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

Backbones of Folded Proteins Reveal Novel Invariant Amino Acid Neighborhoods

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.

Abstract

Folding of naturally occurring proteins has eluded a universal molecular level explanation till date. Rather, there is an abundance of diverse views on dominant factors governing protein folding. Through rigorous analyses of several thousand crystal structures, we observe that backbones of folded proteins display some remarkable invariant features. Folded proteins are characterized by spatially well-defined, distance dependent, and universal, neighborhoods of amino acids which defy any of the conventionally prevalent views. These findings present a compelling case for a newer view of protein folding which takes into account solvent mediated and amino acid shape and size assisted optimization of the tertiary structure of the polypeptide chain to make a functional protein.

Key words: Protein folding; Proteins; Protein structure.

Introduction

Subsequent to the elegant work of Pauling (1), Anfinsen (2, 3), and several others (4-14) spread over half a century, a consensus view on the principles of protein folding is yet to emerge due to lack of a unified concept leading to a folded protein. Can the diversity of protein architectures be captured by some invariant fundamental 'rules', representing a unifying theme? In pursuit of an answer to this question, we rigorously analyzed backbones of several thousand folded proteins from their published crystal structures (15). This was based on our hypothesis that rules of protein folding should be embedded in the (Cα) backbones. Assuming that folded protein (crystal) structures were a consequence of specific amino acid interactions, the backbones of folded proteins would be organized within the constraints of defined 'neighborhoods' for Cα atoms of each amino acid. If two amino acids 'interacted' through their side chains (via the conventionally accepted non-covalent interactions), their respective Cα atoms would be expected to occur in fixed neighborhoods relative to each other, regardless of their actual position in the protein. Further, Cα of an amino acid occurring mostly in the 'center' of folded proteins would always be expected to be surrounded by a higher number of Cα atoms of other amino acids.

Till date, several investigators have analyzed atomic contacts including the spatial distributions of Cαs in folded with at least two objectives: (a) to decipher rules of folding, and, (b) to derive statistical potentials in contrast to the physics based potentials (16) for predicting tertiary structures of proteins or to predict sequences compatible with specified tertiary structures (17-53). While the Cα analyses have not yielded any rules (54-56), the protein tertiary structure prediction attempts with

*Phone: +91-11-26591052
+91-11-26591505
Fax: 091-11-26582037
E-mail: amittal@bioschool.iitd.ac.in
bjayaram@chemistry.iitd.ac.in

statistical potentials have met with varying but limited degrees of success on a variety of protein structures (54-69).

This prompted us to critically examine spatial distributions of Cαs in a large number of crystal structures in an assumption-free and model independent manner (70).

Materials and Methods

Coordinates of all atoms in crystal structures of 3718 proteins were taken from the Protein Data Bank (see Supplementary Table S1 of reference 70). After specifically extracting the Cα coordinates for all the amino acids (*i.e.,* backbone of the folded protein) from a given PDB file, neighborhood analysis was done as described in Figure 1. The analysis was done for all the Cα atoms in the protein, with the neighborhood distances fixed at 0-9 Å, with increments of 1 Å, and 10-90 Å, with increments of 5 Å. Distances of 0-3 Å were chosen as an internal check (since zero neighbors were expected at these distances). Beginning with neighborhood analysis of 4000+ crystal structures of soluble proteins, we finally analyzed 3718 total crystal structures (see supplementary information of Ref. 71) by including only those proteins with 50 or more residues and removing those structures that did not pass the internal check. These proteins had structural resolutions of 2.5 Å or better and we analyzed only the A chains for each protein. For A chain of each protein, a 20 x 20 matrix of number of 'neighbors', within a defined neighborhood distance, resulted by considering each of the amino acids

