http://www.pjbs.org

PJBS

ISSN 1028-8880

Pakistan Journal of Biological Sciences

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Pakistan Journal of Biological Sciences 8 (3): 487-492, 2005
ISSN 1028-8880

Protein Folding: From Hypothesis Driven to Data Mining

Basir Ahmad and Rizwan Hasan Khan
Interdisciplinary Biotechnology Unit, Aligarh Muslim University, Aligarh UP, India, 2002

Abstract: There has been considerable progress made over the past year in both theoretical and experimental studies of protein folding. In this review different aspects of protein folding phenomena is outlined and recent approaches which provide recent understanding of protein folding problem are described. The different types of protein substructures and intermediate states that have possible implication in protein folding have been discussed. The various theoretical models and experimental developments that are being evolved to obtain a more detailed understanding of this complex process are tabulated and some are described. *In vivo* protein folding and human diseases has been covered in a separate section to highlight the possible role of protein folding in various inherited and non-inherited diseases. The *in silico* approach of protein folding has been covered in great detail as because biomolecular simulation to protein folding is expected to shed significant light into this process. Blue gene is expected to enable a tremendous increase in the scale of simulation studies that can be carried out as compared with existing supercomputers. The data mining and analyzing from *in vivo*, *in vitro* and *in silico* approaches of folding may provide certain common trend that will lead to a unified mechanism of protein folding.

Key words: Protein flooding problem, autonomous folding units, energy landscape, molten globule, Blue gene

INTRODUCTION

The genetic information encoded in DNA is expressed only after the unique three-dimensional structure is formed from newly synthesized polypeptide chain. Thus the central dogma DNA↔RNA→protein is all about protein biogenesis. The efforts to answer two central questions: what are the determinants of protein structure and how does a linear polypeptide chain acquires its native structure has been addressed as protein folding problem. From the classical experiment of Anfinsen[1] to recent advancement of biophysical techniques[2] seem to have shifted protein-folding field from hypothesis driven to data mining. The emergence of computational biology has added another dimension to this field. The accumulation of larger data sets from all kind of proteins is needed for unraveling the second-generation genetic code (folding code), which causes the linear sequence of amino acid to acquire unique structure within a reasonable time.

Understanding protein-folding problem is not just a scientific curiosity, it has also application in medicine, biotechnology and synthetic chemistry (Table 1). For example this information could be helpful in designing the protein that are resistance to denaturation or to proteolysis, or in engineering proteins with new functions. It may also provide important clues to the cause of certain disease that appears to involve misfolding or toxic folding of protein. This diverse applications in different field of biology necessitate an understanding of the folding process both in *in vitro*, *in vivo* and also *in silico*. Here we have tried to accumulate the major works, which has been carried to solve this problem.

Protein folding *in vitro*: Major advances in elucidating the folding mechanism of proteins have been made in test tubes. The unfolding/refolding of proteins occurs in submillisecond[3,4], is reversible[1] and various partially folded states accumulate during equilibrium unfolding of a number of proteins[5-9]. These findings gave the concept of native intermediate and denatured states of the protein.

Native state: The three dimensional biologically active conformation that a protein posses in its native environment. It is the state where the interatomic and intermolecular interactions are maximum, the conformation is quite rigid with a very low free energy[10].

Denatured state: The denatured state is an ensemble of energetically comparable conformational state possessing least possible interactions with flexible structure more like a random coil[11].

Intermediate states: A large number of proteins have been shown to accumulate intermediate conformations

Corresponding Author: Rizwan Hasan Khan, Interdisciplinary Biotechnology Unit, Aligarh Muslim University, Aligarh UP, India, 2002 Tel: 0571 2720388 Fax: 0571 2721776 E-mail: rizwanhkhan@hotmail.com

Table 1: Practical application of protein folding

Field	Applications
Biomedicine	(a) Drug design
	(b) Protein misfolding/aggregation
	© Protein toxic folding
Biotechnology	(a) Designing denaturant resistant protein
	(b) Designing protease resistant protein
	© Enzyme design for paper, pulp etc industry
	(d) Recovery of inclusion body
Nanotechnology	Self assembly of nanomachine
Synthetic chemistry	Polypeptide synthesis with desired function

between native and denatured state in their folding pathway. There has been a considerable interest in characterizing these states for gaining insight into the possible determinants of protein folds and mechanism of protein folding[12]. The intermediates have been reported to arise as a consequence of energy barrier[13], incorrect proline-isomerization[14] and non native contacts that need to be broken before folding proceed[15].

