FEBS 17336 FEBS Letters 390 (1996) 249–252

Hypothesis

# Linguistic analysis of protein folding

Michael Groß*

*Oxford Centre for Molecular Sciences, New Chemistry Laboratory, Oxford OX1 3QT, UK*

Received 21 June 1996

**Abstract Folding of nascent chains resembles the decoding of spoken language in that information is emitted as a unidirectional, one-dimensional string of elements, with higher structures and long-distance interactions emerging with time. Applying a 'pseudolinguistic' analysis of structure to a set of all 36 possible six-stranded antiparallel β-sandwich topologies reveals new order principles and reduces the complexity of this family significantly. The simple connectivity diagrams ('linguistic trees') proposed here allow predictions of the speed and cooperativity of β-sheet folding and help understanding the cotranslational folding from the N-terminus.**

*Key words:* β-Sheet; Folding problem; Nascent chain; Tendamistat; Topology

## 1. Introduction

Proteins are synthesized on the ribosome from the N-terminus. There is now a widespread consensus that the nascent chain can fold while it is still bound to the ribosome [1–3]. This implies that, unlike in the classical in vitro refolding experiment starting from the completely unfolded full-length chain, folding in vivo may well be a sequential process, with new pieces of information being added during the process. Often, essential structural elements may not occur until late in the sequence so that the folding process has to 'wait for the clue' and hence appears to be highly cooperative.

In this way, deciphering 'the second half of the genetic code' seems closely related to the decoding of spoken language. In speech, as in protein biosynthesis, a linear string of information emerges with time which slowly builds up to complex structures (phrases, clauses, sentences). However, important parts of these higher order structures may emerge late in the sentence, hence requiring listeners to keep unmatched words or phrases in the back of their minds until a match turns up and the whole finally takes on meaning. 'Solving the folding problem' would correspond to describing the generative grammar of protein sequences. By definition, a generative grammar provides the rules to produce all the 'grammatical' strings (i.e. the protein sequences which form a fold) and no ungrammatical ones.

## 2. 'Bottom up' vs. 'top down' description of protein folds

In the early days of sequence and structure analysis, proteins were indeed represented as linear sequences, e.g. as a string of beads with the amino acids written in the circles. However, since ribbon diagrams (derived from the hundreds of high resolution structures) allow convenient inspection of secondary and tertiary structure, the linear sequence diagrams have fallen out of fashion and connectivity is analysed from the tertiary structure backwards. One tends to look at the structure cartoon, label the strands according to their three-dimensional array, and then think of the loops as their connecting pieces. However, if one wants to understand nascent chain folding, one should start from the N-terminus of the linear sequence and analyse (in a 'bottom up' approach) how individual pieces of the sequence interact to form secondary and tertiary structures. In order to facilitate this type of analysis, we have developed a set of tree-like symbols inspired by the linguistic analysis of sentence structures (Fig. 1b).

There are two ways of representation for the sequence/structure relationship of a protein. For the 'sequence-based' representation needed for our kind of analysis, we start by drawing the sequence of tendamistat as a linear chain of six β-strands numbered 1...6. Then we connect the sheet forming strands with shallow U-shaped arcs if they are next neighbours in the sequence, and with deeper ones, if they are further apart. This results in two 'trees', one for each β-sheet, the equivalents of two phrases consisting of three words each in the linguistic analysis of a sentence. If we now look at the spatial sites A...F (capital letters refer to the location of a β-strand in the spatial array as in Fig. 1a and in [4]) and note down which of the β-strands 1...6 fill them, we arrive at 1-2-5 (front sheet) 4-3-6 (back sheet).

In contrast, the commonly used space-based description would trace the chain through the array A...F and note the order in which the positions are visited. In the case of tendamistat this description reads ABEDCF. Each space-based description $X_1X_2X_3X_4X_5X_6$ of a topology can unambiguously be translated into a sequence-based description $n_A$-$n_B$-$n_C$ $n_D$-$n_E$-$n_F$.