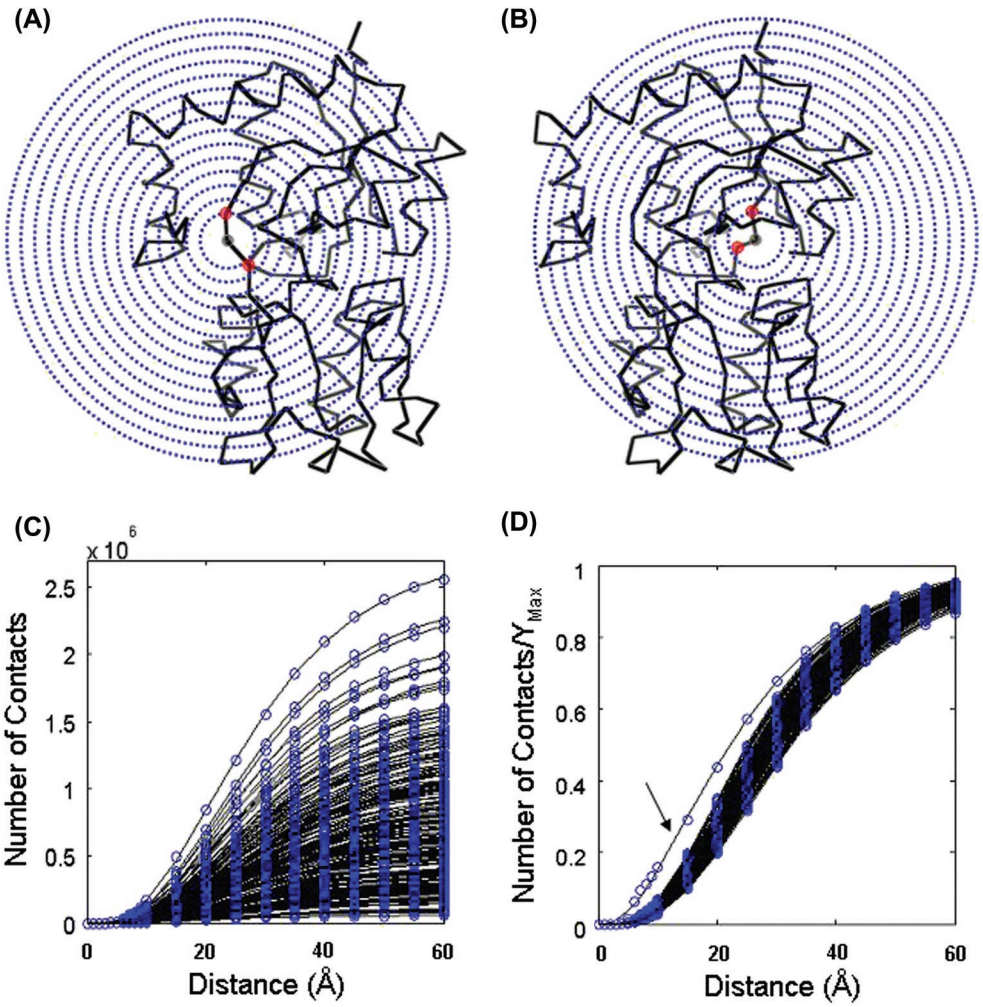

Figure 1: Backbones in 3718 crystal structures reveal a single, amino acid independent, spatial distribution of Cα neighborhoods in folded proteins – (**A**) shows neighborhood analysis of a Cα (grey) in a periplasmic protein (PDB: 1LST). Cα of any amino acid found within a defined distance (2-D blue circles) in the 3-D crystal structure was scored as a neighbor. Cα atoms of peptide-bonded partners (red) were not scored as neighbors. (**B**) shows the neighborhoods for another Cα in the same protein. (**C**) Number of contacts (neighbors) of every Cα in the backbones of 3718 folded proteins are shown as a function of neighborhood distance. 400 total neighborhoods (20x20, for each of the 20 amino acids) are shown. Regardless of the amino acid, all neighborhoods display a similar sigmoidal behavior. (**D**) All the 400 sigmoids collapsed into a very narrow band, except for one (arrow), on normalizing (dividing) each of the sigmoids with its asymptotic value. Smooth lines in (**C**) and (**D**) result from a single equation $Y = Y_{Max}(1-e^{-kX})^n$, that fits the data of all 400 neighborhoods. This shows that all amino acids are arranged according to a single spatial distribution in folded proteins. Note that in (**C**) and (**D**) there are "0" contacts up to 3 Å for all amino acids (see materials and methods, supplementary contact data tables in (71) for details).

individually. Thus, the total number of 20 x 20 matrices was equal to the total number of the defined neighborhood distances. Data of all the 20 x 20 matrices was analyzed in MATLAB (Mathworks Inc., USA). Other details are provided in (70).

Results and Discussion

By analyzing the backbones of 3718 folded proteins (see Supplementary Table S1 in (70) for a listing of all proteins), we extracted neighborhoods of each amino acid in every crystal structure. This yielded 400 'neighborhood' data-sets (see supplementary material in reference 71), each of which provided the number of times each of the 20 amino acids appear as a 'neighbor' of a given amino acid within a defined "neighborhood distance" for each protein (see Figure 1A-B).

Our first observation was that the neighborhoods of all amino acids (with respect to each other) followed the same spatial distribution (70) apparently different only at asymptotic values, as shown in Figure 1C. In fact, based on the consensus views on amino acid classification till date (*e.g.* polar vs. non-polar), one would expect the non-polar amino acids to have a different spatial distribution compared to polar ones. Non-polar residues are expected to be surrounded mostly by other non-polar neighbors only (especially at closer distances). In contrast polar residues are expected to be surrounded by fewer amino acid neighbors (but more aqueous environment). The result observed however was a collapse of all neighborhood distributions into a narrow (amino acid independent) band, when each neighborhood is normalized by its respective asymptote, as shown in Figure 1D. This indicated a 'secular' behavior of amino acids in that contributions of all amino acid 'interactions' to spatial organization were very similar to each other regardless of the conventional classifications (*e.g.* polar vs. non-polar). This discovery based on several thousand crystal structures questioned every existing view on "preferential interactions" in protein folding.

It is interesting to note that only one pair of amino acids, out of the possible 400, was clearly different from the others, shown by the arrow in Figure 1D. This was the Cysteine-Cysteine pair. The cysteines were found to be closest to each other, regardless of their location within a protein (*i.e*. middle vs. periphery) and independent of size of the protein. If a pair of amino acids were preferentially interacting with each other (*e.g.* via side chains), then the crystal structures would show several populations of sigmoids, each population containing specific pairs of amino acids (as Cys-Cys) that are spatially segregated from each other. However, other than the Cys-Cys pair, all the 399 amino acid pair-wise neighborhood distributions apparently belonged to a single family of sigmoids (Figure 1D).

Why did the neighborhood distributions of all amino acids collapse into a narrow band on simply normalizing each with its respective asymptote? To answer this question, we formulated a straightforward hypothesis. Asymptote of the spatial distribution of a single amino acid pair reflects the maximum number of times one amino acid appears as a neighbor for the other within the whole periphery of all folded proteins, independent of (a) the location of the two amino acids in each protein and (b) the size of the protein. Now, if an amino acid occurred maximum number of times in (the primary sequence of) folded protein, it would be expected to be found as a neighbor for all amino acids (including itself) maximum number of times. However, this would be the case only in the absence of any preferential interactions between amino acids in folded proteins. Alternatively, if a pair of amino acids interacted preferentially and were consistently found as neighbors, then the maximum number of times

one would observe the given pair in folded proteins would exceed the number of times they would be found as neighbors simply through their stoichiometries (frequencies of occurrence). To test which of the above two possibilities is correct, we plotted the total number of contacts made by a given amino acid in a folded protein against its percentage occurrence in folded proteins, as shown in Figure 2A. An excellent correlation between the number of neighbors of a given amino acid and its frequency of occurrence in folded proteins is clearly observed. This strongly suggested that protein folding was a direct consequence of simply the stoichiometric occurrences of amino acids in folded proteins rather than any "preferential interactions" between amino acids, contrary to all of the prevalent views.

