Protein Science (1994), *3*:650–668. Cambridge University Press. Printed in the USA.

Protein folding dynamics: The diffusion-collision model and experimental data

MARTIN KARPLUS[1] AND DAVID L. WEAVER[2]

[1] Department of Chemistry, Harvard University, Cambridge, Massachusetts 02138
[2] Department of Physics, Tufts University, Medford, Massachusetts 02155

(RECEIVED August 20, 1993; ACCEPTED November 17, 1993)

Abstract

The diffusion-collision model of protein folding is assessed. A description is given of the qualitative aspects and quantitative results of the diffusion-collision model and their relation to available experimental data. We consider alternative mechanisms for folding and point out their relationship to the diffusion-collision model. We show that the diffusion-collision model is supported by a growing body of experimental and theoretical evidence, and we outline future directions for developing the model and its applications.

Keywords: diffusion-collision model; folding models; folding pathways; kinetic intermediates; microdomains; protein folding dynamics

The biological activity of a globular protein is determined by its average 3-dimensional structure (the native conformation) and by the internal flexibility of this structure (Brooks et al., 1988). It has been shown (Anfinsen, 1973) for a number of proteins that the native structure can fold spontaneously under appropriate conditions without any information other than that contained in the linear sequence of amino acid residues. Because the purpose of the genetic code is to specify the sequence of amino acids, the prediction of the structure of a protein from its amino acid sequence is an essential part of molecular biology. There are now about 1,700 proteins whose structures have been determined (Bernstein et al., 1977) by X-ray crystallography and by NMR. Although these structures have served as the basis of much of our understanding of proteins, they represent an insignificant fraction of the total number of different proteins (on the order of 10^9) in living systems. It will not be possible to determine all of their 3-dimensional structures by experimental methods, though the problem would be simplified if the naturally occurring proteins are constructed from a much more limited number of structural elements (Dorit et al., 1990). Moreover, because a chain of only 100 amino acids has $20^{100} \approx 10^{130}$ different possible sequences, and a significant number of these sequences may have a thermodynamically dominant fold (Chan & Dill, 1990; Shakhnovich & Gutin, 1990), there could be many artificial proteins with novel properties not yet realized in nature. A rational approach to protein design and engineering thus has as one of its elements the prediction of the structure of a protein from its amino acid sequence.

There are 2 aspects to the prediction problem. One is thermodynamic in character and concerns the prediction of the native structure of a protein from its sequence; the other is dynamical in character and concerns the prediction of the mechanism by which a denatured protein folds to the native conformation in solution or in vivo. The dynamic aspect of the folding problem is considered in this review. It is often phrased in terms of the Levinthal paradox (Levinthal, 1966, 1968), which corresponds to the realization that a random search of all possible structures would take longer than the age of the universe to find the native conformation. There would be no search problem if each of the amino acids could find its native conformation independently of the others, or if only nearest-neighbor interactions were involved. This would reduce the protein folding problem to an analog of the helix-coil transition. What makes the search problem difficult is that long-range interactions are involved. It is the presence of long-range effects that make the folding transition cooperative (pseudo-first order), an essential element of the stability and kinetics of proteins (Privalov, 1989; Shakhnovich & Finkelstein, 1989).

In 1976, we discussed the dynamics of protein folding (Karplus & Weaver, 1976) and introduced a possible mechanism by which a protein can make use of a more sophisticated procedure than a simple random search of all conformational possibilities. The proposal, which is related to the experiments of Anfinsen (1973) and the model of Ptitsyn and Rashin (1973, 1975), is that the folding of a protein molecule is divided into parts such that the information contained in the sequence of each part

Reprint requests to: Martin Karplus, Department of Chemistry, Harvard University, Cambridge, Massachusetts 02138; e-mail: marci@tammy.harvard.edu; or David L. Weaver, Department of Physics, Tufts University, Medford, Massachusetts 02155; e-mail: dweaver@jade.tufts.edu.

can be used independently. As one limiting case, we proposed the diffusion-collision model (Karplus & Weaver, 1976) and suggested that the diffusion-collision mechanism had a significant role in the folding dynamics of many proteins. In 1976 there was very little experimental information concerning the detailed kinetics of protein folding or concerning the structures of species along the protein folding pathways. Since that time, the use of a variety of techniques has provided much information on differential rates, on the nature of intermediates, and on the stability of isolated secondary structural elements. A current perspective of the diffusion-collision model and its predictions seems appropriate. In the present paper, we evaluate the status of the diffusion-collision model in light of recent experimental results, primarily based on NMR and CD techniques, that provide information, albeit limited, on the mechanism of protein folding. We also consider its relation to other models that have been proposed since it was developed.

The diffusion-collision model is based on the existence of fluctuating quasiparticles, called microdomains, which may be portions of incipient secondary structure (α-helices or β-strands) or hydrophobic clusters. The microdomains move diffusively, and microdomain–microdomain collisions take place. Collisions can lead to coalescence into multimicrodomain intermediates, which may involve microdomains not necessarily adjacent in the linear sequence. If the microdomains consist of secondary structural elements, they would be expected to appear and disappear transiently before the tertiary structure is formed (individual, marginally stable microdomains in a folding protein are difficult to observe, at present, because of the dead time in experiments, e.g., stopped flow). Folding then proceeds as a series of coalescence steps that might follow a unique order (single folding pathway). Alternatively, there might be parallel folding possibilities (multiple pathways), whose dominant components could depend on the solution conditions for a given protein and on the amino acid sequences of structurally homologous proteins. The exact folding behavior would be governed by a number of properties, such as the stability of the elementary microdomains and the barriers involved in the coalescence steps. Incorrect intermediates would be likely to occur as transient species, and portions of the folding process would be expected to have rates that depend on the solvent viscosity.

The diffusion-collision model reduces the dynamics of the folding process from a consideration of the individual amino acids forming the polypeptide chain to that of the properties of microdomains and their interactions. This provides a possible solution to the search problem that could lead to folding on the observed time scale. Moreover, the diffusion-collision model is sufficiently simple that it is possible to make approximate calculations of the magnitudes of kinetic rate constants describing the various steps of the folding process. These are based on the physical properties of the microdomains, some of which can be estimated from experiments (e.g., size, shape, stability). Thus, the model can be used to estimate folding rates and the concentration of kinetic intermediates. In this respect, the diffusion-collision model differs from many other models of folding, which only provide pictorial descriptions of important folding events. The diffusion-collision model does not address the final stage of the folding process in which the precise atomic positions in the native structure are determined. This aspect of folding could be studied by molecular dynamics simulations starting from folding intermediates that result from the diffusion-collision process. Some progress has been recently made in such refinements (Skolnick et al., 1993).

In what follows, we first review the diffusion-collision model. A description is given of the qualitative aspects and quantitative results of the diffusion-collision model and their relation to available experimental data. We then assess the role of alternative dynamical mechanisms in folding and point out the relationship of other models for the folding process to the diffusion-collision model. In the final section future directions for developing the model and its applications are outlined.

Diffusion-collision model

The diffusion-collision model (Karplus & Weaver, 1976, 1979) views the protein as composed of several parts (elementary microdomains), each short enough for all conformational alternatives to be searched through rapidly, as compared with the time scale of the entire folding process. Thus, the existence of microdomains in the folding polypeptide chain and their importance in the kinetics of folding provides a way for the folding protein to avoid examining the entire set of conformational alternatives. For example, if each residue has 3–5 possible conformational alternatives, and it takes 10^{-12}–10^{-13} s to examine each one, a full conformational search of a 10-residue microdomain would take $\approx 10^{-6}$–10^{-9} s. In some more regular structures (e.g., α-helices), a more limited search is required and the overall time scale is expected to be somewhat shorter. The small size of the elementary microdomains implies that their structure, though accessible by random events, is only marginally stable. Recent studies on the helix–coil transition of the S-peptide of ribonuclease (Shoemaker et al., 1985; Mitchinson & Baldwin, 1986) and other helix-forming peptides (Marqusee et al., 1989; Chakrabartty et al., 1991) and on the existence of nascent α-helices and β-turns (Wright et al., 1988) in aqueous solution have shown that peptide fragments can have reasonable stability (e.g., up to 10% at low temperatures and 1% at room temperature) and that this stability depends on the sequence.

The microdomains move diffusively under the influence of internal and random external forces and microdomain–microdomain collisions take place. Collisions sometimes lead to coalescence into microdomain pairs and higher aggregates (higher-order microdomains or subdomains). This can occur if both microdomains have formed at least part of their secondary structure (i.e., that in the contact region), and the collision involves an appropriate orientation. The diffusion-collision portion of the folding process could start with the entire polypeptide chain in an extended random-coil state or in a more collapsed state. Overall folding would consist of a series of steps leading to a conformation with portions of the backbone close to the native conformation. It is possible that this stage of folding leads to the molten globule state (Ohgushi & Wada, 1983; Ptitsyn, 1987), where some secondary structural elements are loosely associated. The final step of the folding process would then be the formation of the exact tertiary structure, including close packing of the side chains. During this last phase, proline isomerization (Brandts et al., 1975) or other slow events could be rate limiting.

The folding steps might have to follow a unique order to yield the native structure (single correct pathway; Creighton, 1988; Baldwin, 1990); alternatively, particularly in the early stages of folding, there could be different sequences of folding steps (multiple pathways; Harrison & Durbin, 1985) leading toward the native structure. The detailed dynamics and whether 1, a few,

or many pathways are important are expected to depend on the properties of the microdomains for the protein under consideration. If the higher-order microdomains are on a pathway leading to the native structure, they are generally native or correctly folded intermediates. Non-native or incorrectly folded intermediates are also possible in the diffusion-collision model, because microdomains that are not in contact in the native structure are likely to collide in the course of folding. The folding of a protein with disulfide bonds (bovine pancreatic trypsin inhibitor, BPTI) may involve folding pathways through intermediates (Creighton, 1978) that have non-native disulfides, although their kinetic importance is not clear (Staley & Kim, 1992). Structural studies suggest that these intermediates consist of a combination of native and random-coil-like structures, rather than non-native folded structures (States et al., 1980, 1984; Staley & Kim, 1992). Other aspects to be considered in the model are that after partial collapse and/or weak coalescence of microdomains to a more compact structure with a non-native conformation, the attainment of the native conformation might involve surface diffusion in 1 or 2 dimensions. In these cases, the effective diffusion coefficient could be significantly different from that appropriate for an aqueous solution and internal friction effects might be important (Karplus & Weaver, 1979).

As outlined above, the diffusion-collision model provides a pictorial description of the larger scale kinetic event involved in the folding of small globular proteins and in the domains of larger proteins (Fig. 1). It implies that the principal early dynamical events of the folding process are concerned with the properties of microdomains and their interactions rather than the individual amino acids. This reduces the folding times to reasonable values because it simplifies the search of the vast range of configurational space that exists for the unfolded polypeptide chain. However, even with this reduction of the problem, the presence of the many potential energy barriers and the long overall folding time (10^{-3}–10^{2} s) make it difficult to study the detailed motions involved in the folding process at the atomic level. Consequently, it is useful for a semiquantitative analysis of protein folding dynamics to concentrate on the microdomains as the basic entities. This makes it possible to develop a scheme for calculating the folding pathways and the time course of species formed during the folding reaction. In this respect, the diffusion-collision model differs from other descriptions of folding, which are generally limited to the pictorial aspects of the folding process (see the section on other descriptions of folding, below). In the original description of the diffusion-collision model (Karplus & Weaver, 1976) a relatively simple calculation of the folding time of 2 (unstable) microdomains was presented. Based on the experimental estimates of the stability of protein fragments available at that time (Hammes & Schullery, 1968; Gratzer & Beaven, 1969; Brown & Klee, 1971; Epand, 1972; Sachs et al., 1972; Panijpan & Gratzer, 1974; Saski et al., 1975) and the physical properties of idealized microdomains, the calculated folding times were estimated to be in the range of the experimental values for small proteins.

Two methods have been used to calculate the diffusion-collision dynamics involved in the folding of peptides and proteins. In the first method, which is essentially analytic, the dynamics of folding is simulated by a set of diffusion equations that describe the motion of the microdomains in aqueous solution and by coupled boundary conditions that provide for their collision and possible coalescence. Detailed analyses (Weaver, 1984) showed that the diffusion-collision dynamics of a multimicrodomain protein reduces to a network of 2-microdomain steps in which the calculable rate constants depend on physical properties of microdomains (size, shape, reactivity, orientation) and the equilibrium state of the system. The folding kinetics are approximated by solving the kinetic equations that couple the elementary steps; this corresponds to applying the "chemical kinetics" approximation to the diffusion-collision model. The individual rate constants for the coalescence of 2 entities (e.g., 2 elementary microdomains or microdomain complexes) can be written analytically in terms of the physical parameters of the system.

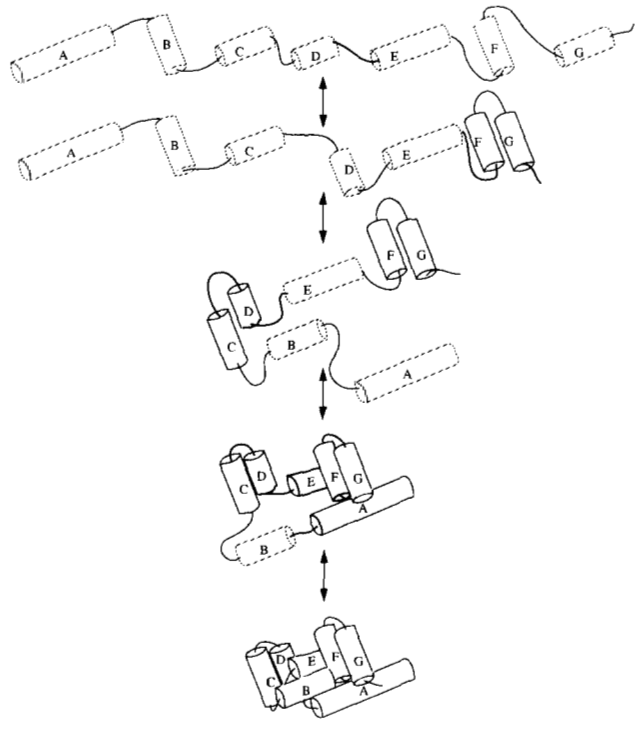

Fig. 1. A cartoon showing 5 "snapshots" of the diffusion-collision folding kinetics of a multimicrodomain protein or protein domain. The time line of the kinetics runs from the top down. The system starts in the "random coil" set of conformations. The microdomains (A–G) are individually unstable (indicated by dashed outlines) and transiently occupy folded secondary-structure conformational states (probably native or near native). Diffusive encounters, when a pair of microdomains are transiently folded, lead to more stable coalescence intermediates (denoted by solid outlines) held together by hydrophobic interactions. Multimicrodomain intermediates collide and coalesce into a loosely folded, more stable structure.

