Simple Models of Biological Macromolecules

E.G. Timoshenko[1], Yu.A. Kuznetsov, K.A. Dawson
Department of Chemistry, University College Dublin, Belfield, Dublin 4,
Ireland

Abstract

We present a brief overview of simple statistical mechanical models of biopolymers such as proteins and nucleic acids. These are studied by the Gaussian self–consistent method and direct Monte Carlo simulations. The equilibrium phase diagrams and kinetics of conformational transitions between different states are elucidated. Amphiphilic copolymers and the persistent homopolymer considered here exhibit a remarkable variety of conformational states with a rather complex multistage kinetics of folding from the extended coil into compact globular states.

1 Introduction

Complexity and huge size of biological macromolecules [1] makes their modelling and analysis an exteremely hard if not an impossible task. The helix of DNA which is closed packed into a cell nucleus of only about 10^3 nm size being unwound into a straight line would be about a meter long! This basic molecule of life contains all of the relevant genetic information about an individual of a species. Due to rather complex interactions between its atoms and atoms of other molecules in the chromatin complex DNA forms a very compact native structure which permits further very efficient unpacking and replication [1]. Changes of atom positions (*the conformation*) of a macromolecule that are accompanied by a significant change in its size and the degree of packing (*the fractal dimension*) known by the name of *conformational transitions* present some of the most important mechanisms of how a living cell functions.

Another example of conformational transitions is given by the folding of a nascent protein molecule. Proteins [2] perform most of working functions in a cell such as chemical (catalysis of chemical reactions by enzymes), mechanical (constriction of muscles), electric (transmission of nerve impulses in neurons), structural (robustness and tissue forming) and many others [1, 2]. Despite their variety and versatility, all proteins are build from only 20 types of building blocks (amino acids). A genetic information for production of a specific protein is stored in the mesenger RNA obtained from the original DNA. Inside a ribosome it is translated into a linear chain (*primary sequence*) of these 20 monomer types according to simple mapping rules. The outgoing protein chain then undergoes a folding process to produce eventually a compact globular *tetrary structure* which is biologically functional. This native state of a protein is characterised by essentially unique relative positions of monomers in the globule entirely predetermined by its primary sequence. Such state by necessity corresponds to the pronounced energy minimum. How exactly the protein is capable to find this state among an astronomically large number of all possible conformations in a rather short time is a mystery known as the Levinthal paradox.

We should also emphasise that one is interested here in the time dependence of tranformations from one equilbrium structure to another, i.e. in kinetics of conformational transitions. However, in fact, most of structures present in a cell are not at true equilbrium but rather are metastable states which could only be produced by a specific kinetic pathway. The non–equilibrium behaviour plays a crucial role in biology and results in an intiricate variety of states which would be otherwise impossible.

Because biopolymers are typically quite large one can use the classical statistical mechanics for their description with the mean–force inter–atomic potentials deduced from quantum chemistry. The computational expenses in direct simulation methods such as Molecular Dynamics are huge for even a few base pairs of DNA on the best modern supercomputers even for reaching a nanosecond timescale. Unfortunately, most of conformational changes take microseconds and longer, not mentioning the problems of maintaining a required temperature and keeping many molecules of the solvent in Molecular Dynamics.

Therefore, it seems important to develop coarse–grained methods [3] of non–equilibrium statistical mechanics that could reliably describe long timescales with a lower spatial resolution, but maintaining the main conformational properties of the macromolecule of interest. This desciption can be deduced from the microscopical theory by introducing larger groups of monomers as new monomers — something

[1]Corresponding author. E-mail: timosh@fiachra.ucd.ie

very similar to the renormalisation procedure proposed by Kadanoff for the Ising model. Having done such a procedure once, one may look at the coarse–grained system more phenomenologically by including only the main relevant interactions and determining the parameters of the model from experimentally measurable properties instead of trying to relate them to the parameters of the microscopic theory.

