МОЛЕКУЛЯРНА БІОФІЗИКА

Clustering Monte Carlo simulations of the hierarchical protein folding on a simple lattice model

S. O. Yesylevskyy[1], A. P. Demchenko[1, 2]

[1]Palladin Institute of Biochemistry of the National Academy of Sciences of Ukraine
Leontovycha str. 9, Kyiv 01030, Ukraine

[1, 2]Research Institute for Genetic Engineering and Biotechnology
TUBITAK, Gebze-Kocaeli, 41470, Turkey

A role of specific collective motions and clustering behavior in protein folding was investigated using simple 2D lattice models. Two model peptides, which have the sequences of hierarchical and non-hierarchical design, were studied comparatively. Simulations were performed using three methods: Metropolis Monte Carlo with the local move set, Metropolis Monte Carlo with unspecific rigid rotations, and the Clustering Monte Carlo (CMC) algorithm that has been recently described by the authors. The latter was developed with particular aim to provide a realistic description of cluster dynamics. We present convincing evidence that the folding pathways and kinetics of hierarchically folding sequence are not adequately described in conventional MC simulations. In this case the account for cluster dynamics provided by CMC algorithm reveals important features of folding of hierarchically organized sequences. Our data suggest that the methods, which enable specific cluster motions, should be used for realistic description of hierarchical folding.

Introduction. The ideas that proteins fold hierarchically, by sequential formation and association of clusters of residues with increasing their size and complexity, are in the minds of many researchers [1—11]. Definitely, the process of folding is not a one-step event, and the acquisition of native structure occurs via formation and ordering its less organized elements. The sequential acquisition of structure can explain the observation of equilibrium and kinetic intermediates [3—6, 10, 11] including those with non-native structural elements [12—15]. This mechanism explains the observed very fast kinetics of the folding process [16] and provides a clear solution of Levinthal paradox [17]. Instead of global unfolding-folding equilibrium, a spectrum of available protein conformations is observed in hydrogen exchange experiments [18—20].

The hierarchical nature of fluctuations in the native state at equilibrium has been established based on these data, and the possibility that they may represent folding intermediates of variable complexity has been suggested. During the process of folding the appearance of stable structures of larger and larger dimension should change the whole broad spectrum of collective motions [20]. When some group of residues forms a cluster stabilized by non-covalent interactions, then there appear new degrees of freedom, which are the rotations and translations of the cluster, with dramatic reduction of conformational space available for individual residues forming the cluster [21]. Since such mechanism seems reasonable and supported by numerous experimental observations, have been many attempts for its modeling and simulation with different objectives and on different level of complexity [7—9, 22, 23].

The extreme complexity of the folding process justifies the development of highly simplified models [24—27]. Lacking the details, these models should be able to observe the role of basic physical principles otherwise hidden by atomic description and capture essential elementary events of the folding. The most popular examples of simplified models are the lattice models, in which the residues are represented by beads connected by rigid «sticks». In these models the motion of a chain is restricted to a lattice, and the only allowed interactions are the interactions with the nearest neighbors. The folding of lattice proteins is usually simulated by different Monte Carlo (MC) algorithms [26]. These methods provide an easy way for global energy minimization of the system, which is thought to be reached in the native state of the folding chain. Monte Carlo simulations play an important role in investigation of essential steps of protein folding and provide the observations of folding nuclei, reaction bottlenecks, misfolded and intermediate states and so on [24—27].

Despite this apparent success the results of these simulations can not be easily accepted as representing the real mechanisms of protein folding. The lattice models allow only a rough course-grained description of the protein conformational space, which keeps the model very simple and makes the folding problem computationally tractable. Despite the simplified treatment of the conformational space the models may still correctly describe the basic folding dynamics of a real protein. This is possible if the motions of the chain on the lattice correspond to the actually occurring motions of the real chain.

However, existing methods of Monte Carlo simulations of lattice proteins seem to oversimplify the dynamics of the chain and thus have several serious weak points. One of them is a limitation on each MC step imposed by the fixed set of allowed elementary moves. The choice of the move set is a widely discussed question, which is not completely solved yet [28]. The «classical» MC studies were performed using the Local Move Set (LMS) [24]. It is a minimal move set for sampling the major part of conformational space of the chain. It includes only three types of moves: pivot and corner single-bead moves and crankshaft moves of two beads. This move set is definitely far remote from physical reality, because it does not allow analyzing possible collective motions. The latter can be very important, especially at the final steps of folding. Another widely used move set is LMS with the unspecific rigid rotations added. We will call it MS2 following Chan and Dill [29]. MS2 allows collective «diffusional» motions and therefore is considered to be much more realistic.

Both abovementioned move sets and their numerous modifications have one serious drawback. For a given bead the move set does not depend on the current state of the chain. It does not depend on the current chain topology and remains the same even if a particular bead forms the contacts with its neighbors. In other words, the move sets LMS and MS2 are structure-independent, or unspecific. In real proteins, however, the conformational mobility of the amino acid residues is always controlled by the local and non-local interactions and the chain topology. It should definitely restrict the motion along certain degrees of freedom and diminish dramatically the available conformational space of the chain.

Collective motions are present in MS2, but they are also structure-independent. Any segment may be involved into rigid rotation regardless of its topology. Particularly, non-compact (even linear) chain segments may be rotated as rigid «sticks», which is obviously far from physical reality. More realistic description of the collective motions should consider the motions which are specific to current topological state of the chain and which depend on the formed contacts.

Thus, a new step in the development of MC algorithms in their application to lattice models of protein folding is required. We need to provide a description of collective motions with regard to increase in their complexity in the course of folding and of their dependence on the current chain topology and energetics. In order to achieve this goal we have inevitably to address the questions on the validity of the micro-reversibility postulate, which is in the background of all conventional MC simulations. Based on this postulate, the energies of two sequencial conformations are compared and each elementary step (the change of chain configuration) is either accepted or rejected according to Metropolis or some other similar criterion. Acceptance criterion is based on the assumption that the folding chain is considered to be in local equilibrium described by Bolzmann energy distribution. This means that the transitions between sequential conformations are micro-reversible. Meantime, in physical reality the collective motions, and the cluster motions in particular, are essentially irreversible, so that the destruction of the large compact structure and its re-assembly may often follow completely different trajectories in the conformational space. It is possible however to subdivide a complex collective motion into small steps, so that each of them may be considered as a micro-reversible transition. Applying Metropolis criterion to each step, in principle one can describe any collective motion. However, this is not possible with the coarse-grained lattice models. On the square lattice, for example, a cluster can be rotated by only 90 degrees at once, and no intermediate cluster positions are allowed. So, in order to simulate possible hierarchical cluster formation there appears the necessity not only to modify basic MC methods but also to introduce new concepts.