In most calculations, the geometrical parameters in the pairwise rate constants have been evaluated in a spherical approximation of the type often used to simplify diffusion calculations. Each microdomain is assigned a radius corresponding to a sphere with the same volume as that obtained from the van der Waals surface of the structure. The relative diffusion of 2 spheres of radii R_1 and R_2 connected by a perfectly flexible inelastic string is equivalent to the motion of a point particle confined to move in the volume between 2 concentric, spherical shells. The inner shell has a radius $R_{min} = R_1 + R_2$ and the outer shell has a radius R_{max} = length of inelastic string (the extended chain length between the microdomains). The folding

rate, τ_f, depends in a simple way on the sizes of the microdomains through R_{min}, on their separation through R_{max}, on their stability through their coil–secondary-structure equilibria, and on their relative coalescence probability (denoted by β) once a collision has taken place. The quantity β is determined by the probability that the 2 microdomains are both folded and correctly oriented when they collide and by the activation energy barrier, if any, to coalescence (Karplus & Weaver, 1979). The time for folding, τ_f, decreases as β increases ($0 \leq \beta \leq 1$). The equation for τ_f can be written (see Appendix A for details):

$$\tau_f = \frac{l^2}{D} + \frac{L}{D}\frac{\Delta V(1-\beta)}{A\beta}. \quad (1)$$

Here D is the microdomain–microdomain relative diffusion coefficient, the parameter l^2 has the dimension of length-squared and is related to the limits of the diffusion space R_{max} and R_{min} and the relative size of the inner and outer spherical shells, $R_{min}/R_{max} = \epsilon$. For example, in 3 dimensions,

$$l^2 = \frac{R_{max}^2}{3}\frac{(1 - \frac{9}{5}\epsilon + \epsilon^2 - \frac{1}{5}\epsilon^6)}{\epsilon(1-\epsilon^3)}. \quad (2)$$

The diffusion volume (denoted by ΔV) is $\frac{4}{3}\pi(R_{max}^3 - R_{min}^3)$; the larger the volume of diffusion space, the harder it is for the microdomains to collide. The collision surface area (denoted by A) is the area of the inner spherical shell in the spherical microdomain approximation, $4\pi R_{min}^2$; the larger the collision surface area, the easier it is for the microdomains to collide. The parameter L has the dimension of length and depends on the size of the diffusion space and on the overall rate at which microdomain pairs switch from possible coalescence states to noncoalescence states (see Equation A-10); the larger the value of L, the harder it is for microdomains to coalesce. Some typical sizes for the parameters in Equation 1 for myoglobin are given in Table 1; the myoglobin helices are treated as microdomains in the spherical approximation.

The above method has been applied to study the dynamics of 2- (Karplus & Weaver, 1979; Weaver, 1980, 1982; Bashford & Weaver, 1986) and 3- (Weaver, 1984) microdomain systems, and to the overall folding kinetics of apomyoglobin (Bashford et al., 1988) and the λ-repressor (Bashford et al., 1984), as well as the initial step in the folding of cytochrome *c* (Bashford et al., 1990).

Table 1. *Diffusion-collision parameters for some myoglobin helix pairs at room temperature*

Parameter	AH pair	FH pair	GH pair
R_{min}	21.0 Å	20.0 Å	21.5 Å
R_{max}	271 Å	98.4 Å	59 Å
D	$6.3(10)^{10}$ Å^2/s	$6.5(10)^{10}$ Å^2/s	$6.1(10)^{10}$ Å^2/s
l^2	$2.72(10)^5$ Å^2	$9.91(10)^3$ Å^2	$1.31(10)^3$ Å^2
l^2/D	$4.3(10)^{-6}$ s	$1.52(10)^{-7}$ s	$2.2(10)^{-8}$ s
L	21.0 Å	11.4 Å	24.7 Å
A	$5.5(10)^3$ Å^2	$5.28(10)^3$ Å^2	$5.8(10)^3$ Å^2
ΔV	$8.3(10)^7$ Å^3	$4.37(10)^6$ Å^3	$8.2(10)^5$ Å^3
$L\Delta V/DA$	$5.0(10)^{-6}$ s	$1.57(10)^{-7}$ s	$5.7(10)^{-8}$ s

Some aspects of the results and their relation to experiments are discussed in the next section. Figure 2 shows an example of the chemical kinetics approximation to the diffusion-collision model. It is applied to the 3-microdomain protein (or portion of a larger protein) described in Appendix B. Figure 2 shows the time dependence of each of the 8 states (see Table 2 for a description of the states and Table 3 for the kinetic parameters). The rate constants in Equations B-10–B-17 have been chosen so that the time dependence of each of the 8 states of the system is unique for visualization. The 1-microdomain pair intermediates (states 2, 3, and 5 in Table 2) peak early and dissipate completely after 0.5 time units. The 2-microdomain pair intermediates (states 4, 6, and 7 in Table 2) arise more slowly (the 1-microdomain intermediates need to appear first) and dissipate more slowly in this example. The final state (state 8 in Fig. 2), with 3 pairs, has an initial time lag while the 1- and 2-pair intermediates form. Its probability (concentration in a unimolecular process) then increases with a half-time of about 1 time unit.

As an alternative to the description of the microdomains as spherical entities, it is possible to introduce a somewhat more detailed model. Each amino acid residue of the peptide chain is approximated by a single spherical interaction center linked by virtual bonds (Flory, 1969; Levitt, 1976; McCammon et al., 1980). The diffusion equation, which is used in the first method and which provides a continuum approximation to the motion of the microdomains, is replaced by the discrete, coupled Langevin equations of motion for the residue spheres that interact according to an effective potential energy function; the effect of solvent is included in the potential and in the frictional drag and random force terms of the Langevin equation. These equations can be solved numerically (Ermack & McCammon, 1978) to treat the full folding problem or to obtain information related to the chemical kinetics approximation. The validity of some of the approximations of the analytical model has been assessed by this approach. This includes the effect of intervening chain segments on the diffusion of microdomains. Also, the evaluation of some of the parameters used to calculate folding rates (e.g., stabilities of helices and sheets) and the kinetic constants for microdomain–microdomain collisions were examined. The method has been applied in simulating the helix–coil transition of an α-helical (McCammon et al., 1980) microdomain, the sheet–coil transitions of a β-hairpin (Yapa et al., 1992) microdomain, and the diffusion and collision of 2 α-helices (microdomains) connected by a coil segment (Lee et al., 1987). In the 2-helix simulation (Lee et al., 1987), the problem was simplified to reduce the cost of the simulations by constraining the helices to remain helical and allow dihedral angle transitions to occur only in the coil segments, i.e., the helix–coil transition of residues in the helical portion was not permitted to occur. Thus, the calculation simulated the diffusion-collision dynamics of 2 separately stable microdomains connected by a chain segment. Helix–coil transitions, coupled to collisional events, will be introduced in subsequent simulations on the basis of studies of the rates of the coil-to-secondary-structure transitions (McCammon et al., 1980; Yapa et al., 1992; Schneller & Weaver, 1993).

Consequences of the diffusion-collision model

In this section, we describe certain results obtained by use of the diffusion-collision model. Some of them are qualitative in character and provide a pictorial view of folding. Others are more

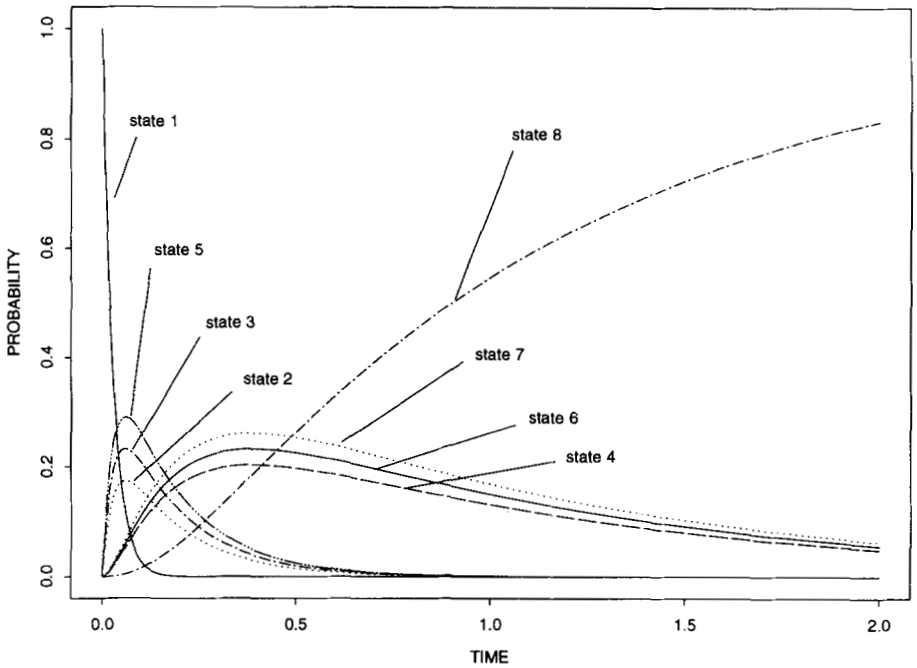

Fig. 2. Folding kinetics of a 3-microdomain (ABC) polypeptide (see Appendix B). The probabilities of each of the 8 states, Equations B-10–B-17, are plotted versus time. Table 2 shows the connection between the state labels (state 1–8) and the microdomain pairs (AB, AC, BC) that are present in that state. The kinetic parameters are in inverse time units and have been chosen so that each state has a distinct time dependence (see Table 3 for parameter list).

quantitative and yield structural and kinetic information that can be compared with experimental information. Comparisons are made with the available experimental data. In this section, we concentrate on the diffusion-collision model and defer discussion of other models of folding and their relation to the diffusion-collision model to the next section. The other models, in general, have not been developed sufficiently by their authors to address the range of consequences/predictions described below for the diffusion-collision model. We comment briefly in the next section on the relation of some of the other models to experiments, where comparisons can be made.

Qualitative aspects

Existence and nature of microdomains

Almost 2,000 protein structures have been determined by X-ray crystallography and NMR (Bernstein et al., 1977). A number of analyses (e.g., Levitt & Chothia, 1976; Lesk & Rose, 1981) have shown that most of the structures are composed of helices, sheets, and connecting turns, with helices and sheets accounting for 50–75% of the amino acids on the average. It is reasonable to suppose that in many cases the segments of secondary structure correspond to microdomains, though hydrophobic clusters (e.g., hydrophobic amino acids close to each other in the amino acid sequence) could also be involved. The utility of microdomains for solving the conformational search problem and their identification with the elements of secondary structure would give secondary structure a role in addition to its contributions to the stability of proteins.

Table 2. *Data for 3-microdomain protein*

State	AB pair	AC pair	BC pair	Pairings
1	No	No	No	None
2	Yes	No	No	AB
3	No	Yes	No	AC
4	Yes	Yes	No	AB, AC
5	No	No	Yes	BC
6	Yes	No	Yes	AB, BC
7	No	Yes	Yes	AC, BC
8	Yes	Yes	Yes	AB, AC, BC

Attempts to predict secondary structure in peptide chains are not completely successful (e.g., Kabsch & Sander, 1983; Holley & Karplus, 1989; Zang et al., 1992; Rost & Sander, 1993). One possible reason for this is that the database of known structures is limited (Rooman & Wodak, 1988). However, it appears more likely that tertiary interactions between amino acid residues not near to each other in the linear sequence are important in stabilizing secondary structural elements (Bryugelson & Wolynes, 1987). This implies that the existence of stable secondary structure, as observed from the native coordinates, requires long-range stabilizing interactions between portions of the polypeptide chain not near each other in the linear sequence. This would clearly be consistent with a diffusion-collision mechanism in which the collision and coalescence of elementary microdomains contributes to the stabilization of the secondary structure. It would also contribute to the cooperative nature of the folding transition.

When the diffusion-collision model was proposed (Karplus & Weaver, 1976), there was little experimental evidence for the existence of secondary structure in relatively short polypeptide fragments in aqueous solution under physiological conditions.

Table 3. *Forward transitions for 3-microdomain protein*

Transition	Rate constant	Microdomain pair formed	Rate constants for Figure 2	Rate constants for Figure 3A	Rate constants for Figure 3B
$1 \rightarrow 2$	k_{12}	AB	9	1	7
$1 \rightarrow 3$	k_{13}	AC	12	9	1
$1 \rightarrow 5$	k_{15}	BC	15	2	9
$2 \rightarrow 4$	k_{24}	AC	3	5	1
$2 \rightarrow 6$	k_{26}	BC	3	5	1
$3 \rightarrow 4$	k_{34}	AB	3	1	3
$3 \rightarrow 7$	k_{37}	BC	3	1	3
$5 \rightarrow 6$	k_{56}	AB	3	5	1
$5 \rightarrow 7$	k_{57}	AC	3	5	1
$4 \rightarrow 8$	k_{48}	BC	1	3	3
$6 \rightarrow 8$	k_{68}	AC	1	5	3
$7 \rightarrow 8$	k_{78}	AB	1	3	3

However, it has now been established that small peptide fragments can have marginally stable secondary structure in aqueous solution. Studies of helix stability of the S-peptide of ribonuclease (Shoemaker et al., 1985; Mitchinson & Baldwin, 1986), the C-peptide of ribonuclease A (Osterhout et al., 1989), and other helix-forming peptides (Marqusee et al., 1989; Chakrabartty et al., 1991; Scholtz et al., 1992) have shown that helices can readily exist in short peptides. For example, Marqusee et al. (1989) found that short peptides that contain only alanine and lysine or alanine and glutamate form surprisingly stable monomeric helices in water, and Chakrabartty et al. (1991) found that the ratio of the helix propensities (identified with s, the helix propagation parameter) of alanine to glycine is about 100 in substitution experiments with a 17-residue reference peptide containing alanine and lysine. Peptide fragments comprising the complete sequences of myohemerythrin, a 4-helix bundle protein (Dyson et al., 1992a), and plastocyanin, a mainly β-sheet protein (Dyson et al., 1992b), have recently been synthesized and their conformational preferences examined using proton NMR and CD. In both cases, there is evidence for some secondary structure in the individual fragments in aqueous solution with considerably less in the latter protein. In plastocyanin, the lack of turn propensity and the evidence for extended structure in fragments that are β-strands in the folded state are consistent with a diffusion-collision kinetic mechanism for folding, with collision coalescence being required for stabilization of individual strands. It is a misconception that the diffusion-collision mechanism requires high secondary-structure content in the microdomains. The intrinsic secondary-structure propensities of microdomains affect the rates, not the basic folding mechanism. Studies of folding of immunogenic peptide fragments of proteins in aqueous solution and related peptide models (Blond & Goldberg, 1987; Dyson et al., 1988a, 1988b) have also shown that peptide fragments can have significant stability. Quantitative determinations of secondary-structure equilibrium constants for peptide fragments could be used in evaluating 1 contribution to β, the microdomain–microdomain coalescence probability used in determining a diffusion-collision folding rate. The stability data have been reviewed recently (Wright et al., 1988; see also Baldwin, 1989). Simulation studies concerning secondary-structure stability supplement the experimental data (Brooks, 1993).

