A backbone-based theory of protein folding

George D. Rose[†‡], Patrick J. Fleming[†], Jayanth R. Banavar[§], and Amos Maritan[¶]

[†]T. C. Jenkins Department of Biophysics,The Johns Hopkins University, Jenkins Hall, 3400 North Charles Street, Baltimore, MD 21218; [§]Department of Physics, 104 Davey Lab, Pennsylvania State University, University Park, PA 16802; and [¶]Dipartimento di Fisica ''G. Galilei'' and Istituto Nazionale di Fisica Nucleare, Università di Padova, Via Marzolo 8, 35131 Padova, Italy

Edited by Jeremy Nathans, Johns Hopkins University School of Medicine, Baltimore, MD, and approved September 21, 2006 (received for review August 15, 2006)

Under physiological conditions, a protein undergoes a spontaneous disorder ⇌ order transition called "folding." The protein polymer is highly flexible when unfolded but adopts its unique native, three-dimensional structure when folded. Current experimental knowledge comes primarily from thermodynamic measurements in solution or the structures of individual molecules, elucidated by either x-ray crystallography or NMR spectroscopy. From the former, we know the enthalpy, entropy, and free energy differences between the folded and unfolded forms of hundreds of proteins under a variety of solvent/cosolvent conditions. From the latter, we know the structures of ≈35,000 proteins, which are built on scaffolds of hydrogen-bonded structural elements, α-helix and β-sheet. Anfinsen showed that the amino acid sequence alone is sufficient to determine a protein's structure, but the molecular mechanism responsible for self-assembly remains an open question, probably the most fundamental open question in biochemistry. This perspective is a hybrid: partly review, partly proposal. First, we summarize key ideas regarding protein folding developed over the past half-century and culminating in the current mindset. In this view, the energetics of side-chain interactions dominate the folding process, driving the chain to self-organize under folding conditions. Next, having taken stock, we propose an alternative model that inverts the prevailing side-chain/backbone paradigm. Here, the energetics of backbone hydrogen bonds dominate the folding process, with preorganization in the unfolded state. Then, under folding conditions, the resultant fold is selected from a limited repertoire of structural possibilities, each corresponding to a distinct hydrogen-bonded arrangement of α-helices and/or strands of β-sheet.

Proteins are linear, unbranched polymers of amino acid residues that can undergo a reversible disorder ⇌ order transition called protein folding. Under suitable conditions, all of the information needed to realize the ordered form of most proteins is encoded in their linear sequence; no auxiliary components are necessary to guide the disordered chain to its unique, biologically relevant three-dimensional structure (1). In water with a little salt and at physiological temperature, most proteins self-assemble spontaneously.

In fact, most biological components self-assemble spontaneously, apart from the three main template-driven process (replication, transcription, and translation). Larger assemblies, such as the ribosome, self-assemble from smaller composites iteratively, a top-down structural hierarchy that terminates ultimately with protein monomers, which assemble themselves. Life is rooted in self-assembly processes, and we seek to explain them, starting with protein folding. However, as Goethe said, "The hardest thing to see is what is in front of our eyes."

This perspective attempts to lay bare the premises that motivated current thinking by tracing their development during the past half-century of research. Then, we question the current mindset and propose a radically different interpretation of the known facts. In brief, it is widely accepted that side-chain interactions are primarily responsible for conformational differences among proteins because residue backbones are chemically equivalent and, therefore, are lacking in discriminatory power. Contrary to this plausible idea, we propose that, in fact, the backbone is primarily responsible for determining the fold because peptide hydrogen bonds dominate the folding process. Even one or two unsatisfied hydrogen bonds in the molecular interior would counterbalance the entire free energy of folding for a typical globular protein. Of course, other factors also favor or disfavor the folded state, but backbone hydrogen bonding outweighs them all. Here, we distinguish between the fold, a scaffold of α-helices and β-strands interconnected by tight turns and loops, and the detailed atomic structure that is elaborated upon this molecular skeleton. For single-domain proteins, only a limited number of scaffolds are possible; others are excluded by steric impossibility and/or the lack of hydrogen bond satisfaction. This backbone-based theory derives support from the solution thermodynamics of protecting osmolytes that promote folding by exerting their effect primarily on the backbone, not the side chains. We now develop these ideas in detail.

The conclusion that proteins can self-assemble spontaneously is based on Anfinsen's Nobel prize-winning experiments showing that the protein ribonuclease can be reversibly denatured/renatured in a test tube (2). Both structure and biological activity are abolished under denaturing conditions but restored spontaneously upon return to physiological conditions. Since that time, variations on the Anfinsen experiment have been repeated successfully for hundreds if not thousands of other proteins. This spontaneous folding transition, from a less-ordered population to a more ordered population, begs explanation.

Anfinsen's own explanation was the thermodynamic hypothesis, which postulates that under physiological conditions the protein population attains a minimum in Gibbs free energy in its native state. In this view, each individual molecule in a protein solution can assume an astronomical number of conceivable conformations under unfolding conditions. Upon shifting to folding conditions, the entire population is driven spontaneously toward the conformation that optimizes the protein's interactions with both itself and its solvent environment. In other words, the folding transition is a consequence of the spontaneous drive to minimize the chemical potential. Gibbs devised the chemical potential to be exactly analogous to other expressions of potential energy, such as the electrical potential, in which current flows spontaneously between two poles so as to minimize any difference in voltage levels. Subsequent interpretations notwithstanding, the thermodynamic hypothesis is simply a statement that

Author contributions: G.D.R., P.J.F., J.R.B., and A.M. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS direct submission.

[‡]To whom correspondence should be addressed. E-mail: grose@jhu.edu.

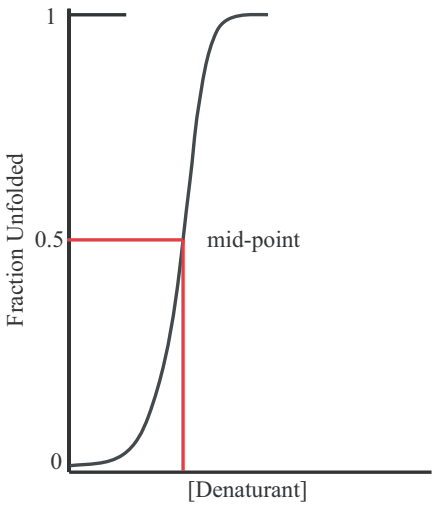

Fig. 1. The folding transition. Many small proteins of experimental interest fold with high cooperativity so that a plot of some structure-disrupting factor, like temperature or a chemical denaturant, against the folded fraction of the population results in a sigmoidal curve. At the transition midpoint, 50% of the ensemble is folded and 50% is unfolded; the population of partially folded molecules is negligible. In this idealized plot of an actual experiment, the population is followed by a conformational probe (e.g., circular dichroism) as a function of denaturant concentration. Upon addition of sufficient denaturant, the probe signal reaches a plateau, indicating that the transition is complete. In experiments using multiple conformational probes (e.g., circular dichroism and fluorescence), all indicators trace the same sigmoidal curve after suitable normalization (6). Thus, one refers to the folding transition, not the circular dichroism folding transition, the fluorescence folding transition, etc.

proteins fold to their native state like a ball rolling down a free energy hill.

Anfinsen framed protein folding in explicit thermodynamic terms, as had both Wu (3) and Mirsky and Pauling (4) many years earlier. In retrospect, this approach still seems entirely appropriate. Many small proteins of experimental interest fold in an all-or-none manner (5). A plot of some structure-disrupting factor, like temperature or a chemical denaturant, against the folded fraction of the population results in a sigmoidal (i.e., highly cooperative) curve (6) (Fig. 1). At the curve's midpoint, half the population is folded, half is unfolded, and the population of partially folded intermediates is negligible. Such behavior is a disappointment to the chemist who seeks to track the reaction by monitoring a succession of intermediate states. But it is a simplifying windfall for the thermodynamicist, who now can represent the folding process as a valid chemical equilibrium, U(nfolded) $\rightleftharpoons$ N(ative), with equilibrium constant $K_{eq} = [N]/[U]$, for which the free energy difference between the folded and unfolded populations is given by $\Delta G^0_{conformational} = -RT \ln K_{eq}$ (R is the gas constant; T is the absolute temperature). $\Delta G^0_{conformational}$ has been measured for hundreds of proteins, and typical values fall within a narrow range between -5 to -15 kcal/mol (7).

Surely, equilibrium thermodynamics is our most powerful discipline for understanding biological systems. However, thermodynamic descriptions are deliberately mechanism-independent. Anfinsen's thermodynamic hypothesis underscores his key observation that under suitable conditions, no input of energy (e.g., from metabolism) is needed to drive the folding reaction, U $\rightleftharpoons$ N. Yet, the hypothesis is silent on the question of how individual molecules negotiate routes from the unfolded state to the folded state. Experimental approaches to this question have been confounded by the small free energy difference that separates folded from unfolded populations; a difference of -5 kcal/mol is the energy equivalent of a single hydrogen bond. Today, we still lack a general theory that can successfully predict a protein's fold from its sequence *a priori* or, for that matter, that can reliably predict whether an arbitrary sequence will fold.

