FOR THE RECORD

Structural argument for N-terminal initiation of protein folding

NICKOLAI ALEXANDROV¹

Protein Engineering Research Institute, 6-2-3, Furuedai, Suita, Osaka, 565, Japan

(RECEIVED July 2, 1993; ACCEPTED August 30, 1993)

It is not yet clear whether proteins fold during biosynthesis or whether translation and folding are separated in time. There are experimental data that support both of the hypotheses (Tsu, 1988; Fredman, 1992).

The most convincing arguments for the independence of folding on the direction of biosynthesis are the following:

1. Many proteins can be denatured and then refolded again. From this it was concluded that the direction of the biosynthesis is not important for protein structure formation, which is determined only by amino acid sequence.
2. It was shown for several proteins that permutations in sequence do not affect the final structure. In these experiments, edges of a polypeptide chain were artificially linked and then cut in different sites. This permutation replaced the terminal regions without changing other parts of the sequence. In spite of the different order of synthesis, the fold is retained.
3. Some proteins cannot be folded properly without the C-terminal portion.

To support the idea that proteins fold simultaneously with the biosynthesis, one should consider that:

1. Speed of the biosynthesis is several amino acids per second, and this leaves enough time for folding of the partially synthesized protein. At the same time, protein activity can be detected within a few minutes after the initiation of translation, but refolding of completely denatured large proteins requires many minutes or even hours.

2. Conformations of the isolated N-terminal fragments in proteins may resemble those in native structures.
3. Some proteins cannot be refolded and require a propeptide for proper folding. This propeptide is located at the N-terminal region of the protein sequence.
4. Secondary structure prediction methods give better results for the N-terminal parts of proteins, indicating that conformation of the N-terminal region is determined mainly by local interactions (Schulz, 1988).

All these facts, however, still cannot be regarded as strict evidence. It may well be that there are several different folding paths leading to the same final structure. This always leaves a possibility that a folding path beginning from the N-terminus will result in the native structure.

In this paper I report, from a statistical analysis of known protein structures, new findings in line with the idea that proteins fold while being synthesized. To realize what features of protein structures arise from the direction of folding, let us consider a two-dimensional lattice representation of proteins (Fig. 1). Assume that the initially synthesized N-terminal region folds first, and any change in its structure requires energy for destroying this primary fold. In such structures the N-terminus has a more compact structure than the rest of the polypeptide chain, which rolls over the already formed core. Each newly synthesized residue interacts with the previous part of the protein. As a result, an essential property of the structures, folded from N to C, is that the number of residues close to one of the previously synthesized residues (N_p) should be larger than the number of residues close to the following residues (N_f) (Fig. 1). Yet, in reality, the structure of nascent proteins is not very stable and can be destroyed with some probability if there is a possibility of a conformation with lower free energy. Furthermore, *in vivo* the process of protein folding incorporates many different cofactors (Gething & Sambrook, 1992). Therefore,

Reprint requests to: Nickolai Alexandrov, Laboratory of Mathematical Biology, FCRF-NCI, Frederick, Maryland 21702.

¹On leave from Institute for Genetics of Microorganisms, 1-st Dorozhnyi pr. 1, Moscow 113545, Russia.

Fig. 1. Two-dimensional lattice representation of proteins. Black residues are close to the previously synthesized part of the chain; white residues are close to one of the following residues. Black/white residues are close to both classes of residues. For the structure on the left, $N_p = 12$ and $N_f = 6$. For the structure on the right, $N_p = 8$ and $N_f = 4$. Note that the N-terminal region is more compact than the C-terminal one.

such properties of three-dimensional structures are hardly to be a strict rule; more likely they should be a tendency.

In my calculations I used two definitions of close residues: (1) the distance between corresponding C^α atoms is less than d_0 ($d_0 = 7.0 \text{ \AA}$); and (2) the distance between the closest atoms in the residues is less than d_0 ($d_0 = 4.5 \text{ \AA}$). When a residue is close to several residues, k_p of which are previous and k_f of which are following in the amino acid sequence, I refer this residue to the class of residues close to the previous ones if $k_p > k_f$, and to the class of residues close to the following ones if $k_p < k_f$.

To measure the local compactness of a protein fold, I divided polypeptide chains into straight fragments. A straight fragment can be defined by its property that the distance between its first C^α atom (C_1) and each subsequent C^α atom from the same fragment increases steadily: $d(C_1, C_i) > d(C_1, C_j)$, if $i > j > 1$. However, this definition results in an abundance of short fragments. In this work I define C^α atom C_{i+1} as the beginning of the next straight fragment (and, correspondingly, C_i as the end of the current straight fragment) if for three sequential C^α atoms (C_i, C_{i+1}, C_{i+2}): $d(C_1, C_i) > d(C_1, C_{i+1})$ and $d(C_1, C_i) > d(C_1, C_{i+2})$. Such a representation of a protein structure as a set of straight fragments neglects the difference in the local main chain conformation. The local compactness of the polypeptide chain can be expressed, for example, by the number of close residues from neighboring straight fragments. To assign the value of local compactness to each residue, I determine boundaries of the two consecutive straight fragments beginning from a given residue, and then count the number of the close residues between these fragments.