2 The Gaussian self–consistent method

The main variables in the coarse–grained description of the polymer chain [3] are the spatial monomer coordinates $\mathbf{X}_n$, where n is the monomer number. The solvent molecules are excluded from the consideration by integrating out their degrees of freedom from the path integral representation for the partition function. The resulting monomer interactions are represented by the effective free energy functional,

$$H = \frac{k_B T}{2l^2} \sum_n (\mathbf{X}_n - \mathbf{X}_{n-1})^2 + \sum_{J=2}^{\infty} \sum_{\{n\}} u_{\{n\}}^{(J)} \prod_{i=1}^{J-1} \delta(\mathbf{X}_{n_{i+1}} - \mathbf{X}_{n_1}). \tag{1}$$

The first term in Eq. (1) describes the connectivity of the chain with l called the statistical segment length. There are also volume interactions represented by the virial–type expansion in Eq. (1). They reflect the hard–core repulsion, the weak attraction between monomers and the effective interaction mediated by the solvent–monomer couplings. The virial coefficients, $u_{\{n\}}^{(J)}$, may in principle have any dependence on the site indices $\{n\} \equiv \{n_1, \ldots, n_J\}$. Here we consider the choice of site–dependent second virial coefficients in Eq. (1) with monomers differing only in the monomer–solvent coupling constants,

$$u_{nn'}^{(2)} = \bar{u}^{(2)} + \Delta \frac{\sigma_n + \sigma_{n'}}{2}, \qquad u_{\{n\}}^{(J)} = u^{(J)} \;\; \text{for} \;\; J > 2. \tag{2}$$

Here $\bar{u}^{(2)}$ and Δ are called the *mean second virial coefficient* (quality of the solvent) and the *degree of amphiphilicity* respectively. The set $\{\sigma_n\}$ expresses the chemical composition, or the *primary sequence* of a heteropolymer.

The long timescale evolution of the conformational state is well represented by the Langevin equation [3],

$$\zeta_b \frac{d}{dt} \mathbf{X}_n = -\frac{\partial H}{\partial \mathbf{X}_n} + \boldsymbol{\eta}_n(t), \tag{3}$$

where ζ_b is the friction constant per monomer and $\boldsymbol{\eta}_n$ is the Gaussian noise with the second momentum,

$$\langle \eta_n^{\alpha}(t)\, \eta_{n'}^{\alpha'}(t') \rangle = 2k_B T \zeta_b\, \delta^{\alpha,\alpha'} \delta_{n,n'} \delta(t - t'), \tag{4}$$

and the Greek indices denote the spatial components of 3-d vectors.

The main idea of the Gaussian self–consistent (GSC) method [4, 5] is to choose the trial Hamiltonian, H_0, as a most generic quadratic form, with matrix coefficients depending on time,

$$H_0(t) = \frac{1}{2} \sum_{nn'} V_{nn'}(t)\, \mathbf{X}_n(t)\, \mathbf{X}_{n'}(t). \tag{5}$$

This corresponds to the Gaussian distribution of the inter–monomer distances with the mean squared,

$$D_{m\,m'}(t) \equiv \frac{1}{3} \left\langle (\mathbf{X}_m(t) - \mathbf{X}_{m'}(t))^2 \right\rangle. \tag{6}$$

Obviously, choosing Eq. (5) as the trial Hamiltonian is equivalent to replacing the nonlinear Langevin equation (3) by a linear stochastic ensemble,

$$\frac{d}{dt} \mathbf{X}_n = -\sum_{n'} V_{nn'}(t)\, \mathbf{X}_{n'} + \boldsymbol{\eta}_n(t). \tag{7}$$

The time–dependent coefficients are chosen at each moment in time according to the criterion,

$$\left\langle \mathbf{X}_n \frac{\partial H}{\partial \mathbf{X}_{n'}} \right\rangle_0 = \left\langle \mathbf{X}_n \frac{\partial H_0}{\partial \mathbf{X}_{n'}} \right\rangle_0, \tag{8}$$

where $\langle \ldots \rangle_0$ denotes the averaging over the trial ensemble. At equilibrium these equations become exactly the extrema conditions for the trial free energy in the Gibbs–Bogoliubov variational principle based on minimising the variational free energy, $\mathcal{A} = -k_B T \log \mathrm{Tr} \exp(-H_0/k_B T) + \langle H - H_0 \rangle_0$, with respect to $V_{nn'}$.

