ELSEVIER

Biophysical Chemistry 109 (2004) 17–40

Biophysical Chemistry

www.elsevier.com/locate/bpc

Towards realistic description of collective motions in the lattice protein folding models

S.O. Yesylevskyy[a], A.P. Demchenko[a,b,*]

[a]*A.V. Palladin Institute of Biochemistry, Leontovicha Street 9, Kiev 01030, Ukraine*
[b]*Research Institute for Genetic Engineering and Biotechnology, TUBITAK, Gebze-Kocaeli 41470, Turkey*

Received 3 September 2003; received in revised form 3 October 2003; accepted 3 October 2003

Abstract

Collective motions and the formation of clusters of residues play an important role in the folding of real proteins. However, existing Monte Carlo (MC) techniques of the protein folding simulations based on highly popular lattice models provide only a schematic representation of collective motions, which is rather far from physical reality. The Clustering Monte Carlo (CMC) algorithm was developed with particular aim to provide a realistic description of collective motions on the lattice. CMC allows modeling the cluster dynamics and the effects of the solvent viscosity, which is impossible in conventional algorithms. In this study two 2D lattice peptides, with the ground states of hierarchical and non-hierarchical design, were investigated comparatively using three methods: Metropolis MC with the local move set, Metropolis MC with unspecific rigid rotations and the CMC algorithm. We present evidence that the folding pathways and kinetics of hierarchically folding clustered sequence are not adequately described in conventional MC simulations, and the account for cluster dynamics provided by CMC allows to capture essential features of the folding process. Our data suggest that the methods, which enable specific cluster motions, such as CMC, should be used for a more realistic description of protein folding.

PACS: 87.15.Cc; 87.15.Aa

Keywords: Protein folding; Lattice models; Collective motions; Cluster dynamics; Clustering Monte Carlo algorithm

1. Introduction

The problem of protein folding is one of the most challenging and most practically important problems in modern biophysics. It is established that the proteins rapidly fold into unique 3D structure, which is determined by their amino-acid sequence. Native structure of the protein globule is believed to be the ground state of the chain with the minimal free energy. This unique structure is rapidly found by self-assembly among the astronomical number of possible chain conformations without prohibitively long exhaustive search. The extreme complexity of the protein folding process justifies the development of highly simplified models [1–6,11]. Lacking the details, these models

*Corresponding author. Tel./fax: +380-444108326.
E-mail address: dem@rigeb.gov.tr (A.P. Demchenko).

0301-4622/04/$ - see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/j.bpc.2003.10.001

should be able to observe the role of basic physical principles otherwise hidden by detailed 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. In lattice models, the protein folding is usually simulated by different Monte Carlo (MC) algorithms [3,5,10]. 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. MC 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 [1–4].

Despite this success the results of these simulations cannot be easily accepted as representing the real mechanisms of protein folding. The lattice models allow only rough course-grained description of the protein conformational space, which keeps the model very simple and makes the folding problem computationally tractable. Despite the highly simplified coarse-grained treatment of the conformational space the lattice 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 local and collective motions of the real chain.

Physically correct description of the collective motions is especially important if the possibility of the so-called hierarchical folding is considered. The ideas that proteins fold hierarchically, by sequential formation and association of clusters of residues with increasing their size and complexity, are for a long time in the minds of many researchers [12–24]. The hierarchical acquisition of structure can explain the observation of equilibrium and kinetic intermediates [14–19,25–28,32–34], frequently observed very fast kinetics of the folding process [29] and provides a clear solution of Levinthal paradox [30]. When some group of residues forms a cluster stabilized by non-covalent interactions, then there appear the new degrees of freedom. They are the rotations and translations of the cluster, which results in dramatic reduction of conformational space available for individual residues forming the cluster [34]. In view that such mechanism seems reasonable and supported by numerous experimental observations, there are many attempts for its modeling and simulation with different objectives and on different level of complexity [20–22,35].

However, existing methods of MC 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 [7–9]. The 'classical' MC studies were performed using the local move set (LMS) [1]. 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 [1,37,38]. This move set is definitely far remote from physical reality because it does not allow performing possible collective motions. The latter can be very important, especially at the final steps of folding when the large correctly folded 'blocks' are already formed. Another widely used move set is the LMS with the addition of unspecific rigid rotations. We will call it MS2, following Chan and Dill [38]. MS2 allows collective 'diffusional' motions and therefore is considered to be much more realistic.

Both abovementioned move sets and their numerous modifications [8,9] have one serious drawback. For given bead the move set does not depend on the current chain topology and remains the same even if this bead forms the contacts with its neighbors. In other words, the move sets LMS and MS2 are unspecific (structure-independent). In real proteins, however, the conformational mobility of the amino-acid residues is always controlled by the local and non-local interactions and the current chain topology. This should definitely restrict the motion along certain degrees of freedom and diminish dramatically the available conformational space of the chain.

Collective motions, which are present in MS2, are also structure-independent. Any chain 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 collective motions should consider the motions which are specific to current topological state of the chain and which depend on the formed contacts.

Thus, the new step of 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 their dependence on the current chain topology and energetics. In order to achieve this goal, we have to inevitably 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 sequential 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 considered structures are in local equilibrium described by Boltzmann energy distribution, and the transitions between sequential conformations are micro-reversible. Meantime, in physical reality the collective motions and the motions of the stable clusters of residues in particular, are 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° at once and no intermediate positions are allowed. Another fundamental difficulty arises when in order to make the model closer to physical reality the attempts are made to include into consideration the effects of viscous media surrounding the folding chain. Since the energy of any collective motion in this situation dissipates due to the viscous friction, this makes collective motions essentially irreversible. Metropolis sampling cannot be applied in the presence of viscous friction even if the collective motions are subdivided into small steps. In general, irreversible effects like viscous friction cannot be adequately introduced into the models based on Metropolis or other equilibrium sampling.

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. Preliminary studies [7] demonstrated that CMC algorithm allows finding correctly the unique energy minimum (the folded state) for a short 12-member peptide. In this communication we report on 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.

2. Clustering Monte Carlo algorithm

The basic idea behind this algorithm is to consider the motion of not only a single residue but of a cluster of residues of variable size ranging from single residue to the whole protein sequence. Then the folding can be described as the process of growth, association and destruction 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 any part of the cluster cannot 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 providing 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 Section 1, the coarse-grained lattice does not allow the rotations of the large clusters to be micro-reversible. We 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 that conforms to Boltzmann distribution

$$P(E_f) = \frac{\exp(-E_f/k_B T)}{T}. \quad (1)$$

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). It is necessary to emphasize that the cluster 'breakage' does not mean that *all* the contacts inside the cluster are destroyed. For each elementary rotation the pivot residue is randomly chosen (see detailed algorithm flow chart in Fig. 2). It divides the cluster into two parts, one of which is rotated. Thus, only the bonds, which connect two parts of the cluster, break (as it is visualized in Fig. 1).

In physical reality the cluster rotation should be controlled by solvent viscosity. CMC algorithm is unique in its ability to simulate the viscosity effects. Introduction of viscosity concept into lattice models should primarily serve to the purpose of capturing the basic fact that the larger clusters are less likely to be rotated. To simulate this effect, we can introduce some constant energy E_{r0}, which is needed to rotate a minimal-size single residue cluster. Larger clusters in order to rotate will need the energy $E_r = nE_{r0}$, where n is a size of the cluster. If $E_f < E_r$, then the cluster (or its part) cannot rotate at all. Thus, E_{r0} can be considered as a measure of the solvent viscosity that allows or does not allow a particular motion. The case $E_{r0} = 0$ corresponds to the absence of the solvent.

Apparently, there might be a possibility to introduce viscosity concept into MS2 algorithm in a similar manner. However, it is not possible in reality because of the Metropolis sampling used in MS2. Introduction of the viscous friction implies introduction of the energy dissipation and thus results in the loss of micro-reversibility of elementary moves. Such system with dissipation is no longer described by the equilibrium Boltzmann distribution, which makes Metropolis sampling invalid.

