## SHORT COMMUNICATION

# Why Are Both Ends of the Polypeptide Chain on the Outside of Proteins?

**Sven Hovmöller*** **and Tuping Zhou**
*Structural Chemistry, Stockholm University, Stockholm, Sweden*

*ABSTRACT* **Protein folding starts before the whole polypeptide has been synthesized by the ribosome. No matter how long the polypeptide is or how intricate the fold, both ends of the chain always end up on the surface. From a topological point of view, this is surprising; one would have expected to find the starting (N-terminal) end inside the core of the folded protein, just as in a ball of yarn. We suggest here that the reason for this apparent paradox is that the first amino acid of the emerging polypeptide chain is gripped during protein synthesis, perhaps by the ribosome, and is not released until the whole polypeptide has been synthesized. This binding would greatly decrease the degrees of freedom for the protein-folding process and could also explain why knots are so rare in proteins. Gripping would also guarantee that the N-terminal is accessible on the protein surface as required for binding of ubiquitin, which regulates the natural degradation of proteins and avoids buildup of protein aggregates, such as those found in Huntington's, Alzheimer's, Parkinson's, and other neurodegenerative diseases. Proteins 2004;55:219–222.**

**Key words: amino acid; gripping; N-terminal; polypeptide chain; ribosome**

## INTRODUCTION

It is generally accepted that a protein's three-dimensional (3D) folded structure is determined solely by its amino acid sequence,[1] but the mechanism of protein folding is still essentially unknown. In vivo, proteins are folded as fast as they are synthesized by the ribosome (i.e., within a few seconds to a minute). Denatured proteins can regain their native structure and enzymatic activity also in vitro, but this may take hours and may also result in misfolded structures, as was shown for RNase A.[2,3] Recently, Taylor[4] investigated the topologies of protein folds and found that knots are extremely rare. Among all the proteins with known 3D structure in the Protein Data Base (PDB) he only found one deeply knotted protein structure, in acetohydroxy acid isomeroreductase (PDB code 1YVE). The knot was formed within the last (C-terminal) half of the 513 amino acid long polypeptide. Another knot was later found[5] in a protein with unknown function from the archaebacterial *M. Thermoautotropicum.* In addition, this knot is found near the C-terminal end of the polypeptide, indicating that the C-terminal but not the N-terminal of the polypeptide is a loose end that might get entangled inside a knot.

The abundant presence of hydrophobic side-chains makes it highly unlikely—not to say unthinkable—that the entire polypeptide first emerges as an unfolded extended chain inside the water-rich solution of the cell and then folds. That would cause a huge entropy term of the water. Thus, it must be assumed that the protein starts to fold immediately as it emerges from the ribosome. Then one would expect that the beginning (i.e., the N-terminal) start of the chain should often be found deeply buried within the folded protein, with the C-terminal end of the polypeptide ending up at the surface. Although it is common knowledge that both ends of the polypeptide are generally found on the protein surface, we believe that the possible significance of this fact has not been recognized. We have here investigated this topology quantitatively and speculate about what can cause the N-terminal to be on the outside of proteins and why this is important for the cell.

## MATERIALS AND METHODS

The locations of the two ends of the polypeptide chain in 1379 nonhomologous protein subunits were investigated. We extracted 2177 protein subunits[6] solved by X-ray crystallography from the Protein Data Bank (PDB) of October 27, 2001. The proteins have <30% amino acid sequence homology and were determined at 3.0 Å resolution or better. Of these, 393 subunits with <100 amino acids were eliminated. Another 282 subunits were excluded because they were hollow or doughnut-shaped and so had no or very few backbone atoms near the center of the molecule; many of these were of the TIM-barrel or β-barrel type. Finally, 103 subunits with too many (>2%) atoms for which no coordinates were given in the PDB

*Correspondence to: Sven Hovmöller, Structural Chemistry, Stockholm University, SE-10691 Stockholm, Sweden. E-mail: svenh@struc.su.se

Received 12 March 2003; Accepted 18 September 2003

Published online 27 February 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20011

