LU TP 97-33
July 9, 1998
Revised version

Monte Carlo Procedure for Protein Design

Anders Irbäck[1], Carsten Peterson[2],
Frank Potthast[3] and Erik Sandelin[4]

Complex Systems Group, Department of Theoretical Physics
University of Lund, Sölvegatan 14A, S-223 62 Lund, Sweden
`http://thep.lu.se/tf2/complex/`

Submitted to *Physical Review* **E**

Abstract:

A new method for sequence optimization in protein models is presented. The approach, which has inherited its basic philosophy from recent work by Deutsch and Kurosky [*Phys. Rev. Lett.* **76**, 323 (1996)] by maximizing conditional probabilities rather than minimizing energy functions, is based upon a novel and very efficient multisequence Monte Carlo scheme. By construction, the method ensures that the designed sequences represent good folders thermodynamically. A bootstrap procedure for the sequence space search is devised making very large chains feasible. The algorithm is successfully explored on the two-dimensional HP model with chain lengths $N = 16$, 18 and 32.

PACS numbers: 87.15.By, 87.10.+e

[1]irback@thep.lu.se
[2]carsten@thep.lu.se
[3]frank@thep.lu.se
[4]erik@thep.lu.se

The "inverse" of protein folding, sequence optimization, is of utmost relevance in the context of drug design. This problem, which amounts to finding optimal amino acid sequences given a target structure, has also been investigated in the context of understanding folding properties of coarse-grained models for protein folding. Such models are described by energy functions $E(r,\sigma)$, where $r = \{\mathbf{r}_1, \mathbf{r}_2, .., \mathbf{r}_N\}$ denotes the amino acid coordinates and $\sigma = \{\sigma_1, \sigma_2, .., \sigma_N\}$ the amino acid sequence.

Good folding sequences fold fast and in a stable way into the desired target structure. A brute force search for sequences meeting these criteria is prohibitively time-consuming even in minimalist models for protein folding. Although it has been possible to apply this type of criteria to a simple helix-coil model [1], it is essential to find more efficient strategies. A fairly drastic simplification was proposed in [2], where the problem is approached by minimizing E with respect to σ with r clamped to the target structure, r_0. This method is very fast since no exploration of the conformational space is involved, but, unfortunately, it fails for a number of examples (see e.g. [1, 3, 4]). Recently, a more generic scheme was suggested [3], which aims at optimizing the conditional probability $P(r_0|\sigma)$, i.e. the Boltzmann weight, rather than $E(r_0,\sigma)$. This approach has the advantage that entropy effects are taken into account, but its usefulness is not obvious since maximizing $P(r_0|\sigma)$ is a non-trivial task. In fact, the calculations in [3] involved simplifying assumptions about both the form of $P(r_0|\sigma)$ and the conformational space. In this letter we present a practical Monte Carlo (MC) procedure for performing the maximization of $P(r_0|\sigma)$.

Thermodynamical characteristics for good folders are that the ground state minima are well separated from other states — at finite T the system spends a long time in the ground state well. In lattice models, where for relatively small chains the states are enumerable, this is often taken as non-degeneracy of ground state and that the latter is well separated from higher energy states. One expects that working with finite T distributions in the matching process singles out those optimal sequences that have good folding properties in terms of non-degeneracy. Indeed, in [3] when exploring the technique on lattice models, superior results are obtained when comparing with what was obtained in [2], where $E(r_0,\sigma)$ was minimized.

Computationally, straightforward MC approaches for maximizing $P(r_0|\sigma)$ are extremely tedious. Our novel MC methodology is based on the *multisequence method* [5], where both sequence and coordinate degrees of freedom are subject to simultaneous moves. The basic idea is to perform a single simulation of a joint probability distribution $P(r,\sigma)$ rather than repeated simulations of $P(r|\sigma)$ for different fixed σ. Hence, our approach is fundamentally different from that of [4].

Our method for maximizing $P(r|\sigma)$ is explored on the two-dimensional HP lattice model [6] with chain lengths N = 16, 18 and 32. For N=16 we study an example used in [4, 2, 3]. The results for both N=16 and 18 are checked against exact enumerations, whereas for N = 32 we use a target structure constructed "by hand". Our method reproduces the exact results extremely rapidly whenever comparisons are feasible. Furthermore, the method has quite some potential to deal with very large chains.

