11-1-2004

# Relaxation to Native Conformation of a Bond-Fluctuating Protein Chain With Hydrophobic and Polar Nodes

Johan Bjursell
*George Mason University*

Ras B. Pandey
*University of Southern Mississippi*, ras.pandey@usm.edu

# Relaxation to native conformation of a bond-fluctuating protein chain with hydrophobic and polar nodes

Johan Bjursell

*George Mason University, MS 5C3, Fairfax, Virginia 22030-4444, USA*

R. B. Pandey

*Department of Physics and Astronomy, University of Southern Mississippi, Hattiesburg, Mississippi 39406-5046, USA*

The conformation and dynamics of a protein chain with hydrophobic and polar nodes are examined by the bond-fluctuation model using Monte Carlo simulations on a cubic lattice. The minimal (nearest neighbor) interaction leads to standard (self-avoiding walk) conformation, i.e., the scaling of the radius of gyration $R_g$ with the molecular weight $N$ $R_g \propto N^\gamma$ with $\gamma \simeq 3/5$. Specific interactions with longer range and higher strength are needed to approach the native globular conformations with $\gamma < 3/5$. Relaxation into the globular ground state shows a weak power-law decay, i.e., $R_g \propto t^{-\alpha}$, $\alpha \sim 0.06 – 0.12$.

The structural stability of protein [1–4] chains has been studied extensively in recent years [5–19], primarily by computational methods. How the protein chain relaxes to its native conformations is one of the main questions addressed by many researchers [5–19]. A native structure evolves into a stable configuration (in steady state or equilibrium) as the chain explores its conformational phase space and is expected to be globular in appropriate solvent conditions. A protein is a large polymer consisting of 20 amino acids in a specific sequence. These amino acid groups are similar except for their side chains that distinguish their characteristics. They are roughly divided into three categories: hydrophobic (*H*), polar (*P*), and charged (*C*) groups. Hydrophobic and polar groups are considered to be the main constituents in most coarse-grained models [20] and as the main constituents to describe the general characteristics of the protein.

In a coarse-grained model, a polymer chain is described by nodes consecutively connected by bonds in a linear fashion. Primary chain models [21,22] are (i) constant bond (CB) chains with consecutive nodes connected by a constant bond length on lattice, (ii) bond-fluctuating (BF) chains with fluctuating bond lengths on lattice, and (iii) bead-spring (BS) chains (and variants) off lattice. While the CB methods (i) are efficient in probing the equilibrium properties such as conformation of polymer chains, some microscopic details are usually missed in many simulations due to limited degrees of freedom with fast but somewhat artificial segmental dynamics. The off-lattice approaches (iii), on the other hand, are excellent in probing the microscopic details but generally too slow to reach equilibrium due to the long relaxation time in many complex systems, primarily because of large (practically infinite) degrees of freedom. In order to examine the approach to asymptotic global properties, resorting to simplifications [20] is almost unavoidable with either method unless one develops hybrid simulation approaches that incorporate the efficiency, effectiveness, and accuracy of both methods [23–25]. The BF model (ii) lies in between (i) and (iii) as it captures more microscopic details with a considerably larger number of degrees of freedom than the CB model

(i) without significantly compromising the efficiency of a discrete lattice. A chain node in the BF model occupies an elementary cube, i.e., a node is represented by eight lattice nodes in contrast to a single node in the CB model on a cubic lattice [21]. Due to excluded volume constraints of the node (cube) the bond length fluctuates between 2 and $\sqrt{10}$ with the exception of $\sqrt{8}$ and involves as many as 108 vectors [21] to access it. Banavar and co-workers [8,9] have recently argued that the thickness of the bonds that tether the nodes are very important to correctly take into account structural features of the protein. The bonds are usually very thin (negligibly small) in both CB (i) and BS (iii) models but they possess fluctuating length and thickness in the BF model (ii) which is our choice here to study the conformation and dynamics of a protein chain model.