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

Detailed flow-chart of the CMC is shown in Fig. 2 (initialization and equilibration stages are not included). On each iteration of the algorithm the random residue is chosen as a pivot point. One of the bonds adjacent to this residue is chosen randomly to indicate the part of the chain, which will be rotated. Rotation direction is also chosen randomly. The energy of the thermal fluctuation is chosen from Boltzmann distribution Eq. (1). Then, the cluster, which contains the pivot point, is found

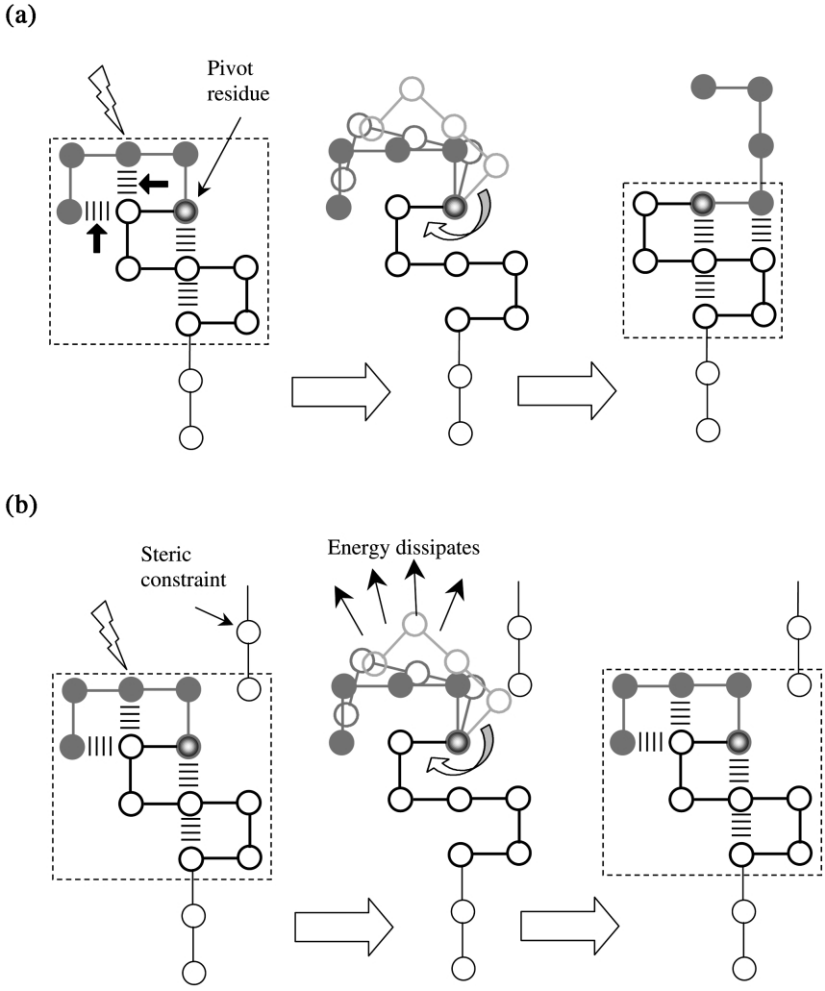

Fig. 1. Cluster disruption in CMC algorithm. The dashed boxes indicate initial and final clusters. The filled circles indicate rotating part of the cluster. The striped bars show bonds inside the cluster. Two bonds, indicated by solid arrows, break during the rotation. In the case (a) there are no steric constraints for rotation, and the large thermal fluctuations rotate the part of the cluster and disrupt it. In the case (b) the rotation is restricted by 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.

according to the procedure described below. Cluster dissociation and rotation energies are calculated and compared with the magnitude of the thermal fluctuation. If destruction or rotation is possible, then the corresponding move is performed. Resulting conformation is checked for self-overlaps and chain breaks. If these checks fail then the old conformation is restored, otherwise new conformation is accepted. The procedure continues until the native state is reached or the given number of folding events is accumulated.

Identification of the cluster, which contain the pivot residue, is performed according to the following scheme:

1. All 2×2 squares, which have all four vertices occupied and contain the target residue, are found on the lattice.

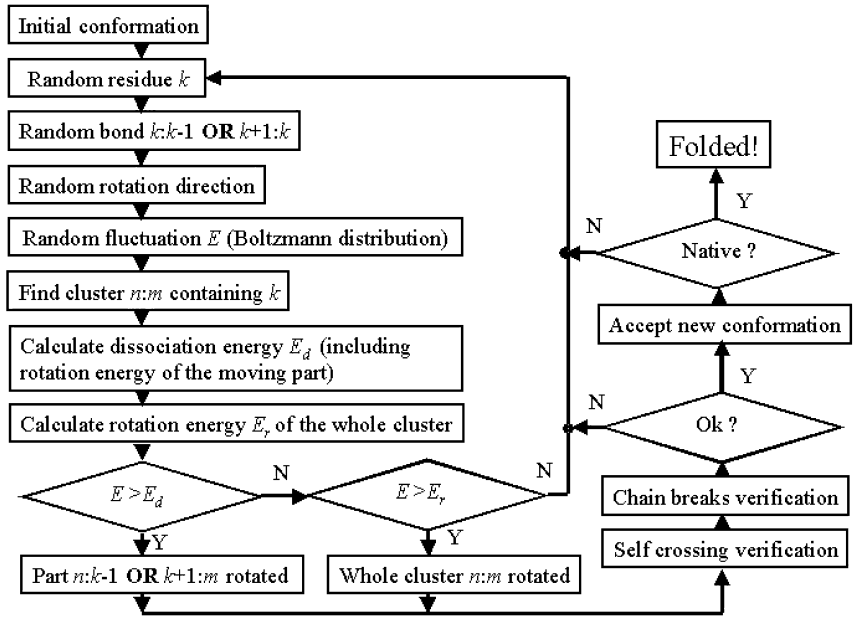

Fig. 2. Flow chart of the CMC algorithm. See text for details.

2. The residues included into the found squares are added to the cluster.
3. The same procedure is recursively performed for each of the newly added residues until the whole cluster is marked.

The residues, which precede the first and follow the last residues in the cluster, are included into the cluster if they follow the cluster during rotation in order to avoid sterical overlaps. This depends on the chosen direction of rotation.

Thus, the CMC algorithm has clear physical interpretation and does not require the questionable assumptions 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.

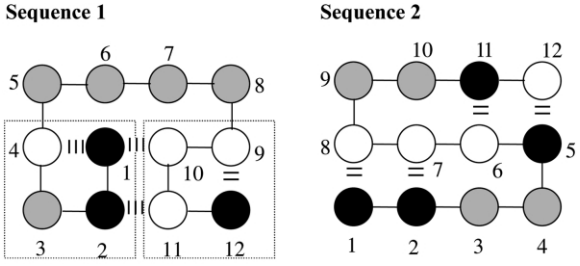

Fig. 3. The structures of the studied 12-member peptides, sequence 1 and sequence 2, in their native conformations. Positively charged residues are marked black, negatively charged are white, neutral residues are gray. Dashed boxes in the case of sequence 1 indicate the clusters of the first hierarchical level. Non-local bonds with non-zero energy are shown by the striped bars.

3. The model

In order to illustrate the applicability and the advantages of CMC algorithm we 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 non-degenerate compact ground states.

Sequence 1 forms a hierarchically organized compact structure (Fig. 3), 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. Two bonds 1–10 and 2–11 combine the clusters. Residues 5–8 form a connecting loop. In contrast, sequence 2 has no hierarchical features (Fig. 3). It folds into a 'β sheet' with three strands stabilized by the bonds 1–8, 2–7, 6–11 and 5–12.

Compared to the most common two-letter HP models [1,38] we used three-letter protein alphabet. The HP models consider only attractive and neutral interactions. In contrast, our three-letter model operates with attractive, repulsive and neutral interactions. They are represented by negatively charged (type 1), positively charged (type 2) and neutral (type 3) residues. This model allows to design a hierarchically organized ground state configuration using very short peptides.

It is necessary to note that our model lacks unspecific hydrophobic interactions, thus it cannot describe the formation of the hydrophobic core of the real protein and larger lattice peptides. The concept of hydrophobic core is hardly applicable to peptides of that small size.

We 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. Each bond between charged residues is assumed to have the energy of 50 abstract dimensionless units in the case of charges of the same sign and -50 in case of charges of opposite sign. All other bonds are of zero energy. Both folded structures (sequences 1 and 2) attain unique native states (Fig. 3) with the energy -200 units each.

4. Simulation methods

Three series of simulations were performed comparatively using CMC algorithm (with various E_{r0}, which correspond to various solvent viscosities) and Metropolis MC algorithms with LMS and MS2 move sets [31,38]. The chains were equilibrated at the high temperature $T=1000$ for 1000 iterations to generate a random unfolded conformation (the term 'temperature' is used in the meaning of k_BT energy measured in abstract dimensionless units). 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 was 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. 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 averaged over 1000 runs for various temperatures and used them for kinetics studies.

Folding thermodynamics was monitored at various temperatures in equilibrium simulations for long times, which last until 3000 folding–unfolding events occur in the system. The probability of the folded state P_{fold} was calculated as $P_{fold}=t_{fold}/t_{total}$, where t_{fold} is the number of iterations spent in the folded state and t_{total} is the total simulation length.

5. Results and discussion

5.1. Energy landscapes of the 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. 4). 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

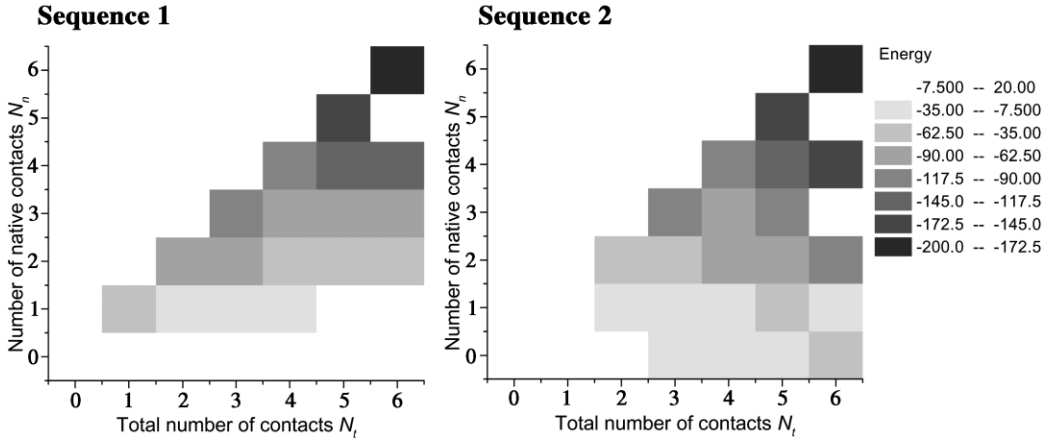

Fig. 4. Energy landscape in coordinates N_n vs. N_t for sequence 1 and sequence 2 obtained by complete enumeration of conformations. The folded state corresponds to the point (6,6).

