Self-Consistent Approaches for the Protein Folding Problem: Kinetics and Equilibrium Properties of Random Copolymers

Edward G. TIMOSHENKO, *) Yuri A. KUZNETSOV, Andrei MOSKALENKO and Kenneth A. DAWSON

Theory and Computation Group, Centre for Colloid Science and Biomaterials Department of Chemistry, University College Dublin, Dublin 4, Ireland

We address the problem of protein folding by studying a model of random copolymers with a Gaussian distribution of quenched disorder in the monomers' hydrophobicity. We develop the Gaussian self-consistent method in kinetics with the effective potentials depending on the disorder variables, which are further integrated out perturbatively. As an alternative, at equilibrium, we also suggest a version of a variational approach in the replica space. These methods allow us to achieve a unified description of the extended coil, liquid-like, frozen and folded globular states, as well as of the kinetic transformations between them. We discuss the role of the finite-size effects and the optimization of the disorder distribution for improving the folding properties of random sequences. We believe that the final resolution of these issues may shed light on the protein folding puzzle.

§1. Introduction

Elucidation of the conformational states of simple models of proteins, along with their attendant kinetic laws, is a crucial step in understanding protein folding (see Refs. 1) - 6) and references therein) and misfolding. One expression of the protein folding problem is to determine how a one-dimensional primary sequence of amino acid residues relates to the three-dimensional structure of the folded protein, and to deduce the kinetics of folding process that brings a statistical ensemble of extended coils into an essentially unique native state. Renewed interest in the misfolding of proteins has arisen recently as a result of evidence that proteins aggregate resulting, possibly, in deceases such scrapie, Alzheimers, BSE, CJD and some others. [7)]

§2. The self-consistent method for kinetics

In this section we shall briefly define the model and write down the self-consistent equations derived in Ref. 8). We also introduce necessary notations and define the important observables. For further explanations and details we refer the reader to Ref. 8).

It is convenient to use the Fourier transforms $\boldsymbol{x}_q$ of the monomer coordinates $\boldsymbol{X}_m$ in the chain index. Note that we use letters of different case in order to distinguish between the two sets of coordinates. Description of kinetics of the model is based upon the Langevin equation,

$$\zeta \frac{d}{dt}\boldsymbol{x}_q(t) = -\frac{\partial H}{\partial \boldsymbol{x}_{-q}} + \boldsymbol{\eta}_q(t),$$

*) E-mail: timosh@fiachra.ucd.ie

$$\langle \eta_q^{\alpha}(t)\eta_{q'}^{\alpha'}(t')\rangle = 2k_BT\zeta\,\delta_{q+q',0}\,\delta^{\alpha\alpha'}\delta(t-t'), \tag{2·1}$$

where $\zeta = N\zeta_b$, ζ_b is the bare friction constant and N is the degree of polymerization.

The model accounts for the connectivity of the chain, excluded volume effects and the random amphiphilicity of the monomers. We choose the Edwards-type effective free energy functional, $H = \bar{H} + H_{\rm dis}$, containing the homopolymeric, $\bar{H}$, and the disordered, $H_{\rm dis}$, terms respectively,

$$\bar{H} = \frac{\kappa}{2}\sum_n (\boldsymbol{X}_{n+1} - \boldsymbol{X}_n)^2 + \sum_{L>2} \bar{u}_L \sum_{\{m\}} \prod_{i=1}^{L-1} \delta(\boldsymbol{X}_{m_i} - \boldsymbol{X}_{m_{i+1}}), \tag{2·2}$$

$$H_{\rm dis} = \frac{1}{2}\sum_{m_1 m_2} (\Lambda_{m_1} + \Lambda_{m_2})\,\delta(\boldsymbol{X}_{m_1} - \boldsymbol{X}_{m_2}). \tag{2·3}$$