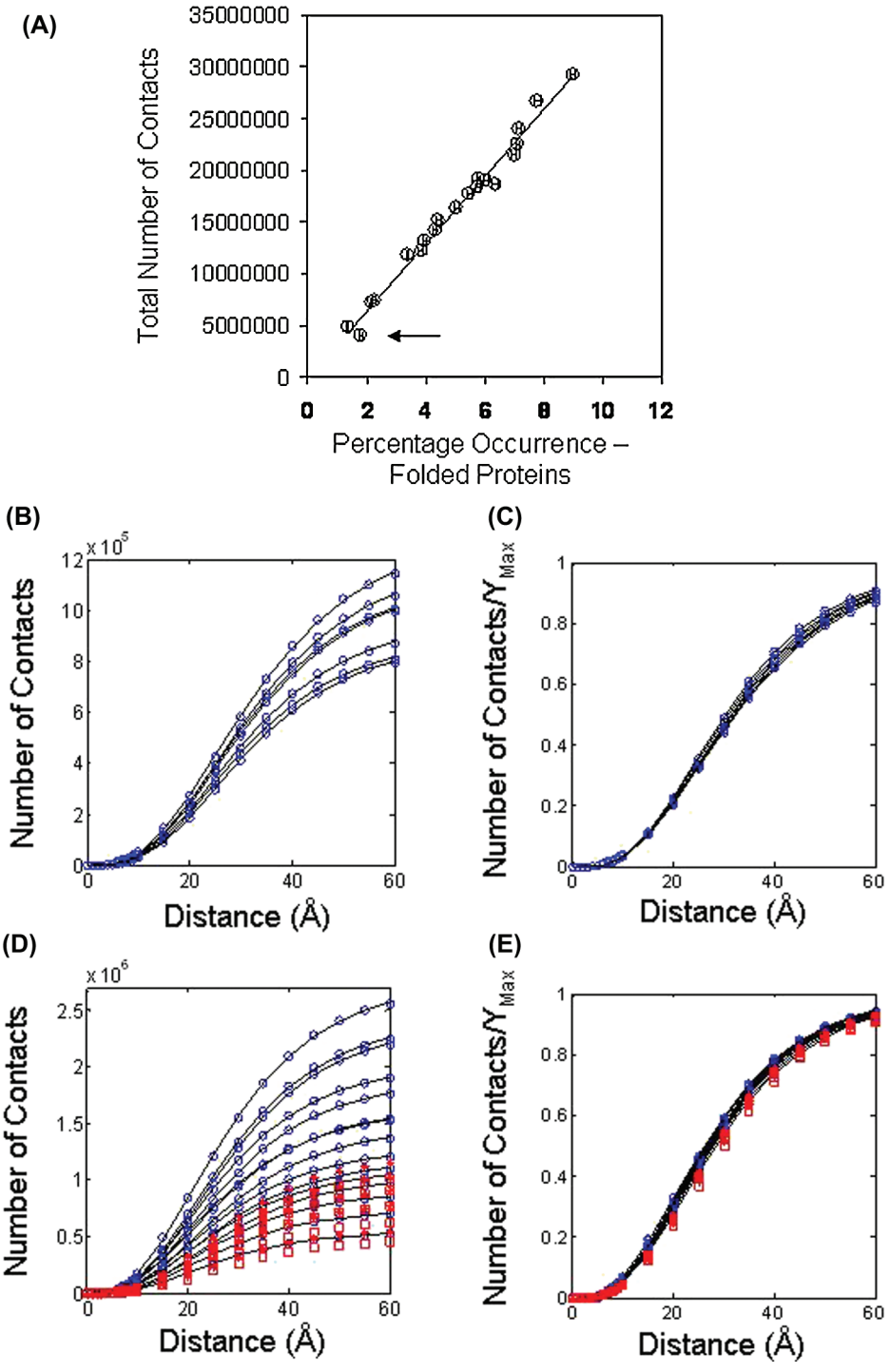

Figure 2: 3718 crystal structures show amino acid stoichiometry in the primary sequence, and not any preferential interactions between amino acids, as a rule of protein folding – (A) Total number of contacts for Cα of a given amino acid (sum of asymptotic values of the 20 neighborhood sigmoids of the amino acid), correlates excellently (r^2 = 0.99) with percentage occurrence (shown as mean ± CI, with alpha = 0.05, n = 3718) of that particular amino acid in 3718 folded proteins. (B) Number of contacts between conventionally classified "positively charged" amino acids R, K with R, K, D, E as a function of neighborhood distance. (C) Normalizing each sigmoid in (b) by its respective asymptote collapses all seven sigmoids into an overlapping single sigmoid irrespective of the neighborhood being defined as R-R, K-K, R-K, R-D, R-E, K-E or K-D. (D) Number of contacts between conventionally classified "hydrophobic" amino acids A, V, I, L, F with each other (blue ○), and, with "polar" amino acids N (red *) and Q (red □). (E) Normalizing each sigmoid in (d) by its asymptote collapses all sigmoids into an overlapping sigmoid regardless of the classification or size of amino acids.

To confirm these results rigorously, we first looked carefully at neighborhoods of the two "positively charged" amino acids as per conventional classification, namely arginine (R) and lysine (K). One would expect that if amino acid neighborhoods were indeed defined by preferential interactions, then the R-R or K-K or R-K (positive charge – positive charge) sigmoids would be substantially different from the R-D (positive charge – negative charge) or R-E or K-D or K-E. Number of the positive-positive pairs would be much lesser in the small distance (short range or medium range) neighborhoods compared to positive-negative pairs. Figure 2B shows the actual number of contacts made by R and K with each R, K, D and E. Normalization of each of the sigmoids with their respective asymptotes remarkably led to a collapse of the sigmoids into literally overlapping curves as shown by Figure 2C. Clearly, no electrostatic interactions were playing any role in the spatial distributions of these amino acids in folded proteins.

We next investigated non-polar pairs. Alanine (A), Valine (V), Isoleucine (I), Leucine (L) and Phenylalanine (F) are known conventionally as non-polar residues in the increasing order of size. Blue circles in Figure 2D show sigmoids representing the number of contacts each of the A, V, I, L, F amino acids make with each other within defined neighborhood distances. Assuming non-polar interactions were a dominant factor in protein folding, one would expect more number of pairs of smaller amino acids compared to pairs of large amino acids, and hence one would expect sigmoids of smaller amino acids to be different from sigmoids of larger amino acids. Figure 2E shows that this is clearly not the case. When each of the sigmoids is normalized with its respective asymptote, all possible pairings of A, V, I, L, F collapse into a single sigmoid. Thus, non-polar pairs were occurring identically in folded proteins, independent of their sizes. While this pointed to a clear disagreement with conventional views of "hydrophobic packing" it still did not show whether non-polar contacts were important in protein folding. To test this, we compared contacts made by A, V, I, L and F with two conventionally classified polar residues, Q (red □, Figure 2D) and N (red *, Figure 2D). Figure 2E shows clearly that neighborhood data of even the pairs of these non-polar and polar residues overlapped with the neighborhood data of non-polar pairs. Thus, "hydrophobic interactions" were not playing any special role in the spatial distributions of amino acids in folded proteins.

In the absence of evidence to support the two strongest schools of thought in protein folding, viz. electrostatics and hydrophobic collapse, we were faced with the grand challenge question again: how do proteins fold? A clear comparison of Figure 1D with Figure 2C and E provided us with the first clue. The narrow band of neighborhood data in Figure 1D was still not as narrow as in Figure 2C and E. Therefore, we hypothesized that several 'families' of neighborhoods could be present in Figure 1D, each family as narrow as in Figure 2C and E, but not necessarily in accordance with the conventional views on protein folding.

