 (1992).
4. S. Miyazawa and R. L. Jernigan. *Macromolecules 18*, 534-552 (1985).
5. P. Sklenovský, M. Otyepka. *J Biomol Struct Dyn 27*, 521-539 (2010).
6. M. J. Aman, H. Karauzum, M. G. Bowden, and T. L. Nguyen. *J Biomol Struct Dyn 28*, 1-12 (2010).
7. I. N. Berezovsky, A. Y. Grosberg, and E. N. Trifonov. *FEBS Letters 466*, 283-286 (2000).
8. I. N. Berezovsky. *Prot Engineering 16,* 161-167 (2003).
9. I. N. Berezovsky and E. N. Trifonov. *J Mol Biol 307*, 1419-1426 (2001).
10. I. N. Berezovsky, A. Kirzhner, V. M. Kirzhner, and E. N. Trifonov. *Proteins 45*, 346-350 (2001).
11. I. N. Berezovsky and E. N. Trifonov. *J Biomol Struct Dyn 20*, 5-6 (2002).
12. E. N. Trifonov and I. N. Berezovsky. *Curr Opin Struct Biol 13*, 110-114 (2003).
13. I. N. Berezovsky, A. Kirzhner, V. M. Kirzhner, V. R. Rosenfeld, and E. N. Trifonov. *J Biomol Struct Dyn 21*, 317-325 (2003).
14. I. N. Berezovsky, A. Kirzhner, V. M. Kirzhner, and E. N. Trifonov. *J Biomol Struct Dyn 21*, 327-339 (2003).
15. Goncearenco and I. N. Berezovsky. *Bioinformatics 26*, i497-i503 (2010).
16. C. B. Anfinsen. *Science 181*, 223-230 (1973).
17. C. B. Anfinsen and H. A. Scheraga. *Adv Prot Chem 29*, 205-300 (1975).
18. E. I. Shakhnovich. *Chem Rev 106*, 1559-1588 (2006).
19. T. Yoda, Y. Sugita, and Y. Okamoto. *Chem Phys Lett 386*, 460-467 (2004).
20. J. S. Yang, W. W. Chen, J. Skolnick, and E. I. Shakhnovich. *Structure 15*, 53-63 (2007).
21. C. Zong, G. A. Papoian, J. Ulander, and P. G. Wolynes. *JACS 128,* 5168-5176 (2006).
22. Irback, S. Mitternacht, and S. Mohanty. *PMC Biophysics* 2, 2 (2009).
23. F. Ding, D. Tsao, H. Nie, and N. V. Dokholyan. *Structure 16*, 1010-1018 (2008).
24. Y. A. Arnautova, A. Jagielska, and H. A. Scheraga. *J Phys Chem B 110*, 5025-5044 (2006).
25. J. W. Ponder and D. A. Case. *Adv Prot Chem 66*, 27-85 (2003).
26. G. Wang and R. L. Dunbrack, Jr. *Bioinformatics 19*, 1589-1591 (2003).

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


***Comment***

# Protein Folding: A Few Random Thoughts

http://www.jbsdonline.com

**Sridevi Akella**
**Chanchal K Mitra***

University of Hyderabad, Hyderabad 500046 India

The basics, or the foundations, of biology, *i.e.,* molecular biology, have been explored in such great detail that one often wonders what is still left to be discovered. This has been particularly fueled by the wide availability of tools and kits for the direct manipulation of living systems at the molecular level. Genetic manipulations are now routine, though still highly labor intensive. Unfortunately, even though we know biology at the molecular level quite well, we are still unsure what makes it work. I often think that the molecular biology is really simple and it plays a relatively minor role in the living system. It is the *organization* and the *hierarchical structure* of the living system that makes the living system unique (and worth the name "living"). The question that begs today is "how the molecular biology at the lowest level dictates the terms and conditions for the higher level?" I *think* the answer is "it doesn't."

Perhaps the question is badly asked. This problem can be seen very clearly in the classic question of protein folding. We have seen in the last quarter century a phenomenal growth in the computing power and we thought that the answer is going to come tomorrow. It is rather high time that we take a fresh look. How does nature figure out how to fold a protein without using a supercomputer? Is our current approach or the understanding wrong? Perhaps we have already reached the dead end, but are afraid to admit? Do we need a radically different approach? What are the right questions to ask? It was presumed that once the sequences of a large number of proteins are determined, the clear answer would come out naturally. When it did not happen, we thought that once we get the 3-D structures of a sufficiently large number of proteins, we should know the answer. But the answer still remains elusive.

In a recent paper in this *Journal*, Mittal *et al*. discard the conventional wisdom and took a new look at the same old problem (1). In brief, they have discarded the R-groups of the amino acids and have taken the bare backbones. Next they have looked at the distribution of the various R-groups from random points. They have finally plotted the results (for the 3718 proteins taken from the PDB database) as a function of the distance. If we normalize the results by the relative abundances of the various R-groups, then we get one sigmodial looking curve. In Figure 2 of Mittal *et al.* the four sets are remarkably same (after the normalization). But then it is an average behavior. How can we conclude the behavior of an individual from the overall population (2)?

***Seeing is Believing***

In Figure 1 (top left), we see the rasmol plots for lysozyme, with all the atoms (water molecules have been removed) in a ball and stick form. The same molecule, plotted for the backbone atoms only (all the backbone

*Corresponding Author:
Chanchal K Mitra
Phone: +91 40 2301 0814
E-mail: c_mitra@yahoo.com

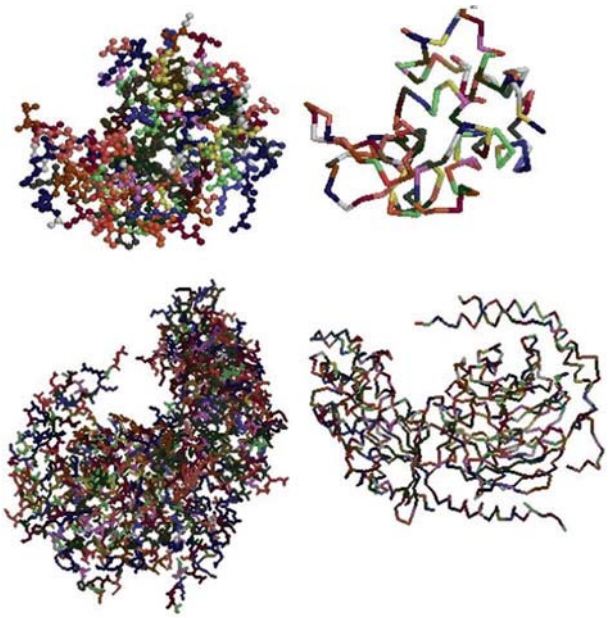

**Figure 1:** Rasmol plots of lysozyme (top panel), a small protein by common biological standards and G-protein (bottom panel), a large membrane bound protein. The full molecules (without the ligands) are shown on the left and the backbones (not the Cα alone) are shown on the right. How bare the backbones look!

atoms, not the Cα alone) can be seen on the right (top right). Lysozyme is a small protein (PDB ID 3IJV)- well studied and well characterized and is considered a hydrophilic protein (found in the cytosol)- a *typical small representative protein*. The other protein, a trimeric G protein, responsible for signal transduction, is shown on the bottom panel (PDB ID 3AH8). It is a monster hydrophobic protein (associated with membranes) seen in a ball-and-stick model on the bottom left and just the backbone on the bottom right. Qualitatively, they share the same features, viz.,

1. loosely packed structures; and
2. poor approximation to a sphere.

Globular proteins need some flexibility in their conformational space for effective function. Binding to their substrates or making a signal is not possible with a static structure (fibrous proteins do not perform any such task and can be very tightly packed). We therefore perhaps understand the loosely packed structure of the proteins. But why do we want them to be spherical at all? The computationally folded structure is often a good approximation to the real structure which is partly dynamic in nature. For a protein to function well, it must have several rigid regions and a few flexible regions. Often it is possible to guess the rigid and flexible regions by a careful observation of the 3-D structure of the protein. It is very difficult to guess the rigid and flexible regions of the protein by looking at the backbone alone. *At some level, however, Mittal et al. are right: the two proteins do share some structural similarity (and with any other common proteins).*

Globular proteins are not perfect spheres but we can make some intelligent guess by considering them so. For a tightly packed sphere, the number of neighbors at a distance **r** from a (any) given point increases as $\mathbf{r^2}$ but as we approach the surface the number (of neighbors) decreases. For large values of **r**, the number is expected to be zero (as all the proteins have reasonable sizes). For a loosely packed sphere the increase is expected to be $\mathbf{r^x}$ where $1 < x < 2$ is some number that depends on the packing density. On a space-filling model, *most proteins appear* to be "reasonably" packed "tightly". The best packing is obtained for a regular structure (but the helix leaves an empty core and the sheet structures leave space for hydrogen bonding) but a structure containing turns and bends wastes space. From the graphs (Figure 2 in reference 1; the authors have plotted the integral of this quantity, i.e., they have considered all contacts lying upto the distance **r**), the value of **x** appears to be close to 1.5. In Figure 2 (below left), we have illustrated this concept in a different way, but as long as the origin is taken at a point well within the protein molecule, the results and conclusions remain valid. The shape of the curves (in Reference 1) is therefore unsurprising. Graphs in Figure 3 (see Ref 1) show very clearly the quadratic dependence on **r** (*i.e.,* the nearest number of contacts increases as $\mathbf{r^2}$ for small values of **r**; again we must allow for the integration that the authors present their results). Does it provide any new insight? I do not know at this point.

In Figure 2 (see inset), we have plotted a theoretical curve for a compact spherical molecule which looks very much similar to the graphs the authors have presented. For a non-compact molecule, the slope at the midpoint of the sigmoid curve will be reduced and this is in line with our basic postulates. In an earlier work we have shown that the amino acids in real protein sequences correlate with a fractal dimension that is broadly consistent with the present observations (3). The 3-D structures of proteins are also reported to be fractal in nature.

### *Textbook Information*

Virtually all text books (4) on biochemistry report that the polypeptide chain starts folding as soon as it emerges from the ribosome (the protein synthetic machinery of the cell). It may well be correct, but then we shall expect one end of the polypeptide chain deeply embedded inside the folded structure. The protein folding then will follow a path similar to a ball of thread wound carelessly (one end completely inaccessible). I find difficulty in this theory when I consider multimeric proteins with the chains "quite" nicely entangled. None of the available computational tools (for protein structure determination) start folding the chain from one end. Another difficulty lies with the role of the "lowest energy conformation"

that may not be the ultimate desired structure and we need "chaperones" to guide the protein to a decent functional form (5).

### *New Lamps for Old*

New hypotheses are always put on a higher level of test: it *must* accommodate all known facts. The very first theories are only expected to explain *most* of the facts with a hope that existing *inconsistencies* will be ironed out with a *better* understanding. We *believe* that every protein is unique and has a given role to perform. We also *believe* that there exists some global unifying principle that govern the process of protein folding. We have the following problems here:

1. Small changes (*e.g.,* replacement of one amino acid by another) can sometimes cause large changes in structure (this is not always true).
2. Most proteins (particularly the globular proteins) have very similar amino acid composition. One wonders how the amino acid composition (see Table I in Reference 1) can drive the overall folded structure.
3. A major part of the protein structure acts as a scaffold for the active site. The active site is composed of a relatively few amino acids. The scaffold is relatively insensitive to small changes in the structure but the active site is. Therefore key features of the active site may get lost in the bulk of the scaffold that holds the active site together.
4. How does the working chaperone(s) fit in this model?

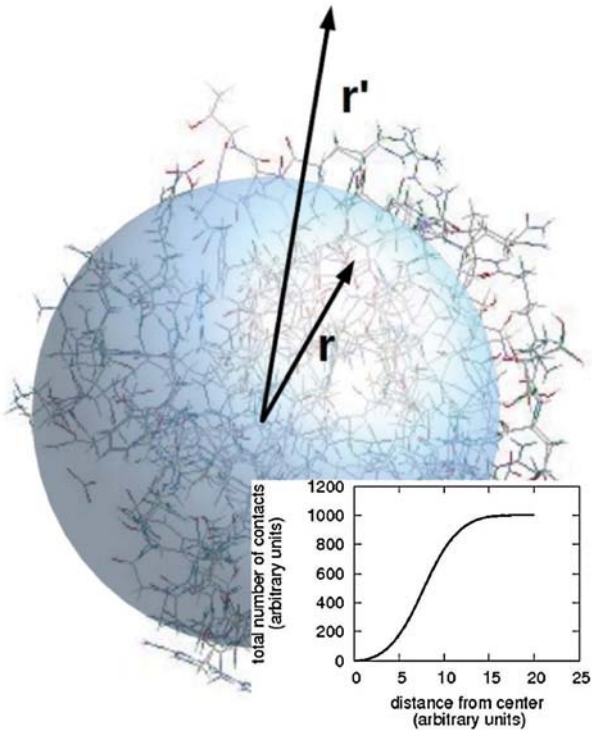


**Figure 2:** A compact approximately spherical molecule will enclose "number of contacts" that is proportional to $\mathbf{r}^3$ but once **r** becomes larger than the size of the molecule this cannot increase further (number of contacts cannot exceed the total number of atoms in the molecule). We therefore get a "sigmoidal looking" graph. The slope at the midpoint of the graph determines the compactness of the molecule and is related to the fractal nature of the molecular structure. The graph is plotted in gnuplot with smooth bezier options.

### *Thermodynamics of Protein Folding*

Protein folding is a spontaneous process (*e.g.,* many denatured proteins, particularly small ones, can be renatured by very slow and careful annealing) and the free energy change must be negative. However, the folded structure is ordered and has lower entropy (denatured proteins are disordered and have higher entropy). Therefore the decrease in entropy must be fueled by a *large* enthalpy change. The enthalpy change is driven by molecular interactions (including solvent). Small proteins have a relatively large surface and the solvent interaction plays a very significant role but for larger proteins (including but not limited to, membrane bound, or hydrophobic proteins) the role of the solvent is less (relatively speaking) and intramolecular interactions must drive the folding process. For larger proteins the total intramolecular interaction is large but is also distributed over a larger number of atoms. Very large proteins usually assemble in subunits. Experimental evidence for the "molten globule", the intermediate partially folded state, support this.

Computational tools (*e.g.,* molecular dynamics) are quite powerful and are widely available today (6, 7 and references therein). They employ a variety of modern techniques for geometry optimizations. However, the force field used for the energy computations is somewhat empirical and has been experimentally optimized. But the results clearly suggest the importance of the intramolecular interaction in the protein folding process (8).

### *Amino Acid Correlations and Cooperativity*

Unfortunately this is a neglected field. Many large ordered system have long range correlations. For protein folding, we define cooperativity as having several independent folding nuclei that come together and fuse into the final structure in a dynamical fashion (9). Experimental observations of the molten globule phase for the protein folding suggest this view. Also the success of the Chou-Fasman algorithm for the prediction of the secondary structures of protein *directly* supports the presence of correlations and cooperativity.

### *Homology Modeling*

That homology modeling (10-12 and references therein) works tells us that similar sequences have similar structures. If the sequences are shuffled, the process fails. According to the authors, any permutation of the sequence will preserve

the structure. Site directed mutagenesis suggests that a few mutations (at some sites) has no significant effect.

### *Conclusions*

The parametric equation used in the curve fitting in Figure 2 (ref: 1) has no sound theoretical basis. Simple geometrical consideration tells us that the number of points at a distance r will be proportional to $\mathbf{r^2}$ for small values (relative to the size of the protein) of **r** for tightly packed spheres and zero for large values of **r**. This can also explain the given curves but the exponent will be dependent on the packing density. The graphs in Figures 2 and 3 (ref:1) will almost perfectly superimpose if properly normalized by the amino acid distribution (either individually for each protein or even for the full set of 3718 proteins studied). As the equation used in the curve fitting the experimental data has no theoretical significance, the parameters **k** and **n** has no physical significance. Nevertheless, sigmoidal behavior (see Figure 4A) has been widely related to cooperative phenomena (*e.g.,* binding of oxygen to myoglobin follows a hyperbolic pattern, similar to the plot for n = 1 in Figure 4A whereas binding of oxygen to hemoglobin follows a sigmoidal curve). It is important to remember that phase transitions and order-disorder transitions also belong to cooperative phenomena. We strongly suspect that the authors have seen the correlations but have failed to notice. Table I, lists the overall percentages (mean and the standard deviation) for the 20 amino acids. However, I think that the standard deviations are not directly comparable and we should use the CV (coefficient of variation, $\sigma/\overline{x}$) for comparison. We notice that the CV for high abundance amino acids are relatively small and those for the low abundance amino acids are relatively high. These are probably related to the metabolic pathways.

Amino acid composition has evolved over a period of time and as more and more hydrophobic proteins are identified, isolated, purified and sequenced, we notice a small but steady increase in the proportions of the hydrophobic amino acids. Several amino acids are less common, (*e.g.,* W and H) perhaps because they place a strain on the metabolic resources. Several amino acids preferentially occur at the active site (S, H, W, *etc*.,) in the sequence. A large number of proteins are known to exist in the lipid phase and they *perform* important functions (13).

I believe the text book version of the hypothesis that the polypeptide chain starts folding as soon as emerges from the ribosome has merit. Protein folding is a dynamical process and dynamics must start the moment it emerges from the ribosome. However, these are the initial conditions and the path traveled (in the phase space and on the energy surface) by the growing polypeptide chain must somehow be constrained. How several polypeptide chains intertwine remains baffling. Perhaps it is high time we take a (or more) step back and take a fresh look at the holistic approach for the protein folding problem.

### *References*


1. A. Mittal, B. Jayaram, S. Shenoy, and T. S. Bawa. *J Biomol Struct Dyn 28*, 133-142 (2010).
2. G. D. Rose, P. J. Fleming, J. R. Banavar, and A. Maritan. *Proc Nat. Acad Sci. US103*, 14646-14651 (2006).
3. M. Rani and C. K. Mitra. *J Biomol Struct Dyn 13,* 935-944 (1996).
4. J. M. Berg, J. L. Tymoczko, Lubert Stryer, in Biochemistry (5th Edition; W. H. Freeman and Co, New York, New York, 2002, Chapter 3, pp. 64-69.
5. K. Huang. *Mod Phys Lett, B24(20),* 2113-2115 (2010).
6. M. J. Aman, H. Karauzum, M. G. Bowden, and T. L. Nguyen. *J Biomol Struct Dyn 28*, 1-12 (2010).
7. Y. Tao, Z. H. Rao, and S. Q. Liu. *J Biomol Struct Dyn 28*, 143-157 (2010).
8. J. Mittal and R. B. Best. *Biophysical J, 99*, L26-L28 (2010).
9. C. I. Lee and N. Y. Chang. *Biophys Chem 151*, 86-90 (2010).
10. A. Mahalakshmi and R. Shenbagarathai. *J Biomol Struct Dyn 28*, 363-378 (2010).
11. K. Bhargavi**,** P. Kalyan Chaitanya**,** D. Ramasree**,** M. Vasavi**,** D. K. Murthy**,** and V. Uma. *J Biomol Struct Dyn 28*, 379-391 (2010).
12. E. F. F. Da Cunha, E. F. Barbosa, A. A. Oliveira, T. C. Ramalho. *J Biomol Struct Dyn 27,* 619-625 (2010).
13. C. K. Mitra and M. Rani. *J Biosci 18,* 213-220 (1993).

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


***Comment***

# Two- and Higher Point Correlation Functions in Proteins

http://www.jbsdonline.com

**Martin Gruebele**

Departments of Chemistry, Physics, and Center for Biophysics and Computational Biology, University of Illinois, Urbana, IL 61801

In their paper in a recent issue of this *Journal*, Mittal and coworkers make some interesting observations about the spatial correlation between pairs of amino acids in proteins (1), by analyzing nearly 4000 proteins from the Protein Data Bank (2). They find that the cumulative pair distance probability distributions are well fitted by a universal sigmoid. The asymptotic value of the sigmoid simply reflects natural amino acid abundances.

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 (3). 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.

But when it comes to the origin of protein compactness, the devil is in the details, and the details do appear in the data. As can be seen in Figure 2 of Mittal *et al.* many of the probability curves intersect, and the detailed view in Figure 3 shows that in the 5-8 Å range, different types of amino acids have different deviations from the fit: Leu (a nonpolar residue) low, Asn (a polar residue) high. The deviations are small, but not unimportant. For a two-point correlation function, one expects the deviations to be small: the hydrophobic residues, most of which are packed in protein interiors (4), have evolved to accommodate about 4 side chain contacts in the folded state, making proteins almost three-dimensional solids (5). Thus 3- or 4-body correlations are required to fully reveal the ordering in the interior of a protein. This complexity is one of the reasons why current *ab initio* structure predictions do well at finding generally reasonable structure candidates, but not so well at getting small root-mean-square deviations from the experimentally measured structure (6). The deviations tell us what fraction of the interactions is specific to sidechains, *vs*. what fraction is generic (*e.g.,* entropically driven solvent exclusion).

Not surprisingly, the standard deviations of the stoichiometric averages in Table I of Mittal *et al.* are fairly large, comparable to the average value for many amino acids. Consider tryptophan for example: 1.3 ± 1.0%. One can find proteins much richer in tryptophan of course, such as the Tryptophan-rich basic protein that has 3.5% tryptophan in its 173 amino acid sequence (7). One can also find many proteins devoid of tryptophan. It may well be that the frequency of tryptophan reflects its late evolution and representation by a single codon, rather than a stoichiometric constraint on folding. Certainly 20 amino acids are not required to produce a functional folding alphabet (8).

Corresponding Author:
Martin Gruebele
E-mail: mgruebel@illinois.edu

In the end, Mittal *et al.* conclude that hydrophobicity drives the structure-nonspecific aspects of folding, as most workers in the field would agree: 'a protein pack[s]/fold[s] in an "exclusion by water" manner,' which basically defines the hydrophobic mechanism that the entropy of the solvent shell is greater than the entropy of water solvating interior residues of a protein.

***References***


1. A. Mittal, B. Jayaram, S. Shenoy, and T. S. Bawa. *J Biomol Struct Dyn 28*, 133-142 (2010).
2. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. *Nucleic Acids Research 28*, 235-242 (2000).
3. G. R. Grimmett and D. R. Stirzaker. 1992. Probability and Random Processes. Clarendon Press, Oxford.
4. A. Godzik. *Structure 4*, 363-366 (1996).
5. P. D. Chowdary and M. Gruebele. *J Phys Chem A 113*, 13139-13143 (2009).
6. J. Moult, K. Fidelis, A. Kryshtafovych, B. Rost, T. Hubbard, and A. Tramontano. Proteins-*Structure Function and Bioinformatics 69*, 3-9 (2007).
7. Murata, S. Degmetich, M. Kinoshita, and E. Shimada. *Development Growth & Differentiation 51*, 95-107 (2009).
8. L. R. Murhpy, A. Wallqvist, and R. Levi. *Protein Eng 13*, 149-152 (2000).

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


***Comment***

# Stoichiometry and Topology in Protein Folding

http://www.jbsdonline.com

**Ruxandra I. Dima**

University of Cincinnati Department of Chemistry, University of Cincinnati, Cincinnati, OH 45221

A fundamental piece of the puzzle, that is the protein folding problem, refers to the balance between energetic and topological frustration. Energetic frustration results from the often conflicting requirements that hydrophobic residues need to reside inside the folded protein structure, while polar and charged residues prefer to reside on the surface of the structure in close contact with water. Moreover, even if the energetic frustration for a particular protein sequence is substantially minimized, the requirement of almost perfect solid-like packing for a protein structure leads to topological frustration manifested in conflicting geometrical orientation of the various parts of the sequence due to the chain connectivity. While topological frustration is relatively straightforward to study by reducing the protein chain to a homopolymeric description, understanding the origin and degree of energetic frustration in proteins is far more challenging.