At an even more basic level, recent theoretical work implicitly raises the question of whether protein systems are compatible with such a theory. A fundamental distinction can be made between two extreme types of systems in statistical thermodynamics, described by Baldwin as the "classical view" and the "new view" (8). In the classical view, the behavior of the population is dictated by a small number of equilibrium states, and in this case, a general predictive theory is feasible. In the contrasting new view, the population of interest is distributed at random across a complex energy landscape, a condition that resists generalization. In this latter case, each protein will, of necessity, fold in its own unique way if, indeed, it folds at all. These contrasting concepts go to the very core of the research directions that have informed the field.

This seven-part perspective seeks to illuminate such issues. We argue that a predictive molecular theory of protein folding should be possible. If so, why has it been so long in coming? And what form would it take? What do we mean by a protein fold? We begin by describing current thinking in the field (*Part 1*). We then attempt to diagnose the main conceptual impediments to progress (*Part 2*), explore some lessons from nature (*Part 3*) that prompt a reassessment (*Part 4*), and an alternative, backbone-based folding model (*Part 5*). At this point, it becomes apparent that an ambiguous use of the word "fold" has hindered understanding and requires clarification (*Part 6*). Finally, we place the backbone-based model in a larger perspective (*Part 7*). A *Supporting Appendix*, which is published as supporting information on the PNAS web site, summarizes a complementary physics-based picture, in which we suggest that proteins may occupy a novel phase of matter.

Part 1. Protein Folding: The Current Perspective

Research has been directed at both sides of the folding reaction, U $\rightleftharpoons$ N. Remarkably, at this writing there are $\approx$35,000 protein structures in the Protein Data Bank (www.rcsb.org), solved at near-atomic resolution by either x-ray crystallography or NMR spectroscopy. This wealth of data has transformed protein chemistry since the early pioneering efforts of Bernal and Crowfoot (9) and Perutz (10). The availability of a structure has made a telling difference in countless studies of biologically important molecules. In addition, structure-based programmatic initiatives now are commonplace, including, for example, a diversity of database analyses (e.g., ref. 11), taxonomic classification at the molecular level (12, 13), estimates of the number of folds (14), and pattern recognition-based approaches to prediction (15). At this point, N rests on firm ground.

Turning now to U, in the prevailing view, the population is distributed at random across a featureless energy landscape under denaturing conditions. Such an intrinsically disordered population is incommensurate with structural characterization. Consequently, the field has resorted to statistical characterization, using concepts from polymer theory.

The term "unfolded protein" is generic and can range from protein solutions in harsh denaturants to protein subdomains that undergo transitory excursions from their native format via spontaneous fluctuations (16). This range is too diverse to be practically useful, and the field has focused more specifically on denatured proteins, the population of unfolded conformers that can be studied at equilibrium under high concentrations of denaturing solvents, high temperature, high pressure, and high/low pH.

In a denaturing solution, the chain paths described by individual protein molecules are thought to be well approximated by self-avoiding random walks. More precisely, the denatured chain behaves like a statistical coil, de-

scribed by the Flory rotational isomeric state model (17), which takes into account constraints on bond rotation imposed by covalent chemistry. In any case, the number of conceivable paths for even a small protein of 100 residues is of order at least 10^{30} and possibly much larger (18). At every time slice, each molecule in the population will have a specific conformation, but with only small energy barriers between them (approximately kT; k is the Boltzmann constant; T is the absolute temperature), so conformations are readily interconvertible.

Accordingly, the structure of any single molecule would not represent the population in any meaningful sense. However, it is possible to measure the degree to which molecules in the population are expanded or contracted, as given by their radius of gyration, R_G, the rms distance of atoms from their common center of gravity:

$$R_G = \sqrt{\frac{\sum_{i=1}^{N} R_{Gi}^2}{N}}, \qquad [1]$$

where R_{Gi} is the distance of atom i from the center of gravity and N is the number of atoms in the molecule (19). The population then can be characterized by its average radius of gyration, which can be determined experimentally under conditions of interest (20, 21).

Flory (17) provided a simple relationship between these coil dimensions and solvent quality. For a statistical-coil polymer with excluded volume, the radius of gyration, R_G, is given by:

$$R_G = R_0 n^{\nu}, \qquad [2]$$

where R_0 is a constant that depends on intrinsic chain stiffness, n is the number of residues, and ν is the exponent of interest that depends on solvent quality. Values of ν range from 0.33 for a collapsed molecule, like a folded protein, to 0.6 for a self-avoiding random walk, like a denatured protein.

Is this theory valid for denatured proteins? Persuasive evidence was provided by Tanford (20), who demonstrated that typical proteins denatured in 6 M guanidinum chloride (a strong denaturant) behave as structureless, statistical coils, with mean radii of gyration that are consistent with theory. Tanford's corroborating studies established a compelling framework for interpreting experimental protein denaturation.

The current view, encapsulated in the compact equation U $\rightleftharpoons$ N, has been developed over the past 40 years or so. Summarized in a sentence: individual molecules are distributed across a vast, undifferentiated energy landscape under denaturing conditions but adopt a unique native conformation spontaneously under folding conditions. At least in a general outline, the folding picture seems to be complete.

The Search Problem. An inescapable search problem is deeply embedded in this view of the folding reaction, one that has stimulated the field since it was first made apparent in a famous back-of-the-envelope calculation (22) that came to be known as the "Levinthal paradox." In a nutshell, how can an unfolded polypeptide chain that is free to sample the vastness of conformational space discover the native conformation in biological real-time after a shift to folding conditions?

In greater detail, Corey and Pauling (23) demonstrated that the peptide bond has partial double-bond character, and, therefore, the six backbone atoms in the peptide unit (-Cα-CO-NH-Cα-) are coplanar, or largely so. Consequently, there are only two primary degrees of freedom in each peptide unit, parameterized by Ramachandran *et al*. (24) as the two torsion angles, ϕ and ψ (Fig. 2*A*). Further, Ramachandran and Sasisekharan (25) showed that only a small subset of these torsions result in clash-free configurations (Fig. 2*B*); other values would experience stiff repulsive forces between the electron clouds of nonbonded atoms within the peptide unit.

In Levinthal's original estimate (22), there are three staggered configurations per torsion, nine (3×3) conformers per peptide unit and, therefore, $9^{100} \approx 10^{95}$ conformers for a 100-residue protein. With a subpicosecond speed limit for bond rotations, the universe would end before chains could encounter the native conformation via an unguided search. Of course, this calculation was oversimplified for dramatic effect (see e.g., ref. 26.

However, proteins are known to fold in the microsecond to millisecond range (27), so even the addition of more realistic constraints cannot explain away the underlying search problem. Under denaturing conditions, the number of conceivable conformations far exceeds the number of actual molecules in a dilute protein solution, and, in the extreme, every molecule might have a different conformation. Whereupon, for a protein with a typical stability of -10 kcal/mol, on average all but one molecule in 17 million adopt the native fold after shifting to folding conditions. This transition from the unfolded population to the folded population can be completed in

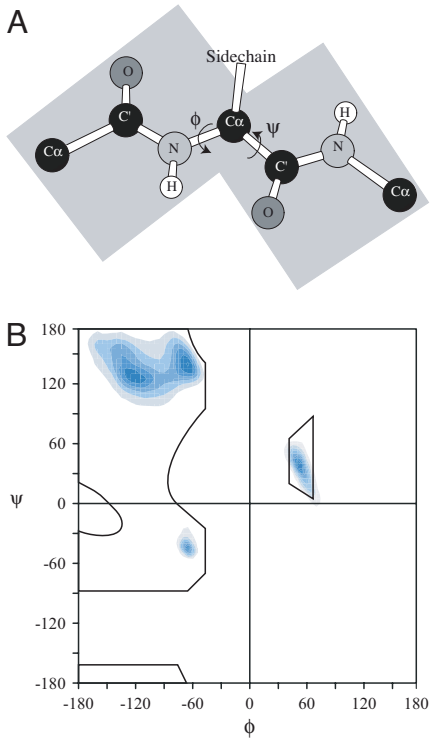

Fig. 2. The peptide unit. (*A*) Degrees of freedom. The peptide bond, C′-N, has partial double-bond character (23), so the six backbone atoms, -Cα-C′O-NH-Cα-, in the peptide unit (shaded rectangles) will be approximately coplanar. Consequently, there are two primary degrees of freedom in each peptide unit, the two torsion angles, ϕ and ψ (24). Assuming complete independence of these angles, there would be three staggered configurations per torsion, $3 \times 3 = 9$ conformers per peptide unit, and $9^{100} \approx 10^{95}$ conformers for a 100-residue protein. (*B*) Residue ϕ,ψ distributions. Sterically allowed ϕ,ψ regions for the alanyl dipeptide, from model studies of Ramachandran and Sasisekharan (25), are shown in dark outline. Other regions are predicted to be unpopulated because their backbone torsion angles would cause a steric clash within the dipeptide unit. ϕ,ψ distributions of experimental data from the major populated regions from the coil library (88) are shown superimposed on the predicted sterically allowed regions.

microseconds in some proteins (27). Estimates like this paint a paradoxical picture in which the ostensible magnitude of conformational space is so large that the native conformation could not be discovered in microseconds, yet it is.

The Levinthal paradox adds a temporal dimension to the basic conundrum wherein an ordered population emerges spontaneously from a disordered population, and we are still left seeking an explanation.

The Folding Funnel. For Levinthal, his back-of-the-envelope calculation was not a paradox at all; rather, it was a vivid demonstration that the native state is attained via a directed search, but how?

