Is protein folding hierarchic? I. Local structure and peptide folding

Robert L. Baldwin and George D. Rose

The folding reactions of some small proteins show clear evidence of a hierarchic process, whereas others, lacking detectable intermediates, do not. Nevertheless, we argue that both classes fold hierarchically and that folding begins locally. If this is the case, then the secondary structure of a protein is determined largely by local sequence information. Experimental data and theoretical considerations support this argument. Part I of this article reviews the relationship between secondary structures in proteins and their counterparts in peptides.

> *Synthetic analogs of globular proteins are unknown. The capability of adopting a dense globular configuration stabilized by self-interactions and of transforming reversibly to the random coil are characteristics peculiar to the chain molecules of globular proteins alone*[1].
>
> P. J. Flory

ALTHOUGH SYNTHETIC ANALOGS of proteins might be developed ultimately, Flory's succinct statement accurately summarizes the remarkable behavior of globular proteins. Today, an intense search is being made for the principles that guide the folding process. Although the work began with experimental studies (see Box 1), it has become increasingly clear that these results cannot be interpreted successfully without a satisfactory theory and simulations based on an adequate physical model. Accordingly, both experimentalists and theorists are now racing to learn what constitutes an adequate physical model.

In pursuit of this goal, we examine a curious paradox. The folding kinetics of small proteins reveal the existence of two classes of molecule (which we call class I and class II) that appear to fold by quite different mechanisms. Class I proteins, typified by α-lactalbumin (α LA), apomyoglobin (apoMb), RNase H, barnase and cytochrome *c* (cyt *c*), fold by a hierarchic process in which native-like secondary structure forms rapidly and is stabilized in molten-globule intermediates. Class II proteins, typified by chymotrypsin inhibitor 2 (CI2)[2] and cold-shock protein B (Csp B)[3], fold rapidly, in a kinetically two-state manner that lacks detectable intermediates. Some workers believe that these conflicting results imply that two different folding mechanisms exist: (1) a hierarchic mechanism that involves proteins with populated intermediates; and (2) a tertiary nucleation mechanism in which intermediates are not detectable. We turn to this question in Part II of this article (which will appear in the February issue of *TiBS*) by considering the structures and properties of observable intermediates in class I proteins, and data that describe the transition states of some class II proteins.

Throughout this review, we are asking a fundamental question: is protein folding hierarchic? We define hierarchic folding as a process in which folding begins with structures that are local in sequence and marginal in stability; these local structures interact to produce intermediates of ever-increasing complexity and grow, ultimately, into the native conformation (see Box 2). Non-hierarchic folding is a process in which tertiary interactions not only stabilize local structures but actually determine them. It follows that protein secondary structure is determined largely by local sequence information if folding is hierarchic, but not if folding is non-hierarchic. We use this difference to distinguish between the two models. Hierarchic folding is an attractive model because it is both conceptually simple and computationally tractable.

We consider several approaches for testing and evaluating hierarchic folding. For the model to be plausible, secondary structures (helices, turns and individual strands of sheet) must have at least borderline stability in peptides in the absence of tertiary interactions. Moreover, the local interactions inferred from inspection of helices, turns and strands in proteins of known structure should be reproducible in peptides, in which they can be measured quantitatively. The stop signals responsible for terminating protein helices should be found in the residue sequences that bracket the helix, not in tertiary interactions, and these local termination signals should operate in suitable peptide helices as well. We cover these topics in Part I of this article, which represents the straightforward, almost classical, part of this article. The results strongly support a hierarchic mechanism, but the evidence extends only to the initial stages of folding.

For further evidence, we turn next to folding intermediates and transition states, the topics reviewed in Part II of this article. The interpretation of such work is far more controversial but, in our opinion, it is here that remaining conflicts between alternative folding models will be resolved. We attempt to show that present evidence can be reconciled with hierarchic folding and to argue that intermediates in class I folding reactions strongly support this view. A novel approach to the problem is provided by the LINUS program[4], which can perform simple folding simulations in which structure is allowed to develop solely on the basis of local interactions.

Unlike our definition, the term hierarchic folding is sometimes used to describe a unique, sequential pathway that progresses to the native state in successive steps through a strict series of ever-larger, native-structure intermediates. By contrast, we suppose that alternative pathways of self-assembly are viable – as in the diffusion–collision model of Karplus and Weaver[5], which is illustrated by the folding kinetics of the λ repressor[6]. None of the five helices of the λ repressor is entirely stable alone, but random collisions between helices are mutually stabilizing and produce higher-order intermediates. Many possible folding routes exist, and the energy landscape (i.e. the differing stabilities of the helices and their combinatorial intermediates) determines the dominant populations.

Local sequence information influences local backbone conformation

Is the secondary structure of protein segments caused primarily by local interactions, or are non-local interactions

R. L. Baldwin is at the Dept of Biochemistry, Beckman Center, Stanford University Medical Center School of Medicine, Stanford, CA 94305-5307, USA; and **G. D. Rose** is at the Dept of Biophysics and Biophysical Chemistry, Johns Hopkins University School of Medicine, 725 N. Wolfe St, Baltimore, MD 21205-2185, USA.

0968–0004/99/$ – See front matter © 1999, Elsevier Science. All rights reserved. PII: S0968-0004(98)01346-2

also essential? At least three arguments support the proposition that non-local interactions play an essential role in secondary-structure formation.

(1) The accuracy of secondary-structure predictions is only 65–70%. This fact is usually interpreted to imply that the remaining variance of 30–35% is caused by non-local interactions, which are neglected in prediction algorithms.

(2) In proteins of known structure, the ratio of local to non-local contacts is small: there are normally fewer local contacts.

