Physics – Uspekhi **40** (2) 125 – 158 (1997)

REVIEWS OF TOPICAL PROBLEMS PACS numbers: **36.20.–r**, **64.70.–p**, **87.10.+e**, **87.10.–v**

Disordered polymers

A Yu Grosberg

Contents

1. Introduction **125**
1.1 Types of disorder in polymer macromolecules; 1.2 Disorder and biopolymers; 1.3 Heteropolymers and the protein folding problem; 1.4 Note about the terminology; 1.5 What is NOT discussed in this article
2. Disorder of branches **127**
2.1 The Zimm – Stockmayer equation; 2.2 Quenched and annealed branches; 2.3 Flory type theory for the quenched branched polymer; 2.4 Flory type theory for the annealed branched polymer; 2.5 Preparation conditions for the quenched case; 2.6 Discussion and comparison with other theoretical and Monte Carlo data
3. Disorder of topology: knots and entanglements **130**
3.1 Why topology? 3.2 Topology and disorder; 3.3 General formulation of the topological problems in polymer statistics; 3.4 Briefly on knot invariants; 3.5 What is known of knot entropy? 3.6 Crumpled globules; 3.7 Knot inflation; 3.8 Tight knots
4. Freezing transition of globular heteropolymers with random sequence **137**
4.1 Globular heteropolymers; 4.2 The random energy model (REM); 4.3 Is REM valid for heteropolymer freezing? 4.4 Annealed heteropolymers; 4.5 Freezing transition; 4.6 Computational tests of freezing
5. Designed heteropolymers **146**
5.1 Design of sequence using canonical ensembles; 5.2 Designing and folding with different interactions
6. Statistics of real protein sequences **153**
6.1 Individual vs. statistical approach; 6.2 The random walk technique to study the statistics of sequences; 6.3 The design of sequences and models of evolution
7. Conclusions **155**
References **156**

Abstract. **A single polymer macromolecule is considered with disorder types such as branches, knots, and heterogeneous sequences of chemical units. In all cases, simple theoretical approaches are employed to gain useful physical insights. For branched polymers, a simple Flory-type theory is described by means of which the difference between the universality classes for molecules with quenched and annealed branches is demonstrated. For knots, another Flory-type theory is suggested to describe the swelling and/or collapse of a quenched topology ring or the size distribution for the annealed case. To consider heteropolymers, the Random Energy Model borrowed from the spin glass theory is systematically employed. This allows a simple yet rigorous description of both the freezing transition of a random sequence globule and the use of the canonical ensemble for designing sequences with energy-optimized ground state conformation. Along with the analytical theory, computer tests for the freezing and design processes are discussed. The sequence design scheme is shown to yield a specific prediction concerning the character of correlations in protein sequences. Statistical tests confirming this prediction are described.**

A Yu Grosberg Institute for Biochemical Physics
117977 Moscow, Russia
Department of Physics and Center for Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA
E-mail: shura@gels.mit.edu

Received 3 September 1996
Uspekhi Fizicheskikh Nauk **167** (2) 129 – 166 (1997)
Translated by A Yu Grosberg; edited by M S Aksent'eva

1. Introduction

Although in the 45 years of his scientific career Il'ya Mikhaĭlovich Lifshits worked fruitfully in many areas of theoretical condensed matter physics, the theory of disordered systems lies at the very heart of his scientific legacy. In particular, he came to polymers from the biopolymers side (see the title of his first work [1]), and his view of polymers was always somewhat tilted toward the disordered systems perspective, as for example, in his concept of a 'linear memory.' The approach of Lifshitz, in this respect, contrasts the approaches of S F Edwards [2] and P-G de Gennes [3], for whom the field-theoretical and critical phenomena aspects were the most important. When, in the mid-sixties, several prominent soviet physicists and mathematicians started to participate in molecular biology gatherings, and some of them were said to 'switch to biology,' friends tried to talk Lifshits out of this move. His friends fears were groundless: he did not switch anywhere, but found a real physical approach to biopolymers, and his works in this direction were an 'analytic continuation' of his other works in physics.

The present author witnessed the enthusiasm and curiosity that Lifshits kept literally till the last day of his life, for the physical understanding of what allows biopolymers to function in the way they do. In recent years, a deeper level of

understanding of these problems has been achieved. It seems very natural to review this progress in the journal dedicated to I M Lifshitz.

– · –

To simplify its reading, we begin the article with a brief glossary of the main terms:

Conformation — the spatial shape of the polymer backbone. In the statistical mechanics of polymers, conformation plays the role of microstate, the partition function represents the sum over conformations.

Construction of sequences —the process of formation of sequences of heteropolymer in such a way that the polymer takes on the desirable conformation (corresponding to self-organization of protein conformations).

Disorder, annealed — the elements of disorder which take part in thermal motion and eventually relax to thermodynamic equilibrium.

Disorder, quenched — the degrees of freedom that are frozen with the preparation of the system and do not take part in any further thermal motion.

Freezing transition — the phase transition between two globular (compact) phases, one of which is similar to homopolymer globules, comprised of an exponentially large number of conformations, while the other is dominated by very few conformations.

Heteropolymer — a polymer with a chemically different monomer species.

Homopolymer — a polymer of identical monomers.

Independent Interaction Model (IIM) — heteropolymer model in which the interaction energy for every pair of monomers is a random value independent of all other interactions.

Knot — a conformation of a closed ring macromolecule embedded in three-dimensional space.

Protein folding — self-organization of a unique native conformation for the protein macromolecule.

Random Energy Model (REM) — a model of disordered systems in which energy of each microstate is assumed random and independent of other energies.

1.1 Types of disorder in polymer macromolecules

In Figure 1, we schematically summarize the types of disorder that can exist in a polymer macromolecule and will be discussed in the present article. They include disorder of branches, topological disorder related to knots and links, and disorder of sequences in heteropolymers. In every case, the distinction must be made between the regimes of quenched and annealed disorder.

While this is very well known and a very important concept of general importance for statistical physics, it is worth reminding the reader about the main points:

(1) Quenched elements are formed during the system's preparation, and cannot be changed by thermal motion;

(2) Annealed elements participate in the thermal motion.

In theoretical jargon, one often says that the difference is what you average: for a quenched system, one has to average the free energy (which is difficult and challenging); for an annealed system, one has to average the partition function (which is much easier). While this is certainly correct, it is only a semi-truth (reminiscent of Bulgakov's 'salmon of a second freshness'). The reason why the free energy is averaged for the quenched system is to do with the principle of self-averaging

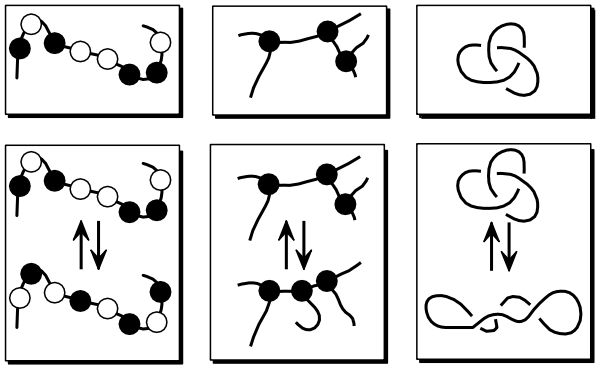

Figure 1. Types of disorder in polymer macromolecule.

of free energy [4]: free energy is distributed normally over the realizations of disorder, and its average is dominated by the 'typical' realizations; meanwhile, the partition function is distributed exponentially more broadly, and its average is dominated by some weird realizations — those that the system selects if it is allowed to thermally (ergodically) sample all the realizations.

1.2 Disorder and biopolymers

The view that to average free energy is the only meaningful thing to do for quenched systems arises from experience with spin glasses and other 'normal' physical systems, where one can never experimentally control all the details of disorder, but only some statistical properties, such as the densities of impurities, etc. Biological systems provide physics with very different possibilities. For example, the sequence of monomers in a protein can play the role of an element of quenched disorder, because as long as the protein chain exists in the solution, its sequence does not change. Using the synthetic apparatus of a living cell, the experimenter is able to produce a macroscopic quantity of identical copies of this disordered system. It is like being able to reproduce, after annealing, the exact spatial arrangement of paramagnetic centers in a spin glass sample. While these questions will be only touched in the present review, and we will mostly average the free energy, one has to keep in mind that other formulations are also possible due to these properties of biological systems.

In fact, this goes back to the fact that while biopolymers are certainly quenched, they are not exactly disordered; their quenched elements are often the product of evolution, as, for example, protein sequences. This leads to very interesting and deep physical questions; some of them will be discussed later in this article.

The lion's share of this review is devoted to heteropolymers, as they are considered a model of one of the most challenging problems of protein folding. It is reasonable to make here few preliminary comments on this.

1.3 Heteropolymers and the protein folding problem

The phenomenon of protein folding is the ability of protein chains to renature, that is, a single protein molecule is able to find its 'correct' native 3D conformation. Many proteins (though not all of them) do not need any assistance and are able to do this in a dilute solution [5]. This phenomenon is a deep challenge to statistical physics (and physicists), as is seen most dramatically from the following statement known as the

Levinthal paradox. A polymer of N monomers can have as many as $\exp(\omega N)$ conformations. To sample all of them would require astronomical time. How can the chain find the correct conformation, if it is unique, and exhaustive sampling is impossible?†

In the last decade, there has been remarkable progress in the theoretical understanding of protein folding. This progress is mainly due to the concept of heteropolymer freezing, which is the phase transition of a heteropolymer chain between two compact globular phases, of which one is dominated by an exponentially large number of conformations ($O(\exp(N))$) while in the other only one or very few of them ($O(1)$) are thermodynamically relevant. To describe heteropolymer freezing, ideas and concepts were borrowed from the statistical mechanics of spin glasses, such as the Random Energy Model (REM) first suggested by Derrida [6].

There are two approaches, which can be considered parallel. One of them, rooted in the seminal work by Bryngelson and Wolynes [7], employs rather simple mathematics, but unfortunately is vague as regards to the formulation of the model considered. Not surprisingly, it is extremely difficult, if not altogether impossible, to either generalize this approach or to find out its conditions of applicability (see also [8]). Another approach, started from the other seminal contribution by Shakhnovich and Gutin [9], considers a model which is very clearly formulated, but unfortunately, this theory remains overshadowed by the complexity of the theoretical machinery employed. Not surprisingly, although this theory is widely considered very important, it remains hardly known beyond some qualitative conclusions.

In the meantime, both theories are related to REM, which in turn, is fairly simple both as regards its physical nature and mathematical treatment. It should therefore be possible to formulate the theory that combines the simplicity of the Bryngelson – Wolynes approach with the sophistication of the Shakhnovich – Gutin theory. Such an approach has been recently developed [10] and will be discussed below.

The first models of heteropolymers considered just random sequences of monomer species. This was consistent with the fact that real protein sequences statistically look very much like random sequences [11]. Moreover, simple estimates of the time and material involved in evolution indicate that the sequences could not have evolved very far from randomness. However, upon a more detailed examination, random sequences were found to be not sufficiently protein-like. In particular, while many chains with sequences taken at random have indeed unique ground states, these ground states are not sufficiently robust, so that a minor perturbation of the interaction energies (induced, for example, by a change in the surrounding solution) leads to a complete change of the ground state conformation. Furthermore, for most of the sequences whose ground states are unique, folding to these states is neither quick nor reliable. This is by no means surprising, as protein sequences are known to have undergone evolutionary optimization. It was hypothesized on the qualitative level [7] that protein evolution has resulted in sequences which obey the 'minimal frustration principle.' Real models of sequence 'design' were suggested very recently: Shakhnovich and Gutin anneal sequences by the Monte Carlo method with the criterion that the sequence minimized the energy of a particular target conformation [12]; another incarnation of the same idea is the so-called Imprinting Model [13]. Interestingly, as soon as these models were suggested, a prediction was made as to which kind of correlations should exist in protein sequences, and those correlations were indeed immediately found [14]. We shall discuss all these ideas in the present work.

We note that the problem of freezing of heteropolymers is not confined to biopolymers such as proteins. Indeed, the freezing of synthetic polymers has attracted great interest, due to potential industrial and biomedical applications. Thus, perhaps in the pursuit of understanding proteins, we have made some progress in the direction of making synthetic *protein-like* heteropolymers as well.

1.4 Note about the terminology

Although the terms quenched and annealed disorder are commonly used in English, the Russian term, especially for annealed case, does not seem to be commonly accepted...

1.5 What is NOT discussed in this article

There are other important examples of disordered polymer systems which are not discussed here, including polymer gels (see the seminal work by S Panyukov and Y Rabin [15]), block-copolymers, with their ability to form domain structures (see the works [16 – 18] and references therein), polymers in disordered external fields, adsorption of heteropolymers on the surface [19 – 25] or on the selective interface [26 – 29], melting of heteropolymeric DNA [30 – 32], uneven flexibility of DNA [33, 34], polyampholytes [35 – 39] and many others. The collapse of annealed heteropolymers was considered in [40], and recently the results were re-discovered in [41]. It is impossible to review everything; this article concentrates on branched polymers, knots, and mostly, heteropolymers.

2. Disorder of branches

Branched polymers are of significant interest both for synthetic polymer chemists and for biophysicists. Most of the synthetic polymers are somewhat branched. Also, RNA molecules form clover leaf structures that can be viewed as branched polymers, with elements of secondary structure forming branches. An even better example of a branched polymer is super-coiled DNA [42, 43], of which the reader can get a good idea by twisting a telephone cord (unfortunately, the interesting behavior begins when it is already pretty dangerous for the telephone).

2.1 The Zimm – Stockmayer equation

To begin with, let us disregard volume interactions and ask what is the size of an ideal randomly branched polymer, R_{id}? This was answered by B Zimm and W Stockmayer as early as in 1949 [44]. Their result reads $R_{\mathrm{id}} \propto N^{1/4}$. There are several ways to derive it; one of the simplest is given in the book [32]. It is based on the estimate of the 'chemical diameter' of the branched structure, that is, the chemical interval, or contour distance, between two arbitrary ends of the structure, and it is done by mapping the branched structure on a Cayley tree graph (see Fig. 2). By choosing an arbitrary origin point O on the Cayley tree, the branched structure can be viewed as a one dimensional random walk with possible steps to and from O. This yields an estimate that the chemical diameter, L, scales as the square root of the number of monomers in the macro-

† There are some proteins which need some help to fold, and the cell is able to provide that assistance. From the physics perspective, however, it is important that at least some proteins are able to fold spontaneously. The first task is to explain that.

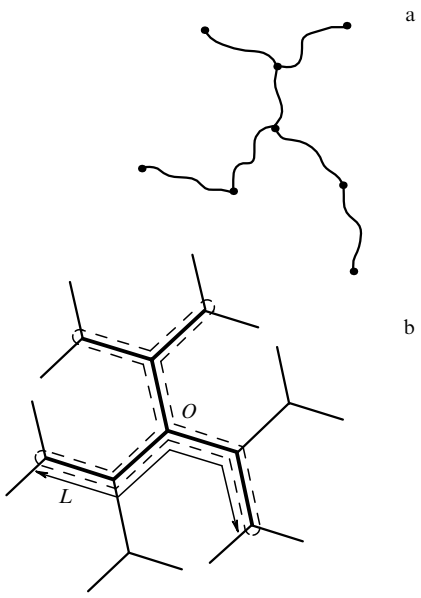

Figure 2. The branched structure without cycles (a) can be mapped onto the Cayley tree graph (b). On the graph, the image can be encircled with the 'ring polymer' (dashed line), its characteristic distance from the origin O scales as the chemical diameter of the structure, L.

molecule, N, or more accurately,

$$L \propto g\sqrt{\frac{N}{g}} \propto \sqrt{Ng}\,, \tag{1}$$

where g is the characteristic number of monomers in the linear part of the structure, between two neighboring branch points. As each diameter in the ideal polymer represents a Gaussian linear chain, we arrive at

$$R \sim aL^{1/2} \sim a(Ng)^{1/4}\,, \tag{2}$$

where a is a monomer length.

Thus, in three dimensions ($d = 3$), branched polymers are very compact and dense, and this is why the excluded volume effect is very strong.

2.2 Quenched and annealed branches

Speaking of the excluded volume for the branched polymer, one has to make the distinction between two extremes. To explain it, let us say that the branched polymer corresponds to the tree-like graph embedded in real space. In one extreme, the branched polymer is far from equilibrium such that the structure of the graph is fixed. The opposite extreme is the situation where the positions of the branches fluctuate and are in thermal equilibrium. These two extremes are called quenched and annealed, respectively. It turns out that polymers with quenched and annealed branches belong to different universality classes and are characterized by different critical indices.

The physical reason for the difference between quenched and annealed branches is simple. The excluded volume leads only to the stretching of subchains in the case of quenched branches, but it is accompanied by the rearrangement of branches in the annealed case. This leads to an extra contribution to the corresponding entropy and changes the universality class of the problem [45].

2.3 Flory type theory for the quenched branched polymer

Flory theory does not pretend to yield exact results, but it is very simple and it is also known that its results are reasonably accurate for linear polymers. In this approach, one says that the characteristic size of a polymer, R, is balanced so as to minimize the free energy

$$F(R) = F_{\text{elast}}(R) + F_{\text{int}}(R), \tag{3}$$

where $F_{\text{elast}}(R)$ and F_{int} stand for the polymeric entropy and interaction parts of the free energy. As for the interaction part, it can be written in the usual second virial form

$$F_{\text{int}} \sim T\frac{vN^2}{R^d}, \tag{4}$$

where v is the excluded volume. For now, let us assume that $v \approx a^d$.

In complete analogy with the Flory derivation for a linear polymer, it is tempting to write the elastic contribution in the form

$$F_{\text{elast}}(R) = T\,\frac{R^2}{R_0^2}, \tag{5}$$

with $R_0 \sim a(gN)^{1/4}$ standing for the unperturbed size of a polymer. From equations (3) and (4), one automatically arrives at

$$R \sim aN^{\nu_{\text{qu}}} g^{\mu_{\text{qu}}} \tag{6}$$

with

$$\nu_{\text{qu}} = \frac{5}{2(d+2)}\,, \qquad \mu_{\text{qu}} = \frac{1}{2(d+2)}\,. \tag{7}$$

This result was obtained in [46, 47]. Obviously, the derivation assumes that the structure of branches is not subject to change in response to swelling, which is correct only for the quenched case. Interestingly, it was not pointed out in [46, 47].

2.4 Flory type theory for the annealed branched polymer†

In order to take into account the additional rearrangement of branches one should notice that characteristic diameter L obeys the Zimm – Stockmayer estimate (1) in an unperturbed polymer only. For a polymer with excluded volume and an annealed system of branches, the branches will rearrange, changing the characteristic number of bonds between ends.

To describe this factor, we can again use Flory theory, this time applying it to the polymer placed on the Cayley tree (see Fig. 2). How many configurations are there for the diameter L? As was already mentioned, these configurations can be mapped onto a one dimensional random walk (to or from the origin O on the Cayley tree), and thus the answer is given by the entropy of a linear polymer stretched out to the end-to-end distance L, because, on the Cayley tree, L plays the role of spatial size. Thus, we get

$$F_{\text{br}} = T\frac{(L/g)^2}{N/g} = T\,\frac{L^2}{Ng}\,. \tag{8}$$

Thus, our Flory theory is now like a Russian doll (one inside the next). First, the elastic free energy now has two

† This section is based on the work [45].

Flory-type contributions of the form:

$$T\frac{R^2}{a^2L}+T\frac{L^2}{Ng}\,. \tag{9}$$

Interestingly, both terms here are of the same type, except that L plays the role of chain length in the first term and of the chain spatial size in the second term. Minimization with respect to L yields [45]

$$L\sim\left(\frac{R}{a}\right)^{2/3}(Ng)^{1/3} \tag{10}$$

and

$$F_{\rm elast}^{\rm ann}\sim T\frac{(R/a)^{4/3}}{(Ng)^{1/3}}\,. \tag{11}$$

We now have to minimize the free energy

$$F=F_{\rm int}+F_{\rm elast}^{\rm ann}\sim Tv\frac{N^2}{R^d}+T\frac{(R/a)^{4/3}}{(Ng)^{1/3}} \tag{12}$$

with respect to R. This gives

$$R\sim aN^{\nu_{\rm ann}}g^{\mu_{\rm ann}} \tag{13}$$

with

$$\nu_{\rm ann}=\frac{7}{3d+4}\,,\qquad \mu_{\rm ann}=\frac{1}{3d+4}\,. \tag{14}$$

It is also instructive to write down the relationship between L and N. This involves new critical exponents ρ and σ:

$$L\sim N^{\rho}g^{\sigma}\,,\qquad \rho=\frac{d+6}{3d+4}\,,\qquad \sigma=\frac{d+2}{3d+4}\,. \tag{15}$$

Clearly, this can be valid only as long as $L<N$, or at

$$g<g^{\star}=N^{(1-\rho)/\sigma}\sim N^{2(d-1)/(d+2)}\;; \tag{16}$$

if $g>g^{\star}$, the polymer is so weakly branched, that rearrangement of its branches brings it to an essentially linear form with $L\simeq N$. Not surprisingly, when $g=g^{\star}$, equation (13) yields

$$R\sim aN^{\nu_{\rm lin}}\sim aN^{3/(d+2)}$$

just as equation (6) does for quenched case when $g=N$.

2.5 Preparation conditions for the quenched case

If one thinks a little deeper about quenched branches, then it becomes apparent that there could be very different patterns of branches, ranging from a regular comb with $L\approx N/2$ to regular Cayley type structures with $L\sim\ln N$. The result above, equation (6), is meant to be valid for random branches, because the majority of branched structures obey equation (1) and have $L\sim N^{1/2}$. Nevertheless, one can indeed prepare different branched structures, and it is reasonable to ask the question of their sizes. One simple way to address this would be to say that the preparation of a branched structure involves its stabilization in the annealed regime in one solvent, then its freezing and putting in another solvent.

To examine this, let us now say that the excluded volume depends on the solvent and thus we prepare our polymer in the solvent where the excluded volume, $v=\tau_p a^d$, and then, when the structure is frozen, we observe it in a solvent where $v=\tau a^d$. In general, for a quenched structure with arbitrary chemical diameter L, the Flory free energy (4) and (5) is given by

$$\frac{R^2}{La^2}+\frac{vN^2}{R^d}\,, \tag{17}$$

yielding upon minimization

$$R\sim a(\tau LN^2)^{1/(d+2)}\,. \tag{18}$$

As regards L, it is given roughly by (15), because our polymer was annealed under the preparation conditions. More precisely, keeping track of $\tau_{\rm p}$, we have

$$L\simeq\tau_{\rm p}^{2/(3d+4)}N^{\rho}g^{\sigma} \tag{19}$$

and thus

$$\begin{aligned}R&\simeq\tau_{\rm p}^{2/(3d+4)(d+2)}\tau^{1/(d+2)}N^{\nu_{\rm ann}}g^{\mu_{\rm ann}}\\&\simeq\left(\frac{\tau_{\rm p}}{\tau}\right)^{2/(3d+4)(d+2)}R_{\rm ann}\,.\end{aligned} \tag{20}$$

As one might have suspected, the critical power describing N dependence is the same here as in the annealed case, but this is valid only as long as τ and $\tau_{\rm p}$ are considered N independent. However, using for example the solution where τ is very small one can potentially prepare a branched polymer whose structure and swelling is very different from both annealed and random quenched polymers.

The main message of this section is that the annealing of the structure has to do with the preparation procedure for quenched systems. We shall consider this point in more detail later.

2.6 Discussion and comparison to other theoretical and Monte Carlo data

The dependencies of critical exponents ν for both quenched and annealed branched polymers on the spatial dimension d can be easily plotted along with a similar dependence for linear polymers given by the Flory formula $\nu=3/(d+2)$. It is known [46], that $d=8$ is the upper critical dimension for branched structures, and thus it is not surprising that Flory theory yields unperturbed values for ν, both for quenched and annealed cases, in $d=8$.

For $d=3$, we are fortunate to have an exact result from the field-theory approach [48] for the annealed case: $\nu_{\rm ann}=1/2$. Flory theory yields the value $1/2$ for the quenched case $\nu_{\rm qu}=0.5$; for the annealed case, it gives $\nu_{\rm ann}=7/13\approx0.54$. This can characterize the accuracy of the Flory approach. Neither of the works [46, 47] mentioned that they were dealing with quenched branches, nor did the work [48] quote that its subject was in the annealed regime, and their results were thought to be in agreement.

What is really important is that always $\nu_{\rm qu}<\nu_{\rm ann}$. This is clear physically: an annealed polymer has some additional freedom, and uses it to escape unfavorable excluded volume effects. This also agrees with the $\epsilon=8-d$-expansion, that yields to the first order

$$\nu_{\rm ann}\simeq\frac{1}{4}\left(1+\frac{\epsilon}{9}+\ldots\right),\qquad \nu_{\rm qu}\simeq\frac{1}{4}\left(1+\frac{\epsilon}{10}+\ldots\right)$$

[51, 52].

The latest computer simulation [52] yielded $\nu_{\rm ann} = 0.49 \pm 0.01$ and $\nu_{\rm qu} = 0.45 \pm 0.01$, which is similar to the Flory-type theory results as regards the relative difference between the two exponents.

The difference between the quenched and annealed cases manifests itself in a variety of physical properties, including osmotic pressure of the semi-dilute solution, permeability through thin capillaries, etc. Many of them can be studied using the standard scaling approach [2] (see [45]).

