Condensed Matter Physics, 1999, Vol. 2, No 2(18), pp. 369–378

CONDENSED MATTER PHYSICS

Master equation approach to protein folding

M.Cieplak [1] , M.Henkel [2] , J.R.Banavar [3]

[1] Institute of Physics, Polish Academy of Sciences, and College of Sciences, 02-668 Warsaw, Poland

[2] Laboratoire de Physique des Matériaux, Université Henri Poincaré Nancy I, F-54506 Vandœuvre, France

[3] Department of Physics and Center for Materials Physics, 104 Davey Laboratory, The Pennsylvania State University, University Park, PA 16802

Received June 25, 1998

The dynamics of two 12-monomer heteropolymers on the square lattice is studied exactly within the master equation approach. The time evolution of the occupancy of the native state is determined. At low temperatures, the median folding time follows the Arrhenius law and is governed by the longest relaxation time. For both good and bad folders, significant kinetic traps appear in the folding funnel and the kinetics of the two kinds of folders are quite similar. What distinguishes between the good and bad folders are the differences in their thermodynamic stabilities.

Key words: *protein folding, kinetic equations*

PACS: *87.15.By, 87.10.+e*

1. Introduction

A denatured protein folds into a compact native state in a time of the order of a millisecond after the physiological conditions are restored [1]. This process is reversible and the native state is believed to coincide with the ground state of the system. Studies of simplified lattice models of proteins have provided many insights into the dynamics of the folding process [2]. The crucial feature that makes lattice models useful is the possibility of enumerating the complete set of conformations and determining which of them is the ground state. This can be accomplished provided the length of the heteropolymer, N, is small. Such studies have elucidated the role of thermodynamic stability [2,3], stability against mutations [4] and the existence of a linkage between the rapid folding and the stability of the native state [3].

The key concept that has been introduced to explain the rapid folding occurring in natural proteins is that of the folding funnel [5,6] – a set of conformations that are smoothly connected to the native state, as indicated schematically at the top of figure 1. It is expected that for random sequences of aminoacids there exist competing basins of attraction which would otherwise trap the system away from the native state. This corresponds schematically to what is shown at the bottom of figure 1. It isn't easy to identify the states belonging to the funnel because of an enormous number of conformations present even for a small N as well as because the problem is a dynamical one. Possible approaches include the monitoring of frequencies of passages between various states in Monte Carlo trajectories [6] or the mapping of states into underlying valleys of effective states [7,8].

Essentially all approaches to the studies of the folding dynamics are restricted to Monte Carlo simulations that start from a few randomly chosen initial conformations [9]. The only exception is an approach due to Chan and Dill [10] in which an enumeration of transition rates between classes of conformations which have the same number of contacts and are a given number of kinetic steps away from the native state.

Figure 1. A schematic representation of the phase space properties of good (top) and bad (bottom) folders. The vertical axis corresponds to energy and the horizontal axis to a "coordinate" in the phase space.

Recently, we presented an exact method to study the dynamics of short model proteins [11] which was based on the master equation [13]. We have specifically considered two N=12 sequences, called A and B, which are placed on a square lattice. The dynamics of these sequences can be studied exactly because the sequences can acquire only $\mathcal{N}$=15037 conformations. The two sequences have the same set of contact energies B_{ij} but their assignment to various monomers i and j is different with the result that A is a good folder and B is a bad folder. The basic finding of [10] is that the qualitative picture corresponding to figure 1 is absolutely correct. By identifying kinetic trap states responsible for the slowest dynamical processes in the system we could demonstrate that the traps for sequence A are within the folding funnel whereas for sequence B the relevant traps form a valley which competes with the native valley similar to the bottom of figure 1. In the current paper we provide a deeper characterization and comparisons of the two sequences and elucidate the nature of the trap states.