There is a homologous structure symmetry-related to tendamistat, which is not found in any protein: CBEFAD [4]. In three-dimensional representations, the CBEFAD topology is the mirror-image of the ABEDCF structure. It translates to 5-2-1 6-3-4 and can be described by the same set of linguistic trees as 1-2-5 4-3-6, namely the one shown in Fig. 1b. A second pair of mirror-images (ABEFCD and CBEDAF, Fig.

*Corresponding author. Fax: (44) (1865) 275674.
E-mail: mgross@bioch.ox.ac.uk

*Abbreviations:* 3D, three-dimensional; CI-2, chymotrypsin inhibitor 2; CspB, cold shock protein B from *Bacillus subtilis*; IgG CH, constant heavy domain of immunoglobulin G; IgG VH, variable heavy domain of immunoglobulin G; *n*, any number 1...6 (refers to position of β-strand in sequence); TBSV, tomato bushy stunt virus; TNF α, alpha domain of tumour necrosis factor; X, any letter A...F (refers to position of β-strand in space).


*PII* S0014-5793(96)00727-2

1c) also relates to the linguistic tree in Fig. 1b, but these, too, have not yet been observed in known protein structures [4].

## 3. The syntax of antiparallel β-sandwiches

Applying this analysis to the whole set of 36 possible structures with the antiparallel 2×3 β-sandwich topology [4], we obtained nine different linguistic trees (Fig. 2), each accounting for two pairs of symmetry-related structures. In none of the cases more than one of the four structures is found in nature. Of these linguistic trees, three are symmetric in themselves, while the other six can be grouped as three pairs of mirror-images.

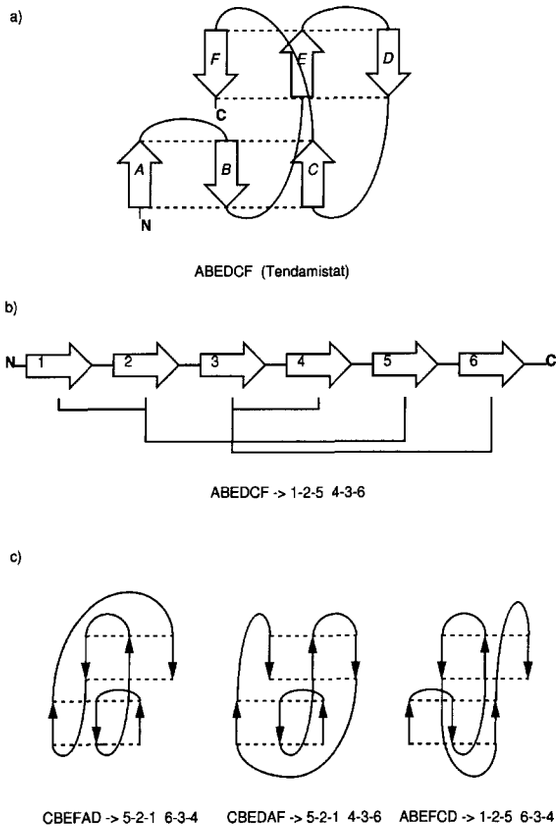


Fig. 1. Space- and sequence-based descriptions of the β-sheet topology of tendamistat. (a) Space-based model of the antiparallel beta-sandwich structure as used by Woolfson et al. The positions of β-strands in the 3D space are named A...F. By following the trace of the polypeptide chain through this fixed spatial array of 2×3 β-strands and noting down the letter one encounters, one arrives at the description ABEDCF for the topology of tendamistat. Note that the array has a horizontal two-fold symmetry axis relating A to F, B to E, and C to D. Hence, the above description is equivalent to FEBCDA. (b) Sequence-based description with 'linguistic tree'. Here the β-strands are labeled 1...6 according to their location in the amino acid sequence and then grouped according to the three-dimensional structure. For Tendamistat this results in the description 1-2-5 4-3-6, where the first triplet corresponds to the β-sheet ABC in (a) and the second triplet to the DEF sheet. However, the above-mentioned symmetry of the topology means that the triplets can be swapped to yield the equivalent description 4-3-6 1-2-5. In order to visualize the β-sheet interactions between the strands in the sequence diagram, we connect the corresponding strands with tree-like diagrams, where the branching is deeper when the interacting strands are farther apart in the sequence. (c) Further theoretical topologies which also relate to the linguistic tree in (b). 5-2-1 6-3-4 is a mirror-image of 1-2-5 4-3-6 (vertical plane through B and E); 5-2-1 4-3-6 and 1-2-5 6-3-4 are a second pair of mirror-images fitting this tree. They can be derived from 1-2-5 4-3-6 by mirroring one of the sheets at a time. In the sequence-based description, mirror-images are easily identified as they present mirrored copies of the triplets.