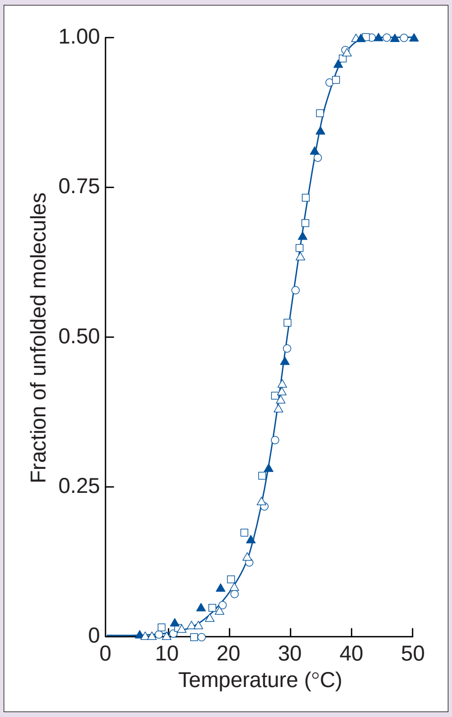
(3) The information needed to specify the conformation of an *n*-atom segment is proportional to $n \times (x_i, y_i, z_i)$, where x_i, y_i, z_i are coordinates of the i^{th} atom ($i \in n$). Secondary-structure propensities cannot encode this much information, which implies that non-local interactions are also involved.

Backbone dihedral angles in high-resolution protein structures, however, cluster almost entirely within two narrow regions (Box 3)[7]. Approximately half of all residues (except glycine and proline) are in the α region, within 30° of the values that specify an α-helix ($\phi = -60°$, $\varphi = -40°$), and another ~40% are in the β region, within 30° of values that specify a β-strand ($\phi = -120°$, $\varphi = +120°$).

Why are these two cluster points observed for non-glycyl, non-prolyl residues? This question was answered more than three decades ago by Sasisekharan and Ramachandran[8,9]: even in a dipeptide, other conformations are disfavored because of steric interference. Their conclusion, which was based on a hard-sphere model, was criticized as simplistic, although, in Richards's view, '…the use of the hard-sphere model has a venerable history and an enviable record in explaining a variety of different observable properties'[10].

Box 1. The classical picture of protein folding

Typically, a dilute solution of purified protein might contain ~10^{15} individual molecules; this population is called an ensemble. At any given moment, each molecule in the ensemble has a conformation, and each fluctuates dynamically over time. Many biophysical measurements – such as fluorescence or circular dichroism – report an ensemble-averaged quantity. Folding/unfolding of this ensemble is usually studied as a function of temperature or perturbing solvents (e.g. urea and guanidinium chloride). The plot shows a classic experiment by Ginsburg and Carroll[44]. Ribonuclease A unfolds (i.e. melts) as temperature increases. The different symbols represent different probes used to assess the conformation. Notably, each such probe falls on the same curve and therefore senses the same shift towards unfolding at a given temperature. This fact implies that there is a single unfolding transition.



Protein folding is a highly cooperative, all-or-none process. Like tipping over a chair, each molecule is either fully folded or fully unfolded and does not linger en route between one state and the other. This fact can be recast in thermodynamic terms. At the transition midpoint (the halfway point in the region of change) in the plot of ribonuclease unfolding, half the ensemble is folded and half is unfolded, and the population of partially folded/unfolded molecules remains negligible. Thus, folding is described as a two-state transition. Most small biophysical proteins undergo a two-state transition, whereas larger proteins that have multiple domains can deviate from strict two-state behavior. At a given temperature, individual molecules flip back and forth between folded and unfolded forms, but the fraction of folded/unfolded molecules does not change, because the probabilities of folding and unfolding are equal.

For only two populated states in equilibrium

$$\text{unfolded} \rightleftharpoons \text{folded} \tag{1}$$

it is valid to write an equilibrium constant:

$$K = \frac{[\text{folded}]}{[\text{unfolded}]} \tag{2}$$

The Gibbs free-energy difference between folded and unfolded forms is:

$$\Delta G^0 = -RT \ln K \tag{3}$$

In the above expression, R is the gas constant, and T is the absolute temperature.

Anfinsen's thermodynamic hypothesis[43] states that the folded conformation of the system will attain a minimum Gibbs free energy under physiological conditions. Above the transition midpoint, the free energy is positive (i.e. unfavorable), yet a population of folded molecules still persists[39]. Therefore, molecular specificity (i.e. the information needed to determine the fold) can be separated from stability (i.e. the set of interactions that resist unfolding).

On-line, see Fig. I.

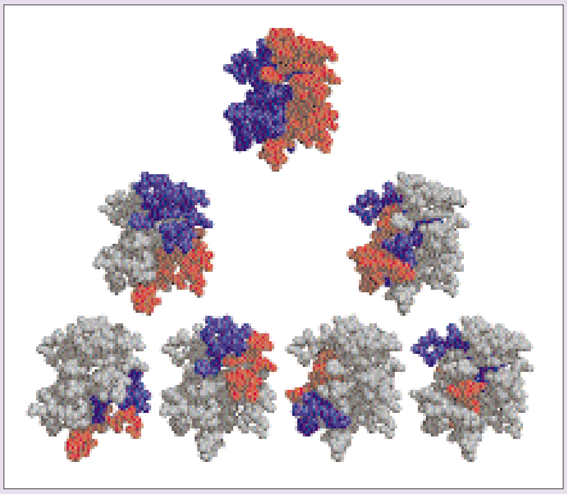
The sterically driven clustering of backbone dihedral angles favors local interactions for the following reasons. In a folded protein, most of the backbone is sequestered from solvent water, and hydrogen bonds for these interior polar groups are necessarily provided by intramolecular partners. (If these polar groups were left without hydrogen-bond partners, the folding equilibrium would be pushed well towards unfolding, where intermolecular hydrogen bonds with water could be realized.) Only two backbone geometries can result in systematic, extensible intramolecular hydrogen bonding: α-helix and β-sheet. Each is attained by repetition of values of backbone dihedral angles from the α and β region, respectively. Thus, the strong tendency of non-glycyl, non-prolyl residues to populate two regions preferentially is further enhanced in the interior of a folded protein.

The intramolecular interactions in a dipeptide are necessarily short-range, as is the physical basis for the two cluster points. The question remains, however: are non-local interactions needed to discriminate between the two allowed regions of the Ramachandran plot? Clues to the answer come from experiments on alanine-based peptides.

Short alanine-based peptides form stable helices in water[11]. This observation implies that helix formation is energetically favored for main-chain atoms, because poly-L-alanine is essentially pure backbone. (The β carbon can be regarded as a backbone atom because it has no additional degrees of rotational freedom, and all chiral amino acid residues have an equivalent β carbon.) This far-reaching conclusion is undiminished by ongoing controversy over whether short helices are stabilized primarily by hydrogen bonds or van der Waals interactions.