Despite the limited degrees of freedom, the constant bond lattice model has the advantages of simplicity and computational efficiency, which are useful for exploring issues such as the energy landscape for stable structures of proteins. Using the CB description of the *HP* protein chain, Dill and co-workers [10,11] have successfully described the core assembly and protein folding into native structure via funnel pathways [12]. Lattice models provide useful insight into some of the basic characteristics of proteins [13,14]. The dynamics of the *HP* chain and its relaxation to the native structure is severely limited due to relatively few choices (degrees of freedom) for the node to move. In addition, it is not clear which combination of local segmental moves (i.e., kink-jump, crankshaft, reptation, etc.) [21] should be used to capture appropriate dynamical modes [23–25]. Off-lattice methods have contributed considerably in understanding the evolution of $\alpha$ helices and $\beta$ sheets [15–18] where microscopic details are crucial. Incorporation of local interactions with large numbers of degrees of freedom is very important in such analysis around stable structures even with crude approximations [20]. The large-scale dynamics involving relaxation from one structure to another is, however, severely restricted due to the long time needed with small movement of nodes in off-lattice simulations [23–25].

Very recently, Chen and Chen [19] have used the bond-fluctuation model to study the folding and native structures of a specific protein, sensory rhodopsin I. They have used interactions between nodes, empty sites (to represent the host medium, i.e., a membrane between the water planes at the opposite sides of the lattice), the bending energy for the chain bonds, and the interaction energy with the water planes. The simulations seem to be performed in three stages with appropriate interactions to avoid the long relaxation time and to capture desired segmental packing. This may appear somewhat *ad hoc* or arbitrary, but necessary to overcome the energy barriers and technical difficulties. We would like to investigate the conformational relaxation and segmental mobility of proteins in a single domain (keeping the same interaction throughout) in detail with a somewhat simpler set of interactions without imposing the constraints of a specific protein. We focus on the general features of a simplified bond-fluctuating protein chain model by a large-scale computer simulation study: we encounter major technical problem in reaching the native globular structure even with large-scale simulations but we are able to show how the radius of gyration relaxes ($R_g \propto t^{-\alpha}$) with a revealing conversion of conformational populations into their native state.

We consider a cubic lattice of size $L^3$ with $L=30$–200. The protein chains of length $N=50$–400 are considered with hydrophobic ($H$) and polar ($P$) nodes connected by fluctuating bonds. A node occupies a cube (eight lattice sites) and the bond length $l$ in units of the lattice constant can vary, $l^2=4, 5, 6, 9, 10$ with 108 vectors connecting consecutive monomers [21]. Initially, a chain of length $N$ is randomly placed in the lattice. Apart from the excluded volume effect, we consider a short-range interaction ($U$) among nodes and between nodes and empty sites which represent the effective solvent ($S$) components,

$$U = \sum_i \sum_k J(i,k), \tag{1}$$

where the index $i$ runs over all constituents ($H,P,S$) and $k$ over all their neighboring sites within a range $r_i$ of site $i$. The interaction energy between the constituents ($A,B$) at these sites

$$J(A,B) = \epsilon_{AB}. \tag{2}$$

The range of interaction is varied, i.e., $r_i^2 = l^2$, where $r_i^2 = 4, 5, 10$ represent nearest neighbor, next nearest neighbor sites, etc. The set of interaction matrix elements

$$\epsilon_{HS} = -\epsilon_{PS} = \epsilon_1, \quad \epsilon_{HP} = \epsilon_2, \quad \epsilon_{HH} = \epsilon_{PP} = \epsilon_3, \tag{3}$$

with a range of interaction strengths $\epsilon_i$, i.e., $\epsilon_1 = \epsilon_2 = 0, 1, 2, 3, \ldots$; $\epsilon_3 = 0$. The energy is measured in units of $k_B T$. The chain nodes are moved randomly to their neighboring sites (i.e., cubes) with the METROPOLIS algorithm and an attempt to move each node once defines one Monte Carlo step (MCS) as the time unit [21]. The simulation is performed for a long time with a number of independent runs for averaging the radius of gyration $R_g$ and mean square displacements of each node ($\langle R_n^2 \rangle$) and of their center of mass ($\langle R_c^2 \rangle$).
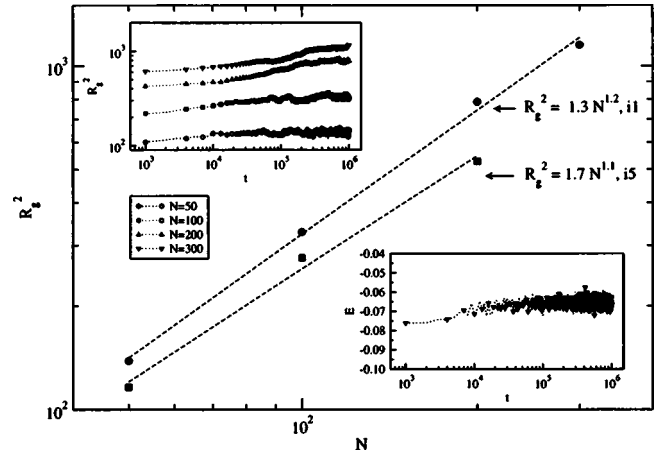


