Comment

The Relevance of Distance Statistics for Protein Folding

http://www.jbsdonline.com

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

Herman J. C. Berendsen

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

Corresponding Author:
Herman J. C. Berendsen
E-mail: H.J.C.Berendsen@rug.nl
\[ N_{\text{aa}} = \sum_i L_i = \sum_i L_i, \] is the total number of aa’s in the set,
\[ f_i = N_i / N_{\text{aa}} \] 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 \( X \):

\[ Y_{kl} = Y_{kl}^{\text{max}} (1 - \exp(-\kappa X))^{\alpha}, \tag{1} \]

where \( \kappa \) and \( n \) are parameters that appear to be independent of the type of aa (\( \kappa = 0.075 \) and \( n = 4.5 \)). 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 \( Y_{kl}^{\text{max}} \), i.e., \( \sum Y_{kl}^{\text{max}} \), 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:

\[ \sum Y_{kl}^{\text{max}} = 3.25 \times 10^8 f_k. \tag{2} \]

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 \( n_k \) occurrences in protein \( i \), so that the total summation over all proteins yields:

\[ \sum Y_{kl}^{\text{max}} = \sum n_k (L_i - 3). \tag{3} \]

Under the assumption that the aa distribution is homogeneous over proteins of all sizes, implying that \( n_k = f_k L_i \), eq. [3] reduces to

\[ \sum Y_{kl}^{\text{max}} = f_k \sum L_i (L_i - 3) = f_k \sum L_i^2. \]

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