Box 2. Hierarchic organization of domains in globular proteins

In a hierarchy, each component is contained within the next larger component, as in a series of nested boxes. Two decades ago, Crippen[45] and Rose[46] showed that protein domains are organized as a structural hierarchy. In their work, they defined a structural domain as a contiguous, compact and physically separable segment of the polypeptide chain.



Though controversial at the time, the hierarchic organization of proteins is now well accepted. The early work[45,46] used analytical methods to identify domains in X-ray structures, but it was later realized that a simple procedure can approximate these results. To divide a protein into separable domains, display the structure with the first $n/2$ residues in red and the remaining $n/2$ residues in purple. Then repeat this procedure, iteratively, as shown for high potential iron protein (a typical example). In each successive stage of the hierarchy, it is apparent at a glance that red and purple regions do not intermingle. (The chain shown in gray is a place holder.)

The top-down, hierarchic organization of folded proteins is an experimental fact; no hypothesis is necessary to extract this result from known structures. The hierarchy suggests a bottom-up folding mechanism[46] in which chain segments form local structures of marginal stability, which then interact to produce intermediates of ever-increasing complexity. In this process, multiple folding routes co-exist, and the stabilities of the intermediates and their combinatorial associations will determine the dominant pathways.

On-line, see Fig. I.

Interactions that promote helix stability are common to almost every residue because all except glycine and proline have identical backbones. Therefore, the backbone is not the repository for information that discriminates between the α and β regions. Local interactions that effect such discrimination originate in side chains.

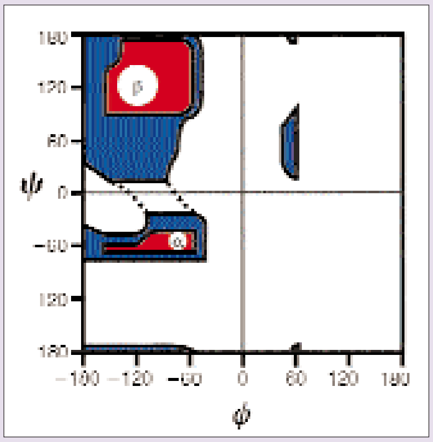
Side chains lose conformational freedom upon helix formation (see Box 4) because the bulky helix backbone is sterically incompatible with some side-chain conformations[12]. Such conformational restrictions are relieved when the backbone is extended. Conformational restriction is always thermodynamically unfavorable, and its extent is measured by differences in conformational entropy between one state (e.g. a helix) and another (e.g. a strand). These side-chain entropy factors vary among residues. For example, helix formation largely restricts a central valine residue to only one of its three possible side-chain conformations because one of the γ carbons bumps into a backbone atom in either of the other two.

Conformational entropy provides one physical mechanism for realizing the discriminatory information stored in residue side chains. The desolvation (i.e. stripping away of water) of backbone polar groups by side chains is another such mechanism (P. Luo and R. L. Baldwin, unpublished). In essence, side-chain steric factors bias selected chain segments away from the α region and therefore toward the β region. Consecutive residues that populate the same region of the Ramachandran plot, either α or β, preferentially become candidates for further stabilization as helices or strands, respectively. An α-helical structure is favored by peptide hydrogen bonds, as we argue below, but, for certain sequences, the attendant price in side-chain entropy is too overwhelming, pushing the segment towards a β-strand structure.

These sterically driven segments of nascent secondary structure will emerge in the equilibrium unfolded state and bias all subsequent folding events. Segments that have strong biases are poised to form persisting structure, especially when they are fortified by additional stabilizing interactions among segment side chains. Such segments become attractive candidates for study as autonomous folding units. We turn next to this topic.

Before proceeding, we must distinguish between β-strand and β-sheet. Residues in a β-strand have consecutive backbone dihedral angles, ϕ and ψ, of ~−120° and ~+120°, respectively. Such segments need not participate in hydrogen-bonded superstructures, although usually they do, often as strands of a β-sheet. β-sheets involve two or more β-strands that are hydrogen bonded to each other and, in this sense, they are tertiary structures. β-strand propensities depend largely upon local side-chain steric factors that bias chain segments away from the α region. β-sheet propensities are more complex. Data suggest[13,14]

Box 3. Allowed conformations of a dipeptide



All possible values of a dipeptide can be represented by a two-dimensional plot of the backbone angles ϕ and ψ. This was first done by Sasisekharan and Ramachandran[8,9], who modeled peptide atoms as hard spheres and mapped the conformations that are sterically allowed. The ϕ,ψ values that result in steric clash (i.e. bring any two atoms closer than the sum of their respective van der Waals radii) are considered to be disallowed; all other values are allowed. Such a plot is shown; sterically allowed regions are shaded (red regions are completely allowed; blue regions are marginally allowed).

Remarkably, most conformational space is disallowed – as first noted by Sasisekharan[13]. There are only two major allowed regions: one near −60°,−40°; the other near −120°,−120°. Significantly, these two regions correspond to the values assumed by backbone dihedral angles of peptides in α-helix and β-sheet (although a dipeptide lacks backbone hydrogen bonds). This finding is generally applicable, because steric constraints that limit a dipeptide are pertinent to any larger peptide.

The Sasisekharan–Ramachandran (S–R) plot is derived from a hard-sphere model. How well do the two allowed regions correspond to actual experimental observations? When values of backbone dihedral angles from 42 high-resolution proteins are overlaid on an S–R plot[7], >50% fall within a narrow region around α, and >40% fall within β (if we ignore glycine and proline, the two residues that are conformationally distinct from the other 18).

On-line, see Fig. I.

that sheet propensities depend both on the isolated β-strand propensity and on the extent of side-chain burial[13], a complicated, context-sensitive function that involves the tertiary structure microenvironment[14]. Now that β-hairpin systems are available for study[20], more information about these issues should be forthcoming.

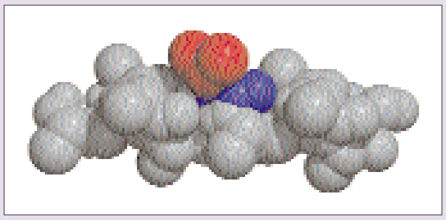
Helix-termination signals

Is the secondary structure adopted by a peptide preserved in all essential respects when the same sequence is present in a protein? This question is posed in particularly clear form for helix endpoints. The homopolymer peptide helices studied 40 years ago in organic solvents are difficult to nucleate but, after a nucleus is formed, helical residues are added easily, and the helix can be extended readily to >100 residues. There are no natural stopping points. In proteins, by contrast, helices have precise boundaries. Are the helix-termination signals encoded in the local sequence, or are tertiary interactions responsible for helix localization?