FIG. 1. $R_g^2$ versus chain length $N$ on $200^3$ samples for random $H$ and $P$ sequences. Nearest neighbor interaction ($r_i^2=4$) with interaction strengths $\epsilon_1=1$ ($i1$), 5 ($i5$) are used. Simulations are performed for up to $10^7$ time steps, each with 100 independent runs. Inset figures show the variation of $R_g^2$ (for $N=50$–300) and energy ($E$ for $N=300$ for clarity) with the time steps.

A variety of random and ordered sequences including diblock copolymers of $H$ and $P$ are considered for comparison. In general, the segmental and global motion of the protein chain depend on the fraction of $H$ and $P$ groups and their sequences. For example, a chain with primarily $H$ groups tends to be less mobile than a chain dominated by $P$ groups. Hydrophobic interactions seem to pin down the configurations leading to reduced mobility. The polar groups, on the other hand, enhance the segmental mobility and therefore accelerate the equilibration. The ordered sequences in blocks of $H$'s and $P$'s have lower energy than random sequences. However, the data presented below are generated from simulations performed with equal numbers of hydrophobic and polar groups in random sequences.

The protein chain equilibrates well with the short-range interaction ($r_i^2=4$); therefore it is easier to study the scaling of the radius of gyration with the molecular weight ($N$). Figure 1 shows $R_g^2$ versus $N$ plots on a log-log scale. Accordingly,

$$R_g^2 \propto N^\gamma, \tag{4}$$

with the exponent $\gamma \simeq 0.6$. Note that the radius of gyration of each chain has relaxed well which is confirmed by the variation of energy with the time steps (see inset figures). The data for higher interaction strength ($\epsilon_i$) are more fluctuating, but equilibration is achieved by increasing the simulation time. Thus, with the minimal interactions, the equilibrium conformation of the $HP$ protein chain is more like a self-avoiding walk than a compact globular form generally expected in a native state.

Figure 2 shows the variation of the gyration radius with the time steps for higher range ($r_i^2=5,10$) with different interaction strengths. For a relatively weak interaction ($\epsilon_1=1$), we see that the protein chain has relaxed well for $r_i^2=5$. The conformation is relatively spread as in Fig. 1. Increasing the interaction strength to $\epsilon_1=2,3,5$ leads to the onset of relax-
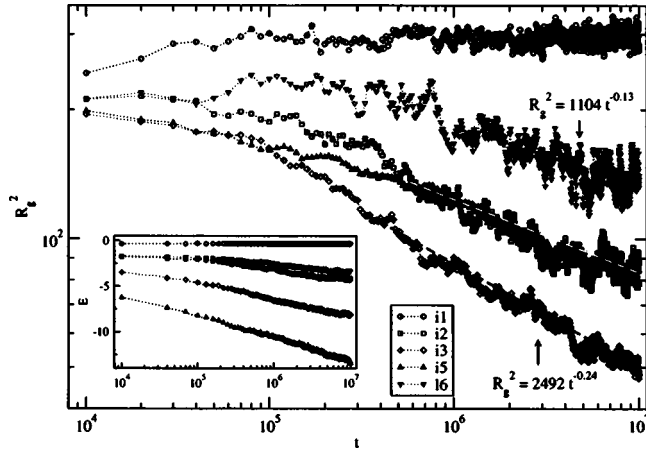
FIG. 2. Variation of $R_g^2$ with time steps on a log-log scale for $N=100$ with next neighbor interaction ($r_i^2=5$) with different interaction strengths [$\epsilon_1=1$ ($i1$), $2$ ($i2$), $3$ ($i3$), $5$ ($i5$)] and longer-range interaction [$r_i^2=10$ ($i6$)] with strength $\epsilon_1=1$ ($i1$) (down triangles). The inset figure shows corresponding energy variation on a semilog scale. Sample size $200^3$ is used with 100 independent runs.

ation into conformations with smaller sizes (presumably the globular). These relaxations show a power-law decay of the radius of gyration,

$$R_g^2 = R_0^2(1 + Ct^{-\alpha}), \qquad (5)$$

where $R_0$ is the radius of gyration of the chain in globular conformation and $C$ is a constant. The decay exponent $\alpha =0.06–0.12$ depends on the range and magnitude of the interaction. The decay of energy with the time steps is also consistent with the relaxation of protein chains into their globular structures.

