Design Equation: A Novel Approach to Heteropolymer Design

Yukito IBA[1*], Kei TOKITA [2,3] and Macoto KIKUCHI[3**]

[1] *The Institute of Statistical Mathematics,*
4-6-7 Minami-Azabu Minato-ku , Tokyo, 106-8569, Japan
[2] *Department of Chemistry and Chemical Biology,*
Harvard University, 12 Oxford Street, Cambridge MA 02138, USA
[3] *Department of Physics, Osaka University, Toyonaka, 560-0043, Japan*

(Received August 7, 1998)

A novel approach to heteropolymer design is proposed. It is based on the criterion by Kurosky and Deutsch, with which the probability of a target conformation in a conformation space is maximized at low but finite temperature. The key feature of the proposed approach is the use of soft spins (fuzzy monomers) that leads to a design equation, which is an analog of the Boltzmann machine learning equation in the design problem. We implement an algorithm based on the design equation for the generalized HP model on the $3 \times 3 \times 3$ cubic lattice and check its performance.

KEYWORDS: design, heteropolymer, protein, optimization, learning, MTP criterion, Boltzmann machine, design equation, HP model

§1. Introduction

Recently, computer design of microscopic objects draws much attention of theoretical physicists. Designing a heteropolymer that fold into a given shape[1–9] is the most challenging one because it is a key technology for protein engineering. This problem is also called "inverse folding" of heteropolymers. While the equilibrium conformation of a polymer with a given sequence of monomers is asked in a *folding* problem, a sequence that folds into a given conformation of a polymer is requested in the corresponding *inverse folding* problem.

The first step in the inverse folding problem is to formulate it as an optimization problem. In the pioneering work[1] of Shakhnovich and Gutin, a sequence $\boldsymbol{\sigma}$ of monomers (or "amino acids") that minimize the energy $E(\widetilde{\boldsymbol{r}}|\boldsymbol{\sigma})$ of a target conformation $\widetilde{\boldsymbol{r}}$ is chosen under the condition of a fixed monomer composition. This criterion is based on the hypothesis that the energies of misfolded conformations depend only on the composition of the polymer. Algorithms with this criterion have been shown to work in many practical problems with computational efficiency. An optimal solution, however, is not always ensured.[4]

Kurosky and Deutsch[3)4] proposed a different criterion. In their framework,[10] the equilibrium probability

$$P_\beta(\widetilde{\boldsymbol{r}}|\boldsymbol{\sigma}) = \frac{\exp(-\beta E(\widetilde{\boldsymbol{r}}|\boldsymbol{\sigma}))}{Z_\beta}\ , \tag{1.1}$$

$$Z_\beta = \sum_{\boldsymbol{r}} \exp(-\beta E(\boldsymbol{r}|\boldsymbol{\sigma})) \tag{1.2}$$

that a polymer is found in a target conformation $\widetilde{\boldsymbol{r}}$ is maximized at a sufficiently low temperature $1/\beta$. Here $\sum_{\boldsymbol{r}}$ means the summation over all possible conformations. Hereafter, we denote this criterion as "MTP criterion" (Maximum Target Probability criterion). The MTP criterion automatically excludes solutions with degenerate ground states when there exists other solution with an unique ground state. In addition, if we keep the temperature $1/\beta$ low but a finite value, it selects a solution with a larger energy gap above the ground state. An advantage of MTP criterion is that it is useful to systematically understand various algorithms as methods for approximate maximization of MTP criterion, for example, the original algorithm by Deutsch and Kurosky,[4] the algorithm by Morrissey and Shakhnovich[5] , and the algorithms by Seno *et al.*[6]

The purpose of this paper is to present a new approach to the inverse folding problem based on the MTP criterion. At the present status, we are mostly concentrated on the design of the sequences that have a given target conformation as the unique ground state. For this purpose, the proposed approach has a considerable advantage over other approaches, e.g., the one in ref. 6, which will be described in the section 3 and 4. Although we are interested in the low temperature limit $\beta \to \infty$, the use of MTP criterion at low but *finite* temperature is essential in the derivation of the proposed algorithm.

A goal of inverse folding of hetropolymers is to control the dynamical property of the designed sequence as well as thermodynamic one. That is, we want to design a sequence that fold fast into the desired conformation. Around this problem, there are a number of issues[7, 11, 12] on the relation between the thermodynamic property and dynamical property of heteropolymers below the folding temperature. In our context, it will be interesting to test the dynamical properties of the sequence designed by the proposed algorithm. We will, however, leave it

* e-mail: iba@ism.ac.jp
** e-mail: kikuchi@godzilla.phys.sci.osaka-u.ac.jp

for a future problem and restrict ourselves to an innovation of the computational technique in this paper. We also hope that the development of an efficient algorithm will contribute investigations into the above-mentioned problem.

