Why Do Proteins Divide into Domains? Insights from Lattice Model Simulations

Aleksandra Rutkowska* and Andrzej Kolinski

Faculty of Chemistry, Warsaw University, Pasteura 1, 02-093 Warsaw, Poland

Received July 12, 2007; Revised Manuscript Received August 20, 2007

It is known that larger globular proteins are built from domains, relatively independent structural units. A domain size seems to be limited, and a single domain consists of from few tens to a couple of hundred amino acids. Based on Monte Carlo simulations of a reduced protein model restricted to the face centered simple cubic lattice, with a minimal set of short-range and long-range interactions, we have shown that some model sequences upon the folding transition spontaneously divide into separate domains. The observed domain sizes closely correspond to the sizes of real protein domains. Short chains with a proper sequence pattern of the hydrophobic and polar residues undergo a two-state folding transition to the structurally ordered globular state, while similar longer sequences follow a multistate transition. Homopolymeric (uniformly hydrophobic) chains and random heteropolymers undergo a continuous collapse transition into a single globule, and the globular state is much less ordered. Thus, the factors responsible for the multidomain structure of proteins are sufficiently long polypeptide chain and characteristic, protein-like, sequence patterns. These findings provide some hints for the analysis of real sequences aimed at prediction of the domain structure of large proteins.

Introduction

Protein folding transition is a very cooperative process. Moreover, for small proteins, folding transition has some features of a first-order phase transition and is usually described as the all-or-none phase transition.[1,2] All-atom classical molecular mechanics studies of the entire folding process remain impractical due to the enormous computational cost of such simulations. Thus, simplified lattice models are useful in studies of protein dynamics and thermodynamics.[1,3,4]

Most proteins, especially in eukaryotic organisms, are composed of two or more domains. Each domain in such protein constitutes a relatively independent folding unit. In some cases (i.e., thioredoxin, elastase),[5] separately folded domains can subsequently associate into the correct biological ternary structure. Other domains cannot exist without the rest of the molecule. Sometimes separate, even properly folded, domains do not associate to the correct structure.

In this work, we applied a simple minimal lattice model to studies of folding of different types of protein-like polymers of varying chain length.

Methods

The model employed here was described in detail in previous work,[6,7] where it has been proven that this model constitutes a minimal one for a protein-like cooperative, two-state folding transition. The polypeptide chain is restricted to the face centered cubic (fcc) lattice, where a single lattice point represents a single residue. The pseudobonds between subsequent residues belong to the set of 12 fcc vectors of type {[±1, ±1, 0]}, with proper permutations of coordinates. Each model residue is characterized by two properties: its hydrophobicity and secondary structure preference. Such definition of a polypeptide chain implies two types of potential. The short-range interactions mimic the conformational stiffness of the chain. In this work, the secondary propensities are limited to these simulating the β-type structures; there is an energetic preference for expanded β-type conformations. Three subsequent residues are an expanded state when two corresponding planar angles are equal to 120°, the first and the third consecutive vectors are identical, and the third residue has assigned a β-type structure preference. For each fragment fulfilling these three criteria, the energy of the chain is decreased by ϵ_b. The long-range interactions, between nonbonded

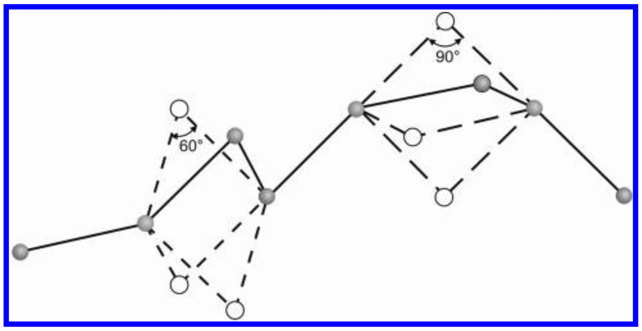

Figure 1. Examples of local micromodifications of the model chain conformations.

Table 1. Force Field Parameters in k_BT Units

parameter	hydrophobic homopolymer	alternating heteropolymer	random heteropolymer
ϵ_{HH}	−0.39	−0.79	−0.79
ϵ_{HP}	n.a.	0.65	0.65
ϵ_{PP}	n.a.	−0.81	−0.81
$\epsilon\beta$	−1.00	−1.00	−1.00

Table 2. Estimated Transition Temperatures

chain length	hydrophobic homopolymer	alternating heteropolymer	random heteropolymer
50	0.50	0.56	0.50
100	0.52	0.61	0.58
200	0.60	0.63	0.72
400	0.65	0.60	0.64

* Corresponding author. E-mail: arutka@chem.uw.edu.pl.