Out of all intermediates identified and characterized so far following are supposed to be more important and significant.

Autonomous Folding Units (AFU's): These are the protein substructures that can fold to native like conformations independent and rest of the polypeptide chain[16]. Identification of AFU of a protein can indicate the minimal amount of information that is both necessary and sufficient for determining a structure and studies of AFU's can help to elucidate how a protein's structural information is stored in its amino acid sequence. The AFU's can be detected by one and or combination of following methods (a) limited proteolysis (b) multistate denaturation (c) prediction based on the structure of the intact proteins and (d) hydrogen exchange Nuclear Magnetic Resonance measurements (NMR).

Molten Globule (MG): The term Molten Globule (MG) was coined by Ohgashi and Wada[17] and has been considered as a general intermediate in all folding pathways[18]. The common structural features of the MG state are (a) significant amount of secondary structure (b) absence of the specific tertiary structure (c) radius of gyration 10 to 30% larger than that of native state (d) loosely packed hydrophobic core. Thus it is a compact globule with molten side chain structure. The MG state is characterized usually by conventional techniques such as (1) far-UV CD that detects the secondary structure (2) near-UV CD and fluorescence measurement for tertiary structure and side chain mobility (3) gel filtration chromatography to determine radius of gyration and (4) ANS binding experiments that detect formation of loose hydrophobic core. However new experimental techniques such as hydrogen exchange Nuclear Magnetic Resonance (NMR), solution X-ray scattering and protein engineering have also been used in characterizing MG state[19,20]. The study of *in vitro* MG state is important because this state has also been reported to exist in *in vivo* folding[21,22].

Burst Phase intermediate (BP): The significant amount of secondary structure formation has been reported in a number of proteins during the first few millisecond of refolding under native condition[23,24]. This structure has been named as Burst Phase (BP). The BP belongs to molten globule type of state but differ in following aspects from the molten globule; (a) the secondary structures formed in burst phase are non native[25] (b) does not contains strongly protected amide protons indicating that the BP's are highly dynamic states without stable hydrogen bonds[26]. The BP intermediates has been detected in Rnase, barnase and lysozyme etc by kinetic CD[27].

Experimental advancement in protein folding: Major advances in elucidating the mechanism of protein folding have been chiefly derived from the development of methods that can monitor fast transitions between structurally dynamic ensembles. The latest experimental techniques, which are used to measure folding, have been described in Table 2. Among these NMR is the most powerful technique for analyzing the conformational properties of protein ensemble[28]. Using high resolution multidimensional NMR, detailed information about the structure, energetic, dynamics and the mechanistic relevance of the denatured states, collapsed species, partially folded intermediates and high-energy transition state is coming up[20,29,30].

An another recent technique, protein engineering methods for Φ value analysis have been used for obtaining the structure of transition states and intermediate states at the level of individual residues, at almost atomic resolution[19]. In this method, an amino acid is replaced by another amino acid from the desired protein and stability of the native protein is determined from equilibrium denaturation study and the stability of intermediate and transition state by using kinetic methods. The relative stability of two proteins (native and mutated) is compared. If the mutation destabilizes folded structure by $\Delta\Delta G_{N-D}$ unit of energy relative to unfolded state. Then, if the free energy of a transition state, measured relative to the unfolded state, change by $\Delta\Delta G_{T-D}$, Φ_F is defined by $\Phi_F = \Delta\Delta G_{N-D}/\triangle\triangle G_{T-D}$ A Φ_F value of 0 means that the structure is unfolded at the site of mutation by the same extent as it is in the denatured state and a Φ_F value of 1.0 means that the structure is folded at the site of mutation as much as in the native state.

