NASA CONTRACTOR REPORT 166318

Protein Folding, Protein Structure and the Origin of Life: Theoretical Methods and Solutions of Dynamical Problems

D. L. Weaver

NASA Purchase Order A 86784B (VAB)
March 1982

NASA

NASA CONTRACTOR REPORT 166318

Protein Folding, Protein Structure and the Origin of Life: Theoretical Methods and Solutions of Dynamical Problems

D. L. Weaver
Department of Physics
Tufts University
Medford, MA 02155

Prepared for
Ames Research Center
under NASA Purchase Order A 86784B (VAB)

National Aeronautics and
Space Administration

Ames Research Center
Moffett Field, California 94035

INTRODUCTION

Proteins are produced in living systems by adding amino acids, one-by-one starting from the NH_2 terminus of the polypeptide chain. It has been established for some time[1] that such chains can fold spontaneously into the native structure (three-dimensional, compact folded structure in which the protein may carry on its structural, regulatory or catalytic function) without any further information than is contained in the linear sequences of amino acid residues and their interaction with the surroundings (water, salts, pH, temperature).

The spontaneity of the folding process may be considered to be a crucial element of Origin of Life studies concerned with the production of functioning biomacromolecules by non-biological methods, in the pre-biological environment that may have led to the origin of self-replicating structures. Thus, the major question to be answered for amino acid polymers is what three-dimensional structure or structures are favored in a particular environment, and what physical mechanisms lead these biopolymers from the set of unfolded conformations to the set of folded conformations.

These questions prompt one to divide the problem of the folding of a protein to its native structure into two parts. The first is the static aspect concerned with the elements in the amino acid sequence that provide the information; the second deals with the dynamics of the folding process itself. Clearly, the answer to the first part of the problem is involved in the second (in particular, the stabilities of intermediate structures may be important in selecting the folding paths) and conversely, it is possible, though less likely, that the second affects the first (that is, that the nature of the folding process results in a non-equilibrium structure).

Available data indicate that proteins "in vitro" can fold into their native structure in times from tenths of seconds to minutes in the absence of S-S bridges[2]; formation of S-S bridges coupled to refolding tends to take longer. Producing the complete, folded protein "in vivo" is also a seconds to minutes process. To appreciate the problems involved in understanding the folding dynamics, these times must be contrasted with the long time required to find the native structure by a random search through all possible conformations. For example, for a protein consisting of 100 amino acids with three independent configurations for each one, there are 3^{100} possible conformations. If each one can be searched in 10^{13} seconds, the total time to search each structure once is $3^{100} \times 10^{13}$ seconds $\sim 10^{37}$ seconds, compared to 10^{17} seconds, the age of the universe. Of course, this estimate neglects excluded volume effects, correlated motions, etc. but underestimates the possible configurations per amino acid and so gives the correct impression, that protein folding must make use of more sophisticated search procedures.

The above comparison of experimental folding times with the simplest possible model of independent random searches by each amino acid indicate that in the folding processes fluctuations and correlated motions among the amino acid residues must play an essential role in searching out the native structure. Unfortunately, the vast range of configurational space that has to be examined, the many potential barriers that are likely to be present, and the long time scale of the overall process (tenths of seconds to minutes) make it very difficult to study the detailed motions of the atoms involved in the folding process. It is necessary, therefore, at present, to introduce models for the dynamical aspects of the folding process.

With the help of the models, it may be possible to theoretically "fold" the protein using computer simulations to the point where energy minimization techniques may be applied to a realistic representation of the amino acid chain, including environmental effects whether of solution or surfaces. The alternative is to simplify the description of the amino acids in order to make a computer simulation from an unfolded state be technically, temporally and economically feasible. This approach has been attempted[3] but without great success[4] since the choice of simplified representation appears to be arbitrary, at present.

A possibly viable alternative to the computer simulation of folding, starting from a realistic representation of the entire polypeptide chain, is to deal with only small fragments of the chain. Then, the forces can be made more realistic and perhaps the computer time element not so overpowering. This will be discussed further, below.

It will be necessary to consider in some detail the elements of a dynamical model to represent the initial stages of folding and to discuss the calculation and/or experimental determination of the parameters of the model. This is done in the next section.

DYNAMICAL FOLDING MODELS

Karplus and Weaver[5] and Baldwin[6] have considered the experimental and theoretical evidence as to the basic folding mechanism and have concluded that a slow random search nucleation followed by rapid folding about the nucleus is unlikely to be the main folding mechanism. Instead, they have concluded that locally ordered intermediates, called microdomains (hereafter denoted as MD), form in several parts of the polypeptide chain then collide and coalesce with the rates of successive steps on the folding pathway generally depending on the stabilities of preceding intermediates.

To describe the dynamical aspects of MD behavior, Karplus and Weaver[7] have introduced the diffusion-collision (DC) model of protein folding. In the DC model, the protein molecule is thought of as divided into parts (the MDs). Folding pathways are then studied by following the diffusive motion of the centers of mass of the MDs (in first approximation the detailed structure of the MDs is not included in the model) as they are subjected to random, external dissipative forces caused by very frequent collisions with solvent molecules. Collisions between MDs sometimes lead to their coalescence into MD pairs and so on into larger MD aggregates and eventually to the native structure of the protein. In the regime of high solvent friction that applies to the motion of protein parts in aqueous solution, the dynamics of the MDs is governed by the diffusion equation[8] describing the spatial and temporal behavior of the probability density of a microdomain.

In the DC model, the MDs themselves are local structures of limited stability (generally thought to be the α-helical and β-strand segments

which have been observed in crystallographic studies of folded proteins[9], although this is not crucial to the model). In a real protein, the entire folding pathway would consist of a sequence of steps of diffusion, collision and eventual coalescence of MDs with the slowest, rate-determining step occurring at the end of the pathway. Particularly in the early stages of folding there may be a number of different sequences of folding steps (different pathways) which converge toward the rate determining slow step leading to the native structure.

Thus the picture of a folding pathway given by a multistep DC mechanism is the following: first, two MDs both of which are in fast equilibrium between some folded conformations (perhaps α-helical or β-strand) and an unfolded set of "random coil" conformations (with an equilibrium constant greatly favoring the coil state) combine by the DC mechanism. The state so formed would probably be somewhat more stable than the free but folded MDs if the state were on the folding pathway rather than being an incorrectly folded intermediate. This MD-MD state can in turn dissociate into the two separate MDs, or it can interact with a third MD to form a still more stable entity with a larger equilibrium constant. Near the end of this pathway, the coalesced state encounters the attractive interactions which produce a very stable state to be identified with the native or near native set of conformations.

It is clear that in this model of the folding process, the computation aspects of the initial stages on a folding pathway may be reduced considerably from consideration of the entire polypeptide chain to consideration of the possible conformations of individual MDs and pairwise interactions among MDs.

In order to make a quantitative study of the above folding picture, one needs to calculate, in the context of the DC model, how the reactions for the elementary folding step

$$(MD)_1 + (MD)_2 \rightleftharpoons \begin{matrix}(MD)_1 \\ | \\ (MD)_2\end{matrix} \qquad (1)$$

affect the (time-dependent) probability of the existence of a MD-MD pair state. A basic assumption of the DC model is that the relative motion of the centers of mass of the MDs is described by the Smoluchowski equation[10] (the diffusion equation with potential energy of interaction) which has the form

$$\frac{\partial \rho}{\partial \rho} = \vec{\nabla} \cdot \{D\,(\vec{\nabla}\rho + \rho\vec{\nabla}V)\} \qquad (2)$$

where $\rho(\vec{x}, \vec{x}_o, t)$ is the MD relative position probability density at position $\vec{x}$ and time t starting from position $\vec{x}_o$ at time zero. D is the (possibly position dependent) diffusion coefficient for the relative motion and $V(\vec{x})$ is the potential energy of interaction between the MDs (in units of K_BT where K_B is Boltzmann's constant and T the absolute temperature). The diffusion is limited in spatial extent by how far away from one another two MDs may get without breaking the polypeptide chain on the one hand, and by how close together two MDs can get before they collide.

