Protein folding: looping from hydrophobic nuclei.

Igor N. Berezovsky¹ *, Valery M. Kirzhner², Alla Kirzhner² and Edward N. Trifonov¹,²

¹Department of Structural Biology, The Weizmann Institute of Science, P.O.B. 26, Rehovot 76100, Israel; ²Genome Diversity Center, Institute of Evolution, University of Haifa, Haifa 31905, Israel.

*Corresponding author:

Dr. I.N.Berezovsky:

E-mail: Igor.Berezovsky@weizmann.ac.il

Telefax: 972-8-9342653

Telephone: 972-8-9343367
Abstract

Protein structure can be viewed as a compact linear array of nearly standard size closed loops of 25-30 amino acid residues irrespective of details of secondary structure. The end-to-end contacts in the loops are likely to be hydrophobic, which is a testable hypothesis. This notion could be verified by direct comparison of the loop maps with Kyte and Doolittle hydropathicity plots. This analysis reveals that most of the ends of the loops are hydrophobic, indeed. The same conclusion is reached on the basis of positional autocorrelation analysis of protein sequences of 23 fully sequenced bacterial genomes. Hydrophobic residues valine, alanine, glycine, leucine, and isoleucine appear preferentially at the 25-30 residues distance one from another. These observations open a new perspective in understanding of protein structure and folding: a consecutive looping of the polypeptide chain with the loops ending primarily at hydrophobic nuclei.

Key words: protein folding/ closed loops/ nuclei/ hydrophobicity/ domains/complete genomes/ major folds.
A fundamental property of protein structure is preferentially hydrophobic interior and hydrophilic exterior. Protein folding, therefore, should involve formation of hydrophobic nuclei. The search for the nucleation centers is one of the most promising ways to solve the problem of protein folding. Among various forces stabilizing the assumed nucleation centers the hydrophobic interactions (more generally, van-der Waals interactions) play special role because of their enthalpy nature. Hence, analysis of distribution of the hydrophobic residues in 3-D and along the protein sequence is highly relevant to the problem of the protein folding.

Universal structural element of the proteins – closed loops of 25-30 amino acid residues – have been recently discovered by exhaustive analysis of crystallized protein structures. In that work the ends of the closed loops were defined as chain-to-chain contacts with the shortest C to C distances, typically less than 10 Å (in almost 70% of cases - less than 7 Å). The typical closed loop size, 25-30 residues, is found to be the same for small and large proteins. The secondary structure was found to be, essentially, in no relation to the loop nature of the protein. There was no obvious correlation between sequence positions of elements on the one hand and the loop ends on the other. More than quarter of the loop’s ends actually reside in the middle of sections.

In the Figure 1 the protein Interleukin-1beta (Trefoil) is presented as combination of the loops according to the map calculated in the earlier paper. For the purpose of clarity at the bottom of the Figure 1 the path of the polypeptide chain is smoothed by a sliding window of seven amino acid residues (two periods of -helixes). The consecutive organisation of the loops (one right after another) exemplified in the Figure 1 is typical for all proteins.
tested. The closing ends of the loops may be considered as the folding nuclei or parts thereof. One would expect, then, that the hydrophobic interactions would make substantial contribution to the loop closure. To verify this expectation we compared sequence locations of the loop ends in nine major folds with positions of hydrophobic clusters in their sequences. These clusters were identified by Kyte & Doolittle procedure as individual peaks in the hydropathicity plots. Figure 2 presents pairwise comparisons of the hydropathicity and the loop maps. Here, the graph 2b corresponds to the Trefoil structure (1i1b) in the Figure 1. Inspection of the Figure 2 reveals that majority of the ends of the loops (85%) are found, indeed, to coincide with the peaks of hydrophobicity (loop mapping error bars: ±3 amino acid residues). Quantitative agreement of the loop ends with the Kyte & Doolittle plots is further demonstrated by the Figure 3, where total 52 sections of the hydropathicity plots around the loop ends (left and right) of the nine folds are synchronised and summed up together. Essentially, the Figure 3 represents an averaged hydrophobicity plot in the vicinity of the loop end. The position zero corresponds to the loop ends. The averaged hydropathicity plot shows clear elevation over mean value with the main maximum only two amino-acid residues off the synchronized ends. The ragged shape of the plot is due, apparently, to a small ensemble size of the structures analysed. While the nearly standard size of the loops can be explained by polymer statistics of the polypeptide chains, the hydrophobicity of the ends of the loops cannot. The amino-acid sequence and its local biases could only be result of evolutionary selection pressures. The results above, therefore, indicate that, apparently, there was such selection pressure. Then a specific prediction follows: a tendency of hydrophobic amino acids to maintain a specific distance between them - 25-30
residues. If, indeed, there is the sequence bias towards the standard distance, then the sequence correlation analysis may indicate such bias. Occurrence of locally enriched hydrophobic sites would be rather high even in random sequence, and clusters of four-five hydrophobic amino acid residues would appear along the random sequence about every 16 to 32 residues (taking for the purpose of the estimate 50% content of the hydrophobic residues). Therefore, if the expected distance correlation exists, it may appear only as a weak bias, with, e. g., avoidance of too short distances between the clusters. To ensure detection of such, presumably, weak signal we carried out positional autocorrelation analysis of protein sequences of 23 fully sequenced bacterial genomes (total over 42 000 sequences). The expected result of the calculation – observation of a preferred distance of 25-30 residues between hydrophobic amino acids. In Figure 4A the positional autocorrelation functions for all 20 amino acids are summed up, calculated for 23 genomes. In addition to strong variations at short distances – peaks at residues 3,4, and 7, responsible for helical components in the proteins\textsuperscript{10,11}, a clear excess of occurrences in the region 24-31 residues is, indeed, observed. Respective plot for randomized (shuffled) sequences of 23 proteomes is shown for comparison (small dots). In the insert B the same curves are presented in smoothed form. Valine, alanine, glycine, leucine, and isoleucine, all hydrophobic residues, as expected, are the main contributors to the observed maximum (Table 1).