10.1021/bm7007718 CCC: $37.00 © 2007 American Chemical Society
Published on Web 10/12/2007

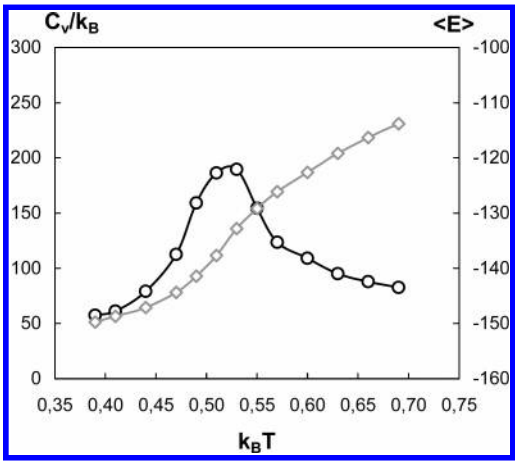

Figure 2. Heat capacity (black) and average energy (gray) as a function of temperature for homopolymer composed of $N = 100$ units.

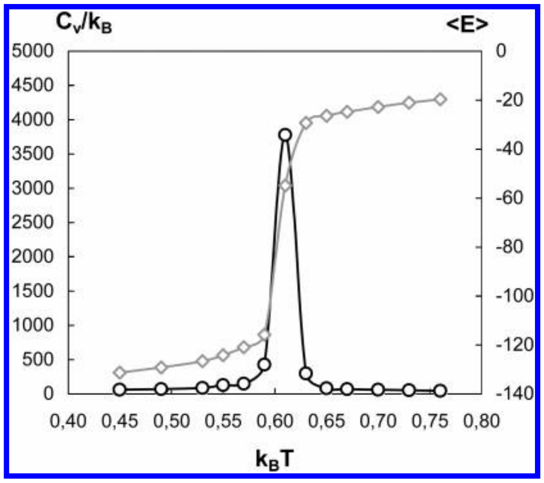

Figure 3. Heat capacity (black) and average energy (gray) as a function of temperature for alternating heteropolymer $N = 100$.

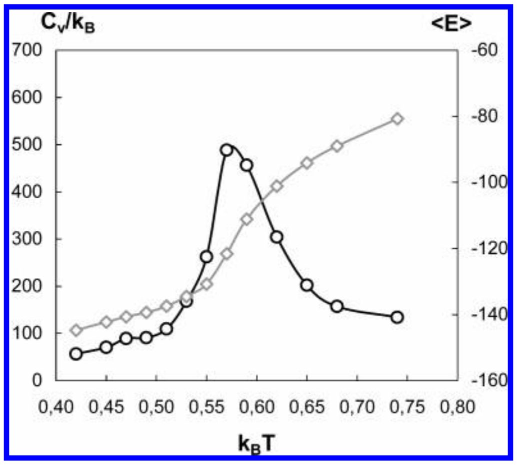

Figure 4. Heat capacity (black) and average energy (gray) as a function of temperature for random heteropolymer $N = 100$.

residues, have the form of a contact potential. There are repulsive interactions between hydrophobic and polar residues, attractive between hydrophobic residues, and orientation-dependent attractive interaction between polar residues. The interactions within a pair of polar residues are effective only when the contact occurs in a parallel fashion. The parallel orientation of the implicit side-chains is determined by the parallel (or antiparallel) orientation of the corresponding two-segment fragments of the model chain. Such design of the potential implicitly encodes preferential interactions between polar amino acids on the exposed surrounding solvent surface of the protein.[7−9]

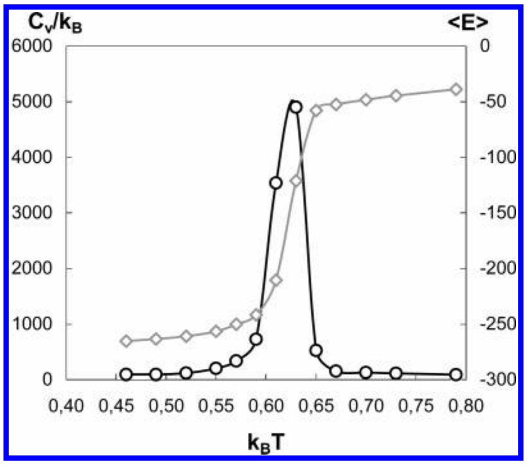

Figure 5. Heat capacity (black) and average energy (gray) as a function of temperature for alternating heteropolymer $N = 200$.