To be specific, some of the details of the solution of Eq.2 will be given for a one-dimensional diffusion space with V(x) = 0 and constant D. With these parameter choices, Eq. 2 becomes

$$\frac{\partial \rho}{\partial t} = D \frac{\partial^2 \rho}{\partial r^2} \tag{3}$$

r being the distance between MD centers. The minimum value of r is the sum of the radii of the two MDs (assumed to be spherically symmetrical in this example) to be denoted by r = a. The maximum value of r, denoted by r = b, is equal to a plus whatever length of polypeptide chain occurs between the MDs. In order to solve Eq. (3), the behavior of ρ at the boundaries of the diffusion space r = a and r = b must be specified. As discussed in Ref. 7, the MDs are unable to attain a value of r greater than b (the MDs must always be connected) so the solution of Eq. (3) must there satisfy a completely reflecting boundary condition, that is

$$\frac{\partial \rho}{\partial r} = 0, \; r = b. \tag{4}$$

Note that $\partial\rho/\partial r$ is proportional to the flux of probability. For the boundary condition at the contact distance r = a, one needs to include the possibility of coalescence or "folding" (which tends to reduce the value of $\rho(a,t)$) and the possibility of dissociation or "unfolding" (which tends to increase the value of $\rho(a,t)$). In the absence of dissociation, the DC model boundary condition at r = a (see Ref. 7 for further details) is

$$\frac{\partial \rho}{\partial r}\Big|_a = \frac{\beta \rho}{\ell \gamma}\Big|_a \tag{5}$$

where β is the probability of reaction, ($0 \leq \beta \leq 1$, if $\beta = 1$, every MD-MD collision leads to folding) γ the probability of reflection at the target

(MD) surface and ℓ is the characteristic length. The interpretation of these parameters has been extensively discussed in Ref. 7.

In the DC model, it is assumed that the motion of the MDs can be described by a diffusion equation. This assumption is based on the forces existing in proteins and their magnitude relative to hydrodynamic damping effects. A theoretical study of the hinge bending mode in lysozyme[11] has shown that in spite of the large force constant ($k \simeq 3 \times 10^{13}$ ergs rad^{-2} mol^{-1}) for bending due to the interactions (covalent plus non-bonded) between the two lobes of the enzyme, the relative motion is diffusive in character; that is, the system is overdamped due to the friction from the solvent. For the present case, a corresponding argument should be valid. Although the smaller size of the MDs would yield a reduced frictional coefficient, the effective force constant for the relative motion is expected to be much smaller as well. If the diffusing units are two MDs that are adjacent in the sequence and have some kind of bend (e.g. a β-turn) between them, the energy stabilizing the turn would be the major contribution to the force constant; the magnitude of this is not known, though the available data and calculations suggest that the energy is not large. Another possibility is two MDs that are further separated, in which case the polypeptide backbone is unlikely to contribute significantly to bringing them together. Then the most important force is of the hydrophobic type. This is expected to be relatively short range; that is, until the two MDs are sufficiently close to exclude water molecules, no hydrophobic attraction exists. Consequently most of their relative motion will involve essentially free diffusion.

Because of the dissociation reaction in Eq. (1), the boundary condition at $r = a$ must be modified to incorporate the contribution to $\rho(a,t)$ from

dissociation of the MD pair into the individual MDs. To this end, let $n_a(t)$ be the number (fraction or probability when the initial probability density integrated over the diffusion space is normalized to one MD pair) of coalesced MD pairs at time t.

Since the number of uncoalesced pairs N(t) is given by

$$N(t) = \int_a^b dr\ \rho(r,t) \tag{6}$$

with N(o) = 1, one finds that

$$n_a(t) = 1 - \int_a^b dr\ \rho(r,t) \tag{7}$$

Furthermore, from the rate at which N(t) changes due to the reaction at r = a:

$$\frac{dN}{dt} = -D \frac{\partial \rho}{\partial r} \Big|_a \tag{8}$$

one may write $n_a(t)$ as

$$n_a(t) = D\int_o^t dt' \frac{\partial \rho}{\partial r} \Big|_a \tag{9}$$

In Ref. 7 it was found that to a very good approximation, $n_a(t)$ followed an exponential increase to its maximum value of one. This relatively simple behavior was quantified in Ref. 7 by using the mean coalescence time τ_c (see also Ref. 5 in which τ_c was introduced for protein folding) to approximate the folding kinetics to the native structure. The mean coalescence time is an extension of the concept of first passage time[12,13,14,15] to the DC model physical situation, and may be used when the probability density goes to zero as $t \to \infty$, the approximate situation during the folding step leading to the native structure.

As discussed in detail in Ref. 7, the mean coalescence time has the general form

$$\tau_c = \frac{\ell \Delta V \gamma}{D \beta A} + \tau_a \qquad (10)$$

where, as mentioned above, ℓ is the characteristic length, ΔV the finite diffusion volume (spherical symmetry), A the target surface area for coalescence, γ the reflection (non-coalescence upon collision) probability, and β the probability that both microdomains (or larger aggregates) are "folded" when they collide. Under the usual folding conditions β is expected to be small compared to unity and, therefore, γ is approximately unity and is usually incorporated into ℓ. τ_a, called the mean absorption time, represents the average time for two microdomains to coalesce if every collision were to result in coalescence. Under normal folding conditions, $\tau_c >> \tau_a$. For the diffusion limits r = a to r = b with coalescence occurring at r = a, one finds in one dimension, the results $\tau_a = (b-a)^2/3D$, $\Delta V = b-a$ and A = 1.

To include the effect of the dissociation (unfolding) reaction in Eq. 1, the probability density at the reaction boundary, r = a, must have a contribution proportional to the number of MD pairs already coalesced at time t, that is, Eq. 5 must be modified to be[16]

$$\left.\frac{\partial \rho}{\partial r}\right|_a = \frac{\beta}{\ell \gamma}\left\{\rho|_a - \frac{n_a(t)}{K_a(b-a)}\right\} \qquad (11)$$

with K_a being the equilibrium constant for Eq. 1, that is,

$$K_a = \lim_{t \to \infty} \frac{n_a(t)}{N(t)} \qquad (12)$$

the ratio of coalesced to uncoalesced MD pairs at equilibrium. In Ref. 16, it was found that use of the boundary condition in Eq. 11 leads one to consider

a quantity called the mean equilibrium time, τ_{eq}, defined to be

$$\tau_{eq} \equiv \int_o^\infty dt \; \frac{\{N(t)-N_{eq}\}}{1-N_{eq}} \tag{13}$$

which, for a uniform initial distribution, may be used to approximately represent $n_a(t)$ according to

$$n_a(t) \tilde{=} n_{a_{eq}} \{1 - e^{-t/\tau_{eq}}\} \tag{14}$$

the goodness of this approximation to $n_a(t)$ depending on the size of τ_{eq} compared (in one dimension) to the time unit $(b-a)^2/D$. τ_{eq} has a particularly simple form. It is

$$\tau_{eq} = \frac{K_a}{(1+K_a)} \; \tau_c \tag{15}$$

with τ_c defined by Eq. 10.

The exponential approximation to the coalescence probability $n_a(t)$ has been discussed in some detail above. This approximation is meant to replace the infinite series of time-dependent exponential terms which characterize diffusional processes in closed systems such as the intramolecular protein folding system. While it is not necessary to make the exponential approximation when using the DC model, it is certainly very convenient to do so and also gives a simplified description (to the extent that the model is valid) of how the gross factors important in folding kinetics combine to determine the folding rate. In addition, numerical studies made to date by the author and others[17], some of which have been discussed above, show that for most values of the relevant parameters, the exponential approximation to $n_a(t)$ is quite good, regardless of the actual application to protein

folding dynamics. Therefore, in discussing the DC model further, it will be assumed that the exponential approximation holds to the extent that it simplifies discussion without compromising the possible validity of the model.

In a real globular protein there will be a number of MDs whose mutual coalescence leads to the native, folded structure of the protein. For example, in α-helical proteins such as hemoglobin[18], the various helical segments into which the native structure is divided would be candidates for MDs (if one assumes that the interactions dominant in the native structure are also important on the folding pathway). In Ref. 7, the final step leading to the native protein structure was described and calculated with the DC model. Above, in the present paper, a preliminary step on the folding pathway leading to an unstable intermediate was described and the calculation of the probability $n_a(t)$ of a MD pair existing at time t was carried out. The next step is to consider the detailed DC dynamics of a hypothetical protein whose observable folding pathway consists of the two steps mentioned above, that is, the first step is the coalescence-dissociation reaction of two unstable MDs: the second step is the coalescence of a third MD with the first pair (when they are in the coalesced state), the final state being the stable native structure.