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.

5.2. Integrated residence time maps

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. 5. The maps obtained by CMC simulations ($E_{r0}=0$) show existence of several types of folding intermediates. For small temperatures ($T=10$) the folding pathway is dominated by intermediates with five contacts, three of which are native. They may be classified as *semi-compact* intermediates. Most probable chain conformation in this region is composed of two-folded clusters of the first level combined by the non-native bonds. A significant part of the folding time is spent in that intermediates or in nearby regions of the map.

For higher temperatures ($T=15$–30), 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. 4 shows that the non-native energy minimum at point (5,5) is occupied rarely. This demonstrates the fact 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. This means that both methods actually sample the same configurations during the folding. The features observed in 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). 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

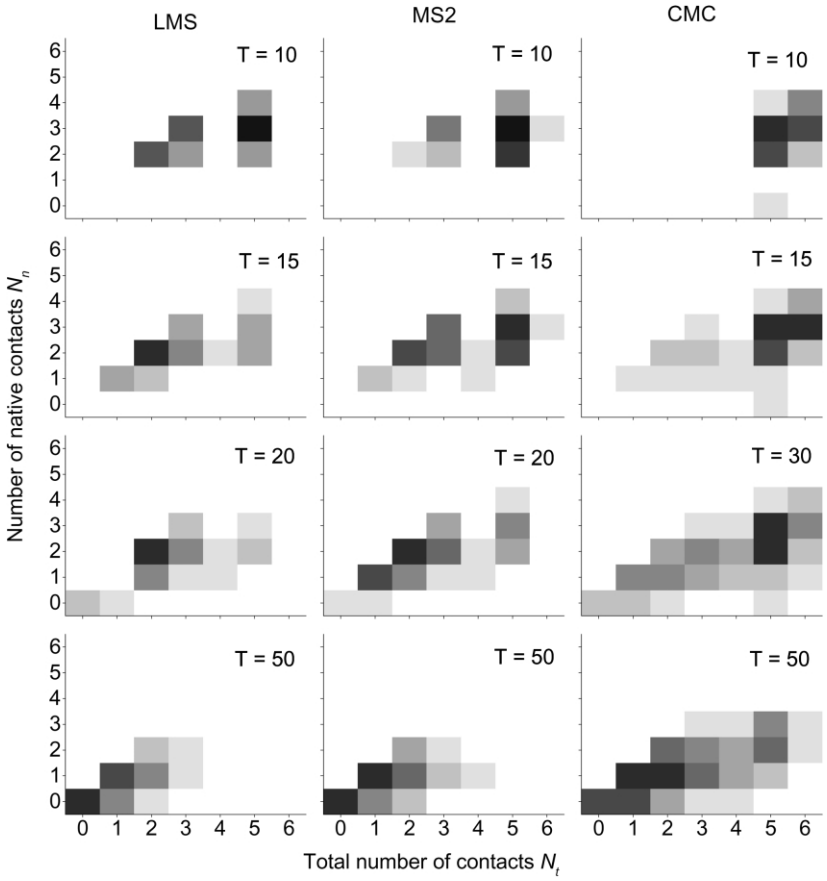

Fig. 5. Integrated residence time maps for sequence 1 for various temperatures obtained in LMS, MS2 and CMC simulations. Color intensity represents the amount of time spent by the sequence in the given point, white corresponds to zero, black to the maximum.

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 becomes dominant for much lower temperatures (40–50) than that in the case of CMC.

Integrated residence time maps for sequence 2 are shown in Fig. 6. The maps obtained by CMC ($E_{r0}=0$) are fundamentally different from sequence 1 in two ways. First of all, for small temperatures the 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 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 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

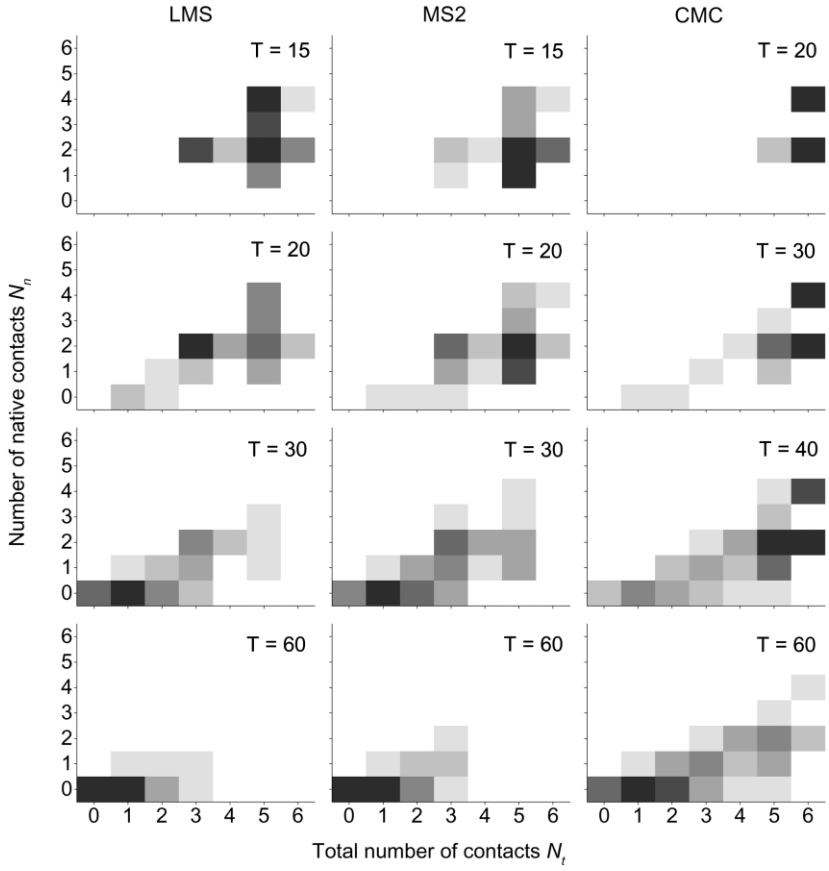

Fig. 6. Integrated residence time maps for sequence 2 for various temperatures obtained in LMS, MS2 and CMC simulations. Color intensity represents the amount of time spent by the sequence in the given point, white corresponds to zero, black to the maximum.

chain is very unlikely to reach it. Instead the chain will form some structures with misfolded cluster arrangement and with 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 the 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. 5). There are few compact intermediates observed in LMS/MS2 simulations even for lowest computationally possible temperatures. Highly populated states, which correspond to the points (5,4), (5,2) 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

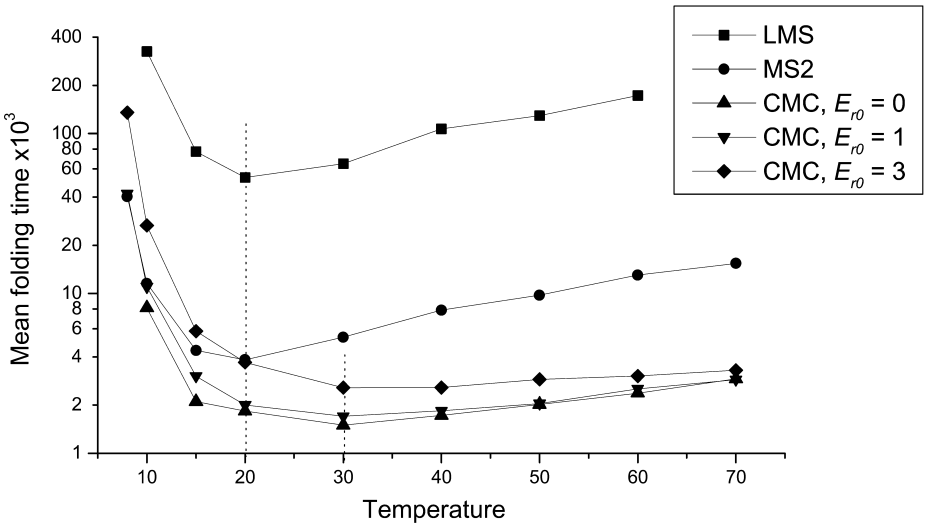

Fig. 7. Mean folding times for sequence 1 obtained in CMC, LMS and MS2 simulations. In the case of CMC, the results for various E_{r0} values, which correspond to various solvent viscosities, are shown. Optimal folding temperatures are indicated by the vertical dashed lines.

finds these minima and allows observing the chain trapping.

5.3. 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 $A\exp(-t/\tau_U)$. The mean folding time τ_U was obtained from histogram fitting.