Table 2: Experimental techniques used to monitor protein folding

Techniques	Information about folding process
Protein engineering	Role of individual amino acid in stability
Laser scattering and gel filtration chromatography	Radius of gyration
DSC	Thermodynamics of folding process
Fluorescence	
Intrinsic	Environment and orientation of Trp
Polarization and Anisotropy	Dynamic of fluorophore
FRET	Distance between two point in a protein
Quenching	Accessibility and environment of fluorophore
REES	Difference between environment of fluorophore
Stopped flow	Time scale of fluorescence changes
Ligand binding (Substrate and Inhibitor)	Formation of native structure at active site
Cicular Dichroism (CD)	
Far-UV	Secondary structure
Near-UV	Tertiary structure
Stopped flow	Time for secondary and tertiary structure formation
Hydrogen exchange	
Native state exchange	Detection of metastable state
NMR	Rate of formation of backbone hydrogen bonds and protection from exchange of amino acid side chain
Real time NMR	Environment of protein side chain
Dynamic NMR	Detect equilibrium species
Laser temperature Jump	Trigger folding/unfolding at nano seconds

Models of protein folding: The vast input of data has led to the origin of many protein folding models each representing a definite pattern of a limited number of intermediate that a folding protein should pass through. Various theories of protein folding with their postulates are presented in the Table 3. Here we are discussing only the new view of protein folding

Table 3: Protein folding models

Protein folding models	Postulates of the model
Framework model	Local secondary structure elements form independent of tertiary structure
Diffusion collision model	These secondary structures diffuse, collided, adhered and coalesced to give tertiary structure.
Nucleation model	Neighboring residues of sequence form native secondary structure that act as a nucleus from which native structure form step wise
Hydrophobic-collapse model	Protein collapse rapidly around hydrophobic side chains then rearrange from restricted conformational space occupied by intermediate state
Jigsaw-puzzle model	Each molecule of the protein folds by a different path

Energy landscape theory: The energy landscape presents one of the most successful models for protein folding. Its takes into account both the thermodynamic and kinetic aspects of protein folding. This model begins with the view that folding kinetic is best considered as a progressive organization of an ensemble of partially folded structure through which a protein passes on its way to folded structure and of the free energies and entropies associated with these ensembles. According to this model energy and entropy parameters of protein form a funnel landscape[31]. The depth of the funnel represents energy and Q (the native state contacts formed) while width represents entropy. The unfolded protein occupies brim of the funnel where both energy and entropy are maximum. The native state with minimum energy and entropy occupies the apex of funnel while intermediate state lies intervening between the two states[32]. The Q value has used to determine the state of protein. The MG state is formed with a Q around 0.27 and transition state occurs around 0.6[33]. Thus it is a conceptual mechanism for understanding the self-organization of a protein and how it avoids the Levinthal paradox.

***In vivo* folding versus human diseases:** Proteins are known to fold *in vivo* to their final conformation either during or after their biosynthesis on ribosome. Very little is known directly about how and when this occurs *in vivo*. Protein folding *in vivo* is aided by molecular chaperones. These are proteins that bind to denatured states of proteins and prevent them misfolding and aggregating. The most interesting protein regarding folding is GroEL, which is a typical member of the Hsp60 or Cpn60 class of molecular chaperonins[34]. It consists of 14 x 58 kDa subunits arranged in two-stacked seven-member rings, which have a large central cavity. *In vivo* a co-chaperonin GroES is also required as is the hydrolysis of ATP[35]. The mechanism of GroEL is controversial. One school proposes that the central cavity acts as a folding cage in which a single molecule of denatured protein can fold in isolation. The opposing school postulates that GroEL is an active unfoldase that catalyses the unfolding of misfolding and aggregated states. A series of experiments examining the folding of barnase and CI2 in the presence of GroEL showed that both proteins can fold from their denatured states when bound to chaperone.

The chaperonin actually slows down folding by binding to the denatured state but it increases the yield of active protein. GroEL *in vivo* appeared to act as a folding and annealing cage.

Protein folding *in vivo* differs from the situation in a test tube in that nascent polypeptides attempt to fold even before their entire length has been synthesized. Since the protein folding in cell is so complex, there is a significant likelihood of defects arising in the process. There are a number of causes that could lead to the loss of function and thus a defective phenotype. These include thermodynamic destabilization of the native or intermediate state, alterations of the folding kinetics, prolonged or inappropriate association with molecular chaperones or folding enzyme preferential formation of off pathway or toxic conformations or folding in improper compartment. Therefore, inability of a essential protein to forms its native conformation under physiological conditions forms the basis of variety of human disease[36]. A list of some human diseases presented in Table 4 have been found to be a ramification of altered protein conformations[37-44].