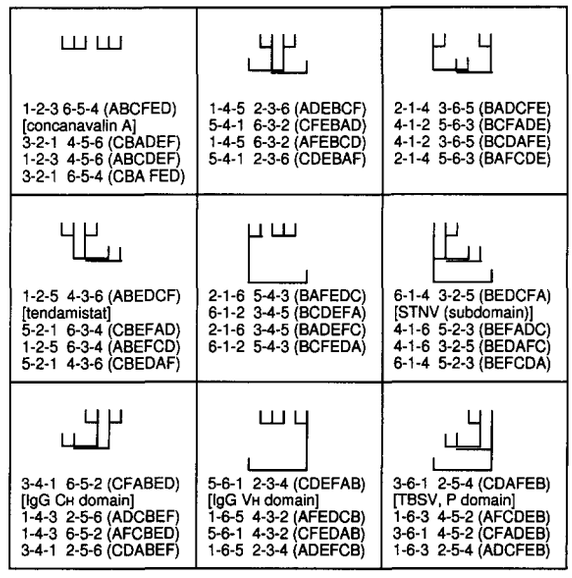


Fig. 2. Complete description of the set of 36 theoretically possible antiparallel six-stranded β-sandwich structures devised by Woolfson et al. [4] by only nine linguistic trees. Each tree accounts for four topologies, i.e. two pairs of mirror-images.

Thus, while Woolfson et al. [4] have grouped the 36 structures into 18 pairs of mirror-images, the present analysis reduces the complexity of this system further by combining pairs of these pairs to families which were not identified before. In order to analyse the way these structures might form from the N-terminus to the C-terminus during nascent chain folding, we have drawn the 'family tree' of nascent chain folding of β-structures shown in Fig. 3. Surprisingly, viable protein structures are found both at the far left end of the tree (i.e. where β-sheets are formed as early as possible) and at the far right end (where they are formed as late as possible). If one looks at the four fundamental three-strand intermediates (third row), it is striking how each of them has exactly one descendent which is viable as a domain.

## 4. Re-interpretation of folding data

More generally, this type of analysis can be applied to the interpretation of the structure/folding characteristics of proteins. In a study of N-terminal fragments of the small model protein chymotrypsin inhibitor II (CI-2), Fersht and co-workers [5] have found that the native-like structure is only observed when the fragments reach near-native length. This is hardly surprising if one looks at the topology of this protein (Fig. 4a). The major β-sheet of the structure needs the C-terminal strand 6 to build up. While 6 is not present, the N-terminal β-strand segment 1 remains unliganded. Hence, what might be interpreted as evidence for a high 'all or nothing' cooperativity in folding is in this case simply a consequence of a late node in the linguistic tree of the structure.

Refolding studies using circular permutants of the all-β SH3 domain of α-spectrin [6] have revealed that the folding ki-