Secondary structure forms before tertiary structure

A diffusion-collision folding mechanism requires that the microdomains, which are elements of secondary structure, while forming transiently from the random-coil polypeptide chain and dissipating, diffuse together and collide in order to be stabilized fully as a part of the tertiary structure. There is now considerable evidence concerning the early formation of secondary structure. Kinetic CD studies have indicated that secondary structure is formed before the final tertiary structure during the folding of RNase S (Labhardt, 1984), α-lactalbumin (Kuwajima et al., 1985; Gilmanshin & Ptitsyn, 1987), lysozyme (Kuwajima et al., 1985), carbonic anhydrase (McCoy et al., 1980; Dolgikh et al., 1984; Semisotnov et al., 1987), cytochrome *c* (Kuwajima et al., 1987), β-lactoglobulin (Kuwajima et al., 1987), and the *Escherichia coli* trp aporepressor (Mann & Matthews, 1993). In many of these proteins, it is presumed that a molten globule intermediate is formed that has some of the secondary structure of the final protein. However, except for some of the NMR studies described below, there is no direct evidence concerning the nature of the secondary structure (e.g., whether it is native-like or not). The recent synthesis of fragments comprising the complete sequences of myohemerythrin (Dyson et al., 1992a) and plastocyanin (Dyson et al., 1992b) has shown that transient secondary structure exists in fragments that have well-defined secondary structure in the native state.

It is of interest that in a molecular dynamics simulation of the folding of crambin (Brunger et al., 1986) in the presence of NOE constraints that could be obtained from NMR spectra, it was observed that only if the secondary structure formed before tertiary structure during the folding process was the correct final structure obtained. This is, of course, an artificial folding simulation, but the demonstration that early secondary-structure formation simplifies the search problem is likely to be germane to the real folding process.

Existence of intermediates

The diffusion-collision model envisions folding to proceed by the formation of microdomain pairs and subsequent formation of higher aggregates of the elementary microdomains. These multimicrodomain clusters are likely candidates for kinetic intermediates in the folding process. A kinetic intermediate is a

well-defined species that accumulates transiently during folding. The accumulation occurs as a consequence of the rate constants for the formation and disappearance of the species. It is usually assumed (see, for example, Bycroft et al., 1990; Serrano et al., 1992) that such species are intermediates in the sense that they are local minima on the free energy surface for the reaction. However, direct evidence for this (e.g., trapping of an intermediate at low temperatures) has been difficult to obtain. In apomyoglobin, where a partly unfolded species has been studied in some detail (Hughson et al., 1990, 1991), it is not clear that the "intermediate" is on the folding pathway. A recent study using CD and NMR (Jennings & Wright, 1993) suggests that an intermediate on the apomyoglobin folding pathway is very similar to that studied by Hughson et al. (1990). Molten globule intermediates have been identified as such (Ptitsyn, 1987; Ptitsyn et al., 1990), but there are few details concerning their structure.

We describe in what follows some of the recent NMR data concerned with the protection of peptide NHs from proton exchange, which is presumed to be related to hydrogen bond formation in secondary structural elements. Although such experiments are providing important structural information concerning intermediates between the folded and unfolded state, it should be noted that there is an inherent limitation to the experiments (Creighton, 1992). They can only study protons that are slowly exchanging in the native state under some conditions. Because most of such protons are in secondary structural elements, the experimental data are limited to determining whether such secondary structure exists. It is more difficult to obtain information about loop regions to demonstrate that secondary structure is formed first.

Roder et al. (1988) have studied the refolding of GuHCl-denatured cytochrome *c* in which the reformation of hydrogen bonds was measured by a pulsed labeling NMR technique based on NH proton exchange and its control by pH (Roder & Wuthrich, 1986). They found (Roder et al., 1988) that the first H bonds seen in refolding are in the N- and C-terminal helices (2 possible microdomains). The protection of the NHs in the 2 helices occurs essentially simultaneously with a time constant of about 20 ms; in addition, there appear to be slower contributions to the protection of these helices. The experiments suggest that the helical microdomains have been stabilized by association to form a folding intermediate involving the 2 helices far apart along the polypeptide chain. Recently a 2-peptide study (1–38 plus heme and 87–104) corresponding to the N- and C-terminal portions of cytochrome *c* has confirmed that helical structure is induced by formation of the binary complex (Kuroda, 1993). The folding of cytochrome *c* is discussed further below.

Briggs and Roder (1992) have investigated the hydrogen-bonded structure in the folding reaction of ubiquitin, a small cytoplasmic protein with an extended β-sheet and an α-helix surrounding a hydrophobic core, using H–D exchange labeling with rapid mixing and 2-dimensional NMR. They found that the amide protons in the β-sheet and the α-helix, as well as protons involved in hydrogen bonds at the helix/sheet interface become 80% protected in an initial 8-ms folding phase. This indicates that the 2 elements of secondary structure form and associate in a common cooperative folding event (coalescence of microdomains). Somewhat slower protection rates for residues 59, 61, and 69 provide evidence for the subsequent stabilization of a surface loop.

Dyson et al. (1992a, 1992b), in their studies of synthesized fragments of myohemerythrin and plastocyanin comprising the entire proteins, found evidence for transient secondary structure in fragments that have secondary structure in the fully folded molecule. This is consistent with a diffusion-collision folding mechanism with the native secondary-structure elements (α-helices in myohemerythrin and mainly β-strands in plastocyanin) as the microdomains.

Jeng and Englander (1991) have investigated cytochrome *c* at low pH and low salt concentration where it is non-compact. By hydrogen exchange labeling, 2-dimensional NMR, and CD measurements, they observed helical structure that they termed "stable submolecular folding units" composed of more than 1 helical segment held together by "soft packing" (presumably hydrophobic in nature, rather than the interlocked side-chain "hard packing" characteristic of the native state). With the helices as the elementary microdomains, these units of structure would be microdomain pairs and multimicrodomain clusters, as expected for a diffusion-collision folding mechanism.

Oas and Kim (1988) designed a single disulfide-bonded peptide pair ($P_\alpha P_\beta$) to mimic the first crucial intermediate along one of the pathways for the folding of BPTI and showed by NMR that it contained secondary and tertiary structure similar to that in the native protein. This study confirmed earlier work by States et al. (1980, 1984) concerning the existence of secondary structure in trapped intermediates of BPTI. In fact, the portions of BPTI removed by Oas and Kim correspond to the random coil portions of the complete BPTI chain in the 1-disulfide species. In the intact protein, the helical (P_α) and sheet (P_β) regions are in non-neighboring sections of the polypeptide chain and would need to diffuse together (P_α has an α-helix and P_β is mainly β-sheet) to collide in order to form the disulfide bond (30–51) present in such an intermediate state.

Hughson et al. (1990) have characterized a partially folded apomyoglobin species at pH 4.2 and determined the slowly exchanging peptide NH protons by NMR. They found that some protons in the A, G, and H helix regions are protected from exchange, whereas protons in the B and E helix regions exchange freely. Recent experimental results show that the AGH helix triplet structure (with a bit of B) folds within 5 ms and appears to be the same as the equilibrium intermediate (Jennings & Wright, 1993). The result is of interest because helices A, G, and H are all in contact in the native structure. In the folded protein, the GH helix pair has a solvent-accessible area (Lee & Richards, 1971) loss of 852 $Å^2$ upon folding, the AH helix pair an area loss of 602 $Å^2$, and the AG pair an area loss of 276 $Å^2$. This shows that the main native contacts in the AGH helix triplet are between helices A and H and between helices G and H. The experiment also demonstrates that contiguity is not necessary for the stability of partially folded structure. As already mentioned, such behavior is a possible consequence of a diffusion-collision folding mechanism. However, the specific partially folded species found by Hughson et al. (1990) does not appear in the diffusion-collision folding simulation of myoglobin (Bashford et al., 1988). In that simulation, because of the absence of detailed information on the individual helix stabilities, equal βs were assumed for the helical microdomains. This leads to near-neighbor interactions first (rather than N-terminus–C-terminus interactions, which would be dominant early with substantially larger βs for helices A, G, and H) with physically appropriate choices of the geometrical parameters. Two major folding path-

ways were found that pass through the BDE and FGH helix clusters, respectively, which subsequently coalesce. Thus, the most important intermediates were formed by coalescence of adjacent regions of the polypeptide chain. The A helix, which interacts hydrophobically with the H helix, coalesces at a late stage, near the end of the kinetic pathway (see Table 1 for some geometrical kinetic parameters concerned with the AH and GH helix pair folding rates).

Fersht and coworkers (Bycroft et al., 1990; Matouschek et al., 1990) have detected a transient species on the folding pathway of barnase using kinetic experiments on engineered mutants. By use of linear free energy-like relations, they inferred which of the interactions present in the native state of barnase were formed in the intermediate. In an ^{1}H NMR peptide proton exchange experiment, Fersht and coworkers (Bycroft et al., 1990) found that the refolding of barnase is a multiphasic process in which the secondary structure in α-helices and β-sheets (and some turns) is formed more prior to the overall folding. These kinetic results are in accord with the diffusion-collision-coalescence model (Serrano et al., 1992).

Miranker et al. (1991), using NMR protection experiments based on competition between hydrogen exchange and refolding, have found transient species in the refolding of lysozyme that are similar to the molten globule state of α-lactalbumin (Kuwajima, 1989). They found that the 2 structural domains of lysozyme are distinct folding domains, in that they differ significantly in the extent to which compact, probably native-like structure is present in the early stages of folding; i.e., the α-helical domain folds first and the β-sheet domain is formed subsequently. More recent stopped-flow experiments indicate that the formation of the α-helical domain has 2 stages, with a loose α-helical cluster preceding the formation of the native-like α-helical domain (Radford et al., 1992). Data concerning the relation between the molten globule state of α-lactalbumin and the transient species observed in lysozyme has been obtained by Kuwajima et al. (1985). There is some indication that formation of stable secondary structure is accompanied by partial assembly of the hydrophobic core.

Existence of pathways

In the diffusion-collision model, folding pathways depend on the properties of microdomains. There is no reason to expect all proteins to have the same types of microdomains and, therefore, there is no reason to expect all proteins to have the same number and types of folding pathways leading to the native structure. Some proteins will have essentially 1 major pathway, whereas other proteins may have several more or less equal pathways. This is a likely consequence of a microdomain-dominated folding mechanism. It is, therefore, a mistake to rely on folding pathway evidence from a single protein to infer properties of folding pathways in general. This caveat applies to the interpretation of the recent results on the folding of barnase (Bycroft et al., 1990; Matouschek et al., 1990) by Baldwin (1990). He suggested that the existence of 1 preferred pathway in the folding of barnase is sufficient to eliminate the jigsaw-puzzle model (Harrison & Durbin, 1985) as a viable folding model. The same caveat applies to the elegant experiments by Creighton (1988) on the folding pathways of BPTI. It has recently been shown (Staley & Kim, 1992) that essentially complete folding of BPTI is obtainable with only the [5–55] disulfide bond intact. This demonstrates that folding can occur without formation of "incorrect" disulfide bonds, though the latter clearly contribute under certain conditions.

One case where multiple folding pathways have been suggested is cytochrome *c*. As followed by H-exchange labeling and proton NMR (Roder et al., 1988), about 40% of the NH protons associated with the N- and C-terminal helices are protected rapidly, whereas 60% of the same protons are protected much more slowly. An interpretation of the results, as pointed out by Roder et al. (1988), is that there are alternative pathways that do not involve coalescence of the N- and C-terminal α-helices as the first step. Other proteins where alternative pathways have been detected are lysozyme (Radford et al., 1992) and ribonuclease (Baldwin, 1990).

With modern protein engineering techniques, it is possible to engineer microdomains to give intermediates that occur more or less prominently than those occurring naturally. Thus, certain pathways to the native structure could be increased, decreased, and otherwise manipulated for study. It would be interesting, for example, to synthesize the G helix–turn–H helix from myoglobin and study its properties. In this regard, a de novo designed 2-helix hairpin has recently been synthesized (Fezoui et al., 1994). Preliminary characterization indicates that the peptide is monomeric over a wide range of concentrations, and that it is mainly helical with NOE interactions between residues that were designed to be on separate helices.

In the recent diffusion-collision folding simulation of myoglobin (Bashford et al., 1988), although all the important helix-helix interactions observed in the crystal were included, 2 pathways dominated the folding because they were more important than alternative pathways. This was mainly a consequence of geometry because the β values for all helices were taken to be equal (but different at different stages in the folding) in the absence of information on the individual helix stabilities (see Table 1 for some of the geometrical parameters). The results show that even with all microdomain stabilities the same, certain pathways may make the dominant contribution. Having differing stabilities for different microdomains would, in many cases, modify the "geometrically" favored pathways.

Possible incorrectly folded intermediates

It is a likely consequence of a diffusion-collision folding mechanism that pairs of microdomains that do not interact in the native structure collide and transiently associate. Ordinarily, such states will be unstable compared to correct microdomain-microdomain pairings and will be kinetically unimportant for folding pathways.

