

# The role of 'pro-sequences' in protein folding

M. Yahiya Khan

Department of Biochemistry, School of Life Sciences, North-Eastern Hill University, Shillong 793 014, India

**Protein folding is a complex process that is believed to proceed through a sequence of events that are under kinetic control. Any deviation from the predetermined folding pathways might lead to the formation of 'non-native' conformations. Recent findings have indicated that 'pro-sequences' play an important role in controlling the rate of folding in many proteins. Whereas a short pro-sequence has been found to accelerate the rate of refolding of alpha-lytic protease by over seven orders of magnitude, it has been shown to have decelerating effect on the folding rate of maltose- and ribose-binding proteins. Although, the increase in the rate of folding by the pro-sequences has been attributed to the stabilization of an 'on-pathway' folding intermediate in alpha-lytic protease, it is not clear if the same mechanism operates for the pro-sequence-mediated folding in other proteins.**

MANY proteins are synthesized as precursors containing pro-sequences which are usually cleaved off before the expression of their biological activities<sup>1,2</sup>. Pro-sequences have generally been believed to mediate the localization of the newly synthesized polypeptide chains to their site of action and/or the regulation of their biological activities<sup>1,2</sup>. Accordingly, these sequences have been vividly named as targeting sequences, signal sequences, leader sequences and cellular 'zipcodes'. Although a signal sequence is the most universal requirement for protein secretion, the exact mechanism of their action is not yet known. Similarly, attempts to find primary structural homology within a given class of their sequences have failed<sup>2,3</sup>. The notion that these sequences are exclusively involved in protein-targeting thus appears to be untenable and it would be more appropriate to call these 'extra bits' of the proteins' primary structure as pro-sequences.

The pro-sequences are generally short (15 to 30 amino acids) in length and usually occur on the amino terminus of proteins. A central stretch of about 10–12 hydrophobic amino-acid residues is frequently flanked by hydrophilic residues at the two termini of the pro-segments which makes the latter conformationally very sensitive towards the environment<sup>2</sup>. The possibility of the involvement of the pro-sequences in the protein folding process can therefore not be ruled out. However, the discussion on this possibility cannot be taken up unless we update our understanding about the protein folding problem itself. The present article, therefore, first addresses the question of the protein-folding mechanism and then argues, with specific

examples, that the pro-sequences go a long way in influencing the rate and the nature of the formation of the folded structures of proteins.

## Thermodynamic versus kinetic control

Two conflicting views of the protein folding mechanism have developed over the years<sup>4</sup>. According to what is popularly known as 'the thermodynamic control of protein folding', only thermodynamic stability determines the conformation of a folded protein<sup>4,5</sup>. It follows that the native structure of a protein is the one which has the minimum Gibbs free energy. This view is supported by the experimental observations that a protein unfolded in various ways refolds to the same structure under native conditions, i.e. there is no path restriction. Moreover, predictive approaches for computation of the native protein structure from its amino-acid sequence have strong thermodynamic connotation<sup>6</sup>.

The alternative view about the folding phenomenon is that the protein structure is determined by kinetic limitations<sup>4,7,8</sup>. In other words, folding to the native state occurs through certain defined routes without necessarily reaching the thermodynamically most stable structure. The argument that led to the development of the 'kinetic hypothesis' stemmed from the late Cyrus Levinthal's calculations showing that there were astronomically large number of conformations of an unfolded protein undergoing folding to the native state by random search<sup>7</sup>. Thus, if all the amino acids in a polypeptide chain having 150 residues are allowed to exist in only two conformations, the polypeptide can acquire  $10^{45}$  structures through random search. The time required to search all these possible conformations would be  $10^{26}$  years<sup>9</sup>. Since *in vivo* synthesis and folding of a protein usually takes place in minutes or less, folding cannot occur by a random search. In other words, protein folding is path dependent and therefore it is governed by kinetic considerations.

## Evidences supporting involvement of pro-sequences

From the arguments presented above, it is conceivable that protein folding is a non-random process (also see refs. 10 and 11). If that is so, one should be able to block folding by interfering with the folding pathway(s)<sup>12</sup>. Unfortunately early attempts to identify stable inter-

mediates on the folding pathways and their correlation with the kinetic-control theory of the assembly process generally failed. As a result, the classical two-state theory of protein folding (based on thermodynamic considerations) not only got well established but also severely influenced the general trend in development of a model for protein folding. However, the advent of modern techniques has now enabled scientists to study more and more proteins which were otherwise not available in adequate quantities earlier. The examples discussed below clearly suggest that protein folding is predominantly a kinetic process in which pro-sequences play an important role.