The aim of our work is to develop a concept that could overcome these difficulties and to elaborate a novel MC method, which explicitly simulates the formation, motion and destruction of clusters appearing during the folding process. This concept can be called the Clustering Monte Carlo (CMC) algorithm.

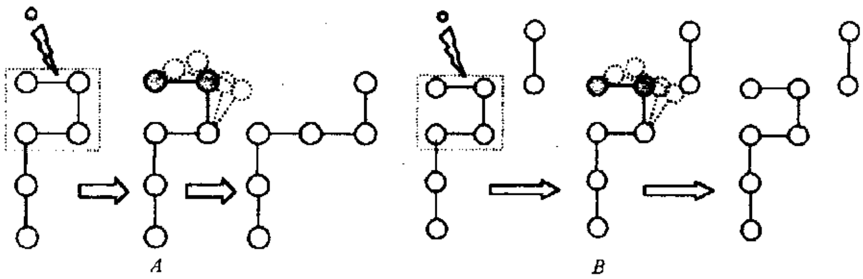

Fig. 1. Cluster disruption in CMC algorithm. Cluster is indicated by dashed box. In case (*a*) there are no steric constraints for rotation, and the large thermal fluctuations rotate the part of the cluster (shown in bold) and disrupt it. In case (*b*) the rotation is restricted by the proximately located part of the chain shown schematically to the right of the cluster. In this case the fluctuation energy dissipates and the cluster remains intact. Rotation of the whole clusters is performed according to the same principle

Preliminary studies [9] have demonstrated that CMC algorithm allows finding correctly the unique energy minimum of the folded state for a short 12-member peptide. In this communication we report on the comparative study of two model sequences which represent hierarchical and non-hierarchical folding pathways using CMC and conventional LMS and MS2 methods. We demonstrate the possibility for a much more realistic description of the folding pathway and kinetics compared to conventional MC simulations with the unspecific move set. In simulations of folding of hierarchical sequences the account for specific collective motions of clusters including their formation and dissociation can be realized.

Materials and Methods. *Clustering Monte-Carlo algorithm.* The basic idea behind this algorithm is to consider the motion of not a single residue, but of a cluster of residues of variable size ranging from a single residue to a whole protein sequence. Then the folding can be described as the process of growth and association of clusters. The cluster is defined as a set of residues connected by non-covalent bonds, which form a sterically rigid structure. This means that no part of the cluster can be moved or rotated without breaking the bonds connecting its elements. An elementary conformational change in our model is a rotation of a cluster or its part (the latter means in fact the cluster breakage).

Linear motions of the clusters are taken into account implicitly, they occur when the rotation of one cluster «pulls» the other one. In the course of folding with the increasing size of the cluster these elementary changes start to involve collective motions of larger number of residues. So, the «scale» of elementary act in our model is variable and it always corresponds to the current size of the formed cluster. This eliminates from the process of folding the «frozen» degrees of freedom inside the clusters and allows to provide a correct description of collective motions on different scales. It is necessary to emphasize that there is no predefined move set in CMC in its conventional meaning because the cluster rotation may involve any number of residues and may cause different chain rearrangements.

As it was stated in introduction, the coarse-grained lattice does not allow the rotations of the large clusters to be micro-reversible. We have found a solution of this problem by considering irreversible cluster rotations triggered by local thermal fluctuations. The process of cluster breakage or rotation is divided into two independent steps. On the first step the cluster is provided with additional energy E_f which is the energy of thermal fluctuation. The second step is a «decision making» process. If the fluctuation energy E_f is larger than the bond-breaking energy E_d (which is the energy of the bonds needed to be broken in order to perform a rotation), the cluster has to break apart. Otherwise the cluster will rotate as a whole. If there are some external steric restrictions, the energy will dissipate with no result in change of chain configuration (Fig. 1).

In the simplest case the fluctuations may be considered as collisions with the solvent molecules. Therefore the energies of fluctuations may be described by Bolzmann distribution

$$P(E_f) = \frac{\exp(-E/k_BT)}{T}.$$

In reality the cluster rotation should be controlled by solvent viscosity. Introduction of some constant energy E_{r0}, which is needed to rotate a single residue cluster (of a minimal size), simulates this effect. Larger clusters in order to be rotated will need the energy $E_r = n \cdot E_{r0}$, where n is a size of the cluster. If $E_f < E_r$, the cluster can not rotate at all. Thus, E_{r0} can be considered as a measure of the energies in CMC that allow or not allow a particular motion.

It is clear, that the clusters with large bond-breaking energies will be stable, while the clusters with small or negative (destructive) energies will break apart almost immediately. In other words, the system will perform an evolution in the search for an energy minimum by selection of clusters with higher stability.

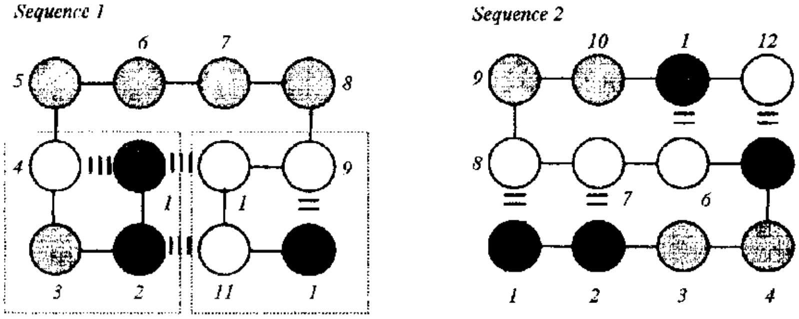

Fig. 2. The structures of the studied 12-member peptides, sequence 1 and sequence 2, in their native conformations. Positively charged residues are marked black, negatively charged are white, neutral residues are gray. Dashed boxes in the case of sequence 1 indicate the clusters of the first hierarchical level. Non-local bonds with non-zero energy are shown by the striped bars

Thus, CMC algorithm has clear physical interpretation and does not require the highly questionable assumptions of micro-reversibility and of the presence of local thermodynamic equilibrium in elementary changes of chain configuration. It suggests a solution for the problem of description of structure-specific collective motions and is applicable to the cases of hierarchical cluster formation. In principle this approach allows the analysis of folding process in the systems of any size and any number of hierarchical levels.