It would be interesting to examine the variation of mean square displacements with the time steps in order to probe the segmental mobility as the protein chains relax into their globular conformation. Figure 3 shows such a variation. We see that for a relatively low interaction strength ($\epsilon_1=1$), both
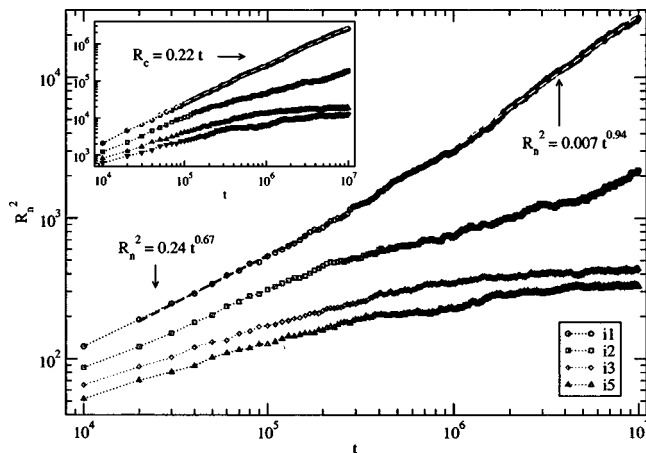


FIG. 3. Mean square displacement of each node $\langle R_n^2 \rangle$ and its center of mass $\langle R_c^2 \rangle$ (inset) versus $t$ for $N=100$ chain with random sequences of $H$ and $P$. Statistics is the same as in Fig. 2.
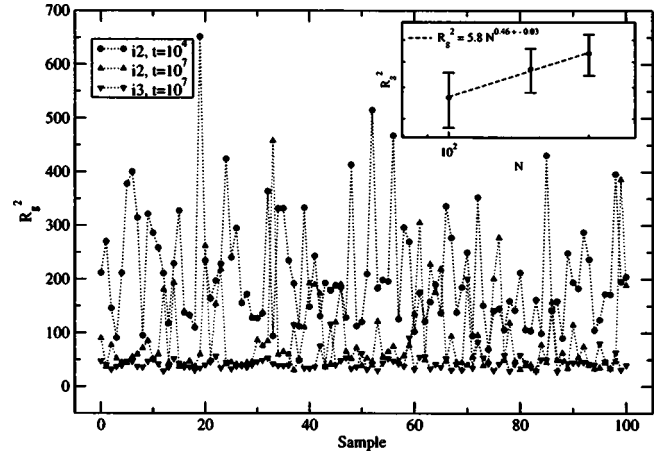


FIG. 4. Conformational histogram, i.e., variation of $R_g^2$ with independent runs for $N=100$ with next neighbor interaction ($r_i^2=5$) at time steps $t=10^4$, strength $\epsilon_1=2$ ($i2$), and $t=10^7$, $\epsilon_1=2$ ($i2$), $3$ ($i3$). Inset figure is $R_g^2$ versus $N$ plot on a log-log scale for $\epsilon_1=2$ ($i2$) with each data point generated for $t=10^7$ MCSs with a selective sampling of 100 independent runs; $y$-axis range is 35–80. Sample size $200^3$.

nodes and their center of mass reach a diffusive power-law behavior,

$$R_{n,c} \propto t^{\nu}, \qquad (6)$$

with $\nu \simeq 1/2$ in the asymptotic regime. At higher interaction strengths ($\epsilon_1=2,3,5$), on the other hand, a critical slowing down seems to occur. It becomes difficult to describe the dependence of $R_{n,c}^2$ on time by a single power law in our observation time. While stronger interactions beyond the nearest neighbor range are needed for the protein conformations to reach their globular states, the relaxation time is too large to reach such states within our computational resources at present. The necessity to resort to more simplifications (or increased computational resources) is unavoidable.