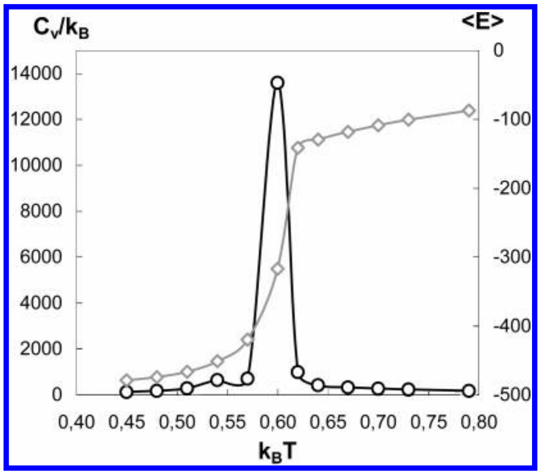

Figure 6. Heat capacity (black) and average energy (gray) as a function of temperature for alternating heteropolymer $N = 400$.

Conformational space has been explored using Replica Exchange Monte Carlo.[10−13] Conformational updating was performed using a random sequence of two-bond micromodifications of the chain geometry (Figure 1).

In Replica Exchange Monte Carlo (REMC), several copies of the system are simulated. Replicas are placed at different temperatures. The set of temperatures covers the range from the denaturing to the folding conditions. Each replica is subject to standard asymmetric Metropolis sampling.[14] Occasionally, replicas are exchanged with the probability:

$$P = \min[1, \exp(-\Delta)]$$

$$\Delta = (1/(k_B T_{i+1} - 1/k_B T_i)(E_i - E_{i+1})$$

REMC is a very efficient sampling scheme.[15,16] The replicas at high temperatures easily surmount the free-energy barriers, while the replicas at low temperatures sample with high accuracy local (and global) minima.

The collapse (folding) transition temperature is identified as a temperature with the highest value of the heat capacity.

Results and Discussion

Calculations were done for three types of protein-like polymers: hydrophobic homopolymer (HHHHHH...); alternating heteropolymer (HPHPHPHP....); and andrandom heteropolymer (HPPPHHP...), where the numbers of hydrophobic and polar residues are equal.

