Statistical Mechanics of Protein Design

Tomoei Takahashi,1 George Chikenji,2 and Kei Tokita1

1Graduate School of Informatics, Nagoya University, Nagoya, 464-8601, Japan.
2Graduate School of Engineering, Nagoya University, Nagoya, 464-8603, Japan.

(Dated: May 12, 2022)

In this study, we propose an analytic statistical mechanics approach to solve a fundamental problem in biological physics called protein design. Protein design is an inverse problem of protein structure prediction, and its solution is the amino acid sequence that best stabilizes a given conformation. Contrary to previous computational physics studies, we used the cavity method, an extension of the mean-field approximation that becomes rigorous when the interaction network is a tree. We found that for small two-dimensional (2D) lattice hydrophobic-polar (HP) protein models, the design by the cavity method yields results almost equivalent to those from the Markov chain Monte Carlo method.

Over the past 25 years, there has been an increasing interest in using statistical mechanics approaches to solve inverse problems in information processing, such as error-correcting codes, combinatorial optimization problems, data analysis, and machine learning 1. The Ising model or a spin glass model, which is the model of disordered systems with heterogeneous interactions, has been commonly used to solve inverse problems 2. The calculation is performed in the opposite direction of the ordinary statistical mechanics, i.e., to identify a microscopic sequence or parameters that minimize specific cost functions. For example, the Boltzmann machine learning identifies synaptic weights which makes learning patterns an equilibrium state; in an inverse Ising problem we infer the coupling strengths between spins given observed spin configurations 3. An essential benefit of these studies is that such approaches have solved the problem of computational explosion in probabilistic inference for information processing using approximation methods, such as mean-field approximation.

In this letter, we apply statistical mechanics to an inverse problem, not to an information processing problem, but to a fundamental problem in biological phenomena: protein design. Protein design is an inverse problem of protein structure prediction. The problem is to find the amino acid sequence that best stabilizes a given conformation 4, 5.

The application of protein design to drug design by designing novel proteins with desired biological functions is significant. Besides, its fundamental scientific significance is understanding the relationship between structures and sequences from the perspective of “design” which cannot be clarified by the forward approach, i.e., structure prediction.

Because a native conformational state of a protein is an equilibrium state determined solely by the amino acid sequence under the physiological conditions (the Anfinsen’s dogma) 6, to solve the protein design problem, it is sufficient to know the sequence that minimizes the free energy of a given native conformation. Therefore, the solution to the protein design problem is a sequence that makes the given native conformation the only ground state at low temperature.

The “design criteria” can be categorized into two main ways: maximizing target probability (MTP) and energy minimization. The lattice HP model 7, a kind of Ising model of proteins is the most commonly used. The MTP is a method to maximize the conditional probability of a native conformation expressed by a canonical distribution (target probability) 8, 9. Moreover, various MTP-based methods have been proposed 10–14. Energy minimization is a more straightforward method and a sequence that minimizes the energy of a given conformation is the solution 15–22. The aforementioned studies are engineering-oriented and computational approaches.

By contrast, we propose an analytical design method using the cavity method. The cavity method is a method that extends the Bethe approximation to other than two-body interactions. If the interaction network is a tree graph, the cavity method is rigorous except far from the tree center. The cavity method has been applied to analyze stochastic models on random graphs that can be regarded locally as trees. In addition, there are a few applications to protein folding problems, such as the computation of phase diagrams for lattice HP models 23 and the prediction of contact maps 24. However, to the best of our knowledge, there are no studies of protein design using the cavity method.

In the lattice HP model, the backbone chain of a protein is represented by a lattice self-avoiding walk, and each site represents an amino acid residue. There are only two types of amino acid residues: hydrophobic (H) and hydrophilic (P). In this study, we consider $N$ residues $\sigma = \{\sigma_1, \sigma_2, \ldots, \sigma_N\}$ on a lattice position $r = \{r_1, r_2, \ldots, r_N\}$, where $i = 1, 2, \ldots, N$ $\sigma_i = 1$ indicates that the $i$-th residue is an H-residue, and $\sigma_i = 0$ indicates that it is a P-residue. Let $\mu$ be the chemical potential of water, we proposed the following Hamiltonian for proteins with structure $r$ and sequence $\sigma$ in our previous work 25:
\[ H(r, \sigma; \mu) = -\sum_{i<j} \sigma_i \sigma_j \Delta(r_i - r_j) - \mu \sum_i (1 - \sigma_i). \] (1)

The function \( \Delta(r_i - r_j) \) in Eq. (1) takes the value 1 only when \( r_i \) and \( r_j \) are not continuous in the backbone chain and are nearest neighbors in the coordinate space, otherwise, it takes the values 0. Such a positional relationship is referred to as being in contact with each other. The first term in Eq. (1) is the Hamiltonian of the lattice HP model usually used. Because \( \sigma_i \) is a P-residue, \( \sigma_i = 0 \), Eq. (1) is a modified Hamiltonian of the original Hamiltonian of the lattice HP model by adding the interaction between P-residues and water.