For the linear polymer, except for swelling, which is related to the excluded volume problem and self-avoidance, there is an opposite regime of collapse, or coil-globule transition. It is relevant for $d > 2$, because in $d = 2$ a Gaussian polymer would have N-independent density of order unity, and there would be no room to collapse. Similarly, branched polymer collapse may exist in $d > 4$. Flory-type theory can be generalized to this case, and also a more sophisticated theory can be developed [53] that follows the lines of the Lifshitz theory for globules of linear polymer [1].

3. Disorder of topology: knots and entanglements

3.1 Why topology?

The image that comes to mind when one thinks about a linear polymer is that of a rope or a thread. The analogy suggests that entanglements and knots may play an important role in the behavior of polymers and indeed, their manifestations are numerous, from the viscoelasticity of polymer liquids to the disentanglement of DNA molecules in the living cell by the action of special topological enzymes (topoisomerases).

As far as the present author knows, the earliest works on knots in physics were due to W Thomson (later Lord Kelvin) [54] and J C Maxwell. They were trying to answer the question: where does the discreteness of chemical elements come from? The idea was to associate each chemical element with a certain knot, thus producing discreteness. Although we know now the quantum explanation of the nature of chemical elements, the idea remains beautiful. Very recently, various knot-related concepts have been applied in statistical physics, field theory, hydrodynamics, astrophysics and magneto-hydrodynamics, etc [55 – 58]. In these areas, however, knots appear in a rather abstract way. By contrast, knots in polymers are very obvious, they are just common knots, like ones on a rope. In the biopolymer context, the first discussions on the relevance of knots were due to H Frisch and E Wasserman [59] and M Delbrück [60].

All of what are called the topological properties of polymers stem from the simple fact that two piece of polymer chain cannot pass through one another. Imagine now that we have a ring polymer: as self-intersections are forbidden, the motions that remain possible are only those that are continuous, without breaking the polymer — and this is exactly what is discussed in mathematics when the term topology is defined. Thus, a ring polymer can arbitrarily change its geometrical shape with only the constraint of unchanged topological class.

The word topology is sometimes overused in physics (for reasons that the present author does not understand). For example, overall shape of a protein is sometimes referred to as protein topology. In the present case, we are speaking about real topology.

Clearly, topological constraints strongly affect all the properties of polymers, both static and dynamic. Many DNA molecules are known to work in the living cell in the form of closed rings, and some of them are indeed knotted. Knots have been reported recently in some proteins [61].

In mathematics, the word knot means an arbitrary closed curve embedded in 3D space. What we would normally call an unknotted ring is called a trivial knot. Similarly, a link is a number of closed loops embedded together in 3D space. A trivial link is one in which rings are not entangled. Obviously, these concepts are directly and perfectly applicable for ring polymers.

It is also worth mentioning, that a macroscopic polymer network, as realized in gels, represents an extreme of very complex topology.

3.2 Topology and disorder

The topological properties of polymers are similar to those of disordered systems in the sense that the topology is formed during the process of polymer preparation and can be then memorized. As in other cases, the difference between quenched and annealed regimes should be made very clear:

(1) Quenched topological disorder is realized in regular polymer rings and/or networks.

(2) The annealed situation can be realized in a DNA solution when a special enzyme, called topoisomerase II, is present in abundance (along with ATP molecules that are required for the functioning of topo-II).

(3) In the system of linear (open) polymers, topology does not impose strict constraints, but as rearrangements take rather a long time, temporary topological constraints can be viewed as similar to annealed disorder.

3.3 General formulation of topological problems in polymer statistics

Let us formulate now in general terms how one should approach the topological properties of polymers. Let us begin with the case of quenched topology. Suppose one wants to describe the thermodynamic equilibrium of, say, a ring polymer that has been prepared in a certain state of knot topology, trivial or non-trivial. To do so, one has to evaluate the partition function taking into account all the conformations that belong to a given topology (or, in a more sophisticated language, to the given homotopy class). Formally, this can be written as

$$Z = \int_{\text{given topology}} \exp\left(-\frac{E}{T}\right) \mathrm{d}\Gamma = \int \exp\left(-\frac{E}{T}\right) \delta\big(G(\Gamma) - G_0\big)\, \mathrm{d}\Gamma, \tag{21}$$

where integration over $\mathrm{d}\Gamma$ means summation over conformations. In the first line, summation is performed over conformations of the given topology. In the second line, this is formally transformed into integration over all conformations, with a δ function which does the job of choosing the given topology: $G(\Gamma)$ is some value which is different for conformations Γ of different topologies, and which does not change as the long as topology stays the same. This value is called the topological invariant.

Thus, the entire problem includes two steps: first, one has to find the topological invariant. In other words, one has to classify possible topologies. Second, when the topological

invariant is already known, one has to evaluate the partition function (21).

In the simplest case when there are no interactions between polymer pieces (except, of course, hard core repulsion that prohibits intersections), the latter problem is reduced to the question of the phase volume available to a given knot $\mathcal{K}$, or *knot entropy* $S(\mathcal{K})$. This question is obviously equivalent to the one of the probability of obtaining a certain knot upon random closure of a polymer, $p(\mathcal{K}) \sim \exp[S(\mathcal{K})]$.

3.4 Briefly on knot invariants

The first attempt to classify knots was due to P G Tait [77]. He drew knots and tried to uncover some regularities. More recently, considerable efforts were spent on making tables of knots (see [76,78]). In these tables, knots are organized according to the minimal number $n(\mathcal{K})$ of crossings that a given knot $\mathcal{K}$ can have on a two-dimensional projection. Unfortunately, the number of possible knots grows exponentially with n, and thus, not surprisingly, 'even the most powerful computers quickly run out of enthusiasm' [79]. Another stream of development here goes back to Alexander [80], with the ideas of algebraic topology. In recent years, there has been remarkable progress in the mathematical theory of classification of knots based on the theory of polynomial topological invariants [81]. A number of new polynomial invariants were invented in recent years, including Jones, Kauffman, Vassiliev polynomials and others [57].

Although the new polynomials are much stronger than Alexander polynomials, they are computationally incomparably less convenient, because to compute the Alexander polynomial for a knot with n intersections on the projection requires about n^3 operations, while for other more sophisticated polynomials this number grows as $\exp n$. This was thought to make a computation virtually intractable. A way to circumvent this problem has been suggested in the recent work [72].

One more approach to classify the knots is based on the following idea. Suppose we have a knotted ring polymer. Let us make it evenly charged all along the contour length and let us gradually increase the charge per unit length. Obviously, as all charges are of one sign, their repulsion will lead to stretching of the polymer. Finally, when charge is very high, the polymer will adopt some conformation that is maximally extended, compatible with the fixed topology. This maximally charged shape depends on the knot type and is hypothesized to be a topological invariant. The simplest invariant is just the minimal Coulomb energy that corresponds to this optimal shape. This later value can be written in the form

$$E = \min_{\mathbf{r}(s)} \left\{ \frac{1}{L} \oint\!\!\oint \mathrm{d}s\, \mathrm{d}s' \left[\frac{1}{|\mathbf{r}(s) - \mathbf{r}(s')|} - \frac{1}{|s - s'|} \right] \right\}, \qquad (22)$$

where s is arc length along the polymer, L is its contour length, and $\mathbf{r}(s)$ gives the polymer shape to be optimized. The second term in the integral is subtracted to remove the pathological divergence of Coulomb energy at $s \to s'$; $1/L$ is introduced to make the value independent of the linear scale, only on the topology.

There are several other constructions of 'knot energy,' but they lack an obvious physical interpretation compared to equation (22).

Very recently, another idea was suggested to classify knots according to their 'maximally inflated' [73] or 'ideal' [82] representation. We shall discuss it in more detail below, in Section 3.7.

3.5 What is known of knot entropy?

Despite the progress in mathematical theory of knot invariants and the better understanding of knot classification, there are only few rigorous results on the entropic properties of knots [62, 63], and further attempts in this direction encounter severe mathematical difficulties [64–66]. In this situation, one has to appeal to some exactly solvable models [67–69], computer simulations [70–72], or seek some simplified Flory-type approach [73]. As to the exactly solvable models, such as a winding around a point-like obstacle on a plane or lattice of obstacles, they are reviewed, for example, in [74, 75]. An excellent review on computer simulations was published in UFN several years ago [76]. This is why only recent simulation results are summarized here and then a simple Flory theory is discussed.

3.5.1 For long chains, the probability of unknotting decreases exponentially with the number of segments. There are very few well established results relating to knot entropy. One of them is due to the works [62, 63]. Consider a broken line of N segments, each of unit length. This is obviously a model of a freely-jointed polymer without excluded volume. What is the probability, $\mathcal{P}_0(N)$, that it will form a trivial knot ($\equiv$ an unknot) upon random closure? The result of [62, 63] is that this probability tends to zero exponentially with N:

$$\mathcal{P}_0(N) \propto \exp\left(-\frac{N}{N_0}\right), \qquad (23)$$

where N_0 is the characteristic scale (it is just a number)†.

While rigorous proof of relation (23) is by no means simple, it is easy to gain an insight into why probability (23) varies exponentially. Indeed, let us divide the entire polymer of N segments into N/g blobs of g segments each. To guarantee an unknotted conformation of the entire chain, one has to make all the blobs unknotted, thus yielding

$$\mathcal{P}_0(N) \leqslant \left[\mathcal{P}_0(g)\right]^{N/g},$$

which leads to exponential behavior in N.

An important consequence of result (23) is that a normal Gaussian model of a polymer chain, that represents chain as a trajectory of Brownian motion, is totally inappropriate for topological problems. Indeed, a Brownian trajectory can be viewed as the limit of a broken line with the segment length tending to zero, $l \to 0$, and the number of segments tending to infinity, $N \to \infty$, in such a way that the contour length of the polymer remains constant: $L = Nl = \text{const}$. Naturally, the knotting probability depends on N, not on L, and this is why it tends to unity in the Brownian limit. This is exactly the 'ultraviolet catastrophe' that is known for simplified models, this is totally due to indefinite knotting of Brownian trajectories on small scales. Thus, a certain sense of granularity, either in the segment length, or in the chain width, or in the lattice character of the underlying space, is necessary for a meaningful approach to the problem of knot entropy.

What is the value of N_0? There are no methods even to estimate it other than to resort to computer models and simulations.

† More precisely, the statement of the works [62, 63] reads: there exists finite limit $\lim_{N\to\infty}\left[N^{-1}\ln \mathcal{P}_0(N)\right]$.

3.5.2 Some Monte Carlo data. There are a number of works on Monte Carlo generation of knots, from [70] through [71] to the most recent [72] (see also the very nice review in [76]). The general scheme is as follows:

(1) First, one has to generate closed loop. This can be done on different lattices, or with various off-lattice models, for instance, consisting of straight segments. To make them closed loops, there is simple prescription to generate next step of the walk from the *conditional* probability distribution, with the condition imposed that the trajectory as a whole represents a loop (or Brownian bridge).

(2) Second, when the loop is generated, one has to decide which type of knot is present. This is the most computationally demanding part, and in most cases it is done using the Alexander polynomial topological invariant. In the recent more advanced work [72], sophisticated Vassiliev polynomial invariants were employed instead of Alexander ones.

Some of the data are summarized in Fig. 3. In earlier works, these data were presented for the probability $\mathcal{P}(N) = 1 - \mathcal{P}_0(N)$ obtaining a non-trivial knot. This is because originally people were thinking of knots as rare events, while theorem (23) shows that they must be typical in the extreme of long chains. The data presented in Fig. 3 agree very well with the asymptotic prediction (23), and yield the estimate $N_0 \approx 335 \pm 5$.

The characteristic chain length appears rather long! It remains unknown why it is so long. It is tempting to compare this number with another, appearing in the reptation theory [2, 3] as the entanglement number N_e, which is usually about 50 – 500. The relation between these two numbers, however, remains a mystery.

Further Monte Carlo studies include the determination of probabilities for various particular non-trivial knots [72]. Not surprisingly, when the chain length, or better to say — the number of segments, becomes larger, the diversity of knots obtained grows as well. As to the probability of finding any particular type of knot, say, $\mathcal{K}$, upon random closure of a loop of N segments, $\mathcal{P}_\mathcal{K}(N)$, it can be guessed to behave as follows:

(1) For small N, $\mathcal{P}_\mathcal{K}(N) \simeq 0$, unless $\mathcal{K} = 0$ is the trivial knot, because short loop cannot form any knot but the trivial one;

(2) For very large N, $\mathcal{P}_\mathcal{K}(N)$ should decay exponentially, with the same characteristic length N_0 that was found for the trivial knot [see (23)]. Indeed, for long enough chains, with growing N, any knot $\mathcal{K}$ becomes finally 'small' and 'local.' For that large N, the probability $\mathcal{P}_\mathcal{K}(N)$ can be thought of as the probability of forming the given knot $\mathcal{K}$ on some relatively small part of the chain N_1 times the probability $\mathcal{P}_0(N - N_1)$ that the rest of the chain remains unknotted. In the large N limit, this later factor dominates, thus yielding exponential decay on the scale N_0 independent of the knot $\mathcal{K}$;

(3) Given the two arguments above, we conclude that $\mathcal{P}_\mathcal{K}(N)$ should have a maximum at some particular chain length $N_\mathcal{K}$, which depends on the type of knot. It is reasonable to assume, that $N_\mathcal{K}$ increases when the knot $\mathcal{K}$ gets more complex.

In the work [72], these statements were beautifully confirmed. It was found that the following relationship is valid within the accuracy of the data:

$$\mathcal{P}_\mathcal{K}(N) = C_\mathcal{K} N^{\nu(\mathcal{K})} \exp\left(-\frac{N}{N_0}\right). \qquad (24)$$

Some numerical data of the work [72] are shown in Table 1.

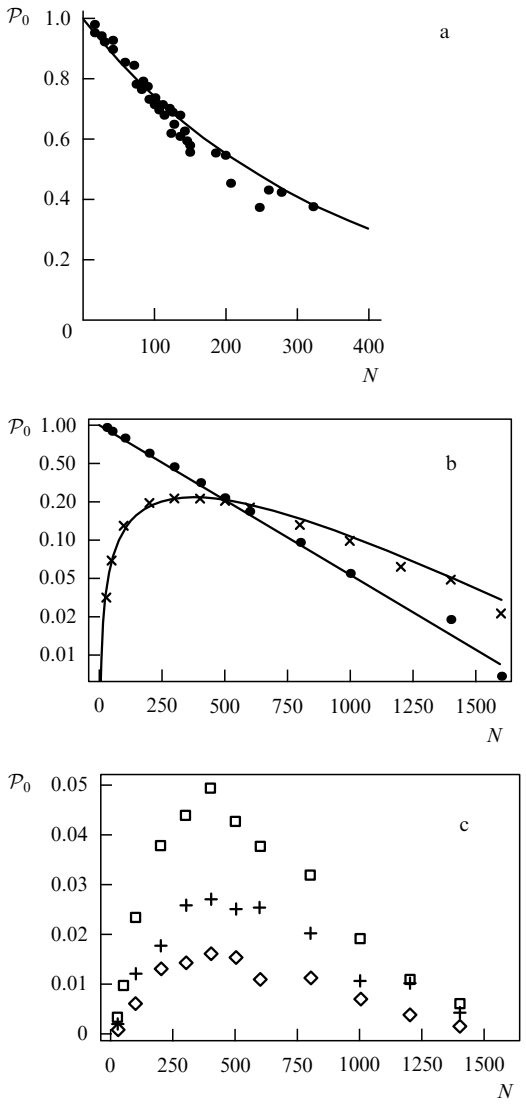

Figure 3. Monte Carlo data for the probabilities of forming trivial and some non-trivial knots upon random closure of the polymer of N segments of zero width: (a) represents the data of [76] for the trivial knot, along with the fitting curve $\exp(-N/335)$. Figures B (in logarithmic scale) and C (in linear scale) present data of the work [72]. Fitting curves are shown according to equation (24). Designations: • - for the trivial knot, × sign for trefoil 3_1, □ for 4_1, + for 5_1, ◇ for 5_2.

It was hypothesized based on numerical data [72], that for a composite knot $\mathcal{K} = \mathcal{K}_1 \# \mathcal{K}_2$, the relationship $\nu_\mathcal{K} = \nu_{\mathcal{K}_1} + \nu_{\mathcal{K}_2}$ is valid. As far as this author knows, nobody has been able to prove it so far.

3.5.3 The role of polymer width. Although this is not proven mathematically, there is little doubt that almost all conformations of the (infinitely) long polymer ring are knotted, whatever the width of segments. The width, however, can and does strongly influence the characteristic length at which knotting becomes most probable.

As to polymers of moderate length, width appears to suppress dramatically their ability to form non-trivial knots. It was found in Monte Carlo simulations performed on a model of freely jointed segments of length l and width d [83]. Obviously, the probability of knotting, $\mathcal{P}$, or the probability of the trivial knot, $\mathcal{P}_0 = 1 - \mathcal{P}$, depends now on both $N = L/l$

Table 1. Numerical data for some knots.

Knot type	$v_{\mathcal{K}}$	$C_{\mathcal{K}}$	$N_{\max} = N_0 v_{\mathcal{K}}$	$\mathcal{P}_{\max}$
0	≈ 0	1.03 ± 0.03	≈ 0	1
3_1	1.11 ± 0.05	$(1.1 \pm 0.3) \times 10^{-3}$	372 ± 15	0.26
4_1	1.34 ± 0.09	$(6.6 \pm 3.0) \times 10^{-5}$	450 ± 30	0.06
5_1	1.35 ± 0.15	$(1.9 \pm 1.5) \times 10^{-5}$	450 ± 50	0.02
5_2	1.40 ± 0.11	$(2.2 \pm 1.3) \times 10^{-5}$	470 ± 30	0.03
$3_1\#3_1$	2.3 ± 0.1	$(2.9 \pm 2.1) \times 10^{-7}$	770 ± 30	0.13
$3_1\#4_1$	2.5 ± 0.2	$(3.4 \pm 4.2) \times 10^{-8}$	840 ± 70	0.06
$3_1\#3_1\#3_1$	3.7 ± 0.2	$(1.5 \pm 1.5) \times 10^{-11}$	1240 ± 70	0.1

and d/l: $\mathcal{P}_0 = \mathcal{P}_0(N, d/l)$. It appears that Monte Carlo data fit reasonably well to the following formula:

$$\mathcal{P}_0\left(N, \frac{d}{l}\right) \approx \exp\left[-\frac{N}{N_0 \exp(27d/l)}\right]. \tag{25}$$

As we see, the chain width makes the characteristic length grow dramatically over its already large value of about $N_0 \approx 335$. For the given number of segments, this means strong suppression of knotting.

The fact that polymer width suppresses knotting is well established, but it is not very well understood. There are two effects of potential importance. One of them is that chain width is nothing but excluded volume, and it leads to overall swelling of polymer coil. Another effect is that width affects the ability of the growing polymer chain to penetrate and go through the small loops of the already formed part of the polymer. To separate these two effects, one can examine computationally a model where the chain does not have width, but is swollen due to an appropriate external field (say, of the form $\varphi(x) \sim -\kappa x^2$, where one chain link is fixed at the origin). While some studies in this direction have been performed, the data are inconclusive.

3.5.4 Experimental observation of knots. There is a rather long history of observation of knots in DNA, it goes back to 1976 [84]. The recent achievement of the works [85, 86] is the possibility to measure probabilities for some particular knots, such as 3_1, 4_1, 5_1, and 5_2. It was decisively important for the comparison of experimental and Monte Carlo data that the non-zero width of real polymers is taken into account in the latter, because in experiments [85, 86], by changing the ionic strength of the solution, it was possible to control the effective diameter of DNA double helices. The experimental dependence of the probabilities for the knots 3_1, 4_1, 5_1, and 5_2 on the ionic strength appears to be in a very good agreement with the Monte Carlo data for the same probabilities as functions of chain diameter, d (see [78] for the standard designations of knots).

3.5.5 The role of polymer compression. In normal polymer physics, the excluded volume effect leads to polymer swelling, and there is also the opposite situation, where the excluded volume can be said to be negative ($B < 0$), and that leads to chain collapse. As far as the topology is concerned, we cannot say that the polymer width is zero or negative even when the solvent is poor and the polymer tends to collapse. Indeed, it is the second virial coefficient that becomes zero at Θ temperature and negative below Θ, but not the geometrical width of the polymer. Thus, one has to ask what is the probability of knotting for a polymer that is maintained in a collapsed globular state, either due to attractive interactions present on top of non-zero width, or due to compression in a restricted geometry or an external field.

Monte Carlo simulations [76, 87] indicate that collapse strongly enhances the appearance of non-trivial knots. It appears, that the probability of trivial knots for globular polymers goes to zero much faster than it does for the free random coil.

Although the strong knotting in the globular state is very well seen in the data, it is difficult to extract some quantitative measure or interpolation expression for the knotting probability. Unfortunately, the problem also appears to be very difficult for an analytic approach, and there has been only one attempt to estimate this probability theoretically [88]. As usual in the mien-field theory of globules, what one does is determine the change of the polymer properties due to collapse. In this case, the estimate of the work [88] gives an additional factor that suppresses the probability of trivial knot:

$$\mathcal{P}^{\text{glob}}(R) \sim \exp\left[-\frac{Nl^2}{R^2}\right] \sim \exp\left[-N^{1/3}\left(\frac{l}{d}\right)^{4/3}\right], \tag{26}$$

where R is the size of the polymer localization, and the latter estimate is given for a maximally compact polymer with R^3 of the order of polymer's own volume, Nld^2. It is unclear, however, whether one can write the estimate for the resulting probability as simply the product $\mathcal{P}_0(N, d/l)\mathcal{P}^{\text{glob}}(R)$ In any case, it appears that even for a rather moderate chain length, almost all of the compact chain conformations are heavily knotted. This fact has important consequences.

3.6 Crumpled globules

3.6.1 Collapse of an 'underknotted' polymer. Surprisingly, the abundance of knots in the collapsed state appears to be of special importance for those polymer systems where formation of knots, for one reason or another, is suppressed or prohibited. There are actually many such systems:

(1) subchains of a polymer network;

(2) the solution of untangled polymer rings;

(3) the long (open) linear chain in the initial stages of collapse, before the ends could penetrate the globule and form an equilibrium quantity of knots.

Thus the question appears: what is the structure of a collapsed, but unknotted (or 'underknotted') polymer? The answer was suggested in [89]. According to the estimates of that work, collapsed unknotted polymers adopt a conformation that is crumpled in the sense that it does not obey Flory theorem. If one takes a piece of polymer of, say, k monomers, then this piece appears to be collapsed, with a size of about $k^{1/3}$. This is in a sharp contrast with a regular globule, where the size of a k-monomer piece of the chain varies as $k^{1/2}$ as long as k is less than one chain 'passage' through the globule,

and becomes k-independent at larger k. This is illustrated schematically in Fig. 4. This is a rather strong statement: although the absence of knots means some global condition, it is stated that this global condition leads to selection of trajectories with very peculiar local fractal properties.

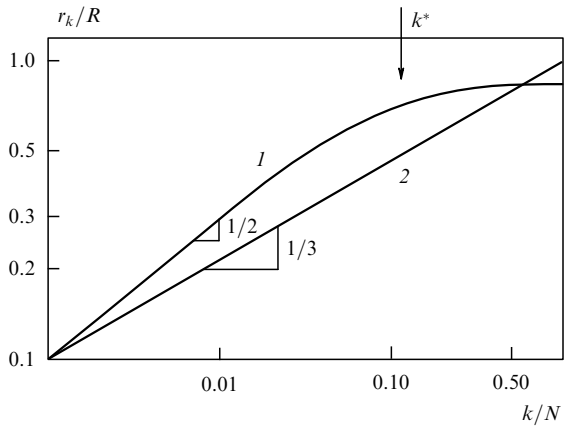

Figure 4. Schematic Log-Log plot of the size of k-monomers chain pieces as a function of k in a regular globule (A) and in a crumpled globule (B). The characteristic scale k^* for the regular globule is defined in association with one Gaussian passage through the globule core, that is, $a(k^*)^{1/2} \sim R \sim (N/n_0)^{1/3}$, or $k^* \sim (N/a^3 n_0)^{2/3}$, where n_0 is the density in the globule. For the maximally compact polymer, $n_0 \simeq 1/v$.

Since it was suggested in 1987, the concept of a crumpled globule has been neither confirmed nor rejected. Recently, some indirect evidence was obtained in favor of this concept both from computer [90, 91] and from real experiments [92, 93], as well as theoretically [94, 95]. Nevertheless, it remains at the status of hypothesis.

3.6.2 The crumpled globule as a model of native DNA. Simple packing considerations tell us that the native spatial structure of DNA has to be of a dense globular type, rather than that of an expanded coil. It was also mentioned that spatial 3D structures of globules, either equilibrium or not quite equilibrium ones, are controlled by volume interactions between chain monomers. As to the native DNA globule, these volume interactions are of tremendous complexity, since they are mediated by proteins and include phenomena such as the recognition of particular sequences by proteins, etc. Nevertheless, the conclusion that most of the dense conformations are heavily knotted holds independently of that, and thus should be valid for DNA.