To seek out the hypothesized families, we utilized a model-independent analysis of the sigmoids (72). Figure 3A shows that by drawing a tangent to the steepest part of the sigmoid, we were able to define three parameters: "Close-Contacts", "Intermediate-Contacts" and "Long-Contacts" for any given pair of amino acids. Note that while these parameters are an arbitrary choice, they consistently define the sigmoidal neighborhoods in a completely model-independent manner. Thus, we obtained each of the three parameters for all of the 400 sigmoids in Figure 1C (Figure 1D also gives the same values). What we had observed so far had not prepared us for the next finding. Since we had observed amino acid independent spatial distributions, we expected that the frequency distributions for the three parameters extracted out of the 400 sigmoids would show a single uniform distribution. However, we were surprised to observe that frequency

distributions of each of the three parameters indicated presence of multiple populations rather than a single population, as shown in Figure 3B. Even more interestingly, these apparent multiple populations were confirmed to be statistically mutually exclusive by the K-means clustering technique, as shown in Figure 3C. We had observed in our work, for the first time, a possible distinction between amino acid pairs. More importantly, these distinctions were certainly pointing towards a completely new set of rules for protein folding.

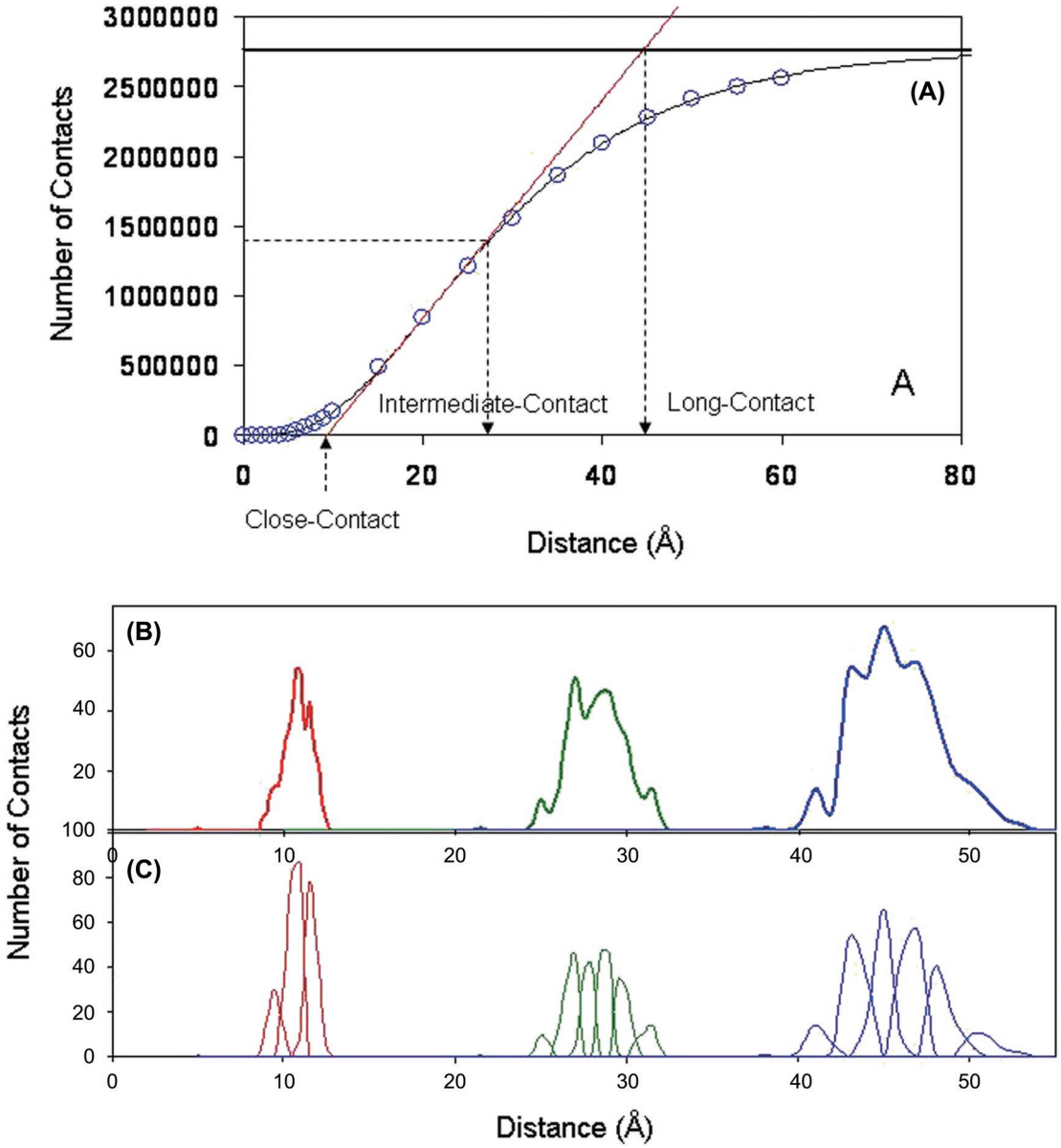

Figure 3: Cα neighborhoods are clustered within the amino acid independent spatial distributions in folded proteins – (A) A model independent characterization of the sigmoids is done, that is also independent of normalization of the sigmoids by their respective asymptotic values. Tangent is drawn to the steepest part of the sigmoid (red line). Intersection of the tangent with the X-axis defines the "Close-Contact" distance and intersection with the asymptote (black line) defines the "Long-Contact" distance. The distance at which there are half the number of maximum possible contacts is defined as "Intermediate-Contact". Note that Close-Contact, Intermediate-Contact and Long-Contact distances, while defined arbitrarily, are consistent and model independent parameterizations of the sigmoids for comparison purposes. (B) Frequency distributions for number of contacts within the distances defined in (a) are shown. The Close-Contact frequency distribution (red) shows number of Cα-Cα contacts between 5-12.5 Å (*i.e.* all Close-Contact distances are in this distance range). The Intermediate-Contact frequency distribution (green) shows number of Cα-Cα contacts between 21-33 Å (indicative of ~half of "globular" size of proteins). The Long-Contact frequency distribution (blue) shows number of Cα-Cα contacts between 38-55 Å (indicative of, but smaller than, "globular" size of proteins). Each frequency distribution results from 400 sigmoids shown in Figure 1C. Presence of convoluted/multiple frequency distributions is clear for all three parameters. (C) De-convolution of frequency distributions by K-means clustering partitions Close-Contacts into 3 (mean silhouette value, MSH = 0.701), Intermediate-Contacts into 6 (MSH = 0.724), and, Long-Contacts into 6 (MSH = 0.734), mutually exclusive clusters. Note that in both (B) and (C), the cysteine-cysteine pair (i.e. one out of the 400 pairs) appears as a small and separate "blip" at ~5 Å.