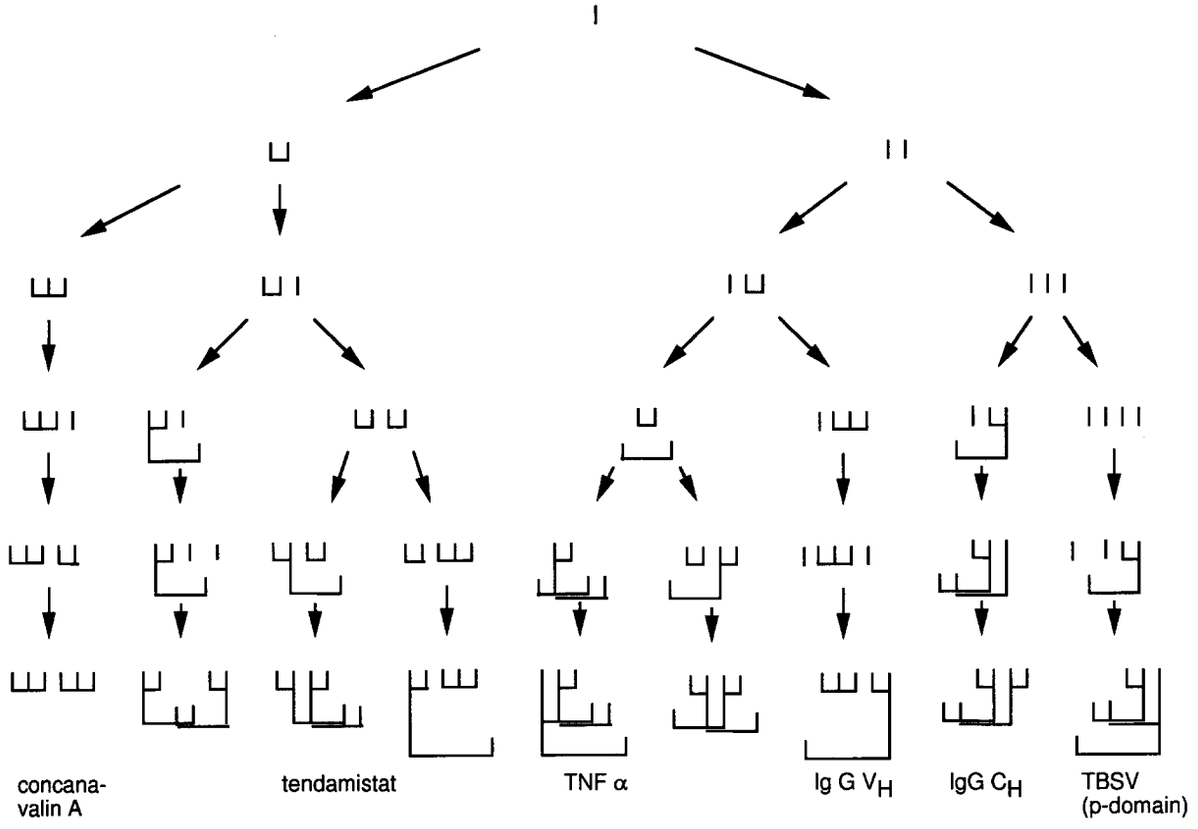


Fig. 3. 'Family tree' demonstrating how the nine linguistic trees derived from the 36 β-sandwich structures build up from their N-termini. Applying the same empirical constraints which Woolfson et al. used for the selection of the 36 structures (e.g. no parallel β-sheets allowed), this picture represents all the possibilities. Starting from the one N-terminal strand segment at the top of the picture, each row accounts for a chain elongation by one potential β-strand, which may or may not form β-sheet interactions with strands present in the row above. Where alternatives are possible, bifurcations were drawn in a way that a descendent with a new β interaction formed stands left of one with just an unmatched strand added, and that a descendent, in which the new interaction is formed with a strand occurring early in the sequence comes out left of one forming a sheet with a late occurring strand.

netics depend strongly on the order of the secondary structure elements, although the overall structure does not [6]. Inspection of the linguistic trees of the protein and its circular permutants (Fig. 4b) reveals that the fastest folding permutant is the one where all the complexity has been straightened out and β-sheets are only formed between strands which are next neighbours in the sequence (1-2-3 4-5). This suggests that a folding mechanism including local structuring in the first phase, while not essential, is an advantage.

Similarly, the all β protein Csp B, which has been found to fold surprisingly fast [7], has a neighbours-only topology (Fig. 4c). However, the $FNIII^{10}$ domain, which folds almost as rapidly as CspB [8], has a rather complex topology (Fig. 4d) Interestingly, it can be regarded as a direct descendent of the tendamistat in the family tree of topologies (Fig. 3).