The sequences that we study are described by the Hamiltonian

$$\mathcal{H} = \sum_{ij} B_{ij}\Delta_{ij} \quad , \tag{1}$$

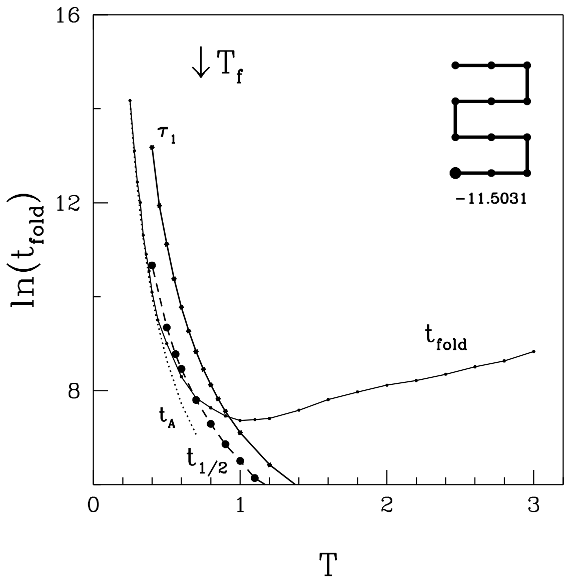

Figure 2. Top: The native conformation and its energy for sequence A. The enlarged circle shows the first monomer. Main: Dynamical data for the folding. The solid line marked by $t_{\rm fold}$ gives the median folding time derived from 1000 Monte Carlo trajectories. The solid line τ_1 is the longest relaxation time. The broken line $t_{\frac{1}{2}}$ with the black circles gives the time for $P_0(t)$ to reach $\frac{1}{2}P_0$ from the uniformly random initial state. The dotted line t_A is a fit of the Monte Carlo data to the Arrhenius law with δE=2.76. The arrow at the top indicates the value of the folding transition temperature.

where Δ_{ij} indicates the contact interaction B_{ij} assigned to monomers which are geometrical nearest neighbours on the lattice but are not neighbours along the sequence. In such arrangements Δ_{ij} is equal to 1, otherwise it is 0. The values of the 25 couplings are chosen as Gaussian numbers which are centred around -1 to provide an overall attraction, as detailed in [10]. The ground state of sequence A is maximally compact and it fills the 3×4 lattice, as shown at the top of figure 2. For sequence B, the ground state, shown at the top of figure 3, is doubly degenerate. Both of the ground states are compact. However they are not maximally compact. They differ merely by a placement of one end monomer and therefore they were considered as an effective single native state.

We have found that the dynamics of A and B are superficially similar: for both, the median folding time, $t_{\rm fold}$, and the longest relaxation time, τ_1 diverge at low T according to an Arrhenius law. The temperature $T_{\rm min}$ at which folding to the native state proceeds the fastest is about the same for both sequences. It is the location of the folding transition temperature, $T_{\rm f}$, with respect to $T_{\rm min}$ which distinguishes between sequences A and B. $T_{\rm f}$ is defined as the temperature at which the equilibrium value of the probability to occupy the native state, P_0, crosses $\frac{1}{2}$ and is a measure of thermodynamic stability. For bad folders, $T_{\rm f}$ is well below $T_{\rm min}$ and thus a substantial occupation probability for the native state is found only in a temperature range in which the dynamics are glassy. For sequence A, the values of $T_{\rm f}$ and $T_{\rm min}$ are 0.71 and 1.0 ± 0.1, respectively, while for sequence B, the corresponding values are 0.01 and 1.1 ± 0.1. Thus the two sequences are dynamically similar but the equilibrium properties differ dramatically.

2. Methods

We study our sequences through an analysis of the master equation and then compare the results to those obtained by the Monte Carlo approach. The master equation does not deal with a specific trajectory but with an ensemble of trajectories and it reads

$$\frac{\partial P_\alpha}{\partial t} = \sum_{\beta\neq\alpha}\Big[w(\beta\to\alpha)P_\beta - w(\alpha\to\beta)P_\alpha\Big], \qquad (2)$$

where $P_\alpha = P_\alpha(t)$ is the probability of finding the sequence in conformation α at time t. The quantity $w_{\alpha\beta} = w(\beta \to \alpha)$ is the transition rate from conformation β to conformation α. Of course, writing a master equation relies on the assumption that a Markov chain description adequately describes protein kinetics. In particular, memory effects are assumed to be negligible. The actual derivation of a master equation remains a formidable problem in itself [12].

