Statistical mechanics of protein design based on Bayesian learning

Tomoei Takahashi

Acknowledgment

This doctoral thesis is the result of my five years of research in Nagoya. I could not have achieved these research results without the support of many people over the past five years. Although it is impossible to cover them all, I would like to express my gratitude in the following.

First and foremost, I would like to express my greatest appreciation to my advisor, Prof. Kei Tokita. We exchanged many words, including small talk, during discussions, seminars, and weekly lab meetings. In the process, I believe that I consciously and unconsciously learned from him how to carry out research and how to think. I believe that what I learned from Prof. Tokita will continue to be very important in my life.

I am also very grateful to my co-researcher, Dr. George Chikenji, for his help. It is thanks to Dr. Chikenji that my study has received some attention from people who specialize in proteins.

I am also grateful Prof. Makoto Kikuchi, Motonori Ota, Osaka University and Prof. Tomoyuki Obuchi, Kyoto University for giving me the opportunity to give seminars and asking many useful comments and questions on my study.

I am also grateful Prof. Shogo Tanimura and Prof. Yasuyuki Nakamura of the Many-Body Systems Science Unit, Prof. Yuki Sugiyama who retired from the laboratory, and Dr. Yuki Izumida who moved to the University of Tokyo. I attended the same seminar with all of them. I am very grateful for their advice in those seminars and for their other administrative support.

I would also like to thank all of students that I met in the Many-Body Systems Science Unit. My friends in the Tokita Lab were all excellent and very inspiring members with their own different personalities. I really enjoyed many discussions with them, which sometimes lasted into the early hours of the morning. The younger students who accompanied me on my rambling talks and hobbies of mountain climbing are also precious experience.

Finally, I am grateful for my parents for bring me up and watching over me. They gave me the greatest possible support as I repeated a year in university, and selfishly declared my intention to enter a doctoral program of graduate school. However, it is also my father and mother who have made me the intellectually curious person I am today. Thank you.

This work was financially supported by JST SPRING, Grant Number JPMJSP2125.

Abstract

The study of this doctoral thesis addressed the inverse protein folding problem called protein design. In general, protein design has an essential significance for applying drug design. In contrast, we have emphasized the fundamental scientific significance of protein design. Statistical mechanical informatics, a fusion field of informatics and statistical mechanics, is one of the most appropriate theories for the fundamental scientific purpose: understanding the "design principle" of proteins. We have applied Bayesian learning to protein design. That is the typical approach in statistical mechanical informatics. We obtained the posterior based on Anfinsen's dogma and the statistical mechanical hypothesis for evolution: "the amino acid sequences which have lower free energy have remained through Darwinian evolution", initially proposed by us. We carried out two estimation methods for the posterior: the Markov chain Monte Carlo (MCMC) method and the cavity method. The latter is a mean-field approximation frequently used in statistical mechanical informatics but has rarely been applied to life phenomena. As a result, on applying the 2D HP model, MCMC successfully finds an amino acid sequence for which the target conformation has a unique ground state. However, the performance was not as good for the 3D lattice HP models compared to the 2D models. The performance of the 3D model improves by using 20-letter lattice proteins. The cavity method yields results almost equivalent to those 2D small HP model from MCMC with lower computational costs. We carried out a "redesign" of lysozyme using the cavity method and achieved 65% of all estimated residues that match the true residues with 2-letter level. However, this result may not be very high. We added some brief analysis about the prior. We find that at particular temperatures and chemical potentials, the evolutionarily selected sequence's free energy is lower than that of the random sequence. Thus, one of the necessary conditions for the prior distribution is proved. Based on these results, as the tentative solution for "what is the design principle of proteins?", we propose energy minimization of the target structure with the addition of chemical potential as a candidate protein design principle. That is only valid for evolutionarily selected temperatures and chemical potentials. We also discussed that it is more likely valid for less frustrated structures.

Contents

Chapter 1

Introduction

This chapter begins with a general description of proteins, followed by a description of the protein design problem and a discussion of its significance. We also introduce the effectiveness of *statistical mechanical informatics*, the method on which our research relies, for inverse problems in protein design and other life phenomena. Finally, we describe the purpose and structure of this doctoral thesis.

1.1 Protein

In this section, we begin with a general description of proteins, the phenomenon of protein folding, the Anfinsen's dogma, an important principle on which all statistical mechanical studies of proteins rely, and a brief description of their structure. The physical aspects of proteins are detailed in [1, Sasai 2008].

1.1.1 Proteins and its function

A protein is a complex three-dimensional (3D) structure formed by 20 types of amino acids linked in a chain. There are about tens of thousands kinds of proteins in the human body, and they express their functions through their unique 3D structures. Proteins are the essential substances in living organisms, playing a central role in almost all life phenomena, such as the transport of oxygen from the lungs to the rest of the body, energy conversion from food, intracellular signal transduction, and immune responses.

1.1.2 Protein folding

Proteins fold from an unfolded state without a specific structure into a specific 3D structure. It is called *protein folding.* The folded state of a protein is called its native state or in some cases, its native structure. As explained in the following subsections, the native structure of a protein is determined by its amino acid sequence. Therefore, if we understand the folding process, we can predict the 3D structure of a protein from the amino acid sequence alone. The problem of predicting the 3D structure from the amino acid sequence is called protein 3D structure prediction. Protein folding and protein structure prediction are crucial issues in structural biology.

1.1.3 Anfinsen's dogma

The function of a protein is determined by its complex 3D structure. The conformation is thermal equilibrium, which is determined by the pattern of the sequence of 20 amino acids. It is called Anfinsen's dogma. In the 1960s, Anfinsen denatured ribonuclease, an enzyme that degrades RNA molecules, by adding urea and mercaptoethanol, a compound used as a reducing agent that maintains protein activity. The structure is denatured, and the ribonuclease spontaneously returns to its original 3D structure when the added substances are removed [2, Anfinsen 1973]. It indicates that the sequence of amino acids uniquely determines the 3D structure of a protein.

In modern times, It is now known that macromolecules called *chaperone*[1] assist in protein folding. It has been shown that the amino acid sequence alone does not determine the conformation of all proteins. Nevertheless, the amino acid sequence remains essential in determining conformation, and Anfinsen's dogma is still fundamentally correct today. Therefore, as with many statistical machanics studies on proteins, this study is based on the dogma.

1.1.4 The spatio scale and structure of proteins

In general, a protein is composed of 50 to 2000 amino acids. The spatial scale is approximately on the order of 10^{-9} m (1 nm) to 10^{-8} m (10 nm).

A protein is a chain of amino acids, which are divided into two parts: one part that is identical to other connected amino acids and one part that is different. The same portion is called the main chain, and the different portion is called the side chain. The diversity of side chains of linked amino acids determines the diversity of amino acid sequences or protein structures. Twenty types of amino acids are the types of side chains. The unit corresponding to one side chain is called an amino acid residue (or simply a residue), which means the remnant of what was an amino acid before it was bound.

[1] a protein whose function is to prevent proteins from forming intermediates that differ from their final 3D structure when they fold or from aggregating with other proteins in the cell. The original definition of a chaperone was given by John Ellis [3, Ellis 1987], who proposed the concept of chaperone in the 1980s, when its function was first understood. However, it is now known that proteins play an active role in various scenes throughout their lives, from synthesis to degradation. However, we will not go into the relationship between proteins and molecular chaperones in this study.

1.2 Protein design

1.2.1 What is protein design?

In principle, the native structure of a protein can be predicted from its amino acid sequence. Then, the following engineering-oriented idea comes to mind. If we could know the amino acid sequence that folds into a certain native structure, we could create a protein with the desired function. In other words, a new problem of "design" a protein. This is the theme of this doctoral thesis: *protein design*.

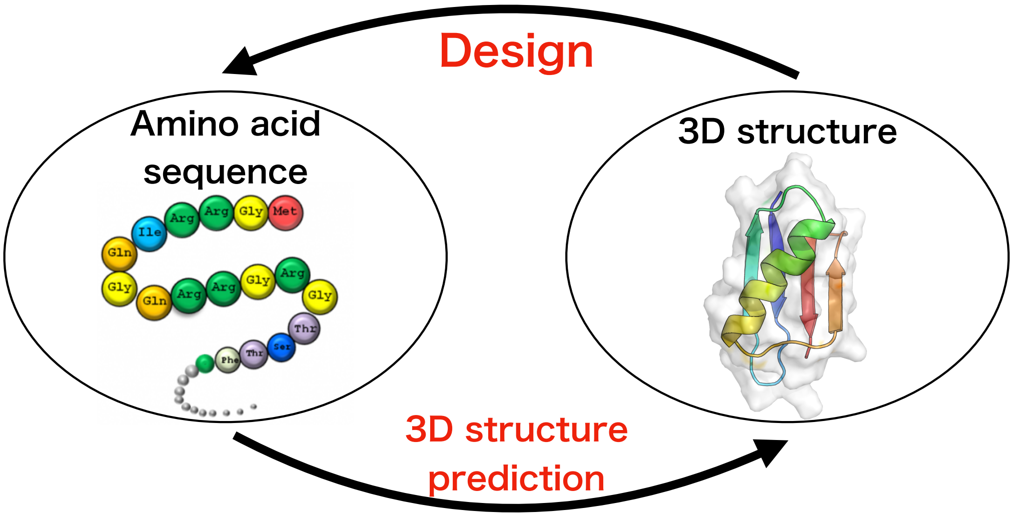

Figure 1.1: Conceptual diagram of protein design. The right represents the 3D structure, and the left represents the amino acid sequence. Prediction of 3D structure from amino acid sequence is 3D structure prediction, while prediction of amino acid sequence from 3D structure prediction is protein design. Each ball represents an amino acid.

Protein design is to predict the amino acid sequence that realizes the conformation of a protein from its conformation. Fig. 1.1 shows the conceptual diagram.

An inverse problem is a problem in which the input is estimated from the output. Conversely, the problem of finding an output from an input is called a forward problem. In the case of protein folding, information on the amino acid sequence can be considered as input and information on the native structure as output. Therefore, the statistical mechanics approach to protein design, such as our study, also has an aspect of a statistical mechanics approach to inverse problems.

The protein design problem is defined in statistical mechanics as "the problem of estimating an amino acid sequence that uniquely folds into a given structure as the equilibrium state. In other words, if we can find an amino acid sequence with a given structure (hereafter called the target structure) as its ground state at low temperatures, it is surely a solution to the protein design problem. Ideally, the system's temperature should be the cell's physiological temperature. However, in this study, we mainly use a very theoretical analysis called the lattice protein model, in which the three-dimensional structure of a complex protein is represented by a coarse-grained model such as the Ising model (described in detail in Chapter 2). For this reason, we propose a design method at as low a temperature as possible to ensure that the target structure is the only ground state of the system that is not degenerate with other similar structures, if not at zero temperature.

The work on protein design from a statistical mechanics standpoint is summarized in the review papers [4, 5, Coluzza 2017; Cocco *et al.* 2018]. More extensive protein design methods, including computational *ab initio* methods, are summarized in [6–9, Street and Mayo 1999; Lippow and Tidor 2007; Samish *et al.* 2011; Samish 2017] for more details.

1.2.2 Significance of protein design studies

Protein design studies generally has two major significances. One is the engineering and social significance of the application to macromolecular drug discovery. Even if a protein structure prediction problem is solved, it is not enough to create a new protein with the desired function. In order to create a protein with a desired function, it is necessary to "design" the protein, i.e., design it. Specifically, if the 3D structure and the amino acid sequence on a protein surface are known to fit the shape of the biosynthetic enzyme of a virus particle or a substance abnormally produced in the body by some disease, which is the target of interaction, it can be designed using protein design methods. If we can actually design antibodies against various pathogens by computer, we can realize inexpensive and speedy development of therapeutic drugs for intractable diseases that would require a large amount of development cost using current methods. In fact, (artificial) protein design was featured alongside the discovery of gravitational waves and AlpahaGo, the AI that defeated mankind's strongest Go player, in 2016 as one of the 10 biggest science news stories selected annually by Science[2].

Protein design has great social significance. However, it also has significance as a basic science. It is, in general, to elucidate the relationship between protein structure and sequence in the opposite direction to the analysis of protein folding. It is not easy to discuss this significance in detail. However, suppose, for example, that one assumes specific rules for the 3D protein structure, designs the protein under those rules, and successfully designs the structure. From the result, we can understand that the amino acid sequence has evolved to satisfy the assumed rule. Such knowledge is difficult to obtain by forward analysis (analysis of protein folding problems). Thus,

[2] https://www.science.org/content/article/watch-breakthrough-year-2016

protein design is essential as fundamental science.

That significance is not confined to protein design. Organisms are purposefully designed by evolution. They behave like autonomous, flexible machines. That characteristic not found in conventional physics is perhaps the most remarkable difference that distinguishes organisms from ordinary materials. One practical approach to understanding the principles of operation and designing machines with complex mechanisms is the constructive approach of “understanding by creating”. Protein design is a way to understand proteins by creating them. It is precise because of this approach that we can make the discoveries described above. The “understanding by creation” approach is called synthetic biology.

1.2.3 The fundamental difficulty of protein design

Protein design aims to obtain an amino acid sequence that folds into the target structure as its “unique” ground state. Therefore, even if we can estimate the sequence to minimize the energy of a given target structure, it does not mean that we ensure that the target structure with the estimated sequence is the only ground state among all possible structures.

Based on the problem, in order to obtain a solution to protein design problem, it is necessary to check whether the estimated sequence folds into the target structure each time one estimates a sequence. If the sequence folds into the target structure, it is the correct sequence; otherwise, it is necessary to re-estimate the sequence and perform the same check. One has to repeat the process until the sequence reliably folds into the target structure. In other words, in principle, protein design involves a dual optimization of sequence space and conformational space.