I used two compilations of protein structures. One consists of 170 protein chains without significant sequence homology that were submitted to the Protein Data Bank (PDB) (Bernstein et al., 1977) before October 1991. Another compilation includes 45 nonhomologous proteins from preliminary entries and the structures submitted between October 1991 and October 1992.

Table 1. Preference in a data base of protein structures for residues to be close to previously synthesized residues^a

	C ^α proximity		Residue proximity	
	$N_p > N_f$	$N_p < N_f$	$N_p > N_f$	$N_p < N_f$
Compilation 1 (170 structures)	100	48	93	52
Compilation 2 (45 structures)	26	15	29	11
Total (215 structures)	126	63	122	63

^a N_p , number of residues close to one of the previously synthesized residues; N_f , number of residues close to the following part of the molecule. The number of protein structures in which residues tend to be closer to the previously synthesized part of the molecule ($N_p > N_f$) is significantly larger than the number of those structures for which $N_p < N_f$.

I have calculated the number of residues close to one of the previously synthesized residues, N_p , and the number of residues close to the following part of the molecule, N_f , for all the proteins from my compilations (Table 1). It appeared that the number of proteins for which $N_p > N_f$ is about twice the number of proteins for which $N_f > N_p$. This proportion was obtained for both compilations, as well as for both definitions of close residues. The probability of such a ratio occurring by chance is less than 0.001 according to the χ^2 criterion.

I have calculated compactness of the N- and C-terminal regions, considering the first and the last pairs of straight fragments in the polypeptide chain correspondingly. The results reveal a definite tendency of the N-terminus to be more compact than the C-terminus (Table 2).

I have investigated whether this difference in compactness is just an edge effect or whether it decreases steadily along the chain. The dependence of the density of close residues on the position of the beginning of a pair of straight fragments (Fig. 2) indicates that the compactness decreases steadily. The plot is shown for the first 150 residues. After this point the results are less clear because of the domain structure of the proteins: the N-terminus of the second domain can have the same compactness as the beginning of the protein.

The results of this analysis fit surprisingly well with what was expected assuming that proteins fold during

Table 2. Compactness of N- and C-terminal regions^a

	Measurements	Close residues	Compactness
N-terminus	22,112	1,069	4.83 ± 0.40
C-terminus	26,201	782	2.98 ± 0.38

^a N- and C-terminal data for the number of residue pairs in neighboring straight fragments (measurements), the number of close pairs, and the compactness.

Fig. 2. Dependence of the local compactness of protein structure on the number of the residue. The compactness is defined as a normalized number of the close residues between a pair of neighboring straight fragments, beginning from a given position. Each point of the plot is a result of more than 9,000 measurements of distances in 170 independent protein structures. The linear approximation is shown with a correlation coefficient of 0.70.

their synthesis, beginning from the N-terminus. Residues tend to interact with the previously synthesized part of the molecule, and the N-terminal region on average is more compact than the C-terminal.

Does the folding from the N-terminus mean that the native structure is not in a global minimum state? Not necessarily. It is known that not all amino acid sequences

are able to adopt stable conformations. Evolutionary proteins in which the local conformation is consistent with the global energy minimum were selected. This principle of consistency (Go, 1983) may be a solution of the Levinthal paradox (Levinthal, 1968). Folding from the N-terminus enhances the principle: the structure of the global minimum should also be consistent with the direction of biosynthesis.

Acknowledgments

I am grateful to Dr. A. Kidera for fruitful discussions and for his interest in this research and to Drs. H. Nakamura and M. Swindells for reading the manuscript and making valuable remarks.

References

- Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Jr., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., & Tasumi, M. (1977). The Protein Data Bank: A computer-based archival file for macromolecular structures. *J. Mol. Biol.* 112, 535-542.
- Fredman, R.B. (1992). Protein folding in the cell. In *Protein Folding* (Greighton, T.E., Ed.), pp. 455-539. W.H. Freeman and Company, New York.
- Gething, M.-J. & Sambrook, J. (1992). Protein folding in the cell. *Nature* 355, 33-45.
- Go, N. (1983). Theoretical studies of protein folding. *Annu. Rev. Biophys. Bioeng.* 12, 183-210.
- Levinthal, C. (1968). Are there pathways for protein folding? *J. Chim. Phys.* 65, 44-45.
- Schulz, G.E. (1988). A critical evaluation of methods for prediction of protein secondary structures. *Annu. Rev. Biophys. Biophys. Chem.* 17, 1-21.
- Tsu, C.-L. (1988). Folding of the nascent peptide chain into a biologically active protein. *Biochemistry* 27, 1809-1812.