UNIVERSALITIES IN PROTEIN TERTIARY STRUCTURES: SOME NEW CONCEPTS

B.JAYARAM[1,2]*, ADITYA MITTAL[1]*, AVINASH MISHRA[1], CHANCHAL ACHARYA[1], GARIMA KHANDELWAL[2]

[1]Kusuma School of Biological Sciences, Indian Institute of Technology Delhi, Hauz Khas, New Delhi 110016, India.[2]Department of Chemistry & Supercomputing Facility for Bioinformatics & Computational Biology, Indian Institute of Technology Delhi, Hauz Khas, New Delhi 110016, India.

**Email: bjayaram@chemistry.iitd.ac.in ; amittal@bioschool.iitd.ac.in*

We discuss the recent extraction of signatures of stoichiometry driven universal spatial organization of backbones of folded proteins regardless of their size, shape/structure and function. We present further evidence for secularity of amino acids in protein structures from the perspectives of surface area and energy. While conceptual fragmentation to gain insights into the diversity of protein structures appears to be a popular approach, we believe that the secrets to solving the protein folding problem lie in appreciating concepts that are universally applicable.

1. Introduction

Historically, the ideas of Pauling[1-3] and Ramachandran[4] proposed in 1950s and 1960s established universality among secondary structures in proteins. Pauling's work led to our understanding that proteins, irrespective of their structure and function, are made up of regular secondary structural elements called alpha helices and/or beta sheets and irregular regions connecting these called loops with Ramachandran's work providing a raison d'etre in terms of a stereochemical interpretation for these. Each secondary structural element (alpha helix or beta sheet), is characterized by a well-defined allowed region in the dihedral angle (ϕ, ψ) space of the backbones of proteins. Not surprisingly, even the so called "irregular regions", i.e. the loops, assume either helical or sheet like dihedral values[5]. A protein consisting of n peptide linkages shows up as n points in the 2D- (ϕ, ψ) Ramachandran plot, exhibiting a clustering of points as per its secondary structural composition. Structural studies via crystallography and NMR (RCSB[6]) have verified the hypotheses of Pauling and Ramachandran time and again.

How to extend these ideas of commonalities among proteins to tertiary structures remains a pending question however. A specification of the 2n Ramachandran angles (n ϕs and n ψs) leads to a coarse-grained description of the tertiary structure of a protein. The overall conformation of a protein thus corresponds to a single point in this 2n dimensional Hyper-Ramachandran plot. Of course, one could add more dimensions to account for the degrees of freedom associated with the side chains. If free energy is added to this 2n-dimensional surface, native structure of the protein, according to Anfinsen[7], corresponds to the bottom most point or the global minimum in free energy on this (2n+1)

dimensional surface. Thus ensued several proposals on energy landscapes which overall conform to the concept of minimum free energy for the native structure[8-10]. Several other physico-chemical parameters based on size, shape, area, energy (intrinsic as well as transfer) have been investigated[11-14] but few universal ideas applicable to all proteins have emerged. This has led to extensive classifications of both secondary and tertiary structures of proteins such as various flavors of helices, turns, super-secondary structural motifs, folds etc.[15-19]. The corollary of all such classifications which chronicle the architectural splendor of proteins is to give up on universality.

We revisit proteins from a new perspective here and show evidence for very compelling clues to the existence of some universal principles, not on the folding pathways but, on the organization of protein tertiary structures. We hope that the perspective presented here paves the way for re-embracing unifying principles of protein structures rather than develop numerous subtle fragmentations.

2. Stoichiometry

It has been demonstrated recently[20-24] that amino acid space of proteins is not infinite rather proteins have well defined stoichiometries with bounds set on amino acid compositions (Table 1) as seen from the sequence data available in Swissprot/Uniprot[25]. The compositions are non-random and the deviations from the averages (called the margin of life[20]) account for the diversity of proteins.

It is logical to expect that these stoichiometries are the essence of protein size, shape, structure and function. In fact, a sequence analysis reveals the requirement of all 20 amino acids in all known protein sequences, as shown in Figure 1. This, coupled with the complete absence of naturally occurring anagramic protein sequences, strongly supports the idea that protein structure and function is governed by the stoichiometric ratios of amino acids.