§2. The Design Equation Approach

Let us discuss the proposed approach – the design equation approach. Our starting point is the analogy to the learning in artificial neural networks. In fact, the maximization of the probability eq. (1.1) is analogous to Boltzmann machine learning[13, 14] in neural network theories (and also maximum likelihood estimation in statistics[15]), where the parameters that maximize the probability of a given data are selected as an optimal solution. Although this analogy to Boltzmann machine learning was already mentioned in refs. 3 and 4, it has not been fully explored in the literature. Specifically, no direct analog of *Boltzmann machine learning equation* has been discussed in connection with the inverse folding problem. Here we further pursue the analogy between the MTP criterion and Boltzmann machine learning and give a new approach based on a *design equation*, an analog of Boltzmann machine learning equation. For this purpose, we introduce "soft spins" or "fuzzy monomers" as a tool of computation.

We will discuss our approach in terms of the generalized HP model[16–19] of protein. This model consists of a self-avoiding polymer chain on a lattice with two types of monomers indexed by 1 and -1. As a model of protein, the indices ± 1 indicate hydrophobic (H) and polar (P) residues, respectively. The interaction energy U between a pair of monomers (the contact energy) is defined as $U(1,1)=\epsilon_1$, $U(1,-1)=U(-1,1)=\epsilon_2$, and $U(-1,-1)=\epsilon_3$. It acts only between monomers on nearest neighbor sites but not consecutive along the chain. The energy $E(\boldsymbol{r}|\boldsymbol{\sigma})$ of a conformation $\boldsymbol{r}=\{r_i\}$ (r_i represents the vector of coordinates of the ith monomer) of a polymer with a sequence $\boldsymbol{\sigma}=\{\sigma_i\}$ is written as

$$E(\boldsymbol{r}|\boldsymbol{\sigma}) = \frac{1}{2}\sum_{ij} U(\sigma_i,\sigma_j)\,\delta(|r_i-r_j|-1)\,\eta_{ij} \qquad (2.1)$$

where $\sigma_i \in \{-1,1\}$ represents the type of ith monomer. The factor $\eta_{ij}\in\{0,1\}$ takes the value 0 if and only if monomers i and j are consecutive along the chain. For later convenience, we note that the interaction energy $U(\sigma_i,\sigma_j)$ of the model is written as

$$U(\sigma_i,\sigma_j) = \epsilon_a\sigma_i\sigma_j + \frac{\epsilon_b}{2}(\sigma_i+\sigma_j)+\epsilon_c \qquad (2.2)$$

$$\epsilon_a = \frac{\epsilon_1-2\epsilon_2+\epsilon_3}{4},\ \epsilon_b = \frac{\epsilon_1-\epsilon_3}{2},\ \epsilon_c = \frac{\epsilon_1+2\epsilon_2+\epsilon_3}{4}\ . \qquad (2.3)$$

At this point, we introduce "soft spins" $\boldsymbol{m}=\{m_i\}$ each of that takes a continuous value $-1\le m_i \le 1$ and substitute the original binary variables $\boldsymbol{\sigma}=\{\sigma_i\}$. Non-integer values of the variable m_i ("fuzzy monomer") have no physical meaning in the generalized HP model. They are, however, convenient tools for the computation, as they are in the Hopfield-Tank method[20] for combinatorial optimization problems. We take the following form eq. (2.4) of the energy function, which is a straightforward extension of eq. (2.1) to the soft spin model.

$$E^*(\boldsymbol{r}|\boldsymbol{m}) = \frac{1}{2}\sum_{ij} U^*(m_i,m_j)\,\delta(|r_i-r_j|-1)\,\eta_{ij} \qquad (2.4)$$

The only constraint imposed on the modified energy function $E^*(\boldsymbol{r}|\boldsymbol{m})$ with continuous variables $\boldsymbol{m}=\{m_i\}$ is that it coincides with the original energy function eq. (2.1) when $|m_i|=1$ for all i. In this paper, we consider two possible choices for U^*,

$$U^1(m_i,m_j) = \epsilon_a m_i m_j + \frac{\epsilon_b}{2}(m_i+m_j)+\epsilon_c\ , \qquad (2.5)$$

and

$$U^2(m_i,m_j) = \epsilon_a m_i m_j + \frac{\epsilon_b}{2}(m_i|m_i|+m_j|m_j|)+\epsilon_c\ . \qquad (2.6)$$

The former expression eq. (2.5) is an obvious extension of the original energy function eq. (2.2) and it is easy to see that the latter expression eq. (2.6) is also coincides with the original energy function when $|m_i|=|m_j|=1$.