The almost infinite degrees of freedom of the protein conformation make this dual optimization a difficult task. Furthermore, protein folding is a quantum many-body dynamics with a complex Hamiltonian, so even today, it is impossible to perform the folding simulation, even with a supercomputer. Moreover, it must be done multiple times in the double loops. Thus, the protein design problem is, in principle, a more computationally difficult problem involving protein structure prediction, which is the forward problem of protein design.

However, widely used protein design packages, such as Rosetta [3], have succeeded in designing de novo proteins. Moreover, they are not performing the above-mentioned double optimization but can design a protein by only sampling the sequence in many cases. Why is this? Thus, there is a large gap between the principle problem of protein design and the actual protein design algorithms, and even today we do not know any protein design“principle” that can explain this gap. There is a possibility that important issues related to protein evolution are hidden here.

[3] We will describe Rosetta in detail in Chapter 2

The most critical issue of this doctoral thesis is a fundamental scientific problem that is not encountered in solving the forward problem but is encountered only in the constructivist approach of "understanding by creating". In this study, we aim to solve this problem by proposing a statistical mechanics hypothesis of evolution based on the prior distribution of protein design formulated by Bayesian learning, which is an a priori probability distribution of the likelihood of the appearance of amino acid sequences. The prior profile, which will be discussed in detail in Chapter 3 and further chapters, is the basis of the research for this doctoral thesis.

1.3 Statistical mechanical informatics

What is the most effective approach to solve the above questions? The author believes that statistical mechanical informatics is one of the most effective approaches. The formulation of the problem by Bayesian learning and its statistical mechanics analysis used in this doctoral thesis can be regarded as the cornerstone method in statistical mechanical informatics.

Statistical mechanical informatics[4] is an interdisciplinary area of physics (statistical mechanics) and informatics that provides efficient methods and systematic performance evaluation for problems such as error-correcting codes, combinatorial optimization, machine learning, and high-dimensional statistics[5] using statistical mechanics of disordered systems such as spin glasses [10–13, Mézard *et al* 1987; Nishimori 2001; Kabashima 2007; Mézard and Montanari 2009].

All of the objects mentioned above of statistical mechanical informatics are in the form of inverse problems. In statistical mechanics, one starts from microscopic states to obtain a thermodynamic macroscopic quantity such as entropy. However, statistical mechanical informatics targets problems in the opposite direction, such as estimation problems. One of the main objectives of statistical mechanical informatics is said to be two-fold. The first is to develop a method for estimation by the replica method[6] systematically. The other is the proposal of efficient estimation methods based on mean-field approximation methods, such as the cavity method, which extends the Bethe approximation to three-body interactions and beyond. In other words, statistical mechanical informatics is a theory with the character of a constructivist approach, even though it has a physics aspect. That is compatible with the constitutive nature of synthetic biology, as described in section 1.2.2, which

[4] This name is probably only used in Japan. In the West, it is called statistical mechanics of information processing, etc. In this doctoral thesis, we use this terminology to unify the field of statistical mechanics, which deals with informatics, into a single field.

[5] Statistics for the case where the number of dimensions of the data is larger than the number of data points. For example, the number of genes in a single sample is larger than the number of samples for gene expression data.

[6] An analytic method for obtaining double averages of random interactions and physical variables for disordered systems.

is “understanding by creating”. In other words, statistical mechanical informatics has the potential to play a significant role in the constructivist approach to biological phenomena such as protein design.

There are only a few examples of applying statistical mechanical informatics to life phenomena. Some examples include the study of the phase diagram of protein folding using the cavity method [14, Montanari *et al.* 2004], but no other notable studies have been reported. However, many problems in life phenomena, such as interaction networks between amino acids in proteins, chemical reaction networks for intracellular signal transduction, and neural networks, involve the overall state or typical behavior of networks of complex interactions between individual elements. Therefore, the spin glass model on which statistical mechanical informatics relies is well suited for life phenomena.

In fact, researchers have used spin glass theory to analyze the biological systems: associative memory model of neural nets [15–17, Amit *et al.* 1985; Coolen 2001; Hertz 2018](the study [Amit *et al.* 1985] can be regarded as the birth of statistical mechanical informatics) and protein folding transition [18–21, Bryngelson and Wolynes 1987; Garel and Orland 1988; Goldstein *et al.* 1992; Gutin and Shakhnovich 1993].

As described above, we have confirmed that statistical mechanical informatics has an affinity with the constructivist approach to life and has been proven in analyzing several important biological systems, although not in the context of constructivist research. This study follows the methods of statistical mechanical informatics, i.e., Bayesian learning formulation and statistical mechanics analysis. Therefore, this study is different from other studies in that it is close to fundamental scientific theoretical studies, although it is also one of the machine learning-based protein design studies, especially those based on deep learning, which has been rapidly developing in recent years.

1.4 The aim and structure of this doctoral thesis

1.4.1 What should be clarified?

As explained in Sec. 1.2.3, despite the fundamental difficulty that protein design requires performing a dual optimization of conformational space and sequence space over and over again, existing design methods do not do so and basically, only require a single optimization of sequence space for a given structure (henceforth this one-time computation of parameters (output to input) from the data is called ”inverse computation” to distinguish it from the overall inference of the inverse problem itself). Among protein researchers, the question “Why is it possible (or more precisely, why are there so many such proteins) to design a protein even if one skips the inner optimization (structure-space optimization)?” The question “Why is it possible to design proteins even if inner optimization (structure-space optimization) is skipped (precisely, are there many such proteins)?" However, is it not a nontrivial fact that, despite the infinite number of degrees of freedom in protein structure, one can exclude the possibility of inferring sequences that fold into other structures just by looking at information on a given structure? Perhaps there is a hidden problem of protein evolution behind this.

This problem is also relevant to inverse problems in general. Inverse problems generally contain a forward problem. For example, backpropagation[7] in deep learning considers the neural network and its weights as parameters or inputs (note that it is not the input layer or the data to be put into the input layer) and the estimation results as outputs. Then, forward propagation from the input layer to the output layer is a forward problem; adjusting the parameters in one set of forward and backpropagation is the inverse calculation. In backpropagation, one has to repeat the forward problem and inverse calculation many times. That is precisely an example where the inverse problem contains the forward problem inside of it.

From an engineering point of view, this is a huge benefit because a significant reduction in computational complexity allows for very efficient estimation. Moreover, for problems that require multiple iterations of double optimization, the speedup is not on the order of a few times. However, it can be tens or even hundreds of times faster, depending on the problem. The enhancement of the corresponding fundamental science is indispensable for significant progress in engineering, and this acceleration is exactly the correct response. In addition, design methods based on such fundamental problems may lead to "protein function design," which researchers have not yet achieved in protein design. Of course, this is too difficult and a big challenge, but it is precisely at such times that we should go back to the principles instead of thinking of complex algorithms and their combinations for engineering purposes only.

1.4.2 The goal of this doctoral thesis

From the discussion so far, we consider the purpose of this doctoral thesis. It is to consider the question, "Why are some proteins designable even if we skip the internal optimization?" Then, in the spirit of synthetic biology, we propose a novel design method by actually creating proteins (theoretically) and evaluating them quantitatively.

Note that this doctoral thesis also has a fundamental science aspect. However, research with fundamental scientific significance is probably rarely seen in the field

[7] Backpropagation is an optimization algorithm that determines the error with the correct output at the output layer of a neural network and updates the weights of synaptic connections from the output layer to the input layer to output more accurate predictions based on this error. One repeats forward and backpropagation until the error converges.

of protein design or protein engineering, as it is called.

1.4.3 The structure of this doctoral thesis

This chapter is an introduction to the general framework, and in Chapter 2, we summarize the previous studies. In Chapter 3, we formulate the protein design problem using Bayesian learning and statistical mechanics, which is the theoretical framework of this doctoral thesis research. This theory is the basis for all the following chapters. In Chapters 4 and 5, we then run the protein design with a concrete algorithm and evaluate its quantitative accuracy within the scope of computer simulations. Chapter 4 presents the results of the Markov chain Monte Carlo (MCMC) method, and Chapter 5 presents the results of the cavity method. Chapter 6 summarizes the discussion of the theory and results up to that point. Chapter 7 is a summary.

Chapter 2

A review of conventional protein design studies

This chapter begins with a brief review of the general history of previous protein design studies, including experiments, followed by a summary of previous studies by equilibrium statistical mechanics. After introducing each topic, a short discussion of the author's opinion on the fundamental scientific significance of the findings is given. Protein design research by equilibrium statistical mechanics relies on two "design criteria": energy minimization and target probability maximization. In particular, we will see that the latter is precisely the statistical mechanics counterpart of the "double optimization" described in the previous chapter. The description presented in this section is detailed in [22, Shiraki 2019].

2.1 General studies of protein design

2.1.1 Directed evolution: Empirical design

Protein design or protein engineering[1] research began in the late 1980s. The discovery of genetic recombination technology in the 1970s triggered the development of experimental techniques that allowed DNA to be cut and joined at will. The pioneering work in protein design was the proposal of the directed evolution method by Frances Arnold and her colleagues, a method of artificially designing proteins by evolving them in vitro using such genetic engineering techniques [23, 24, Chen and Arnold 1992; Chen and Arnold 1993]. The directed evolution method was the subject of the 2018 Nobel Prize in Chemistry.

[1] Protein engineering is a field of study that aims at mass production and functional modification of proteins using genetic engineering methods. Therefore, the term"protein design" does not necessarily mean exactly the same thing as "protein engineering," but it is used here because the history of protein design research began with protein engineering motivation, as explained below.

Directed evolution follows Darwinian evolution and creates a protein with a desired function by repeatedly evolving the protein through mutation and natural selection by modifying the gene [25, Packer and Liu 2015]. Specifically, the protein (enzyme) is produced by introducing random mutations to DNA into bacteria, selecting only those enzymes that exhibit the desired reaction, and then repeating the cycle of adding random mutations to the corresponding bacterial DNA to design a protein that exhibits the desired reaction. It is a method of directly reproducing Darwinian evolution in the laboratory.

Directional evolution has been successfully applied to drug development. For example, based on directed evolution, the phage display method[2] has been successfully applied to developing antibody drugs. *Phage* is a virus that infects bacteria. When a phage gene is mixed with another gene, the protein encoded by the gene is presented on the surface of the phage, and the phage surface protein is evolved like the directed evolution method to design a phage surface protein that is specific for antibodies [26, Parmley and Smith 1988]. It is the phage display method. Adalimumab, approved by the U.S. Food and Drug Administration in 2002 as an excellent drug for rheumatoid arthritis, is developed using such evolutionary engineering methods.

The fundamental scientific significance of the success of the empirical design, which has been highly successful in engineering, is that it has confirmed that proteins are products of Darwinian evolution. Of course, not all proteins have been designed by directed evolution, so there is still a possibility that there is some framework of evolutionary theory that goes beyond Darwinian evolution. However, the empirical design has not revealed any new nontrivial facts about the relationship between protein structure and sequence.

2.1.2 *Ab initio* design method: Theoretical design

In this section, we introduce Rosetta, a protein design package briefly mentioned in Chapter 1, which was developed by David Baker et al. and is a computer algorithm for computing the various interactions that act on proteins [27, Leman 2020]. However, Rosetta is not a package dedicated solely to protein design; it can also predict conformation, protein-protein docking, and protein-ligand docking.

Rosetta predicts sequences directly without reference to structures in databases, unlike earlier methods that used structure and sequence databases to find sequences similar to the sequence for which the target structure is being designed. That makes it possible to design artificial protein structures that do not exist in nature. Such protein designs are called *de novo* designs. Baker and his group have designed a vast protein design consisting of 60 units with a diameter of about 25 nm [28, Hsia *et al.* 2016], and a dodecahedral conformation of about 40 nm in diameter [29, Bale *et al.* 2016]. The group has confirmed that those conformations are stable under high

[2] Phage display methods were also eligible for the Nobel Prize in Chemistry in 2018.

temperatures of 80°C. They have also succeeded in creating not only proteins but also mimetic viruses containing their coding RNA [30, Betterfield *et al.* 2017].

The outline of Rosetta's methodology is as follows. First, the conformation of the protein or other biopolymer to be designed, called the *Pose*, is identified. Next, *ResidueSelector* determines amino acid residues, and *TaskOperations* optimizes side chains. Then, *Specific Movers* change the structure of the Pose. Finally, the energy of the sequence and structure designed by the above procedure is calculated. That process is the one step of Rosetta's protocol. One performs the Metropolis method [31, Metropolis *et al.* 1953], a typical Markov-chain Monte Carlo (MCMC) method, with the process after RedisueSeletor as one MC step (Fig. 2.1).

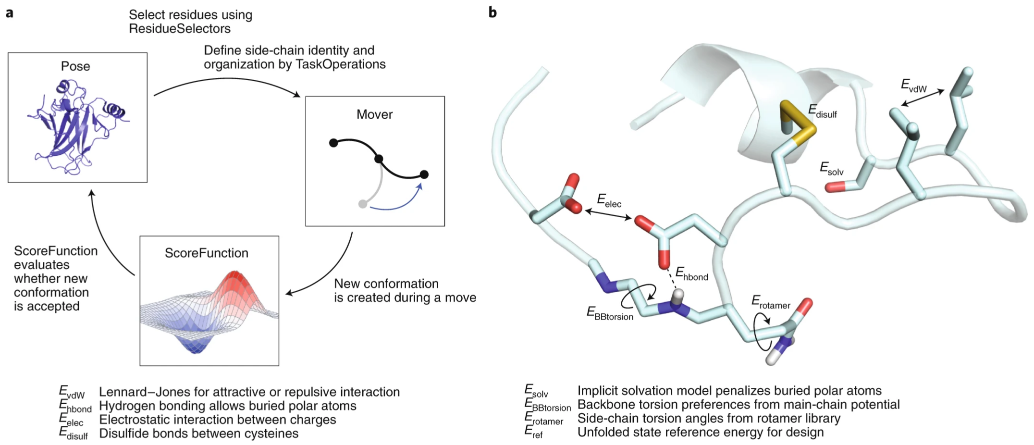
a
Select residues using ResidueSelectors
Define side-chain identity and organization by TaskOperations
Pose
Mover
ScoreFunction
ScoreFunction evaluates whether new conformation is accepted
New conformation is created during a move
E_{vdW} Lennard–Jones for attractive or repulsive interaction
E_{hbond} Hydrogen bonding allows buried polar atoms
E_{elec} Electrostatic interaction between charges
E_{disulf} Disulfide bonds between cysteines
b
E_{vdW}
E_{disulf}
E_{solv}
E_{elec}
E_{hbond}
$E_{\text{BBtorsion}}$
E_{rotamer}
E_{solv} Implicit solvation model penalizes buried polar atoms
$E_{\text{BBtorsion}}$ Backbone torsion preferences from main-chain potential
E_{rotamer} Side-chain torsion angles from rotamer library
E_{ref} Unfolded state reference energy for design

Figure 2.1: Schematic of Rosetta's protocol, reprinted from [27, Leman *et al.* 2020]. a: The representation of the overall protocol. b: The representation of each energy term.