On the other hand, it seems that complex knotted conformations cannot dominate the native state of a functioning biopolymer since entanglements will dramatically reduce its ability to respond to bio-chemical influences. Indeed, if the number of entanglements in the globular structure of a high molecular weight polymer is comparable to the number of segments, the structure will become glass-like (i.e., kinetically frozen), with the result that many monomer units will be out of reach for any biological system involved in DNA processing. This is why it was assumed in work [96] that, *in a statistical sense, the DNA globule is practically unknotted*. The argument was also given that this conclusion holds in spite of the existence of topoisomerases and other proteins which can cut DNA; being small compared to DNA dimensions they cannot even recognize the global topology of DNA and thus they can not have a statistically significant effect on the number of entanglements in the globule.

The crumpled globule model of DNA native structure explains naturally the observed hierarchy of structural levels of DNA spatial organization, starting from the double helix, to nucleosomes, and all the way up to chromatin, as these levels are associated with self-similar crumples of different scales.

An alternative view point was presented in the works [97, 98], where a model of phantom DNA was suggested. The idea was that the topoisomerases could be present in abundance, along with a practically unlimited supply of ATP. If that were the case, intersections of DNA would occur easily, and DNA would indeed behave effectively like an phantom polymer. Then, in particular, it could have the *in vivo* conformation of a usual globule, instead of a crumpled globule.

The dispute remains unresolved, and it is for an experiment to decide which model is closer to reality.

3.6.3 Speculation about proteins. Peculiar fractal properties of crumpled globules gave rise to various speculations dealing both with suspected fractal properties of protein chains and attempts to explain the appearance of secondary structure. These ideas will not be discussed here in further detail.

3.7 Knot inflation†

3.7.1 The topological invariant: maximally inflated knot representation. To build up a simple theory, the following construction was suggested [73, 82]. Consider a polymer chain in some spatial conformation and denote by L the contour length of the chain. Let us first construct a tube that contains the polymer chain and is sufficiently narrow such that the topology of the tube as a whole is the same as that of the polymer. We now inflate the tube such that its length L is preserved, while its cross-section is roughly the same everywhere along the tube (we assume that different tube portions cannot penetrate each other). This inflation will eventually end when the inflated tube fills the main part of the volume within its loops. Let us denote by D the diameter of the maximally inflated tube. We say that the aspect ratio of the maximally inflated tube

$$p \equiv \frac{L}{D} \tag{27}$$

is a topological invariant, albeit it rather weak (since there may be many topologically different knots that have the same value of p)‡. Nevertheless, p is a topological invariant, in the sense that if we take two closed curves in three dimensional space that are geometrically different but identical with regard to their topology, then tube inflation, as described above and as illustrated in Fig. 5 , will work independently of initial geometries and will result in identical ('maximally inflated') geometrical shapes of the corresponding tubes, and therefore, will yield identical p values. This happens because the redistribution of the 'stored' length among the

† This section is based on work [73].

‡ There are some inherent uncertainties in the definition of p, related to the definition of the 'diameter' of a sharply curved tube, as well as assumptions made about the flexibility of the central tube axis during inflation. For example, if one attempts to approximate this axis with a broken line of some n straight segments, the value of p will be (slightly) dependent on n. This is why p is indeed a rather weak topological invariant. However, in practical terms, p appears to be jairly good invariant [82].

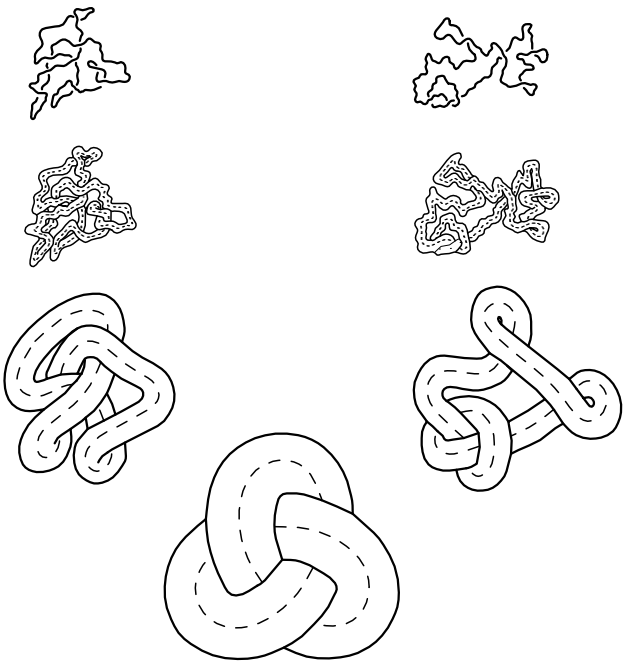

Figure 5. Illustration of tube inflation. Two different yet topologically equivalent loop conformations are shown. The thin tube centered around the polymer closely resembles the conformation of the polymer itself. When the tube is inflated, with the contour length of its axis preserved, all the small scale 'jigglings' of the polymer conformation are gradually eliminated, and finally we arrive at the maximally inflated shape that is independent of the initial conformation, but only on the knot topology. Courtesy of G Berritz.

loops is unrestricted in this single linear chain problem, and thus inflation does not encounter spin-glass-type frustrations and leads to some well defined optimum. A closely-related definition of a topological invariant was introduced in reference [58], in the context of vortex tubes in fluid mechanics.

If the ring is not knotted in the conventional sense, or if it forms the trivial knot, its inflation leads to a simple torus with $D \sim L$ and thus $p \simeq 1$. On the other hand, the more complex the knot we have, the less is its 'inflation capability'. Physically, since real polymers have a finite thickness, there is a maximal knot complexity that can be achieved which corresponds to a knot that is so dense that any inflation over an already existing diameter would be impossible. If we imagine a polymer chain made of monomers whose length is equal to their thickness, we conclude that the maximal value of $p = L/D \simeq N$ is of the order the number of monomers per polymer chain. Thus, our topological invariant can take values in the interval

$$1 \leqslant p \leqslant N, \tag{28}$$

and it provides a rough measure of knot complexity: more complex knots correspond generally to higher p-values. To illustrate the later conjecture, it is worth mentioning that the topological invariant p has the property of additivity: for the composite knot $\mathcal{K} = \mathcal{K}_1 \# \mathcal{K}_2$, the topological invariant is given by

$$p(\mathcal{K}_1 \# \mathcal{K}_2) = p(\mathcal{K}_1) + p(\mathcal{K}_2)\,. \tag{29}$$

We note that the topological invariant p is closely related to the primitive path of the chain in a lattice of obstacles [65]. In order to clarify this point we introduce a self-consistent representation of a complex knot in terms of an 'effective' lattice of obstacles. The polymer trajectory on this lattice is represented by a primitive path which is measured in units of the lattice constant (this makes the topological invariant independent of lattice deformation). Maximal inflation is equivalent to lattice enlargement up to the point when the polymer chain becomes completely stretched along the primitive path, and therefore p can be interpreted as the chain length measured in units of the expanded lattice constant. The above analogy allows one to estimate the number of topologically different knots with a given p value. Since a lattice of obstacles can be mapped onto the Cayley tree [69], this quantity should grow exponentially with p:

$$K(p) \sim \exp(\lambda p)\,, \tag{30}$$

where λ is some numerical constant.

3.7.2 Flory-type theory for the swelling and collapse of a knot. We proceed to estimate the polymer chain size dependence on solvent quality and topology, the latter represented by the topological invariant p. We begin with the assumption that the chain conformation can be characterized with a single length scale R. Our goal is to find the equilibrium value of R for the given solvent conditions and the given p. Following the classical Flory approach, the equilibrium polymer size R is given by the balance of rubber-like elasticity and interactions between monomers dispersed in the polymer volume. This is described by the minimization of the free energy

$$F = F_{\text{elast}} + F_{\text{interact}}\,. \tag{31}$$

Let us assume first that all monomers are more or less evenly distributed over the volume of the system R^3 (actually, we shall show later that this assumption is not universally valid). For the uniformly smeared cloud of monomers, the interaction term for the ring is identical to that of the linear chain,

$$\frac{F_{\text{interact}}}{T} \simeq \frac{BN^2}{R^3} + \frac{CN^3}{R^6} \simeq \frac{B}{a^3}\frac{\sqrt{N}}{\alpha^3} + \frac{C}{a^6}\frac{1}{\alpha^6}\,, \tag{32}$$

where $\alpha = R/a\sqrt{N}$, B and C are the second and the third virial coefficients, respectively, and a is the monomer size.

The problem is to obtain a plausible estimate for the entropic part which would account for the frozen topology of the polymer ring. To this end we suggest the following approximation based on the construction of the maximally inflated tube which occupies a volume of order LD^2. Let us deform it affinely so that it occupies the volume of the chain R^3, but preserves the geometrical shape that it has in the maximally inflated state, and call this deformed tube an 'R-size tube'. Let L_R and D_R be the length and the diameter of the R-size tube, respectively. Since the R-size tube is obtained by an affine transformation of the maximally inflated tube, we have $L_R/D_R = p$, and since it occupies the whole volume of the polymer, $L_R D_R^2 = R^3$. We obtain:

$$L_R \simeq Rp^{2/3}\,, \qquad D_R \simeq Rp^{-1/3}\,. \tag{33}$$

Now the central assumption comes: in order to estimate the entropic (i.e., elastic) free energy, one can consider our polymer as a phantom chain, but confined within the R-size tube. The evidence in favor of this assumption was obtained

by computer simulation [82]. The entropy of a phantom polymer confined in a tube is independent of the way this tube is embedded in $3D$ space; one can estimate the entropy of a phantom polymer in a simple torus-shaped tube, or even that of a linear polymer in a straight tube, with polymer ends attached to the tube ends. This gives

$$\frac{F_{\text{elast}}}{T} \simeq \frac{L_R^2}{Na^2} + \frac{Na^2}{D_R^2} = \alpha^2 p^{4/3} + \alpha^{-2} p^{2/3}\,, \tag{34}$$

where the first and the second term describe the chain elongation along the tube and the squeezing within the tube diameter, respectively (see, for example, [32]). We balance the free energy contributions (32) and (34) and obtain the following equation for α:

$$\alpha^5 p^{4/3} - \alpha p^{2/3} - \frac{B}{a^3}\sqrt{N} - \frac{C}{a^6}\frac{1}{\alpha^3} = 0\,. \tag{35}$$

This equation is similar to the equation [see [32], Eqm (13.5)] for linear or phantom polymers, except for the inclusion of p-dependent factors. This is a good news: it means that the 'correspondence principle' is obeyed, at $p = 1$ we are back at the 'old' theory. For simplicity, in the following we present results in terms of polymer chain size $R = \alpha a\sqrt{N}$ and take $B \simeq a^3\tau$ and $C \simeq a^6$, where τ denotes the dimensionless deviation from the Θ-temperature. Eqn (35) implies the following regimes:

Good solvent regime is realized when $\tau > (p/N)^{1/2}$; in this regime entropic elasticity associated with chain elongation along the tube [first term in eq (35)] competes with two-body-repulsion (third term), yielding

$$R \sim aN^{3/5}\tau^{1/5}p^{-4/15}\,. \tag{36}$$

The N and temperature–dependence of polymer size ($R \sim N^{3/5}$) is identical to the linear or phantom polymer case, but there is an important prefactor that gets smaller for complex knots with large p values. Note that the chain size in this regime can be smaller than the size of a Gaussian phantom chain, $aN^{1/2}$; this happens in the range $p^{1/2}N^{-1/2} < \tau < p^{4/3}N^{-1/2}$, or, in other words, for sufficiently complex knots with $p > (\tau N^{1/2})^{3/4}$. Only a relatively simple knot in a truly good solvent $\left[p < (\tau N^{1/2})^{3/4}\right]$ is swollen compared to the Gaussian phantom size.

At smaller τ or higher p (more complex knots) the polymer crosses over to the quasi-Gaussian regime.

The quasi-Gaussian regime arises when $-(p/N)^{1/2} < \tau < (p/N)^{1/2}$; in this regime the elasticity associated with chain elongation along the tube competes with the three-body repulsion and the chain elasticity across the tube, the latter two being of the same order of magnitude. This gives

$$R \sim aN^{1/2}p^{-1/6}\,, \tag{37}$$

i.e., the N-dependence of the chain size remains of the Gaussian type, but the coefficient gets smaller for more complex knots.

At even smaller (more negative) τ or for smaller p (simpler knots) the polymer crosses over to the poor solvent regime.

The poor solvent regime is realized when $\tau < -(p/N)^{1/2}$; in this regime the two-body attractive term in Eqn (35) competes with the three-body repulsion, yielding

$$R \simeq a|\tau|^{-1/3}N^{1/3}\left[1 + |\tau|^{-4/3}\left(\frac{p}{N}\right)^{2/3}\right]. \tag{38}$$

The main term here is an obvious result, because in the presence of strong inter–monomer attraction the polymer must collapse into a dense sphere (globule) with N-independent density. We included the correction due to the next most important term, which is chain compression in the tube. This indicates that a heavily knotted globule is less compact compared to its phantom counterpart.

The maximally tightened knot regime is realized when $p \sim N$ and, interestingly, it does not depend on solvent quality and interactions. In this regime,

$$R \simeq aN^{1/3}\,. \tag{39}$$

and, thus, a tightly knotted ring is always compact.

Our results for the scaling exponents, (36), (37) agree with the computational data of reference [99], where ring sizes, $R(\mathcal{K})$, were studied for several different ring topologies $\mathcal{K}$, including the unknot, trefoil, figure eight, and double trefoil. It was found that critical indices $\nu(\mathcal{K})$ in $R(\mathcal{K}) \sim N^{\nu(\mathcal{K})}$ are the same (within statistical errors) for all tested knots $\mathcal{K}$, namely, close to 0.6 or to 0.5 for rings with or without excluded volume, respectively.

The simplest theory presented above does not distinguish between a trivial knot ($p = 1$) and a phantom ring. This defect can be fixed by considering the confinement in the R-size tube of a non-phantom unknotted polymer instead of a phantom one. The exclusion of knots for a chain confined within an R-size tube gives rise to an additional term in the elastic free energy (34) that scales as α^{-6}. Indeed, expression (34) has to be modified in the case when the chain is compressed both across and along the tube; in the latter case, the scaling form of the free energy can be obtained from the fact that the resulting osmotic pressure must depend on N and the tube volume $L_R D_R^2 \sim R^3$ only through the polymer density, N/R^3. This leads only to the redefinition of the C/a^6 coefficient in equation (35) and does not affect the scaling forms of all our main results (36)–(39). More subtle corrections may be needed if one is to incorporate delicate properties of the collapsed state [100] instead of a simple crumpled globule scheme.

Let us return now to our assumption of the even distribution of monomers in the volume R^3 and discuss once again the good solvent regime. Suppose that the chain adopts a conformation in which a part of the length of about p monomers forms a dense region where all of the knots are located, and another part of $N - p$ monomers which swells freely in the good solvent. As long as $p \ll N$, our theory gives for this 'phase segregated' state a free energy of about p for the 'collapsed' knotted part plus about $(N-p)^{1/5} \approx N^{1/5}$ for the freely swollen loop part, yielding about $N^{1/5}$ in total (assuming $p \ll N^{1/5}$). On the other hand, a state with a uniform distribution of knots gives, after substitution of Eqn (36) into Eqns (32) and (34), a much higher free energy, of the order $N^{1/5}p^{4/5}$ ($p \gg 1$). Thus, while for $N^{1/5} \ll p \ll N$, thermodynamics favors a uniform distribution of knots along the chain contour, our theory predicts that segregation of knots will take place for less knotted chains, with $p \ll N^{1/5}$. Of course, if the topological state of the polymer represents a composite knot, then simple knot-components will diffuse independently from each other along the chain contour. In this sense, knot segregation gives rise to a picture that is very similar to the idea of 'local knots,' suggested in [101].

It would be interesting to test the prediction of knot segregation by computer simulations and experiments.

3.7.3 Probabilities for knots. It is tempting to relate the free energy F_{elast} (34) to the probability distribution of the knots that form in the process of formation a ring by random contacts between the ends of a linear polymer (assuming that the ends remain glued upon contact). An identical distribution is obtained from the collection of the instantaneous configurations of a phantom (freely passing through itself) ring. Indeed, F_{elast} is determined by the volume in configuration space which is available to the non-phantom polymer with a given quenched knot topology, and this volume is obviously proportional to the probability of getting this same knot topology in a phantom system that goes freely from one topology to another. Thus, the (normalized) probability of obtaining a knot with a given p value can be written as

$$\mathcal{P}_p(R) = Q_p(R)\left[\int_1^N \mathrm{d}p Q_p(R)\right]^{-1},$$

$$Q_p(R) = K(p)\exp\left[-\frac{F_{\text{elast}}(R)}{T}\right]. \tag{40}$$

Unfortunately, one cannot directly plug in here expression (34) for $F_{\text{elast}}(R)$, because we are already aware of the segregation of knots for the swollen polymer. To make a simple estimate, we can say that for a swollen polymer, knots will be totally segregated, and thus the length p will not be able to swell; in other words, the effective length of the polymer will be reduced to $N-p$, and thus for the swollen regime $R > aN^{1/2}$ we have to use

$$F_{\text{elast}}(R) \simeq \frac{TR^2}{a^2(N-p)}.$$

This yields

$$Q_p(R) \sim \begin{cases} \exp\left[\lambda p - \dfrac{R^2}{a^2(N-p)}\right], & R \gg a\sqrt{N}, \\ \exp\left[\lambda p - \alpha^2 p^{4/3} - \alpha^{-2}p^{2/3}\right], & R \ll a\sqrt{N}, \end{cases} \tag{41}$$

where we have used estimate (30) for the number $K(p)$ of different knots with a given p.

Inspection of equation (41) indicates first of all that for each polymer size R, there is an optimal quantity of knots, where the probability distribution (40) peaks (if $\lambda > 4\sqrt{2}/3 \approx 1.9$):

$$p_{\text{opt}}(R) \sim \begin{cases} N - \dfrac{R}{a\sqrt{\lambda}}, & R \gg a\sqrt{N}, \\ N\left(\dfrac{Na^3}{R^3}\right)^2, & R \ll a\sqrt{N}. \end{cases} \tag{42}$$

Thus, most likely degree of knotting grows with chain compression and reaches its maximal value $p \sim N$ for a maximally compact globule with $R \sim aN^{1/3}$.

If we now resort to the saddle point approximation to evaluate the normalization integral in equation (40) (which yields, of course, simply $Q_{p_{\text{opt}}}(R)$), then we obtain the leading terms in the form

$$\mathcal{P}_p(R) \propto \begin{cases} \exp\left[\lambda(p-N)\right], & R \gg a\sqrt{N}, \\ \exp\left[-p^{2/3}\dfrac{Na^2}{R^2} - \lambda N\left(\dfrac{Na^3}{R^3}\right)^2\right], & R \ll a\sqrt{N}. \end{cases} \tag{43}$$

This result agrees very well with equation (23), as it yields an exponential decay of the trivial knot probability with N. Moreover, this sheds light on a possible relation between $N_0 \approx 335$ and λ. Furthermore, our result agrees with the data of numerical analysis (24) in the large N region (we could not conceivably pretend to get power law corrections to the main exponential term). Finally, we have also reproduced exactly estimate (26).

3.8 Tight knots

Knots formed by usual ropes or threads are easily tighten. This can be unpleasant when you are fishing, or can be helpful for your shoes. It was hypothesized by P G de Gennes in [102] that the same can potentially happen in polymers: if stored length is pulled out of the knot region on the polymer chain, then the reptation relaxation of this knot is completely suppressed, and relaxation takes very long time. It was also suggested that tight knots could potentially be of importance for some condensed polymer systems, such as crystals.

How small can a completely tightened polymeric knot be? It depends, of course, on local chemistry. The stereo model [102] shows that the minimal number of monomers necessary to make a tight knot in a typical flexible polymer is about 33. Interestingly, this number is very close to the length of shortest walk that can be knotted on a three dimensional cubic lattice; this latter number was first found by M Delbrück [60] to be 27.

In the recent work [103], the possibility was suggested to realize tight knots in the course of the decollapse of polymer globules. Indeed, if we begin with a globular polymer and let it settle for a while, it will adopt a conformation with an abundance of knots. If we now quench the polymer in a good solvent condition, the knots will not have time to diffuse from the polymer through its ends and will appear tightened and effectively frozen in the chain. It would be interesting to test this prediction experimentally.

4. Freezing transition of globular heteropolymers with random sequence†

4.1 Globular heteropolymers

From now on, we switch to heteropolymers, that is, to the polymers where disorder is present in the form of a sequence of chemically different monomers connected into a single chain. We shall concentrate on the single chain problem. While considering various simplified models, we shall keep in mind the protein folding problem as a sort of super-goal.

In this section, we consider heteropolymer chains which are in the maximally compact, globular state. This means, in particular, that the density of the globule cannot fluctuate and is evenly distributed in space. Thus, the volume approximation [1] is applicable. In the simplest lattice model case, a polymer of N monomers occupies a region with exactly N lattice sites and therefore, visits every site once and only once (for instance, a polymer of 27 monomers occupies the $3 \times 3 \times 3$ region in the cubic lattice). The Hamiltonian that envelops all three important ingredients of the problem, namely, the sequence of the certain set of monomer species,

† The material of this section, as well as Sections 5.1.5 and 5.2 is based on the work [10].

arbitrary interactions between them, and conformations, has the following form:

$$\mathcal{H}(\text{seq}, \text{conf}) = \sum_{I,J}^{N} B_{s_I s_J} \Delta(\mathbf{r}_I - \mathbf{r}_J) , \qquad (44)$$

where capital Latin indices count the monomers along the chain, $s_I \in \{1, \ldots, q\}$ is the *species* of monomer I along the chain (and, thus $\{s_I\}$ represent 'sequence'), q is the number of species, and $\mathbf{r}_I$ is the *position* of monomer I ($\{\mathbf{r}_I\}$ represent 'conformations'). $\Delta(r)$ is a function concentrated on the nearest neighboring points in space; on the lattice, $\Delta(a) = 1$ and $\Delta(r > a) = 0$, where a is the lattice spacing. Thus, our model simply says that the energy of a polymer conformation is determined by the matrix of species-species energies B_{ij} for the monomers in contact. In writing the energy in the form (44), we implicitly assume that the conditions of chain connectivity (points $\mathbf{r}_I$ and $\mathbf{r}_{I+1}$ are always next to each other in space for all I), excluded volume ($\mathbf{r}_I \neq \mathbf{r}_J$ for $I \neq J$), and dense packing are all met.

Clearly equation (44), however general, is still an approximation. For example, one could also consider heteropolymeric three body interactions, which depend on the species of the three monomers in contact, etc. Nevertheless, it does include many essential components of the problem and we shall restrict ourselves to this model. Also, one has to keep in mind that the 'monomers' of the Hamiltonian (44) are really renormalized 'quasi-monomers' [32]. This means, in particular, that many small scale details of reality are coarse grained out in our model. For example, we cannot even attempt with this model to approach the low temperature properties of proteins, where another glass type transition is known to occur due to the freezing of small vibrations around the native conformations [104, 105].

Nevertheless, we believe that our model is sufficient to understand the large scale properties of protein globules, including the particular phenomenon of unique folding. We shall restrict our discussion to this model.

As to the interaction matrix B_{ij}, natural proteins include $q = 20$ species of monomers, and thus, the B_{ij} matrix should be 20×20. Neither the values of its matrix elements nor of models with smaller number of species are agreed on among experts. The so-called 20×20 MJ matrix is extracted from the statistics of proteins database [109]. There have also been other attempts to derive realistic amino acid interaction energies [110]. Actually, even the very procedure of extraction of the energies from the statistics of protein structure is conceptually neither simple nor reliable (see, e.g. [106–108]). Clearly, these energies cannot be perfect, as the very idea of a 'contact' is somewhat approximate or semi-qualitative for such bulky molecules as amino acids. On the opposite extreme, as hydrophobicity is believed to be the main driving force of protein collapse, various models are used with just two monomeric species, hydrophobic and polar. Somewhat special is Independent Interaction Model (IIM) [9], where the number of monomer species is as large as the total number of monomers, such that each matrix element B_{ij} enters in the energy of any conformation never more than once; matrix elements are then taken independently from a Gaussian distribution. This model is convenient for theorists (as we will see in later sections). The most natural and often used interaction matrices along with some comments are given in Table 2.

4.2 The random energy model (REM)

In this section, we digress from the polymer problem and discuss the properties of REM. The reader who is interested in getting straight to polymer freezing can skip to Section 4.2.5, where all the necessary properties of REM are briefly summarized.

4.2.1 What is REM? Formally, to compute the partition function of an arbitrary system, one only needs the list of all microstates (conformations), $1, 2, \ldots, \mathcal{M}$ with their respective energies $E_1, E_2, \ldots E_{\mathcal{M}}$. Generally, $\mathcal{M}$ is huge, as it scales exponentially with the number of particles (monomers), N:

$$\mathcal{M} \simeq \exp(\omega N) , \qquad (45)$$

where $\omega \sim 1$ depends on the conformations available, i.e. on chain flexibility, packing conditions, lattice geometry in the

Table 2. Commonly employed models of heteropolymer interactions. For studies of folding and design, 2×2 matrices can be parametrized in terms of a single parameter θ without loss of generality (see Section 5.2.3).

