Brazilian Journal of Medical and Biological Research (2001) 34: 419-435
ISSN 0100-879X

Protein folding: a perspective for biology, medicine and biotechnology

J.M. Yon

Institut de Biochimie, Biophysique Moléculaire et Cellulaire, UMR CNRS, Université de Paris-Sud, Orsay, France

Abstract

Correspondence
J.M. Yon
Institut de Biochimie, Biophysique
Moléculaire et Cellulaire
UMR CNRS, Université de Paris-Sud
Orsay
France
Fax: +1-6985-3715
E-mail: jeannine.yon@mip.u-psud.fr

Presented at the XIV Annual Meeting of the Federação de Sociedades de Biologia Experimental, Caxambu, MG, Brazil, August 25-28, 1999.

Received June 1, 2000
Accepted February 5, 2001

At the present time, protein folding is an extremely active field of research including aspects of biology, chemistry, biochemistry, computer science and physics. The fundamental principles have practical applications in the exploitation of the advances in genome research, in the understanding of different pathologies and in the design of novel proteins with special functions. Although the detailed mechanisms of folding are not completely known, significant advances have been made in the understanding of this complex process through both experimental and theoretical approaches. In this review, the evolution of concepts from Anfinsen's postulate to the "new view" emphasizing the concept of the energy landscape of folding is presented. The main rules of protein folding have been established from *in vitro* experiments. It has been long accepted that the *in vitro* refolding process is a good model for understanding the mechanisms by which a nascent polypeptide chain reaches its native conformation in the cellular environment. Indeed, many denatured proteins, even those whose disulfide bridges have been disrupted, are able to refold spontaneously. Although this assumption was challenged by the discovery of molecular chaperones, from the amount of both structural and functional information now available, it has been clearly established that the main rules of protein folding deduced from *in vitro* experiments are also valid in the cellular environment. This modern view of protein folding permits a better understanding of the aggregation processes that play a role in several pathologies, including those induced by prions and Alzheimer's disease. Drug design and *de novo* protein design with the aim of creating proteins with novel functions by application of protein folding rules are making significant progress and offer perspectives for practical applications in the development of pharmaceuticals and medical diagnostics.

Key words
- Protein folding
- Energy landscape
- Misfolding and disease

Introduction

An important challenge in molecular biology is finding the rules that determine how a nascent polypeptide chain acquires its three-dimensional and functional structure. Indeed, the rapid progress in genome sequencing, including that of the human genome, which has aroused great interest in medical circles, makes the solution of the problem all the more urgent. Sequencing the gene is only the first step; it is essential to know the nature, the structure and the function of the protein that is coded by the gene. For diagnosis of the true cause of a disease, and for an approach to a rational treatment, after having located the gene responsible for a disease, it is essential to know the structure of the protein for which it codes.

Protein folding, the second translation of the genetic message, completes the information transfer from DNA to the active protein product. In other words, for a complete understanding of this process, it is necessary to decipher the folding code, the second part of the genetic message: DNA→RNA→polypeptide chain — ?→active protein.

There are about 100,000 proteins in the human genome and about 10^{11} in all organisms. To become biologically active these proteins must fold into a stable three-dimensional structure. In spite of the great diversity of proteins, the number of folds is relatively small, less than 700 observed to date, while protein domains exhibit only 32 different architectures according to a recent classification (1). Nature has created complexity through the combination of a small number of simple elements, such as the two most common elements of secondary structure observed in proteins: α-helices and ß-strands.

The question of the mechanism of protein folding has intrigued scientists for many decades. As far back as 1929, a time when nothing was known about protein structure, Wu had analyzed the reverse of the folding process, denaturation. Shortly thereafter, Northrop in 1932, and Anson and Mirsky in 1934-35 succeeded in reversing the denaturation of several proteins, such as hemoglobin, chymotrypsinogen, and trypsinogen. During the decade between 1950 and 1960, several thermodynamic studies of the unfolding-refolding process and the role of noncovalent interactions in protein stability were published by several authors, among them, Linderstrøm-Lang, Lumry, Klotz and Kauzmann. In 1959, Kauzmann proposed that the hydrophobic effect is the driving force in directing the folding process. Later, the determination of the three-dimensional structures of proteins by X-ray diffraction provided a new basis on which to study the folding process. Significant progress began to be made when Anfinsen successfully refolded denatured and reduced ribonuclease into the fully active enzyme. He stated the fundamental principle of protein folding in 1973: the folding of a protein is determined by its amino acid sequence. These historical aspects have been reviewed by Ghélis and Yon (2). The evidence gathered over many years supports this principle even for folding *in vivo* assisted by molecular chaperones.

The fundamental questions are the following. How does the sequence code for the fold, given that the backbone of all proteins has the same composition; in other words, how do the side chains dictate the overall fold? How does a given sequence find its specific native structure in a finite time among the enormous number of possible conformations that a polypeptide chain could adopt? How is the folding process initiated and what is (are) the pathway(s) of folding? And last, are the main rules of protein folding deduced from *in vitro* studies valid for folding *in vivo*?

At the present time, protein folding is an extremely active field of research involving aspects of biology, chemistry, biochemistry, computer science and physics. The fundamental principles have practical applications in the exploitation of the recent advances in genome research, in the understanding of several pathologies and in the design of novel proteins with special functions. Although the detailed mechanisms of folding are not completely known, significant advances have been made in the understanding of this complex process through both experimental and theoretical approaches (3).

The fundamentals of protein folding from the Anfinsen postulate to the new view

The Anfinsen postulate and the Levinthal paradox

The remarkable achievement of C. Anfinsen and his group in refolding denatured and reduced ribonuclease into a fully active enzyme marked the beginning of the modern era of the protein folding problem. From his results, the author concluded that “all the information necessary to achieve the native conformation of a protein in a given environment is contained in its amino acid sequence” (4). The thermodynamic control of protein folding is a corollary to the Anfinsen postulate; it means that the native structure is at a minimum of the Gibbs free energy. This statement was discussed by Levinthal in a consideration of the short time required for the folding process *in vitro* as well as *in vivo*. Indeed, for a 100-amino acid polypeptide chain, if we assume only two possible conformations for each residue, there are 10^{30} possible conformations for the chain as a whole. If only 10^{-11} second is required to convert one conformation into another, a random search of all conformations would require 10^{11} years, an irrealistic time in a biological context where the folding time is of the order of seconds or minutes. Thus, it is clear that evolution has found an effective solution to this combinatorial problem. This is referred to as the Levinthal paradox and has dominated discussions for the last 30 years. Different mechanisms have been suggested in order to solve the Levinthal paradox. Among them is Wetlaufer’s proposed model in which protein folding is under kinetic control rather than thermodynamic control. This states that the protein is trapped in an energetic minimum which is not the global minimum, a high energetic barrier preventing the protein from reaching the latter (see Ref. 2).

Folding models and pathways of protein folding

In order to understand how the polypeptide chain could overcome the Levinthal paradox, different folding models arising from theoretical considerations (5, and references therein), folding simulations, or experimental observations (6, and references therein) have been proposed.

The classical nucleation-propagation model, which applies to helix-coil transitions, involves a nucleation step followed by a rapid propagation, the limiting step being the nucleation process. This model has been proposed to explain the folding of ribonuclease A but was forsaken after new kinetic studies of the refolding of ribonuclease were performed (6). More recently a nucleation-condensation model, different from the classical one, has been proposed by Fersht (7). This model proposes a mechanism involving a weak local nucleus which is stabilized by long range interactions.

A stepwise sequential and hierarchical folding process, in which several stretches of structure are formed and assemble at different levels following a unique route, has been supported by several authors for many years (6,8). According to this model, the first event, nucleation, is followed by the formation of secondary structures which associate to generate supersecondary structures, then domains and eventually the active monomer; the association of domains induces the last conformational refinements which generate the functional properties. Such a hierarchy of protein folding corresponds to the hierarchy of protein structure. Similarly, the framework model assumes that the secondary structure is formed in an early step of folding, before the tertiary structure, emphasizing the role of short range interactions in directing the folding process (9).

