MECHANISM OF FOLDING AND DISULPHIDE BOND FORMATION IN GLOBULAR PROTEINS: A PROPOSAL FOR A UNIFIED THEORY

N. SESHAGIRI

National Information Centre, E-Wing, Pushpa Bhawan, Chiragdelhi – Madangir Road, New Delhi – 110 062, India

Structure of the Medium of Folding

Several hypotheses have so far been advocated in an effort to explain and simulate the folding of protein chains. On the assumption that these hypotheses are based on partial sets of experimental results representing features of a possible unified theory of protein folding, this paper attempts to provide a conceptual framework which may contribute to the evolution of such a unification.

Perhaps one of the least understood mechanisms is that involving the interaction of the chain with the surrounding aqueous medium. The average diameter of the water molecule of approximately 4 Å indicates that the number of atoms in the water molecules in the region of importance to the folding will be comparable to the number of atoms in the protein molecule. Simulation models should therefore include water molecules individually as far as possible, though statistical and thermodynamic analysis may guide such a simulation.

Frank and Wen [1] and Frank [2] postulated the existence of flickering clusters of water molecules which form hydrogen bonds in a co-operative manner. This idea was utilized by Nemethy and Scheraga [3, 4] in characterizing the medium of protein folding. However, the representation here is basically different in that we assume a continuum of water molecules with a variety of distorted hydrogen bonds in such a manner that the continuum responds to small local fluctuations of energy by varying the distortion in the bonds. A mathematical description of such a continuum is given in the following:

The total effective pair potential between two water molecules i and j with positional co-ordinates ($\mathbf{R}_i$, $\mathbf{R}_j$) is given in the usual form,

$$U(i,j) = U_{LJ}(R_{ij}) + U_{EL}(i,j) + U_{HB}(i,j), \tag{1}$$

where, $U_{LJ}(R_{ij})$ is the Lennard–Jones part ($R < 2$ Å), $U_{EL}(i,j)$ is the dipole–dipole part ($R > 5$ Å), and $U_{HB}(i,j)$ is the hydrogen bond part (2 Å $\leqslant R \leqslant$ 5 Å).

As a regular hydrogen bond occurs only along the four directions pointing to the vertices of a regular tetrahedron, four unit vectors, $i(L_\alpha)$, $i(H_\alpha)$, $j(L_\beta)$, and $j(H_\beta)$, are introduced in each molecule. Also, $G(x)$ refers to an unnormalized Gaussian function equal to $\exp(-x^2/2\sigma^2)$, where σ^2 is the variance. A bond will form whenever the direction of the O–H bond of one molecule lies along the direction $\mathbf{n} = \mathbf{R}_{ij}/R_{ij}$ and the direction O–L of the second molecule simultaneously lies along the direction $(-\mathbf{n})$. From the point of view of Monte Carlo

simulations of the class considered by Ben Naim [5], the configuration favouring H-bond formation for any pair (α, β) may be taken to satisfy,

$$\left.\begin{aligned} -\sigma' &< R_{ij} - R_H < \sigma', \\ -\sigma &< \mathbf{i}(A)\cdot\mathbf{n} - 1 < \sigma, \\ -\sigma &< \mathbf{j}(B)\cdot\mathbf{n} + 1 < \sigma. \end{aligned}\right\} \tag{2}$$

Here $\sigma = 2.85$ Å and σ' is the rate at which the energy drops at the distance $|R_{ij} - R_H|$, where R_H is the intermolecular distance at which maximum binding energy is obtained. (A, B) is either (H_α, L_β) or (L_α, H_β). A configuration of the pair (i, j) is hydrogen-bonded if and only if it satisfies conditions (2).