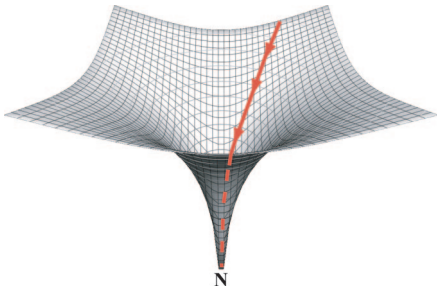

Fig. 3. A folding funnel. The funnel landscape depicts protein folding as a process that proceeds from a high entropy, disorganized state lacking in intramolecular interactions (mouth), to a low entropy, organized state with native intramolecular interactions (spout). Evolution has selected sequences that avoid frustrating traps en route from mouth to spout, smoothing what might otherwise be a rugged landscape. Under folding conditions, individual molecules can follow any route from mouth to spout, like a ball rolling down a free energy hill. One such trajectory is shown here. For a gallery of variant funnel landscapes, see ref. 126.

Wolynes and colleagues (28, 29) proposed that the overall energy landscape is funnel-shaped under folding conditions (Fig. 3). Regardless of where on this surface any particular molecule may happen to be, it follows the gradient toward a lower energy state, like a ball rolling down a free energy hill. Wending its way downhill, the protein accumulates favorable interactions that lower the energy, thereby promoting chain compaction and reducing the search volume. On a funnel landscape, molecules are driven toward the native state reliably, cooperatively, and with gathering speed as the U $\rightleftharpoons$ N folding reaction progresses.

But there's another problem. Why doesn't the ball get stuck in a ditch en route to the bottom of the hill (i.e., in a kinetic trap)? Clearly, that doesn't happen in known cases where proteins fold to completion on rapid timescales (27).

Restating the issue in physical-chemical terms, one might expect that a folding polypeptide chain, with its heterogeneous amino acid sequence, would inevitably encounter favorable but nonnative interactions at metastable energy levels. Such frustrating interactions would etch a rugged free-energy landscape, with multiple minima and greater-than-kT energy barriers between them. To rationalize this situation, Bryngelson and Wolynes (30) imported a key concept from physical systems: the idea of a spin glass.

A spin glass, originally introduced in the context of frustrated, random magnetic systems by Anderson (31), is an important paradigm in physics for many classes of problems. Classic optimization problems, such as the optimal placement of circuit elements on a chip and the determination of the best route of a traveling salesman, models of content-addressable memories in the brain, and models of prebiotic evolution, all require an energy or fitness landscape with multiple minima and barriers between them. As pointed out by Anderson (32), such a landscape confers both stability and diversity: stability because each minimum is locally stable and diversity because there are multiple minima in such systems. It is notoriously difficult to find the true ground state of a spin glass, owing to the rugged nature of the energy/fitness landscape.

As it folds, a protein is stabilized by a large number of weak (i.e., noncovalent) interactions and can visit numerous, sequence-dependent minima, a classic spin-glass landscape. How then does the protein fold rapidly and reproducibly within such a landscape but evade frustrating, glassy behavior? In Wolynes' model, proteins avoid such metastable traps because accessible interactions are selected by nature to be minimally frustrating, resulting in smoother funnel walls, more akin to an unfrustrated ferromagnet than a spin glass.

What forces or factors can reduce the ruggedness of funnel walls? In an extreme constructed example of a minimally frustrated system devised by Go (33), native contacts are assumed to interact favorably, whereas nonnative contacts do not interact at all. In the more realistic case, the landscape is shaped by evolutionary pressure to select those amino acid sequences that minimize the energy of the native fold while avoiding potentially frustrating alternatives as well.

The spin-glass model has spread from condensed matter physics into many disparate fields. Imported into protein chemistry as the funnel model, it provides an answer to the Levinthal paradox and is in satisfying accord with statistical thermodynamics (i.e., the new view); no special equilibrium states need be invoked.

The Funnel Landscape Is Explicitly Sequence-Dependent. The funnel model describes the behavior of a population of proteins of identical sequence as they wend their way downhill from U to N under folding conditions. Every unique sequence has its own funnel. For example, the globin fold is attained by thousands of different known globin sequences, many of which have only a small fraction of their residues in common (34). Each such globin sequence is associated with its own characteristic folding funnel. All globin sequences are presumed to have evolved so as to adopt the globin fold and to maintain similar overall structural characteristics (35), while simultaneously avoiding frustrating traps and dead ends in transit. The need to avoid unintentional, impeding interactions has long been recognized in protein-design research, where it is called "negative design" (36, 37). Neither folding theorists nor protein designers can ignore the inadvertent pitfalls of frustration.

Part 2. Questioning the Current Perspective: A Tale of Two Landscapes

The current view of folding is grounded in an explicit, amino acid sequence-dependent funnel landscape, as just described. However, the population also is regulated by a second, structure-dependent but sequence-indifferent landscape, although only by implication. Both landscapes impose major constraints on any sequence-dependent folding model.

The Sequence-Indifferent Landscape. As depicted above, the unfolded free energy landscape is vast and featureless. Most proteins unfold under rather similar conditions of temperature or denaturant concentration, consistent with $\Delta G^0_{\text{conformational}}$ values within the typical range of -5 to -15 kcal/mol (7). In other words, the unfolded free energy landscape is sequence-indifferent because structure is abolished under approximately the same conditions, regardless of sequence.

Given the independence of backbone torsion angles under unfolding conditions (17), this landscape spans all conceivable conformations of the polypeptide chain, including all possible native folds and subfolds. To be specific, under unfolding conditions, a lysozyme molecule could happen upon the ribonuclease fold. Although the likelihood of such an encounter is negligibly small, it is essentially no smaller than the probability that the molecule would chance upon its own native fold.

Upon shifting to folding conditions, distinct minima must ultimately emerge from this previously featureless landscape, each corresponding to a stable domain (i.e., a simple fold of $\approx$100 residues) (14, 38–41).

To see this situation, consider the shift from U to N. Here, it is important to realize that the folding reaction, U $\rightleftharpoons$ N, is not an ordinary chemical reaction; no covalent bonds are made or broken. For individual proteins, the reaction is all-or-none: proteins are either folded or unfolded, with a negligible population of partially folded intermediates, as noted above. For a population of proteins, the folded fraction is simply dialed up or

down in response to physical and/or chemical conditions (temperature, pressure, solvation, pH, etc.). Just as most proteins unfold under similar conditions, they also fold under similar conditions (although extremophiles that function at either very high or very low temperatures represent separate classes), regardless of sequence. Remarkably, upon switching to conditions that favor N, the repertoire of stable minima will emerge coordinately from the undifferentiated terrain of the unfolded landscape.

A simple calculation is sufficient to show that there are only a few thousand stable folded domains that correspond to these minima. Proteins are constructed from segments of α-helix and β-strand (42, 43), interconnected by turns (44) and loops (45). A typical domain of $\approx$100 residues might contain $\approx$10 such segments, either helix or strand, resulting in $2^{10} = 1{,}024$ possible constructs, amplified by the complexity of their interconnecting turns and loops, which are often restrictively short. This simple calculation is in good agreement with other estimates ranging between 10^3 and 10^4 distinguishable domains (14, 46), with larger proteins being modular constructs of these fundamental folds (12).

Accordingly, it can be concluded that the free energy landscape is presculpted (47, 48) into a few thousand intrinsically stable domains which are unmasked upon switching from unfolding to folding conditions. Specific sequences would be needed to discriminate among the possibilities in this repertoire, and exactly how this happens remains to be explained.

Embedding the Sequence-Dependent Funnel Landscape in the Sequence-Indifferent Structure Landscape. A folding protein must satisfy the dual constraints imposed by the absence of frustration within its own private funnel and, at the same time, those that stabilize its fold relative to other possible domains. A sequence that minimizes frustration cannot come at the expense of those interactions needed to guide the resultant fold to a unique stable domain. Given the limited repertoire of available domains, this restriction is especially stringent. Stable domains are built on scaffolds of α-helices and β-strands (42, 43), and it is plausible that minor changes in sequence could tip the energy balance, redistributing the population across multiple domains simultaneously. Were this dilemma to occur, a sequence would no longer have a unique conformation under physiological conditions, in violation of a central tenet of both protein biochemistry and life itself.

In the context of a funnel landscape, Anfinsen's hypothesis, that proteins fold by rolling down the equivalent of a free energy hill, comes with a significant, unsuspected negative design constraint. How can proteins fold in an inherently glassy landscape, avoid frustrating traps, and reach a unique minimum? The usual answer is as follows: given the 20 natural amino acids, there are 20^{100} possible sequences for a short 100-residue chain, a more-than-astronomical number. Although the constraints may seem overwhelming, sequence space is more than equal to them. In this view, a biologically viable protein sequence is the successful end-product of a trial-and-error experiment, conducted by evolution at the molecular level. To borrow a phrase from Darwin, "there is grandeur in this view of" proteins, but, grandeur notwithstanding, there are no general folding rules.

Folding Proteins One at a Time. Two corollaries are implicit in a funnel landscape. Corollary 1: Residue side chains are primarily responsible for such discrimination because all backbones (except glycine and proline) are chemically equivalent and, therefore, lacking in discriminatory power. Corollary 2: If the unfolded state is featureless, then all structural discrimination necessarily takes place under folding conditions. In essence, every protein sequence is a universe unto itself, folding via a constellation of detailed side-chain interactions that accumulate structural definition en route from U to N.