Here κ is the connectivity constant, $\bar{u}_L$ are the virial coefficients of the excluded volume interactions. The random variables Λ_m, and consequently their Fourier transforms λ_q, are assumed to possess a Gaussian distribution with the second momentum,

$$\overline{\lambda_q\lambda_{q'}} = \bar{\Delta}^2\,\delta_{q+q',0}, \quad \bar{\Delta}^2 \equiv \Delta^2/N, \tag{2·4}$$

where Δ has the meaning of the dispersion of disorder. Here and throughout we use the brackets $\langle A\rangle$ to denote the statistical averages over the noise and initial ensemble of monomer positions $\{\boldsymbol{x}(t=0)\}$ and the bar $\overline{A}$ to denote averages over the quenched distribution of disorder $\{\Lambda\}$.

In our previous work in Ref. 8) we have derived in some approximation closed kinetic equations for the two types of correlation functions. Thus, let us introduce the mean squared amplitudes of the normal modes,

$$\mathcal{F}_q(t) \equiv \overline{F_q(t)}, \quad F_q(t) = \frac{1}{3}\langle |\boldsymbol{x}_q|^2(t)\rangle, \tag{2·5}$$

and the disorder correlation functions,

$$\varphi_{qp}(t) \equiv \overline{\phi_{qp}(t)}, \quad \phi_{qp}(t) = \frac{1}{3}\lambda_{q-p}\langle \boldsymbol{x}_{-q}(t)\boldsymbol{x}_p(t)\rangle. \tag{2·6}$$

These satisfy the following self-consistent equations,

$$\frac{\zeta}{2}\frac{d}{dt}\mathcal{F}_q(t) = -\frac{2}{3}\left(\mathcal{F}_q\frac{\partial\mathcal{A}}{\partial\mathcal{F}_q} + \sum_p \varphi_{qp}\frac{\partial\mathcal{A}}{\partial\varphi_{qp}}\right), \tag{2·7}$$

$$\zeta\frac{d}{dt}\varphi_{qp}(t) = -\frac{2}{3}\left(\varphi_{qp}\left(\frac{\partial\mathcal{A}}{\partial\mathcal{F}_q} + \frac{\partial\mathcal{A}}{\partial\mathcal{F}_p}\right) + \bar{\Delta}^2\left(\mathcal{F}_q + \mathcal{F}_p\right)\frac{\partial\mathcal{A}}{\partial\varphi_{qp}}\right), \tag{2·8}$$

where $\mathcal{A}$ is the variational free energy functional given in Ref. 8). This form of the kinetics could be understood as a motion representing the flow of the whole statistical ensemble in the phase space of the averaged dynamic variables.

To understand the phase behaviour of the system we need to identify the important order parameters. First, the overall size of the chain is given by the average

$\overline{R_g^2}$ of the squared radius of gyration, $R_g^2 = \sum_{q\neq 0} F_q$. Second, the glassy behaviour is reflected in the cumulant of the squared radii of gyration,[9),10)]

$$\overline{R_g^2 R_g^2}^{(c)} = \bar{\Delta}^{-2}\Upsilon^2, \quad \Upsilon = \sum_{q\neq 0} \varphi_{qq}. \tag{2·9}$$

And third, the phase separation order parameter is

$$\Psi = \frac{1}{6N^2} \sum_{mm'} \overline{(\Lambda_m + \Lambda_{m'} - 2\lambda_0) D_{mm'}} = \sum_{q\neq p,\ q,p\neq 0} \varphi_{qp}. \tag{2·10}$$

§3. Phase diagram of the model

Traditionally[11)-13)] we work with the following combinations $\mathcal{L} = (k_B T/\kappa)^{1/2}$ and $\mathcal{T} = \zeta_b/\kappa$ as the units of size and time in the system. In the sequel we have used the following particular choice of parameters: $k_B T = 1$, $\kappa = 1$ and $\zeta_b = 1$, which fix $\mathcal{L}$ and $\mathcal{T}$ to be equal to unity.