Name	Number of letters	Matrix	Ref.
MJ	20	Realistic energies for	[109]
MHB	N	Independent random energies	[9]
Potts	q	$B_{ij} = 1 - 2\delta_{ij}$	[111]
BWM	2	$B_{ij} = \overline{B} + \frac{\delta B}{2} \begin{pmatrix} -\sqrt{2}\cos\theta - \sin\theta & \sin\theta \\ \sin\theta & \sqrt{2}\cos\theta - \sin\theta \end{pmatrix}$	Sec. 5.2.3
HP	2	$\theta = \arccos\left(\sqrt{\frac{2}{3}}\right) \approx 35.3^\circ , \quad B_{ij} = \begin{pmatrix} -1 & 0 \\ 0 & 0 \end{pmatrix}$	[112]
Ising	2	$\theta = \frac{\pi}{2} = 90^\circ , \quad B_{ij} = -\sigma_i\sigma_j = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$	[113]
	2	$\theta = 0 , \quad B_{ij} = \sigma_i + \sigma_j = \begin{pmatrix} +1 & 0 \\ 0 & -1 \end{pmatrix}$	[114]
NEC	2	$\theta \approx 0.092 \approx 5.27^\circ , \quad B_{ij} = \begin{pmatrix} -2.3 & -1 \\ -1 & 0 \end{pmatrix}$	[115]
Electroweak mixing	2	$\theta \approx 0.485 \approx 27.8^\circ , \quad B_{ij} = \begin{pmatrix} 0.762 & 0.479 \\ 0.479 & -1.720 \end{pmatrix}$	[116]

case of lattice models, etc. As the system is disordered, all of its energies $E_1, E_2, \ldots, E_{\mathcal{M}}$ depend in general on the realization of disorder, that is, on the sequence. In the REM, one says that the energy of each conformation, say, E_1, is distributed over the realizations of disorder in the same way as the energies of all other conformations and is *statistically independent of them*. If we call $P(E)$ the probability distribution of the energy of some particular conformation over disorder, then REM implies that

$$P(E_1, E_2) = P(E_1)P(E_2)\,. \tag{46}$$

It is also usually supposed that the $P(E)$ distribution is Gaussian:

$$P(E) = (2\pi N\mathcal{E}^2)^{-1/2} \exp\left[-\frac{E^2}{2N\mathcal{E}^2}\right], \tag{47}$$

where $\mathcal{E}$ is the characteristic width of the distribution.

It should be stressed that the Gaussian character of a single-energy distribution (47) is far less important than the statistical independence of states expressed in (46), which is the hallmark property of REM.

4.2.2 The density of states for REM. In order to discuss latter the REM freezing phase transition, let us look at the energy spectrum of a typical realization of disorder. It is easy to generate realizations of this spectrum computationally, and two of them are shown as examples in Fig. 6. The figure shows that typical spectra consist of a very dense region, with many states at high and relatively modest energies, and a low energy part of the spectrum, which is discrete and comprised of only a few levels. Furthermore, the continuous part looks identical for all realizations of the disorder, while the discrete part is very individual and looks completely different for different realizations.

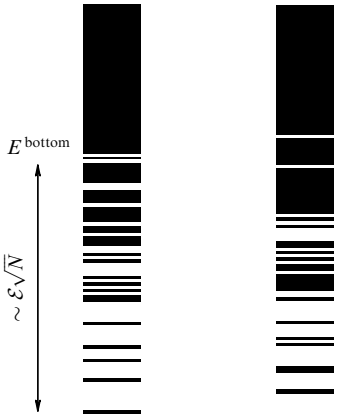

Figure 6. Two typical energy spectra for REM. Each consists of continuous and discrete parts. The two realizations demonstrate the property that the continuous part does not depend on realization, while the discrete part depends strongly.

We can get an insight into these properties of the REM energy spectrum by examining the density of states, $n(E)$. We remember that $n(E)$ is defined such that $n(E)\Delta E$ is the number of states with energies between E and $E + \Delta E$. It is very easy to write the expectation value for $n(E)$:

$$\langle n(E)\rangle = \mathcal{M}P(E)\,. \tag{48}$$

This value is huge (due to $\mathcal{M}$) whenever E is not far from the central part of the spectrum; in other words, an astronomically large set of states form an almost continuous spectrum at all energies where probability (47) is not very small. When the density of states is so large, it is about the same in all particular realizations, such that $n(E) \simeq \langle n(E)\rangle$. This argument, however, works only as long as $\mathcal{M}P(E) > 1$, or $E > E^{\text{bottom}}$, where

$$\mathcal{M}P(E^{\text{bottom}}) \sim 1 \to E^{\text{bottom}} \simeq -N\mathcal{E}\sqrt{2\omega}\,. \tag{49}$$

If we go to low energies $E < E^{\text{bottom}}$ where this breaks down, then the expectation number of the energy levels in an interval ΔE becomes less than unity. This means, that sometimes, for some realizations, there is one energy level, while for others there is not even a single one. Thus, we come to the important conclusion that REM in a typical realization has a practically continuous spectrum of states above a certain energy, and a discrete spectrum below it:

$$n(E) = \begin{cases} \mathcal{M}P(E)\,, & \text{when } E > E^{\text{bottom}}\,, \\ \text{random peaks}\,, & \text{when } E < E^{\text{bottom}}\,. \end{cases} \tag{50}$$

It is important that the continuous part of the spectrum is practically independent of the particular realization of disorder, while the discrete part (comprised of very few energy levels) is absolutely individual for each new realization.

4.2.3 Typical and atypical realizations in REM. To gain a deeper insight into the properties of REM, let us look at the energy differences between low energy states. First of all, one can easily write the probability distribution for the ground state energy. Indeed, for some state with energy E to be the ground state, all other states must be of higher energy; for each state, the corresponding probability is

$$\int_E^\infty P(E)\,\mathrm{d}E\,,$$

and most importantly, as all other $\mathcal{M} - 1$ states are independent, we get

$$P_{\text{ground}}(E) = \mathcal{M}P(E)\left[\int_E^\infty P(E)\,\mathrm{d}E\right]^{\mathcal{M}-1}. \tag{51}$$

Given that $\mathcal{M}$ is astronomically large, this amounts to†

$$P_{\text{ground}}(E) \simeq \begin{cases} \mathcal{M}P(E), & \text{when } E < E^{\text{bottom}}\,, \\ 0, & \text{when } E > E^{\text{bottom}}\,. \end{cases} \tag{52}$$

Thus, for the overwhelming majority of the realizations, the ground state energy is about $\mathcal{E}N^{1/2}$ below the boundary of the continuous spectrum, E^{bottom}. All discrete levels are in this interval; they are, therefore, very close to each other. Note that the differences between them, that scale as $N^{1/2}$, are negligible in the thermodynamic limit: the discrete energy levels correspond to states that are almost identical energetically and the system can be kinetically trapped in

† As long as $N \gg 1$, we can use saddle point approximation to get $\int_{E^{\text{bottom}}}^{+\infty} P(E)\,\mathrm{d}E \simeq 1 - k\exp\left[-(E^{\text{bottom}})^2/2N\mathcal{E}^2\right] \simeq 1 - k\exp(-N\omega)$ (k is a numerical constant). This value is very close to unity, but its (negative) deviation from unity is about $1/\mathcal{M}$. Therefore, immediately above E^{bottom}, the value $\int_{E^{\text{bottom}}}^{+\infty} P(E)\,\mathrm{d}E$ becomes small enough such that its $\mathcal{M}$-th power practically vanishes.

any of them. Note that the boundary of the continuous spectrum is about $\mathcal{E}N$ below the mean energy [see Eqn (49) above].

It is vitally important for what follows that besides the typical realizations of disorder, there are some rare atypical realizations of disorder for which the spectrum looks quite different. In particular, there are some realizations, albeit exponentially rare, for which the ground state energy is an order N below the threshold E^{bottom}; they will be of importance below.

4.2.4 Thermodynamics of REM. Consider now the thermodynamics of REM. While, in principle, one may wish to compute the partition function and the free energy for each particular realization of disorder, this is clearly impractical for most of the applications, and what one does instead is note that the averaged free energy is dominated by the typical realizations of disorder. Thus, we are first of all interested in an average of the form

$$F(T) = \left\langle F_{\text{seq}}(T) \right\rangle = -T \left\langle \ln Z_{\text{seq}}(T) \right\rangle . \tag{53}$$

To average the logarithmic function is a tedious mathematical task, and this is precisely why disordered systems are so difficult for theoretical examination. This is the place where the famous replica trick [117] enters. The main good news about REM is that one does not need to resort to these big theoretical guns.

Indeed, $Z_{\text{seq}}(T)$, the partition function for the given realization of disorder, is just the sum over all states $i = 1, 2, \ldots, \mathcal{M}$ and it can be always rewritten in terms of the density of states:

$$Z_{\text{seq}}(T) = \sum_{i=1}^{\mathcal{M}} \exp\left(-\frac{E_i}{T}\right) = \int_{-\infty}^{\infty} n(E) \exp\left(-\frac{E}{T}\right) \mathrm{d}E . \tag{54}$$

At high enough temperatures, this sum is dominated by the states of high entropy (large $n(E)$), where the spectrum is continuous and independent of sequence. This means that all the complications connected with the difference between individual realizations of disorder do not arise in this temperature region and the disorder is, in a way, irrelevant. Indeed, as long as the saddle point of the integral

$$\begin{aligned} Z &= \int_{-\infty}^{\infty} \mathcal{M} P(E) \exp\left(-\frac{E}{T}\right) \mathrm{d}E \\ &\simeq \mathcal{M} P(E_{\text{saddle}}) \exp\left(-\frac{E_{\text{saddle}}}{T}\right) \end{aligned} \tag{55}$$

belongs to the continuous spectrum region $E > E^{\text{bottom}}$, the first line of the equation (50) is valid, and thus we get a *partition function* that is independent of the disorder. For the Gaussian distribution (47), $E_{\text{saddle}} = -N\mathcal{E}^2/T$, we get that $E_{\text{saddle}} > E^{\text{bottom}}$ is valid at $T > T_{\text{glass}} = \mathcal{E}(2\omega)^{-1/2}$. Thus, at $T > T_{\text{glass}}$ we can safely average the partition function over the disorder (as it does not depend on disorder!) and arrive at the free energy which is also independent of the disorder.

This is not valid at lower temperatures. What happens there is that one or a few low energy states dominate the partition function. In principle, one could expect that at this low temperature, the thermodynamics of a particular sample will strongly depend on the disorder. Note, however, that typical differences between low energy states are only about $N^{1/2}$ and they are negligible in the thermodynamic limit.

As the free energy is a continuous function of temperature, we arrive at the following powerful conclusion for REM (see also [118]):

$$F(T) = \begin{cases} -T \ln \left\langle Z_{\text{seq}}(T) \right\rangle , & T > T_{\text{glass}} , \\ -T_{\text{glass}} \ln \left\langle Z_{\text{seq}}(T_{\text{glass}}) \right\rangle , & T \leqslant T_{\text{glass}} . \end{cases} \tag{56}$$

We shall comment in more detail later, that to take the average of the partition function means to take the 'annealed average.' Thus, equation (56) shows that for REM, the real quenched average of the free energy coincides, above the temperature T_{glass}, with the annealed average (see Section 4.4.2 below). This powerful conclusion is valid for every REM type model and it will provide us with a tool for further consideration.

4.2.5 Summary of REM properties. We summarize here the main properties of REM:

(1) The defining property of REM is the statistical independence of states (46).

(2) The REM energy spectrum consists of a continuous part which is independent of disorder and a few discrete energy levels that are placed very individually for each realization of disorder.

(3) The REM ground state for typical realizations is of the order $\sqrt{N}$ below the edge of the continuous spectrum, which in turn, is of the order N below the mean energy. Also, for typical realizations, the discrete levels are of the order $\sqrt{N}$ from each other.

(4) There is certain temperature for REM, T_{glass}, such that at $T > T_{\text{glass}}$ the system explores the high entropy continuous part of its spectrum, while at $T < T_{\text{glass}}$ it is locked into discrete individual states.

(5) The free energy of the REM is given by the equation (56).

Note, that these properties are independent of the Gaussian form of the single energy distribution (47).

4.3 Is REM valid for heteropolymer freezing?

The question posed in the title of this section has been addressed in more detail in the work [119]. The reader who ready to trust REM can skip this section and go straight to the next one.

4.3.1 REM cannot be exact for heteropolymers. In the work [7], Bryngelson and Wolynes postulated the applicability of REM for protein globules. In the work [9], Shakhnovich and Gutin showed that REM is applicable for compact heteropolymers with independent interactions (see below).

In the meantime, the statement of REM applicability is often met with understandable distrust. Indeed, REM obviously cannot be exact for heteropolymers. To understand that, let us imagine two conformations, say α and β, each of which represents some small local rearrangement of the other. As energies E_α and E_β are given as sums over all pairs of contacting monomers (we are speaking now about short-range interactions), they are dominated by identical contributions and differ only due to the small region of difference between α and β. Clearly, these two energies are strongly dependent.

The simplest quantitative measure of statistical interdependence between the energies of two given conformations α and β over the set of sequences can be obtained by taking correlation

$$\langle E_\alpha E_\beta\rangle - \langle E_\alpha\rangle\langle E_\beta\rangle = \langle \delta E_\alpha \delta E_\beta\rangle\,,$$

where $\langle \ldots \rangle$ denotes averaging over sequences. REM invalidity can be demonstrated by the non-vanishing of this correlator. To average, we take the probability for each sequence in the form of the product

$$P^{(0)}_{\rm seq} \equiv P^{(0)}(\{s_i\}) = \prod_{I=1}^{N} p_{s_I}\,, \tag{57}$$

which corresponds to the monomer species, $\{i\}$, occurring independently with probabilities $\{p_i\}$, (see also equation (62) below)†. We also define the mean and variance of the interaction matrix, $\overline{B}$ and δB^2, as

$$\begin{aligned}\overline{B} &= \sum_{ij} p_i B_{ij} p_j\,,\\ \delta B^2 &= \sum_{ij} p_i (B_{ij})^2 p_j - \left(\sum_{kl} p_k B_{kl} p_l\right)^2\\ &= \sum_{ij} p_i (B_{ij} - \overline{B})^2 p_j\,.\end{aligned} \tag{58}$$

Then, for the Hamiltonian (44), a straightforward calculation yields

$$\langle \delta E_\alpha \delta E_\beta\rangle = \langle \delta B^2\rangle \mathcal{Q}_{\alpha\beta} + K_{\alpha\beta} \sum_{ijk} p_i \delta B_{ij} p_j \delta B_{jk} p_k\,, \tag{59}$$

where

$$\mathcal{Q}_{\alpha\beta} \equiv \sum_{I\neq J} \Delta(\mathbf{r}_I^\alpha - \mathbf{r}_J^\alpha)\Delta(\mathbf{r}_I^\beta - \mathbf{r}_J^\beta) \tag{60}$$

is the conventionally defined overlap between conformations; it is proportional to the number of bonds that the conformations α and β have in common, and

$$K_{\alpha\beta} = \sum_{I\neq J\neq K} \Delta(\mathbf{r}_I^\alpha - \mathbf{r}_J^\alpha)\Delta(\mathbf{r}_J^\beta - \mathbf{r}_K^\beta)\,. \tag{61}$$

In fact, as the polymer is maximally compact, so that each monomer has z space neighbors, the value of $K_{\alpha\beta}$ in the volume approximation does not depend on the conformations and is equal simply to Nz^2. For some other cases, the value of $K_{\alpha\beta}$ plays the role of an important order parameter.

Thus, there are indeed correlations between states, and Eqn (59) shows that the correlations depend on the interactions (B_{ij}), the space of available conformations ($\mathcal{Q}_{\alpha\beta}$), and the sequences ($p_i$). The first (conformation dependent) term of the energy correlator, Eqn (59), corresponds precisely to our qualitative arguments, as it is proportional to the number of bonds in common.

Interactions enter into the conformational dependent term of Eqn (59) through the variance δB of the elements of the interaction matrix, and thus this term is present for all types of interactions. However, interactions play a more dramatic role in the conformational independent term in Eqn (59), as it vanishes for many models, but not, for example, if there is one monomer species that interacts particularly strongly with all others; then a correlated contribution comes to the energies even when there is no common pair, but just one common monomer of this peculiar species that interacts with anything else. The appearance of the conformation independent term signals a departure from REM, as even states with vanishing $\mathcal{Q}$ are statistically dependent. One notable example is the HP model (see Table 2).

For even overall composition ($p_i = 1/q$) and a symmetric contribution from monomer species, such as Potts or Independent Interaction Model (IIM) [9], the conformation independent term in Eqn (61) vanishes and the only statistical dependence comes from conformational overlap.

4.3.2 Why REM can be a good approximation. While REM cannot be exact, it appears the very good approximation in many cases (though not always). Its validity is due to the geometry of conformation space, which allows only relatively few local rearrangements. Typically, this happens because of severe constraints imposed on the conformations when a polymer is maximally compact; this is especially obvious if one thinks of the compact polymer on the lattice. Formally, to establish REM validity or invalidity for a particular model, one has to compute how many pairs of states (conformations), α and β, there are with the given value of the overlap $\mathcal{Q}_{\alpha\beta}$; as this is usually done using Monte Carlo technique, this number is associated with the probability distribution for $\mathcal{Q}_{\alpha\beta}$. REM is valid if this probability distribution is bimodal, with the peaks at small and maximal $\mathcal{Q}$ with the minimum in between. REM is invalid otherwise.

Moreover, as the heteropolymer freezing transition occurs between a phase consisting of exponentially many, unrelated conformations to a phase consisting of one conformation, any corrections to REM due to statistical dependence of states will have no effect on the thermodynamics. Such corrections are important for describing protein folding kinetics.

4.3.3 Examples of REM and non-REM logic. The REM-like assumption of statistical independence of states is implicit in the motivation of several experimental works; for example, *de novo* protein design [120] makes an REM-like assumption that the selection of sequences which lower the energy of a desired conformation will not also lower the energies of other conformations.

An *opposite* intuition is also prevalent in many works, such as the computational generation for a given sequence of a low, but not the lowest, energy conformation [121]; if REM were valid, then a low but not the lowest energy conformation would tell us nothing about the ground state.

Throughout this paper, we assume that REM is applicable. We discuss REM violations elsewhere [119]. We also explore non-REM generalizations of the theory presented here [122], but in this work we concentrate on the situations where REM is doing well.

4.4 Annealed Heteropolymers

4.4.1 What is an annealed heteropolymer? The result (56) of Section 4.2.4 involves the average value of the partition function over all possible sequences. Let us look at this value more closely:

$$\langle Z_{\rm seq}(T)\rangle = \frac{1}{\mathcal{N}} \sum_{\rm seq} Z_{\rm seq}(T)\,. \tag{62}$$

where $\mathcal{N}$ is the total number of sequences. As $Z_{\rm seq}(T)$ itself represents the sum over conformations, the bigger sum

† If the polymer is synthesized in some preparation bath, then p_i are proportional to $\exp(\mu_i/T)$, where μ_i are the corresponding chemical potentials of the monomers in the bath.

$\sum_{\text{seq}} Z_{\text{seq}}(T)$ has the physical meaning of the partition function of the system in which both conformation and sequence take part in thermal motion on an equal footing. This hypothetical system is called an annealed heteropolymer. Although there are some systems more or less resembling annealed heteropolymers [40], the interest of this system is deeper.

Note, that the Hamiltonian of annealed heteropolymers is given by the same equation (44) as for their quenched counterparts. The difference between the two is not in the form of the Hamiltonian, but rather in the way we compute the partition function. For the quenched system, we sum over all conformations. For the annealed system we sum over both conformations and sequences.

An annealed heteropolymer is in principle by far simpler compared to the real quenched counterpart. This is why the relationship (56) is so powerful, as it allows one to express directly all properties of a real quenched heteropolymer in terms of the much simpler annealed free energy. Although there is no universal exact solution even for the annealed free energy in terms of B_{ij}, relationship (56) allows the use of a variety of approximations or heuristic phenomenological formulae for the annealed free energy. This is similar to the approach of standard polymer theory. Indeed, a polymer fluid is *a priori* more difficult to study compared to its counterpart of regular small molecules. Given that there is not (and cannot be) a simple and satisfactory theory for the latter, one does not try to create such a theory for the former. Instead, one typically expresses the properties of polymeric liquid in terms of the macroscopic statistical properties of the appropriate low-molecular-weight fluid (which is usually the system of disconnected quasi-monomers [1, 32]). This was the program suggested by I M Lifshitz for polymers [1]. Similarly, our program here is to employ some qualitatively plausible interpolation for the annealed free energy to gain insight into the freezing behavior of quenched heteropolymers. We stress that the power of the results obtained is not undermined by the approximate character of the annealed free energy that we shall use. By contrast, as long as REM is valid, our method allows for the easy incorporation of any potential improvement of the impression for annealed free energy, whether taken from computer simulations, or numerical computations for further terms of high temperature expansion, etc.

Accordingly, before proceeding to the quenched case, we first examine the annealed free energy.

4.4.2 The annealed averaged free energy of IIM. The only model that allows for an exact solution for the annealed free energy is the Independent Interaction Model (IIM), as it is mappable onto the ideal gas problem. In IIM, we assume that there are at least as many monomer species as monomers $q \geqslant N$ such that the interaction energies between the monomers are chosen *independently* from a Gaussian distribution

$$P(B_{ij}) = \frac{1}{\sqrt{2\pi\delta B^2}} \exp\left[-\frac{(B_{ij}-\overline{B})^2}{2\delta B^2}\right], \tag{63}$$

where $\overline{B}$ and δB^2 are the mean and the variance, respectively. In this case, the annealed problem is solved by the following simple argument. To take the partition function over both conformations and sequences, let us first fix some arbitrary conformation and consider the summation (or averaging) over sequences. This is illustrated schematically in Fig. 7. In

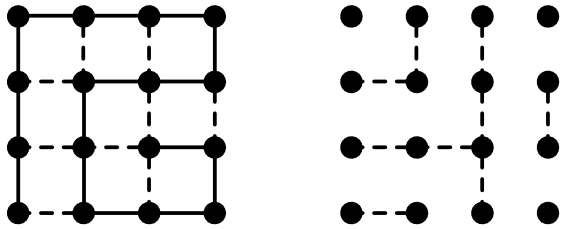

Figure 7. In this two dimensional figure, we illustrate that a fixed conformation means a fixed set of bonds between monomers.

IIM, we do not assign species to every monomer, but rather we assign energy to every bond that exists in a given conformation; as we take these energies independently from each other, averaging over sequences is reduced to independent averaging over all interaction energies, and this transforms a heteropolymer with a variety of monomers ('colorful pattern') into a homopolymer ('grey background') with even interaction energy given by

$$\exp\left(-\frac{B_{\text{eff}}}{T}\right) = \int \exp\left(-\frac{B}{T}\right) P(B)\,\mathrm{d}B\,; \tag{64}$$

given Gaussian distribution (63), we arrive at

$$B_{\text{eff}} = \overline{B} - \frac{\delta B^2}{2T}\,. \tag{65}$$

As long as we consider only maximally compact conformations, the total number of bonds, each with energy B_{eff}, is the same for all $\mathcal{M}$ conformations, and thus we end up with the annealed average free energy of the form

$$\begin{aligned} F_{\text{ann}}^{\text{av}} &= -T\langle Z_{\text{seq}}(T)\rangle = -T\ln\left[\mathcal{M}\exp\left(-\mathcal{Q}\frac{B_{\text{eff}}}{T}\right)\right] \\ &= \mathcal{Q}\left[\overline{B} - \frac{\delta B^2}{2T}\right] - TN\omega\,, \end{aligned} \tag{66}$$

where $\mathcal{Q} = \sum_{I\neq J}\Delta(\mathbf{r}_I - \mathbf{r}_J) \equiv \mathcal{Q}_{\alpha\alpha}$ is the (independent of the conformation α) number of bonds (or contacts) between monomers in any particular compact conformation, and $\omega = -\ln\mathcal{M}/N$ is the polymer entropy per monomer. Alternatively, we can arrive at the same answer formally, by calculating the Gaussian integral over B_{ij} in

$$\langle Z_{\text{seq}}(T)\rangle = P(B_{IJ}) \sum_{\text{confs}} \exp\left[-\sum_{I,J}^{N} \frac{B_{IJ}}{T}\ \Delta(\mathbf{r}_I - \mathbf{r}_J)\right]. \tag{67}$$

4.4.3 The annealed averaged free energy in terms of high temperature expansion. Unfortunately, for all other models there is no exact solution. Instead, one commonly employs a high temperature expansion to perturbatively calculate the annealed partition function. It may seem unjustified *a priori* to use a high temperature expansion to study freezing, which seems to be a 'low temperature' effect. However, we have to consider that freezing is caused by frustrations which prohibit the system to reach lower energy microstates of the unfrustrated system. In the polymeric case, the monomers may wish to rearrange themselves into a lower energy configuration, but the polymeric bonds prohibit this. Thus, the system

is 'frozen' at some temperature $T_{\rm glass}$. The validity of the high temperature expansion to describe freezing resides on the value of $T_{\rm glass}$ compared with the annealed phase transition temperature $T_{\rm c}$.