Protein model. In order to illustrate the applicability and the advantages of CMC algorithm we have selected two «minimal» sequences, the first one exhibiting hierarchical behavior and the other-not. We studied two pre-designed lattice peptides with 12 residues each, they are referred below as sequence 1 and sequence 2. The native states of these sequences are unique compact states with the minimal energy. Sequence 1 forms a hierarchically organized compact structure (Fig. 2, *a*), which folds into a 3 × 4 bar on the lattice. Residues 1—4 and 9—12 form two small clusters of the first hierarchical level (dashed boxes) which can assemble into the cluster of the second level. The latter contains all the residues. Single non-covalent bonds 1—4 and 9—12 stabilize the structure of the first-level clusters respectively. Two bonds 1—10 and 2—11 combine clusters. Residues 5—8 form a connecting loop. In contrast, sequence 2 has no hierarchical featurres (Fig. 2, *b*). It folds into a «β sheet» with 3 strands stabilized by the bonds 1—8, 2—7, 6—11 and 5—12.

Compared to the common two-letter model (HP model) we have made a step towards a more realistic description of interaction between the residues. The HP model considers only attractive and neutral interactions. In contrast, our three-letter model operates with three types of interactions: attractive, repulsive and neutral. They are represented by negatively charged (type 1), positively charged (type 2) and neutral (type 3) residues. We have introduced attraction and repulsion energies for the charged residues located in adjacent vertices of the lattice. Neutral residues do not interact with the residues of any type. The rotation energy of a single residue cluster E_{r0} is taken as a unit of energy. Each bond between charged residues is assumed to have the energy 50 in the case of charges of the same sign and −50 in case of charges of the opposite sign. All other bonds are of zero energy. Both folded structures (sequences 1 and 2) attain unique native states (Fig. 2) with the energy −200 E_{r0} units each.

Simulation methods. Three series of simulations have been performed comparatively using CMC algorithm and Metropolis Monte-Carlo algorithms with LMS and MS2 move sets. The chains were equilibrated at the high temperature $T = 1000$ for 1000 iterations to generate a random unfolded conformation (the temperature is measured in dimensionless units $k_B T_{abs}/E_{r0}$). Then the temperature was abruptly lowered to the desired level and the simulation proceeded up to the first folding to the native structure. The number of averaged independent runs for each temperature is 1000.

The folding process was monitored by three progress parameters: the number of native contacts N_n, the total number of contacts N_t and the energy of the structure E. We constructed a set of all possible chain conformations by their exhaustive enumeration. The energy landscapes of the studied sequences were constructed by calculating the average energies of conformations from this set, which attain particular values of N_n and N_t.

Integrated residence time maps were constructed by monitoring the number of iterations spent by the sequence in the state with given progress parameters averaged over 1000 independent runs and normalized to unity.

We calculated the values of the progress parameters for the final 100 iterations (1000 for LMS simulations of the sequence 1) averaged over 1000 runs for various temperatures and used them for kinetics studies.

Results and Discussion. *Energy landscapes of the*

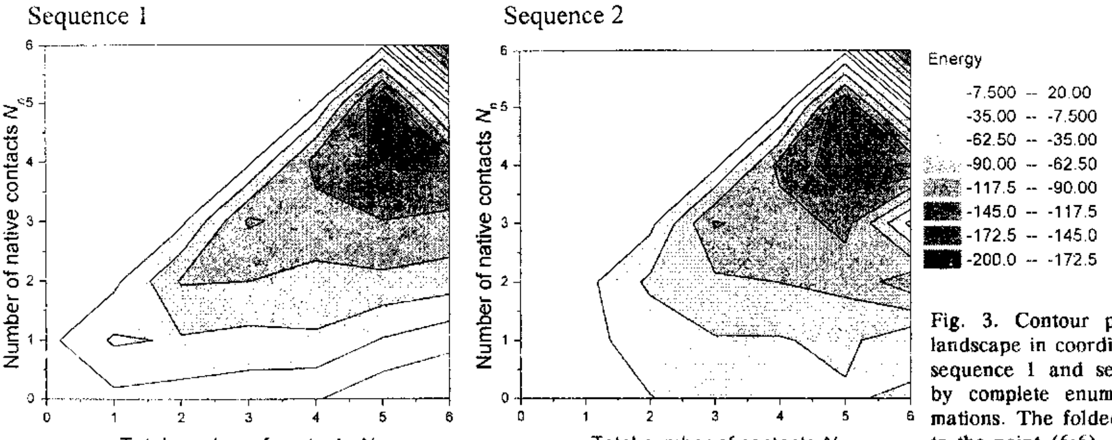

Fig. 3. Contour plot of the energy landscape in coordinates Nn vs. Nt for sequence 1 and sequence 2 obtained by complete enumeration of conformations. The folded state corresponds to the point (6:6)

studied sequences. The 2D energy landscapes for our sequences in coordinates N_n vs. N_t were constructed by exhaustive enumeration of all chain conformations (Fig. 3). For both sequences the native state corresponds to the upper right corner of $N_n - N_t$ diagram with coordinates (6:6). Both sequences have pronounced non-native energy minima at the points (5; 5) with the energy −150. The energy landscape for sequence 2 is more rugged. It has two additional minima at points (6:4) and (6:2) corresponding to fully compact but misfolded conformations with the average energies −150 and −108 respectively.

Integrated residence time maps provide important information about the folding pathway, especially about the intermediates and misfolded conformations emerged during the folding process.