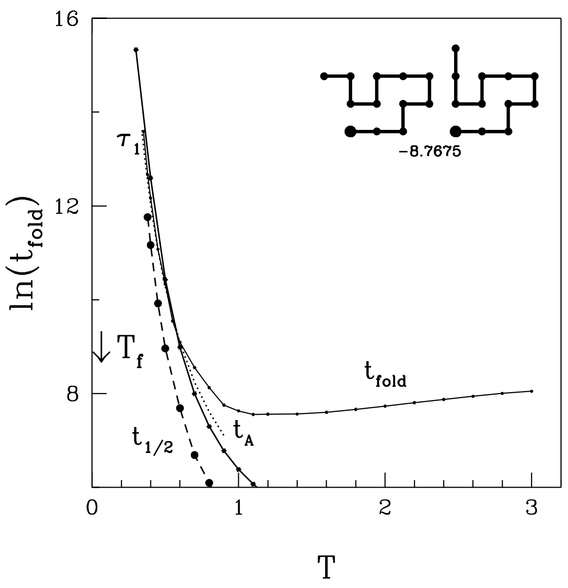

Figure 3. Same as figure 2, but for sequence B. For the curve t_A, δE=3.55. There are two native conformations with the same energy.

One may bring this into a matrix form by letting $\vec{P} = (P_1, \ldots, P_{\mathcal{N}})$ and

$$m_{\alpha\beta} = -w_{\alpha\beta} \leqslant 0, \;\; \alpha\neq\beta; \;\; , \;\; m_{\alpha\alpha} = \sum_{\beta\neq\alpha} w_{\beta\alpha} \;\; . \qquad (3)$$

The master equation then takes the form of an imaginary-time Schrödinger equation [13,14]

$$\partial_t \vec{P} = -\hat{M}\vec{P} \;\; , \qquad (4)$$

where the $m_{\alpha\beta}$ are the matrix elements of $\hat{M}$.

The time-dependence state vector $\vec{P}(t)$ at time $t = n\tau_0$ can be obtained by applying n times the recursion $\vec{P}((n+1)\tau_0) = (1-\tau_0\hat{M})\vec{P}(n\tau_0)$ to $\vec{P}_{\rm in}$, where τ_0 is a microscopic time associated with a single move and $\vec{P}_{\rm in}$ is some initial probability distribution. On the other hand, the spectrum of relaxation times $\tau_\lambda = 1/(\mathrm{Re}\ M_\lambda) \geqslant 0$ follows directly from the eigenvalues M_λ of $\hat{M}$.

The transition rates $w_{\alpha\beta}$ are designed so that a stationary solution of the master equation corresponds to equilibrium with the Hamiltonian given by equation (1). This can be accomplished provided the detailed balance condition

$$w_{\alpha\beta}P_\beta^{\rm eq} = w_{\beta\alpha}P_\alpha^{\rm eq} \qquad (5)$$

is satisfied. Here, $P_\alpha^{\mathrm{eq}} \sim \mathrm{e}^{-E\alpha/T}$ is a stationary solution of the master equation and E_α is the energy of the sequence in conformation α. Equation (5) is satisfied by any $w_{\alpha\beta} = f_{\alpha\beta} \exp\left[-(E_\alpha - E_\beta)/2T\right]$ provided only that $f_{\alpha\beta} = f_{\beta\alpha}$. Here, we choose $w_{\alpha\beta} = w_{\alpha\beta}^{(1)} + w_{\alpha\beta}^{(2)}$, where

$$w_{\alpha\beta}^{(\sigma)} = \frac{2}{\tau_0} R_\sigma \cosh\left[\left(E_\alpha - E_\beta\right)/T\right] \tag{6}$$

with $R_1 + R_2 = 1$. Here, $\sigma = 1$ and 2 refer to the single- and double-monomer moves respectively. It is understood that $w_{\alpha\beta}^{(\sigma)} = 0$ if there is no move of type σ linking β with α. The choice (6) guarantees that transition rates are finite and bounded for all temperatures. Similar to [3], we focus on $R_1 = 0.2$ and take the single and two-monomer (crankshaft) moves as in [10].