It is important to emphasize that CMC algorithm has additional parameter E_{r0}, which may describe the control by solvent viscosity. There are no correspondent parameters in LMS and MS2 algorithms, and therefore they cannot account for solvent viscosity effect. Meantime the introduction of control by the solvent viscosity is an important step, which can make the lattice model to become much closer to the physical reality. That is why in CMC simulations we studied the influence of E_{r0} on the folding rate. Since this possibility does not exist in LMS/MS2 algorithms, a correct direct comparison of CMC with these algorithms can be made only if $E_{r0}=0$, which correspond to the absence of solvent.

5.3.1. Sequence 1

Temperature dependencies of the mean folding times obtained for sequence 1 are shown in Fig. 7. The shapes of the curves obtained by all simulation techniques are similar. An optimal temperature is approximately 20 in the case of LMS and MS2 and approximately 30 in the case of CMC. However, the mean folding times are quite different. LMS shows the slowest folding, which is more than ten times slower than that 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'. In order to escape the trap the chain has to perform several unfavorable local moves, which can retard the folding dramatically. In MS2 and CMC the escape from such states can be made by a single collective move and therefore the system does not get trapped. Both latter methods show

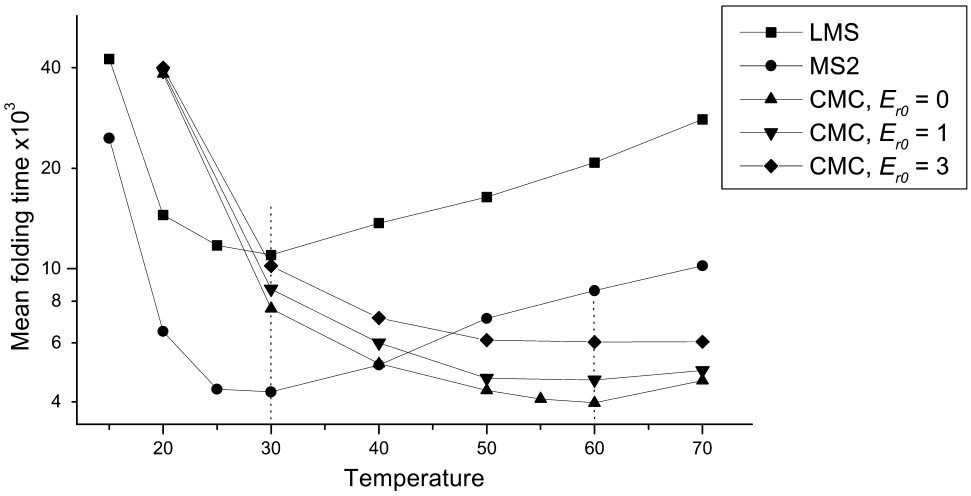

Fig. 8. Mean folding times for sequence 2 obtained in CMC, LMS and MS2 simulations. In the case of CMC, the results for various E_{r0} values, which correspond to various solvent viscosities, are shown. Optimal folding temperatures are indicated by the vertical dashed lines.

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 much shorter 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.

5.3.2. Sequence 2

The temperature dependencies of the mean folding times obtained for sequence 2 are shown in Fig. 8. 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 approximately 30. Meantime, the temperature dependence obtained by CMC simulations is very different. An optimal temperature is twice higher ($\sim$60). Minimal folding time for CMC is only slightly shorter than that given by MS2.

5.4. 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. 6 and 7, the mean folding time of sequence 1 in LMS is ten 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 could have 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 conditions, in which these clusters maintain integrity during the folding process. In contrast, the optimal folding temperature for non-hierarchical sequence 2 is much higher and exceeds the strength of the single bond. 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 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 20–30 for both sequences. LMS/MS2 simulations disregard the 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 quite close for both sequences.

Since hierarchical sequence 1 forms two most proximal contacts in the sequence (contacts 1–4 and 9–12, see Fig. 3) 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. 6 and 7). In contrast, MS2 simulations show almost identical folding times for 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 having 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 beta-sheet structure in CMC occurs on a much longer time scale than the formation of compact clusters, which is physically more realistic. This explains the fact that in CMC the hierarchical sequence 1 folds much faster than the non-hierarchical sequence 2.

5.5. Influence of solvent viscosity

Real proteins are folded in the solvent, and solvent viscosity may strongly influence the folding mechanisms and rates [39]. The possibility of accounting for solvent viscosity is a unique feature of CMC algorithm. Solvent viscosity in CMC is modeled by E_{r0} parameter, which is the energy needed to rotate a single-residue cluster. With the increase of E_{r0} a higher energy is needed to rotate a cluster. As a result, the probability of successful move decreases and the mean folding time becomes longer. This behavior is clearly seen in Figs. 7 and 8. The larger is the solvent viscosity the higher lies the corresponding curve. Optimal folding time seems not to be affected by the viscosity changes.

Fig. 9 shows dependencies of the mean folding times on solvent viscosity for both studied sequences at optimal temperature ($T=30$ in the case of sequence 1 and $T=60$ in the case of sequence 2). Both sequences show monotonous increase of the mean folding time with the increase of viscosity, and this time in the case of the Sequence 1 increases much sharper for large viscosities. This peculiarity may be explained qualitatively in the following manner. The last stage of the folding of hierarchical sequence 1 is the aggregation of the preformed clusters of the first level.

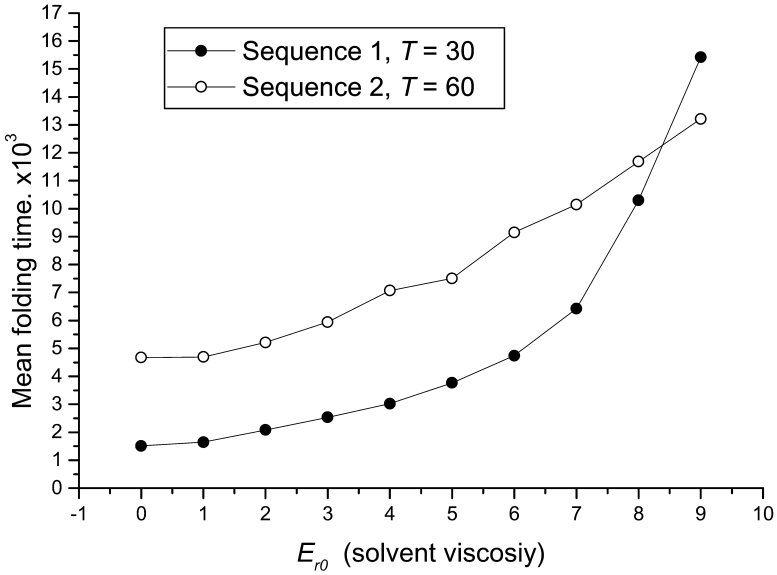

Fig. 9. Mean folding times for sequence 1 and sequence 2 obtained in CMC simulations for various E_{r0} values, which correspond to various solvent viscosities. Simulations were performed at optimal folding temperatures (indicated).

In the case of high-solvent viscosity this process is effectively retarded by the high-energy cost of cluster rotation, which leads to sharp increase of the mean folding time. In contrast, the folding of non-hierarchical sequence 2 does not depend on the rearrangements of the preformed clusters and thus shows less pronounced increase of the mean folding time.

5.6. Folding thermodynamics in different simulation techniques

As it was stated above, CMC algorithm does not require the assumption of micro-reversibility. The detailed balance is not achieved on individual steps of CMC simulation. Instead, the folding protein is considered as located in the equilibrium heat bath, which is a source of thermal fluctuations. These fluctuations trigger the cluster motions. That is why it is necessary to investigate the thermodynamic properties of CMC to ensure that it does not contradict the physical reality.

At low temperatures the population of the folding molecules is dominated by the folded chains and partially folded intermediates. The set of intermediates formed for each given temperature and their relative probabilities depend strongly on the folding algorithm implemented. In contrast, for very high temperatures the chain is mostly unfolded and any simulation technique samples essentially the same set of the unfolded conformations.

LMS and MS2 algorithms are based on Metropolis sampling criterion, which ensures Boltzmann energy distribution for high temperatures. CMC should give the same distribution in order to be in line with thermodynamic requirements. Comparison of the high temperature energy distributions obtained in MS2 and CMC simulations is shown in Fig. 10. CMC produces essentially the same distributions as MS2 for both sequences. This means that it samples the conformations in the unfolded ensemble with the Boltzmann probabilities despite the fact that individual folding steps are irreversible.