Studies on BPTI (Creighton, 1978) have shown that formation of non-native disulfide bonds can be an important part of the folding process in this protein, and, thus, that misfolded intermediate states occur in at least 1 protein. However, in the misfolded intermediates observed by Creighton, it is not known whether 2 microdomains are making incorrect contacts. Also, it has been shown that the disulfide bonds themselves are not necessary for the formation of native-like tertiary structure in BPTI (Marks et al., 1987; Nilsson et al., 1990; Staley & Kim, 1992) and that proteins homologous to BPTI do not need to pass through misfolded intermediates to reach the native structure (Hollecker & Larcher, 1989). In fact, BPTI itself does not need to go through the misfolded states, though it does under certain conditions (Creighton, 1992; Weissman & Kim, 1992). These experiments are also important in showing that the folding path-

way or pathways found in 1 protein do not necessarily apply to other proteins, even when they are homologous (see discussion of pathways, above).

Molten globule state

After a series of diffusion-collision steps, the microdomains are expected to have packed loosely against one another, but without the complete local interdigitation of side chains required for the native state. Consequently, the backbone of the protein would have a conformation close to the native structure, and the side chains would be relatively free to move about. This stage of diffusion-collision folding is suggestive of the molten globule state (Ptitsyn, 1987). Examples are cytochrome *c* (Ohgushi & Wada, 1983; Jeng & Englander, 1991; Kuroda et al., 1992), α-lactalbumin (Kuwajima, 1989), and lysozyme (Miranker et al., 1991). It should be noted that in the diffusion-collision model, the correct secondary structure forms as part of the transition to the molten globule state because it involves association of a number of elementary microdomains. However, the detailed structure of the molten globule (i.e., whether the secondary structural elements have exactly the native orientation) is not known in most cases. For α-lactalbumin (Baum et al., 1989) and ubiquitin (Harding et al., 1991) NMR data on the molten globule state are available. In the former, at least 2 helices (B and C) are present and they have the relative orientations of the native state as indicated by aromatic chemical shifts and NOEs; in the latter, several β-strands have been shown to have their native conformation.

Quantitative results

Folding rate

Kinetic experiments have shown that proteins are able to find their native structures in the time range 10^{-3}-10^{2} s. Although the long time phenomena involve processes such as proline isomerization, the basic folding mechanism appears to be on the order of a millisecond or longer. Karplus and Weaver (1976, 1979) have shown that Equation 1 yields correct estimates of folding times with β values in the range 10^{-3}-10^{-4} (see below).

Equation 1 has recently been applied by Bashford et al. (1990) to the refolding of GuHCl-denatured cytochrome *c* (Roder et al., 1988). Application of the diffusion-collision model to this process with estimates of the parameters in Equation 1 yielded a value of 28 ms for τ_f if a value of $\beta = 10^{-4}$ is used; the experimental value is 20 ms. In this treatment, all βs were set equal. This means that each helix would have a 1% chance of being correctly folded and oriented when a collision that yields coalescence occurs. The model gives a satisfactory interpretation of the rate of formation of the interaction of the N- and C-terminal helices. It does not explain why they are the first to come together and be stabilized by coalescence (there are 2 other helices and 3 helix–helix contacts). A possible reason is variation in the parameter β. If the other helices are substantially less stable, the β values for the folding steps associated with them may be small enough to cause those steps to be slower than the C and N helix pair coalescence. This behavior is illustrated in Figure 3, which shows the folding kinetics of a hypothetical protein (or portion of a larger protein) with 3 microdomains (see Appendix B). The microdomains are labeled A, B, and C in order along the chain. The time course of the appearance of early intermediates depends on the stabilities (β) of the elementary microdomains. In particular, the microdomain pair AC, involving the 2 end microdomains A and C (analogous to the N- and C-terminal helices in cytochrome *c*), can appear before the nearest-neighbor pairs AB and BC (Fig. 3A). In previous applications of the model to actual proteins (Bashford et al., 1984, 1988), no information was available with which to differentiate the stabilities of the microdomains in nascent form. Thus, all elementary microdomains were assumed to have identical values of β. To make more quantitative studies, it would be of considerable interest to have measurements of the individual helix stabilities for use in analyzing the folding kinetics of a protein that has been studied experimentally.

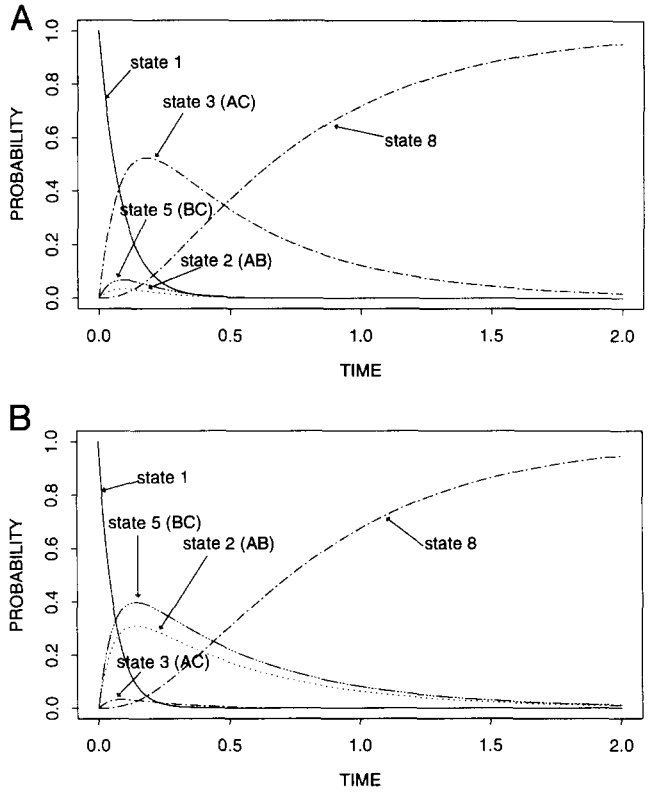

Fig. 3. Folding kinetics of a 3-microdomain (ABC) polypeptide (see Appendix B). The probabilities of states 1, 2, 3, 5, and 8, Equations B-10, B-11, B-12, B-14, and B-17 are plotted versus time. Table 2 shows the connection between the state labels and the microdomain pairs (AB, AC, BC) that are present in that state. The kinetic parameters are in inverse time units. **A:** The case in which the probability β_B that microdomain B has its secondary structure is much smaller than the probabilities β_A and β_C. The coalescence probability dominates the nearest-neighbor geometry to enhance the early production of the microdomain pair AC. **B:** The case in which the probabilities are equal and the geometry effects of nearest-neighbor coalescence dominate to enhance the early production of the microdomain pairs AB and BC.

An application of the diffusion-collision model to myoglobin folding (Bashford et al., 1988) used the helices observed in the crystal structure as the microdomains (see below). The rates of formation of intermediate states as well as of the final state were studied as a function of β.

Solvent viscosity dependence

The size of the elementary units (i.e., the microdomains) is such that Brownian dynamics should be adequate for the de-

scription of their motions. The forces governing the motion of microdomains are the specific interactions between microdomains and the random collisions with the solvent and with other parts of the polypeptide chain. Thus, the kinetics of the diffusion-collision model are expected to be dependent on the solvent viscosity; the specific dependence is through the parameters D and L in Equation 1. At long distances, which includes most of the diffusion space, the details of the specific forces can be neglected. At close range, they become important and may give rise to attractive interactions and to barriers prior to coalescence. Except for the latter, which can be subsumed in β (Karplus & Weaver, 1979), the diffusive motion is expected to govern the time scale of coalescence.

In the chemical kinetics approximation to diffusion-collision folding, the time dependence of the probabilities of the states of the system (initial state, intermediates, and folded state) have the form of a sum of decreasing time-dependent exponentials modulated by time-independent amplitudes. The exponential time constants are linear combinations of rate constants and the amplitudes are ratios of combinations of products of rate constants (see Appendix B for an example involving 3 microdomains). The effect of solvent viscosity is determined by the viscosity dependence of the rates. The rate parameter τ_f in Equation 1 is dependent on solvent viscosity through the diffusion coefficient, D, which is inversely proportional to η and through the parameter α (Equation A-11), which appears in the expression for L (Equation A-10). If α is independent of viscosity, that is, if λ_1 and λ_2 are inversely proportional to viscosity (as found in Brownian dynamics simulations of helix–coil and sheet–coil transitions; McCammon et al., 1980; Yapa et al., 1992; Schneller & Weaver, 1993), then τ_f is directly proportional to solvent viscosity. The solvent viscosity dependence of backward rates is probably small, because microdomain–microdomain dissociation (unfolding) due to thermal fluctuations would depend on internal frictional effects (Karplus & Weaver, 1979). An example of the effect of viscosity on intermediates is shown in Figure 4. This figure shows the time dependence of the intermediate state 2 (microdomain pair AB) in the 3-microdomain system described in Appendix B as a function of solvent viscosity. As the viscosity increases, the peak is both expanded (the intermediate lives longer) and moved to later times, as expected from the analytical time dependence of this state given in Appendix B (Equation B-11).

There are several experiments that address the question of the viscosity dependence of folding rates. Haas et al. (1978) investigated the kinetics of the fluorescence decay of the energy donor in a homologous series of oligopeptides, each containing at its ends a donor and acceptor of electronic excitation energy. They found a clear dependence of the rate on solvent viscosity as well as an "internal friction" contribution. This is in accord with expectations for the diffusion-collision mechanism.

Tsong and Baldwin (1978) studied the kinetics of folding of the 2 forms of denatured ribonuclease A (with all disulfide bonds intact) as a function of solvent viscosity by adding either sucrose or glycerol. They originally interpreted their results as showing no solvent viscosity dependence of the kinetics of folding. This experiment has been cited by a number of authors (Kim & Baldwin, 1982; Go, 1983) as evidence against the diffusion-collision model. Their interpretation of the experiment was incorrect (R.L. Baldwin, pers. comm.) and the conclusion from their experiment is no longer valid. In a subsequent experiment on the same system, Tsong (1982) found a fast reaction whose kinetics depend strongly on the solvent viscosity.

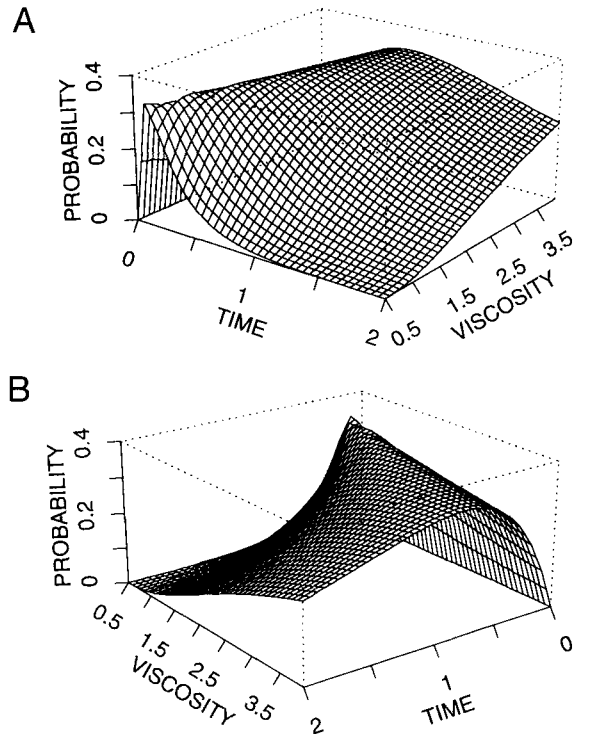

Fig. 4. Folding kinetics of state 2 (AB) of a 3-microdomain (ABC) polypeptide (see Appendix B and Table 2) as a function of solvent viscosity. The same results are plotted from 2 different viewpoints in **A** and **B**. The probability of state 2 (Equation B-11) is plotted versus time for different values of the solvent viscosity in normalized units (a viscosity of 1 corresponds to room temperature and water). The kinetic parameters in Equation B-11 are in inverse time units and are taken to be inversely proportional to the solvent viscosity.

Hurle et al. (1987) have studied the domain association reaction in the urea-induced unfolding and refolding of the α subunit of tryptophan synthase of *E. coli*. They found a viscosity dependence in the unfolding and refolding after corrections for the stabilizing effect of the viscosity enhancer, sucrose, on the folding reaction. Chrunyk and Matthews (1990) have also studied the role of diffusion in the rate-limiting step in the folding of the α subunit. When the effects of the cosolvent on protein stability were taken into account, the rates were found to show an inverse dependence on the viscosity of the solvent, clearly demonstrating that the rate-limiting folding unit association/dissociation reaction in the α subunit of tryptophan synthase involves a diffusional process.

In summary, all the experiments performed (and correctly interpreted) to determine the viscosity dependence of folding have shown that one is present, in accord with the diffusion-collision model. Because other models have not been formulated on a kinetic basis, their viscosity dependence is not available, though it is not unreasonable that certain stages (e.g., folding after nucleation) would be diffusive in character. It clearly would be of considerable interest to extend such viscosity studies to kinetic measurements by NMR and CD from which more detailed structural information is available.

Concentration of kinetic intermediates

When the diffusion-collision model is applied to the folding of a protein, the output of the simulation is a set of time-dependent probabilities for all states of the system that are included in the model. This makes it possible to determine the concentrations of the kinetic intermediates as a function of time. A model simulation of this type has been made for myoglobin (Bashford et al., 1988; see Fig. 2 for a 3-microdomain example). It shows that different intermediates have very different time-dependent concentrations. Moreover, the concentrations were shown to depend on the degree to which the folding reaction goes to completion (formation of the native state), as well as on the microdomain probability parameter β. The latter is an essential parameter, which can change relative rates and concentrations if the different microdomain pairs have different β values.