If we substitute these energy functions in eq. (1.1) of the MTP criterion, it gives

$$P_\beta(\widetilde{\boldsymbol{r}}|\boldsymbol{m}) = \frac{\exp(-\beta E^*(\widetilde{\boldsymbol{r}}|\boldsymbol{m}))}{Z_\beta}\ , \qquad (2.7)$$

$$Z_\beta = \sum_{\boldsymbol{r}} \exp(-\beta E^*(\boldsymbol{r}|\boldsymbol{m}))\ . \qquad (2.8)$$

The variables $\{m_i\}$ that maximize the expression eq. (2.7) take, in general, non-integer values and not necessary satisfy the relations $-1\le m_i\le 1$. Then we introduce a penalty term

$$V_p(\boldsymbol{m}) = \frac{\lambda}{4}\sum_i (m_i^2-1)^2 \qquad (2.9)$$

to force them towards 1 or -1. The value of the parameter λ controls the strength of the penalty. Using this penalty term, we arrive at the cost function

$$V_{cost}(\boldsymbol{m}) = -\log P_\beta(\widetilde{\boldsymbol{r}}|\boldsymbol{m}) + V_p(\boldsymbol{m}) \qquad (2.10)$$

to be minimized in the present approach.

The use of the soft spins and the penalty term are key features of our approach. By virtue of them, we can differentiate the cost function (2.10) with m_i and write down a set of equations,

$$\tau\frac{dm_i}{dt} = -\frac{\partial V_{cost}}{\partial m_i} = f_i(\beta,\boldsymbol{m}) - \lambda m_i(m_i^2-1)\ , \qquad (2.11)$$

$$f_i(\beta,\boldsymbol{m}) = \beta\sum_j \frac{\partial U^*(m_i,m_j)}{\partial m_i} \times \{\delta(|\widetilde{r}_i-\widetilde{r}_j|-1) - \langle\delta(|r_i-r_j|-1)\rangle_\beta\} \qquad (2.12)$$

that minimize the cost function eq. (2.10) with a gradient decent method. Here the variable t is a fictitious time and the constant τ controls the time scale. The average $\langle\delta(|r_i-r_j|-1)\rangle_\beta$ indicates the canonical average of $\delta(|r_i-$

$r_j|-1)$ at the inverse temperature β, i.e.,

$$\langle\delta(|r_i-r_j|-1)\rangle_\beta=\sum_{\boldsymbol{r}}\delta(|r_i-r_j|-1)P_\beta(\boldsymbol{r}|\boldsymbol{m})\ . \quad (2.13)$$

We omit the factors η_{ij} in the expression of f_i in eq. (2.12) because their effects cancel between the first and second terms in the brace { } in eq. (2.12).

In this paper, we refer to the set of equations, eqs. (2.11) and (2.12), as the *design equation* for this problem. The design equation is nothing but an analog in the inverse folding problem of the Boltzmann machine learning equation. When the value of the control parameter λ is gradually increased to $+\infty$ as the fictitious time $t\to\infty$, the value of each soft spin m_i converges to ±1, which defines a sequence with proper meaning in the original problem. It is easy to see that a sequence whose unique ground state coincides with the desired conformation satisfies the equation $\partial V_{cost}/\partial m_i=0$. Thus, for sufficiently low temperature $1/\beta$, the output of the present procedure is a candidate for the solution of the original problem with the MTP criterion.

Details of our implementation of the design equation are shown in the followings:

1. Initialization.
   Set $k:=1$, $m_i:=m_i^0$, $\lambda:=\lambda_0$.
2. Calculation of the canonical averages.
   Calculate $\langle\delta(|r_i-r_j|-1)\rangle_\beta$ by the exact enumeration or by a Monte Carlo procedure. The former is possible only for short chains.
3. An iteration of the discretized design equation.

$$\begin{aligned}f_i&:=\beta\sum_j\frac{\partial U^*(m_i,m_j)}{\partial m_i}\\&\quad\times\{\delta(|\tilde{r}_i-\tilde{r}_j|-1)-\langle\delta(|r_i-r_j|-1)\rangle_\beta\}\\m_i&:=m_i+\Delta\left\{f_i-\lambda m_i(m_i^2-1)\right\}\ .\end{aligned}$$

4. Clipping.
   If $m_i>1$, set $m_i:=1$.
   If $m_i<-1$, set $m_i:=-1$.
5. Increment the variables and check the convergence.
   Set $k:=k+1$ and $\lambda:=\lambda+a$. If a prescribed stopping criterion is satisfied or the counter k exceeds a prescribed maximum of the iteration, end up the calculation. Else, return to step 2.