## *The case of alpha-lytic protease*

The alpha-lytic protease is a bacterial protease synthesized as inactive pro-protease having a 166-residue long pro-segment which is proteolytically removed to generate the active enzyme<sup>13</sup>. Like many other proteins, alpha-lytic pro-protease can be unfolded and refolded *in vitro* whereas the mature form does not refold in the absence of the pro-region<sup>14</sup>. Several probable explanations have been given for this type of unfolding-refolding behaviour of alpha-lytic and other proteolytic enzymes in the past. Since these proteases are activated after the removal of the pro-segments, it has been suggested that the lack of the complete folding information (c.f. Anfinsen's conclusion that primary structure governs the formation of tertiary structure<sup>5</sup>) in terms of the amino-acid sequence is the primary cause of irreversible unfolding of activated proteases<sup>15</sup>. Similarly, owing to the inhibitory action of most of these pro-segments, it has been suggested that pro-sequences 'protect' the folded conformation of proteases which are otherwise susceptible to autolysis<sup>14</sup>. Hence folding to active forms cannot proceed unless the pro-segment inhibitor is present.

The above explanations about the folding of alpha-lytic protease were however, not completely true. By omitting the pro-region of alpha-lytic protease in an *in vitro* refolding experiment, Baker *et al.*<sup>16</sup> were able to trap an inactive but folding competent intermediate state, I, that had native like secondary structure<sup>16</sup>. They found that the I state was stable for weeks at physiological pH and folded rapidly to the active native state on addition of the pro-region as separate polypeptide chain. It was further demonstrated that the pro-sequence of the enzyme helped in its folding by lowering the height of a limiting energy barrier (kinetic barrier) by more than 27 kcal per mole and thereby enhancing the rate of conversion of the intermediate state, I, to the active native state by over  $1.7 \times 10^7$  times<sup>16</sup>. These findings thus clearly suggest the presence of kinetic control in protein folding wherein the pro-sequence influences the rate of acquisition of the

protein's tertiary structure by interfering with a kinetic intermediate in the folding pathway.

## *Maltose- and ribose-binding proteins*

Maltose binding protein is a monomeric protein which exists in both 'pro' precursor as well as mature active forms<sup>3</sup>. A systematic study of the unfolding-refolding behaviour of the two forms of the protein by Park *et al.*<sup>3</sup> revealed that the presence of the pro-sequence decreased the rate of refolding but had little effect on the rate of unfolding of the precursor polypeptide chain. Similar kinetic interference from the pro-sequence was also demonstrated towards the refolding of the precursor polypeptide of the ribose-binding protein<sup>3</sup>.

## *Lysosomal proteinases*

Many lysosomal proteases like cathepsins B, H and L have been found to be synthesized in a precursor form from which the mature enzymes are produced by removal of the pro-sequences through several proteolytic cleavages. *In vitro* studies on unfolding-refolding behaviour of the mature forms of the lysosomal proteinases cathepsins B<sup>15</sup> and D<sup>17</sup> confirmed the presence of kinetic intermediates. Although similar data on the precursor 'pro-forms' of these enzymes are not available, the non-reversibility of the unfolding transitions of the mature enzymes do suggest a role for the pro-sequences in the folding of these enzymes<sup>15,17</sup>. Indeed the correlation observed between increased size of the pro-sequence deletion and decreased recovery of enzymatic activity of cathepsin L following its renaturation suggests that the pro-peptide has a crucial role in the folding of this protein<sup>18</sup>.

## **Mechanism of action**

Although the detailed mechanism by which proteins fold is not yet fully understood, we are reasonably sure about many conclusions that have been reached recently. For instance, protein folding, which is no more considered to be a single step process, is believed to proceed through a sequence of events that are under kinetic control. Similarly it is now amply clear that the folding pathways can be manipulated by changing the unfolding-refolding conditions. In other words, the 'product' of the folding phenomenon would depend on the environment at large and it would continue to be modulated by interactions with specific ligands present in the 'assembly plant'<sup>19</sup>.

Since folding occurs through a definite pathway, only deviation from the latter is expected to lead to the formation of 'non-native' conformations whose thermodynamic stability might be equal to, lesser or even more

than, the 'native' state of a given protein. The molecular chaperonins, a ubiquitous class of proteins, are believed to bind to and stabilize the partially folded intermediates in the folding pathways and thus promote folding by decreasing the rate of off-pathway folding reactions<sup>19,20</sup>. Alternatively, the rate of folding reactions of a protein can be affected by altering the rate of a limiting on-pathway reaction. In principle, pro-sequences may influence the overall rate of protein folding by either of these two mechanisms. The latest findings of Baker *et al.*<sup>6</sup> indicate that the pro-sequence in alpha-lytic protease increases the rate of folding by over seven orders of magnitude by directly stabilizing the rate-limiting on-pathway intermediate state<sup>16</sup>. Whether this is a universal mechanism for the action of pro-sequences is not clear at the moment and this should be an interesting area of research in coming years.