It is also interesting to compare the energy distributions obtained in CMC and MS2 simulations at the temperatures that correspond to the fastest folding. Each technique samples an ‘optimal’ set of conformations at this optimal folding temperature, which allows providing the fastest

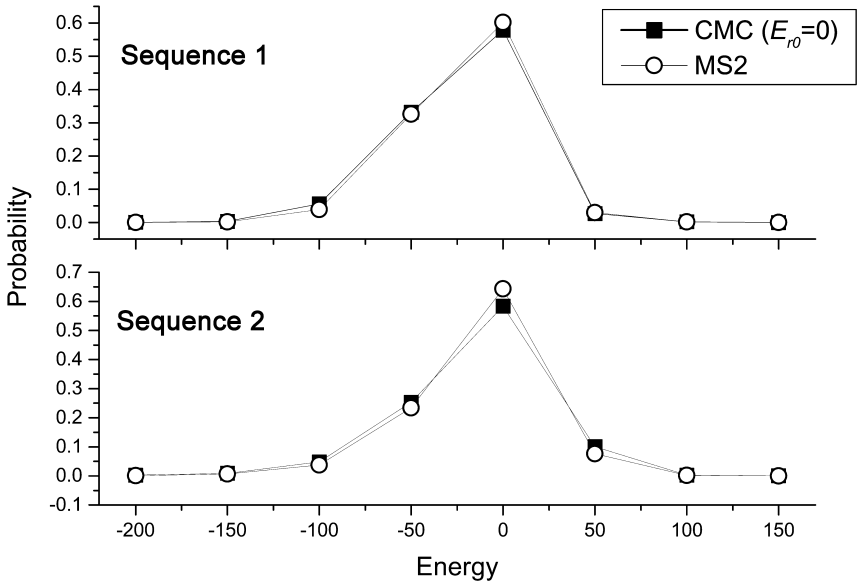

Fig. 10. Energy distribution functions for sequences 1 and sequence 2 at high temperature $T=500$ obtained in MS2 and CMC simulations.

folding of the sequence (as it was shown above, the optimal temperatures are quite different for different simulation techniques). Results shown in Fig. 11 indicate that the energy distributions obtained at optimal temperatures for CMC and MS2 are quite different for both studied sequences. This reflects the differences in the sampling criteria used in CMC and MS2. Moves in CMC are structure-specific and irreversible, while MS2 (and LMS) uses unspecific micro-reversible move sets. With the increase of solvent viscosity in CMC the energy of the contacts inside the clusters starts to play smaller role in controlling collective motions. In the limit of large viscosities ($E_{r0} \gg E_d$) the energy of thermal fluctuation, which triggers cluster destruction in CMC, is 'spent' almost entirely to overcome the viscous friction, and only a small amount of this energy is needed to break the cluster bonds. Stable and unstable clusters have almost identical probabilities of rotation and destruction in this case and the move set becomes almost unspecific. So, it is expectable that the energy distribution for CMC will resemble that for MS2 in the case of large viscosities. This effect is clearly seen in Fig. 11.

Although high-temperature thermodynamics of MS2 and CMC is the same, the characteristics of the folding transition are quite different. Fig. 13 shows the probabilities of the folded state in the long equilibrium runs as a function of temperature. The folding transition occurs at the temperature corresponding to 50% probability of the folded state. It is clearly seen that for both sequences the folding transition is much sharper in the case of MS2 simulations. This feature may also be deduced from the mean folding time curves (Figs. 7 and 8) that are much steeper in the case of MS2 in comparison with CMC. The temperature of the folding transition is higher in the case of CMC, which correlates with the higher optimal folding temperature in the case of CMC (Figs. 7 and 8).

The most probable explanation of these facts is the following. According to Metropolis criterion, in the case of MS2 the probability of the elementary chain move depends on the energies of initial and final conformations. The moves, which decrease the overall energy of the chain, are always accepted even if they break some energetically favorable bonds. In contrast, in CMC no moves are accepted with the unit probability. If particular

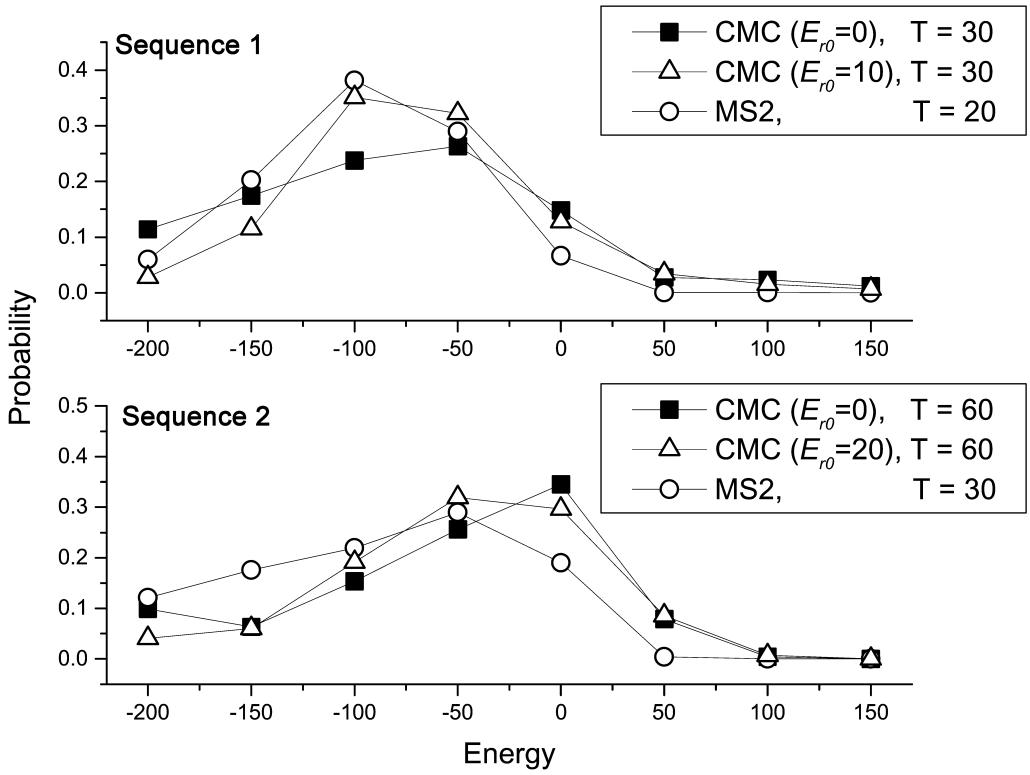

Fig. 11. Energy distribution functions for sequences 1 and sequence 2 at optimal folding temperatures (indicated) obtained in MS2 and CMC simulations with various solvent viscosities expressed as E_{r0}.

cluster rotation destroys existing bonds, which have the energy E_d, then according to Eq. (1) this move is performed with the probability $P \approx (\exp(-E_d/k_B T)/T)$. In this process it is not important, whether this energy is compensated by the formation of the new bonds after rotation or not. Therefore, the moves, which lead to destruction of the folded state, are much less probable in the case of CMC. This explains the origin of higher folding transition temperatures.

Probabilities of the folded state obtained for the fixed temperature but various viscosities are almost the same (data not shown). This correlates with the fact, that the optimal folding temperature is independent of the solvent viscosity.

5.7. Transmission coefficients in different simulation techniques

Transmission coefficients (TC) of different conformations are very important characteristics of the folding of particular sequence. TC is usually defined as the probability of given conformation to reach the native state until it unfolds. High TC characterize conformations, which are quite close to the native state on the folding pathway (particularly native state itself has TC = 1), while low TC usually indicates unfolded or misfolded states. The set of partially folded conformations, which has TC = 0.5 is called transition state ensemble (TSE). Conformations of TSE have equal probabilities to fold toward the native state and unfold to the random coil. It is believed that conformation of TSE possess a critical set of bonds, which have to be formed in order to allow rapid 'downhill' collapse to the native state. Conformations, which follow the TSE on the folding pathway, are called post-critical conformations (PCC). They can be viewed as direct route toward the native state. It is quite expectable that the transition coefficients depend strongly on the simulation techniques used

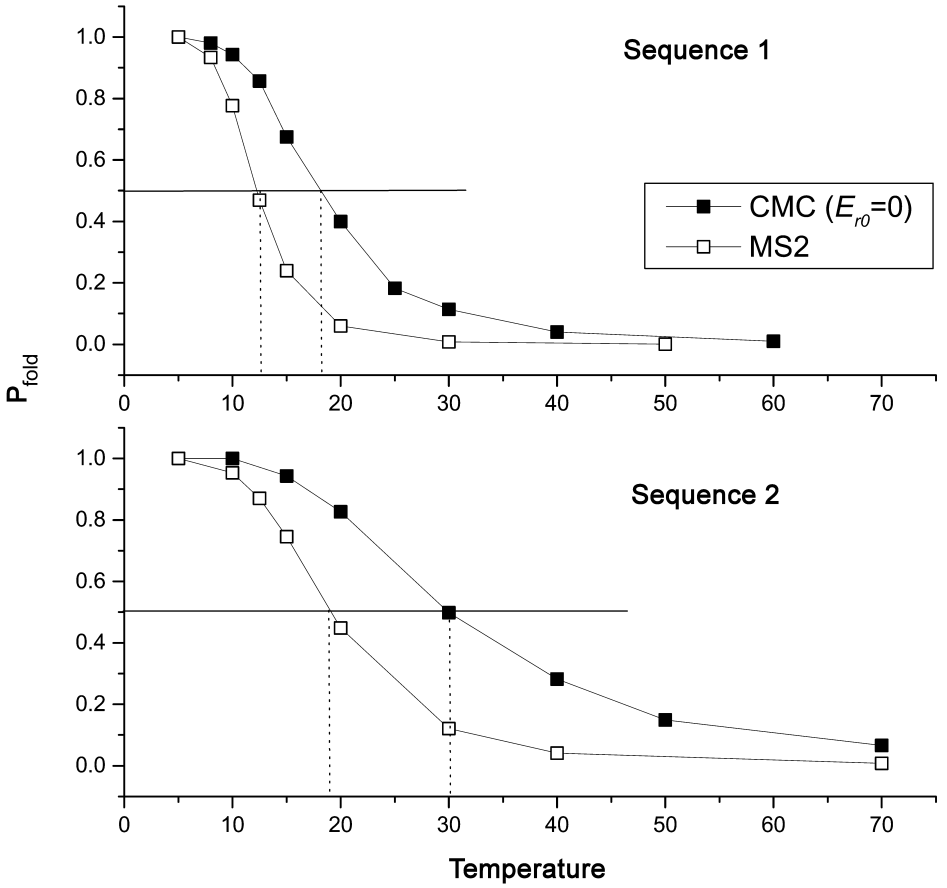

Fig. 12. Probability of the folded state for sequences 1 and 2. The dashed lines indicate folding transition temperatures.

and may help to visualize the differences between them. It is necessary to note that the number of conformations for 12-mer is so small that there may be no conformations with TC = 0.5. Thus, it may be impossible to determine the TSE reliably.