Mittal *et al.'s* study (1) addresses the energetics of the protein folding problem and therefore the energetic frustration issue (2). Starting from a large and diverse set of protein structures present in the PDB (3), the authors aim to uncover the driving force between the formation of amino acid pairs in functional protein states: is it the stoichiometry or the chemistry (mainly the hydrophobicity) of the members? Their study lands in the middle of a long standing controversy in the protein folding field that started with Kauzmann's 1959 review in which the author argued that the stabilization of a protein structure is largely due to the hydrophobic effect (4). By contrast, according to this view, hydrogen bonds have little or maybe even opposing influence on the protein folding reaction. While this is still an influential point of view, alternative proposals are gaining ground. Recent experimental findings and theoretical modeling indicate that osmolytes, which can dramatically influence the folding/unfolding processes, target primarily the protein backbone (not the side-chains) and act on the unfolded state rather than on the native state (5, 6), and that most mesophilic proteins unfold or fold under very similar denaturating/renaturating conditions (temperature or chemical denaturation) (7). It is also now well known that the number of stable domains is limited. Recently, Rose and collaborators (8) showed that two-state folding implies that conformation and stability are separable. Based on these findings, as well as the success of the tube-like model of proteins (9, 10) in proving that the native conformations of proteins can emerge on the basis of geometry and symmetry, the view that hydrogen bonding, rather than hydrophobicity, plays the central role in the protein folding process received an unprecedented push (8). According to this viewpoint, "side chains serve to select conformations from the limited repertoire of possible backbone conformations: alpha-helix, beta-strand, turns, and loops." This viewpoint puts special emphasis on the protein backbone for the protein folding process resulting in the proposal that water is a poor solvent for the protein backbone. This proposal received backing from a study of intrinsically disordered proteins (11)

Corresponding Author:
Ruxandra I. Dima
Phone: +1 513 5563961
Fax: +1 513 5569239
E-mail: Ruxandra.Dima@uc.edu

which revealed that, in water, even polyglycine chains that lack side-chains adopt compact but disordered structures due to the preferential formation of backbone-backbone hydrogen bonds over the formation of backbone-water hydrogen bonds. However, the same study suggested that additional factors, beyond backbone hydrogen bonding, are needed to rationalize their findings. The authors proposed that the negative entropy term seen in the collapse of short polypeptide chains becomes increasingly unfavorable for the hydration of long, flexible chains. Thus, long protein chains collapse to minimize the entropic cost of solvent organization around a heterogeneous ensemble of loosely packed conformations making the intrinsic chain flexibility a central player in setting the length scale of the collapse. Another protein folding viewpoint is that 'reduced-alphabet solvation-based' codes correctly encode native protein structures (12). This viewpoint is supported for example by the finding that protein structures can be designed using non-biological backbones (13).

Mittal *et al.'s* mathematically appealing study (1) adds support to the hydrogen-bonding centric viewpoint. Namely, their central finding that the parameters of the mathematical function that describes the spatial distribution of the total number of pairs of amino acids in native protein structures are independent of the chemical identity of the residues and of the size of the protein strongly suggests that side-chains do not play the main role in achieving packing in proteins. Unfortunately, the predictive power of this study is severely limited. For example, one limitation of their approach is that it does not distinguish between orientation-dependent and orientation-independent contacts, *i.e.,* between hydrogen bonds and other non-covalent contacts. As a result, the authors cannot provide direct insight into the factors that drive the collapse of a protein sequence into a compact structure. A different approach is needed to this end, such as one that accounts for the dynamics of the folding process as employed by Hubner and Shakhnovich (14). These authors showed that, in a model protein system, classical potentials accounting for directional hydrogen bonds formation and van der Waals interactions that promote overall compaction lead to parts of the sequence folding into alpha-helices. By contrast, exclusion of the hydrogen bond contribution while retaining the van der Waals contribution does not lead to the formation of secondary structure (alpha-helices) in their model (14). Another limitation of the Mittal *et al.'s* study is that their method cannot be extended to predicting the role played by side-chains in the selection of the final, well-folded conformation, which is supported by the hydrogen-bond centric viewpoint (8). Studies have revealed that relying entirely on the poor solvent quality of water for the protein backbone leads to overly compact structures that deviate from the native conformations of real proteins thus indicating that side-chains are required to achieve the final folded structure (15). One reason why Mittal *et al.'s* methodology (1) cannot reveal the role of side-chains results from one of the main assumptions behind their approach: by pulling together data about spatial distribution of residue pairs in a large number of diverse protein structures, the authors implicitly assume that any native protein structure can be assembled from a "soup" of amino acid pairs at various spatial separations obtained by the decomposition of many structures (16). While the backbone hydrogen-bond centric viewpoint recognizes that proteins are build on scaffolds of secondary structure elements (alpha-helices and beta-strands) (8), further decomposition of this scaffold into amino acid pairs is not supported by the available experimental and theoretical literature. This is not surprising as this assumption completely neglects the chain connectivity which is known to give rise to the topological frustration in protein folding. In conclusion, Mittal *et al.'s* work is a nice exercise in the use of a large set of protein structures to map out characteristics of the underlying amino acid distributions, but unfortunately with limited predictive power.

***Acknowledgement***

This work was partially supported in part by the National Science Foundation grant MCB-0845002.

***References***

1. A. Mittal, B. Jayaram, S. Shenoy, and T. S. Bawa. *J Biomol Struct Dynam 28*, 133-142 (2010).
2. C. Hyeon, and D. Thirumalai, *J Phys Cond Mat 19*, 113101-113127 (2007).
3. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. *Nucleic Acids Res 28*, 235-239 (2000).
4. W. Kauzmann. *Adv Protein Chem 14*, 1-63, (1959).
5. M. Auton, and D. W. Bolen. *Biochemistry 43*, 1329-1342 (2004).
6. T. Y. Lin, and S. N. Timasheff. *Biochemistry 33*, 12695-12701 (1994).
7. M. D. Kumar, K. A. Bava, M. M. Gromiha, P. Prabakaran, K. Kitajima, H. Uedaira, and A. Sarai. *Nucleic Acids Res 34*, D204-D206 (2006).
8. G. D. Rose, P. J. Fleming, J. R. Banavar, and A. Maritan. *Proc Nat. Acad Sci USA 103*, 16623-16633 (2006).
9. A. Maritan, C. Micheletti, A. Trovato, and J. R. Banavar. *Nature 406*, 287-290 (2000).
10. T. X. Hoang, A. Trovato, F. Seno, J. R. Banavar, and A. Maritan. *Proc Natl Acad Sci USA 101*, 7960-7964 (2004).
11. H. T. Tran, A. Mao, and R. V. Pappu. *J Am Chem Soc 130*, 7380-7392 (2008).
12. K. A. Dill, S. B. Ozkan, T. R. Weikl, J. D. Chodera, and V. A. Voelz. *Curr Op Struct Bio 17*, 342-346 (2007).
13. B. C. Lee, R. N. Zuckermann, and K. A. Dill. *J Am Chem Soc 127*, 10999-11009 (2005).
14. I. A. Hubner and E. I. Shakhnovich. *Phys Rev E 72*, 022901-022901 (2005).
15. R. I. Dima and D. Thirumalai. *J Phys Chem B 108*, 6564-6570 (2004).
16. R. I. Dima, G. Settanni, C. Micheletti, J. R. Banavar, and A. Maritan. *J Chem Phys 112*, 9151-9166 (2000).

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


***Comment***

# Stoichiometry and Preferential Interaction: Two Components of the Principle for Protein Structure Organization

http://www.jbsdonline.com

**Bin-Guang Ma***
**Hong-Yu Zhang***

National Key Laboratory of Crop Genetic Improvement, College of Life Science and Technology, Huazhong Agricultural University, Wuhan 430070, P. R. China

A paper recently published in *J Biomol Struct Dyn* by Mittal and coworkers proposed an interesting view that protein folding is driven by the stoichiometry of amino acids rather than by the "preferential interactions" between them (1). By analyzing about 3700 backbones of folded proteins, the authors argued that something like Chargaff's rules exist in protein folding similarly as in the organization of DNA structures. The above conclusion was made by counting the numbers of contacts between two specific residues at a variable range of neighborhood distance and by showing that they all follow a similar sigmoid trend and the total number of contacts formed by one residue is excellently correlated with the percentage occurrence of that residue. The authors' conclusion highlights the importance of amino acid composition in the determination of protein folding as well as in the organization of the folded protein structures. This is generally true and in line with some previous reports where the correlation between protein folding rates and the amino acid composition was found as high as 0.7 (2) and the prediction accuracy of protein folding type (two-state or multi-state folding) based merely on the occurrence numbers of amino acids could be more than 80% (3). Considering the limited information (only the frequencies of amino acids) used in the predictions, the achieved accuracy is surprising. All these results emphasized the important roles played by the stoichiometry of amino acids in protein folding.

However, the above results are not a reason to deny the roles played by preferential interactions of amino acids in the organization of protein structures. A wealth of evidence showed that the protein functional conformations are accomplished by particular arrangement of amino acids. Firstly, the neighborhood occurrence of amino acids in the primary sequence is not random (4). Secondly, it is well known that amino acids preferentially occur in different types of secondary structures (α-helix, β-sheets and coils, *etc.,*) and the successful prediction of protein secondary structures just utilized this kind of information (5, 6). Thirdly, the stability of tertiary structures depends on the preferential interactions between specific types of residues such as aromatic, charged and hydrophobic residues for stacking, salt bridge and internal packing effects (7-10). Fourthly, the formation of quaternary structures owes to preferential contacts between amino acids in the subunit interface (11). Our recent work on the thermal stability of prokaryotic protein complexes also demonstrated the strategical use of charged residues in the subunit interface to adapt to an elevated environment temperature (12). A database called AAindex summarized the amino acid properties and many of the contained indexes or matrices involve the preferential interactions between amino acid residues (13).

*Corresponding Authors:
Bin-Guang Ma
Hong-Yu Zhang
Email: mbg@mail.hzau.edu.cn
zhy630@mail.hzau.edu.cn

Actually, the authors' methodology has complicated the scenario by a sigmoid fitting. To show if there are preferential contacting patterns between residues, some direct statistical tests (for example, $\chi^2$ test) are proper to check the deviation of amino acid pairing (contacting) frequencies from a random uniformity. Seen from the Figures 2 and 3 of the paper, the numbers of contacts between one particular residue and the other 20 (including itself) residues are not uniformly distributed. If $\chi^2$ tests were used, the results might be different. On the other hand, it has to be noticed that whatever the result is, it's just statistical, meaning that it is an averaged consequence over different types of protein classes, folds, secondary structure contents and primary sequences. However, for the functional conformation of each particular protein, the amino acids are arranged in a strategical or even subtle manner, far from promiscuous pairing. Universality obtained from averaged diversity doesn't mean it holds for every single molecule individually, which just shows the limitation of statistical approach.

Considering the roles of amino acid composition in the determination of protein folding rates (2), folding types (3) and protein-protein interactions (14), the importance of the stoichiometry information could never be neglected. The authors' finding of universal backbone spatial organization just enriched the store of evidence for this. Even with this finding, the "grand challenge" of accurate *ab initio* prediction for protein folding trajectory and final structure from the sole primary sequence still remains and cannot be solved all of a sudden. In their paper, the authors stressed the principle of minimizing the surface-to-volume ratio in protein folding. Although this principle might be true (at least for some globular proteins), the picture given by the authors that protein folding looks like fitting "Lego Blocks" while precluding any preferential interactions between these blocks is unrealistic. The authors mentioned in the last section of their conclusion that "This (protein folding, as we understand) must be done while satisfying the structural constrains of the primary sequence composition and constitution (*i.e.*, the order in which a given stoichiometry of amino acids appear)". Actually, as suggested earlier, the primary sequence can only provide the two sorts of information: composition and permutation (or in the authors' term, constitution) (15, 16). The permutation information is in fact a manifestation of the preferential amino acid interaction (adjacency) in the primary sequence. Unfortunately, in the given picture of protein folding, the authors missed this point (constitution constraint) that they just mentioned and arrived at a biased conclusion of stoichiometry driving protein folding without the contribution from preferential amino acid interactions. Considering the roles of amino acid composition in determining protein folding behaviors (2, 3) and protein-protein interactions (14) together with the importance of preferential amino acid interactions in the organization of different levels of protein structures (4-12, 17, 18), we can conclude that stoichiometry and preferential interactions of amino acids are two indispensable components of the principle for protein folding and structure organization. Such a conclusion makes a lot of sense because the preferential interaction between amino acids is the basis for the development of knowledge-based potentials, which in turn form the underpinning of modern day protein structure prediction by modeling and simulation (19-22 and references therein).

This work was supported by the National Key Project for Basic Research (2010CB126100) and the National Natural Science Foundation of China (30870520).

***References***

1. A. Mittal, B. Jayaram, S. Shenoy, and T. S. Bawa. *J Biomol Struct Dyn 28*, 133-142 (2010).
2. B.-G. Ma, J.-X. Guo, and H.-Y. Zhang. *Proteins 65*, 362-372 (2006).
3. B.-G. Ma, L.-L. Chen, and H.-Y. Zhang. *J Mol Biol 370*, 439-448 (2007).
4. X. Xia and Z. Xie. *Mol Biol Evol 19*, 58-67 (2002).
5. D. Frishman and P. Argos. *Protein Engineering 9*, 133-142 (1996).
6. B. Rost. *J Struct Biol 134*, 204-218 (2001).
7. S. K. Burley and G. A. Petsko. *Science 229*, 23-28 (1985).
8. C. Vetriani, D. L. Maeder, N. Tolliday, K. S. Yip, T. J. Stillman, K. L. Britton, D. W. Rice, H. H. Klump, and F. T. Robb. *Proc Natl Acad Sci USA 95*, 12300-12305 (1998).
9. E. Querol, J. A. Perez-Pons, and A. Mozo-Vilarias. *Protein Engineering 9*, 265-271 (1996).
10. P. J. Haney, J. H. Badger, G. L. Buldak, C. I. Reich, C. R. Woese, and G. J. Olsen. *Proc Natl Acad Sci USA 96*, 3578-3583 (1999).
11. A. Anashkina, E. Kuznetsov, N. Esipova, and V. Tumanyan. *Proteins 67*, 1060-1077 (2007).
12. B. Ma, A. Goncearenco, and I. N. Berezovsky. *Structure 18*, 819-828 (2010).
13. S. Kawashima, P. Pokarowski, M. Pokarowska, A. Kolinski, T. Katayama, and M. Kanehisa. *Nucleic Acids Res 36*, D202-D205 (2008).
14. S. Roy, D. Martinez, H. Platero, T. Lane, and M. Werner-Washburne. *PLoS ONE 4*, e7813 (2009).
15. B.-G. Ma. *BioSystems 90*, 20-27 (2007).
16. B.-G. Ma. *Nature Precedings* <http://hdl.handle.net/10101/npre.2008.2223.1> (2008).
17. J. C. Biro. *Theoretical Biology and Medical Modelling 3*, 15 (2006).
18. M. Berrera, H. Molinari, and F. Fogolari. *BMC Bioinformatics 4*, 8 (2003).
19. P. Sklenovský and M. Otyepka. *J Biomol Struct Dyn 27*, 521-539 (2010).
20. M. J. Aman, H. Karauzum, M.G. Bowden, and T. L. Nguyen. *J Biomol Struct Dyn 28*, 1-12 (2010).
21. C. Koshy, M. Parthiban, and R. Sowdhamini. *J Biomol Struct Dyn 28*, 71-83 (2010).
22. Y. Tao, Z. H. Rao, and S. Q. Liu. *J Biomol Struct Dyn 28*, 143-157 (2010).

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


***Comment***

# Is Stoichiometry-Driven Protein Folding Getting Out of Thermodynamic Control?

http://www.jbsdonline.com

**Xing-Lai Ji[1,3]**
**Shu-Qun Liu[1,2,3*]**

[1]Laboratory for Conservation and Utilization of Bio-Resources & Key Laboratory for Microbial Resources of the Ministry of Education, Yunnan University, Kunming 650091, P. R. China
[2]National Laboratory of Macromolecules, Institute of Biophysics, Chinese Academy of Science, Beijing 100101, P. R. China
[3] Sino-duomedial and Information Engineering School, Northeastern University, Shenyang 110003, P. R. China

Understanding the mechanism by which monomeric proteins fold under *in vitro* conditions is fundamental to describing their functions at molecular level. Significant advances in theory, experiment and simulation have been achieved (1), making it possible to solve the three mostly focused aspects of protein folding problems (2):

(i) The thermodynamic question of how a native structure results from inter-atomic forces acting on an amino acid sequence - the folding code;
(ii) The kinetic problem of how a native structure can fold so fast - the folding rate;
(iii) The computational problem of how to predict the native structure of a protein from its amino acid sequence – the protein structure prediction.

The views on protein folding have evolved from simple force-driven folding (3), *i.e.,* the sum of many different small interactions (such as van der Waals interactions, hydrophobic interactions, hydrogen bonds, electrostatic interactions and ion pairs), to complex, free energy-driven "folding funnel" model (4) based on the energy landscape theory of protein folding (5). The latter is essentially a thermodynamically controlled process and emphasizes that folding is driven by complex balance of enthalpy and entropy leading to global free energy minimum for the protein-solvent system, rather than by simple optimization of inter-atomic forces only within the protein.

A recent article in this *Journal* (6) by Mittal *et al*. presents interesting statistical results based on 3718 folded protein structures, showing a simple principle of backbone organization of protein structure, which is interpreted as Chargaff's Rules related to protein folding, *i.e.,* a stoichiometry-driven protein folding. One of the interesting finding is that the total number of possible contacts for $C_\alpha$ of a given amino acid correlates excellently with its occurrence percentage in primary sequences, leading the authors to conclude that protein folding is a direct consequence of a narrow band of stoichiometric occurrences of amino-acids in primary sequences, regardless of the size and the fold of a protein (6). However, if this is true, what is the mechanism by which the percentage occurrences of amino acids determine protein folding? The authors do not answer definitely this question, although the folding manner of "exclusion by water" to minimize the surface-to-volume ratio, which is essentially equal to hydrophobic collapse hypothesis (7) for global protein folding, is proposed to relate the stoichiometric occurrences of amino acids to protein folding. It is hard to understand how the shape characteristics of individual residues can minimize the surface-to-volume ratio through constraints imposed by amino acid occurrence frequencies. One possibility is that

*Corresponding Author:
Shu-Qun Liu
Phone: +86 871 5035257
Fax: +86 871 5034838
E-mail: shuqunliu@ynu.edu.cn
shuqunliu@gmail.com

the higher occurrence frequency an amino acid has, the more probably it will occupy the core of a folded protein? Further work is needed to examine the relationship between the burial extent of amino acids and their occurrence percentages.

The "n" and "k" values for all the neighborhood sigmoids are independent of the occurrence percentages of amino acids, suggesting that there is no preferential interaction in short and medium distance ranges. Accompanying the absence of long range interactions, the authors concluded that the "preferential interactions" between amino-acids do not drive protein folding (6). It must be mentioned that preferential interactions between amino acids were the basis for introducing knowledge-based potentials, which in turn provided the underpinning for present day 3D protein structure prediction by modeling and simulation (8-10 and references therein). However, the authors' (6) conclusion of lack of preferential interaction between amino acids is drawn from analyses of 3718 already folded protein crystal structures. Therefore, it is easy to conclude the overall structural stability of already folded proteins can be maintained by the non-preferential/random inter-residue interactions, but the question that to what extent the non-preferential/random interactions contribute to the forces that drive protein folding is hard to answer based on current data.

The process of folding of a polypeptide chain, either newly synthesized from mRNA or denatured/unfolded from its native state, must be driven by certain forces such as hydrophobic side-chain interaction (11) or backbone hydrogen-bonding interaction (12). Interestingly, the hydrogen bond between protein backbone >C=O···H-N< has a potential to form between any two amino acids and can be considered as non-preferential interactions. Rose and colleagues (12) have recently proposed that the energetics of the backbone hydrogen bonds dominates the folding process, supporting the role of non-preferential interactions in driving protein folding. Nevertheless, the interactions between side-chain groups of two amino acids can be considered preferential because different amino acids have distinct side chains. At first glance, the "preferential interaction" of a given amino acid with another amino acid seems to come from specifically favorable side chain contacts such as hydrophobic stacking, hydrogen bonding (side chain-side chain hydrogen bond and side chain-backbone hydrogen bond) and electrostatic interactions. However, a further deep-thinking reveals that the so-called "preferential interaction" is to a large extent the consequence of protein desolvation effect (solute exclusion by water as mentioned in (6)) rather than specifically favorable side chain contacts. When an unfolded polypeptide chain interacts with the aqueous solvent under physiological conditions, the massive water molecules will exclude and squeeze the polypeptide to bring about the hydrophobic collapse (7, 13, 14). Upon collapse, the number of hydrophobic side chains exposed to water is minimized and the entropy of the solvent is maximized, thus lowering the total free energy of the protein-solvent system. Therefore the burial and packing of hydrophobic side chains, as the consequence of the hydrophobic collapse, increase the probability of observing hydrophobic interactions. Furthermore, it is interesting to note that not all polar or electrostatically charged side chains/groups are exposed to water, many of which can be buried inevitably in the interior of the folded structures. It has been suggested that the loss of stability by burying polar or charged groups can be gained back through forming hydrogen bonds (side chain-side chain hydrogen bonds or side chain-backbone hydrogen bonds) or salt bridges within the protein interior (15). The strength of hydrogen bonds depends on their environment; and therefore hydrogen bonds enveloped in a protein interior contribute more than those exposed to the aqueous environment to the stability of the native structure (16). This has led to the proposal that the protein folding is associated with a systematic desolvation of hydrogen bonds by surrounding hydrophobic groups (17). At stages after hydrophobic collapse but before reaching the native state, *i.e.,* the molten globular state (18) and the glass transition state (5, 19), further conformational rearrangements, which are obtained through favorable energetic contacts or preferential interactions between certain groups, are required to further lower the free energy of the protein-solvent system. The interactions between surface-exposed polar side chains and water molecules are not a negligible contributor to energetic enthalpy term of free energy change. Conclusively, the process of protein folding, which is driven by decrease in total free energy, is dictated by a delicate balance of the mechanisms of opposing effects involving entropic and/or enthalpic contribution.

For the folding process of an individual protein, the favorable/preferential interaction between any two amino acids would undoubtedly contribute to the enthalpic term of the free energy. If such a preferential interaction is "correct" (which means that the interaction is preserved in the final native structure), it contributes really to lowering free energy; if such a preferential interaction is "incorrect" (which means that the interaction is not presented in the final native structure), it contributes to the "trapped" free energy in the folding funnel of energy landscape. The entropic effect from solute and solvent and the competitive interactions (enthalpic effect) can help the protein jump out the "trap".

The statistical absence of preferential interaction between amino acids can be explained. On the one hand, the result is based on a large sample set of folded structures and therefore the simple count of number of $C_\alpha$ contacts within varied neighborhood distances would shield preferential interactions between side chain groups of amino acids. On the other hand, it is possible that the contacts between any two

amino acids, regardless of hydrophobic or hydrogen bonding/electrostatic interactions, can satisfy to some extent the requirement of lowering free energy during protein folding. Therefore, the observation of lack of preferential interactions can only be considered as a consequence of protein folding — a process that is driven by combined effect of enthalpy and entropy of the system — rather than the cause of protein folding.

Elucidating the folding mechanism is crucial for development of effective protein structure prediction methods, which in turn will improve our understanding of protein structure-function relationship and facilitate drug discovery and development. A very recent work by Sasisekharan and coworkers (20) reveals that the folding code is actually a network of inter-atomic interactions within the core regions of protein domains, and that the application of such a network signature to structure prediction has achieved great successes (for details, see (20)). This work also shows that each protein fold family has its own unique protein core atomic interaction network (PCAIN), implying that there must be preferential inter-atomic interactions. Such specific PCAIN is also the consequence of thermodynamically controlled folding process, and is not contradictory with the statistical result of the lack of preferential inter-residue interactions found by Mittal *et al.* (6) because most of the so-called preferential inter-atomic interactions can be observed between any two residues when the sample set is large enough.

In summary, we conclude that:

(i) The statistical method used by Mittal *et al*. is not sensitive to identify preferential/specific inter-atomic interactions.
(ii) The statistical phenomena observed in this work are the consequences of thermodynamics-driven folding rather than the driving force of protein folding.
(iii) It seems impossible to apply the "stoichiometry-driven" folding principle to protein structure prediction unless stoichiometric occurrences of residues can be translated into position constraint information.