***In silico* folding:** The inherent difficulties in solving experimentally a protein's tertiary structure only amplify the problem. Whereas it takes only hours to days to determine an amino acid sequence, for example, it would take months to years to discover its corresponding 3D shape by X-ray crystallography or nuclear magnetic resonance experiments. Equally challenging are experiments that explore folding process kinetics and dynamics. In short, efficient computational methods could help us tackle the protein-folding problem.

A number of rapidly folding proteins have been 1characterized in recent years[45-50]. These small proteins can provide the first direct comparisons between simulated and experimental protein folding kinetics and pathways. Proteins have been characterized through thermodynamic sampling methods, unfolding simulations and folding simulations using simple potentials. Computational predictions are in excellent agreement with the experimentally determined mean folding times and equilibrium constants. The convergence of experimentally and computationally accessible time scales will allow the comparison of absolute quantities characterizing *in vitro* and *in silico*.

BlueGene: On account of the complexity of the folding problem, computational requirement for such a prediction is a major issue which can be assessed by the estimate of time required to fold a 200 amino acid protein which involves $\sim 10^{-11}$sec per day per processor according to Newton's law of motion. This will take a million years for a single protein to fold. If one can use a million processors working together, protein can be folded in one-year computer time. In this sprit, IBM has launched the Blue Gene project. The Blue Gene project is a five-year effort started by IBM on Dec 1999 to build a massively parallel computer to be applied to the study of protein folding and other biomolecular phenomena. Atomistic simulation of protein folding in explicit solvent is the application they are aiming at.

CONCLUSIONS

Despite the fact that a number of theories have been proposed, none of these is able to explain the mechanism of protein folding completely and the folding problem is still a problem to solve. We require more experimental methods for the detection and characterization of dynamic structures in unfolding/folding pathway of the protein as well as more powerful computers. We are waiting, with

Table 4: Protein folding diseases

Diseases	Protein involved	Cause
Crueltzfeldt-jacob	Prion protein	Toxic fold/aggregation
Alzheimer's	Beta-amyloid	Toxic folding/aggregation
Cystic fibrosis	CFTR	Misfolding
Cancer	P 53	Misfolding
Familial amyloid polyneuropathy	Apolipoprotein	Aggregation
Parkinson	Alpha-synuclein	Amyloid fibril formation
Huntington's	Huntington	Amyloid fibril formation
Familia visceral amyloidosis	Lysozyme	Aggregation
Cataract	Crystallins	Aggregation
Marfan syndrome	Fibrillin	Misfolding
Scurvy	Collagen	Misfolding
Osteigenesis imperfecta	Type 1 procollagen pro alpha	Misassambly
Amyotropic Lateral sclerosis	SOD	Misfolding
Finish type Familia amyloidosis	Gesolin	Amyloid fibril formation
Kelanddic cerebral angiopathy	Cystatin C	Amyloid fibril formation
Type II diabetes	Islet anyloid polypeptide	Amyloid fibril formation

excitement, when we will be able to detect the first structural element that form during folding from extended amino acid sequence It is thought that the data mined from vast array of individual experiments on different kinds of protein systems may shed light on a common pattern in the folding mechanism. This far is beginning to emerge.

