Statistical mechanics model of protein folding : short and long chains have different folding transitions

J.-R. Garel, T. Garel, Henri Orland

Classification
Physics Abstracts
05.90 — 61.40D — 87.10

Statistical mechanics model of protein folding : short and long chains have different folding transitions

J.-R. Garel ([1]), T. Garel ([2, *]) and H. Orland ([2])

([1]) Laboratoire d'Enzymologie, C.N.R.S., 91190 Gif-sur-Yvette, France
([2]) Service de Physique Théorique, Institut de Recherche Fondamentale, CEA-CEN Saclay, 91191 Gif-sur-Yvette Cedex, France

(*Reçu le 17 avril 1989, accepté sous forme définitive le 30 juin 1989*)

Résumé. — Nous considérons un modèle de mécanique statistique pour simuler le repliement d'une chaîne polypeptidique. Chaque résidu de la chaîne est défini par un ensemble de caractères indépendants. Les simulations Monte-Carlo pour une géométrie de champ moyen (dimension infinie) montrent que (i) les chaînes courtes ne se replient pas, (ii) la transition de repliement des chaînes de longueur intermédiaire est du type « verre de spins » (où tous les caractères doivent être partiellement satisfaits), (iii) la transition de repliement des chaînes longues est du type « Mattis » (où un seul caractère dominant est presque complètement satisfait). Ce modèle est peut-être applicable au cas de protéines multidomaines.

Abstract. — A statistical mechanics model of a polypeptide chain is used to simulate the folding process. Each residue is defined by a set of independent characters within a given sequence. Simulations of this model in a mean-field (infinite-dimensional) geometry show that (i) short chains do not fold, (ii) medium chains fold according to a spin-glass-like « transition » (where most characters must be partially satisfied) and (iii) long chains fold according to a Mattis-like « transition » (where only one dominant character is almost completely satisfied). This change in folding mechanism associated to chain length may be relevant to the existence of multidomain proteins.

1. Introduction.

There is enough information in the amino-acid sequence of a protein to direct its folding into its native conformation [1]. Deciphering the stereochemical code which transforms the one-dimensional information contained in the amino-acid sequence into a unique three-dimensional structure has become a major challenge, since this would allow to predict the conformation of a protein from its sequence. The folding code appears to be largely degenerate since (i) very different sequences can fold into almost identical conformations, and (ii) only a limited subset of all possible sequences is able to fold into a stable conformation

(*) On leave of absence from Physique des Solides, Orsay, France.

[2]. The folding of a polypeptide chain is a thermodynamically driven process in favorable conditions : the chain minimizes its energy in reaching an ordered conformation. It is the stereochemical information coded inside the chain sequence which determines the energy associated with a given configuration of the chain. This configuration results from the balance between the (attractive as well as repulsive) interactions that the chain has with itself and its surrounding solvent. This balance may have dynamical implications due to the possible existence of long-lived metastable states. In this article we refine a statistical mechanics model of a chain that we have recently studied [3], and perform Monte Carlo simulations of its thermodynamics, in a simplified mean-field geometry. This model allows us to study how the information present in a linear sequence is utilized to direct the folding of the chain into a stable and organized structure, and it predicts that short and long chains will follow different folding transitions.

2. The model of a polypeptide chain.

The chain consists of N links, with a contact interaction between links i at r_i and j at r_j given by $v_{ij}\,\delta(r_i - r_j)$ where i and j specify the positions of the links in the sequence. This contact interaction is an extreme limit of the more realistic situation where links i and j interact only through short range interactions, such as steric, screened Coulombic, induced dipolar, etc., interactions. Furthermore, the absence of angular dependent terms in the interaction suppresses the existence of secondary structures (helices, sheets, ...). Since the sequence of links is given, the v_{ij} are quenched variables relative to the geometry of the chain. The Hamiltonian reads :

$$H = \sum_{i,j=1}^{N} v_{ij}\,\delta(r_i - r_j)\,. \tag{1}$$

Two extreme limits can be studied, depending on the probability distribution of the v_{ij} :

1) the v_{ij} may be taken as independent random variables : this supposes that the interactions between a pair of links, i and j, is not related to the interactions between another pair, i' and j', even when these two pairs share a common link or when these pairs are made by the same types of links but at different positions. This case, which cannot take the nature of links i and j into account and considers only their positions, leads to a spin-glass-like transition between unfolded and folded states [3a]. This analogy with spin-glasses is consistent [5a] with long-lived metastable states, slow relaxations, etc. A more phenomenological model [5b], based on a random-energy model, leads to similar conclusions ;