***References***


1. D. Thirumalai, E. P. O'Brien, G. Morrison, and C. Hyeon. *Annu Rev Biophys 39*, 159-183 (2010).
2. K. A. Dill, S. B. Ozkan, T. R. Weikl, J. D. Chodera, and V. A. Voelz. *Curr Opin Struct Biol 17*, 342-346 (2007).
3. C. B. Anfinsen and H. A. Scheraga. *Adv Protein Chem 29*, 205-300 (1975).
4. P. E. Leopold, M. Montal, and J. N. Onuchic. *Proc Natl Acad Sci USA 89*, 8721-8725 (1992).
5. J. N. Onuchic, Z. Luthey-Schulten, and P. G. Wolynes. *Ann Rev Phys Chem 48*, 545-600 (1997).
6. A. Mittal, B. Jayaram, S. Shenoy, and T. S. Bawa. *J Biomol Struct Dyn 28*, 133-142 (2010).
7. P. Sklenovský and M. Otyepka. *J Biomol Struct Dyn 27*, 521-539 (2010).
8. M. J. Aman, H. Karauzum, M. G. Bowden, and T. L. Nguyen. *J Biomol Struct Dyn 28*, 1-12 (2010).
9. Y. Tao, Z. H. Rao, and S. Q. Liu. *J Biomol Struct Dyn 28*, 143-157 (2010).
10. V. R. Agashe, M. C. Shastry, and J. B. Udgaonkar. *Nature 377*, 754-757 (1995).
11. W. Kauzmann. *Adv Protein Chem 14*, 1-63 (1959).
12. G. D. Rose, P. J. Fleming, J. R. Banavar, and A. Maritan. *Proc Natl Acad Sci USA 103*, 16623-16633 (2006).
13. M. Brylinski, L. Konieczny, and I. Roterman. *Comput Biol Chem 30*, 255-267 (2006).
14. M. S. Cheung, A. E. Garcia, and J. N. Onuchic. *Proc Natl Acad Sci USA* 99, 685-690 (2002).
15. C. N. Pace, B. A. Shirley, M. Mcnutt, and K. Gajiwala. *FASEB J 10*, 75-83 (1996).
16. S. Deechongkit, H. Nguyen, E. T. Powers, P. E. Dawson, M. Gruebele, and J. W. Kelly. *Nature 430*, 101-105 (2004).
17. A. Fernandez, T. R. Sosnick, and A. Colubri. *J Mol Biol 321*, 659-675 (2002).
18. M. Ohgushi and A. Wada, *FEBS Lett 164*, 21-24 (1983).
19. J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes. *Proteins 21*, 167-195 (1995).
20. V. Soundararajan, R. Raman, S. Raguram, V. Sasisekharan, and R. Sasisekharan. *Plos One* 5, e9391 (2010).

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


***Comment***

# Discriminatory Power of Stoichiometry-Driven Protein Folding?

http://www.jbsdonline.com

**Mihaly Mezei**

Department of Structural and Chemical Biology, Mount Sinai School of Medicine, New York, New York 10029

There is a saying that when life hands you a lemon, you make lemonade out of it. I found that this has a parallel in molecular modeling: when your test of a putative diagnostic property fails to show discriminatory power, make an invariant out of it. The paper by Mittal, Jayaram, Shenoy and Bawa (1) is a brilliant example of this. Based on three independent observations, comparison of $C_\alpha$ distance distributions, small standard deviation (STD) of the amino-acid propensities and the comparison of amino-acid propensities in structured and unstructured proteins, the paper concluded that the relative frequencies each amino acid occurs in natural proteins is an important contributor to protein stability. This may not necessarily be surprising, but certainly it has not been thought to be the case. Given that these relative frequencies are found to be (more or less) constant, the authors call this set of relative frequencies stoichiometry. I will comment on these observations in the order of strength (as I see it).

In my opinion, the comparison of folded and unfolded proteins in (1) is direct evidence of the importance of the close adherence to the right stoichiometry (as defined by the experimentally observed relative frequencies of amino acids). However, this comparison also points to the fact that the right stoichiometry is only a necessary condition for protein stability but probably not a sufficient one, since it is known that the probability that a random sequence will fold is vanishingly small and restricting random sequences to the right stoichiometry may still leave many that would not fold.

As for the observed STDs of propensities, I was first mildly skeptical that they indicate the importance of this particular stoichiometry, since the values were not that small. This led me to the thought experiment: what values of STDs would one expect if the residues were chosen randomly, with the only restriction that the probability of selecting a give residue is its observed propensity? Selection of residue $r_i$ with probability $p_i$ is described with the binomial distribution whose STD is given as $np(1\text{-}p)$ where $n$ is the sample size. It turns out that the observed STDs are consistently smaller than the STD from the binomial distribution. Table I shows the data from Table I of (1) extended with the STD's calculated from the binomial distribution. On the average, the STD's from random sequences are 2.1 times larger than the observed values; the ratios range from 1.2 to 2.8.

The distribution of the Cα distances has been examined in great detail and it was found that they follow a sigmoidal distribution that can be described by three parameters only. While different residues suggest higher likelihood of interactions with specific partners, no such distinctions were found in the distributions. This was also true when the short range part of the distributions were compared – an

Corresponding Author:
Mihaly Mezei
Phone: (212) 659 5475
E-mail: Mihaly.Mezei@mssm.edu

**Table I**
The average percentage occurrence of each amino acid, their STD as observed and as calculated from the binomial distribution.

| | P (%) | STD (observed) | STD (random) |
|---|---|---|---|
| A | 7.8 | 3.4 | 7.2 |
| V | 7.1 | 2.4 | 6.6 |
| I | 5.8 | 2.4 | 5.5 |
| L | 9.0 | 2.9 | 8.2 |
| Y | 3.4 | 1.7 | 3.3 |
| F | 3.9 | 1.8 | 3.7 |
| W | 1.3 | 1.0 | 1.3 |
| P | 4.4 | 2.0 | 4.2 |
| M | 2.2 | 1.3 | 2.2 |
| C | 1.8 | 1.5 | 1.8 |
| T | 5.5 | 2.4 | 5.2 |
| S | 6.0 | 2.5 | 5.6 |
| Q | 3.8 | 2.0 | 3.7 |
| N | 4.3 | 2.2 | 4.1 |
| D | 5.8 | 2.0 | 5.5 |
| E | 7.0 | 2.7 | 6.5 |
| H | 2.3 | 1.4 | 2.2 |
| R | 5.0 | 2.3 | 4.8 |
| K | 6.3 | 2.8 | 5.9 |
| G | 7.2 | 2.8 | 6.7 |

important point since the long-range part is not expected to show significant residue-dependence. The sigmoidal shape in itself is not surprising – it is a consequence of the finite size of proteins.

Given that the importance of the right stoichiometry for protein stability has been amply demonstrated by the two arguments quoted above, I think that the residue neighborhood distributions are worth revisiting to see if there is a way to tease out data related to the contributions of specific interactions to protein stability. It is important to note what is at stake here: the development of knowledge-based potentials is based on preferential interaction between amino acids, and the determination of protein structure by modeling and simulation, based on preferential contacts and interactions, has become very common (2, 3 and references therein). There are two reasons why the $C_\alpha$ distance distribution is not the best measure of interaction specificity: (a) the cumulative distributions are dominated by the effect of the cubic dependence of volume with distance and (b) different side chains are of different size and thus the contact distances are different for different residue pairs. This suggests two additional ways of analyzing the residue-residue distances: (a) instead of the cumulative distributions, calculate radial distributions, *i.e.,* normalize by the volume available - this would generate distributions with peaks and troughs whose height and depth would be rather sensitive to small changes in propensities; and (b) calculate the number of residues of different kind that are in *contact* (*i.e.,* have at least one pair of heavy atoms within VdW distance) with the test residue.

The authors draw a parallel of their observation to the seminal observation of Chargaff related to the fixed ratio of nucleotides in DNA. That ratio soon found its explanation in the double helix structure. Completing the parallel between the Chargaff rules would require the detection of the structural or mechanistic origin of the importance of the right stoichiometry. The alternative analysis of residue neighborhoods suggested above would be one option. It would be also informative to compare the STD of the ratio of different type of residues with the STD expected from the individual residue propensity STDs.

***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).

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


***Comment***

# Proteins will Fold Anyway!!

**Indira Ghosh**

School of Computational & Integrative Sciences, Jawaharlal Nehru University, New Delhi 11 0067, India

http://www.jbsdonline.com

Protein folding is one of the most challenging physiological phenomena which has intrigued every discipline of research in science and several thousands researchers are working in the field to unveil the reason of such a complex formation in nature. One of the most discussed behavior of protein is its compactness in globular structure and the cause of the same (1, 2). Among the many different methods and conclusions, the most popular concept is that the protein interior is compact and has been so due to the presence of several selective amino acids, like polar and apolar groups, making it different in case of thermophillic vs. mesophillic proteins (3-7). Considering different physico-chemical parameters like residue hydrophobicity vs. atomic hydrophobicity most of the researchers (8, 9) has been providing a common guideline towards the protein's folding; that the balance of hydrophobicity vs. charge distribution inside the protein makes it to fold in stable form of the 3D structure. It is the accessibility of polar residues to water vs. the buried apolar group in one or other form (alpha vs beta or their combinations) that makes the protein to fold in its thermodynamically stable structure. Considering these simple two types of residues, Dill *et al.* (2) summarize the principle of protein folding: "The folding code is mainly binary and delocalized throughout the amino acid sequence. The secondary and tertiary structures of a protein are specified mainly by the sequence of polar and nonpolar monomers."

The recent paper by Mittal *et al.* (10) has come up with a database driven conclusion which has challenged this concept of protein folding to form a stable folded 3D structure. Their calculations at atomic level (counting neighborhood residues of each Cα atom within a spherical boundary) shows that " a direct consequence of a narrow band of stoichiometric occurrences of amino-acids in primary sequences, regardless of the size and the fold of a protein" drives protein folding. This raises several questions: Is there any significance influence of the sequence of amino-acids on the structure? Will any sequence fold in functional protein as long as it satisfies the contact between residues? What about single mutation and unfolding of proteins? Will any homopolymer fold only in one form of a conformer? But according to Dill *et al*. (2) "Homopolymers do not collapse to unique states".

More interesting finding is that the $C_\alpha$ neighborhood is dependent only on the frequency of occurrence of the aminoacids and not on any preferential interactions. This conclusion that "Preferential interactions between aminoacids" do not drive protein folding is incapable to explain that most of the hydrophobic residues come inside and hydrophilic residues come on the surface of Globular proteins in particular (11). In this connection it may be noted that the development of knowledge-based potentials is based on preferential interaction between amino acids, and the determination of protein structure by modeling and simulation, based on preferential contacts and interactions, has become very common (13-16 and references therein). The preference of occurrence of aminoacids in local secondary structures,

Corresponding Author:
Indira Ghosh
E-mail: indirag@mail.jnu.ac.in
ighdna@yahoo.com

hydrogen bonding, torsional angle preferences along the backbone, preferred folding patterns (class, architecture and topologies) appear to be redundant or just has to fall in line as long as the "Chargaff's rules for Protein folding" is satisfied.

The paper (10) mentioned the detailed method of considering the number of neighborhoods and specific residues as long as the Cα of the pair are within a spherical distance. The authors have considered only Cα atoms and assumed that every spherical volume will include the sidechain atoms of the counted Cα which may not be the case when they are polar sidechains. This overestimates the polar residues to be within the neighborhood and underestimates the apolar sidechains.

Truly the Cα positions considered in folding patterns need to take into account of spatial distribution *i.e.,* spherical coordinate space (R, theta, phi) but the authors here considered only R space (2D distance space, R). Spatial distribution of Cα positions from the centre of mass of the fold would be ideal rather taking any Cα as its centre and its neighbors. Cα does not indicate any type of aminoacid and its preferences for another aminoacids in the neighborhood unless one considers directional interactions (hydrogen bonds, electrostatic interactions) as well as the molecular interactions (hydrophobic and vanderwaals). These forces drive the folding patterns taking account of microenvironment of aminoacids, size, shape and compact packing to form the structural and functional domains in a protein. Many proteins may not be as compact as expected and might be dependent on their secondary structure as a recent study has shown (12).

Sigmoidal trends that were observed for neighborhoods of the twenty amino acids by the method of calculation described in the paper and dependence on their percentage of occurrences are expected observations and support the hypothesis of "Chargaff's rules". Though the considered dataset of 3718 proteins has covered all the protein classes, in depth analysis of the results reveals that this might be due to average effect of highly dispersed frequency of occurrence of amino acids in the total protein data bank considered. Using the data provided by the authors (2548 pdb data ), when one re-calculates the Table I (10), interestingly one finds that some of the apolar sidechain have a large variance, more than average values (shown in Table I below), mainly in Leu, Val & Ala. Also some discrepancies (frequency of Leu: 7.06 , Val: 8.92 & Ala: 7.8 as calculated using 2548 database) in the average values are noted; but beyond this one starts to think: is this result obtained by Mittal *et al.* due to averaging of high variance data?

Folding of protein is a thermodynamic phenomenon which can not be simplified at the atomic level and limited to Cα counting. Moreover, the database driven result has biased the conclusion towards the number of residues rather than physico-chemical characteristics of the residues in the packed protein. Nonetheless it brings out an obvious conclusion that the packing of residues makes the soluble proteins to have compact structures. But the propensity of amino acid in a protein is not the cause rather the result of folding in appropriate 3D structure required by its function.

**Table I**
Modifications to Table I in Mittal *et al.* (10).

Frequency of Amino Acid in Folded Proteins – Margin of Life (mean ± std, n = 3718)

| AA | freq ± std | variance of 2548 proteins |
|---|---|---|
| A | 7.8 ± 3.4 | 11.67 |
| V | 7.1 ± 2.4 | 9.24 |
| I | 5.8 ± 2.4 | 5.59 |
| L | 9.0 ± 2.9 | 6.03 |
| Y | 3.4 ± 1.7 | 2.89 |
| F | 3.9 ± 1.8 | 3.06 |
| W | 1.3 ± 1.0 | 1.2 |
| P | 4.4 ± 2.0 | 3.82 |
| M | 2.2 ± 1.3 | 1.65 |
| C | 1.8 ± 1.5 | 5.75 |
| T | 5.5 ± 2.4 | 4.83 |
| S | 6.0 ± 2.5 | 6.01 |
| Q | 3.8 ± 2.0 | 3.66 |
| N | 4.3 ± 2.2 | 4.39 |
| D | 5.8 ± 2.0 | 3.66 |
| E | 7.0 ± 2.7 | 7.54 |
| H | 2.3 ± 1.4 | 2.25 |
| R | 5.0 ± 2.3 | 5.61 |
| K | 6.3 ± 2.8 | 9.03 |
| G | 7.2 ± 2.8 | 6.89 |

***References***


1. M. Behe, E. Lattman, and G. Rose. *Proc Natl Acad Sci USA 88*, 4195-4199 (1991).
2. K. Dill, S. Bromberg, K. Yue, S. Fiebig, D. Yee, P. Thomas, and H. Chan. *Protein Sci 4*, 561-602 (1995).
3. S. Rackovsky and H. Scheraga. *Proc Natl Acad Sci USA 74*, 5248-5251 (1977).
4. L. Xiao and B. Honig. *J Mol Biol 289*, 1435-1444 (1999).
5. R. Das and M. Gerstein. *Funct Integr Genomics 1*, 76-88 (2000).
6. N. Kannan and S. Vishveshwara. *Protein Eng 1*, 753-761 (2000).
7. R. Banerjee, M. Sen, D. Bhattacharya, and P. Saha. *J Mol Biol* 333, 211-226 (2003).
8. P. Haney, J. Badger, G. Buldak, C. Reich, C. Woese, and G.Olsen. *Proc Natl Acad Sci USA 96*, 3578-3583 (1999).
9. A. Banerji and I. Ghosh. *Eur Biophys J* 38, 577-587 (2009).
10. A. Mittal, B. Jayaram, S. Shenoy, and T. S. Bawa. *J Biomol Struct Dyn 28*, 133-142 (2010).
11. J. Janin, S. Miller, and C.Chothia. *J Mol Biol 204*, 155-164 (1988).
12. A. Banerji and I. Ghosh. *PLoS ONE 4*(10) (2009).
13. P. Sklenovský and M. Otyepka. *J Biomol Struct Dyn 27*, 521-539 (2010).
14. M. J. Aman, H. Karauzum, M. G. Bowden, and T. L. Nguyen. *J Biomol Struct Dyn 28*, 1-12 (2010).
15. C. Koshy, M. Parthiban, and R. Sowdhamini. *J Biomol Struct Dyn 28*, 71-83 (2010).
16. Y. Tao, Z. H. Rao, and S. Q. Liu. *J Biomol Struct Dyn 28*, 143-157 (2010).

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


***Comment***

# Protein Structures-based Neighborhood Analysis vs Preferential Interactions Between the Special Pairs of Amino acids?

**Jihua Wang[1,2*]**
**Zanxia Cao[1,2]**
**Jiafeng Yu[1,2]**

[1]Key Lab of Biophysics in Universities of Shandong (Dezhou University), Dezhou 253023, China
[2]Department of Physics, Dezhou University, Dezhou 253023, China

http://www.jbsdonline.com

In the paper titled "A stoichiometry driven universal spatial organization of backbones of folded proteins: are there chargaff's rules for protein folding?" (1), extensive statistic works were performed on 3718 structured proteins to exploit the potential inherent laws in protein folding problems. Subsequently, protein structure-based neighborhood analysis (see Figure 1 in Mittal *et al.* paper) was conducted to the backbones of these protein dataset. The results show that there are similar sigmoidal trends for neighborhoods of all the 20 amino acids, and all the sigmoidals were fitted as a generalized, single, sigmoidal equation (see Figure 2 in Mittal *et al.* paper). Based on the results, the total number of contacts made by amino acids was found to correlate excellently with average occurrence of that amino acid in the folded proteins shown in Figure 2E. Thereout the authors deduced that there is no any preferential interactions between amino acids, and protein folding is a direct consequence of a narrow band of stoichiometric occurrences of amino acids in primary sequences regardless of the size and the fold of a protein. In this connection, it should be noted that the 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).

As is well known, the basic frame of protein folding (6-8) has been constructed for about half century researches since Pauling and Anfinsen. In 1997, the folding funnel hypothesis is depicted as a specific version of the energy landscape theory of protein folding (9). More and more experimental and theoretical studies have validated that the energy funnel model is a correct pathway to understand protein folding (10-14), which indicates that there is a preference pathway (energy driving or energy leaning) for protein folding that can explain Levinthal's paradox very well. It is unlikely that there is a single mechanism for protein folding (15). If there is a single mechanism of folding of module, it is nucleation-condensation, with framework and hydrophobic collapse mechanisms being extreme cases. According to energy funnel model and nucleation-condensation mechanism of protein folding, it is necessary that there are preference interactions between the special pairs of amino acids that firstly drive protein folding. It is apparent that this basic result on protein folding is inconsistent with the conclusions of Mittal *et al.* paper.

In our opinions, several conclusions in Mittal *et al.* paper need further deliberating. Firstly, the statistic methods and results are reasonable (Figure 1-5 in Mittal *et al.* paper), but the main problem is that the results obtained from the 3718 proteins are used to deduce the conclusions of an individual protein folding; Secondly, there are some logical shortcomings in preparing the dataset; Thirdly, the total

*Corresponding Author:
Jihua Wang
Phone: (+86)534-8985933
Fax: (+86)534-8985884
E-mail: jhwyh@yahoo.com.cn

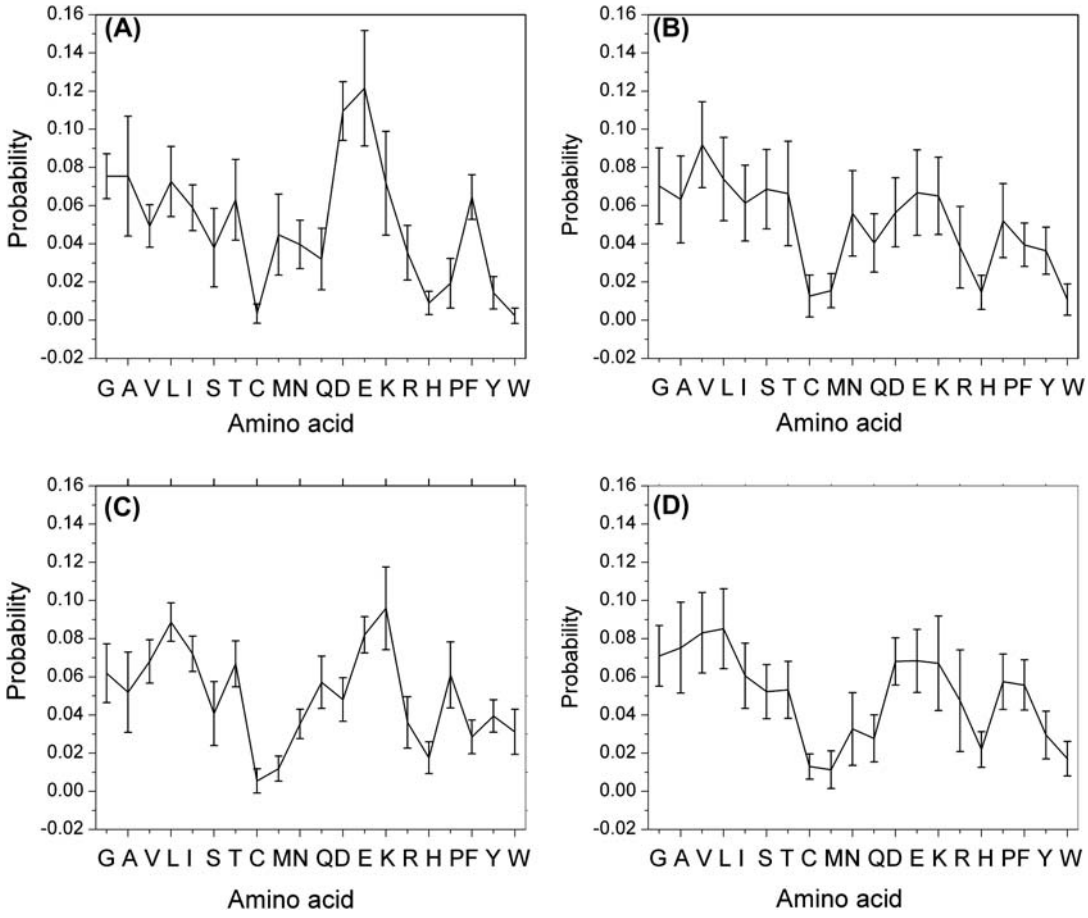


**Figure 1:** Average occurrence probabilities of the twenty kinds of amino acids in the primary sequence. Graphs (**A-D**) correspond to the structure class of α, β, α + β, α/β, respectively.

number of contacts has no clear biological functions, which is not enough to denote the complete information of protein folding. So it is unlikely to conclude that protein folding is a direct consequence of a narrow band of stoichiometric occurrences of amino acids in primary sequences from the excellently linear relation between the total number of contacts and the average occurrence of the amino acids.

### *On the Methods and Materials*

There are some logical shortcomings in preparing the dataset in Mittal *et al.* paper. In this paper, a final dataset composed of 3718 folded proteins are constructed, based on which extensive statistic works are performed. However, there are about 67, 981 proteins in current Protein Data Bank (PDB) (16), what is the criterion that the authors choose the 3718 proteins as statistic samples? There are thousands of families of proteins whose unique native structures have been determined. The total number of families in nature is much larger, and is estimated to be around 23,100. The large majority of these proteins are composed of complex patterns of the 20 kinds of amino acids. This pattern and compositional complexity has proven to be a significant hurdle to scientists attempting to crack the protein folding code or explore the sequence–structure relationship. In fact, it is well known that some amino acid residues are similar in physicochemical properties and their substitutions are tolerated in many regions of a protein sequence. Some previous works have demonstrated that large proteins can be designed with a reduced set of amino acids (17). On the other hand, the specific physiochemical properties of each kind of amino acid determined by its side chains play important roles in the diverse protein structures, different kinds of secondary structures, such as α-helix, β-sheet, *etc.,* have their preferential amino acids compositions. Moreover, amino acids compositions have been successfully used to discriminate different kinds of issues for proteins, such as protein subcellular location (18), protein-protein interactions (19), IDPs prediction (20, 21), *etc*., However, these 3718 items are calculated together, the results should be farfetched. Among these proteins, there are diverse functions and structure types. To give objective results, the adopted 3718 proteins should be classified into several subgroups according to their structure type.