The problem of finding thermodynamically optimal sequences given a target structure r_0 is simple to formulate mathematically — maximize with respect to σ the conditional probability

$$P(r_0|\sigma) = \frac{1}{Z(\sigma)} \exp(-E(r_0,\sigma)/T) \tag{1}$$

$$Z(\sigma) = \sum_r \exp(-E(r,\sigma)/T) \tag{2}$$

where Eq. (1) can be rewritten in terms of the free energy $F(\sigma) = -T \ln Z(\sigma)$ as

$$P(r_0|\sigma) = \exp[-(E(r_0,\sigma) - F(\sigma))/T] \tag{3}$$

Hence, for each σ one needs to estimate $P(r_0|\sigma)$ which in turn involves a sum over all possible r. The situation is shown in Fig. 1, where the horizontal line represents the region probed in protein folding. In the simplified approach to the inverse problem [2], minimizing $E(r_0,\sigma)$, one works along the vertical line. Maximizing $P(r_0|\sigma)$ is a real challenge since it requires sampling of the entire (r,σ)-plane. Refs. [3, 4] approached this problem by using simulated annealing in sequence space.

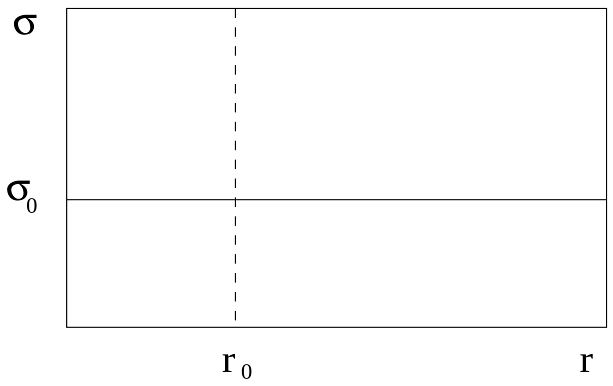

Figure 1: The (r,σ)-plane (see text).

The key difficulty then is to estimate the partition function $Z(\sigma)$ (Eq. (2)). In [3] this was done using the lowest-order term in the cumulant expansion of $F(\sigma)$. This approximation is valid at high temperature but it is unclear how good it is in the temperature regime of interest here. Ref. [4], on the other hand, used a chain growth MC method to estimate $Z(\sigma)$. In this case one has a nested MC, where the inner part by itself is far from trivial.

In [3, 4] these methods were successfully tested on examples where a simple minimization of $E(r_0,\sigma)$ along the vertical line in Fig. 1 fails. The chains were short enough for the results to be tested against exact enumerations. Another difference between optimizing $P(r_0|\sigma)$ versus $E(r_0,\sigma)$ is that the latter requires an optimization constrained to a preset net hydrophobicity.

In this letter we take a quite different path capitalizing on the multisequence method [5]. Here the basic strategy is to create an enlarged configuration space; the sequence σ becomes a dynamical variable. Hence, r and σ are put on a more equal footing, which, in particular, enables us to avoid a nested MC.

Our starting point is the joint probability distribution

$$P(r,\sigma) = \frac{1}{Z}\exp(-g(\sigma) - E(r,\sigma)/T) \tag{4}$$

$$Z = \sum_\sigma \exp(-g(\sigma))Z(\sigma) \tag{5}$$

where $\{g(\sigma)\}$ is a set of tunable parameters, which govern the marginal distribution

$$P(\sigma) = \sum_r P(r,\sigma) = \frac{1}{Z}\exp(-g(\sigma))Z(\sigma) \tag{6}$$

From the Bayes relation $P(r,\sigma) = P(r|\sigma)P(\sigma)$ one obtains the desired conditional probabilities, Eq. (1), which of course are independent of the choice of $g(\sigma)$.

The choice of $g(\sigma)$ is crucial for the efficiency of the method. At first sight, it may seem that one would need to estimate $Z(\sigma)$ in order to obtain reasonable $g(\sigma)$. However, a convenient choice is

$$g(\sigma) = -E(r_0,\sigma)/T \tag{7}$$

Maximizing $P(r_0|\sigma)$ in Eq. (1) corresponds to minimizing the quantity

$$\Delta F_0 = -T\ln P(r_0|\sigma) = E(r_0,\sigma) - F(\sigma) \tag{8}$$

which for the choice in Eq. (7) can be rewritten as

$$\Delta F_0 = T\ln P(\sigma) + T\ln Z \tag{9}$$

Hence, neglecting an unimportant constant, ΔF_0 can be obtained directly from the marginal distribution $P(\sigma)$.