Recapitulating the current framework, under unfolding conditions, the energy landscape is featureless. A shift to folding conditions unmasks several thousand energy minima, each corresponding to a stable domain. Evolution selects for protein sequences that can overcome potential frustration and populate these domains uniquely without spilling over into other competing stable structures. Smaller proteins correspond to individual domains, and larger proteins are modular constructs of these fundamental folds. With 20^N possible sequences, an N-residue protein can satisfy these multiple constraints simultaneously.

An underappreciated aspect of the funnel landscape is that almost all conclusions are based on studies using a Go model, often in lattice simulations. Such approaches enjoy the benefits of simplicity. But, as mentioned, the unrealistic Go model was deliberately contrived to be a minimally frustrated system in which nonnative contacts make no contribution. To our knowledge, there is only one experimentally determined free-energy landscape in the literature at present (49), and it is not consistent with a funnel landscape, although some may disagree (50). Questions raised in *Part 2* prompt us to look to nature for further guidance.

Part 3. Lessons from Nature: Organic Osmolytes

Organic osmolytes are ubiquitous in living systems. These small organic compounds affect protein stability dramatically, and nature utilizes them to counteract the adverse effects of physical and chemical factors that might otherwise promote denaturation. Examples include desiccated plant seeds in desert conditions that can remain viable for centuries, animals that function at extremes of pressure in the deep ocean, and even proteins in human kidney that resist denaturation despite high urea concentration. Such phenomena are not mere curiosities, but rather they reflect a vital adaptive mechanism that makes life possible (51).

An organic osmolyte is a small molecule that affects protein stability. In the equilibrium protein-folding reaction, $N \rightleftharpoons U$, protecting osmolytes push the equilibrium toward N, whereas denaturing osmolytes push the equilibrium toward U. Examples of protecting osmolytes include several amino acids, trimethylamine *N*-oxide, glycerol, and many sugars. Urea, a denaturing osmolyte found naturally in mammalian kidney, has been a key reagent throughout the long history of solvent denaturation studies (52). It has been a mystery how such compounds could affect diverse proteins in similar ways, but recently it was shown that the predominant osmolyte effect is on the unfolded state (53) and it is exerted primarily on the backbone (54), which is the component in common to all proteins.

The osmolyte effect is universal throughout all three kingdoms of life, operating on proteins in general (51). Protein molecules do not have explicit built-in binding sites for osmolytes, and, therefore, this universal ability of osmolytes to modulate folding is tantamount to an existence proof: A universal folding mechanism must exist. Significantly, the mode of osmolyte action is the exact opposite of the two corollaries in the previous section: Osmolytes operate on U, not N, and their primary effect is on the backbone, not side chains.

Universality Implies a Backbone-Based Mechanism. The osmolyte effect has far-reaching implications for protein folding. Previously, it seemed plausible that the folding mechanism would depend

solely on those chemical components that differ from one protein to another, i.e., their side chains. Contrary to this supposition, the osmolyte effect reveals the existence of a universal mechanism. Intuitively, the most likely way to realize such a mechanism is via the protein backbone (55), which is component in common to all residues.

Other Factors also Suggest a Backbone-Based Mechanism. Other aspects of protein folding also point to a backbone-associated mechanism. As mentioned above, stable domains are built on scaffolds of α-helices and β-strands (42, 43). These two backbone structures are unique: they can be extended indefinitely without steric interference, and the resultant structures satisfy their own peptide hydrogen bonds (intrastructure hydrogen bonds in the case of α-helices and β-hairpins or interstructure hydrogen bonds in the case of β-sheet). No other choice of backbone torsion angles has these two properties for L-amino acids (25).

In further detail, proteins are organized as a structural hierarchy in which large contiguous-chain regions can be iteratively decomposed into smaller contiguous-chain regions (Fig. 4; refs. 56 and 57). Plausibly, this top-down architecture is the consequence of a bottom-up self-assembly process (40). Repeated values of backbone torsion angles from the two major sterically allowed dipeptide regions (Fig. 2*B*) generate the two hydrogen-bonded scaffold structures: α-helix and β-strand. These two structural elements can interact favorably in all combinations, giving rise to familiar supersecondary structure assemblies: $\alpha\alpha$, $\beta\beta$, and $\beta\alpha\beta$ (42, 43).

In the hierarchic process described here, self-organization takes place at a backbone level, with side chains playing the more limited role of selecting among helix, strand, or neither one. Providing that intramolecular peptide hydrogen bonds are favored over peptide-water hydrogen bonds, these hydrogen-bonded backbone structures will be energetically favorable. Consequently, these structural composites will become increasingly stable as their complexity grows, stemming from the covalent level: Energetically favorable dipeptide regions generate marginally stable elements of secondary structure that, in turn, associate to form yet more stable supersecondary structure assemblies. In other words, as the ball rolls down the free energy hill, it pushes the protein toward greater self-organization. Notably, hierarchic self-assembly of backbone segments does not engender a glassy landscape, and frustration is not a concern.

Small differences notwithstanding, typical globular proteins fold and unfold under approximately the same conditions, as discussed above. This experimental observation points once again to the existence of a general folding mechanism. Proteins are highly compact under folding conditions, with an average packing density resembling that of small organic solids (58). It has been shown that the conformational freedom of a compactly folded tube of polypeptide chain dimensions is severely limited by excluded volume restrictions in the marginally compact phase (47, 59). Such chain organization is an automatic consequence of compact geometry and does not result in a glassy landscape.

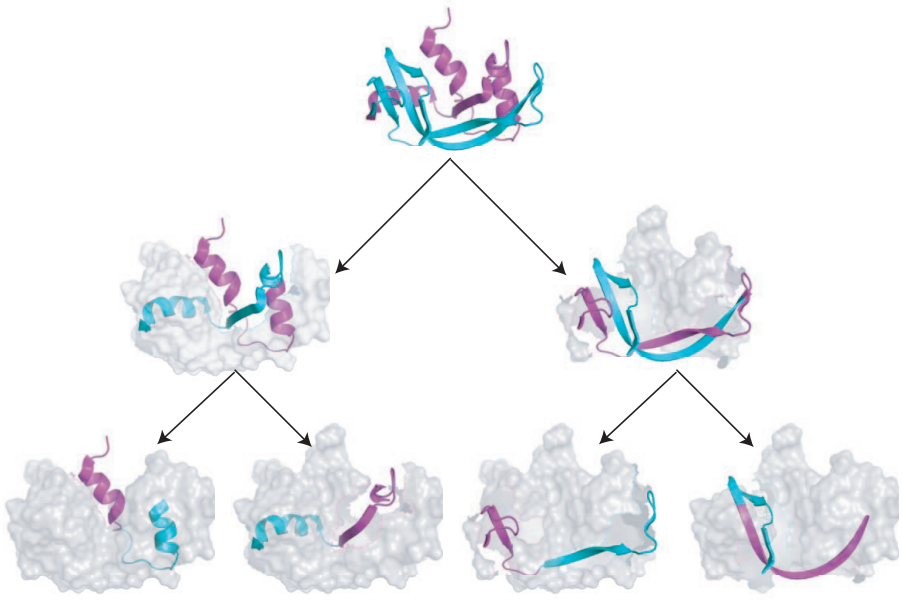

Fig. 4. Hierarchic organization of proteins. In a hierarchy, each component is contained within the next larger component, like a series of nested boxes. Hierarchic architecture, illustrated here for ribonuclease, can be verified easily for any protein of known structure by using a simple procedure: display the structure with the first N/2 residues in magenta and the remaining N/2 residues in cyan. Then repeat this procedure, iteratively. It is apparent at a glance that at each successive level of the hierarchy, magenta and cyan regions do not intermingle. (Surface shown in gray is a place holder.) This top-down, hierarchic architecture is an experimental fact, and no hypothesis is needed to extract this result from known structures. Hierarchy suggests a bottom-up folding mechanism (40, 56, 57) in which chain segments form local structures of marginal stability, which then interact iteratively to produce intermediates of ever-increasing complexity. In this process, multiple folding routes coexist, and the stabilities of intermediates and their combinatorial associations determine the dominant pathways (49). Picture was rendered by using Pymol (127).

Part 4. Reassessment

Our current perspective tracks back to Anfinsen's hypothesis that proteins fold so as to minimize free energy. To characterize this hypothesis: under folding conditions, each protein slides down an energy gradient to the conformation that optimizes its constellation of favorable interactions, the native state. For example, were the lysozyme sequence to adopt the ribonuclease fold, it would have a high energy, but it has a low energy in the lysozyme fold; lower, in fact, than in any other conceivable fold. In this exquisitely energy-sensitive conception, stability alone determines native fold.

What Anfinsen Could Not Have Known. However, Anfinsen could not have foreseen that there are only a few thousand stable domains, a crucial realization that changes the very nature of the folding problem. The millions of folding-competent sequences in the biosphere need not solve this conventional formulation of the folding problem. On the contrary, a given sequence need only discriminate among a few thousand potential possibilities to fold correctly (48). Changes in stability that are sufficient to dislodge the protein from its native basin will promote unfolding, not alternative folding. Surprisingly, this conception of folding leads to the conclusion that protein conformation and protein stability are separable issues.