The GSC equations can be written in terms of instantaneous gradients of the variational free energy, $\mathcal{A}(t) = \mathcal{E} - T\mathcal{S}$ [6],

$$\frac{\zeta_b}{2}\frac{d}{dt}D_{mm'}(t) = -\frac{2}{3}\sum_{m''}(D_{mm''}(t) - D_{m'm''}(t))\left(\frac{\partial\mathcal{A}}{\partial D_{mm''}(t)} - \frac{\partial\mathcal{A}}{\partial D_{m'm''}(t)}\right). \tag{9}$$

The energetic and the entropic contributions in the free energy can be completely expressed in terms of the mean squared distances $D_{mm'}(t)$,

$$\mathcal{E} = \frac{3k_BT}{2l^2}\sum_n D_{n\,n-1,n\,n-1} + \sum_{J=2}^{\infty}\sum_{\{n\}'}\frac{u^{(J)}_{\{n\}}}{(2\pi)^{3(J-1)/2}}(\det\Delta^{(J-1)})^{-3/2}, \tag{10}$$

$$\mathcal{S} = \frac{3}{2}k_B \log\det R^{(N-1)}, \qquad R_{nn'} = \frac{1}{N^2}\sum_{mm'} D_{nm,n'm'}, \tag{11}$$

$$D_{mm',nn'} \equiv \frac{1}{2}(D_{m'n} + D_{mn'} - D_{mn} - D_{m'n'}), \quad \Delta^{(J-1)}_{ij} \equiv D_{n_1 n_{i+1}, n_1 n_{j+1}}. \tag{12}$$

In Eq. (11) we have the determinant of the truncated matrix $R^{(N-1)}$ to exclude the zero eigenvalue related to the translational invariance for the centre–of–mass of the system. In the second term in Eq. (10) the summation is taken over not coinciding indices, $n_1 \neq n_2 \neq \ldots \neq n_J$.

Let us also introduce the mean squared radius of gyration, R_g^2, and the micro–phase separation (MPS) order parameter, Ψ,

$$R_g^2 = \frac{1}{2N^2}\sum_{mm'} D_{mm'}, \qquad \Psi = \frac{1}{N^2 R_g^2}\sum_{mm'}\frac{\sigma_m + \sigma_{m'}}{2\Delta_\sigma}D_{mm'}, \qquad (\Delta_\sigma)^2 = \frac{1}{N}\sum_m \sigma_m^2. \tag{13}$$

The MPS parameter describes the degree of correlation between matrices of the relative two–body interaction, $(\sigma_m + \sigma_{m'})/2$, and the mean squared distances, $D_{mm'}$.

This method has been successfully applied by us [4, 6, 7] and others [5] for study of a number of interesting problems, of which we shall only consider two here. Its weakest point is obviously the assumption of the Gaussian distribution for the trial dynamics. However, direct simulations show that for polymers the distribution is close to Gaussian at large separation of monomers. As long distances are dominant in determining the conformational structure of a polymer, the Gaussian self–consistent approach is well justified. However, to achieve a better description for the coil one has to take into account next non–Gaussian corrections which may be important in certain situations.

3 Conformational states of amphiphilic copolymers

A typical phase diagram in terms of the mean second virial coefficient, $\bar{u}^{(2)}$, and the amphiphilicity, Δ, in Eq. (2) is presented in Fig. 1a. In the region $\bar{u}^{(2)} > 0$ and for small values of amphiphilicity typical conformations of copolymers are akin to the extended coil (see Fig. 2a). By decreasing $\bar{u}^{(2)}$ to the negative region the chain undergoes the continuous collapse transition, which is characterised by a rapid fall of the radius of gyration, R_g^2, and the change of the fractal dimension. The collapse transition for larger values of amphiphilicity turns out to be more complicated, and essentially dependent on the sequence. The globular state for large values of Δ is different from the liquid–like globule and consists of a hydrophobic core surrounded by a shell of hydrophilic monomers (see Fig. 2c). It is characterized by a somewhat larger value of the radius of gyration and an extremely large value of the MPS order parameter. That is why we call this state the MPS globule. The phase diagram in the intermediate region of $\bar{u}^{(2)}$ is much more complicated. Starting from some value of Δ there appear additional solutions corresponding to local minima of the free energy (the region between curves I' and II" in Fig. 1a). With increasing Δ the