## 5. Conclusion

Although linguistic metaphors have invaded the terminology of biochemistry (*translation*, *transcription*, *message*, genetic *code*, *reading* frame...), linguistic methods have not yet been used to 'decode' the second half of the genetic code, i.e. to address the folding problem. Biochemical applications of linguistic analysis so far have addressed the 'first half of the genetic code', mainly to decide which DNA sequences have a message and which do not [9], and the analysis of linguistic

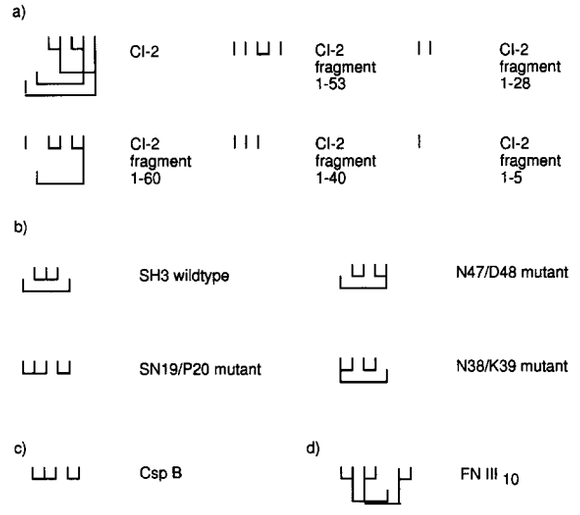


Fig. 4. Linguistic trees of some proteins whose folding properties are discussed in the text. (a) Chymotrypsin inhibitor 2, the N-terminal fragments of which were found to lack structure until they reach near-native length [5]. (b) Circular permutants of the SH3 domain. The SN19/P20 mutant was found to fold much faster than both the wild-type and other circular permutants [6]. (c) Csp B was found to fold surprisingly fast for an all-β protein [7]. (d) $FNIII^{10}$ also folds fast, in spite of a complex β topology and eight proline residues [8].

elements such as palindromes has provided useful insight into RNA folding [10].

This paper is the first attempt at understanding of the grammatical structure of proteins, which is seen as equivalent to understanding their folding properties. The fact that, of each of the sets of four possible symmetry-related topologies that can be described by a single linguistic tree, there is never more than one topology found in actual proteins, clearly indicates that this classification is meaningful and that there are as yet undiscovered 'grammatical rules' that exclude the other three.

Further interdisciplinary investigation of the folding problem, including the primary structure ('semantic') level is anticipated to be particularly fruitful. The author feels that the enormous body of data on protein structure and folding acquired to date may be sufficient to solve the folding code, if only suitable methods are developed and applied.

*Acknowledgements:* Interdisciplinary discussions with Victoria Martin (St. Anne's College, Oxford) and intradisciplinary ones with Kevin Plaxco and Chris Dobson (OCMS) are gratefully acknowledged. This is a contribution from the Oxford Centre for Molecular Sciences which is supported by the BBSRC, EPSRC and MRC. M.G. held a FEBS long-term fellowship from May 1994 to April 1996.

## References

[1] Friguet, B., Djavadi-Ohaniance, L., King, J. and Goldberg, M.E. (1994) J. Biol. Chem. 269, 15945–15949.
[2] Frydman, J., Nimmesgern, E., Ohtsuka, K. and Hartl, F. (1994) Nature 370, 111–117.
[3] Kudlicki, W. et al. (1995) J. Biol. Chem. 270, 10650–10657.
[4] Woolfson, D.N., Evans, P.A., Hutchinson, E.G. and Thornton, J.M. (1993) Protein Eng. 6, 461–470.
[5] Prat Gay, G., Ruiz-Sanz, J., Neira, J.L., Itzhaki, L.S. and Fersht, A.R. (1995) Proc. Natl. Acad. Sci. USA 92, 3683–3686.
[6] Viguera, A.R., Blanco, F.J. and Serrano, L. (1995) J. Mol. Biol. 247, 670–681.
[7] Schindler, T., Herrler, M., Marahiel, M.A. and Schmid, F.X. (1995) Nat. Struct. Biol. 2, 663–673.
[8] Plaxco, K.W., Spitzfaden, C., Campbell, I.D. and Dobson, C.M. (1996) Proc. Natl. Acad. Sci. USA 93, in press.
[9] Pesole, G., Attimonelli, M. and Saccone, C. (1994) Trends Biotechnol. 12, 401–408.
[10] Searle, D.B. (1992) Am. Sci. 80, 579–591.