3. Sizes & Shapes

3.1 Radius of gyration

Physical chemistry of polymers is rich with scaling laws, the most celebrated one being that of Flory[26-29]. Root mean square end to end distances and equivalently the radius of gyration, R_G, varies as N^{ν} where N is the number of monomeric units and ν is the scaling exponent. It is now accepted that $\nu = 0.60$ for homopolymers in good solvents, $\nu = 0.33$ for homopolymers in poor solvents, while for proteins, values of ν ranging from 0.30 to 0.40 have been observed (Figure 2).). It may be noted that $\nu = 1.00$ for a fully extended chain, $\nu = 0.50$ for random chains, the lower the ν value, the higher the compaction or

exclusion by solvent[20, 21]. The low values of ν seen for proteins clearly point to a 'precipitation' type phenomenon upon folding.

Many of these early discoveries on radii of gyration transited to advancing hydrophobicity and fractal dimensions for proteins and did not stay focused on molecular origins of such high levels of compaction.

Table 1. Margin of Life

AMINO ACID	Protein sequences confirmed by annotation and experiments (Mean ± Std. dev.; 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 ± 1.1
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

3.2 Surface area

It is believed for long that folded proteins follow a simple axiom: hydrophobic residues 'in' and hydrophilic 'out'. Structural analyses of 1000 globular proteins[30-31] suggested that the ratio of loss in surface area of nonpolar residues to that of polar residues is close to unity (~ 1.1). The axiom is thus obeyed more as an exception than as a rule. The nonpolar and polar areas of proteins have been examined hundreds of times.

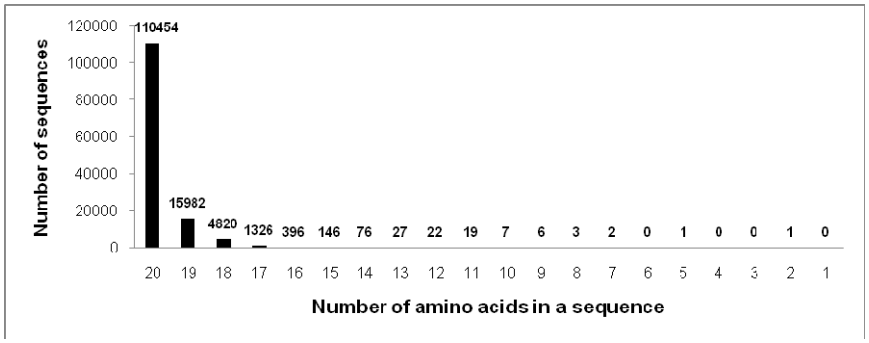

Figure 1: In the 133388 sequences (> 75 amino acids) examined, all the 20 amino acids occur in 82.87%, only 19 in 11.99%, 18 in 3.62% and 17 in 0.99% of the sequences.

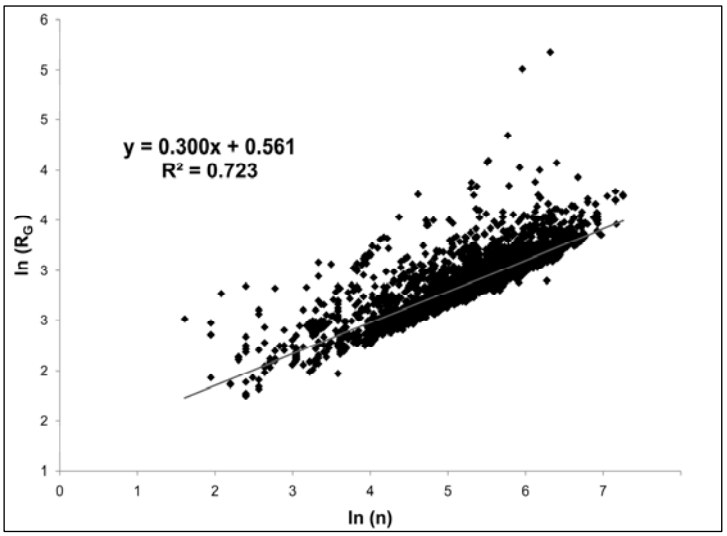

Figure 2: Radius of gyration plotted against number of residues as a log-log plot for ~ 6750 proteins. Proteins are seen to be extremely compact compared to random chains and synthetic polymers in good solvents. In the parlance of Flory, water is not a “good solvent” for proteins.