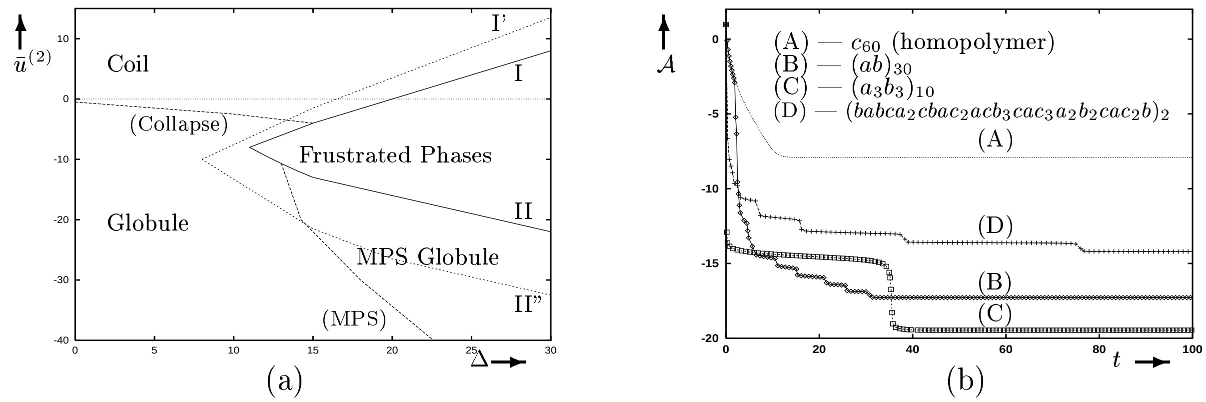


Figure 1: The phase diagram (Fig. (a)) of the sequence '$(babca_2cbac_2acb_3cac_3a_2b_2cac_2b)_2$' in terms of the mean second virial coefficient, $\bar{u}^{(2)}$, and the amphiphilicity, Δ. Time evolution (Fig. (b)) of the free energy, $\mathcal{A}$, for different sequences after an instantaneous quench from the coil state, $\bar{u}^{(2)} = 15$ and $\Delta = 0$, to the region with $\bar{u}^{(2)} = -25$ and $\Delta = 30$. Monomers 'a', 'b' and 'c' correspond to the values of $\sigma_n = -1$ (hydrophobic), 1 (hydrophilic) and 0 ("neutral") respectively. We fix the units of temperature, size and time by choosing $k_BT = 1$, $l = 1$ and $\zeta_b = 1$. We also fix higher virial coefficients: $u^{(3)} = 10$ and $u^{(J)}_{\{n\}} = 0$ for $J > 3$. In Fig. (a) curves (Collapse) and (MPS) correspond respectively to the collapse and the MPS continuous transitions. Curves (I) and (II) correspond to discontinuous transitions to the frustrated phases. "Spinodal" curves (I') and (II") bound the regions of metastability of the frustrated states.

number of such solutions grows quickly. Significantly, in the region of the phase diagram, between curves I and II in Fig. 1a, some of these possess the lowest free energy value, thus being the thermodynamically stable state. Since the number of such solutions is rather high even for short sequences and their number grows quickly with the chain length, we do not attempt to draw all their boundaries of (meta)stability. We shall call them collectively as *frustrated phases*. In Fig. 2b we exhibit a typical polymer conformation corresponding to these phases.

In the series of pictures in Fig. 3 we present the pattern of the matrix of mean squared distances, $D_{mm'}$, for the 'ab' copolymer at some values of the mean second virial coefficient, $\bar{u}^{(2)}$. For positive $\bar{u}^{(2)}$ (see Fig. 3a) the mean squared distances possess the structure typical for the extended coil, increasing monotonically on moving away from its diagonals towards the distance of half–ring along the chain. Decreasing the mean second virial coefficient, $\bar{u}^{(2)}$, causes the copolymer to pass through frustrated states (Fig. 3b), finally reaching the MPS globule state (see Fig. 3c). The characteristic feature of the $D_{mm'}$ matrix in a frustrated state is that it possesses some number of monomer groups having smaller distances between each other than between monomers from other groups. Clearly, such a group represents a *cluster* of monomers, so that here the copolymer chain forms a set of micro–phase separated clusters. In the regime of nearly compensating repulsion and attraction between monomers it is more preferable to achieve phase separation on a smaller, than the globular, scale by forming clusters.

The kinetics of folding from the coil to the MPS globule for copolymers takes much longer than for the homopolymer since it is strongly affected by the presence of the transient frustrated states (akin to Fig. 2b) along the kinetic pathway [6]. This leads to a complicated kinetic process (see Fig. 1b) consisting of multiple steps with pronounced slowing down and then acceleration in the folding rate. The folding kinetics also strongly depends on the primary copolymer sequence. Typically, the kinetics for copolymers consisting of long blocks proceeds in a smaller number of steps, but not necessarily faster, than kinetics for short block copolymers. Small modifications of the sequence may change crucially the overall kinetic behaviour and even the final kinetic state itself.