Consider three MDs connected in a linear fashion by relatively flexible portions of the polypeptide chain and interacting only by the trio of intramolecular reactions given by the two reactions in Eq. 1 and the following reaction

$$\begin{matrix}(MD)_1\\|||\\(MD)_2\end{matrix} + (MD)_3 \rightarrow \begin{matrix}(MD)_1\\|||\\(MD)_2\end{matrix}\!\!\diagdown\!\!\diagup (MD)_3 \text{ (native structure).} \quad (16)$$

Since the right hand side of Eq. 16 is the native structure of this protein the dissociation (unfolding) reaction is negligible and will not be considered. To obtain a folding pathway for which the probability of the "native structure" as a function of time may be calculated analytically, suppose that incorrectly folded intermediates are unimportant and that MD_3 does not interact significantly with either MD_1 or MD_2 separately but interacts strongly with the pair MD_1-MD_2 when they are in the coalesced state, the right-hand side of Eq. 1. These assumptions restrict the dynamics to a particular pathway leading to the native structure, and lead to the decoupling of the relative position probability density of the MD_1, MD_2 pair, call it ρ_{12}, from the relative position probability density of MD_3, call it ρ_3, except at the reaction boundary. That is, one may calculate the probability $n_{a12}(t)$ of the pair MD_1-MD_2 being folded together (as done above) independently of MD_3, and then use the results for $n_{a12}(t)$ in the β_3 parameter to calculate ρ_3 and $n_3(t)$ exactly, in principle. To summarize the physical picture of this assumed definite folding pathway for a three MD hypothetical "protein", imagine MD_2 to be fixed in space and MDs one and three diffusing about it but not interacting with one another. Then, MD_3 will not coalesce until MD_1 has, first, coalesced with MD_2. Because of the requirement of a definite order of coalescence, the diffusion problems for MD_1 and MD_3 are independent except at the boundary with MD_2 where the probability of MD_1 being coalesced influences the probability of MD_3 coalescing but not the reverse, since, by assumption, the triplet is the native structure and does not dissociate appreciably.

For this example, the following set of equations must be satisfied:

$$\frac{\partial \rho_{12}}{\partial t} = D_{12} \frac{\partial^2 \rho_{12}}{\partial r_{12}^2} \qquad a < r_{12} < b \tag{17}$$

$$\frac{\partial \rho_{12}}{\partial r_{12}} \Big|_{r_{12} = b} = 0 \tag{18}$$

$$\frac{\partial \rho_{12}}{\partial r_{12}} \Big|_{r_{12} = a} = \frac{\beta_{12}}{\ell_{12}\gamma_{12}} \{\rho_{12}|_{r_{12} = a} - \frac{n_{a_{12}}(t)}{K_{12}(b-a)} \} \tag{19}$$

which are Eqs. 3, 4, and 11 with the notation modified for the present example, and

$$\frac{\partial \rho_3}{\partial t} = D_3 \frac{\partial^2 \rho_3}{\partial r_3^2} \qquad a' < r_3 < b' \tag{20}$$

$$\frac{\partial \rho_3}{\partial r_3} \Big|_{b'} = 0 \tag{21}$$

$$\frac{\partial \rho_3}{\partial r_3} \Big|_{a'} = \frac{\beta_3}{\ell_3 \gamma_3} \; n_{a_{12}}(t) \; \rho_3 \Big|_{a'} \tag{22}$$

the equations satisfied by the relative position probability density of MD_3. Until MD_3 collides with the already coalesced pair MD_1-MD_2, there is no interaction (by assumption) between MD_3 and either MD_1 or MD_2. Thus, the equations for ρ_{12} and $n_{a_{12}}$ may be solved independently of MD_3. This, in fact, has already been carried out in the preceding section. The result for $n_{a_{12}}(t)$, which is the probability that MDs one and two are coalesced,

is then needed in the boundary condition, Eq. 22, describing the coalescence of MD_3 with the pair.

The time-dependent factor $n_{a_{12}}(t)$ in Eq. 22 makes the solution of the set of equations for $n_3(t)$ difficult. However, under many circumstances relevant to protein folding and within the framework of the DC model, an approximate solution for $n_3(t)$ may be derived. The results may be summarized in terms of the characteristic time constants τ_{a_3}, τ_{c_3}, and $\tau_{eq_{12}}$ with

$$\tau_{a_3} = \frac{(b'-a')^2}{3D_3} \tag{23}$$

and

$$\tau_{c_3} = \frac{\ell_3\gamma_3(b'-a')}{D_3\beta_3 n_{a_{12}eq}} + \frac{(b'-a')^2}{3D_3} \quad . \tag{24}$$

Also of importance is the ratio

$$y \equiv \frac{\tau_{c_3} - \tau_{a_3}}{\tau_{a_3}} \quad . \tag{25}$$

In terms of these time constants, the approximate analytical result for $n_3(t)$ is

$$n_{a_3}(t) \simeq 1 - e^{-t/\tau_{c_3}} \{[1 + \frac{1}{y}(1 - e^{-t/\tau_{eq_{12}}})]^y\}^{\tau_{eq_{12}}/\tau_{c_3}} \tag{26}$$

It is often true under protein folding conditions that τ_{a_2} and τ_{a_3} are very small compared to their overall folding time counterparts, τ_{c_2} and τ_{c_3}, respectively. In this case, as an approximation to Eq. 26, one may look at the limiting case in which $y \to \infty$. One obtains

$$n_{a_3}(t) \simeq 1 - e^{-t/\tau_{c_3}} \quad e^{(1-e^{-t/\tau_{eq_{12}}})} \frac{\tau_{eq_{12}}}{\tau_{c_3}} \qquad (27)$$

One notes that as $\tau_{eq_{12}}/\tau_{c_3} \to 0$, Eq. 27 approaches the mean coalescence time approximation to $n_{a_3}(t)$ which assumes that the preceding intermediates have come to equilibrium with the individual MDs before the next reaction leading to the native structure occurs to any significant degree. Significant corrections to the $(1-e^{-t/\tau_{c_3}})$ approximation for $n_{a_3}(t)$ will be required whenever $\tau_{eq_{12}}$ is not substantially smaller than τ_{c_3}. As mentioned above, under protein folding conditions, their ratio has the approximate value

$$\frac{\tau_{c_3}}{\tau_{eq_{12}}} = \left(\frac{\ell_3}{\ell_{12}}\right)\left(\frac{(b'-a')}{(b-a)}\right)\left(\frac{D_{12}}{D_3}\right)\left(\frac{(1+K_{12})}{(K_{12})^2}\right)\left(\frac{\beta_{12}}{\beta_3}\right) \qquad (28)$$

To the extent that the first three sets of brackets on the right hand side of Eq. 28 are of order unity, the ratio of folding times depends on the various equilibrium constants. As discussed in Ref. 7,

$$\beta_{12} = \frac{K_{12}}{(1+K_1)} \frac{K_2}{(1+K_2)}$$

where K_1 and K_2 are the individual MD folding equilibrium constants (coil-helix equilibrium constants for helical MDs). Similarly,

$$\beta_3 = \frac{K_3}{1+K_3}$$

(note that whereas β_{12} is a product of two terms, each of which refers to a single MD, the product $\beta_3\, n_{a_{eq_{12}}}$ occurs in the rate of approach to the native structure and it depends on all three MDs). In general, one will find the rates of subsequent reactions on the folding pathway depending in the above way on the equilibrium constants of the preceding reactions.

Since all of the equilibrium constants are expected to be much less than one in this example, Eq. 28 may be approximately written

$$\frac{\tau_{c_3}}{\tau_{eq_{12}}} \cong \frac{K_1 K_2}{(K_{12})^2 K_3} \tag{29}$$

As a further reasonable approximation for this estimate, one may set the MD folding equilibrium constants equal, that is $K_1 = K_2 = K_3$. This leaves the ratio of folding times to be determined by the quantity $K_1/(K_{12})^2$. Since $K_1 \sim 10^{-3}$ (as estimated in Ref. 5) and K_{12} is not expected to be more than an order of magnitude greater than K_1, the ratio of times in Eq. 25 could be greater than ten and thus a prior equilibrium approximation may be justified.

STATIC ASPECTS OF FOLDING

The structure of a monomaic protein is conveniently divided into three parts. There is first the primary structure which is a description of the ordered sequence of amino acid residues in the polypeptide chain. It is thought that this sequence, along with environmental effects, determines the thermodynamic or kinetic native structure of the protein. However, study of the sequence does not directly provide detailed information about the three-dimensional structure and function of the protein in most cases.