A calculation can clarify this idea (60). In a population with $\Delta G_{\text{conformational}} = -5$ kcal/mol, all but one molecule in $\approx$10,000 on average would adopt the native fold. After a shift to conditions where $\Delta G_{\text{conformational}} = 0$ kcal/mol, only half the population would retain this

fold. Nevertheless, despite greatly diminished stability, those factors that determine the native fold still persist in two-state folding (Fig. 1). By way of macroscopic analogy, it is possible to stabilize a house against denaturation from a windstorm by strengthening connections between the roof and the walls, but the floor plan remains unchanged. Conformation and stability are separable.

Summarizing, the current view holds that residue side chains are the principal agent of chain organization (Corollary 1), and their effect is exerted under folding conditions (Corollary 2). In contrast to this view, the accumulating experimental evidence is focused on the peptide backbone in the unfolded state. That evidence is compelling: (*i*) osmolytes fold proteins by influencing the unfolded population, (*ii*) regardless of sequence, most proteins fold/unfold under similar conditions, (*iii*) the number of stable domains is limited, and (*iv*) two-state folding implies that conformation and stability are separable. All of these experimental facts prompt a reexamination of the denatured state.

Reassessing the Denatured State. Under denaturing conditions, protein molecules are typically regarded as randomly coiled polymers (like long strands of cooked spaghetti), and a population of such molecules (like a plate of pasta) is distributed over an undifferentiated energy landscape without pronounced conformational preferences. This view was developed in quantitative detail by Flory (17) and corroborated in experiment by Tanford (20): the radius of gyration (Eq. **1**) of a typical protein in strong denaturant (e.g., 6 M guanidinium chloride) (21) is well predicted by the Flory equation for a self-avoiding random walk (Eq. **2**).

Flory's treatment of unfolded chains stems from his isolated pair hypothesis, the idea that a protein's backbone torsion angles (ϕ,ψ angles), are independent of each other (17). Specifically, the hypothesis posits that the only local steric constraints on a residue are those imposed by immediately adjacent chain neighbors (Fig. 2). Such behavior simplifies the formal treatment of chain statistics; from residue independence, it can be inferred that the statistical behavior of the chain is given by the product of individual residue statistics. For more than four decades, the field's treatment of the left side of the folding equation, $U \rightleftharpoons N$, under denaturing conditions has been anchored by the isolated pair hypothesis (17).

We now realize that the isolated pair hypothesis breaks down in some regions of the dipeptide map (61). For instance, three or more residues in the α-region of ϕ,ψ space cannot be followed immediately by a residue from the β-region without encountering a steric clash (62). Such local steric restrictions extend beyond the linked dipeptide, and they eliminate conformational hybrids of α-helices and β-strands, thereby promoting chain organization. Moreover, the fact that unfolded proteins satisfy random-coil statistics need not imply that they are featureless; structured chains with flexible links also satisfy random-coil statistics (63). Indeed, even a steel rod behaves as a self-avoiding random coil if it is long enough.

Polyproline II in the unfolded state. In addition to organization imposed by systematic local steric restrictions, there is also a substantial population of left-handed polyproline II (P_{II}) conformation in unfolded proteins (64–67), as proposed by Tiffany and Krimm (68) more than three decades ago. Even earlier, Schellman and Schellman (69) had already argued that the spectrum of unfolded proteins was unlikely to be that of a true random coil. After these early studies, the ensuing literature hinted at a noticeable similarity between the spectra of P_{II} and unfolded proteins, but these ideas lay fallow for many years until Creamer's recent work (70, 71) stimulated renewed interest.

Residual structure in the denatured state. The limited success of early attempts to detect residual structure under denaturing conditions fortified a conviction that denaturation abolishes structure and reinforced the notion that the unfolded state is a featureless random coil. Unchallenged, these ideas eventually became dogma. But this dogma has been overturned by recent experimental evidence. For example, Kallenbach and coworkers (72) analyzed a blocked peptide containing seven consecutive alanine residues for the presence of residual structure. This peptide is too short to form a stable α-helix and therefore should be a random coil. Contrary to this expectation, the peptide is largely in P_{II} conformation, in agreement with predictions from theory (73), although the issue is not without controversy (74).

Loss of conformational entropy on folding. If accessible conformational space is vast and undifferentiated, the entropic cost of populating the native basin exclusively will be large. However, if the unfolded state is largely restricted to a few basins, the entropic cost is far less severe. For example, a residue in P_{II} is within a room-temperature fluctuation of any sterically allowed ϕ,ψ value in the upper left quadrant of the dipeptide map (64), and, consequently, ϕ,ψ values from this entire region would be thermodynamically indistinguishable. Suppose that a residue can visit any allowed region of the upper left quadrant in the unfolded state, but upon folding, it is constrained to lie within $\pm 30°$ of ideal β-sheet ϕ,ψ values. The reduction in ϕ,ψ space would only be a factor of 5.58, an energy cost of $\approx$1 kcal/mol at physiological temperature.

Preorganization in the unfolded state. At present, the unfolded state is undergoing reevaluation, and the full implications of the issues mentioned here are still evolving. It is clear, however, that the $U \rightleftharpoons N$ folding picture is being altered in radical ways by these three issues: (*i*) the breakdown of the isolated pair hypothesis, (*ii*) the presence of residual structure under denaturing conditions, and (*iii*) the detection of a significant P_{II} population. All three factors contribute to preorganization in the unfolded state.

If the unfolded state is preorganized, then the magnitude of accessible conformational space is not as vast as previously believed, and the corresponding entropy loss upon folding is not as large. It now seems likely that thermodynamic behavior is, in fact, dictated by a limited number of equilibrium states in both U and N, in accordance with the "classical view."

Part 5. A Backbone-Based Theory of Folding

This perspective has described 10 seemingly disparate aspects of protein folding. In particular:

1. The native fold is unique. The folding reaction is $U \rightleftharpoons N$, not $U \rightleftharpoons N_1 + N_2 + \ldots = \Sigma N_i$.
2. Folding is reversible.
3. No covalent bonds are made or broken in the folding reaction, $U \rightleftharpoons N$. Only weak bonds are involved.
4. Folding conditions and unfolding conditions are similar, respectively, for most mesophilic proteins, regardless of sequence.
5. The $U \rightleftharpoons N$ reaction is highly cooperative. Most single-domain proteins fold in an all-or-none manner (Fig. 1).
6. The fold is built on a scaffold of hydrogen-bonded α-helices and β-strands.
7. The number of stable domains is limited to a few thousand.
8. Proteins typically avoid metastable kinetic traps under native folding conditions.
9. Protecting/denaturing osmolytes fold/unfold proteins by operating predominantly on the backbone in the unfolded state, dialing folding

up/down but leaving the fold itself unaltered.

10. Stability and conformation are not synonymous. The native conformation can still be attained under grossly destabilizing conditions. Such conditions shift the U ⇌ N equilibrium toward either N or U, but not toward N* (i.e., an alternative folded state).

We hypothesize that these disparate properties are all a direct consequence of a central, underlying protein-folding mechanism: backbone hydrogen bonding.

The backbone hydrogen-bonding hypothesis follows immediately from the insight afforded by osmolytes: a universal protein folding mechanism must exist, and the only plausible candidate for its realization is the peptide backbone. In essence, the hypothesis recognizes that proteins are built on scaffolds of α-helices and β-strands. Conditions that favor intramolecular hydrogen bonding stabilize this scaffold, whereas conditions that disfavor intramolecular hydrogen bonds destabilize the scaffold (75). The two main denaturing agents used in folding studies are temperature or chemical denaturants, and they disfavor intramolecular hydrogen bonds either by heat-induced disorder or by competing with them, respectively.

Ostensibly, this backbone-based hypothesis lacks specificity. Backbones are the same, but conformations differ. Can such a hypothesis successfully account for the 10 properties listed here?

The most remarkable attribute of globular proteins is their capability to adopt a unique conformation. Many polymers undergo a coil ⇌ globule transition, but the condensed phase is not a unique structure. What factors or forces in proteins allow for the formation of a unique folded conformation to the exclusion of stable alternatives? The literature abounds with ideas regarding this topic. In the backbone-based hypothesis, the principal formative elements are α-helices and β-strands. A single protein domain might contain ≈10 such elements, so only a limited number of distinguishable constructs ($\approx 2^{10}$) is possible (see *What Anfinsen Could Not Have Known*) (76).

Side chains serve to select conformations from the limited repertoire of possible backbone conformations (48): α-helix, β-strand, turns, and loops. Discrimination among these categories appears to be exercised quite locally. For example, almost all backbone surface area in proteins is buried within, not between, elements of secondary structure (see figure 2 in ref. 77). The side-chain:backbone interactions in capping motifs that bracket helices are all within a few residues of the helix termini (78, 79). Once established, these backbone elements determine the tertiary structure, an old idea (80) extended recently (75, 81, 82).

Other characteristic properties also follow directly from the backbone-based hypothesis, such as the reversibility of the folding reaction. With a typical $\Delta G_{\text{conformational}} \approx -10$ kcal/mol, the energy equivalent of approximately two hydrogen bonds (83), the structure is poised near the margin of stability, where the U ⇌ N equilibrium can be successfully modulated by only small changes in intramolecular hydrogen bonding. Moreover, if backbone hydrogen bonding is key, structure formation/dissolution is expected to track with intramolecular hydrogen bond formation, thereby accounting for the observation that both folding conditions and unfolding conditions are largely sequence-indifferent. The high cooperativity of folding is explained by the fact that stabilizing conditions affect all backbone hydrogen bonds simultaneously, rather like an on/off switch in the cooperative folding unit.