The phase diagram[22),19)] is presented in Fig. 1. Phase (I) corresponds to the extended Flory coil. This normal homopolymer-like coil with increasing dispersion of disorder becomes what we call a random coil (III) after a rather soft continuous transition. On passing this transition the squared radius of gyration, $\overline{R_g^2}$, decreases somewhat while the "glassy" order parameter, $\overline{R_g^2 R_g^2}^{(c)}$, increases significantly. We believe that phase (III) is composed of relatively open coils with numerous loops.

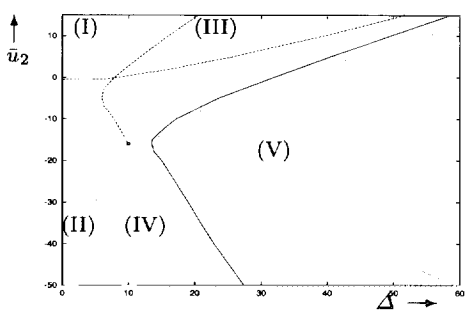

Fig. 1. Sketch of the phase diagram of the model in terms of the second virial coefficient, $\bar{u}_2$ (in units $k_B T\mathcal{L}^3$), and the dispersion of disorder, Δ ($k_B T\mathcal{L}^3$). Solid lines represent first order-like transitions, dashed lines — continuous transitions, and dotted lines — "spinodal" curves. The Roman numerals correspond consequently to: Flory coil, liquid-like globule, random coil, "glassy" phase and folded globule. Here and below $N = 30$, and $\bar{u}_3 = 10$ (in units $k_B T\mathcal{L}^6$).

The collapse transition (the curve separating I and III from all other phases) is second order in the whole range of Δ. Beginning from the homopolymer-like coil (I) for small dispersions of disorder it leads to a homopolymer-like globule (II). That phase is often referred to as a liquid-like globule in the literature because the connectivity constraints there are manifested only at short distances along the chain.

In the next series of figures we draw the behaviour of the conformational entropy (Fig. 2(a)), the squared radius of gyration (Fig. 2(b)), and the glass order parameter (Fig. 2(c)) vs the dispersion of disorder Δ for three different values of the second virial coefficient.

With increasing Δ the globule (II) undergoes a freezing transition. As evident from Fig. 2 the frozen globule (IV) has a smaller entropy, a larger size and a

pronounced glass order parameter. The freezing transition is continuous above the tricritical point and becomes first-order-like below it. In Fig. 1 we show the spinodal curve beyond which the homopolymer-like globule solution ceases to exist. The freezing transition curve will be determined in §4 using the replica formalism.

According to Fig. 2 the folded state (V) is characterized by very low entropy, compact size, small glass order parameter and optimal phase separation of hydrophobic/hydrophilic units (see Figs. 2, 7 in Ref. 8)). We believe that the frozen globule is akin to the *molten* globule and the folded globule to the *native* state in proteins respectively. This conjecture is somewhat justified also by the kinetics of the folding process discussed in Ref. 8).

An interesting observation here is that there is a pronounced region of the metastable frozen globule. The kinetic evolution after a rapid quench to the region bounded by the transition and spinodal curves will remain trapped in a frozen misfolded state for a long time related to the barrier height. This is in agreement with observation in numerous Monte Carlo simulations [14)-18)] that there is poor kinetic accessibility of the folded conformation for a generic class of sequences.

The situation deteriorates dramatically with increasing degree of polymerization. Thus, for sufficiently long chains the spinodal curve cannot be reached until the dispersion of disorder value Δ_{sp}, which tends to infinity exponentially quickly. This means that in our model, having all possible sequences of monomers characterized by a Gaussian distribution, the kinetic accessibility is very poor indeed! The way to make folding more efficient is to optimize the distribution by restricting acceptable sequences to a more narrow subclass of sequences possessing good folding properties.