2) the v_{ij} may be decomposed into a sum of separable (Mattis-like) interactions [3b] :

$$v_{ij} = -\frac{1}{N}\sum_{p=1}^{M} v_p\,\xi_i^p\,\xi_j^p \tag{2}$$

where the $\{\xi_i^p\}$ are independent random variables. In the following, we take $\{\xi_i^p = \pm 1\}$ with equal probability. Note that the factor $\frac{1}{N}$ in equation (2) is present to avoid a complete collapse of the chain, due to the absence of hard-core repulsion. This form of v_{ij} supposes that the interactions of a given link i with the others depend on some particular features of this residue i, defined by the set of M variable $\{\xi_i^p\}$. M is the number of independent characters

(1) associated with a given link, and this set of $\{\xi_i^p\}$ corresponds to the « nature » of link i. Links of a different chemical type (such as the different aminoacids) will be described by different sets of M values of the $\{\xi_i^p\}$. In this case, the nature of the link present at position i in the sequence is explicitly considered through a set of M independent characters ; these characters can be viewed as the Van der Waals volume, partial charge, hydrogen-bonding ability, hydrophobicity, solvatation, etc., of link i. Note that both the attractive and repulsive interactions of a given link i are considered through the signs of the corresponding M values v_p : for instance the hydrophobic character yields a positive v_p (see the appendix), whereas the short-range Coulombic character yields a negative v_p since the interaction is given $q_i\, q_j\, \delta(r_i - r_j)$. With the two simplifications mentionned above (zero-range and no angular dependence of the interactions), the frustration of the system is clearly decreased. However, our model still has some important frustration effects in its interactions, and can be viewed as a first step towards a better understanding of the folding process.

3. An example : the single character chain ($M = 1$).

In order to illustrate the physical content of this model we will briefly describe a chain in which the residues are defined by a single character, i.e. $M = 1$. For instance, this chain can be viewed as composed of hydrophilic ($\xi = -1$) and hydrophobic ($\xi = +1$) residues. The interactions between the chain and the solvent molecules are described by a microscopic Hamiltonian which leads, through integration over the positions of solvent molecules (see the appendix), to an intra-chain interaction which is indeed described by equation (1). Although the solvent does not appear explicitly in the expression of v_{ij} given by equation (1), it is taken into account through the variables ξ_i.

In the thermodynamic limit $N \to \infty$, the « folding » transition of such a chain corresponds to a simple condensation governed by the hydrophobic character ; a hydrophilic residue likes to be close to another hydrophilic residue or to a solvent molecule, while an hydrophobic residue prefers to be away from an hydrophilic residue or a solvent molecule (this simple picture breaks down if one considers the chain with hard-core repulsion. In that case, one expects microdomain formation). This transition can be described by an order parameter (à la Mattis) :

$$m(r) = \frac{1}{N} \overline{\sum_i \xi_i \langle \delta(r - r_i) \rangle} \tag{3}$$

where $\langle \ldots \rangle$ and $\overline{\ldots}$ denote respectively thermal and disorder averages. This order parameter $m(r)$ is a measure of the correlation between the distribution of the ξ_i along the chain (its « sequence ») and that of the r_i (its « conformation »). As such, it is a measure of the quality of the coding of this « conformation » by this « sequence ».

4. The physical definition of a chain.

A key parameter of this model is the number M of characters needed to describe the interactions of a given type of residue in the chain. In a broader view of statistical mechanics models, the M independent characters associated with a link bear some similarity to the patterns defined in neural networks theories [4]. As M increases, one expects the folding transition of a chain of N links to evolve from a Mattis-like to a spin-glass-like behavior [5]. So

(1) This definition of M is slightly different from the one used in reference [3b] ; see the discussion in section 5.

far, the model has been defined in any dimension. From now on, we will restrict ourselves to the following (high-dimension or mean-field) approximation : the chain is constructed on a set (*simplex*) such that the distance between any pair of points is constant (a simplex with three points (resp. four points) is an equilateral triangle (resp. a tetrahedron), etc.). Note that for $M = 1$, a chain on a simplex always folds on at most four points (due to the chain constraint), and different chemical sequences correspond to different paths going through these four points. One could also study model (2) with $M = \alpha N$ in the thermodynamic limit, as in reference [4b]. The resulting equations being rather intricate, we have turned to a different approach. For a given value of M, the folding of a « long » chain will correspond to a Mattis-like transition, and that of a « short » chain to a spin-glass-like transition. We have first chosen the value $M = 8$ for the number of characters using a semi-biological argument, and we have then studied the possible folding transitions of chains of various lengths ($20 < N < 100$) on the simplex.