Because of the detailed balance condition, the eigenvalues E_α are not calculated by diagonalizing $\hat{M}$ directly, but by diagonalizing an auxiliary matrix $\hat{K}$ with elements

$$k_{\alpha\beta} = \frac{v_\beta}{v_\alpha} m_{\alpha\beta} = k_{\beta\alpha} \ \ , \tag{7}$$

where $v_\alpha = \mathrm{e}^{-E_\alpha/2T}$. The master equation can be then rewritten as

$$\partial_t \vec{Q} = -\hat{K}\vec{Q} \ \ , \tag{8}$$

where $\hat{K}$ is real symmetric and $Q_\alpha = P_\alpha/v_\alpha$. This equation is then diagonalized using the standard symmetric Lanczos algorithm without reorthogonalization, as described in [15,16] and then the eigenmodes of $\hat{M}$ are determined. In particular, since $\hat{K}$ and $\hat{M}$ are related by a similarity transformation, all eigenmodes M_λ are real.

The eigenvector corresponding to $M_\lambda = 0$, i.e. to the infinite relaxation time, determines the equilibrium occupancies of the conformations. This eigenvector needs to be normalized to satisfy its probabilistic interpretation. The longest finite relaxation time $\tau_1 = 1/M_1$ is found from the smallest non-zero eigenvalue M_1.

Our Monte Carlo simulations were done in a way that satisfies the detailed balance conditions and were devised along the lines described in [10]. For polymers, getting these conditions satisfied is not trivial because each conformation has its own number, K, of allowed moves that the conformation can make. Thus the propensities to make a move in a unit time vary from conformation to conformation and the effective “activities” of the conformations need to be matched. This can be accomplished by first determining the maximum value of K, $K_{\max}$. For the 12-monomer chain on the square lattice, $K_{\max}$ is equal to 14. We associate a single time unit with the conformations in which K=$K_{\max}$. This means that each move allowed is attempted always with probability $1/K_{\max}$. For a conformation with K allowed moves, the probability to attempt any move is then $K/K_{\max}$ and the probability not to carry out any attempt is $1 - K/K_{\max}$. The attempted moves are then accepted or rejected as in the standard Metropolis procedure. The probability of a single monomer move (an end flip or a switch in a pair of bonds that make the 90° angle) is additionally reduced by the factor of 0.2 and an allowed double monomer crankshaft move by 0.8. The time is then measured in terms of the total number of the Monte Carlo attempts divided by $K_{\max}$. This scheme not only establishes the detailed balance conditions but it also uses less CPU compared to a process in which moves are attempted regardless of their being allowed or not allowed.

3. Results

Figure 4 shows the 4 longest finite relaxation times for sequence A as a function of temperature. It is seen that they are all roughly proportional to each other and they all appear to diverge at T=0, albeit with different energy barriers. The longest relaxation time follows the Arrhenius law, $\tau_1 \sim e^{\delta E_1/T}$ with δE_1 of about 4.1. The Monte Carlo derived median folding time also follows the Arrhenius law at low temperatures but with $\delta E = 2.76 \pm 0.06$. The overall Arrhenius-like behaviour suggests that the physics of folding at low T's is dominated by processes that establish equilibrium. Their longest time scale is then set by τ_1. At temperatures above $T_{\min}$, t_{fold} no longer follows τ_1. Here, the physics of folding is dominated by the statistics of rare events: the system establishes equilibrium rapidly and then folding is reached by a search in equilibrium. The search becomes more and more random when T becomes larger and larger. For sequence B, similar results are obtained but the Arrhenius barrier in t_{fold} is 3.55 ± 0.06, whereas $\delta E_1 = 4.0 \pm 0.06$. Figures 2 and 3 show fits to the Arrhenius law for sequences A and B respectively.