The joint distribution $P(r,\sigma)$ can be simulated by using separate ordinary r and σ updates. This single simulation of $P(r,\sigma)$ replaces simulations of $P(r|\sigma)$ for a number of different fixed σ. This is quite convenient. However, it should be stressed that the major motivation for using this scheme is its efficiency. In fact, it was demonstrated in [5] that the simulation of $P(r,\sigma)$ can be much faster than the simulation of $P(r|\sigma)$ even for *a single* σ; the exploration of the conformational space becomes more efficient when the sequence is allowed to fluctuate.

The number of sequences that can be studied in a multisequence simulation is of course limited. It is therefore desirable to incorporate a step in which "bad" sequences are removed. This elimination step can be formulated in different ways. In our calculations a sequence σ is removed as soon as some structure $r \neq r_0$ is encountered for which $E(r,\sigma) \leq E(r_0,\sigma)$. This process is continued until the remaining sequences can be studied through a final multisequence run. If the elimination proceeds for a sufficiently long time, then the surviving sequences are, by construction, those that have the desired structure as their non-degenerate ground state. It should be pointed out that the elimination process serves two purposes. In addition to bringing down the number of sequences to a manageable level, it also tends to make the distribution $P(\sigma)$ more uniform, which is instrumental for the efficiency of the final multisequence run. Note that for a set of sequences all having r_0 as their unique ground state, one has $\exp(-g(\sigma))Z(\sigma) \approx 1$ to leading order at low T, independent of σ, which implies that $P(\sigma)$ becomes uniform in the zero temperature limit.

We have performed extensive numerical explorations on the HP model [6] for a variety of sizes and target structures and find that the approach consistently identifies the appropriate sequences in an efficient way. In this letter we report on results for $N = 16$ and 18, which have been previously used for evaluating design algorithms [3, 4]. Also results for $N = 32$ are presented in some detail.

The HP model [6] is defined by

$$E(r,\sigma) = -\sum_{i<j}\sigma_i\sigma_j\Delta(r_i - r_j) \tag{10}$$

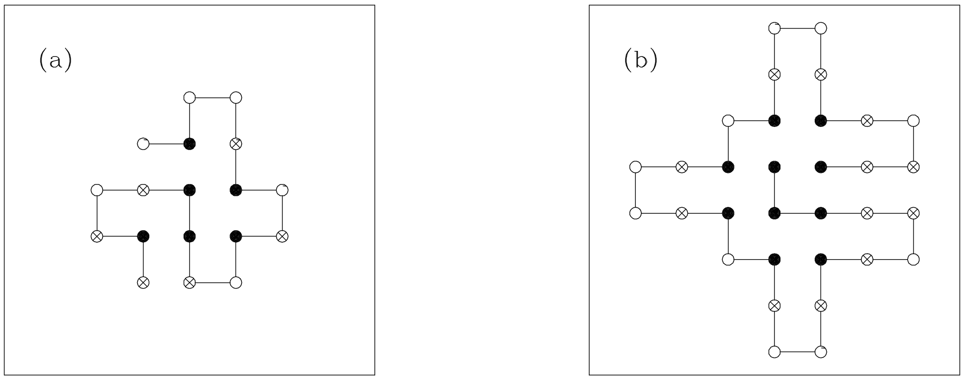

Figure 2: Target structures for (a) $N = 18$ and (b) $N = 32$. Symbols are explained in the text. The best sequence found for (b) is: `HHPPHHPPPPHPHPPPPHPHPPPPHHPPHHHH`.

where $\Delta(r_i - r_j) = 1$ if r_i and r_j are nearest neighbor monomers but non-adjacent along the chain and zero otherwise. Dependent upon whether σ_i is hydrophobic (H) or polar (P), one has σ_i=1 and 0, respectively. We work on the square lattice, for which it is known that sequences with non-degenerate ground states are not too rare [7]. All simulations are performed using standard Metropolis steps [8] in σ. In r we use three types of elementary moves: one-bead, two-bead and pivot [9, 10]. A sweep refers to a combination of these three move types followed by one attempt to update σ.

We first test our method for the $N = 16$ target structure studied in [4] (see Fig. 1 in [4]). There is one sequence which has this structure as its non-degenerate ground state, as can be shown by exact enumeration. It turns out that the design procedure in [4] is able to find this sequence, while the methods in [2, 3] fail to do so. Our calculation is carried out starting from the set of all 2^{16} possible sequences. After 8000 MC sweeps, corresponding to a few CPU seconds on a DEC Alpha 200, all sequences except the correct one have been removed.