5. 8 characters may be sufficient to describe a polypeptide chain.

Each type of residue being defined by a set of M variables $\{\xi_i^p\}$ with values ± 1, the r.h.s. of equation (2) can take 2^M different values, and therefore, there are 2^M different possible values of v_{ij}. With 20 amino-acids there are 210 possible different pairs of residues ($20 \times 21/2$) and thus 210 different interaction terms v_{ij}, if we assume that the interaction between residues i and j depends only on the « nature » of these residues and not on their position in the sequence. To obtain the 210 interaction terms needed for the 210 different pairs, we must take M so that $2^M > 210$. The smallest integer value possible is $M = 8$ (a similar argument in the case of RNA or DNA chains would lead to $M = 4$ independent characters for each base).

Is $M = 8$ a sufficient number of characters to describe all the interactions between the 20 amino-acids ? Individual amino-acids can probably be discriminated using more than 8 features : polar or apolar, neutral or charged, large or small, helix former or breaker, hydrogen-bond donor or acceptor, etc., but these features do not correspond to truely *independent* characters : for instance, properties such as hydrogen-bonding abilities, polarity, secondary structure preference, and sidechain bulkiness are likely to be related. It is difficult to specify more precisely what these 8 independent characters could be, but this value of 8 is sufficient to distinguish all interactions between pairs of amino-acids. The simulations given below correspond to the case $M = 8$, a fixed set of 8 characters, and to chains of variable lengths ($20 < N < 100$).

6. Simulations.

We have performed Monte Carlo simulation for chains of N links ($20 < N < 100$) constructed on a simplex of N points. Starting with a random initial configuration of the chain, we performed single link moves (cost in energy ΔE) : an elementary Monte Carlo step moves a link i from its actual position r_i on the simplex to a new position r_i^{new} (keeping the rest of the chain unchanged), where the new position is different from r_{i-1} and r_{i+1} to satisfy the chain constraint. These moves are accepted according to the Metropolis algorithm [6]. The outcome of the calculation is :

(i) the internal energy $U = \langle H \rangle$ with H defined in equation (1).

(ii) the order parameters $m_p(r)$ as defined in equation (2) :

$$m_p(r) = \frac{1}{N} \sum_i \xi_i^p \langle \delta (r - r_i) \rangle \tag{4}$$

which is a measure of the correlation of the chain conformation with character p. It is technically more convenient to use global order parameters :

$$S_p = \sum_r m_p^2(r) \,. \tag{5}$$

Since on a given chain, the chemical sequence is fixed, we have not averaged equation (4) and (5) over the disorder.

We have taken $v_1 = 1$, $v_2 = 0{,}5$, $v_3 = 0.2$, $v_4 = 0.15$ and $v_{i+4} = -v_i$ $(i = 1, \ldots, 4)$ in equation (2), and the number of Monte Carlo steps per link was 40 for $N = 40$, 60, 100 (which turned out to be sufficient to reach thermal equilibrium ; this was checked by using different initial conditions), and 100 for $N = 20$. In the last case, it will become clear that one has to increase this factor to reach thermal equilibrium (see below). The internal energy U is shown in figure 1 as a function of the temperature T. A finite size « critical temperature » can be defined for $N = 40$, 60, 100, corresponding to the folding of the chain (this is not the case for $N = 20$, due to finite size fluctuations). This critical temperature is defined as the temperature at which some order parameters S_p start to grow. The theoretical critical temperatures (Ref. [3b]) are shown in figure 1, for comparison ; the agreement can be considered as reasonable. The order parameters S_p have been measured just below the (finite size) transition (T_1, T_1', T_1'') and at low (finite size) temperature (T_0, T_0', T_0'') for $N = 40$, 60, 100. We have also studied the case $N = 20$ at two similar temperatures. In that case, there is a progressive freezing of the system, due to the presence of many metastable states (since one is in a glassy situation) separated by small (since N is small) barriers. The results are displayed in table I, where only the S_p larger than 10^{-2} are shown. According to the simulations, the folding « transitions » seem to be first order.