The protein function is dependent on its three-dimensional structure and dynamics. Most proteins exist in unique conformations exquisitely suited to their function. Some methods (22-25) have been developed to predict the three-dimensional structure of protein which did not resolve by experimental

method based on the information of known structure protein in PDB. The quality of the training dataset strongly affects the accuracy of the methods being implemented (26). However, the Mittal *et al.* article did not describe the detail of how to generate the dataset, this made the result imprecise.

Generally, side chains of protein play important role in protein folding (27-29), however the side chains of the 3718 proteins have been removed in the neighborhood analysis of Mittal *et al.* paper, so some important information on protein folding has been lost.

Previous studies have indicated that protein folding is consistent with two-state folding (30) or three-state (31) or multi-state folding (32) and the last two models exhibit the obvious intermediate state during the folding process under the experimental conditions. The folding intermediate has some characters different from final structures, it is improper to use only the final structure to prove the folding problem.

On the other hand, many researches (33-37) have indicated that the passage of the folded state is mainly guided by hydrophobic interactions, intramolecular hydrogen bonds, van der Waals force and electrostatic interaction. Some key interaction residues play most important role in the form and stability of protein. It is improper to treat all contacts equally.

***On the Results and Conclusions***

Figures 2-5 in the Mittal *et al.* paper were obtained using the neighborhood analysis, which is logical and reasonable. However, the causality is confused in this paper. From the original primary sequence to a structured protein with biological function, the physiochemical interactions such as electrostatics, hydrogen bonding and hydrophobic interactions provide essential driving force. When a mature protein is born, all the amino acids are restrained in fixed regions. In Mittal *et al.* paper, the authors claimed that protein folding is a direct consequence of a narrow band of stoichiometric occurrences of amino acids in primary sequences regardless of the size and the fold of a protein. It is obviously not the case. Since the primary sequence of a bioactive protein is not randomly composed by the 20 amino acids, the usage bias of amino acids is expected. In other words, occurrences of amino acids are not the necessary and sufficient conditions for protein folding, but necessary conditions.

According to the definition in Mittal *et al.* paper, the total number of contacts of a amino acid (for short: $N_{tc}$) is sum of all 20 $Y_{max}$ values (the general, single equation of all the sigmoids: $Y = Y_{max}(1- e^{-kx})^n$), reflected by the sum of all asymptotes of the 20 sigmoids of that amino acid. What does the $N_{tc}$ means? the $N_{tc}$ only shows the total number of contacts in backbone of all the 3718 proteins, which has no clear biological function and did not depict all the complete information of protein folding, which could not reflect whether there are preferential interactions between two amino acids in a protein even though there are the interactions. Because the neighborhood distance (the most distance is 90 Å) almost cover the size area of all the 3718 proteins, it is clear that $N_{tc}$ should have close relationship with the average occurrence of the amino acid in the protein dataset. So it is certain that the total number of contacts were correlated excellently with the average occurrence of the amino acid from its definition (see Figure 2E in Mittal *et al.* paper). On the other hand, the total number of contacts has few biological function especially having no direct relationship with protein folding, $N_{tc}$ depicts some global property of protein and Figure 2E reveals global relationship between $N_{tc}$ and the average occurrence of the amino acid in the 3718 proteins, the result can not be used to explain an individual protein folding. Therefore, Figure 2E only reveals the fact that there is the defined total number of contacts has excellently linearity relationship with the average occurrence of the amino acid in the 3718 proteins, but it has not indicated whether there are preferential interactions between amino acids in an individual protein.

In the statistical results in Figures 2-4, number of contacts does not appear to support the authors' conclusions objectively. It is conceivable that the more percentage occurrence, the more contacts they have. If substituting the number of contacts with the ratio of number of contacts and occurrence numbers of corresponding amino acids, the figures may display some useful information better, and the results may be more objective.

Comparing Figure 5B with Figure 2E, the correlation of the former figure is weaker than the latter. However, the number of the adopted unstructured proteins is only 212, which is not sufficient enough to make an objective comparison. On the other hand, it seems that most of the 212 unstructured proteins are not complete structural disorder, but have some unstructured regions in part. The intrinsically disordered regions of "unstructured" proteins possess a highly flexible, malleable random coil-like structure. They lack specific tertiary structure or secondary structure, and are composed of an ensemble of highly heterogeneous conformations. The states of the "unstructured" proteins might be considered to represent the "liquid phase" of proteins (38). Intrinsically disordered proteins (IDPs) are highly, but not uniformly flexible. Different mobility in different regions may be linked to various functions, making the characterization of dynamics in such proteins highly desirable. Therefore, the highly flexible, malleable random coil-like structure of the "unstructured" proteins bring on the Ca of the backbone of the proteins have more fluctuant than the well-structure proteins, inducing that the correlation of the "unstructured" proteins is weaker than the well-structure proteins. Therefore, the statistics needs further discussion.

### *Analysis of Sequence Character of Four Different Classes of Protein Structure*

The Mittal *et al.* paper shows that the protein folding has a directly correlation with the frequency of occurrence of amino acids in the primary sequence. In order to verify this conclusion, we performed some analysis to see whether the same frequency of occurrence of amino acids in the primary sequence for proteins which have same structure character. We have selected four different proteins (PDB ID: 2NLN, 2VO8, 2K8E, 1ZZO) which corresponding to four different structure class ($\alpha$, $\beta$, $\alpha$+$\beta$, $\alpha/\beta$), respectively. Then Dali server (39) was used to find structures that have similar fold of the four different proteins in PDB. The frequencies of occurrence of twenty different amino acids for the four different structure classes have been analyzed (see Figure 1). It is obvious that the high standard deviations of the percentage occurrences of amino acids in the four different structure class proteins for some amino acids.

The essential fact of protein folding remains that the amino acid sequence of each protein contains the information that specifies both the native structure and the pathway to attain that state. However, this is not to say that nearly identical amino acid sequences always fold similarly (40). Furthermore, surroundings can influence conformations in most cases.

### *Summary*

Although extensive statistic works were performed in Mittal *et al.* paper, there some shortcomings and questionable conclusions need clarifying. As have pointed above, the main problem is that the results obtained from the 3718 proteins can not support their conclusions objectively. An appropriate dataset is prerequisite for statistic works, while there are some logical shortcomings in preparing the dataset, and the total number of contacts was not enough to show the complete information of protein folding. Therefore it is unlikely for the results to conclude that protein folding is a direct consequence of a narrow band of stoichiometric occurrences of amino acids in primary sequences, and it is unlikely to obtain that there is no any preferential interactions between amino acids for protein folding from the results in Mittal *et al.* paper.

We thank the financial support from the Chinese Natural Science Foundation and Shandong Natural Science Foundation (grant number 30970561 and 31000324, 2009ZRA14027 and 2009ZRA14028).

#### *References*

1. A. Mittal, B. Jayaram, S. Shenoy, and T. S. Bawa. *J Biomol Struct Dyn 28*, 133-142 (2010).
2. P. Sklenovsy 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. C. B. Anfinsen, E. Haber, M. Sela, and F. H. White, Jr. *Proc Natl Acad Sci USA 47*, 1309-1314 (1961).
7. C. B. Anfinsen. *Science 181*, 223-230 (1973).
8. C. B. Anfinsen. *Biochem J 128*, 737-749 (1972).
9. K. A. Dill and H. S. Chan. *Nat Struct Mol Biol 4*, 10-19 (1997).
10. S. Govindarajan and R. A. Goldstein. *Proc Natl Acad Sci USA 95*, 5545-5549 (1998).
11. K. A. Dill. *Protein Sci 8*, 1166-1180 (1999).
12. C. J. Tsai. S. Kumar, B. Ma, and R. Nussinov. *Protein Sci 8*, 1181-1190 (1999).
13. J. M. Sorenson and T. Head-Gordon. *J Comput Biol 7*, 469-481 (2000).
14. B. A. Shoemaker and P. G. Wolynes. *J Mol Boil 287*, 657-674 (1999).
15. A. R. Fersht and V. Daggett. *Cell 108*, 573-582 (2002).
16. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. *Nucleic Acids Res 28*, 235-242 (2000).
17. K. Fan and W. Wang. *J Mol Boil 328*, 921-926 (2003).
18. K. C. Chou and H. B. Shen. *Anal Biochem 370*, 1-16 (2007).
19. I. Ezkurdia, L. Bartoli, P. Fariselli, R. Casadio. A. Valencia, and M. L. Tress, *Brief Bioinform 10*, 233-246 (2009).
20. B. He, K. Wang, Y. Liu, B. Xue, V. N. Uversky, and A. K. Dunker. *Cell Res 19*, 929-949 (2009).
21. J. C. Hansen, X. Lu, E. D. Ross, and R. W. Woody. *J Biol Chem 281*, 1853-1856 (2006).
22. J. L. Klepeis, Y. Wei, M. H. Hecht, and C. A. Floudas. *Proteins 58*, 560-570 (2005).
23. X. Xiao, P. Wang, and K. C. Chou. *J Theor Biol 254*, 691-696 (2008).
24. H. Lin and Q. Z. Li. *J Comput Chem 28*, 1463-1466 (2007).
25. A. Yan, A. Kloczkowski, H. Hofmann, and R. L. Jernigan. *J Biomol Struct Dyn 25*, 275-288 (2007).
26. G. Armano and A. Manconi. *BMC Res Notes 2*, 202 (2009).
27. S. Yasuda, T. Yoshidome, H. Oshima, R. Kodama, Y. Harano, and M. Kinoshita. *J Chem Phys 132*, 065105 (2010).
28. A. J. Doig and M. J. Sternberg. *Protein Sci 4*, 2247-2251 (1995).
29. V. Z. Spassov, L. Yan, and P. K. Flook. *Protein Sci 16*, 494-506 (2007).
30. R. Zwanzig. *Proc Natl Acad Sci USA 94*, 148-150 (1997).
31. S. Khorasanizadeh, I. D. Peters, and H. Roder. *Nat Struct Biol 3*, 193-205 (1996).
32. Y. Bai and S. W. Englander. *Proteins 24*, 145-151 (1996).
33. R. Zhou, X. Huang, C. J. Margulis, and B. J. Berne. *Science 305*, 1605-1609 (2004).
34. H. J. Dyson, P. E. Wright, and H. A. Scheraga. *Proc Natl Acad Sci USA 103*, 13057-13061 (2006).
35. C. M. Roth, B. L. Neal, and A. M. Lenhoff. *Biophys J 70*, 977-987 (1996).
36. A. Azia and Y. Levy. *J Mol Boil 393*, 527-542 (2009).
37. M. Oliveberg and A. R. Fersht. *Biochemistry 35*, 2726-2737 (1996).
38. J. Wang, Z. Zhang, H. Liu, and Y. Shi. *Phys Rev E 67*, 061903 (2003).
39. L. Holm and P. Rosenstrom. *Nucleic Acids Res 38 Suppl*, W545-549 (2010).
40. P. A. Alexander, Y. He, Y. Chen, J. Orban, and P. N. Bryan. *Proc Natl Acad Sci USA 104*, 11963-11968 (2007).

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


***Comment***

# Insenstivity to Close Contacts and Inability to Predict Protein Foldability

http://www.jbsdonline.com

**Karel Berka***
**Michal Otyepka***

Department of Physical Chemistry,
Faculty of Science, Palacky University
tr. 17. listopadu 12, 771 46 Olomouc,
Czech Republic

Mittal and coworkers recently (1) analyzed the spatial organization of folded proteins backbone and advocated an alternative hypothesis on protein folding stating that specific interactions among amino acids do not drive protein folding. The authors analyzed spatial distributions of $C_\alpha$ atoms for all amino acid pairs in about four thousands proteins available in PDB database and they found that the total number of contacts of an amino acid is correlated with the average occurrence of the amino acid. Their analysis led to a conclusion that there is no preferential neighborhood for an amino acid.

The correlation that they present between the amino acids occurrence in folded proteins and the number of total contacts is not surprising, because it is trivial. Their finding that there is not any preferential neighborhood for an amino acid is, however, strange, unconventional and appears to be not entirely correct.

There are similar observations where amino-acid pairs were analyzed (2-5). From these studies one concludes that there are statistically significant preferences between amino acid pairs, in addition to specific spatial arrangements. Furthermore such preferences are extensively used to prepare knowledge-based potentials for successful prediction of protein structures (6, 7). The theoretical determination of protein structure and function by modeling and simulation, based on preferential contacts and interactions, has become routine (8, 9, and references therein).

The disagreement between these observations and the report by Mittal *et al*. (1) may indicate that the analysis of spatial distribution of $C_\alpha$ atoms does not provide enough information. The sigmoid curves, employed to analyze the numbers of contacts in relation to the distance, are insensitive to close contacts as the majority of the curve is defined by the long distance interactions. This can be seen in the poor interleaving of the sigmoids in close contacts where the error bars would be approximately in tens of percents. These errors would probably be even stronger for more interaction-specific amino acids like cysteine or charged residues. Therefore the analysis of Mittal *et al.* (1) is actually detecting the preferential interactions in close contacts even though they are saying that such preferences are not present.

Main reason for such a disagreement is probably hidden in the fact that most of the contacts are random and as such defined only by amino acid frequencies and only a small minority of contacts is spatially and energetically distinguishable as was shown previously by Berka *et al*. (5). This can also explain why only a small number of amino acids is "critical" for folding of a protein (10, 11); why only a minimal change of amino acid sequence with 88% sequence identity leads to a completely different fold (12).

*Corresponding Authors:
Karel Berka
Michal Otyepka
E-mail: karel.berka@upol.cz
michal.otyepka@upol.cz

Last but not least, the work of Mittal and coworkers (1) cannot, unfortunately, be used to predict protein foldability, nor the protein structure, as the standard deviations in their "Chargaff's rules" are quite large and therefore not informative enough for any prediction.

***References***


1. A. Mittal, B. Jayaram, S. Shenoy, and T. Bawa. *J Biomol Struct Dyn 28*, 133 (2010).
2. S. Miyazawa and R. L. Jernigan. *J Mol Biol 256*, 623-644 (1996).
3. J. B. Mitchell, R. A. Laskowski, and J. M. Thornton. *Proteins: Struct Funct Genet 29*, 370-380 (1997).
4. H. Lu and J. Skolnick. *Proteins: Struct Funct Bioinf 44*, 223-232 (2001).
5. K. Berka, R. A. Laskowski, P. Hobza, and J. Vondrášek. *J Chem Theor Comput 6*, 2191-2203 (2010).
6. M. Shen and A. Sali. *Protein Science 15*, 2507 (2006).
7. R. Das and D. Baker. *Ann Rev Biochem 77*, 363-382 (2008).
8. P. Sklenovský and M. Otyepka. *J Biomol Struct Dyn 27*, 521-539 (2010).
9. M. J. Aman, H. Karauzum, M. G. Bowden, and T. L. Nguyen. *J Biomol Struct Dyn 28*, 1-12 (2010).
10. M. Vendruscolo, E. Paci, C. M. Dobson, and M. Karplus. *Nature 409*, 641-645 (2001).
11. J. Shimada and E. I. Shakhnovich. *Proc Natl Acad Sci 99*, 11175-11180 (2002).
12. Y. He, Y. Chen, P. Alexander, P. N. Bryan, and J. Orban. *Proc Natl Acad Sci 105*, 14412-14417 (2008).

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


***Comment***

# Stoichiometry of Amino acids Drives Protein Folding?

http://www.jbsdonline.com

**Carlos H. T. P. Silva[1*]**
**Carlton A. Taft[2*]**

[1]Departamento de Ciências Farmacêuticas, Faculdade de Ciências Farmacêuticas de Ribeirão Preto, Universidade de São Paulo, Av. do Café, s/n, Monte Alegre, 14040-903, Ribeirão Preto-SP, Brazil
[2]Centro Brasileiro de Pesquisas Físicas, Rua Dr. Xavier Sigaud, 150, Urca, 22290-180, Rio de Janeiro-RJ, Brazil

Mittal *et al.* (1) propose that the protein folding process is a direct consequence of a narrow band of stoichiometric occurrences of amino acids in primary sequences, regardless of the size and the fold of a protein. Based on rigorous analyses of several thousand crystal structures of folded proteins, the authors suggest that "preferential interactions" between amino acids do not drive protein folding, even though, such preferential interactions currently form the basis of the development of our knowledge based potentials and determination of protein structure by modeling and simulation (2, 3 and references therein).

Understanding the process in which a primary amino acid sequence folds into a biologically active protein with a well-established three-dimensional conformation that drives the function, still remains as one of the main challenges of modern biochemistry. In considering the 125 most important unsolved problems in science, Science magazine reported: "Can we predict how proteins will fold? Of infinite possibilities of ways to fold, a protein picks one in only tens of microseconds. The same task takes 30 years of computer time" (4, 5).

Analysis and extrapolation of information in the literature on protein folding indicate that certain amino acids exert specific influence on the local conformation; Kang *et al.* (6) thus discuss how the physicochemical properties of the amino acids preferentially result in specific structural arrangements. Although specific potentials of all 20 amino acids in forming secondary structures cannot be consistently predicted, the most and least favorable helix formers can be successfully figured out. As an example, aliphatic residues with branched carbon beta side chains (Val, Ile, Thr) are unfavorable helix formers due to the large volume of the functional groups that can cause steric hindrances (6). In contrast, alanine residues have a high helix-forming propensity.

The importance of the amino acid sequence in determining secondary structure elements is well illustrated by intrinsically unstructured proteins. Unlike globular proteins that in general form hydrophobic cores, disordered proteins possess a comparatively lower proportion of ''order-promoting'' bulky hydrophobic residues and aromatic residues, but are significantly richer in ''disorder-promoting'' residues including proline and glycine (6).

Nevertheless, it´s worth discussing what actually stabilizes secondary structures. Although different amino acids have different propensities to form specific secondary structures, there are also many "chameleon" sequences in natural proteins, which are peptide segments that can assume either helical or β-sheet conformations depending on their tertiary context (5). Studies of lattice and tube

*Corresponding Authors:
Carlos H. T. P. Silva
Carlton A. Taft
Phone: 55 16 3602 4717
55 21 2141 7201
E-mail: tomich@fcfrp.usp.br
catff@terra.com.br

models have shown that secondary structures in proteins are substantially stabilized by the chain compactness, an indirect consequence of the hydrophobic force to collapse, since helices and sheets are the only regular ways to pack a protein sequence into a tight space (5).

Considering such protein packing and a possible role of the intermolecular forces in the protein folding process, it is worthwhile remembering that, before the mid-1980s, the predominant view was that the protein folding code is the sum of many different small interactions (hydrogen bonds, salt bridges, van der Waals interactions) that occur locally (among neighbor residues) in the sequence, through secondary structures (7). However, from a statistical mechanical point of view, a different idea emerged in the late 1980s, with a dominant component to the "folding code" (the hydrophobic interaction). Such folding code would be distributed both locally and non-locally in the sequence, and native secondary structures would be more a consequence than a cause of protein folding (7). Today, there is considerable support that hydrophobic interactions must play a major role in protein folding, since proteins have hydrophobic cores and the apolar amino acids are then guided along the process to be sequestered from water (5).

In such folding process, despite the inherent complexity of the reactions, proteins seem to fold via a limited number of accumulated intermediates, where several proteins show a two-state behavior (8). As a result, the folding transition states (TS) play an important role in protein folding. Moreover, several proteins have also been suggested to fold via parallel routes, such as lysozyme and c-type cytochromes. Additional studies have reported proteins to switch their preferred folding pathways depending on the conditions. Finally, the different TS observed in homologous proteins also corroborate the notion of parallel folding pathways (8).

In considering the contribution of Mittal and collaborators (1) to the protein folding problem, it is especially significant with respect to the statistical underpinning that interactions between amino acids do not drive protein folding or, at least, they play a secondary role in the overall process. In considering a likely hierarchical order in protein folding, the authors have found that such event would be primarily dictated by the frequencies of occurrence (stoichiometry) of amino acids in the primary sequence, regardless of the length/size. Moreover, the results pointed out that to achieve a folded protein, the stoichiometric ratios of individual amino acids have to be defined by the known "Chargaff's Rules". Such rules indicate the average percentage occurrence of each one of the 20 natural amino acids in folded proteins, where the low standard deviations represent the "margin of life" (1).

Nevertheless, stoichiometry for example does not explain very well why there are many cases of nearly identical structures sharing no sequence similarity (9-12). Furthermore, not only in the cases where, the sequence identity shared by two proteins is low and the protein modeling is made by "threading", but also in the homology modeling cases the average frequency of occurrence of amino acids in the primary sequence is not enough information or rule. In homologous proteins, it is not rare to see substitution of leucine (the most frequent residue, 9%) by the least frequent, tryptophan (1.3%) in the protein data bank -PDB, particularly if such residues are buried in a hydrophobic pocket. In addition, simple stoichiometry also does not predict/explain post-translational modifications that, in most of the cases, have structural implications for the folding and conformation of the proteins, with functional consequences.

In reality, the "Chargaff's Rules" only translates the structural and physicochemical constraints that amino acids of a sequence must have in order to fold, event that occurs in predominantly aqueous environment. In this respect, and such as well discussed by Mittal and collaborators, the aqueous environment would force a protein to fold in "exclusion by water", thus minimizing its overall surface-to-volume ratio, which is provided by the individual shapes of the residues along the sequence (1). As previously mentioned here, helices and sheets are thus formed as an indirect consequence of the hydrophobic force to collapse, allowing the protein sequence to pack into a tight space, with likely subsequent formation of a folding transition state (TS) (5, 8). In the last step of the protein folding process, the packing would be then completed with the optimization of the interactions among residues positioned in specific chemical environments (13).

***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. D. Kennedy. *Science 309*, 75 (2005).
5. K. A. Dill, S. B. Ozkan, M. S. Shell, and T. R. Weikl. *Annu Rev Biophys 37*, 289-316 (2008).
6. T. S. Kang and R. M. Kini. *Cell Mol Life Sci 66*, 2341-2361 (2009).
7. K. A. Dill, S. B. Ozkan, T. R. Weikl, J. D. Chodera, and V. A. Voelz. *Curr Opin Struct Biol 17*, 342-346 (2007).
8. Y. Ivarsson, C. Travaglini-Allocatelli, M. Brunori, and S. Gianni. *Eur Biophys J 37*, 721-728 (2008).
9. S. E. Brenner and M. Levitt. *Protein Sci 9*, 197-200 (2000).
10. D. Devos and A. Valencia. *Proteins 41*, 98-107 (2000).
11. W. A. Koppensteiner, P. Lackner, M. Wiederstein, and M. J. Sippl. *J Mol Biol 296*, 1139-1152 (2000).
12. W. D. Tian and J. Skolnick. *J Mol Biol 333*, 863-882 (2003).
13. D. Eisenberg, R. Lüthy, and J. U. Bowie. *Methods Enzymol, 277*, 396-404 (1997).