In an early study[16], Perutz, Kendrew and Watson noted that proline residues occur commonly at the ends of helices in myoglobin and hemoglobin, and they suggested that proline is a termination signal. Today, proline residues are known to be highly effective in terminating peptide helices, but proline is found in the interior of some protein helices.

Analysis of helix capping in proteins of known structure has provided a wealth of information. Practically every helix has recognizable termination signals. A recent analysis of 1316 protein helices[17] found that hydrophobic capping at the N- and C-termini is evident in ⩾80% of the cases studied and that hydrogen bonding to backbone NH or CO groups is evident in nearly half of these cases. Helix ends are weakened by a deficit of backbone hydrogen bonds: the initial four NH groups and final four CO groups lack intrahelical hydrogen-bonding partners. These end effects are substantial, encompassing two-thirds of the residues present in an α-helix of average length (12 residues). Further, solvent access to amide groups is already severely hindered by helix geometry at the helix N-terminus. Thus, a helix can be stabilized as well as terminated by hydrogen bonding between a side-chain or main-chain group and backbone peptide groups at the helix ends. Seven commonly occurring helix-capping motifs are listed in Table 1. Their existence is clear evidence that the signals for helix localization are encoded in the local sequence.

Box 4. Conformational entropy and helix formation



The flexibility of almost all residue side chains is reduced upon helix formation, because the bulky helix backbone is sterically incompatible with some side-chain conformers. Such restriction is energetically disfavored and is expressed as a loss of conformational entropy[12]. For example, helix formation largely restricts the central valine of an otherwise polyalanyl helix to only one (*trans*) of its three staggered (*gauche*$^-$, *gauche*$^+$ and *trans*) side-chain conformers, because one of the γ carbons bumps into the backbone in either of the other two (on-line, see Fig. I). Alanines are shown in gray; the valine backbone is shown in blue; the valine side-chain atoms – in *trans* – are shown in red. Note that the volume occupied by the valine side chain is sterically restricted.

If all three side-chain conformers are populated equally in the unfolded state, and if restrictions imposed by the backbone limit the side chain to just one conformer in the helix, then the loss in conformational entropy will be Rln3 (where R is the gas constant). In actuality, these suppositions are slightly exaggerated; the computed entropy cost of putting a valine side chain in the middle of a polyalanine α-helix[12] is closer to Rln2 (slightly more than 0.4 kcal mol^{-1} at physiological temperature).

On-line, see Fig. I.

Peptide folding

The first peptide helix studied in water, C peptide (residues 1–13 of RNase A), has the properties[18] expected from the parent protein structure, in which this helix extends from Thr3 through His12. Because the helix still terminates near His12 in the longer S peptide (residues 1–20), this finding focused attention on local side-chain interactions that serve as helix-stop signals. Two such signals, which also serve to stabilize the helix, were found in C peptide[18]. The Glu2–Arg10 salt bridge had been recognized much earlier in the X-ray structure of RNase S, but studies of the C-peptide helix in solution identified the Phe8–His12$^+$ pseudo hydrogen bond, providing an early experimental example of this novel class of interactions. Baldwin and co-workers[18] deduced the existence of these two pairwise side-chain interactions by substituting alanine for various C-peptide residues and then measuring changes in the pH dependence of the helix content. However, these two pairwise interactions were not easily separated from charge–helix dipole effects exerted between formal charges on peptide groups at helix termini and nearby charged side chains.

The compound effects seen in studies of C peptide led to the conclusion that a simpler model system was needed – one in which interactions could be singled out and measured individually. Furthermore, helix formation is an inherently statistical phenomenon that must be analyzed by helix-coil theory; the two-state model (in which the peptide is either entirely helical or entirely non-helical) is a poor approximation because helix ends are frayed. Fortunately, one amino acid, alanine, has a high enough helix propensity to form a stable helix on its own in the absence of fortifying side-chain interactions[11,18]. Alanine-based peptides provide a sufficiently simple model system to allow quantitative and straightforward measurement of specific interactions between pairs of side chains[18–20]. These measurements provide insight into the contribution of side-chain interactions to protein stability, including some interactions that are controversial, such as hydrogen bonds and salt bridges.

Table 2 lists some representative side-chain interactions that have been measured in peptide helices. Typically, only one pairwise rotamer conformation can realize a given interaction (i.e independently, each side chain can adopt multiple rotamers, but only one of these will result in an interaction between the two side chains): for example, Gln–Asp or Gln–Asn residues spaced at an i – i+4 interval in an α-helix maintain an interaction only if residue i and residue i+4 are in the *trans* and *gauche*$^+$ conformation, respectively[20]. When corrected for specific rotamers, the strength of the side-chain interactions listed in Table 2 is ~−1 kcal mol^{-1}. Studies of this kind are complemented by NMR studies of peptide helices, which provide direct evidence for a putative side-chain interaction. For example, this latter approach verified the structure of the capping box[19]. In addition, Monte Carlo simulations can estimate side-chain interactions between pairs of nonpolar residues[21]. Successful measurement and/or calculation of side-chain interactions is closely tied to determining helix propensities accurately (see below).