Integrated residence time maps for sequence 1 are shown in Fig. 4, *a*. The maps obtained by CMC simulations show existence of several types of folding intermediates. For small temperatures (T = 10) the folding pathway is dominated by intermediates with 5 contacts, 3 of which are native. They may be classified as semi-compact intermediates. The most probable chain conformation in this region is composed of two folded clusters of the first level combined by non-native bonds. A significant part of the folding time is spent in those intermediates or in nearby regions of the map.

For higher temperatures (T = 15—20) there appears the second broad region of non-compact intermediates located at the region (1:1—3:2). It corresponds to one or two correctly formed clusters of the first level connected by unfolded loop. With the increase of temperature the amount of time spent in these less compact configurations progressively increases. When the temperature reaches 80—100 (data not shown), the strength of the single bond (50) becomes too small to stabilize the clusters of the first level. This temperature corresponds to denaturation conditions, so the totally unfolded state dominates the folding pathway.

Comparison with the energy landscape in Fig. 3 shows that non-native energy minimum at point (5:5) is occupied rarely. This means that the folding pathway does not always follow the gradient of energy, but rather goes through the kinetically accessible conformations.

For LMS and MS2 the integrated residence time maps are almost identical. It means that both methods during the folding actually sample the same configurations. The features observed in the integrated residence time maps for LMS/MS2 are qualitatively similar to that obtained in CMC simulations. With the increase of temperature the amount of time spent in semi-compact states decreases, while the non-compact states with one or two formed first level clusters become more probable (Fig. 4, *a*). However, there are several pronounced differences between these maps for CMC and LMS/MS2. For LMS/MS2 there is an additional region of intermediates at (3:3). At small temperatures these intermediates coexist with the semi-compact intermediates. The non-compact intermediates for LMS/MS2 are well separated and the totally unfolded state become dominant for much lower temperatures (40—50) than in the case of CMC.

The integrated residence time maps for sequence 2 are shown in Fig. 4, *b*. The maps obtained by CMC are fundamentally different in two ways. First of all, for small temperatures two deep non-native minima at points (6:4) and (6:2) are occupied with high probability (compact intermediates), whereas the single non-native minimum for sequence 1 with the same energy −150 is never occupied. With the increase of

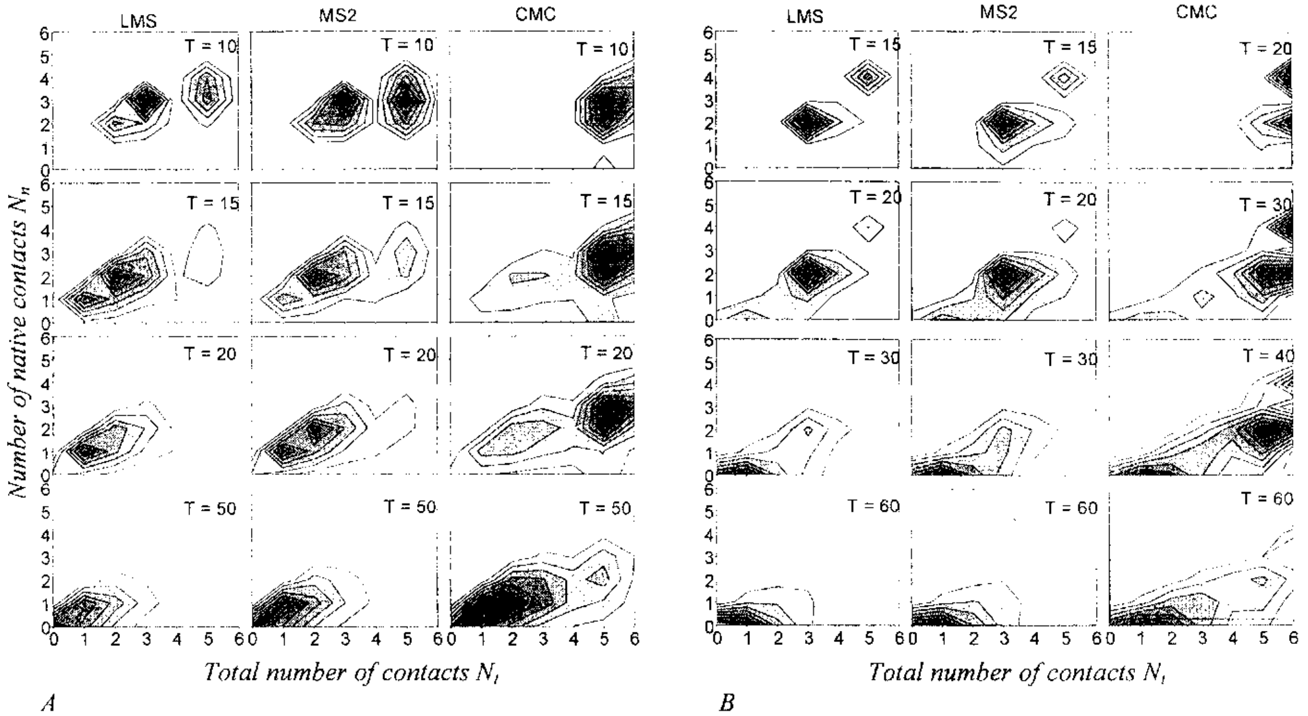

Fig. 4. Integrated residence time maps for sequence 1 (*A*) and sequence 2 (*B*) for various temperatures obtained in LMS, MS2 and CMC simulations. Colors represent the amount of time spent by the sequence in the given point, white corresponds to zero, black to the maximum

temperature a second region of non-compact intermediates with 1 or 2 contacts appear, but these contacts are all not native. This is in contrast with the results obtained for sequence 1, which show a non-compact intermediates with the native contacts corresponding to the first level clusters.

These features are easily explained by considering the hierarchical character of folding of sequence 1. During the folding of sequence 1 the clusters of the first level are likely to appear first. If the temperature is small enough the clusters will be very stable, and their persistence will not allow the chain to sample certain parts of conformational space. This space can be accessible only after clusters' destruction. Non-native minimum at the point (5:5) belongs to these conformations, so the chain is very unlikely to reach it. Instead the chain will form some structures with a misfolded cluster arrangement and with a higher energy. With the increase of temperature these conformations will dissociate into individual clusters: there appears the region of non-compact intermediates. Since all contacts are inside the clusters, they are native. The clusters break apart only at very high temperatures leading to complete unfolding.

