DENSITY OF STATES FOR HP LATTICE PROTEINS*

Michael Bachmann† and Wolfhard Janke‡

Institut für Theoretische Physik, Universität Leipzig
Augustusplatz 10/11, 04109 Leipzig, Germany

(Received August 5, 2003)

The density of states contains all informations on energetic quantities of a statistical system, such as the mean energy, free energy, entropy, and specific heat. As a specific application, we consider in this work a simple lattice model for heteropolymers that is widely used for studying statistical properties of proteins. For short chains, we have derived exact results from conformational enumeration, while for longer ones we developed a multicanonical Monte Carlo variant of the nPERM-based chain growth method in order to directly simulate the density of states. For simplification, only two types of monomers with respective hydrophobic (H) and polar (P) residues are regarded and only the next-neighbour interaction between hydrophobic monomers, being nonadjacent along the chain, is taken into account. This is known as the HP model for the folding of lattice proteins.

PACS numbers: 05.10.–a, 87.15.Aa, 87.15.Cc

1. Introduction

Proteins perform numerous functions in a biological cell system, *e.g.* controlling of transport processes of organelles, stabilisation of the cell structure, enzymatic catalysis of chemical reactions, *etc.* It is well established that the three-dimensional conformation of a protein within an aqueous environment determines its biological function. Due to the enormous number of tasks to be necessarily fulfilled to ensure the stability of a biological system, a large number of various proteins exists. All of them are built up of chains of amino acid residues, linked by peptide bonds. Since 20 different amino acids are known from nature, a protein with N monomers is, in principle, formed from 20^N possible sequences. Only a small number of so-called designing

* Presented at the Workshop on Random Geometry, Kraków, Poland, May 15–17, 2003.
† e-mail: michael.bachmann@itp.uni-leipzig.de
‡ e-mail: wolfhard.janke@itp.uni-leipzig.de

sequences, however, is actually realised in equilibrium. The reason is that the protein must be stable against thermal fluctuations and may not fold into a different shape leading to a loss of its associated function. Therefore, real proteins are supposed to possess a funnel-like deep global minimum in a rough free energy landscape [1]. It is one of the essential goals of computational protein research to identify the native state associated with the global free energy minimum of a protein with a given sequence of amino acid residues. Since the sequence of amino acids is known to be responsible for the resulting fold, it is also interesting to analyse what properties sequences of such favoured proteins have.

Unfortunately, computer simulations of real proteins are extremely difficult due to the relatively big number of degrees of freedom influenced by electrostatic, Lennard–Jones, hydrogen bond, torsional, and environmental interactions (for a review see, *e.g.*, Ref. [2]). In order to qualitatively study the folding behaviour of proteins and also for sequence analysis, simple lattice models seem to be very practical. Nevertheless, the determination of the lowest-energy states and their degeneracies remains challenging. In fact, it was shown [3] that folding proteins within the HP model [4], the most simple lattice model for proteins, is an NP-complete problem. On the numerical side, one technical problem is that the polymers are required to be self-avoiding. Thus, updating the conformation in a Monte Carlo simulation is quite involved. Two completely different methods are widely used, first the application of a move set consisting of transformations that allow the change of a conformation of total length N, while in the second method, chain growth, a new monomer is attached to the end of a partial chain of length $n < N$ until the total chain length is reached. Both techniques work well in computer simulations of polymers at comparatively high temperatures, for example the investigation of the Θ-point transition between compact globule polymer states and random coils [5]. For studying the low-temperature behaviour of heteropolymers, however, the application of move sets is not very suitable, since transformations that usually belong to a move set, *e.g.* end and corner flips, crankshafts, and pivot rotations are inefficient for the creation of very dense conformations. The transition between lowest-energy states and compact globules represents a "conformational barrier" at low temperatures that is much better circumvented with chain-growth based algorithms such as PERM [6] and its new variants nPERM$_{\mathrm{ss}}^{\mathrm{is}}$ [7].