We now focus on the time evolution of the probability, $P_0(t)$, to occupy the native state, where t denotes time. This probability depends on the initial conditions. The solid lines in figure 5 show the evolution of $P_0(t)$ when in the initial state all conformations have equal probability of $1/\mathcal{N}$ to be occupied. The main figure is for sequence A and the time evolution is shown for three temperatures: 1.2, 0.9 and 0.6. The data for the lowest of these temperatures are also compared with the Monte Carlo evolution. The Monte Carlo data, obtained by starting from random conformation, agree with the exact evolution. The time scales at which the equilibrium saturation values of P_0 are reached are of order τ_1. The top portion of figure 5 shows $P_0(t)$ for sequence B. It is seen that the saturation levels are significantly

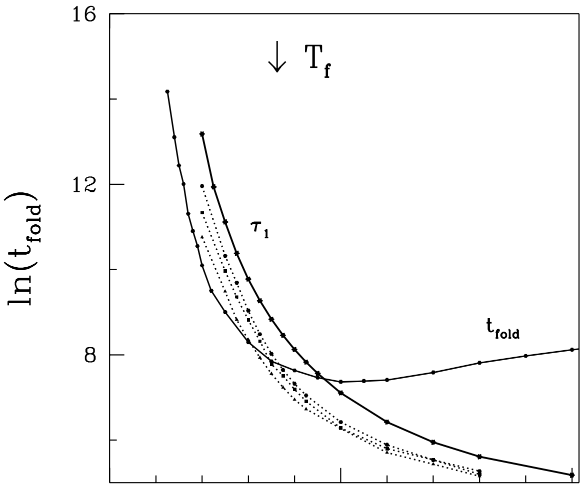

Figure 4. The four longest relaxation times versus T for sequence A. The very longest, τ_1, is drawn by a solid line whereas the remaining three – by the dotted lines. The value of T_{f} is indicated by an arrow and the Monte Carlo data on t_{fold} are repeated from figure 2 for a comparison.

lower for B than for A at the same temperatures. This reflects a significantly lower value of T_f for B. We also observe, by looking at figures 2 and 3 that the median folding time at low temperatures appears to coincide with the time needed for $P_0(t)$ to reach half of its equilibrium value.

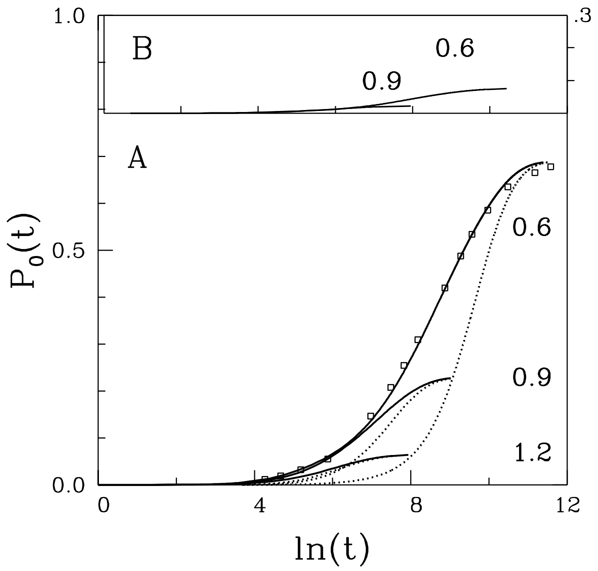

Figure 5. Probability of occupation of the native state, $P_0(t)$, of sequence A, for three values of the temperatures, indicated on the right. $P_0(\infty)$ agrees with the equilibrium value. The solid lines correspond to the uniformly random initial condition and the dotted lines correspond to an initial placement of the system in the trap state. The squares correspond to Monte Carlo results, based on 200 random starting conformations. The top figure shows $P_0(t)$ for sequence B. The initial condition is that of uniform randomness. The y-axis scale is indicated on the right.