For simplicity of simulation, the system of m-sequential hydrogen bonds may be assumed to favour the presence of the OH...OH...OH and related chain structures. Assuming a co-operative bonding in the m-sequence it is possible to derive the following condition for its feasibility utilizing the results of Pople [6, 7], Donohue [8], and Goel *et al.* [9]:

$$\left.\begin{aligned} -m\sigma' f_1(G_m) &< \sum_{s=1}^{m} (R_{ijs} - R_{Hs}) < m\sigma' f_1(G_m), \\ -m\sigma f_2(G_m) &< \sum_{s=1}^{m} (\mathbf{i}_s(A_s)\cdot\mathbf{n}_s - 1) < m\sigma f_2(G_m), \\ -m\sigma f_2(G_m) &< \sum_{s=1}^{m} (\mathbf{j}_s(B_s)\cdot\mathbf{n}_s + 1) < m\sigma f_2(G_m), \end{aligned}\right\} \tag{3}$$

$$G_m = 1 + \frac{1}{mU}\sum_{s=1}^{m} k_{sm}(1 - \cos\phi_s\delta_s). \tag{4}$$

Here θ_s is the angle of bend between the respective OH and H...O. k_{sm} is the effective hydrogen bond bending force constant which depends on m, and the relative bending of the sth bond in the sequence relative to its neighbours taking into consideration non-additivity factors. $\delta_s = 1$ if $\phi_s > \phi_0$ and 0 if $\phi_s \leqslant \phi_0$. Though the calculations of Goel *et al.* [9] indicate that $\phi_0 \simeq 25°$, better results are obtained by assuming $\phi_0 < 15°$. k_{sm} may be more conveniently written as $k_0 k_{sm}$, where k_{sm} is a factor depending on the relative bending in the neighbourhood and k_0 $(= 3.78 \times 10^{-13}$ erg rad$^{-2})$ is the bond bending force constant. U is the minimum energy of an open dimer, $f_1(G_m)$ and $f_2(G_m)$ are functions of G_m which can be approximately expressed as $C_1 G_m$ and $C_2 G_m$ respectively, where C_1 and C_2 are semi-empirical constants. The single H-bond case, which is a modification of the conditions (2), is derived by letting $m = 1$ and $k_{sm} = k_0$.

Mixed Hydrophobic-Hydrophilic Cavity Effect

The possibility that the introduction of the protein chain in a water medium decreases the density in the neighbourhood and creates conditions favouring the formation of cavities is examined as a prelude to the quantification of hydrophobic and hydrophilic behaviour of side chains.

Based on the two structure model, the total differential of the volume V, with constant T and P, can be expressed as,

$$dV = V_s\,ds + V_L\,dl + V_H\,dh, \tag{5}$$

where

$$V_s = \left(\frac{\partial V}{\partial s}\right)_{l,h}, \quad V_L = \left(\frac{\partial V}{\partial l}\right)_{h,s}, \quad V_H = \left(\frac{\partial V}{\partial h}\right)_{s,l}. \tag{6}$$

Here, s, l, and h are the number of solute molecules, average number of L-cules of water and average number of H-cules of water respectively.

Imposing the condition $dl + dh = 0$, eqn. (5) can be rewritten as

$$dV = V_s\,ds + (V_L - V_H)\,dl. \tag{7}$$

As T, P, and the number of water molecules w ($s < w$) are constants and l is not an independent variable,

$$dV = V_s\,ds + (V_L - V_H)\left(\frac{\partial l}{\partial s}\right)_w ds. \tag{8}$$

The derivative is taken along the equilibrium line for the reaction $L \rightleftharpoons H$. With the addition of ds molecules to the system keeping l and h fixed, the measurable change in volume is given by the first term of eqn. (8). Releasing the constraint imposed by fixing l and h relaxes the system to a new equilibrium position with the measurable change in volume given by the second term of eqn. (8). Similar expression can be derived for enthalpy which gives the degreee of hydrogen bonding.

