Testing a New Monte Carlo Algorithm for Protein Folding

Ugo Bastolla,[1] **Helge Frauenkron,**[1] **Erwin Gerstner,**[1,2] **Peter Grassberger,**[1,2]* **and Walter Nadler**[1]
[1]*HLRZ, c/o Forschungszentrum Jülich, Jülich, Germany*
[2]*Department of Physics, University of Wuppertal, Wuppertal, Germany*

ABSTRACT We demonstrate that the recently proposed pruned-enriched Rosenbluth method (PERM) (Grassberger, Phys. Rev. E 56:3682, 1997) leads to extremely efficient algorithms for the folding of simple model proteins. We test it on several models for lattice heteropolymers, and compare it to published Monte Carlo studies of the properties of particular sequences. In all cases our method is faster than the previous ones, and in several cases we find new minimal energy states. In addition to producing more reliable candidates for ground states, our method gives detailed information about the thermal spectrum and thus allows one to analyze thermodynamic aspects of the folding behavior of arbitrary sequences. Proteins 31:52–66, 1998. © 1998 Wiley-Liss, Inc.

Key words: heteropolymers; lattice models; lattice polymers; Monte Carlo; protein folding; protein structure prediction

INTRODUCTION

Protein folding[1–4] is one of the most interesting and challenging problems in polymer physics and mathematical biology. It is concerned with the problem of how a heteropolymer of a given sequence of amino acids folds into precisely that geometrical shape in which it performs its biological function as a molecular machine.[5,6] Currently, it is much simpler to find coding DNA—and thus also amino acid—sequences than to elucidate the 3D structures of given proteins. Therefore, solving the protein folding problem would be a major breakthrough in understanding the biochemistry of the cell and in designing artificial proteins.

In this article we are concerned with the direct approach, namely, when given a sequence of amino acids, a molecular potential, and no other information, to find the ground state and the equilibrium state at physiological temperatures. Note that we are not concerned with the kinetics of folding, but only in the final outcome. Also, we will not address the problems of how to find good molecular potentials,[7–9] and what is the proper level of detail in describing proteins.[8] Instead, we will use simple coarse-grained models that have been proposed in the literature and have become standards in testing the efficiency of folding algorithms.

A plethora of methods have been proposed to solve this problem, ranging from simple Metropolis Monte Carlo simulations at some nonzero temperature[10] over multicanonical simulation approaches[11] to stochastic optimization schemes based, e.g., on simulated annealing[12] and genetic algorithms.[13,14] Alternative methods use heuristic principles,[15] information from databases of known protein structures,[16] sometimes in combination with known physicochemical properties of small peptides.

The algorithms we apply here are variants of the pruned-enriched Rosenbluth method (PERM).[17] This is a chain growth approach based on the Rosenbluth-Rosenbluth (RR)[18] method. Preliminary results have been published before.[19] Here we will provide more details on the algorithm and on the analyses that can be performed, and we will present more detailed results on ground-state and spectral properties, and on the folding behavior of the sequences analyzed.

MODELS

The models we have studied are heteropolymers that live on two- and three-dimensional regular lattices. They are self-avoiding chains with attractive or repulsive interactions between neighboring nonbonded monomers.

The majority of authors considered only two kinds of monomers. Although different interpretations are also possible for such a binary choice, e.g., in terms of positive and negative electric charges,[20] the most important model of this class is the HP model.[21,22] There, the two monomer types are assumed to be

---

TABLE I. Newly Found Lowest Energy States for Binary Sequences With Interactions $\epsilon = (\epsilon_{HH}, \epsilon_{HP}, \epsilon_{PP})$*

N	d	ϵ	Sequence configuration	E	Reference
100	2	$(-1, 0, 0)$	$P_6HPH_2P_5H_3PH_5PH_2P_2(P_2H_2)_2PH_5PH_{10}PH_2PH_7P_{11}H_7P_2HPH_3P_6HPH_2$	-44	28
			$r_6ur_2u_3rd_5luldl_2drd_2ru_2r_3(rulu)_2$ $urdrd_2ru_3lur_3dld_2rur_5d_3l_5uldl_2d_3ru_2r_3d_3l_2urul$	-47	
			$rdldldrd_2r_2d_3$ $l_3drdldr_2dl_2$ $dl_2(urul)_2urur_2ul_2u_{22}drd_2lul_2uru_2r_2u_4rul_3$ $drd_3l_2d_2ldllu_2ru_2lu_3rd_2rdr$	-47	
			$u_3r_3ur_5dl_4drd_2r_2ulur_4dr_2dld_2lu_2luld_2rdl_2ul_3dr_2dr_2drurdr_3d_2luld_2l_2drd_2lulu_3$ $l_2ul_2u_3ru$	-47	
			$rdr_3dldl_3drdrur_2ur_2ulur_2urd_2(ldrd)_2l_2u_2ldl_2dr_2dr_3dl_2dlulul_2ul_2d_3ldr_3$ $u_2rdrdldldr_2urur_4u_3rul$	-47	
			$rd_3rdldl_2uld_3ld_2rur_3dl_2dr_6ul_3u_2l_3ur_3ur_2dld_2rur_2ulu_2l_2u_3lu_3r_2d_5r_2d_2rdru_2lu_2r_2$ u_2ldl_2dl	-47	
			$r_3u_3ru_2ru_3l_2ur_2ul_2urul_4dr_2dldrdld_2rur_2dld_2luld_2rdl_3uru_2ldld_3l_3ururu_2$ $rur_2ululdldl_2u_3ld_4r$	-47	
100	2	$(-1, 0, 0)$	$P_3H_2P_2$ $H_4P_2H_3(PH_2)_3H_2P_8$ $H_6P_2H_6P_9HPH_2PH_{11}P_2H_3PH_2PHP_2HPH_3P_6H_3$	-46	28
			$ru_2ldlu_2ld_2lu_2lurulur_2d_2ru_4r_2dld_4ru_3rdrdld_6rdru_3rul_2u_2rdr_2ululu_2(rd)_3$ $rur_2dldld_4lulu_2ru$	-49	
			$u_3rdru_2rd_2$ $ru_2r_2u_2ldluldlu_5ld_6l_2d_2lu_3r_2u_6l_2d_3ldr_2dl_2dldlu_2ru_2ldl_2drdl_2$ $d_2rurdrd_3ru_3ru$	-49	
			$ul_2drdl_2u_3ld_4ldrdl_2u_2l_2d_3l_2uru_3r_2u_3rd_3ru_4rul_5dldr_2d_2luldldrdldldlu_3lul_2ulur_2dr_2$ u_3rd_4l	-49	
60	2	$(-1, 0, 0)$	$P_2H_3PH_8P_3H_{10}PHP_3H_{12}P_4H_6PH_2PHP$	-34	13
			$r_5d_2lul_3$ $dld_2(ru)_2rd_2ldldrdr_2uluru_2rd_2rdldr_2u_3lu_3rd_2rur$	-36	
80	3	$(-1, 0, -1)$	$PH_2P_3(H_3P_2H_3P_3H_2P_3)_3H_4P_4(H_3P_2H_3P_3H_2P_3)H_2$	-94	24, 29
			$lbruflbl_2br_2drur_2dldl_3ulfrdr_3urfldl_3ulurur_3drblul_3br_3bl_3dldrdr_3urul_2dlu$	-98	

*Configurations are encoded as sequences of r (ight), l (eft), u (p), d (own), f (orward), and b (ackward).

hydrophobic (H) and polar (P), with energies $\epsilon_{HH} = -1$ and $\epsilon_{HP} = \epsilon_{PP} = 0$ for interaction between not covalently bound neighbors. Since this parameter set leads to highly degenerate ground states, alternative parameters were proposed, e.g., $\epsilon = (-3, -1, -3)$[23] and $\epsilon = (-1, 0, -1)$.[24] Note, however, that in these latter parameter sets, since they are symmetric upon exchange of H and P, the intuitive distinction between hydrophilic and polar monomers gets lost.

An interesting extension to the HP model can be obtained by allowing the interactions to be anisotropic. This is done by introducing amphipatic (A) monomers that have hydrophobic as well as polar sides.[25] Such a generalization is possible for all lattice types, but we confine ourselves to two dimensions (2D) here. It can be shown that in this HAP model for a wide range of interaction parameters, the inverse folding problem, i.e., the determination of a sequence that has a particular conformation as ground state, can be solved by construction.[25] While three-dimensional chiral amphipatic monomers can be considered as well as nonchiral (the sides are allowed to rotate in the latter), in 2D only nonchiral monomers are possible.

In the other extremal case of models, all monomers of a sequence are considered to be different, and interaction energies are drawn randomly from a continuous distribution.[26,27] These models correspond, effectively, to assuming an infinite number of monomer types.

SEQUENCES

For the above models various sequences were analyzed in the literature, and in Frauenkron et al. (1998)[19] these analyses were taken as a test for our algorithm. Here we will take a closer look at the properties of some of the sequences that were considered there.

2D HP Model

Two-dimensional HP chains were used in several papers as test cases for folding algorithms. We shall discuss the following two.

First, several chains of length 20 to 64 were studied by means of a genetic algorithm in Unger and Moult (1993)[13], in which the authors quote supposedly exact ground-state energies, and lowest energies obtained by simulations. While these coincide for the shorter chains ($N \le 50$), the authors of that article were unable to fold the longer chains with $N = 60$ and 64.