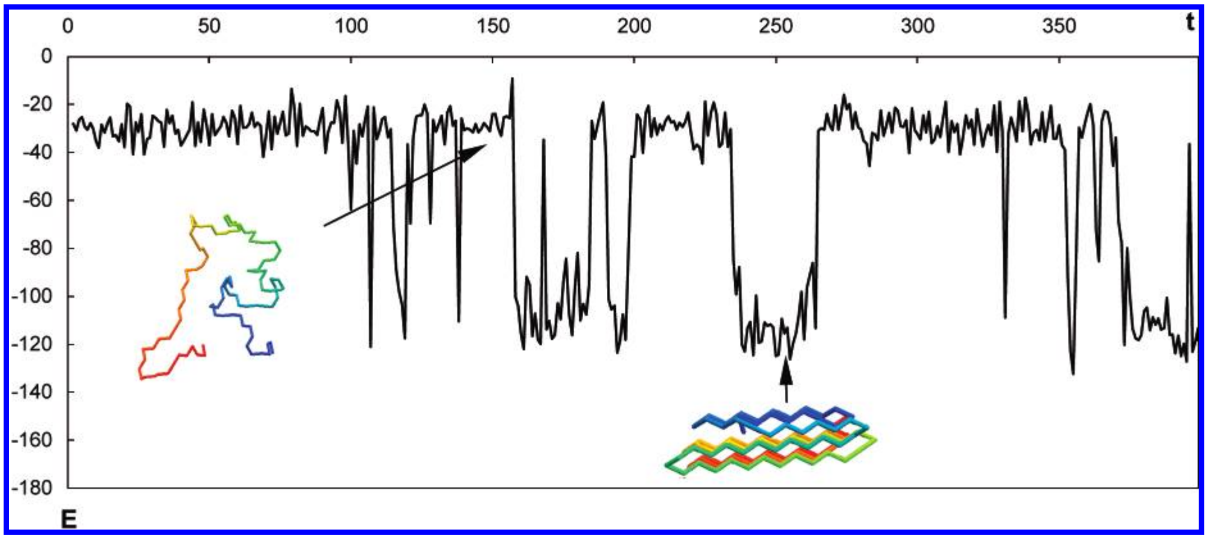

Figure 7. Energy as a function of time for alternating heteropolymer $N = 100$. The data collected at the transition temperature. The time is in arbitrary units, proportional to the number of cycles of the MC simulations.

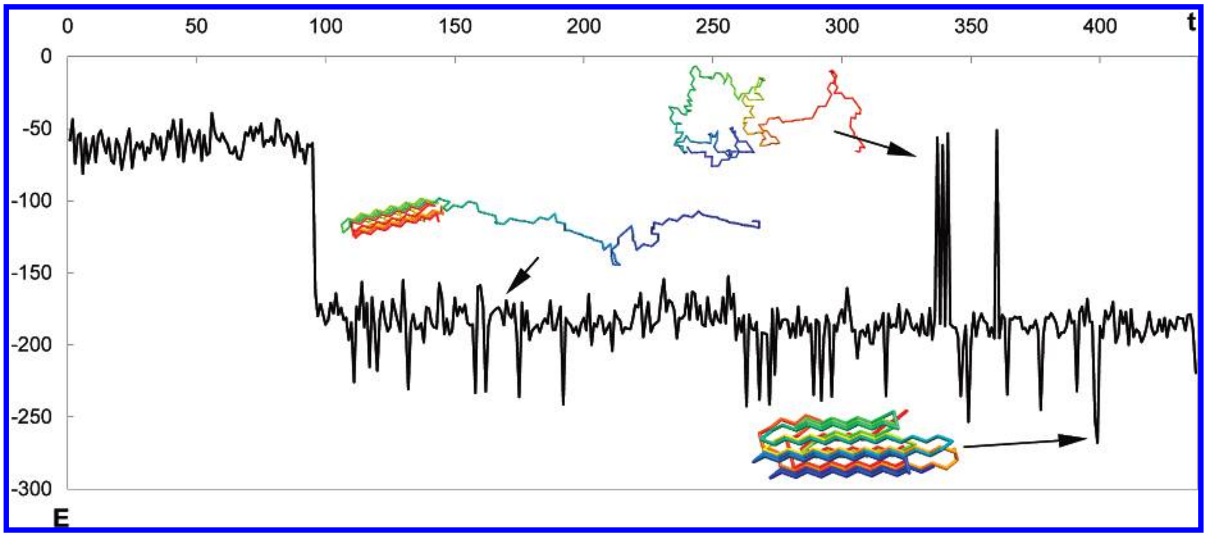

Figure 8. Energy as a function of time for alternating heteropolymer $N = 200$. The data collected at the transition temperature. The time is in arbitrary units, proportional to the number of cycles of the MC simulations.