An interesting observation from Figure 3B and C is that while Close-Contacts show only three distinct clusters, the Intermediate- and Long- Contacts show six clusters each. This is intriguing, since it would be expected that short-range (occurring within distances of 5-12 Å) pairings of amino acids would show more distinct populations in the presence of specific side-chain interactions. On the other hand, the medium-range (~20-30 Å) and the long-range pairings (~40-60 Å) of amino acids would be expected to show much lesser number of pairings due to existence of very few known chemical interactions at these distances. In simpler words, if neighborhoods of amino acids were governed by side-chain interactions, then one would expect to see greater "splitting" at short range compared to the Intermediate- and Long ranges. On the other hand, if solvent were playing a major role then the Long and Intermediate range neighborhoods would show greater splitting. This is because at the long-range distances, solvent (predominantly aqueous) is the primary constituent rather than amino acids of a protein. Thus, our results clearly indicated the important, and quite under-rated, role of aqueous surroundings in protein folding. The aqueous environment was evidently playing a much bigger role in packing amino acids rather than amino acid side-chain interactions. From an energetic perspective one possible inference is that solvation and desolvation balance in a manner such that all amino acids behave similarly.

To closely inspect the multiple populations of amino acid pairs from the de-convolutions in Figure 3C, we decided to compile the neighbors for each amino acid found in each of the distinct populations. Table I, shows the listing of 'nearest' neighbors for each amino acid found in each of the three Close-Contact clusters.

The first observation is that none of the clusters have any common members. Each of the 20 amino acids appears as a neighbor for a given amino acid in a unique cluster only. The second observation is the absence of any neighbors in the first cluster for some amino acids. All neighbors for these occur in second or third cluster only. This indicates absence of a 'shell' of amino acids immediately around P, Q, N, D, E, R and K. Furthermore, even T, S, H and G belong to this group, if it were not

Table I
Identification of Cα atoms of amino acids in Close-Contact clusters (from Figure 3C), for the Cα of every amino acid.

	Cluster Number		
AA	1	2	3
A	V, I, L, C	A, Y, F, W, M, T, S, Q, H, R, G	P, N, D, E, K
V	A, V, I, L, Y F, M, C	W, P, T, S, Q, N, D, E, H, R, K, G	---
I	A, V, I, L, Y, F, W, M, C	P, T, S, Q, N, D, E, H, R, K, G	---
L	A, V, I, L, F, M, C	Y, W, P, T, S, Q, N, D, E, H, R, K, G	---
Y	V, I, C	A, L, Y, F, W, P, M, T, S, Q, H, R, G	N, D, E, K
F	V, I, L, F, M, C	A, Y, W, P, T, S, Q, N, E, H, R, K, G	D
W	I, F, C	A, V, L, Y, W, P, M, T, S, Q, N, H, R, G	D, E, K
P	---	V, I, L, Y, F, W, C	A, P, M, T, S, Q, N, D, E, H, R, K, G
M	V, I, L, F, M, C	A, Y, W, T, S, Q, N, H, R, G	P, D, E, K
C	A, V, I, L, Y, F, W, M, C, T, S, H, G	P, Q, N, D, E, R, K	---
T	C	A, V, I, L, Y, F, W, M, T, S, H, G	P, Q, N, D, E, R, K
S	C	A, V, I, L, Y, F, W, M, T, S, H, G	P, Q, N, D, E, R, K
Q	---	A, V, I, L Y, F, W, M, C	P, T, S, Q, N, D, E, H, R, K, G
N	---	V, I, L, F, W, M, C	A, Y, P, T, S, Q, N, D, E, H, R, K, G
D	---	V, I, L, C	A, Y, F, W, P, M, T, S, Q, N, D, E, H, R, K, G
E	---	V, I, L, F, C	A, Y, W, P, M, T, S, Q, N, D, E, H, R, K, G
H	C	A, V, I, L, Y, F, W, M, T, S, H, G	P, Q, N, D, E, R, K
R	---	A, V, I, L, Y, F, W, M, C	P, T, S, Q, N, D, E, H, R, K, G
K	---	V, I, L, F, C	A, Y, W, P, M, T, S, Q, N, D, E, H, R, K, G
G	C	A, V, I, L, Y, F, W, M, T, S, H, G	P, Q, N, D, E, R, K

for C. The third observation is the absence of any neighbors for some amino acids in the third cluster. This indicates that A, V, I, L, Y, F, W, M and C are closely surrounded by a shell of amino acids. The fourth observation is that regardless of the conventional view of amino acids, the Cα of cysteine is always found in close proximity to all amino acids, especially itself (see also Figure 1D, sigmoid indicated by an arrow).

The first deduction out of Table I is a possible distinction between the conventionally classified polar and non-polar residues (except C, Y and possibly W). It is extremely important to appreciate that Table I is independent of the location (*i.e.* center vs. periphery) of the amino acid in a folded protein. While Figure 2F clearly demonstrates the absence of hydrophobic interactions in protein folding, our current results definitively show a possible grouping of non-polar residues (except P). The only reconciliation for this comes from a "water-centric" view on protein folding. In a folded protein, an amino acid excluded by water must have neighbors in terms of other amino acids only.

To further explore the water-centric rule of protein folding emerging out of the data (in contrast to protein-centric or residue-centric), we compiled the clusters of Intermediate-Contacts and Long-Contacts in Tables II and III. While some observations from Table I were remarkably replicated even for medium and long range amino acid pairings, the most interesting observation was the shuffling of some amino acids as neighbors in the extracted clusters for a given amino acid. Additionally, a few amino acids were always present in either the first or the last clusters, regardless of the neighborhoods being defined as short-, medium- and long-range.

An integrated 'wheel' of neighborhoods derived from Tables I, II and III defining the probability of a given amino acid as a neighbor for any amino acid in folded proteins is presented in Figure 4.

Table II

Identification of Cα atoms of amino acids in Intermediate-Contact clusters (from Figure 3c), for the Cα of every amino acid.