The change in volume in eqn. (8) propagates a pressure on the continuum whose magnitude can be expressed through

$$\left(\frac{\partial V}{\partial P}\right)_{T,w} = s\frac{\partial V_s}{\partial P} + l\frac{\partial V_L}{\partial P} + h\frac{\partial V_H}{\partial P} + (V_L - V_H)\left(\frac{\partial l}{\partial P}\right)_w. \tag{9}$$

Here it can be shown that along the equilibrium line for $L \rightleftharpoons H$

$$(V_L - V_H)\left(\frac{\partial l}{\partial P}\right)_{T,w} = \frac{-V(V_L - V_H)^2\, x_L x_H \eta}{kT}, \tag{10}$$

where x_L and x_H are the respective mole fractions and η is the volume density. The increased pressure is partly absorbed by bent m-sequences in the continuum in accordance with eqns. (3) and (4).

A measure of the hydrophobic driving force was made by Tanford [10], Brandts [11], and Epstein [12]. The hydrophobic residues have a tendency proportional to this hydrophobicity factor H_{fA}, for creating cavities in the continuum by virtue of their lack of affinity for water molecules. The work required to form a cavity in a hydrophobic environment can be quantified by known thermodynamic methods. The probability density of observing any specific configuration $\mathbf{R}^w$ is given in the usual form,

$$P(\mathbf{R}^w) = \frac{\exp[-\beta U_w(\mathbf{R}^w)]}{\int \ldots \int d\mathbf{R}^w \exp[-\beta U_w(\mathbf{R}^w)]}. \tag{11}$$

The probability of finding the centres of all particles in a region S is

$$P(S) = \int_S \ldots \int d\mathbf{R}^w P(\mathbf{R}^w). \tag{12}$$

If all the centres are to be excluded from a volume $V(\mathbf{R}_0, r)$ formed by a spherical cavity of radius r and centre at a fixed position $\mathbf{R}_0$,

$$P_{\text{cav}}(\mathbf{R}_0, r) = \int_{V-V(\mathbf{R}_0-r)} \ldots \int d\mathbf{R}^w P(\mathbf{R}^w) = \exp[-\beta W(\mathbf{R}_0, r)], \tag{13}$$

where $W(\mathbf{R}_0, r)$ is the work required for the creation of the cavity. Conversely, given W the radius r of the cavity can be obtained from eqn. (13).

In the folding process, if a tendency for cavity formation exists because of decrease in density of hydrophilic forces nearby, the volume of the cavity may be accentuated by $W \propto H_{fA}$ for the side chain of a given amino acid A. The volume contribution of this W to the overall cavity can be obtained by eqn. (13).

The foregoing derivations point to a reduction in the density of the medium around the protein chain as well as in the interior of the folds as compared to the bulk medium. As the number of atoms in the chain and those of the water molecules in the domain of folding are comparable, the volume of cavity will be appreciably high. As the hydrophilic side chains attract water molecules and the hydrophobic ones do not, the structure of the continuum in the interior, as the chain folds, is likely to be in the form of a molecular grid with the continuum spanning from one hydrophilic part to another, the connectivity depending upon the spatial distribution of the various hydrophilic and hydrophobic parts. At any stage of the folding we can determine the approximate cavity-to-continuum volume ratio in the following manner.

The increase of volume in the presence of the neighbourhood of a folding chain is determined by a non-linear interpolation between that for the linear chain and the fully folded globular form. For the linear chain the net increase in volume due to the addition of solute particles alone is given by integrating eqn. (8) over the entire volume of the linear chain. This is given as

$$V' = \int (V_L - V_H) \left(\frac{\partial l}{\partial s}\right)_w ds \tag{14}$$

computed over the entire volume V_c of the linear chain. It is further noted that the hydrophobic behaviour of the side chains will create a *tendency* for cavity formation enhancing the volume increase of eqn. (14). This tendency can be taken to be proportional to $P_{\text{cav}}(\mathbf{R}_0, r)$ of eqn. (13) integrated and normalized over the entire volume V_c. The work $W(\mathbf{R}_0, r)$ is taken to be a function $C_0 H_{hA}$ computed for each amino-acid residue $A = 1, 2, \ldots, N$ for creating a cavity of radius $r = 1$ and represented as $W(\mathbf{R}_A, 1)$. Here C_0 is an undetermined constant which is evaluated from trial simulations using known protein structures of small chain lengths. The net increase of volume is given by

$$V_1 = C_1 \int_{(V_c, A)} \left[(1 + \exp(-\beta C_0 H_{hA}) (V_L - V_H) \left(\frac{\partial l}{\partial s}\right)_w\right] ds. \tag{15}$$

Here the integration is carried out over all the solute molecules representing the chain taking care to see that H_{hA} refers to different A, corresponding to different sets of solute molecules. C_1 is a proportionality constant which is determined similar to C_0.