REFERENCES

1. Anfisen, C.B., 1973. Principle that govern the folding of protein chain. Science, 181: 223-230.
2. Callender, R.H., R.B. Dyer, R. Gilmanshin and W.H. Woodruff, 1998. Fast events in protein folding. Annu. Rev. Phys. Chem., 49: 173- 202.
3. Nolting, B., R. Golbic and A.R. Fersht, 1995. Sub millisecond events in protein folding. Proc. Natl. Acad. Sci. USA, 92: 10668-10672.
4. Plaxco, K.W. and C.M. Dobson, 1996. Time resolved biophysical methods in the study of protein folding. Curr. Opin. Struc. Biol., pp: 6630-636.
5. Haq, S.K., S. Rasheedi, P. Sharma, B. Ahmad and R.H. Khan, 2005. Influence of salts and alcohols on the conformation of partially folded states of stem bromelain at low pH. Intl. J. Biochem. Cell Biol., 37: 361-374
6. Ptitsyn, O.B, 1994. Kinetic and equilibrium intermediate in protein folding. Protein Eng., 7: 593-596.
7. Radford, S.E., C.M. Dobson, 1995. Insights into protein folding using physical techniques: Studies of lysozyme and α-lactalbumin. Phil. Trans. R. Soc. Lond. B., 348: 17-25.
8. Privalov, P.L., 1996. Intermediate states in protein folding. J. Mol. Biol., 258: 707-725.
9. Ahmad, B., M.K.H. Khan, S.K. Haq and R.H. Khan, 2004. Intermediate formation at lower urea concentration in B isomer of human serum albumin: A case study using domain specific lignads. Biochem. Biophys. Res. Commun., 314: 166-73.
10. Imoto, T., 1997. Stabilization of Protein. Cell. Mol. Life Sci., 53: 215.
11. Creighton, T.E., 1994. In Mechanism of Protein Folding. Oxford Univ. Press (Pain, R.H., Ed.) pp: 1-22.
12. Ptitsyn, O.B. and V.N. Uversky, 1994. The molten globlue is a third thermodynamical state of protein molecule. FEBS Lett., 341: 15-8.
13. Xie, D., V. Bhakuni and E. Freire, 1991. Calorimetric determination of the energetic of the molten globlue intermediate in protein folding. Apo α-lactalbumin. Biochemistry, 30: 10673-78.
14. Schmid, F.X. and H. Blaschek, 1981. A native-like intermediate on the ribonuclease A folding pathway. 2. Comparison of its properties to native ribonuclease Eur. J. Biochem., 114: 111-117.
15. Chan H.S. and K.A. Dill, 1998. Protein folding in the landspace perspective: Chevron plots and non-Arthenius kinetics. Proteins, 30: 2-33.
16. Zheng, Y.U.P. and C.W.U. Lawren, 2000. Autonomous protein folding units. Adv. Protein Chem., 53: 1-47.
17. Ohgushi, M. and A. Wada, 1983. Molten globlue state: A compact form of grobular protein with mobile side chains. FEBS Lett., 164: 21-24.
18. Ptitsyn, O.B., R.H. Pain, G.V. Semisotnov, E. Zerovnik and O.I. razgulyaev, 1990. Evidence for a molten globlue state as a general intermediate in protein folding. FEBS Lett., 262: 20-24.
19. Campos, L.A., M. Bueno, J. Lopez-Llano, M.A. Jimenez and J. Sancho, 2004. Structure of stable protein folding intermediate by equilibrium phi-analysis: The apoflavodxin thermal intermediate. J. Mol. Biol., 344: 239-255.
20. Engel, M.F., A.J. Visser and C.P. van Mierlo, 2004. Conformation and orientation of a protein folding intermediate trapped by adsorption. Proc. Natl. Acad. Sci., 101: 11316-21.
21. Arai, M. and K. Kuwajima, 2000. Role of the molten globlue state in protein folding. Adv. Protein Chem., 53: 209-271.
22. Schartz, G., 1993. The protein import mechinery of mitochondria. Protein Sci., 2: 141-146.
23. Kuwajima, K., Y. Hiraoka, M. Ikeguchi and M. Sugai, 1985. Comparison of the transient intermediate in lysozyme and α-lactalbumin. Biochemistry, 24: 874-81.
24. Jenning and Wright, 1993. Formation of molten glouble intermediate early in the kinetic folding pathway of apomyoglobin. Science, 262: 892-96.
25. Sawyer, L. and C. Holt, 1993. The secondry structure of milk proteins and their biological function. J. Dairy Sci., 76: 3062.
26. Jeng, M.F., S.W. Englander, G.A. Elove, A.J. Ward and H. Roder, 1990. Structural description of acid dentured cytochrome C by hydrogen exchange and 2D NMR. Biochemistry, 29: 10433-37.
27. Roder, H. and G.A. Elove, 1994. In Mechanism of Protein Folding. Oxford University Press (Pain, R.H., Eds.) pp: 26-54.
28. Dayson, H.J., and P.E. Wright, 1998. Equilibrium NMR studies of unfolded and partially folded proteins. Natl. Struc. Biol., 5: 499-503.