Although situations have been reported where proteins get stuck in metastable traps (84), the backbone-based mechanism would tend to inhibit such occurrences. In a protein's coil ⇌ globule transition, backbone polar groups are sequestered unavoidably from solvent access and must be satisfied instead by intramolecular hydrogen bonds. The Boltzmann-weighted frequency of occurrence of a completely unsatisfied hydrogen bond can be estimated as $P_{\text{unsatisfied}} = e^{(-\Delta E_{hb}/RT)}$, where $P_{\text{unsatisfied}}$ is the probability of an unsatisfied hydrogen bond, ΔE_{hb} is the hydrogen bond energy, R is the gas constant and T is the temperature in Kelvin. At an energetic cost of ≈5 kcal/mol (85) for a completely unsatisfied hydrogen bond (either by an intramolecular partner or by water), the relative probability of finding one, $P_{\text{unsatisfied}}$, is ≈0.0002 at room temperature (83). In other words, even one unsatisfied hydrogen bond is unlikely. This constraint alone is sufficient to limit a protein solution to a few native-like populations when conditions favor collapse (81, 82).

The realization that stability and conformation are not synonymous comes as a surprise to some, although it is a direct consequence of two-state folding behavior (60). This realization is driven home in dramatic fashion by the demonstration that the osmolyte effect can shift the U ⇌ N folding equilibrium without affecting the fold (86, 87).

Of course, the fold is more than just a scaffold of α-helices and β-strands. These isodirectional segments account for approximately half of protein structures on average (88), and incorporating them into a three-dimensional structure requires turns and loops that can reverse the overall chain direction. However, β-turns are hydrogen-bonded backbone elements too (44), and they comprise at least half of the remaining protein structure (89). These hydrogen-bonded structural components diffuse, collide (90), and anneal (91) to form larger, ever more stable structural composites (40). The annealing process can be slow and complex (91), particularly for proteins having long loops or those comprised predominantly of α-helices in which scaffold elements adhere via hydrophobic interactions among side chains. Although such dynamics can dominate the folding behavior of individual proteins (92), they are only details in a larger perspective, where the osmolyte effect attests to the existence of a universal-folding mechanism (86, 87). The fine points of folding for any one protein are an insufficient basis for conclusions regarding the folding process in general.

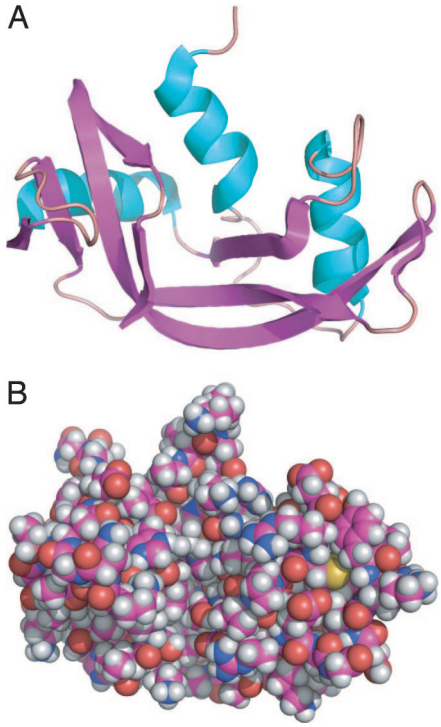

Fig. 5. Ribonuclease. (*A*) The fold. The molecule is depicted as a color-coded ribbon diagram: α-helices are shown in cyan, β-strands are shown in magenta. (*B*) The all-atom structure. Atoms are depicted as space-filling spheres with radii proportional to their van der Waals radii. Color-code is carbon, magenta; nitrogen, blue; oxygen, red; hydrogen, white; sulfur, yellow. Picture was rendered by using Pymol (127).

Part 6. The Protein Fold: What's in a Name?

Two definitions of the word protein fold are used interchangeably in the litera-

ture, and up to this point we have finessed the difference. One definition uses fold to mean the cartoon-level backbone conformation (Fig. 5*A*), whereas the other uses it to mean the detailed three-dimensional atomic structure (Fig. 5*B*). This distinction has been unintentionally blurred by the simplified ribbon representations of proteins (93, 94) that trace the through-space course of the backbone in a visually comprehensible way, to such an extent that many have come to view the cartoon and the structure as synonymous.

The difference between the backbone fold and the three-dimensional structure corresponds to our previously described distinction between sequence-indifferent conformation and sequence-dependent stability. A similar distinction was made some time ago by Head-Gordon *et al.* (95), who demonstrated that the polyalanine backbone, devoid of side chains, retains local minima that correspond to atom-level native structures. They concluded that the "mechanical stability of protein native structures does not depend on side-chain details but absolute energy (or free energy) stability is indeed controlled by those details." To avoid further ambiguity, we use the word fold to mean the hydrogen-bonded scaffold, reserving the word structure to emphasize atomic-level details. This distinction is not just a semantic quibble, as described next.

It is likely that a fold, as defined here, corresponds to the molten globule intermediate seen in experimental protein-folding studies. The molten globule is a collapsed state of the protein in which secondary structure has formed, or mostly so, but tightly packed side-chain interactions have yet to develop (96, 97). This state can be captured and observed in some proteins at low concentrations of denaturant or at acid pH under conditions that destabilize the native protein but not the intermediate. More than 30 years ago, Ptitsyn (98) guessed the existence of such intermediates, and confirmatory experimental evidence has been accumulating slowly since that time.

Ptitsyn's conjecture is now on solid ground. For example, hydrogen exchange protection factors indicate that the hydrogen-bonded scaffolding has already formed in the molten globule, although it lacks the stability of the final native structure, where core protection factors are many orders of magnitude greater (99). But, stability *per se* is not the critical issue here. The very existence of a molten globule equilibrium intermediate demonstrates that the core scaffolding and the native structure are thermodynamically separable populations during protein folding.

The hydrogen-bonded scaffold is a structural construct evident in individual molecules, whereas the corresponding molten globule is a thermodynamic entity measured in a population. They are two sides of the same coin, and each informs the other. The scaffold is robust to changes in the protein's amino acid sequence. For example, there are ≈130 structures of phage lysozyme mutants in the protein databank (100); all have conformations that closely resemble the parent structure (101). Presumably then, the molten globule also is robust to changes in sequence. Additionally, it has been shown that the fraction of the protein involved in local hydrogen-bonded scaffold elements, α-helices, β-hairpins, and β-turns, is highly correlated with the folding rate: the larger this fraction, the faster the protein folds (102). This correlation implies that formation of the molten globule is unlikely to be a slow step in folding.

In essence, the backbone-based theory presented here might be more accurately represented as a process in which the transition from U to N involves two stages. Scaffold formation occurs hierarchically during an initial stage, followed by a later stage that involves side-chain annealing, solvent exclusion, and further rigidification (103). If the protein is large enough, it is possible to capture this initial stage as a molten globule intermediate under suitable experimental conditions.

This two-stage folding strategy facilitates natural protein bioengineering. Once a scaffold is established, evolution is at liberty to spawn species-adapted orthologs and paralogs that tune the sequence for optimal performance in specific microenvironments or to explore new functions.

Part 7. Protein Folding: Analog Mechanism vs. Digital Mechanism

The thermodynamic hypothesis established the mindset that launched current thinking: Under folding conditions, proteins roll down a free energy hill to the bottom. On closer examination, this hypothesis bears the burden of an implicit paradox, one with many facets: how do molecules in transit from U to N avoid kinetic traps and alternative folds; given the speed limit set by single-bond rotation, how can the U to N transition be completed in microseconds; and how can N pay for lost conformational entropy on departure from a vast and featureless U?

A radically different perspective is suggested by the backbone-based hypothesis: only a few thousand distinguishable stable domains are possible, built on hydrogen-bonded scaffolds of α-helices and β-strands that fold/unfold under similar conditions. Here, the fold is selected from a limited repertoire of discrete, presculpted possibilities (48).

The thermodynamic hypothesis implies an analog (i.e., continuous) mechanism, whereas the backbone-based hypothesis implies a digital (i.e., discrete) mechanism, rooted in hydrogen bonding. Two conditions must be met for this digital mechanism to work: (*i*) a completely unsatisfied backbone polar group must be energetically expensive enough to be rare (and at ≈5 kcal/mol, it is; see *Part 3*) (84), and (*ii*) intramolecular peptide hydrogen bonds must be energetically favored over corresponding peptide-water hydrogen bonds. This second condition is only now being resolved, as discussed next.

A Half-Century Controversy About the Peptide Hydrogen Bond. Accurate assessment of the contribution that hydrogen bonds make to protein stability is now a central issue in protein folding (104). As stated by Baldwin, "the drive for continued rapid progress in protein structure prediction..., which requires a fuller understanding of protein-folding energetics, brings peptide H-bonds and peptide solvation into central focus." The energetic role of hydrogen-bonding in protein stability has experienced a literal revolution during the past half-century. In his model of the α-helix, Pauling *et al.* (105) estimated the strength of the peptide hydrogen bond to be of order −8 kcal/mol. Soon after, Schellman's measurements, from solution studies of urea dimers, showed that an intrapeptide hydrogen bond is enthalpically favored over a peptide-water hydrogen bond by ≈1.5 kcal/mol (106), remarkably close to contemporary values (107–110). These and similar early studies led to the conclusion that the peptide hydrogen bond contributes significantly to protein stability.