For $N = 100$, it is clear that one is almost in the $M = 1$ case, at least close to the transition. There is a single character showing up at the transition (T_1''). As N decreases, more and more characters appears at the transition (if any). At low temperatures, the frustration grows and

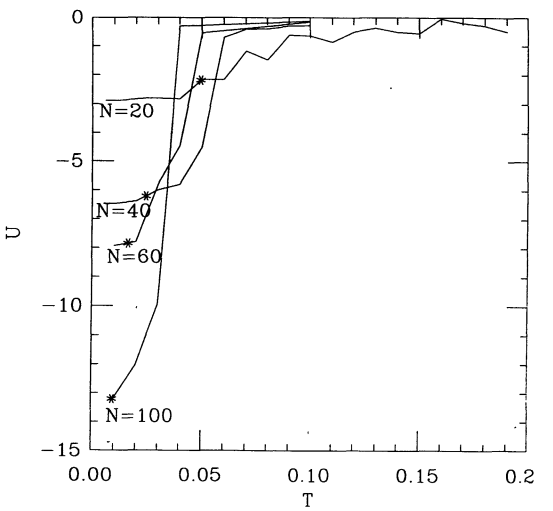

Fig. 1. — Internal energy U *vs.* temperature T for various chain length N. The stars denote the theoretically calculated critical temperatures.

Table I. — *For each value of N* ($N = 20, 40, 60, 100$), *the various* S_p ($p = 1, \ldots, 8$) *are shown at two different temperatures. For* $N = 20$, *where there is no « transition », the* S_p *are seen to grow simultaneously as the temperature T is decreased. For* ($N = 40, 60, 100$), *the first temperature* (T_1, T_1', T_1'') *is chosen to be close to the folding « transition », while* (T_0, T_0', T_0'') *illustrates how the other characters grow at low temperature. A bar* (-) *denotes a value of* S_p *smaller than* 10^{-2}.

N \ p		1	2	3	4	5	6	7	8
20	$t_1 = 5 \times 10^{-2}$	0.133	0.088	0.018	0.023	—	0.019	0.018	—
	$t_0 = 3 \times 10^{-2}$	0.19	0.136	0.033	0.047	0.011	0.03	0.02	0.01
40	$T_1 = 5 \times 10^{-2}$	0.104	0.025	0.029	—	—	—	—	—
	$T_0 = 2 \times 10^{-2}$	0.268	0.093	0.086	0.015	0.014	0.014	—	—
60	$T_1' = 4 \times 10^{-2}$	0.068	0.017	—	—	—	—	—	—
	$T_0' = 2 \times 10^{-2}$	0.227	0.038	0.014	—	—	—	—	—
100	$T_1'' = 3 \times 10^{-2}$	0.12	—	—	—	—	—	—	—
	$T_0'' = 2 \times 10^{-2}$	0.20	0.013	—	—	—	—	—	—

other dominated characters appear. In an analytical calculation such as the one mentionned in section 4, one would say that for $\frac{M}{N} > \alpha_c$, the folding transition is of the spin-glass type, whereas it is of the Mattis type for $\frac{M}{n} < \alpha_c$. Our simulations are in qualitative agreement with this picture.

7. Conclusions.

The present model has considered chains of N links of 20 different types, the « amino-acids », each one defined by 8 independent characters. For very short chains ($N \leqslant 20$), the chain cannot fold into an ordered state. For medium chains ($20 \leqslant N \leqslant 60$), folding into an ordered state involves the condensation of several characters. Many (if not all) different interactions between pairs of links contribute to the decrease in internal energy associated with folding. The cooperative unit over which the chain finds an energetic compromise which minimized its frustration corresponds to the entire chain itself : folding resembles a spin-glass transition. For long chains ($N \geqslant 100$), the condensation of a single dominant character is sufficient to promote folding into an ordered state. Folding resembles a Mattis-like transition, and there is a strong correlation between the « conformation » of the chain and the distribution of this dominant character along its « sequence ». The change in the regime which governs the folding of chains of increasing lengths may be relevant to the behavior of real proteins : smaller proteins fold into a single compact structure upon a highly cooperative transition, whereas in large proteins, different segments along the chain behave as independent folding units to form the compact substructures called « domains » [7]. The present model certainly