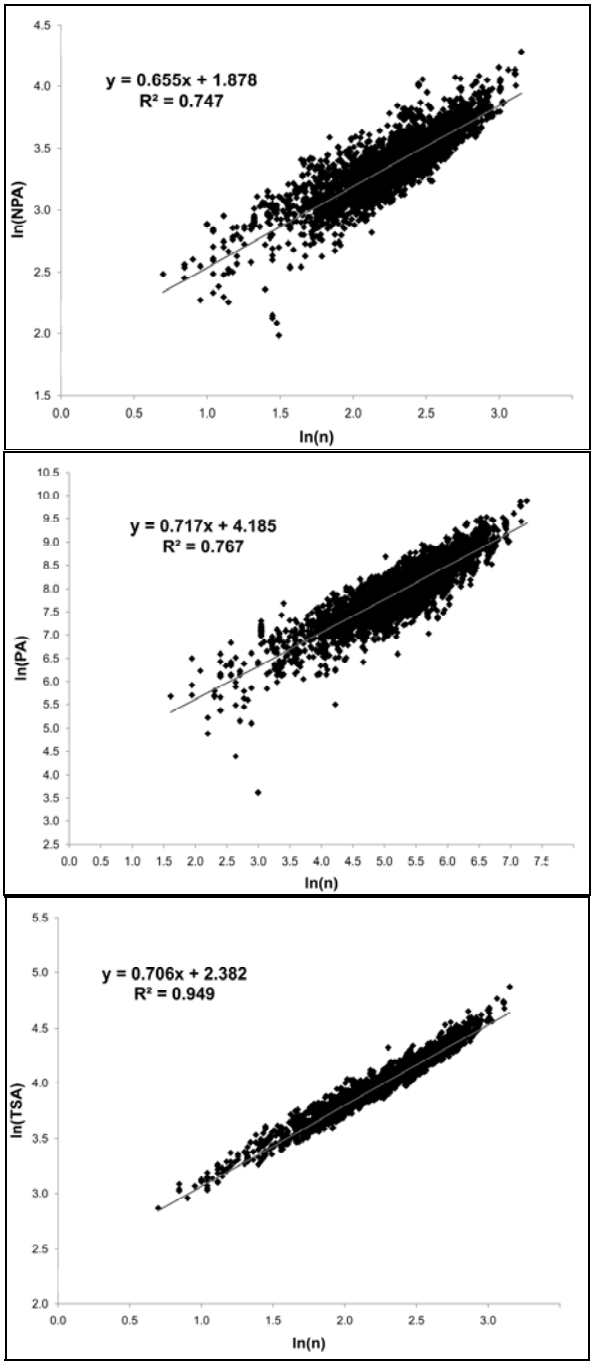

Figure 3: Solvent accessible surface areas Nonpolar (top panel), polar (middle panel), total (bottom panel) versus number of residues (n) in ~6750 proteins shown as log-log plots.

Figure 3 demonstrates a simple fact that it does not profit to divide the areas into nonpolar or polar to understand the molecular phenomena. Total area (with a very high correlation coefficient of 0.95) serves as a simple and elegant variable to estimate the solvent exposed surface area of proteins. Furthermore, it is strongly suggestive of an inherent shape/area conservation principle or an invariant area per residue metric.

4. Spatial distribution of backbone Cα atoms

A recent study[20], reported that regardless of the size, shape and composition of proteins, the spatial distribution of their backbone Cα atoms obey a simple sigmoidal equation: $Y = Y_{Max}(1-e^{-kX})^n$, Y is the cumulative number of Cα neighbors, X is the radial coordinate and, n and k are two parameters characterizing the sigmoid. It may be expected just as the (ϕ, ψ) space of proteins for secondary structures, that n and k should display allowed regions for tertiary structures of proteins.

A computer simulation was carried out on solid objects with idealized geometries (sphere, cylinder, ellipsoids, hourglass etc.)[32]. The n,k space of each was characterized (Figure 4). This was followed by an analysis of 13550 crystal structures. Considering each of the amino acids in the crystal structures individually, the geometrical analysis gave rise to neighbourhood data for a given protein backbone (i.e. Cα coordinates) in the form of a 20x20 matrix at each neighbourhood distance and the n, k values were extracted from this data. The above simulation results and the crystal structure analyses provide a direct and definitive proof confirming the earlier observations[20-22] that soluble protein indeed have a universal spatial organization regardless of size, fold, structure and function.

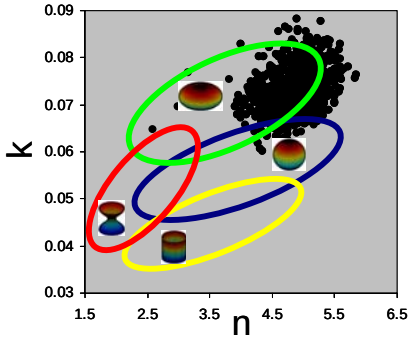

Figure 4: n, k values of 13,550 proteins superposed on allowed n, k values for some regular solid geometrical objects. Regardless of the structural classification of the proteins, and regardless of the amino acid neighbourhoods investigated, Cα neighbourhoods in folded proteins form a definite cluster in the n-k space. Proteins structures fit ellipsoidal geometries better[32-33].