A modular model of folding was suggested on the basis of the three-dimensional structures of proteins. This model assumes that not only domains, but also subdomains can be considered as folding units which fold independently into a native structure, forming structural modules that assemble to yield the native protein (10,11).

The diffusion-collision model of folding was developed in 1976 by Karplus and Weaver and reconsidered in 1994 in the light of more recent experimental data (12). In such a model, nucleation occurs simultaneously in different parts of the polypeptide chain generating microstructures which diffuse, associate and coalesce to form substructures with a native conformation. These microstructures have a lifetime controlled by segment diffusion, so the folding of a polypeptide chain containing 100 to 200 amino acids can occur within a very short time, less than a second. According to this model, folding occurs through several diffusion-collision steps.

The hydrophobic collapse model implies that the first event of protein folding consists of a collapse via long range hydrophobic interactions and occurs before the formation of a secondary structure (13). The idea originates from the work of Kauzmann considering the hydrophobic effect as the driving force in protein folding and stabilization. Later, Dill and co-workers proposed that the formation of stretches of secondary structures occurs simultaneously with the hydrophobic collapse.

The jigsaw puzzle model was introduced in 1985 by Harrison and Durbin (14). This model admits the existence of multiple folding routes to reach a single solution. According to this hypothesis, the identification of folding intermediates represents a kinetic description rather than a structural one, each intermediate consisting of heterogeneous species in rapid equilibrium. It was the subject of controversy until recently but presents some similarities with the diffusion-collision model proposed by Karplus and Weaver.

Detection and characterization of intermediates in protein folding

The unfolding-refolding transition under equilibrium transition has often been described as a two-state process in which only the unfolded and the native species are significantly populated. The intermediates are generally unstable and poorly populated under equilibrium conditions. The two-state approximation, which applies to very cooperative transitions, is frequently valid for small proteins. Such an approximation allows the determination of ΔG_o, the free energy of denaturation. This value varies between 5 and 15 kcal/mol for most proteins studied thus far, indicating a relatively low stability (for reviews, see 2,15).

The existence of intermediates has been shown from kinetic studies for most proteins even when the two-state approximation describes the overall denaturation process. For many proteins, monophasic unfolding kinetics and multiphasic refolding kinetics are observed. This is one of many ways in which the experimental evidence proves the occurrence of intermediates in the folding pathway. The structural characterization of such intermediates is a prerequisite to solving the folding problem. However, while kinetic studies demonstrate the existence of intermediates and their location on the folding pathway, these intermediates are often too poorly populated to obtain detailed structural information. Two major impediments to characterizing these species are thus the high cooperativity and the rapidity of the process, especially in the early events of protein folding.

In spite of these inherent difficulties, much effort has been devoted to characterizing these transient species. Kinetic trapping of intermediates during the refolding of disulfide-bridged proteins was developed for lysozyme (16) and used for bovine pancreatic trypsin inhibitor (BPTI) (17,18). An elegant method using differential chemical labeling has been elaborated by Ghélis (19) and applied to the refolding of elastase.

In the past decade, technological advances have improved our approaches to the study of the protein folding process. Several methods have been developed allowing some of the intermediates to be characterized, particularly stopped-flow mixing devices coupled to circular dichroism, and NMR using rapid hydrogen-deuterium exchange associated with a mixing system allowing for the pulse labeling of transient species. This method is very informative, yielding residue-specific information (20-22). Protein engineering has also been successfully employed to stabilize intermediates or to probe particular regions of a protein during the folding process (23-25). Another approach frequently used is the study of protein fragments (10,26). Transient folding species have been found to accumulate at low pH for several proteins, and especially for α-lactalbumin, permitting the study of their structural properties.

The formation of secondary structures in the early steps of protein folding has been observed for many proteins (24,27). Such early species with a high content of secondary structures were named "the molten globule" by Ohgushi and Wada (28). Ptitsyn (27) has suggested that it is a general intermediate in the folding pathway of proteins. The literature being rather confusing concerning the structural characteristics of the molten globule state, Goldberg and colleagues (29) have introduced the term "specific molten globule", and defined its characteristics. The specific molten globule is a rather compact intermediate with a high content of native secondary structure, but a fluctuating tertiary structure. It contains an accessible hydrophobic surface susceptible to binding a hydrophobic dye, aniline naphthalene sulfonate. Since the tertiary structure is not stabilized, the aromatic residues can rotate in a symmetrical environment and are accessible to the solvent, as assessed by the absence of near UV circular dichroism. The formation of a molten globule as an early folding intermediate has been reported for several proteins, among them α-lactalbumin, carbonic anhydrase, ß-lactamase, the α- and $ß_2$-subunits of tryptophan synthase, bovine growth hormone, and phosphoglycerate kinase (24,27,29).

However, the secondary structures observed in the early steps of the folding process are not always identical to those observed in the native structure. In the refolding of the $ß_2$-subunit of tryptophan synthase (29), a rapid formation of non-native secondary structures precedes a reorganization yielding the native structure. Similarly, non-native secondary structures are formed during the refolding of ß-lactoglobulin (30). Even in the case of the paradigmatic molten globule of α-lactalbumin, some differences have been observed by NMR spectroscopy. This molten globule appears as a heterogeneous species. The helical domain is structured, but has highly fluctuating hydrophobic interactions, whereas the ß-sheet is rather disordered. Furthermore, the side chains of Tyr103, Trp104 and His107 are packed within a hydrophobic cluster in a region that differs in structure from the native one (31). The folding pathway of hen egg white lysozyme includes transient steps corresponding to a reorganization of secondary structures (32).

An intermediate preceding the molten globule state has been identified by Ptitsyn (27), and Uversky and Ptitsyn (33). This species, less compact than a molten globule, has a significant secondary structure content but smaller than that of a molten globule, and displays hydrophobic regions accessible to a solvent. It has been called a "pre-molten globule" by Jeng and Englander (34). Its occurrence has been observed during the cold denaturation of ß-lactamase and carbonic anhydrase, and also during the refolding of several proteins (35). The rapid formation of transient intermediates, either molten or pre-molten, or both, with a high content of secondary structure and a small amount of fluctuating tertiary structure, is supported by a great number of observations (22,35). However, since these transient states are formed within the dead-time of a stopped-flow device, it is possible that their formation might be preceded by an earlier event.

According to another point of view, the

first event in the folding of a polypeptide chain consists of a hydrophobic collapse preceding the formation of secondary structure or occurring simultaneously, and being followed by a rearrangement of a small number of condensed states. This view emphasizes both the hydrophobic effect and the role of long range interactions in the initiation of the folding process. Several experimental data are consistent with a hydrophobic collapse during the early stages of folding. For example, residual microstructures persisting during the denaturation process have been characterized for several proteins. These microstructures may be involved in the folding process as hydrophobic nucleation centers. They have been observed in the folding of 434 repressor, in FK506 binding protein, in staphylococcal nuclease, in dihydrofolate reductase, and in a peptide from BPTI (for a review, see 15). In phosphoglycerate kinase and two of its mutants containing only one of the two tryptophan residues, residual microstructures around W308 and W333 have been detected by fluorescence emission spectroscopy during the guanidinium/HCl-induced unfolding of the protein; these microstructures consist of hydrophobic clusters (25).