We performed exhaustive enumeration of all 7600 unique 12-mer conformations unrelated by rotations, reflections and reversal labeling symmetries. Conformations with no contacts were excluded from consideration because they cannot belong to TSE. For each of the remaining 5696 conformations calculation of the TC was performed. TC was estimated from 1000 independent CMC or MS2 runs. Each run starts from the selected conformation, which has N contacts, and proceeds until the number of contacts drops below N (unfolding) or until the native state is reached (folding). TC was calculated as $TC = N_{folding}/N_{total}$, where $N_{folding}$ is the number of runs, which reach the native state before unfolding and N_{total} is total number of runs. Calculations were performed at the temperature, which correspond to the folding transition temperature for given sequence and simulation technique (Fig. 12) (for sequence 1: 17.5 in the case of CMC and 12.5 in the case of MS2; for sequence 2: 30.0 in the case of CMC and 18.0 in the case of MS2). PCC were defined as conformations, which have $TC > 0.5$.

TC calculations for sequence 1 are shown in Fig. 13. In the case of CMC simulations, there are only two PCC conformations, C39 and C109. Conformation C39 is essentially the native state with one broken bond. C109 has two large clusters and can be transformed to the native state by the

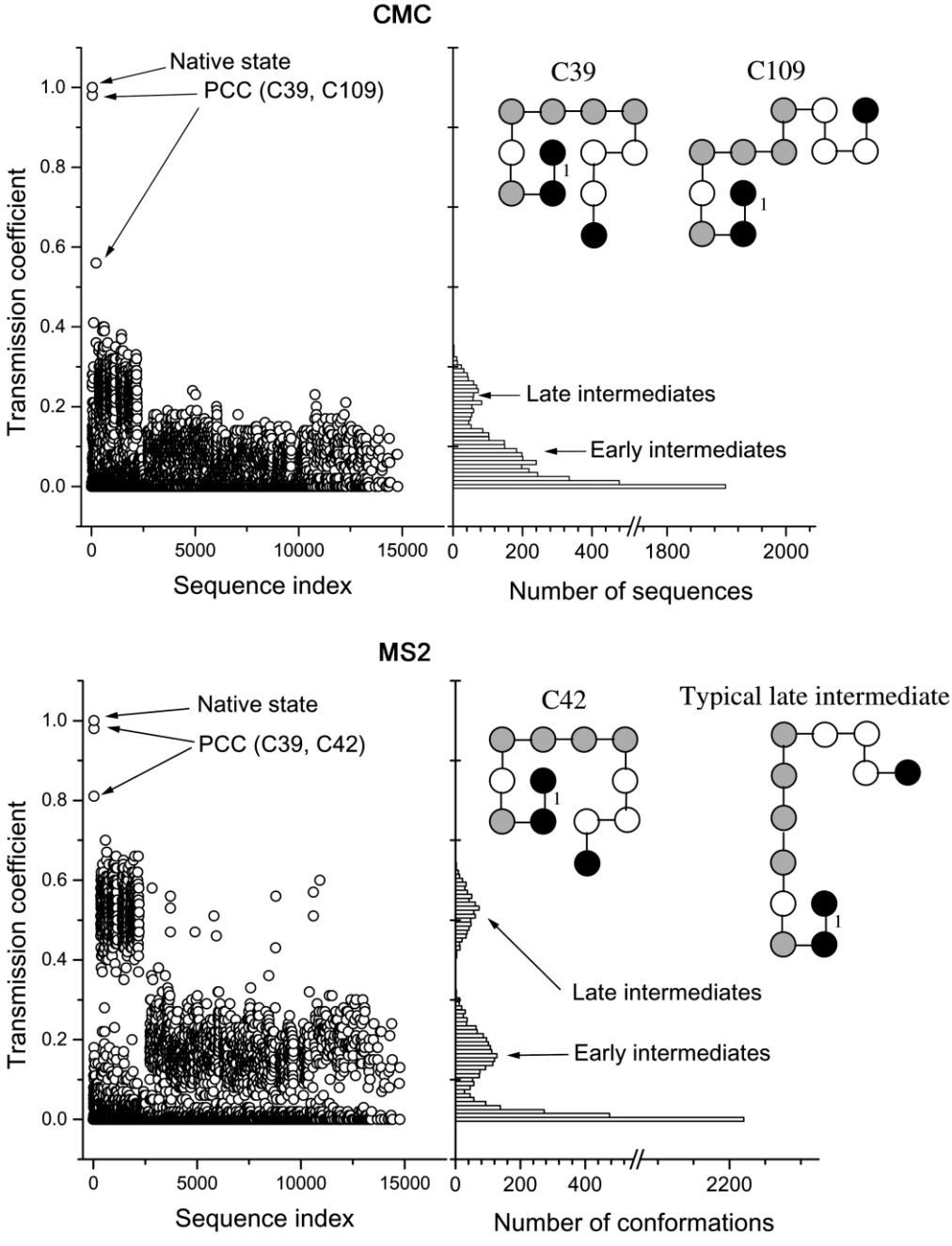

Fig. 13. Transition probabilities of 5696 conformations possessing at least one contact for sequence 1. Indexes are assigned to conformations during the enumeration procedure and should be viewed only as unique identification labels. Histograms attached to plots show the number of conformations, which fall into small intervals of TC values. Selected conformations from PCC set and intermediates are shown next to the histograms.

single rotational move. Two maxima are clearly seen on the histogram. We call corresponding conformations the early and late intermediates. Majority of these conformations have one or two formed clusters of the first level separated by the completely unfolded loop.

Results obtained in MS2 simulations of the sequence 1 are qualitatively similar to those obtained in CMC (Fig. 13). There are only two well-separated PCC and two sets of intermediates. First PCC is also C39, however, second PCC is not C109 but C42, which has completely different structure. In contrast with C39, it does not possess two correctly folded clusters. There is an extended chain segment with non-native topology, which can be brought to its place in the native structure by one corner and one end move. This fact is remarkable. It shows that the latest folding steps in MS2 are local moves, which change the chain topology, while in CMC the correct topology is reached on the previous steps, and the last step is the aggregation of the preformed blocks. Late intermediates in MS2 simulations typically contain long extended chain segments, which may be folded into PCC or to the native state by single 'stick move'. These unphysical moves are forbidden in CMC, and the corresponding structures in CMC simulations belong not to the late intermediates but to completely misfolded states.

TC calculations for sequence 2 are shown in Fig. 14. PCC set is identical for CMC and MS2 simulations and consists of two very similar conformations—C3880 and C3881. These conformations have disrupted bonds at the chain terminals. This reflects the fact that sequence 2 has non-hierarchical design, and the final folding steps are local in both simulation techniques. However, the analysis of intermediates reveals significant differences in the early steps of folding. In the case of CMC there are very few intermediates, which cannot be reliably classified into an early or late group. Typically the intermediates possess the central strand of the 'beta sheet' and the correct topology of both turns, but have 'open' contacts at the ends. In contrast, in the case of MS2 the number of intermediates is much larger and the early and late ones may be reliably separated. In the latter case, the typical late intermediates have two correctly formed strands and one turn of the beta sheet. The second turn is not formed and the third strand is misfolded or extends as a 'stick'. Early intermediates also often have long 'sticks'.

It can be seen that in the case of CMC the formation of beta sheet starts from the central strand. The chain forms the central strand and two turns and attains the native topology *before* energetically favorable contacts at the ends are formed. Once the correct topology is achieved, the free ends of the chain bind to the formed 'core'. In contrast, in the case of MS2 the formation of the beta sheet starts from the arrangement of two unbound straight 'sticks'. These sticks become connected after the series of 'stick' moves forming one end of the beta sheet. Remaining strand is then affixed to this structure forming PCC or directly the native state.