Table 1. Helix-capping motifs in globular proteins

N-capping motifs		
N′ → N3 or N4	Capping box	Hydrophobic interaction between residues N′ and N3/N4; H-bonds between Ncap side chain and N3 backbone and, reciprocally, Ncap backbone and N3 side chain
N″ → N3 or N4	Big box	Hydrophobic interaction between residues N″ and N3/N4; H-bonds between Ncap side chain and N3 backbone and, reciprocally but staggered, N′ backbone and N3 side chain
N‴ → N3 or N4	β box	Hydrophobic interaction between residues N‴ and N3/N4; backbone H-bond between Ncap and N‴
C-capping motifs		
C″ → C3/C′Gly	Schellman motif	Hydrophobic interaction between C″ and C3; C′ is glycine; backbone H-bonds between C″ and C3, and between C′ and C2
C‴ or C′′′′ → C3/C′n	Pseudo-Schellman motif	Hydrophobic interaction between C‴ or C′′′′ and C3; C′ is non–β-branched (designated n) in lieu of glycine; backbone H-bonds between C″ and C3, and between C′ and C2
C‴ or C′′′′ → C3/C′Gly	α_L motif	Hydrophobic interaction between C‴ or C′′′′ and C3; C′ is glycine and C″ is not proline; backbone H-bond between C′ and C3
C′′′′ or C′′′′′ → C3/C′Pro	Proline motif	Hydrophobic interaction between C′′′′ or C′′′′′ and C3; C′ is trans-proline; three-center backbone H-bonds between amide hydrogens at C‴ and C′′′′ and carbonyl oxygen at Ccap

Conformational constraints at helix ends result in a small number of structures that can provide intramolecular hydrogen-bonded (H-bonded) partners while maintaining the hydrophobic interaction. Most common among them are the seven motifs listed[17]. Nomenclature for helices and their flanking residues is as follows:

…N′′′′-N‴-N″-N′-Ncap-N1-N2-N3- … -C3-C2-C1-Ccap-C′-C″-C‴-C′′′′…

N1 through C1 belong to the helix proper; the primed residues belong to turns that bracket the helix at either end. Ncap and Ccap are bridge residues that belong both to the helix and an adjacent turn.

Each motif is named for the closest pair of interacting hydrophobic residues that straddles the helix terminus. A hydrophobic interaction between residues A and B is written as A → B, where the arrow points from the hydrophobic residue external to the helix to the hydrophobic residue within the helix. For example, N′ → N4 signifies a hydrophobic interaction between residue N′ and N4 (the capping box). Several motifs are further qualified by the presence of a particular residue found preferentially at a given position. Such cases are annotated by appending a slash (/), and then the position and residue name. For example, C″ → C3/C′Gly signifies the Schellman motif, which has a characteristic hydrophobic interaction between C″ and C3 and, preferentially, a glycine residue at the C′ position.

Muñoz and Serrano[22] have developed a different approach to predicting helix content. Their AGADIR algorithm[22] fits experimental peptide-helix data to helix-coil theory, using many contributing parameters. Specifically, they used data for 323 peptides taken from the literature to derive the following: (1) all 20 single-residue helix propensities; (2) pairwise side-chain interaction parameters at both *i* – *i*+3 and *i* – *i*+4 spacings (324 pairwise interactions are included in an 18 × 18 matrix; Pro and Cys are omitted because of lack of data); (3) values for the peptide hydrogen bond; (4) values for all amino acids at N-cap and C-cap positions; and (5) values for capping-box residues. Given that these parameters represent weighted statistical averages, they are less accurate than corresponding values given by experiments designed to measure only a single parameter. Nevertheless, AGADIR does a remarkable job of reproducing the helical content of natural-sequence peptides – to within ±10% in most cases and often better. The inclusion of side-chain interactions is particularly important because, without them, natural-sequence peptides tend to lack measurable helix content in water.

Table 2. Side-chain interactions in peptide helices

Interaction	Examples
Charge–aromatic	Phe–His$^+$ Phe–Met Trp–His$^+$
Charged hydrogen-bond	Gln–Asp$^-$ Glu^0–Lys$^+$ His$^+$–Asp^0
Hydrogen bond	Gln–Asp^0 Gln–Glu^0
Nonpolar	Leu–Ile Leu–Leu Leu–Val

These side-chain interactions, and other examples of the same types, have been measured quantitatively in alanine-based peptide helices[18,20]. When the measured strength of the interaction is corrected for frequencies of the two rotamers forming the interaction[20], values of ~–1 kcal mol^{-1} are found. In the examples given here, the pairwise spacing is *i*, *i* + 4, and the rotamers are *trans, gauche*$^+$. In most cases, the frequency of occurrence of interacting pairs in protein helices is above random, and a high proportion of *trans, gauche*$^+$ rotamers are evident. The Phe–Met interaction can be classified as either nonpolar or charge–aromatic.

There is an unresolved discrepancy in reconciling parameters derived from alanine-based peptides with those from natural-sequence peptides. Although the rank order of helix propensities is the same in both systems, ratios of the corresponding helix propensities (e.g. relative to glycine) are about six times higher in alanine-based peptides. An actual helix propensity (defined as the helix-propagation parameter of helix-coil theory) can be determined in alanine-based peptides, but only the ratio of two helix propensities can be determined in natural-sequence peptides. A study of a 17-residue helix excised from RNase T1 shows the differing behavior of the two systems clearly[23]. Helix-propensity ratios for nonpolar substitutions at a central position in this peptide helix agree well with corresponding substitutions made in intact RNase T1, similar data for barnase and T4 lysozyme, and predictions made by AGADIR. However, these values are a sixth of their counterparts for alanine-based peptides. A possible explanation is that the magnitude of the helix propensities depends on whether or not side-chain shielding desolvates CO and NH groups in the helix backbone. Unlike most other residues, alanine is too short to contribute to backbone desolvation (P. Luo and R. L. Baldwin, unpublished).

AGADIR has proven to be quite successful in reproducing the helix content of natural-sequence peptides by using the parameters described above. However, an important problem remains: using peptides to predict whether protein sequences will form α-helix, β-sheet or neither. Because β-sheet is tertiary structure, it is difficult to reproduce its formation in peptide models, although Serrano and co-workers[20] are making progress, using β-hairpins. In Part II of this article, we describe the use of folding simulations (performed by LINUS) to predict local conformational biases for α-helices, β-strands and peptide-chain turns. To determine which helices are likely to form first in the folding process, the helix-forming behavior of peptides from all the helical segments of a protein can be ascertained; several groups have analyzed hen lysozyme[24,25] and myoglobin[26] in this way.