We are interested in the energetic thermodynamic properties of heteropolymers for all temperatures and therefore we proposed a multicanonical chain growth algorithm [8] which allows an explicit sampling of the density of states. The density of states is identical with the canonical distribution at infinite temperature. Nevertheless, we also obtain very accurate results in the low-temperature region which in effect is due to the capacity of the

multicanonical sampling [9] which spreads the canonical distribution to a flat histogram, such that all energetic states are, in principle, equally probable within the simulation. At the end, the canonical distribution at any temperature and thus all thermodynamic functions can be obtained by a simple reweighting procedure. This is only possible since the multicanonical method allows a sampling of the entire space of states, including such events that are canonically suppressed by many orders of magnitude. In our simulations of the HP model for lattice proteins with more than 40 monomers, we were also required to sample the lowest-energy states having a probability of realization in the density of states of the order of 10^{-25}, since these states dominate the low-temperature behaviour of the protein. Another problem is that the conformational transition between ground states and globules just appears in this temperature region, causing a conformational barrier that is avoided best, as described above, by using an adequate chain growth algorithm. Therefore we combined the multicanonical method with the new PERM variants for simple and importance sampling, nPERMss and nPERMis [7], respectively, to obtain densities of states with high and uniform accuracies for all energies.

2. Density of states of HP lattice proteins

For simplicity, we investigate lattice proteins that consist of only two types of monomers: hydrophobic (H) and polar (P). This choice is made since most of the amino acids occurring in nature can be grouped into these two classes. Moreover it is assumed that the protein mainly folds due to an effective hydrophobic interaction. This means that a core of hydrophobic monomers is formed which is screened from the aqueous solvent by a shell of polar (or hydrophilic) residues. The simplest form of the HP model takes into account only the attractive interaction between next-neighbouring H monomers being nonadjacent along the chain [4]:

$$E = -\sum_{\langle i,j<i-1\rangle} \sigma_i \sigma_j, \tag{1}$$

where $\sigma_i = 0$ (1) if the ith monomer is polar (hydrophobic). The partition sum of a HP lattice protein with fixed sequence at temperature T is then given by $Z = \sum_{\{\boldsymbol{x}\}} \exp\{-E(\{\boldsymbol{x}\})/k_{\mathrm{B}}T\}$, where the sum is taken over all admissible conformations of the polymer. Sorting all conformational states with respect to their energies, the partition sum can also be expressed in terms of the density (or degeneracy) $g(E)$ of states with energy E:

$$Z = \sum_i g(E_i) \exp\left\{-\frac{E_i}{k_{\mathrm{B}}T}\right\}. \tag{2}$$

Knowing $g(E)$, the mean energy $\langle E\rangle$ of the system can be calculated by

$$\langle E\rangle(T) = \frac{\sum_i E_i\, g(E_i) \exp\{-\frac{E_i}{k_\mathrm{B}T}\}}{\sum_i g(E_i) \exp\{-\frac{E_i}{k_\mathrm{B}T}\}} \tag{3}$$

and the specific heat is given by the fluctuation formula

$$C_V(T) = \frac{1}{k_\mathrm{B}T^2} \left(\langle E^2\rangle - \langle E\rangle^2\right). \tag{4}$$

Other energetic quantities being related to the density of states are the Helmholtz free energy

$$F(T) = -k_\mathrm{B}T \ln \sum_i g(E_i) \exp\left\{-\frac{E_i}{k_\mathrm{B}T}\right\} \tag{5}$$

and the entropy

$$S(T) = \frac{1}{T}\left[\langle E\rangle(T) - F(T)\right]. \tag{6}$$

3. Exact enumeration of 14mers

As a first example we have investigated HP proteins with 14 monomers by enumerating all possible conformations. This study is quite interesting, because there is only one sequence ($\mathrm{HPHPH_2PHPH_2P_2H}$, in the following denoted as 14.1) that is *designing*, *i.e.* the ground state of the associated lattice protein is unique (up to translational, rotational, and one reflection symmetry). It possesses $n_\mathrm{H} = 8$ hydrophobic monomers and the ground-state energy is $E_\mathrm{min} = -8$, since there are 8 hydrophobic contacts (see Fig. 1). In order to understand the particular properties of such a protein