We consider the following probability of a target conformation \( r = R \) with a sequence \( \sigma \):

\[
p(R|\sigma) = \frac{1}{Z(\sigma; \beta, \mu)} e^{-\beta H(R, \sigma; \mu)}, \quad (2)
\]

\[
Z(\sigma; \beta, \mu) = \sum e^{-\beta H(r, \sigma; \mu)}. \quad (3)
\]

The partition function (3) is a sum over the states for all possible conformational patterns \( r \) that a given sequence can fold. Henceforth, Eq. (2) is referred to as the target probability or likelihood function. For the target probability (2), the maximum likelihood estimation (in previous studies, using only the first term of Eq. (1)) is the MTP. However, Eq. (2) contains a partition function (3) for the conformation space, which causes a computational explosion problem, making the MTP extremely difficult in practice.

In our previous work \[25\], we used the posterior \( p(\sigma|R) \) for designing based on Bayesian learning. In the derivation of the posterior \( p(\sigma|R) \), we proposed a prior \( p(\sigma) \) reflecting the hypothesis explained below. There, we assumed that evolved sequences have a certain statistical mechanical tendency. We hypothesize that sequences with smaller free energies have evolved and become more likely to appear. The prior distribution reflecting this hypothesis is as follows:

\[
p(\sigma) = \frac{Z(\sigma; \beta_p, \mu_p)}{\Xi(\beta_p, \mu_p)}. \quad (4)
\]

The denominator \( \Xi(\beta_p, \mu_p) \) is the partition function given by \( \sum_\sigma \sum_r e^{-\beta_p H(r, \sigma)} \) and does not depend on conformations and sequences. Let \( \beta_p \) and \( \mu_p \) be the inverse temperature and water chemical potentials in the prior distribution, respectively. The statistical mechanical explanation of this prior is that the lower the free energy \( F(\sigma; \beta_p, \mu_p) = -(1/\beta_p) \log Z(\sigma; \beta_p, \mu_p) \), the higher the probability (4).

Substituting the target probability (2) and prior (4) into the following Bayes’ theorem, we obtain the following posterior:

\[
p(\sigma|R) = \frac{p(R|\sigma)p(\sigma)}{\sum_\sigma p(R|\sigma)p(\sigma)} e^{-\beta H(R, \sigma; \mu)} Z(\sigma; \beta, \mu). \quad (5)
\]

\[
\Xi(\beta_p, \mu_p) Z(\sigma; \beta_p, \mu_p) \equiv \Xi(\beta_p, \mu_p). \quad (6)
\]

If \( \beta_p = \beta \) and \( \mu_p = \mu \), then the denominator and numerator \( Z(\sigma; \beta, \mu) \) in Eq. (6) cancel each other out. The partition function \( \Xi(\beta_p, \mu_p) \) does not depend on the sequence \( \sigma \), so it cancels with the one appearing in the normalization constant in Eq. (5). Consequently, the following posterior distribution is obtained:

\[
p(\sigma|R) = \frac{1}{Z(R; \beta, \mu)} e^{-\beta H(R, \sigma; \mu)} Z(\sigma; \beta_p, \mu_p). \quad (7)
\]

Sampling from the partition function (8) is computationally easier than sampling from the partition function (3). We perform the Markov chain Monte Carlo (MCMC) method for generating the optimal sequence from the posterior (7) in our previous work \[25\].

Notably, the derivation of the posterior (7) has interesting theoretical implications. This derivation of the above is consistent with the fact that in calculations for physical quantities of the spin glass \( +J \) model, the partition function for the spin configuration cancels out on the Nishimori line. The Nishimori line is the hypersurface in the parameter space \[26\] and achieves the Bayes-optimality, an upper bound on accuracy in error-correcting codes \[27\]. Therefore, in other words, our method can be called protein design on the Nishimori line. In addition, our method provides a correspondence between protein design and error-correcting codes in terms of Bayes-optimality. Thus, our design theory not only overcomes the computational bottleneck, but also presents a surprising relationship between the “evolutionary problem”, the analytic theory of the spin glass model, and error-correcting code.