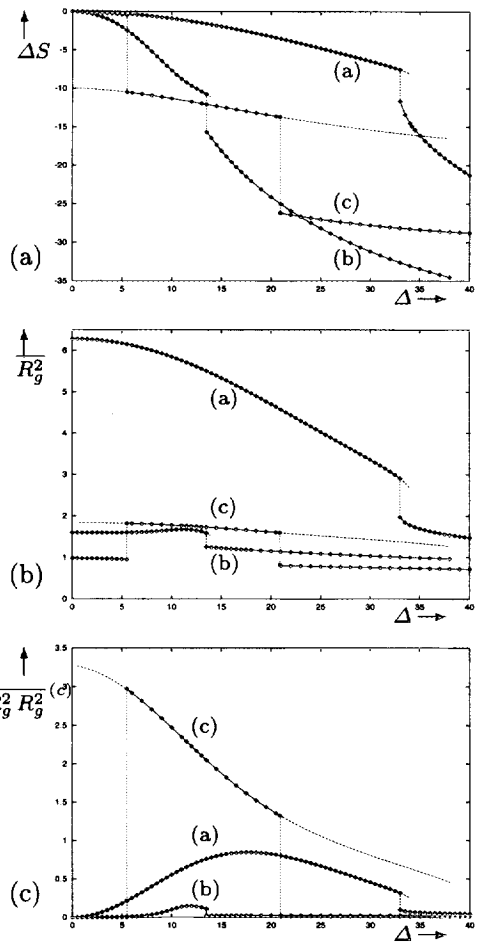

Fig. 2. Plots of observables vs the dispersion of disorder Δ (in units $k_BT\mathcal{L}^3$) for different values of $\bar{u}_2$ (in units $k_BT\mathcal{L}^3$): (a) $\bar{u}_2 = 0$; (b) $\bar{u}_2 = -16$; and (c) $\bar{u}_2 = -35$. In Fig. 2(a) we present the conformational entropy change $\Delta S \equiv S_c(\Delta) - S_c(0)$ (in units k_BT); In Fig. 2(b) — the mean squared radius of gyration $\overline{R_g^2}$ (in units $\mathcal{L}^2$); and in Fig. 2(c) — the glass order parameter $\overline{R_g^2 R_g^2}^{(c)}$ (in units $\mathcal{L}^4$).

An intriguing point however is that, although the kinetic accessibility is impeded for long polymers, the folded state still remains the main free energy minimum for systems with quite small dispersions of disorder. The more nontrivial issue of whether this remains true in the thermodynamic limit ($N \to \infty$) in answered in the affirmative in Ref. 22).

§4. Replica variational approach

We derive the effective free energy using the one-step Parisi scheme [9] and analyse it numerically, [23]

$$\begin{aligned}\lim_{n\to 0} \beta \mathcal{A}_{\rm eff}/n = &-\frac{3}{2}\sum_q \left[\frac{1}{x}\log\left(\tilde{\mathcal{F}}_q - (1-x)\mathcal{F}_q^{(1)} - x\mathcal{F}_q^{(0)}\right)\right] \\ &-\frac{3}{2}\sum_q \left[\frac{x-1}{x}\log\left(\tilde{\mathcal{F}}_q - \mathcal{F}_q^{(1)}\right) + \frac{\mathcal{F}_q^{(0)}}{\tilde{\mathcal{F}}_q - (1-x)\mathcal{F}_q^{(1)} - x\mathcal{F}_q^{(0)}}\right] \\ &+\frac{3\kappa N}{2}\tilde{D}_{01} + \hat{u}_2 \sum_{ij}\left(\tilde{D}_{ij}\right)^{-3/2} + (\hat{u}_3 - \hat{\Delta}^2)\sum_{ijk}\tilde{\Pi}_{ijk}^{-3/2} \\ &+\hat{\Delta}^2 \sum_{ijk}\left((1-x)\Pi_{ijk}^{(1)-3/2} + x\Pi_{ijk}^{(0)-3/2}\right) + \hat{u}_4 \sum_{i_0 i_1 i_2 i_3} \tilde{\Pi}_{i_0 i_1 i_2 i_3}^{-3/2},\end{aligned} \tag{4·1}$$