Next, we turn to the $N = 18$ target structure shown in Fig. 2a. The seven sequences listed in Table 1 were found by the design procedure. Non-degeneracy is tested and confirmed as in the $N = 16$ case above. The multisequence simulation is then continued with only the seven sequences left in order to estimate $P(\sigma)$ and $P(r_0|\sigma)$. As can be seen from Table 1, the results agree very well with the exact results. Since all seven sequences have r_0 as their unique ground state, $P(\sigma)$ is constant for $T = 0$. At $T = 1/3$, the temperature studied here, the $P(\sigma)$'s are not perfectly equal, but similar enough to allow for good mobility in sequence space. In summary, our method has removed those sequences that do not have r_0 as their unique ground state, and it also provides $P(r_0|\sigma)$ for all the surviving sequences. Note that the remaining sequences all have the same monomer type at 12 of the 18 positions. These are marked by filled (H) and open (P) circles in Fig. 2a. In the second part of our simulations, where $P(r_0|\sigma)$ is estimated, it is clear that the stochastic sequence moves are essential. How useful these moves are in the first part, the elimination process, is less clear. To investigate this, we performed calculations both with and without these moves. In the latter case the simulated sequence is replaced only if it is to be removed from the simulation, and is then replaced by a randomly chosen sequence among the remaining ones. In Fig. 3 we show the number

Sequence	MC	exact
PHPPPHPPHPPHHPPPHP	0.2101(08)	0.2112
PHPPHHPPHPPHHPPPHP	0.0625(09)	0.0617
PHPPPHPPHPHHHPPPHP	0.3104(18)	0.3113
PHPPPHPPHPPHHHPPHP	0.0495(03)	0.0495
PHPPPHPHHPPHHHPPHP	0.1765(19)	0.1757
PHPPPHPPHPPHHHPHHP	0.1102(07)	0.1110
PHPPPHPPHPHHHHPPHH	0.0807(19)	0.0797

Table 1: $P(\sigma)$ for those seven $N = 18$ sequences that design the structure shown in Fig. 2a ($T = 1/3$). Listed are both the results from our multisequence simulation (MC) and the exact results, obtained by enumeration.

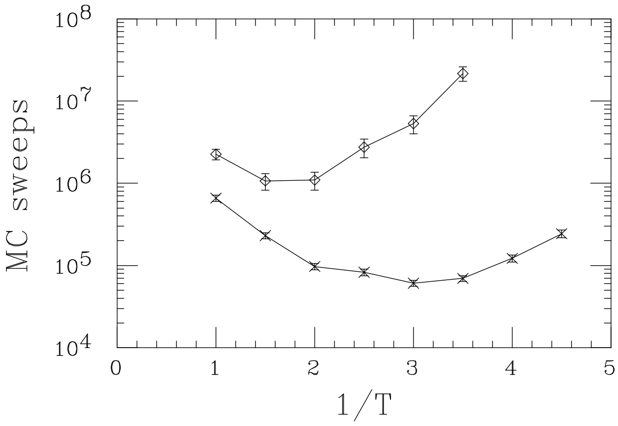

Figure 3: The number of MC sweeps needed to single out the seven sequences in Table 1. Each data point is an average over 50 experiments, each started from the full set of all 2^{18} sequences. Shown are both results obtained with ($\times$) and without ($\diamond$) stochastic sequence moves.

of MC sweeps needed to remove all sequences except those in Table 1, as a function of $1/T$. The results show that the required number of sweeps can be reduced by more than a factor of 10 by adding the stochastic sequence moves. Furthermore, the efficiency is less T dependent. The cost of the sequence moves is negligible.

We next turn to the $N = 32$ target structure shown in Fig. 2b, which is designed by hand since exhaustive enumeration is impracticable for this problem size. It is readily verified that this structure represents the minimum energy for the sequence with H at the filled circles and P at all the other positions along the chain (see Fig. 2b). As with any other method, it is for large chains not feasible to explore the entire sequence space with our multisequence method. However, as in the $N = 18$ example above, a given structure typically exhibits several positions where σ_i is effectively frozen to H or P (see e.g. [11, 12]). It turns out that such positions can be easily detected by means of a trial run. This leads us to a two-step bootstrap procedure.