The protein energy used in Rosetta is as follows:

$$E = E_{\text{vdW}} + E_{\text{hbond}} + E_{\text{elec}} + E_{\text{disulf}} + E_{\text{solv}} + E_{\text{BBtorsion}} + E_{\text{rotamer}} + E_{\text{ref}}. \quad (2.1)$$

Each energy term of Eq. (2.1) is following:

- E_{vdW}: Lennard-Jones for attractive or repulsive interaction
- E_{hbond}: Hydrogen bonding allows buried polar atoms
- E_{elec}: Electrostatic interaction between charges

- E_{disulf}: Disulfide bonds between cysteines
- E_{solv}: Implicit solvation model penalizes buried polar atoms
- $E_{\mathrm{BBtorsion}}$: Backbone torsion preferences from main-chain potential
- E_{rotamer}: Side-chain torsion angles from rotamer library
- E_{ref}: Unfolded state reference energy for design

The eighth, average unfolded state reference energy, is the average energy of the protein when it moves through thermal denaturation. The reference state is the energy such that subtracting the energy in that state from the energy of the native state yields an energy more specific to that native structure. The average unfolded state reference energy is the average energy of the ensemble of denatured states [32, Liu and Gong 2012].

Rosetta is a protein design package that is widely used around the world, with many successes, as described above. However, as explained above, its algorithms are rather complex and require specific protein knowledge to understand and use. Thus, for all its engineering successes, Rosetta is not very insightful into questions such as what exactly is essential in the structure-sequence relationship. However, the critical point of Rosetta's computational protocol is that it does not significantly change the main chain structure. As explained in Chapter 1, this point is nontrivial considering the existence of the inner conformational search, which is a fundamental difficulty of protein design.

2.1.3 Design method using deep learning: AI-based design

In recent years, there have been many studies on protein design by deep learning [33, Ding *et al.* 2022]. They are largely based on multilayer neural networks (MNN) [34–36, Li *et al.* 2014; O'Connell *et al.* 2018; Wang *et al.* 2018], convolutional neural networks (CNN) [37–40, Chen *et al.* 2019; Zhang *et al.* 2019; Qi and Zhang 2020; Huang *et al.* 2017], and graph neural networks (GNN) [41–46, Defferrard *et al.* 2016; Kipf and Welling 2016; Velič ković *et al.* 2017; Ingraham *et al.* 2019; Jing *et al.* 2020; Strokach *et al.* 2020]. At the current stage (January 2023), the most accurate method is it ADesign, which is a GNN-based method using AlphaFold, a deep learning method which predicts protein 3D conforamotions with very high accuracy than conventional method [47, Jumper *et al.* 2021]. ADesign uses a large database of more than 200 million protein 3D structures predicted by AlphaFold, AlphaFold database[3]. ADesign has achieved an accuracy of more than 60%. The "accuracy" here is the percentage of correctly predicted residues among the 20 amino acid residues that make up a protein. Therefore, even with deep learning, there is

[3] https://alphafold.ebi.ac.uk/

still much room for improvement. Using Rosetta, the accuracy of this “redesign” is about 30%.

Above accuracy, the percentage of correctly predicted residues among the 20 amino acid residues that make up a protein is also used in Chapter 5 of this doctoral thesis. It should be noted, however, that this accuracy measure for redesign is different from the original criterion for the solution of the protein design problem, which is the sequence as the ground state only for a given target structure. The solution to the protein design problem is not necessarily unique. Rosetta’s redesign accuracy is only 30%, but as discussed in section 2.1.2, Rosetta has successfully designed many *de novo* protein structures.

These deep learning design methods use feature values[4] such as the dihedral angle of the native protein structure and the residue-residue distance. The more elements of the input information, the higher the accuracy.

When one uses deep learning, it is not easy to discuss what deep learning protein design research implies about the principles of protein design because we do not know why the learning converges to a good solution in deep learning. However, if we consider deep learning as an AI that can make perfect predictions by capturing some essential features in the relationship between input and output data, the information on what features are used is essential from a fundamental scientific point of view. However, it is evident that the dihedral angle and the residual period distance are essential information for determining the structure of a protein. Deep learning does not seem to have provided any significant insight into the structure-sequence relationship at this stage. Therefore, there is still much room for elucidating “what is essential in the relationship between structures and sequences?” using fundamental scientific studies.

2.2 Statistical mechanical studies

Since the following sections are descriptions of equilibrium statistical mechanics studies that are intrinsically related to this study, we will exaplain each in more detail. Note that most of the works presented below were done before the Rosetta and deep learning methods discussed in this chapter and thus do not include most of the fundamental scientific questions that the author has emphasized several times so far.

[4] A data element that is considered essential in the relationship between input and output and that influences accuracy. For image recognition, it is the color of the pixels that make up the image data; for speech recognition, it is the sound waveform; for natural language processing, it is the frequent words in a sentence, etc.

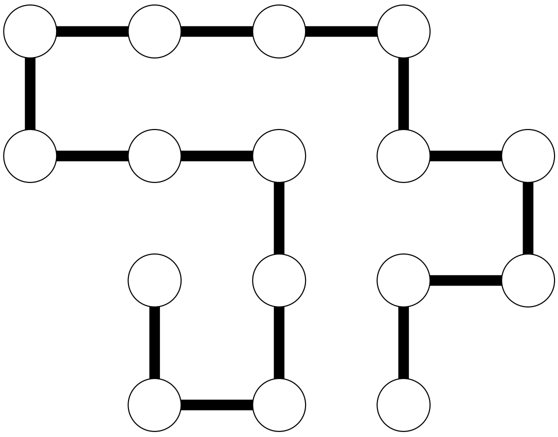

Figure 2.2: An example of 2D HP model with 16 residues . Each site is either H or P, not distinguished here.

2.2.1 The lattice HP model

Proteins are too complex to be treated as they are, and statistical mechanics cannot be applied to them. Therefore, some coarse-grained model is used, especially in theoretical studies. Among these, the simplest and most often analyzed statistically is the lattice HP model proposed by Lau and Dill [48, Lau and Dill 1989]. The lattice HP model (from now on referred to as the HP model) is a Ising model-like model in which the backbone chain of a protein is represented by a lattice-shaped self-avoiding walk, where each lattice point represents an amino acid residue. As the name suggests, the HP model simplifies the 20 amino acids into two types: hydrophobic (hydrophobic) and hydrophilic (polar). Fig. 2.2 shows an example of an HP model.

Compared to the complex realistic protein structures, the HP model may seem overly simplistic. However, lattice models have been used to elucidate many problems such as characterizing the free energy landscape of protein folding [**?**,49, Cieplak and Banavar 2013; Shi *et al.* 2016], an explanation of the cold denaturation [50, Dijk *et al.* 2016], the effect of mutation of amino acid sequence for the native structure [51, 52, Holzgrafe *et al.* 2011; Shi *et al.* 2014], and the analysis of RNA folding energy landscape [53, Chen and Dill 2000]. The minimal model such as the lattice model is, therefore, still adequate for discussing why the natural protein design methods succeed in designing proteins without the conformational search.

We here introduce the normal Hamiltonian of the HP model. We consider N residues $\boldsymbol{\sigma} = \{\sigma_1, \sigma_2, \ldots, \sigma_N | \forall i, \sigma_i = \pm 1\}$ on a lattice position $\boldsymbol{r} = \{r_1, r_2, \ldots, r_N\}$, where $i = 1, 2, \ldots, N$ $\sigma_i = 1$ indicates that the i-th residue is an H residue, and

$\sigma_i = -1$ indicates that it is a P residue.

We assume the energy of the lattice protein is given by

$$H(\boldsymbol{r};\boldsymbol{\sigma}) = \sum_{i<j} U(\sigma_i, \sigma_j)\Delta(r_i - r_j), \tag{2.2}$$

where $U(\sigma_i, \sigma_j)$ denotes the interaction potential between the monomers i and j. We moreover assume the simplest functional form of U: $U(1,1) = \epsilon_1, U(1,-1) = U(-1,1) = \epsilon_2, U(-1,-1) = \epsilon_3$, using two types of interaction set, $(\epsilon_1, \epsilon_2, \epsilon_3) = (-1,0,0)$ and $(-2.3,-1,0)$. The definition of the contact energy $\Delta(r_i - r_j)$ is

$$\Delta(r_i - r_j) = \begin{cases} 1 & \text{if } r_i \text{ and } r_j \text{ contact each other,} \\ 0 & \text{otherwise,} \end{cases} \tag{2.3}$$

where contact between two residues is defined as the case where $|r_i - r_j| = 1$ but $|i - j| \neq 1$. In this model, therefore, the energy given by Eq. (2.2) of denatured conformations is always higher than the energy of compact conformations.

2.2.2 The energy minimization method

Shakhnovich and Gutin proposed a simple design method, which is called energy minimization [54, 55, Shakhnovich and Gutin *Protein Eng.* 1993; Shakhnovich and Gutin *PNAS* 1993;]. As the name implies, it is a method for finding the amino acid sequence that minimizes the energy of a folded protein.

n the studies of statistical mechianical protein design, a native structure $\boldsymbol{R}$ with a sequence $\boldsymbol{\sigma}$ folds with a some kind of probability. The probability is the conditional probability $p(\boldsymbol{R}|\boldsymbol{\sigma})$. Based on the Anfinsen's dogma, the state in which the native structure is realized as its equilibrium state, so $p(\boldsymbol{R}|\boldsymbol{\sigma})$ is expressed as a probability distribution in statistical mechanics. Thus, let β the inverse temperature of a heat bath, i.e., physiological temperature, $p(\boldsymbol{R}|\boldsymbol{\sigma})$ should be

$$p(\boldsymbol{R}\,|\,\boldsymbol{\sigma}) = \frac{e^{-\beta H(\boldsymbol{R};\boldsymbol{\sigma})}}{Z_\beta(\boldsymbol{\sigma})}, \tag{2.4}$$

$$Z_\beta(\boldsymbol{\sigma}) = \sum_{\boldsymbol{r}} e^{-\beta H(\boldsymbol{r};\boldsymbol{\sigma})}. \tag{2.5}$$

Shakhnovich and Gutin theoretically derived by a random energy model that the energy of the native state of a protein is significantly lower than other denatured states and that these denatured states are all similar and self-averaging, independent from the sequence [21, Gutin and Shakhnovich 1993]. Therefore, Eq. (2.5) is almost constant regardless of the sequence. Therefore, the sequence with the highest probability Eq. (2.4) is the one with the lowest Hamiltonian $H(\boldsymbol{R};\boldsymbol{\sigma})$.

However, this would lead to an obvious solution in which every residue is an H-residue, regardless of its structure. Shakhnovich and Gutin, therefore, determined the ratio of H-residues and P-residues to be close to those of real proteins, and minimized the energy by MCMC with the constraint that the number of contacts between H-residues and P-residues is less than or equal to 1.

Shakhnovich and Gutin did not test the accuracy of the design in their research [55, Shakhnovich and Gutin 1993]. They only compared the probability (2.4) of the designed sequence and the random sequence that can have a non-degenerate ground state. They also showed that the probability (2.4) of designed sequence is higher even in the high-T region. However, another group (Deutsch and Kurosuky, we will review their study next) has compared the design accuracy. According to their results, the accuracy is about 50% even for small 2D HP models with $N = 10$ to 18 residues. The design accuracy is the percentage of structures that achieve a non-degenerate ground state out of all possible structures.

Thus, the design method of Shakhnovich and Gutin (from now on referred to as the SG method), while theoretically suggestive, is not reliable from an engineering standpoint. However, the SG method is qualitatively convincing because it can generate sequences with a high probability of folding to a given structure, even among sequences with no degeneracy. It means that we can confirm that $Z_\beta(\boldsymbol{\sigma})$ is almost constant regardless of the sequence, which is the assumption in the SG method described above. It is what is meant by making a fundamental scientific discovery through a synthetic biology approach, as explained in Chapter 1. Such a result is difficult to achieve in a study such as protein folding simulation.

2.2.3 The MTP-based methods

2.2.3.1 Approximation of free energy

Deutsch and Kurosuky proposed a fundamentally different method from the SG method. It is target probability maximization (MTP), which maximizes Eq.(2.4) (from now on, called the target probability). In other words, this is maximum likelihood estimation in mathematical statistics.