As we did for the IIM above, we begin with the averaging over sequences for some given compact conformation and the performing of a high temperature expansion keeping terms of order $O(1/T^2)$ in the annealed average:

$$\begin{aligned} -T\ln W^{\rm conf}(T) &= -T\ln\left[\left\langle \exp\left[-\frac{\mathcal{H}(\text{seq},\text{conf})}{T}\right]\right\rangle\right] \\ &\simeq \langle\mathcal{H}\rangle - \frac{1}{2T}\left[\langle\mathcal{H}^2\rangle - \langle\mathcal{H}\rangle^2\right] \\ &= \mathcal{Q}\left[\overline{B} - \frac{\delta B^2}{2T}\right]. \end{aligned} \tag{68}$$

To average is straightforward because, as in the IIM case above, the result does not depend on a particular conformation, as the number of contacts $\mathcal{Q}$ is the same in all compact conformations. When we finally sum (the partition function) over all $\mathcal{M}$ compact conformations, we arrive at

$$F_{\rm ann}^{\rm av} = \mathcal{Q}\left[\overline{B} - \frac{\delta B^2}{2T}\right] - TN\omega\,, \tag{69}$$

which is much the same as for the IIM (66), except for the more general definitions (58) for $\overline{B}$ and δB. Thus, to this order, we are essentially approximating $P(B_{ij})$ by a Gaussian and thus reforming the model into the IIM. Deviations from this behavior will be seen by examining terms in the expansion to higher order. We also note the difference between the free energy of the annealed system and the 'annealed average.' In the annealed average, the realizations of disorder behave like states, but we average, not sum over them. Thus, the entropy $T\ln\mathcal{M}$ is present in the annealed system, but not the annealed average.

There are several general properties which derive from aspects of this free energy, which have been previously worked out [123] and we simply repeat here:

(1) Heteropolymeric effects are independent of the mean of B_{ij}.

(2) Changing the variance of B_{ij} is equivalent to changing the temperature; the characteristic temperature scale in the heteropolymer problem is set by the variance δB.

(3) **Reduction theorems:** There are matrices which are formally different, but physically identical. For example, one can create a new 'clone' of species, but as the long as the interactions are identical nothing should happen physically. It was indeed shown that this requirement is obeyed.

4.5 Freezing transition

We are now equipped to describe the phase behavior of real quenched heteropolymers with random sequences. Our tools are equation (56) that expresses the quenched free energy in terms of the annealed average and expression (69) for the annealed averaged free energy. For further reference, we collect here these two to obtain the free energy of a real quenched system (averaged over sequences, as in (53)):

$$F(T) \simeq \begin{cases} \mathcal{Q}\left[\overline{B} - \dfrac{\delta B^2}{2T}\right] - TN\omega\,, & T > T_{\rm glass}\,, \\ \mathcal{Q}\left[\overline{B} - \dfrac{\delta B^2}{2T_{\rm glass}}\right] - T_{\rm glass}N\omega\,, & T \leqslant T_{\rm glass}\,. \end{cases} \tag{70}$$

4.5.1 Glass-like freezing in REM. Our discussion of REM suggests that something important happens when, due to temperature decrease, the average energy becomes lower than the boundary of the continuous spectrum. Above the corresponding temperature $T_{\rm glass}$, the REM-represented system explores many (of order $O(\exp N)$) states and behaves practically independently of the particular realization of disorder. Below this temperature, on the other hand, the equilibrium is dominated by a very few discrete states of low energy, and they are extremely individual for every realization of disorder. At $T > T_{\rm glass}$, the entropy of mixing over the continuous spectrum wins; at $T < T_{\rm glass}$, the energy of fluctuational low lying energy levels wins. The temperature $T_{\rm glass}$ is called the glass temperature, and the transition is called freezing. While it is very easy to find that $T_{\rm glass} = \mathcal{E}(2\omega)^{-1/2}$ for the example of a Gaussian single energy distribution (47), we are now more interested to apply the idea of freezing to heteropolymers. To this end, we note that the freezing transition is marked by the temperature at which entropy becomes $O(1)$. In the thermodynamic limit, we can therefore calculate the freezing transition temperature by looking at the point where the entropy vanishes. Thus, to find the freezing temperature, we can simply examine the relation

$$S(T_{\rm glass}) = -\left.\frac{\partial F}{\partial T}\right|_{T=T_{\rm glass}} = 0\,. \tag{71}$$

4.5.2 The freezing of random heteropolymers. We use our main relationship (70) and find the entropy of the quenched heteropolymer

$$S(T) = -\frac{\mathrm{d}F}{\mathrm{d}T} \simeq N\omega - \mathcal{Q}\frac{\delta B^2}{2T^2}\,,\ T \geqslant T_{\rm glass}\,. \tag{72}$$

Thus, $S(T) = 0$ at the temperature $T_{\rm glass}$ such that

$$T_{\rm glass}^2 = \frac{\delta B^2}{2s}\,, \tag{73}$$

where $s = \ln a^3/v$ is the conformational entropy per bond (and is therefore related to the entropy per monomer ω by the relation $s = N\omega/\mathcal{Q}$). This result, as well as higher order corrections agrees perfectly with the previous results of the (much more difficult) replica calculations [9, 123, 124]. Equation (73) also has a perfectly clear physical meaning: freezing is generally due to frustrated interplay between the energy gained by arranging favorable contacts, which amounts to about δB per monomer, and entropy loss due to polymer linear memory, which is governed by s.

It is instructive to rewrite expression (70) for the free energy of the globule in terms of $T_{\rm glass}$ instead of ω or s; using (73), we get

$$F(T) \simeq \begin{cases} \mathcal{Q}\left[\overline{B} - \dfrac{\delta B^2}{2T}\left(1 + \dfrac{T^2}{T_{\rm glass}^2}\right)\right], & T \geqslant T_{\rm glass}\,, \\ \mathcal{Q}\left[\overline{B} - \dfrac{\delta B^2}{T_{\rm glass}}\right], & T \leqslant T_{\rm glass}\,. \end{cases} \tag{74}$$

4.5.3 The order parameter for the freezing transition. The more delicate characteristic of freezing is the parameter $x(T)$ that can be defined for each sequence as

$$x_{\rm seq}(T) = 1 - \sum_{\gamma=1}^{\mathcal{M}} \mathcal{P}_\gamma^2\,, \tag{75}$$

where $\mathcal{P}_\gamma$ is the Boltzmann probability for the state (conformation) γ; for the given sequence, $\mathcal{P}_\gamma \sim \exp(-\mathcal{H}_\gamma/T)$. One can also define the sequence average $\langle x(T)\rangle_{\text{seq}} \equiv x(T)$. On the one hand, this value has a simple physical meaning in the replica approach (describing a grouping of replicas due to spontaneous replica symmetry breaking; out of n replicas, there are n/x groups with x replicas in each group). On the other hand, this value is known to be related to the number of thermodynamically relevant states, $M_{\text{seq}}(T)$: $x_{\text{seq}}(T) = [(1 - M_{\text{seq}}(T)]^{-1}$. This value plays the role of order parameter for the freezing transition [117]. For REM one can show that

$$x(T) = \begin{cases} 1\,, & T \geqslant T_{\text{glass}}\,, \\ \dfrac{T}{T_{\text{glass}}}\,, & T \leqslant T_{\text{glass}}\,. \end{cases} \tag{76}$$

4.5.4 Is freezing typical for only some particular types of interactions? To conclude, let us stress that a polymer is considered a heteropolymer if it is composed of differing monomeric species, mathematically expressed by $\widehat{\Delta}\widehat{B} \neq \widehat{0}$, where $\widehat{\Delta}_{ij} = p_i p_j - p_i \delta_{ij}$. All interaction matrices of this form lead to a finite freezing temperature for random sequences. Thus, the particular details of the interaction matrix are vital neither to the *existence* of the freezing transition nor to the qualitative aspects of its properties. Needless to say, they are very important in what concerns the choice of a particular conformation to be the ground state. As to the freezing itself, it is the robust property of virtually any heteropolymer, provided the interaction energies are sufficiently diverse to bring the transition into a reasonable temperature interval.

4.6 Computational tests of freezing

4.6.1 Why computer simulations are needed. There are several reasons why analytic results of the previous sections should be complemented by computational studies:

(1) Real proteins are relatively small, and analytic models never allow anything but the thermodynamic limit;

(2) REM applicability is not very solidly argued;

(3) More complicated models are also of interest.

4.6.2 Enumeration of Hamiltonian walks and other conformations†. To perform computational studies of the freezing transition necessarily requires an exhaustive list of conformations. Indeed, as the system is going to freeze to one conformation, which gives the overwhelming contribution to the partition function, Monte Carlo sampling over conformational space can easily miss this particular conformation. For a maximally compact globule, this leads to the problem of the enumeration of Hamiltonian walks on parts of a (cubic) lattice.

A Hamiltonian walk is defined to be a walk over some graph such that each vertex is visited once and only once. This was first performed by Shakhnovich and Gutin [125] when enumerating all the 103346 (unrelated by symmetry) Hamiltonian walks on a $3 \times 3 \times 3$ cubic sub-lattice. One of them is shown in Fig. 8. The use of a massively parallel computer (128 node Thinking Machines CM-5) yielded sufficient computational power to enumerate the Hamiltonian walks on $3 \times 3 \times 4$ and $3 \times 4 \times 4$ sub-lattices [125]. The results are summarized in Table 3.

In Figure 9, the natural logarithm of the number of Hamiltonian walks $\mathcal{M}$ is plotted versus N and the data are

† This section is based on the work [126].

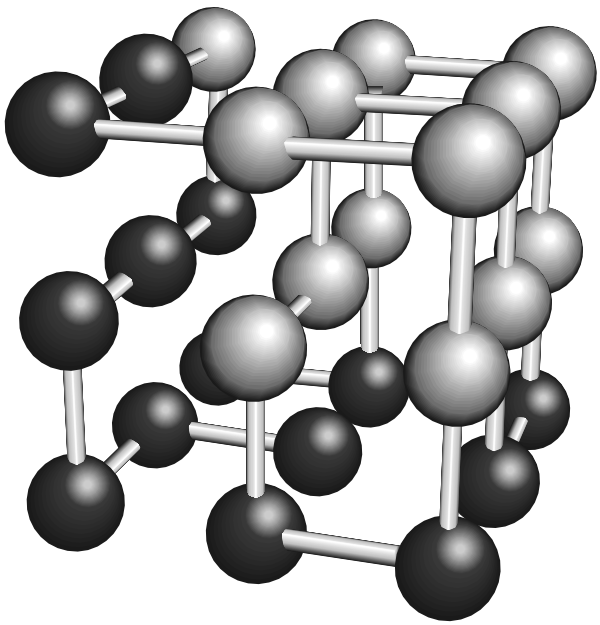

Figure 8. One of the compact conformations of the 27-mer on a $3 \times 3 \times 3$ piece of the cubic lattice. Due to its computational tractability, 27-mer has become one of the most popular models for protein folding studies. The 27-mer shown here consists of two types of monomers (black and white) and the conformation shown here is a ground state conformation for Ising interactions.

Table 3. Summary of enumeration data. N is the number of sites, $\mathcal{M}$ is the number of Hamiltonian walks (maximally compact conformations unrelated by symmetry), $\mathcal{M}_{\text{tot}}$ is the total number of conformations.

N	Comment	$\mathcal{M}$	$\mathcal{M}_{\text{tot}}$
18		1,085	5,577,317,124
27		103, 346	
36		84,731,192	
48	64 hours of CPU time on SM-5	134,131,827,745	
26	empty site	564,368	
64	crumpled	261,496,832	

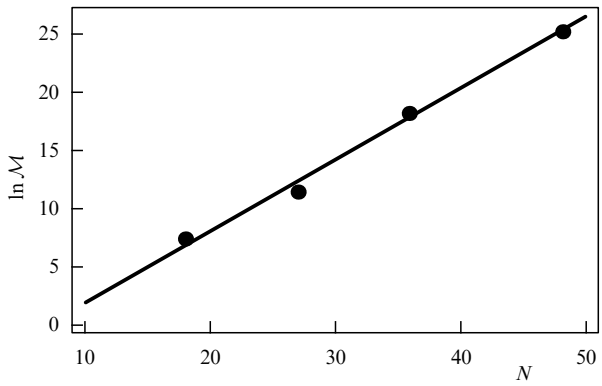

Figure 9. Log-Log plot of the number of Hamiltonian walks $\mathcal{M}$ vs. chain length N. Interestingly, the point for $N = 27$ is somewhat below the line. It is possibly due to the right cubic shape; if this is the case, then the point for $N = 64$ should also be below the line.

seen to fit well to a straight line, or

$$\mathcal{M} \approx ab^N \approx 0.02 \cdot (1.86)^N \tag{77}$$

(more accurately, $\ln a = -4.3 \pm 1.2$ and $\ln b = 0.62 \pm 0.04$). This fit works well for the region of small $N \leqslant 48$.

On the other hand, Flory mean field calculation of the entropy of polymer melts [127] is known to be applicable to the estimation of the number of compact globular conformations in the $N \to \infty$ limit. Indeed, the conceptual foundation of the Flory treatment is the restriction imposed on the addition of new monomers within the constraints of the avoidance of occupied sites and chain connectivity. This kind of argument is equally applicable to both a macroscopic melt of different long chains, and a large globule of one single chain, as two systems differ only in the contributions of independent chains mixing entropy, which is negligible in the long-chains melt, and of surface effects, which are negligible in the thermodynamic limit. Therefore, in the $N \to \infty$ limit we have the estimate

$$\mathcal{M} \approx \left(\frac{z-1}{\exp 1}\right)^N \approx (1.84)^N , \tag{78}$$

where z is the coordination number of the lattice, and we take $z = 6$ for the simple cubic lattice.

We see that equations (77) and (78) agree well; we suggest, therefore, that equation (77) may be used to derive the number of walks for arbitrary N to logarithmic accuracy.

In the recent work [128], the claim was made that the Flory result (78) should be improved and the correct one looks like $\mathcal{M} \simeq (z/\exp 1)^N$; the enumeration results [126] do fit the Flory formula (78), but do not fit this improved formula.

For $N = 64$, equation (77) yields $\mathcal{M} \approx 2 \times 10^{15}$. The enumeration of the $4 \times 4 \times 4$ sub-lattice is, therefore, several orders of magnitude out of reach using our current computer power. Also, the case $N = 48$, while possible to enumerate, is still extremely consuming of CPU time and therefore cannot be used routinely in any current polymer modeling scheme. However, enumeration of $N = 36$ is not very consuming of CPU time. Furthermore, there are fundamental differences between the cases of $N = 27$ and $N = 36$, such as the presence of pseudo-knots in the later.

As to the other examples where conformations can be exhaustively enumerated, we mention here the enumeration of the conformation of 26-mer on the $3 \times 3 \times 3$ sub-lattice with one forbidden site [13], the recently performed enumeration of all conformations for $N \leqslant 18$, and the enumeration of all crumpled conformations on a cubic sub-lattice of size $2^k \times 2^k \times 2^k$†. For the later case, a very good approximation is given by the $k \to \infty$ asymptotic

$$\mathcal{M}_k^{\text{crumpled}} \simeq \gamma A^N , \tag{79}$$

where $\gamma \approx 0.87300$, $A \approx 1.3565$; for example, it yields $\mathcal{M}_2^{\text{crumpled}} \approx 2.605 \times 10^8$ as compared to the exact answer 2.61496832×10^8.

† Crumpled conformations (see Section 3.6 above) on the lattice can be defined in the following way: as the $2^k \times 2^k \times 2^k$ cube can be viewed as 8 smaller cubes $2^{k-1} \times 2^{k-1} \times 2^{k-1}$ each, and each smaller-subcube can be further divided in a similar way, etc, down to the smallest $2 \times 2 \times 2$ cubes, we define the trajectory to be crumpled if it visits all the vertices within given subcube before entering next subcube of the same level.

Note, that according to (78), the fraction of crumpling among all compact trajectories decreases exponentially with N:

$$\frac{\mathcal{M}^{\text{crumpled}}}{\mathcal{M}} \approx \left(\frac{A \exp 1}{z-1}\right)^N \approx \exp(-0.3N) .$$

4.6.3 The freezing transition is indeed observed computationally. The first computational test of freezing was performed in the work [125] for the Independent Interaction Model. The value $x(T)$ (75) was employed to monitor freezing. The type of behavior predicted by equation (76) was indeed found in [125] and later in several other works for $N = 18$, 27, and 36. Of course, as the system is finite, there could not be a breaking point on the $x(T)$; but the curves are really sharp. The fact that $x(T)$ becomes less than unity tells us directly that the entropy of the system vanishes, and that means precisely freezing.

4.6.4 Computational tests of REM validity. One of the most obvious questions that can be addressed by computer experiment is to test the applicability of REM. It was first done in [125] by looking at the thermal probability distribution of the overlap order parameter $\mathcal{Q}_{\alpha\beta}$:

$$P(\mathcal{Q}) = \sum_{\alpha\beta} P_\alpha P_\beta \delta\big(\mathcal{Q} - \mathcal{Q}_{\alpha\beta}\big) . \tag{80}$$

If REM is valid, this distribution must be bimodal, with peaks at zero or complete overlap. Indeed, probability distribution (80) is dominated by the low energy states which in REM do not overlap. Therefore, all pairs with $\alpha \neq \beta$ have $\mathcal{Q}_{\alpha\beta} = 0$ and contribute to the $\mathcal{Q} = 0$ peak of the probability distribution $P(\mathcal{Q})$. On the other hand, all pairs with $\alpha = \beta$ obviously give $\mathcal{Q} = \mathcal{Q}_{\max}$, thus yielding

$$P(\mathcal{Q}) = x\delta(\mathcal{Q}) + (1-x)\delta(\mathcal{Q} - \mathcal{Q}_{\max}) ,$$

where parameter x depends on the temperature and tells us how deeply the system is frozen; this is, of course, exactly the $x(T)$ parameter defined above (75).

The bimodal distribution $P(\mathcal{Q})$ was indeed found in [125]. Of course, the peaks were not exactly at $\mathcal{Q} = 0$ and $\mathcal{Q}_{\max}$, and this was interpreted as the manifestation of final size effects. Thus, the conclusion of the work [125] was that REM appears valid for heteropolymer freezing.

In the recent work [119], we returned to this question and have scrutinized REM applicability. We found that REM is valid in many cases, but is far from being applicable universally. It is well justified for maximally compact conformations and the Independent Interaction Model (for which it was originally tested in [125]), but not for other cases. The conformation dependent aspect is illustrated by Fig. 10. As we see, $P(\mathcal{Q})$ for all conformational spaces studied are peaked at small $\mathcal{Q}$. As $\mathcal{Q}$ increases, $P(\mathcal{Q})$ decreases exponentially, and above a particular $\mathcal{Q}_{\text{d}}$, there are at most $O(1)$ conformations available. Thus, for a space with small $\mathcal{Q}_{\text{d}}$, there are few states with large overlap and REM is favored. Crumpled conformations have the greatest $\mathcal{Q}_{\text{d}}$ as they allow a greater possibility of rearrangement on small scales (large $\mathcal{Q}$). Unfortunately, there is no known way to even estimate $\mathcal{Q}_{\text{d}}$ analytically.

There are other aspects to the problem of REM applicability, besides the geometry of conformations and conforma-

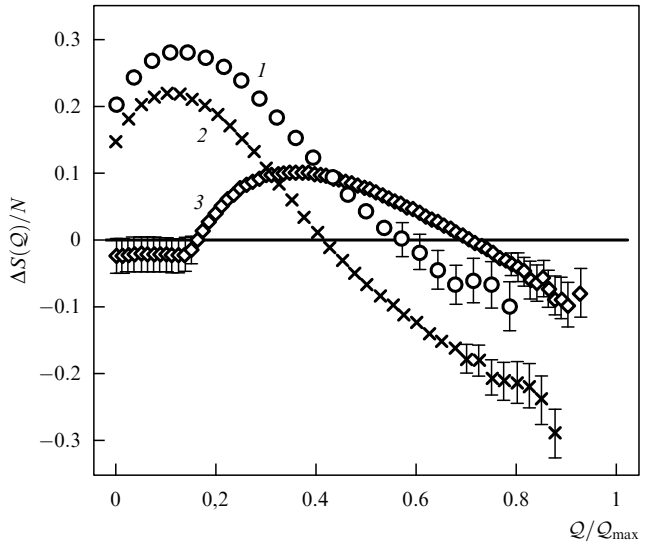

Figure 10. $\Delta S(\mathcal{Q}) \equiv \ln[P(\mathcal{Q})/P(\mathcal{Q}=\mathcal{Q}_{\max})]$ for compact 27-mers (○), compact 36-mers (×), and compact & crumpled 64-mers (◇). The discrete region boundary varies greatly: $\mathcal{Q}_d/\mathcal{Q}_{\max} \approx 0.6, 0.4, 0.7$ for 27-mers, 36-mers, and crumpled 64-mers respectively. When considered from right to left, this figure can be viewed as the dependence of the 'sphere surface area' on the sphere radius in conformational space. Indeed, $1 - \mathcal{Q}/cal Q_{\max}$ measures the distance from a given conformation, and we may see how many conformations there are at a given distance from a given conformation. At zero distance, there is always one conformation - the given one, thus $P(\mathcal{Q}_{\max}) = 1$ for all examples. When the distance is small $\mathcal{Q} > \mathcal{Q}_d$, there are typically no conformations at all, only some discrete events for some particular centers. This is why $\Delta S < 0$ in this region. Only after that, the 'surface area' starts to grow with the radius.

tional spaces. For example, degenerate ground states (as is common in models with 'discrete' interactions [129]) do not violate REM if they do not overlap. This holds for 27-mers and 36-mers with Potts interactions (see Table 2), whose ground states (i.e. $T = 0$) yield a $P(\mathcal{Q})$ which is indistinguishable from that of conformations taken at random [125, 129]. In the light of our previous discussion, REM appears to be valid for these cases because $\mathcal{Q}_{\rm d}$ is sufficiently small. The situation is different for crumpled 64-mers: upon enumerating [119] the energies of all conformations for 1000 sequences with Ising interactions and comparing the ground states, we have shown that the increase in $\mathcal{Q}_{\rm d}$ for crumpled 64-mers is sufficient such that REM fails for this conformational space.

Although we have thus demonstrated that REM validity is clearly not an *a priori* property of conformation spaces in general, even in three dimensions, and in particular, that REM breaks for crumpled 64-mers, we shall keep considering the REM approximation as it is still valid in many cases.

5. Designed heteropolymers

5.1 Design of sequence using canonical ensembles

5.1.1 Why sequences should be selected. When the freezing transition for random heteropolymers was first discovered [9], it was believed by many that this was already a good model for protein folding, as it yields a unique ground state with reasonable (N independent) probability. It was later realized, that this ground state, although formally unique, is not sufficiently robust. As the typical energy difference between low energy states scales as $N^{1/2}$, and in REM they are structurally very different and even unrelated, every slight change of parameters (about $N^{-1/2}$) or solvent conditions leads to a complete alteration of the ground state conformation [130]. Obviously, this is not what happens in nature, where protein native states are remarkably stable.

Another aspect of the problem is that the choice of ground state for the random sequence cannot be controlled, and thus it is problematic to obtain any desirable properties of the native state out of random choice of sequences.

One can draw the following analogy, which is actually quite deep. Protein folding phenomena can be viewed similarly to the reading of a message; or better, to reading with *understanding*. When we read a message and understand it, an image appears in the mind, which is in a sense 'induced' by the meaning of the message, but its physical nature is, of course, completely different from, say, the string of ink letters. Similarly, folding of the protein chain means the appearance of a three dimensional structure, that is, of a structure that is very different in nature from a one dimensional string of chemical 'letters.' The important point about reading is that one has to know the language. In other words, one cannot just read an arbitrary (random) string of letters. For the message to be readable it must have been written in the first place by somebody who knows the same language. Similarly, for a heteropolymer chain to fold reasonably its sequence should be first 'written,' or designed. From the analogy we conclude that the same language should be used for design as is used for 'reading = folding.'

This discussion suggests that the design of sequences should employ interactions between monomers ('language') and be directed at choosing atypical realizations from the REM ensemble, such that the ground state energy is sufficiently below the REM bottom of the continuous spectrum. This was realized in procedures called sequence annealing [12] and imprinting [13]; they were suggested independently, and later realized to be identical to the mean field approximation. Both can be said to be the realization of the 'minimal frustration' principle [7] (see also [131]). While they can be formulated in a rather abstract way, it is easier to begin with a simple representation.

5.1.2 Imprinting. The idea of imprinting is illustrated in Fig. 11. The process includes a few stages. There are two independent inputs for the process. Firstly, we take a mixture of disconnected monomers, mix them, and let them reach equilibrium at some preparation temperature T_p. We have to mix as many different monomers as we want to have later in our polymer chain. We also assume that these disconnected monomers are confined in a small cavity, such that the density of the 'monomer fluid' is close to maximally dense packing. Finally, the monomers interact in the preparation mixture through the interactions B_{ij}. Secondly, we choose one conformation from the list of all compact conformations of the given length and size. We want this conformation to be the renaturable ground state. This is why we call it the target conformation $\star$.

When both of the mentioned ingredients are ready, we start cooking the dish: we instantly apply the chosen conformation to the equilibrium configuration of monomers and we instantly and irreversibly connect monomers along the path prescribed by the target conformation $\star$. Thus, the chain appears, and we consider that its sequence of monomers is quenched and does not change any more.