Then, we show that posterior (7) can be strictly divided into independent probabilities for each sequence by the cavity method. This topic is novelty point of this work compared with our previous study. Posterior (7) can be expressed as follows:

\[
p(\sigma|R) = \frac{1}{Z(R; \beta, \mu)} \left( \prod_a \psi_a(\sigma_a) \right) \prod_i \phi_i. \quad (9)
\]

In Eq. (9), let \( \sigma_a \) be the set of residues related to the \( a \)-th contact, and \( \psi_a(\sigma_a) \) be a function of it. That is, if \( \sigma_a = \{ \sigma_i, \sigma_j \} \), \( \psi_a(\sigma_a) = \exp(\beta \sigma_i \sigma_j) \). The factor \( \phi_i \) is defined by \( \phi_i = \exp(\beta \mu(1 - \sigma_i)) \).

We aim to marginalize the posterior (9). Some isolated residues do not interact with any other residues in a lattice protein. For such isolated residues, marginalization
is easy. The summation of any other residues (not isolated) cancels in the denominator and numerator of Eq. (9); hence the marginal posterior of Eq. (9) is obtained as follows:

$$p(\sigma_i | R) = \frac{e^{\beta \mu (1 - \sigma_i)}}{\sum_{\sigma_i} e^{\beta \mu (1 - \sigma_i)}}. \quad (10)$$

For each residue $\sigma_i$ in contact with other residues than the isolated one, if the residue-residue interaction network is a tree, one can derive explicitly the recursion formula of the belief propagation (BP), which is an algorithm to compute marginal distributions. Therefore, we created an interaction network (hereinafter, a contact graph) for all small 2D lattice protein models used to test our method. We found that for the group of 2D lattice protein models, the fraction of structures containing even a single loop of any size was negligible ($N = 3 \times 3, 3 \times 4, 4 \times 4, 5 \times 5$, and $6 \times 6$ versus $0\%, 0\%, 1.61\%, 2.13\%$, and $5.67\%$, respectively).

The BP algorithm is derived by the expectation of Eq. (9) under the probability distribution of the system excluding the residue $\sigma_i$: $p(\sigma_i | R)$ (which is called cavity distribution). We leave the details to the supplemental material (SM), only the results are given below:

$$\nu_{a \rightarrow i}^{(t)}(\sigma_i) = C_{a \rightarrow i} \sum_{\sigma_j} e^{\beta \mu (1 - \sigma_j)} \nu_{j \rightarrow a}^{(t)}(\sigma_j), \quad (11)$$

$$\nu_{i \rightarrow a}^{(t+1)}(\sigma_i) = \nu_{i \rightarrow a}^{(t)}(\sigma_i) \prod_{b \in \partial_i \setminus a} \nu_{b \rightarrow i}^{(t)}(\sigma_i). \quad (12)$$

In Eqs. (11) and (12), let $a$ and $b$ be indices on contacts and $i$ and $j$ are indices on residues. The symbol $\partial_i$ denotes the index set of contacts related to residue $\sigma_i$. $\nu_{a \rightarrow i}^{(t)}(\sigma_i)$ is the belief from the $a$-th contact to the $i$-th residue, $\nu_{i \rightarrow a}^{(t)}(\sigma_i)$ is the belief from the $i$-th residue to the $a$-th contact, and the upper right subscript is the number of steps in the BP algorithm. The constants $C_{a \rightarrow i}$ and $C_{i \rightarrow a}$ are the normalizing constants of each distribution function. If one properly defines $\nu_{a \rightarrow i}^{(t \rightarrow 0)}(\sigma_i)$ as the initial condition and computes Eqs. (11) and (12) at each step for all combinations $(i, a)$ excluding the isolated residues, after sufficient iterations $t_{\text{max}}$, the following belief:

$$\nu_{i}^{(t)}(\sigma_i) = C_i \prod_{a \in \partial_i} \nu_{a \rightarrow i}^{(t-1)}(\sigma_i), \quad (13)$$

converges to the marginal distribution $p(\sigma_i | R) = \sum_{\sigma_j \setminus a} p(\sigma | R)$. In Eq. (13) where $C_i$ is the inverse of the normalization constant of $\nu_{i}^{(t)}(\sigma_i)$, we set the initial condition as a uniform distribution $\nu_{i}^{(t \rightarrow 0)}(\sigma_i = 1) = \nu_{i}^{(t \rightarrow 0)}(\sigma_i = 0) = 1/2$. If $\nu_{i}^{(t)}(\sigma_i = 1) > 1/2$ residue $\sigma_i$ is H and P otherwise. Because the above calculations are equivalent to the Ising model of two-body interaction, they are strictly equivalent to the Bethe approximation [29].