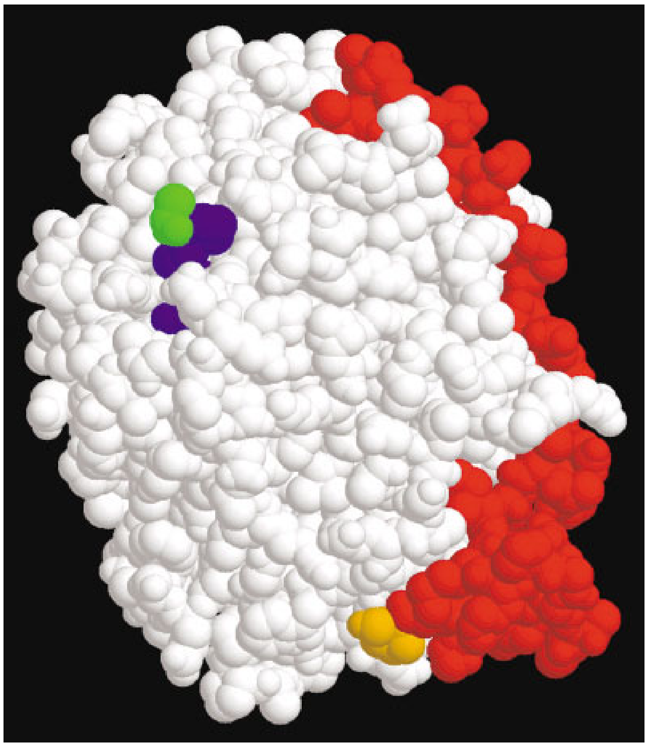

Fig. 1. An example of a protein showing the typical case with both the N- and C-terminal ends on the surface. Although the N-terminal part quickly disappears into the core of the protein, the C-terminal lies with many residues on the surface. The protein is human aryltransferase A[13], PDB code 1AUK. Of the first (N-terminal) 10 amino acids, only 3 are visible from the outside. The first residue is colored green; the next ones are blue. Of the last (C-terminal) 48 amino acids, 503 are colored orange and 460–502 are red. The very last four amino acids (504–507) were disordered in the crystal structure, so no coordinates for those were given in the PDB. The picture was made with use of RASMOL.[14]

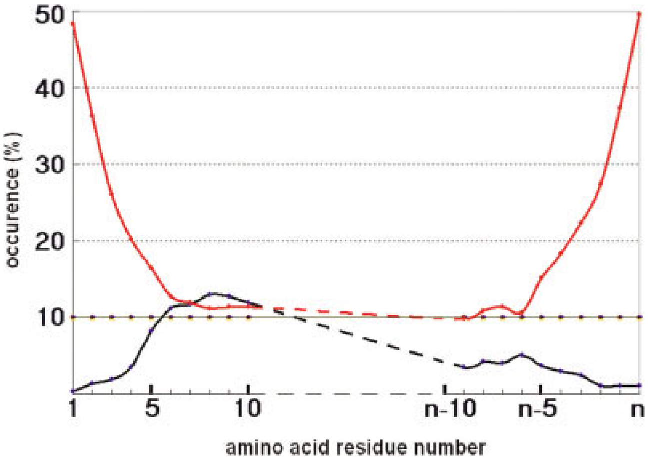

Fig. 2. The spatial occurrence of the first and last 10 amino acids in 634 globular proteins. The red curve shows how many percents of the residues that were found among those 10% of all residues farthest from the center of the protein. The black curve shows how many percents of the residues are found among the inner 10%. The dotted line indicates that all values would be at 10% if the locations were random. Notice that both ends are found on the surface of the protein but that the N terminal tends to go deep into the protein within a few residues, whereas the C-terminal does not.

were also excluded. The remaining 1379 subunits were split into two groups: 634 globular and 745 nonglobular. Those subunits were defined as globular that had >75% of all atoms within a sphere large enough to accommodate the whole protein, if it had been completely spherical. Given an average volume of 19.5 Å$^3$ per non-hydrogen atom,[7] we estimated the volume for each protein from its amino acid sequence and then calculated the radius of a corresponding sphere with that volume.

## RESULTS

A quantitative analysis of the positions of the amino acids along the polypeptide chains shows that both ends are equally often found on the protein surface or indeed as loose ends hanging out from the otherwise compact protein molecule (Fig. 1), whereas all other parts of the chains on average are equally often found inside or on the surface (Table I and Fig. 2). The reason for this is not the amino acid composition: the amino acid distribution of the first few amino acids do not differ much from the average distribution. Except for methionine, only Ala and Ser are overrepresented at the N-terminal position. From the second position and on, the amino acid compositions do not differ significantly from the average composition of proteins.