where $\hat{u}_L \equiv (2\pi)^{-3(L-1)/2}\bar{u}_L$ and $\hat{\Delta} \equiv (2\pi)^{-3/2}\Delta$. Here we introduced the correlation functions of coordinates in Fourier and real space for monomers in the same replica, $\{\tilde{\mathcal{F}}_q, \tilde{D}_{ij}\}$, the same group of replicas, $\{\mathcal{F}_q^{(1)}, D_{ij}^{(1)}\}$ and from different groups of replicas $\{\mathcal{F}_q^{(0)}, D_{ij}^{(0)}\}$, and

$$\tilde{\Pi}_{ijk} = \tilde{D}_{ij}\tilde{D}_{jk} - \left(\tilde{D}_{ijk}\right)^2, \quad \Pi_{ijk}^{(0,1)} = \tilde{D}_{ij}\tilde{D}_{jk} - \left(D_{ijk}^{(0,1)}\right)^2, \tag{4·2}$$

$$\tilde{\Pi}_{i_0 i_1 i_2 i_3} = \det{}_{3\times 3}\, \tilde{D}_{l:k}, \quad \tilde{D}_{l:k} = \tilde{D}_{i_l i_0 i_k}, \quad l,k = 1,2,3. \tag{4·3}$$

Note that the correlation functions of the Fourier modes are related to the correlations of the monomer spatial positions by formulas $\mathcal{D}_{mm'} \equiv \mathcal{D}_{mm'm}$:

$$D_{mm'm''} \equiv \frac{1}{3}\langle (\boldsymbol{X}_m - \boldsymbol{X}_{m'})(\boldsymbol{X}_{m''} - \boldsymbol{X}_{m'})\rangle = \sum_q d_{mm'm''}^{(q)} F_q, \tag{4·4}$$

$$d_{mm'm''}^{(q)} = \frac{1}{2}\left(d_{mm'}^{(q)} + d_{m''m'}^{(q)} - d_{mm''}^{(q)}\right),$$

$$d_{mm'}^{(q)} = 2\left(1 - \cos\frac{2\pi q(m-m')}{N}\right). \tag{4·5}$$

The result of the numerical analysis of the effective free energy shows that the collapse can occur in two different ways, see Fig. 3. For small values of the dispersion of disorder $\hat{\Delta}$, the coil collapses to the liquid-like globule as the effective two-body

interaction parameter $\hat{u}_2$ becomes approximately negative. For large $\hat{\Delta}$ the effective three-body interaction changes sign from negative to positive and collapse occurs due to the three-body effects, whilst $\hat{u}_2$ remains positive, but small. There is also a transition from liquid-like globule to the frozen globule.

Above the freezing temperature, the monomer coordinates correlation functions closely coincide with those for the annealed disorder. We find that at the freezing transition line scales approximately as, $\hat{\Delta}_f \sim |\hat{u}_2|^{-\gamma}$, $1 \leq \gamma^{-1} \leq 2$, for large negative $\hat{u}_2$, and scales linearly in $\hat{u}_3$. This scaling is essentially independent of the size of the chain for sufficiently large degree of polymerization N.

§5. Conclusion

We proposed self-consistent methods for kinetics and equilibrium of a Gaussian random amphiphilic copolymer. One of these methods is based on the explicit averaging of the disorder dependent kinetic self-consistent equations. The second one is a version of a variational approach based on the use of the replica trick and Parisi scheme.

An interesting conclusion of our consideration is that the Gaussian distribution of sequences is too wide to possess any good folding properties in the average. Despite the thermodynamic stability of the folded state for long chains, its kinetic accessibility is very impeded. Thus, selection and design of good folding sequences are among the main issues in the fundamental problem of protein folding.