Classical rapid mixing techniques such as stopped-flow, continuous flow and quenched-flow are limited to the millisecond time scale, preventing analysis of the events occurring within the initial burst phase of folding. Most of the secondary structures are formed within the dead-time of a stopped-flow device. This is particularly inconvenient as a crucial part of the folding problem is the characterization of the very early intermediates. Recent technical advances to improve the resolution time of kinetic studies have been made (36). Submillisecond mixing techniques have been developed and applied to the refolding of cytochrome *c*. Non-mixing techniques such as the classical T-jump, nanosecond infrared laser-induced T-jump and picosecond T-jump have been used to study the refolding of cold-denatured proteins. Other rapid techniques such as nanosecond laser photolysis, optical electron transfer and dynamic NMR methods have been reported. They allow detection of very fast events which occur on a time scale of less than a microsecond. For comparison, it has been shown that the time scale for helix formation is 10^{-7} s, and of the order of 10^{-6} for the formation of a ß-hairpin or a 10-residue loop (37). The upper limit for the rate of protein folding has been evaluated to be around 1 µs (38), in agreement with theoretical estimates (39). The initial collapse of apomyoglobin into a compact state is complete in less than 20 µs. Moreover, the use of a laser T-jump method has allowed the detection of two very fast phases during the folding of this protein. The first one occurring in 250 ns is a local collapse around Trp14 and consists of non-native hydrophobic contacts and helix formation. During the second phase within 5 µs, helix A makes contact with two other helices, H and G. These two rapid phases are followed by a slower final phase of 0.9 s generating the native protein as shown by Ballew et al. (40). The refolding of cytochrome *c* also starts by a rapid collapse occurring within 50 µs followed by a massive chain condensation (41).

It appears therefore that the very fast events of protein folding consist of a hydrophobic collapse accompanied or not by the formation of secondary structures, depending on the protein. Intermediate events on the folding time scale occur after the formation of the molten globule and before the rate-limiting step of the folding process which generates the native and functional structure. In these intermediate phases, the appearance of substrate- or ligand-binding sites can be observed. For example, in lactalbumin, Ca^{2+}-binding sites appear before completion of the native structure (42).

In the final rate-limiting step, the protein achieves its native conformation with the emergence of functional properties. These final events correspond to the precise ordering of the elements of secondary structure, the correct packing of the hydrophobic core, the correct domain pairing in multidomain proteins, the reshuffling of disulfide bonds, cis-trans proline isomerization, and subunit assembly in oligomeric proteins. For several proteins, the rate-limiting step consists of the reorganization of misfolded species (for reviews, see 2,6,15,43). Figure 1 illustrates a folding pathway with kinetic competition between correct folding and a side reaction leading to the formation of aggregates.

Domains and subdomains in protein folding

Domains are compact substructures within a protein molecule. They have been considered as folding units by Wetlaufer, forming structural modules that fold independently and assemble to generate the native structure. Many experimental data have shown that domains behave as autonomous folding units, as reviewed by Ghélis and Yon (2), Yon (15) and Jaenicke (8,43). In our laboratory, we have studied the role of structural domains in phosphoglycerate kinase whose structure is organized into two domains of approximately the same size (Figure 2). The independently expressed engineered N- and C-domains have quasi-native structures and are capable of refolding cooperatively. The C-domain binds the ATP substrate with the same affinity as the native protein. However, even though these isolated domains can refold independently, it has not proved possible via a variety of experimental procedures to reproduce functionally active protein from the two separate domains. The ability of independently folded domains to associate to yield a functional protein has been observed only with a few proteins, such as thioredoxin, elastase, and methionyl-t-RNA synthetase (reviewed in Ref. 15). Thus, it seems that, for many proteins, the correct folding of isolated domains is not sufficient to yield a functional ensemble. Rather, the interactions between domains are required during the folding process to allow the structural adjustments that yield a functional protein.

In many proteins, the N and C termini are spatially adjacent in the folded state. In phosphoglycerate kinase, for example, circular permutations have been obtained by protein engineering that introduce a discontinuity into either one or the other domain. Such a change can modify the sequence of events in the folding process, but does not prevent the protein from achieving a native and functional conformation. However, it significantly decreases the stability of the enzyme (by 4 kcal/mol), suggesting that the continuity of

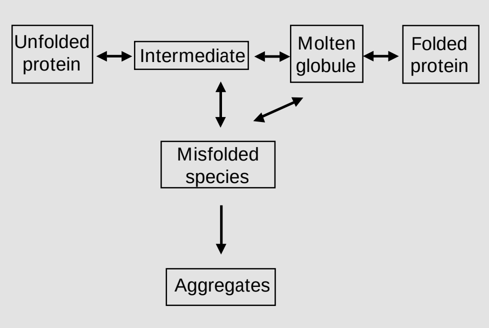

Figure 1 - Illustration of a schematic pathway for protein folding with a side reaction and the formation of aggregates.

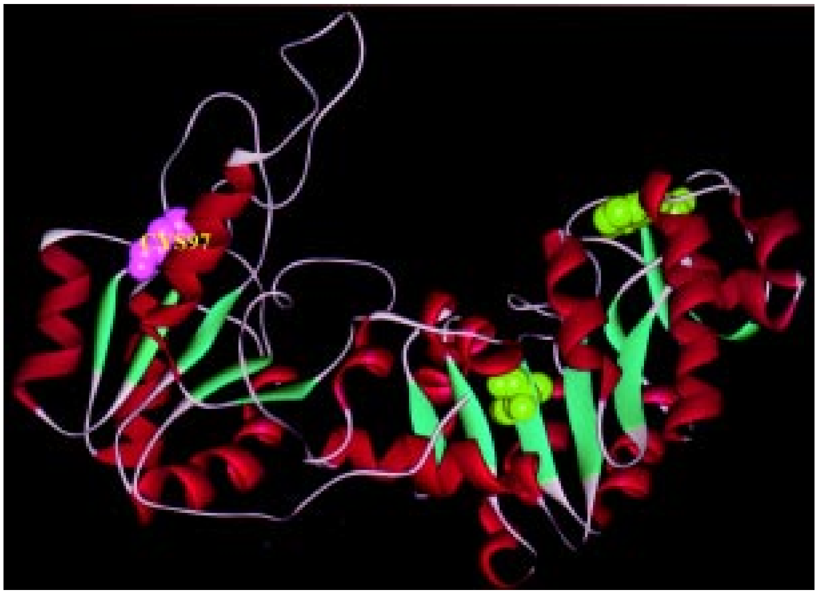

Figure 2 - Structure of yeast phosphoglycerate kinase. The two tryptophans are colored in green and Cys97 in magenta.

the domains is required for the stability but not for the correct folding of phosphoglycerate kinase (15). It is likely that during evolution structural patterns of proteins consisting of continuous domains have been selected on the basis of thermodynamic stability. However, proteins have enough plasticity to find their functional fold, even when the continuity of the domains has been artificially disrupted.

It has been proposed that subdomains, which represent compact regions smaller than domains within a protein, might also fold autonomously, forming folded modules that assemble to generate the native protein (11). These subdomains are proposed to be condensed states in which the close packing of the atoms in the hydrophobic core of the molecule has not yet taken place. Oas and Kim have reported that fragments from BPTI corresponding to subdomains fold autonomously and then associate. The results obtained on the folding and complementation of fragments from barnase are also consistent with a modular model. In contrast, this model cannot account for the folding of barley chymotrypsin inhibitor 2. Likewise, fragments smaller than a domain cannot be considered as folding units in the SH2 domain of proteins p60 and p85, in staphylococcal nuclease, in tryptophan synthase, or in cytochrome *c*. Several pairs of structural fragments from phosphoglycerate kinase have been obtained. Some of them, smaller than a domain, have a quasi-native structure; but their unfolding transition exhibits a very low cooperativity, as discontinuities in the ß-sheet regions perturb the folding process. Furthermore, several pairs of adjacent fragments have been found to give a functional complementation. Among them, at least one of the two fragments was not significantly folded when isolated. This demonstrates that the association of individual fragments with a non-native structure can favor subsequent rearrangements to functional native structures through long range interactions (for a review, see Ref. 15).