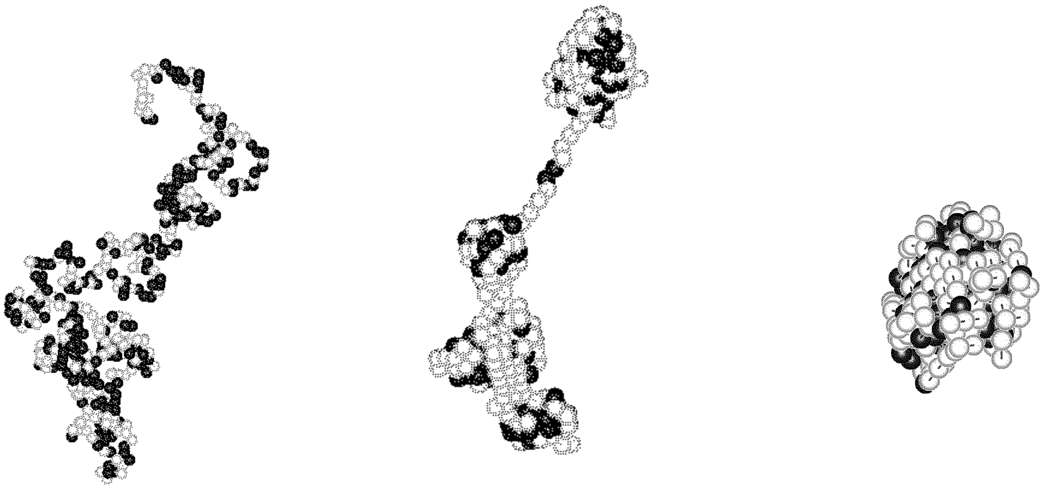


Figure 2: Typical copolymer conformations from lattice Monte Carlo simulations [8] for the coil, the frustrated and the MPS globule.

4 Conformational states of a stiff homopolymer

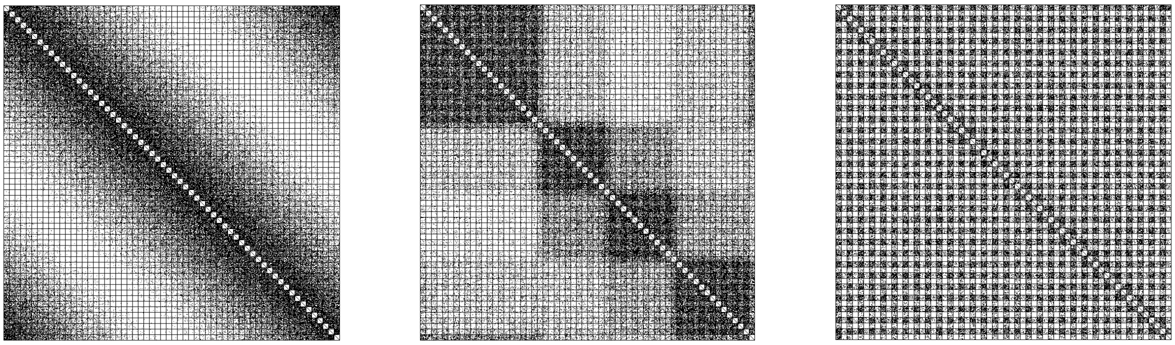
In this section we shall be interested in how the homopolymer conformations are modified due to the effect of stiffness. There is a significant amount of theoretical [9] and experimental [10] literature dealing with various questions about equilibrium properties of rigid chains, the most important practical example of which is DNA. Experimentally, it is well known that DNA can acquire a torus–like shape in its globular state, and that condensation of DNA induced by various agents could lead to even more complicated phases [10]. The physical reason for a torus is clear — a persistent chain has no desire to bend, so it tends to have as large a radius of curvature as possible, consistent with quite close packing of the chain. In the framework of the bead–and–spring model to account for the stiffness it is sufficient to introduce the bending energy, $\sum_n (d^2\mathbf{x}_n/dn^2)^2$, as first proposed in Ref. [11]. This leads to the following bending contribution to the mean energy in Eq. (10),

$$\mathcal{E}_{bend} = \frac{3k_BT\lambda}{2l^3}\sum_n (2D_{n+1\,n,n+1\,n} + 2D_{n-1\,n,n-1\,n} - D_{n-1\,n+1,n-1\,n+1}), \tag{14}$$

where λ is called the persistent length of the polymer chain.