Second, two chains with $N = 100$ were studied in Ramakrishnan et al. (1997)[28], in which the authors claimed that their native configurations were compact, fitting exactly into a 10×10 square, and had energies -44 and -46 (see Table I for the sequences and Figs. 1 and 2 for the respective proposed ground-state structures). These conformations were found by a specially designed Monte Carlo (MC) algorithm, which should be particularly efficient for compact configurations.
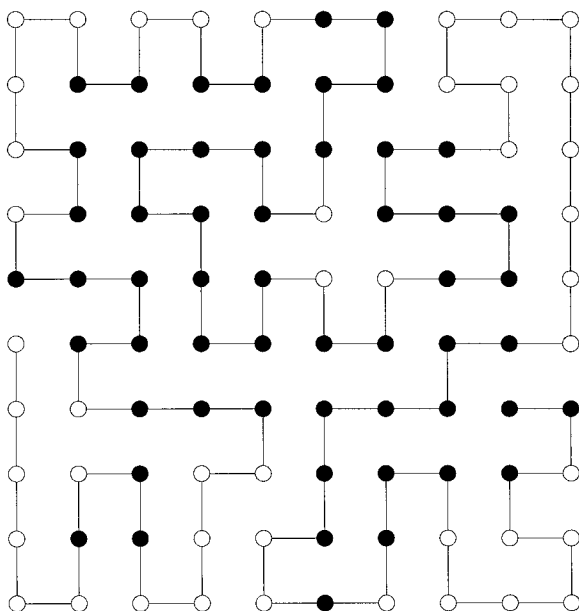
Fig. 1. Putative compact native structure of sequence 1 from Table I ($E = -45$) according to Ramakrishnan et al. (1997).[28] Filled circle denotes H monomers; open circle, P monomers. .
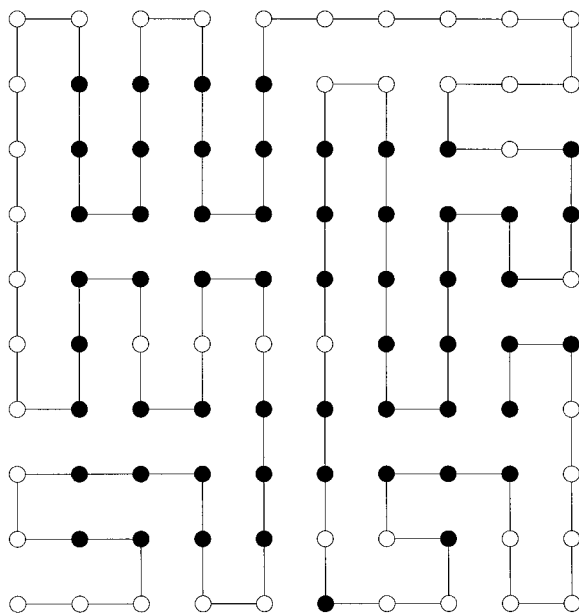

Fig. 2. Putative compact native structure of sequence 2 from Table I ($E = -46$) according to Ramakrishnan et al. (1997).[28]

2D HAP Model

To analyze the performance of PERM on the HAP model, we used $\epsilon = (-4, -1, 2)$ as energy parameters, for which set the inverse folding problem is solvable.[25] We choose a three-helix structure with $N = 42$ (see Fig. 3), which is the ground state of the sequence $pA(HAAH)_3PA_3H_{10}A_3P(HAAH)_3 Ap$, where
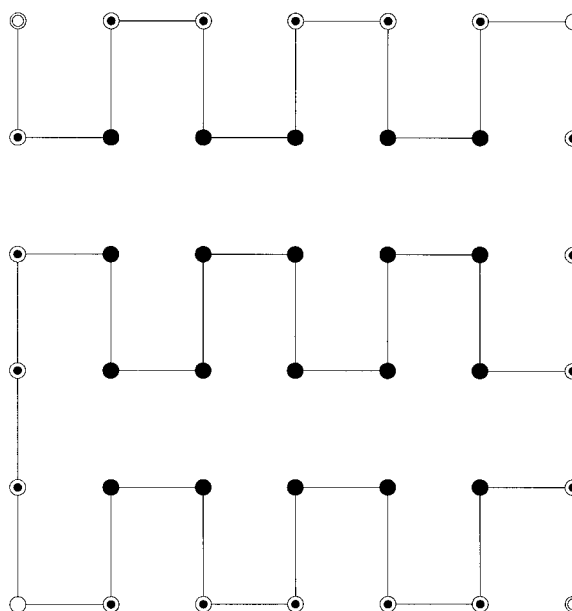

Fig. 3. Ground-state structure of the HAP sequence. Filled circle denotes H monomers; structured circle, A monomers; open circle, P monomers.

A denotes an amphipatic intrachain group with one hydrophobic and one polar side, while p denotes an amphipatic end group with one hydrophobic and two polar sides.

3D HP Model

Ten sequences of length $N = 48$ were given in Yue et al. (1995).[30] Each of these sequences was designed by minimizing the energy of a particular target conformation in sequence space under the constraint of constant composition.[31] The authors tried to find the lowest energy states with two different methods, one being a heuristic stochastic approach,[15] the other based on exact enumeration of low energy states.[32] With the first method they failed in all but one case to reach the lowest energy. With the second method in all but one cases they obtained conformations with energies that were even lower than the putative ground states the sequences were designed for, while for another case the ground-state energy was confirmed. Precise CPU times were not quoted.

3D Modified HP Model

A most interesting case is a two-species 80-mer with interactions $(-1, 0, -1)$ studied first in O'Toole and Panagiotopoulos (1992).[24] These particular interactions were chosen instead of the HP choice $(-1, 0, 0)$ because it was hoped that this would lead to compact configurations. Indeed, the sequence was specially designed to form a "four-helix bundle" that fits perfectly into a $4 \times 4 \times 5$ box (see Fig. 4). Its energy in this putative native state is -94. Although the
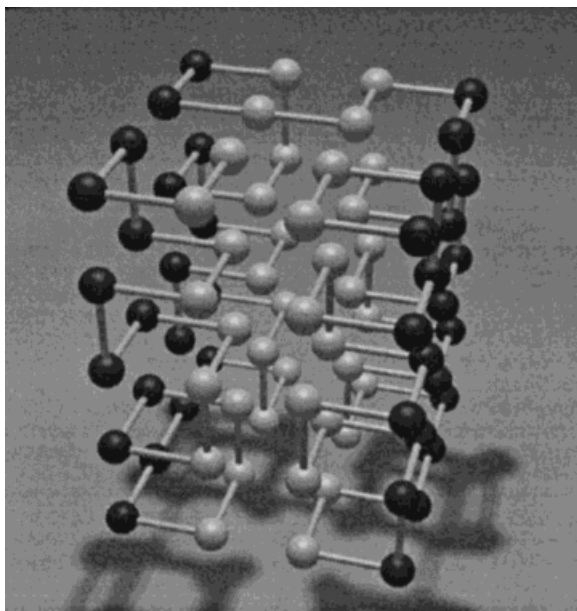
Fig. 4. Putative native state of the "four helix bundle" sequence (see Table I) as proposed by O'Toole and Panagiotopoulos.[24] It has $E = -94$, fits into a rectangular box, and consists of three homogeneous layers. Structurally, it can be interpreted as four helix bundles.

authors of O'Toole and Panagiotopoulos (1992)[24] used highly optimized codes, they were not able to recover this state by MC. Instead, they reached only $E = -91$. Supposedly, a different state with $E = -94$ was found in Ramakrishnan et al. (1997)[28], but figure 10 of this paper, which is claimed to show this configuration, has a much higher value of E. Configurations with $E = -94$ but slightly different from that in O'Toole and Panagiotopoulos (1992)[24] and with $E = -95$ were found in Deutsch (1997)[29] by means of an algorithm similar to that in Ramakrishnan et al. (1997).[28] For each of these low energy states the authors needed about one week of CPU time on a Pentium.

3D ∞ Monomer Types

Sequences with $N = 27$ and with continuous interactions were studied in Klimov and Thirumalai (1996).[27] Interaction strengths were sampled from Gaussians with fixed nonzero mean and fixed variance. These $N(N-1)/2$ numbers were first attributed randomly to the monomer pairs, then randomly permuted, using a Metropolis accept/reject strategy with a suitable cost function, to obtain good folders. Such "breeding" strategies to obtain good folders were also developed and employed in other studies for various models,[31,33,34] and seem necessary to eliminate sequences that fold too slowly and/or unreliably. It is believed that during biological evolution optimization processes also took place with similar

effects, so that actual proteins are better folders than random amino sequences.

ALGORITHM

The algorithms we apply here are variants of the pruned-enriched Rosenbluth method (PERM),[17] a chain growth algorithm based on the Rosenbluth-Rosenbluth (RR)[18] method. There, monomers are placed sequentially at vacant sites either with uniform probability, or with some nonuniform probability distribution. In either case it leads to weighted samples where each configuration carries a weight W. For long chains or low temperatures, the spread in weights can become very wide, which then leads to serious problems.[35] But since the weights accumulate as the chains grow, one can interfere during the growth process by pruning configurations with low weights and replacing them by copies of high-weight configurations. This is in principle similar to population-based methods in polymer simulations[36,37] and in quantum Monte Carlo,[38] but the implementation is different. Pruning is done stochastically: if the weight of a configuration has decreased below a threshold $W^<$, it is eliminated with probability $\frac{1}{2}$, whereas it is kept and its weight is doubled in the other half of cases. Copying (enrichment[39]) is done independently of this. If W increases above another threshold $W^<$, the configuration is replaced by n copies, each with weight W/n. Technically, this is done by putting onto a stack all information needed about configurations that still have to be copied. This is most easily implemented by recursive function calls, by which the need for keeping large populations of configurations is avoided.[36–38] PERM has proven extremely efficient for studies of lattice homopolymers near the θ point.[17] It has also been successfully applied to phase equilibria,[40] to the ordering transition in semistiff polymers,[41] and to spiraling transitions of polymers with interactions depending on relative orientation of monomers.[42] We refer to these papers for more detailed descriptions of the basic algorithm.