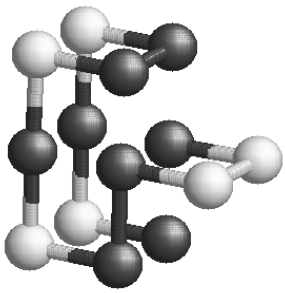

Fig. 1. Unique ground state of the 14mer with designing sequence 14.1 (dark spheres: hydrophobic residues, light spheres: polar monomers).

with the lowest ground-state degeneracy among all the 2^{14} different 14mers, we compare it with three other ones having similar properties ($n_{\mathrm{H}} = 8$, $E_{\min} = -8$), but different sequences and therefore different ground-state degeneracies. The degeneracy of the lowest-energy state of the sequences 14.2 ($H_2P_2HPHPH_2PHPH$) and 14.3 ($H_2PHPHP_2HPHPH_2$) is twice that of the designing sequence 14.1, while sequence 14.4 ($H_2PHP_2HPHPH_2PH$) is even four times higher degenerated. Figure 2 shows the densities of states for the four sequences. Since the densities of the excited states do not considerably differ (see Table I), the low-temperature behaviour of these proteins can only vary due to the different ground-state degeneracies. Indeed, the specific heat shown in Fig. 3 exhibits a pronounced low-temperature peak only for the designing sequence 14.1, while it is largely suppressed for the other proteins. This peak indicates the transition from the ground states to compact globule states. At higher temperatures, the globules unfold and form random coil conformations.

TABLE I

Exact total densities of the states with energy E for the 14mers. The entries of the table include all states that contribute to the partition function Z_∞ for 14mers at infinite temperature (except translations) which counts the number of self-avoiding random walks with $(14 - 1) = 13$ steps.

	sequence			
E	14.1	14.2	14.3	14.4
−8	48	96	96	192
−7	12 576	10 560	9 576	11 136
−6	162 120	140 496	126 240	160 536
−5	1 349 808	1 089 792	1 053 744	1 259 040
−4	8 434 536	6 661 032	6 028 944	7 831 752
−3	36 120 840	29 943 792	28 329 504	38 367 360
−2	118 052 520	100 663 488	109 433 232	129 351 360
−1	312 691 992	273 343 176	305 911 056	314 705 352
0	467 150 070	532 122 078	493 082 118	452 287 782
Z_∞	943 974 510	943 974 510	943 974 510	943 974 510

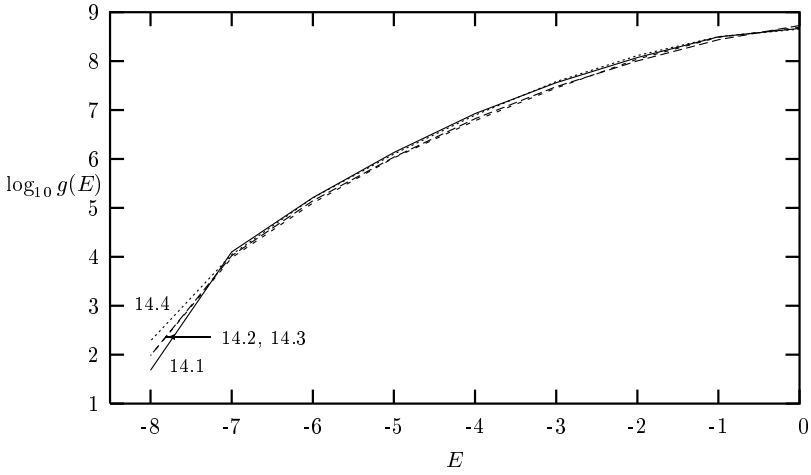

Fig. 2. Exact densities of states of exemplified 14mers with similar properties ($n_{\mathrm{H}} = 8$, $E_{\min} = -8$) but different sequences.