The constant a controls the increase of the strength λ of the penalty and the constant Δ controls the size of steps in the gradient decent. The tunable control parameters are λ_0, a, Δ, β and initial conditions $\{m_i^0\}$.

There are two candidates of the stopping criterion. A possible criterion is "If, for all i, the value of m_i is sufficiently close to 1 or -1 and does not change in several consecutive iterations, end up the calculation". If we can check at every step whether the output is a solution of the problem, we can use another criterion based on the "forced discretization" $\hat{\sigma}_i$ of m_i defined as follows:

$$\begin{aligned}\hat{\sigma}_i&=\ \ 1\qquad\text{if}\ \ m_i>0,\\\hat{\sigma}_i&=-1\qquad\text{otherwise}.\end{aligned}\qquad(2.14)$$

With this stopping criterion, we end up the calculation before reaching prescribed maximum number of the iteration if and only if the target conformation is the unique ground state of the polymer with the sequence $\hat{\sigma}_i$.

§3. Related Works

In this section, we discuss some of earlier works in connection with the design equation approach. First, the use of the soft spin variables is, in itself, not a new technique for the protein design. They are also introduced as optimization tools in the refs. 8, 9, 21. In these references, however, criteria similar to that of Shakhnovich and Gutin[1] are used and the MTP criterion is not employed. The design equation is not appeared in these works.

On the other hand, there have been proposed several algorithms based on the MTP criterion, which use different approximations and computational techniques. In the original algorithm proposed by Deutsch and Kurosky[4] (see also Kurosky and Deutsch[3]), the logarithm of the partition function Z_β is approximated by the first order cumulant expansion and equally-weighted average over all compact self-avoiding conformations of the polymer. The resultant expression is optimized through a simulated annealing procedure. For the generalized HP model, the high temperature approximation leads to the expression apparently similar to the the right hand side of the design equation. There is, however, a major difference, because the right hand side of the design equation eq. (2.12) can automatically incorporate the effect of higher order correlations in the conformation space as the magnitudes of $\{|m_i|\}$ are increased. Morrissey and Shakhnovich[5] also gives an algorithm based on MTP criterion, which uses a higher order cumulant expansion of the free energy. These cumulants are evaluated by a mean field approximation in conformation space, so that it still keeps the computational economy of the algorithms based on the Shakhnovich and Gutin's criterion.[1] Its performance seems mostly dependent on the validity of the approximation in the conformation space. We will discuss another aspect of their work at the end of the paper.

Seno *et al.*[6] developed a dual Monte Carlo algorithm that is most faithful to the MTP criterion. In their algorithm the partition function Z_β is calculated by an important sampling in the conformation space and the calculated value of $P(\tilde{\boldsymbol{r}}|\boldsymbol{\sigma})$ is optimized through a simulated annealing procedure in the sequence space. They also test the performance of the algorithm where the Monte Carlo calculation of the partition function Z_β is replaced by the exact enumeration of conformations. In the next section, we compare the performance of an algorithm based on the design equation with this version of algorithm. A significant difference between our approach and theirs is that the average $\langle\delta(|r_i-r_j|-1)\rangle_\beta$ over the Gibbs distribution eq. (2.7) is required in our approach, instead of the partition function Z_β required in the algorithms by Seno *et al.*

For any algorithm based on the simulated annealing in the sequence space, the value of a cost function should be evaluated once per *a spin flip* in the simulated annealing procedure. For example, in the algorithms of Seno *et* *al.*, the value of the partition function Z_β is calculated once per a trial of changing the type of one monomer in the sequence. In the design equation approach, the canonical averages $\langle \delta(|r_i - r_j| - 1) \rangle_\beta$ is calculated only once per *an iteration* of the discretized design equation. As will be shown later, this difference causes a significant advantage of the algorithms based on the design equation when the computational cost of the cost function and its derivatives are comparable and both intensive.