Thus, not only the loop sizes show the preference to the standard 25-30 residues, but the sequence as well maintains this standard distance between, primarily, hydrophobic residues (Figure 5). Interestingly, the minimum at 15-22 residues is also common for both curves, that, apparently, means that both the loop sizes and the hydrophobic character of their ends had been under the same
selection pressure. In our previous work the only criteria used for the loop mapping were end-to-end distances and closeness to nearly standard loop size. The hydrophobic interactions were not at all taken into account. Above correlations suggest, that they should have been considered as a major criterion for the mapping of the loops. As it follows from comparisons with hydrophobicity plots not all hydrophobic sites participate in the linear arrangement of the nearly standard loops and their ends. The additional hydrophobic sites correspond, apparently, to secondary loop-to-loop interactions, that is interactions between ends of loops, substantially larger than standard 25-30 residue size. On the other hand, some end-to-end contacts may involve also polar residues. Such sites would not appear as hydrophobic ones, though they may well correspond to strong contacts if van-der-Waals interactions are accounted for. Additional analysis is required to further elucidate distributions of various amino acids (clusters, neighbour preference, density etc.).

The correlations we found out demonstrate fundamental importance of hydrophobic nuclei on one hand and of the loops as protein structure primary building blocks on the other. This immediately suggests a protein folding scenario that has not been considered in earlier works. A new knowledge that enters the scene is the loop fold nature of proteins and existence of standard size closed loops (not in the sense of traditional definition\textsuperscript{12-15}. The analysis above also reveals an important new structural meaning for the maxima of the Kyte & Doolittle hydropathicity plots. These plots as well as positional sequence correlation, can be used now as a specific guide in the studies of protein folding indicating likely positions of the closing contacts. An obvious route for the protein folding would be, thus, formation of the standard size
loops with the closed ends (nucleation centres) as an initial stage. The resulting looping structure is similar to theoretically derived flower-like design of internally correlated globules with potential wells\textsuperscript{16}. Sequential co-translational formation of these loops may be an essential initial step in the biogenesis of 3D structure of protein molecule. The final shape of the molecule is further stabilized by secondary interactions: loop-to-loop contacts, formation of helices, $\beta$-sheets etc.

**Acknowledgements**

We are grateful to A.Y.Grosberg for illuminating discussions and E.A. Yakobson for reading the manuscript and valuable comments. I.N.B. is Post-Doctoral Fellow of the Feinberg Graduate School, Weizmann Institute of Science.
References


7. Berezovsky IN. Protein structure: chapters which have been missed. Recent Research Developments in Protein Folding, Stability and Design, 2001;in press.


**Figure legends**

**Figure 1.** Loop structure of Interleukin-1beta (Trefoil, 1i1b) in traditional backbone presentation (top) and in smoothed form (bottom, see text). The loops of five different colours correspond to the map shown on Figure 2B.