Effect of β on rates and pathways

Folding rates depend on the coalescence probability β through Equation 1. Among the physical parameters affecting the rate, β can potentially have the largest variation. β is also the parameter that is least well known experimentally (most of the other parameters in the diffusion-collision folding picture are geometrical in nature). Therefore, in model calculations of folding, β has been used as an essentially free parameter. Values of β near 1 lead to folding on a time scale of nanoseconds. Because the processes leading to stable secondary-structure formation appear to be in the millisecond and submillisecond time range from recent experiments (e.g., Roder et al., 1988), this suggests that β is less than unity. As already discussed, small β values are in accord with the stability of isolated microdomains. Also, different values for different microdomains affect the individual folding pathways, as shown in the model simulations for myoglobin (see above). Recent studies on the helix-coil transition of the S-peptide of ribonuclease (Shoemaker et al., 1985; Mitchinson & Baldwin, 1986) have shown that β can be highly variable and sequence specific. This suggests that pathways in different proteins (even from the same family) can be quite different. It is therefore very important to find specific values for the stability contribution to β by experiment or calculations for proteins whose kinetics have been studied. This would permit a more quantitative analysis of folding pathways.

Other models of folding

A range of phenomenological models for protein folding have been proposed. We describe here their relation to the diffusion-collision model to make clear that there are both similarities and differences. This is of particular importance because in some cases experimental papers have tried to distinguish between models whose consequences are the same for the experiments reported.

Six years after the diffusion-collision model of folding was published (Karplus & Weaver, 1976), Kim and Baldwin (1982) suggested what they called the "framework" model in which the H-bonded secondary structure is formed before the tertiary structure. They proposed the framework model as a working model and indicated that it was based on the experimental evidence. As presented, their model was a generalization of the work of Ptitsyn and Rashin (1973, 1975), who considered myoglobin as composed of stable helices and, on the basis of a simple model for hydrophobic interactions between the helices, compared different paths to the most likely folded structure. In a 1989 survey of folding, Baldwin (1989) updated the framework model to incorporate more recent experimental results. He proposed that folding begins with the formation of individual, transient secondary structures (elementary microdomains) that are stabilized by packing against each other (collision and coalescence), and that folding is a hierarchical process in which simple structures are formed first and these interact to give more complex structures. In this form, the description of the framework model is essentially that inherent in the diffusion-collision model proposed in 1976 (Karplus & Weaver, 1976). However, the framework model has not been elaborated to allow for quantitative calculations. Nevertheless, cases cited in the previous section where secondary structure is observed early in folding are consistent with the framework model.

Harrison and Durbin (1985) proposed the jigsaw puzzle model as a qualitative mechanism for folding of the compact domains of proteins. They suggested that proteins fold by large numbers of different, parallel pathways rather than by a single defined sequence of events, and that this would make folding more robust to mutations that do not adversely affect the native structure. These authors point out that folding kinetics of this type may be obtained with the diffusion-collision model (Karplus & Weaver, 1976) if all the elementary microdomains have similar properties and multimicrodomain intermediates of the same size have similar folding and unfolding rates. As is true for the framework model, the jigsaw puzzle model was not used by its authors to investigate quantitative aspects of folding. In accord with the ideas expressed in the jigsaw puzzle model, the number of diffusion-collision pathways increases rapidly with the number of possible interactions between elementary microdomains. For example, a protein with 3 elementary microdomains has 3 possible initial pairings and $3! = 6$ possible separate pathways. With 4 elementary microdomains, there are 6 possible initial pairings and $6! = 720$ pathways. Thus, even a small protein has, in principle, many alternative pathways by which it could reach the native structure. However, many pathways in a typical protein are expected to be unimportant. An example of this behavior is given in the apomyoglobin calculations (see above).

Udgaonkar and Baldwin (1988) stated that the framework (Kim & Baldwin, 1982) and jigsaw puzzle (Harrison & Durbin, 1985) models are fundamentally different models of protein folding. That is not true when one views a protein as a collection of coupled microdomains, as in the diffusion-collision model. In fact, as discussed above, both models emphasize phenomenological aspects of folding that could be limiting cases of a diffusion-collision mechanism. Their applicability to a specific protein would depend on the properties of the individual microdomains and their interactions. In barnase it appears, though the evidence is not direct, that a limited number of pathways contribute to the folding reaction (Matouschek et al., 1990).

The sequential model (Levinthal, 1968) suggests that a portion of the polypeptide chain, unstable by itself and possibly equivalent to a microdomain, serves as a nucleus for chain propagation to obtain the native structure. The model requires the nucleation subunit to be small enough for a random search to find its native structure, and that once this nucleus has the native structure, it makes possible a sequential, essentially independent folding of subsequent amino acids. Such a mechanism

was outlined by Levinthal (1966, 1968). Wetlaufer (1973) and Tsong et al. (1972a, 1972b) have provided a kinetic model for such sequential protein folding. It is not clear how this analog of crystallization would work, because for a nucleus that is small enough, the remainder of the polypeptide chain would still appear to suffer from the Levinthal paradox. In the simplest version of this model, folding would proceed along a unique pathway in which formation of the nucleus would dominate the time dependence. It has been elaborated and modified by Go and coworkers (Go et al., 1980; Go & Abe, 1981; Go, 1983) into the noninteracting local structure, or growth-merge model, and by Kanehisa and Tsong (1978, 1979a, 1979b, 1979c) into the cluster model. A cluster or "embryo," grows until it merges with another growing nucleus. The growth-merge mechanism has elements in common with the diffusion-collision model; a difference appears to be in whether the growth of a single embryo (microdomain) or the coming together (diffusion-collision-coalescence) of 2 or more of these elements governs the overall kinetics of the folding process. In both models, nearest-neighbor microdomains are more likely to coalesce first, whether by growth or via collisions. This is in accord with the distribution of contacts observed between secondary structural elements in proteins (Levitt & Chothia, 1976; Lesk & Rose, 1981). In the diffusion-collision model, it is likely that nearest neighbors will collide first. However, as already mentioned, it is not mandatory that they coalesce first. In fact, in the cytochrome *c* folding experiment (Roder et al., 1988), the N- and C-terminal helices seem to coalesce first.

Moult and Unger (1991) have continued the development of the sequential model with a model of protein folding based on the assumptions that (1) burying of hydrophobic area is the dominant contribution to the relative free energy of a conformation, (2) a record of the folding process is largely preserved in the final structure, and (3) the denatured state is a random coil. They suggest that folding begins by "nucleation" of regions 8–16 residues long (clearly microdomains), followed by propagation and diffusion-collision. They further developed a simulation algorithm, using Monte Carlo methods, to emulate kinetic folding behavior assuming that propagation is the dominant folding mechanism. They nevertheless conclude that diffusion-collision plays a role in the coming together in the final structure of some of their nucleation sites.

It has been suggested (Levitt & Warshel, 1975; Levitt, 1978; Dill, 1985) that a "hydrophobic" collapse takes place before the formation of specific structure. The occurrence of a "collapse" would be an artifact of a change of the solvent conditions at the beginning of a folding experiment. When physiological folding takes place, presumably the protein is formed in a collapsed state. Entropy considerations indicate that the equilibrium probability distribution of a polypeptide chain is not uniform. In particular, fully extended chains will be less likely than more compact (and more abundant) conformations. Equilibrium distributions of end-to-end distances of a series of oligopeptides containing at their ends a donor and an acceptor of electronic excitation energy have been determined by Haas et al. (1975). They found the distributions to be peaked approximately in the center of their range with the longer chains having peaks at larger distances. This is in accord with the simulation of the diffusion and collision of coupled helices (Lee et al., 1987) and the ends of a β-hairpin (Yapa et al., 1992) using Langevin dynamics techniques. We also find that the equilibrium probability distribution is not uniform. Rather it has 1 or more peaks in probability space and falls off at the 2 ends, where the microdomains are in contact and at the limit of the fully extended chain. Although the number of conformations in such a collapsed state would be significantly reduced (Dill, 1985), it is not sufficient to overcome the Levinthal paradox. Consequently, some search mechanism is still necessary to find the native structure. One possibility is that some of the compact conformations are sufficiently native-like to allow rapid rearrangement to give the precise native structure. More likely is the alternative that diffusion-collision-type folding takes place at this stage.

In addition to the use of phenomenological models, another attempt to examine the details of the folding mechanism is based on simulations of a quasiparticle (bead) model of the polypeptide chain on a regular lattice. Such lattice models permit sampling of larger ranges of conformations and time scales than off-lattice models, though they are more limited than the diffusion-collision model. Skolnick and coworkers (Sikorski & Skolnick, 1990; Skolnick & Kolinski, 1990a, 1990b; Skolnick et al., 1990) have used tetrahedral (diamond) lattices as well as more complex lattices to represent the protein. Most of their work on the folding process has been concerned with model systems that form β-barrels or α-helical bundles (Sikorski & Skolnick, 1990; Skolnick & Kolinski, 1990a); they have examined a more realistic model for plastocyanin (Skolnick & Kolinski, 1990b). The simulations of the folding pathways of 6-stranded β-barrels (Skolnick & Kolinski, 1990a) were restricted to local moves (3 and 4 bond flips) to avoid "producing a distorted time scale." Secondary-structure propensities (β-strand, β-turn) and distributions of hydrophobic and hydrophilic residues for a 74-amino acid chain were used that, in previous equilibrium studies (Skolnick et al., 1990), had been shown to yield an all-or-none transition to the "native" state. Because full enumeration to find the lowest energy conformation and determine its uniqueness was not possible, there is a question as to whether a metastable minimum is being considered. The results of a number of folding simulations showed no initial hydrophobic collapse and folding began at one or another of the predetermined β turns of the native structure. "Collapse" took place simultaneously with the formation of tertiary structure. An intermediate was formed that contained part of the native structure (e.g., 4 of 6 correctly positioned strands with 50 of 74 native contacts). In the early stages of folding (i.e., from the initiation step to the formation of the intermediate), several pathways were found that led to nearly the same intermediate.

The authors emphasize that the β-sheets are formed by "on-site" construction rather than diffusion-collision, but that is implicit in the model because strand diffusion is not allowed in the local move Monte Carlo algorithm (Skolnick & Kolinski, 1990a). In a subsequent paper on 4-helix bundles (Sikorski & Skolnick, 1990) that uses analogously biased sequences (i.e., helix and turn regions plus choices of hydrophobic and hydrophilic residues that yield all-or-none transitions to the 4-helix bundle as the native state in equilibrium simulations), they introduce helix translation and rotation steps so as to provide a more realistic test of the diffusion-collision model versus the on-site assembly model. The α-helical bundle results are described as being similar to those forming β-barrels. On-site construction is found to be the dominant mechanism of helix–bundle formation. However, the probability of the translation and rotation steps was made to be so small (by the choice of parameters in the Monte

Carlo algorithm) relative to the local moves, that such a result is not surprising. In an attempt to show that their lattice Monte Carlo simulation methods are realistic, Rey and Skolnick (1991) carried out a Brownian dynamics simulation of the folding of an α-hairpin. However, their simulation methods raise a number of questions because they used time steps of $\approx$3 ps rather than time steps of 0.03 ps previously found to be necessary in Brownian dynamics of simplified residue chains (McCammon et al., 1980; Lee et al., 1987; Yapa et al., 1991; Schneller & Weaver, 1993), in order to have the forces be approximately constant during a time step. They assigned an attractive force between residues in the 2 helices that would be in contact in the folded hairpin, whether or not the helices are formed. In addition, they restricted the diffusion-collision mechanism to refer only to diffusing together of preformed helices rather than unstable, transient ones as the diffusion-collision model (Karplus & Weaver, 1976) proposes. The first point could lead to unreliable results overall because the time step is of the same order as transition rates to and from the coil state (Yapa et al., 1991; Schneller & Weaver, 1993). The second point favors on-site construction, as in their Monte Carlo lattice simulations, and their third point omits an essential aspect of the diffusion-collision model. Thus, this work cannot be considered a "test" of on-site construction versus diffusion-collision as a kinetic folding mechanism nor a validation of the on-site construction mechanism itself. Recently Godzik et al. (1992) have applied the same type of structurally biased Monte Carlo model, though the details are somewhat different from the earlier simulations, to the folding of TIM barrels. The folding process is described as "starting from several independent initiation sites that grow by an on-site mechanism and are later joined by diffusion."

It should be noted that it is very difficult to draw conclusions from this type of simulation concerning the details of the folding mechanism because no attempt is made to determine the relative probability of different Monte Carlo moves in accord with the type of protein motions that they represent. Also, the results obtained depend strongly on the interactions that stabilize interresidue contacts. In the simulations discussed, individual residues have a tendency to coalesce, independent of whether the chain has the correct secondary structure or not. This contrasts with the diffusion-collision model, in which the correct secondary structure in the coalescence region has to exist before a stable contact can be formed. Thus, although the simulations of Skolnick et al. (1990) are of interest as a model of protein folding, they do not provide a "test" of the diffusion-collision model.

An alternative lattice approach makes use of even simpler lattices that permit exhaustive enumeration of the possible states of the polypeptide chain (Sali et al., 1994). This makes possible the investigation of models that are not connected to a particular protein because the lowest energy state for a given model can be determined directly. Monte Carlo folding studies with a random interaction model of this type for a 27-bead polymer and simple cubic lattice have been used to investigate the requirements for rapid folding. There are a total of 10^{18} possible conformations in this lattice system so that the Levinthal paradox is present; i.e., a complete search would take an astronomical number of Monte Carlo moves. It has been found that a necessary and sufficient requirement for rapid folding is that the ground state be significantly lower in energy than the neighboring states. The system collapses rapidly to a globular state and is then able to find one of the many transition states that fold rapidly to the unique native state by a random search of the space accessible in the compact globule. In real protein systems of more than 100 residues, the collapsed globule still is too large for such a search procedure (Dill, 1985). It is likely therefore that there is a necessary hierarchy in the folding of proteins with an earlier search step being nascent secondary-structure formation and collapse. In the small model system, the probability of secondary structure is small and appears not to be necessary for folding.

Goldenberg and Creighton (1985) proposed the cardboard box model as a qualitative mechanism for the folding and unfolding of small proteins. They suggested that the rate limiting transition state in both unfolding and folding is a high-energy distortion of the fully folded state, and that partially folded intermediates are important, but their formation is not rate limiting. The analog is the distortion required for the interleaving of the 4 flaps to close the top of a cardboard box. The model is based, in part, on the observations by Creighton (1978) of nonnative disulfide intermediates occurring in the folding of BPTI and the conclusion that the rate-limiting step in both folding and unfolding proceeds through either of 2 intermediates with nonnative second disulfide bonds (30–51, 5–14) and (30–51, 5–38). However, it has recently been demonstrated (Weissman & Kim, 1992) that all the well-populated BPTI folding intermediates contain only native disulfide bonds, and that essentially complete folding of BPTI is obtainable with only the [5–55] disulfide bond intact (Staley & Kim, 1992). This shows that folding can occur without the formation of "incorrect" disulfide bonds, so the main rationale for the cardboard box model appears no longer to be applicable. Nevertheless, the model raises an important question about the folding mechanism: does the rate-determining step in folding occur early enough so that diffusion-collision dominates the time scale, or does the rate-determining step occur after the collision and coalescence of the microdomains, e.g., by a rearrangement step that has a relatively high potential barrier (a step that could involve surface diffusion of microdomains on other microdomains, as well as interdigitation). In the latter case, the diffusion-collision mechanism would still provide a solution to the Levinthal paradox (Levinthal, 1968), but it would not govern the overall rate of folding.