In maximum likelihood estimation, one estimates the parameter θ that maximizes the probability $p(D|\theta)$ (called likelihood function) when the data D with probability $p(D|\theta)$. That is, the estimated value of θ $\hat{\theta}$ is given by

$$\hat{\theta} = \arg\max_{\theta} p(D|\theta). \tag{2.6}$$

Therefore, the MTP is a maximum likelihood estimation for Eq. (2.4). The maximum likelihood estimate $\boldsymbol{\sigma}_{\mathrm{MTP}}$ is given by

$$\boldsymbol{\sigma}_{\mathrm{MTP}} = \arg\max_{\boldsymbol{\sigma}} p(\boldsymbol{R}|\boldsymbol{\sigma}). \tag{2.7}$$

Taking the natural logarithm of both sides of Eq. (2.3), we obtain as follows:

$$-\frac{1}{\beta}\log p(\boldsymbol{R}|\boldsymbol{\sigma}) = H(\boldsymbol{R};\boldsymbol{\sigma}) - F_\beta(\boldsymbol{\sigma}), \tag{2.8}$$

where $F_\beta(\boldsymbol{\sigma}) := (1/\beta)\log Z_\beta(\boldsymbol{\sigma})$. Deutsch and Kurosuky expanded this free energy to the lowest order term in the cumulant expansion, i.e., the average energy: $\langle H(\boldsymbol{r};\boldsymbol{\sigma})\rangle_{\boldsymbol{r}}$, and then minimized the right-hand side of Eq. (2.8). The average $\langle\cdot\rangle_{\boldsymbol{r}}$ means the expectation value under

$$p(\boldsymbol{r}\,|\,\boldsymbol{\sigma}) \quad = \quad \frac{e^{-\beta H(\boldsymbol{r};\boldsymbol{\sigma})}}{Z_\beta(\boldsymbol{\sigma})}. \tag{2.9}$$

That is, Deutsch and Kurosky's method computes $H(\boldsymbol{R};\boldsymbol{\sigma}) - \langle H(\boldsymbol{r};\boldsymbol{\sigma})\rangle_{\boldsymbol{r}}$ for each sequence sample, and the sequence that maximizes it is the correct one. They used simulated annealing to optimize the sequence space.

The design results were generally more than 60% for small 2D HP models with $N = 10$ to 18 residues, where all lattice conformation patterns can be enumerated so that one can computes $\langle H(\boldsymbol{r};\boldsymbol{\sigma})\rangle_{\boldsymbol{r}}$ rigorously. For $N = 12$, the accuracy was particularly good at about 90%. In addition, the method outperformed the SG method in all cases.

The MTP criterion is plausible as a criterion for protein design. However, to implement the MTP method, one has to calculate the partition function $Z_\beta(\boldsymbol{\sigma})$ (hereafter called the conformational partition function). This is a hopeless task given the astronomical patterns of native protein structures. This difficulty is the statistical mechanical counterpart of the difficulty of protein design described in the previous section, the exhaustive conformational search inside the protein. Therefore, all of the methods described below that rely on the MTP criterion need a long time to design, and in addition, they have only been successful for HP models with a maximum number of residues of $N = 50$.

2.2.3.2 Dual Monte Carlo method

Seno *et al.* proposed the dual Monte Carlo method, in which one optimizes not only sequence space but also conformational space [56, Seno et al. 1996]. However, it is very difficult to carry out MCMC sampling in the conformational space of the HP model. Seno *et al.* therefore performed inportance samling, which is one of the weighted sampling for the conformational space. They also carried out simulated annealing for the sequence space.

As a result, the dual MC method succeeded to obtain all non-degenerate ground state for designable 456 conformations of 2D $N = 16$ HP model.

2.2.3.3 Multi sequence Monte Carlo method

In contrast to the above two studies, Irbakäck *et al.* proposed a method called the multi sequence MC method, in which the sequence space is also MC sampled at each MC step of the conformational space [57, 58, Irbäck *et al.* 1998; Irbäck *et al.* 1999]. The main feature of the multi sequence MC method is that it does not search for a solution from the total number of patterns in the sequence space of 2^N, but rather it samples the target structure according to some criteria. With this sequence elimination step, they achieved a significant speed-up over other methods.

Their elimination method is as follows. Rather than dealing directly with Eq. (2.4), they considered the joint distribution of the structures $\boldsymbol{r}$ and $\boldsymbol{\sigma}$ given below

$$p(\boldsymbol{r}, \boldsymbol{\sigma}) = \frac{1}{Z} \exp[-g(\boldsymbol{\sigma}) - \beta H(\boldsymbol{r}; \boldsymbol{\sigma})], \tag{2.10}$$

$$Z = \sum_{\boldsymbol{\sigma}} \exp[-g(\boldsymbol{\sigma})] Z_\beta(\boldsymbol{\sigma}). \tag{2.11}$$

The function $g(\boldsymbol{\sigma})$ is a controlling function for probability of $\boldsymbol{\sigma}$. They assumed $g(\boldsymbol{\sigma})$ as follows:

$$g(\boldsymbol{\sigma}) = -\beta H(\boldsymbol{R}; \boldsymbol{\sigma}). \tag{2.12}$$

The definition (2.12) is a reasonable assumption: a sequence which likely to fold into $\boldsymbol{R}$ is likely to appear. Then, the target probability $p(\boldsymbol{R}|\boldsymbol{\sigma})$ becomes

$$p(\boldsymbol{R}|\boldsymbol{\sigma}) = \frac{p(\boldsymbol{R}; \boldsymbol{\sigma})}{p(\boldsymbol{\sigma})} = \frac{1}{Zp(\boldsymbol{\sigma})}. \tag{2.13}$$

Eq. (2.12) shows that maximizing $p(\boldsymbol{R}|\boldsymbol{\sigma})$ and minimizing $p(\boldsymbol{\sigma})$ are equivalent. Therefore, we can search the sequence space efficiently if we can eliminate large $p(\boldsymbol{\sigma})$ sequences. The $p(\boldsymbol{\sigma})$ can be obtained by marginalizing Eq. (2.10),

$$p(\boldsymbol{\sigma}) = \frac{1}{Z} \exp[-g(\boldsymbol{\sigma})] Z_\beta(\boldsymbol{\sigma}). \tag{2.14}$$

However, since Eq.(2.14) contains $Z_\beta(\boldsymbol{\sigma})$, the sequence elimination by $p(\boldsymbol{\sigma})$ is computationally difficult. Therefore, Irbäck *et al.* also devised another sequence elimination based on the Hamiltonian $H(\boldsymbol{R}; \boldsymbol{\sigma})$ and $H(\boldsymbol{r}; \boldsymbol{\sigma})$ of the target structure $\boldsymbol{R}$ and any structure $\boldsymbol{r}$ sampled in the dual MC step of structure space and sequence space. If a sequence $\boldsymbol{\sigma}$ makes any structure (other than $\boldsymbol{R}$) lower or equal energy state than tagerget structure as follows:

$$H(\boldsymbol{r}; \boldsymbol{\sigma}) \leq H(\boldsymbol{R}; \boldsymbol{\sigma}), \tag{2.15}$$

one eliminates such sequence satisfying Eq. (2.15). This elimination method can get desired sequence. The advantage of the latter elimination method using Eq. (2.15) is that Eq. (2.15) does not have $Z_\beta(\boldsymbol{\sigma})$.

Irbäck *et al.* tested the design accuracy of $N = 16$ and $N = 18$ HP models using the SG method (Energy minimization), the approximation of free energy $F_\beta(\boldsymbol{\sigma})$ by Deutsch and Kurosky, the dual MC method by Seno et al. , and the multi sequence MC method to compare the performance between these methods. The details of the results are given in Table 2 of the original paper [58, Irbäck *et al.* 1999]. The summery of the result is as follows. For $N = 16$, the design accuracies are 87% (Energy minimization), 70% (Approximation of $F_\beta(\boldsymbol{\sigma})$), 100% (Dual MC), and 100% (Multi sequence MC), respectively. The computation times are $\mathcal{O}(1)$ (Energy minimization), $\mathcal{O}(1)$ (Approximation of $F_\beta(\boldsymbol{\sigma})$), $\mathcal{O}(10^3)$ (Dual MC), and $\mathcal{O}(10)$ (Multi sequence MC), respectively. These numbers represent the CPU seconds (using DEC Alpha 200). The important point is the comparison with Dual MC. Although the percentage of correct answers is the same for both Dual MC and Multi sequence MC, the computation time for the latter is significantly reduced to 1/100 of the former.

2.2.3.4 Dynamical approach using the Boltzmann machine learning

The study by Iba *et al.* was the first in the world to apply the Boltzmann machine learning method to protein design [59, 60, Iba *et al.* 1998; Tokita *et al.* 2000]. They proposed a method to minimize a cost function by converting the sequence variable σ_i to the continuous value m_i and adding a term that prevents m_i from converging to a non-integer value. Using $p(\boldsymbol{R}|\boldsymbol{m})$ with $\boldsymbol{\sigma}$ replaced by $\boldsymbol{m}$ ($\boldsymbol{m} := \{m_1, m_2, \cdots, m_N\}$) in the target probability, its cost function is given by

$$V(\boldsymbol{m}) = -\log p(\boldsymbol{R}|\boldsymbol{m}) + \frac{\lambda}{4}\sum_i (m_i^2 - 1)^2, \tag{2.16}$$

where λ is the controlling parameter of the penalty term. If one differentiates Eq. (2.16) with respect to m_i, the dynamical equation to get optimal sequence is obtained as follows:

$$\tau \frac{dm_i}{dt} = -\frac{\partial V}{\partial m_i} = f_i(\beta, \boldsymbol{m}) - \lambda m_i(m_i^2 - 1), \tag{2.17}$$

$$f_i(\beta, \boldsymbol{m}) := \beta \sum_j \frac{\partial U(m_i, m_j)}{\partial m_i}[\Delta(R_i - R_j) - \langle \Delta(r_i - r_j)\rangle_{\boldsymbol{r}}]. \tag{2.18}$$

In Eq. (2.17), t denotes the pseudo time and τ denotes the time constant that determine the scale of the dynamics. The fixed point of the dynamical system (2.17) corresponds to the design solution. In Eq. (2.18), $\langle \Delta(r_i - r_j)\rangle_{\boldsymbol{r}} = \sum_r \Delta(r_i - r_j) p(\boldsymbol{r}|\boldsymbol{m})$.

Boltzmann machine is a machine that learns interactions to reproduce neural firing patterns given as data. In this case, Eqs. (2.17) and (2.18) learn the continuous amino acid sequence that best fits the target structure $\boldsymbol{R}$ given as data. Thus, this method is a Boltzmann machine learning for protein design method. Iba *et al.* named Eq. (2.16) the “design equation”.

We will not give a detailed explanation, but they tested the method on five randomly selected conformations of a cubic HP model with $N = 3 \times 3 \times 3$, where one can enumerate all conformational patterns. The success rate is the percentage of the number of correct sequences generated by the simulation from several initial sequences. They achieved a high percentage of 82% correct for one structure, but at most 30% correct for the other four structures. However, they also obtained the percentage of correct answers per "unit" of time, which is calculated by dividing the percentage of correct answers by the time required for a single conformational search ($\langle\Delta(r_i - r_j)\rangle_{\boldsymbol{r}} = \sum_r \Delta(r_i - r_j)p(\boldsymbol{r}|\boldsymbol{m})$). The rate of correct answers per time is about ten times higher than the method by Deutsch and Kurosky.

This method was the first in the world to use machine learning methods in protein design. By definition, the MTP-based methods can always find a solution if an appropriate algorithm is used and a long time is spent on it. Therefore, the "number of correct answers per time" is an important indicator in evaluating the performance of algorithms. In Chapter 4, we also use the measure to test the performance our method.

2.2.4 Other studies

After 2000, the competition for accuracy in protein design by proposing equilibrium statistical mechanics algorithms for the HP model, as described in the previous sections, has been slowing down. In particular, the momentum of research on MTP-based methods declined. The reason may be that the computational complexity of the partition function $Z_\beta(\boldsymbol{\sigma})$ for the conformational space is so large that it is completely impractical for realistic proteins. However, there are some important studies, which are briefly summarized below.

Jiao *et al.* proposed a design method based on relative entropy minimization between $p(\boldsymbol{R}|\boldsymbol{\sigma})$ and the probability of occurrence of a structure $p(\boldsymbol{r}|\boldsymbol{\sigma})$ [61,62, Wang *et al.* 2004; Jiao *et al.* 2006]. Since relative entropy is (-1) times the Kullback-Leibrer divergence, this means that this is a Boltzmann machine learning method[5] In other words, the general mathematical framework is the same as the design equation method described above, but Jiao et al. do not make the amino acid residue variables continuous. They used the HNP model in which the 20 amino acids are divided into three classes, H (Hydrophobic), N (Neutral), and P (Polar), and defined the percentage of correct answers as how well the three classes of HNP in the real protein sequence are guessed. In other words, the accuracy of a classifier that classifies 20 amino acid residues into the three classes of HNP was measured. As a result, they designed 20 proteins with a correct answer rate ranging from 40% to 55% [62, Jiao *et al.* 2006]. Although it is difficult to discuss a correct comparison because such a measure of design accuracy is not widely used and lacks a benchmark, it can be said that the method still needs improvement because the accuracy is expected to decrease as the number of amino acid types is increased.

Coluzza *et al.* proposed a design method by energy minimization with conditions

[5] the signs of relative entropy and Kullback - Leibrer divergence are often explained without distinguishing between them, and indeed (In fact, what Jiao et al. call relative entropy is generally regarded as the Kullback-Leibler divergence, so their study is not about maximizing relative entropy but minimizing it.

that ensure heterogeneity of sequence types [63, 64, Coluzza *et al.* 2003; Coluzza 2011]. A condition that ensures sequence type heterogeneity is the quantity obtained by dividing the permutation of the total number of residues N by the product of the permutations of the number of residues $n_a, a = 1, 2, \cdots, 20$ for each amino acid type:

$$N_p = \frac{N!}{n_1!n_2!\cdots n_{20}!}. \tag{2.19}$$

Coluzza devised his own off-lattice model of proteins and used the Hamiltonian to design a small realistic protein with about 50 to 70 residues, which approximately reproduces the 3D structure. The model was found to approximately reproduce the 3D structure of the protein [64, Coluzza 2011].