The main freedom when applying PERM consists in the a priori choice of the sites for placing the next monomer, in the thresholds $W^<$ and $W^>$ for pruning and copying, and in the number of copies made each time. All these features do not affect the formal correctness of the algorithm, but they can greatly influence its efficiency. They may depend arbitrarily on chain lengths and on local configurations, and they can be changed freely at any time during the simulation. Thus the algorithm can "learn" during the simulation.

In order to apply PERM to heteropolymers at very low temperatures, the strategies proposed in Grassberger (1997)[17] are modified as follows.

(1) For homopolymers near the θ point, it had been found that the best choice for the placement of

monomers was not according to their Boltzmann weights, but uniformly on all allowed sites.[17,40] This might be surprising since the Boltzmann factor has then to be included into the weight of the configuration, which might lead to large fluctuations. Obviously, this effect is counterbalanced by the fact that larger Boltzmann factors correspond to higher densities and thus to smaller Rosenbluth factors.[35]

For a heteropolymer this has to be modified, as there is no longer a unique relationship between density and Boltzmann factor. In a strategy of anticipated importance sampling, we should preferentially place monomers on sites with mostly attractive neighbors, let us assume that we have two kinds of monomers, and that we want to place a type-A monomer. If an allowed site has m_B neighbors of type B ($B = H$, P), we would select this site with a probability $\propto 1 + a_{AH}m_H + a_{AP}m_P$. Here a_{AB} are constants, with $a_{AB} > 0$ for $\epsilon_{AB} < 0$ and vice versa.

(2) Most naturally, $W^<$ and $W^<$ are chosen proportional to the estimated partition sum Z_n.[17] This becomes inefficient at very low T since Z_n will be underestimated as long as no low-energy state is found. When this finally happens, $W^>$ is too small. Thus too many copies are made that are all correlated but cost much CPU time.

This problem can be avoided by increasing $W^>$ and $W^<$ during particularly successful tours (a tour is the set of configurations derived by copying from a single start[17]). But then the average number of long chains is also decreased in comparison with short ones. To reduce this effect and to create a bias toward a sample that is flat in chain length, we multiply by some power of M_n/M_1, where M_n is the number of generated chains of length n. With $\mathcal{N}(n)$ denoting the number of chains generated during the current tour, we used therefore

$$W^< = CZ_n[(1 + \mathcal{N}(n)/M)(M_n + M)/(M_1 + M)]^2$$

and $W^> = rW^<$. Here C is a constant of order unity, $r \approx 10..100$, and M is a constant of order 10^4–10^5.

(3) Creating only one new copy at each enrichment event (as done in Grassberger (1997)[17]) cannot prevent the weights from exploding at very low T. Thus we have to make several copies if the weight is large and surpasses $W^>$ substantially. A good choice for the number of new copies created when $W > W^>$ is int$[\sqrt{W/W^>}]$.

(4) Two special tricks were employed for compact configurations of the 2D HP model filling a square. First of all, since we know in this case where the boundary should be, we added a bias for polar monomers actually to be on that boundary, by adding an additional energy of -1 per boundary site. Note that this bias has to be corrected in the weights, thus

the final distributions are unaffected by it and unbiased. Secondly, in two dimensions we can immediately delete chains that cut the free domain into two disjoint parts, since they never can grow to full length. In the present simulations, we checked for this by looking ahead one time step. In spite of the additional work this was very efficient, since it reduced considerably the time spent on dead-end configurations.

(5) In some cases we did not start to grow the chain from one end but from a point in the middle. We grew first one half, and then the other. Results were averaged over all possible starting points. The idea behind this is that real proteins have folding nuclei,[43] and it should be most efficient to start from such a nucleus. In some cases this trick was very successful and speeded up the ground-state search substantially, in others not. We take this observation as an indication that in various sequences the end groups already provide effective nucleation sites. This is the case, for instance, for the 80-mer with modified HP interactions of O'Toole and Panagiotopoulos (1992).[24] We also tried to grow the chain on both sides simultaneously. However, it turned out that this is not effective computationally.*

(6) In the case of the HAP model, it turned out that, while the ground-state configuration of the chain geometry could be reached easily, the ground-state configurations of the side groups (i.e., of the H/P bond attributions of amphiphilic monomers[25]) could note be reached effectively using PERM alone. Therefore, after building the chain conformation to its full length, we let the side groups of the amphipatic monomers rotate thermally using a Metropolis algorithm. This approach utilizes the short relaxation time of side group fluctuations within the subphase of a fixed chain conformation and leads to the desired ground states.[†]

(7) For an effective sampling of low-lying states, the choice of simulation temperature T appears to be of importance. If it is too large, low-lying states will have a low statistical weight and will not be sampled reliably. On the other hand, if T is too low, the algorithm becomes too greedy: configurations that look good at first sight but lead to dead ends are

---

*At first sight, one might believe that allowing the chain to grow at both ends should decrease the attrition rate and hence be advantageous. For homopolymers one can see easily that this is not true. Attrition is actually decreased, since chains that have one end blocked can still grow at the other end. But such chains have small Rosenbluth factors and thus extremely low W in average. Therefore, they just cost efforts without efficiently contributing to statistical averages. For heteropolymers this argument no longer holds, however, and it is not clear why growing the chain at both ends is not efficient either.

†Note, however, that there exists an alternative approach: for a fixed chain geometry, a computationally effective, i.e., $O(N)$ instead of $O(2^N)$, complete enumeration of all side group conformations is possible (Gerstner et al., unpublished data).

sampled too often, while low-energy configurations, whose qualities become apparent only at late stages of the chain assembly, are sampled rarely. Of course they then get huge weights (since the algorithm is correct after all), but statistical fluctuations become huge as well. This is in complete analogy to the slow relaxation hampering more traditional (Metropolis type) simulations at low T. Note, however, that "relaxation" in the proper sense does not exist in the present algorithm.

In the cases we considered, it turned out to be most effective to choose a temperature that is below the collapse transition temperature (note, however, that this transition is smeared out, see the results below) but somewhat above the temperature corresponding to the structural transition that leads to the native state. This observation corresponds qualitatively to the considerations of Finkelstein et al. (1995),[44] although a quantitative comparison appears not to be possible.

(8) For two-dimensional HP chains, we performed also some runs where we further restricted the search for native configurations by disallowing non-bonded HP neighbor pairs. The idea behind this was that such a pair costs energy, and is thus less likely to appear in a native state. But this is only a weak heuristic argument. Forbidding such pairs certainly gives wrong thermal averages, and it might prevent the native state to be found, if it happens to contain such a pair. But in two cases this restriction did work, and gave states with lower energies than those we could reach without this trick.

RESULTS

Let us now discuss our results. All CPU times quoted below refer to SPARC Ultra machines with 167 MHz.

2D HP Model

For all chains of Unger and Moult (1993)[13] we easily reached the ground state, except for the longest chain ($N = 64$). For this chain the ground state is very regular, with all polar monomers on the outside, and with no nonbonded HP neighbors. Its energy is -42 (see Fig. 5). The authors of Unger and Moult (1993)[13] were unable to recover by simulations this state and any other state with $E < -37$. Although we could not reach $E = -42$ either without special tricks, we obtained at least $E = -40$ after about four hours of CPU time. Forbidding non-bonded HP neighbor pairs, we found the native state easily, but this cannot be really counted as a success since we knew in advance that such pairs do not appear in the native state. The difficulty posed by this sequence for PERM is obvious from Figure 5: there is no folding center in this chain. No matter where one starts the assembly, one first has to construct a large part of the boundary before the interior behind this boundary is filled up. This
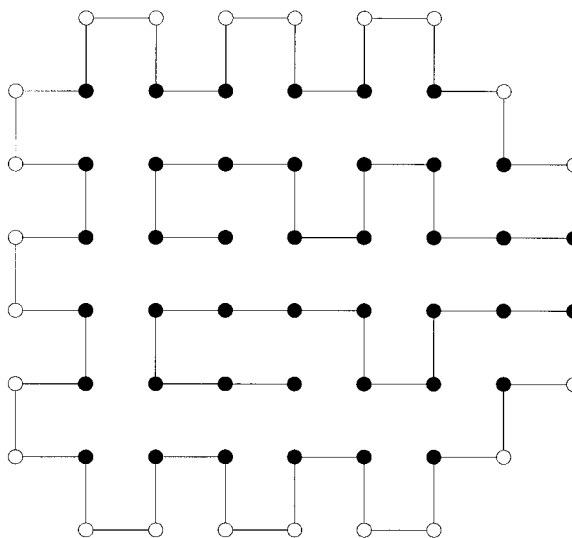

Fig. 5. One of the ground states of the $N = 64$ sequence of Unger and Moult (1993).[13] The other ground states differ in the configurations of the chain ends filling the interior of the structure, but have the same boundary and overall shape.

intermediate state has a very low Boltzmann weight and acts thus as a bottleneck for PERM.

Note that the configuration of Figure 5 cannot be obtained by popular local moves, including, e.g., bond exchange or crankshaft, or by reptation. But it can be reached if pivot moves are included. This illustrates that folding properties can depend strongly on the chosen kinetics.

In contrast to the sequence with $N = 64$, we had no problem with the shorter sequences of Unger and Moult (1993).[13] In particular, for the $N = 60$ sequence we found a configuration with $E = -36$ (see Table I), although the authors had quoted $E = -34$ as the supposedly exact ground-state energy.