Concluding discussion and future directions

The diffusion-collision model was suggested in 1976 as a model for the process of protein folding based on the collective multiresidue (microdomain) dynamical interactions that appear in a series of diffusion-collision steps as a small globular protein folds toward its native structure. It is of particular interest because it has consequences that can be observed experimentally. Several of these are important elements of the mechanism that were initially thought to be incorrect. In the intervening years as experimental studies of protein folding have improved, many of these elements have in fact been confirmed. They include (1) the existence of secondary structure in small peptides and protein fragments, (2) the dependence of some folding rates on the solvent viscosity, and (3) the importance of native microdomain–microdomain interaction pathways (e.g., early helix–helix interactions) in folding.

Nevertheless, much more experimental information is needed on the nature of the folding process to resolve questions that re-

main concerning the details of the folding mechanism. One aspect concerns the events that occur on a submillisecond time scale. By the time that earliest measurements are made (due to limitations of the stop-flow apparatus used in most cases), a compact globule with at least partially formed secondary structure appears to exist. Thus, the question of how secondary structure and collapse are linked remains unclear. It is unlikely that an unstructured molten globule can be formed first because it would be difficult to satisfy hydrogen bonding donors and acceptors. However, the converse, corresponding to a collapse process of the diffusion-collision type, is possible. Of course, in the in vivo formation of proteins, where they are synthesized under folding conditions, it is unlikely that a fully unfolded random coil state ever exists. Thus, some of the questions being examined are limited to what happens in in vitro experiments. There it is necessary to find a trigger (an example might be a pH jump via a compound that has a different pK_a in a laser excitable state) that works faster than stop-flow mixing experiments. In addition, the individual steps in folding are not well characterized because intermediates on the folding pathways of globular proteins are difficult to study experimentally. The intermediates are only slightly populated (if at all) in equilibrium folding experiments and are transiently populated in kinetic folding experiments. One approach that is being used is to study fragments as model systems (Dyson et al., 1988a, 1988b; Oas & Kim, 1988; Scholtz & Baldwin, 1992). From such experiments, information that can be used to determine parameters in the diffusion-collision model has been obtained. A stable peptide containing two 17-amino acid α-helices connected by a 4-residue turn has been designed and synthesized (Fezoui et al., 1994). It contains an elementary unit of the diffusion-collision model (i.e., an α-helical hairpin), and there is evidence that it has the desired structure. Both structural and kinetic studies on this system are in progress. It should be possible to determine both the individual helix and the microdomain pair stabilities from experimentation. The other geometrical parameters in Equation 1 are calculable for this system. In parallel with the experimental studies to determine the kinetics of association of the helices, simulations of the folding and unfolding of the helical hairpin will be made. These go beyond the work with an α-helical hairpin composed of rigid helices of polyvaline (Lee et al., 1987). The refined model simulates helix–coil transitions, in addition to the interhelical collisional events, and uses the specific amino acid sequence of Fezoui et al. (1994). Two types of representations of the residues are being used in the model simulations. One model uses a simplified sphere representation with parameters chosen to correspond to the designed helical hairpin. The other model uses an all-atom representation for the residues. The former can be followed long enough to obtain very good statistics and information about larger scale correlations, and the latter provides complementary detailed information about atom–atom interactions, both intramolecularly and between peptide and solvent. For the simplified Langevin kinetics model, the rates of coil-to-secondary-structure transitions from our previous results suggest transitions on a 10–30-ps time scale. This is in accord with all-atom simulation results (Daggett et al., 1991; Daggett & Levitt, 1991; Brooks, 1993). Collisions between helices connected by a coil segment have been shown to occur on a 1–10-ns time scale. Thus, simulations on a time scale of 1 μs will be required to obtain details of the kinetics. There will be many interactions of helices undergoing helix–coil transitions while diffusing and sometimes colliding, and both folding and unfolding processes will be investigated. The full-atom representation studies concentrate on processes originating from a folded structure and thus explore unfolding kinetics. The combined experimental and simulation studies of this peptide can be used to investigate the interplay of the forces involved in the formation of secondary and tertiary structures and the contributions of the hydrophobic interface, the turn, and the individual helices to the stability of the peptide.

Aside from the study of the 2-helix hairpin described above, it will be important to investigate experimental systems with several microdomains to determine the kinetic pathways and their dependence on the properties of the microdomains of the system. For example, if individual microdomains (identified as secondary structural elements) can be modified so that their intrinsic stability is changed (either by introducing mutants or by naturally occurring variants), the folding rates and even the dominant pathways can be changed for proteins that reach the same final fold. This is a direct consequence of the dependence of the kinetics on the properties of the individual microdomains (e.g., their stability) and their interactions. By such changes, an important pathway in 1 protein will not necessarily be as important in a related protein. NMR experiments on various lysozymes are in progress to examine the dependence of the folding pathways in the helical domain on the amino acid sequence (C. Dobson, pers. comm.). The results of such experiments should provide the information necessary for quantitative applications of the diffusion-collision model.

The packing together of relatively large substructures (helices, sheets) can occur only in a limited number of ways if the instantaneous structure in the subunits is important to the stability of kinetic intermediates, as in the diffusion-collision model. This is in contrast to collapse-type folding mechanisms, because collapsing in itself will not lead to a significant reduction in structural possibilities. Thus, a diffusion-collision folding mechanism is consistent with the observation (see, for example, Chothia, 1992) that the number of protein structural motifs is limited. It provides a mechanistic explanation that supplements the evolutionary argument for the observation.

In addition to its direct relation to the kinetics of protein folding, the diffusion-collision model has applications for protein prediction algorithms. In particular, it suggests a 3-stage approach to prediction. The first stage is the reduction of the combinatorics of conformational possibilities to the microdomains of the system rather than the individual amino acids. Obviously, this requires good secondary-structure prediction methods. The second stage is the diffusion-collision coalescence of the first stage microdomains to the loosely associated state at the end of the kinetic folding pathway, which could be a molten globule state. Various search algorithms in addition to dynamical methods could be used here. The third stage is molecular dynamics and energy minimization on an all-atom representation of the molecule to attain the interdigitation required for close association of the hydrophobic core. It is not clear that the presently available methodologies are capable of accomplishing this step. With increased computer power and new ideas, improvements in the near future are likely.

Acknowledgments

We thank the National Institute of General Medical Sciences (D.L.W.) and the National Science Foundation (M.K.) for support, and our col-

leagues Donald Bashford, Youcef Fezoui, Suhail Islam, Sang Youb Lee, Tom Ngo, John Osterhout, Eugene Shakhnovich, and Kanthi Yapa for many useful discussions. We thank Emily Neel for assistance with the references.

References

Anfinsen CB. 1973. Principles that govern the folding of protein chains. *Science 181*:223-230.

Baldwin RL. 1989. How does protein folding get started? *Trends Biochem Sci 14*:291-294.

Baldwin RL. 1990. Pieces of the folding puzzle. *Nature 346*:409-410.

Bashford D. 1986. Fluctuation diffusion-influenced monomolecular reactions. *J Chem Phys 85*:6999-7010.

Bashford D, Cohen FE, Karplus M, Kuntz ID, Weaver DL. 1988. Diffusion-collision model for the folding kinetics of myoglobin. *Proteins Struct Funct Genet 4*:211-227.

Bashford D, Karplus M, Weaver DL. 1990. The diffusion-collision model of protein folding. In: Gierasch LM, King J, eds. *Protein folding. Deciphering the second half of the genetic code*. Washington, DC: American Association for the Advancement of Science. pp 283-290.

Bashford D, Weaver DL. 1986. Unmolecular diffusion-mediated reactions with a nonrandom time-modulated absorbing barrier. *J Chem Phys 84*:2587-2592.

Bashford D, Weaver DL, Karplus M. 1984. Diffusion-collision model for the folding kinetics of the λ-repressor operator-binding domain. *J Biomol Struct Dynam 1*:1243-1255.

Baum J, Dobson CM, Evans PA. 1989. Characterization of a partly folded protein by NMR methods: Studies on the molten globule state of guinea pig alpha-lactalbumin. *Biochemistry 28*:7-13.

Bernstein FC, Koetzle TF, Williams GJB, Meyer EF, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. 1977. The protein data bank: A computer based archival file for macromolecular structure. *J Mol Biol 112*:535-542.

Blond S, Goldberg M. 1987. Partly native epitopes are already present on early intermediates in the folding of tryptophan synthase. *Proc Natl Acad Sci USA 84*:1147-1151.

Brandts JF, Halvorson HR, Brennan M. 1975. Consideration of the possibility that the slow step in protein denaturation reactions is due to cis-trans isomerism of proline residues. *Biochemistry 14*:4953-4963.

Briggs MS, Roder H. 1992. Early hydrogen-bonding events in the folding reaction of ubiquitin. *Proc Natl Acad Sci USA 89*:2017-2021.

Brooks CL. 1993. Molecular simulations of protein and protein unfolding in quest of a molten globule. *Curr Opin Struct Biol 3*:92-98.

Brooks CL, Karplus M, Pettitt BM. 1988. *Proteins. A theoretical perspective of dynamics, structure, and thermodynamics*. New York: Wiley.

Brown JE, Klee WA. 1971. Helix-coil transition of the isolated amino terminus of ribonuclease. *Biochemistry 10*:470-476.

Brunger AT, Clore GM, Groneborn AM, Karplus M. 1986. Three-dimensional structure of proteins determined by molecular dynamics with interproton distance restraints: Application to crambin. *Proc Natl Acad Sci USA 83*:3801-3805.

Bryugelson JD, Wolynes PG. 1987. Spin glasses and the statistical mechanics of protein folding. *Proc Natl Acad Sci USA 84*:7524-7528.

Bycroft M, Matouschek A, Kellis JT Jr, Serrano L, Fersht AR. 1990. Detection and characterization of a folding intermediate in barnase by NMR. *Nature 346*:488-490.

Chakrabartty A, Schellman JA, Baldwin RL. 1991. Large differences in the helix propensities of alanine and glycine. *Nature 351*:586-588.

Chan HS, Dill KA. 1990. Origins of structure in globular proteins. *Proc Natl Acad Sci USA 87*:6388-6392.

Chothia C. 1992. One thousand families for the molecular biologist. *Nature 357*:543-544.

Chrunyk BA, Matthews CR. 1990. Role of diffusion in the folding of the alpha subunit of tryptophan synthase from *Escherichia coli*. *Biochemistry 29*:2149-2154.

Creighton TE. 1978. Experimental studies of protein folding and unfolding. *Prog Biophys Mol Biol 33*:231-297.

Creighton TE. 1988. Towards a better understanding of protein folding pathways. *Proc Natl Acad Sci USA 85*:5082-5086.

Creighton TE. 1992. The disulfide folding pathway of BPTI. *Science 256*:111-114.

Daggett V, Kollman PA, Kuntz ID. 1991. A molecular dynamics simulation of polyalanine: An analysis of equilibrium motions and helix-coil transitions. *Biopolymers 31*:1115-1134.

Daggett V, Levitt M. 1991. Molecular dynamics simulations of helix denaturation. *J Mol Biol 223*:1121-1138.

Dill KA. 1985. Theory for the folding and stability of globular proteins. *Biochemistry 24*:1501-1509.

Dolgikh DA, Kolomiets AP, Bolotina IA, Ptitsyn OB. 1984. "Molten-globule" state accumulates in carbonic anhydrase folding. *FEBS Lett 165*:88-92.

Dorit RL, Schoenbach L, Gilbert W. 1990. How big is the universe of exons? *Science 250*:1377-1382.

Dyson HJ, Merutka G, Waltho JP, Lerner RA, Wright PE. 1992a. Folding of peptide fragments comprising the complete sequence of proteins. Models for initiation of protein folding I. Myohemerythrin. *J Mol Biol 226*:795-817.

Dyson HJ, Rance M, Houghten RA, Lerner RA, Wright PE. 1988a. Folding of immunogenic peptide fragments of proteins in water solution. I. Sequence requirements for the formation of reverse turns. *J Mol Biol 201*:161-200.

Dyson HJ, Rance M, Houghten RA, Wright PE, Lerner RA. 1988b. Folding of immunogenic peptide fragments of proteins in water solution. II. The nascent helix. *J Mol Biol 201*:201-217.

Dyson HJ, Sayre JR, Merutka G, Shin H, Lerner RA, Wright PE. 1992b. Folding of peptide fragments comprising the complete sequence of proteins. Models for initiation of protein folding II. Plastocyanin. *J Mol Biol 226*:819-835.

Epand RM. 1972. Conformational properties of cyanogen bromide-cleaved glucagon. *J Biol Chem 247*:2132-2138.

Ermak DL, McCammon JA. 1978. Brownian dynamics with hydrodynamic interactions. *J Chem Phys 69*:1352-1361.

Faber HR, Matthews BW. 1990. A mutant T4 lysozyme displays five different crystal formations. *Nature 348*:263-266.

Fezoui Y, Weaver DL, Osterhout JJ. 1994. De novo design and structural characterization of an α-helical hairpin peptide ($\alpha t \alpha$): A novel model system for the study of protein folding intermediates. *Proc Natl Acad Sci USA*. Forthcoming.

Flory PJ. 1969. *Statistical mechanics of chain molecules*. New York: Wiley.

Gilmanshin RI, Ptitsyn OB. 1987. An early intermediate of refolding alpha-lactalbumin forms within 20 ms. *FEBS Lett 223*:327-329.

Go N. 1983. Theoretical studies of protein folding. *Annu Rev Biophys Bioeng 12*:183-210

Go N, Abe H. 1981. Noninteracting local-structure model of folding and unfolding transition in globular proteins. I. Formulation. *Biopolymers 20*:991-1011.