A crucial issue in peptide folding is the energetic role of the peptide hydrogen bond. The subject has been controversial since 1955, when Schellman[27] used the enthalpy of urea-dimer formation in water to estimate the enthalpy of the peptide hydrogen bond to be −1.5 kcal mol^{-1}. Baldwin and co-workers[18] measured the enthalpy of formation of a 50-residue alanine-based helix calorimetrically. The value (−1.1 ± 0.2 kcal mol^{-1} residue) agrees with results obtained from fitting helix-coil theory to the thermal unfolding curves of alanine-based peptides. The measurements are reliable, but the values are puzzling because estimating the enthalpy of unfolding as 1 kcal mol^{-1} gives rise to a large enthalpy deficit[28] when the unfolding enthalpies of proteins are examined. The puzzle can be resolved upon the realization that the folding process is accompanied by wholesale desolvation of backbone polar groups. Desolvation commences during helix formation in a peptide of heterogeneous composition, where side chains shield the backbone from water molecules and reduce the net enthalpy of helix formation (P. Luo and R. L. Baldwin, unpublished). This effect is attenuated in an alanine-based peptide because alanine, unlike most residues, is too short to interfere with backbone–water interactions. However, the desolvation effect is enhanced in a folded protein, where hydrogen-bonded secondary structure in the interior substitutes for some, but not all, protein– water hydrogen bonds. We note that, in a hierarchic folding reaction, the desolvation penalty (i.e. the unfavorable desolvation of backbone polar groups) of helical segments is largely prepaid (i.e. accomplished) upon helix formation, prior to formation of tertiary structure.

Another related topic in peptide-folding studies is the role of the helix-enhancing reagent trifluoroethanol (TFE). Luo and Baldwin ascribed the TFE effect to strengthening of peptide hydrogen bonds[29]. In particular, the strength of the intramolecular hydrogen bond formed by salicylic acid is augmented in TFE–H_2O mixtures, in the same manner as the average helix propensity measured for alanine-based peptides[29]. Moreover, TFE–H_2O mixtures also stabilize a β-hairpin in a similar manner[15] – as expected if TFE acts by strengthening peptide hydrogen bonds. Reverse turns can form in water in peptides that have favorable sequences, and NMR data confirm the existence of a peptide hydrogen bond between turn residues 1 and 4 (Ref. 30). These results are consistent with an active role for peptide hydrogen bonds in the formation of all classes of secondary structure.

Peptide simulations

If folding is hierarchic then the folding reaction should be rooted in chain segments that have native-like conformations. Do peptide fragments corresponding to these segments exhibit a dominant backbone conformation in simulations? If so, does that conformation resemble the native one?

Surprisingly, the field lacks a physicochemical theory of protein secondary structure. Typical prediction methods are based on statistical likelihoods[31] or neural nets[32]. Eisenberg and co-workers[33], and Hecht and co-workers[34], have documented distinctive patterns of hydrophobicity that are consistent with amphipathic helices and strands, but this observation applies to tertiary structures[21]. The lack of a satisfactory theory surely contributes to the suspicion that no such theory exists because tertiary-structure formation is needed to induce secondary structure.

Although peptides seem to be natural candidates for simulation, they have not been investigated as thoroughly as proteins. One reason is that even those peptides that do exhibit a dominant backbone conformation in simulations also visit other conformations as well. This ensemble (i.e. set of conformations) lends itself to the study of conformational transitions[35] but resists evaluation in predictive work. In particular, ensemble behavior is difficult to evaluate, given the field's pervasive tendency to assess predictive success by using a single scalar figure of merit – the root-mean-square deviation (RMSD) from an X-ray structure. We return to these issues shortly.

Rooman and Wodak[36], Abagyan and Totrov[37], and Pedersen and Moult[38] have performed simulations to predict the conformation of selected peptides. Although encouraging, their results are limited to carefully chosen peptides suspected to have unusual stability. It can be argued that the conformational tendencies of occasional peptides are insufficient to underwrite a general folding mechanism.

For hierarchic folding, we would expect pronounced conformational biases to be distributed throughout the entire amino acid sequence. That is exactly what Pedersen, Braxenthaler and Moult find (J. T. Pedersen, M. Braxenthaler and J. Moult, pers. commun.). In a study that included every 12-residue fragment of a small protein, they analyzed the distribution of conformers within each fragment's free-energy spectrum. Although the number of conformations sampled by each fragment is not large, all but five of the spectra include energetically favorable conformations that are within a 3-Å RMSD of their respective counterparts in the protein; the five outliers are within a 3.5-Å RMSD. This suggests that conformational biases are both pronounced and ubiquitous.

What is the principal source of conformational bias in peptide fragments? The conventional answer is hydrogen bonding and the hydrophobic effect. However, the forces that stabilize a protein against denaturation need not be synonymous with the forces that select for a particular conformation, if folding is hierarchic[39]. In fact, hydrogen bonds and hydrophobic interactions are universal precisely because they lack specificity. The backbone hydrogen bonds that knit residues into secondary structure are probably individually stabilizing (Table 2), but they are also promiscuous, involving atoms common to almost all residues. Similarly, the protein interior has sufficient plasticity to ensure stabilizing hydrophobic interactions, even when challenged by wholesale mutation of the core[1,40,41]. Mutations that alter size, shape and chain length have been made in numerous proteins and have a negligible effect on overall conformation. To a surprising degree, grease is simply grease.

Repulsive interactions in proteins are a familiar source of conformational bias[10,42]. The fact that two atoms cannot be in the same place at the same time imposes severe limits on the configurational freedom of adjacent residues[9,10], and additional excluded volume constraints are prevalent in space-filling protein models. However, the degree to which steric constraints induce ubiquitous chain organization has yet to be realized fully in fact or used in simulations.

