LU TP 99-38
December 17, 1999

Sequence Design in Coarse-Grained Protein Models

Anders Irbäck[*)]

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

To appear in *Proceedings of the 5th International Conference on Computational Physics*, Kanazawa, Japan, October 11–13, 1999

Abstract:
Designing amino acid sequences that are stable in a given target structure amounts to maximizing a conditional probability. A straightforward approach to accomplish this is a nested Monte Carlo where the conformation space is explored over and over again for different fixed sequences. In this paper we discuss an alternative Monte Carlo approach, multisequence design, where conformation and sequence degrees of freedom are simultaneously probed. The method is explored on hydrophobic/polar models. A statistical analysis of sequence correlations is also discussed. It is found that hydrophobic/polar model sequences and enzymes display hydrophobicity correlations that are qualitatively similar.

[*)] irback@thep.lu.se

§1. Introduction

The protein design problem amounts to finding an amino acid sequence given a target structure, which is stable in the target structure, and is able to fold fast into this structure. In a typical model the second, kinetic requirement implies that stability must set in at not too low a temperature. Hence, one is led to consider the problem of finding sequences that maximize the stability of the target structure at a given temperature. In a model described by an energy function $E(r, \sigma)$, where $r = \{\boldsymbol{r}_1, \boldsymbol{r}_2, .., \boldsymbol{r}_N\}$ denotes the conformation and $\sigma = \{\sigma_1, \sigma_2, .., \sigma_N\}$ the amino acid sequence, this can be expressed as maximizing the conditional probability

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

where r_0 denotes the target structure, T the temperature and the partition function $Z(\sigma)$ is given by

$$Z(\sigma) = \sum_r \exp[-E(r, \sigma)/T] \ . \tag{1·2}$$

Different ways of handling this sequence optimization problem have been proposed and partly explored in the context of coarse-grained (or minimalist) models [1)]. Some of these methods [2)-4)] are quite fast but approximate, and can therefore fail [5)]. In order to avoid introducing uncontrolled approximations, one is forced to turn to Monte Carlo (MC) methods. The most straightforward MC approach is to use simulated annealing in σ, and a normal fixed-σ MC in r for estimating (1·1). This gives a nested MC with a highly non-trivial inner part. Although correct results have been reported for toy-sized problems [5)], this approach is prohibitively CPU time-consuming for larger problem sizes. Another method that has been proposed [6)] is to use an analog of the Boltzmann machine learning equation. In this case, the σ_i are replaced by soft variables which evolve in a fictitious time t. It has been shown [6)] that this can be a more efficient strategy. Each step in t requires, however, a normal MC run in r.

In this paper we discuss an alternative MC methodology, multisequence design [7)]. This method does not attempt to maximize $P(r_0|\sigma)$ over all possible sequences. Rather it maximizes $P(r_0|\sigma)$ over a large but limited number of selected sequences. The advantage of this restricted search is that the calculations can be carried out with the sequence as a truly dynamical variable. In particular, this means that the use of a nested MC can be avoided.

The multisequence design method has been successfully applied [7)] to chains with up to $N = 50$ monomers in the HP model [8)] on the square lattice, and to $N = 20$ chains in a simple hydrophobic/polar three-dimensional off-lattice model. The examples discussed in this paper are from the minimalistic HP model. This model has two types of monomers, H (hydrophobic) and P (polar), and is defined by the energy function $E = -N_{\mathrm{HH}}$, where N_{HH} denotes the number of HH pairs that are nearest neighbors on the lattice but non-adjacent along the chain.

The last part of the paper deals with sequence correlations. In recent years important insights have been gained into the question of what characterizes sequences that show fast folding into a unique native state; by using coarse-grained models, physical characteristics of good folding sequences have been identified[9)-11)]. In this paper we discuss the results of a purely statistical analysis of "good" sequences in the HP model. All sequences with unique native states are considered as good, which means that the kinetic folding requirement is neglected. We expect this to be a reasonble approximation because almost all these sequences have the same energy gap. In the present model, it is well-known[12)] that the fraction of sequences with unique native states is a few per cent for small N. The question addressed is how these sequences differ statistically from random sequences.

Hydrophobicity correlations have been studied previously for both model and real proteins[13)-15)]. One method employed in these studies, and in studies of correlations in DNA[16)], is the blocking method. In this paper we briefly discuss some recent results[17)] obtained by this method for good HP sequences and for enzymes from the CATH database[18)]. The results for these two groups of sequences are contrasted with those reported by Khokhlov and Khalatur[19)] for designed copolymers with certain protein-like features.