The new view: the energy landscape and the folding funnel

Many phenomenological models have been thus proposed over the years; more recently a unified model of protein folding based on the effective energy surface of the polypeptide chain has emerged. This so-called new view of protein folding arises from theoretical studies. Although simplified to take into account the computational limitations, several models have been proposed to overcome the Levinthal paradox by simulation of folding from random coil to the native structure. There are essentially two approaches, lattice models and molecular dynamics simulations.

In the lattice models, the protein chain is represented as a string of beads on a two-dimensional square lattice, or on a three-dimensional cubic lattice (Figure 3). The interactions between residues (the beads) provide the energy function for Monte Carlo simulations. In such models, the aim is not to examine the folding of a particular amino acid sequence. Instead, the idea is to include the most essential features of proteins, i.e., the heterogeneous character of the interactions (hydrophobic and polar) and the existence of long range interactions to explore the general characteristics of the possible folds. Such an approach has played an important role in polymer science. Lattice models were first applied to protein folding by Go and co-workers and simple exact models for proteins were initiated by Dill and co-workers (for a review, see 44) and have been used by several researchers. In these models, the native state corresponding to the energy minimum is a compact globule with a native fold which has the characteristics of the experimentally observed molten globule with secondary structure, but not tertiary interactions. From the lattice simulations, insights into possible folding scenarios have been obtained, providing a basis for exploring the general characteristics of folding for real proteins. The exploration of such models supplies useful information that can be submitted to experimental tests.

Using such a 3 x 3 x 3 cubic lattice model, Karplus and co-workers (45) have studied the folding of a 27mer (a polypeptide of 27 amino acids). Two hundred random amino acid sequences were generated, each characterized by an interaction matrix for all pairs of beads. Each was submitted to Monte Carlo simulations. The results indicated that 30 of the 200 sequences found their native state very easily, whereas 146 never found the native state. The difference between these two sets of sequences, strongly folding and non-folding, was not to be found in their structural features. Instead, it was found in the large energy gap between the native and the excited states in the strongly folding sequences. Under these conditions, the folding can be very fast. The energy gap concept focuses on small fragments and their role as early folding units. Thus, the Levinthal paradox can be overcome by the simultaneous and fast formation of microstructures in several regions of the protein, restricting the number of possible conformations during their subsequent assembly to the final structure.

The so-called "new view" has evolved during the past few years from both experiment and theory through the use of simplified mechanical models and is illustrated by the concept of the folding funnel introduced by Wolynes and co-workers (46). The model is represented in terms of an energy landscape and describes the thermodynamic and kinetic behavior of the transformation of an ensemble of unfolded molecules to a predominantly native state (Figure 4). As underlined by Wolynes et al. (46): "to fold, a protein navigates with remarkable ease through a complicated energy landscape". A wide variety of folding behavior emerges from the energy landscape depending on the energetic parameters and conditions. The authors have suggested that the folding rate is slowed by ripples in the energy landscape corresponding to local minima populated by transiently stable intermediates. According to this model, there are parallel micropathways, each individual polypeptide chain following its own route. Towards the bottom of the folding funnel, the number of protein conformations decreases as does the chain entropy. The steeper the slope, the faster the folding. In Figure 5 (left), an idealized smooth funnel is represented in three dimensions; in this case, the protein can fold very quickly exhibiting a two-state behavior. In contrast, in Figure 5 (right), a rugged energy landscape with kinetic traps formed by energy

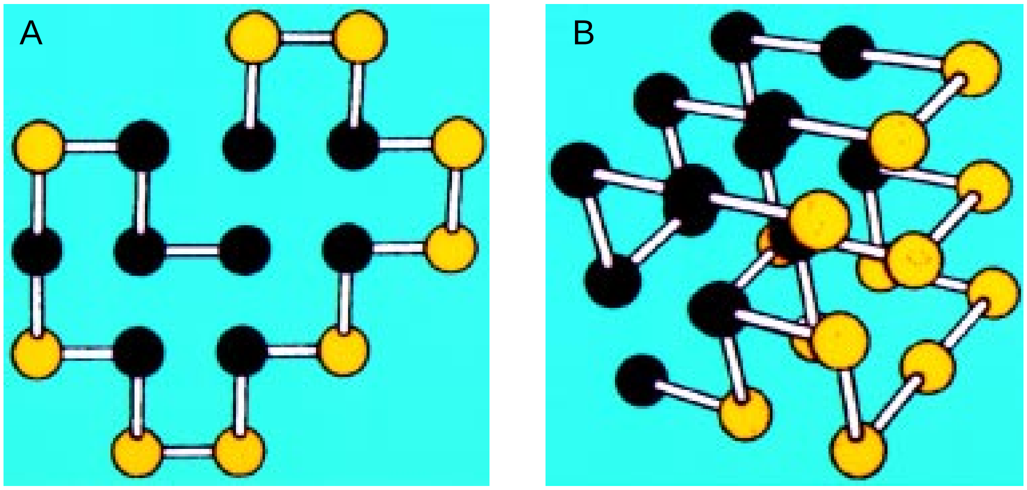

Figure 3 - Lattice models. *A*, Square; *B*, cubic.

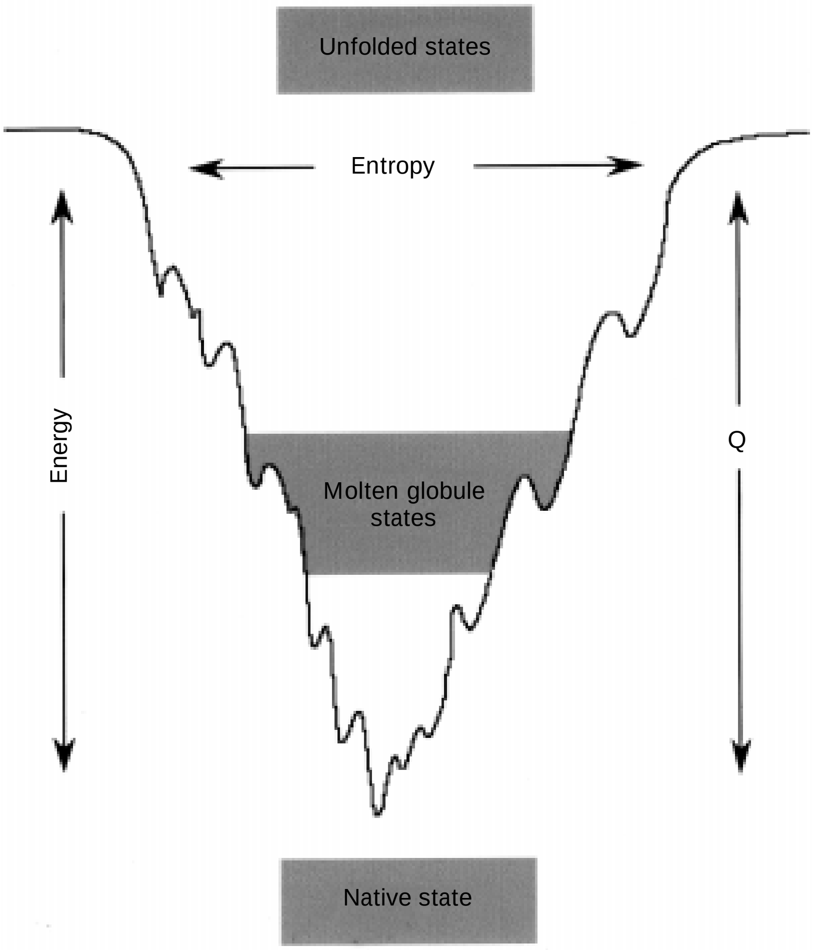

Figure 4 - A schematic two-dimensional representation of the folding funnel. The width of the funnel represents entropy, and depth the energy. Q is the fraction of native contacts indicated for each collection of states.

barriers is represented; in this case folding can be multistate and slower (47). When local energy barriers are high enough, protein molecules can be trapped and possibly aggregate.