A different picture is observed for sequence 2 that lacks clustering behavior. In this case the whole conformational space remains accessible at small temperatures. As a result, the non-native compact states are frequently occupied. With the increase of temperature these conformations dissociate directly to completely unfolded state with 1 or 2 accidental non-native contacts.

LMS and MS2 maps for sequence 2 are also identical and essentially different from CMC maps (Fig. 4, *b*). There are no compact intermediates observed in LMS/MS2 simulations even for lowest computationally possible temperatures. Highly populated states at the points (5:3) and (3:2) are observed instead. With the increase of temperature these states disappear and the completely unfolded state dominates. These results suggest that both LMS and MS2 move sets are not efficient in finding local energy minima for sequence 2, thus the chain is not trapped in the lowest non-native minima. CMC finds these minima and allows observing the chain trapping.

Kinetics. For kinetics studies the first folding times of 1000 independent MC runs were grouped into 20—50 bins forming the histograms. All obtained histograms were accurately fitted by the single exponential functions $A \cdot \exp(-t/\tau_U)$. The mean folding time τ_U was obtained from histogram fitting.

Sequence 1. Temperature dependencies of the mean folding times obtained for sequence 1 are shown in Fig. 5, *a*. The shapes of the curves obtained by all simulation techniques are similar. An optimal temperature is about 20 in the case of CMC and LMS and

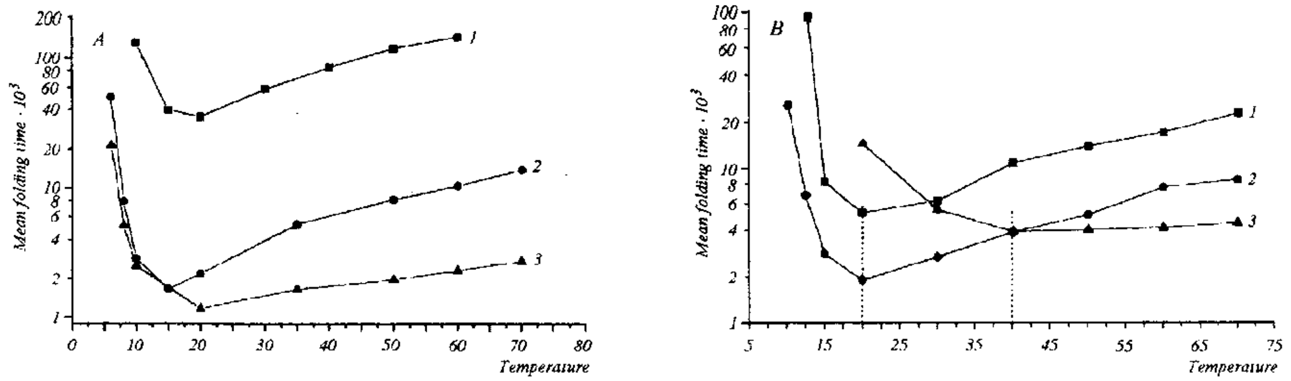

Fig. 5. Mean folding times for sequence 1 (*A*) and sequencq 2 (*B*) obtained in LMS (*1*), MS2 (*2*) and CMC (*3*) and MS2 simulations

about 15 in the case of MS2. However the mean folding times are quite different. LMS shows the slowest folding, which is more than 10 times slower than in the cases of MS2 and CMC. It is quite possible that such a large difference in the folding times is caused (at least partially) by the native topology of the sequence 1 which belongs to the so-called «buried-end» sequences. Lacking collective motions, LMS is likely to get into the «topological trap» which can retard the folding dramatically.

The MS2 and CMC modeling of chain folding is not sensitive to the chain topology because of the collective motions allowed. Both methods show fast folding, however, CMC is faster in terms of the minimal mean folding time.

Temperature dependence of the mean folding time for CMC is more shallow which leads to mush smaller folding times at high temperatures. This correlates with the peculiarities on the integrated residence time maps: the completely denatured state becomes dominant under CMC only at the temperatures 80—100, which are much higher than 40—50 observed for MS2.

Sequence 2. The temperature dependencies of the mean folding times obtained for sequence 2 are shown in Fig. 5, *b*. Since sequence 2 does not have buried ends, there is no topological trap in the case of LMS simulations. However, LMS still shows the slowest folding. Both LMS and MS2 give qualitatively similar temperature dependencies of the mean folding times with the optimal temperature about 20. Meantime, the temperature dependence obtained by CMC simulations is very different. An optimal temperature is twice higher (about 40) and the high temperature tail of the curve is much more shallow, which makes the folding time almost temperature independent for high temperatures. The minimal folding time for CMC is larger than given by MS2 but smaller than given by LMS.

Hierarchical and non-hierarchical folding in different simulation techniques. It is necessary to emphasize that both sequences that are analyzed here have identical number of native contacts and the same energies of the native state, so the differences in kinetics of folding appear only due to the differences in folding mechanisms. In order to fold correctly, the hierarchical sequence 1 has to form stable clusters, while the non-hierarchical sequence 2 has to attain the sheet topology that does not require the formation of clusters.

As it is evident from Figs 4, *b* and 5, *a*, the mean folding time of sequence 1 in LMS is 10 times longer in comparison with sequence 2. We believe that this is because of topological trapping in the former sequence possessing buried ends. Unfortunately there are no open-end hierarchical sequences of length 12 which have a non-degenerate ground state (it was verified by exhaustive enumeration), so it is not possible to test the possibility of trapping by direct comparison. Since the collective motions are important in this case and LMS does not describe them, we concentrate on comparison of MS2 and CMC data.