The optimal hyperparameter, the chemical potential of water $\mu = \mu^*$, is the same value obtained in our previous study [22]. The optimal chemical potential $\mu^*$ was determined to be the value with the highest design accuracy by repeated computational experiments.

We now show the design results. In this study, we use 2D small lattice proteins for which all possible compact conformations are enumerable, allowing us to determine whether the design was successful or not rigorously because one can calculate the energies of all pairs of the generated sequence and every conformation. Specifically, we use $N = 3 \times 3, 3 \times 4, 4 \times 4, 5 \times 5$, and $6 \times 6$ lattices. For $N = 3 \times 3, 3 \times 4$, and $4 \times 4$ we design all desirable conformations. Because $N = 5 \times 5$ and $6 \times 6$ have many total conformations (1,075 and 52,667, respectively), we chose 100 conformations randomly for each size due to computational cost. The meaning of "desirable" here is the number of sequences that make a target conformation the only ground state (designability) is nonzero. A more detailed description of the conformations, including the conformation figure, is given in the SM.

We did not use three-dimensional (3D) lattice proteins. The reason is that the structures $N = 2 \times 2 \times 3$ and $N = 3 \times 3 \times 3$, for which one can determine strict design success, are not typical examples of proteins because the number of core residues relative to the number of surface residues is deficient compared with natural proteins.

There are three types of sequences: good sequence, which has the target conformation as a unique ground state; medium sequence, which has the target conformation as one of the degenerated ground states; and bad sequence, which has the ground state conformation(s) that does not include the target conformation.

We summarize our design results in Table I. Table I shows the sequences generated using the cavity method and our previous work by MCMC, where $N_c^{(g)}$, $N_c^{(m)}$, and $N_c^{(b)}$ are the number of structures that successfully obtained a good, medium, and bad sequences, respectively. Therefore, the design success rate, SR is the ratio of $N_c^{(g)}$ to the total number of structures $N_c$. The optimal value of chemical potential $\mu^*$ was determined as explained above, and the same value was used for the cavity method and MCMC. The value $\mu^*$ may differ for each conformation even if the size is the same, but we use the same $\mu^*$ for the same size without considering this issue. The total number of conformations designed is the same for the cavity method and MCMC.

Table I shows that the cavity method and MCMC differ slightly in the percentage of correct answers at $6 \times 6$, otherwise, they perform precisely the same. This result shows almost no difference between the cavity method and MCMC in design accuracy, at least for small 2D lattice proteins.

What does it mean that the results from the cavity method, a generalization of the Bethe approximation, and the numerical results from MCMC are almost identical? That is, simply put, the lattice protein contact graph has a graph structure suited to the Bethe approxi-
TABLE I. Comparison of the cavity method and MCMC de-

sign results. The hyperparameter $\mu^*$ was calculated many
times for each size in MCMC case, and the values achieve the
highest success rate, SR. The same values of $\mu^*$ were used for
the cavity method.

| Size | Cavity method | MCMC | Cavity method | MCMC |
|------|---------------|------|---------------|------|
|      | $N_{(g)}^{(a)}$ | $N_{(g)}^{(b)}$ | SR (%) | $N_{(g)}^{(a)}$ | $N_{(g)}^{(b)}$ | SR (%) | $\mu^*$ |
| 3 x 3 | 4 | 0 | 0 | 100 | 4 | 0 | 0 | 100 | 0.55 |
| 3 x 4 | 20 | 7 | 0 | 74.1 | 20 | 7 | 0 | 74.1 | 0.6 |
| 4 x 4 | 46 | 16 | 0 | 74.2 | 46 | 16 | 0 | 74.2 | 0.62 |
| 5 x 5 | 68 | 32 | 0 | 68 | 68 | 32 | 0 | 68 | 0.74 |
| 6 x 6 | 63 | 37 | 0 | 63 | 63 | 37 | 0 | 63 | 0.8 |

It is unclear whether the contact graph of any real protein is a tree. In addition, although the loop effect can be ignored in the thermodynamic limit, it is unclear whether the loop effect can be ignored for the contact graphs of real proteins with a finite number of amino acid residues (hundreds to thousands). However, because our approach is equivalent to the Bethe approximation, we can expect it to be a good approximation if the fluctuations in the mean-field of the next-nearest neighbor residues are sufficiently small. Therefore, we believe that extending our approach to real proteins is worthwhile. We are currently verifying this using data on the 3D structure and amino acid sequence of real proteins and will report on our findings soon.