**Figure 2.** Superposition of nearly standard size loops (1) on the Kyte & Doolittle hydropathicity profiles (6) for nine major protein folds (5). a, Globin (1thb), b, Trefoil (1i1b), c, Up-down (256b), d, Immunoglobulin folds (2rhe), e, Sandwich, f, Jelly roll (2stv), g, Doubly Wound (4fxn), h, UB roll (1ubq), i, TIM barrel (7tim). The Kyte & Doolittle plots are calculated by ProtScale routine of the ExPASy Proteomic Tools. Hydropathicity units correspond to original definition⁹. Smoothing window for the hydropathicity plots is taken equal five residues. The sequence coordinates for the protein folds are indicated at the x-axes, and the hydropathicity values are indicated at the y-axes.

**Figure 3.** Sum of synchronised hydropathicity plots around the ends of the mapped loops. To eliminate the sequence end effect the loop ends closer than 30 amino acid residues to the ends of the sequences were excluded from the calculation.

**Figure 4. A.** Autocorrelation function for all 20 amino acids, calculated for 23 genomes (Archaea: *A. pernix, A. fulgidus, M. thermoautotrophicum, P. horikoshii*; Eubacteria: *A. aeolicus, B. burgdorferi, C. jejuni, C. muridarum, D. radiodurans, E. coli, H. influenzae, H. pylori, M. tuberculosis, M. pneumoniae,
N. meningitidis, R. prowazekii, Synechocystis, T. maritima, T. pallidum, U. urealyticum, V. cholerae, X. fastidiosa) available in July 2000 through the Entrez Browser and provided by the National Center for Biotechnology Information. The plot demonstrates a clear excess of occurrences in the region 24-31 residues. Respective plot for randomized (shuffled) sequences of 23 genomes is shown for comparison (small dots). Insert B shows same curves in smoothed form. Smoothing is done by running window of three residues. Several sequence shufflings are made both of close range shuffling and long range shuffling with the same general result – disappearance of the minimum at about 16-23 residues and maximum at 24-31 residues. The shuffling plot shown is derived from the sequences where every residue is replaced by randomly chosen residue within neighboring ten amino acids. The statistical error bar (+/- 515) around the peak 25-30, calculated as square root of expected occurrences (3 *795 000), does not exceed the size of the dots on the curve.

**Figure 5.** A comparison of autocorrelation function derived by the analysis of all 23 bacterial genomes (A) and distribution of loop sizes (B) in 302 representative crystallized protein structures (96 eukaryotic, 151 prokaryotic, 28 fungal, 18 viral, and nine archaeabacterial). Plot B corresponds to Figure 1C from 1.
Table 1. Absolute contributions to the maximum at 25-30 residues.

<table>
<thead>
<tr>
<th>Amino acid</th>
<th>Total excess at 25-30</th>
<th>Error bar*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Val</td>
<td>4280</td>
<td>568</td>
</tr>
<tr>
<td>Ala</td>
<td>4160</td>
<td>714</td>
</tr>
<tr>
<td>Gly</td>
<td>3360</td>
<td>615</td>
</tr>
<tr>
<td>Leu</td>
<td>2495</td>
<td>814</td>
</tr>
<tr>
<td>Ile</td>
<td>2015</td>
<td>517</td>
</tr>
<tr>
<td>Arg</td>
<td>1595</td>
<td>441</td>
</tr>
<tr>
<td>Pro</td>
<td>1545</td>
<td>364</td>
</tr>
<tr>
<td>Lys</td>
<td>1445</td>
<td>456</td>
</tr>
<tr>
<td>His</td>
<td>510</td>
<td>168</td>
</tr>
<tr>
<td>Cys</td>
<td>335</td>
<td>96</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>24665</td>
<td>1993</td>
</tr>
</tbody>
</table>

*) Error bars are calculated as square roots of average occurrences.
a. Globin (1thb):

![Globin Graph](image)

b. Trefoil (1i1b):

![Trefoil Graph](image)

c. Up-down (256b):

![Up-down Graph](image)
d. Immunoglobulin folds (2rhe):

![Graph showing Immunoglobulin folds (2rhe)](image)

e. Sandwich (1aps):

![Graph showing Sandwich (1aps)](image)

f. Jelly roll (2stv):

![Graph showing Jelly roll (2stv)](image)
g. Doubly Wound (4fxn):

h. UB / roll (1ubq):

i. TIM barrell (7tim):