5.8. Energetics of the final folding steps

In order to better understand the folding mechanism and kinetics of the studied sequences, we calculated the averaged energies of the final 100 iterations for sequences 1 and 2 (Fig. 15). Remarkable feature of the obtained relations is the presence of barriers on the average folding pathways. 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 a 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. 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. In this case the curve is essentially 'flat' up to the critical point where there appears 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 barrier position. This behavior is a direct consequence of the existence of semi-compact and compact misfolded intermediates, which are detected on the

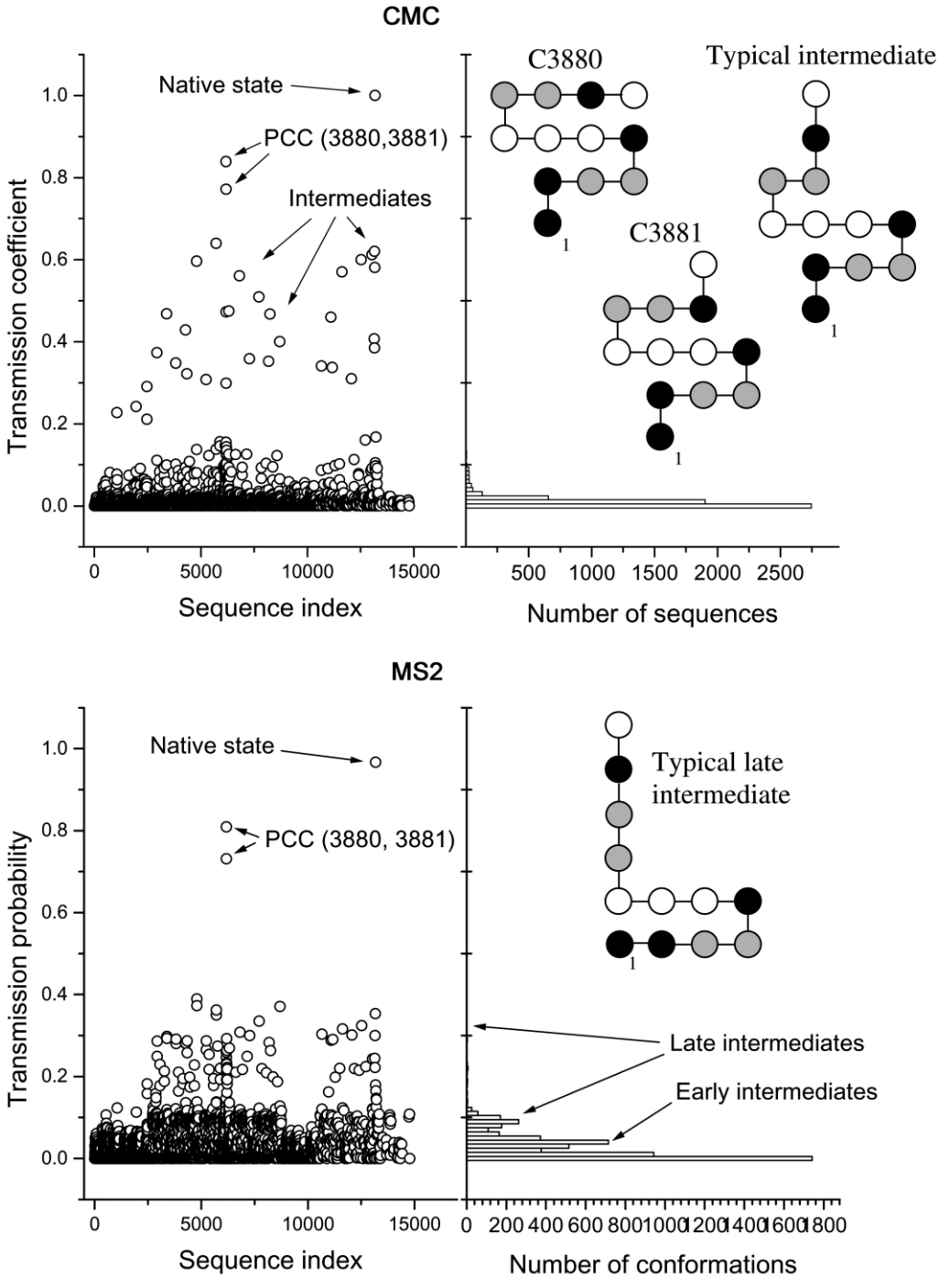

Fig. 14. Transition probabilities of 5696 conformations possessing at least one contact for sequence 2. Indexes are assigned to conformations during the enumeration procedure and should be viewed only as unique identification labels. Histograms attached to plots show the number of conformations, which fall into small intervals of TC values. Selected conformations from PCC set and intermediates are shown next to the histograms.

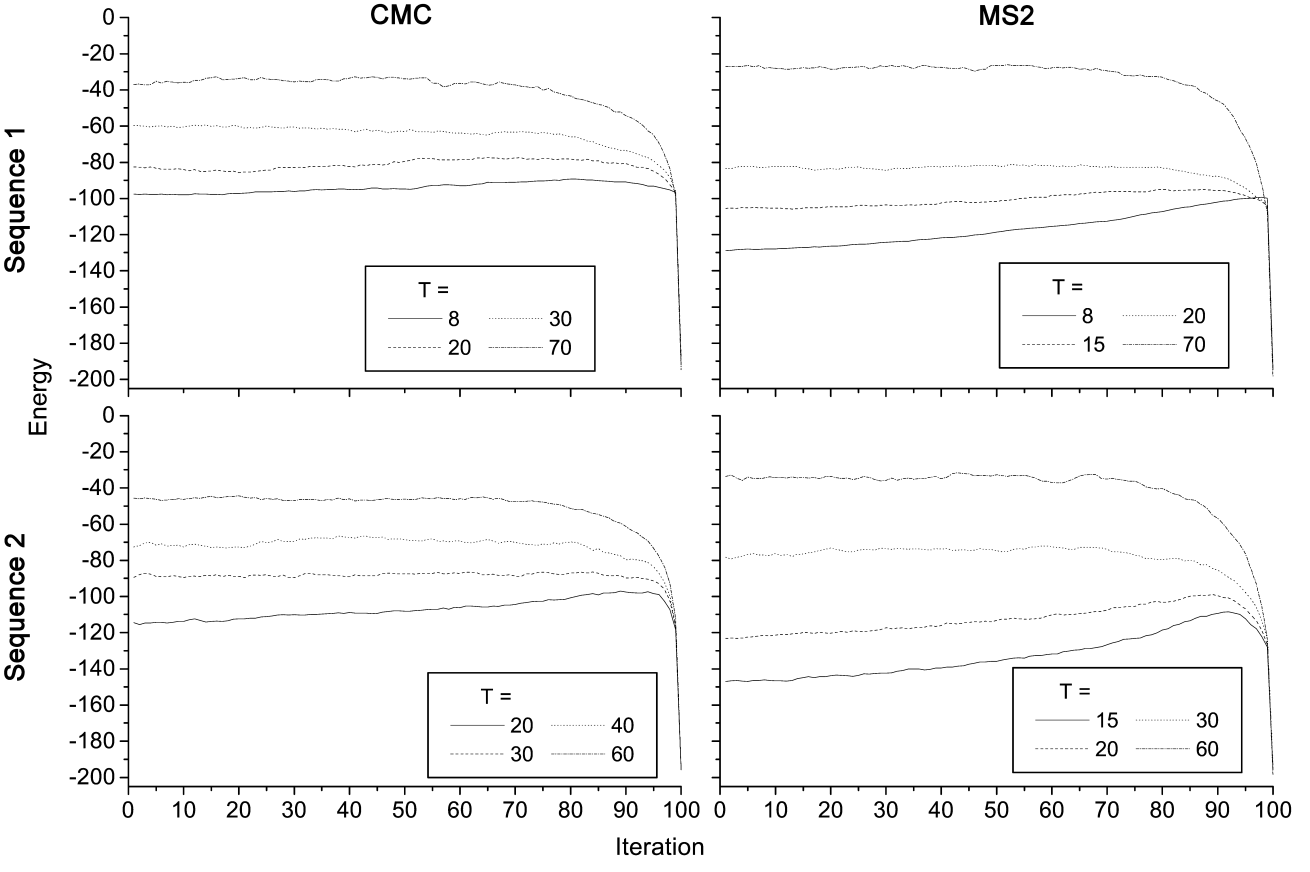

Fig. 15. Average energies for the final 100 steps of folding of sequence 1 and sequence 2 at different temperatures obtained in MS2 and CMC simulations. The folding proceeds from left to right. The last point corresponds to the native state.

integrated residence time maps (Figs. 4 and 5). 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.

6. Conclusions

The results of this work demonstrate the difference in folding behavior between non-hierarchical and hierarchical sequences, which are revealed by all three used simulation methods. In the case of 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. Since LMS disregards collective motions, it underestimates the folding rate. In contrast, MS2 could overestimate it due to unspecific nature of allowed collective motions and the necessity to involve the physically unjustified 'stick moves'. In contrast, all collective motions in CMC are structure-dependent (specific). Only the clusters can be rotated as a whole, while the open chain segments are not treated as the rigid 'sticks'. The probability of the cluster rotation depends on the cluster size and the solvent viscosity, which is much closer to the physical reality. Since CMC provides the most accurate description of the collective motions, it probably describes the most justified folding rates.