The fundamental notion behind the continuum is that the average density of the hydrogen-bonded water molecules is the same whether it is the bulk water medium or the vicinity of the chain. The increase in volume V_1 will, therefore, be possible only if we assume the existence of cavities in the interior and the exterior neighbourhood of the chain. In view of this, V_1 itself will be the volume of the cavity in the neighbourhood of the chain.

The volume of the cavity $V_{(f)}$ in the neighbourhood of a folding chain in the intermediate stage is approximately interpolated as

$$V_f = C_2 \frac{N_{(f)}}{N_1} V_1. \tag{16}$$

Here N_1 and $N_{(f)}$ are the minimal number of water molecules in the vicinity of the chain that is of relevance to folding for the linear chain and the intermediate folded chain respectively.

The distribution of the cavity volume within the domain of interest to folding is taken as a non-linear function increasing towards the instantaneous centre of mass of the intermediate folded chain. For purposes of the present simulation model, we take

$$V(r) = C_4 + \frac{C_5}{(r+C_3)} + \frac{C_6}{(r+C_3)^2}, \tag{17}$$

$$\iiint_V V(r)\, dV = V_{(f)}, \tag{18}$$

where V is the folding domain and C_3 to C_6 are constants. The folding domain can itsel be divided into an interior domain and an exterior domain for purposes of simulation. The interior domain is taken as the three-dimensional envelope of the straight lines drawn from each residue to the others. For more accurate simulations, it is necessary to define a kernel of the backbone. The kernel for the backbone for a linear chain is approximately the straight line joining the NH_2 and COOH terminii and that for the fully folded protein is the centre of mass. An interpolation between the two can be made for intermediate stages using well-known geometrical methods for obtaining the intermediate kernels. The integration in eqn. (18) is worked with respect to the nearest point on the kernel to the point under consideration.

The determination of the structure of the grid along with path connectivities and path cross-sections require a knowledge of the manner in which polar residues will affect the grid paths.

The simplest approximation assumes that each hydrophilic side chain can be represented in the domain of the intermediate folding stage by a set of positive and negative charges bearing the same polarity index as given by Epstein [12] and drawing the flux lines by well-known electrostatic methods. For more accurate computations, methods based on the well-known Hellmann–Feynman theorem can be utilized. The continuum is accommodated around the kernels of the flux field joining the hydrophilic side chains. The cross-section of the path is made proportional to the flux density along the kernels. A better approximation

recognizes the ionic, dipole factors present in the water medium and draws upon certain techniques normally employed in the thermodynamics of ionic hydration as well as calculations of the above type.

With the determination of path connectivities and path cross-sections in the above form, we can distribute the bonded water molecules along the grid paths keeping the deviation of cavity volume distribution from that given eqns. (17) and (18) as minimum, i.e.

$$\iiint_V [V(\mathbf{R}) - \bar{V}(\mathbf{R})]^2 \, d\mathbf{R} \rightarrow \text{minimum} \tag{19}$$

subject to

$$\iiint_V V(\mathbf{R}) \, d\mathbf{R} = \iiint_V \bar{V}(\mathbf{R}) \, d\mathbf{R} = V_{(f)}. \tag{20}$$

With the grid structure postulated above it is possible to associate various means of energy transfer like electron migration in conductivity bands, transfer by proton, hydride ion or hydrogen radical migration, transport by orientational defects, etc. A survey of possible mechanisms of energy transfer has been made by Steinberg [13] and von Hippel [14]. In this background we postulate the existence of a vector current (energy carrier) through the grid paths.