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


***Comment***

# Protein Folding: Grand Challenge of Nature

http://www.jbsdonline.com

**Bishnu P. Mukhopadhyay***
**Hridoy R. Bairagya**

Department of Chemistry,
National Institute of Technology,
Durgapur – 713 209, W. B. India

Protein folding is a longstanding challenge in structural biology. No one knows how to predict the correct native structure or folding of a protein from the primary amino acid sequence unless that sequence has significant homology with another protein of known 3D structure. Concerning 1D-3D structure/folding prediction in proteins, one remarkable research work "A Stoichiometry Driven Universal Spatial Organization of Backbones of Folded Proteins: Are there Chargaff's Rules for Protein Folding?" has recently been published by Mittal *et al*. in 2010 in this *Journal* (1). The work opens an unanticipated window into the new biostatistics methods on protein folding. After rigorous investigation on close to fourthousand crystal structures (of protein), the authors conclude "Protein folding is a direct consequence of a narrow band of stoichiometric occurrences of amino-acids in primary sequences, regardless of the size and the fold of a protein". Contrary to all prevalent views, the authors also emphasized: "preferential interactions" between amino-acids do not drive protein folding. In this connection, it should be noted that that preferential interactions between amino acids are the basis for introducing knowledge-based potentials, which in turn provide the underpinning for present day 3D protein structure prediction by modeling and simulation (2-5 and references therein). Perhaps the actual rules of protein folding still remain elusive; however, this work contributes new information on bio-mathematical expression and adds a new dimension that could lead to solutions to problems presented by structural biology.

Though the research work of Mittal *et al*. may provide some light on the recognition of protein fold from their primary amino acid sequences, it has several limitations. Several points in Mittal *et al*. appear to conflict with existing concepts. In the prediction of protein folding, the contribution of side chains (polar/non-polar, acidic and basic) in the primary sequence seems missing in this work, and this should be clarified. The hydrophobic, hydrophilic, van der Waals and H-bonding interaction may control the folding mechanism by fixing the orientation of side chains in proteins (6-8). The side chains (polar, acidic or basic) interact with the aqueous medium and play a major role in shaping the protein, the hydrophobicity of amino acids tend to drive them from exterior to interior. The charge – charge, charge – dipole and dipole – dipole interaction of side chains may calibrate the fold recognition. So, the essence of side chain could be to gear the folding topology of protein, *i.e.,* folding topology is not geared by protein backbone alone as advocated by Mittal *et al*. The dynamical results reveal that the backbone is incapable to build the actual fold of a protein, perhaps the side chain with backbone amino acids produce the folds (9). During dynamics, the fold of protein structure undergoes the conformational changes which may be identical or not with its x-ray structure. All the atoms (backbone and side chain) move anisotropically, and the direction of motion in protein is determined by non-bonded interaction of sidechains. This motion is anharmonic at biological temperature, usually due to

*Corresponding Author:
Dr. Bishnu P. Mukhopadhyay
Phone: +91- 03432547074
Fax: 91-(0343) 2547375
E-mail:bpmk2@yahoo.com

side chain hopping between two or more alternative minima (10-12). Moreover, the loop–flap dynamics demand that the backbone behave as rigid bodies (minimum RMSF) whereas the displacement of loop is possible due to the reorientation of sidechains (13). Hence, in addition to the C-alpha atoms, the sidechain could influence the shape, size and fold of a protein. On the basis of the above arguments, it is doubtful or not clear, whether the correct fold recognition is provided by the stoichiometry driven folding thesis of Mittal *et al*.

Again, an another important point, the influence of topology to protein folding has not been mentioned by the authors. Protein-folding rates and mechanisms are largely determined by a protein's topology rather than its inter-atomic interactions. In the several crystal structures, the large changes in amino-acid sequence do not alter the overall topology of a protein and evolution has not optimized protein sequences for rapid folding (14). The observation that some structurally related proteins with little or no sequence similarity have very similar type of fold, makes the situation more complicated.

Nevertheless, some single-point mutation in a protein perturbs the secondary, tertiary, and quaternary structures, thus inducing new folding (15). The backbone of protein could not control the folding mechanism, perhaps a conflict in side chain orientation causes a misfold (16, 17). Again in some multi nuclear metalloproteins, the occupancy of metal ions within the protein are controlled by side chains, not by backbone, and sometime the selectivity of metal ions (18) are also governed by protein folding. In addition, continuum models of water do not account for the discrete nature of water molecules, which may lead to differences in protein folding dynamics (19) such as a cooperative expulsion of water upon folding. The authors' view on this aspect is not clear.

Undoubtly, the work of Mittal *et al*. will be very effective to provide impetus for further investigation in protein research.

***References***


1. A. Mittal, B. Jyaram, 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. Y. Tao, Z. H. Rao, and S. Q. Liu. *J Biomol Struct Dyn 28*, 143-157 (2010).
5. C. Koshy, M. Parthiban, and R. Sowdhamini. *J Biomol Struct Dyn 28*, 71-83 (2010).
6. D. Higgins and W. Taylor. Bioinformatics: Sequence, Structure, and Databanks. Oxfford University Press, New York (2000).
7. R. Zhou, B. J. Berne, and R.Germain. *PANS USA 26*, 14931-14936 (2001).
8. C. Branden, J. Tooze. Introduction to Protein Structure. Garland Publishing, New York and London (1991).
9. K. H.Wildman and D. Kern. *Nature 450*, 964-972 (2007).
10. M. Karplus and G. A. Petsko. *Nature 347*, 631-639 (1990).
11. A. R. Dinner, A.Šalib, L. J. Smith, C. M. Dobson, and M.Karplus. *TIBS 25,* 331-339 (2000).
12. P. G. Wolynes, J. N. Onuchic, and D. Thirumalai. *Science* 267,1619-162 (1995).
13. H. R. Bairagya, B. P. Mukhopadhyay, and A. K. Bera. *J Mol Recognit*. online published (2010).
14. D. Baker. *Nature 405*, 39-42 (2000).
15. A. Banerjee, H. R. Bairagya, B. P. Mukhopadhyay, T. K. Nandi, and A. K. Bera. *Indian Journal of Biochemistry and Biophysics* 47, 197-202 (2010).
16. X. Hou, M. I. Aguilar, and D. H. Small, *FEBS J 274*, 1637-1650 (2007).
17. C. M. Dobson. *Nature 426*, 884-890 (2003).
18. K. J. Waldron, J. C. Rutherford, D. Ford, and N. J. Robinson. *Nature Review Article 460*, 823-830 (2009).
19. Y. M. Rhee, E. J. Sorin, G. Jayachandran, E. Lindahl, and V. S. Pande. PANS USA 101, 6456-6461 (2004).

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


***Comment***

# Is Protein Folding Still a Challenge?

**R. Nagaraj**

Centre for Cellular and Molecular Biology, CSIR, Uppal road, Hyderabad 500 007, India

http://www.jbsdonline.com

Mittal *et al*. in their paper entitled "A stoichiometry driven universal spatial organization of backbones of folded proteins: Are there Chargaff's rules for protein folding?" have asked whether there is a unifying theme or concept underlying the magnificent diversity of folded protein structures (1). They have observed that preferential interactions between amino acids do not drive protein folding, contrary to all prevailing views. This view is so prevailing that preferential interactions between amino acids are the basis for introducing knowledge-based potentials, which in turn provides the underpinning for present day 3D protein structure prediction by modeling and simulation (2-5 and references therein).

Every student of protein biochemistry is instilled with the idea that the protein folding process is complex and modulated by several factors. Levanthal's paradox (6) is often stated to emphasize the complexity of protein folding. Delineating the pathway of protein folding and identification of folding intermediates has been and still is a favourite research topic. Structural propensities of amino acids to occur in a specific conformation such as α-helix, β-sheet or β-turns were proposed based on analysis of protein structures in the crystalline state way back in the 1970's by Chou and Fasman (7, 8). Prediction of protein folding was attempted based on structural propensities of amino acids and probability values (8). Extensive analysis of protein crystal structures, have provided insights into interaction between aromatic amino acids (9), disulfide conformation (10) and occurrence of β-hairpin structures (11, 12).

Mittal *et al*. have analyzed 3718 protein X-ray structures and delineated interacting partners of amino acids based on Cα distances (excluding those adjacent to each other connected by a peptide bond). Their reasoning is that if the two methyl groups in Leu and Ile are proximal to each other and interact via hydrophobic interactions, this would be reflected in the Cα distances between them. Their analysis also suggests that any one of the 20 coded amino acids could interact with any of the other 19 coded amino acids. Also, interactions between cationic or anionic amino acids are not precluded. Their intriguing observation is that, in the crystalline state, the interaction between amino acids is independent of the nature of the amino acid. The frequency of occurrence rather than chemical nature determines the interaction between amino acids. Their conclusion is that folding of a protein is simply dictated by frequency of occurrence of amino acids and not the sequence.

An important conclusion by Mittal *et al*. is that for a folded protein, stoichiometric ratios of individual amino acids should have values indicated in Table I. The implication of this is that it would not be possible for a protein to fold if the percentage of Ala deviates from 7.8+/-3.4 or that of His deviates from 2.3+/-1.4. This information would be of immense help to protein engineers in determining whether a protein encoded by an open reading frame can fold or possibly exist.

Corresponding Author:
R. Nagaraj
Phone: +91-40-2716022
Fax: +91-40-27160591/031
E-mail: nraj@ccmb.res.in

Mittal *et al*. have proposed in their paper that the most important parameters for protein folding are (i) exclusion of water and (ii) shape characteristics of individual amino acids along the sequence that would minimize surface-to-volume ratio. However, they summarize that protein folding is primarily dictated by frequencies of occurrence of amino acids. This appears to be at variance with their proposal in (ii).

Chargaff's rule stems from the association of Adenine (A) with Thymine (T) and Guanine (G) with Cytosine (C) via hydrogen bonding. A similar association of amino acids does not emerge from the analysis by Mittal *et al*. In fact, their analysis stresses the absence of any preferential relationship between amino acids. Hence, the basis for stochiometry-driven spatially organized double helical structure of DNA does not apply to organization of amino acids in proteins. The authors are using the term Chargaff's rule to emphasize the stoichiometric relationship that exist among the amino acids in Table I, this stoichiometry being the underpinning of protein folding. From species to species, source to source, the proportions of A, G, C and T in DNA may vary widely, but universally, the molar ratios of A and T and that of G and C in DNA were not far from unity. In the same manner, from protein to protein, the proportions of amino acids may vary, but in folded globular proteins, the stoichiometry of the amino acids should not be far from what is in Table I in Mittal *et al*. (1).

The process of protein folding in solution can be modulated by several factors such as pH, temperature, salt concentration and the presence of organic solvents. The study by Mittal *et al*. gives a fascinating and new insight into organization of amino acids in folded proteins in the crystalline state. However, knowledge of only amino acid composition may not provide insights into how the folded conformation is achieved, particularly in solution. Protein folding continues to be a challenge and a problem yet to be solved.

***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. Y. Tao, Z. H. Rao, and S. Q. Liu. *J Biomol Struct Dyn 28*, 143-157 (2010).
5. C. Koshy, M. Parthiban, and R. Sowdhamini. *J Biomol Struct Dyn 28*, 71-83 (2010).
6. C. Levinthal. *J Chim Phys 65*, 44-45 (1973).
7. P. Y. Chou, and G. D. Fasman. *Biochemistry 13*, 211-222 (1974).
8. P. Y. Chou, and G. D Fasman. *Biochemistry 13*, 222-245 (1974).
9. S. K. Burley, and G.A. Petsko. *Science 229*, 23-28 (1985).
10. C-C. Chuang, C-Y. Chen, J-M. Yang, P-C. Lyu, and J-K Hwang. *PROTEINS: Structure, Function and Genetics 52*, 1-5 (2003).
11. B. L. Sibanda and J. M. Thornton. *Methods Enzymol 202*, 59-82 (1991).
12. K. Gunasekaran, C. Ramakrishnan, and P. Balaram. *Protein Engineering 10*, 1131-1141 (1997).

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


***Comment***

# Is Stoichiometry a New Metric for Evaluating Folded Proteins?

**James C. Burnett**
**Tam Luong Nguyen***

Target Structure-Based Drug Discovery Group, SAIC Frederick, Inc., NCI-Frederick, Frederick, MD 21702, USA

http://www.jbsdonline.com

The folding of randomly coiled polypeptide chains into biologically active protein structures has been intensely studied for many years. In early studies by Anfinsen (1, 2), the unfolding and refolding of disulfide-rich proteins initially led to the formation of mispaired disulfide bonds (*i.e.,* relative to the wild-type). However, in an oxidative environment, the mispaired disulfide bonds returned to their native disulfide pairings, thereby restoring normal enzyme activity. These results suggested that non-bonded interactions drive protein folding rather than the random pairing of stronger covalent disulfide bridges.

Due to experimental observations like those of Anfinsen, the current opinion is that the sequence identity along a polypeptide chain contains the essential code to understanding and predicting protein folding. However, the article by Mittal *et al*. "A Stoichiometry Driven Universal Spatial Organization of Backbones of Folded Proteins; Are there Chargaff's Rules for Protein Folding?" (3) is set to add an entirely new metric to these prevailing views.

In the work by Mittal *et al.* the crystal structures of 3718 folded proteins were analyzed. For each amino acid in a protein, the authors determined the occurrence of neighboring amino acids by measuring between Cα carbons within specified spherical distances. In this way, neighborhoods for the 20 amino acids were defined over increasing distance ranges. The 20 amino acid neighborhoods were found to be characterized by sigmoidal plots that could be parameterized by a single, general equation (Figure 2 in the publication). Significantly, the total number of contacts made by an amino acid correlated with its average occurrence in the folded proteins (Figure 4 in the publication), suggesting that the amino acid neighborhoods were governed by the frequency of individual amino acids rather than their preferential interactions with other amino acids. This proved to be true even when short sub-10 Å neighborhood distances were examined (Figure 3 in the publication). Taken together, these results led Mittal *et al*. to propose that protein folding is stoichiometry-driven. In this context, it should be noted that preferential interactions between amino acids were the basis for introducing knowledge-based potentials, which provided a foundation for modern day protein structure prediction by modeling and simulation (4, 5 and references therein).

The concept of stoichiometry-driven protein folding is in sharp contrast to the predominant view in modern biochemistry, which holds that the rapid clustering of large hydrophobic side chains is responsible for driving the conversion of random polypeptide chains into biologically relevant proteins. At first glance, a stoichiometric basis for protein folding may appear to be incongruent

*Corresponding Author:
Tam Luong Nguyen
Phone: 301-846-6035
Fax: 301-846-6106
E-mail: nguyent5@mail.nih.gov

with established concepts in biochemistry, but the work of Mittal *et al*. captures a distinct aspect of protein folding. Specifically, the end result of protein dynamics and folding is a fundamental 'constant' that appears as a single sigmoid equation.

Because diverse proteins appear to converge on a fundamental constant, this single equation may be used as a quantitative measure for 'protein fold space' (Table 1 in the publication), and much like Ramachandran plots, may establish a metric for evaluating the quality of 3D protein structures, whether they be experimentally determined structures or a *de novo* folds.

However, the impact of side chain-side chain interactions on the protein folding process may have been 'lost in translation' during this study. In particular, residue neighborhoods were generated by identifying amino acids within specified distance ranges of a given amino acid, and these distance parameters radiated in a spherical fashion from the Cα carbon atom. As a result, residues with side chains that do not engage in close atom-atom interactions within a sub- 10 Å distance may be identified as close contact neighbors. For instance, in an amphipathic helical structure, a large hydrophobic residue such as leucine at position *i* may be identified as a short, 5-10 Å neighbor of a highly acidic residue such as glutamic acid at position *i*+2 – even though the side chains of these two residues are oriented on opposite sides of the helical structure. In this example, any interactions between the side chains of leucine (*i*) and glutamic acid (*i* + 2) would be highly unfavorable (*i.e.,* hydrophobic-polar clashes), and yet, these two residues are nearest neighbors. Since protein folding must follow the fundamental rules of chemistry and the hydrophobic effect, these rules are not addressed by a purely stoichiometric approach.

In closing, the work by Mittal *et al*. contributes to the protein folding story, but is by no means the final chapter. Rather, it is another piece of the puzzle. The reported sigmoidal behavior of amino acid neighborhoods is quite amazing, and illustrates the enormous complexity underlying the forces that drive protein folding. Indeed, Nature has built a large number of diverse protein structures, and in doing so, has managed to apply a single equation to array amino acids in three-dimensional 'fold' space.

***References***

1. C. B. Anfinsen, E. Haber, M. Sela, and F. H. White Jr. *Proceedings of the National Academy of Sciences of the United States of America 47*, 1309-1314 (1961).
2. C. B. Anfinsen. *Science 181*, 223-230 (1973).
3. A. Mittal, B. Jayaram, S. Shenoy, and T. S. Bawa. *J Biomol Struct Dyn 28* (2010).
4. P. Sklenovský, and M. Otyepka. *J Biomol Struct Dyn 27*, 521-539 (2010).
5. M. J. Aman, H. Karauzum, M. G. Bowden, and T. L. Nguyen. *J Biomol Struct Dyn 28*, 1-12 (2010).

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


***Comment***

# Stoichiometry-driven Protein Folding: A Comment

**Paul S. Agutter**

Theoretical Medicine and Biology Group, 26 Castle Hill, Glossop, Derbyshire, U.K

http://www.jbsdonline.com

Mittal *et al*. (1) have examined the frequency distributions of amino acid residue proximities in the secondary and tertiary structures of almost 4000 proteins, and have concluded that there are no preferences related to side-chain chemistry. Irrespective of the size and three-dimensional shape of a protein, the probability that two residues will be close together depends mainly, if not entirely, on how many of each there are within that protein. This result sits uncomfortably with major current ideas about protein folding, such as the evolution of wonderfolds (2), the proteomic code (3), and the 'first-in-last-out' hypothesis (4). It may not be logically inconsistent with these ideas, but it will be considered surprising. On the simplest level: it is common knowledge that polar residues cluster with other polar residues rather than hydrophobic ones, and vice versa, and this almost-universal pattern is crucial for the spatial structuring of globular proteins. Preferential interaction between amino acids is the basis of the development of knowledge-based potentials, which in turn form the underpinning of protein structure prediction. Nowadays, determination of protein structure by modeling and simulation, based on preferential contacts and interactions, has become a common practice (5, 6 and references therein).

On the face of it, the findings of Mittal *et al*. (1) seem to conflict with this common knowledge, so the study will be inspected for methodological flaws.

One likely target for criticism will be the 'broad-brush' approach used by these authors, which could make genuine patterns invisible. For example, a glutamate residue on the surface of a globular protein may be no more distant from a leucine below the surface than it is from a serine or arginine that is also exposed to the aqueous environment. Thus, physical proximity alone does not distinguish hydrophilic surface from hydrophobic core. Therefore, what Mittal *et al*. (1) have found need not be contrary to received wisdom after all. That may lead to wider doubts about the validity of their method.

There is a simple test for validity, and it may merit a follow-up study. In non-structural secreted proteins such as mammalian digestive enzymes and some plasma proteins, three-dimensional structural stability depends more on disulphide bridges than on hydrophobic interactions. Therefore, in this class of globular proteins, a high percentage of cysteine residues are close neighbours of other cysteine residues (being covalently linked to them). The distribution of cysteine neighbours should therefore appear significantly less random than the distributions reported in (1). If that prediction is confirmed, it will be reasonable evidence that the work of Mittal *et al*. is methodologically sound, and those who are studying the patterns and mechanisms of protein folding can then face the challenge of reconciling these findings with the dominant models in the field.

Corresponding Author:
Paul S. Agutter
E-mail: psa@tmedbiol.com

***References***


1. A. Mittal, B. Jayaram, S. Shenoy, and T. S. Bawa. *J Biomol Struct Dyn 28*, 133-142 (2010).
2. I. N. Berezovsky and E. N. Trifonov. *Comp Funct Genomics 3*, 525-534 (2002).
3. J. C. Biró. *Theor Biol Med Model 4*, 45 (2007).
4. M. Preuss and A. D. Miller. *FEBS Lett 466*, 75-79 (2000).
5. P. Sklenovský and M. Otyepka. *J Biomol Struct Dyn 27*, 521-539 (2010).
6. M. J. Aman, H. Karauzum, M. G. Bowden, and T. L. Nguyen. *J Biomol Struct Dyn 28*, 1-12 (2010).

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


***Comment***

# Thermodynamic Framework of the Interaction between Protein and Solvent Drives Protein Folding

http://www.jbsdonline.com

**Teodorico C. Ramalho***
**Elaine F. F. da Cunha***

Department of Chemistry, Federal University of Lavras, Caixa Postal 37, CEP 37200-000 Lavras, Minas Gerais, Brasil

In the last three decades, simulation studies have driven the development of the modern perspective on protein folding and have provided certain key insights on the relationship between protein topology and the folding mechanism which may emerge from folding free energy surface calculations (1). To date, however, it should be kept in mind that the complete information on folding mechanism has not emerged from free-energy surface projections. Therefore, the detailed aspects of protein folding is not easily accessible even from the folding landscape theory (2-6). In this respect, new perspectives on this fundamental biological phenomenon are welcome.

The Mittal-Jayaram thesis (7) reveals uncommon results and unfolds a novel perspective on protein folding in the light of Chargaff's Rules. Sir Chargaff discovered a quantitative relationship (stoichiometry) among the nucleic acid bases of DNA (8). This finding was crucial and led to the DNA double helical structure discovery a few years later. Nevertheless, since the computational facilities and mathematic algorithms have emerged as a powerful tool to investigate the matter at the molecular level, Chargaff's Rules have been subjected to an intense theoretical investigation (1, 8, 9). Currently, it is known that the Chargaff's rule can be derived from a thermodynamic framework, because GC and AT base pairs are the most stable complexes energetically formed (9). Therefore, the thermodynamic contribution is relevant in order to address the present question at hand. As stated in Mittal *et al.* (7), the interaction between solvent molecules and amino acid residues plays an important role in protein folding. In this scenario, studies directed at a family of small, single-domain, cooperative, fast-folding proteins indicated that relatively exposed peptide-hydrogen bonds, as they occur in helices and turns, are significantly weakened due to competition from water-hydrogen bonds (1). However, in β sheet hydrogen bonds, even in the simplest β-sheet model comprised of an antiparallel alanine "dipeptide," the alanine side chain provides significant shielding of the peptide interactions, resulting in stronger interactions between the peptide chains (10, 11). These strong interactions were attributed to lead to a slower dissolution of β sheet structure in solution. In addition, some studies involving molecular dynamics simulation delineated the role of hydrogen-bonding interactions in stabilizing the secondary structure of proteins in water (12-14). It is worth noting that there is an intense discussion about the balance between polar electrostatic and nonpolar cavity interactions calculated by implicit and explicit solvent models. Recent findings suggest that the folding free energy landscape results, using the generalized Born (GB) implicit solvent model, is quite different from those using an explicit solvent model (15). The most complete theoretical description of these

*Corresponding Authors:
Teodorico C. Ramalho
Elaine F. F. da Cunha
Phone: +55-35-38291891
Fax: +55-35-38291271
Email: teo@ufla.br
elaine_cunha@dqi.ufla.br

systems is reached when the protein and its solvent environment are represented by explicit atoms, whose motions are evaluated using molecular dynamics or Monte Carlo approaches (16-20).

Thus, these kinds of studies demonstrated that relatively exposed peptide-hydrogen bonds, as they occur in helices and turns, were significantly weakened by competition from water-hydrogen bond donor and acceptor groups. From Mittal *et al*. (7), the amino acids leucine, alanine, valine and glycine show the highest percentage occurrences in folded proteins (see Table I, (7)). Interestingly, this information is in good agreement with the basic thermodynamics of the interaction between protein and solvent. In fact, from our point of view, Chargaff's Rules applied to protein folding might be related to a stoichiometry of amino acids in order to create hydrophobic cavities in the solvent environment, which could then maximize the hydrophobic/electrostatic interactions among the amino acids resulting in the protein folding. Actually, as commented by Mittal and his colleagues (7), there is no evidence that "preferential interactions" between amino-acids drive the protein folding, despite the fact that determination of protein structure by modeling and simulation, based on preferential contacts and interactions, has become a common practice (21, 22 and references therein). However, from a theoretical perspective, the creation of these afore mentioned "hydrophobic cavities" might take place as an initial step for the process.

We would like to mention the very productive interplay between experiment and theory that has shaped much of the work in the discipline. The Mittal-Jayaram protein folding thesis is very interesting and promising, because it highlights that stoichiometric occurrences of amino acids in primary sequences is closely related to protein folding. It is important to mention, however, that the research field in protein folding has its *genesis* in the physical organic chemistry, and hence it is not surprising that it may be evaluated by physicochemical approaches. These approaches involve the physical and energetic characterization of the several states of the protein in the folding process (23).