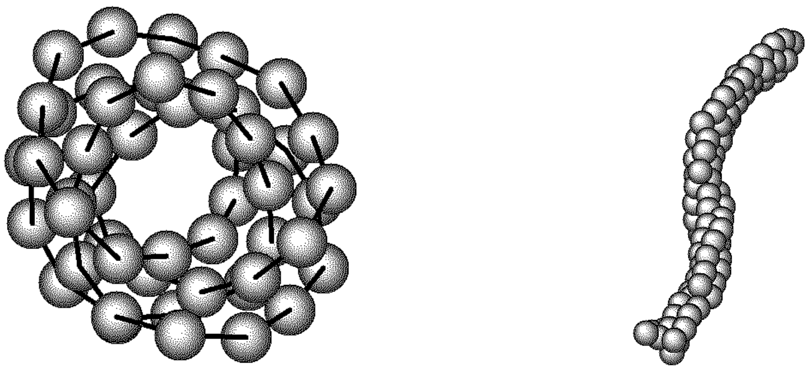
Analysis of the GSC equations for a stiff homopolymer shows [7] that the toroidal conformation exists in some intermediate region in the second virial coefficient at comparatively large values of the persistent length. Thus, decreasing $u^{(2)}$ quasistatically brings the rigid chain first into the toroidal globule which then collapses filling the interior of the torus. Our analysis also shows that this phase diagram and the toroidal conformation itself are rather sensitive to the chain length. The toroidal conformation can be clearly identified from the typical oscilating behaviour of the mean squared distances, D_{0m}, with the number of peaks corresponding to the number of windings [7].

Study of such systems by Monte Carlo method [8] shows that a variety of metastable states is also present here. In fact, amongst toroidal states (Fig. 4a) with different number of windings there also exist metastable states of *hairpins* (Fig. 4b). In a hairpin the chain forms a nealy linear conformation turning around several times thereby minimising contacts with solvent. Such states can be easily obtained by an abrupt quench from the extended coil.



(a) (b) (c)

Figure 3: Patterms of the mean squared distances matrix, $D_{mm'}$ for '$(ab)_{30}$' copolymer and amphiphilicity $\Delta = 20$. Diagrams (a-c) correspond respectively to the values of the mean second virial coefficient $u^{(2)} = 15$, -21 and -40. Indices m, m' are counted from the upper left corner. Each matrix element, $D_{mm'}$ is denoted by a quadratic cell with varying degree of the black colour, the darkest and the lightest cells corresponding respectively to the smallest and to the largest mean squared distances.



(a) (b)

Figure 4: Typical conformations of a rigid chain from Monte Carlo simulations for the toroidal globule and the hairpin state.

5 Conclusion

As we have seen even simple models of biopolymers that take into account the connectivity and stiffness of the chain as well as the monomer specific volume interactions (attractions and repulsions) lead to extremely complicated phase diagrams and even more complex non–equilibrium structures and processes. This is due to a subtle interplay of competing interactions with different ranges acting in two different metrics (of the 1-d chain and the 3-d embedding space).

We would like to note that although an individual DNA can form a toroidal globule which is described by our simple model, DNA in a cell normally exists inside complexes involving many special proteins and other ligands. The corformational structure of such complexes is not completely understood at the moment and obviously requires much more detailed models.

The amphiphilic copolymer model that we have considered here only describes the hydrophobic interactions in proteins. There are other factors that are important for stabilising the native structure such as electrostatics, hydrogen bonding, secondary structure and others. These, in principle, may be added by including additional terms into the model.

However, a more important conceptual point is that protein sequences are very special as they were selected during billions of years of biological evolution. The statistical properties of protein sequences should reflect their optimal character in the stability of the native state and efficiency of folding. A typical protein is about several hunders monomers long and with 20 different types of amino acids there may be roughly 20^{500} different proteins, and yet only hundreds of thousands of proteins are known to exist. So any reasonable model of proteins [12] should take the special properties of their sequences into account.

Thus, we hope we have convinced the reader that the methods that are used in statistical mechanics and high energy physics can be also successful in attacking biological problems. The skill of a theorist in this area still remains in being able to understand the most important factors in complex systems and to build adequate models that would be solvable and sufficiently non–trivial.