§2. Sequence design

2.1. *The multisequence method*

A straightforward MC approach to the problem of maximizing the conditional probability (1·1) is to use simulated annealing in σ and a normal fixed-σ MC in r. This gives a nested MC which is prohibitively time-consuming except for very small systems.

The multisequence method offers a fundamentally different approach. In this method one replaces the simulations of $P(r|\sigma)$ for different fixed σ by a single simulation of the joint probability distribution

$$P(r,\sigma) = \frac{1}{Z}\exp[-g(\sigma) - E(r,\sigma)/T] \ , \tag{2·1}$$

where

$$Z = \sum_{\sigma}\exp[-g(\sigma)]Z(\sigma) \ . \tag{2·2}$$

The simulation of this joint distribution can be carried out by using alternated ordinary

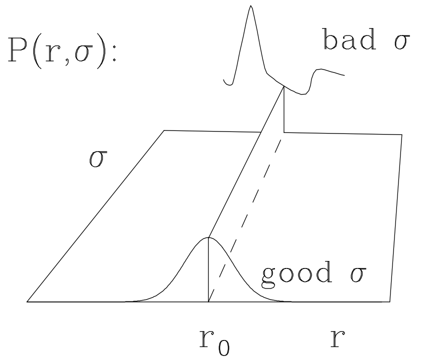

Fig. 1. The distribution $P(r,\sigma)$. The choice (2·4) makes $P(r_0,\sigma)$ flat in σ. A sequence having r_0 as its unique ground state has a unique maximum at $r = r_0$ in $P(r|\sigma)$, which for low T contains most of the probability.

updates of r and σ. The parameters $g(\sigma)$ determine the marginal distribution

$$P(\sigma) = \frac{1}{Z}\exp[-g(\sigma)]Z(\sigma) \tag{2·3}$$

and must therefore be chosen carefully. 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

$$g(\sigma) = -E(r_0,\sigma)/T\ , \tag{2·4}$$

for which one has

$$P(r_0|\sigma) = \frac{P(r_0,\sigma)}{P(\sigma)} = \frac{1}{ZP(\sigma)}\ . \tag{2·5}$$

In other words, maximizing the conditional probability $P(r_0|\sigma)$ is, for this choice of $g(\sigma)$, equivalent to minimizing the marginal probability $P(\sigma)$. The situation is illustrated in Fig. 1.

The fact that bad sequences are visited more frequently than good ones in the simulation may seem strange at a first glance, but can be used to eliminate bad sequences. In our calculations we have used two different methods for elimination of bad sequences, which are referred to as $P(\sigma)$- and $E(r,\sigma)$-based elimination, respectively.

In $P(\sigma)$-based elimination, which can be applied to lattice as well as off-lattice chains, those σ with highest $P(\sigma)$ are removed. A new multisequence run is then started for the remaining sequences. Obviously, this step can be repeated.

In $E(r,\sigma)$-based elimination, which is practical for lattice models only, one removes sequences that do not have the target structure r_0 as their ground state. For each conformation $r \neq r_0$ visited in the simulation, it is checked, for each remaining sequence σ, whether $E(r,\sigma) \leq E(r_0,\sigma)$. Those σ for which this is true are removed. With this method it may happen that one removes the sequence with highest $P(r_0|\sigma)$, but this should not be viewed as a shortcoming. If it happens, it rather means that the temperature studied is too high.

The multisequence method is a dynamical-parameter algorithm of the same type as the original simulated-tempering method[20], in which the temperature is dynamical. Simulated tempering is used below to check design results for $N = 50$. It should be stressed that for this system size the verification is a non-trivial task, which requires an efficient algorithm such as simulated tempering, the multicanonical method[21] or the multi-self-overlap ensemble method[22].

2.2. *Restricted search*

Before the multisequence simulation can be started, it is necessary, except for very short chains, to restrict the sequence set. The question then is how to do that without losing those typically very few sequences that design the desired target structure. In the HP model this turns out to be relatively easy; a simple but useful strategy is to identify σ_i with a strong preference for either H or P, and then clamp these degrees of freedom.