We define a current path between two fixed-charge locations on side chains as paths connecting them along the grid path with constant scalar currents, each scalar representing a distinct mode of energy transport. Along a grid path between two nodes of the grid two or more current paths may coalesce in the continuum giving a net current along the segment. Without going into the proof, it is stated here that the current along a grid path is approximately a superposition of the currents along the (constant) current paths. From this the net current at any point in the grid can be determined.

The Grid-path Simulation of Folding

A broad outline of the simulation is presented below:

(1) As the tendency of folding is such that the cavity volume in the domain decreases from the linear chain to the globular chain, the folding pathway is such that this decreasing trend is maintained. Further, the polar residues are the starting or terminal points for the currents both to the interior and exterior of the domain.

(2) Apart from the grid-path concept proposed here, all other aspects concerning the steric forces and constraints are considered using established methods.

(3) A set of special folding tendency functions have been defined for α-helices, β-bends, etc., which are included in the simulation as indicators of sensitivity of folding.

(4) In the present routine, certain approximate gross representation of the features of vector currents with orientational features found between the cysteines acting through the grid-path medium have been assumed. This simulates the tendency for disulphide bond formation.

(5) A feedback mechanism based on the following steps has been incorporated:

   (a) The grid-path actuated mechanism mentioned in (1) and (4) above introduces mechanical forces of different magnitudes with a tendency to place different side chains towards or away from the instantaneous grid kernel.

(b) The mechanical forces contribute to the folding of the chain in addition to the forces mentioned in (2) and (3) above.
(c) The folding reduces the cavity volume as well as squeeze out a determinate number of water molecules out of the domain with associated redistribution of the cavity volume.
(d) The grid paths reorient themselves to accommodate the new situation described in (c). Also, the vector currents over the current path are altered.
(e) Step (a) is initiated again to begin the next iteration.

Conclusion

Concepts like water-molecular grid path and vector current have been advocated in the proposal for evolving a theory of protein folding for unifying the thermodynamic and nucleation pathway dependent theories. The simulation experiments, which are at present in a preliminary stage, have been carried out with β_2-microglobulin and BPTI with a moderate degree of success. Until the simulation is carried out with bigger and more complex chains using more accurate derivations, it is too early to compare the grid-path theory with others. Such an elaborate simulation package is currently under development.

Summary

Introducing the concept of "water grid pathways", a theory is advocated for the mechanism of folding and the formation of disulphide bonds in the tertiary structure of globular proteins. The proposal is intended to give a framework for unifying the thermodynamic and kinetic approaches.

References

1. H. S. Frank and W. Y. Wen, *Disc. Faraday Soc.* **24,** 133 (1957).
2. H. S. Frank, *Proc. Roy. Soc. Lond.* A, **247,** 481 (1958).
3. G. Nemethy and H. A. Scheraga, *J. Chem. Phys.* **36,** 3382, 3401 (1962).
4. G. Nemethy and H. A. Scheraga, *J. Phys. Chem.* **66,** 1773 (1962).
5. A. Ben-Naim, *J. Chem. Phys.* **54,** 3382 (1971).
6. A. Pople, *Proc. Roy. Soc.* A, **202,** 323 (1950).
7. A. Pople, *Proc. Roy. Soc.* A, **205,** 163 (1951).
8. J. Donohue in *Structural Chemistry and Molecular Biology* (A. Rich and N. Davidson, eds.), Freeman, San Franscisco, 1968.
9. A. Goel, A. S. W. Murthy, and C. N. R. Rao, *J. Chem. Soc. Lond.* A, **1971,** 190 (1971).
10. C. Tanford, *Am. Chem. Soc.* **84,** 4240 (1962).
11. J. Brandts, *Biological Macromolecules*, Marcel Dekker, New York, 1965.
12. C. J. Epstein, *Nature* **215,** 355 (1967).
13. I. Z. Steinberg, *Ann. Rev. Microbiology* **747,** 83 (1971).
14. A. von Hippel, *J. Chem. Phys.* **54,** 145 (1971).