5. Energetics: Secularity

Empirical energy functions with an implicit consideration of solvent have been in vogue for quite some time in the protein structure prediction field. The functional form and parameters therein are commonly anchored to a well-established force field. One such all atom based energy function originating in Amber force field reported by us earlier[34] accounts for all non-bonded interactions in proteins including van der Waals, electrostatics and solvation. It was seen that this function was able to separate the native from thousands of decoys in 67 of the 69 protein systems studied[34], thus demonstrating that it captures the essential features of a free energy function. The total energies of 6750 proteins are shown in Figure 5. The high correlation obtained (0.99) and the stability of the energy per residue clearly imply the compensatory nature of the diverse intra- and inter-molecular energy components and the resultant secularity or energetic equivalence of amino acids across a large number of tertiary structures of proteins of varying sizes and sequence compositions.

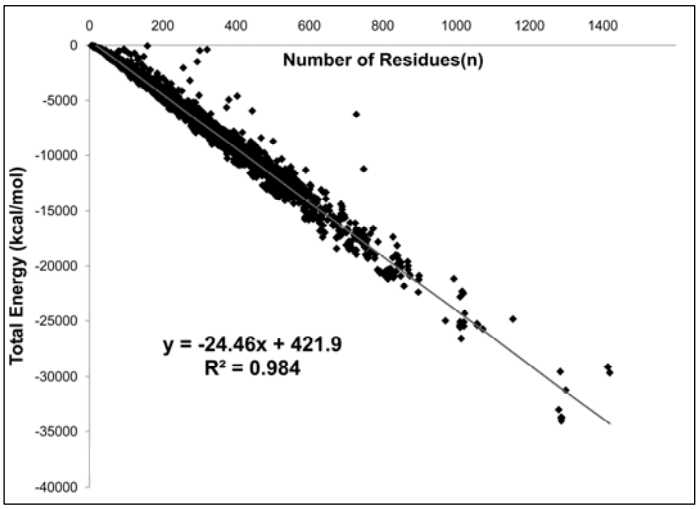

Figure 5: Total energy of 6750 proteins shown as a function of number of residues.

6. Perspectives & Conclusions

A preliminary application of the above ideas was implemented (BhageerathH www.scfbio-iitd.res.in/bhageerath/bhageerath_h.jsp) during the first ever successful completion of an Indian entry for tertiary structure prediction of proteins in CASP10 (1st May to 17th July, 2012). CASP - Critical Assessment of

Techniques for Protein Structure Prediction (http://predictioncenter.org/casp10/) refers to a worldwide biennial experiment to gauge the current abilities in the area of protein tertiary structure prediction. Initiated in the year 1994 by the University of Maryland under the leadership of John Moult, CASP has provided a platform to the prediction community all over the world to test their methodologies, ideas and "true predictions" on soon to be known/released proteins structures. Over the years, CASP has managed to evolve major benchmarks and computational breakthroughs in the field of protein structure prediction. Further, in spite of CASP being almost the equivalent of 'Olympics' of protein folding, no group in India has managed to participate in this competition for predicting tertiary structures of proteins. Till the time this article has been compiled, 38 experimental structures have been released since the beginning of CASP10. Our group (BhageerathH) has been able to predict a structure within 3Å rmsd from the native in 20 cases. These are remarkably promising results especially considering the participation of 261 prediction groups which include 125 registered servers comprising some of the best structure prediction groups for 113 blind prediction targets.

We are encouraged by the fact that universal insights into the rich diversity of protein structures, rather than development of numerous micro-subclassifications to account for the face value of the observed diversity, have led to the biggest breakthroughs in protein folding. We take our inspiration, in large part, from the universally applicable work of G. N. Ramachandran published 50 years ago, and, continue our march towards addressing one of the most important grand challenges in science today, i.e. protein folding.

Acknowledgments

BJ acknowledges the programme support from the Department of Biotechnology (DBT), Govt. of India, to the supercomputing facility for bioinformatics & Computational biology. CA is a senior research fellow of the Council of Scientific and Industrial Research (CSIR), Govt. of India. This manuscript is dedicated to the 50th anniversary of the first publication on "Ramachandran maps", a ubiquitous text-book concept now, by the great Indian scientist Prof. G. N. Ramachandran with his colleagues.