Let us consider a simple example[17] of such a clamping scheme. The rule is in this case to clamp all sites with two or three non-trivial nearest neighbors ("core" sites) to H, and all those with zero such neighbors to P. The remaining sites are left open. Let N_t denote the total number of such sequences, and let N_c be the number of these that actually design the target structure (in the sense that they have this structure as their unique ground state). What is then the probability that $N_c > 0$, and what is the typical fraction of correct sequences, N_c/N_t? We checked this for $N = 18$, by calculating N_c and N_t for each of the 1475 structures with designability $N_r > 0$ (the designability N_r is defined as the number of sequences that design the given structure). By complete enumerations, it was found that $N_c > 0$ for 91% of these structures, and that N_c/N_t is 0.09 on average. For designabilities $N_r \geq 10$, the fraction with $N_c > 0$ increases to 100%, and the average N_c/N_t to 0.31. This example shows that, at least for this system size, clamping selected σ_i is a useful way to restrict the sequence set, especially for highly designable structures.

An alternative way to decide which σ_i to clamp is to use some short trial runs[7], each started from a set of random sequences. Sequences are then removed using one of the two methods described earlier. For the surviving sequences, the average hydrophobicity profile $\{h_i\}$ is calculated, and sites with high or low h_i are clamped to H and P, respectively.

2.3. *Results*

The multisequence design method has been tested[7] on the two-dimensional HP lattice model and a hydrophobic/polar three-dimensional off-lattice model. In this section we give a brief description of the calculations for the largest target structure studied in the HP model, the $N = 50$ structure shown in Fig. 2.

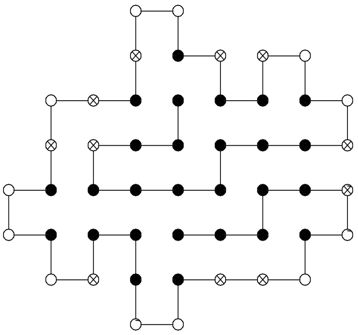

Fig. 2. Target structure with $N = 50$ monomers. Based on trial runs, 27 sites were clamped to H (filled circles) and 12 to P (open circles). The 11 remaining sites (crosses) were left open.

We began the sequence design for this structure by performing ten short runs, each started from 10^5 random sequences. Based on these, 27 σ_i were clamped to H and 12 to P, as indicated in Fig. 2. Having restricted the search this way to 2^{11} sequences, we proceeded in two steps. First, we performed a run with $E(r, \sigma)$-based elimination. The number of sequences surviving this run was 832, indicating that there are quite a few sequences that design this structure. The second step was a run with $P(\sigma)$-based elimination. This step was repeated three times using different random number seeds, each time starting from the same 832 sequences. The best sequence was the same in all the three runs. This sequence has four H and seven P at the sites left open after clamping. The four H positions are $i = 10$, 11, 18 and 28 ($i = 1$ corresponds to the lower of the two end points in Fig. 2).

This designed $N = 50$ sequence was checked by means of a long, independent simulated-tempering run. In fact, more CPU time was spent on this run than on the design itself. In this run there were about 30 "independent" visits of the target structure, whereas no other structure with the same or lower energy was encountered. We take this as strong evidence that this sequence indeed has the target structure as its unique ground state.

How much of a difference does it make to use the multisequence method rather than a normal fixed-σ MC in the final calculation of $P(r_0|\sigma)$? This was carefully tested for a $N = 32$ target structure [7)]. First, a multisequence simulation of the 180 remaining sequences was performed. Additional fixed-σ runs, with the same r update, were then carried out for three of these sequences. Each of these three runs required approximately the same amount of CPU time as the multisequence run. The statistical errors from the multisequence run turned out to be smaller than those from the fixed-σ runs. Simulating 180 sequences with the multisequence method was, in other words, faster than simulating a single sequence with fixed-σ MC.

2.4. *Comments*

Although the precise number of sequences that design this $N = 50$ structure is unknown, it is clear that these sequences constitute a very tiny fraction of all possible 2^{50} sequences. At a first glance, the sequence design problem may therefore seem like a hopeless task, but this is, as the multisequence calculation shows, not the case.

In the sequence design problem it is implicitly assumed that the given structure is designable, which in our example turned out to be true. However, structures vary widely in designability [23)], which makes finding designable structures a highly relevant problem. Whether the multisequence design method is useful in this context too is unclear; it can be used to scan random structures for designable ones, but the feasibility of this approach is far from obvious. This is currently being investigated [17)] for HP chains with $N = 25$ and 36.

§3. Hydrophobicity correlations

3.1. *The blocking method*

Forming block variables is a widely used technique for studying correlations. In this section the blocking method is applied to binary hydrophobicity strings $\{\sigma_i\}$, where $\sigma_i = 1$ and -1 correspond respectively to hydrophobic and hydrophilic amino acids.