§4. Numerical Experiments

Here, we test the design equation approach for a three dimensional generalized HP model (2.1) on the cubic lattice with 27 monomers. In the experiments, we restrict the conformations of the polymer to maximally compact self-avoiding ones filling the $3 \times 3 \times 3$ lattice. Numerical experiments with compact conformations on small lattices are common in the study of protein folding[22] and inverse folding[1, 4, 6] because then the exact enumeration of the conformations is possible.

In the following experiments, the canonical averages $\langle \delta(|r_i - r_j| - 1) \rangle_\beta$ are calculated by the exact enumeration. The parameters $(\epsilon_1, \epsilon_2) = (-1, 0)$ and $(-2.3, -1)$ are used. We set $\epsilon_3 = 0$ throughout the experiments. Note that if we restrict our attention to the maximally compact conformations,[23] the result is not affected by the addition of a common constant to all of the parameters $\epsilon_1, \epsilon_2, \epsilon_3$ (equivalently, by the value of the parameter ϵ_c). Thus, we can set $\epsilon_3 = 0$ without any loss of generality. We test both of the modified energy function U^1 , U^2 defined in eq. (2.6) and eq. (2.5).

In Fig.1, an example of the designing process is shown. In this case, the energy function U^1 is used and the algorithm successfully found a sequence that has the target conformation as the unique ground state (a "good" sequence).

To check the performance of the present algorithm, we execute the test calculations for five different target conformations in Table I. The conformation #1 is "the most designable one" for the generalized HP model on the $3\times3\times3$ lattice according to the definition in ref.24. It seems to be one of the easiest target also in the present cases. The conformations #2 - #5 are randomly chosen ones and some of them are much more difficult as a target than #1. For each target conformation, 200 trials with different initial sequences $\{m_i^0\}$ are performed. The number of the successful attempts that result in a "good" sequence is recorded. We also record the number of the iterations needed to reach a good sequence when we use "forced dicretization" as a stopping criterion. The initial values $\{m_i^0\}$ of $\{m_i\}$ are generated randomly within the range $[-w, w]$ using a uniform random number. The temperature $1/\beta$ in MTP criterion is set to 0.01. The values of the control parameters in the design equation are $\Delta = 0.5, \lambda_0 = 0.0, a = 0.5$.

To get a feeling of the difficulty of the problem, we also implement an algorithm based on simulated annealing in the sequence space. This algorithm, to which we will refer as "SA", is essentially an algorithm by Seno *et al.*[6] with the calculation of Z_β by the exact enumeration. In the experiments of SA, 10 trials with different initial sequences are performed for each target conformation. The fictitious temperature T_{SA} of the simulated annealing is initially set as $T_{SA}^0 = 2.0$ and updated by $T_{SA} := 0.8 \times T_{SA}$ in every Monte Carlo step per spin. Note that the fictitious temperature T_{SA}, which is introduced for the optimization in the sequence space, is independent of the temperature $1/\beta$ that determines the Gibbs distribution eq. (1.1) in the conformation space. The value of $1/\beta$ is kept 0.01 throughout the experiments.

The results of the experiments with $(\epsilon_1, \epsilon_2) = (-1, 0)$ and $(-2.3, -1)$ are shown in Table II and Table III, respectably.

#1	$brf_2l_2u_2b_2r_2f_2ldrb_2l_2dfuru$
#2	$r_2f_2ubufl_2d_2brfubulbr_2dl_2f$
#3	$r_2f_2ublulfd_2brfu_2rb_2dluldf$
#4	$r_2f_2ul_2ub_2druf_2rb_2dfldflbu$
#5	$r_2f_2ubldfu_2rblbrdl_2uf_2d_2bu$

Table I. The target conformations in the experiments. Each conformation is encoded by the symbols r(ight), l(eft), u(p), d(own), f(forward), b(ackward). The expressions r_2, f_2 .. are contracted forms of rr,ff etc.

The design equation with the modified energy function U^2 successfully finds at least one good sequence for each of the five targets both for $(\epsilon_1, \epsilon_2) = (-1, 0)$ and $(-2.3, -1)$. The comparison to SA shows that the cost S of the calculation in each successful run is much less in the proposed algorithm than in SA. This result is highly dependent on the computational advantage of the design equation approach, i.e, the economy of the evaluation of the cost function that we have already mentioned at the end of the previous section. Note that this advantage will be larger as the length of the polymer increases. On the other hand, the rate R of the success is lower in the proposed algorithm in most cases. Then overall performance seems comparable in both algorithms. If we define an index $P = 100 \cdot R/S$ as a measure of efficiency, the proposed algorithm with U^2 makes better scores than SA in most cases.