The dotted line in figure 5 corresponds to the initial state being the kinetic "trap" conformation [17] which is the strongest obstacle in reaching equilibrium. If the system starts in the trap state, it will still reach the final equilibrium in a time scale set by τ_1 but at shorter times the occupancy of the native state is significantly smaller compared to the start from a totally random state.

In general, it is not easy to determine which conformation forms the most potent trap. One may attempt to determine it by monitoring the occupation of the local energy minima in a Monte Carlo process performed at very low temperatures. This approach has been successfully applied in [8]. Here, the trap state is determined in an exact way by studying the eigenvector corresponding to the longest relaxation time and by identifying the local energy minima which contribute the most to this eigenvector at low T. The largest weight is associated with the native state. Usually, the second largest weight corresponds to the most relevant trap. This happens provided the second largest occupancy is in a local energy minimum state. A non-minimum state, immediately preceding the native state kinetically, may also possess a significant occupancy but it does not constitute a trap. Thus the search for the most relevant trap goes after non-native local energy minima which have the biggest occupancy. In the limit of $T \to 0$, weights associated with all other local energy minima states become small.

We now focus on the interpretation of the trap states to demonstrate the qualitative validity of figure 1. We place the system in the trap state and ask what is the best trajectory which leads from the trap state to the native state. The best trajectory is defined to be one in which the highest energy state is lower than the highest energy state on all other trajectories.

Figure 6 shows the trap state (of energy -8.9627) for sequence A and the best corresponding trajectory that connects it to the native state. The trajectory reaches the native state in 8 steps and it elevates in energy by 4.5323, relative to the the energy of the trap state, after accomplishing the second step. Thus it takes only two uphill steps to enable a flow to the native state. The first of these steps breaks two contacts and it costs 2.8823 in energy. Thus this biggest single step energy cost is what determines the value of the energy barrier in the Arrhenius law for t_{fold}, as derived through the Monte Carlo technique. On the other hand, δE_1, that determines τ_1, appears to be governed by the energy costs to make the initial two-step climb to the point of the highest elevation. In a Monte Carlo process, two consecutive climbs up are not very likely at low T's and can be missed. However, the exactly derived longest relaxation time indicates the true nature of the barrier.

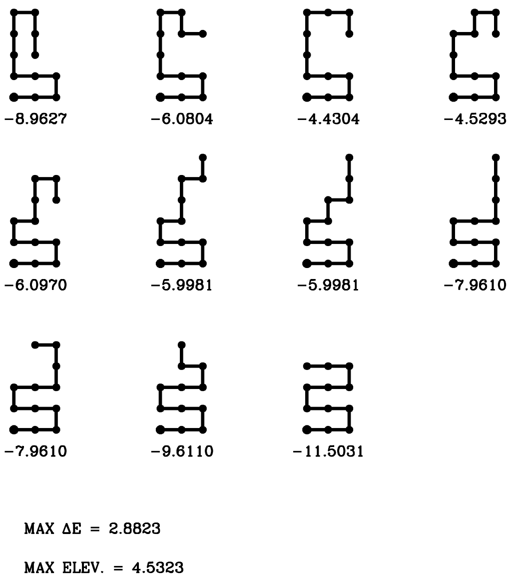

Figure 6. The best trajectory linking the trap state of sequence A (shown at the top left) and the native state. The energies involved are shown below each conformation. The biggest single step and global energy expenses are indicated at the bottom.