The block variable

$$\sigma_k^{(s)} = \sum_{i \in \text{block } k} \sigma_i \qquad (k = 1, \ldots, N/s) \tag{3·1}$$

is the sum of s consecutive σ_i along the chain. A useful quantity is the mean-square fluctuation

$$\psi^{(s)} = \frac{1}{K}\frac{s}{N}\sum_{k=1}^{N/s}\left(\sigma_k^{(s)} - \frac{s}{N}\sum_{k'=1}^{N/s}\sigma_{k'}^{(s)}\right)^2 \ , \tag{3·2}$$

where K is a normalization factor. A convenient choice is

$$K = \frac{N^2 - M^2}{N^2 - N}(1 - s/N) \ , \tag{3·3}$$

where M denotes the total hydrophobicity of the sequence,

$$M = \sum_{i=1}^{N} \sigma_i \ . \tag{3·4}$$

The average of $\psi^{(s)}$ over all possible sequences with the same N and M, denoted by $\psi_0^{(s)}$, can be easily computed. With the choice (3·3), it takes the simple form

$$\psi_0^{(s)} = s \ , \tag{3·5}$$

independent of N and M.

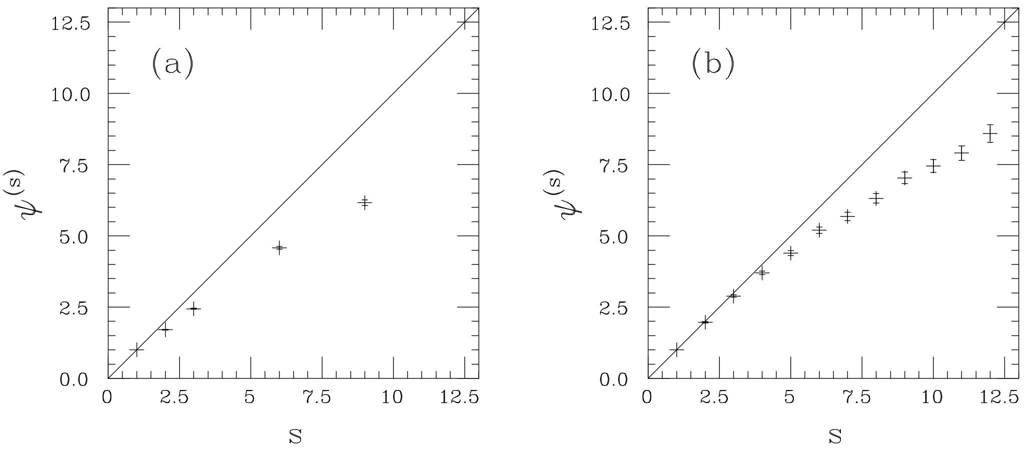

Fig. 3. The average $\psi^{(s)}$ against s for (a) good $N = 18$ sequences in the HP model and (b) enzymes. The line represents random sequences, see (3·5).

3.2. *Results*

The behavior of the mean-square block fluctuation $\psi^{(s)}$ has previously been studied [15] for sequences with good folding properties in a two-dimensional hydrophobic/polar off-lattice model. To test the model dependence of these results, which were based on a fairly small sample of 37 sequences, the same analysis was repeated [17] for the HP model.

In Fig. 3a we show the average $\psi^{(s)}$ against s for good $N = 18$ sequences in the HP model. By exhaustive enumeration, it can be shown that there are 6349 such sequences. We see that the data points fall well below the line representing random sequences, (3·5). The conclusion that $\psi^{(s)}$ is suppressed for good sequences is in perfect agreement with the results obtained in the previous model.

The blocking method provides information about the distribution of hydrophobicity along the chains. One may also ask about the distribution of total hydrophobicity, M. It turns out that the variance of M is smaller for good $N = 18$ HP sequences than for random sequences of the same length [17]. So fluctuations in M are, like block fluctuations, suppressed for the model sequences.