Our experiments show that the performance with the energy function U^1 is rather poor in the case of $(\epsilon_1, \epsilon_2) = (-2.3, -1)$. On the other hand, the algorithm works well with the energy function U^2 in both cases $(\epsilon_1, \epsilon_2) = (-1, 0)$ and $(-2.3, -1)$. A possible reason of this difference is that the energy U^2 is a better representation of the original energy function eq. (2.2) in the sense that the factor $m_i m_j$ in the first term and the factor $m_i|m_i| + m_j|m_j|$ in the second term of eq. (2.6) are both in the same order of magnitude for small values of $|m_i|$s. In the energy function U^1, the second term $\frac{\epsilon_b}{2}(m_i + m_j)$ in eq. (2.5) dominates when the values of $|m_i|$s are small and the early stage of the designing process is mostly driven by the second term when we start

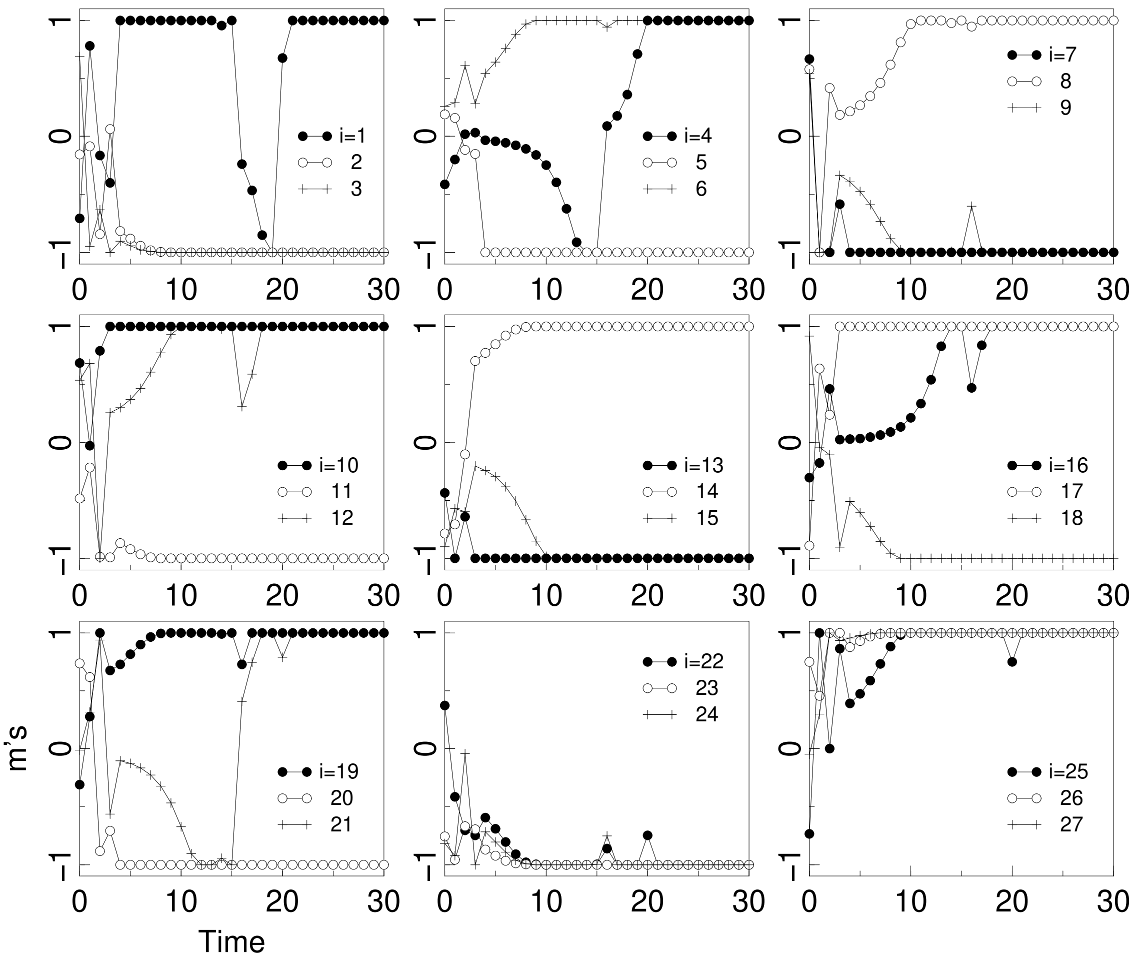

Fig. 1. An example of the designing process.
The value of each m_i is plotted versus the number of the iteration. A successful convergence to a good sequence of the target conformation #1 (see Table I) is shown. Note that "forced discretization" provides a good sequence with 20 iterations of the discretized design equation. Here $(\epsilon_1, \epsilon_2) = (-1, 0)$ and $\Delta = 0.5, \lambda_0 = 0.0, a = 0.5$.

from an initial condition with small $|m_i|$s.