An optimal folding temperature for sequence 1 in CMC simulations is well below the point where the clusters of the first level begin to dissociate. This means that the fastest folding is achieved in the conditions, in which these clusters maintain integrity during the folding process. In contrast, the optimal folding temperature for the non-hierarchical sequence 2 is much higher. This is the result of formation on the folding pathway of misfolded intermediates, which have to unfold in order to proceed toward the native state. So an optimal temperature in this case should be high enough to effectively break the non-native

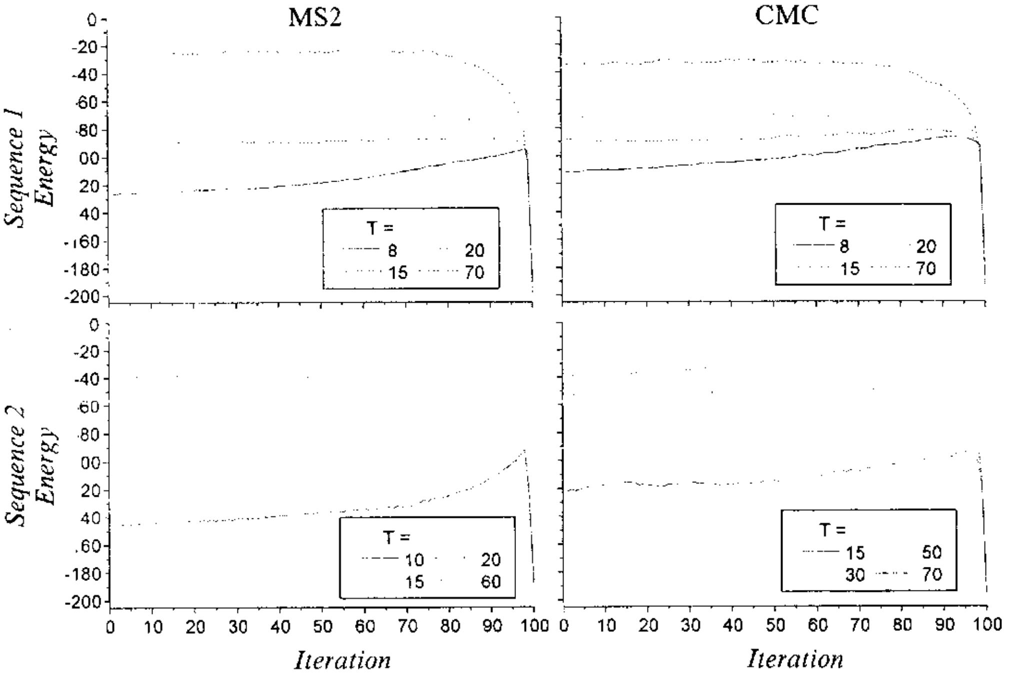

Fig. 6. Average energies for the final 100 steps of folding of sequence 1 and sequence 2 at different temperatures obtained in MS2 and CMC simulations. The folding proceeds from left to right. The last point corresponds to the native state

contacts. These misfolded states are well seen on the CMC integrated residence time maps for sequence 2 as compact intermediates. They correspond to deep local energy minima, which are efficiently sampled by CMC.

A different picture is observed with LMS/MS2 simulations. They show similar optimal temperature of 15—20 for both sequences. LMS/MS2 simulations disregard formation of the stable clusters, so when simulated by these methods the sequence 2 does not get trapped in the misfolded states. Corresponding integrated residence time maps show that indeed the deep local energy minima are not sampled. This explains why the optimal temperatures are identical for both sequences.

Since hierarchical sequence 1 forms two most proximal contacts in the sequence (contacts 1—4 and 9—12, see Fig. 2) it is expectable that it will fold faster than sequence 2, which has only the contacts between remote regions of the chain. This is really observed in the case of CMC simulations. They show that hierarchical sequence 1 is a fast folder (Figs 4, *b* and 5, *a*). In contrast, MS2 simulations show almost identical folding times for the both sequences.

So, how can it happen that according to MS2 simulations the non-hierarchical sequence 2 folds as fast as hierarchical sequence 1? In our view this behavior is a direct consequence of the fact that MS2 move set includes unspecific, and thus unnatural, collective motions. Particularly MS2 allows rigid rotations of the long linear segments, which have a pivot point at the end move as rigid «sticks». In reality such long segments will never behave as sticks but rather as highly flexible soft ropes, which tend to form a compact coil. The native beta-sheet topology of sequence 2 can be easily reached in MS2 in just several «stick» moves, which lead to overestimation of the folding rate. In contrast, CMC allows only the rotations of clusters, which are indeed compact rigid structures stabilized by internal bonds. Linear chain segments in CMC will never be translocated as a whole. The formation of a beta-sheet structure in CMC occurs on a much longer time scale than the formation of compact clusters, which is physically more realistic.

Energetics of the final folding steps. In order to better understand the folding mechanism and kinetics of the studied sequences, we have calculated the averaged energies of the final 100 iterations for sequences 1 and 2 (Fig. 6). The remarkable feature of the obtained relations is the presence on the average folding pathway of the definite energy barrier. For the temperatures smaller than the optimal folding temperature the average energy of the chain increases slowly with time and reaches the maximum of approximately −100.

This process is accompanied by the decrease in the total number of contacts and by increase in the number of native contacts (not shown). This signifies the appearance of the states that are non-compact but enriched in native contacts. The state with the maximal energy is observed only 2 iterations before complete folding in the case of MS2 and 7—10 iterations before in the case of CMC. The maximum observed in MS2 simulations is very sharp, while in the case of CMC a smooth broad region is observed at the top of the barrier. After reaching the maximum value the energy decreases rapidly on the motion toward the native state. With the increase of temperature the barrier becomes less pronounced and finally disappears at the temperature close to the optimal folding temperature. In this case the curve is essentially «flat» up to the critical point where there appear a rapid collapse to the native state. For higher temperatures the average energy decreases smoothly up to the native state starting from the point, which is 20—30 iterations before the complete folding. The observed behavior is a direct consequence of the existence of semi-compact and compact misfolded intermediates, which are detected on the integrated residence time maps (Fig. 4). Destruction of these misfolded conformations requires the overall decrease of the chain compactness and increase of its energy.