The new view has progressively replaced the idea of folding pathways. The energy landscape metaphor provides a conceptual framework for understanding both two-state and multistate kinetics, and also protein misfolding and subsequent aggregation. This view is similar to the jigsaw puzzle model (14).

Many experimental results are consistent with this view. There is an increasing amount of evidence showing that the initially extended polypeptide chain folds through a heterogeneous population of partially folded intermediate species in fluctuating equilibrium. It has been reported that hen egg lysozyme and cytochrome *c* refold according to parallel alternative pathways (45). Heterogeneous species have also been detected during the refolding of phosphoglycerate kinase. Furthermore, for the first time, it was shown that transient oligomeric species are formed very rapidly, but that they nevertheless yield monomeric native protein during the slow step of folding. The associations occur via the N-terminal domain (48). More

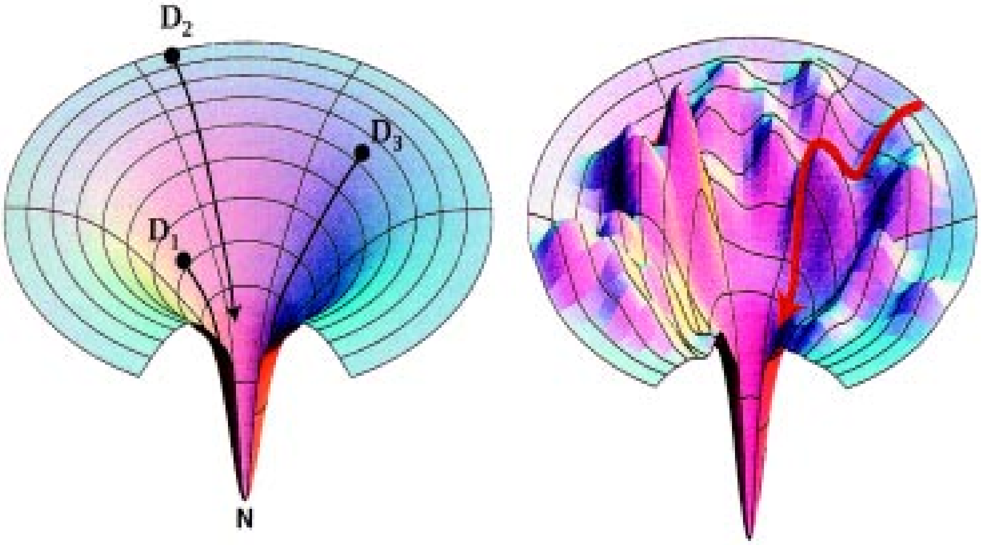

Figure 5 - Three-dimensional representation of folding funnels (see text) according to Dill and Chan (47; reproduced with permission).

recently, transient oligomeric species have also been reported during the folding of the 102-residue monomeric protein U1A, and chymotrypsin inhibitor 2 (49). A transient dimer has been characterized by small angle X-ray scattering in the case of the refolding of cytochrome *c* (50).

From the convergence of theoretical and experimental studies, a unified view of the folding process has progressively emerged, providing also an explanation of the aggregation processes observed in several pathologies. This view has replaced the classical one of a hierarchical and sequential pathway and is now quite generally accepted. Even Baldwin recognized in 1994 (51): “In retrospect, it seems that Harrison and Durbin may have had the right idea”.

Folding in the biological environment

The main rules of protein folding have been established from *in vitro* experiments. It has been accepted that the *in vitro* refolding process is a good model for understanding the mechanisms by which a nascent polypeptide chain reaches its native conformation in the cellular environment. Indeed, many denatured proteins, even those whose disulfide bridges have been disrupted, are able to refold spontaneously. This view was nevertheless challenged by the discovery of molecular chaperones in 1987. From that moment on, a great amount of information has been gathered. The structural data now available together with functional studies led to significant progress in the understanding of the mechanisms used by the chaperone machinery to assist protein folding. There are now more than 20 protein families that have been described as molecular chaperones. It is now well established that neither further information nor external energy is required to achieve the correct folding in chaperone-assisted systems. The ATP hydrolysis by GroEL is only used to induce the conformational change of the chaperone which permits the release of the folded protein. The molecular chaperones by their transient association with nascent, stress-destabilized or translocated proteins, have a role in preventing improper folding and subsequent aggregation. They do not interact with folded proteins. They transiently associate with an early intermediate on the folding pathway, probably a molten globule or a pre-molten globule, by hydrophobic interactions. They do not carry information capable of directing a protein to assume a structure different from that dictated by the amino acid sequence. Furthermore, they increase the yield but not the rate of folding reactions; in this respect they do not act as catalysts.

Most small chaperones bind transiently to small hydrophobic regions of nascent polypeptide chains, preventing aggregation and premature folding. In contrast, large chaperones such as the GroEL-GroES system in prokaryotes or TriC in eukaryotes completely sequester partially folded molecules in a central cage. This cage is provided by the heptameric double ring of GroEL and is capped by GroES to prevent the premature release of the folding protein. This cage is large enough to accommodate proteins up to about 70 kDa. Figure 6, according to Wang and Weissman (52), illustrates the GroEL reaction cycle. The non-native protein binds to the apical domains of the unoccupied upper ring of GroEL-GroES. ATP and GroES bind to the same ring sequestering the protein. The binding of GroES induces a large conformational change in GroEL and ATP hydrolysis induces a conformational change in the bottom ring allowing it to bind a non-native protein. This promotes subsequent binding of ATP and GroES in the lower ring, disrupting the upper complex and ejecting GroES and releasing the protein. If the protein has not reached the native state, it is subjected to a new cycle. The hydrolysis is required in some cases but not all for the release of the protein. The energy of ATP is used to provoke a conformational change in the chaperone allowing the release of the protein. As concluded by Hartl: “The role of ATP hydrolysis is mainly to induce conformational change in GroEL that results in GroES cycling at a physiologically relevant rate”. Therefore molecular chaperones assist the *in vivo* folding without violation of Anfinsen’s postulate. The same mechanisms direct the *in vivo* and *in vitro* folding processes.

In vitro experiments on protein folding have shown that misfolding and subsequent aggregation result from a kinetic competition between correct folding and an off-pathway process. It results from the occurrence of local minima in the folding funnel described in the previous section. When the formation of the correct structure is kinetically favored *in vivo*, molecular chaperones are not required. This was shown not only for monomeric, but also for dimeric proteins such as the p22 Arc repressor studied by Sauer and co-workers. The probability of incorrect interactions will be greater if the rate of biosynthesis is much slower than the rate of folding. In such a case, the partially synthesized polypeptide chain could undergo incorrect folding missing long range interactions. *In vivo*, fewer than 15% of biosynthesized proteins need the assistance of molecular chaperones to achieve their native structure (reviewed in Ref. 53).

Other accessory molecules are able to play a helper role in the folding of proteins *in vivo*. Protein disulfide isomerase, an abundant component of the lumen of the endoplasmic reticulum, catalyzes the formation of disulfide bonds in secretory proteins. Another enzyme, peptidyl-prolyl-cis-trans isomerase, catalyzes the cis-trans isomerization of X-Pro peptide bonds. These enzymes only accelerate the process, which can also be achieved in their absence *in vitro* under adequate conditions and are thus adequately classified as enzymes.

Protein aggregation and its implications in human diseases

Protein aggregation is a widespread phenomenon that can occur under particular conditions or following certain mutations in the polypeptide chain (54). It occurs commonly when proteins are overexpressed in foreign organisms, with the high concentration of material leading to the formation of inclusion bodies. The role of molecular chaperones in cells is to prevent aggregation and the ability of normal proteins to fold rapidly is an important evolutionary development to minimize aggregation.