1. Newrath, H., *Trends Biochem. Sci.*, 1989, **14**, 268–271.
2. Gierasch, L. M., *Protein Folding: Deciphering the Second Half of the Genetic Code* (eds. Gierasch, L. M. and King, J.), American Association for the Advancement Science, Washington, 1990, pp. 211–219.
3. Park, S., Liu, G., Topping, T. S., Cover, W. H. and Randall, L. L., *Protein Folding: Deciphering the Second Half of the Genetic Code* (eds. Gierasch, L. M. and King, J.), American Association for the Advancement of Science, Washington, 1990, pp. 220–224.

4. Baldwin, R. L. and Eisenberg, D., *Protein Engineering* (eds. Oxender, D. L. and Fox, C. F.), Alan R. Liss, New York, 1987, pp. 127–148.
5. Anfinsen, C. B., *Science*, 1973, **181**, 223–230.
6. Chou, P. V. and Fasman, G. D., *Annu. Rev. Biochem.*, 1978, **47**, 251–276.
7. Levinthal, C., *J. Chim. Phys.*, 1968, **65**, 44–45.
8. Wetlaufer, D. B., *Proc. Natl. Acad. Sci. USA*, 1973, **70**, 697–701.
9. Anfinsen, C. B. and Scheraga, H. A., *Adv. Protein Chem.*, 1975, **29**, 205–300.
10. Kim, P. S. and Baldwin, R. L., *Annu. Rev. Biochem.*, 1990, **59**, 631–660.
11. Braakman, I., Helenius, J. and Helenius, A., *Nature*, 1992, **356**, 260–262.
12. Creighton, T. E., *Nature*, 1992, **356**, 194–195.
13. Silen, J. L., McGrath, C. N., Smith, K. R. and Agard, D. A., *Gene*, 1988, **69**, 237–244.
14. Baker, D., Silen, J. L. and Agard, D. A., *Proteins*, 1992, **12**, 339–399.
15. Agarwal, S. K. and Khan, M. Y., *Biochem. J.*, 1988, **256**, 609–613.
16. Baker, D., Sohl, J. L. and Agard, D. A., *Nature*, 1992, **356**, 263–265.
17. Lah, T., Drobnic-Kosorok, M., Turk, V. and Pam, R. H., *Biochem. J.*, 1984, **218**, 601–608.
18. Smith, S. M. and Gottesman, M. M., *J. Biol. Chem.*, 1989, **264**, 20487–20495.
19. Khan, M. Y., *Indian J. Biochem. Biophys.*, 1992, **29**, 311–314.
20. Khan, M. Y., *Curr. Sci.*, 1990, **59**, 723–724.

ACKNOWLEDGEMENT. I thank the Department of Science and Technology, New Delhi for financial assistance.

Received 7 July 1992; accepted 5 March 1993

## RESEARCH ARTICLES

# Genetic differentiation at *Adh* locus in Indian natural populations of *Drosophila melanogaster*

Ravi Parkash and Shamina

Department of Biosciences, Maharshi Dayanand University, Rohtak 124 001, India

Ten Indian geographical populations of *D. melanogaster* collected along 20°N latitudinal range revealed significant clinal variation (3% for 1° latitude) at the *Adh* locus, and *Adh*<sup>F</sup> allelic frequency correlated significantly with increase in latitude. The data on interpopulational genotypic and allelic frequency heterogeneity as well as  $F_{ST}$  value of 0.25 revealed significant genic divergence at the *Adh* locus. Patterns of ethanol utilization and ethanol tolerance in larval and adult individuals revealed significant genetic divergence.  $LC_{50}$  values revealed clinal variation in the range of 9.25 per cent to 15.8 per cent, i.e. southern populations of *D. melanogaster* depicted significant lower ethanol tolerance compared with north

Indian populations. The parallel occurrence of latitudinal genetic divergence at the *Adh* locus and for ethanol tolerance in colonizing populations could be maintained by balancing natural selection varying spatially along the north-south axis of the Indian subcontinent. The present data further support and validate the hypothesis that occurrence of parallel or complementary latitudinal clines across different continental populations provide strong evidence of natural selection maintaining such clinal variation.

COLONIZING species populations offer excellent material for micro-evolutionary studies<sup>1,2</sup>. Studies on biogeo-