For the two HP chains of Ramakrishnan et al. (1997)[28] with $N = 100$ (see Table I), we found several compact states (within about 40 hours of CPU time) that had energies lower than those of the compact putative ground states proposed in Ramakrishnan et al. (1997).[28] Figures 6 and 7 show representative compact structures with $E = -46$ for sequence 1 and $E = -47$ for sequence 2. Moreover, we found (again within 1–2 days of CPU time) several noncompact configurations with energies even lower: $E = -47$ and $E = -48$ for sequences 1 and 2, respectively. Forbidding nonbonded HP pairs, we obtained even $E = -49$ for sequence 2. Figures 8 and 9 show representative noncompact structures with these energies; a nonexhaustive collection of these is listed in Table I. These results reflect the well-known property that HP sequences (and those of other models) usually have ground states that are not maximally compact (see, e.g., Yue et al. (1997),[30]) although there is a persistent prejudice to the contrary.[24,28,45]
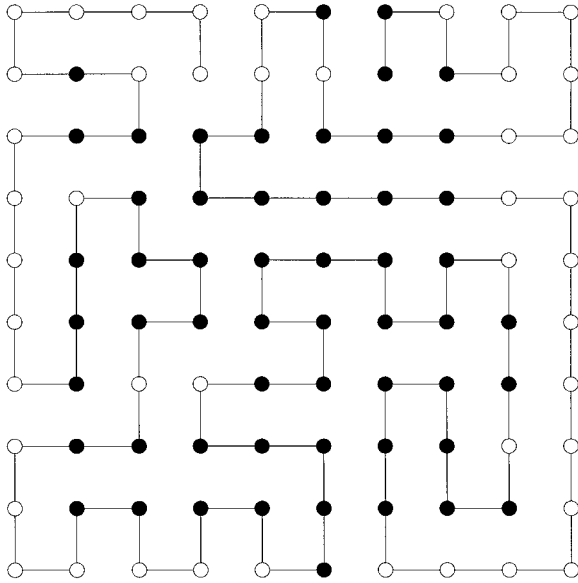
Fig. 6. One of the compact structures of sequence 1 with energy ($E = -46$) lower than the native state proposed by Ramakrishnan et al.[28]
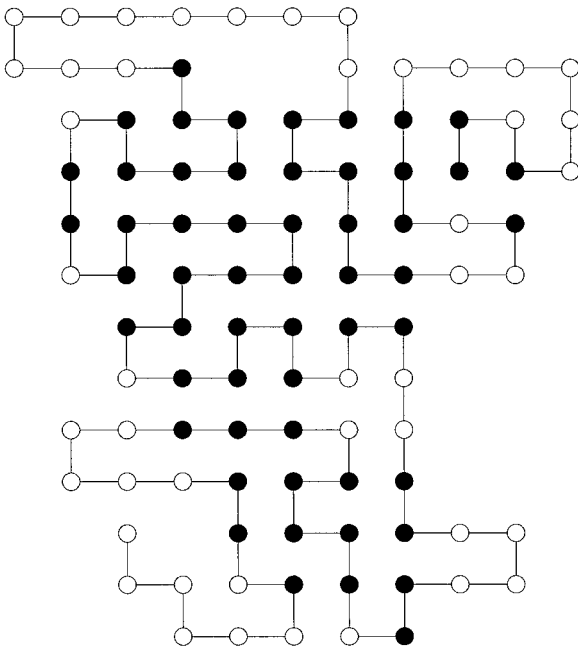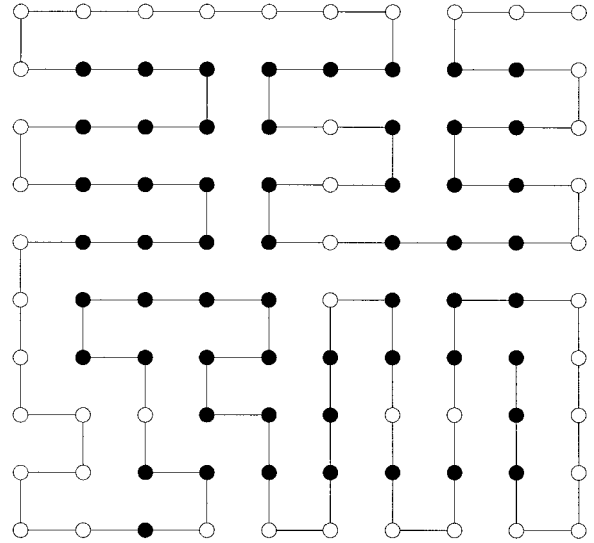

Fig. 7. One of the compact structures of sequence 2 with lower energy ($E = -47$).

Fig. 8. One of the (noncompact) lowest energy sequences for sequence 1 ($E = -47$).
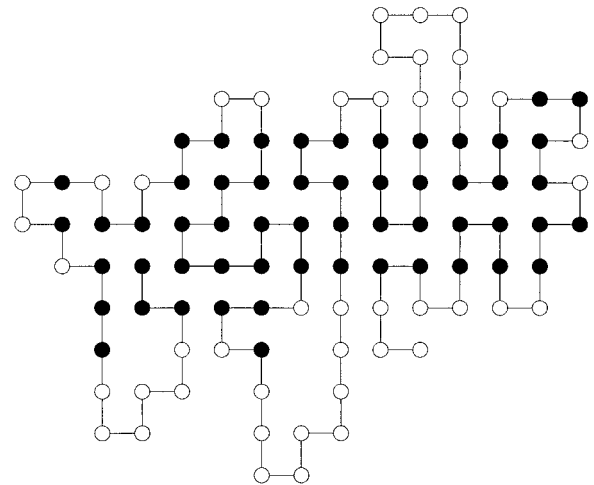
Fig. 9. One of the (noncompact) lowest energy sequences for sequence 2 ($E = -49$).

2D HAP Model

The ground state of the HAP triple helix was found within several minutes of CPU time using the PERM-Metropolis hybrid algorithm. It was not found to be degenerate.

In order to obtain information about the folding transition, energy and contact matrix‡ histograms were determined at $T = 1.25$, 2, and 4. Free energy differences, necessary for combining the histograms,[46] were determined using Bennet's acceptance ratio method.[47]

In Figure 10, the thermal behavior of the mean number of contacts, number of native contacts, and the mean energy is shown. While the structural transition to the ground-state phase is best monitored by the number of native contacts and takes
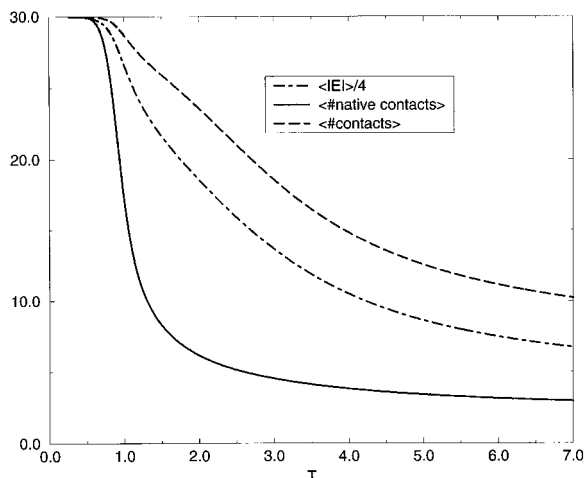
‡For a definition see Fig. 14.

Fig. 10.   Number of native contacts, total number of contacts, and energy E vs. temperature T for the HAP sequence of Figure 3.
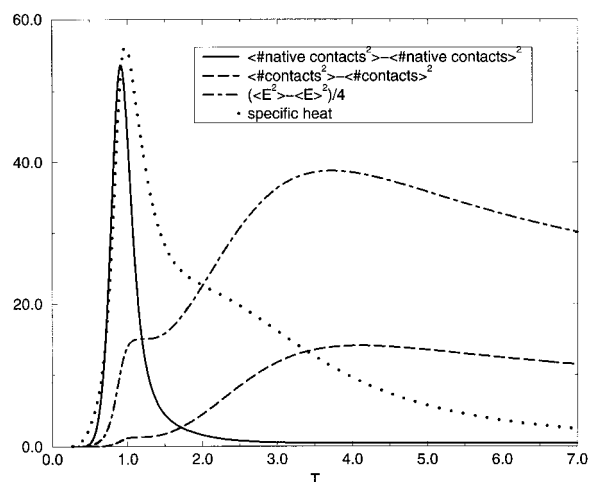

Fig. 11.   Fluctuations of the number of native contacts, total number of contacts, and energy E vs. temperature T for the HAP sequence of Figure 3. Since energy fluctuations and specific heat emphasize the polymer collapse and structural transition differently, we have included the specific heat, too.

place around $T = 1$, the compactification of the polymer chain is most clearly seen in the mean number of all contacts and in the radius of gyration (not shown here). It takes place already at much higher temperatures and is smeared out over a wide temperature range. Note that the number of all contacts follows closely the behavior of the energy.

These two transitions are seen more clearly when the fluctuations of the above order parameters are considered (see Fig. 11). Energy and contact number fluctuations exhibit a broad maximum around $T = 3.5$, while the structural transition is indicated by a narrow peak of the fluctuations of the number of native contacts and of the specific heat near $T = 1.0$. Note that specific heat and energy fluctuations emphasize the two transitions differently due to the factor T^2 by which they differ.