Another consequence of abnormal protein folding yields aggregates with the appearance of amyloid fibrils, leading to the spongiform encephalopathies. These severe pathologies include prion-associated diseases

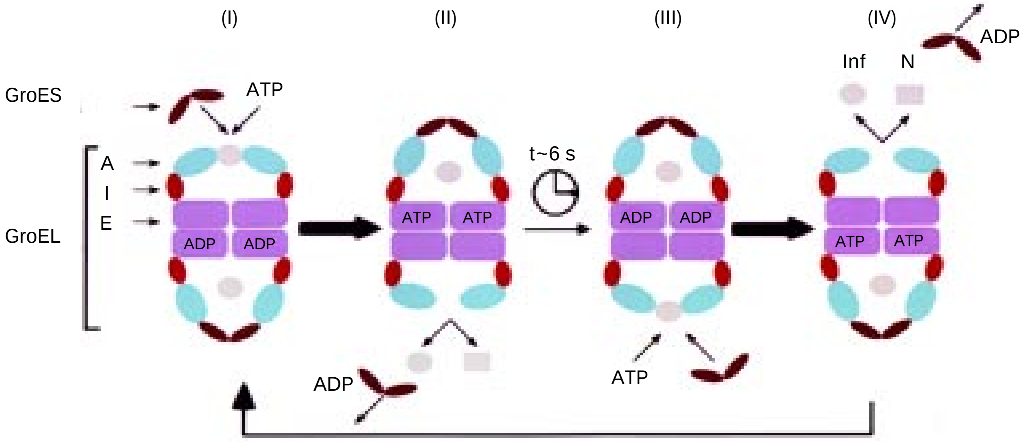

Figure 6 - The GroES-GroEL cycle according to Wang and Weissman (52). Inf is the unfolded protein, N the folded one. A is the apical domain (in blue) which binds the unfolded protein and GroES; I is the intermediate domain (in red) and E is the equatorial domain (in magenta) which binds and hydrolyzes ATP (reproduced with permission).

such as scrapie in sheep, mad cow disease in cattle and Creutzfeld-Jacob in humans. Alzheimer's disease is also characterized by the presence of amyloid fibrillar deposits in the brain tissue. However, Alzheimer's disease is a more complex process during which the amyloid protein precursor is initially cleaved by α or ß secretase, and then by γ secretase generating peptides of 40 and 42 amino acids, Aß(40) and Aß(42), which aggregate into amyloid structures. There are 16 human amyloidogenic proteins known to abnormally self-assemble into fibrils 60-100 Å in width and of variable length. They are characterized by a cross-ß repeat structure (55). Table 1 summarizes these different amyloidogenic proteins and the corresponding pathologies.

The tremendous number of results obtained on the prion protein has served to firmly establish the concept of "protein only" introduced in 1982 by Prusiner who discovered the protein, and emphasized the role of protein conformational change and misfolding in human pathologies (for a review, see 56). Prion diseases may result from genetic, infectious or sporadic disorders, all involving the conformational change of the prion protein. The normal cellular protein PrP^c, whose function remains unknown, is converted into the pathological one PrP^{sc}, through a structural transformation in which parts of the α-helices of the native structure are converted into ß-strands. The three-dimensional structures of normal prion protein from hamsters, mice, and recently humans have been solved by NMR studies (56,57). Prions are transmissible particles devoid of nucleic acid and composed of PrP^{sc}. The data are consistent with a mechanism in which PrP^{sc} acts as a template for the conversion of nascent PrP^c into further molecules of PrP^{sc}. Many investigations have led to the determination of the amino acids involved in the species barrier, i.e., in the impossibility of transmission from a species to another (Figure 7). This unusual disease mechanism has shown how an aberrant conformational change in a protein can be propagated, and underlines the importance of protein folding for understanding the cause and propagation of a disease.

Table 1 - Amyloidogenic proteins and the corresponding diseases (according to Kelly (55)).

Clinical syndrome	Precursor protein	Fibril component
Alzheimer's disease	Amyloid protein precursor	ß-Peptide 1-40 to 1-43
Primary systemic amyloidosis	Immunoglobulin light chain	Intact light chain or fragments
Secondary systemic amyloidosis	Serum amyloid A	Amyloid A (76-residue fragment)
Senile systemic amyloidosis	Transthyretin	Transthyretin or fragments
Familial amyloid polyneuropathy I	Transthyretin	Over 45 transthyretin variants
Hereditary cerebral amyloid angiopathy	Cystatin C	Cystatin C minus 10 residues
Hemodialysis-related amyloidosis	$ß_2$-Microglobulin	$ß_2$-Microglobulin
Familial amyloid polyneuropathy III	Apolipoprotein A1	Fragments of apolipoprotein A1
Finnish hereditary systemic amyloidosis	Gelsosin	71-Amino acid fragments of gelsosin
Type II diabetes	Islet amyloid polypeptide (IAPP)	Fragment of IAPP
Medullary carcinoma of the thyroid	Calcitonin	Fragments of calcitonin
Spongiform encephalopathies	Prion	Prion or fragments thereof
Atrial amyloidosis	Atrial natriuretic factor (ANF)	ANF
Hereditary non-neuropathic systemic amyloidosis	Lysozyme	Lysozyme or fragments thereof
Injection-localized amyloidosis	Insulin	Insulin
Hereditary renal amyloidosis	Fibrinogen	Fibrinogen fragments

Protein folding in biotechnology: protein engineering and design

Two main applications concerning protein folding are involved in biotechnologies: protein engineering and the *de novo* design of proteins with novel functions. Recombinant proteins of pharmaceutical interest, such as growth hormone, insulin, and antihemophilic factor VIII, are commonly expressed and overproduced in *E. coli* or other cells. Since the overexpression of genes in foreign hosts often results in the formation of inclusion bodies, further treatments including unfolding and refolding are required. These proteins can also be modified by genetic engineering to increase their stability for storage which is an important industrial problem.

De novo protein design has recently emerged with the hope of constructing proteins with functions unprecedented in nature. This research is based on our understanding of the principles of protein folding. The conception of new proteins represents a burgeoning field of research. Genetic engineering provides a powerful methodology to redesign existing proteins. The aim of *de novo* protein design is to create completely new proteins with determined activities. The first step consists of choosing a function, and finding the amino acids with a favorable spatial arrangement capable of generating the desired function. Then it is necessary to find a polypeptide scaffold capable of supporting the reactive groups in the appropriate orientation. In the following step, it is necessary to determine an amino acid sequence able to fold into an adequate and stable three-dimensional structure, which presents the desired geometry of the binding site. Some folds such as triose phosphate