The next step in the organization of a protein (both logically and probably physically, as well) is the propensity of amino acid chains to form a number of well-ordered local structures with definite symmetry. Principal among these local structures is a helix (commonly a right-handed α-helix). Because of its symmetry a helix is easy to recognize in three-dimensional protein structures. Helices are stabilized by interactions among neighboring amino acid residues. In particular for a right-handed α-helix, there are 3.6 residues per turn, a translation of 1.5 A° per residue along the helix axis. Neglecting end effects, each peptide carbonyl in the α-helix is a hydrogen bond acceptor for the peptide N-H donor four residues away. Among the commonly occuring amino acids, only proline is sterically prohibited from fitting into an α-helix.

Among the other possible secondary structural elements, the most common is the 2-fold helix or pleated structure known as the β-strand. This structure is not stable independently, because although the peptides form hydrogen bonds, they do so with other β-strands along the polypeptide chain to form a β-sheet. This calls, however, for a higher level of organization than is required for

a simple helix with intra-helical hydrogen banding such as an α-helix. The β-sheet structures may occur in parallel or antiparallel forms depending on the relative orientation of the two β-strands. Both are known to occur in proteins.

Also found in short stretches of globular proteins is the 3_{10} helix with three residues per turn instead of the 3.6 residues of the α-helix.

Another definite secondary structural feature that is common in globular proteins is the β-bend. This is a chain-reversal turn in the polypeptide chain in which four peptides participate with a hydrogen bond between the first and fourth peptides. Thus, this structure may be classified as a helix with zero pitch. Further classification of chain turns may be carried out to broaden the scope of this type of secondary structure.

A separate type of secondary structure, the polyproline helices, occurs commonly in proteins of the collagen family but is quite rare in globular proteins of the single chain variety.

When one examines the three-dimensional structure of a globular protein as determined by its x-ray crystal structure (approximately 10^2 structures of this type are known), one observes that elements of secondary structure, as discussed above, are arranged in space in definite three-dimensional patterns. These patterns are termed the tertiary structure of the protein, and it is to understand and be able to predict these structures that theoretical and experimental studies of proteins are aimed.

PREDICTING PROTEIN STRUCTURE AND FUNCTION

The aims of theoretical research on protein structure and dynamics are two-fold. On the one hand, one has the fundamental aim of understanding at the molecular level the properties of protein molecules since they play such an important role in the life processes. On the other hand, one has the "practical" aim of being able to predict protein structure and dynamics in order to design drugs, catalytic enzymes, etc. Thus, one needs to develop methods for predicting protein conformation, conformational pathways to a given conformation from the unfolded structures as well as among folded conformations, the relative populations of these conformations, the rates of transitions among them and for bimolecular transformations (allosteric transitions, enzyme catalysis, etc.) their rates and changes in rates due to structural or environmental changes.

In summary, the principle aims of theoretical research on protein structure are to predict the tertiary structure of globular proteins and to predict the functional aspects of the protein from its tertiary structure with the prediction including any dynamical aspects of the structure which are relevant. The starting point for any tertiary structure prediction is the primary structure of the protein, the linear sequence of amino acids that makes up the polypeptide chain, and which in conjunction with the particular environment in which the polypeptide finds itself determines the tertiary structure.

Since the tertiary structure of most proteins is made up of the secondary structural elements mentioned above and since some, at least transient, secondary structural elements are required to account for the kinetic aspects of folding from the primary to the tertiary structure as discussed above, it

is important to have an understanding of and be able to predict theoretically the principal secondary structures observed in globular protein, namely α-helices, β-strands (and sheets) and β-bends. There are a large number of methods of predicting secondary structure from the amino acid sequence,[19-28] all of which assume that only the short-range interactions between residues near one another in the primary sequence determine the local secondary structure. The methods of prediction may be classified as statistical[19-25] (analyzing known structures for the propensities of individual amino acids) and "physical" where the sizes and hydrophobicities of residues are considered in determining their local structure.[26-28] As shown by several comparative studies,[29-31] no individual method of predicting secondary structure is clearly superior to the others. As is also shown by such studies, the success of such predictive schemes is not increasing with time as the data base of crystal structures of protein increases. Thus, it appears that the formation of secondary structure, at least to the extent that the structure remains an element of the tertiary crystal structure, is determined in part by long-range interactions between secondary structural elements. Therefore, the production of secondary structure appears to require tertiary structure information and probably the production of tertiary structure requires secondary structure information so that to a certain extent, the native structure of a protein represents a self-consistent boot-strap kind of final state, and, thus, probably requires some dynamical scheme to be used as part of the folding algorithm to predict the tertiary structure. Such a scheme may be provided by the DC model discussed above.

In order to clarify the problems encountered in predicting the tertiary structure of a small globular protein, consider staphylococcal nuclease which contains 149 amino acid residues and no disulfide bonds or other restrictive interactions (e.g. heme group) to restrict the possible conformations of the native structure. The secondary structure consists of three α-helices: (residues 54-67, 99-107, 122-134) and three β-sheets (three-strand; residues 38-41, 108-113). If one assigns three possible energy minima to each amino acid residue, then a simple statistical count gives $3^{149} \simeq 10^{71}$ possible conformational states for this small protein. There are also more than 1500 atoms to be considered in any kind of atomic resolutions energy minimization scheme to obtain the global energy minimum of this system (plus solvent interaction). Each of the atoms is, in principle, interacting with the other atoms via electrical forces, and since each atom has one or more polarizable electrons the net potential energy of interaction which must be minimized to obtain the global energy minimum is very complex.

If the molecule were completely static, then the potential energy would be a sum of the individual Coulomb interaction energy terms for each atom with its associated electrons. Because of the polarizability of the electrons and also the polarizability of some of the hydrogen atom protons (hydrogen bond formers), a more complex potential energy of interaction emerges. In practice, any description of the protein molecule potential energy function to be used in a tertiary structure calculation will be put in somewhat different terms, that is, in terms of the bond directions and bond angles between atoms which are covalently bonded and the various non-bonded interactions which make up the greater part of the stabilization energy of the tertiary

structure (hydrogen bonds, van der Waals interactions, hydrophobic interactions). Such a description may be found, in principle, for a folded protein with a known crystal structure from the crystallographic coordinates, but, of course, the essential non-bonded interactions are not known for most proteins. To start from a completely described polypeptide chain in a random conformation and attain the folded structure with lowest energy by energy minimization is not possible at the moment for two reasons. First, the size of the computational problem is too great, and second the time to compute the structure even if possible is too long for an individual lifetime.

An alternative approach is to reduce the computational problem so that in the earlier stages of folding,groups of atoms are treated simultaneously. This has been done by several groups with limited success, since the simple structural representation and energy minimization techniques used appear to proclude the possibility of agreement with the known crystal structures. However, this method in modified form may have promise for further development.

Some problems involving protein structure and dynamics are outlined below along with background material and potential methods of attack.

1. Myoglobin Folding Kinetics

The three-dimensional structure of myoglobin was shown by Kendrew, et. al.[32] to be composed of a number of α-helical segments, connected to one another in a linear fashion by short lengths of polypeptide chain, and interacting via non-covalent bond forces in a well-defined way so that a globular container is formed for the heme group. As the first protein for

which a high resolution structure was determined, and with most of the amino acid residues occupied in quite regular helical arrays, myoglobin ought to be a prime candidate for having its folding pathway be well established and even successfully computer simulated. However, the lack of kinetic information on the experimental side and lack of success to date on the theoretical side in folding simulations brings strongly to the fore the difficulties which one faces in unraveling the folding mechanism.

The problem to be faced in understanding the folding of myoglobin is to understand how the native structure is found from the $3^{153} \cong 10^{72}$ conformations available to the polypeptide chain if each amino acid has three independent configurations. As mentioned above, the solution to a problem of this type is to reduce the space of configurations to be searched by considering the collective concerted motion of a number of subunits of the polypeptide, the number being small compared to the number of amino acid residues,in order to reduce the search problem to tractable size. When this approach to the folding problem is considered, it is necessary to define the subunits (MDs) which collectively interact. The general characteristics of MDs have been discussed above, and it is clear that prime candidates for MDs would be secondary structural elements which are observed in the crystal structures of most proteins, namely, α-helices, β-strands and β-bends being careful to recall that secondary structural prediction methods are not completely reliable because some crystal secondary structure elements are stabilized by tertiary interactions. Therefore, in the initial stages of a folding simulation,it is necessary to consider a variety of MD sets and to test which set or sets lead to viable tertiary structures. If one uses only the MD set observed in the

crystal, then the result (which ought to resemble the crystal structure) is built into the simulation.