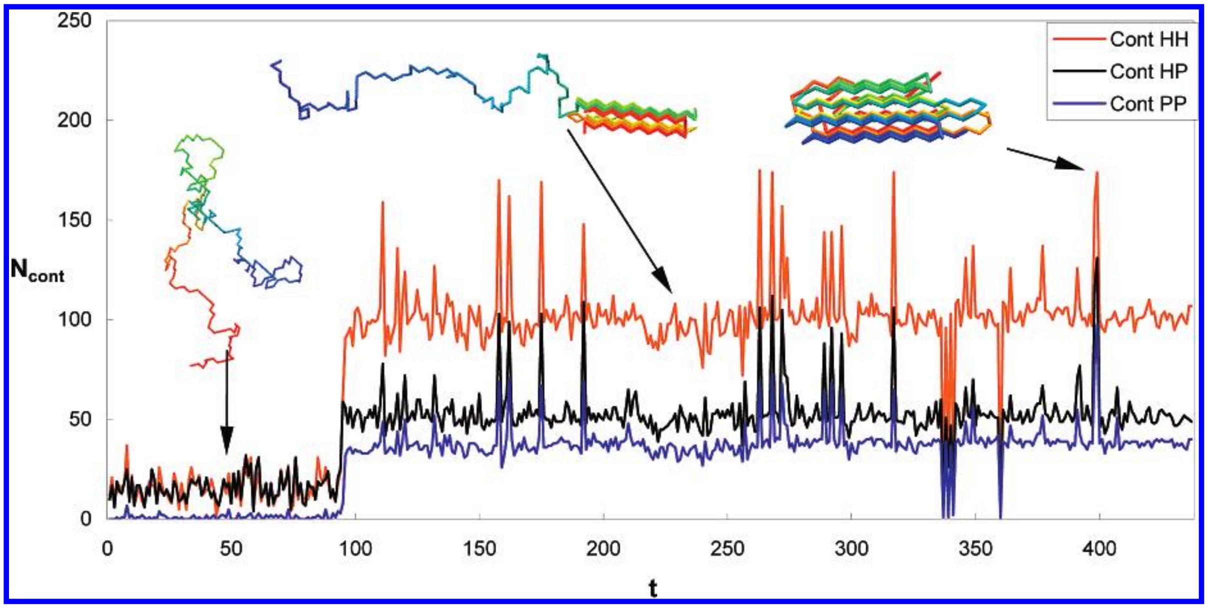

Figure 9. Number of contacts as a function of time for alternating heteropolymer at estimated transition temperature $N = 200$ (one-domain structure).

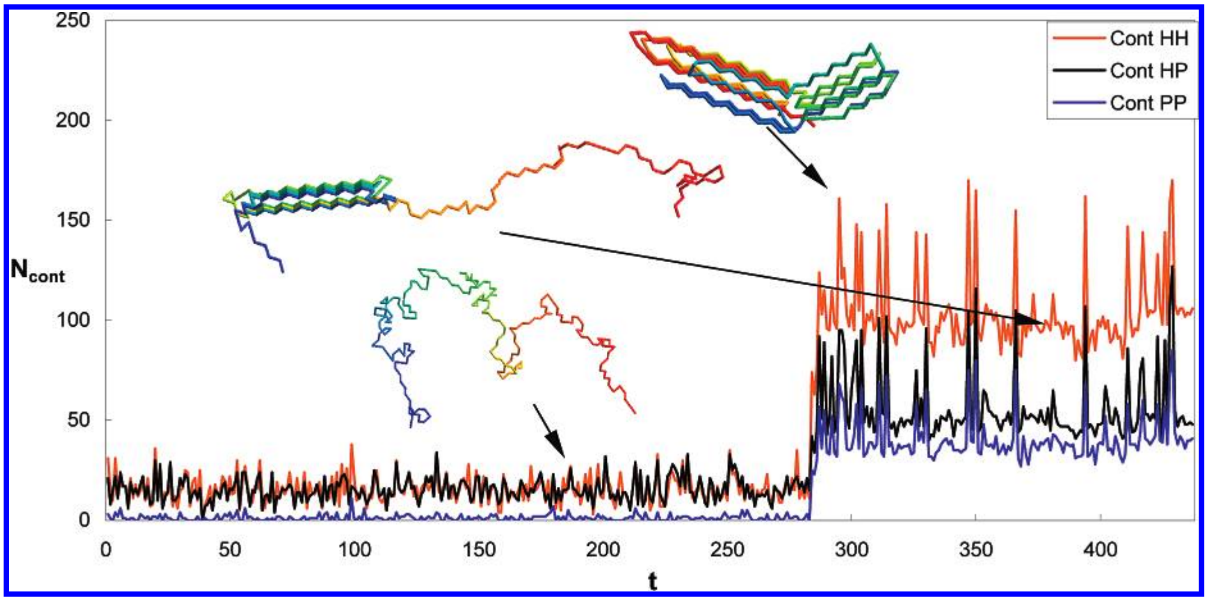

Figure 10. Number of contacts as a function of time for alternating heteropolymer at estimated transition temperature $N = 200$ (two-domain structure).

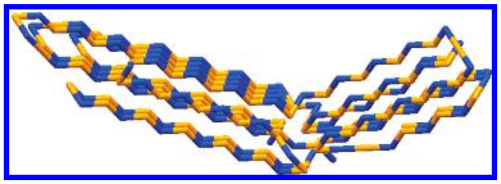

Figure 11. Two-domain structure of alternating heteropolymer ($N = 200$); blue, polar residues; orange, hydrophobic residues.