Thus, an energy barrier appears on the folding pathway, which has to be surmounted by the chain in order to reach the native state. Very sharp collapse to the native state, which lasts only 1—3 iterations, suggests that there is a bottleneck on the folding pathway. It is determined by chain conformations on the top of the energy barrier, the majority of which have the energy of −100. In the case of sequence 1 this corresponds to two correctly formed clusters of the first level, which then combine to form the native state. In the case of sequence 2 the bottleneck conformations correspond to the single correctly formed beta-hairpin (containing either the residues 1—8 or 5—12) stabilized by two native contacts. The rest of the chain remains unfolded. Native state is reached when the second hairpin forms. Sharpness of the barrier observed in MS2 simulations reflects unspecific nature of the collective motions described in MS2. This fact can be easily explained in the case of sequence 1. After the ends of the chain attain correct topology, they can be assembled to the native structure by the single «stick» move. This can happen with the majority of conformations, which from the top of the energy barrier, and an abrupt collapse toward the native state makes the barrier sharp. In contrast, in the case of CMC the number of conformations containing two correctly formed first-level clusters on the pathway to the native state is much smaller. Once the clusters are formed, they will diffuse for a while until the unfolded linker segment attains the correct conformation after that single rotation can assemble a native structure. The diffusion step, which does not involve the changes in the chain energy, makes the barrier broad. Similar considerations are applicable to the sequence 2.

Conclusions. The results of this work demonstrate the difference in folding behavior between non-hierarchical and hierarchical sequences, which have been revealed by all used simulation methods. In the case of the non-hierarchical sequence all simulations result in similar values of the folding rates. However, CMC is much more efficient in finding local energy minima, which leads to much higher optimal folding temperature in comparison with LMS and MS2. Due to the fact, that LMS disregards collective motions, it underestimates the folding rate. In contrast, MS2 overestimates it due to the unspecific nature of allowed collective motions and the necessity to involve the physically unjustified «stick» moves. CMC seems to provide the most accurate description of the collective motions and thus shows intermediate folding rates.

In the case of the hierarchically organized sequence the LMC algorithm fails to produce realistic folding rates, which is probably a consequence of topological trapping in the sequence with buried ends. Meantime, MS2 and CMC, which allow collective motions, demonstrate very similar maximal folding rates CMC is a bit faster. Comparison of the energetics of the last folding steps shows that CMC describes diffusive motions of the stable clusters and the dynamics of the open segments accurately. Collective motions in MS2 are described in a less realistic way because this method makes no difference between open segments and stable compact clusters. This leads to overestimation of the clusters' mobility and to unnatural dynamics of the open segments.

The CMC algorithm shows directly the reduction of effective conformational space caused by cluster formation. Due to this process even at initial steps of folding certain regions of conformational space become unavailable for the sequence. We show that CMC and LMS/MS2 algorithms sample very different parts of conformational space. This is a direct consequence of unspecific nature of LMS and MS2 move sets, which provide an inadequate description of collective dynamics.

The amino acid sequences can be roughly classified into fast and slow folders. Should this classification depend on the simulation method used? Probably not, and since our study shows that the applied methods are not equivalent in following the folding kinetics, we have to make a choice in favor of the method that suggests a physically more realistic picture of folding events. We observe that the hierarchical sequence in CMC is a fast folder, while in MS2 both hierarchical and non-hierarchical sequences show very similar folding times. This provides us the reason to believe that hierarchically organized sequences fold in a hierarchical manner, and the suggested CMC algorithm provides a step for a more realistic description of this process.

In our simulations we were able to demonstrate that the average folding pathway is not a continuous decrease of the chain energy toward the native state. For low temperatures the folding sequence has to surmount some energy barrier breaking the non-native contacts of initially formed misfolded conformations, and this slows down the folding process. With the increase of temperature the compact and semi-compact misfolded states become unstable and the energy barrier disappears providing the maximal folding rate. Further increase of temperature destabilizes non-compact intermediates, which are essential for correct folding, and thus decreases the folding rate again.

Thus, our simulations demonstrate not only an exceptional importance of collective motions in the folding simulations but also an importance of physically correct description of collective motions. In this case the folding process can not be adequately described neither in MC simulations with the local move set (such as LMS) nor in simulations with unspecific collective motions (such as MS2). We believe that in these cases the accounting for specific collective motions i. e. for cluster dynamics (as it is implemented in CMC) provides much more realistic description of the folding pathways and kinetics.

С. О. Єсилевський, О. П. Демченко

Моделювання ієрархічного фолдингу білків у простій гратковій моделі за допомогою кластерного методу Монте-Карло

Резюме

Досліджено роль специфічних колективних рухів та кластерної поведінки у фолдингу білків з використанням простих двовимірних граткових моделей. Проведено порівняльний аналіз пептидів з ієрархічною та неієрархічною будовою. Моделювання здійснювали за допомогою трьох методів: стандартного методу Монте-Карло з локальним набором рухів, стандартного методу з неспецифічними колективними обертаннями та кластерного методу Монте-Карло (СМС) запропонованого авторами для реалістичного моделювання динаміки кластерів. Показано, що шляхи та кінетика ієрархічного фолдингу не можуть бути адекватно описані звичайними методами. У цьому випадку врахування кластерної динаміки у методі СМС виявляє важливі риси ієрархічного фолдингу. Визначено, що для реалістичного моделювання ієрархічного фолдингу потрібно використовувати розрахункові методи, які враховують специфічні колективні рухи.

С. А. Есилевский, А. П. Демченко

Моделирование иерархического фолдинга белков в простой решеточной модели с помощью кластерного метода Монте-Карло

Резюме

Исследована роль специфических коллективных движений и динамики кластеров в фолдинге белков с использованием простых двухмерных решеточных моделей. Проведен сравнительный анализ фолдинга пептидов с иерархической и неиерархической структурой. Моделирование осуществляли с помощью трех методов: стандартного метода Монте-Карло с локальным набором движений, стандартного метода с неспецифическими коллективными вращениями и кластерного метода Монте-Карло (СМС), предложенного авторами для реалистичного описания динамики кластеров. Показано, что пути и кинетика иерархического фолдинга не могут быть адекватно описаны стандартными методами. В этом случае учет кластерной динамики в методе СМС выявляет важные особенности иерархического фолдинга. Обнаружено, что для реалистичного моделирования иерархического фолдинга должны использоваться методы, учитывающие специфические коллективные движения.