By using the secondary structure prediction algorithms, one generates a set of myoglobin MDs[29-31] which mainly reproduce the known myoglobin secondary structure. Of course, in the unfolded myoglobin molecule most of the polypeptide chain will be in random coil conformations and the predicted secondary structural elements will occur only transiently. The next step in the folding simulation will be to generate pathways of multi-MD interactions. Again, one must be extremely careful not to bias the resulting tertiary structures by allowing only these MD-MD interactions that appear in the crystal structure. Instead, all possible interactions among the various helical segments must be considered to determine whether or not the observed crystalline set is dominant.[33,34] The tertiary structure or set of tertiary structures generated by the folding simulation may be further refined by energy minimization techniques[35-37] (see below) to produce the final predicted structure. Since the full energy minimization only takes place on an already roughly folded protein molecule, the constraints of time and financial resources, which prohibit minimization of an unfolded structure without gross reduction of the parameter space, are negated.

The algorithm outlined above represents one method for attacking the folding problem. In the next section an alternative method is outlined which also avoids the time-money computer crunch.

2. Protein Folding Using Simplified Representations of Residues

A protein of 100 residues has about 1500 atoms and 400 degrees of

freedom (single bond torsion angles). Calculating its free energy (particularly when considering interactions with rapidly moving solvent molecules) is an impossible task at present. Even if one considers only small protein fragments such as an α-helical segment the direct computation of conformational alternatives is too formidable. Nevertheless, since MDs (α-helical, β-strand, β-bend) clearly play an extremely important role in folding, it will be necessary to attain a fairly complete theoretical understanding of them.

Therefore, for direct study, one must, at present reduce the number of degrees of freedom by using a simplified model of the polypeptide chain. A model which has had some success is to represent each amino acid residue by a soft sphere with the volume of the true residue being maintained in this spherical approximation. This kind of representation has been shown by Richmond and Richards[34] to describe the packing of the residues in myoglobin, and it has been used by Levitt and Warshel[35] in computer simulation studies of the folding of trypsin inhibitor. Although this type of model will not describe detailed atomic interactions, it will give an adequate approximation to the overall structure and mobility of peptide fragments with regard to the more general interresidue constraints such as steric effects and hydrophobic interactions.

In detail, in this simplified representation of a polypeptide chain, the spheres representing near neighbor residues will be linked by virtual bonds (see e.g. Flory[38]) with harmonic restoring forces. Further neighbors will interact via an excluded volume spherical potential and an attractive van der Waals and solvent potential. Stabilization of helices, β-sheets or β-bends due to hydrogen bonding will be simulated with an additive potential chosen to have a maximum size for the secondary structure under consideration and to fall off rapidly to zero at other angles, the correct parameters

being chosen by comparison to peptides in aqueous solution (see e.g. Ref. 39 for an application to the α-helix).

3. Internal and External Friction Effects in Diffusion of Connected Polypeptide Segments

The main driving force in the intramolecular motion of MDs comes from random collisions with solvent molecules and with other parts of the polypeptide chain. Thus, the kinetics of folding is expected to be viscosity-dependent through diffusion coefficients for the various MDs and MD aggregates. The precise dependence of folding rates on solvent viscosity is controlled by the extent to which internal friction effects due to portions of the polypeptide chain moving over one another rather than through solvent only play a role in the folding pathways.

In earlier papers,[5,7] since the emphasis has been on obtaining the essential ingredients of the early stages of a DC model folding pathway, it has been assumed that the MDs diffuse freely until they are close together as defined by the interaction radius $r = a$. Their interaction to coalesce has been contained in the boundary condition at this radius (Eq. 11), and the diffusion coefficient D has been assumed to be constant, leading to Eqs. 13-15 for the MD-pair coalescence probability. An alternative approach involving potential barrier effects and variable diffusion coefficients was discussed briefly in Ref. 7. In either treatment, a general calculation allowing for a variable diffusion coefficient (for example, a transition from external to internal friction effects at a particular distance of separation of MDs) leads to time constants that depend both on the external and internal viscosity.

There are several experiments which bear upon the question of viscosity dependence of folding rates. Haas et. al., [40] 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 an acceptor of electronic excitation energy in solvent mixtures of different viscosities. With an assumed theoretical analysis, diffusion coefficients were derived which increased systematically upon decreasing the solvent viscosity. The values obtained for the diffusion coefficients were about an order-of-magnitude smaller than the values expected for the diffusion coefficients of the free chromophores in solvents of comparable viscosity, and appear to have a solvent viscosity independent part when one considers D^{-1}, that is, the friction coefficient, although this effect may be model dependent. In any case, there is a clear dependence of diffusion coefficient on solvent viscosity in this intra molecular, diffusion mediated interaction.

Tsong and Baldwin,[21] on the other hand, in their study of the kinetics of folding of the two forms of unfolded ribonuclease A (with all disulfide bonds intact) as a function of solvent viscosity, by adding either sucrose or glycerol, found no dependence on solvent viscosity, the rates of both folding reactions being either unchanged or slightly faster in the presence of sucrose or glycerol.

In the same system, Tsong[42] has recently found a reaction which is strongly dependent on solvent viscosity and somewhat faster than the reactions observed in Ref. 41. Tsong[42] also observed the two solvent viscosity independent reactions. Thus, in all systems studied to date, there is a strong solvent viscosity dependence to the reaction rate, as well as a solvent viscosity independent contribution.

The precise interpretation of these experimental results is not completely clear, at present, due to the complexity of the systems involved. However, it appears that diffusion mediated reactions are playing a significant role, and thus the reactions could be interpreted with the DC model. Further experimental studies on other systems are necessary to obtain a complete understanding of this effect.

In previous work[5-7] and above in this paper, it has been assumed that the diffusive motion of the MDs is essentially free. That is, until they approach to the distance r = a, their relative motion satisfies Eq. 2 with $V(\vec{x}) = 0$ and constant D. All the MD-MD interaction has been placed, in the above calculation, in the boundary conditions, that is, the forces influencing the motion of the MDs are assumed to be short-range. While probably true of the MD-MD hydrophobic interactions, the nature of the polypeptide chain between diffusing MDs may well contribute to the potential energy function in Eq. 2 at longer ranges. For example, in a theoretical study of the hinge bending mode in lysozyme, McCammon, et. al.[43] found that the potential energy function for the relative motion of the two lobes was relatively harmonic, but that, nevertheless, the motion was overdamped due to the frictional drag of solvent, that is, the relative motion satisfied the Smoluchowski equation (eq. 2).

To treat potential energy effects due to the polypeptide chain between interacting MDs in a rigorous way via a potential energy function is outside of the scope of analytical calculations since for most potentials, it, then, becomes impossible to obtain even approximate analytical results. Two points may be made however, concerning the effect of an intervening potential. First, if the intervening potential may be approximated by a

periodic potential, then the diffusive motion is essentially the same as that for a free Brownian particle,[44] the difference being that the diffusion coefficient is renormalized to a smaller value. The second point is that the results of the viscosity experiments of Haas, et.al.[40] may be interpreted to indicate that an intervening polypeptide chain introduces, in addition, a term independent of the external solvent viscosity into D^{-1}. If this turns out to be generally true for proteins, than the effect of the intervening chain may be fairly readily introduced into the DC model.

Further work along these lines may be carried out by 1. studying a bead-spring type of model polymer to approximate a polypeptide chain in order to get a feeling for the effect of the intervening potential on the interaction rate of the polymer ends (for example); 2. examining the DC model with interaction contained in the potential (rather than exclusively in the boundary conditions) and with a variable diffusion coefficient.

4. Evolutionary Implications of DC Folding Mechanism

The DC model envisions a protein molecule to be made up of several connected MDs which interact with a sequence of diffusion-collision-coalescence-dissociation steps until eventual coalescence in a cooperative manner into the relatively stable native structure. Particularly in the early stages of folding there would be possible a number of different sequences of folding steps which later converge toward the rate determining slow step.

The stabilities of multi-MD intermediates are determined by the non-covalent attractive interactions among the MDs which are, in turn, controlled by the sequences of amino acid residues. One would expect the viability of individual pathways to be sensitive to changes in the non-covalent interactions and thus to be amenable to evolutionary advances via mutational changes in the residue sequence of MDs.