* * *

The authors acknowledge interesting discussions with Professor K. Kawasaki, Professor G. Parisi, Dr A.V. Gorelov, Dr R.V. Polozov and Dr D.A. Tikhonov.

References

[1] H. Lodish, D. Baltimore, A. Berk, S.L. Zupirsky, P. Matsudaira, J. Darnell, Molecular Cell Biology (Scientific American Books, NY, 1995); G.M. Blackburn, M.J. Gait, Nucleic Acids in Chemistry and Biology (Oxford University Press, NY, 1996).

[2] T.E. Creighton (ed.), Protein Folding (Wiley, New York 1992); R. Elber, (ed.), New Developments in Theoretical Studies of Proteins (World Scientific, Singapore 1994).

[3] P. G. de Gennes, Scaling Concepts in Polymer Physics (Cornell Univ. Press, NY, 1988); J. des Cloizeaux, G. Jannink, Polymers in Solution (Clarendon Press, Oxford, 1990); M. Doi, S. F. Edwards, The Theory of Polymer Dynamics (Oxford Science, NY, 1989); A. Yu. Grosberg, A. R. Khokhlov, Statistical Physics of Macromolecules (AIP, NY, 1994).

[4] E. G. Timoshenko, Yu. A. Kuznetsov, K. A. Dawson, J. Chem. Phys. 102 (1995) 1816; Yu. A. Kuznetsov, E. G. Timoshenko, K. A. Dawson, J. Chem. Phys. 104 (1996) 3338; E. G. Timoshenko, Yu. A. Kuznetsov, K. A. Dawson, Phys. A 240 (1997) 432.

[5] G. Allegra, F. Ganazzoli, J. Chem. Phys. 83 (1985) 397; G. Raos, G. Allegra, J. Chem. Phys. 104 (1997) 1626.

[6] E. G. Timoshenko, Yu. A. Kuznetsov, K. A. Dawson, Phys. Rev. E 53 (1996) 3886; E. G. Timoshenko, Yu. A. Kuznetsov, K. A. Dawson, Phys. Rev. E 54 (1996) 4071; E. G. Timoshenko, Yu. A. Kuznetsov, K. A. Dawson, in Proceedings of the International Conference on Morphology and Kinetics of Phase Separating Complex Fluids, edited by F. Mallamace (Messina, Italy, 1997); E. G. Timoshenko, Yu. A. Kuznetsov, K. A. Dawson, submitted to Phys. Rev. E (1997).

[7] Yu. A. Kuznetsov, E. G. Timoshenko, K. A. Dawson, J. Chem. Phys. 105 (1996) 7116; Yu. A. Kuznetsov, E. G. Timoshenko, K. A. Dawson (1997) unpublished.

[8] Yu. A. Kuznetsov, E. G. Timoshenko, K. A. Dawson, J. Chem. Phys. 103 (1995) 4807; Yu. A. Kuznetsov, E. G. Timoshenko, K. A. Dawson, J. Chem. Phys. 104 (1996) 336.

[9] V.A. Bloomfield, Biopolymers 31 (1991) 1471; J. Ubbink, T. Odijk, Biophys. J. 68 (1995) 54; N.V. Hud, K.H. Downing, R. Balhorn, Proc. Natl. Acad. Sci. USA 92 (1995) 3581.

[10] L.S. Lerman, Proc. Natl. Acad. Sci. USA 68 (1971) 1886; U.K. Laemmli, Proc. Natl. Acad. Sci. USA 72 (1975) 4288; Yu.M. Evdokimov *et al*, Nucl. Acids Res. 3 (1976) 2353; G.E. Plum, P.G. Arscott, V.A. Bloomfield, Biopolymers 30 (1990) 631; V.V. Vasilevskaya, A.R. Khokhlov Y. Matsuzawa, K. Yoshikawa, J. Chem. Phys. 102 (1995) 6595.

[11] O. Kratky, G. Porod, Rec. Trav. Chim. 68 (1949) 1106.

[12] K. Yue, K.A. Dill, Proc. Natl. Acad. Sci. USA 89 (1992) 4163; E.I. Shakhnovich, A.M. Gutin, Proc. Natl. Acad. Sci. USA 90 (1993) 7195; V.S. Pande, A.Yu. Grosberg, T. Tanaka, J. Chem. Phys. 103 (1995) 9482; A. Irbäck, C. Peterson, F. Potthast, Phys. Rev. E 55 (1997) 4260.