Bianco *et al.* considered that water molecules around the protein and in the inner pocket contribute to the stabilization of the native structure of the protein [65, 66, Bianco *et al.* 2012; Bianco *et al.* 2015], they proposed a design method considering those effects. They considered the enthalpy added to the Hamiltonian of the volume change of the protein due to the entry of water molecules inside the protein during folding. They then proposed a design method by enthalpy minimization that minimizes the expected value of the enthalpy described above for the configuration of water molecules and by maximizing the enthalpy difference between the denatured and native states [67, Bianco *et al.* 2017]. Although they did not carry out an explicit test of their design method because the purpose of their study was not to evaluate the accuracy of their design method, they did confirm that the designed lattice proteins retain stability close to that of real proteins. Notably, their results show that the stability of the protein is reduced when the segregation of hydrophilic residues on the protein surface and hydrophobic residues in the inner core is too high. They found that the region of temperature T and pressure P where the protein can exist stably is narrowed under such conditions.

2.3 Summary

As the above reviews show, empirical, theoretical, and AI-based protein design have been successful in designing many realistic proteins. In contrast, statistical mechanics studies have not succeeded in designing proteins of realistic sizes, even for coarse-grained models such as the HP model. Therefore, the fundamental scientific significance of protein design studies by statistical mechanics is increasing. That will lead to the solution of significant problems such as the protein function design and the design of intrinsically disordered proteins[6] (IDP), considering that engineering breakthroughs require the development of basic science.

[6] Intrinsically disordered proteins are proteins that do not have a specific conformation, which has attracted much attention in recent years.

In statistical mechanical protein design studies, it is clear from the previous reviews that, from an engineering point of view, the energy minimization design method has relatively low accuracy with a short computation time. In contrast, the MTP-based design method has relatively high accuracy with long computation time[7].

However, if we classify which of the three methods introduced in section 2.2.4 relies on which design criterion, we can say that the first one, relative entropy minimization, is based on the MTP. The other two on energy minimization[8]. Thus, the statistical mechanical protein design research is gradually returning to energy minimization by SG from the MTP-based methods. The main reason is that the MTP-based methods involve $Z_\beta(\boldsymbol{\sigma})$, which requires a vast amount of computation. However, not only such negative reasons, but also a positive reason may be that energy minimization by SG is sufficient to some extent if the diversity of amino acid residues is well secured or the effect of water molecules bound to the protein is introduced explicitly.

However, there is no discussion in the work of Coluzza *et al.* and Bianco *et al.* about why energy minimization produces good results. Therefore, it is still relevant today to test statistical mechanics design methods on HP models in a way that explains why, and to evaluate their accuracy more exhaustively. Such a study may also explain why Rosetta is able to correctly design many protein structures, given that Rosetta's protocol is similar to the energy minimization.

[7] In the comparison of accuracy in the study by Irbäck et al. reviewed in 2.3.4, the design accuracy is 87% for energy minimization and 70% for $F_\beta(\boldsymbol{\sigma})$ approximation, so the accuracy relationship described here is reversed. This is because, unlike other MTP-based methods, $F_\beta(\boldsymbol{\sigma})$ is only an approximation, and its accuracy is considered to be affected by the accuracy of $F_\beta(\boldsymbol{\sigma})$

[8] Bianco *et al.*'s design method is based on energy minimization because it minimizes the expected value of enthalpy but not the expected value in the configuration of the structure.

Chapter 3

Bayesian formulation

In this chapter, we formulate the protein design problem by Bayesian learning. Specifically, we examine what the three important probability distributions in Bayesian learning, i.e., likelihood function, prior distribution, and posterior distribution, are for protein design. In particular, we propose our own prior distribution, a statistical mechanics hypothesis on the evolution of proteins. The formulation presented in this chapter will be used in exactly the same way in the subsequent chapters 4 and 5.

3.1 Bayesian learning

In this section, we explain Bayesian learning shortly. Bayesian learning is a theory of machine learning based on Bayesian statistics. Although it is also called Bayesian statistics or Bayesian machine learning, we use the term Bayesian learning in this doctoral thesis.

Bayesian learning considers that data D is generated according to a conditional probability distribution $p(D|\theta)$ under an observed parameter θ. It is called the likelihood function. In addition, Bayesian learning assumes a prior probability distribution $p(\theta)$ for the parameter θ, unlike the usual frequentist statistics. It is called the prior distribution, or more simply, prior. We can obtain $p(\theta|D)$, which is "the probability of the parameter θ given the observed data D," from the following Bayes' theorem. When θ is a continuous variable, Bayes' theorem is given by

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{\int p(D|\theta)p(\theta)d\theta}. \tag{3.1}$$

The left-hand side of Eq. (3.1), $p(\theta|D)$, is called the posterior distribution, or more simply, posterior. Bayesian learning estimates the true parameters from this posterior, or predicts unobserved data.

3.2 Why do we use Bayesian learning?

The main reasons for using Bayesian learning instead of other machine learning methods are as follows:

1. Easy to incorporate domain knowledge for each field.

2. It is similar to the cognitive and reasoning processes of living organisms.

3. Can naturally apply statistical mechanical informatics methods.

1. is because Bayesian learning allows humans to freely determine probability distributions for variables and parameters appearing in the problem in a patchwork fashion. There does not have to be a single prior distribution, and multiple distributions can be used to reflect existing knowledge or hypotheses, depending on the data generation process.

Concerning 2. the free energy principle, for example, a hypothesis about the brain's perceptual and inferential processes, states that "the brain unconsciously infers the data-generating process behind sensory input to minimize surprises from environmental information [68, Friston *et al.* 2006]. That is exactly in line with the Bayesian learning scheme described above. The process by which cells perceive and correctly respond to information from the outside world is similar. In quantitative studies, Bayesian inference has been used to decode information from noisy receptor signals [69, Kobayashi 2010].

As for 3., in statistical mechanical informatics, it is a standard practice to formulate the problem using Bayesian statistics and perform statistical mechanics analysis. This ease of application of statistical mechanical analysis is a great advantage when considering the integration of physics and machine learning, as described in Chapter 1.

3.3 Bayesian learning for protein design

In the following, we will apply Bayesian learning to protein design. Since this formulation is not dependent on any particular protein model, we will describe the case with the HP model described in the previous section after explaining the model-independent formulation for a specific protein.

The observed data correspond to the native structure $\boldsymbol{R}$ of a given protein whose sequence is unknown, and the parameters correspond to the amino acid sequence $\boldsymbol{\sigma}$ that realizes it. The structural data $\boldsymbol{R}$ is a set of coordinate variables that allow us to identify the protein backbone structure. The amino acid sequence $\boldsymbol{\sigma}$ is a set of variables representing each amino acid type. The target structure $\boldsymbol{R}$ is considered to be realized by $p(\boldsymbol{R}|\boldsymbol{\sigma})$ given the amino acid sequence $\boldsymbol{\sigma}$. If the prior of the amino acid sequence is $p(\boldsymbol{\sigma})$, the probability of the sequence $\boldsymbol{\sigma}$ given the target structure $\boldsymbol{R}$ as observed data, i.e. the posterior $p(\boldsymbol{\sigma}|\boldsymbol{R})$, can be expressed by the following formula from Bayes' theorem:

$$p(\boldsymbol{\sigma}|\boldsymbol{R}) = \frac{p(\boldsymbol{R}|\boldsymbol{\sigma})p(\boldsymbol{\sigma})}{\sum_{\boldsymbol{\sigma}} p(\boldsymbol{R}|\boldsymbol{\sigma})p(\boldsymbol{\sigma})}. \tag{3.2}$$

The next step is to obtain posterior $p(\boldsymbol{R}|\boldsymbol{\sigma})$ by assuming a likelihood function $p(\boldsymbol{R}|\boldsymbol{\sigma})$ and a prior $p(\boldsymbol{R}|\boldsymbol{\sigma})$.

First, what is the form of the likelihood function $p(\boldsymbol{R}|\boldsymbol{\sigma})$? Following the protein design study by statistical mechanics introduced in the previous chapter, we consider the following canonical distribution.

$$p(\boldsymbol{R}\,|\,\boldsymbol{\sigma}) = \frac{e^{-\beta H(\boldsymbol{R},\boldsymbol{\sigma};\mu)}}{Z_{\beta,\mu}(\boldsymbol{\sigma})}, \tag{3.3}$$

$$Z_{\beta,\mu}(\boldsymbol{\sigma}) = \sum_{\boldsymbol{r}} e^{-\beta H(\boldsymbol{r},\boldsymbol{\sigma};\mu)}. \tag{3.4}$$

As a Hamiltonian $H(\boldsymbol{R},\boldsymbol{\sigma};\mu)$, we consider one that consists of the following interactions between amino acid residues and between amino acid residues and water molecules surrounding the protein. The Hamiltonian of a protein with native structure $\boldsymbol{r}$ amino acid sequence $\boldsymbol{\sigma}$ is expressed as follows:

$$H(\boldsymbol{r},\boldsymbol{\sigma};\mu) = H_{\mathrm{rr}}(\boldsymbol{r},\boldsymbol{\sigma}) + \mu H_{\mathrm{rw}}(\boldsymbol{\sigma}). \tag{3.5}$$

The first term represents the residue-residue interaction, and the second represents the interaction between amino acid residues and water molecules. The parameterμ is the chemical potential of water. The Hamiltonian is such that the larger μ is, the more bound states with water and the smaller μ is, the less bound states with water.

This Hamiltonian is an energy function of the HP model Hamiltonian Eq. (2.2) with the addition of the interaction with water. The solvation term was added because the entropy related to the arrangement of water molecules bound to proteins has been driving protein folding since the series of statistical protein design studies introduced in the previous chapter, mainly in the late 1990s [70, Kinoshita 2009]. In addition to Eq. (3.5), a Hamiltonian that includes the effect of water on the interaction between water molecules in bulk and on the surface of the protein has been developed [65–67, Bianco *et al.* 2012; Bianco and Franzese 2015; Bianco *et al.* 2017]. In this study, however, we use a simpler Hamiltonian. We call the second term in Eq. (3.5) the water effect, but the first term in (2.5) also reflects the water effect since the hydrophobic interaction contains the effect of repulsion from water in the first term. Therefore, Eq. (3.5) can be considered as double counting the effect of water. However, recent studies have theoretically shown that the attraction between hydrophobic residues which is independent of the hydrophobic effect, is more effective for the stabilization than the hydrophobic effect in protein folding [71, Sumi and Imamura]. Therefore, we consider the first term in Eq. (3.5) to be an attraction independent of the hydrophobic effect between amino acid residues.

In the case of the HP model, the Hamiltonian Eq. (3.5) becomes

$$H(\boldsymbol{r}, \boldsymbol{\sigma}; \mu) = -\sum_{i<j} \sigma_i \sigma_j \Delta(r_i - r_j) - \mu \sum_i (1 - \sigma_i). \tag{3.6}$$

Here, $\sigma_i = 1$ indicates that the i-th residue is an H-residue, and $\sigma_i = 0$ indicates that it is a P-residue. Because σ_i is a P-residue, $\sigma_i = 0$, Eq. (3.6) is a modified Hamiltonian of the original Hamiltonian of the lattice HP model by adding the interaction between P-residues and water. In this study, for simplicity, it is assumed that the residues interact with water molecules only in the P-residue.

The next step is the prior $p(\boldsymbol{\sigma})$. The formulation of $p(\boldsymbol{\sigma})$ is a highly non-trivial problem compared to the likelihood $p(\boldsymbol{R}|\boldsymbol{\sigma})$. That is because, since the amino acid sequence is determined from the base sequence written in DNA, the question is the same as "What bias does the genetic information have in the genotypic space as a result of (or in the process of) the evolution of life?

We consider the following hypothesis for this problem. That is, the amino acid sequences that remain in nature, i.e., that have evolved, lower the following free energy:

$$F_{\beta,\mu}(\boldsymbol{\sigma}) = -\frac{1}{\beta} \log Z_{\beta,\mu}(\boldsymbol{\sigma}). \tag{3.7}$$

This hypothesis is identical to that sequences (genes) have evolved with Eq. (3.6) as a kind of fitness. The idea that the amino acid sequences that makeup proteins are a special kind of evolution among all sequence patterns is shared by many protein researchers. However, what is special about such sequences as opposed to random sequences is still unknown. Our hypothesis answers this question by asserting that such special sequences are those with low free energy (3.6). We propose a protein design method based on this hypothesis in this study. In Section 6.2, we present an analysis to verify this hypothesis.

We reflect the above hypothesis in the prior $p(\boldsymbol{\sigma})$. Since the lower the free energy $F_{\beta,\mu}(\boldsymbol{\sigma})$ is higher the partition function $Z_{\beta,\mu}(\boldsymbol{\sigma})$ is high, we propose $p(\boldsymbol{\sigma})$ as follows:

$$p(\boldsymbol{\sigma}) = \frac{Z_{\beta_p,\mu_p}(\boldsymbol{\sigma})}{\Xi_{\beta_p,\mu_p}}, \tag{3.8}$$

$$\Xi_{\beta_p,\mu_p} = \sum_{\boldsymbol{\sigma}} \sum_{\boldsymbol{r}} e^{-\beta_p H(\boldsymbol{r};\boldsymbol{\sigma};\mu_p)}. \tag{3.9}$$

It is clear that Eq. (3.7) satisfies the normalization condition if summed over $\sum_{\boldsymbol{\sigma}}$. Eq. (3.9) is the partition function that sums over both conformational and sequence space and is constant independent of conformation and sequence.