This work was financially supported by JST SPRING, Grant Number JPMJSP2125. The author (Initial) would like to take this opportunity to thank the “Interdisciplinary Frontier Next-Generation Researcher Program of the Tokai Higher Education and Research System”. The authors are grateful to T. Obuchi, Kyoto University for illuminating discussions. This work was also supported by KAKENHI Nos. 19H03166 (G.C.) and 19K03650 (K.T.).

[1] M. M´ ezard and A. Montanari, Information, Physics, and
Computation (Oxford University press, 2009).
[2] M. M´ ezard, G. Parisi, and M. A. Virasoro, Spin-Glass
Theory and Beyond (World Scientific, Singapore, 1987).
[3] H. Chau, R. Zecchina, and J. Berg, arXiv:1702.01522
(2017).
[4] I. Coluzza, J. Phys. Condens. Matt. 29, 143001 (2017).
[5] S. Cocco, C. Feinauer, M. Figliuzzi, R. Monasson, and
M. Weigt. Rep. Prog. Phys. 81, 032601 (2018).
[6] C. B. Anfinsen, Science 181, 223 (1973).
[7] K. F. Lau and K. A. Dill, Macromolecules 22, 181 (1989).
[8] T. Kuroskey and J. M. Deutsch, J. Phys. A 28, L387
(1995).
[9] J. M. Deutsch and T. Kuroskey, Phys. Rev. Lett. 76, 323
(1996).
[10] F. Seno, M. Vendruscolo, A. Maritan, and J. R. Banavar,
Phys. Rev. Lett. 77, 1901 (1996).
[11] A. Irbi`ec, C. Peterson, F. Potthast, and E. Sandelin,
Phys. Rev. E 58, R5249 (1998).
[12] A. Irbi`ec, C. Peterson, F. Potthast, and E. Sandelin,
Structure 7, 347 (1999).
[13] Y. Iba, K. Tokita, and M. Kikuchi, J. Phys. Soc. Jpn.
67, 3985 (1998).
[14] K. Tokita, M. Kikuchi, and Y. Iba, Prog. Theor. Phys.
Suppl. 138, 378 (2000).
[15] E. I. Shakhnovich and A. M. Gutin, Proc. Natl. Acad.
Sci. U.S.A. 90, 719 (1993).
[16] G. Salvi, S. M¨ olbert, and P. De Los Rios, Phys. Rev. E
66, 061911 (2002).
[17] S. Abeln and D. Frenkel, PLoS Compt. Biol. 4, e1000241
(2008).
[18] R. Ni, S. Abeln, M. Schor, M. A. Cohen Stuart, and P.
G. Bolhuis, Phys. Rev. Lett. 111, 058101 (2013).
[19] S. Abeln, M. Vendruscolo, C. M. Dobson, and D. Frenkel,
PLoS ONE 9, e85185 (2014).
[20] V. Bianco, G. Franzese, C. Dellago, and I. Coluzza, Phys.
Rev. X 7, 021047 (2017).
[21] V. Bianco, M. Alonso-Navarro, D. Di Silvio, S. Moya, A.
L. Cortajarena, and I. Coluzza, J. Phys. Chem. Lett. 10,
4800 (2019).
[22] V. Bianco, G. Franzese, and I. Coluzza, ChemPhysChem
21, 377 (2020).
[23] A. Montanari, M. M¨ eller, and M. Mezard, Phys. Rev.
Lett. 92, 185509 (2004).
[24] M. Weight, R. A. White, H. Szurmant, J. A. Hoch, and
T. Hwa, Proc. Natl. Acad. Sci. U.S.A. 106, 67 (2009).
[25] T. Takahashi, G. Chikenji, and K. Tokita, Phys. Rev. E
104, 014404 (2021).
[26] H. Nishimori, J. Phys. C : Solid State Phys. 13, 4071
(1980).
[27] Y. Iba, J. Phys. A : Math. Gen. 32, 3875 (1999).
[28] B. Rafferty, C. F. Zachary, and A. Martini, (2020),
"Protein Contact Maps", https://nanohub.org/
resources/contactmaps. (DOI: 10.21981/02DQ-MT84)
[29] Y. Kabashima and D. Saad, Europhys. Lett. 44, 668
(1998).