REFERENCES

1. *Murphy K. P., Bhakuni V., Xie D., Freire E.* Molecular basis of co-operativity in protein folding. III. Structural identification of cooperative folding units and folding intermediates // J. Mol. Biol.—1992.—227.—P. 293—306.
2. *Karplus M., Weaver D. L.* Protein folding dynamics: the diffusion-collision model and experimental data // Protein Sci.—1994.—3.—P. 650—668.
3. *Jamin M., Antalik M., Loh S. N., Bolen D. W., Baldwin R. L.* The unfolding enthalpy of the pH 4 molten globule of apomyoglobin measured by isothermal titration calorimetry // Protein Sci.—2000.—9.—P. 1340.
4. *Akiyama S., Takahashi S., Ishimori K., Morishima I.* Stepwise formation of alpha-helices during cytochrome c folding // Nat. Struct. Biol.—2000.—7.—P. 514.
5. *Raschke T. M., Kho J., Marqusee S.* Confirmation of the hierarchical folding of RNAse H: a protein engineering study // Nat. Struct. Biol.—1999.—6.—P. 825—831.
6. *Hua Q. X., Nakagawa S. H., Jia W., Hu S. Q., Chu Y. C., Katsoyannis P. G., Weiss M. A.* Hierarchical protein folding: asymmetric unfolding of an insulin analogue lacking the A7-B7 interchain disulfide bridge // Biochemistry.—2001.—40.—P. 12299—12311.
7. *Tsai C. J., Nussinov R.* Transient, highly populated, building blocks folding model // Cell Biochem. and Biophys.—2001.—34.—P. 209—235.
8. *Tsai C. J., De Laureto P. P., Fontana A., Nussinov R.* Comparison of protein fragments identified by limited proteolysis and by computational cutting of proteins // Protein Sci.—2002.—11.—P. 1753—1770.
9. *Islam S. A., Karplus M., Weaver D. L.* Application of the diffusion-collision model to the folding of three-helix bundle proteins // J. Mol. Biol.—2002.—318.—P. 199—215.
10. *De Jong D., Riley R., Alonso D. O., Daggett V.* Probing the energy landscape of protein folding/unfolding transition states // J. Mol. Biol.—2002.—319.—P. 229—242.
11. *Paci E., Vendruscolo M., Karplus M.* Native and non-native interactions along protein folding and unfolding pathways // Proteins.—2002.—47.—P. 379—392.
12. *Hamada D., Kuroda Y., Tanaka T., Goto Y.* High helical propensity of the peptide fragments derived from beta-lac-

toglobulin, a predominantly beta-sheet protein // J. Mol. Biol.—1995.—254.—P. 737—746.

13. *Hamada D., Segawa S., Goto Y.* Non-native alpha-helical intermediate in the refolding of beta-lactoglobulin, a predominantly beta-sheet protein // Nat. Struct. Biol.—1996.—3.—P. 868—873.

14. *Kuwata K., Shastry R., Cheng H., Hoshino M., Batt C. A., Goto Y., Roder H.* Structural and kinetic characterization of early folding events in beta-lactoglobulin // Nat. Struct. Biol.—2001.—8.—P. 151—155.

15. *Capaldi A. P., Kleanthous C., Radford S. E.* Im7 folding mechanism: misfolding on a path to the native state // Nat. Struct. Biol.—2002.—9.—P. 209—216.

16. *Wagner C., Kiefhaber C.* Intermediates can accelerate protein folding // Proc. Nat. Acad. Sci. USA.—1999.—96.—P. 6716—6721.

17. *Karplus M.* The Levinthal paradox: yesterday and today // Fold. Des.—1997.—2.—P. 69—75

18. *Englander S. W., Kallenbach N. R.* Hydrogen exchange and structural dynamics of proteins and nucleic acids // Quart. Rev. Biophys.—1983.—16.—P. 521—655.

19. *Englander S. W., Mayne L.* Protein folding studied using hydrogen-exchange labeling and two-dimensional NMR // Annu. Rev. Biophys. and Biomol. Struct.—1992.—21.—P. 243—265.

20. *Sadqi M., Casares S., Abril M. A., Lopez-Mayorga O., Conejero-Lara F., Freire E.* The native state conformational ensemble of the SH3 domain from alpha-spectrin // Biochemistry.—1999.—38.—P. 8899—8906.

21. *Brooks C. L. III, Karplus M., Pettitt B. M.* Proteins: a theoretical perspective of dynamics, structure and thermodynamics.—New York: Wiley Intersci., 1988.

22. *Demchenko A. P.* Concepts and misconcepts in the analysis of simple kinetics of protein folding // Curr. Protein Peptide Sci.—2001.—2.—P. 73—98.

23. *Micheletti C., Banavar J. R., Maritan A.* Conformations of proteins in equilibrium // Phys. Rev. Lett.—2001.—87.—P. 88—102.

24. *Erman B.* Analysis of multiple folding routes of proteins by a coarse-grained dynamics model // Biophys. J.—2001.—81.—P. 3534—3544.

25. *Sali A., Shakhnovich E., Karplus M.* How does a protein fold? // Nature.—1994.—369.—P. 248—251.

26. *Pande V. S., Rokhsar D. S.* Folding pathway of a lattice model for proteins // Proc. Nat. Acad. Sci. USA.—1999.—96.—P. 1273—1278.

27. *Mirny L., Shakhnovich E.* Protein folding theory: from lattice to all-atom models // Annu. Rev. Biophys. and Biomol. Struct.—2001.—30.—P. 361—396.

28. *Klimov D. K., Thirumalai D.* Lattice models for proteins reveal multiple folding nuclei for nucleation-collapse mechanism // J. Mol. Biol.—1998.—282.—P. 471—492.

29. *Yesylevskyy S. O., Demchenko A. P.* Modeling the hierarchical protein folding using clustering Monte-Carlo algorithm // Protein and Peptide Lett.—2001.—6.—P. 437—442.

УДК 577.322
Надійшла до редакції 17.03.03