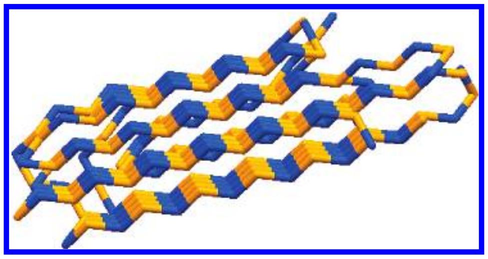

Figure 12. One-domain structure of alternating heteropolymer ($N = 200$); blue, polar residues; orange, hydrophobic residues.

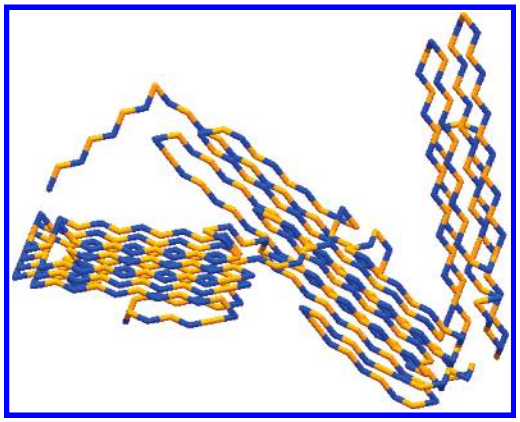

Figure 13. Low-temperature structure of alternating heteropolymer ($N = 400$); blue, polar residues; orange, hydrophobic residues.

All residues were assigned the β-type preference. The force field parameters were chosen from earlier works[7] and are presented in Table 1. When simulating, hydrophobic homopolymer parameter ϵ_{HH} was halved due to lack of repulsive interactions in such system.

To examine the relation between the length of polymer and differences in folding process, calculations were done for four lengths: 50, 100, 200, and 400 residues. For shorter chains (50 and 100), we used 13 replicas; for the longer ones, the number of copies was increased to 15 or 20. The set of temperatures covers temperatures below and above the transition temperature. Transition temperature was estimated for each system in preliminary simulations. The program performed about 10^9 iterations per single amino acid.

Replica Exchange Monte Carlo simulations provided data for folding process analysis. Transition temperature can be estimated from scaled heat capacity curve as a function of k_BT (where T is temperature and k_B is Boltzmann's constant):

$$\frac{C_v}{k_B} = \frac{\langle\langle E^2\rangle - \langle E\rangle^2\rangle}{(k_BT)^2}$$

In Table 2, the values of estimated transition temperatures are presented (in k_BT dimensionless units).

The average energies, average square energies, and the average square radius of gyration are measured from the second part of simulation data. Energy values are in dimensionless k_BT units.

Analysis of the heat capacity curve delivers information about transition cooperativity. With a highly cooperative process, the peak should be well defined, sharp, and narrow. Comparison of Figures 2−4 shows that folding of alternating heteropolymer is the most cooperative, while for another two types of chains cooperativity is rather weak. For well-defined cooperativity and two-state transition, all competitive interactions are needed. Furthermore, it was proved that the presence of both types of interactions (short- and long-range) is necessary for protein-like behavior.[7] An interplay between these interactions leads to cooperative effects. If one of these interactions is weaker or absent, the other one will be weaker too.[17] Thus, a significant change in the energy of the system should occur at the transition temperature. Such a significant change is observed only for heteropolymers. Cooperativity of folding depends on the length of a chain. The longer is the chain, the less cooperative is the folding process (Figures 5 and 6).

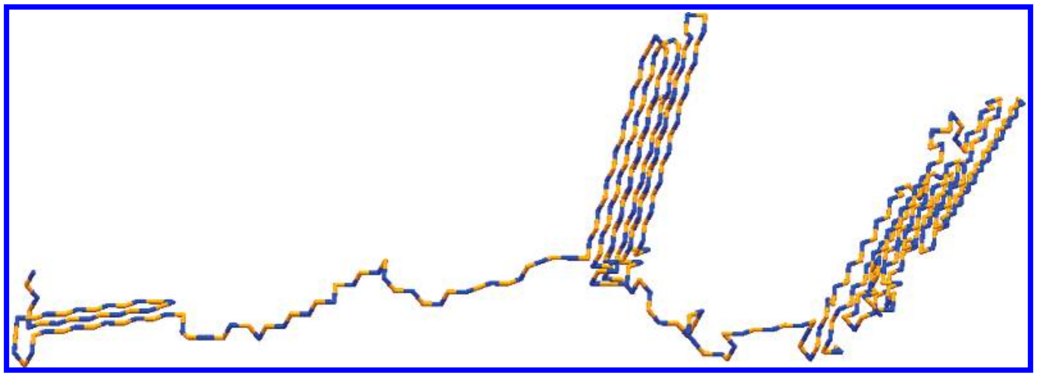

Figure 14. Two-domain structure of alternating heteropolymer ($N = 400$); blue, polar residues; orange, hydrophobic residues.