Substituting Eq. (3.2) and Eq. (3.7) into the Eq. (3.1), we obtain the following posterior:

$$
\begin{aligned}
p(\boldsymbol{\sigma}|\boldsymbol{R}) &= \frac{p(\boldsymbol{R}|\boldsymbol{\sigma})p(\boldsymbol{\sigma})}{\sum_{\boldsymbol{\sigma}} p(\boldsymbol{R}|\boldsymbol{\sigma})p(\boldsymbol{\sigma})} && (3.10)\\
&\propto \frac{e^{-\beta H(\boldsymbol{R},\boldsymbol{\sigma};\mu)}}{Z_{\beta,\mu}(\boldsymbol{\sigma})} \cdot \frac{Z_{\beta_p,\mu_p}(\boldsymbol{\sigma})}{\Xi_{\beta_p,\mu_p}}. && (3.11)
\end{aligned}
$$

If $\beta_p = \beta$ and $\mu_p = \mu$, then the denominator and numerator $Z_{\beta,\mu}(\boldsymbol{\sigma})$ in Eq. (3.10) cancel each other out. The partition function Ξ_{β_p,μ_p} does not depend on the sequence $\boldsymbol{\sigma}$, so it cancels with the one appearing in the normalization constant in Eq. (3.9). Consequently, the following posterior distribution is obtained:

$$
\begin{aligned}
p(\boldsymbol{\sigma}|\boldsymbol{R}) &= \frac{e^{-\beta H(\boldsymbol{R},\boldsymbol{\sigma};\mu)}}{Z(\boldsymbol{R};\beta,\mu)}, && (3.12)\\
Z_{\beta,\mu}(\boldsymbol{R}) &= \sum_{\boldsymbol{\sigma}} e^{-\beta H(\boldsymbol{R},\boldsymbol{\sigma};\mu)}. && (3.13)
\end{aligned}
$$

The new partition function $Z_{\beta,\mu}(\boldsymbol{R})$ is the sum of Boltzmann factors $e^{-\beta H(\boldsymbol{R},\boldsymbol{\sigma};\mu)}$ for all sequences, with structure fixed to the target structure. An important point in the derivation of the posterior (3.12) is that the partition function $Z_{\beta,\mu}(\boldsymbol{\sigma})$ cancels and does not appear in the results. It avoids the computational difficulties of the MTP-based method introduced in the previous chapter. In other words, Eq. (3.8) not only has the nontrivial claim that the native protein sequence is the lower one in Eq. (3.7) but also has the extremely large engineering benefit of omitting the exhaustive conformational search corresponding to the difficulty in the protein design problem. We can perform some estimation methods to infer the sequence from the posterior (3.12), such as MCMC. The computation of the partition function (3.13) is much easier than $Z_{\beta,\mu}(\boldsymbol{\sigma})$.

The design method we propose is very similar to the SG method described in 2.2.2 in that it does not include the partition function $Z_{\beta,\mu}(\boldsymbol{\sigma})$, except for the Hamiltonian form and the μ adjustment. This point is discussed in Chapter 6.

3.4 Parameters

Although the temperature should essentially be the same as the physiological temperature, we set β to $\beta = 10$. We chose such a low temperature in order to ensure the ground state as explained in 1.2.1.

The optimal value $\mu = \mu^*$ for the water chemical potential μ is determined by design many times with different values in this study, and is the value that achieved the most accurate design. In addition, $\mu^* = \mu_p$. The criterion for the optimization of μ is not obvious. For example, minimizing the following free energy:

$$F_{\beta,\mu}(\boldsymbol{R}) = -\frac{1}{\beta}\log Z_{\beta,\mu}(\boldsymbol{R}), \tag{3.14}$$

is one promising candidate for this criterion. However, it is not clear whether minimizing the free energy (3.14) maximizes the design accuracy. The free energy when the normalization constant of Bayes' theorem as in the free energy (3.14) is sometimes referred to as *Bayes free energy*.

Chapter 4

Sequence estimation by MCMC

We show the design results of our protein design method in this and the following chapters. In this chapter, we use MCMC to estimate the amino acid sequence and test it to relatively small 2D lattice proteins. The results shown in this chapter have been published in the article [72, Takahashi *et al.* 2021].

4.1 Applying to the posterior

In the previous section, we obtained the posterior distribution $p(\boldsymbol{\sigma}|\boldsymbol{R})$. The final step is to obtain an optimal HP sequence using Eq. (3.12). Nevertheless, the exact calculation of Eq. (3.13) is difficult if the number of residues N is large. We thus utilize one of the simplest MCMC methods, Gibbs sampling (GS) [73, S. Geman and D. Geman 1984]. One obtains realized values of random variables from the conditional probability distribution with random variables other than the random variable of interest in GS. In this method, the sampling probability of σ_i of each Monte Carlo step (MCS) is a conditional probability of σ_i given other random variables. We thus obtain the following sampling probability. Accordingly, the sampling probability of an H residue ($\sigma_i = 1$) or P residue ($\sigma_i = 0$) is given by

$$p(\sigma_i = \pm 1|\boldsymbol{R};\boldsymbol{\sigma}_{\setminus i}) = \frac{1}{1+\mathrm{e}^{\pm\beta\{\Delta E_i(\boldsymbol{R};\boldsymbol{\sigma})+\mu\}}}, \tag{4.1}$$

where $\boldsymbol{\sigma}_{\setminus i} := \{\sigma_1, \ldots, \sigma_{i-1}, \sigma_{i+1}, \ldots, \sigma_N\}$, a vector of all random variables of residues except for the i-th residue σ_i, and the double signs correspond. Let $\Delta E_i(\boldsymbol{R};\boldsymbol{\sigma}) := \sum_{j\in n(i)} (U(1,\sigma_j) - U(0,\sigma_j))$, where $n(i)$ denotes the set of sites j that are the nearest neighbors of i-th site except for those along the chain ($j \neq i-1, i+1$). The random variables $\boldsymbol{\sigma}_{\setminus i}$ have fixed realizations in the denominator and the numerator of the right-hand side of Eq. (3.12) at every MCS. Thus, the random variables that interact with the i-th residue σ_i remain only on the right-hand side of Eq. (4.1), because those fixed realizations, except for the residues that interact with σ_i, are canceled out in Eq. (3.12). We decide whether each residue is H or P using the expectation $\langle\sigma_i\rangle$; that is, σ_i is H if $\langle \sigma_i \rangle > 0.5$ and P otherwise. We also take the number of MCSs until the estimated value does not change and let the burn-in be the leading 1/5 of all MCSs. As mentioned in the previous chapter, the inverse temperature is set to $\beta = 10$ for all conformations of all lattice models. On the other hand, we heuristically set the chemical potential μ in order to design a unique ground state by repeating the design experiment many times. The necessary and sufficient condition for successful design is that the energy given by Eq. (2.2) of the target conformation and the sequence designed corresponds to a unique ground state of all possible compact conformations.

4.2 Results

4.2.1 Enumerable conformations

First, we tested our design method with comparatively small lattice protein models, for which all compact conformations were enumerable. We designed 2D $N = 3 \times 3$, 3×4, 4×4, 5×5, and 6×6 lattice models, and 3D $N = 2 \times 2 \times 3$ and $3 \times 3 \times 3$ lattice models.

Native conformations are not necessarily maximally compact. This is because proteins can have low energy if the hydrophobic core is compact enough [74, Yue and Dill 1995]. Therefore, we designed non-maximally compact conformations of 2D $N = 9$, 12, and 16 used in the study by Irbäck and Troein [75, Irbäck and Troein 2022] to compare the statistical property between maximally compact and not maximally compact conformations. Following [75, Irbäck and Troein 2022], we do not design approximately unfolded conformations without a core. The examples of both $N = 16$ maximally compact and non-maximally compact conformations are shown in Fig. 4.1.

The numbers of all conformations N_c including those that are not maximally compact conformations and the numbers of all HP sequences N_s of these lattice models are shown in Table 4.1.

The number of maximally compact conformations is the number of all compact self-avoiding walks from which all kind of rotational, reflection, and head-tail symmetrical conformations have been eliminated. The total number of conformations of these lattice models is enumerable; hence, one can confirm whether or not the designed sequence folds into the target conformation as a unique ground state.

The number of sequences that fold into the target conformation as a unique ground state differs among target conformations and is called designability. In general, a conformation with higher designability is easier to design. This is because high designability means a large solution space in sequence space.

Designability is a significant quantity that relates to the thermodynamic stability of proteins; however, we do not address issues of designability in depth here. In order to calculate the exact success rate (SR) of the overall conformation, one needs to

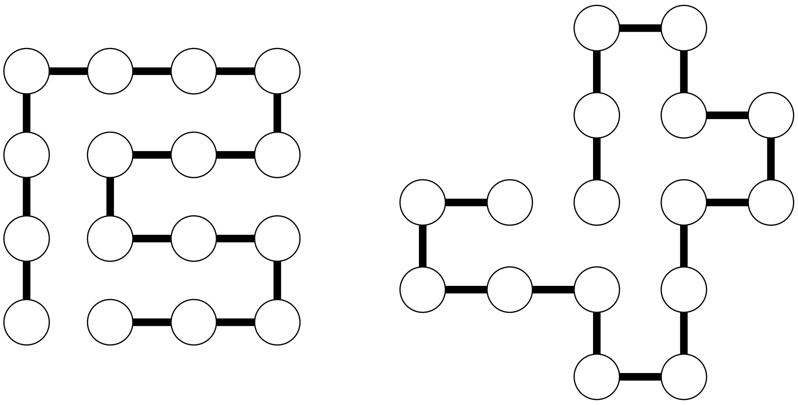

Figure 4.1: Examples of $N = 16$ maximally compact conformation (left) and non-maximally compact conformation (right). The maximally compact conformation has nine contacts, whereas the non-maximally compact conformation has seven contacts.

select designable target conformations with designability greater than zero; however, to enumerate the designabilities of each conformation, one would need to enumerate the energy of every combination of conformations and sequences. This would require vast computation time for models with comparatively large size, such as the 5×5, 6×6, and $3 \times 3 \times 3$ models (Table 4.1), even though they are compact. Therefore, in this study, we carried out the enumeration of designabilities only for the $N = 9$, $N = 12$, $N = 16$, and $2 \times 2 \times 3$ lattice models.

For the models with $N = 5 \times 5$, 6×6, and $3 \times 3 \times 3$, the number of conformations was too large. Thus, we randomly chose 100 target conformations and determined the SR, that is, the number of successfully designed conformations (Table 5.1). For the models with $N = 6 \times 6$ and $3 \times 3 \times 3$, we moreover identified the most highly designable conformation (MHDC) (Figs. 4.2, 4.3, and 4.4), in which designabilities were exactly enumerated [76, Li *et al.* 1996], to test whether our method could be used to design the easiest instance.

The results of the application of our method are summarized in Table 5.1. All designed sequences were classified into three types: good, medium, and bad sequences. The good sequences had the target conformation as a unique ground state, medium sequences had the target conformation as one of the degenerated ground states, and bad sequences had ground state conformation(s) that did not include the target conformation. In the table, SR, $N_{\mathrm{c}}^{(\mathrm{g})}$, $N_{\mathrm{c}}^{(\mathrm{m})}$, and $N_{\mathrm{c}}^{(\mathrm{b})}$, denote the percentage of good sequences and the number of conformations that were designed with good, medium,

Table 4.1: Number of conformations and HP sequences. $N = 9$, $N = 12$, and $N = 16$ involves compact 3×3, 3×4, and 4×4 as their maximally compact conformations, respectively. The numbers in brackets represent the numbes of maximally compact conformations, respectively.

Size	N_{c}	N_{s}	Conformations designed
$N = 9$	12 (8)	512	All
$N = 12$	52 (27)	4096	All
$N = 16$	518 (62)	65536	All
5×5	1075	33554432	Random 100
6×6	52667	68719476736	Random 100 and MHDC
$2 \times 2 \times 3$	69	4096	All
$3 \times 3 \times 3$	103346	134217728	Random 100 and MHDC

and bad sequences, respectively. We also calculated the average degeneracy, d_{av}, for all $N_{\mathrm{c}}^{(\mathrm{g})} + N_{\mathrm{c}}^{(\mathrm{m})}$ ground states. We repeated the calculations with various values of μ and obtained the optimal value μ^* that gave the maximum success rate. The values of μ^* for each lattice size are listed in Table 5.1. The values of the energy parameters are also listed in Table 5.1. The energy parameters $(\epsilon_1, \epsilon_2, \epsilon_3) = (-2.3, -1, 0)$ were also used for a $3 \times 3 \times 3$ lattice model in previous work [76, Li *et al.* 1996] in order to avoid the degeneracy of ground states. We used the same energy parameters for 3D lattice models for the same reason. The total MCSs were set to 10^5 for all target conformations listed in Table 5.1. The $N = 9$ and $2 \times 2 \times 3$ lattices included several non-designable conformations; we excluded such conformations when calculating SR.

According to the results shown in Table 5.1, the SRs were relatively high for small 2D HP models, but they decreased as N increased. Nevertheless, in the case of $N = 16$, the SR is higher than that for the smaller case, $N = 12$. The $N = 16$ case has an extremely high percentage of non-maximally compact conformations than the $N = 12$ one. Hence, this result shows that the proposed design method is more efficient for non-maximally compact conformations.

The average degeneracy d_{av} was low for 2D models. By contrast, the success rate for 3D models

was low compared with that of 2D models. For $2 \times 2 \times 3$, d_{av} was low, but for $3 \times 3 \times 3$, it was comparatively high. Thus, designed sequences did not appear to be likely to fold into the target conformations for the $3 \times 3 \times 3$ cubic lattice. In addition, μ^* increased as the number of residues increased for both the 2D and 3D lattices.

Nevertheless, to design the 3D lattice conformations of the HP model efficiently is difficult because the logarithm of the number of types of the amino acid (alphabet size) is smaller than the conformational entropy of a residue [77, Pande *et al.* 2020

Table 4.2: Design results and the optimal chemical potential μ^*

Size	$N_\mathrm{c}^{(\mathrm{g})}$	$N_\mathrm{c}^{(\mathrm{m})}$	$N_\mathrm{c}^{(\mathrm{b})}$	SR	d_av	μ^*	$(\epsilon_1, \epsilon_2, \epsilon_3)$
$N = 9$	7	1	0	87.5	1.25	0.55	$(-1, 0, 0)$
$N = 12$	29	11	0	72.5	1.475	0.6	
$N = 16$	393	89	0	81.5	1.26	0.62	
5×5	68	32	0	68	1.48	0.74	
6×6	63	37	0	63	1.58	0.8	
$2 \times 2 \times 3$	17	30	1	35.4	2.94	1.7	$(-2.3, -1, 0)$
$3 \times 3 \times 3$	8	80	12	8	10.67	2.33	

Sec.IV]. Therefore, for $N = 2 \times 2 \times 3$ and $3 \times 3 \times 3$, design accuracy would be low when using the HP model. We thus show the design results by increasing the alphabet size for the 3D lattice cases in the next subsection.