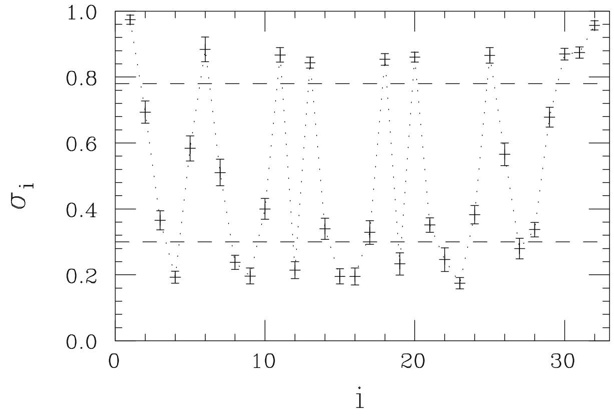

Figure 4: Average of σ_i against i for the surviving sequences from 10 runs, each started with a set of 10^5 random sequences ($N = 32$). The upper and lower lines represent $\sigma^{(1)}$ and $\sigma^{(2)}$, respectively (see text). The length of each run is 20000 MC sweeps (45 CPU seconds) and the number of surviving sequences varies between 15 and 57.

The first step amounts to picking sets of random sequences and gauging σ_i for the surviving sequences. The σ_i profile obtained this way for the target structure in Fig. 2b is shown in Fig. 4, from which it is clear that indeed many σ_i exhibit a clear preference for either P or H. Based on this, we divide the positions along the chain into three groups corresponding to $\sigma_i > \sigma^{(1)}$ (filled circles in Fig. 2b), $\sigma_i < \sigma^{(2)}$ (open circles) and $\sigma^{(2)} < \sigma_i < \sigma^{(1)}$ (crosses), as indicated in Fig. 4. We then rerun the algorithm with those σ_i in the first two groups clamped to H ($\sigma_i = 1$) and P ($\sigma_i = 0$), respectively, and those in the third group left open, which corresponds to a set of 2^{12} sequences. In 20 CPU minutes ($5 \cdot 10^6$ MC sweeps) this set was reduced to 200 sequences. These can be readily studied in a final multisequence run, and the best sequence found is given in Fig. 2. Stability was confirmed by repeating the procedure for different random seeds. We also performed runs where the elimination process was continued for much longer, which finally contained 167 surviving sequences. While it could be that this is still not the proper asymptotic value, we feel confident that the best sequence is correctly identified by our method.

It is clear that this method can be generalized to a corresponding multi-step procedure for very large chains.

In summary, we have developed a new efficient MC method for protein design by maximizing conditional probabilities using the multisequence method. The method circumvents calculations of partition functions by a judicious choice of the multisequence sample weights. Large chains are feasible with the approach by means of a bootstrap procedure that limits the search in sequence space. The method, which is successfully explored on two-dimensional lattice models, can easily be used in off-lattice models [13]. An alternative to prune σ-space by removing $E(r,\sigma) \le E(r_0,\sigma)$ sequences is to discard high $P(\sigma)$ sequences (see Eq. (9)).

References

[1] M. Ebeling and W. Nadler, *Proc. Natl. Acad. Sci. USA* **92**, 8798 (1995).

[2] E.I. Shakhnovich and A.M. Gutin, *Proc. Natl. Acad. Sci. USA* **90**, 7195 (1993).

[3] T. Kurosky and J.M. Deutsch, *J. Phys.* **A 27**, L387 (1995); J.M. Deutsch and T. Kurosky, *Phys. Rev. Lett.* **76**, 323 (1996)

[4] F. Seno, M. Vendruscolo, A. Maritan and J.R. Banavar, *Phys. Rev. Lett.* **77**, 1901 (1996).

[5] A. Irbäck and F. Potthast, *J. Chem. Phys.* **103**, 10298 (1995).

[6] K.F. Lau and K.A. Dill, *Macromolecules* **22**, 3986 (1989).

[7] K.A. Dill, S. Bromberg, K. Yue, K.M. Fiebig, P.D. Thomas and H.S. Chan, *Protein Sci.* **4**, 561 (1995).

[8] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth and A.H. Teller, *J. Chem. Phys.* **21**, 1087 (1953).

[9] M. Lal, *Molec. Phys.* **17**, 57 (1969).

[10] A.D. Sokal, in *Monte Carlo and Molecular Dynamics Simulations in Polymer Science*, ed. K. Binder (Oxford University Press, New York, 1995).

[11] K. Yue and K.A. Dill, *Proc. Natl. Acad. Sci. USA* **92**, 146 (1995).

[12] H. Li, R. Helling, C. Tang and N. Wingreen, *Science* **273**, 666 (1996).

[13] A. Irbäck, C. Peterson, F. Potthast and O. Sommelius, *J. Chem. Phys.* **107**, 273 (1997).