The kinetics of sequence B is quite similar to that of sequence A. Thus what distinguishes the two sequences are their thermodynamic stabilities as measured by T_{f}. Figure 7 shows the steps of the best trajectory from the the most important trap state for sequence B (of energy – 8.4431) to the native state. The trap state is almost identical in shape to the two native conformations: it is only the positioning of the middle portion of the conformation that is not quite right. And yet it takes 10 steps, like in figure 6, to accomplish the transition between the trap and native states. The overall barrier that is climbed is equal to 4.4 which is close to the Arrhenius barrier δE_1 in τ_1 of 4.0. The biggest single step

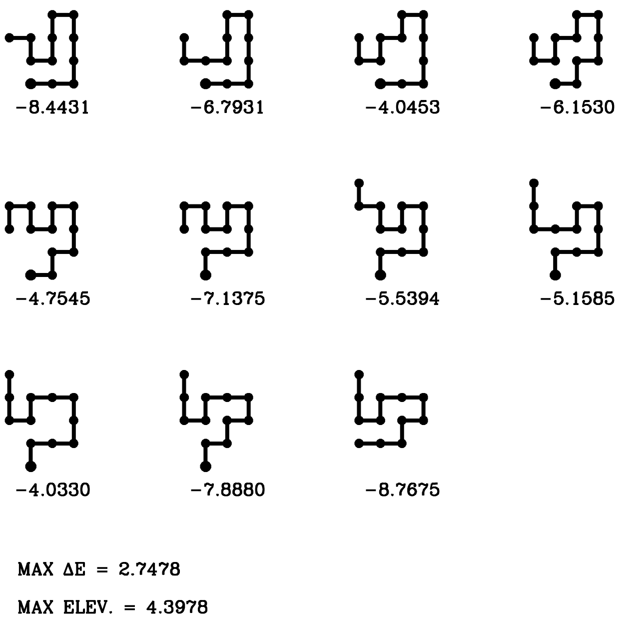

Figure 7. As in figure 6 but for sequence B.

energy cost of 2.7478 on the trajectory is, however, further away from δE in $t_{\rm fold}$ of about 3.6. The important observation is that for the bad folder B the most relevant trap conformation is still in the folding funnel of the native state. The presence of other low lying valleys that compete with the native valley does not really affect the trapping effects in the native funnel. Instead it generates low stability of the native state through a substantial equilibrium probability of being in some other valleys.

Even though our exact results have been obtained in a toy model of only 12 monomers, it is expected that the essential physics of protein folding and the distinction between the good and bad folders are as illustrated here. Tools to study larger and off-lattice systems in this spirit remain to be developed.

We thank Jan Karbowski for collaboration and T. X. Hoang for discussions. This work was supported by KBN (Grant No. 2P03B-025-13), Polonium, CNRS-UMR 7556, NASA, and the Applied Research Laboratory at Penn State.

References

1. Creighton T.E. Proteins. Freeman, New York, 1993.
2. Wolynes P.G., Onuchic J.N., Thirumalai D. Navigating the folding routes. // Science, 1995, vol. 267, p. 1619; Dill K.A., Bromberg S., Yue S., Fiebig K., Yee K.M., Thomas D.P., Chan H.S. Principles of protein folding – A perspective from simple exact models. // Protein Sci. 1995, vol. 4, p. 561; Camacho C.J., Thirumalai D. Kinetics and thermodynamics of folding in model proteins. // Proc. Nat. Acad. Sci. USA, 1993, vol. 90, p. 6369; Chan H.S. and Dill K.A. The protein folding problem. // Phys. Today, 1993 (February), vol. 46, p. 24.
3. Sali A., Shakhnovich E., Karplus M. How does a protein fold? // Nature, 1994, vol. 369, p. 248.
4. Li H., Helling R., Tang C., Wingreen N. Emergence of preferred structures in a simple model of protein folding. // Science, 1996, vol. 273, p. 666; Vendruscolo M., Maritan A., Banavar J.R. Stability threshold as a selection principle for protein design. // Phys. Rev. Lett. 1997, vol. 78, p. 3967.
5. Onuchic J.N., Wolynes P.G., Luthey-Schulten Z. Towards an outline of the topography of a realistic protein-folding funnel. // Proc. Natl. Acad. Sci., 1995, vol. 92, p. 3626; Bryngelson J.D., Onuchic J.N., Socci N.D., Wolynes P.G. Funnels, pathways and the energy landscape of protein folding: A synthesis. // Proteins: Struct. Funct. and Genet., 1995, vol. 21, p. 167.
6. Leopold P.E., Montal M., Onuchic J.N. Protein folding funnels: a kinetic approach to the sequence-structure relationship. // Proc. Natl. Acad. Sci. USA, 1992, vol. 89, p. 8721.
7. Cieplak M., Vishveshwara S., Banavar J.R. Cell dynamics of model proteins. // Phys. Rev. Lett., 3681 vol. 77, p. 3681; Cieplak M., Banavar J.R. Cell dynamics of folding in two dimensional model proteins. // Fold. Des., 1997, vol. 2, p. 235.
8. Cieplak M., Hoang T.X. Coarse grained description of the protein folding. // Phys. Rev. E, 1998, vol. 58, p. 3589.
9. Socci N.D., Onuchic J.N. Folding kinetics of proteinlike heteropolymers. // J. Chem. Phys., 1994, vol. 101, p. 1519.