§5. Summary and Future Problems

In Summary, we have proposed a novel approach for heteropolymer design that is based on the Maximum Target Probability criterion by Kurosky and Deutsch. The essential point of the proposed approach is the introduction of soft spins (fuzzy monomers) that leads to a *design equation*, which is an analog of the Boltzmann machine learning equation in heteropolymer design. We have tested this approach, which we call *the design equation approach*, for a generalized HP model on the $3 \times 3 \times 3$ lattice and have shown that it could successfully find good sequences for target conformations with different degrees of the difficulty. With these examples, our implementation of the design equation approach shows at least comparable performance to an algorithm based on simulated annealing in sequence space, when a suitable form of the modified energy function U^* is chosen.

In this paper, we have evaluated the performace of the algorithm by the ability to give a "good sequence", a sequence whose unique ground state coincides with a desired connformation. As has been mentioned in the introduction of the paper, it is also interesting to test dynamical property of the outputs of the algorithm below their folding temperatures. On the other hand, a recent work[5)] of Morrissey and Shakhnovich arises a problem of selecting sequences whose equilibrium state at a finite, not necessarily low, temperature is dominated by a given shape. The design equation approach might also be useful in this problem. For this purpose, however, further study of the behavior of the design equation at finite temperature $1/\beta$ is required.

The design equation approach is fairly general and various modifications are possible. Here we touch on a few important extensions. First, we can substitute a Monte Carlo simulation for the exact enumeration in the present algorithm. This is important because exact enumeration of the conformations is impossible for longer or off-lattice polymers. Research in this direction is now in progress and we can successfully implement an algorithm based on the design equation with a dynamical Monte Carlo simulation. The results will be reported in the forthcoming paper.[25)]

Target	U^1			U^2			SA		
	R	S	P	R	S	P	R	S	P
#1	$0.82^{(*)}$	12.2	6.7	$0.56^{(\dagger)}$	16.4	3.4	1.0	162.0	0.62
#2	$0.215^{(*)}$	15.0	1.4	$0.16^{(*)}$	14.8	1.1	0.9	294.3	0.31
	$0.20^{(\dagger)}$	18.5	1.1	$0.25^{(\dagger)}$	14.0	1.8			
#3	$0.30^{(*)}$	16.0	1.9	$0.355^{(\dagger)}$	15.9	2.2	0.8	162.0	0.49
#4	$0.04^{(*)}$	17.6	0.23	$0.055^{(\dagger)}$	14.4	0.38	0.4	256.5	0.16
#5	$0.08^{(*)}$	12.1	0.66	$0.11^{(\dagger)}$	14.7	0.75	0.1	232.2	0.043

Table II. The results with $(\epsilon_1, \epsilon_2) = (-1, 0)$ for the target conformations in Table I. Rates (R) of finding good sequences, average numbers (S) of iterations (or spin flips in SA) and "efficiency" $P = 100 \cdot R/S$ are shown in the cases of (1) the proposed algorithm with the energy function U^1, (2) the proposed algorithm with the energy function U^2, and (3) an algorithm based on simulated annealing in the sequence space (SA). In SA, total number of spin flips until a "good" sequence is used as a correspondence of the number of the iteration of the design equation, because the calculation of the partition function is required in each spin flip in SA. The initial condition m_i^0 is generated randomly within $(-w, w)$, where (*) $w = 1.0$, (†) $w = 0.1$.