3D HP Model

With PERM we succeeded to reach ground states of the ten sequences of length $N = 48$ given in Yue et al. (1995)[30] in *all* cases, in CPU times between a few minutes and one day (see Table II). In these simulations we used a rather simple version of PERM, where we started assembly always from the same end of the chain. We found that the sequences most difficult to fold were also those which had resisted previous Monte Carlo attempts.[30] In those cases where a ground state was hit more than once, we verified also that the ground states were highly degenerate. In no case were there gaps between ground and first excited states (see Fig. 12). Therefore, none of these sequences is a good folder, though they were designed specifically for this purpose.

TABLE II. PERM Performance for Binary Sequences From Yve et al.[30]

Sequence number	$-E_{min}$[a]	$-E_{MC}$[b]	$n_{success}$[c]	CPU time, minutes
1	32	31	101	6.9
2	34	32	16	40.5
3	34	31	5	100.2
4	33	30	5	284.0
5	32	30	19	74.7
6	32	30	24	59.2
7	32	31	16	144.7
8	31	31	11	26.6
9	34	31	1	1420.0
10	33	33	16	18.3

[a]Ground-state energies.[30]
[b]Previously reached energies with Monte Carlo methods.[30]
[c]Number of independent tours in which a ground state was hit.

3D Modified HP Model

For the two-species 80-mer with interactions $(-1, 0, -1)$, even without much tuning our algorithm gave $E = -94$ after a few hours, but it did not stop there. After a number of rather disordered configurations with successively lower energies, the final candidate for the native state has $E = -98$. It again has a highly symmetric shape, although it does not fit into a $4 \times 4 \times 5$ box (see Fig. 13). It has twofold degeneracy (the central $2 \times 2 \times 2$ box in the front of Fig. 13 can be flipped), and both configurations were actually found in the simulations. Optimal parameters for the ground-state search in this model are $\beta = 1/kT \approx 2.0$, $a_{PP} = a_{HH} \approx 2$ and $a_{HP} \approx -0.13$. With these, average times for finding $E = -94$ and
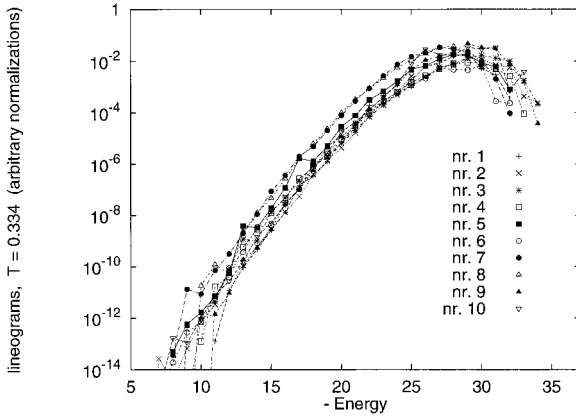
Fig. 12.   Energy spectrum of the 10 sequences given in Yue et al. (1995).[30] More precisely, to emphasize the low-energy part of the spectrum, we show the histograms obtained from the spectra by multiplying with $e^{E/T}$, where $T = 0.334$. Note that there are no energy gaps in any of these spectra.
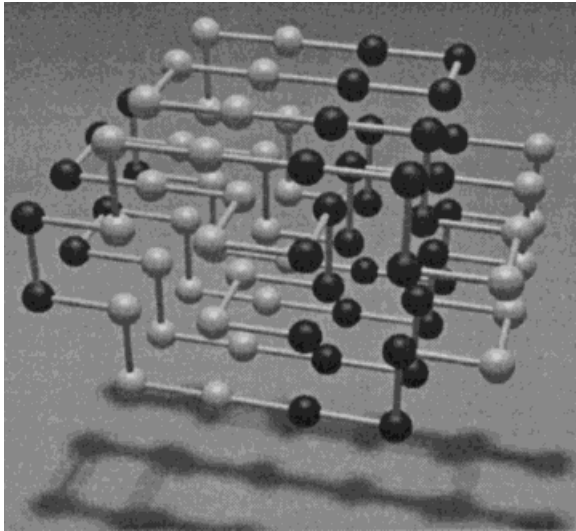

Fig. 13.   Conformation of the "four helix bundle" sequence with $E = -98$. We propose that this is the actual ground state. Its shape is highly symmetrical although it does not fit into a rectangular box. It is not degenerate except for a flipping of the central front $2 \times 2 \times 2$ box.

$E = -98$ in new tours are about 20 minutes and 80 hours, respectively.

A surprising result is that the monomers are arranged in four homogenous layers in Figure 13, while they had formed only three layers in the putative ground state of Figure 4. Since the interaction should favor the segregation of different type monomers, one might have guessed that a configuration with a smaller number of layers should be favored. We see that this is outweighed by the fact that both monomer types can form large double layers in the new configuration. Again, our new
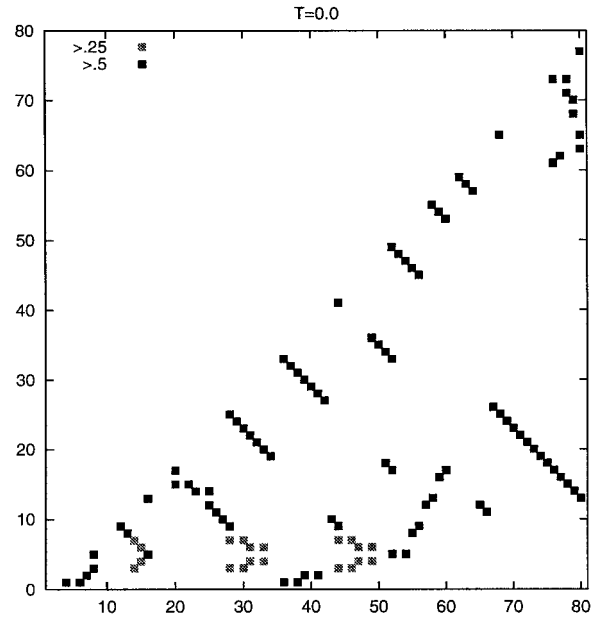

Fig. 14.   Contact matrix of the structure in Figure 13. A black point at (i, j) indicates that there is a contact between monomer i and monomer j; gray points indicate contacts in only one of the two native states, corresponding to the twofold degeneracy of the central $2 \times 2 \times 2$ box. Note that the lines orthogonal to the main diagonal correspond to antiparallel β sheet secondary-structure elements. See, e.g., Chan and Dill (1990).[48]

ground state is not "compact" in the sense of minimizing the surface, and hence it also disagrees with the widespread prejudice that native states are compact.

In terms of secondary structure, the new ground state is fundamentally different from the putative ground state of Ramakrishnan et al. (1997).[28] While the new structure (Fig. 13) is dominated by β sheets, which can most clearly be seen in the contact matrix (see Fig. 14), the structure in Figure 4 is dominated by helices (see also the corresponding contact matrix in Fig. 15).

In order to analyze the folding transition of this sequence, we again constructed histograms of the distribution of energy, end-to-end distance, and radius of gyration by combining the results obtained at various temperatures between $T = 0.45$ and 3. Figure 16 shows these distributions, reweighted so that it corresponds to $T = 0.75$. The thermal behavior of these order parameters as functions of T is obtained by Laplace transform, and is shown in Figure 17. The behavior of energy, end-to-end distance, and radius of gyration follow each other closely and exhibit clearly only the smeared-out collapse of the chain from a random coil to some unstructured compact phase. In contrast, the specific heat exhibits more structure: the shoulder around $T = 1$ corresponds again to the coil–globule collapse, but there are additional transitions seen around $T =$
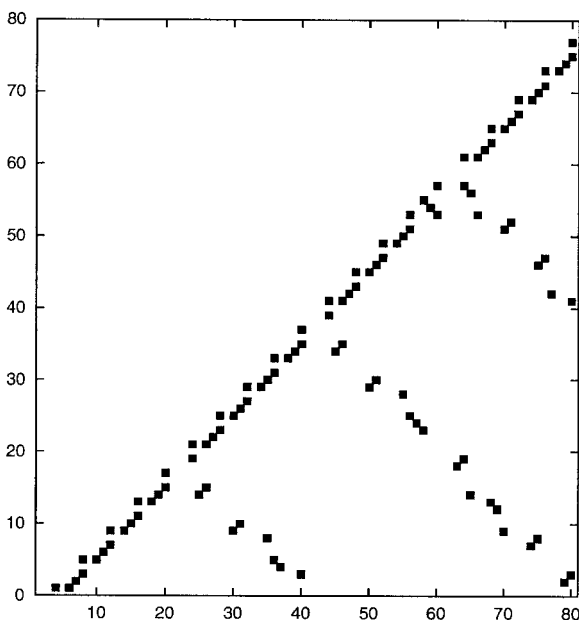
Fig. 15. For comparison, the contact matrix of the putative ground state of Ref. 28 in Figure 4. Note that point triples close to the diagonal parallel as well as orthogonal to it are signatures of 3D helical secondary-structure elements (see, e.g., Chan and Dill (1990).[48] The other points denote tertiary contacts between helices.

Fig. 16. Histograms of thermal weight (**top**), radius of gyration, R_G^2 (**middle**), and end-to-end distance, R_{ee}^2 (**bottom**), vs. energy E for the 80-mer "four helix bundle" at $T = 0.75$.

0.62 and $T = 0.45$. The last one is the transition to the β-sheet-dominated native phase. However, the transition at $T = 0.62$ is from an unstructured globule to an intermediate phase that is helix-dominated but exhibits strong tertiary fluctuations. These structural transitions are illustrated in Figures 18 to 22, where the thermally averaged contact matrices are shown for the respective phases.