29. Kamatari, Y.O., R. Kitahara, H. Yamada, S. Yokoyama and K. Akasaka, 2004. High-pressure NMR spectroscopy for characterizing folding intermediate and denatured state of protein. Methods, 34: 133-143.
30. Henning, M., W. Bermel, A. Spencer, C.M. Dobson, L.J. Smith and H. Schwalbe, 1999. Side chain conformation of unfolded proteins. J. Mol. Biol., 288: 705-723.
31. Leopold, P.E, M. Montal and J.N. Onuchic, 1992. Protein folding funnels: A kinetic approach to the sequences sturcutre relatiocship. Proc. Natl. Acad. Sci. USA., 89: 8721-8725.
32. Laughlin, R.B., D. Pines, J. Schemalian, B.P. Stojkovic and P. Wolynes, 2000. The middle way. Proc. Natl. Acad. Sci., 97: 32-37.
33. Ferst, A.R., 1998. A Guide to Enzyme Catalysis and Protein Folding (Freeman, W.H. and Co.).
34. Corrales, F.J. and A.R. Fersht, 1995. The folding of GroEL-bound barnase as a model for chaperonin-mediated protein folding. Proc. Natl. Acad. Sci., USA., 92: 5326-5330.
35. Corrales, F.J. and A.R. Fersht, 1996. Towards a mechanism for GroEL. GroES chaperone activity: An ATPase-gated and -plused folding and annealing cage. Proc. Natl. Acad. Sci., USA., 93: 4509-13.
36. Carrel, R.W., D.A. Lomas, 1997. Conformational disease. Lancet, 350: 134.
37. Aguzzi, A. and C. Haass, 2003. Games played by rogue protein in prion disoreded and Alzheimer's disease. Science, 302: 814-8.
38. Otzen, D.E., S. Miron, M. Akke and M. Oliveberg, 2004. Transient aggregation and stable dimerization induced by introduced an Alzheimer sequence into a water-soluble protein. Biochemistry, 43: 12964-78.
39. Landles, C. and G.P. Bates, 2004. Huntingtin and the molecular pathogenesis of Huntington's disease. EMBO Rep., 5: 958-63.
40. Guggino, W.B. and S.P. Banks-Schlegel, 2004. Macromolecular interactions and ion transport in cystic fibrosis. Am. J. Respir. Crit. Care Med., 170: 815-20.
41. Foguel, D. and J.L. Silva, 2004. New insight into the mechanism of protein misfolding and aggrgation in amyloidogenic diseases derivied form pressure studies. Biochemistry, 43: 11361-70.
42. Ross, C.A. and M.A. Poirier, 2004. Protein aggregation and nuerodegenerative disease. Natl. Med., S10-7.
43. Nayeem, M.S. and R.H. Khan, 2004. Misfolded protein and human diseases. Protein Pept. Lett., 11: 593-600.
44. Rochet, J.C. and J.P.T. Lansbury, 2000. Amyloid fibrillogenesis: Themes and variations. Corr. Opin. Stuct. Biol., 10: 60-8.
45. Dehouck, Y., D. Gilis and M. Rooman, 2004. Database-derived potentials dependent on protein size for *in silico* folding and design. Biophys. J., 87: 171-81.
46. Watters, A.L. and D. Baker, 2004. Searching for folded proteins *in vitro* and *in silico*. Eur. J. Biochem., 271: 1615-22.
47. Brylinski, M., W. Jurkowski, L. Konieczny and I. Roterman, 2004. Limited conformational space for early-stage protein folding simulation. Bioinformatics, 20: 199-205.
48. Lu, L., A.K. Arakaki, H. Lu and J. Skolnick, 2003. Multimeric threading-based prediction of protein-protein interaction on a genomic scale: Application to the saccharomyces cerevisiae proteome. Genome Res., 13: 1146-54.
49. Aita, T., M. Ota and Y. Husimi, 2003. An *in silico* exploration of the neutral network in protein sequence space. J. Theor. Biol., 221: 599-613.
50. Attwood, T.K. and C.J. Miller, 2002. Progressive in bioinformatics and the importance of being earnest. Biotechnol. Annu. Rev., 8: 1-54.