In the case of hierarchically organized sequence the LMC algorithm fails to produce realistic folding rates, which is probably a consequence of topological trapping that is characteristic for the sequence with buried ends. Meantime, MS2 and CMC, which allow collective motions, demonstrate very similar maximal folding rates. Comparison of the TC shows that the folding pathways of hierarchical sequence in CMC and MS2 are quite different. Folding in CMC is achieved after formation of large clusters with the correct topology, which then aggregate due to cluster rotation. In contrast, MS2 simulations reveal formation of intermediates with non-native topology, which are transformed to the native state by local moves at the latest steps of folding. In the case of non-hierarchical sequence, the CMC simulations also tend to form correct chain topology before the formation of all energetically favorable contacts. In contrast, MS2 in the latter case seems to form favorable contacts before the formation of native topology. It can be seen that CMC describes accurately diffusive motions of the stable clusters and the dynamics of the open segments. 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 certain regions of conformational space become unavailable for the sequence even at initial steps of folding. We show that CMC and LMS/MS2 algorithms sample 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 of real proteins can be roughly classified into fast and slow folders. This classification can be extended to model sequences in the lattice. Should it depend on 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 in CMC the hierarchical sequence in contrast to non-hierarchical one is a fast folder, while in MS2 no essential differences between these sequences are detected. 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 continu-

ous 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.

We believe that CMC algorithm has a powerful potential for the future development. This will allow extending the possibilities of lattice models that are presently viewed as very limited. Particularly, we show that CMC can simulate the effects of the solvent viscosity, which is not possible in the other algorithms due to presumption of microreversibility of elementary moves. Extension of CMC to 3D lattice is straightforward. In principle, CMC can fold the chains of any length. Our recent tests showed that CMC successfully folds 2D 28- and 36-mers. Although the search for clusters in CMC is rather time consuming, the physical time spent in CMC and MS2 simulations of the same sequence are almost the same, even for the longest chains tested. We observe no restrictions for implementing the concept of CMC to the off-lattice models. The main idea of the algorithm could remain the same, however, identification of the clusters should require quite different criteria. It is not excluded that the implementation of CMC to off-lattice models will substantially reduce the computational expenses due to the freezing of clusters' internal degrees of freedom.

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. We believe that the folding process cannot be adequately described neither in MC simulations with the LMS nor in simulations with unspecific collective motions (such as MS2). In these cases, the account 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.

References

[1] A. Sali, E. Shakhnovich, M. Karplus, How does a protein fold?, Nature 369 (1994) 248.

[2] V.S. Pande, D.S. Rokhsar, Folding pathway of a lattice model for proteins, Proc. Natl. Acad. Sci. USA 96 (1999) 1273.

[3] L. Mirny, E. Shakhnovich, Protein folding theory: from lattice to all-atom models, Annu. Rev. Biophys. Biomol. Struct. 30 (2001) 361–396.

[4] D.K. Klimov, D. Thirumalai, Lattice models for proteins reveal multiple folding nuclei for nucleation-collapse mechanism, J. Mol. Biol. 282 (1998) 471.

[5] C.M. Dobson, A. Sali, M. Karplus, Protein folding: a perspective from theory and experiment, Angew. Chem. Int. Ed. 37 (1998) 868.

[6] K.A. Dill, S. Bromberg, K. Yue, et al., Principles of protein folding—a perspective from simple exact models, Protein Sci. 4 (1995) 561.

[7] S.O. Yesylevskyy, A.P. Demchenko, Modeling the hierarchical protein folding using clustering Monte-Carlo algorithm, Protein Peptide Lett. 6 (2001) 437.

[8] N.L. Nunes, K. Chen, J.S. Hutchinson, A flexible lattice model to study protein folding, J. Phys. Chem. 100 (1996) 10443.

[9] J. Shin, W.S. Oh, J. Phys. Chem. 102 (1998) 6405.

[10] S.S. Sung, Monte Carlo simulations of beta-hairpin folding at constant temperature, Biophys. J. 76 (1999) 164.

[11] A.R. Dinner, M. Karplus, The thermodynamics and kinetics of protein folding: a lattice model analysis of multiple pathways with intermediates, J. Phys. Chem. 103 (1999) 7976.

[12] K.P. Murphy, V. Bhakuni, D. Xie, E. Freire, Molecular basis of co-operativity in protein folding. III. Structural identification of cooperative folding units and folding intermediates, J. Mol. Biol. 227 (1992) 293.

[13] M. Karplus, D.L. Weaver, Protein folding dynamics: the diffusion–collision model and experimental data, Protein Sci. 3 (1994) 650.

[14] R.L. Baldwin, G.D. Rose, Is protein folding hierarchic?, Trends Biochem. Sci. 24 (1999) 26.

[15] M. Jamin, M. Antalik, S.N. Loh, D.W. Bolen, R.L. Baldwin, The unfolding enthalpy of the pH 4 molten globule of apomyoglobin measured by isothermal titration calorimetry, Protein Sci. 9 (2000) 1340.

[16] S. Akiyama, S. Takahashi, K. Ishimori, I. Morishima, Stepwise formation of alpha-helices during cytochrome *c* folding, Nat. Struct. Biol. 7 (2000) 514.

[17] T.M. Raschke, J. Kho, S. Marqusee, Confirmation of the hierarchical folding of RNase H: a protein engineering study, Nat. Struct. Biol. 6 (1999) 825.

[18] Q.X. Hua, S.H. Nakagawa, W. Jia, et al., Hierarchical protein folding: asymmetric unfolding of an insulin analogue lacking the A7–B7 inter-chain disulfide bridge, Biochemistry 40 (2001) 12299.
[19] C.J. Tsai, R. Nussinov, Transient, highly populated, building blocks folding model, Cell Biochem. Biophys. 34 (2001) 209.
[20] C.J. Tsai, L.P. Polverino, A. Fontana, R. Nussinov, Comparison of protein fragments identified by limited proteolysis and by computational cutting of proteins, Protein Sci. 11 (2002) 1753–1770.
[21] S.A. Islam, M. Karplus, D.L. Weaver, Application of the diffusion–collision model to the folding of three-helix bundle proteins, J. Mol. Biol. 318 (2002) 199.
[22] D. De Jong, R. Riley, D.O. Alonso, V. Daggett, Surfing on protein folding energy landscapes, J. Mol. Biol. 319 (2002) 229.
[23] E. Paci, M. Vendruscolo, M. Karplus, Native and non-native interactions along protein folding and unfolding pathways, Proteins 47 (2002) 379.
[24] D. Hamada, Y. Kuroda, T. Tanaka, Y. Goto, High helical propensity of the peptide fragments derived from beta-lactoglobulin, a predominantly beta-sheet protein, J. Mol. Biol. 254 (1995) 737.
[25] D. Hamada, S. Segawa, Y. Goto, Non-native alpha-helical intermediate in the refolding of beta-lactoglobulin, a predominantly beta-sheet protein, Nat. Struct. Biol. 3 (1996) 868.
[26] K. Kuwata, R. Shastry, H. Cheng, et al., Structural and kinetic characterization of early folding events in beta-lactoglobulin, Nat. Struct. Biol. 8 (2001) 151.
[27] A.P. Capaldi, C. Kleanthous, S.E. Radford, Im7 folding mechanism: misfolding on a path to the native state, Nat. Struct. Biol. 9 (2002) 209.
[28] C. Wagner, C. Kiefhaber, Intermediates can accelerate protein folding, Proc. Natl. Acad. Sci. USA 96 (1999) 6716.
[29] S.W. Englander, N.R. Kallenbach, A signature of the TR transition in human hemoglobin, Quart. Rev. Biophys. 16 (1984) 521.
[30] S.W. Englander, L. Mayne, Protein folding studied using hydrogen-exchange labeling and two-dimensional NMR, Annu. Rev. Biophys. Biomol. Struct. 21 (1992) 243.
[31] M. Sadqi, S. Casares, M.A. Abril, O. Lopez-Mayorga, F. Conejero-Lara, E. Freire, The native state conformational ensemble of the SH3 domain from alpha-spectrin, Biochemistry 38 (1999) 8899.
[32] C.L. Brooks III, M. Karplus, B.M. Pettitt, Proteins: A Theoretical Perspective of Dynamics, Structure and Thermodynamics, Wiley, New York, 1988.
[33] A.P. Demchenko, Concepts and misconcepts in the analysis of simple kinetics of protein folding, Curr. Protein Peptide Sci. 2 (2001) 73.
[34] C. Micheletti, J.R. Banavar, A. Maritan, Conformations of proteins in equilibrium, Phys. Rev. Lett. 87 (2001) 088102.
[35] B. Erman, Analysis of multiple folding routes of proteins by a coarse-grained dynamics model, Biophys. J. 81 (2001) 3534.
[37] T.X. Hoang, M. Cieplak, Protein folding and models of dynamics on the lattice, J. Phys. Chem. 109 (1998) 9192.
[38] H.S. Chan, K.A. Dill, Transition states and folding dynamics of proteins and heteropolymers, J. Phys. Chem. 100 (1994) 9238.
[39] M. Jacob, F.X. Schmid, Protein folding as a diffusional process, Biochemistry 38 (1999) 13773.