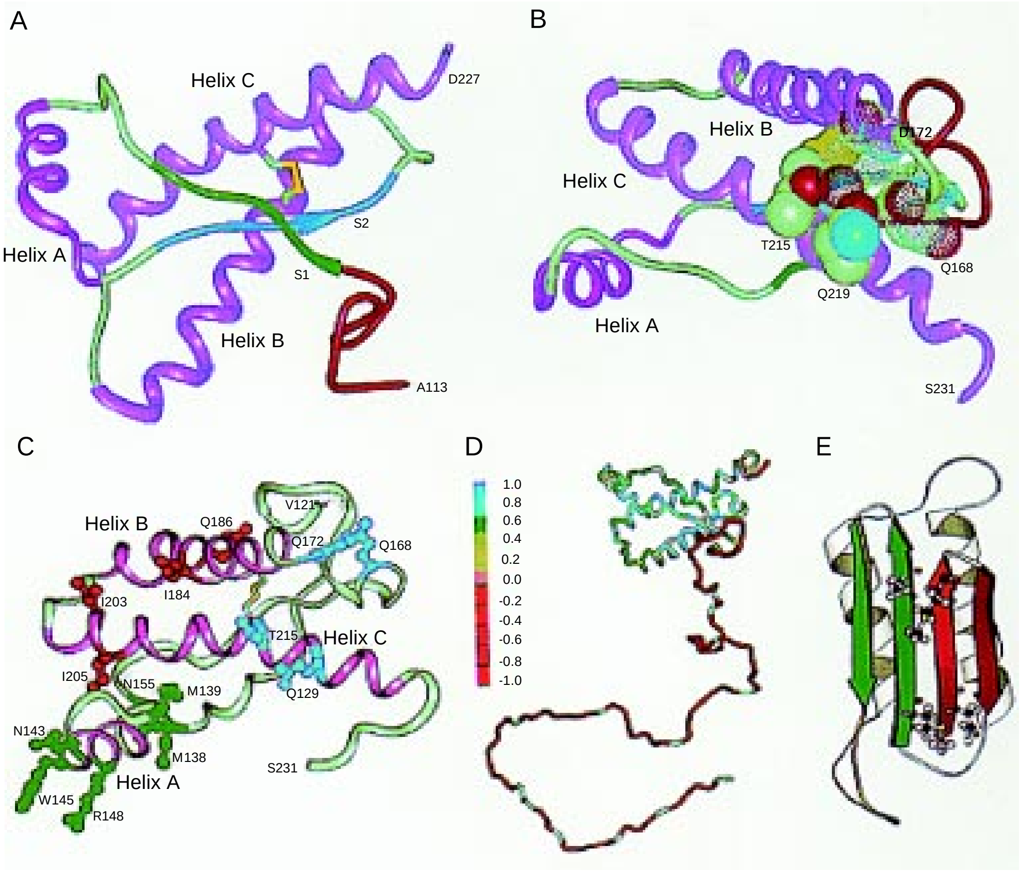

Figure 7 - Structure of prion proteins. *A*, NMR structure of hamster recombinant PrP (90-231); helices are colored in pink, the disulfide bond in yellow and conserved hydrophobic regions in red. *B*, NMR structure of recombinant PrP (90-231) indicating a van der Waals rendering of the residues thought to bind PrP^{C} with a protein X. *C*, PrP residues governing the transmission of prions across species are indicated. *D*, Schematic diagram of PrP (29-231) showing the flexibility of the N-terminal region. *E*, Plausible model for the tertiary structure of PrP^{Sc} (reproduced from Ref. 56 with permission of the author and the owner National Academy of Sciences, USA).

isomerase barrel, three- and four-helix bundles, and immunoglobulin fold appear frequently in proteins with highly divergent sequences. They are highly designable and may be easily modified without perturbing their three-dimensional structure.

At this stage, several different strategies may be used. One consists of using as a scaffold a protein of known structure with properties close to those desired. This is called the local conception. The other, the global conception, consists of the design of a structure by analogy with one of the classical folds in the protein data bank. However, since in general a large number of sequences can fold into the same three-dimensional structure, it is only necessary to arrive at one of them. Genetic methods can be used to screen a great number of randomized sequences to find those that fold into a given three-dimensional structure. Combinatorial computational algorithms provide a powerful complementary approach to genetic methods for exploring the sequence space. They consist of the exploration of a large number of side-chain combinations that can fit together to stabilize a given backbone fold and necessarily include a potential energy function. Automated design of functional proteins capable of generating a sequence compatible with the template fold and specific for some purpose is being developed. Once the sequence is conceived, the recombinant protein can be produced from the corresponding gene. Another strategy uses peptide synthesis, for example in the design of monomeric ß-sheets, or helical bundles. The different methods for *de novo* protein design have been reviewed by de Grado et al. (58).

Most accomplishments have concerned the synthetic design and structural characterization of secondary and supersecondary structures such as helices, helix bundles, ß-hairpins, and ß-sheets. They have widely contributed to the understanding of secondary structures in proteins. A large number of parallel or antiparallel helix bundles have been designed, resulting in the desired fold but showing a marginal stability, some of them displaying the characteristics of molten globules. Several designed helix-bundle peptides that adopt multiple conformations in solution have been crystallized in only one of these conformations. These motifs can serve as a starting point in protein design.

Functionalization of designed polypeptides has been successfully obtained in the field of catalysis, metal ion and heme binding, and introduction of cofactors (for a review, see 59). A 14-residue polypeptide that forms a bundle-like structure catalyzes the decarboxylation of oxaloacetate with a low catalytic activity, a cysteine residue acting as nucleophile. Hydrolysis and transesterification reaction of paranitrophenyl esters have been accomplished by designed four-helix bundles formed from 42-residue polypeptides in which histidine residues have been introduced. The successful design of a four-helix bundle protein that binds four heme groups with high affinity has been reported. The structure is well defined as shown by NMR spectroscopy.

A number of natural proteins have been redesigned with important changes in their sequence. The strategies generally used are based on genetic selection with the help of computational methods and the construction of consensus sequences. Phage display offers a powerful method to select the highest affinity binders.

It is clear that *de novo* protein design represents a growing field of research that will be useful both in testing the principles of protein folding and in offering the perspective to design new proteins with practical applications for pharmaceuticals and diagnostics.

I will conclude this review with a citation from Dobson, Šali and Karplus (45): “An understanding of folding is important for the analysis of many events involved in cellular regulation, the design of proteins with novel functions, the utilization of sequence information from the various genome projects, and the development of novel therapeutic strategies for treating or preventing human diseases that are associated with the failure of proteins to fold correctly."

Acknowledgments

The author thanks Charles Robert for carefully reading the manuscript and revising the English, and Charis Ghélis for helpful discussions.