Steric interplay between side chains and the backbone is thought to be an

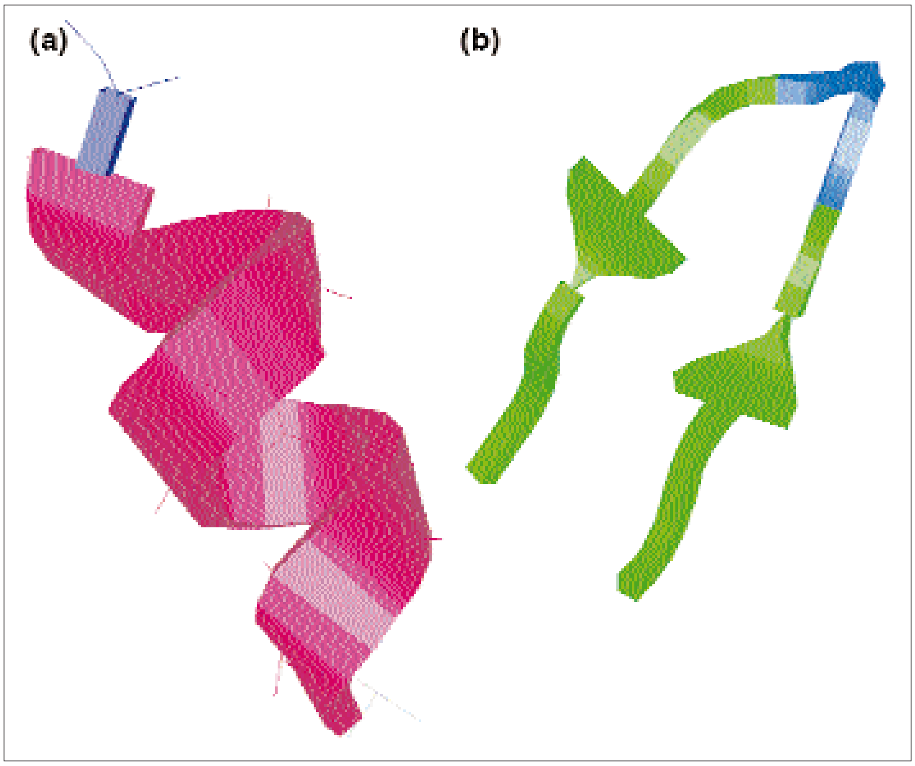


Figure 1
LINUS[4] simulations of polyalanine and polyvaline. Simulations were performed for a short (500 cycle) interval (see Ref. 4). Interactions are necessarily local, because both peptides are short. Secondary-structure elements are color coded (helix, pink; strand, green; turn, blue; coil, cyan). **(a)** Simulation of an alanyl decapeptide. During 500 cycles of simulation, residues 4–10 were in helical conformation >80% of the time (residues 5–7 were in helical conformation >90% of the time). The minimum-energy structure associated with this simulation is shown. **(b)** Simulation of the tetradecapeptide V_6GGV_6. During 500 cycles of simulation, residues 3–6 and 9–13 were in a strand conformation >88% of the time, and residues 7–8 remained exclusively in turn conformation. Again, the minimum-energy structure is shown.

important factor in discriminating between helix and strand[21] (see above). Such factors are difficult to represent explicitly in a polypeptide chain of heterogeneous composition. A given residue both affects and is affected by its chain neighbors, and conformational biases established in this way will propagate along the chain. Simulations are well suited to treat fluctuating linkages of this kind and, to this end, LINUS was created[4].

LINUS is a Monte Carlo routine that was devised, in large part, to explore the influence of sterics on protein folding. The program introduced a smart move set (a jargon term used in Monte Carlo simulations) in which the torsion angles of three consecutive residues are perturbed simultaneously. For backbone dihedrals, four overall move types are allowed: helix, strand, turn and coil. This search strategy enables a helix, turn or strand to be nucleated in a single Monte Carlo move. LINUS ascends the folding hierarchy in discrete stages. Only local interactions are allowed in the initial stage; then, increasingly, non-local interactions are phased into successive stages. The chain folds under the influence of a primitive energy function that has only three terms – small attractive contributions from hydrogen bonds and hydrophobic interactions, and an infinite repulsive contribution from any steric clash (i.e. two atoms in the same place at the same time). Dispersion forces and electrostatics are ignored deliberately. Within only a few hundred Monte Carlo cycles, and with all non-local attractive interactions suppressed (which gives the initial stage in the hierarchy), pronounced conformational biases emerge throughout the molecule. These biases – towards helix, strand, turn or coil – are caused by local steric effects and can be quantified readily.

LINUS simulations support the idea that secondary-structure biases arise locally, are sterically based and are distributed throughout the chain. To illustrate this, we performed two simulations: one of polyalanine; the other of a polyvaline chain that has two glycine residues in the middle (Fig. 1). Within a few simulation cycles, all but the terminal residues of the polyalanyl peptide spend at least 80% of their trials in a helical conformation, which is stabilized by backbone hydrogen bonding. With the same protocol, the polyvalyl peptide arranges quickly into a hairpin, which is stabilized by hydrogen bonding and hydrophobic interactions between the two strands. Importantly, the polyvalyl peptide still adopts a hairpin when hydrophobic interactions are switched off (an easy experiment in a simulation, but an impossible one in reality), although excursions from this preferred conformation become more frequent. In all simulations, each residue samples conformational space uniformly – that is, the protocol itself does not introduce biases: each residue is given the same opportunity to visit helix or strand.

Mechanistically, how do sterics drive a polyalanyl peptide towards an α-helix and a polyvalyl peptide towards a β-hairpin? In the polyalanyl peptide, most Monte Carlo moves that sample (i.e. try out) the α region of ϕ,ψ space (near $-60°$, $-40°$) are accepted (i.e. allowed), thus increasing the probability that a backbone hydrogen bond will form and, in turn, stabilizing a helical conformation – a continuing cycle of helix-promoting reinforcement. In the polyvalyl peptide, almost all Monte Carlo moves that sample the α region are rejected (i.e. disallowed) because of the high probability of steric clash between a side-chain γ carbon and the backbone. However, most Monte Carlo moves in the β region (near $-120°$, $+120°$) are accepted because side-chain–backbone steric clash is relieved in this conformation. The resultant effect pushes the polyvalyl peptide towards an extended conformation, thus increasing the probability that an intersegment hydrogen bond will form between juxtaposed strands and, in turn, stabilizing this conformation – a continuing cycle of β-sheet-promoting reinforcement. Interstrand hydrophobic interactions serve to shift the equilibrium even further towards sheet formation.

In the picture of the folding reaction that emerges from these and other examples, an unfolded protein – under folding conditions – will experience pronounced conformational biases distributed throughout the polypeptide chain. Local, steric interactions arise as a consequence of the covalent sequence, an extension of the allowed conformations in a dipeptide[8,9]. As such, they provide a plausible starting point for hierarchic folding.