In many cases, one or more amino acids at the N-terminal and/or C-terminal ends are disordered in the crystal structures (659 and 495 cases, respectively) because they are hanging as loose ends away from the bulk protein molecule. In those cases, we used in our analysis the outermost residue for which atomic coordinates were given in the PDB as the first and last, respectively.

For each protein, we first sorted the amino acids into 10 equally large groups (similar to onion shells), depending on their distance to the center of gravity (or geometric centroid) of the whole protein subunit. The first group contained those 10% of the amino acids that were closest to the centroid and so on. Each amino acid was represented by a point at the center of gravity of all its non-hydrogen atoms. Similarly, we split up the sequence of every protein subunit into 10 groups of equally many amino acids. The matrix of 10 by 10 points, obtained after summing up the contributions from all the protein subunits, would have 10% in each element, if the different parts of the polypeptide chain had equal probabilities for being in the center, on the surface, or any other place in between. For all but the first and last few residues of the polypeptide chain, the distribution was very even. However, the first and last 10% of the chain had a significantly different distribution, with only 5% in the central bin but >20% in the outermost bin.

With an even more detailed analysis, looking at one amino acid at a time, along the polypeptide chain, the result is even more striking. For the 634 globular subunits, 49% of both the very first and the very last amino acids were among those 10% of the amino acids farthest from the centroid of the protein, but <1% were in the inner 10% (Table I and Fig. 2). For the 745 nonglobular subunits, the trends were the same, but less pronounced, with 42% of both the first and the last amino acid among the 10%

**TABLE I. Frequencies With Which the First and Last 10 Amino Acids Are Found in Different Regions of 634 Globular Proteins**

| | Amino acid # | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Region | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| | The first 10 amino acids | | | | | | | | | |
| 0–10% | **0.5** | **1.4** | **1.9** | **3.6** | **7.9** | **12.2** | **12.3** | **13.4** | **12.6** | **11.0** |
| 10–20% | 1.0 | 0.8 | 3.5 | 6.8 | 7.1 | 8.2 | 8.4 | 9.6 | 10.1 | 9.6 |
| 20–30% | 2.7 | 3.2 | 5.5 | 7.7 | 8.7 | 7.4 | 8.5 | 9.0 | 9.5 | 8.8 |
| 30–40% | 2.5 | 5.1 | 7.3 | 6.6 | 7.4 | 7.7 | 7.9 | 9.0 | 8.2 | 8.0 |
| 40–50% | 2.7 | 4.7 | 7.7 | 8.0 | 9.3 | 11.4 | 10.4 | 8.4 | 9.8 | 9.8 |
| 50–60% | 4.9 | 9.0 | 11.5 | 10.7 | 8.5 | 8.5 | 10.4 | 7.3 | 7.9 | 10.6 |
| 60–70% | 8.5 | 10.9 | 9.2 | 9.8 | 11.8 | 9.5 | 9.6 | 10.9 | 10.4 | 10.4 |
| 70–80% | 9.8 | 12.9 | 11.7 | 12.6 | 10.7 | 9.0 | 9.6 | 10.3 | 9.9 | 10.1 |
| 80–90% | 18.9 | 17.0 | 14.8 | 15.8 | 11.7 | 12.3 | 11.0 | 10.7 | 10.4 | 9.9 |
| 90–100% | **48.6** | **35.0** | **27.0** | **18.3** | **16.9** | **13.9** | **11.8** | **11.5** | **11.2** | **11.7** |
| SUM | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| | Amino acid # | | | | | | | | | |
| Region | N-9 | N-8 | N-7 | N-6 | N-5 | N-4 | N-3 | N-2 | N-1 | N |
| | The last 10 amino acids | | | | | | | | | |
| 0–10% | **6.2** | **5.1** | **3.8** | **4.4** | **3.5** | **2.2** | **1.7** | **0.8** | **0.8** | **0.8** |
| 10–20% | 7.1 | 8.2 | 9.5 | 7.9 | 5.7 | 5.5 | 3.0 | 1.1 | 1.1 | 1.1 |
| 20–30% | 8.8 | 9.8 | 8.4 | 6.5 | 8.0 | 7.3 | 4.7 | 3.9 | 2.2 | 1.0 |
| 30–40% | 11.0 | 15.8 | 9.3 | 10.3 | 9.2 | 7.9 | 7.4 | 6.3 | 3.5 | 2.1 |
| 40–50% | 13.1 | 10.1 | 8.4 | 8.8 | 11.2 | 10.3 | 7.6 | 6.8 | 5.5 | 3.0 |
| 50–60% | 10.4 | 9.8 | 9.5 | 9.8 | 10.7 | 9.5 | 9.9 | 9.0 | 8.0 | 4.6 |
| 60–70% | 12.8 | 12.3 | 12.6 | 11.2 | 10.7 | 12.6 | 10.7 | 13.3 | 10.4 | 9.3 |
| 70–80% | 11.2 | 8.0 | 13.3 | 15.5 | 12.8 | 11.5 | 16.3 | 13.3 | 13.4 | 13.1 |
| 80–90% | 10.1 | 10.3 | 13.9 | 11.8 | 12.6 | 14.7 | 15.3 | 17.4 | 18.6 | 16.1 |
| 90–100% | **9.3** | **10.7** | **11.5** | **13.9** | **15.6** | **18.6** | **23.3** | **28.2** | **36.4** | **49.1** |
| SUM | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |