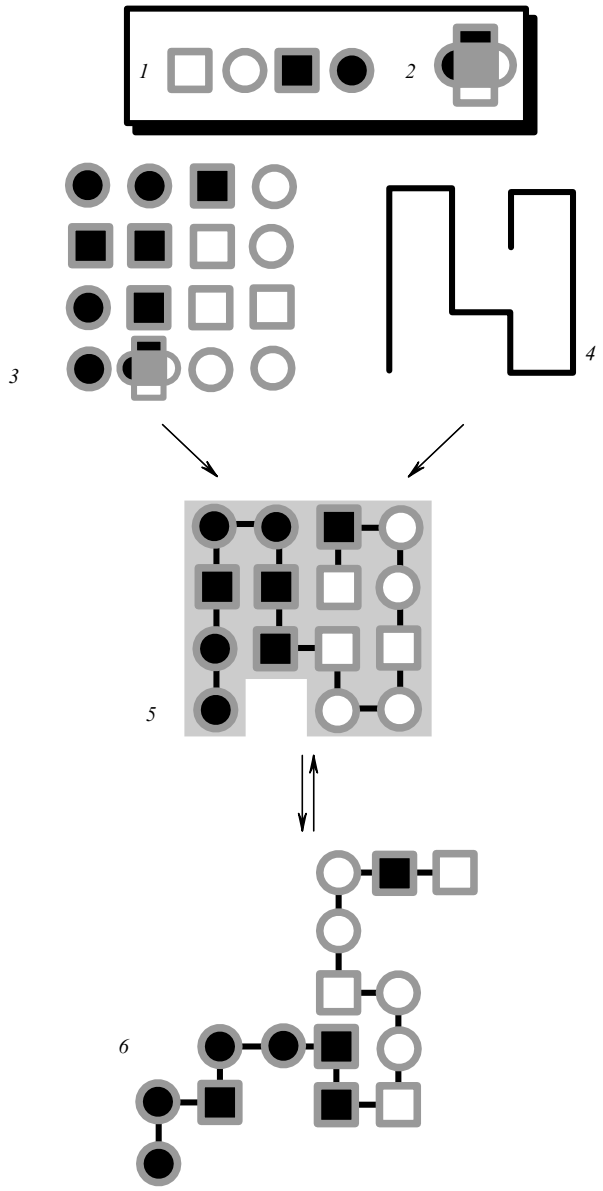

Figure 11. Schematic two-dimensional representation of imprinting. We begin with chemical constituencies, in this case with four distinct chemical species (1) and a target molecule (2). It is meant that rectangles attract other rectangles, circles attract circles, black attracts black, and white attracts white. We then prepare mixture of monomers and a taget molecule (3), keep it in a close volume, and allow to equilibrate there at the temperature $T_{\rm des}$. Independently, we choose target conformation (4). When both components are done, we instantly connect them together (5). Now we have a polymer. The polymer can expand (6), but its sequence remains quenched, and it is able to renature back to the globular conformation where it has a pocket complementary to the target molecule.

The hope is that the conformation $\star$ will be the ground state conformation. Why can we hope so? Because while the monomers were disconnected they optimized their configuration in space, forming preferably low energy contacts, or, as we call it in our theoretical jargon, forming lots of 'good bonds.' Clearly, some of those bonds are removed for polymerization. However, the fraction of these bonds is relatively small. Indeed, even on a cubic lattice, each particle has 6 bonds, of which 2 are utilized for polymerization; thus, 2/3 remain unchanged. For off-lattice models with bulky monomers this fraction maybe even higher. In general, the fraction of two along-the-chain connections among all monomer neighbors in the condensed state is the parameter v/a^3, well known in polymer theory [32]. The mean-filed approximation corresponds to the lowest order in v/a^3, and thus at least in the mean filed approximation we hope that the $\star$ conformation is indeed well optimized for the polymer, not only for the set of disconnected monomers.

The name of this approach is such because the target molecule gets 'imprinted' on the sequence.

5.1.3 Sequence annealing. Although it is unclear how to realize the idea of imprinting chemically, it is at least formulated in such a way that ignoring chemistry one can think it possible. This author thinks that this problem is difficult, but not hopeless. An alternative scheme of design has been suggested in the work [12] and it is formulated in a more abstract (or honest) way, as a prescription for computer simulation. It begins with the polymer whose conformation is frozen at $\star$ and then monomers are allowed to move arbitrarily, with only the condition of excluded volume: each monomer occupies one and only one place on the chain. In other words, conformation is considered quenched here, while sequence is annealed. This sounds, of course, somewhat counterintuitive, and indeed, it is unclear how this process can be realized other than in computer simulation. Nevertheless, it is obvious that the only difference between the two design schemes is that imprinting works with disconnected monomers, while the sequence annealing scheme works with annealed sequences. Hence the difficulties of the both: in one case, how to anneal the sequence; in the other, how to perform polymerization such that monomer low energy configuration is not totally destroyed. In any case, computationally the two schemes are much the same.

5.1.4 Microcanonical and canonical design. In principle, one can try to select sequences directly based on their lowest native state (NS) energy, $E_{\rm NS}$. The corresponding ensemble of sequences is similar to the microcanonical ensemble in regular statistical mechanics.

In statistical mechanics, it is technically more convenient to use the canonical ensemble, where the temperature is fixed instead of the energy. A similar idea is also valid for the sequence design. We use an analog of the canonical ensemble where the native state energy E_{NS} is not fixed, but rather controlled through an artificial temperature $T_{\rm des}$. Equivalently, we choose some 'target' conformation $\star$ that we want to be the native state of the designed sequence and constrain its energy $E_\star = \mathcal{H}({\rm seq}, \star)$ with a Lagrange multiplier $1/T_{\rm des}$. In this canonical ensemble, each sequence appears with Gibbs distributed probability:

$$P^{\star}_{\rm seq} = P^{(0)}_{\rm seq} \exp\left[-\frac{\mathcal{H}({\rm seq},\star)}{T_{\rm des}}\right]\left\{\sum_{\rm seq} P^{(0)}_{\rm seq} \exp\left[-\frac{\mathcal{H}({\rm seq},\star)}{T_{\rm des}}\right]\right\}^{-1}, \tag{81}$$

where $P^{(0)}_{\rm seq}$ is the probability for the sequences made randomly from independent monomer species, with occurrence probabilities p_i; we have already used these probabilities earlier, see (57). Obviously, both imprinting and sequence annealing are the realizations of the general idea of canonical design.

Thus, we characterize a given canonical ensemble of designed sequences by the value of $T_{\rm des}$: for lower $T_{\rm des}$, we model sequences whose native states are better optimized

energetically, while for higher $T_{\rm des}$ we are left with an unaltered ensemble of random sequences.

There is an interesting question as to whether all possible target conformations $\star$ are equally suitable for design. This question was addressed in the works [106 – 108, 132], where it was formulated in the following way: why many proteins, that are very different both as regards their evolutionary origin and their function, still have somewhat similar ternary structures, albeit similar only in a coarse grained sense. The authors argued that nature has selected those structures for which it is easier to design the sequences. More recently, this view was elaborated in the work [115] by means of the following tremendous computational effort. For one particular interaction matrix (marked as NEC in the Table 2), the authors computed the energies of all the 103346 conformations for all 2^{27} possible sequences. They showed that the conformations are indeed very different as regards the number of sequences for which they serve as ground states. Theoretical analysis of this question indicates, however, that this effect is either due to the final size of the system [115], or to the crumpled character of the conformations [132]. In any case, it is beyond what we shall consider analytically; we shall consider all compact $\star$ conformations 'democratically,' on an equal basis.

5.1.5 Energy of the target conformation. In this section, we shall compute the energy of the target conformation, averaged over sequences. Using the probability distribution for the designed sequences (81), we write:

$$\begin{aligned}\left\langle E_\star(\mathrm{seq})\right\rangle &= \sum_{\rm seq} P^\star_{\rm seq} E_\star(\mathrm{seq})\\ &= \frac{\sum_{\rm seq} P^{(0)}_{\rm seq} \exp\left[-\mathcal{H}(\mathrm{seq},\star)/T_{\rm des}\right]\mathcal{H}(\mathrm{seq},\star)}{\sum_{\rm seq} P^{(0)}_{\rm seq} \exp\left[-\mathcal{H}(\mathrm{seq},\star)/T_{\rm des}\right]}\,.\end{aligned} \tag{82}$$

The very structure of this equation suggests the following simple trick (similar to what is done regularly on the first several pages of any statistical mechanics text book). Let us return to definition (68) of the annealed average partition function as a function of temperature:

$$\begin{aligned}W^\star(T) &= \left\langle \exp\left[-\frac{\mathcal{H}(\mathrm{seq},\star)}{T}\right]\right\rangle\\ &= \sum_{\rm seq} P^{(0)}_{\rm seq} \exp\left[-\frac{\mathcal{H}(\mathrm{seq},\star)}{T}\right].\end{aligned} \tag{83}$$

Then, in terms of this partition function, we can immediately write:

$$\left\langle E_\star(\mathrm{seq})\right\rangle = -\frac{1}{W^\star}\,\frac{\partial W^\star}{\partial(1/T)}\bigg|_{T\to T_{\rm des}} = -\frac{\partial \ln W^\star}{\partial(1/T)}\bigg|_{T\to T_{\rm des}}. \tag{84}$$

Up to this point, we have made no approximations. To obtain some concrete result, however, we use our lowest order high temperature expansion for the annealed averaged free energy (68) and find that in this approximation $\ln W^\star(T)$ does not depend on $\star$ and is

$$-\ln W^\star(T) = \frac{\mathcal{Q}\overline{B}}{T} - \frac{\mathcal{Q}\delta B^2}{2T^2}\,. \tag{85}$$

Therefore $\langle E_\star\rangle$ is given by

$$\left\langle E_\star\right\rangle = -\frac{\partial \ln W^\star}{\partial(1/T)}\bigg|_{T\to T_{\rm des}} = \mathcal{Q}\left[\overline{B} - \frac{\delta B^2}{T_{\rm des}}\right]. \tag{86}$$

This is a very important result. We see directly here how the design temperature affects the target state energy. Note, that the 'susceptibility' of the target state energy to design is proportional to the variance δB of the interaction matrix: it is natural that it is purely a heteropolymeric effect, as it is based on the energy optimization for the target conformation.

REM implies a very peculiar spectrum of energies for a heteropolymer with a designed sequence. Note that design affects only the energy of the target conformation, while the statistics of all other energies remain unaffected. In this sense, one should understand the common jargon of design described as the 'pulling down' of just one energy level. This idea is illustrated by Fig. 12. Thus, design means selection of very atypical realizations out of the REM ensemble (see Section 4.2.3).

We note that more general analysis beyond the REM framework indicates that design in fact affects many conformations, especially ones with $\mathcal{Q} > \mathcal{Q}_{\rm d}$; this is very important for kinetics.

5.1.6 Folded (native) phase and folding temperature. This peculiar energy spectrum implies very special freezing behavior for designed sequences. Indeed, when we design at infinite temperature ($T_{\rm des} = \infty$), we are not selective as to the energy of the sequence in the target conformation $\star$ at all and the sequences we obtain are random. As we lower $T_{\rm des}$, we choose sequences in which $\star$ has lower energy. At the point where $T_{\rm des}$ is sufficiently low such that the energy of $\star$ is less than that of the REM typical ground state, then $\star$ is the ground state and we expect freezing to $\star$. This occurs for $T_{\rm des} < T_{\rm glass}$, as the REM ground state is stable at $T_{\rm glass}$. Furthermore, freezing of the sequences designed with $T_{\rm des} < T_{\rm glass}$ occurs at a temperature above $T_{\rm glass}$, because the ground state for those sequences is more stable and of lower energy compared to a typical REM ground state.

Also, for $T_{\rm des} < T_{\rm glass}$, the target conformation will be better optimized than the REM ground state and will freeze at some folding temperature $T_{\rm fold}$ which is greater than the glass temperature $T_{\rm glass}$. We can find $T_{\rm fold}$ by comparing which is lower: the target energy or the random globule free energy.

Thus, we compare the target state energy $\langle E_\star\rangle$ (86) to the random globule free energy expressed most conveniently by the first line of equation (74). We find for the folding transition temperature $T_{\rm fold}$:

$$\frac{1}{T^2_{\rm fold}} + \frac{1}{T^2_{\rm glass}} = \frac{2}{T_{\rm fold}\,T_{\rm des}}\,. \tag{87}$$

Note that this relation is independent of the specific aspects of $\widehat{B} = B_{ij}$, although the variance δB^2 enters through $T_{\rm glass}$. This is an approximation inferred by truncation of the high temperature series, and it is valid for the case in which the number of monomer species is large and the matrix elements of $\widehat{B}$ are uncorrelated. For other cases, higher order terms of the high temperature expansion must be used [133, 134].

The results of this section are summarized in the phase diagram, Fig. 13. All phase boundaries are determined directly from the annealed average partition function. This allows quick calculation of even exotic heteropolymer interaction models, within the validity of REM.

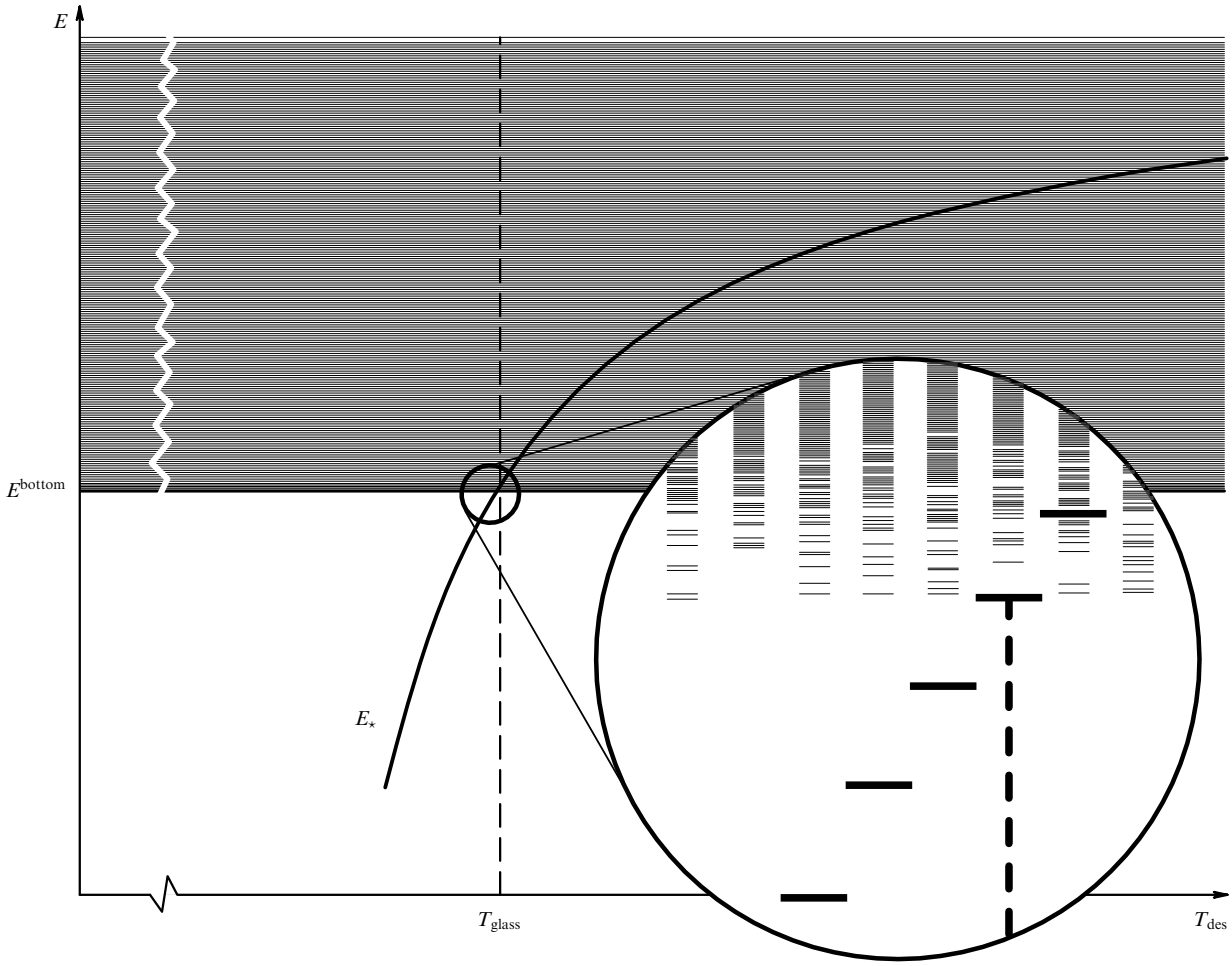

Figure 12. Sample energy spectra for sequences imprinted at different polymerization temperatures (T_{des}). The energy of the target conformation ($E_\star$) vs polymerization temperature (T_{des}) is plotted. As T_{des} is increased to T_g, $E_\star$ increases. In the region $T_{\text{des}} \approx T_g$ (magnified section), we see that $E_\star$ is equal to E^{bottom}, the average ground state energy of a random chain. This is related to the phase transition between the folded and glassy phases (see phase diagram, Fig. 13 below). It is instructive to see a realistic representation of the very bottom part of the energy spectrum, as is shown here in the magnified section.

We stress that REM is important not just in the aid of calculations, but in enabling this simple design scheme (i.e. selecting sequences which minimize the energy in a desired conformation) to work at all. Due to the statistical independence of states, we can alter the energy of a particular state without influencing the others.

The phase diagram in Fig. 13 allows the formulation of the principle of sequence optimization in a new form: one can say that the optimized sequences are those that have the maximal possible ratio of the two characteristic temperatures, $T_{\text{fold}}/T_{\text{glass}}$. This formulation is becoming increasingly popular. Nevertheless, in the present author's opinion, it is not very useful: there is no way to determine T_{fold} or T_{glass} by looking at the sequence. To determine them, one has to go to 3D, but in this case one can directly determine renaturability or any equivalent property, without resorting to any indirect criteria.

5.1.7 Where are the replicas? Although we intentionally avoid in this paper the use of the replica trick, some readers may be familiar with it and may be interested in the connections between our approach and the more standard, albeit more heavy, replica treatment. If the reader is not interested in this question, he/she should skip this section and go directly to the next Sections 5.2 and 6.

To uncover the parallelism with the replicas, let us consider the annealed average partition function $W^\star(T)$ at the special temperature defined as

$$\frac{1}{T_{\text{eff}}} = \frac{1}{T_{\text{des}}} + \frac{n}{T}, \tag{88}$$

and rewrite equation (84) in the form:

$$\left\langle \frac{E_\star(\text{seq})}{T} \right\rangle = -\left. \frac{\partial W^\star(T_{\text{eff}})}{\partial n} \right|_{n\to 0}. \tag{89}$$

As one would expect, W has the analogous form to the replicated partition function Z^n in the replica trick. Indeed, we have an annealed average and n replicated Hamiltonians. Here, we have no need to interpret n as replicas, but merely as some 'external source field' which we eventually set to zero. Therefore, not surprisingly, the results of the non-replica method agree well with those of the previous replica calculation for design in the matrix formalism [134].

Of course, to avoid difficulties is never without cost. In this case, we avoid the difficult replica calculation at the

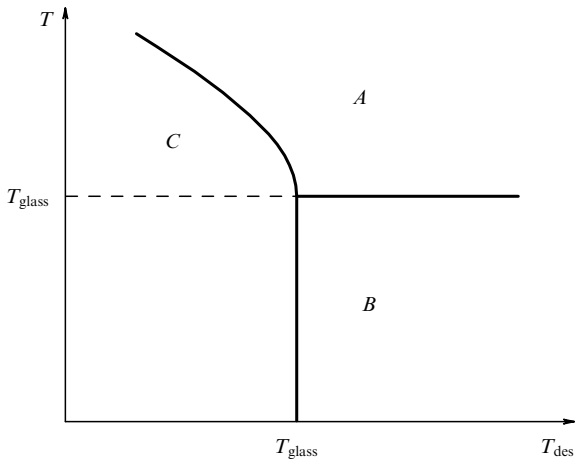

Figure 13. For the freezing of globular heteropolymers, there are three phases: A) *Random*: an exponential number of globular conformations dominate the equilibrium (similar to a homopolymer globular state); B) *Glassy*: for sequences which are not well optimized (sufficiently high T_{des}), only order one conformations dominate below the glass temperature T_{glass}, but these conformations are not the target conformation; one can consider random sequence ground states to be optimized at $T_{des} = T_{glass}$; C) *Folded*: the target conformation $\star$ dominates the equilibrium; for $T_{des} < T_{glass}$, $\star$ is better optimized than the ground state of random sequences. Note that regions $T_{des} > T_g$ and $T_{des} < T_g$ correspond to random and designed sequences, respectively. The phase boundary between phases B (glassy) and C (folded) is vertical because both are characterized by vanishing entropy and thus the transition between them cannot be caused by a temperature change.

expense of the use of REM. Also, we do not derive expression (76). In the full replica formalism, one can, at least in principle, try to go beyond the REM framework. Some attempts to apply more general models, such as GREM, are discussed in [136, 137]. In the meantime, the approach we suggest here also allows for some generalizations, as we show in the work [122].

5.1.8 Computational tests of design. There are at least two reasons to test the conclusions above with computer simulations:

1. All our results rely heavily on REM, while the validity of REM itself may be questionable.

2. As in every statistical mechanics approach, our consideration uses the thermodynamic limit. Real polymers are long, but the typical number of monomers, often as small as hundreds, is far less than in conventional applications of thermodynamics. Thus, it is desirable to test our conclusions for relatively small systems.

Extensive computational tests of both design schemes have been performed [12, 13, 138]. Computational tests indicate that design works very well indeed. What specifically should be tested computationally? If one has a prepared sequence, then the questions are as follows:

1. Is the $\star$ conformation the ground state, i.e. is it of lowest energy among all the conformations?

2. If $\star$ is the ground state, is it also a non-degenerate ground state?

3. Finally, if $\star$ is a non-degenerate ground state, is it kinetically accessible?

Not surprisingly, the situation depends significantly on the type of interactions B_{ij}. In the works [13, 138], we have systematically examined q-Potts interactions of the type $B_{ij} = J(1 - 2\delta_{ij})$ (similar monomers attract with energy $-J$, dissimilar monomers repel with energy $+J$). Note, that both the $q = 1$ and $q = N$ cases correspond to homopolymers (the latter because all monomers repel evenly)†. Therefore, one expects an optimum at some intermediate q. It was found that the optimal q is 7 for 27-mers and 9 for 36-mers. For optimal q, the yield of the chains that have $\star$ as their non-degenerate ground state was of order unity (above 50%).

If one believes that optimal q varies linearly with the chain length N, and extrapolates our results for typical protein domain length of about 100, then one arrives at the encouraging result $q \approx 20$, which is exactly the number of amino acids employed by nature. This could, of course, be just a coincidence.

As to kinetics, extensive Monte Carlo simulations have been performed. Generally, designed sequences demonstrate good ability to kinetically fold to the conformation $\star$. Of course, an appropriate temperature has to be chosen.

An impressive result was achieved in the work [135]. The author worked there with rather long chains, of up to 100 monomers. For these, there is no way to know exactly what is the ground state conformation, as enumeration of the conformations is impossible (see Section 4.6.2). Nevertheless, sequences could be designed, and Monte Carlo kinetics, starting from the expanded coil state, could be performed. The chains were indeed found to more in quite short Monte Carlo times to the conformation $\star$.

Another interesting result was as follows. It has already been mentioned, that 36-mer can have conformations with a knot (or, better to say, with a quasi-knot, as the chain does have free ends). Just for fun, we designed the sequence for which the knotted conformation was the ground state (for 36-mer, as it is enumerable, we knew the ground state exactly). It was interesting to see whether the chain would be able to fold into the knotted ground state conformation. And it did! Furthermore, the relaxation time remained moderate, only about twice as long as that for regular unknotted conformations.

In this paper, we shall not go into the simulation details and only present Fig. 14, which shows the successful comparison between the theoretical and computational phase diagrams. Note, that there are no free parameters involved in this comparison: the 7 letters Potts model was chosen for this figure because $q = 7$ is the most 'heteropolymeric' for 27-mer [13]. (Remember that both $q = 1$ and $q = 27$ correspond to homopolymers for Potts interactions: an optimum must be somewhere in between.) For other models, the behavior is similar overall, but, for example, for the Ising model ($q = 2$) more terms of the high temperature expansion are needed for qualitative results.

5.2 Designing and folding with different interactions

5.2.1 Why interactions can be different. Consider, for example, a computer simulation of protein folding or design. We understand now that folding is very sensitive to how well the native state energy is optimized, or how the sequence in question has been designed. On the other hand, one must make some approximations as to the nature of the interaction potentials involved. This directly leads to a problem: we

† To simplify matters, we do not speak of the general mean attraction effect that is always present and maintains polymers in the dense collapsed globular shape.