Conclusions

The basic requirement for hierarchic folding is satisfied: α-helices, β-hairpins and β-turns populate the flickering clusters of local structure that Anfinsen[43] hypothesized 25 years ago. Simulations using LINUS indicate that biases towards these several local structures are established primarily by steric effects and hydrogen bonds, and then further enhanced by hydrophobic interactions in some cases. In particular, helix biases are caused by the steric interplay between side-chain rotamers and the bulky helix backbone, and by cooperative formation of helices. Residues biased away from helix and turn are biased toward β-strand. These secondary-structure biases emerge during the earliest stages of the folding reaction, and they condition subsequent folding events. In Part II of this article, we will examine the question of whether early bias anchors a hierarchic folding process or is merely adventitious. Here, we have already provided an important, preliminary clue: helix-termination signals are localized near the boundaries of helix sequences. This fact favors the argument that folding is hierarchic.

Acknowledgements

We thank our co-workers, past and present, for their discussion of these issues, and we apologize to our colleagues whose contributions could not be cited because of the limit on references. We are indebted to L. Mario Amzel for discussion of several issues in this review. We thank the NIH for support (grants GM 31475 to R. L. B. and GM 29458 to G. D. R.).

References

1 Flory, P. J. (1969) *Statistical Mechanics of Chain Molecules*, p. 301, Wiley

2 Itzhaki, L. S., Otzen, D. E. and Fersht, A. R. (1995) *J. Mol. Biol*. 254, 260–288

3 Perl, D. *et al*. (1998) *Nat. Struct. Biol*. 5, 229–235

4 Srinivasan, R. and Rose, G. D. (1995) *Protein Struct. Funct. Genet*. 22, 81–99

5 Karplus, M. and Weaver, D. L. (1976) *Nature* 260, 404–406

6 Burton, R. E., Myers, J. K. and Oas, T. G. (1998) *Biochemistry* 37, 5337–5343

7 Creamer, T. P., Srinivasan, R. and Rose, G. D. (1997) *Biochemistry* 36, 2832–2835

8 Sasisekharan, V. (1962) in *Collagen* (Ramanathan, N., ed.), pp. 39–78, Wiley

9 Ramachandran, G. N. and Sasisekharan, V. (1968) *Adv. Protein Chem*. 23, 283–438

10 Richards, F. M. (1977) *Annu. Rev. Biophys. Bioeng.* 6, 151–176

11 Marqusee, S., Robbins, V. H. and Baldwin, R. L. (1989) *Proc. Natl. Acad. Sci. U. S. A.* 86, 5286–5290

12 Creamer, T. P. and Rose, G. D. (1992) *Proc. Natl. Acad. Sci. U. S. A.* 89, 5937–5941

13 Minor, D. L., Jr and Kim, P. S. (1994) *Nature* 371, 264–267

14 Smith, C. K., Withka, J. M. and Regan, L. (1994) *Biochemistry* 33, 5510–5517

15 Ramirez-Alvarado, M., Blanco, F. J. and Serrano, L. (1996) *Nat. Struct. Biol.* 3, 604–611

16 Perutz, M. F., Kendrew, J. C. and Watson, H. C. (1965) *J. Mol. Biol.* 13, 669–678

17 Aurora, R. and Rose, G. D. (1998) *Protein Sci.* 7, 21–38

18 Baldwin, R. L. (1995) *Biophys. Chem.* 55, 127–135

19 Kallenbach, N. R., Lyn, P. and Zhou, H. (1996) in *Circular Dichroism and Conformational Analysis of Biomolecules* (Fasman, G. D., ed.), pp. 201–259, Plenum

20 Stapley, B. J. and Doig, A. J. (1997) *J. Mol. Biol.* 272, 465–473

21 Creamer, T. P. and Rose, G. D. (1995) *Protein Sci.* 4, 1305–1314

22 Muñoz, V. and Serrano, L. (1994) *Nat. Struct. Biol.* 1, 399–409

23 Myers, J. K., Pace, C. N. and Scholtz, J. M. (1997) *Proc. Natl. Acad. Sci. U. S. A.* 4, 2833–2837

24 Segawa, S-I., Fukuno, T., Fujiwara, K. and Noda, Y. (1991) *Biopolymers* 31, 497–509

25 Yang, J. J. *et al*. (1995) *J. Mol. Biol.* 252, 483–491

26 Reymond, M. T., Merutka, G., Dyson, H. J. and Wright, P. E. (1997) *Protein Sci.* 6, 706–716

27 Schellman, J. A. (1955) *C. R. trav. lab. Carlsberg Ser. chim*. 29, 223–229

28 Yang, A-S., Sharp, K. A. and Honig, B. (1992) *J. Mol. Biol.* 227, 889–900

29 Luo, P. and Baldwin, R. L. (1997) *Biochemistry* 36, 8413–8421

30 Dyson, H. J. *et al*. (1988) *J. Mol. Biol.* 201, 161–200

31 Fasman, G. (1989) *The Development of the Prediction of Protein Structure,* Plenum

32 Rost, B. and Sander, C. (1994) *Protein Struct. Funct. Genet.* 19, 55–72

33 Eisenberg, D., Weiss, R. M. and Terwilliger, T. C. (1984) *Proc. Natl. Acad. Sci. U. S. A.* 81, 140–144

34 Kamtekar, S. *et al*. (1993) *Science* 262, 1680–1685

35 Karplus, M. and Shakhnovich, E. (1992) in *Protein Folding* (Creighton, T. E., ed.), pp. 127–195, W. H. Freeman

36 Rooman, M. J. and Wodak, S. J. (1991) *J. Mol. Biol.* 221, 961–979

37 Abagyan, R. and Totrov, M. (1994) *J. Mol. Biol.* 235, 983–1002

38 Pedersen, J. T. and Moult, J. (1997) *J. Mol. Biol.* 269, 240–259

39 Lattman, E. E. and Rose, G. D. (1993) *Proc. Natl. Acad. Sci. U. S. A.* 90, 439–441

40 Lim, W. A. and Sauer, R. T. (1991) *J. Mol. Biol.* 219, 359–376

41 Matthews, B. W. (1995) *Adv. Prot. Chem.* 46, 249–278

42 Saven, J. G. and Wolynes, P. G. (1996) *J. Mol. Biol.* 257, 199–216

43 Anfinsen, C. B. (1973) *Science* 181, 223–230

44 Ginsburg, A. and Carroll, W. R. (1965) *Biochemistry* 4, 2159–2174

45 Crippen, G. M. (1978) *J. Mol. Biol.* 126, 315–332

46 Rose, G. D. (1979) *J. Mol. Biol.* 134, 447–470