The intermediate, helix-dominated phase is particularly interesting. It possesses some of the usual characteristics of a molten globule state[49]: compactness, large secondary-structure content (although not necessarily native), and strong fluctuations. This qualifies it as a candidate for a molten globule state, a phase that is absent in the folding transition of the HAP sequence.[§]

3D ∞ Monomer Types

For all sequences with $N = 27$ from Klimov and Thirumalai (1996),[27] we could reach the supposed ground-state energies within less than one hour. In

---

[§]In Deutsch (1997),[29] indications for the β sheet phase were found, too, and on this basis an analogy to the competition of structures in the prion problem (see, e.g., Nguyen et al. (1995)[50] was drawn. However, our results showing that the helical phase is strongly structurally disordered demonstrate that this analogy cannot be perfect. In a model for prions, both competing phases should be much more well defined.
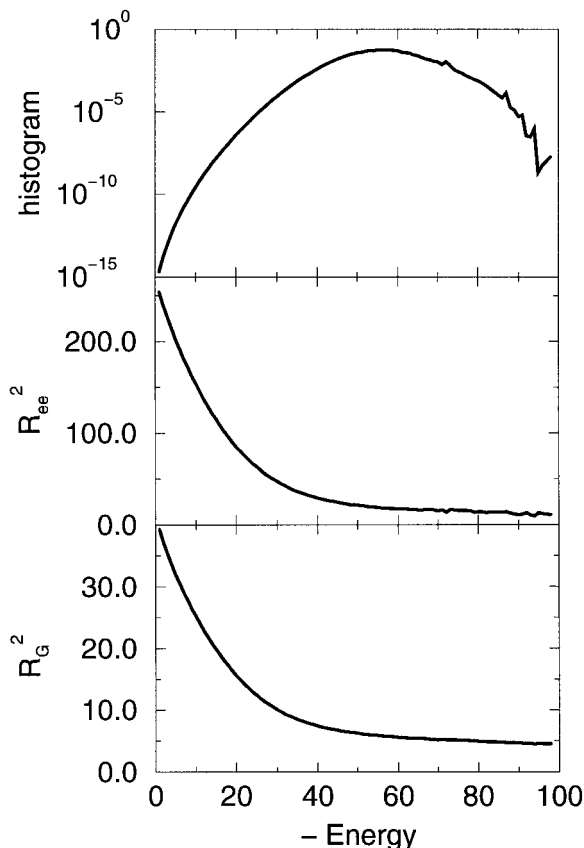
no case have we found energies lower than those quoted in Klimov and Thirumalai (1996),[27] and we verified also the energies of low-lying excited states given in Klimov and Thirumalai (1996).[27] Notice that these sequences were designed to be good folders by the authors of Klimov and Thirumalai (1996).[27] This time the design had obviously been successful, which is mainly due to the fact that the number of different monomer types is large. All sequences showed some gaps between the ground state and the bulk of low-lying states, although these gaps are not very pronounced in some cases.

More conspicuous than these gaps was another feature, namely, all low-lying excited states were very similar to the ground state, as measured by the fraction of contacts that existed also in the native configuration. Stated differently, if the gaps were not immediately obvious, this was because they were filled by configurations that were very similar to the ground state and can therefore easily transform into the native state and back. Such states therefore cannot prevent a sequence from being a good folder. For none of the sequences of Klimov and Thirumalai
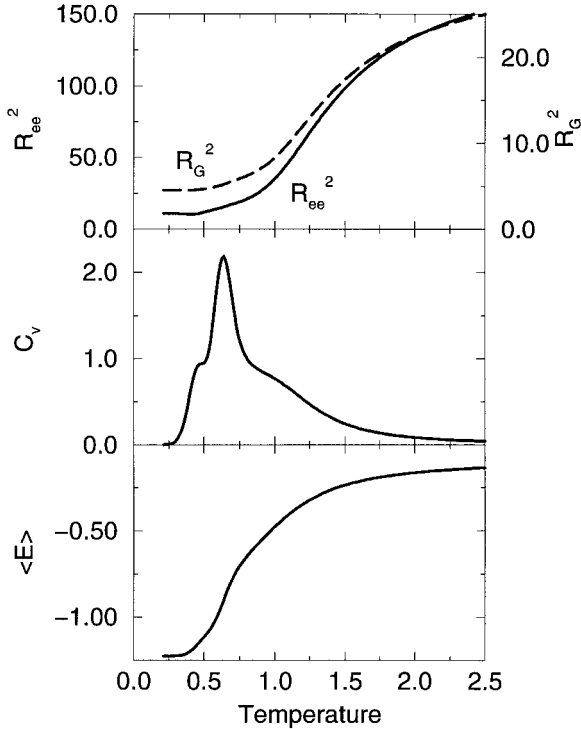
Fig. 17.   Average end-to-end distance, R_{ee}^2, and radius of gyration, R_G^2 (**top**); specific heat per monomer, C_v (**middle**); and average energy per monomer, $\langle E \rangle$ (**bottom**), vs. temperature T for the 80-mer four helix bundle.

have we found truly misfolded low-lying states with small overlap with the ground state.

Figure 23 illustrates this feature for one particular sequence. There we show the *overlap Q,* defined as the fraction by nonbonded nearest-neighbor ground-state contacts, which exist also in the excited state, against the excitation energy. For this and for each of the following figures, the 500 lowest-lying states were determined. We see no low-energy state with a small value of Q. To demonstrate that this is due to design, and is not a property of random sequences with the same potential distribution, we show in Figure 24 the analogous distribution for a random sequence.

To demonstrate that this difference is not merely due to a statistical fluctuation, we show in Figure 25 the distributions for 10 sequences from Klimov and Thirumalai (1996)[27] collected in a single plot. Since the ground-state energies differ considerably for different sequences, we used normalized excitation energies $(E - E_0)/E_0$ on the x-axis. Analogous results for 10 random sequences are shown in Figure 26. While there is no obvious correlation between Q and excitation energies for the random case, all low-energy states with small Q have been eliminated in
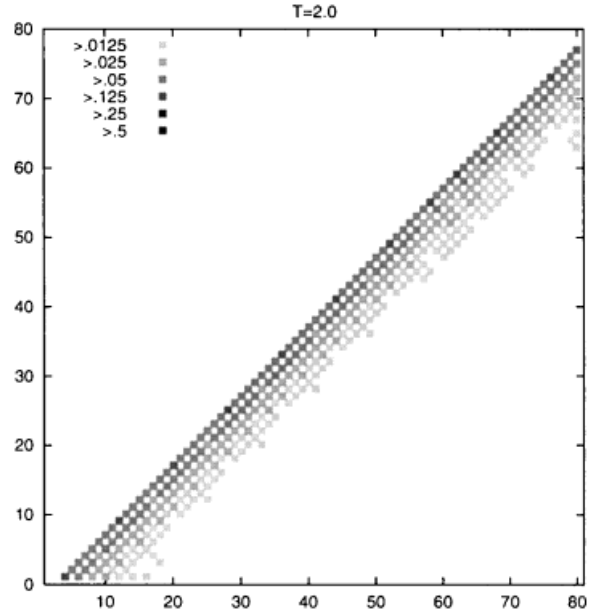

Fig. 18.   Thermally averaged contact matrix for the 80-mer four helix bundle in the random coil phase ($T = 2.0$). Different shades of grey denote different probabilities for the contact to exist.
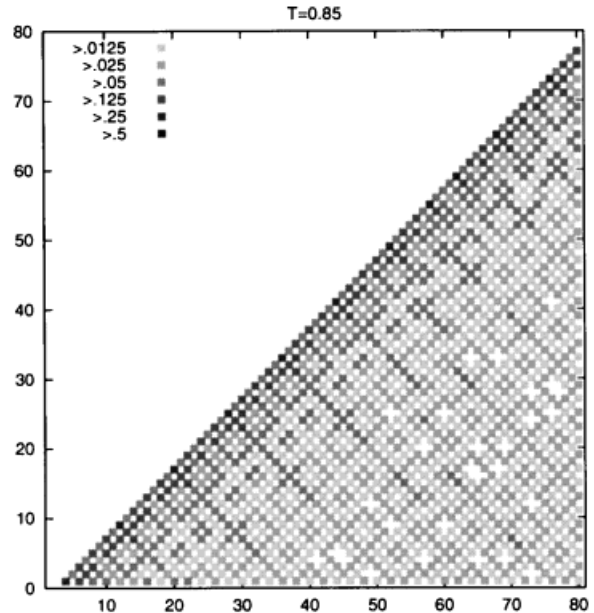

Fig. 19.   Thermally averaged contact matrix for the 80-mer four helix bundle in the collapsed but unstructured phase ($T = 0.85$).

the designed sequences (note the different ranges of Q in Figs. 25 and 26).