corresponds to an oversimplified image of a polypeptide chain. In real chains, links of the same chemical type may not have identical sets of $\{\xi_i^p\}$ if they are at different positions along the chain ; indeed, although two glycine residues are chemically identical, they may have different abilities to form turns because they are preceeded or followed by different sequences. The definition of a « character » assigned to a given link is open enough to accomodate not only intrinsic physicochemical properties of individual amino-acids, but also semi-empirical data derived from the analysis of known structures [8]. Such simple models may be helpful to adapt the concepts of statistical mechanics to the correlation between sequence and conformation of proteins and to elucidate the stereochemical folding code [9].

Note added : After submission of this work, we have received preprints by E. I. Shakhnovich and A. M. Gutin dealing with similar models.

Appendix.

Let the short-range interaction between link i of the chain (at r_i) and a water molecule (at R_α) be $\xi_i\,\delta(r_i - R_\alpha)$. Thus $\xi = -1$ (resp. $\xi = +1$) denotes an hydrophilic (resp. hydrophobic) link. The grand partition function of the chain plus water system reads :

$$Z = \sum_{M=0}^{\infty} \frac{\lambda^M}{M!} \mathrm{Tr}_{\{r_i, R_\alpha\}} \exp\left(-\beta \sum_{i=1}^{N} \sum_{\alpha=1}^{M} \xi_i\,\delta(r_i - R_\alpha)\right) \tag{A1}$$

where λ is the fugacity of a water molecule. We have :

$$Z = \sum_{M=0}^{\infty} \frac{\lambda^M}{M!} \mathrm{Tr}_{\{r_i\}} \left(\mathrm{Tr}_{\{R\}} \exp\left(-\beta \sum_{i=1}^{N} \xi_i\,\delta(r_i - R)\right)\right)^M \tag{A2}$$

that is :

$$Z = \mathrm{Tr}_{\{r_i\}} \exp\left(\lambda \left(\mathrm{Tr}_{\{R\}} \exp\left(-\beta \sum_{i=1}^{N} \xi_i\,\delta(r_i - R)\right)\right)\right). \tag{A3}$$

Expanding (A3) to second order in β yields :

$$Z = \mathrm{Tr}_{(r_i)} \exp\left(\lambda \frac{\beta^2}{2} \sum_{i,j} \xi_i\,\xi_j\,\delta(r_i - r_j)\right) \tag{A4}$$

of the form of equation (2) with $v_1 = N\lambda\beta/2$ (the fact that the interaction has the same value for two hydro-philic or -phobic links is due to the zero range of the interaction).

References

[1] ANFINSEN C. B., *Science* **181** (1973) 223-230.
[2] CREIGHTON T. E., Proteins : Structure and Molecular properties, Ed. W. H. Freeman (New-York) 1984 *Chapters 3 and 6.*
[3] GAREL T. and ORLAND H., (a) *Europhys. Lett.* **6** (1988) 307-310 ; (b) *Europhys. Lett.* **6** (1988) 597-601.
[4] (a) HOPFIELD J. J., *Proc. Natl. Acad. Sci. USA* **79** (1982) 2554-2558 ;
(b) AMIT D. J., GUTFREUND H. and SOMPOLINSKY H., *Phys. Rev. Lett.* **55** (1985) 1530-1533.
[5] (a) ANSARI A. *et al., Proc. Natl. Acad. Sci. USA* **82** (1985) 5000-5004 ;
(b) BRYNGELSON J. B. and WOLYNES P. G., *Proc. Natl. Acad. Sci. USA* **84** (1987) 7524-7528.
[6] See for instance LI Z. and SCHERAGA H. A., *Proc. Natl. Acad. Sci. USA* **84** (1987) 6611-6615 and references therein.
[7] JANIN J. and WODAK S. J., *Prog. Biophys. Molec. Biol.* **42** (1983) 21-78.
[8] THORNTON J. M., *Nature* (London) **335** (1988) 10-11.
[9] JAENICKE R., *Prog. Biophys. Molec. Biol.* **49** (1987) 117-137.