There are (at least) two different ways in which such an evolutionary scheme could manifest itself. First, suppose that the mutations cause residue changes that do not change the basic character of the MDs but only their potential for attractive interaction. The result could then be a new native structure in which the same MDs are organized in a somewhat different way. Since the MDs are the same, it might be possible for the protein to retain some of its previous function and, at the same time, carry on a new function. Another possible result could be two native structures (one metastable but long-lived) which could, in turn, be separated by gene duplication and mutation to evolve separately. Second, suppose that the secondary structures of individual MDs are fluctuating among alternative forms (say α-helical and β-strand) and that mutation changes the stability from one to the other. Then a new protein could be formed with different MDs (although the same number as before). Again, a more viable possibility would be a pair of folded conformations with gene duplication leading to divergent evolution. A careful study of known sequence and structure for globular proteins might produce some evidence for these possibilities.

5. General Description of Biomolecular Diffusion-Mediated Processes

In the above discussions of protein folding, the dynamics is supplied by the DC model which discusses the diffusion of connected polypeptide segments. In many biologically important circumstances it is the diffusion-mediated interaction via collision and coalescence which is important dynamically. Therefore, it is necessary to formulate the bimolecular analogue of the DC model, which will have relevance in several of the topics outlined in subsequent sections.

Consider, for example, a spherically symmetrical infinite system of molecules of initial uniform concentration ρ_o surrounding a target of radius a centered at r = 0. The molecules diffuse according to Eq. 2 (assume V = 0 for simplicity) with concentration $\rho(r,t)$. Instead of Eq. 11, one has the boundary condition at the target surface r = a

$$aL \left.\frac{\partial \rho}{\partial r}\right|_a = \rho|_a - \frac{4\pi a^2 \rho_o D}{K} \int_o^t dt' \left.\frac{\partial \rho}{\rho r}\right|_a$$

For the other boundary condition, one may choose $\rho(r,t) \to \rho_o$ as $r \to \infty$ The diffusion equation with the above boundary conditions must be solved to obtain the rate of association $4\pi a^2 D \left.\partial\rho/\partial r\right|a$ as well as the probability $n_a(t)$ that a target-molecule pair is formed before time t. Preliminary results[45] indicate that for a concentration of targets also equal to ρ_o that

$$n_a(t) \underset{=}{\sim} \frac{K(1 - e^{-t/\zeta})}{1+K(1 - e^{-t/\zeta})}$$

with

$$\zeta \equiv \frac{K(1+L)}{4\pi a D \rho_o} = K(1+L)T$$

where $T \equiv 1/(4\pi a D \rho_o)$ is the basis time unit for this type of diffusional system.

It will be necessary to analyze this basic bimolecular process further in order that application may be made to systems of biological interest. Several possible applications are outlined below including glucagon interactions and collagen folding.

6. Glucagon Trimer Formation and Glucagon-Receptor Interaction

Glucagon is a polypeptide hormone with 29 amino acid residues[46] which is synthesized and stored in the α cells of the islets of Langerhans of the pancreas. The hormone activates glycogenolysis and gluconeogenic pathways resulting in raised blood glucose levels, by specific binding to a plasma membrane receptor site on the regulatory component of adenylate cyclase of liver and other cells, which give rise to an increase of intracellular levels of the second messenger cyclic AMP.[47] To understand the glucagon-receptor interactions, one must know the conformation of the hormone when bound to the receptor. Solution studies[48, 49] show that glucagon exists as unordered structure in dilute solution but self-associates at high concentration in a trimer with a high α-helical content. X-ray analysis[50] shows that in crystals the polypeptide adopts a mainly helical conformation, which is stabilized by hydrophobic interactions between molecules related by threefold symmetry.

Since it appears that the glucagon ordered structure requires the additional interaction provided by another hydrophobic surface such as another glucagon molecule or a receptor site, a DC mechanism both "in vitro" and "in vivo" is a reasonable possibility for the molecular folding mechanism of this small polypeptide.

Application of the DC mechanism to glucagon-glucagon interactions will require some modification of the formalism described in Ref. 7 and above because the interacting segments are no longer in the same molecule but residue on different glucagon molecules or in the receptor site. The basic idea would still, however, be the same. That is, α-helical MDs in a glucagon molecule are in equilibrium between the "random coil" conformations and the ordered helical conformation. When, by diffusive movements the MDs of two molecules come into contact when both are ordered, they may have additional

stabilizing interactions to form a dimer intermediate state. The dimer may in turn further interact with another monomer to form the stable trimer, the equivalent of the native structure in the usual folding model. Alternatively, the mechanism might involve the simultaneous diffusion-collision-coalescence of all three monomers to form the trimer, depending on the stabilities of the individual MDs. A similar mechanism in which the α-helical MDs of glucagon are stabilized by interaction with the receptor site may be envisioned. Elucidation of the folding pathway to the trimer by NMR techniques may be feasible, at least to distinguish between the two proposed pathways.

In order to utilize the DC model dynamics for the above physical situation, the formulation described in the previous section must be used since the interaction is a bimolecular one.

7. Dynamics of Protein-Nucleic Acid Interactions

On the basis of single crystal X-ray diffraction and circular dichroism studies of protamine binding to a t-RNA, it has been suggested[51] that the protamine molecule changes its conformation from a random coil to a structure containing α-helices on binding to t-RNA. This could be an example of the DC model mechanism operating in a protein-nucleic acid system and one could speculate that such a mechanism is important in hormone-receptor and nucleic acid operator-repressor interactions.

Particularly in the interactions with nucleic acids at specific sites, the ligand is faced with the difficult problem of finding a site along the rather lengthy nucleic acid chain. It has been suggested[52] that a diffusional search in space can be considerably speeded up by confining the search to a space of lower dimensionality. For example, in nucleic acid-ligand interactions, the space could be reduced from three to one by having the ligand bind loosely to the nucleic acid and then diffuse along the chain until the specific site is found at which point tight binding would occur, perhaps with a change of conformation. A study of this hypothesis will require knowledge of diffusion along a helical path, probably using methods similar to those used in spherically symmetrical and spheroidal systems.[53]

8. Membrane Dynamics

The motions of molecules in biological membranes is a subject of considerable current attention.[54] One aspect of this subject is the rates of lateral diffusion of lipids and membrane proteins in the same system[55, 56] with the former generally diffusing much faster than the latter. These differences in the measured diffusion coefficients are too large to be explained on the basis of size differences alone according to the current description of size effects in membranes.[57]

In many cases, lateral diffusion measurements correspond to one-dimensional motion so for a specific example, consider motion in one-dimension in the lateral direction in a membrane in which the position r of a specific kind of molecule is followed as a function of time t. Then concentration will then satisfy Eq. 2 in one-dimension and the diffusion coefficient D is defined by the equation

$$D \equiv \lim_{t \to \infty} \overline{(r - \bar{r})^2}/2t \tag{33}$$

where the bar indicates an average of the quantity over the diffusion space weighted with the normalized concentration. The signature of diffusion is then the existence of D as defined by Eq. 33. When V(r) in Eq.2 is not constant and, in fact, goes to infinite at some point in space, there is no true diffusional motion. It seems unreasonable to expect that V(r) = 0 everywhere in the membrane. Nevertheless, the interpretation of the experiments mentioned above[54-56] is carried out on that basis. This leads to two problems. First, an incorrect diffusion coefficient may be extracted from the experimental results, and, second, even if the correct diffusion coefficient is extracted because the definition of Eq.33 is essentially used,

the diffusion coefficient may be interpreted as being for a free molecule rather than a molecule under the influence of an external force due to the membrane structure.

In fact, since membranes are regular structures, one would expect that the external force and hence the potential V(r) in Eq.2 would be periodic in space. In this case it is possible to derive a closed form expression for D which might apply to diffusion in membranes. In any case, it is clearly necessary to inspect the methods for extracting D from the raw experimental results and to calculate D for a periodic one dimensional potential. The aim of this kind of analysis would be to learn something about the structure of the membrane (its periodicity in space as seen by a diffusing protein, perhaps) by measuring D and interpreting it in terms of potential barriers to free diffusion.