Proteins folding is a complex problem with many variables. This outlook is aggravated by the existence of few experimental handles for understanding the individual steps that lead a polypeptide chain to adopt unique and relatively stable three-dimensional conformations. Maybe Chargaff's Rules might be applied in conjunction with the solvent effect in order to elucidate the protein folding problem. Actually, the specific occurrences of amino acids in primary sequences, pointed out by Mittal *et al.* (7), after analyzing close to four thousand folded protein crystal structures and their interactions with water, cannot be ignored, because these facts might play a much more important role in protein folding than we thought. This paper points out reliable interesting results and deserves a deeper evaluation (24-26). We strongly feel that the study reported by Mittal, Jayaram and their colleagues could be helpful for exploring protein engineering as well.

***Acknowledgment***

Authors thank FAPEMIG and CNPq for the financial support of this research. CNPq are also gratefully acknowledged for the fellowships and studentships.

***References***

1. J. E. Shea and C. L. Brooks. *Annu Rev Phys Chem 52*, 499-535 (2001).
2. J. E. Shea, J. N. Onuchic, and C. L. Brooks. *Proc Natl Acad Sci USA 96*, 12512–12517 (1999).
3. J. E. Shea, J. N. Onuchic, and C. L. Brooks. *J Chem Phys 113*, 7663-7671 (2000).
4. C. Clementi, H. Nymeyer, and J. N. Onuchic. *J Mol Biol 298*, 937-946 (2000).
5. H. Nymeyer, N. D. Socci, and J. N. Onuchic. *Proc Natl Acad Sci USA 97*, 634-639 (2000).
6. A. M. Gutin, V. I. Abkevich, and E. I. Shakhnovich. *Fold Des 3*, 183-194 (1998).
7. A. Mittal, B. Jayaram, S. Shenoy, and T. S. Bawa. *J Biomol Struct Dyn* 28, 133-142 (2010).
8. V. Daggett and M. Levitt. *Curr Opin. Struct Biol 4*, 291–959 (1994).
9. A. Furmanchuk, O. Isayev, O. V. Shishkin, L. Gorb, and J. Leszczynski. *Phys Chem Chem Phys 12*, 3363-3375 (2010).
10. E. F. F. da Cunha, T. C. Ramalho, and R. C. Reynolds. *J Biomol Struct Dyn 25*, 377-385 (2008).
11. D. J. Tobias, E. Mertz, and C. L. Brooks. *Biochem 30*, 6054-58 (1991).
12. D. J. Tobias and C. L. Brooks. *Biochem 30*, 6059-6070 (1991).
13. W. S. Young and C. L. Brooks. *J Mol Biol* 259, 560-72 (1996).
14. F. B. Sheinerman and C. L. Brooks. *J Am Chem Soc* 117, 10098-103 (1995).
15. R. Zhou, G. Krilov, and B. J. Berne. *J Phys Chem B 108*, 7528-7530 (2004).
16. C. L. Brooks, M. Karplus, and B. M. Pettitt. Proteins: A Theoretical Perspective of Dynamics, Structure and Thermodynamics. New York: Wiley & Sons (1988).
17. J. A. McCammon and S. Harvey. Dynamics of Proteins and Nucleic Acids. Cambridge Univ. Press (1987).
18. E. F. F. da Cunha, E. F. Barbosa, A. A. Oliveira, and T. C. Ramalho. *J Biomol Struct Dyn 27*, 619-625 (2010).
19. Y. Tao, Z. H. Rao, and S. Q. Liu. *J Biomol Struct Dyn 28*, 143-157 (2010).
20. H. R. Bairagya, B. P. Mukhopadhyay and K. Sekar, *J Biomol Struct Dyn 27*, 149-158 (2009).
21. P. Sklenovský and M. Otyepka, *J Biomol Struct Dyn 27*, 521-539 (2010).
22. M. J. Aman, H. Karauzum, M. G. Bowden, and T. L. Nguyen. *J Biomol Struct Dyn 28*, 1-12 (2010).
23. P. H. Figueiredo, M. A. Moret, P. G. Pascutti, E. Nogueira, S. Coutinho. *Physica A 389*, 2682-2686 (2010).
24. E. Weinan and E. Vanden-Eijnden. *Annu Rev Phys Chem* 61, 391-420 (2010).
25. T. Tian. *Chem Soc Rev* 39, 2071-2082 (2010).
26. C. A. Royer. *Arch Biochem Bioph* 469, 34-45 (2008).

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


***Comment***

# Stoichiometry and Folding of Linear Polypeptides – Reading Between the Lines

http://www.jbsdonline.com

**Richard A. Bryce***

School of Pharmacy and Pharmaceutical Sciences, University of Manchester, Oxford Road, Manchester, M13 9PT, UK

The article "A Stoichiometry Driven Universal Spatial Organization of Backbones of Folded Proteins: Are there Chargaff's Rules for Protein Folding?" by Mittal *et al*. (1) seeks to relate the ability of linear amino acid sequences to form folded protein structures and the frequency of occurrence of those amino acids in the polypeptide chains. Understanding the rules that govern how these simple lines of amino acids fold – *i.e.,* reading between the lines – is a key goal in modern biology.

Indeed, such linear polypeptides self-assemble into a wide array of elegant 3D protein structures, with great variation in size and in the size of their substrates. Proteins can adopt monomeric or oligomeric structures, as well as forming elaborate higher order multi-protein complexes. Beneath this complexity observed across protein structures (*e.g.,* currently 62957 protein structures populate the PDB (2)), there is an underlying simplicity: 1393 protein folds are known (using the SCOP classification (3)), eight possible secondary structural elements (4), with an underlying alphabet of 20 amino acids. The primary sequences of proteins are somewhat more complex than that of DNA, with only four bases, but considerably less complex than the primary sequences of carbohydrates, with their ability to branch in many ways (5). Proteins are large molecules, folding into sufficiently rigid scaffolds to perform their function effectively; with a low enough surface area to volume ratio to prevent aggregation in the crowded cell; and with a suitably sculpted surface to direct substrate to its active site (6), the latter often being the deepest cavity in the protein (7).

In the study of Mittal *et al.* (1), analysis of a set of nearly 4000 folded protein crystal structures finds a rather narrow range of variation in amino acid composition, suggesting that this specific ratio of amino acids is a prerequisite for a successfully folded protein (Table 1 of reference (1)). As a control, the authors observe a different stoichiometry for proteins which populate significantly unfolded states (Figure 5B of reference (1)). Intriguingly, for the stoichiometry of the folded protein set, the highest abundances are observed for Ala, Val, Leu, Gly and Glu, with the lowest for Met, Cys, Trp and His. Associated with these averages are standard deviations across the set of ~4000 crystal structures which span 1.0–3.4%. This range is colourfully referred to by the authors as the "margin of life".

The authors draw an analogy between their work and that of Erwin Chargaff, who demonstrated via elegant experiments the 1:1 stoichiometry of G:C as well as A:T in DNA. Chargaff also used the four nucleotide bases to deduce that the composition of DNA is organism-specific, showing their relative proportions varied between organisms. The latter rather simple observation hides the tremendously intricate differences in genomic sequence that define species. In the same way,

*Corresponding Author:
Richard A. Bryce
Phone: (0)161-275-8345
Fax: (0)161-275-2481
E-mail: R.A.Bryce@manchester.ac.uk

the "margin of life" though apparently narrow must be sufficiently broad, and the ordering of amino acids within the confines of this stoichiometric range sufficiently flexible, to permit the complexity in the folded shape and behaviour of functional proteins.

The underlying physics which dictate protein folding is the subject of much debate, as Mittal *et al.* acknowledge in their introduction (1). The authors examine the average radial distribution of the 20 amino acid types around a given type of amino acid, as measured by $C_\alpha C_\alpha$ distances. From their analysis of the extent of these contacts and the similarity in profile of their distribution for a given amino acid type, the authors conclude that protein folding is dictated simply by frequencies of occurrences of individual amino acids. This leads the authors to highlight solvation and packing effects as the key guiding influences in the protein folding process.

Interestingly, the authors' analysis of the radial $C_\alpha C_\alpha$ contacts in crystal structures bears some resemblance to the derivation of knowledge-based potentials used in protein folding prediction (8), although the latter are normalized with respect to a hypothetical reference state, *e.g.* frequency of contacts expected from random mixing of amino acids and solvent. Despite being low in resolution, the latter potentials have shown some success in predicting 3D structures of small proteins but typically require augmentation with other terms and subsequent force field refinement in order to yield results accurate to a resolution of ~1.5 Å (9-13). Of course, we note that knowledge-based potentials and the $C_\alpha C_\alpha$ neighbourhoods calculated in (1) involve spherical averaging, whereas proteins are inherently asymmetric molecules, a key to understanding their biological function.

Nevertheless, the insights derived from the work of Mittal *et al.* represent an important contribution to the debate surrounding the basis of protein folding; extension of this analysis to a greater number of the known protein structures, as well as to detailed examination of sub-sets such as structural proteins, will provide further understanding of how linear chains of amino acids adopt their exquisite 3D structures.

***References***


1. A. Mittal, B. Jayaram, S. Shenoy, and T. S. Bawa. *J Biomol Struct Dyn 28*, 133-142 (2010).
2. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. *Nucleic Ac Res 28*, 235-242 (2000).
3. A. Andreeva, D. Howorth, J.-M. Chandonia, S.E. Brenner, T. J. P. Hubbard, C. Chothia, and A. G. Murzin. *Nucl Acids Res 36*, D419-D425 (2008).
4. W. Kabsch and C. Sander. *Biopolymers 22*, 2577-637 (1983).
5. V. S. R. Rao, P. K. Qasba, P. V. Balaji, and R. Chandrasekaran. *Conformation of Carbohydrates*, Harwood Academic Press (Amsterdam, 1998).
6. D. S. Goodsell and A. J. Olson. *TIBS 18*, 65-68 (1993).
7. R. A. Laskowski, N. M. Luscombe, M. B. Swindells, and J. M. Thornton. *Protein Sci 5*, 2438-2452 (1996).
8. J. P. Kocher, M. J. Rooman, and S. J. Wodak. *J Mol Biol 235*, 1598-1613 (1994).
9. P. Bradley, K. M. S. Misura, and D. Baker. *Science 309*, 1868-1871 (2005).
10. P. Sklenovský and M. Otyepka. *J Biomol Struct Dyn 27*, 521-539 (2010).
11. M. J. Aman, H. Karauzum, M. G. Bowden, and T. L. Nguyen. *J Biomol Struct Dyn 28*, 1-12 (2010).
12. C. Koshy, M. Parthiban, and R. Sowdhamini. *J Biomol Struct Dyn 28*, 71-83 (2010).
13. Y. Tao, Z. H. Rao, and S. Q. Liu. *J Biomol Struct Dyn 28*, 143-157 (2010).

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


***Comment***

# "Chargaff's Rules" for Protein Folding: Stoichiometric Leitmotif Made Visible

http://www.jbsdonline.com

*"Everything should be as simple as it is, but not simpler!"*
-Albert Einstein

Protein folding! The first thing that almost always comes to mind when someone hears this term is Anfinsen's hypothesis. So much so that protein folding and Anfinsen's hypothesis have long since been considered synonyms of each other.

Anfinsen's hypothesis laid the ground rule for protein folding. The rule does not even need rote learning, rather common sense: sequence determines structure which, in turn, determines function.

Then arrived several models and theories which tried to explain exactly how a protein folds into a perfect functional form commencing as soon as the amino-terminal chain emerges from the ribosome or after being fully synthesized. The models were all woven around a basic set of rules: no new bond making or breaking in the folded protein and global minimum-most free energy for the native structure. None of the methods, however, provide adequate instructions for the exact folding process. This can be seen by the simple fact that we are still far away from predicting an accurate, valid three-dimensional structure right from an arbitrary sequence using computational tools, with absolutely no mediator structures in-between. It is largely known that *in vivo*, the folding process for some proteins requires the aid of molecular chaperones.

In the manifold attempts towards elucidating the real mechanism, a recently published report by Mittal *et al*. (1) comes as a fresh breeze with a novel perspective. In accordance with Einstein's remarks above, they report a 'surprisingly simple unifying principle' in that the protein folding appears to be dictated by 'a narrow band of stoichiometric occurrences of amino-acids in primary sequences, regardless of the size and the fold of a protein'. The authors further maintain that contrary to all known perceptions, 'preferential interactions' between amino acids do not drive protein folding.

The authors devised an elegant and simple way of approaching the problem. Using the crystal structures of 3718 proteins present in Protein Data Bank (PDB), they analyzed the backbones in terms of the neighbourhood of every Cα atom. This was based on the assumption that if two amino acids interact *via* side chains, their respective Cα atoms would be 'neighbours' fixed in space, disregarding their actual position along the protein sequence. The Cα atoms occurring in the protein's 'center' would be expected to be surrounded by a higher number of Cα atoms of other amino acids. Considering each of the amino acids individually, a

**Seema Mishra**

National Institute of Biologicals
Ministry of Health and Family Welfare
A-32, Sector 62, NOIDA, U. P.,
India 201307

Corresponding Author:
Seema Mishra
E-mail: seema_nib@yahoo.com

20 x 20 matrix of the number of neighbours within a defined neighbourhood distance was reported for each protein. Plots for the number of times a specific amino acid appeared as a neighbour within a fixed distance were derived to investigate the presence of any 'preferential interactions'. Here, it should be noted that the non-covalent interactions between amino acids are the basis of the development of knowledge-based potentials, which in turn form the underpinning of protein structure prediction by modeling and simulation (2, 3 and references therein). The authors investigated whether amino acids prefer certain neighbourhoods arising out of non-covalent electrostatic, hydrogen bonds, hydrophobic and van der Waals interactions.

A sigmoidal trend was observed for the neighbourhoods of all the 20 amino acids and was found to be independent of the nature and identity of an amino acid. The immediate implications were that either several spatial distributions exist individually or there is one unifying spatial distribution pattern. A generalized, single equation fitting all the sigmoids was found and it brought forth the latter implication in full clarity. Further, a simple reasoning was applied that assuming there are no 'preferential interactions' involved, an amino acid that occurs as the highest number of times in the primary sequence would be expected to occur as a neighbour of all the 20 amino acids, including itself, the highest number of times. As expected, good correlation was established between the total number of contacts made by an amino acid and its stoichiometry (frequency of occurrence) in folded proteins. By direct implication, for the unstructured proteins, the authors found the stoichiometry to deviate from that of amino acids in folded proteins.

These studies preclude 'preferential interactions' and define stoichiometry of amino acids in the primary sequence governing the folding process. The very fact that the rules governing the folding of these proteins are universal makes it all the more inspiring.

The protein folding thesis by Mittal *et al.* is further strengthened by recent external evidence (4). Several protein properties, that range from size to folding cooperativity to dispersions in melting temperature, were found to depend universally on the number of residues, N. Even the timescale to reach the 'native basin of attraction' (native state) can be estimated in terms of 'N' residues. These predictions have been well-supported by experiments. But there is one major difference here, that it is the stoichiometry of amino acids rather than their number that has been found to govern.

The field of protein folding has been so intriguing that the ideas to understand the basic tenets have been sought even from the laymen. An online game, Foldit (5), recently-developed, serves to further this cause.

Until the 1990's, protein folding was, in terms of computation, an impregnable entity and has gradually been gaining ground since. One might be over-optimistic in that, given this universal principle based on stoichiometry, biologically meaningful *ab initio* modelling will indeed be possible in near future. An entirely 'new' structural space of many hypothetical as well as trans-membrane proteins will be visible, 'new' here referring to the fact that the basic premise will be applicable to one and all of the proteins and hence that, 'there is nothing new under the sun!'. With this, widely acknowledged Darwinian evolution rules might be re-written with some clauses added and others deleted.

It is well-known that a structure is more conserved than a protein's sequence. Mittal *et al*. have taken backbone atoms to investigate the folding process which clearly leads to universal applicability. It would mean that the basic fold is the same across the protein kingdom. Specificity and uniqueness are, but, imparted by the side chains in proper rotameric state. This conformational specificity is achieved as a result of packing and many weak interactions (6). Hence, in order to achieve a proper, functional and unique structure, stoichiometry does appear to play a role, albeit not *per se.*

While these findings are based entirely on theoretical approach, it is essential to consider these in the light of real scenario. Proper folding is usually thought to occur after a part or a domain of the protein is translated by the ribosomes. Apart from molecular chaperones to aid, existence of RNA-mediated chaperone type for folding *de novo* is documented (7). Co-translational kinetics may also play a role (8).

Protein folding *in vivo* is subjected to a different environment than *in vitro*. While *in vivo*, a cramped environment is provided to the protein; *in vitro*, it is generally a much more relaxed one. The cramped environment of the cell may force the protein to misfold causing aggregation (9). It is equally plausible that given relatively less number of solvent molecules due to the presence of more solutes around *in vivo*, the excluded volume effects have to be taken into consideration by the protein and fold correctly. Mittal *et al*. propose that folding has to be done while considering the structural constraints of the amino acid sequence and the order in which a given stoichiometry appears. Further, the authors opine that another parameter for folding is the individual shape characteristics of amino-acids along the sequence minimizing the surface-to-volume ratio. Entropic and energetic considerations of the folding in this manner would also seem favourable. When constituent molecules are used up in stoichiometric ratios as an organized principle, it is logical that the ordered folding may proceed at the expense of increased entropy because of the disorder induced by the unfolded ensembles in the surroundings.

The stoichiometric ratios for each amino acid in folded proteins give rise to "Chargaff's rules" for protein folding, and these rules suggest that leucine has the greatest average percentage occurrence (9.0 ± 2.9) and tryptophan (1.3 ± 1.0) and cysteine (1.8 ± 1.5) the least. This seems plausible since leucine, being a hydrophobic residue and, being among the most abundant in proteins, is expected to play a major role in the folding or packing arrangements of cytosolic proteins that tend to fold in a compact manner away from the solvent as well as the folding of trans-membrane proteins in a hydrophobic environment. No wonder that leucine zippers of transcriptional activators are an ideal system to study the folding dynamics (10). However, this is not to assume that tryptophan and cysteine residues have little role in the folding process because of their lowest percentage occurence. On the contrary, these residues have been shown to be necessary for proper folding of neprilysin (NEP)/endothelin-converting enzyme (ECE) family of metalloproteases, to cite one example (11). Further, if the folding process is uniform for both, intuitively, the question is: what exactly distinguishes the cytosolic and trans-membrane proteins with regard to their structure and hence, specific fate?

If we take this mechanism to be true, then it is entirely explainable that chaperones, obviously, have an indirect, supportive role in either coordinating or overseeing the process. It appears that along the path to a proper fold, if there were mediator structures mediating between any two flanking native-like conformations attempting to generate a consensus, they might force decisions into one direction that may not generate a proper fold. So, if a protein is to get its desired structure and function, the trick is to let it fold naturally without the help of any mediators; and it is here that the inter-play of stoichiometric ratios of amino acids driving the folding is widely apparent. The question remains, how disrupting a part of the protein by mutations, resulting in insertion, deletion or substitution of amino acids and hence, disruption of stoichiometric occurrences, is expected to let the protein be viable in its remaining fold as is sometimes observed.

Protein misfolding may also be well-explained in the context of this new finding. Other proteins within a milieu may compete for specific amino acids thereby leading to misfolding of those deprived of the amino acids in required stoichiometric ratios.

Do the proteins prefer just the native fold? Or variations leading to an alternate functional folded state are plausible? Two proteins were designed with 88% sequence identity and their different structures and functions were found to be encoded by only 12% of amino acids (12). It was further suggested that the free energy change of alternative folded states may be much more favourable. Depending upon stoichiometry and thermodynamics, the alternative folded states may be formed and preferred over the native.

Needless to say, common sense dictates that protein folding is environment- or context-dependent and therefore, folding may not be dependent upon stoichiometry alone. Many proteins are known to fold only in the presence of a ligand. Paralogs, common examples of which are myoglobin and haemoglobin with poor sequence identity but similar secondary structures, reside in different cellular micro-environments. Some proteins such as haemoglobin function only in a multi-domain form. If we consider a particular bacterium which has to live in an environment where some amino acids are limiting, there must be some mechanism for ensuring the availability and hence, maintenance of adequate stoichiometry of amino acids.

Due consideration should also be given to the possibility of the effect of codon bias in different organisms. In the case of silent mutations, another amino acid having the same physico-chemical properties is used to form a similar fold with virtually no phenotypic consequences. Providing an exception to Anfinsen's dogma, a silent polymorphism due to a single-nucleotide change that altered P-glycoprotein's function, was determined to be resulting from a fold different from the wild type (13). A rare codon for isoleucine was substituted in the place of wild type codon. In such a situation, exactly how the stoichiometric occurence of the amino acids encoded by rare and/or cognate codons is expected to drive the correct fold-formation is worth exploring further.

In disulfide-rich proteins, disulfide bond formation is found linked to protein folding activity thermodynamically. Structural disulfide bonds are formed only when there is correct backbone-to-backbone distance and proper orientation of cysteine residues. A mean distance of 5.6 Å is required between two alpha-carbons of cysteine residues of adjacent peptide backbones (14). In the case of oxidative folding of antibody Fab fragment, formation and re-arrangement of disulfide bonds is involved, and so, enzymes such as protein disulfide isomerase need to enter into the picture. Thus, it is apparent that it is not just stoichiometry (two cysteine molecules for one disulfide bond formation) that plays a role in oxidative protein-folding process, enzymes are essential here in the larger scheme of things.

The studies by Mittal *et al*. take into consideration only the proteins with more than 50 amino acid residues. It will be worthwhile to explore the folding of small proteins and peptides that are also crucial to the proper functioning of the cell in light of the new findings. For example, 15-amino acids-long helper T lymphocyte epitopes or the nonameric cytotoxic T lymphocyte (CTL) epitopes need to contact both major histocompatibility complex (MHC) and T cell receptors (TCR)

at any given point of time, and so, how do they adopt their particular fold? An idea about the involvement of stoichiometry in their folding is supported by a theoretical instance (15). In this study, most of the CTL epitopes shown to be interacting with MHC molecules have leucine or valine at the two primary anchor positions and also at the positions where TCR is expected to make a contact. Other examples of peptides for possible further studies are the five-amino acids-long met-enkephalin and the nine-amino acids-long hormone vasopressin.

Rather uniquely, the "Chargaff's rule" for protein folding reported by Mittal *et al.* simplifies both the gross and subtle mechanisms of protein folding and paves novel pathways for many more discoveries, yet to be unravelled. The immediate medical implication seems to be, obviously, a thorough understanding of protein misfolding and aggregation in diabetes, cancer, Alzheimer's and Parkinson's diseases among others ultimately leading towards the discovery and development of successful drugs. A recently developed highly effective prodrug with enhanced stability and long-term bioavailability, supramolecular insulin assembly (SIA-II) of insulin oligomers for diabetes treatment, has been developed taking into account the notions of protein folding (16). And what is more, productive and meaningful *ab initio* protein structure prediction is in our hands now-well, almost! Interestingly, these findings by Mittal *et al*. have been derived using standard mathematical principles for a biological problem. Who knows, there might be a magical number, similar to the universal 'Pi' ($\pi$) or the golden mean 'Phi' ($\Phi$) presiding over many natural phenomena, that governs even the protein folding process!