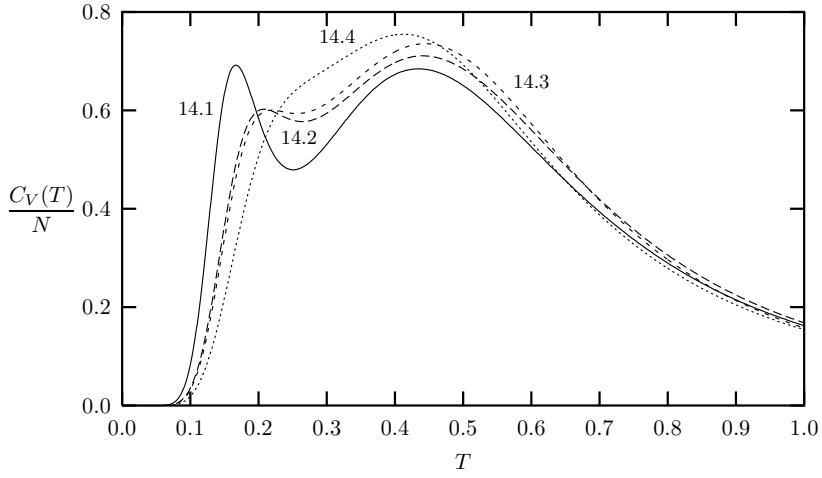

Fig. 3. Specific heat of the 14mers.

4. Simulation of a 42mer: lattice model of *pectate lyase C*

For lattice proteins with more than 20 monomers, enumeration becomes exhausting, since the number of conformations grows exponentially with the number of monomers [10]. More sophisticated search algorithms are required to sample the phase space. For this reason, we developed a multicanonical chain growth algorithm [8] that combines the advantages of avoiding confor-

mational barriers by using a PERM-based chain growth method [6,7] and the capacity of a flat histogram technique allowing the sampling of the entire energy space [9]. In order to achieve this, the canonical distributions provided by PERM at each intermediate length of the growing chain must be flattened. As usual, the multicanonical weights are determined by an iterative procedure [9]. We applied this method to calculate the density of states of a lattice 42mer with sequence $PH_2PHPH_2PHPHP_2H_3PHPH_2PHPH_3P_2HPH$-$PH_2PHPH_2P$ which was designed to simulate the ground-state properties of the parallel β helix of the protein *pectate lyase C* [11–13]. The ground state is known to be low-degenerated. Up to translations, rotations, and reflections there are only 4 ground-state conformations with energy $E_{\min} = -34$. The density of states ranges over 25 orders of magnitude, and the ground states were hit frequently with our simulation method such that the low-temperature properties of this protein could be investigated with good accuracy. In Fig. 4, we show the specific heat and the mean energy of the 42mer. The specific heat has two peaks, the low-temperature ground-state-globule transition occurs near $T_0 \approx 0.27$ and the transition between globules and random coils at $T_1 \approx 0.53$.

In Ref. [8], we have also compared two 48mers with different ground-state degeneracies and found also there that a pronounced low-temperature peak in the specific heat only appears for the example with the lower degeneracy of the ground state (which was about 5000 in that case).