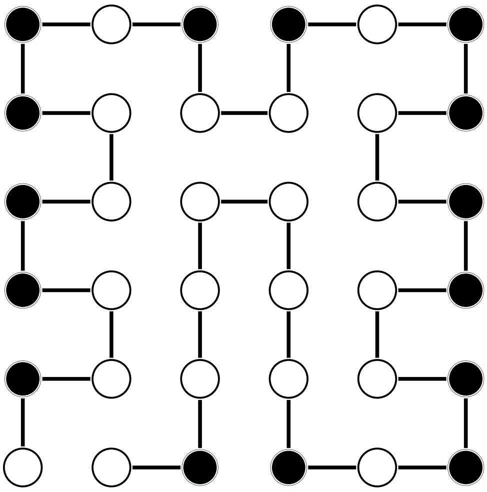

Figure 4.2: Designed sequence of the MHDC of 6×6 HP model with $(\epsilon_1, \epsilon_2, \epsilon_3) = (-1, 0, 0)$, $\beta = 10$, and $\mu^* = 0.8$. The white and black balls denote H and P residues, respectively (the same applies in the following figures.

Note that we did not enumerate designabilities of all conformations for the 5×5, 6×6, and $3 \times 3 \times 3$ models; hence, there may have been non-designable conformations among the 100 randomly chosen conformations. However, it is likely that this was not the case for the 5×5 and 6×6 models, because the smaller $N = 9$ HP model did not lead to any non-designable conformation. On the other hand, the fraction of

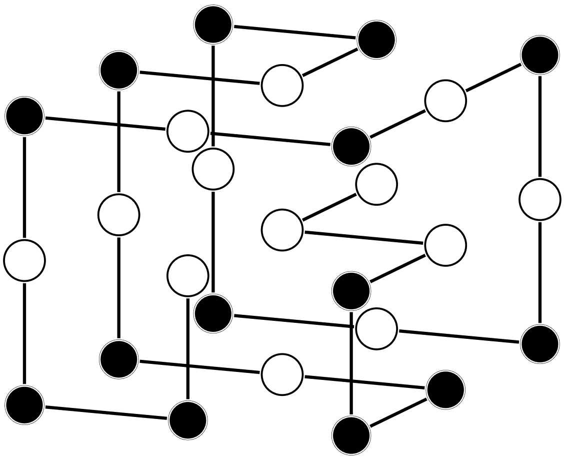

Figure 4.3: Designed sequence of the MHDC $3 \times 3 \times 3$ HP model with $(\epsilon_1, \epsilon_2, \epsilon_3) = (-2.3, -1, 0)$, $\beta = 10$, and $\mu^* = 2.33$.

non-designable conformations out of all conformations for the $2 \times 2 \times 3$ model was $21/69$; the fraction for the $3 \times 3 \times 3$ model is expected to be less than that because the fraction decreased as the size increased in the 2D cases. Thus, there may have been a considerable number of non-designable conformations among the randomly chosen 100 conformations for the $3 \times 3 \times 3$ model; hence, the real success rate of the $3 \times 3 \times 3$ model increased when non-designable conformations were excluded.

Concerning the MHDC of the 6×6 and $3 \times 3 \times 3$ HP models, we obtained a good sequence (Figs. 4.2, 4.3, and 4.4). This is the first example of design of a 6×6 MHDC without enumerating all HP sequences [76]. For the MHDC of the $3 \times 3 \times 3$ HP model, we successfully designed a good sequence for the energy parameters $(\epsilon_1, \epsilon_2, \epsilon_3) = (-2.3, -1,\ 0)$ (Fig. 4.3) and (-1, 0, 0) (Fig. 4.4). We executed 10^4 MCSs for these three cases. The results obtained here demonstrate the features of general globular proteins, with H residues on the inside of the protein and P residues on the surface exposed to the surrounding water molecules. We observed four residues (surrounded by dotted black circles in Figs. 4.3 and 4.4) that were different from each other, possibly owing to the presence or absence of H-P (P-H) contact energies.

The larger μ^* of the MHDC of $3 \times 3 \times 3$ HP model with $(\epsilon_1, \epsilon_2, \epsilon_3) = (-2.3, -1,\ 0)$ compared with the case of $(\epsilon_1, \epsilon_2, \epsilon_3) = (-1, 0, 0)$ could have been due to the lower H-H interaction $\epsilon_1 = -2.3$, leading to a greater increase in the number of H-residues than in the case of $(\epsilon_1, \epsilon_2, \epsilon_3) = (-1, 0, 0)$. Therefore, one needs to increase μ^* in

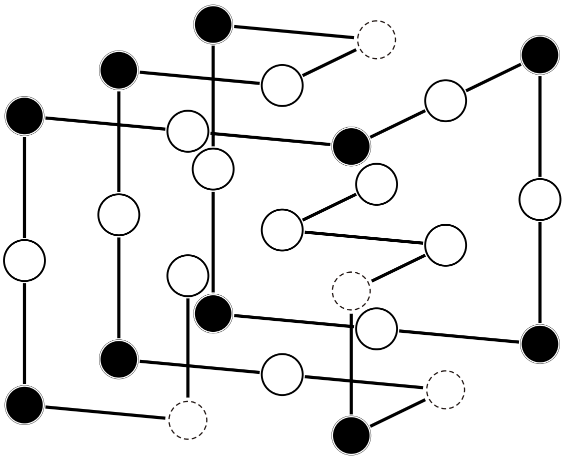

Figure 4.4: Designed sequence of the MHDC of $3 \times 3 \times 3$ HP model with $(\epsilon_1, \epsilon_2, \epsilon_3) = (-1, 0, 0)$, $\beta = 10$, and $\mu^* = 1.0$.

order for the surface residues to be P residues.

4.2.2 Results of 20-letter 3D lattice proteins

In the results shown in Table II, the SRs of 3D cases are quite low. As mentioned above, this is because the logarithm of the alphabet size is smaller than the conformational entropy of a residue in the case of 3D lattice models [77, Pande *et al.* 2020 Sec.IV].

Therefore, we show the design results of the 3D lattice conformations with increasing alphabet size here. The 3D lattice target conformations designed here are the same as those given in the previous subsection. The alphabet size is 20, and we

Table 4.3: Design results of small 3D compact lattice conformations with Miyazawa-Jernigan matrix [78, Miyazawa and Jernigan 1985] and the optimal chemical potential μ^* with the results of HP model in Table. 4.2

Size	Alphabet size	$N_c^{(g)}$	$N_c^{(m)}$	$N_c^{(b)}$	SR	d_{av}	μ^*
$2 \times 2 \times 3$	20 (MJ)	16	18	14	33.3	1.60	1.1
	2 (HP)	17	30	1	35.4	2.97	1.7
$3 \times 3 \times 3$	20 (MJ)	19	63	18	19	5.74	1.55
	2 (HP)	8	80	12	8	10.67	2.33

use the original Miyazawa-Jernigan (MJ) matrix [78, Miyazawa and Jernigan 1985 upper half of Table V] for the contact energy of all the pairs of amino acids. For simplicity, we set all interactions between water molecules and polar amino acids to be equal. The procedure of optimizing μ is identical to the case that of the HP model. We assume that the amino acids Y, F, W, L, V. I, A, P, and M are hydrophobic [79, Monera *et al.* 1995].

It is impossible to calculate the expectation value $\langle \sigma_i \rangle$ in the same way as the HP model because the 20 types of amino acids cannot be represented using the Ising variables. The optimal $\langle \sigma_i \rangle$, therefore, is given by the type sampled the most after the burn-in period.

Table 4.3 depicts the obtained results. To obtain the precise results, one has to calculate the designablities of all conformations using the MJ matrix. This is computationally difficult because calculating the 20^N energy patterns for all conformations is necessary. Nevertheless, the designability of the 20-letter model correlates with the designability of the 2-letter model [80, Li et al. 2002]. Hence, in the case of $N = 2 \times 2 \times 3$, as given in Table II described in previous subsection, we excluded 21 non-designable conformations when we assumed the energy parameter $(\epsilon_1, \epsilon_2, \epsilon_3) = (-2.3, -1.0, 0)$ of the HP model.

According to results summarized in Table 4.3, for $N = 2 \times 2 \times 3$, the value of SR of the 20-letter is a slightly less than the SR of the HP model. By contrast, in the case of $N = 3 \times 3 \times 3$, SR of the 20-letter is more than twice as large as the SR of the 2-letter. The ground state degeneracy typically breaks upon increasing the alphabet size. We believe that an increase in SR for $N = 3 \times 3 \times 3$ is a result of the aforementioned degeneracy breaking. The difference in the changing of SR between the above two cases is, we consider, because of the presence or absence of the core residue.

We discuss how the presence or absence of the core residue affects the above difference. In Fig. 4.5, we represent the change in the number of contacts for the three types, hydrophobic-hydrophobic, hydrophobic-polar (polar-hydrophobic), and polar-polar contacts, when the energy parameter changes from (-2.3, -1, 0) (HP model) to the MJ matrix. However, in the case of $N = 2 \times 2 \times 3$, the number of hydrophobic - polar contacts almost vanishes while changing the energy parameter from (-2.3, -1, 0) (HP model) to the MJ matrix, as shown in Fig. 4.5. In. contrast, in the case of $N = 3 \times 3 \times 3$, even though the number of polar-polar contacts also increases, the balance among the distribution of the three types of contacts does not significantly change. We consider that this difference in the distribution of the three types of contacts shown in Fig. 4.5 is the reason for a difference in the change in SR between the two cases. In the MJ matrix [78, Miyazawa and Jernigan 1985 upper half of Table V], contact energy increases in the following order: of hydrophobic-hydrophobic, hydrophobic-polar (polar-hydrophobic), and polar-polar contacts. Thus, using the

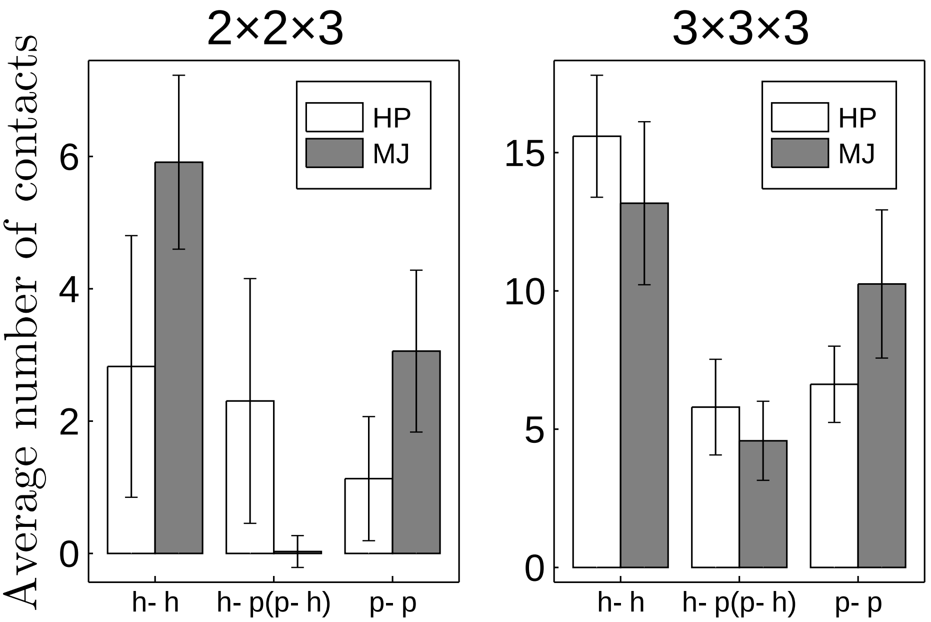

Figure 4.5: The number of contacts for the three types, hydrophobic-hydrophobic (h-h), hydrophobic-polar (polar-hydrophobic) (h-p, p-h), and polar-polar (p-p) contacts, for the two types of energy parameters, the HP model with $(\epsilon_1, \epsilon_2, \epsilon_3) = (-2.3, -1, 0)$ and MJ. The value of each bar is an average of the number of contacts for each type of all the designed conformations. The each error bar represents a standard deviation. The total number of contacts is 9 for $N = 2 \times 2 \times 3$, and 28 for $3 \times 3 \times 3$.

MJ matrix, a conformation with no hydrophobic-polar contacts rarely becomes a ground state. The value of SR for $N = 2 \times 2 \times 3$ is almost constant because this vanishing of the polar-polar contacts and the degeneracy breaking offset each other.

In addition, $N_{\rm c}^{\rm (b)}$ (the number of conformations designed by bad sequences) increase upon increasing the alphabet size in the cases of both $N = 2 \times 2 \times 3$ and $3 \times 3 \times 3$. The degeneracy breaking affects this result as well as increases $N_{\rm c}^{\rm (g)}$. Additionally, one of the reasons for an increase in $N_{\rm c}^{\rm (b)}$ is that the designabilities of the 20-letter model are less than the designabilities of the 2-letter model for many conformations [80, Li *et al.* 2002].

The average degeneracies $d_{\rm av}$ of the two cases decrease with increasing the alphabet size. Thus, this is an evidence of the degeneracy breaking.

4.2.3 Comparison to the dual MC method

In this subsection, using the lattice HP model, we present the results of a comparison between our Bayesian design method and a conventional method involving the exhaustive conformational search. We choose the dual MC method described in Sec.

2.2.3.

The dual MC method maximizes the target probability as given by Eq. (2.4); hence, one can intuitively predict that the SR of the dual MC method is larger. The dual MC method calculates $Z_\beta(\boldsymbol{\sigma})$, which corresponds to the exhaustive conformational search; therefore, the calculation time becomes typically longer.