The rows in bold are the points plotted in Figure 2.

farthest away from the center. This is, of course, just as expected for the last amino acids if the folding is essentially completed when the last amino acid comes out of the ribosome. But what is truly remarkable is that the very first amino acids in the chain had exactly the same distribution as the last ones!

Although the first and last four amino acids have essentially identical patterns of location within the protein, there are some clear differences for amino acids 5–20 away from the ends (Table I). The N-terminal often points rather straight into the center of the protein, so that it soon becomes invisible from the outside (see the green and blue residues in Fig. 1), whereas the C-terminal often has a long stretch lying on the surface (red atoms in Fig. 1). This pattern fits well with our gripping hypothesis: the N-terminal amino acid is fixed to the ribosome, so it cannot become the center of the folding protein, but the next 5 or 10 amino acids will often form a seed for the folding polypeptide (black line in Fig. 2). When the C-terminal end finally emerges from the ribosome, the protein is already essentially folded, so the last 10–20 amino acids will not have access to the inner core but will have to make do with ending up on the surface.

The possibility that the observed patterns were due to effects of engineering the proteins to make them feasible for crystallization was ruled out in the following way. The 634 globular proteins were split into two groups. One group contained all the 253 proteins that had exactly the same amino acids in the crystals (as specified in SEQRES in the PDB files) as those derived from the gene sequences (according to SwissProt), after deletion of any signal sequences and/or preprotein peptides. The other group contained the remaining 381 proteins that had been engineered in one way or another (often by cutting away part of the molecule but occasionally by adding extra amino acids). The resulting tables were essentially identical to the one shown in Table I, which includes all the 634 proteins.

## DISCUSSION

The simplest and clearest explanation for these observations is that something holds on to the N-terminal end of the polypeptide chain as it emerges out of the ribosome and does not let go until the whole polypeptide has come out of the ribosome. Now that several of the large structures involved in protein synthesis and translocation have been solved by X-ray crystallography, including the ribosome,[8] translocon,[9] signal recognition particle,[10] and chaperone,[11] it should soon become possible to test the gripping hypothesis by looking at crystal structures of, for example, ribosomes caught in the action of protein synthesis.

It is known that at least some small proteins can refold into their native conformations after they have been denatured. Whether this is because they are not completely unfolded during denaturation or because they can fold completely without being guided by the gripping proposed here remains to be understood.

This gripping will have a profound effect on the protein-folding mechanism, because the degrees of freedom for the folding are radically reduced if both ends of the growing polypeptide are fixed during the protein synthesis. It might also help to explain how the proteins practically always fold correctly, even for very long polypeptide chains, although it seems impossible to explore all the trillions of possible conformations in the few seconds that it takes to synthesize a protein in vivo. The folding process becomes discretized, in the sense that for every one newly synthesized amino acid residue, the polypeptide chain so far synthesized has a fixed time frame to fold. If no stable 3D arrangement is found within that time span ($\sim$0.1 s), then a new round of possible conformations will be tried because the polypeptide has been elongated by one amino acid. Such a procedure of protein folding appears more attractive (both for the cell and for the computer) than one where all possible conformations have to be tested simultaneously.