10. Chan H.S., Dill K.A. Energy landscape and the collapse dynamics of homopolymers. // J. Chem. Phys. 1993, vol. 99, p. 2116; Transition states and folding dynamics of proteins and heteropolymers. // J. Chem. Phys., 1994, vol. 100, p. 9238.
11. Cieplak M., Henkel M., Karbowski J., Banavar J.R. Kinetic traps in exactly solvable models of proteins. // Phys. Rev. Lett., 1998, vol. 80, p. 3654.
12. Kreuzer H.J. Nonequilibrium Thermodynamics and its Statistical Foundations. Oxford, Oxford University Press, 1981, ch. 10.
13. van Kampen N.G. Stochastic Processes in Physics and Chemistry. Amsterdam, North Holland, 1981; Schnakenberg J. Network theory of microscopic and macroscopic behaviour of master equation systems. // Rev. Mod. Phys., 1976, vol. 48, p. 571.
14. Alcaraz F.C., Droz M., Henkel M., Rittenberg V. Reaction-diffusion processes, critical dynamics and quantum chains. // Ann. of Phys., 1994, vol. 230, p. 250.
15. Henkel M. Conformal Invariance and Critical Phenomena. Heidelberg, Springer, 1999, ch. 9.
16. Dagotto E., Correlated electrons in hig-temperature superconductors. // Rev. Mod. Phys., 1994, vol. 66, p. 763
17. Mirny L.A., Abkevich V., Shakhnovich E.I. Universality and diversity of the protein folding scenarios: A comprehensible analysis with the aid of a lattice model. // Fold. Des., 1996, vol. 1, p.103.

Метод керуючого рівняння у явищі скручування білків

М.Цєпляк [1], М.Генкел [2], Дж.Р.Беневер [3]

[1] Інститут фізики Польської Академії Наук та Коледж природничих наук, 02–668 Варшава, Польща

[2] Лабораторія фізики матеріалів, Університет Генрі Пуанкаре Нансі I, F–54506 Вандоувр, Франція

[3] Факультет фізики і Центр фізики матеріалів, лабораторія Дейвей 104, університет штату Пенсильванія, Юнівесіті Пак, PA 16802, США

Отримано 25 червня 1998 р.

За допомогою методу керуючого рівняння точно проаналізована динаміка двох 12-мономерних гетерополімерів на квадратній гратці. Визначена часова еволюція зайнятості нативного стану. При низьких температурах середній час скручування підлягає закону Ареніуса і визначається найдовшим часом релаксації. Для білків, що добре скручуються, з'являються суттєві кінетичні пастки у наборі послідовних конформацій, у той час як для білків, що погано скручуються, пастки присутні також і в ділянках, що не відповідають нативній конформації.

Ключові слова: *скручування білків, кінетичні рівняння*

PACS: *87.15.By, 87.10.+e*