***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. D. Thirumalai, E. P. O'Brien, G. Morrison, and C. Hyeon. *Annu Rev Biophys 39*, 159-183 (2010).
5. S. Cooper, F. Khatib, A. Treuille, J. Barbero,J. Lee, M. Beenen, A. Leaver-Fay, D. Baker, Z. Popović, and F. Players. *Nature 466*,756-760 (2010).
6. E. E. Lattman and G. D. Rose. *Proc Nat Acad Sci USA 90*, 439-441 (1993).
7. S. I. Choi, K. Ryu, and B. L. Seong. *RNA Biol 6*, 21-24 (2009).
8. A. A. Komar. *Trends Biochem Sci 34*,16-24 (2009).
9. F. U. Hartl and M. Hayer-Hartl. *Nat Struct Mol Biol 16*, 574-581 (2009).
10. J. C. M. Gebhardt, T. Bornschlögla, and M. Riefa. *Proc Nat Acad Sci USA 107*, 2013-2018 (2010).
11. K. J. MacLeod, R. S. Fuller, J. D. Scholten, and K. Ahn. *J Biol Chem 276*, 30608-30614 (2001).
12. P.A. Alexander, Y. He, Y. Chen, J.Orban, and P. N. Bryan. *Proc Nat Acad Sci USA 104*, 11963-11968 (2007).
13. C. Kimchi-Sarfaty, J. M. Oh, I. W. Kim, Z. E. Sauna, A. M. Calcagno, S. V. Ambudkar, and M. M. Gottesman. *Science 315*, 525-528 (2007).
14. P. J. Hogg. *J Thromb Haemost 7*, 13-16 (2009).
15. S. Mishra and S. Sinha. *J Biomol Struct Dyn 24*, 109-121 (2006).
16. S. Gupta, T. Chattopadhyaya, M. P. Singh, and A. Surolia. *Proc Nat Acad Sci USA 107*, 13246-13251 (2010).

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


***Comment***

# New Direction to the Solution of Protein Folding Problem

**Angshuman Bagchi[1]**
**Tapash Chandra Ghosh[2*]**

[1]The Buck Institute for Age Research, Novato, California-94945, USA
[2]Bioinformatics Center, Bose Institute, P-1/12, C.I.T. Scheme VII-M, Kolkata-700 054, India

http://www.jbsdonline.com

Recently we saw an interesting paper entitled "A Stoichiometry Driven Universal Spatial Organization of Backbones of Folded Proteins: Are there Chargaff's Rules for Protein Folding?" by Aditya Mittal, B. Jayram, Sandhya Shenoy and Tejdeep Singh Bawa which appeared in the Journal of Biomolecular Structure and Dynamics (1). The paper comes up with a very revolutionary idea about protein folding. The authors point out that the frequencies of amino acids in the primary structures of proteins guide the folding patterns of the proteins. This very idea is quite contrary to the existing belief of the theory behind protein folding. According to the current understanding behind the mechanism of protein folding, the process of protein folding relies predominantly on non-covalent interactions between the side chains of amino acid residues in proteins (2-10) as well as with the surrounding environment in which the proteins reside. This preferential interaction between amino acids is the basis of the development of knowledge-based potentials and protein structure prediction which is common place nowadays by modeling and simulations. (11, 12 and references therein).

The authors in the paper observed, on the basis of their analysis of the crystal structures of proteins in the protein data bank (PDB), that the frequencies of amino acid residues in proteins govern the folding patterns of proteins. This is quite interesting and, if substantiated in future research, would change the whole scenario of the computational prediction of protein folding. Prediction of protein folding from the amino acid sequence of proteins depends basically on the extent of sequence similarity between the query protein and the template (13-15). As per the observations of the paper in question, proteins with similar amino acid frequency would likely to have similar folding patterns. We have a few questions regarding the hypothesis.

Is there any direct correlation between the amino acid residue frequency and the secondary structures of proteins? If so, then it would be a very essential step towards in-silico protein development. Next question is how does the amino acid sequence frequency determine the interaction patterns in proteins? The pattern of amino acid sequence frequency, while governing the folding patterns of proteins, should also interfere with the interaction patterns of proteins. If indeed this is the case, then this would shed light in the elucidation of the underlying mechanisms of different diseases. But one big question then is what are the effects of the cellular environments on protein folding? According to the results of the paper in question, minimization of the surface to volume ratio of proteins (based on the side chains of the constituting amino acid residues of the proteins) by exclusion of surrounding water molecules, is the most important parameter of the folding process. This correlates with the existing views of the mechanism of protein folding. As per the results of this paper, proteins

[*]Corresponding Author:
Tapash Chandra Ghosh
Phone: +91-33-23554766/
23552816/23556626
E-mail: tapash@bic.boseinst.ernet.in

are found to be enriched in hydrophobic amino acid residues (Leu, Ala, Val) and Glycine and depleted in aromatic amino acid residues (Phe, Trp, Tyr). It is well known that hydrophobic amino acid residues generally constitute the structural scaffolds of proteins, but aromatic amino acid residues also stabilize proteins by cation-pi interactions with basic amino acid residues (Lys, Arg) present in proteins (4-10). The average percentage of Cys residues is very low (second to the minimum which is Trp). Therefore the number of covalent disulfide bonds in proteins is much less. This supports the existing view of the molecular mechanism of protein folding, which suggests the preponderance of non-covalent interactions in proteins. Also the average percentages of polar and charged amino acid residues are much less. The less number of charged or polar amino acid residues in proteins means less interactions with the surrounding aqueous milieu and that means less surface-to-volume ratio. However, we are curious to know what the correlation is between the frequency of amino acid residues in proteins and the numerous intermediates that proteins have before getting converted to the natively folded form from the unfolded states; whether it would be possible to predict the structural aspects of the intermediates from the amino acid frequencies of the proteins. If that is possible then the mystery behind protein folding mechanisms could be solved. But it should also be noted that crystal structures are just some snapshots of the dynamic behavior of the ever changing macromolecules in cells (16-18). Therefore it is still far from certain what is actually happening inside the cells. The paper in question suggests an interesting and novel way of analyzing the mystery behind protein folding. Further validation of the hypothesis presented in the paper may be achieved by determining the structures of proteins from their amino acid residue frequency.

***References***


1. A. Mittal, B. Jayaram, S. Shenoy, and T. S. Bawa. *J Biomol Struct Dyn 28,* 133-142 (2010).
2. C. B. Anfinsen, E. Haber, M. Sela, and F. H. White Jr. *Proc Natl Acad Sci USA 47*, 1309-1314 (1961).
3. C. B. Anfinsen. *Biochem J 128,* 737-749 (1972).
4. C. B. Anfinsen. *Science 181*, 223-230 (1973).
5. C. Chothia. *Nature 357,* 543-544 (1992).
6. C. Pace, B. Shirley, M. McNutt, and K. Gajiwala. *FASEB J 10,* 75–83 (1996).
7. S. Deechongkit, H. Nguyen, P. E. Dawson, M. Gruebele, and J. W. Kelly. *Nature 403,* 101-105 (2004).
8. M. Karplus and J. Kuriyan. *Proc Natl Acad Sci USA 102*, 6679-6685 (2005).
9. G. Rose, P. Fleming, J. Banavar, and A. Maritan. *Proc Natl Acad Sci USA 103*, 16623-16633 (2006).
10. V. Sharma, V. R. I. Kaila, and A. Annila. *Physica A 388,* 851-862 (2009).
11. P. Sklenovský and M. Otyepka. *J Biomol Struct Dyn 27*, 521-539 (2010).
12. M. J. Aman, H. Karauzum, M. G. Bowden, and T. L. Nguyen. *J Biomol Struct Dyn 28*, 1-12 (2010).
13. C. Chothia and A. M. Lesk. *EMBO J 5,* 823-826 (1986).
14. M. J. Sippl. *Proteins 17,* 355-362 (1993).
15. S. Kaczanowski and P. Zielenkiewicz. *Theoretical Chemistry Accounts 125,* 543-550 (2010).
16. J. Drenth. New York: Springer-Verlag (1999).
17. C. Giacovazzo, H. L. Monaco, D. Viterbo, F. Scordari, G. Gilli, G. Zanotti, and M. Catti. Oxford University Press (1992).
18. J. P. Glusker. M. Lewis, and M. Rossi. VCH Publishers New York (1994).

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


***Comment***

# Does Stoichiometry Drive Protein Folding?

**Salvador Ventura**

Institut de Biotecnologia i de Biomedicina and Departament de Bioquímica i Biologia Molecular, Universitat Autònoma de Barcelona, 08193 Bellaterra (Barcelona), Spain.

http://www.jbsdonline.com

Proteins are synthesized in the cell as sequential strings of amino acids. However, ribosome-emerging polypeptide chains are not active and they have to fold into a usually unique three-dimensional (3D) structure to become functional. Many proteins have been shown to attain spontaneously *in vitro* their native state from an initially complex ensemble of unfolded conformations in the absence of the cellular protein quality control. This implies that the information determining how proteins gain their active 3D structures should be imprinted somehow in their primary sequences. Nevertheless, the specific connection between protein sequence and function has remained lost for almost six decades in the black-box of protein folding. Understanding how linear amino acids chains are pre-programmed to fold into their specific functional architectures remains as one of the most challenging and important tasks in molecular biology. Now, in a recent theoretical work Mittal *et al.* put forward a simple principle to explain backbone organization in protein folding (1). Essentially, the authors propose the absence of preferential long-, medium- or short-range interactions between amino acids during the folding of globular proteins. It must be mentioned that preferential interactions between amino acids were the basis for introducing knowledge-based potentials, which in turn provided the underpinning for present day 3D protein structure prediction by modeling and simulation (2-4 and references therein).

The process of folding would be determined instead by the stoichiometric distribution of the twenty proteinogenic amino acids in the primary sequences of proteins in a similar way that the stoichimetry of the four nucleotides, or Chargaff's rules, drive the spatial organization of the DNA double helix. The authors delineate these conclusions upon analyzing the distances between Cα atoms in 3718 proteins in the PDB concluding that the most important factors for the folding of a protein are exclusion by water, the shape of the individual amino acids (not their physicochemical properties) and the way these shapes are distributed along the sequence. There is no doubt that this is a revolutionary idea that if comes to be true would change our present view of protein folding and have many important implications in related areas like structural prediction, protein design or protein engineering.

Unfortunately, the theoretical framework proposed by Mittal *et al.* at least in its present formulation, clashes with different experimental evidences, like the folding behaviour of retro-proteins. Both the classical view and the new proposal agree that a polypeptide with exactly the same stoichiometry as a naturally occurring protein, but with a randomized distribution of residues or, in the authors' formulation, shapes would not get the same native conformation and very likely would not fold at all. But what would happen if we read the primary sequence backwards generating thus a retro-protein? Because native proteins are usually not palindromic in sequence, this exercise would result in a new polypeptide

Corresponding Author:
Salvador Ventura
Phone: 34-93-5868956
Fax: 34-93-5811264
E-mail: salvador.ventura@uab.es

that does not align with its parent sequence and therefore, according to the prevalent view of folding, it will not fold into the same 3D structure, in case it can fold. Instead, if one visualizes protein folding as proposed by Mittal *et al.* because they display identical stoichiometry and the same sequential distribution of amino acids shapes, their packing can potentially attain the same minimal surface-to-volume ratio and there is not obvious reason to doubt that the two protein will attain the same 3D structure. The reality is that, as a general rule, retro-proteins do not fold into the compact, stable and soluble conformations characteristic of natural globular proteins (5).

An important concept in the article by Mittal *et al.* is what the authors define as the "margin of life" or the specific distribution of amino acids compatible with folded proteins. The proposed stoichiometry displays low standard deviations, implying that significant divergences from this distribution would have deleterious structural effects. These so said Chargaff's rules of proteins would impose thus strong compositional constraints for the design of new non-natural proteins. However, several protein design exercises have succeeded in producing new proteins using reduced amino acid alphabets. Although the sequences of these proteins deviate significantly in composition from the one proposed by Mittal *et al.* they encode for stable, and topologically complex native conformations which are able to fold in a biologically relevant time frame (6). In addition, as shown for the paradigmatic case of bovine pancreatic trypsin inhibitor, extensively simplified proteins in which over one-third of the residues are alanines retain the overall backbone and side-chain configurations of the natural inhibitor (7). Together, these observations support a view in which most protein structure determinants occur at a few sites and involve the establishment of selected side chain-side chain interactions, questioning thus the role of strict stoichiometric laws in protein folding.

The discrepancies between Mittal *et al.* calculations and the above mentioned experimental evidences likely arise from the fact that the authors measure distances between Cα atoms in already folded globular proteins. Unfortunately, this says little about the process of folding, since the fact that Cα atoms are separated in the native structure by a short distance does not necessarily implies that the corresponding residues would be interacting in the native conformation, nor that they established contacts during the process of protein folding, since they could be brought in close vicinity simply by the interactions established by neighbouring residues. This brings us to the concept of folding nucleus, defined as key residues in the sequence, which contacts can lead the folding of a polypeptide chain towards the acquisition of its unique functional structure. The number of residues involved in such interactions is usually small. In this way, for acylphosphatase it appears that an extensive long range native-like contact network established by only three residues during the folding reaction is sufficient to determine the overall conformation of the protein (8). Importantly, in the native structure these three key residues are not in contact. Therefore, approximations like the one of Mittal *et al*. will underscore their crucial role in folding, while providing more relevance to residues that are in the vicinity simply because the interaction between residues in the folding nucleus make them come together, not being necessary that they interact directly with each other during or after folding.

Overall, although experimental data do not provide strong evidences for stoichiometry being a major factor controlling the folding reactions of natural proteins, we should not dismiss the relevance of the work of Mittal *et al.* It is clear that the field needs new ways of addressing the folding problem that by generating debate on the mechanisms underlying this process might allow us to advance towards the formulation of general and simple rules to understand how proteins fold.

***References***


1. A. Mittal, B. Jayaram, S. Shenoy, and T. S. Bawa. *J Biomol Struct Dyn 28*, 133-42 (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. Y. Tao, Z. H. Rao, and S. Q. Liu. *J Biomol Struct Dyn 28*, 143-157 (2010).
5. E. Lacroix, A. R. Viguera, and L. Serrano. *Folding and Design 3*, 79-85 (1998).
6. K. W. Plaxco, D. S. Riddle, V. Grantcharova, and D. Baker. *Current Opinion in Structural Biology 8*, 80-5 (1998).
7. M. M. Islam, S. Sohya, K. Noguchi, M. Yohda, and Y. Kuroda. *Proceedings of the National Academy of Sciences of the United States of America 105*, 15334-9 (2008).
8. M. Vendruscolo, E. Paci, C. M. Dobson, and M. Karplus. *Nature 409*, 641-5 (2001).

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


***Comment***

# Is the Folding Topology of a Protein Related to its Amino Acid Occurrence?

http://www.jbsdonline.com

**Juan J. Perez**

Dept d'Enginyeria Quimica (UPC)
ETSEIB Av. Diagonal, 647
08028 Barcelona, Spain

Proteins are molecular machines built from a polypeptide chain with capability to self-assemble into a compact 3D structure. This amazing capability is only exhibited by a subset of all the possible polypeptide chains built using the 20 natural amino acids so that, proteins in the living organisms are the result of years of natural selection (1). The physical basis of protein self-assembling or folding as is commonly referred, has puzzled researchers for the last fifty years and it is still a matter of debate (2, 3). Protein folding has been analyzed using methodologies that range from biophysical methods including structural, spectroscopic or computational to biochemical, including those of molecular biology. Present knowledge is the result of important research efforts basically focused on either understanding the structural features of known folded proteins, on the dynamics of the folding process or on the design of algorithms to predict the 3D structure of a protein from its sequence. These three aspects are closely related and need to be considered simultaneously, since prediction of the 3D structure of a protein involves a deep understanding of the structural features of folded proteins together with a solid knowledge of the driving forces involved in the folding process, including the role played by the solvent (4).

Pioneering experiments performed on Ribonuclease at the beginning of the seventies and followed by hundreds of similar experiments with different proteins, indicate that polypeptide chains fold and unfold reversibly and that all the necessary information of the folding process is embedded in their amino acid sequence (5). These experiments provide a solid basis to consider that the native structure is a minimum in the free energy landscape of a protein.

The first structures of proteins were released from X-ray diffraction studies in the early sixties (6) and confirmed that protein structure is organized in structural subunits, called α-helices and β-sheets. This result had been previously hypothesized from the analysis of the hydrogen bonding capability of polypeptide chains (7). In addition, the first crystal structures also revealed that these structural elements appear much more compact than previously anticipated. Since then, a large number of structures solved from either X-ray diffraction, NMR or cryo-electron microscopy have been deposited in the protein data bank, playing a pivotal role for understanding the structural features exhibited by proteins in their native states. Much of the work has been devoted to identify secondary structure trends of sequences (8), to assess the number of folds expected for proteins to adopt (9) or to identify rules about the neighboring trends of the different amino acids (10).

In the last ten years site directed mutagenesis has also provided useful information about protein structure. One of the most striking results was to demonstrate

*Corresponding Author:
Juan J. Perez
Phone: +34934017150
Fax: +34934017150
E-mail: juan.jesus.perez@upc.edu

that the structure of a protein is only determined by a reduced number of amino acids (11, 12). These experiments suggest that there are key residues in the sequence that dictate the 3D structure and that already formed secondary structures are not necessarily determinants of the structure.

Analysis of the folding kinetics is another importance source of information. Protein folding takes place from micro to milliseconds at room temperature. If folding is considered as the process of finding the right pathway to the lowest free energy minimum, taking into account the large number of energy minima of the potential energy surface, a polypeptide chain cannot be expected to explore all possible configurations in reasonable time (13). Instead, the folding landscape must be funneled: after going through valleys and ridges where they may get trapped, proteins collapse into their native state very quickly through a cooperative process. This process should be understood in terms of a collective behavior in which individual structures of an ensemble achieve their final state following different pathways: "all roads lead to Rome". In other words, the folding process should be viewed as a phase transition and not as a chemical reaction (14).

Spite the level of maturity achieved there is still a debate in the scientific community about the driving force of protein folding. Present knowledge favors thinking that hydrogen bonding interactions account for the formation of secondary structure elements, but do not provide a full explanation of the folded, compact structure of globular proteins. In contrast, hydrophobic interactions are considered as the driving force of protein folding (3). The latter hypothesis gained further support after the shown prediction capabilities of threading models developed in the eighties (15). Calorimetric measurements indicate that the free energy change of the folding process is about 10-15 kcal/mole and it is size dependent. This values, of the order of a single hydrogen bond suggest that the net result of the folding process must be a delicate balance of electromagnetic forces and entropy losses (4).

At present, the availability of new biophysical techniques, the capability to perform longer computer simulations together with the improved capabilities of protein engineering have provided a renewed interest in getting a deeper understanding of the underlying principles of protein folding, together with a considerable effort devoted to develop new methods to predict the 3D structure of a protein from its sequence, with the accuracy necessary to understanding its biological function. In addition, the protein data bank currently contains more than 65,000 entries with a rate of incorporation of more than 5,000 new structures per year, offering a renewed opportunity to revise our ideas of the structure of folded proteins.

Taking advantage of the information accumulated in the protein data bank, a very stimulating paper has been recently published in this journal (16). The authors conducted a thorough analysis using around 4,000 protein structures aimed at identifying patterns of preferential clustering of specific residues around each other, concluding that the amino acid occurrence is a necessary condition for a protein to fold. In their study, the authors carried out a neighborhood analysis for each of the amino acids in the sequence for each of the proteins in the set, using different distance cutoffs. They found that the total number of neighbors for each amino acid correlates well with its average occurrence in the set of proteins studied independently of the cutoff considered. Indeed, the asymptotic behavior of the number of neighboring residues of an amino acid in a protein with the distance, necessarily yields a measure of the occurrence of the different amino acids in folded proteins and this needs to be consistent independently of the amino acid used as reference. The amino acid occurrence numbers provided in the paper agree well with other estimates already reported in the literature using different protein sets (8). The more surprising result of the report is that this result is independent of the cutoff distance.

Although these results are very stimulating they need to be contrasted with other observations. First, to provide support to this proposal in a recent report using a set of 1612 non-redundant protein structures, it was found that the amino acid occurrence is an important discriminant of the protein folding topology (17). Altogether these results suggest that folded proteins exhibit a narrow range of amino acid occurrences and this range can discriminate for the fold attained. Accordingly, unstructured proteins should have amino acid occurrences outside this margin, as the authors suggest in their report. However, it should be noted that with the exception of arginine, a random translation of the genetic code yields amino acid occurrences within to the occurrences found in folded proteins (18). On the other hand, independent analysis carried out recently using a set of 1736 non-redundant proteins reports contact numbers for the different amino acids that do not show any correlation with their occurrence in folded proteins at different cutoff distances between 8 and 20 Å (10). In the same report, the authors show a correlation between residue accessibility to the solvent and its contact number, with hydrophobic residues exhibiting the highest number of contacts and charged the least. In our opinion the evidence provided cannot be considered as definitive, since the standard deviations of the average numbers seem to be dramatically important to infer the necessary conclusions. Accordingly, a more detailed analysis is required where the statistical significance of the numbers reported can be assessed.

Understanding protein folding has still a way to go. New algorithms need to be designed for protein structure prediction. A fundamental question still to be answered is about how folding is initiated. Is it due to a close packing of secondary structure elements already formed or by specific residues or turns

that act as nucleation sites? A deeper understanding needs to be build around the compensation between energy and entropy contributions in the process and also the role played by the solvent, in other words the role play by the different players and the hierarchy of the process.

***References***


1. T. E. Creighton. *Proteins Structures and Molecular principles.* 2nd Ed. W. H. Freeman and Co, New York, NY 1992.
2. G. D. Rose, P. J. Flemming, J. R. Banavar, and A. Maritan. *Proc Natl Acad Sci USA 103*, 16623-1633 (2006).
3. K. A. Dill. *Biochemistry 29,* 7133-7153 (1990).
4. K. A. Dill, S. B. Ozkan, M. S. Shell, and T. R. Weikl. *Ann Rev Biophys 37*, 289-316 ( 2008).
5. C. B. Anfinsen. *Science 181*, 223-230 (1973).
6. J. C. Kendrew *et al*. *Nature 181*, 662 (1958).
7. L. Pauling and R. B. Corey. *Proc Natl Acad Sci USA 37*, 235 (1951).
8. 8. J. M. Otaki, M. Tsutsumi, T. Gotoh, and H. Yamamoto. *J Chem Inf Model 50*, 690-700 (2010).
9. Y. Zhang, I. A. Hubner, A. K. Arakaki, E. Shakhnovich, and Jeffrey Skolnick. *Proc Natl Acad Sci USA 103*, 2605-2610 (2006).
10. G. Faure, A. Bornot, and A. G. de Brevern. *Biochimie 90*, 626-639 (2008).
11. D. Baker. *Nature 405*, 39-42 (2000).
12. Y. He, Y. Chen, P. Alexander, P. N. Bryan, and J. Orban. *Proc Natl Acad Sci USA 105,* 14412-14417 (2008).
13. C. Levinthal. *J Chim Phys 65*, 44-45 (1968).
14. K. A. Dill and H. S. Chan. *Nat Struct Biol 4*, 10-19 (1997).
15. J. Skolnick, D. Kihara, and Zhang Y. *Proteins 56*,502-518 (2004).
16. A. Mittal, B. Jayaram, S. Shenoy, and T. S. Bawa. *J Biomol Struct Dyn 28*, 133-142 (2010).
17. Y-h Taguchi and M. M. Gromiha. *BMC Bioinformatics 8*, 404-414 (2007).
18. J. L. King and T. H. Jukes. *Science 164*, 788-798 (1969).

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


***Comment***

# Are there Still Surprises Buried Inside Statistical Analysis of Protein Structure?

http://www.jbsdonline.com

**Ravishankar Ramanathan**
**Abhinav Verma***

Centro de Investigaciones Biológicas (CSIC) Calle de Ramiro de Maeztu 9, 28040 Madrid, Spain

The Protein Folding Problem has been a long standing challenge towards achieving a complete molecular understanding of life (1). Since the days of Anfinsen, scientists over the world have been trying to understand the basic laws governing protein folding and have come a long way. We currently understand the general forces determining the protein folding such as the hydrophobic effect, secondary structure formation and the role of preferential interactions among amino-acids. However we don't know their relative contributions at the quantitative level. Thus we are not capable of predicting the folding properties of proteins from their amino-acid sequence yet.