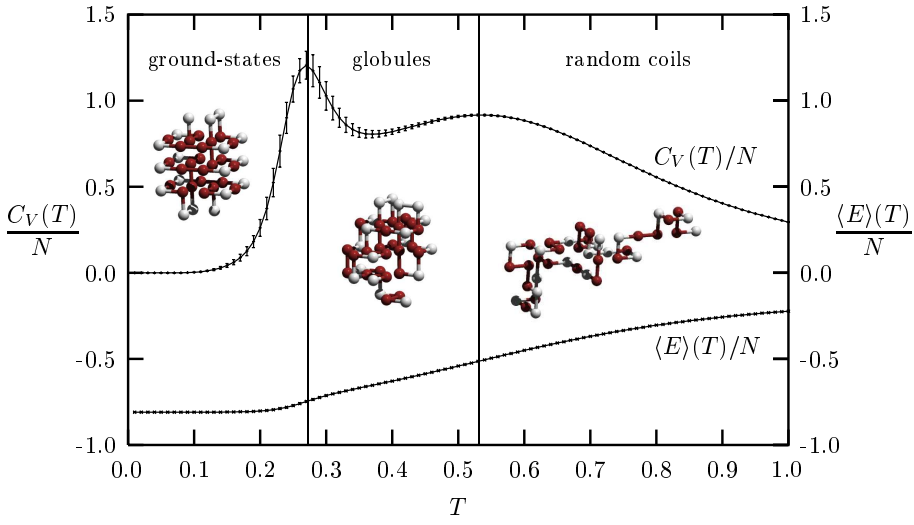

Fig. 4. Specific heat and mean energy of the 42mer.

5. Summary

We have discussed the relation between the low degeneracy of the low-lying energy states and the appearance of a low-temperature transition between compact globules and ground states of HP lattice proteins with 14 and 42 monomers, respectively. For this purpose, we calculated the density of states of the 14mers by exact enumeration of all possible conformations. In order to simulate the density of states of the 42mer with necessarily high accuracy, we developed a multicanonical chain growth algorithm that enabled us to sample the density of states over the entire energy space. As the main qualitative conclusion we find a correlation between the degeneracy of low-lying states and the sharpness of the transition to compact globule states.

This work is partially supported by the German–Israel-Foundation (GIF) under grant No. I-653-181.14/1999 and the EU-Network HPRN-CT-1999-000161 "Discrete Random Geometries: From Solid State Physics to Quantum Gravity".

REFERENCES

[1] K.A. Dill, *Prot. Sci.* **8**, 1166 (1999).

[2] U.H.E. Hansmann, Y. Okamoto, The Generalized-Ensemble Approach for Protein Folding Simulations, in: *Annual Reviews of Computational Physics VI*, ed. D. Stauffer, World Scientific, Singapore 1999, p. 129.

[3] B. Berger, T. Leighton, *J. Comp. Biol.* **5**, 27 (1998); P. Crescenzi, D. Goldman, C. Papadimitriou, A. Piccolboni, M. Yannakakis, *J. Comp. Biol.* **5**, 423 (1998).

[4] K.A. Dill, *Biochemistry* **24**, 1501 (1985); K.F. Lau, K.A. Dill, *Macromolecules* **22**, 3986 (1989).

[5] See, *e.g.*, P.G. de Gennes, *Scaling Concepts in Polymer Physics*, Cornell University Press, Ithaca 1979.

[6] P. Grassberger, *Phys. Rev.* **E56**, 3682 (1997); H. Frauenkron, U. Bastolla, E. Gerstner, P. Grassberger, W. Nadler, *Phys. Rev. Lett.* **80**, 3149 (1998); U. Bastolla, H. Frauenkron, E. Gerstner, P. Grassberger, W. Nadler, *Proteins* **32**, 52 (1998).

[7] H.-P. Hsu, V. Mehra, W. Nadler, P. Grassberger, `cond-mat/0209366`; *J. Chem. Phys.* **118**, 444 (2003).

[8] M. Bachmann, W. Janke, `cond-mat/0304613`, to appear in *Phys. Rev. Lett.* (in print).

[9] B.A. Berg, T. Neuhaus, *Phys. Lett.* **B267**, 249 (1991); *Phys. Rev. Lett.* **68**, 9 (1992); W. Janke, *Physica A* **254**, 164 (1998).

[10] A. Irbäck, C. Troein, *J. Biol. Phys.* **28**, 1 (2002).

[11] M.D. Yoder, N.T. Keen, F. Jurnak, *Science* **260**, 1503 (1993).

[12] K. Yue, K. A. Dill, *Phys. Rev.* **E48**, 2267 (1993); *Proc. Natl. Acad. Sci. USA* **92**, 146 (1995).

[13] G. Chikenji, M. Kikuchi, Y. Iba, *Phys. Rev. Lett.* **83**, 1886 (1999).