This elimination of truly misfolded low-energy states without elimination of native-like low-energy states might be an unphysical property of the design procedure used in Klimov and Thirumalai (1996),[27]
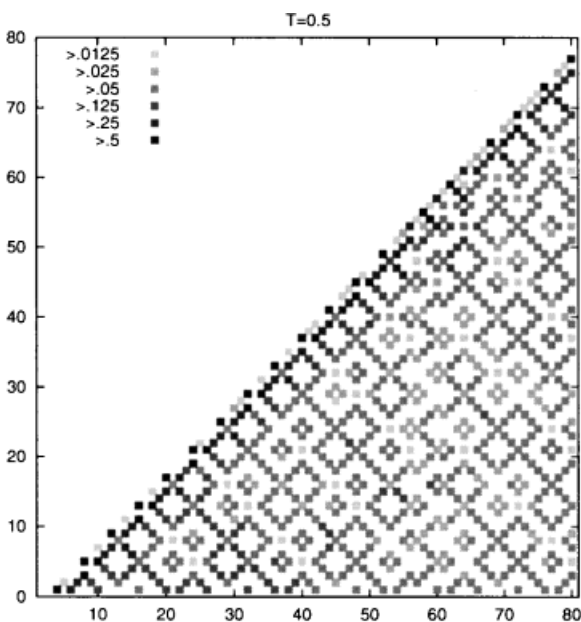
Fig. 20. Thermally averaged contact matrix for the 80-mer four helix bundle in the intermediate helix-dominated phase ($T = 0.5$).
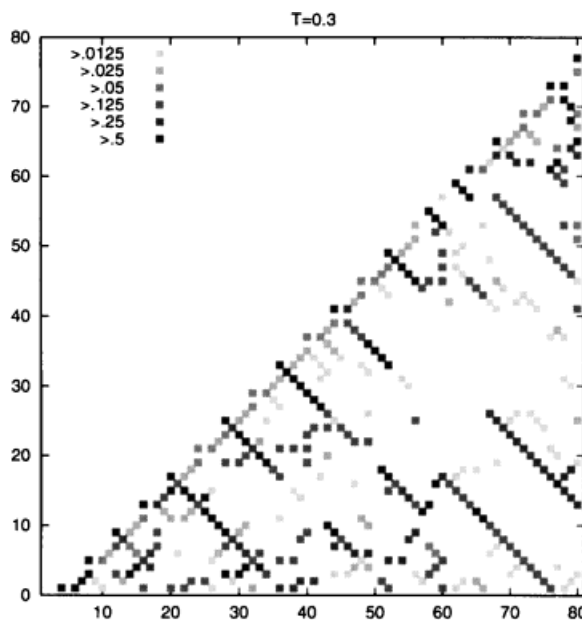

Fig. 22. Thermally averaged contact matrix for the 80-mer four helix bundle in the β-sheet-dominated phase ($T = 0.3$).
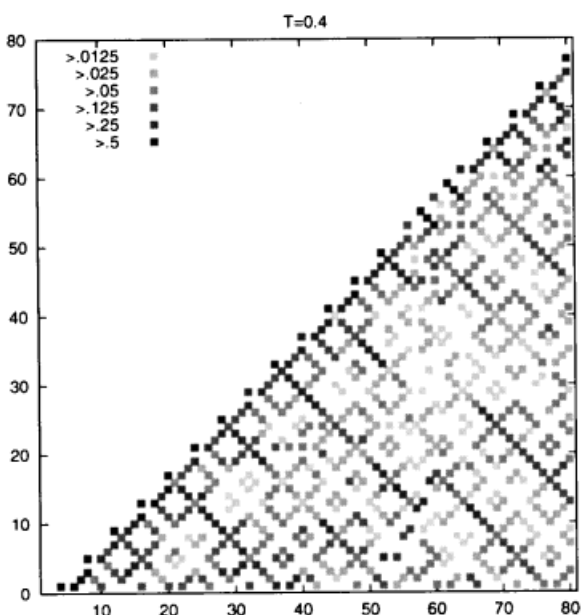

Fig. 21. Thermally averaged contact matrix for the 80-mer four helix bundle at the transition from the intermediate helix-dominated phase to the β-sheet-dominated phase ($T = 0.4$).
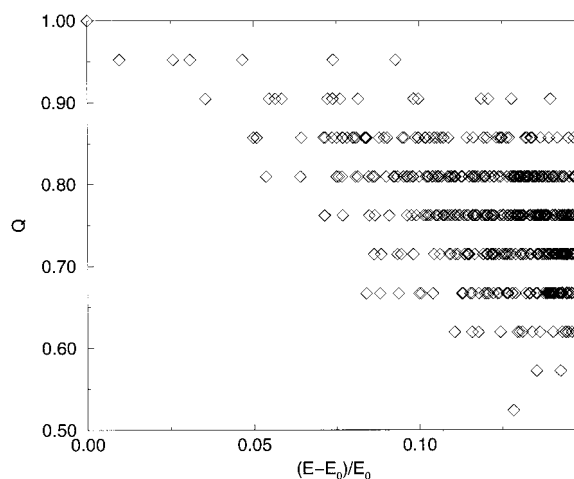

Fig. 23. Overlap with ground state, Q, vs. energy E of the lowest energy conformations for sequence 70 of Klimov and Thirumali (1996).[27]

but we do not believe that this is the case. Rather, it should be a general feature of any design procedure, including the one due to biological evolution. It contradicts the claim of Sali et al. (1994)[26] that it is only the gap between native and first excited state which determines foldicity. On the other hand, our results are consistent with the "funnel" scenario for the protein folding process,[51] where the folding pathway consists of states successively lower in energy and closer to the native state.

We note that for random sequences there are also excited states that have unit overlap with the native state, a feature not present in the folding sequences. These are cases where the native state has open loops and/or dangling ends, so that more compact conformations have all contacts of the native state,
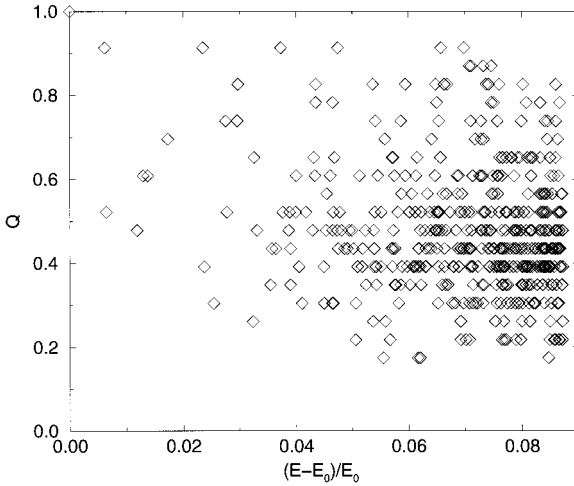
Fig. 24. Overlap with ground state, Q, vs. energy E of the lowest energy conformations for a single random sequence. Note that the range of Q is different from Fig. 23.
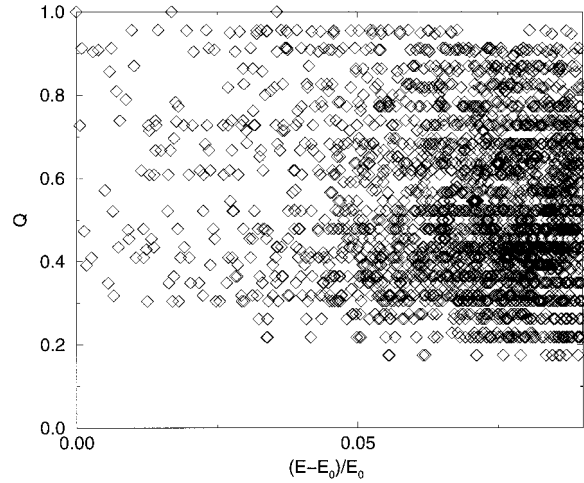

Fig. 26. Overlap with ground state, Q, vs. energy of the lowest energy conformations for 10 random sequences. For better visibility, the same symbol is used for all sequences. Note that the range of Q is different from Fig. 25.
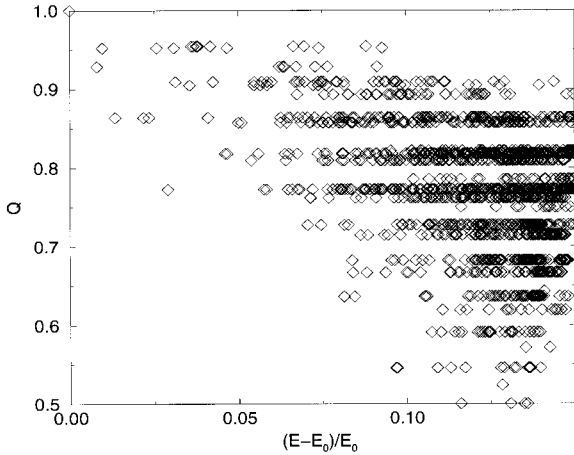

Fig. 25. Overlap with ground state, Q, vs. energy E of the lowest energy conformations for sequences 61 to 70 of Klimov and Thirumalai (1996).[27] For better visibility, the same symbol is used for all sequences.

and have, in addition, energetically unfavorable contacts resulting in a higher total energy.

DISCUSSION

We showed that the pruned-enriched Rosenbluth method (PERM) can be very effectively applied to protein structure prediction in simple lattice models. It is suited for calculating statistical properties and is very successful in finding native states. In all cases it did better than any previous MC method, and in several cases it found lower energy states than those which had previously been conjectured to be native.

We verified that ground states of the HP model are highly degenerate and have no gap, leading to bad folders. For sequences that are good folders we have established a funnel structure in state space: low-lying excited states of well-folding sequences have strong similarities to the ground state, while this is not true for nonfolders with otherwise similar properties.

In particular, we have presented a new candidate for the native configuration of a "four-helix bundle" sequence that had been studied before by several authors. The ground-state structure of the four-helix-bundle sequence, being actually β-sheet-dominated, differs strongly from the helix-dominated intermediate phase. This sequence, therefore, should not be a good folder.

Although we have considered only lattice models in this article, we should stress that this is not an inherent limitation of PERM. Straightforward extensions to off-lattice systems are possible and are efficient for homopolymers at relatively high temperatures.[17] Preliminary attempts to study off-lattice heteropolymers at low T have not yet been particularly successful, but the inherent flexibility of PERM suggests several modifications that have not yet been investigated in detail. One of them are hybrid PERM-Metropolis approaches similar to that used for the HAP model in the present study. Its success also suggests that similar hybrid approaches could be useful for models with more complicated monomers. Another improvement of PERM, which could be particularly useful for off-lattice simulations, might consist in more sophisticated algorithms for positioning the monomers when assembling the chain. Work along these lines is in progress, and we hope to report on it soon.