Target	U^1			U^2			SA		
	R	S	P	R	S	P	R	S	P
#1	$1.0^{(*)}$	8.5	11.8	$0.75^{(\dagger)}$	10.0	7.5	1.0	105.3	0.95
#2	$0.005^{(*)}$	1.0	0.5	$0.005^{(*)}$	12.0	0.04	0.9	186.3	0.48
	$0.01^{(\dagger)}$	2.5	0.4	$0.1^{(\dagger)}$	10.8	0.93			
#3	$0.002^{(*)}$	5.0	0.04	$0.225^{(\dagger)}$	13.0	1.7	0.8	210.6	0.38
#4	$0.00^{(*)}$	—	—	$0.015^{(\dagger)}$	4.8	0.31	0.0	—	—
#5	$0.00^{(*)}$	—	—	$0.045^{(\dagger)}$	11.7	0.38	0.3	251.3	0.12

Table III. The results with $(\epsilon_1, \epsilon_2) = (-2.3, -1)$. The set of the targets is the same as that in the Table II (see Table I). The meaning of the notations are shown in the caption of Table II.

Another important challenge is the extension to the cases with monomers of many different types (many letter cases). Although a formal extension to many letter cases is not difficult, the test of the performance of the design equation in such cases is also left for the future study.

Acknowledgements

This work was supported by a Grant-in-Aid from the Ministry of Education Science and Culture of Japan, Japan Society for the Promotion of Science. A part of the computation in this work has been done using the facilities of the Supercomputer Center, Institute for Solid State Physics, University of Tokyo.

1) E. I. Shakhnovich and A. M. Gutin: Proc. Natl. Acad. Sci. USA, **90** (1993) 7195.
2) M. Sasai: Proc. Natl. Acad. Sci. USA, **92** (1995) 8438.
3) T. Kurosky and J. M. Deutsch: J. Phys. **A27** (1995) L387.
4) J. M. Deutsch and T. Kurosky: Phys. Rev. Lett.**76** (1996) 323.
5) P. Morrissey and E. I. Shakhnovich: Folding & Design **1**(1996) 391.
6) F. Seno, M. Vendruscolo,A. Maritan and J. R. Banavar: Phys. Rev. Lett. **77** (1996) 1901.
7) A. M. Gutin, V. I. Abkevich and E. I. Shakhnovich: Proc. Natl. Acad. Sci. USA **92** (1995) 1282.
8) M. Delarue and P. Koehl: in *Pacific Symposium on Biocomputing '97* Eds. R. B. Altman, A. K. Dunker, L. Hunter, T. E. Klein (World Scientific, 1996) 109.
9) H. Kono and J. Doi: PROTEINS: Structure, Function and Genetics, **19** (1994) 244.
10) In fact, Kurosky and Deutsch proposed a more general framework in refs. 3 and 4. We, however, do not discuss it further here.
11) A. Sali, E. Shakhnovich and M. Karplus: Nature **369** (1994) 248.
12) D. K. Klimov, and D. Thirumalai: Phys. Rev. Lett. **76** (1996) 4070.
13) G. E. Hinton and T. J. Sejnowski: in *Parallel Distributed Processing* Vol. 1, eds. E. Rumelhart and J. L. McClelland, (MIT Press, Cambridge, Mass. 1986)
14) T. J. Sejnowski, P. Kienker, and G. E. Hinton: Physica **D22** (1986) 260.
15) Maximum likelihood estimation of unknown parameters in probability distributions with complicated normalization constants (patition functions) is discussed in the following references: Y. Ogata and M. Tanemura: The Journal of the Royal Statistical Society B **46** (1984) 496. C. J. Geyer and E. A. Thompson: Journal of Royal Statistical Society B **54** (1992) 657. C. J. Geyer: in *Markov Chain Monte Carlo in Practice* , eds. W. R. Gilks, S. Richardson, and

D. J. Spiegelhalter (Chapman and Hall, London, 1996) 241.
16) K. A. Dill: Biochemistry **24** (1985) 1501.
17) K. F. Lau and K. A. Dill: Macromolecules **22** (1989) 3986.
18) E. I. Shaknovich and A. M. Gutin: J. Chem. Phys. **93** (1990) 5967.
19) H. S. Chan and K. A. Dill: J. Chem. Phys. **95** (1991) 3775.
20) J. J. Hopfield and D. W. Tank: Biological Cybernetics **52** (1985) 141.
21) P. Koehl and M. Delarue: Current Opinion in Structural Biology **6** (1996) 222.
22) M. Karplus and E. I. Shakhnovich: in *Protein Folding*, eds. T. Creighton (W. H. Freeman and Company, New York, 1992)
23) In a sense, it corresponds to the situation where the constant ϵ_c has a large positive value, which causes compactification of the polymer.
24) H. Li, R. Helling, C. Tang, and N. Wingreen: Science **273** (1996) 666.
25) Y. Iba, M. Kikuchi, and K. Tokita: in preparation.