If the N-terminal amino acid is gripped, it becomes very difficult to make knots in proteins. This may be the simple explanation for the near absence of knots in proteins. It is noteworthy that both proteins that are found to contain knots[4,5] have the knots near the C-terminal end, which is certainly gripped by the ribosome until the whole polypeptide has been synthesized.

This gripping is also essential in providing a mechanism that secures that the N-terminal amino acid is always on the surface of proteins. This is necessary for a normal life of a cell, because the N-terminal amino acid must be accessible to ubiquitin. Ubiquitin binds to the N-terminal with different affinity, depending on the first amino acid, and brings the protein to the proteasome for degradation—the N-end rule.[12] Neurodegenerative diseases, such as Huntington's, Alzheimer's, Parkinson's, and so forth are caused by building up of protein aggregates. These proteins cannot be degraded as normal proteins are (i.e., by the proteasome). It has been speculated that this could be due to protein misfolding (e.g., so that the polypeptides form knots). Another possibility might be a much smaller misfolding, where the N-terminal becomes hidden inside the protein molecule, making it inaccessible to degradation, leading to a buildup of undegradable protein aggregates.

## ACKNOWLEDGMENTS

We thank Roger Ison and Bob Kretsinger for discussions and encouragement.

## REFERENCES

1. Anfinsen C. Principles that govern the folding of protein chains. Science 1973;181:223–227.
2. Ruoppolo M, Torella C, Kanda F, Panico M, Pucci P, Marino G, Morris HR. Identification of disulphide bonds in the refolding of bovine pancreatic RNase A. Fold Design 1996;1:381–390.
3. Torella C, Ruoppolo M, Marino G, Pucci P. Analysis of Rnase A refolding intermediates by electrospray/mass spectrometry. FEBS Lett 1994;352:301–306.
4. Taylor W. A deeply knotted protein structure and how it might fold. Nature 2000;406:916–919.
5. Zarembinski TI, Kim Y, Peterson K, Christendat D, Dharamsi A, Arrowsmith CH, Edwards AM, Joachimiak A. Deep trefoil knot implicated in RNA binding found in an archaebacterial protein. Proteins 2002;50:177–183.
6. Dunbrack R. Culling the PDB by resolution and sequence identity. http://www.fccc.edu/research/labs/dunbrack/culledpdb.html/cullpdb_pc30_res3.0_R1.0_d010927_chains2177.
7. Andersson KM, Hovmöller S. The average atomic volume and density of proteins. Zeitschrift Kristallogr 1998;213:369–373.
8. Nissen P, Hansen J, Ban N, Moore PB, Steitz TA. The structural basis of ribosome activity in peptide bond synthesis. Science 2000;289:920–930.
9. Sun YJ, Forouhar F, Li HM, Tu SL, Yeh YH, Kao S, Shr HL, Chou CC, Chen C, Hsiao CD. Crystal structure of pea toc34—a novel GTPase of the chloroplast protein translocon. Nat Struct Biol 2002;9:95–100.
10. Birse DE, Kapp U, Strub K, Cusack S, Åberg A. The crystal structure of the signal recognition particle Alu RNA binding heterodimer, SRP9/14. EMBO J 1997;16:3757–3766.
11. Xu Z, Horwich AL, Sigler PB. The crystal structure of the asymmetric GroEL-GroES-(ADP)7 chaperonin complex. Nature 1997;388:741–750.
12. Varshavsky A. The N-end rule: functions, mysteries, uses. Proc Natl Acad Sci USA 1996;93:12142–12149.
13. Lukatela G, Krauss N, Theis K, Selmer T, Gieselmann V, Von Figura K, Saenger W. Crystal structure of human arylsulfatase A: the aldehyde function and the metal ion at the active site suggest a novel mechanism for sulfate ester hydrolysis. Biochemistry 1998; 37:3654–3664.
14. RASMOL http://www.umass.edu/microbio/rasmol