References

1. L. Pauling, R. B. Corey and H. R. Branson, *Proc. Natl. Acad. Sci. U.S.A.*, **37,** 205 (1951).
2. L. Pauling and R. B. Corey, *Proc. Natl. Acad. Sci. U.S.A*, **37**, 251 (1951).
3. L. Pauling and R. B. Corey, Proc. Natl. Acad. Sci. U.S.A, 37, 235 (1951).
4. G. N. Ramachandran, C. Ramakrishnan and V. Sasisekharan, *J. Mol. Biol.*, **7,** 95 (1963).
5. L. Thukral, S. R. Shenoy, K. Bhusan and B. Jayaram, *Journal of Biosciences*, **32(1)**, 71 (2007).
6. H. Berman, K. Henrick, H. Nakamura and J. L Markley, *Nucl. Acid Res.*, **35**, D301 (2007).
7. C. B. Anfinsen, *Science*, **181**, 223 (1973).
8. K. A. Dill, *Biochemistry*, **29(31)**, 7133 (1990).
9. H. Frauenfelder, S. Sligar and P. G. Wolynes, *Science*, **254,** 1598 (1991).
10. S. Auer, M. A. Miller, S. V. Krivov, C. M. Dobson, M. Karplus and M. Vendruscolo, *Phys. Rev. Lett.*, **26**, 178104 (2007).
11. D. W. Bolen and G. D. Rose, *Annual Review of Biochemistry*, **77** 339 (2008).
12. D. Thirumalai, E. P. O'Brien, G. Morrison and C. Hyeon, *Annu. Rev. Biophys.*, **39,** 159 (2010).
13. H. S. Chan and K. A. Dill, *Annu. Rev. Biophys. Biomol. Struct.*, **27,** 425 (1997).
14. J. Janin, *Nature*, **277,** 491 (1979).
15. A. G. Murzin, S. E. Brenner, T. Hubbard and C. Chothia, *J. Mol. Biol.*, **247**, 536 (1995).
16. A. Andreeva, D. Howorth, J. M. Chandonia, S. E. Brenner, T. J. Hubbard, C. Chothia and A. G. Murzin, *Nucl. Acid Res.*, **36**, D419 (2008).
17. Structural Classification of Proteins [SCOP : http://scop.mrc-lmb.cam.ac.uk/scop/]
18. J. S. Richardson, *Nature*, **268**, 495 (1977).
19. P. D. Sun, C. E. Foster and J. C. Boyington, in *Curr. Protoc. Prot. Sci.*, (John Wiley & Sons, 2004), p. 17.1.1.
20. A. Mittal, B. Jayaram, S. R. Shenoy and T. S. Bawa, *J. Biomol. Struc. Dyn.*, **28(2)**, 133 (2010).
21. A. Mittal and B. Jayaram, *J. Biomol. Struc. Dyn.*, **28(4)**, 669 (2011).
22. A. Mittal and B. Jayaram, *J. Biomol. Struc. Dyn.*, **28(4)**, 443 (2011).
23. R. H. Sarma, *J. Biomol. Struct. Dyn.*, **28**, 587 (2011).
24. P. S. Agutter, *J. Biomol. Struct. Dyn.*, **28**, 643 (2011).
25. Swissprot/Uniprot.[ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/complete/uniprot_sprot.fasta.gz].
26. P. J. Flory, *Statistical Mechanics of Chain Molecules* (Wiley, New York, 1969).
27. H. S. Chan and K. A. Dill, *Annu. Rev. Biophys. Biophys. Chem.*, **20**, 447 (1991).

28. L. Hong and J. Lie, *J. Polymer Science: Part B: Polymer Physics*, **47**, 207 (2009).
29. C. F. Tsai and K. J. Lee, *Int. J. Mol. Sci.*, **12**, 8449 (2011).
30. B. Jayaram, K. Bhushan, et al., *Nucl. Acids Res.*, **34**, 6195 (2006)
31. P. Narang, K. Bhushan, S. Bose and B. Jayaram, *Phys. Chem. Chem. Phys.*,**7**, 2364 (2005)
32. A. Mittal [http://precedings.nature.com/documents/6038/version/1/files/npre20116038-1.pdf].
33. B. D. Silverman, *Proc. Natl. Acad. Sci. U.S.A.*, **98**, 4996 (2001).
34. P. Narang, K. Bhushan, S. Bose and B. Jayaram, *J. Biomol. Str. Dyn.*, **23**, 385(2006).