During all simulations, we stored pseudo-trajectories, containing successive snapshots of the conformations of the model replicas visiting the transition temperature. Trajectory analysis reveals that homopolymers do not part into domains. Folding of such chains is neither cooperative nor does it have features of the all-or-none phase transition. Mainly it is a result of the lack of repulsive interactions. There are only short-range interactions and uniform attractive interactions between hydrophobic residues. Energy decrease can be achieved by increase of the number of contacts or by fitting the local geometry. The average β-type fragment is composed of three beads, while the H−H contact needs only two beads at neighboring positions. Maximizing the number of contacts is more effective but requires a round shape of the collapsed structure.

Random heteropolymers do not divide into domains as well. In contrast to homopolymers, folded random heteropolymers have better pronounced elements of a secondary structure. For all explored chain lengths, the whole structure is rather compact with a hydrophobic core and a polar surface and no subunits can be seen. Multidomain structures of random heteropolymers do not exist because of their sequence. Natural proteins sequences constitute only a small fraction of all of the possible sequences. Well-defined secondary structure requires specific order of polar and hydrophobic amino acids. In previous works,[6,7] sequences were carefully designed so as to obtain specific structure motifs (six-stranded, antiparallel, β-barrel motif with a Greek-key topology and four-helix bundle). To obtain a unique, protein-like fold, these sequences were relatively short, and included flexible, putative loop regions. Cooperative folding transitions were observed. It remains to be tested if extending these well-defined sequences (for example, multiplying their length by a factor of 2) would lead to two-domain structures. The purpose of this work was different; for the cost of less specific sequences (no loop regions introduced), we intended to inspect the very fundamental phenomenon of splitting the longer sequences into separate domains. Indeed, for the 200 units alternating HP sequences, two domain structures were observed, while 400 units formed two or (more frequently) three domains. Thus, it could be inferred that the average size of domains for such model is in the range of 100−150 units, which corresponds very nicely with the average size of real domains. Simulation of longer chains (say consisting of 1000 units) would probably lead to a larger number of domains of a similar size. In principle, such computations are feasible, although very costly and less accurate. Besides, we think that the results for the chains studied in this work are sufficiently general. Randomly generated sequence does not guarantee existence of a unique globular structure,[18] even when the number of different amino acids is reduced from 20 to 2.

The presence of multidomain structures of alternating heteropolymers depends on the length of a chain. This effect is demonstrated in Figures 7 and 8. Folding transition of the chain composed of 100 residues exhibits features of the first-order phase transition. Essentially, only two types of states are observed at the transition temperature: the compact folded state and completely unfolded random coils.

Folding of longer alternating heteropolymers is qualitatively different. For instance, near the transition temperature, for the chain consisting of 200 residues, an additional intermediate state can be observed. In all of such intermediates, more than one-half of the chain is folded, while the rest of the chain remains unfolded. As shown in Figure 8, significant changes in the energy are associated with the transition from a random coil to an intermediate state and then from the intermediate state to the fully folded structure. The entire process of folding consists of two subsequent transitions that exhibit features of the all-or-none phase transitions. It can be confirmed by the flow charts of the number of contacts of various types between the model residues (Figure 9). During the simulation, three types of conformations are observed: nearly without contacts, with a significant (well-defined) fraction of them, and with a maximal number of contacts, characteristic for a compact globule.

Moreover, in some simulations the lowest energy structure is composed of two domains as shown in Figure 10. Although such structures do not occur in all simulations, the intermediate state is always present. Both types of fully folded structures of the 200 residues (Figures 11 and 12) chain are the lowest energy states at the transition temperature in each simulation. Two-domain structure has insignificantly higher energy. Both structures have a well-defined hydrophobic core (cores) and a polar surface. In contrast to the two-domain structure, inside the one-domain structure some polar residues are buried. Thus, it is possible that in fact it could be considered as not one but two strongly interacting domains.