We now turn to real proteins. Using binary hydrophobicity assignments ($\sigma_i = 1$ for Leu, Ile, Val, Phe, Met and Trp, and $\sigma_i = -1$ for the others), the blocking method was applied to the 173 non-homologous single domain enzymes found in the October 1998 release of the CATH database [18]. A similar analysis had previously been performed for general proteins [15], without any restriction on protein type. In this earlier study the sequences were divided into groups corresponding to different (normalized) total hydrophobicities M. For the largest group, corresponding to typical values of M, the average $\psi^{(s)}$ was found to be suppressed, whereas the behavior turned out to be the opposite for sequences from the tails of the M

distribution. The new analysis of enzymes[17] was performed without any restriction on M. The results of this analysis are shown in Fig. 3b. We see that the behavior of the enzymes is qualitatively similar to that found for the model sequences (Fig. 3a) and for proteins with typical M.

3.3. *Comments*

It is interesting to compare these results to those for copolymers generated by the method of Khokhlov and Khalatur[19]. These sequences are, for instance, not intended to have unique native states. Nevertheless, they show certain protein-like thermodynamic properties[19]. However, they differ markedly from the sequences studied in this paper in terms of statistical properties. For example, it has been shown[19] that the average lengths of hydrophobic and hydrophilic segments are larger for these sequences than for random sequences, which corresponds to enhanced rather than suppressed block fluctuations.

Acknowledgements

The results presented here were obtained in collaboration with Carsten Peterson, Frank Potthast and Erik Sandelin. This work was supported by the Swedish Foundation for Strategic Research.

References

1) For a review, see E.I. Shakhnovich, Fold. Des. **3** (1998), R45.
2) E.I. Shakhnovich and A.M. Gutin, Protein Eng. **6** (1993), 793; Proc. Natl. Acad. Sci. USA **90** (1993), 7195.
   E.I. Shakhnovich, Phys. Rev. Lett. **72** (1994), 3907.
3) T. Kurosky and J.M. Deutsch, J. Phys. **A27** (1995), L387; Phys. Rev. Lett. **76** (1996), 323.
4) M.P. Morrissey and E.I. Shakhnovich, Fold. Des. **1** (1996), 391.
5) F. Seno, M. Vendruscolo, A. Maritan and J.R. Banavar, Phys. Rev. Lett. **77** (1996), 1901.
6) Y. Iba, K. Tokita and M. Kikuchi, J. Phys. Soc. Jpn. **67** (1998), 3985.
7) A. Irbäck, C. Peterson, F. Potthast and E. Sandelin, Phys. Rev. **E58** (1998), R5249; Structure **7** (1999), 347.
8) K.F. Lau and K.A. Dill, Macromolecules **22** (1989), 3986.
9) A. Săli, E. Shakhnovich and M. Karplus, J. Mol. Biol. **235** (1994), 1614.

10) J.D. Bryngelson, J.N. Onuchic, N.D. Socci and P.G. Wolynes, Proteins: Struct., Funct., Genet. **21** (1995), 167.

11) D.K. Klimov and D. Thirumalai, J. Chem. Phys. **109** (1998), 4119.

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

13) S.H. White and R.E. Jacobs, Biophys. J. **57** (1990), 911.

14) V.S. Pande, A.Y. Grosberg and T. Tanaka, Proc. Natl. Acad. Sci. USA **91** (1994), 12972.

15) A. Irbäck, C. Peterson and F. Potthast, Proc. Natl. Acad. Sci. USA **93** (1996), 9533; Phys. Rev. **E55** (1997), 860.

16) C.-K. Peng, S.V. Buldyrev, A.L. Goldberger, S. Havlin, F. Sciortino, M. Simons and H.E. Stanley, Nature **356** (1992), 168.

17) A. Irbäck and E. Sandelin, Lund preprint LU TP 99-40 (1999).

18) C.A. Orengo, A.D. Michie, S. Jones, D.T. Jones, M.B. Swindells and J.M. Thornton, Structure **5** (1997), 1093.

19) A.R. Khokhlov and P.G. Khalatur, Physica **A249** (1998), 253; Phys. Rev. Lett. **82** (1999), 3456.

20) A.P. Lyubartsev, A.A. Martsinovski, S.V. Shevkunov and P.N. Vorontsov-Velyaminov, J. Chem. Phys. **96** (1992), 1776.
E. Marinari and G. Parisi, Europhys. Lett. **19** (1992), 451.
A. Irbäck and F. Potthast, J. Chem. Phys. **103** (1995), 10298.

21) B.A. Berg and T. Neuhaus, Phys. Rev. Lett. **68** (1992), 9.
U.H.E. Hansmann and Y. Okamoto, J. Comp. Chem. **14** (1993), 1333.

22) G. Chikenji, M. Kikuchi and Y. Iba, Phys. Rev. Lett. **83** (1999), 1886.

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