Go N, Abe H, Mizuno H, Taketomi H. 1980. Local structures in the process of protein folding. In: Jaenicke R, ed. *Protein folding*. Amsterdam/New York: Elsevier/North-Holland Biomedical Press. pp 167-181.

Godzik A, Skolnick J, Kolinski A. 1992. Simulations of the folding pathway of triose phosphate isomerase-type alpha/beta barrel proteins. *Proc Natl Acad Sci USA 89*:2629-2633.

Goldenberg DP, Creighton TE. 1985. Energetics of protein structure and folding. *Biopolymers 24*:167-182.

Gratzer WB, Beaven GH. 1969. Relation between conformation and association state. *J Biol Chem 244*:6675-6679.

Haas E, Katchalski-Katzir E, Sternberg IZ. 1978. Brownian motion of the ends of oligopeptide chains in solution as estimated by energy transfer between the chain ends. *Biopolymers 17*:11-31.

Haas E, Wilchek M, Katchalski-Katzir E, Steinberg IZ. 1975. Distribution of end-to-end distances of oligopeptides in solution as estimated by energy transfer. *Proc Natl Acad Sci USA 72*:1807-1811.

Hammes GG, Schullery SE. 1968. Structure of macromolecular aggregates. I. Aggregation-induced conformational changes in polypeptides. *Biochemistry 7*:3882-3887.

Harding MM, Williams DN, Woolfson DN. 1991. Characterization of a partially denatured state of a protein by two-dimensional NMR: Reduction of the hydrophobic interaction of ubiquitin. *Biochemistry 30*:3120-3128.

Harrison SC, Durbin R. 1985. Is there a single pathway for the folding of a polypeptide chain? *Proc Natl Acad Sci USA 82*:4028-4030.

Hollecker M, Larcher D. 1989. Conformational forces affecting the folding pathways of dendrotoxins I and K from black mamba venom. *Eur J Biochem 179*:87-94.

Holley LH, Karplus M. 1989. Protein secondary structure prediction with a neural network. *Proc Natl Acad Sci USA 86*:152-156.

Hughson FM, Barrick D, Baldwin RL. 1991. Probing the stability of a partly folded apomyoglobin intermediate by site-directed mutagenesis. *Biochemistry 30*:4113-4118.

Hughson FM, Wright PE, Baldwin RL. 1990. Structural characterization of a partly folded apomyoglobin intermediate. *Science 249*:1544-1548.

Hurle MH, Michelotti GA, Crisanti MM, Matthews CR. 1987. Characterization of a slow folding reaction for the alpha subunit of tryptophan synthase. *Proteins Struct Funct Genet 2*:54-63.

Jeng M, Englander SW. 1991. Stable submolecular folding units in a non-compact form of cytochrome *c*. *J Mol Biol 221*:1045-1061.

Jennings PA, Wright PE. 1993. Formation of a molten globule intermediate early in the kinetic folding pathway of apomyoglobin. *Science 262*:892–896.

Kabsch W, Sander C. 1983. How good are predictions of protein secondary structure? *FEBS Lett 155*:179–182.

Kanehisa MI, Tsong TY. 1978. Mechanisms of the multiphasic kinetics in the folding and unfolding of globular proteins. *J Mol Biol 124*:177–194.

Kanehisa MI, Tsong TY. 1979a. Dynamics of the cluster model of protein folding. *Biopolymers 18*:1375–1388.

Kanehisa MI, Tsong TY. 1979b. Kinetic analysis of local structure formations in protein folding. *Biopolymers 18*:2913–2928.

Kanehisa MI, Tsong TY. 1979c. Slow equilibration of a denatured protein: Comparison of the cluster model with the proline isomerization model. *J Mol Biol 133*:279–283.

Karplus M, Weaver DL. 1976. Protein-folding dynamics. *Nature 260*:404–406.

Karplus M, Weaver DL. 1979. Diffusion-collision model for protein folding. *Biopolymers 18*:1421–1437.

Kim PS, Baldwin RL. 1982. Specific intermediates in the folding reactions of small proteins and the mechanism of folding. *Annu Rev Biochem 51*:459–489.

Kuroda Y. 1993. Residual helical structure in the C-terminal fragment of cytochrome *c*. *Biochemistry 32*:1219–1224.

Kuroda Y, Kidokoro S, Wada A. 1992. Thermodynamic characterization of cytochrome *c* at low pH: Observation of the molten globule state and of the cold denaturation process. *J Mol Biol 223*:1139–1153.

Kuwajima K. 1989. The molten globule state as a clue for understanding the folding and cooperativity of globular-protein structure. *Proteins Struct Funct Genet 6*:87–103.

Kuwajima K, Hiraoka Y, Ikeguchi M, Sugai S. 1985. Comparison of the transient folding intermediates of lysozyme and alpha-lactalbumin. *Biochemistry 24*:874–881.

Kuwajima K, Yamaya H, Miwa S, Sugai S, Nagamura T. 1987. Rapid formation of secondary structure framework in protein folding studied by stopped-flow circular dichroism. *FEBS Lett 221*:115–118.

Labhardt AM. 1984. Kinetic circular dichroism shows that the S-peptide alpha-helix of ribonuclease S unfolds fast and refolds slowly. *Proc Natl Acad Sci USA 81*:7674–7678.

Lee B, Richards FM. 1971. The interpretation of protein structures: Estimation of static accessibility. *J Mol Biol 55*:379–400.

Lee S, Karplus M, Bashford D, Weaver DL. 1987. Brownian dynamics simulation of protein folding: A study of the diffusion-collision model. *Biopolymers 26*:481–506.

Lesk AM, Rose GD. 1981. Folding units in globular proteins. *Proc Natl Acad Sci USA 78*:4304–4308.

Levinthal C. 1966. Molecular model building by computer. *Sci Am 214*:42–52.

Levinthal C. 1968. Are there pathways for protein folding? *J Chem Phys 65*:44–45.

Levitt M. 1976. A simplified representation of protein conformations for rapid simulation of protein folding. *J Mol Biol 104*:59–107.

Levitt M. 1978. *Models of protein dynamics*. Universite de Paris, Orsay: CECAM Workshop Report.

Levitt M, Chothia C. 1976. Structural patterns in globular proteins. *Nature 261*:552–558.

Levitt M, Warshal A. 1975. Computer simulation of protein folding. *Nature 253*:694–698.

Mann CJ, Matthews CR. 1993. Structure and stability of an early folding intermediate of *Escherichia coli* trp aporepressor measured by far-UV stopped-flow circular dichroism and 8-anilino-1-naphthalene sulfonate binding. *Biochemistry 32*:5282–5290.

Marks CB, Naderi H, Kosen PA, Kuntz ID, Anderson S. 1987. Refolding of bovine pancreatic trypsin inhibitor mutants lacking cysteines 30 and 51. *Protein Struct Folding Design 2*:335–340.

Marqusee S, Robbins VH, Baldwin RL. 1989. Unusually stable helix formation in short alanine-based peptides. *Proc Natl Acad Sci USA 86*:5286–5290.

Matouschek A, Kellis JT Jr, Serrano L, Bycroft M, Fersht AR. 1990. Transient folding intermediates characterized by protein engineering. *Nature 346*:440–445.

McCammon JA, Northrup SH, Karplus M, Levy RM. 1980. Helix–coil transitions in a simple polypeptide model. *Biopolymers 19*:2033–2045.

McCoy LF, Rowe ES, Wong KP. 1980. Multiparameter kinetic study on the unfolding and refolding of bovine carbonic anhydrase B. *Biochemistry 19*:4738–4743.

Miranker A, Radford SE, Karplus M, Dobson CM. 1991. Demonstration by NMR of folding domains in lysozyme. *Nature 349*:633–636.

Mitchinson C, Baldwin RL. 1986. The design and production of semisynthetic ribonucleases with increased thermostability by incorporation of S-peptide analogues with enhanced helical stability. *Proteins Struct Funct Genet 1*:23–33.

Moult J, Unger R. 1991. An analysis of protein folding pathways. *Biochemistry 30*:3816–3824.

Nilsson B, Kuntz ID, Anderson S. 1990. Expression and stabilization: Bovine pancreatic trypsin inhibitor folding mutants in *Escherichia coli*. In: Gierasch LM, King J, eds. *Protein folding: Deciphering the second half of the genetic code*. Washington, DC: American Association for the Advancement of Science. pp 117–122.

Oas TG, Kim PS. 1988. A peptide model of a protein folding intermediate. *Nature 336*:42–48.

Ohgushi M, Wada A. 1983. "Molten-globule state": A compact form of proteins with mobile sidechains. *FEBS Lett 164*:21–24.

Osterhout JJ, Baldwin RL, York EJ, Stewart JM, Dyson HJ, Wright PE. 1989. H NMR studies of the solution conformations of an analogue of the C-peptide of ribonuclease A. *Biochemistry 28*:7059–7064.

Panijpan B, Gratzer WB. 1974. Conformational nature of monomeric glucagon. *Eur J Biochem 45*:547–553.

Privalov PL. 1989. Thermodynamic problems of protein structure. *Annu Rev Biophys Biophys Chem 18*:47–69.

Ptitsyn OB. 1987. Protein folding: Facts and models. *J Protein Chem 6*:272–293.

Ptitsyn OB, Pain RH, Semisotnov GV, Zerovnitz E, Razgulyaev OI. 1990. Evidence for a molten globule state as a general intermediate in protein folding. *FEBS Lett 262*:20–24.

Ptitsyn OB, Rashin AA. 1973. Stagewise mechanism of protein folding. *Dokl Akad Nauk SSSR 213*:473–475.

Ptitsyn OB, Rashin AA. 1975. A model of myoglobin self-organization. *Biophys Chem 3*:1–20.

Radford SE, Dobson CM, Evans PA. 1992. The folding of hen lysozyme involves partially structured intermediates and multiple pathways. *Nature 358*:302–307.

Rey A, Skolnick J. 1991. Comparison of lattice Monte Carlo dynamics and Brownian dynamics folding pathways of alpha-helical hairpins. *Chem Phys 158*:199–219.

Roder H, Elove GA, Englander SW. 1988. Structural characterization of folding intermediates in cytochrome *c* by H-exchange labelling and proton NMR. *Nature 335*:700–704.

Roder H, Wuthrich K. 1986. Protein folding kinetics by combined use of rapid mixing techniques and NMR observation of individual amide protons. *Proteins Struct Funct Genet 1*:34–42.

Rooman MJ, Wodak SJ. 1988. Identification of predictive sequence motifs limited by protein structure data base size. *Nature 335*:45–49.

Rost B, Sander C. 1993. Prediction of protein secondary structure at better than 70% accuracy. *J Mol Biol 232*:584–599.

Sachs DH, Schechter AN, Eastlake A, Anfinsen CB. 1972. An immunologic approach to the conformational equilibria of polypeptides. *Proc Natl Acad Sci USA 69*:3790–3794.

Sali A, Shakhnovich E, Karplus M. 1994. Kinetics of protein folding. *J Mol Biol 235*:1614–1636.

Sasaki K, Dockerill S, Adamiak DA, Tickle IJ, Blundell T. 1975. X-ray analysis of glucagon and its relationship to receptor binding. *Nature 257*: 751–757.

Schneller W, Weaver DL. 1993. Simulation of alpha-helix–coil transitions in simplified polyvaline: Equilibrium properties and Brownian dynamics. *Biopolymers 33*:1519–1535.

Scholtz JM, Marqusee S, Baldwin RL, York EJ, Stewart JM, Santoro M, Bolen DW. 1991. Calorimetric determination of the enthalpy change for the α-helix to coil transition of an alanine peptide in water. *Proc Natl Acad Sci USA 88*:2854–2858.

Semisotnov GV, Rochonova NA, Kutyshenko VP, Ebert B, Blanck J, Ptitsyn OB. 1987. Sequential mechanism of refolding of carbonic anhydrase B. *FEBS Lett 224*:9–13.

Serrano L, Matouschek A, Fersht AR. 1992. The folding of an enzyme. VI. The folding pathway of barnase: Comparison with theoretical models. *J Mol Biol 224*:847–859.

Shakhnovich EI, Finkelstein AV. 1989. Theory of cooperative transitions in protein molecules. I. Why denaturation of globular protein is a first-order phase transition. *Biopolymers 28*:1667–1680.

Shakhnovich EI, Gutin AM. 1990. Implications of thermodynamics of protein folding for evolution of primary sequences. *Nature 346*:773–775.

Shoemaker KR, Kim PS, Brems DN, Marqusee S, York EJ, Chaiken IM, Stewart JM, Baldwin RL. 1985. Nature of the charged-group effect on the stability of the C-peptide helix. *Proc Natl Acad Sci USA 82*:2349–2353.