The article by Mittal *et al.* (2) investigates this issue with a fresh twist of the classic approach of extracting rules from the analysis of protein 3D structures. Particularly the researchers perform a detailed analysis of Cα- Cα contacts at various neighborhood distances, looking for correlations up to unconventionally high distance cut-offs. Various efforts have been made to study the effect of interactions among amino-acids, but for the first time the authors study the effect and importance of these interactions in protein folding over long distances. Their hypothesis is that if there are any preferential interactions between different amino-acids as it is usually assumed, then the contacts distribution at different distance cut-offs should exhibit specific biases. But if there are no preferential interactions then the contacts distribution would just follow the amino acid frequency distribution in natural sequences. From their detailed analysis, they discover that the composition of the structural environment for any amino-acid in folded proteins is directly proportional to the natural frequency of occurrence of the amino-acid rather than any preferential interactions. In other words, the distributions of amino-acid in proteins follow simple stoichiometric relationships that they call "Chargaff's rules" for protein folding.

Over the years, there have been numerous efforts pertaining to determination of amino-acid interactions and to the development of pair-wise interaction potentials and in their application in coarse-grained computer models of proteins (3). In fact, a popular method assumes a quasi-chemical (4) approximation to determine pair-wise interaction strengths from the frequency of amino-acids pairs in contact. The Mittal *et al.* study departs from this general idea because it looks for correlations at very long distances (up to 60 Å). Their observation that spatial distribution of amino-acids is identical to that dictated by the amino-acid composition is a provocative result that opens up many questions which needs to be addressed in the future.

For example, the authors have performed an analysis for neighborhood distances up to 60 Å, whereas correlations are usually investigated at much shorter dis-

*Corresponding Author:
Abhinav Verma
E-mail: abhinav@cib.csic.es

tances. Therefore there is an issue regarding whether when one goes up to such distances, is it still reasonable to expect specific correlations. When one calculates pairwise correlations over such long distances the number of possible pairs is so large that it is not unreasonable for them to average out resulting in a Gaussian distribution with its peak around the average radius of the proteins within the dataset (20-25 Å). The cumulative sum of a Gaussian distribution produces a sigmoidal curve that may be indistinguishable from those observed by the authors.

On the other hand, in a recent article, Jha *et al.* (5) have shown the presence of preferential interactions among amino-acids in protein folding by using an analysis based on a scoring matrix of Cα-Cα contacts in different structural environments. In this connection, it should be noted that 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 (6-9 and references therein). Another article by Berka *et al.* (10) investigates the intramolecular energies for pairs of amino-acid side-chains using *ab initio* quantum mechanical calculations. These articles observe skewed distributions that indicate the influence of amino-acid interactions, which is contrary to the observations of Mittal *et al.* and therefore it is important to complement the Mittal *et al.* study investigating the interactions at the level of specific amino-acids pairs, look for effects at shorter neighborhood distances and possibly include side-chains in the analysis.

Also, there could be astronomically high number of sequences with the natural distribution of amino-acids, yet not all of them are observed in nature and fold into proteins. In fact, experiments suggest that most random sequences are likely to be dysfunctional folding-wise. So, the amino-acid composition of natural proteins may already include long-range correlations selected by evolution that would be missed out with the Mittal *et al.* approach since the natural composition is used as reference state.

Summarizing, the authors here present a novel, unconventional approach to study protein folding from the statistical analysis of protein structures and make a surprising discovery that suggests the absence of specific amino-acid pairwise interactions in protein folding. If confirmed, these findings would challenge the currently accepted view, but more research is needed to clarify several outstanding issues.

***Acknowledgements***

RR and AV acknowledge the support of Marie Curie Actions through Marie Curie Excellence Grant MEXT-CT-2006-042334.

***References***

1. D. Thirumalai, E. P. O'Brien, G. Morrison, and C. Hyeon. *Ann Rev Biophys 39,* 159-183 (2010).
2. A. Mittal, B. Jayaram, S. Shenoy, and T. S. Bawa. *J Biomol Struct Dyn 28,* 133-142 (2010).
3. R. Dima, G. Settanni, C. Micheletti, J. R. Banavar, and A. Maritan. *J Chem Phys 112,* 9151-9166 (2000).
4. S. Miyazawa and R. L. Jernigan. *Macromolecules 18*, 534-552 (1985).
5. A. N. Jha, S. Vishveshwara, and J. Banavar. *Prot Sci 19,* 603-616 (2010).
6. P. Sklenovský and M. Otyepka. *J Biomol Struct Dyn 27*, 521-539 (2010)
7. M. J. Aman, H. Karauzum, M. G. Bowden, and T. L. Nguyen. *J Biomol Struct Dyn 28*, 1-12 (2010).
8. C. Koshy, M. Parthiban, and R. Sowdhamini. *J Biomol Struct Dyn 28*, 71-83 (2010).
9. Y. Tao, Z. H. Rao, and S. Q. Liu. *J Biomol Struct Dyn 28*, 143-157 (2010).
10. K. Berka, R. Laskowski, K. E. Riley, P. Hobza, and J. Vondrasek. *J Chem Theory Comput 5,* 982-992 (2009).

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


***Comment***

# The Yeast Prion Case: Could There be a Uniform Concept Underlying Complex Protein Folding?

http://www.jbsdonline.com

**Youtao Song[1,2]***
**Yao Song[1]**
**Xing Chen[1]**

[1]School of Life Science, Liaoning University, Shenyang 110036, China
[2]Province Key Laboratory of Animal Resource and Epidemic Disease Prevention, Shenyang 110036, China

To date, there have been many hypotheses on protein folding and certain progress surely has been made, such as Anfinsen's dogma (1), Levinthal paradox (2). Recently, Mittal *et al*. analyzed close to 4,000 folded proteins from their published crystal structures in the protein data bank (PDB) with a bioinformatics and computational method, and innovatively proposed that protein folding is a direct consequence of a narrow band of stoichiometric occurrences of amino-acids in primary sequences, regardless of the size and the fold of a protein (3). Contrary to all prevalent views, this hypothesis actually negated the roles of specific amino-acid interactions and the sequence order of the amino acids in 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 (4-7 and references therein). Although their data analysis approaches seem scientifically correct, the "unified conclusion" drawn from a quantity of statistic data may not explain principles complied by every individual protein for its folding. Nevertheless, spatial distribution of neighborhoods for all amino-acids, rather than residues adjacent along the primary sequence, determine the protein folding, as proposed by Mittal *et al.* is a meaningful finding.

It is the study of the properties of protein folding in certain types of proteins that is being used to deduce the common properties shared by other proteins; such studies has already laid the foundation for almost every hypothesis on protein folding. For example, the reason why Anfinsen's dogma is widely accepted is that it is deduced from the study on the features of ribonuclease molecule (1). On the other hand, the inspiration of Mittal *et al*. was originated from Chargaff's Rules, a statement on DNA composition properties. It is well known that DNA is composed by only 4 kinds of nucleotides while the types of amino acids in proteins are as many as 20. Furthermore, DNA composition is relatively simple and conservative in all species; but the structure of protein is far more complicated in that different proteins have different structures even in the same species, and the structure of a protein with the same function is different in different species. Therefore, it is extremely difficult trying to use one uniform concept to explain all kinds of protein foldings and structural features.

For a decade, our group has been constantly devoted to the study on protein misfolding diseases. The conformational conversion of amyloid proteins, especially prions, is associated with numerous protein aggregation pathologies and infectious properties. We will comment on this issue based on the data obtained from molecular biological and molecular dynamic prion protein studies.

*Corresponding Author:
Youtao Song
Phone: 86-24-6220-2280
Fax: 86-24-6220-2280
E-mail: ysong@lnu.edu.cn

Each prion protein contains one prion-forming domain (PrD), essential to its aggregation. It is found that in the PrDs of Ure2p, Sup35p and Rnq1p, the first three yeast prion proteins to be identified, the glutamine (Q) and asparagine (N) content is unusually high (48% for Ure2p-1–89, 46% for Sup35p and 43% for Rnq1p) (8-10). Mutagenesis studies of Sup35p and Ure2p have confirmed the important role played by the Q/Ns content in prion formation (8, 9). Based on the special amino acid Q/Ns' content and other factors to compose prion, Lindquist and her co-workers conducted a bioinformatic proteome-wide survey to score every PrD in yeast and revealed that ~ 200 proteins have candidate PrDs, in which the amyloid and prion-forming properties of the 100 highest-scored PrDs were tested (11). They identified 24 proteins that satisfied the criterion for prion behavior, of which several were proved to be prion (New1p, Swi1p, Mot3p and Cyc8p). Their study has suggested that specific amino-acids could contribute to the structural feature of a certain protein.

Subsequently, Wickner and his team used Ure2p as a model system and randomly shuffled the order of amino acids of its PrD while keeping the amino acid composition (12). Five Ure2p variants were generated. Test performed *in vitro* shows that the prion domains of all five have readily formed amyloid fibers under native conditions. Meanwhile, four of them formed stable prions *in vivo*, and the fifth formed unstable prions that could only be maintained and transmitted under selective conditions. Liu *et al.* studied Sup35p and shuffled its PrD without altering its amino acid composition, which even formed amyloid fiber *in vitro* (13). The study on the prion formation of shuffled PrD of Sup35p *in vivo* was conducted by Wickner's group (14). Seven shuffled variants were generated, five of which expressed normally *in vivo*. And for these five, it was found that four of them formed stable prions and the fifth formed unstable prion. It needs to be pointed out that their results suggested that the Sup35p oligopeptide repeats (PQGGYQQYN, which is repeatedly expressed in prion domain of Sup35p) were not indispensable for prion formation. These oligopeptide repeats were first found in mammalian prion protein and were thought to influence their conformational conversion to the prion state (15).

Moreover, our group has used the molecular dynamics simulation to study the aggregation characters of a short 7 peptide fragment (GNNQQNY) in yeast prion Sup35p which could form amyloid fibrils (16, 17). The seven amino-acid residues were reorganized randomly into 9 different fragments (shown in Table I), without changing any amino acid content. We performed 20ns simulation for each fragment system at pH 7 and temperature 330 K. The RMSD (Root Mean Square Derivative) value of each fragment system after the simulation was quite close, between 3.214 and 4.473, which indicated that the structural diversity of each fragment were very similar. Hereafter we calculated aggregation time of each fragment system and found that these nine systems aggregated into one cluster eventually despite different time they spent. Although changing the permutation of the seven amino-acid residues has made impacts on the aggregation speeds of nine systems, their aggregation properties have not been influenced. Both Wickner and our groups' results indicated that the aggregation properties of prion proteins are independent to the order of amino-acid sequences.

**Table I**
The sequences of 9 randomly reorganized 7 peptides and the computation after simulation.

| | Sequence | RMSD (Å) | Aggregation Time* (ns) |
|---|---|---|---|
| wt | GNNQQNY | 3.214 | 5.91 |
| 1 | GQNQNNY | 4.391 | 4.95 |
| 2 | GNQNQNY | 3.494 | 1.97 |
| 3 | GNNQNQY | 4.073 | 11.89 |
| 4 | GQNNQNY | 4.473 | 1.89 |
| 5 | GQNNNQY | 3.634 | 4.87 |
| 6 | GNQNNQY | 4.399 | 3.99 |
| 7 | GQQNNNY | 3.539 | 9.24 |
| 8 | GNQQNNY | 3.686 | 7.58 |
| 9 | GNNNQQY | 3.833 | 6.39 |

*Aggregation time, the time that each fragment system aggregate to one cluster.

Our opinion for the thesis of Mittal *et al*. comes from data on only one kind of specific proteins. Results obtained cannot be employed to prove that all the protein folding features of all proteins may be in accordance with this principle. We believe that special amino-acids should be crucial for the structural features of certain type of proteins, and it might only be in a limited fixed type of proteins that the occurrence of amino-acids (stoichiometry) determines the structural features. However, it is still open to question whether stoichiometry driven protein folding is a universal concept that applies to all proteins.

***References***


1. C. B. Anfinsen. *Science 181*, 223-230 (1973).
2. C. Levinthal. *J Chim Phys 65*, 44-45 (1968).
3. A. Mittal, B. Jayaram, S. Shenoy, and T. S. Bawa. *J Biomol Struct Dyn 28*, 133-142 (2010).
4. P. Sklenovský and M. Otyepka. *J Biomol Struct Dyn 27*, 521-539 (2010).
5. M. J. Aman, H. Karauzum, M. G. Bowden, and T. L. Nguyen. *J Biomol Struct Dyn 28*, 1-12 (2010).
6. C. Koshy, M. Parthiban, and R. Sowdhamini. *J Biomol Struct Dyn 28*, 71-83 (2010).
7. Y. Tao, Z. H. Rao, and S. Q. Liu. *J Biomol Struct Dyn 28*, 143-157 (2010).

8. M. L. Maddelein and R. B.Wickner. *Mol Cell Biol 19*, 4516-4524 (1999).
9. A. H. DePace, A. Santoso, P. Hillner, and J. S.Weissman. *Cell 93*, 1241-1252 (1998).
10. I. L. Derkatch, M. E. Bradley, J. Y. Hong, and S. W. Liebman. *Cell 106*, 171-182 (2001).
11. S. Alberti, R. Halfmann, O. King, A. Kapila, and S. Lindquist. *Cell 137*, 146-158 (2009).
12. E. D. Ross, U. Baxa, and R. B. Wickner. *Mol Cell Biol 24,* 7206-7213 (2004).
13. Y. Liu, H. Wei, J. Wang, J. Qu, W. Zhao, and T. Hung. *Biochemical and Biophysical Research Communications 353,* 139-146 (2007).
14. E. D. Ross, H. K. Edskes, M. J. Terry, and R. B. Wickner. *Proc Natl Acad Sci USA 102*, 12825-12830 (2005).
15. I. S. Shkundina, V. V. Kushnirov, M. F. Tuite, and M. D. Ter-Avanes yan. *Genetics 172*, 827-835 (2006).
16. Y. O. Chernoff. *FEBS Letters 581*, 3695-3701 (2007).
17. M. Balbirnie, R. Grothe, and D. S. Eisenberg. *Proc Natl Acad Sci USA 98*, 2375-2380 (2001).

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


***Comment***

# Stoichiometry in Protein Folding?: Deeper Insights may be Useful

http://www.jbsdonline.com

**Rajendra Joshi**

Group Coordinator: Bioinformatics
Centre for Development of
Advanced Computing (C-DAC)
Ganeshkhind, Pune 411007,
Maharashtra, India

Understanding the mechanisms of protein folding has remained a mystery since many decades and there have been a number of models or views which have been proposed over the period of time (1-4). As of today, there has been a largely accepted statistical view of protein folding which suggests that a protein starts from an ensemble of unfolded states and folds through thousands of independent microscopic pathways, and finally converging to the native state (5-8). It has been widely believed and accepted that amino acids side chains play a very important role in the folding process through electrostatics, hydrogen bonding, hydrophobic interactions *etc.*, Numerous experimental and theoretical reports are available to support this view as well (9-10). In fact 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 (11-14 and references therein).

The authors of the paper "A Stoichiometry Driven Universal Spatial Organization of Backbones of Folded Proteins: Are there Chargaff's Rules for Protein Folding?", have made a fresh attempt to relook at this age old problem of protein folding with a completely different view point. In the paper published in this Journal (15), Mittal *et al*. bring forth the view that protein folding is a direct consequence of stoichiometric occurrences of amino acids, regardless of the size and fold of the protein. Mittal *et al*. also claim that the preferential interactions between amino acids do not drive protein folding. Going a step further, the authors also claim that the protein folding is a consequence of the interaction of the C-alpha atoms in backbone rather than the side chains. This obviously means that the side chains of the amino acids do not play any role in the protein folding process. This is contrary to the well accepted view that polar and non-polar amino acids are the primary driving force in the protein folding process.

This paper brings in a fresh insight into the protein folding problem, and adds another factor of "stoichiometry" into the numerous factors which govern the mechanisms of protein folding. Though stoichiometric based analysis seems interesting, there may be a few points which need to be addressed which may bring more light into this:

1. The authors have carried out an analysis of spatial distribution of amino acids in 3D space. The analysis has been carried out for a range between 0 to 9 Å and subsequently between 10 to 90 Å. The amino acids lying in this range have been defined as "contacts". Since the distance of 90 Å is very large one needs to see how significant such interactions could be to affect the entire folding process. It would be more interesting to carry out the

Corresponding Author:
Rajendra Joshi
Phone: +91-20-25694084
Fax: +91-20-25694084
E-mail: rajendra@cdac.in

analysis using "native contacts" which would be in the range 5 to 6 Å and which are well known to play an important role in protein folding. Having done an analysis of spatial distribution of amino acids, it would also be important to see the effects of crystal packing and symmetry on this analysis.

2. The plots of number of contacts versus neighborhood distances show a sigmoidal curve. One needs to explain why a sigmoidal curve is obtained in such an analysis. The analysis claims that a single amino acid independent spatial distribution is obtained for all cases. Does this necessarily mean that the amino acid side chains do not have any role to play in the folding process?
3. If the stoichiometric based hypothesis is true, it would be interesting to examine two proteins having different sequences but having the same stoichiometry. Do they have the same folding pattern? Such cases need to be examined and can be a proof for the hypothesis. To take things further, it would be good to experimentally synthesize such molecules and study their folding pattern.
4. Similarly, can one explain the phenomena of misfolding wherein the amino acid stoichiometry remains same but still the folding pattern changes? In a number of cases, it is well known that chaperones assist in protein folding process. If stoichiometry is the main factor in protein folding, then why does a protein require chaperones for folding?
5. Can we study cases of functional proteins like enzyme, wherein a mutation can change the entire conformation of an active site? It would be interesting to study the effect of a mutation on the protein fold, in one case which changes the stoichiometry and another which does not.
6. Lastly, can we formulate stoichiometric rules for formation of an alpha-helix, beta-sheet, coil *etc.,* This, in my opinion would be the most interesting aspect of this work.

It may be possible that amino acid stoichiometry could be another important factor contributing to protein folding. The paper by Mittal *et al*. is very interesting and can add further value to our understanding of protein folding.

***References***


1. C. B. Anfinsen. *Science 181*, 223-230 (1973).
2. C. Levinthal. *J Chim Phys 65*, 44-45 (1973).
3. D. Thirumalai, E. P. O'Brien, G. Morrison, C. Hyeon. *Annu Rev Biophys 39*, 159-183 (2010).
4. C. M. Dobson, A. Sali, and M. Karplus. *Angew Chem Int Ed 37*, 868-93 (1998).
5. K. A. Dill, S. Bromberg, K. Z. Yue, K. M. Fiebig, D. P. Yee, P. D. Thomas, and H. S. Chan. *Protein Science 4*, 561-602 (1995).
6. K. A. Dill and H. S. Chan. *Nat Struct Biol 4*,10-19 (1997).
7. K. A. Dill, S. B. Ozkan M. S. Shell, and T. R. Weikl. *Annu Rev Biophys 37,* 289-316 (2008).
8. G. D. Rose, P. J. Fleming, J. R. Banavar, and A. Maritan. *Proc Nat Acad Sci USA 103*, 16623-16633 (2006).
9. V. Z. Spassov, L. Yan, and P. K. Flook. *Protein Science 16,* 494-506 (2007).
10. H. J. Dyson, P. E. Wright, and H. A. Scheraga. *Proc Natl Acad Sci USA* 103, 13057-13061 (2006).
11. P. Sklenovský and M. Otyepka. *J Biomol Struct Dyn 27*, 521-539 (2010).
12. M. J. Aman, H. Karauzum, M. G. Bowden, and T. L. Nguyen. *J Biomol Struct Dyn 28,* 1-12 (2010).
13. C. Koshy, M. Parthiban, and R. Sowdhamini. *J Biomol Struct Dyn 28*, 71-83 (2010).
14. Y. Tao, Z. H. Rao, and S. Q. Liu. *J Biomol Struct Dyn 28*, 143-157 (2010).
15. A. Mittal, B. Jayaram, S. Shenoy, and T. S. Bawa. *J Biomol Struct Dyn 28*, 133-142 (2010).

*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 $C_\alpha$-$C_\alpha$ 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 $C_\alpha$-$C_\alpha$ 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 $C_\alpha$-pairs of amino acids whose side chains interact would be distinct from the spatial organization of $C_\alpha$-pairs of amino acids whose side chains do not interact. Number of $C_\alpha$-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

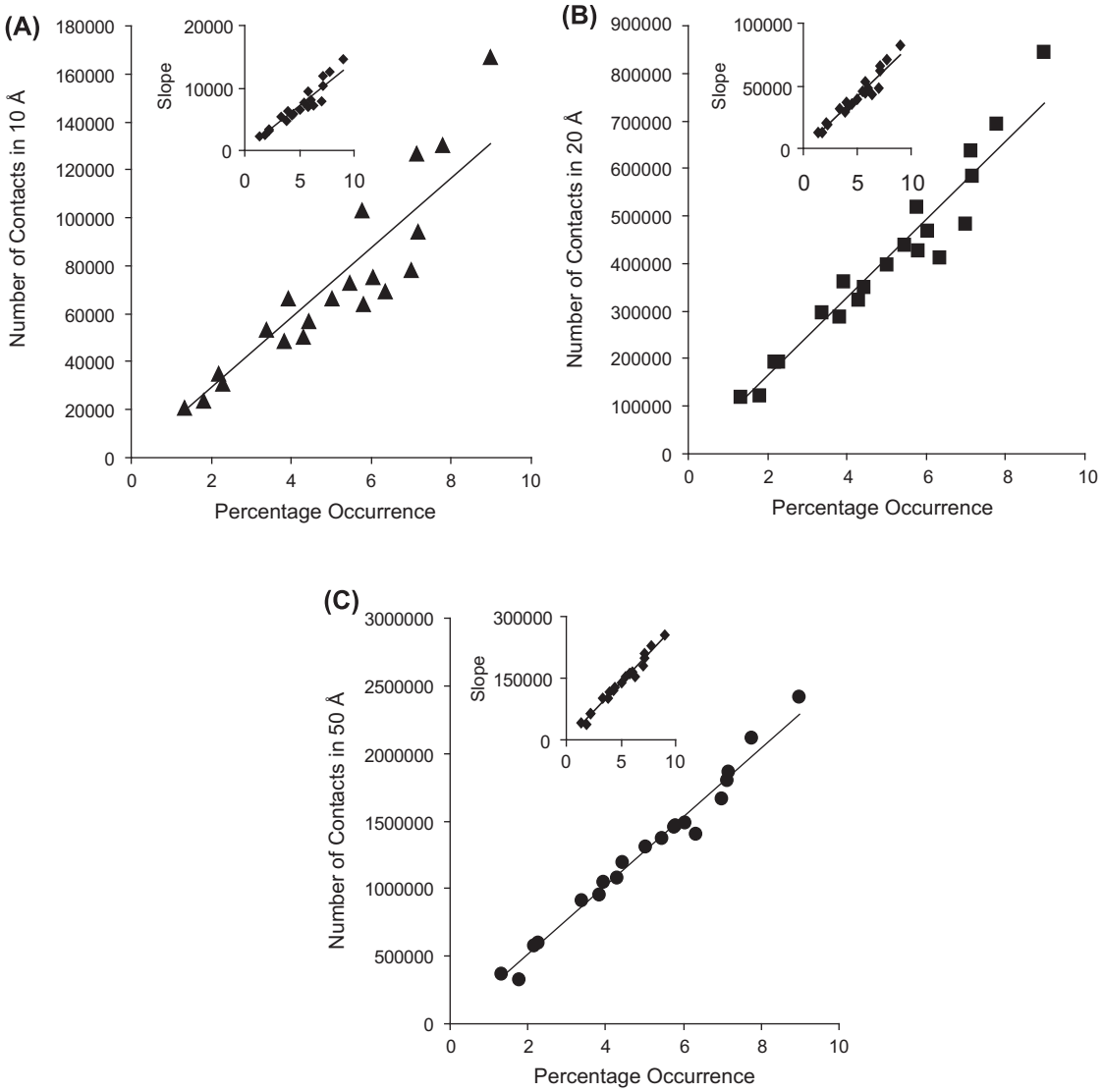


**Figure 1:** Neighborhoods of amino-acids in folded proteins are determined by simply their stoichiometry in primary sequences, regardless of the definition of neighborhood distance – (**A**) Neighbors of leucine within a 10 Å neighborhood are correlated well with their frequency of occurrence in folded 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 $C_\beta$ 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).