References

1. Thornton JM, Orengo CA, Todd AE & Pearl FM (1999). Protein folds, functions and evolution. *Journal of Molecular Biology*, 293: 333-342.
2. Ghélis C & Yon JM (1982). *Protein Folding*. Academic Press, New York.
3. Dobson CM & Karplus M (1999). The fundamentals of protein folding: bringing together theory and experiment. *Current Opinion in Structural Biology*, 9: 92-101.
4. Anfinsen CB (1973). Principles that govern the folding of protein chains. *Science*, 181: 223-230.
5. Karplus M & Sali A (1995). Theoretical studies of protein folding and unfolding. *Current Opinion in Structural Biology*, 5: 58-73.
6. Kim PS & Baldwin RL (1990). Intermediates in protein folding reactions of small proteins. *Annual Review of Biochemistry*, 59: 631-660.
7. Fersht AR (1997). Nucleation mechanisms in protein folding. *Current Opinion in Structural Biology*, 7: 3-9.
8. Jaenicke R (1987). Folding and association of proteins. *Progress in Biophysics and Molecular Biology*, 49: 117-237.
9. Ptitsyn OB & Rashin AA (1973). Stagewise mechanism of protein folding. *Doklady Akademii Nauk SSSR*, 213: 473-475.
10. Wetlaufer DB (1981). Folding of protein fragments. *Advances in Protein Chemistry*, 34: 61-92.
11. Chothia C (1984). Principles that determine the structure of proteins. *Annual Review of Biochemistry*, 53: 537-572.
12. Karplus M & Weaver DL (1994). Protein folding dynamics: the diffusion-collision models and experimental data. *Protein Science*, 3: 650-668.
13. Dill KA (1985). Theory for the folding and stability of globular proteins. *Biochemistry*, 24: 1501-1509.
14. Harrison SC & Durbin R (1985). Is there a single pathway for the folding of a polypeptide chain? *Proceedings of the National Academy of Sciences, USA*, 82: 4028-4030.
15. Yon JM (1997). Protein folding: concepts and perspectives*. Cellular and Molecular Life Sciences*, 53: 557-567.
16. Wetlaufer DB & Ristow S (1973). Acquisition of the three-dimensional structure of proteins. *Annual Review of Biochemistry*, 42: 135-158.
17. Creighton TE (1974). Experimental studies of protein folding and unfolding. *Progress in Biophysics and Molecular Biology*, 33: 231-297.
18. Weissman JA & Kim PS (1992). Reexamination of the folding of BPTI: predominance of native intermediates. *Nature*, 336: 42-48.
19. Ghélis C (1980). Transient conformational states in proteins followed by differential labeling. *Biophysical Journal*, 32: 503-514.
20. Roder H, Elöve GA & Englander SW (1988). Structural characterization of folding intermediates in cytochrome c by H-exchange labeling and proton NMR. *Nature*, 335: 700-704.
21. Baldwin RL (1993). Pulse H/D exchange studies of folding intermediates. *Current Opinion in Structural Biology*, 3: 84-91.
22. Dobson CM (1991). Characterization of protein folding intermediates. *Current Opinion in Structural Biology*, 1: 22-27.
23. Matouschek A & Fersht AR (1991). Protein engineering in analysis of protein folding and stability. *Methods in Enzymology*, 202: 82-112.
24. Ballery N, Desmadril M, Minard P & Yon JM (1993). Characterization of an intermediate in the folding pathway of phosphoglycerate kinase; chemical reactivity of genetically introduced cysteinyl residues during the folding process. *Biochemistry*, 32: 708-714.
25. Garcia P, Desmadril M, Minard P & Yon JM (1995). Evidence for residual structures in the unfolded form of yeast phosphoglycerate kinase*. Biochemistry*, 34: 397-404.
26. Pecorari F, Minard P, Desmadril M & Yon JM (1993). Structure and functional complementation of engineered fragments from yeast phosphoglycerate kinase. *Protein Engineering*, 6: 313-325.
27. Ptitsyn OB (1995). Molten globule and protein folding. *Advances in Protein Chemistry*, 47: 83-229.
28. Ohgushi M & Wada A (1983). Molten globule state: a compact form of protein with mobile side-chains. *FEBS Letters*, 164: 21-24.
29. Chaffotte AF, Cadieux C, Guillou Y & Goldberg ME (1992). A possible folding initial intermediate: the C-terminal proteolytic domain of tryptophan synthase ß-chain folds in less than 4 milliseconds into a condensed state with non-native-like secondary structure. *Biochemistry*, 31: 4303-4308.
30. Shiraki K, Nishikawa K & Goto Y (1995). Trifluoroethanol induced stabilization of the α-helical structure of ß-lactoglobulin: implication for non-hierarchical protein folding. *Journal of Molecular Biology,* 245: 180-194.
31. Alexandrescu AT, Evans PA, Pitkeathly M, Baum J & Dobson CM (1993). Structure and dynamics of the acid-denatured molten globule state of α-lactalbumin: a two-dimensional NMR study. *Biochemistry*, 32: 1707-1718.
32. Radford SE, Dobson CM & Evans PA (1992). The folding of hen lysozyme involves partially structured intermediates and multiple pathways. *Nature*, 358: 302-307.
33. Uversky VN & Ptitsyn OB (1996). Further evidence on the equilibrium "pre-molten globule state": four-state guanidinium chloride unfolding of carbonic anhydrase B at low temperature. *Journal of Molecular Biology,* 255: 215-228.
34. Jeng MF & Englander SW (1991). Stable submolecular folding units in a non-compact form of cytochrome c. *Journal of Molecular Biology,* 221: 1045-1061.
35. Fink AL (1995). Compact intermediate states in protein folding. *Annual Review of Biophysics and Biomolecular Structure*,

24: 495-522.

36. Plaxco KW & Dobson CM (1996). Time-relaxed biophysical methods in the study of protein folding. *Current Opinion in Structural Biology*, 6: 630-636.
37. Eaton WA, Muñoz V, Thompson PA, Henry ER & Hofrichter J (1998). Kinetics and dynamics of loops, α-helices, ß-hairpins, and fast-folding proteins. *Accounts of Chemical Research*, 31: 745-753.
38. Hagen SJ, Hofrichter J, Szabo A & Eaton WA (1996). Diffusion-limited contact formation of an unfolded cytochrome c: estimating the maximum rate of protein folding. *Proceedings of the National Academy of Sciences, USA*, 93: 11615-11617.
39. McCammon JA (1996). A speed limit of protein folding. *Proceedings of the National Academy of Sciences, USA*, 93: 11426-11427.
40. Ballew RM, Sabelko J & Gruebele M (1996). Observation of distinct nanosecond and microsecond protein folding events. *Nature Structural Biology*, 3: 923-926.
41. Shastry MCR, Sander JM & Roder H (1998). Kinetic and structural analysis of submillisecond folding events in cytochrome c. *Accounts of Chemical Research*, 31: 717-725.
42. Kuwajima K (1996). The molten globule state of α-lactalbumin: a review. *FASEB Journal*, 10: 102-109.
43. Jaenicke R (1999). Stability and folding of domain proteins*. Progress in Biophysics and Molecular Biology*, 71: 155-241.
44. Dill KA, Bromberg S, Yue K, Fiebig KM, Yee DP, Thomas PD & Chan HS (1995). Principles of protein folding: a perspective from simple exact models. *Protein Science*, 4: 561-602.
45. Dobson CM, Šali A & Karplus M (1998). Protein folding: A perspective from theory and experiment. *Angewandte Chemie, International Edition in English*, 37: 868-893.
46. Wolynes PG, Onuchic JN & Thirumalai D (1995). Navigating the folding routes. *Science*, 267: 1619-1620.
47. Dill KA & Chan HS (1997). From Levinthal paradox to pathways to funnel*. Nature Structural Biology*, 4: 10-19.
48. Pecorari F, Minard P, Desmadril M & Yon JM (1996). Occurrence of transient multimeric species during the refolding of a monomeric protein. *Journal of Biological Chemistry*, 271: 5270-5276.
49. Silow M, Tan YJ, Fersht AR & Oliveberg M (1999). Formation of short-lived protein aggregates directly from the coil in two-state folding. *Biochemistry*, 38: 13006-13012.
50. Segel DJ, Eliezer D, Uversky V, Fink AL, Hodgson KO & Doniac S (1999). Transient dimer in the refolding kinetics of cytochrome c characterized by small-angle X-ray scattering. *Biochemistry*, 38: 15353-15359.
51. Baldwin RL (1994). Matching speed and stability. *Nature*, 369: 183-184.
52. Wang JD & Weissman JS (1999). Thinking outside the box: new insights into the mechanisms of GroEL-mediated protein folding. *Nature Structural Biology*, 6: 597-600.
53. Ellis RJ & Hartl FU (1999). Principles of protein folding in the cellular environment. *Current Opinion in Structural Biology*, 9: 102-110.
54. Yon JM (1996). Protein aggregation. In: Meyers RA (Editor), *Encyclopedia of Molecular Biology and Molecular Medicine.* Vol. V. VCH, Weinheim, 73-93.
55. Kelly JW (1996). Alternative conformations of amyloidogenic proteins govern their behavior. *Current Opinion in Structural Biology*, 6: 11-17.
56. Prusiner SB (1998). Prions (Nobel Lecture). *Proceedings of the National Academy of Sciences, USA*, 95: 13363-13383.
57. Zahn R, Liu A, Lühre T, Riek R, von Shroetter C, Wider G & Wüthrich K (2000). NMR solution structure of the human prion protein. *Proceedings of the National Academy of Sciences, USA*, 97: 145-150.
58. De Grado WF, Summa CM, Pavone V, Nastri F & Lombardi A (1999). *De novo* design and structural characterization of proteins and metalloproteins. *Annual Review of Biochemistry*, 68: 779-819.
59. Balzer L (1998). Functionalization of designed folded polypeptides. *Current Opinion in Structural Biology*, 8: 466-470.