Sikorski A, Skolnick J, 1990. Dynamic Monte Carlo simulations of globu-

lar protein folding/unfolding pathways. II. Alpha-helix motifs. *J Mol Biol 212*:819-836.
Skolnick J, Kolinski A. 1990a. Dynamic Monte Carlo simulations of globular protein folding/unfolding pathways. I. Six-member, Greek key beta-barrel proteins. *J Mol Biol 212*:787-817.
Skolnick J, Kolinski A. 1990b. Simulations of the folding of a globular protein. *Science 250*:1121-1125.
Skolnick J, Kolinski A, Brooks CL III, Godzik A, Rey A. 1993. A method for predicting protein structure from sequence. *Curr Biol 3*:414-423.
Skolnick J, Kolinski A, Sikorski A. 1990. Dynamic Monte Carlo simulations of globular protein folding, structure and dynamics. *Comments Mol Cell Biophys 6*:223-247.
Staley JP, Kim PS. 1992. Complete folding of bovine pancreatic trypsin inhibitor with only a single disulfide bond. *Proc Natl Acad Sci USA 89*: 1519-1523.
States DJ, Dobson CM, Creighton TE, Karplus M. 1984. A new two-disulphide intermediate in the refolding of reduced bovine pancreatic trypsin inhibitor. *J Mol Biol 174*:411-418.
States DJ, Dobson CM, Karplus M, Creighton TE. 1980. A conformational isomer of bovine pancreatic trypsin inhibitor protein produced by refolding. *Nature 286*:630-632.
Tsong TY. 1982. Viscosity-dependent conformational relaxation of ribonuclease A in the thermal unfolding zone. *Biochemistry 21*:1493-1498.
Tsong TY, Baldwin RL. 1978. Effects of solvent viscosity and different guanidine salts on the kinetics of ribonuclease A chain folding. *Biopolymers 17*:1669-1678.
Tsong TY, Baldwin RL, Elson EL. 1972a. Properties of the folding and unfolding reactions of ribonuclease A. *Proc Natl Acad Sci USA 69*:1809-1812.
Tsong TY, Baldwin RL, McPhie P. 1972b. A sequential model of nucleation-dependent protein-folding: Kinetic studies of ribonuclease A. *J Mol Biol 63*:453-475.
Udgaonkar JB, Baldwin RL. 1988. NMR evidence for an early framework intermediate on the folding pathway of ribonuclease A. *Nature 335*: 694-699.
Weaver DL. 1980. Nonequilibrium decay effects in diffusion-controlled processes. *J Chem Phys 72*:3483-3485.
Weaver DL. 1982. Microdomain dynamics in folding proteins. *Biopolymers 21*:1275-1300.
Weaver DL. 1984. Alternative pathways in diffusion-collision controlled protein folding. *Biopolymers 23*:675-694.
Weissman JS, Kim PS. 1992. Reexamination of the folding of BPTI: Predominance of native intermediates. *Science 253*:1386-1393.
Wetlaufer DB. 1973. Nucleation, rapid folding and globular intrachain regions in proteins. *Proc Natl Acad Sci USA 70*:697-701.
Wright PE, Dyson HJ, Lerner RA. 1988. Conformation of peptide fragments of proteins in aqueous solution: Implications for initiation of protein folding. *Biochemistry 27*:7167-7175.
Yapa K, Weaver DL, Karplus M. 1992. Beta-sheet coil transitions in a simple polypeptide model. *Proteins Struct Funct Genet 12*:237-265
Zhang X, Mesirov JP, Waltz DL. 1992. Hybrid system for protein secondary structure prediction. *J Mol Biol 225*:1049-1063.

Appendix A: Elementary diffusion-collision step

A simple analytical model allows one to calculate the folding rate of 2 connected microdomains, the elementary step in the diffusion-collision model.

Consider 2 microdomains labeled A and B. Each microdomain has a variety of conformational possibilities and is short enough so that a random search is possible on a physiological time scale:

$$A + B \leftrightarrow AB. \tag{A-1}$$

With the dynamical behavior of the microdomains modeled by the diffusion equation, we have for the equation of motion of this system the coupled equations (Bashford, 1986; Bashford & Weaver, 1986; Bashford et al., 1988):

$$\frac{\partial}{\partial t}\begin{pmatrix}\rho_1\\ \rho_2\end{pmatrix} = D\nabla^2\begin{pmatrix}\rho_1\\ \rho_2\end{pmatrix} + \begin{pmatrix}-\lambda_1 & \lambda_2\\ \lambda_1 & -\lambda_2\end{pmatrix}\begin{pmatrix}\rho_1\\ \rho_2\end{pmatrix}. \tag{A-2}$$

Here D is the relative diffusion coefficient (see Equation 1). The probability density ρ has 2 elements, ρ_1, which refers to the state in which both microdomains are folded, and ρ_2, which refers to all other possibilities. The rate constants are λ_1 from the both-folded state to all others and λ_2 for the reverse reaction.

The pair of connected microdomains A and B is limited in the diffusion space available for their relative motion. To simplify the subsequent calculation of the folding rate, we idealize the microdomains as spheres and the polypeptide chain connecting them as a perfectly flexible featureless string. Our model calculations and simulations of helices connected by random coil chains suggest that this leads to the correct orders of magnitude. Collision and coalescence of the microdomains are governed by the boundary conditions for Equation A-2. The inner boundary is the spherical shell of closest approach (governed by the van der Waals envelopes of the microdomains). In this approximation it has a radius that is the sum of the radii of the microdomains, denoted by $R_{\min}$. The outer boundary is the maximum radial separation of the microdomains determined by the length of the string (the polypeptide chain between the microdomains), denoted by $R_{\max}$. The boundary conditions on the probability density fluxes are

$$\left.\frac{\partial \rho_{1,2}}{\partial r}\right|_{R_{\max}} = 0, \tag{A-3}$$

that is, the net outward flux of probability density is 0. This corresponds to complete reflection at the outer boundary, which means that the microdomains cannot get further away from one another than $R_{\max}$:

$$\left.\frac{\partial \rho_2}{\partial r}\right|_{R_{\min}} = 0, \tag{A-4}$$

that is, the net flux of the nonfolded states is 0 at the inner boundary; and,

$$\left.\frac{\partial \rho_1}{\partial r}\right|_{R_{\min+}} = 0, \tag{A-5}$$

that is, the outward flux of the both folded state is zero at the inner (collision) boundary, indicating that coalescence takes place at the inner boundary in this state of the system.

Define the unreacted fraction of spherical pairs at an initial position r_0, $N_{1,2}(r_0,t)$:

$$N_{1,2}(r_0,t) = \int dV \rho_{1,2}(r,r_0,t). \tag{A-6}$$

The vector $\begin{pmatrix}N_1\\ N_2\end{pmatrix}$ satisfies the differential equation adjoint to Equation A-1.

We have shown that $N(t)$ is well approximated by

$$N(t) \cong e^{-t/\tau_f} \tag{A-7}$$

with $1/\tau_f$ the folding rate constant in a first-passage time approximation. The folding time τ_f is the weighted average of τ_1 and τ_2:

$$\tau_f = \frac{\lambda_2}{\lambda_1 + \lambda_2}\tau_1 + \frac{\lambda_1}{\lambda_1 + \lambda_2}\tau_2, \tag{A-8}$$

that is, the individual first-passage times weighted by the probability $\beta = \lambda_2/(\lambda_1 + \lambda_2)$ that the 2 microdomains are in the folded, oriented state when they collide so that there is no barrier to coalescence, and $\lambda_1/(\lambda_1 + \lambda_2) = 1 - \beta$ the probability of noncoalescence.

The general form of τ_f is

$$\tau_f = \frac{l^2}{D} + \frac{L\Delta V(1-\beta)}{\beta DA}. \tag{A-9}$$

The volume available for diffusion of each microdomain pair ΔV, their relative target surface area for collisions A, their relative diffusion coefficient D, and their relative geometry parameter l^2 are calculated for diffusion in a spherical space. The parameter L has units of length. Its value is

$$\frac{1}{L} = \frac{1}{R_{\min}} + \alpha\frac{\alpha R_{\max}\tanh[\alpha(R_{\max} - R_{\min})] - 1}{\alpha R_{\max} - \tanh[\alpha(R_{\max} - R_{\min})]} \tag{A-10}$$

where

$$\alpha \equiv \left(\frac{\lambda_1 + \lambda_2}{D}\right)^{1/2}. \tag{A-11}$$

For the backward, unfolding rate $1/\tau_b$, the actual contact surface area A_{AB} between microdomains A and B is used, as well as the free energy change per unit area f appropriate for this surface area measure. The parameter ν determines the degree to which the folding process goes to completion.

$$\tau_b = \nu^{-1}e^{f\frac{A_{AB}}{k_BT}}. \tag{A-12}$$

Appendix B

It has been shown (Weaver, 1984) that the diffusion-collision dynamics (Karplus & Weaver, 1976, 1979) of a multimicrodomain protein can be treated as a set of 2-microdomain processes. This leads to coupled first-order rate equations for the probabilities (corresponding to concentrations in unimolecular processes) $p_i(t)$ of the possible intermediate states:

$$\frac{dp_i}{dt} = \sum_{j=1}^{m} R_{ij}p_j. \tag{B-1}$$

In Equation B-1, R_{ij} are the elements of the rate matrix to be determined from the model (Equations A-9 and A-10), and m is the number of independent states of the folding protein. If there are n pairwise interactions between the microdomains, then there are 2^n states (including the initial unfolded state and the final folded state) and $n!$ pathways to the final state. For example, consider a protein (or a portion of a larger protein) with 3 elementary microdomains, each of which interacts in a pairwise manner with the other 2 microdomains. With 3 pairwise interactions, m is 8. In this case, if forward rates only (Equation A-9) are considered, the system of rate equations can be solved analytically. Calling the microdomains A, B, and C (for example, they could be 3 α-helices, 3 β-strands, or a combination of β-strands and α-helices), state 1 has each of A, B, C unpaired, state 2 has only the AB pair, state 3 has only the AC pair, state 5 has only the BC pair, state 4 has the AB and AC pairs, state 6 has the AB and BC pairs, state 7 has the AC and BC pairs, and state 8 has all 3 pairwise interactions (see Table 2 and Fig. 2). In terms of the elementary forward rate constants k_{ij} linking the states (Table 3), the set of coupled, first-order differential equations satisfied by the states is

$$\frac{dp_1}{dt} = -(k_{12} + k_{13} + k_{15})p_1 \tag{B-2}$$

where $R_{11} = -(k_{12} + k_{13} + k_{15})$,

$$\frac{dp_2}{dt} = k_{12}p_1 - (k_{24} + k_{26})p_2 \tag{B-3}$$

where $R_{21} = k_{12}$, $R_{22} = -(k_{24} + k_{26})$,

$$\frac{dp_3}{dt} = k_{13}p_1 - (k_{34} + k_{37})p_3 \tag{B-4}$$

where $R_{31} = k_{31}$, $R_{33} = -(k_{34} + k_{37})$,

$$\frac{dp_4}{dt} = k_{24}p_2 + k_{34}p_3 - k_{48}p_4 \tag{B-5}$$

where $R_{42} = k_{24}$, $R_{43} = k_{34}$, $R_{44} = -k_{48}$,

$$\frac{dp_5}{dt} = k_{15}p_1 - (k_{56} + k_{57})p_5 \tag{B-6}$$

where $R_{51} = k_{15}$, $R_{55} = -(k_{56} + k_{57})$,

$$\frac{dp_6}{dt} = k_{26}p_2 + k_{56}p_5 - k_{68}p_6 \tag{B-7}$$

where $R_{62} = k_{26}$, $R_{65} = k_{56}$, $R_{66} = -k_{68}$,

$$\frac{dp_7}{dt} = k_{37}p_3 + k_{57}p_5 - k_{78}p_7 \tag{B-8}$$

where $R_{73} = k_{37}$, $R_{75} = k_{57}$, $R_{77} = -k_{78}$, and the conservation of probability equation for the unimolecular process is

$$p_1 + p_2 + p_3 + p_4 + p_5 + p_6 + p_7 + p_8 = 1. \tag{B-9}$$

Note that only the rate constants, k_{ij}, appearing in Equations B-2–B-8 contribute to the sum in Equation B-1. Thus, the rate constants, R_{ij}, in Equation B-1 are all sums of rates that are the inverse of Equation 1 (or Equation A-9) for the particular reaction (see Table 3). The example applies to conditions that strongly favor folding, in which the native structure is stable and unfolding reactions are unimportant. Equations B-2–B-8 have been solved using the Laplace transform method (details to be published elsewhere). The initial conditions used were $p_1(0) = 1$ and $p_2(0) \ldots p_8(0) = 0$. This corresponds to the entire protein chain being synthesized before folding commences. Other choices would lead to somewhat different probabilities, e.g., folding beginning at the N-terminus before all the subunits are synthesized. The solutions for $p_1 \ldots p_8$ are

$$p_1 = e^{-at} \tag{B-10}$$

$$p_2 = b\,\frac{(e^{-ct} - e^{-at})}{a - c} \tag{B-11}$$

$$p_3 = d\,\frac{(e^{-et} - e^{-at})}{a - e} \tag{B-12}$$

$$p_4 = \frac{fbe^{-ct}}{ha - ac - ch + c^2} + e^{-ht}\,\frac{(hfb + hdg - dgc - bef)}{h_p} + \frac{-dge^{-et}}{he - ha + ea - e^2} + \frac{e^{-at}(dga + fba - bef - dgc)}{a_p} \tag{B-13}$$

where $a_p = a^3 - a^2c - a^2h + ach + hea - a^2e + cea - ceh$; and $h_p = h^3 - h^2c - h^2a + hca + aeh - h^2e + che - cea$;

$$p_5 = i\,\frac{(e^{-jt} - e^{-at})}{a - j} \tag{B-14}$$

$$p_6 = \frac{bke^{-ct}}{ma - ac - cm + c^2} + e^{-mt}\,\frac{(mil + mbk - ilc - bkj)}{m_p} + \frac{-ile^{-jt}}{ja - ma - j^2 + jm} + \frac{e^{-at}(bka + ail - bkj - ilc)}{j_p} \tag{B-15}$$

where $m_p = jma - m^2a + m^3 - m^2j - cja + cma - cm^2 + jcm$; and $j_p = a^3 - ja^2 - ma^2 + jma - a^2c + cja + cma - jcm$;

$$p_7 = \frac{e^{-qt}(eip + djn - ipq - dnq)}{q_p} - \frac{ipe^{-jt}}{ja - qa - j^2 + qj} + \frac{-dne^{-et}}{ea - e^2 - qa + qe} + \frac{e^{-at}(aip + dan - djn - eip)}{e_p} \tag{B-16}$$

where $q_p = qea + jea + q^2e - qje + q^2a - q^3 - qja + q^2j$; and $e_p = a^3 - ja^2 - qa^2 + qja - ea^2 + jea + qea - qje$;

$$p_8 = 1 - p_1 - p_2 - p_3 - p_4 - p_5 - p_6 - p_7. \tag{B-17}$$

The coefficients $a \ldots q$ are given in terms of the elementary rate constants by the following relations: $a = k_{12} + k_{13} + k_{15}$; $b = k_{12}$; $c = k_{24} + k_{26}$; $d = k_{13}$; $e = k_{34} + k_{37}$; $f = k_{24}$; $g = k_{34}$; $h = k_{48}$; $i = k_{15}$; $j = k_{56} + k_{57}$; $k = k_{26}$; $l = k_{56}$; $m = k_{68}$; $n = k_{37}$; $p = k_{57}$; $q = k_{78}$.

It should be noted that the solutions for $p_1 \ldots p_8$ are dimensionless and that the time units cancel in the exponential terms and in the amplitudes. For example, if the rates, k_{ij}, are on the order of 10^9/s, then the relevant time scale for $p_1 \ldots p_8$ is nanoseconds.