9. Folding Dynamics of Collagen

In some respects, the folding of collagen to the triple helical state ought to resemble the folding mechanism suggested above for the formation of trimers of glucagon molecules. But, on the other hand, one might expect considerable differences in the underlying mechanism for two reasons. First, the individual chains of collagen molecules form helical secondary structures which have no hydrogen bonds to promote stability, but rather depend on repulsive interactions (steric constraints) that cause each side chain to get far away from the others. Second, the tertiary structure is formed by twisting individual helices about one another, much more like double helical nucleic acids than like globular proteins.

These differences allow collagen to some extent to be treated theoretically more like coil $\rightleftharpoons$ helix type of transitions rather than like coil $\rightleftharpoons$ native structure type of transitions as in globular proteins. Nevertheless, the actual folding mechanism is not completely clear[58, 59] and much work both experimentally and theoretically needs to be carried out to understand fully the dynamics of this important biochemical process. On the experimental side, detection of helical regions by chiroptical methods would appear to be useful at least for the longer time aspects of the transition, and perhaps, also for the faster processes which are probably not being observed at present. In particular, if the reaction proceeds through the trimer, one ought to find some spectroscopic evidence of double helical structures.

Utilizing the general theory of bimolecular processes outlined above, one may proceed with the formulation of a dynamical model for triple helix

formation by utilizing the dimer formation probability in a calculation of the trimer formation probability, as was outlined in the discussion of the DC model for protein folding above. This can be compared with a direct triple helical formation calculation and with the existing experimental evidence mentioned above, although it appears that faster detection methods are necessary as indicated above, and so this particular project would require both theoretical and experimental efforts.

10. Molecular Dynamics of Folded Proteins

As mentioned previously, most globular proteins have a well-defined equilibrium structure in the nature state with, however, their flexibility and structural fluctuations playing an essential role in their biological activity. One direct theoretical method for studying fast (subnanosecond movements)is molecular dynamics.[60] In this method, one assigns initial positions and velocities to each of the atoms in the system and then solves the classical equations of motion simultaneously for all the atoms with the forces driving the motion being determined from the potential energy of interaction of the constituents. The potential energy is approximated in these systems by an empirical potential energy function with the form of a sum of terms corresponding to the interactions among the elements of the protein itself and separately the interactions of the protein with its environment. The former consists of terms for bonds, bond angles, torsional angles, van der Waals interactions, electrostatic interactions and hydrogen bonds, and the latter would contain only van der Waals, electrostatic and hydrogen bond terms. In general, the extended atom approach is used in actual calculations, in which each non-hydrogen atom and any hydrogens bonded to it are replaced by one extended atom. An example of this mapping is given in Ref. 61 along with a discussion of the specification of the potential energy functions for bovine pancreatic trypsin inhibitor. One point of interest to protein folding studies which has been observed in molecular dynamics simulations is the existence of sizable frictional effects in the displacements of atoms and groups of atoms, i.e., fluctuations from the

average structure were found to be subject to rapid damping. This would clearly bear on the question of internal versus external friction discussed above. It was found, in fact, in an extension of the study initiated in Ref. 60 that torsional fluctuations of buried tyrosine residues in tryspin inhibitor obeyed the Langevin equation for an harmonic oscillator.[62]

Although attractive, in principle, the direct molecular dynamics method described above has the great drawback that large scale computation efforts are required to obtain even 100 ps simulations of internal protein dynamics. Unfortunately, many of the most interesting processes biologically occur as activated processes with rates of 10^9 sec^{-1} or less, considerably slower than the molecular dynamics time frame (e.g. the chemical events associated with enzyme catalysis). Therefore, it is necessary to develop more specific dynamical methods appropriate for particular problems (e.g. for the initial stages of protein folding, highly simplied potential functions describing the forces between MDs as mentioned above). Particularly interesting from this point of view would be the modeling of domains and domain dynamics in certain enzymes such as lysozyme[63] which has two lobes and, in particular, the kinases[64, 65] where large structural changes involving closing of the active site cleft occur on substrate binding. One approach to the study of this kind of motion is to model the domains by geometrical figures (spheres in lowest approximation) and the potential energy of interaction obtained from the potential described above for molecular dynamics simulations when the domains are rotated with respect to one another. This approach has met with some success with lysozyme[63] but has not been applied to other systems such as the kinases or immunoglobins (which are known to have a very well-defined hinged domain structure). This is clearly a problem with many ramifications and should be pursued further.

11. Small Molecule Penetration of Protein Interiors

Globular proteins have closely packed interiors in the sense that the packing is as dense as that found for crystals of small organic molecules,[66] amino acids[67] and small peptides. Nevertheless, it is accepted that the basically static property of close packing is the average in space and time of a wide variety of instantaneous structures and fluctuations the former rapidly interconnecting due to the latter. Since individual proteins are small compared to the macroscopic structures for which there would be extremely small fluctuations in thermodynamic properties,[68] one expects to have, on the contrary, relatively large fluctuations.[69] Such fluctuations can lead to a series of "holes" or channels to the protein interior sufficient to allow entry of water and other small molecules and to local variations in temperature.

As has been amply demonstrated by hydrogen exchange,[70] O_2 quenching of fluorescence[71] and CO binding at the heine group in myoglobin[72] small molecules do reach the interior of globular proteins by some mechanisms, with rates, in the case of hydrogen exchange retarded over a range of eight orders of magnitude compared to unstructured, random coil polypeptides. The mechanisms for such processes are of great interest because of the importance of intraprotein dynamics in protein function. A possible model of this process is outlined below.

Considered from an abstract, physical point of view (and in one dimension for simplicity), a small molecule starting from a surface point on a globular protein must get from one side to the other of an irregular potential energy barrier in order to attain a particular location of the protein interior. There are two known mechanisms for getting from

one side of a potential barrier to the other side, penetration through the barrier by quantum mechanical tunneling and classical diffusion over the barrier. Higher energy (temperature) favors the latter mechanism so it would be classical diffusion over the barrier which would be responsible for the penetration at room temperature. An alternative point of view is that a large scale unfolding fluctuation occurs exposing the target to the ligand without the necessity of overcoming any barriers except in the protein itself during the unfolding process. For a diffusional model, the rate of penetration to the target site and in a rough approximation would be equal to the inverse of the first passage time for overcoming a series of potential barriers. An additional factor which ought to be included in such an analysis is the probable time-dependent nature of the potential barrier, both in height and width. Thus, the rate calculation would have to include an averaging of the potential (or perhaps the rate limiting step might be the barrier height fluctuation step itself).

12. Protein Folding and Interaction Processes with Asymmetrical Geometry

In discussions of diffusion-controlled processes in biological systems including protein folding models,[5, 7] domain dynamical models[73] and protein-ligand interactions,[74] it is generally assumed that the reactants, boundary conditions and initial conditions are symmetrical so that only one coordinate is required to describe the spatial behavior (quasi one-dimensional diffusion). This is clearly not the most realistic assumption under most circumstances and one that ought to be relaxed when better approximations to the real physical situation are contemplated. Nevertheless, almost no work appears to have been done in this regard on

biological problems of interest particularly for intramolecular diffusion which is relevant in protein folding models and protein domain movement models. Some general theoretical principles for biomolecular processes have been discussed by Sole and Stockmayer[75] and a particular process studied numerically by Samson and Deutch.[76] The basic procedure to follow is after a particular interaction model is formulated and translated into a diffusion equation with some asymmetrical elements, numerical solution at some point is required, the particular details depending on the particular physical situation.

References

1. Sela, M., White, F.H. and Anfinsen, C.B., Science 125, 691 (1961).

2. Baldwin, R.L., Ann. Rev. Biochem. 44, 453 (1975).

3. Levitt, M. and Warshel, A., Nature 253, 694 (1975).

4. Hagler, A.T. and Honig, B., Proc. Natl. Acad. Sci. (USA) 75, 554 (1978).

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

6. Baldwin, R.L., "Protein Folding" (ed. R. Jaenicke) Elsevier, Amsterdam, p. 369 (1980).

7. Karplus, M. and Weaver, D.L., Biopolymers 18, 1421 (1979).

8. See, for example, Chandrasekhar, S., Rev. Mod. Phys. 15, 1 (1945).

9. See, for example, Feldmann, R.J., "Atlas of Macromolecular Structure on Microfiche", Tracor Jitco Inc., Rockville, Md. (1976).