With such a large amount of data, it is desirable to dig further into the conformational relaxation. In Fig. 4, we present the histogram of the radius of gyration. At a relatively short time ($t=10^4$) from the beginning of simulation, we see a rather large spread in the magnitude of $R_g$ from one sample to another ($r_i^2=5$, $\epsilon_1=2$). Toward the end of the simulation ($t=10^7$), on the other hand, values of $R_g$ in most samples have fallen to a very low value ($R_g^2 \simeq 30–40$). Such a trend in population inversion from extended conformation into a globular form is also seen with higher interaction strengths. Thus, in order to analyze the native configuration, one has to selectively use those configurations which have reached their ground state. Such a crude sampling of the ground state conformations (after $10^7$ time steps) for different chain lengths shows signs of globular conformations (see the inset of Fig. 4). In summary, the conformation of the protein chain depends on the interaction (i.e., nature of the solvent) and the sequence. We are not able to distinguish differences in data with different random sequences due to large fluctuations. However, we have verified the changes by examining blocked sequences. While the appropriate interac-

tions are necessary to reach the native structures, the relaxation depends on the quality of the solvent. Despite a major problem with a very long relaxation time, our simulation reveals a clear population conversion ("funneling") of *HP* protein chains into their globular ground state. Relaxation of the radius of gyration shows a slow power-law decay [Eq. (5)] with a nonuniversal power-law exponent ($\alpha$) into a native structure with $R_g \propto N^\gamma$, $\gamma < 3/5$.

[1] C.K. Mathews, K.E. van Holde, and K.G. Ahern, *Biochemistry*, 3rd ed. (Addison-Wesley/Longman, Reading, MA, 2000).

[2] P.C. Turner, A.G. McLennan, A.D. Bates, and M.R. H. White, *Molecular Biology*, 2nd ed. (BIOS/Springer-Verlag, New York, 2000).

[3] C. Branden and J. Tooze, *Introduction to Protein Structure*, 2nd ed. (Garland, New York, 1999).

[4] D. Baker, Nature (London) **405**, 39 (2000).

[5] A.I. Jewett, V.S. Pande, and K.W. Plaxco, J. Mol. Biol. **326**, 247 (2003).

[6] B. Gillespie and K.W. Plaxco, Annu. Rev. Biochem. **73**, 837 (2004).

[7] J.E. Kohn *et al.* (unpublished).

[8] J.R. Banavar, O. Gonzalez, J.H. Maddocks, and A. Maritan, J. Stat. Phys. **110**, 35 (2003).

[9] J.R. Banavar and A. Maritan, Rev. Mod. Phys. **75**, 23 (2003), and references therein.

[10] K.M. Fiebig and K.A. Dill, J. Chem. Phys. **98**, 3475 (1993).

[11] K. Yue *et al.*, Proc. Natl. Acad. Sci. U.S.A. **92**, 312 (1995).

[12] J.D. Bryngelson, J.N. Onuchic, N.D. Socci, and P.G. Wolynes, Proteins **21**, 167 (1995).

[13] C. Tang, Physica A **288**, 31 (2000).

[14] E.I. Shakhnovich, Phys. Rev. Lett. **72**, 3907 (1994).

[15] U.H.E. Hansmann and L.T. Wille, Phys. Rev. Lett. **88**, 068105 (2002).

[16] U.H.E. Hansmann and Y. Okamoto, Curr. Opin. Struct. Biol. **9**, 177 (1999).

[17] U.H.E. Hansmann, Physica A **321**, 152 (2003).

[18] S. Takada, Proteins **42**, 85 (2001).

[19] C.-M. Chen and C.-C. Chen, Biophys. J. **84**, 1902 (2003).

[20] U.H.E. Hansmann, Comput. Sci. Eng. **5**, 64 (2003).

[21] *Monte Carlo and Molecular Dynamics Simulations in Polymer Science*, edited by K. Binder (Oxford University Press, New York, 1995).

[22] A. Baumgaertner, in *The Monte Carlo Method in Condensed Matter Physics*, 2nd ed., edited by K. Binder, Topics in Applied Physics Vol. 71 (Springer, Berlin, 1995).

[23] G.M. Foo and R.B. Pandey, Phys. Rev. E **57**, 5802 (1998).

[24] G.M. Foo and R.B. Pandey, J. Chem. Phys. **109**, 1162 (1998).

[25] G.M. Foo and R.B. Pandey, Macromol. Theory Simul. **8**, 571 (1998).