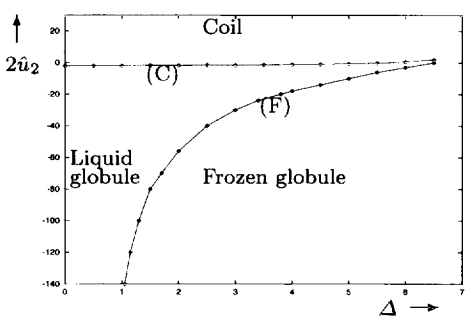

Fig. 3. Part of the phase diagram of a random copolymer obtained using the replica variational approach in variables of the dispersion, $\hat{\Delta}$ (in units $k_BT\mathcal{L}^3$), and mean value, $\hat{u}_2$ (in units $k_BT\mathcal{L}^3$), of the second virial coeficient. Lines (C) and (F) correspond to the collapse and the freezing transitions respectively. Here $N = 20$, the third and the fourth virial coefficients are equal to $\hat{u}_3 = 10$, $\hat{u}_4 = 2$. No substantial qualitative changes are observed for larger N.

References

1) H. Frauenfelder, in *Structure and Dynamics of Nucleic Acids, Proteins and Membranes*, ed. E. Clementi and S. Chin (Plenum, 1986).
2) P. G. Wolynes, in *Spin Glass Ideas in Biology*, ed. D. Stein (World Scientific, Singapore, 1991).
3) T. E. Creighton (ed.), *Protein Folding* (Wiley, New York, 1992).
4) R. Elber (ed.), *New Developments in Theoretical Studies of Proteins* (World Scientific, Singapore, 1994).
5) F. M. Richards, Scientific American (January 1991), p.34.
6) A. Sali, E. Shakhnovich and M. Karplus, Nature **369** (1994), 248.
7) G. Taubes, Science **271** (1996), 1493.
R. A. Bessen et al., Nature **375** (1995), 698.

8) E. G. Timoshenko, Yu. A. Kuznetsov and K. A. Dawson, Phys. Rev. **E54** (1996), 4071.
9) M. Mezard, G. Parisi and M. Virasoro, *Spin Glass Theory and Beyond* (World Scientific, Singapore, 1987).
10) M. Mezard and G. Parisi, J. de Phys. **I1** (1991), 809.
11) E. G. Timoshenko, Yu. A. Kuznetsov and K. A. Dawson, J. Chem. Phys. **102** (1995), 1816.
12) Yu. A. Kuznetsov, E. G. Timoshenko and K. A. Dawson, J. Chem. Phys. **104** (1996), 3338.
13) E. G. Timoshenko, Yu. A. Kuznetsov and K. A. Dawson, Phys. Rev. **E53** (1996), 3886.
14) P. G. Wolynes, J. N. Onuchic and D. Thirumalai, Science **267** (1995), 1619.
15) K. A. Dill, K. M. Fiebig and H. S. Chan, Proc. Natl. Acad. Sci. USA **90** (1993), 1942.
16) C. J. Camacho and D. Thirumalai, Proc. Natl. Acad. Sci. USA **90** (1995), 1277.
17) J. D. Honeycutt and D. Thirumalai, Biopolymers **32** (1992), 695.
18) N. D. Socci and J. N. Onuchic, J. Chem. Phys. **103** (1995), 4732.
19) E. G. Timoshenko, Yu. A. Kuznetsov and K. A. Dawson, J. Stat. Phys. (1997), to be published.
20) V. S. Pande, A. Yu. Grosberg and T. Tanaka, Phys. Rev. **E51** (1995), 3381.
21) C. D. Sfatos, A. M. Gutin and E. I. Shakhnovich, Phys. Rev. **E48** (1993), 465.
22) E. G. Timoshenko, Yu. A. Kuznetsov and K. A. Dawson, Phys. Rev. **E55** (1997), to be published.
23) A. Moskalenko, Yu. A. Kuznetsov and K. A. Dawson, J. de Phys. **II7** (1997), 409.