ACKNOWLEDGMENTS

REFERENCES

1. Gierasch, L.M., King, J. (eds.). "Protein Folding, Deciphering the Second Half of the Genetic Code." New York: AAAS, 1990.
2. Creighton, T.E. (ed.). "Protein Folding." New York: Freeman, 1992.
3. Merz, K.M., Jr., LeGrand, S.M. (eds.). "The Protein Folding Problem and Tertiary Structure Prediction." Boston: Birkhäuser, 1994.
4. Bohr, H., Brunak, S. (eds.). "Protein Folds: A Distance Based Approach." Boca Raton, FL: CRC Press, 1996.
5. Drexler, K.E. "Engines of Creation." New York: Anchor Books, 1986; available also on the web at the http://www.asiapac.com/EnginesOfCreation/.
6. Groß, M. "Expeditionen in den Nanokosmos." Basel: Birkhäuser, 1995.
7. Crippen, G.M., Maiorov, V.N. Contact potential for global identification of correct protein folding. In: "The Protein Folding Problem and Tertiary Structure Prediction." Merz, K.M., Jr., LeGrand, S.M. (eds.). Boston: Birkhäuser, 1994: 231–277.
8. Kolinski, A., Skolnick, J. "Lattice Models of Protein Folding, Dynamics and Thermodynamics," New York: Chapman & Hall, 1996.
9. Mirny, L.A., Shakhnovich, E.I. How to determine protein folding potentials? A new approach to the old problem. J. Mol. Biol. 264:1164, 1996.
10. Sali, A., Shakhnovich, E.I., Karplus, M. Kinetics of protein folding: A lattice model study of the requirements for folding to the native state. J. Mol. Biol. 235:1614, 1994.
11. Hansmann, U.H.E., Okamoto, Y. Prediction of peptide conformation by multicanonical algorithm: New approach to the multiple-minima problem. J. Comp. Chem. 14:1333, 1993; Comparative study of multicanonical and simulated annealed algorithms in the protein folding problem. Physica A212:415, 1994; Monte Carlo simulations in generalized ensemble: Multicanonical algorithm versus simulated tempering. Phys. Rev. E54:5863, 1996.
12. Wilson, S.R., Cui, W. Conformation searching using simulated annealing. In: "The Protein Folding Problem and Tertiary Structure Prediction." Merz, K.M., Jr., LeGrand, S.M. (eds.). Boston: Birkhäuser, 1994:43–70.
13. Unger, R., Moult, J. Genetic algorithms for protein folding simulations. J. Mol. Biol. 231:75, 1993.
14. LeGrand, S.M., Merz, K.M., Jr. The genetic algorithm and protein tertiary structure prediction. In: "The Protein Folding Problem and Tertiary Structure Prediction." Merz, K.M., Jr., LeGrand, S.M. (eds.). Boston: Birkhäuser, 1994:109–124.
15. Dill, K.A., Fiebig, K.M., Chan, H.S. Cooperativity in protein-folding kinetics. Proc. Natl. Acad. Sci. U.S.A. 90:1942, 1993.
16. Eisenhaber, F., Persson, B., Argos, P. Protein structure prediction–Recognition of primary, secondary and tertiary structural features from amino-acid sequence. Crit. Rev. Biochem. Mol. Biol. 30:1, 1995.
17. Grassberger, P. Pruned-enriched Rosenbluth method: Simulation of θ polymers of chain length up to 1,000,000. Phys. Rev. E, in press, 1997.
18. Rosenbluth, M.N., Rosenbluth, A.W. Monte Carlo calculation of the average extension of molecular chains. J. Chem. Phys. 23:256, 1955.
19. Frauenkron, H., Bastolla, U., Gerstner, E., Grassberger, P., Nadler, W. New Monte Carlo algorithm for protein folding. Phys. Rev. Lett. 80:3149, 1998.
20. Kantor, Y., Kardar, M. Collapse of randomly self-interacting polymers. Europhys. Lett. 28:169, 1994.
21. Dill, K.A. Theory for the folding and stability of globular proteins. Biochemistry 24:1501, 1985.
22. Lau, K.F., Dill, K.A. A lattice statistical mechanics model of the conformational and sequence spaces of proteins. Macromolecules 22:3986, 1989; Chan, H.S., Dill, K.A. "Sequence space soup" of proteins and copolymers. J. Chem. Phys. 95:3775, 1991.
23. Socci, N.D., Onuchic, J.N. Folding kinetics of proteinlike heteropolymrs. J. Chem. Phys. 101:1519, 1994.
24. O'Toole, E., Panagiotopoulos, A. Monte Carlo simulation of folding transitions of simple model proteins using a chain growth algorithm. J. Chem. Phys. 97:8644, 1992.
25. Krausche, T., Nadler, W. Protein folding in the HAP model.
26. Sali, A., Shakhnovich, E.I., Karplus, M. How does a protein fold? Nature 369:248, 1994.
27. Klimov, D.K., Thirumalai, D. Factors governing the foldability of proteins. Proteins 26:411, 1996; sequences are available from http://www.glue.umd.edu/~klimov.
28. Ramakrishnan, R., Ramachandran, B., Pekney, J.F. A dynamic Monte Carlo algorithm for exploration of dense conformational spaces in heteropolymers. J. Chem. Phys. 106:2418, 1997.
29. Deutsch, J.M. Long range moves for high density polymer simulations. J. Chem. Phys. 106:8849, 1997.
30. Yue, K., et al. A test of lattice protein folding algorithms. Proc. Natl. Acad. Sci. U.S.A. 92:325, 1995.
31. Shakhnovich, E.I., Gutin, A.M. Engineering of stable and fast-folding sequences of model proteins. Proc. Natl. Acad. Sci. U.S.A. 90:7195, 1993; Shakhnovich, E.I. Proteins with selected sequences fold into unique native conformation. Phys. Rev. Lett. 72:3907, 1994.
32. Yue, K., Dill, K.A. Forces of tertiary structural organization in globular proteins. Proc. Natl. Acad. Sci. U.S.A. 92:146, 1995.
33. Ebeling, M., Nadler, W. On constructing folding heteropolymers. Proc. Natl. Acad. Sci. U.S.A. 92:8798, 1995; Protein Folding: Optimized sequences obtained by simulated breeding in a minimalist model. Biopolymers 41:165, 1997.
34. Deutsch, J.M., Kurosky, T. New algorithm for protein design. Phys. Rev. Lett. 76:323, 1996.
35. Batoulis, J., Kremer, K. Statistical properties of biased sampling methods for long polymer chains. J. Phys. A21:127, 1988.
36. Garel, T., Orland, H. Guided replication of random chains: A new Monte Carlo method. J. Phys. A23:L621, 1990.
37. Velikson, B., Garel, T., Niel, J.-C., Orland, H., Smith, J.C. Conformational distribution of Heptaalanine: Analysis using a new Monte Carlo chain growth method. J. Comput. Chem. 13:1216, 1992.
38. Umrigar, C.J., Nightingale, M.P., Runge, K.J. A diffusion Monte Carlo algorithm with very small time step errors. J. Chem. Phys. 99:2865, 1993.
39. Wall, F.T., Erpenbeck, J.J. New method for the statistical computation of polymer dimensions. Statistical computation of radii of gyration and mean internal dimensions of polymer molecules. J. Chem. Phys. 30:634, 637, 1959.
40. Frauenkron, H., Grassberger, P. Critical unmixing of polymer solutions. preprint cond-mat/9707101, 1997.
41. Bastolla, U., Grassberger, P. Phase transitions of single semi-stiff polymer chains. preprint cond-mat/9705178, 1997.
42. Barkema, G.T., Bastolla, U., Grassberger, P. Two-dimensional oriented self-avoiding walks with parallel contacts. preprint cond-mat/9707312, 1997.
43. Matheson, R.R., Scheraga, H.A. A method for predicting nucleation sites for protein folding based on hydrophobic contacts. Macromolecules 11:819, 1978.
44. Finkelstein, A.V., Gutin, A.M., Badretdinov, A.Y. Why do protein architectures have Boltzmann-like statistics? Proteins 23:151, 1995.

45. Shakhnovich, E.I., Gutin, A.M. Enumeration of all compact conformations of copolymers with quenched disordered sequence of links. J. Chem. Phys. 93:5967, 1990; Gutin, A.M., Shakhnovich, E.I. Ground state of random copolymers and the discrete random energy model. J. Chem. Phys. 98:8174, 1993.

46. Ferrenberg, A.M., Swendsen, R.H. New Monte Carlo technique for studying phase transitions. Phys. Rev. Lett. 61:2635, 1988; Optimized Monte Carlo data analysis. Phys. Rev. Lett. 63:1195, 1989.

47. Bennet, C.H. Efficient estimation of free energy differences from Monte Carlo data. J. Comp. Phys. 22:245, 1976.

48. Chan, H.S., Dill, K.A. Compact polymers. Macromolecules 22:4559, 1989; The effects of internal constraints on the configurations of chain molecules. J. Chem. Phys. 92:3118, 1990.

49. Christensen, H., Pain, R.H. Molten globule intermediates and protein folding. Europ. Biophys. J. 19:221, 1991.

50. Nguyen, J. et al. Prion protein peptides induce α-helix to β-sheet conformation transitions. Biochemistry 34:4186, 1995.

51. Wolynes, P.G., Onuchic, J.N., Thirumalai, D. Navigating the folding routes. Science 267:1619, 1995.