	Cluster Number					
AA	1	2	3	4	5	6
A	---	V, I, C	A, L, Y, F, W, M, S, H, G	T, N, R	P, Q, D, E, K	---
V	C	A, V, I, L, Y, F, W, M, T, S, H, G	P, N, R	Q, D, E, K	---	---
I	C	A, V, I, L, Y, F, W, M, T, S, H, G	P, N, R	Q, D, E, K	---	---
L	C	V, I, L, F, M, H	A, Y, W, T, R, G	P, Q, N, D, R	E, K	---
Y	C	V, I, Y, F, W, M, H	A, L, T, S, G	P, Q, N, D, R	E, K	---
F	C	V, I, L, Y, F, W, M, H, G	A, P, T, S, N, R	Q, D, E, K	---	---
W	C	V, I, Y, F, W, M, H	A, L, T, S, G	P, Q, N, D, R	E, K	---
P	---	C	V, I, F	L, Y, W, M, H, G	A, P, T, S N, R	Q, D, E, K
M	C	V, I, L, Y, F, W, M, H	A, T, S, N, G	P, Q, D, R	E, K	---
C	V, I, L, Y, F, W, M, C, H, G	A, P, T, S, Q, N, R	D, E, K	---	---	---
T	---	V, I, C	L, Y, F, W, M, H, G	A, T, S, N, R	P, Q, D, E, K	---
S	---	V, I, C	A, L, Y, F, W, M, H, G	T, S, R	P, Q, N, D, E, K	---
Q	---	C	---	V, I, L, Y, F, W, M, H	A, T, S, N, R, G	P, Q, D, E, K
N	---	C	V, I, F, M	A, L, Y, W, T, H, G	P, S, Q, N, R	D, E, K
D	---	---	C	V, I, L, Y, F, W, M, H	A, T, S, G	N, D, E, R, K
E	---	---	C	V, I, F	A, L, Y, W, M, T, S, H, G	P, Q, N, D, E, R, K
H	C	V, I, L, Y, F, W, M, H	A, T, S, G	P, Q, N, D, R	E, K	---
R	---	C	V, I, F	A, L, Y, W, M, T, S, H, G	P, Q, N, R	D, E, K
K	---	---	C	V, I, F	A, L, Y, W, M, T, S, H, G	D, E, R, K
G	C	V, I, F	A, L, Y, W, M, T, S, H, G	P, N, R	Q, D, E, K	---

How may the results presented here advance the field of statistical potentials derived from atomic/residue contacts? The answer to this is indeed anticipated by the critical appraisal provided by Jernigan and coworkers a few years ago (49). It is conceivable that these potentials could incorporate the stochiometric dependence of the neighborhoods together with the observed clustering, with the overall

Table III
Identification of Cα atoms of amino acids in Long-Contact clusters (from Figure 3c), for the Cα of every amino acid.

	Cluster Number					
AA	1	2	3	4	5	6
A	---	C	V, I, L, , F, W, P, S, H, G	A, P, T, N, R	Q, D, E, K	---
V	C	V, I, L, Y, F, W, M, T, S, H, G	A, P, N, R	Q, D, E, K	---	---
I	C	V, I, L, Y, F, W, M, T, S, H, G	A, P, N, R	Q, D, E, K	---	---
L	C	V, I, F, M, H	A, L, Y, W, T, S, N, G	P, Q, D, R	E, K	---
Y	C	V, I, L, F, W, M, H	A, L, T, S, G	P, Q, N, D, R	E, K	---
F	C	V, I, L, Y, F, W, M, S, H, G	A, P, T, N, R	Q, D, E, K	---	---
W	C	V, I, Y, F, W, M, H	A, L, T, S, N, G	P, Q, D, R	E, K	---
P	---	C	V, I, F, H	A, L, Y, W, M, T, S, G	P, Q, N, D, R	E, K
M	C	V, I, L, Y, F, W, M, H, G	A, T, S, N, R	P, Q, D, E	K	---
C	V, I, L, Y, F, W, M, C, T, S, H, G	A, P, Q, N, D, R	E, K	---	---	---
T	C	V, I	L, Y, F, W, M, H, G	A, P, T, S, N, R	Q, D, E, K	---
S	C	V, I, F	A, L, Y, W, M, H, G	P, T, S, N, R	Q, D, E, K	---
Q	---	C	---	V, I, L, Y, F, W, M, H, G	A, P, T, S, N, R	Q, D, E, K
N	---	C	V, I, L, F, W, M, H	A, Y, T, S, G	P, Q, N, D, E, K	K
D	---	C	---	V, I, L, Y, F, W, M, H, G	P, T, S, N, R	Q, D, E, K
E	---	---	C	V, I, F, M, H	A, L, Y, W, T, S, N, G	P, Q, D, E, R, K
H	C	V, I, L, Y, F, W, M, H, G	A, P, T, S, N, R	Q, N, E, K	---	---
R	---	C	V, I, F, M, H	A, L, Y, W, T, S, G	P, Q, N, D, R	E, K
K	---	---	C	V, I, F, H	T, S, G	P, Q, N, D, E, R, K
G	C	V, I, F, M, H	A, L, Y, W, T, S, G	P, Q, N, D, R	E, K	---

sigmoidal distribution of the cumulative contacts in folded proteins providing an additional check of consistency.

Conclusions

We carried rigorous, model-independent, analyses of backbones of the crystal structures of 3718 folded proteins. This was necessitated by the need to take a fresh look at the unsolved problem of protein folding (70, 71). Our results allow evolution of the following rules for protein folding:

1. There is an underlying single spatial distribution of the backbone Cα atoms regardless of the fold and size, with Cysteine-Cysteine pairs being the sole exception.

2. The number of contacts an amino acid makes with other amino acids in a folded protein is a direct result of its frequency of occurrence (stoichiometry) in the primary sequence.

3. Cysteine, in spite of its low frequency of occurrence in primary sequences, is a “space-filler”. It is found closest to itself, and to all other amino acids.

4. Exclusion by water is the predominant factor for the protein fold. This implies a “water-centric” view on protein folding rather than a “protein-centric” or a “residue-centric” view. Within the constraints of the primary

sequence (*i.e.* composition and constitution), water excludes the polypeptide chain to assume a minimum surface-to-volume ratio.

5. In the absence of any 'driving' chemical interactions, the only other possibility that can pack the amino acids into a folded protein, with minimum surface-to-volume ratio in water, is the individual shapes of the amino acids.