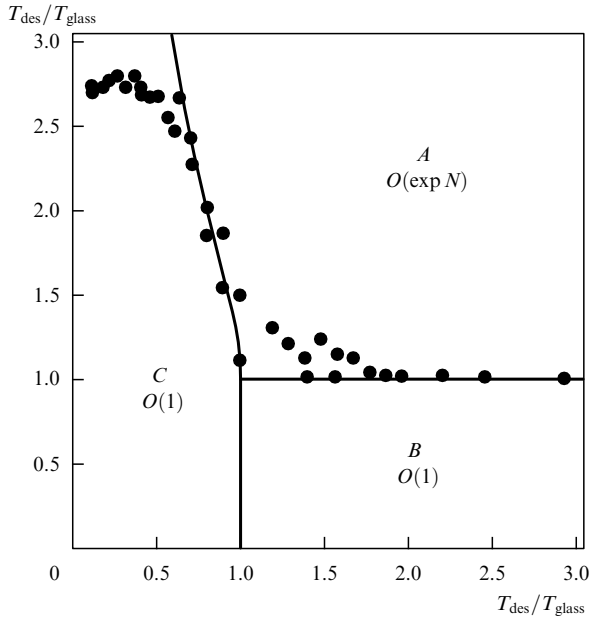

Figure 14. Phase Diagram for designed 7-Potts model heteropolymers: a computer simulation of compact 27-mers on the $3 \times 3 \times 3$ cubic lattice and an analytic prediction with no free parameters. The computer simulations generate chains using Monte Carlo annealing at a given T_{des}. Next, the partition function for maximally compact conformations was calculated exactly. The folding temperature was determined by the temperature at which the REM order parameter $x(T) = 0.9$. Due to finite system effects, one sees that the folding temperature becomes constant for $T_{des} < 0.5$, as we have reached the maximum degree of optimization for 27-mers. Also, near $T_{des}/T_{glass} \sim 1$, there are large fluctuations due to small size effects which lead to small quantitative deviations from our theory. To join the computer simulations with the analytic predictions, we measured the freezing temperature T_{glass} for random sequences directly from the simulation data at high T_{des}. With the measured value of T_{glass} and the calculated value of the variance of the interaction matrix, we plotted the theoretical curves using only the lowest order term $\mathcal{O}(1/T^2)$. Phases A, B, C are the same as in the previous Fig. 13.

believe that the sequence has been designed by nature, and this natural design was governed, obviously, by natural interactions, and we are now trying to fold this same polymer using somewhat distorted or 'noisy' interactions; or equivalently we wish to design using our approximated potentials proteins which we want to fold experimentally. If we make a good approximation for the potentials, then the results will be good, but if the potentials are sufficiently different, one would expect to arrive at spurious results. To this end, there are several questions we can ask:

(1) How different can these two sets of potentials be while still roughlyaccurately modeling protein folding?

(2) In what way do we define 'how similar' are sets of interactions?

(3) What is the phase behavior of this system?

These questions are also relevant for some other situations. For example, we can imagine that the polymer is folding under somewhat different solvent conditions compared to where it was designed.

To examine this problem explicitly, we use two different Hamiltonians: one for design

$$\mathcal{H}^{\mathrm{des}}(\mathrm{seq},\mathrm{conf}) = \sum_{I,J}^{N} B^{\mathrm{des}}_{s_I s_J}\, \Delta(\mathbf{r}_I - \mathbf{r}_J) \tag{90}$$

and another for folding

$$\mathcal{H}^{\mathrm{fold}}(\mathrm{seq},\mathrm{conf}) = \sum_{I,J}^{N} B^{\mathrm{fold}}_{s_I s_J}\, \Delta(\mathbf{r}_I - \mathbf{r}_J)\,. \tag{91}$$

We consider that (canonical) design is performed toward the target state $\star$ according to equation (81) where $\mathcal{H}^{\mathrm{des}}$ is taken for the Gibbs statistical weight.

5.2.2 Target state energy. In complete analogy with what we did in the Section 5.1.5, we can calculate the target state energy, averaged over sequences, by

$$\begin{aligned}\langle E_\star(\mathrm{seq})\rangle &= \sum_{\mathrm{seq}} P^\star_{\mathrm{seq}} E_\star(\mathrm{seq}) \\ &= \frac{\sum_{\mathrm{seq}} P^{(0)}_{\mathrm{seq}} \exp[-\mathcal{H}^{\mathrm{des}}(\mathrm{seq},\star)/T_{\mathrm{des}}]\mathcal{H}^{\mathrm{fold}}(\mathrm{seq},\star)}{\sum_{\mathrm{seq}} P^{(0)}_{\mathrm{seq}} \exp[-\mathcal{H}^{\mathrm{des}}(\mathrm{seq},\star)/T_{\mathrm{des}}]}\,.\end{aligned} \tag{92}$$

The only difference from equation (82) is that two different Hamiltonians are involved, $\mathcal{H}^d$ and $\mathcal{H}^f$. Nevertheless, we can still use our approach: we define the annealed average partition function [similar to equation (83)] for the effective Hamiltonian $\mathcal{H}_{\mathrm{eff}}$

$$\begin{aligned}W^\star &= \left\langle \exp\left[-\frac{\mathcal{H}_{\mathrm{eff}}(\mathrm{seq},\star)}{T}\right]\right\rangle \\ &= \sum_{\mathrm{seq}} P^{(0)}_{\mathrm{seq}} \exp\left[-\frac{\mathcal{H}_{\mathrm{eff}}(\mathrm{seq},\star)}{T}\right],\end{aligned} \tag{93}$$

where

$$\frac{\mathcal{H}_{\mathrm{eff}}(\mathrm{seq},\star)}{T} = \frac{\mathcal{H}^{\mathrm{des}}(\mathrm{seq},\star)}{T_{\mathrm{des}}} + n\frac{\mathcal{H}^{\mathrm{fold}}(\mathrm{seq},\star)}{T}\,. \tag{94}$$

Then, it is easy to check explicitly that the following relation is valid:

$$\frac{\langle E_\star(\mathrm{seq})\rangle}{T} = -\frac{1}{W^\star}\left.\frac{\partial W^\star}{\partial n}\right|_{n\to 0} = -\left.\frac{\partial \ln W^\star}{\partial n}\right|_{n\to 0} \tag{95}$$

[compare with (89)].

This represents the complete REM solution for the problem. Although in principle we can use a variety of approximations for the annealed average partition function $W^\star$, to be specific, we use as usual the high temperature expansion. We proceed as we did in equation (68) except with a new interaction matrix

$$\frac{\widehat{\mathcal{B}}_{\mathrm{eff}}}{T} = \frac{\widehat{B}^{\mathrm{des}}}{T_{\mathrm{des}}} + n\frac{\widehat{B}^{\mathrm{fold}}}{T}\,. \tag{96}$$

To lowest order in $1/T$ and $1/T_{\mathrm{des}}$, the annealed average free energy is independent of $\star$ (again, because all compact conformations have the same number of monomer contacts, $\mathcal{Q}$) and is given by

$$-\ln W^\star = \frac{\mathcal{Q}\overline{\mathcal{B}_{\mathrm{eff}}}}{T} - \frac{\mathcal{Q}\delta\mathcal{B}^2_{\mathrm{eff}}}{2T^2}\,, \tag{97}$$

where

$$\frac{\overline{\mathcal{B}_{\mathrm{eff}}}}{T} = \frac{\overline{B^{\mathrm{des}}}}{T_{\mathrm{des}}} + n\frac{\overline{B^{\mathrm{fold}}}}{T}\,, \tag{98}$$

$$\frac{\delta\mathcal{B}_{\rm eff}^2}{T^2}=\sum_{ij}p_i\left[\frac{B_{ij}^{\rm des}-\overline{B^{\rm des}}}{T_{\rm des}}+n\frac{B_{ij}^{\rm fold}-\overline{B^{\rm fold}}}{T}\right]^2 p_j$$
$$=\frac{\delta B_{\rm des}^2}{T_{\rm des}^2}+2n\sum_{ij}p_i\frac{\delta B_{ij}^{\rm des}}{T_{\rm des}}\frac{\delta B_{ij}^{\rm fold}}{T}p_j+n^2\frac{\delta B_{\rm fold}^2}{T^2}\,, \quad (99)$$

and thus,

$$\langle E_\star\rangle=\mathcal{Q}\left[\overline{B^{\rm fold}}-\frac{1}{T_{\rm des}}\sum_{ij}p_i\delta B_{ij}^{\rm des}\delta B_{ij}^{\rm fold}p_j\right]$$
$$=\mathcal{Q}\left[\overline{B^{\rm fold}}-\frac{\delta B_{\rm eff}^2}{T_{\rm des}}\right], \quad (100)$$

where

$$\delta B_{\rm eff}^2=\sum_{ij}p_i\widehat{\delta B}_{ij}^{\rm des}\widehat{\delta B}_{ij}^{\rm fold}p_j\,. \quad (101)$$

In this form, we draw an obvious analogy with equation (86). Moreover, we see that a particular correlator of the matrix elements is important. We next investigate a geometrical interpretation of this correlator.

5.2.3 Geometrical interpretation. The very form of our results suggests the following interpretation. Let us treat $\widehat{B}$ matrices as vectors, albeit with the components B_{ij} numbered with pair of indices. Subtracting the mean interaction, $\delta B_{ij}=B_{ij}-\overline{B}$ geometrically means that the δB_{ij} vector has zero projection along the 'main diagonal' (vector $(1,1,\ldots,1)$) in this vector space. For any two vectors in this space, say $\mathcal{A}$ and $\mathcal{B}$, we can define the scalar product as

$$\mathcal{A}\cdot\mathcal{B}=\sum_{ij}p_i\mathcal{A}_{ij}\mathcal{B}_{ij}p_j\,.$$

Note that this has nothing to do with the matrix product of the corresponding matrices.

From this point of view, both the glass transition temperature $T_{\rm glass}$ (73) and the 'susceptibility' of the target state energy for design (86) are defined by the (squared) length of the δB vector.

Now, what we found in equation (100) means that in the general case of two different matrices for folding and design, $B^{\rm fold}$ and $B^{\rm des}$, the situation depends on the angle between $\delta B^{\rm fold}$ and $\delta B^{\rm des}$. Indeed, this angle is given by $\cos\theta=g$, where

$$g\equiv\frac{\widehat{\delta B}^{\rm des}\widehat{\delta B}^{\rm fold}}{\sqrt{\left(\widehat{\delta B}^{\rm des}\widehat{\delta B}^{\rm des}\right)\left(\widehat{\delta B}^{\rm fold}\widehat{\delta B}^{\rm fold}\right)}}\,. \quad (102)$$

'Parallel' (completely correlated) interactions yield $g=1$ and 'orthogonal' (completely uncorrelated) yield $g=0$.

This geometrical view explains why we wrote the generic BWM interaction matrix in the form shown in Table 2. Indeed, in the 2×2 symmetric matrix, there are three independent elements. By setting the mean to 0 and the variance (length of the vector) to 1, we are left with only one variable, and this has a natural interpretation as an angle (θ) in the vector space. In terms of this angle θ, matrices with zero mean, unit variance, and even composition (equal fraction of all monomers, $p_i=1/2$) have the form

$$B_{ij}(\theta)=\overline{B}+\frac{\delta B}{2}\begin{pmatrix}-\sqrt{2}\cos\theta-\sin\theta & \sin\theta\\ \sin\theta & \sqrt{2}\cos\theta-\sin\theta\end{pmatrix} \quad (103)$$

which satisfies the following convenient condition

$$\widehat{\delta B}(\theta_1)\widehat{\delta B}(\theta_2)=\delta B^{(1)}\delta B^{(2)}\cos(\theta_1-\theta_2)\,. \quad (104)$$

Thus the angle θ indeed defines the metrics in this space of matrices. These results can be generalized for arbitrary p_i, but they are then cumbersome.

We can now address the following question: how good and reasonable are the models with two types of monomers, and which of them is better? We can answer this question quantitatively. Indeed, as the similarity between matrices is measured by the angle between them, we have to compute the angle $\alpha(\theta)$ between the BWM matrix that corresponds to θ and the realistic MJ matrix. To do that, we first have to resort to the reduction theorem (in the 'wrong' direction). We have to write a 20×20 matrix where the upper 10 lines are identical, and the lower ten lines are identical, and similarly for the columns. It can be shown by computation, that $\alpha(\theta)$ reaches a minimum when $\theta\approx0$, and thus the best of the BWM is HP. However, even the minimal α is as large as about 60°. It is hard to imagine the approximation which considers vectors at the angle 60° parallel, is very good.

5.2.4 Phase diagram. To find out if a polymer designed with $B^{\rm des}$ interactions will still fold correctly under $B^{\rm fold}$ interactions, we now compare the target state energy (100) with the free energy of the random globule (74). They are equal at the folding phase transition temperature $T_{\rm fold}$ which is given by

$$\frac{\widehat{\delta B}^{\rm fold}\widehat{\delta B}^{\rm fold}}{T_{\rm fold}^2}+\frac{1}{T_{\rm glass}^2}=2\frac{\widehat{\delta B}^{\rm fold}\widehat{\delta B}^{\rm des}}{T_{\rm fold}T_{\rm des}}\,. \quad (105)$$

Thus, the original phase diagram (Fig. 13) is modified simply by rescaling of $T_{\rm des}$.

5.2.5 How accurate must potentials be for successful modeling of protein folding? An article under this title was recently published [139]. (An alternative title could 'How different can the languages used by the writer and reader be?') Now we can derive the results in much simpler way. The result is that the folding remains correct ('the reader understands') as long as $T_{\rm des}$ and T remain in the native state phase below the transition (105). When the error angle increases, this condition eventually breaks down. Thus, the maximal acceptable error value depends on both $T_{\rm des}$ and T, but does not depend on N. This is a fundamental advantage of designed sequences. For random sequences, the ground state changes when parameters are shifted by a tiny amount, about $N^{-1/2}$ [130].

Let us characterize the degree of optimization of a protein with the value $T_{\rm fold}/T_{\rm glass}=x$, and the conditions at which this protein is supposed to function with $T/T_{\rm glass}=y$. Then the necessary value of g is given by

$$g\geqslant\frac{1+y^2}{y}\frac{x}{1+x^2}\,. \quad (106)$$

For instance, if $x=1.1$ and $y=1.05$ (which is plausible), then $g\geqslant0.997$. Even if $x=1.2$ and $y=1.1$, then $g\geqslant0.988$. The corresponding angles are not to exceed 1° or 9°, respectively.

Of course, the message here is not in numbers (which are somewhat arbitrary), but in physics: the necessary accuracy is controlled, on the one hand, by how far the functioning protein is from its freezing (y), and on the other, by how good the sequence is (x).

6. Statistics of real protein sequences

6.1 Individual vs. statistical approach

6.1.1 The statistical approach does not mean that sequences are generated letter by letter. Chemical synthetic procedures imply that the sequences of chemical units in copolymers follow simple Markovian statistics [127]. Clearly, Markovian processes produce a variety of different sequences; they have identical mean compositions and local correlations, but in detail they are different. On the other hand, it is the α and ω of molecular biology that the sequence of each type of biopolymer, be it protein, DNA, or RNA, is completely specified and is under strict genetic control. Apart from some relatively rare mistakes of genetic apparatus, all, say, human hemoglobin chains are of the same sequence. They are specified to the same extent and in the same sense as the sequences of letters in novels, such as *War and Peace*, or *Crime and Punishment*. This gives rise to the approach that is universally adopted among biologists — to study each particular biopolymer individually. Many people also believe that as biological function is extremely sensitive to complex molecular details, simple physical models based on universal considerations cannot provide useful guidelines for biologists.

In the meantime, when we consider the physical properties of biopolymers, it is tempting to resort to methods and approaches that proved fruitful in the physics of disordered systems such as spin glasses. The central idea there is just the opposite to what was just said about biopolymers. As individual structure of a piece of glass is totally out of control, and can never be reproduced, the only sensible questions are those about appropriate average values. We are not interested in individual properties of a given sample, but in those properties that will be reproduced in virtually every sample of the material prepared under similar conditions.

We have demonstrated in the previous sections that the 'averaging' approach indeed appears fruitful. Although its possibilities should not be overestimated, it can answer general questions, like why unique native state can exist and which sequences should be selected for them.

6.1.2 The design of sequences does not work on the level of local correlations along the sequence. The statistical approach does imply that the sequences are thought of as randomly fished out of a pool. If the pool consists of all possible sequences, then random fishing from this pool is obviously equivalent to random symbol-by-symbol generation, as in the famous example of the typing monkey†. We have demonstrated that the overwhelming majority of these sequences do not mimic proteins reasonably. Moreover, we have suggested other statistical ensembles of designed sequences (81) — those obtained by canonical design at the temperature T_{des}. Note that design schemes employ the three dimensional properties of polymers, and thus cannot be reduced to any Markovian model that implies local correlations along the chain. Nevertheless, the ensemble of designed sequences differs, of course, from the ensemble of all (random) sequences even when viewed purely sequentially, without thinking of a three dimensional folding.

The theory of sequence design presented above yields certain predictions as to what type of non-randomness we should look for, and this prediction can be tested using data banks of DNA and protein sequences. What is the prediction? Qualitatively, it is very simple. First of all, the prediction is that the properties of amino acids that exhibit correlations should be related to the various 'charges' defining physical interactions between protein chain links. Secondly, if we remember the way, say, imprinting works, we see that those monomers that attract each other in volume interactions have a good chance of becoming close along the chain, and vice versa, those repelling each other have a good chance of not becoming chemical neighbors. These predictions seem to be trivial, as they correspond to normal physical intuition. We stress that they are not trivial at all. Indeed, the sequences are the product of evolution, which seems to have nothing to do with any physical interactions between chain monomers. Nevertheless, we predict that the chain correlations should look like the sequence is formed by simple physics. In the next section, we describe the way to test this prediction.

6.1.3 Protein sequences: are they random? This question has a long history. It was first believed that the proteins with their sequences represent a unique best fit to their respective functions. It was later realized, that neither the time nor the material was sufficient for evolution to make this optimization. It is so clear that it was explained even in the popular book [140]. Thus the opposite conclusion has been formulated: proteins were said to be 'slightly edited random copolymers,'‡ with the 'editing' done in the vicinity of the active site [11] (see also [141]). This conclusion was in agreement with simple statistical tests that portray protein sequences as almost random. However, it is not at all easy to formulate *a priori* a statistical test that will be able to differentiate random and non-random sequences [142, 143].

One may ask if real protein sequences look random, or if they bear some fingerprint of the evolutionary design they have undergone.

6.2 The random walk technique to study the statistics of sequences

A good and powerful technique to examine the statistics of sequences is to map the sequence onto the trajectory of an appropriate random walk. As far as this author knows, this idea was suggested for the first time by I M Lifshitz in the context of his study of the helix-coil transition in DNA [144]. Statistical analysis of DNA sequences is simpler than for proteins (because the alphabet is of 4 letters only, and because the sequences are longer), and this is why we first explain the technique for this example.

6.2.1 The random walk technique for DNA sequences. The following procedure can be used to map a DNA sequence onto a trajectory of a 1D random walk. Suppose that a walker

† By the way, nobody seems to have carried out the experiment with a real monkey. Meanwhile, there are grounds to expect that the monkey, once given an opportunity, will produce fractal, not random sequences.

‡ 'A Poet wants to give the reader clean water from the spring of Poetry, but cannot do it before the editor washes there' (*M Svetlov*).

steps down at the 'time' t if the monomer number t in DNA is purine, and it steps up otherwise; $\xi_t = -1$ if t is purine (Pu), i.e. adenine (A) or guanine (G), and $\xi_t = +1$ if t is pyrimidine (Py), i.e. cytosine (C) or thymine (T). The value of the end-to-end distance for the walker trajectory between points s and s' on the DNA is then defined as†

$$x(s, s') = \sum_{t=s}^{s'} \xi_t, \tag{107}$$

where $\xi_t = \pm 1$ are 'monomer contributions' defined as explained above. A useful function to be defined in terms of the DNA walk is

$$F(l) = \left[\left\langle x^2(s, s+l)\right\rangle_s - \left\langle x(s, s+l)\right\rangle_s^2\right]^{1/2}, \tag{108}$$

where $\langle ... \rangle_s$ means the average over the entire length of DNA, $1 \leqslant s \leqslant L$ (the so-called 'sliding window' average). Of course, $F(l)$ is simply the root mean square purine-pyrimidine difference in the l-length part of the sequence.

A lot of *different* DNA four-letter texts can be imagined with the same Pu-Py sequence. Generally, a DNA text can be one-to-one mapped onto a 2D plane, or onto the complex plane: for example, steps right, down, left and up, or $\xi_t = +1, -\mathrm{i}, -1, +\mathrm{i}$ correspond to T, G, A, C, respectively. In this case, the 1D map is simply the projection of our plane on the main diagonal.

The claim of the work [145] was that

$$F(l) \sim l^{\alpha}, \tag{109}$$

and two distinct sets of sequences were found, with $\alpha \approx 0.5$ (which is trivially expected) and 0.6–0.7. If this is true, it brings the entire problem of the DNA structure, function, and evolution in the realm of fractal science.

6.2.2 Tne Brownian bridge technique for protein sequences‡. As there are 20 types of monomer species in protein sequences, there is considerable freedom in how to map a sequence onto the walk. In other words, the question is how to choose monomer contributions ξ_i depending on the chemical entity of the monomer i. Another aspect of the problem is that proteins are rather short, and the sliding window average technique, which is employed for DNA, is inapplicable. Instead, it was suggested in the work [14] to carry out an averaging over the different proteins in the data base. More specifically, it was done in the following manner.

Three distinct mappings were tested, related to major physical interactions between monomers in proteins. For each of the amino-acid sequences obtained from the Data Bank [146], were the trajectories of three different artificial walkers, each related to a kind of physical interactions between residues — hydrophobic (A), hydrogen bonds (B), and Coulomb (C). The subsequent steps of each walker are given by the numbers $\{\xi_i\}$ defined as

A. $\xi_i = +1$ if monomer number i in the given sequence is highly hydrophilic (Lys, Arg, His, Asp, Glu) or $\xi_i = -1$ otherwise;

B. ξ_i may be $+1$ or -1 for monomers capable (Asn, Gln, Ser, Thr, Trp, Tyr) or not capable (all others) of hydrogen bonding;

C. ξ_i may be $+1$, -1 or 0 for positively (Lys, Arg, His) or negatively charged (Asp, Glu) and neutral (all others) monomers i, respectively.

In order to look for correlations by comparing the trajectories, one has to exclude the dependencies on protein length, overall composition and the step size of the walker. This is done by the following definition of trajectories. We start with a given ensemble of protein sequences. The decoded sequence $\{\xi_1, \xi_2, \ldots, \xi_L\}$is mapped onto the trajectory as

$$x(l) = \sum_{t=1}^{l} \xi_t\,. \tag{110}$$

The walker defined by equation (110) may have a strong drift, so that the leading term in $x(l)$ might be linear in l; this is simply related to the mean composition of the chain considered. Since the overall composition is beyond our interest here, we define the reduced trajectory:

$$y(l) = x(l) - \frac{l}{L}\,x(L)\,, \tag{111}$$

L being the total number of links in the entire polymer chain. Obviously, the y-walker returns to the origin after the entire 'trip.' The corresponding trajectory $y(l)$ is called a 'Brownian bridge.'

In order to collect all the data in a comparable form, it is necessary to rescale all the Brownian bridges compensating for different proteins with different lengths and variances of ξ distribution, by

$$z^2(\lambda) = \frac{y^2}{L\overline{(\xi - \overline{\xi})^2}}\,, \tag{112}$$

where $\overline{(\ldots)}$ means averaging over a given protein sequence e. g.

$$\overline{\xi} = \frac{1}{L}\sum_{i=1}^{L} \xi_i\,, \tag{113}$$

and to exclude L-dependence, we rescale the number of steps taken (l) as $\lambda = l/L$, where $0 \leqslant \lambda \leqslant 1$.

With the rescaled trajectories $z^2(\lambda)$, averaging can be performed over the ensemble of proteins:

$$r(\lambda) = \left\langle z^2(\lambda)\right\rangle_{\text{ensemble}}\,. \tag{114}$$

Combining equations (110) through (114), yields

$$r(\lambda) \equiv \left\langle \left(\sum_{i=0}^{[\lambda L_p]} \frac{\Delta\xi_i^{(p)}}{\sigma^{(p)}} \right)^2 \right\rangle_p\,, \tag{115}$$

where p denotes a given protein, $\langle \ldots \rangle_p$ means average over the set of proteins, $[\ldots]$ means take the next highest integer, and L_p is the total number of amino acids in p.

Looking at the walk representations of the sequences, how can one judge about their randomness or non-randomness? A purely random walker, which corresponds to a random sequence, is expected to travel about $\sigma\sqrt{L}$ from the origin on mean-square-average and it is easy to find

$$r_{\text{rand}}(\lambda) = \left[\lambda^{-1} + (1-\lambda)^{-1}\right]^{-1}. \tag{116}$$

As it must come back in the end, to reach farther, it must go mainly in one direction for the first half-time ($i < L/2$) and

† Note that the values ξ_t for complementary bases differ in sign only, and thus the $F(l)$ are the same for complementary threads.

‡ This section is based on the work [14].

mainly back in the second half-time ($i > L/2$) thus approaching the maximal distance of $\sigma L/2$. On the other hand, to keep as close to the origin as possible, it must compensate each step to one direction by a subsequent opposite step. Therefore, persistent types of correlations in protein sequences would be manifested in trajectories which go beyond the random one, while alternating correlations would lead to trajectories which do not travel as far.

The trajectories $r_A(\lambda)$, $r_B(\lambda)$, $r_C(\lambda)$, along with the theoretically found trajectory (116) for the purely random case, are shown in the Fig. 15 for a set of globular proteins (those coded as catalysts in the Data Bank). The $r_A(\lambda)$ and $r_B(\lambda)$ bridges are clearly over $r_{rand}(\lambda)$ manifesting pronounced persistent correlations in the distribution of hydrophobicity. Alternating correlations are found between electrical charges on protein chains because $r_C(\lambda)$ is definitely below $r_{rand}(\lambda)$.