10. Von Smoluchowski, M., Ann. Phys. (Berlin) 48, 1103 (1915).

11. McCammon, J.A., Gelin, B.R., Wolynes, P.G. and Karplus, M., Nature 262, 325 (1976).

12. Weiss, G.H., Adv. Chem. Phys. 13, 1 (1967).

13. Weaver, D.L., Phys. Rev. B 20, 2558 (1979).

14. Szabo, A., Schulten, K. and Schulten, Z., J. Chem. Phys. 72, 4350 (1980).

15. Deutch, J.M., J. Chem. Phys. 73, 4700 (1980).

16. Weaver, D.L., J. Chem. Phys. 72, 3483 (1980).

17. See, for example, Ref. 14 and Weaver, D.L., unpublished results.

18. Perutz, M.F., J. Mol. Biol. 13, 646 (1965).

19. Ptitsyn, O.B. and Finkelshtein, A.F. Biofizika (USSR) 15, 757-767 (1970.

20. Chou, P. and Fasman, G.D. Biochem. 13, 212, 222 (1974).

21. Burgess, A.W., Ponnuswamy, P.K. and Scheraga, H.A. Israel J. Chem. 12 239 (1974).

22. Robson, B. and Suzuki, E., J. Molec. Biol. 107, 327 (1976).

23. Maxfield, F.R. and Scheraga, H.A., Biochemistry 15, 5138 (1976).

24. Nagano, K. J. Molec. Biol. 109, 251 (1977).

25. Kabat, E.A. and Wu, T.T. Biopolymers 12, 751 (1973).

26. Kotelchuck, D. and Scheraga, H.A. Proc. Natn. Acad. Sci. U.S.A. 61, 1163 (1968).

27. Lewis, P.N. Go, N., Go, M., Kotelchuck, D. and Scheraga, H.A. Proc. Natn. Acad. Sci. U.S.A. 65, 810 (1970).

28. Lim, V.I., J. Molec. Biol. 88, 857, 873 (1974).

29. Schulz, G.E. et al. Nature 250, 140 (1974).

30. Matthews, B.W. Biochem. Biophys. Acta. 405, 442 (1975).

31. Argos. P., Schwartz, J. and Schwartz, J. Biochem. Biophys. Acta 439, 261 (1976).

32. Kendrew, J.C., Dickerson, R.E., Standberg, B.E., Hart, R.G., Davies, D.R., Phillips, D.C. and Shore, V.C. Nature 185, 422 (1960).

33. Ptitsyn, O.B. and Rashin, A.A., Biophys. Chem. 3, 1 (1975).

35. See, for example, Ref. 3 and references contained therein.

36. Kuntz, I.D., Crippen, G.M., Kollman, P.A. and Kimelman, D., J. Mol. Biol. 106, 983 (1967).

37. Tanaka, S. and Scheraga, H.A., Proc. Natl. Acad. Sci. (USA) 74, 1320 (1977).

38. Flory, P.J., "Statistical Mechanics of Chain Molecules", Wiley, New York (1969).

39. McCammon, J.A., Northrup, S.H., Karplus, M., and Levy, R.M., Biopolymers.(t.b.p.)

40. Haas, E., Katchalski-Katzir, E. and Steinberg, I.Z., Biopolymers 17, 11 (1978).

41. Tsong, T.Y. and Baldwin, R.L., Biopolymers 17, 1669 (1978).

42. Tsong, T.Y., private communication.

43. See Ref. 11 and Karplus, M. in "Biomolecular Stereodynamics", Vol. II, p. 211 (1981).

44. Weaver, D.L., Physica 98A, 359 (1979).

45. MacElroy, R.D. and Weaver, D.L., unpublished results.

46. Bromer, W.W., Sinn, L.G. and Behrens, O.K., J. Am. Chem. Soc. 79, 2807 (1957).

47. Pohl, S.L., Birnbaumer, L. and Rodbell, M., Science 164, 566 (1969).

48. Gratzer, W.B. and Beaver, G.H., J. Biol. Chem. 244, 6675 (1969).

49. Panijpam, B. and Gratzer, W.B., Eur. J. Biochem. 45, 547 (1974).

50. Sasaki, K., Dockerill, S., Adamiak, D.A., Tickle, I.J. and Blundell, T., Nature 257, 751 (1975).

51. Warrant, R.W. and Kim, S.H., Nature 271, 130 (1978).

52. Adam, G. and Delbruck, M. in "Structural Chemistry and Molecular Biology" (eds. A. Rich and N. Davidson), Freeman, San Francisco (1968).

53. Weaver, D.L., Biophys. Chem. 10, 245 (1979).

54. Cherry, R.J., Biochem. Biophys. Acta 559, 289 (1979).

55. Kornberg, R.D. and McConnell, H.M., Proc. Natl. Acad. Sci. (USA) 68, 2564 (1971).

56. Edidin, M., Ann. Rev. Biophys. Bioeng. 3, 179 (1974).

57. Saffman, P.G. and Delbruck, M., Proc. Natl. Acad. Sci. (USA) 72, 3111 (1975).

58. Bruckner, P., Bachinger, H.P., Timpl, R. and Engel, J., Eur. J. Biochem. 90, 595 (1978).

59. Bachinger, H.P., Bruckner, P., Timpl, R. and Engel, J., Eur. J. Biochem. 90, 605 (1978).

60. See, for example, McCammon, J.A., Gelin, B.R. and Karplus, M., Nature 267, 585 (1977).

61. Gelin, B.R. and Karplus, M., Proc. Natl. Acad. Sci. (USA) 72, 2002 (1975).

62. McCammon, J.A., Wolynes, P.G. and Karplus, M., Biochemistry 18, 927 (1979).

63. See Refs. 11 and 43 and Karplus, M. and McCammon, J.A., CRC Crit. Rev. of Biochem. 9, 273 (1981).

64. McDonald, R.C., Steitz, T.A. and Engelman, D.M., Biochemistry 18 338 (1979).

65. Anderson, C.M., Zucker, F.H. and Steitz, T.A., Science 204, 375 (1979).

66. Richards, F.M., J. Mol. Biol. 82, 1 (1964).

67. Finney, J.L., J. Mol. Biol. 96, 721 (1975).

68. See, for example, Landau, L.D. and Lifshitz, E.M., "Statistical Physics", Pergamon Press, London (1958).

69. Cooper, A., Proc. Natl. Acad. Sci. (USA) 73, 2740 (1976).

70. See, for example,Englanter, S.W., Downer, N.W. and Teitelbaum, H., Ann. Rev. Biochem. 41, 903 (1972).

71. See, for example, Lakowicz, J.R. and Weber, G., Biochem. 12, 4171 (1973).

72. See, for example, Alben, J.O., et al., Phys. Rev. Lett. 44, 1157 (1980).

73. McCammon, J.A. and Karplus, M., Nature 268, 765 (1977).

74. See, for example, Cantor, C.R. and Schimmel, P.R., "Biophysical Chemistry" Part III, Freeman, San Francisco (1980).

75. Solc, K. and Stockmayer, W.H., J. Chem. Phys. 54, 2981 (1971).

76. Samson, R. and Deutch, J.M., J. Chem. Phys. 68, 285 (1978).

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
NASA CR- 166318		

4. Title and Subtitle	5. Report Date
Protein Folding, Protein Structure and the Origin of Life: Theoretical Methods and Solutions of Dynamical Problems	March 1982
	6. Performing Organization Code
7. Author(s)	8. Performing Organization Report No.
D. L. Weaver	
	10. Work Unit No. T5282
9. Performing Organization Name and Address	11. Contract or Grant No.
Department of Physics Tufts University Medford, MA 02155	P.O.#A 86784B (VAB)
	13. Type of Report and Period Covered
12. Sponsoring Agency Name and Address	Contractor Report
National Aeronautics and Space Administration Washington, D.C. 24056	14. Sponsoring Agency Code 199-60-62

15. Supplementary Notes

Robert D. MacElroy, Technical Monitor, Mail Stop 239-10, Ames Research Center, Moffett Field, CA 94035 (415) 965-5573 FTS 448-5573.

16. Abstract

This report considers theoretical methods and solutions of the dynamics of protein folding, protein aggregation, protein structure and the origin of life. The elements of a dynamic model representing the initial stages of protein folding are presented. The calculation and experimental determination of the model parameters are discussed. The use of computer simulation for modelling protein folding is considered.

17. Key Words (Suggested by Author(s))	18. Distribution Statement
Protein Aggregation Protein Folding Protein Structure Computer Simulation Modelling	Unclassified - Unlimited STAR Catergory 51

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price*
Unclassified	Unclassified	51	

*For sale by the National Technical Information Service, Springfield, Virginia 22161

End of Document