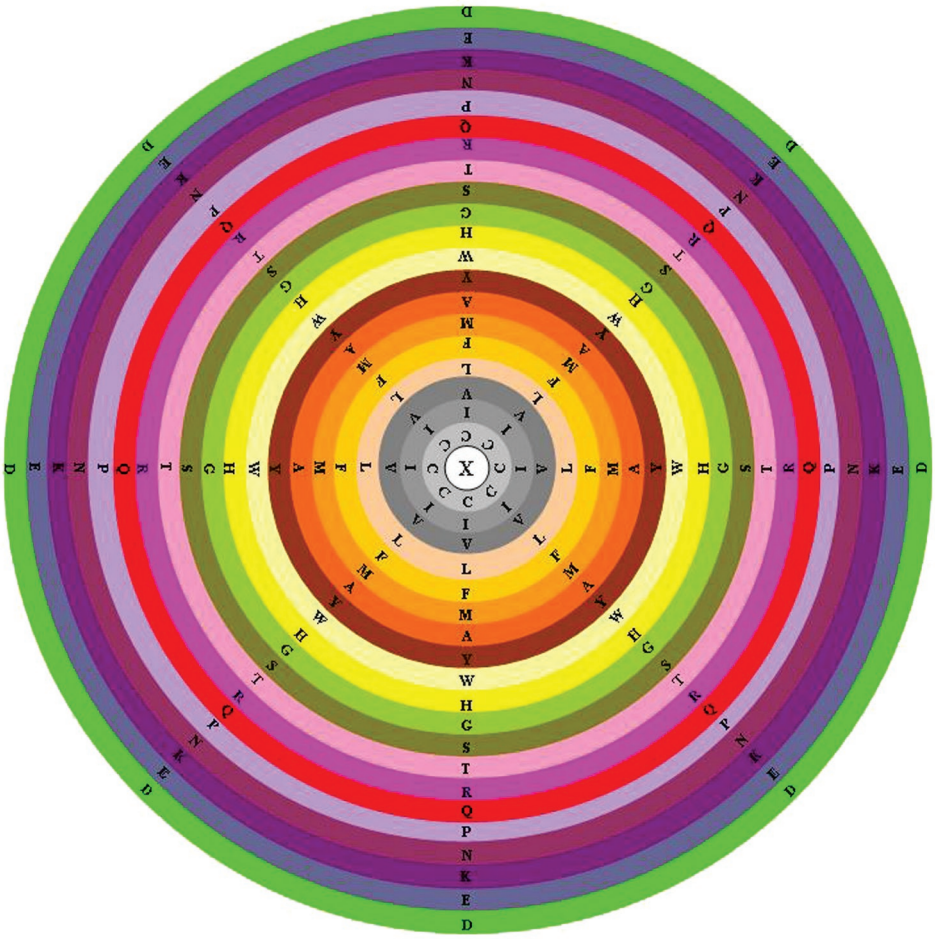

Figure 4: An integrated "wheel" of neighborhoods defining the probability of a given amino acid as a neighbor for any amino acid in folded proteins. The results in Tables I, II and III can be recovered by removing the boundaries between the various concentric circles. The white surrounding the "X" (any one of the amino acids), indicates presence of either solvent (water) or other non- Cα atoms in the protein. Note that the figure is not to scale and only the order in which the amino acids appear is important. Also note that the wheel is location independent for amino acids, *i.e.* "X" can be anywhere in the protein (*e.g.* center or periphery).

6. The conventional classifications of amino acids (*e.g.* polar, non-polar) and their interactions do not play any significant role in protein folding. Rather, they may be merely *post-facto* inferences.

It is interesting to note that the above rules mimic the simplicity of chemical reactions based on only the stoichiometries and solvation/desolvation characteristics of reacting species.

Acknowledgements

The authors acknowledge efforts of S. R. Shenoy and T. S. Bawa for their data collection efforts (70). BJ is grateful for funding support from the Department of Biotechnology, and, Department of Information Technology, Govt. of India, to the Supercomputing facility at IIT Delhi. The authors are grateful to the Editor in Chief, Prof. R.H. Sarma and the three anonymous reviewers for providing valuable constructive criticisms and suggestions allowing significant improvements to the quality of the manuscript.