However, this view was overturned in Kauzmann's famous review (111), which argued that protein stabilization arises largely from the hydrophobic effect. Consistent with Kauzmann's proposal, Klotz and Franzen (112) measured the enthalpy of the interamide hydrogen bond of *N*-methyl acetamide in water at ≈0 kcal/mol and concluded that "the intrinsic stability of interpeptide hydrogen bonds in aqueous solution is small." Susi and Ard also (113) also reached a similar conclusion from a different system. Bolstered by these later studies, Kauzmann's proposal led to the widely held view that the hydrophobic effect provides the major free energy contribu-

tion to protein stability, with hydrogen bonds contributing little to the folding process and, perhaps, even opposing it.

Coming full circle, this view was to change yet again when Scholtz *et al.* (107) inferred that the enthalpy of polyalanyl helix formation in water is favorable by ≈1 kcal/mol per hydrogen bond. Similar values for other peptides were reported also (114). Leading the charge back toward this earlier view, Pace used experimentally determined free energy differences from numerous single-residue polar to apolar mutations to argue that "hydrogen bonds stabilize proteins and that the average net stabilization is −1 to −2 kcal/mol per intramolecular hydrogen bond" (109) with buried residues contributing as much as −3.5 kcal/mol (108).

Many still disagree. Honig and colleagues (115) used finite difference Poisson–Boltzmann methods to calculate the energetics of hydrogen bonding of *N*-methyl acetamide in water and organic solvent, concluding "that the formation and burial of a hydrogen bond opposes protein folding." These conflicting ideas continue to provoke controversy regarding whether peptide hydrogen bonds favor or oppose protein stability (115–118).

Summarizing this ongoing discussion, the weight of present evidence from peptides and proteins supports the conclusion that an intrapeptide hydrogen bond stabilizes a protein by 1- to 2-kcal/mol, although not all agree. This issue may be resolved finally by very recent work of Kiefhaber and coworkers (119) by using FRET, which provides persuasive evidence that intrapeptide hydrogen bonds promote chain collapse and, therefore, must be favored over corresponding peptide-water hydrogen bonds.

Hydrogen-Bonding, Hydrophobic Collapse, and Confinement. Which energetic factor predominates during folding: hydrogen bonding or the hydrophobic effect? Current experimental evidence points to hydrogen bonding. Specifically, at least four natively unfolded proteins can be forced to fold upon addition of organic osmolytes (87, 120, 121). In such cases, the total free energy of osmolyte-induced folding can be dissected into individual group transfer-free energies (86), whereupon it becomes evident that the peptide backbone is the dominant contributor. But the polar backbone unit lacks hydrophobic groups. Consequently, proteins can be driven to fold without any additional hydrophobic contribution, at least in these four cases.

Additionally, these group transfer-free energies can successfully predict the folding cooperativity (*m* values) for whole proteins by using a thermodynamic cycle (86, 122); again, the backbone is the dominant contributor in these predictions. In simulations, correct secondary structure assignments together with hydrogen bonding are sufficient to fold a collapsed backbone chain, devoid of side chains, to its native conformation (81, 82, 123). All of this suggests that the backbone plays the dominant role in protein folding.

It has been argued that a backbone-based model of protein folding is insufficient to account for chain collapse and compaction under folding conditions (124). Whereas solvent-squeezing of hydrophobic groups does facilitate chain collapse (111, 125), it is also true that under folding conditions, water is a less effective hydrogen-bonding solvent for the peptide backbone than the backbone itself, as discussed above.

Concluding Remark. In the current view, a folding protein is funneled toward a global free-energy minimum along a continuum of possible trajectories, each honed by evolution to be free of frustrating traps. Folding is an inherently analog process in which the formative interactions are among side chains. In the contrasting backbone-based model, the side-chain/backbone paradigm is inverted. A folding protein selects its fold from a limited repertoire of stable scaffolds, each built from a composite of hydrogen-bonded α-helices and/or β-strands. Folding is an inherently digital process in which the formative interactions are among backbone elements.

The concepts presented here have deliberately glossed over many details. The definition of a protein domain is fuzzy, both in this paper and in the literature. Does a 40-residue protein have the same stable minima as a 400-residue protein? What about proteins that do not fold in a two-state manner, those with obligatory chaperones, or hyperthermophiles? In seeking to prompt a fresh mindset, we were motivated to frame the whole picture in sweeping strokes. If that proves useful, the details are likely to follow.

G.D.R. dedicates this perspective to Carl Frieden on the occasion of his symposium at 77: a restless intellect, a gentle heart, a cherished friend. We are indebted to our colleagues who read drafts of this paper and offered valuable comments: Mario Amzel, Russell Doolittle, Arthur Lesk, Timothy Street, and, especially, Buzz Baldwin, Wayne Bolen, and Neville Kallenbach. This work was supported by grants from the Mathers Foundation (to G.D.R.), National Aeronautics and Space Administration (to J.R.B.) and Progetti di Ricerca Di Rilevante Interesse Nazionale 2005 (to A.M.).

1. Haber E, Anfinsen CB (1961) *J Biol Chem* 236:422–424.
2. Anfinsen CB (1973) *Science* 181:223–230.
3. Wu H (1931) *Chin J Physiol* 5:321–344.
4. Mirsky AE, Pauling L (1936) *Proc Natl Acad Sci USA* 22:439–447.
5. Myers JK, Pace CN, Scholtz JM (1995) *Protein Sci* 4:2138–2148.
6. Ginsburg A, Carroll WR (1965) *Biochemistry* 4:2159–2174.
7. Kumar MD, Bava KA, Gromiha MM, Prabakaran P, Kitajima K, Uedaira H, Sarai A (2006) *Nucleic Acids Res* 34:D204–D206.
8. Baldwin RL (1995) *J Biomol NMR* 5:103–109.
9. Bernal JD, Crowfoot D (1934) *Nature* 133:794–795.
10. Perutz MF, Rossmann MG, Cullis AF, Muirhead H, Will G, North ACT (1960) *Nature* 185:416–422.
11. Schuler GD, Epstein JA, Ohkawa H, Kans JA (1996) *Methods Enzymol* 266:141–162.
12. Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) *J Mol Biol* 247:536–540.
13. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM (1997) *Structure (London)* 5:1093–1108.
14. Chothia C (1992) *Nature* 357:543–544.
15. Rohl CA, Strauss CE, Misura KM, Baker D (2004) *Methods Enzymol* 383:66–93.
16. Bai Y, Sosnick TR, Mayne L, Englander SW (1995) *Science* 269:192–197.
17. Flory PJ (1969) *Statistical Mechanics of Chain Molecules* (Wiley, New York).
18. Fitzkee NC, Fleming PJ, Gong H, Panasik N, Jr, Street TO, Rose GD (2005) *Trends Biochem Sci* 30:73–80.
19. Fleming PJ, Rose GD (2005) in *Protein Folding Handbook*, eds Kiefhaber T, Buchner J (Wiley-VCH, Weinheim), Vol 2, pp 710–736.
20. Tanford C (1968) *Adv Protein Chem* 23: 121–282.
21. Kohn JE, Millett IS, Jacob J, Zagrovic B, Dillon TM, Cingel N, Dothager RS, Seifert S, Thiyagarajan P, Sosnick TR, *et al.* (2004) *Proc Natl Acad Sci USA* 101:12491–12496.
22. Levinthal C (1969) in *Mössbauer Spectroscopy in Biological Sytems*, eds Debrunner P, Tsibris JCM, Münck E (Univ of Illinois Press, Urbana), pp 22–24.
23. Corey RB, Pauling L (1953) *Proc R Soc London Ser B* 141:10–20.
24. Ramachandran GN, Ramakrishnan C, Sasisekharan V (1963) *J Mol Biol* 7:95–99.
25. Ramachandran GN, Sasisekharan V (1968) *Adv Protein Chem* 23:283–438.
26. Zwanzig R, Szabo A, Bagchi B (1992) *Proc Natl Acad Sci USA* 89:20–22.
27. Myers JK, Oas TG (2002) *Annu Rev Biochem* 71:783–815.
28. Frauenfelder H, Sligar SG, Wolynes PG (1991) *Science* 254:1598–1603.
29. Bryngelson JD, Onuchic JN, Socci ND, Wolynes PG (1995) *Proteins* 21:167–195.
30. Bryngelson JD, Wolynes PG (1987) *Proc Natl Acad Sci USA* 84:7524–7528.
31. Anderson PW (1990) *Phys Today* 43:9.
32. Anderson PW (1983) *Proc Natl Acad Sci USA* 80:3386–3390.
33. Go N (1983) *Annu Rev Biophys Bioeng* 12:183–210.
34. Bashford D, Chothia C, Lesk AM (1987) *J Mol Biol* 196:199–216.
35. Lesk AM, Chothia C (1980) *J Mol Biol* 136:225–270.
36. Bryson JW, Betz SF, Lu HS, Suich DJ, Zhou HX, O'Neil KT, DeGrado WF (1995) *Science* 270:935–941.