The low-temperature structures of an alternating heteropolymer of length of 400 residues (Figure 13) are almost always composed of three domains. At the transition temperature, two-domain structures are also observed (Figure 14), sometimes with a small nuclei for the third domain. At higher temperatures, a random coil is observed.

Conclusions

In this Article, we studied the effect of chain length on the collapse transition of various types of HP-type polymers. Such a minimal model of proteins was simulated using an efficient REMC algorithm. It has been found that a partition into separate ordered domains occurs upon the collapse transition of sufficiently long chains with specific sequence patterns and short-range conformational stiffness. Similarly to some natural proteins, protein-like alternating heteropolymers form multidomain structures only when the chain is long enough. The size of such domains is between 100 and 200 residues. On the other hand, even relatively long homopolymeric or random heteropolymeric chains do not form visible multidomain globular structures. Specific sequence patterns of hydrophobic and polar residues are essential for the formation of ordered collapsed structures. Such structures have well-defined hydrophobic cores and polar surfaces. The partition into separate domains facilitates such separation of hydrophobic residues from the polar ones.

Consistently with the finding of previous[6,7] work, we found typical for proteins is that interplay between the short-range conformational stiffness and the long-range attractive interactions that simulate also orientational effects of the surface polar residues is essential for cooperative two-state (or multi-state for longer chains) collapse transition to the ordered globular structures. Thus, the minimalistic protein model explored here reproduces major features of the protein folding process, including the spontaneous formation of multidomain structures, demonstrated for the first time in the simulations described above.

Acknowledgment. The computational part of this work was done using the computer cluster at the Computing Center of Faculty of Chemistry, Warsaw University.

References and Notes

(1) Hao, M. H.; Scheraga, H. A. *J. Mol. Biol.* **1998**, *277*, 973−983.

(2) Scheraga, H. A.; Hao, M. H.; Kostrowicki, J. Theoretical studies of protein folding. *Methods in Protein Structure Analysis*; Plenum Press: New York, 1995.

(3) Dill, K. A.; Bromberg, S.; Yue, K.; Fiebig, K. M.; Yee, D. P.; Thomas, P. D.; Chan, H. S. *Protein Sci.* **1995**, *4*, 561−602.

(4) Thirumalai, D.; Klimov, D. K. Introducing protein folding using simple models. *Encyclopedia of Chemical Physics and Protein Chemistry*; IoP Publishing: Bristol, 2001.

(5) Yon, J. M. *Braz. J. Med. Biol. Res.* **2001**, *34*, 419−435.

(6) Pokarowski, P.; Droste, K.; Kolinski, A. *J. Chem. Phys.* **2005**, *122*, 214915.1−214915.6.

(7) Pokarowski, P.; Kolinski, A.; Skolnick, J. *Biophys. J.* **2003**, *84*, 1518−1526.

(8) Ilkowski, B.; Skolnick, J.; Kolinski, A. *Macromol. Theory Simul.* **2000**, *9*, 523−533.

(9) Kolinski, A.; Gront, D.; Pokarowski, P.; Skolnick, J. *Biopolymers* **2003**, *69*, 399−405.

(10) Swendsen, R. H.; Wang, J. S. *Phys. Rev. Lett.* **1986**, *57*, 2607−2609.

(11) Geyer, C. J.; Thompson, E. A. *J. Am. Stat. Assoc.* **1995**, *90*, 909−920.

(12) Hansmann, U. H. E. *Chem. Phys. Lett.* **1997**, *281*, 140−150.

(13) Hakushima, K.; Nemoto, K. *J. Phys. Soc. Jpn.* **1996**, *65*, 1604−1608.

(14) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. *J. Chem. Phys.* **1953**, *21*, 1087−1092.

(15) Gront, D.; Kolinski, A.; Skolnick, J. *J. Chem. Phys.* **2000**, *113*, 5065−5071.

(16) Hansmann, U. H. E.; Okamoto, Y. *Curr. Opin. Struct. Biol.* **1999**, *9*, 177−181.

(17) Creighton, T. E. *Proteins. Structures and Molecular Properties*; W. H. Freeman and Co.: New York, 1993.

(18) Hao, M. H.; Scheraga, H. A. *J. Phys. Chem.* **1994**, *98*, 9882−9893.

BM7007718