1. L. Pauling, R. B. Corey, and H. R. Branson. *Proc Natl Acad Sci USA 37*, 205-210 (1951).
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. *Science 181*, 223-230 (1973).
4. W. Kauzmann. *Adv Protein Chem 14*, 1-63 (1959).
5. G. N. Ramachandran, C. Ramakrishnan, and V. Sasisekharan. *J Mol Biol 7*, 95-99 (1963).
6. C. Levinthal. *J Chim Phys 65*, 44-45 (1968).
7. C. Levinthal, in Mossbauer Spectroscopy in Biological Systems. Proceedings of a meeting held at Allerton house, Monticello, Illinois (eds P. Debrunner, J. Tsibris, & E. Munck). University of Illinois Press, Urbana, Illinois. 1969. pp. 22-24.
8. C. Chothia. Proteins. *Nature*, *357*, 543-544 (1992).
9. R. L. Baldwin. *J Biomol NMR 5*, 103-109 (1995).
10. P. G. Wolynes, J. N. Onuchic, and D. Thirumalai. *Science*, *267*, 1619-1620 (1995).
11. B. Honig and F. E. Cohen. *Fold Des 1*, R17-R20 (1996).
12. K. A. Dill and H. S. Chan, *Nat. Struct. Biol 4*, 10-19 (1997).
13. M. Karplus and J. Kuriyan. *Proc Natl Acad Sci USA 102*, 6679-85 (2005).
14. G. D. Rose, P. J. Fleming, J. R. Banavar, and A. Maritan. *Proc Natl Acad Sci USA 103*, 16623-16633 (2006).
15. H. Berman, K. Henrick, H. Nakamura, and J. L. Markley. *Nucleic Acids Res 35*, D301-3 (2007).
16. P. Narang, K. Bhushan, S. Bose, and B. Jayaram. *J Biomol Struct Dyn 23,* 385-406 (2006).
17. S. Tanaka and H. A. Scheraga. *Macromolecules 9*, 945-950 (1976).
18. B. Robson and D. J. Osguthorpe. *J Mol Biol 132*, 19-51 (1979).
19. P. K. Ponnuswamy, M. Prabhakaran, and P. Manavalan. *Biochim Biophys Acta 623*, 301-316 (1980).
20. M. J. Sippl. *J Mol Biol 156*, 359-388 (1982).
21. S. Miyazawa and R. L. Jernigan. *Macromolecules 18*, 534-552 (1985).
22. M. J. Sippl. *J Mol Biol 213*, 859-883 (1990).
23. M. Hendlich, P. Lackner, S. Weitckus, H. Floeckner, R. Froschauer, K. Gottsbacher, G. Casari, and M. J. Sippl. *J Mol Biol 216*, 167-180 (1990).
24. J. Heringa and P. Argos. *J Mol Biol 220*, 151-171 (1991).
25. V. N. Maiorov and G. M. Crippen. *J Mol Biol 227*, 876-888 (1992).
26. S. H. Bryant and C. E. Lawrence. *Proteins, 16*, 92-112 (1993).
27. D. A. Hinds and M. Levitt. *J Mol Biol 243*, 668-682 (1994).
28. K.-C. Chou and C.-T. Zhang. *J Biol Chem 269*, 22014-22020 (1994).
29. A. Godzik, A. Kolinski, and J. Skolnick. *Protein Sci 4,* 2107-2117 (1995).
30. P. D. Thomas and K. A. Dill. *Proc Natl Acad Sci USA 93*, 11628-11633 (1996).
31. L. A. Mirny and E. I. Shaknovich. *J Mol Biol 264*, 1164-1179 (1996).
32. S. Miyazawa and R. L. Jernigan. *J Mol Biol 256*, 623-644 (1996).
33. S. K. Panjikar, M. Biswas, and S. Vishveshwara. *Acat Crystallogr 53*, 627-637 (1997).
34. K. T. Simons, C. Kooperberg, E. Huang, and D. Baker. *J Mol Biol 268*, 209-225 (1997).
35. M. Vendruscolo and E. Domany. *J Chem Phys 109*, 11101-11108 (1998).
36. E. S. Huang, R. Samudrala, and J. W. Ponder. *Protein Sci 7*, 1998-2003 (1998).
37. S. Miyazawa and R. L. Jernigan. *Proteins 34*, 49-68 (1999).
38. M. R. Betancourt and D. Thirumalai. *Proteins Sci 8*, 361-369 (1999).
39. K. T. Simons, I. Ruczinski, C. Kooperberg, B. A. Fox, C. Bystroff, D. Baker. *Proteins 34*, 82-95 (1999).
40. D. Tobi, G. Shafran, N. Linial, and R. Elber. *Proteins 40*, 71-85 (2000).
41. C. Zhang and S. H. Kim. *Proc Natl Acad Sci USA 97*, 2550-2555 (2000).
42. N. Kannan, T. D. Schenieder, and S. Vishveshwara. *Acta Crystallo 56*, 1156-1165 (2000).
43. U. Bastolla, J. Farwer, E. W. Knapp, and M. Vendruscolo. *Proteins 44*, 79-96 (2001).
44. C. Micheletti, F. Seno, J. R. Banavar, and A. Maritan. *Proteins 42*, 422-431 (2001).
45. H. Lu and J. Skolnick. *Proteins 44*, 223-232 (2001).
46. O. Carugo and S. Pongor. *J Mol Biol 315*, 887-898 (2002).
47. O. Carugo. *J Mol Struct (Theochem) 676*, 161-164 (2004).
48. P. Pokarowski, A. Kolczkowski, R. L. Jernigan, N. S. Kothari, M. Pokarowska, and A. Kolinski. *Proteins 59*, 49-57 (2005).
49. R. Rajgaria, S. R. McAllister, and C. A. Floudas. *Proteins: Structure, Function and Bioinformatics 65*, 726-741 (2006).
50. Y. Wu, M. Lu, M. Chen, J. Li, and J. Ma. *Proteins Sci 16*, 1449-1463 (2007).
51. D. M. Bolser, I. Filippis, H. Stebr, J. Duarte, and M. Lappe. *BMC Struc Biol 8*, 53 (2008).
52. R. Rajgaria, Y. Wei, and C. A. Floudas. *Proteins 78*, 1825-1846 (2010).
53. A. N. Jha, S. Vishveshwara, and J. R. Banavar. *Protein Sci 19*, 603-616 (2010).
54. D. W. Bolen and G. D. Rose. *Annu Rev Biochem 77*, 339-362 (2008).
55. K. A. Dill, S. B. Ozkan, M. S. Shell, and T. R. Weikl. *Annu Rev Biophys 37*, 289-316 (2008).

56. D. Thirumalai, E. P. O'Brien, G. Morrison, and C. Hyeon. *Annu Rev Biophys 39*, 159-183 (2010).
57. K. Bhargavi, P. Kalyan Chaitanya, D. Ramasree, M. Vasavi, D. K. Murthy, and V. Uma. *J Biomol Struct Dyn 28,* 379-391 (2010).
58. A. Mahalakshmi and R. Shenbagarathai. *J Biomol Struct Dyn 28,* 363-378 (2010).
59. J. Wiesner, Z. Kriz, K. Kuca, D. Jun, and J. Koca. *J Biomol Struct Dyn* 28, 393-403 (2010).
60. Z. Cao, L. Liu, and J. Wang. *J Biomol Struct Dyn 28,* 343-353 (2010).
61. L. Zhong. *J Biomol Struct Dyn 28,* 355-361 (2010).
62. M. J. Aman, H. Karauzum, M. G. Bowden, and T. L. Nguyen. *J Biomol Struct Dyn 28,* 1-12 (2010).
63. L. K. Chang, J. H. Zhao, H. L. Liu, J. W. Wu, C. K. Chuang, K. T. Liu, J. T. Chen, W. B. Tsai, and Y. Ho. *J Biomol Struct Dyn 28,* 39-50 (2010).
64. Y. Yuan, M. H. Knaggs, L. B. Poole, J. S. Fetrow, and F. R. Salsbury, Jr. *J Biomol Struct Dyn 28,* 51-70 (2010).
65. C. Koshy, M. Parthiban, and R. Sowdhamini. *J Biomol Struct Dyn 28,* 71-83 (2010).
66. Y. Tao, Z. H. Rao, and S. Q. Liu. *J Biomol Struct Dyn 28,* 143-157 (2010).
67. Y. Yu, Y. Wang, J. He, Y. Liu, H. Li, H. Zhang, and Y. Song. *J Biomol Struct Dyn 27,* 641-649 (2010).
68. Z. Cao and J. Wang. *J Biomol Struct Dyn 27,* 651-661 (2010).
69. P. Sklenovsky and M. Otyepka. *J Biomol Struct Dyn 27,* 521-539 (2010).
70. A. Mittal, B. Jayaram, S. R. Shenoy, and T. S. Bawa, *J Biomol Struc Dyn 28*, 133-142 (2010).
71. A. Mittal and B. Jayaram. *J Biomol Struct Dyn 28,* 669-674 (2011).
72. A. Mittal, E. Leikina, L. V. Chernomordik, and J. Bentz. *Biophys J 85*, 1713-24 (2003).

Date Received: April 28, 2010

Communicated by the Editor Ramaswamy H. Sarma