37. Richardson JS, Richardson DC (2002) *Proc Natl Acad Sci USA* 99:2754–2759.
38. Wetlaufer DB (1973) *Proc Natl Acad Sci USA* 70:697–701.
39. Crippen GM (1978) *J Mol Biol* 126:315–332.
40. Rose GD (1979) *J Mol Biol* 134:447–470.
41. Doolittle RF (1995) *Annu Rev Biochem* 64:287–314.
42. Levitt M, Chothia C (1976) *Nature* 261:552–558.
43. Kamat AP, Lesk AM (2006) *Proteins Struct Funct Bioinformatics*, in press.
44. Rose GD, Gierasch LM, Smith JA (1985) *Adv Protein Chem* 37:1–109.
45. Leszczynski JF, Rose GD (1986) *Science* 234:849–855.
46. Chothia C, Finkelstein AV (1990) *Annu Rev Biochem* 59:1007–1039.
47. Hoang TX, Trovato A, Seno F, Banavar JR, Maritan A (2004) *Proc Natl Acad Sci USA* 101:7960–7964.
48. Banavar JR, Hoang TX, Maritan A, Seno F, Trovato A (2004) *Phys Rev E Stat Nonlin Soft Matter Phys* 70:041905.
49. Mello CC, Barrick D (2004) *Proc Natl Acad Sci USA* 101:14102–14107.
50. Weinkam P, Zong C, Wolynes PG (2005) *Proc Natl Acad Sci USA* 102:12401–12406.
51. Hochachka PW, Somero GN (2002) *Biochemical Adaptation* (Oxford Univ Press, Oxford).
52. Schellman JA (2002) *Biophys Chem* 96:91–101.
53. Lin TY, Timasheff SN (1994) *Biochemistry* 33:12695–12701.
54. Auton M, Bolen DW (2004) *Biochemistry* 43:1329–1342.
55. Street TO, Bolen DW, Rose GD (2006) *Proc Natl Acad Sci USA* 103:13997–14002.
56. Baldwin RL, Rose GD (1999) *Trends Biochem Sci* 24:77–83.
57. Baldwin RL, Rose GD (1999) *Trends Biochem Sci* 24:26–33.
58. Richards FM (1977) *Annu Rev Biophys Bioeng* 6:151–176.
59. Maritan A, Micheletti C, Trovato A, Banavar JR (2000) *Nature* 406:287–290.
60. Lattman EE, Rose GD (1993) *Proc Natl Acad Sci USA* 90:439–441.
61. Pappu RV, Srinivasan R, Rose GD (2000) *Proc Natl Acad Sci USA* 97:12565–12570.
62. Fitzkee NC, Rose GD (2004) *Protein Sci* 13:633–639.
63. Fitzkee NC, Rose GD (2004) *Proc Natl Acad Sci USA* 101:12497–12502.
64. Mezei M, Fleming PJ, Srinivasan R, Rose GD (2004) *Proteins* 55:502–507.
65. Hamburger JB, Ferreon JC, Whitten ST, Hilser VJ (2004) *Biochemistry* 43:9790–9799.
66. Drozdov AN, Grossfield A, Pappu RV (2004) *J Am Chem Soc* 126:2574–2581.
67. Shi Z, Chen K, Liu Z, Sosnick TR, Kallenbach NR (2006) *Proteins* 63:312–321.
68. Tiffany ML, Krimm S (1968) *Biopolymers* 6:1379–1382.
69. Schellman JA, Schellman CG (1964) in *The Proteins*, ed Neurath H (Academic, New York) Vol 2, pp 1–37.
70. Stapley BJ, Creamer TP (1999) *Protein Sci* 8:587–595.
71. Creamer TP (1998) *Proteins* 33:218–226.
72. Shi Z, Olson CA, Rose GD, Baldwin RL, Kallenbach NR (2002) *Proc Natl Acad Sci USA* 99:9190–9195.
73. Pappu RV, Rose GD (2002) *Protein Sci* 11:2437–2455.
74. Makowska J, Rodziewicz-Motowidlo S, Baginska K, Vila JA, Liwo A, Chmurzynski L, Scheraga HA (2006) *Proc Natl Acad Sci USA* 103:1744–1749.
75. Honig B, Cohen FE (1996) *Fold Des* 1:R17–R20.
76. Przytycka T, Aurora R, Rose GD (1999) *Nat Struct Biol* 6:672–682.
77. Creamer TP, Srinivasan R, Rose GD (1997) *Biochemistry* 36:2832–2835.
78. Presta LG, Rose GD (1988) *Science* 240:1632–1641.
79. Richardson JS, Richardson DC (1988) *Science* 240:1648–1652.
80. Cohen FE, Richmond TJ, Richards FM (1979) *J Mol Biol* 132:275–288.
81. Gong H, Fleming PJ, Rose GD (2005) *Proc Natl Acad Sci USA* 102:16227–16232.
82. Fleming PJ, Gong H, Rose GD (2006) *Protein Sci* 15:1829–1834.
83. Fleming PJ, Rose GD (2005) *Protein Sci* 14:1911–1917.
84. Krishna MM, Lin Y, Englander SW (2004) *J Mol Biol* 343:1095–1109.
85. Mitchell JBO, Price SL (1991) *Chem Phys Lett* 180:517–523.
86. Auton M, Bolen DW (2005) *Proc Natl Acad Sci USA* 102:15065–15068.
87. Baskakov I, Bolen DW (1998) *J Biol Chem* 273:4831–4834.
88. Fitzkee NC, Fleming PJ, Rose GD (2005) *Proteins* 58:852–854.
89. Panasik N, Jr., Fleming PJ, Rose GD (2005) *Protein Sci* 14:2910–2914.
90. Karplus M, Weaver DL (1976) *Nature* 260:404–406.
91. Frieden C (2003) *Biochemistry* 42:12439–12446.
92. Sadqi M, Fushman D, Munoz V (2006) *Nature* 442:317–321.
93. Richardson JS (1981) *Adv Prot Chem* 34:168–340.
94. Lesk AM, Hardman KD (1982) *Science* 216:539–540.
95. Head-Gordon T, Stillinger FH, Wright MH, Gay DM (1992) *Proc Natl Acad Sci USA* 89:11513–11517.
96. Ptitsyn OB (1995) *Adv Protein Chem* 47:83–229.
97. Kuwajima K (1996) *FASEB J* 10:102–109.
98. Ptitsyn OB (1995) *Trends Biochem Sci* 20:376–379.
99. Hughson FM, Wright PE, Baldwin RL (1990) *Science* 249:1544–1548.
100. Kouranov A, Xie L, de la Cruz J, Chen L, Westbrook J, Bourne PE, Berman HM (2006) *Nucleic Acids Res* 34:D302–D305.
101. Matthews BW, Remington SJ (1974) *Proc Natl Acad Sci USA* 71:4178–4182.
102. Gong H, Isom DG, Srinivasan R, Rose GD (2003) *J Mol Biol* 327:1149–1154.
103. Hoeltzli SD, Frieden C (1998) *Biochemistry* 37:387–398.
104. Baldwin RL (2006) *Adv Protein Chem* 72:ix–xi.
105. Pauling L, Corey RB, Branson HR (1951) *Proc Natl Acad Sci USA* 37:205–210.
106. Schellman JA (1955) *C R Trav Lab Carlsberg [Chim]* 29:230–259.
107. Scholtz JM, Marqusee S, Baldwin RL, York EJ, Stewart JM, Santoro M, Bolen DW (1991) *Proc Natl Acad Sci USA* 88:2854–2858.
108. Shirley BA, Stanssens P, Hahn U, Pace CN (1992) *Biochemistry* 31:725–732.
109. Myers JK, Pace CN (1996) *Biophys J* 71:2033–2039.
110. Morozov AV, Kortemme T (2005) *Adv Protein Chem* 72:1–38.
111. Kauzmann W (1959) *Adv Protein Chem* 14:1–63.
112. Klotz IM, Franzen JS (1962) *J Am Chem Soc* 84:3461–3466.
113. Susi H, Ard JS (1969) *J Phys Chem* 73:2440–2441.
114. Richardson JM, Lopez MM, Makhatadze GI (2005) *Proc Natl Acad Sci USA* 102:1413–1418.
115. BenTal N, Sitkoff D, Topol IA, Yang AS, Burt SK, Honig B (1997) *J Phys Chem B* 101:450–457.
116. Fersht AR, Serrano L (1993) *Curr Opin Struct Biol* 3:75–83.
117. Honig B, Yang AS (1995) *Adv Protein Chem* 46:27–58.
118. Lazaridis T, Archontis G, Karplus M (1995) *Adv Protein Chem* 47:231–306.
119. Möglich A, Joder K, Kiefhaber T (2006) *Proc Natl Acad Sci USA* 103:12394–12399.
120. Henkels CH, Oas TG (2005) *Biochemistry* 44:13014–13026.
121. Mello CC, Barrick D (2003) *Protein Sci* 12:1522–1529.
122. Tanford C (1970) *Adv Prot Chem* 24:1–95.
123. Zhang Y, Hubner IA, Arakaki AK, Shakhnovich E, Skolnick J (2006) *Proc Natl Acad Sci USA* 103:2605–2610.
124. Dill KA (1999) *Protein Sci* 8:1166–1180.
125. Dill KA (1985) *Biochemistry* 24:1501–1509.
126. Dill KA, Chan HS (1997) *Nat Struct Biol* 4:10–19.
127. DeLano W (2003) *The PyMOL Molecular Graphics System* (DeLano Scientific LLC, San Carlos, CA).