The general qualitative behavior is seen in Fig. 15, showing persistent correlations for hydrophobic and hydrogen bond-related mappings, and anti-correlations for Coulomb mapping, and is in perfect agreement with the predictions drawn from the design of sequences model. The data for r_A, r_B, and r_C are shown to fit very well to the phenomenological interpolation

$$r(\lambda) = \frac{L_0^{2\alpha-1}}{\lambda^{-2\alpha} + (1-\lambda)^{-2\alpha}}, \qquad (117)$$

where we chose $L_0 = 110$ and obtained $\alpha_A = 0.520 \pm 0.05$, $\alpha_B = 0.520 \pm 0.05$, $\alpha_C = 0.470 \pm 0.05$. The corridors shown demonstrate that the fit does indeed yield this small error for α. Formally, α in Eqn (117) is similar to the power in equation (109). However, fractal interpretation seems fruitless, as we are speaking of the average over many different proteins. Nevertheless, the fact that α appears close to 0.5 reflects the fact that the sequences are not very far from random; they are edited only slightly. It is thus not a surprise that these deviations from randomness were difficult to find. It is therefore even more important that the character of the correlations does follow theoretical prediction.

Further interesting data are given in the work [14] on Brownian bridges for different groups of proteins. Further development in this direction is reported also in [148].

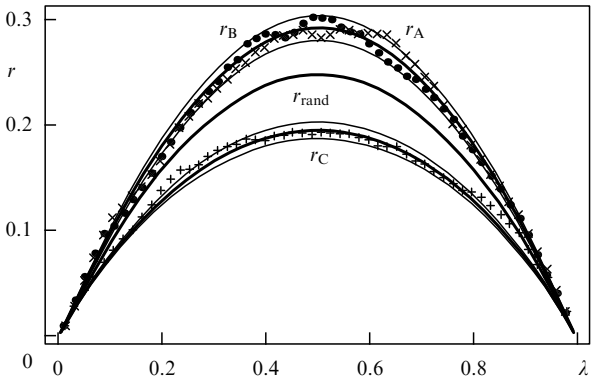

Figure 15. Brownian bridges for hydrophilic (×), hydrogen bonding (•), and coulomb (+) mappings of sequences of prokaryote proteins with catalytic activity, and therefore globular structure. The thick curve corresponds to complete randomness r_{rand} (116).

6.3 The design of sequences and models of evolution

One intriguing application of the design of sequences is to try to model possible scenarios of early prebiotic evolution. It was clear a long time ago, that the problem of protein folding is deeply connected to the problem of early prebiotic evolution and the formation of the first 'biopolymers' in the primordial soup [11, 149, 150]. It was discussed in [151] that this development could be governed by conflicting requirements, such as those of compactness at the same time as solubility. These two requirements are indeed conflicting: if the polymer is compact, we expect that its monomers should attract, but then the polymer chains tend to aggregate; if the polymers are easily dissolved, we think that the monomers repel, but then each chain swells. Until now, we have not been able to materialize this idea. The first step in this direction was attempted very recently in the works [152]. The idea of implementing two conflicting requirements was realized in the following way: a polymer was considered with two types of monomers, hydrophobic and polar (HP model), and the sequences were designed such that the percentage of hydrophobic monomers was given an upper boundary, mimicking solubility, while the sequence was required to yield relatively fast contraction. It was implemented in a computer evolution-like experiment: starting from a random sequence, the authors made random mutations, with the condition of preservation of the overall composition, and looked at the Monte Carlo compaction kinetics: if the kinetics improved, the mutation was accepted, and it was rejected otherwise. An interesting finding is that this procedure led to the sequences with unique ground states and well pronounced energy minima in that state.

Finally, let us also mention another impressive way to look at the evolution data using the idea of design. Suppose we design the sequence for the ground state conformation $\star$. Clearly, we can prepare many sequences for that same conformation. We can then look at which monomer units are preserved throughout those sequences, and which monomers are variable. We can then compare this with what biologists see about variability in real protein sequences taken from similar proteins in different species. A very promising similarity between the two patterns was found in the recent work [153].

7. Conclusions

To conclude, let us return to the question of the relation between the models considered and real biopolymers. First, let us repeat that all the small scale details, however important [104, 105], are renormalized out of the simple models. Furthermore, if one uses lattice models of proteins, typically one considers that the lattice polymer is some sort of 'renormalized' protein, with secondary structure coarse grained out; indeed, the lattice polymer conformation therefore describes purely tertiary structure, i.e. an arrangement of secondary structural elements. This is why 27-mer [154] and perhaps even 18-mer can adequately model short protein folding behavior. This is despite the fact that these domains are typically comprised of about 70 to 100 residues. Moreover, *de novo* designed 4-helix bundles [120] look somewhat similar to lattice 12-mers ($2 \times 2 \times 3$). In the meantime, lattice 36- and 48-mers can possibly model multi-domain proteins [155]. This may be associated with the fact that 36-mer is the shortest for which pseudo-knots are possible. Not surprisingly, 36-mer is also the shortest (of lattice polymers) for

which the globule can have a domain structure, with two 18-mers filling the $2 \times 3 \times 3$ domains. Thus, the thermodynamic limit is by no means the most interesting regime here.

Clearly, simple models cannot pretend to explain all the properties of biopolymers. As an important example we mention that proteins are known to function in a machine-like fashion [156], as systems that have their own constructions [157]. Is it sufficient to design sequences in the way described earlier in the paper to get molecules capable of this machine-style functioning, or does one need to do something more? This important question remains an unanswered challenge. In the meantime, the models described do seem to capture at least some of the exciting properties of biopolymers. For example, it is an *experimental fact* (albeit a fact of computer experiment) that lattice 100-mer is able to renature if its sequence is properly designed [135] — just like natural proteins do. Therefore, the Levinthal paradox exists for lattice model polymers to the same extent and in the same form, as for natural proteins, and it is equally challenging. It is hard to believe that the solution of the Levinthal paradox would be different for proteins and for lattice polymers. Thus, the models we were speaking about are at least interesting.

Has the Levinthal paradox been resolved at least at the level of models? In the present author's opinion — no, not yet. Even though we believe that the idea of sequence design with the optimization of the ground state energy moves us in the right direction, and the discovery of correlations in protein sequences favors this optimism, sequence design so far optimizes only the thermodynamic properties of model chains. We do not have sufficient understanding of folding kinetics. In principle, one might expect that another optimization would be needed to overcome the kinetic barrier. An alternative scenario, however, is that thermodynamic optimization somehow allows for kinetic optimization as well. The resolution of this problem seems to be one of the most interesting questions. The present author believes that topological constraints and secondary structures could be two closely related keys to open this lock.

Acknowledgements. This review is largely based on the works that the author performed over recent years together with A Gutin, V Pande, Y Rabin, E Shakhnovich, T Tanaka. I am deeply grateful to all of them for sharing their insights and enthusiasm.

References

1. Lifshits I M *Zh. Eksp. Teor. Fiz.* **55** 2408 (1968) [*Sov. Phys. JETP* **28** 1280 (1968)]
2. Doi M, Edwards S F *The Theory of Polymer Dynamics* (Oxford: Oxford University Press, 1987)
3. De Gennes P G *Scaling Concepts in Polymer Physics* (Ithaca, NY: Cornell University Press, 1979) [Translated into Russian (Moscow: Mir, 1982)]
4. Lifshits I M, Gredeskul S A, Pastur L A *Vvedenie v Teoriyu Neuporyadochennykh Sistem* (Introduction to the Theory of Disordered Systems) (Moscow: Nauka, 1982) [Translated into English (New York: Wiley & Sons, 1988)]
5. Anfinsen C *Science* **181** 223 (1973)
6. Derrida B *Phys. Rev. Lett.* **45** 79 (1980)
7. Bryngelson J D, Wolynes P G *Proc. Nat. Acad. Sci. USA* **84** 7524 (1987)
8. Wolynes P G, in *Spin Glass Ideas in Biology* (Ed. D Stein) (Singapore: World Scientific, 1991) p. 1
9. Shakhnovich E, Gutin A *Biophys. Chem.* **34** (3) 187 (1989)
10. Pande V S, Grosberg A Yu, Tanaka T *Biophys. J.* (in press)
11. Ptitsyn O B, Volkenstein M V *J. Biomolec. Struct. and Dynamics* **4** 137 (1986)
12. Shakhnovich E, Gutin A *Proc. Nat. Acad. Sci. USA* **90** 7195 (1993)
13. Pande V S, Grosberg A Yu, Tanaka T *Proc. Nat. Acad. Sci. USA* **91** 12976 (1994)
14. Pande V S, Grosberg A Yu, Tanaka T *Proc. Nat. Acad. Sci. USA* **91** 12972 (1994)
15. Panyukov S, Rabin Y *Phys. Rep.* **269** 1 (1996)
16. Bates F, Fredrickson G H *Ann. Rev. Phys. Chem.* **41** 525 (1990)
17. Erukhimovich I, Khokhlov A *Polymer Sci.* **35** 1522 (1993)
18. Binder K *Adv. Polymer Sci.* **112** 181 (1994)
19. Grosberg A Yu, Kaganova E M, Molchanov S A *Biofizika* **29** 30 (1984) [*Biophysics* **29** 27 (1984)]
20. Grosberg A Yu *J. Stat. Phys.* **38** 149 (1985)
21. Grosberg A Yu, Kaganova E M *Dokl. Acad. Nauk* **294** 838 (1987); [*Sov. Phys. Dokl.* **36** 474 (1987)]
22. Grosberg A Yu, Shakhnovich E I *Zh. Exp. Theor. Fiz.* **91** 837 (1986) [*Sov. Phys. JETP* **64** 493 (1986)]
23. Grosberg A Yu, Shakhnovich E I *Zh. Eksp. Teor. Fiz.* **91** 2159 (1986) [*Sov. Phys. JETP* **64** 1284 (1986)]
24. Grosberg A Yu, Shakhnovich E I *Biofizika* **31** 1045 (1986) [*Biophysics* **31** 1139 (1986)]
25. Joanny J-F *J. de Phys II* **4** 1281 (1994)
26. Garel T et al. *Europhys Lett.* **8** 9 (1989)
27. Grosberg A Yu, Izrailev S F, Nechaev S K *Phys. Rev. E* **50** 1912 (1994)
28. Sommer J-U, Daoud M *Europhys. Lett.* **32** 407 (1995)
29. De Gennes P G "*Chrysanthemums*": *Weak Micellisation* Preprint (1995)
30. Vedenov A A, Dykhne A M, Frank-Kamenetskiĭ M D *Usp. Fiz. Nauk* **105** 479 (1971) [*Sov. Phys. Usp.* **14** 715 (1972)]
31. Azbel M Ya *Biopolymers* **19** 61, 81, 95, 1311 (1980); Azbel M et al. *Biopolymers* **21** 1687 (1982)
32. Grosberg A Yu, Khokhlov A R *Statisticheskaya Fizika Makromolekul* (Statistical Physics of Macromolecules) (Moscow: Nauka, 1989) [Translated into English (New York: AIP Press, 1994)]
33. Trifonov E N, Tan R K Z, Harvey S C, in *Structure and Expression* (Eds M H Sarma, R H Sarma) (Schenectady, NY: Adenin Press, 1988) p. 243
34. Bednar J et al. *J. Mol. Biol.* **254** 579 (1995)
35. Higgs P G, Joanny J-F *J. Chem. Phys.* **94** 1543 (1991)
36. Kantor Y, Li, Kardar M *Phys. Rev. Lett.* **69** 61 (1992); Kantor Y, Kardar M *Europhys. Lett.* **27** 643 (1994); Kantor Y, Kardar M, Li H *Phys. Rev. E* **49** 1383 (1994); Kantor Y, Kardar M *Phys. Rev. E* **51** 1299 (1995); Kantor Y, Kardar M *Phys. Rev. E* **52** 835 (1995); Ertas D, Kantor Y *Phys. Rev. E* (1996) (in press)
37. Wittmer J, Johner A, Joanny J F *Europhys. Lett.* **24** 263 (1993)
38. Dobrynin A V, Rubinstein M *J. de Phys. II* **5** 677 (1995)
39. Gutin A M, Shakhnovich E I *Phys. Rev. E* **50** R3322 (1994)
40. Grosberg A Yu *Biofizika* **29** 569 (1984) [*Biophysics* **29** (1984)]
41. Garel T, Leibler L, Orland H *J. de Phys. II* **4** 2139 (1994)
42. Marko J F, Siggia E D *Macromolecules* **27** 981 (1994); *Science* **265** 506 (1994); *Phys. Rev. E* **52** 2912 (1995)
43. Frank-Kamenetskiĭ M D *Samaya Glavnaya Molekula* (Unraveling) (Moscow: Nauka, 1989) [Translated into English (New York: DNA, VCH, 1993)]
44. Zimm B H, Stockmayer W H *J. Chem. Phys.* **17** 1301 (1949)
45. Gutin A M, Grosberg AYu, Shakhnovich E I *Macromolecules* **26** 1293 (1993)
46. Isaacson J, Lubensky T C *J. de Phys. Lett.* **41** L469 (1980)
47. Daoud M et al. *Macromolecules* **16** 1833 (1983)
48. Parisi G, Sourlas N *Phys. Rev. Lett.* **46** 871 (1981)
49. Daoud M, Joanny J-F *J. de Phys.* **42** 1359 (1981)
50. Gaunt D, Flesia S *J. Phys. A* **24** 3655 (1991); Janse van Rensburg E, Madras N *J. Phys. A* **25** 303 (1992)
51. Shi-Min Cui, Zheng Yu Chen *Phys. Rev. E* **52** 5084 (1995)
52. Shi-Min Cui, Zheng Yu Chen *Phys. Rev. E* **53** 6238 (1996)
53. Grosberg A Yu, Gutin A M, Shakhnovich E I *Macromolecules* **28** 3718 (1995)
54. Kelvin, Lord *Trans. Roy. Soc. of Edinburg* **25** 217 (1868)
55. Wu F Y *Rev. Mod. Phys.* **64** 1099 (1992); **65** 577 (1993)
56. Kauffman L H *Knots and Physics* (Singapore: World Scientific, 1991)

57. Atiyah M *The Geometry and Physics of Knots* (Cambridge: University Press, 1990) [Translated into Russian (Moscow: Mir, 1995)]
58. Moffatt H K *Nature* (London) **347** 367 (1990)
59. Frisch H L, Wasserman E *J. Am. Chem. Soc.* **83** 3789 (1961)
60. Delbrück M *Mathematical Problems in the Biological Sciences, Proc. of Symposium in Applied Mathematics* (American Mathematical Society, Providence RI) **14** 55 (1962)
61. Liang C, Mislow K *J. Am. Chem. Soc.* **116** 11189 (1994); **116** 3588 (1994); *J. Math. Chem.* **15** 245 (1994)
62. Sumners D W, Whittington S G *J. Phys. A* **21** 1689 (1988)
63. Pippenger N *Discrete Appl. Math.* **25** 273 (1989)
64. Nechaev S K *J. Phys. A* **21** 3659 (1988)
65. Nechaev S *Statistics of Knots and Entangled Random Walks* (Singapore: World Scientific, 1996)
66. Nechaev S K, Grosberg A Yu, Vershik A M *J. Phys. A* **29** 2411 (1996)
67. Edwards S F *Proc. Phys. Soc.* **91** 513 (1967)
68. Prager S, Frisch H L *J. Chem. Phys.* **46** 1475 (1967)
69. Khokhlov A R, Nechaev S K *Phys. Lett. A* **112** 156 (1985)
70. Vologodskiĭ A V, Lukashin A V, Frank-Kamenetskiĭ M D *Zh. Eksp. Teor. Fiz.* **67** 1875 (1974) [*Sov. Phys. JETP* **40** 932 (1975)]
71. Koniaris K, Muthukumar M *Phys. Rev. Lett.* **66** 2211 (1991)
72. Deguchi T, Tsurusaki K *J. of Knot Theory and its Ramifications* **3** 321 (1994)
73. Grosberg A Yu, Feigel A, Rabin Y *Phys. Rev. E* (in press)
74. Grosberg A Yu, Nechaev S K *Adv. Polymer Sci.* **106** 1 (1993)
75. Drossel B, Kardar M *Phys. Rev. E* **53** 5861 (1996)
76. Frank-Kamenetskiĭ M D, Vologodskiĭ A V *Phys. Usp.* **134** 641 (1981) [*Sov. Phys. Usp.* **24** 679 (1981)]; Vologodskiĭ A V *Topology and Physics of Circular DNA* (Boca Raton: CRC Press, 1992)
77. Tait P G *Scientific Papers* **1** 273 (1898)
78. Rolfsen D *Knots and Links* (Berkeley: Publish or Perish Press, 1976)
79. Atiyah M *Rev. Mod. Phys.* **67** 977 (1995)
80. Crowell R H, Fox R H *Introduction to Knot Theory* (Ginna, Boston, 1963) [Translated into Russian (Moscow: Mir, 1965)]
81. Jones V F R *Bull. Am. Math. Soc.* **12** 103 (1985)
82. Katritch V et al. *Nature* (1996) (in press)
83. Klenin K V et al. *J. Biomolec. Struct. and Dynamics* **5** 1173 (1988)
84. Liu L F, Depew R E, Wang J C *J. Molec. Biology* **106** 439 (1976)
85. Rybenkov V V, Cozzarelli N R, Vologodskii A V *Proc. Nat. Acad. Sci. USA* **90** 5307 (1993)
86. Shaw S Y, Wang J C *Science* **260** 533 (1993)
87. Tesi M C et al. *J. Phys. A* **27** 347 (1994)
88. Grosberg A Yu, Nechaev S K *J. Phys. A* **25** 4659 (1992); *Europhys. Lett.* **20** 613 (1992)
89. Grosberg A Yu, Nechaev S K, Shakhnovich E I *J. de Phys.* **49** 2095 (1988)
90. Jianpeng Ma, Straub J E, Shakhnovich E I *J. Chem. Phys.* **103** 2615 (1995)
91. Timoshenko E et al. *J. Chenm. Phys.* (in press)
92. Chu B, Ying Q, Grosberg A Yu *Macromolecules* **28** 180 (1995)
93. Chu B, Ying Q *Macromolecules* **29** 1824 (1996)
94. Grosberg A Yu, Kuznetsov D V *Macromolecules* **26** 4249 (1993)
95. Ganazzoli F, Ferla R La, Allegra G *Macromolecules* **28** 5285 (1995)
96. Grosberg A Yu et al. *Europhys. Lett.* **23** 373 (1993)
97. Sikorav J -L, Jannink G *C.R. Acad. Sci. Paris* **316** 751 (1993)
98. Sikorav J -L, Church G M *J. Mol. Biol.* **222** 1085 (1991)
99. Janse van Rensburg E J, Whittington S G *J. Phys. A* **24** 3935 (1991)
100. Cates M E, Deutsch J M *J. de Phys.* **47** 2121 (1986)
101. Semenov A N *J. de Phys.* **49** 175 (1988)
102. De Gennes P G *Macromolecules* **17** 703 (1984); reprinted in *Simple Views on Condensed Matter* (Singapore: World Scientific, 1992) p. 215
103. Rabin Y, Grosberg A Yu, Tanaka T *Europhys. Lett.* **32** 505 (1995)
104. Frauenfelder H, Wolynes P G *Physics Today* **47** (2) 58 (1994)
105. Goldanskii V I *Tunneling Phenomena in Chemical Physics* (New York: Gordon and Breach Sci. Publ., 1989)
106. Finkelstein A V, Gutin A M, Badretdinov A Ya *FEBS Lett.* **325** 23 (1993)
107. Finkelstein A V, Badretdinov A Ya, Gutin A M *Proteins* **23** 142 (1995)
108. Finkelstein A V, Gutin A M, Badretdinov A Ya *Proteins* **23** 151 (1995)
109. Miyazawa S, Jernigan R *Macromolecules* **18** 534 (1985)
110. Godzik A, Kolinski A, Skolnick J *Protein Sci.* (*BNW*) **10** 2107 (1995)
111. Goldenfeld N *Lectures on phase transitions and the renormalization group* (Reading, Mass.: Addison-Wesley Press, 1992)
112. Lau K F, Dill K A *Macromolecules* **22** 3986 (1989)
113. Obukhov S P *Phys. Rev. A* **42** 2015 (1990)
114. Garel T, Leibler L, Orland H *J. de Phys. II* **4** 2139 (1994)
115. Li H et al. *Science* **273** 666 (1996)
116. 27.8° is indeed the correct value of electro-weak mixing angle. The author hopes that the reader did not lose his sense of humor...
117. Mezard M, Parisi G, Virasoro M *Spin Glass Theory and Beyond* (Singapore: World Scientific, 1987)
118. Koukiou F *J. Phys. A* **26** L1207 (1993)
119. Pande V S, Grosberg A Yu, Tanaka T *Phys. Rev. Lett.* **76** 3987 (1996)
120. Hecht M H et al. *Science* **249** 884 (1990)
121. Editorial commentary *Science* **269** 1821 (1995)
122. Pande V S, Grosberg A Yu, Tanaka T *J. Chem. Phys.* (in press)
123. Pande V S, Grosberg A Yu, Tanaka T *Phys. Rev. E* **51** 3381 (1995)
124. Sfatos C D, Gutin A M, Shakhnovich E I *Phys. Rev. E* **48** 465 (1993)
125. Shakhnovich E, Gutin A *J. Chem. Phys.* **93** 5967 (1990)
126. Pande V S et al. *J. Phys. A* **27** 6231 (1994)
127. Flory P *Principles of Polymer Chemistry* (Ithaca: Cornell University Press, 1953)
128. Garel T, Orland H, Thirumalai D Preprint (1995)
129. Gutin A M, Shakhnovich E I *J. Chem. Phys* **98** 8174 (1993)
130. Bryngelson J D *J. Chem. Phys.* **100** 6038 (1994)
131. Yue K, Dill K A *Proc. Nat. Acad. Sci. USA* **89** 4163 (1992)
132. Abkevich V, Gutin A, Shakhnovich E *J. Mol. Biol.* **252** 460 (1995)
133. Ramanathan S, Shakhnovich E I *Phys. Rev. E* **50** 1303 (1994)
134. Pande V S, Grosberg A Yu, Tanaka T *Macromolecules* **28** 2218 (1995)
135. Shakhnovich E I *Phys. Rev. Lett.* **72** 3907 (1994)
136. Derrida B *J. de Phys. Lett.* **46** L401 (1985)
137. Plotkin S S, Wang J, Wolynes P G *Phys. Rev. E* **53** 6271 (1996)
138. Pande V S, Grosberg A Yu, Tanaka T *J. Chem. Phys.* **101** 8246 (1994)
139. Pande V S, Grosberg A Yu, Tanaka T *J. Chem. Phys.* **103** 9482 (1995)
140. Grosberg A Yu, Khokhlov A R *Fizika v Mire Polimerov* (Physics in the World of Polymers) (Moscow: Nauka, 1989)
141. Monod J *Chance and Necessity: an Essay on the Natural Philosophy of Modern Biology* (New York: Vintage Books, 1971)
142. Tupygin A Yu, Chechetkin V R *Zh. Eksp. Teor. Fiz.* **106** 335 (1994) [*JETP* **79** 186 (1994)]
143. Chechetkin V R, Turygin A Yu *Phys. Lett. A* **199** 75 (1995); *J. Theor. Biol.* **175** 477 (1995); **178** 205 (1996)
144. Lifshitz I M *Zh. Eksp. Teor. Fiz.* **65** 1100 (1973) [*Sov. Phys. JETP* **38** 546 (1974)]
145. Peng C -K et al. *Nature* **356** 168 (1992)
146. Bairoch A, Boeckmann B *Nucleic Acids Res.* **20** 2019 (1992)
147. Pande V S Ph. D. Thesis (MIT, 1995)
148. Strait B J, Dewey G *Phys. Rev. E* **52** 6588 (1995)
149. Eigen M, Schuster P *The Hypercycle, a Principle of Natural Self-Organization* (Berlin, New York: Springer Verlag, 1979); Eigen M, Winkler R *Laws of the Game: How the Principles of Nature Govern Chance* (New York: Harper & Row, 1983); Eigen M, Winkler-Oswatitsch R *Steps Towards Life: a Perspective on Evolution* (Oxford New York: Oxford University Press, 1992)
150. Dyson F *Origins of Life* (Combridge: Cambridge University Press, 1985)
151. Grosberg A Yu, Khokhlov A R *On the Unsolved Problems in Statistical Physics of Macromolecules* (Pouschino: NCBI, 1985)
152. Gutin A, Abkevich V, Shakhnovich E *Proc. Nat. Acad. Sci. USA* **92** 1282 (1995); Abkevich V, Gutin A, Shakhnovich E *Proc. Nat. Acad. Sci. USA* **93** 830 (1996)
153. Shakhnovich E, Abkevich V, Ptitsyn O *Nature* (London) **379** 96 (1996)
154. Onuchic J N et al. *Proc. Nat. Acad. Sci. USA* **92** 3626 (1995)
155. Shakhnovich E *Phys. Rev. E* (in press)

156. Romanovskiĭ Yu M, Stepanova N V, Chernavskiĭ D S *Matematichesraya Biofizika* (Mathematical Biophysics) (Moscow: Nauka, 1984)
157. Blumenfeld L A, Tikhonov A N *Biophysical Thermodynamics of Intracellular Processes* (Berlin: Springer-Verlag, 1994)