We compare the SR, calculation time (design time), and the SR per design time (design efficiency). Using design methods based on the MTP criterion, one can solve the design problem if design time is as long as possible. Therefore, comparing only the SR is unreasonable if the design time is not limited. Therefore, in this study, we compare the SR as well as the design efficiency. This viewpoint is especially significant for real applications.

The enumerable conformations shown in Table 4.1 are the ones which were designed for this comparison. Nevertheless, we did not design the 2D $N = 6 \times 6$ and the 3D $N = 3 \times 3 \times 3$ conformations except for the MHDC of $3 \times 3 \times 3$, because their sequence spaces are extremely large to be be designed using the dual MC method. The design time of a conformation of $N = 6 \times 6$ and $3 \times 3 \times 3$ is about $10 - 30$ days using a normal PC (1.2 GHz dual-core Intel Core m3 and 8 GB memory) by our estimation. Note that we carried out the exact calculation of $Z_\beta(\boldsymbol{\sigma})$ for the dual MC method because all the target conformations are enumerable. Thus, we carried out the MC sampling (simulated annealing) of the sequence spaces only (therefore, we name this method the dual MC method for convenience).

The conditions for the design calculations are as follows. For our Bayesian method, for each size, μ^* is identical to the value in Table 5.1, and the number of MCSs is set as low as possible to achieve the same SR as reported in Table 4.2. For the simulated annealing of the dual MC method, the terminal temperature is $T = 0.1$; hence, it equals the terminal temperature of our Bayesian method, $\beta = 10$. The cooling schedule of T is not the linear function used in [56, Seno *et al.* 1996] but is an inverse power function of the MC step because the latter function avoids getting trapped into a metastable state in the amino acid sequence space as compared to the linear function (we show the equation of this cooling schedule in the caption of Table 4.4 below).

As shown in Table 4.4, for all the cases, the SRs of the dual MC method surpass the SRs of the Bayesian method, especially for the 3D $N = 2 \times 2 \times 3$. On the contrary, the DTs of the Bayesian method are all significantly less than the DTs of the dual MC method: the former ones are of the order of 1/100 to 1/1000 of the latter ones. Thus, each DE of the Bayesian method is about 100 or 1000 times each DT in the of the dual MC method. Our Bayesian method is quite efficient compared to the dual MC method. Furthermore, the DT of the MHDC of the 3D $N = 3 \times 3 \times 3$ using the Bayesian method is 0.9244s, but it is about 434600s (about five days) for the dual MC method. It indicates that the DTs' difference between the Bayesian

Table 4.4: We carried out the exact calculations of $Z_\beta(\boldsymbol{\sigma})$ for the dual MC method because the target conformations are all enumerable. We designated the name the dual MC method for the sake of convenience. For the dual MC method, the temperature of k-th MC step is given by $T_k = T_0/(1+\alpha k^2)$ where T_0, k, T_k, and $\alpha > 0$ denotes the initial temperature, MC step, temperature of a MC step k, and the controlling parameter respectively. For all the lattice protein sizes, the initial temperature is represented by $T_0 = 10$, and α is the value of the above equation in which we substitute the terminal temperature $T = 0.1$ for T_k and the MCSs for k. DT and DE denote the design time and the design efficiency, respectively. The MCSs of the Bayesian method are set as low as possible to achieve the same SR value as reported in Table II as described in the main text. The last MCS, which is mentioned in the last column of the dual MC method, denotes the average last MC step for all conformations. All MCSs values for the dual MC method range from 500 to 70000. These MCSs differ depending on the lattice protein size as well as the conformations of the same size.

		Bayesian method		
Size	SR (%)	DT (s)	DE (%/s)	MCSs
$N = 9$	87.50	0.02615	3103	1000
$N = 12$	72.50	0.04356	1670	2000
$N = 16$	81.53	0.1422	5708	5000
5×5	68.00	1.179	577.3	30000
$2 \times 2 \times 3$	35.42	0.2513	136.9	10000

		Dual Monte Carlo method		
Size	SR (%)	DT (s)	DE (%/s)	Last MCS
$N = 9$	100.0	1.043	96.12	492
$N = 12$	90.00	4.300	23.53	2509
$N = 16$	92.74	167.6	0.7182	17924
5×5	92.00	477.5	0.2281	19842
$2 \times 2 \times 3$	85.41	40.77	6.407	10169

method and the design methods that have $Z_\beta(\boldsymbol{\sigma})$ increases as the number of residues increases.

4.2.4 Large 2D conformations

Here, we chose 2D HP models with comparatively large size ($N = 32, 50$) models studied by Irbäck *et al.* [57,58, Irbäck *et al.* 1998; Irbäck *et al.* 1999] described in Sec. 2.2.4., the multi sequence MC method. This confirmed that the designed sequence was likely to fold into the target conformation with simulated tempering. For the model with $N = 32$ (respectively 50), the parameters were set to $\mu^* = 0.7$ (0.85) and the MCSs were 10^4 (10^5). The energy parameters were set to $(\epsilon_1, \epsilon_2, \epsilon_3) = (-1, 0, 0)$ in both cases. The simulation was executed by a normal PC with 1.2 GHz dual-core Intel Core m3 and 8 GB memory, and the calculation time was approximately 0.5–1 s (11–12 s) for $N = 32$ (50). Thus, our method ran faster than those used in previous studies. As a result, we successfully designed the same sequences reported by Irbäck *et al.* [57,58, Irbäck *et al.* 1998; Irbäck *et al.* 1999] (Figs. 4.6 and 4.7). Our method also demonstrates the features of globular proteins.

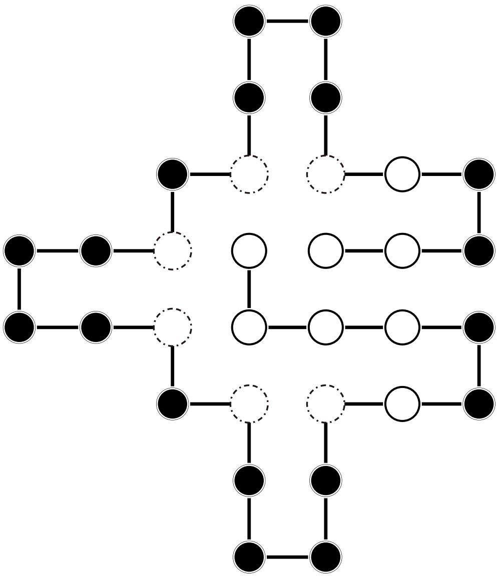

Figure 4.6: Designed sequence of the $N = 32$ 2D HP model with $(\epsilon_1, \epsilon_2, \epsilon_3) = (-1, 0, 0)$, $\beta = 10$, and $\mu^* = 0.7$.

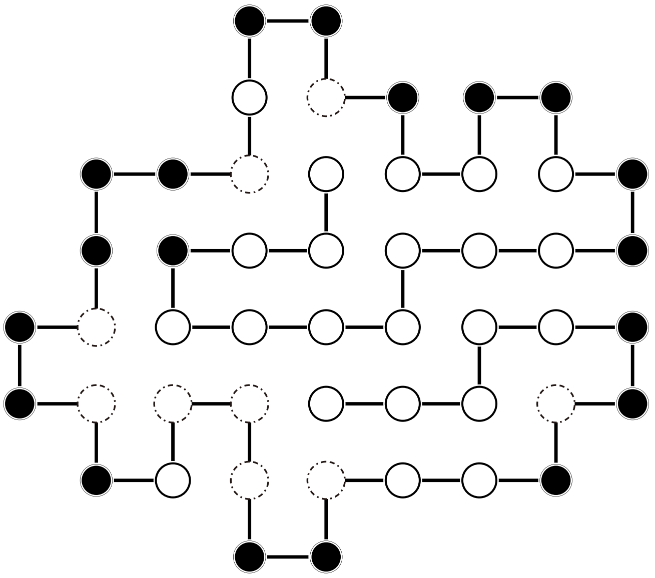

Figure 4.7: Designed sequence of the $N = 50$ 2D HP model with $(\epsilon_1, \epsilon_2, \epsilon_3) = (-1, 0, 0)$, $\beta = 10$, and $\mu^* = 0.85$.

4.2.5 Optimal μ^* and number of surface residues

We represent the relation between the optimal μ^* and the number of surface residues N_{sur} in Fig. 4.8. We show only the results for $(\epsilon_1, \epsilon_2, \epsilon_3) = (-1, 0, 0)$ because the optimal μ^* varies depending on the energy parameters for a given conformation. We therefore plotted the results for all 2D models and the $3 \times 3 \times 3$ MHDC model with $(\epsilon_1, \epsilon_2, \epsilon_3) = (-1, 0, 0)$ The residues that were bent 90 degrees inward (indicated by a dashed black circle in Figs. 4.6 and 4.7) were not counted for N_{sur} because a water molecule is unlikely to combine with such residues (see Fig. 4.8).

We observed noticeable linearity between μ^* and N_{sur}. The outlier $(\mu^*, N_{\mathrm{sur}}) = (0.70, 20)$ was obtained in the 2D $N = 32$ case (Fig. 4.6), in which the target conformation was not fully compact and the number of surface residues was much larger than those of other target conformations tested. According to these results, the optimal μ^* can be estimated by the number of surface residues of a target conformation.

4.2.6 Probability of a P residue

Finally, in order to clarify why 3D models performed less well than 2D models, we consider the probability P_{P} that a residue is P for all residues of the $3 \times 3 \times 3$ and 6×6 MHDC models (Figs. 4.9 and 4.10). We use $p(\sigma_i = -1 | \boldsymbol{R}; \boldsymbol{\sigma}_{\setminus i})$ in Eq. (4.1) as P_{P}.

Each P_{P} in Figs. 4.9 and 4.10 is the average of $1/\left(1 + \exp\left[-\beta\{\Delta E_i(\boldsymbol{R}; \boldsymbol{\sigma}) + \mu\}\right]\right)$

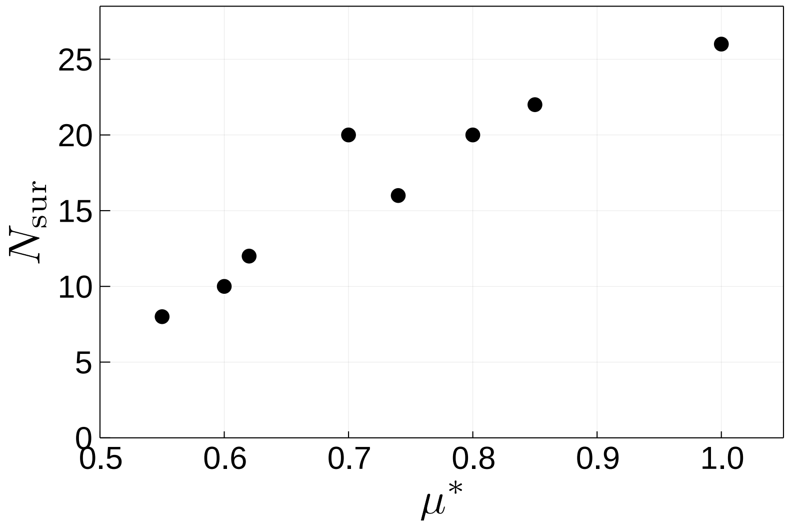

Figure 4.8: Relation between μ^* and number of surface residues, N_{sur}.

over the last 1/5 MCSs in both cases. In the case of the 3D $3 \times 3 \times 3$ lattice, the values of P_{P} for residues 2, 16, 22, and 24 were not very high. These residues were located in the center of each cube side. By contrast, all P_{P} values greater than 0.5 were almost equal to 1 in the case of the 2D 6×6 lattice. Accordingly, it can be seen that the clear division of all P_{P} values into 1 or 0 is an index of successful design. Thus, the 3D models are difficult instances for our design method.

4.3 Conclusion

On applying the 2D lattice HP model, our design method successfully finds an amino acid sequence for which the target conformation has a unique ground state. However, the performance was not as good for the 3D lattice HP models compared to the 2D models. The performance of the 3D model improves on using a 20-letter lattice proteins. Furthermore, we find a strong linearity between the chemical potential of water and the number of surface residues, thereby revealing the relationship between protein structure and the effect of water molecules. The advantage of our method is that it greatly reduces computation time, because it does not require long calculations for the partition function corresponding to an exhaustive conformational search.

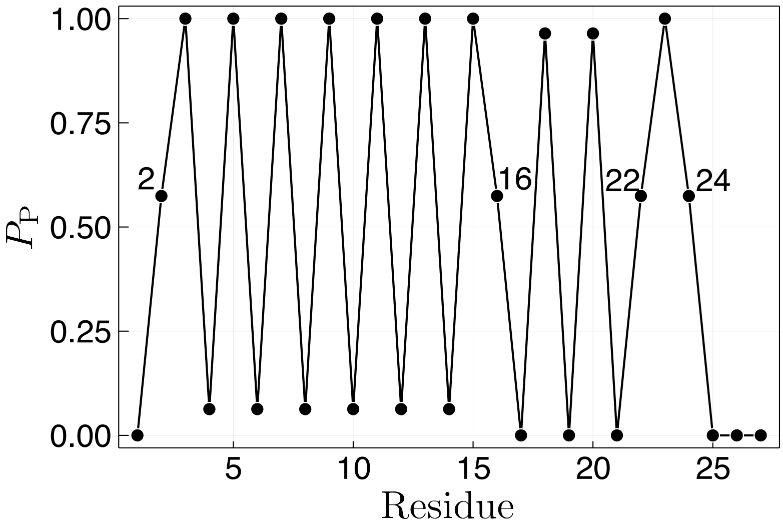

Figure 4.9: P_{P} of each residue of $3 \times 3 \times 3$ MHDC with $(\epsilon_1, \epsilon_2, \epsilon_3) = (-2.3, -1, 0)$ and $\mu = 2.33$ (Fig. 4.3). The residue number starts from the center residue of the front side of Fig. 4.3.

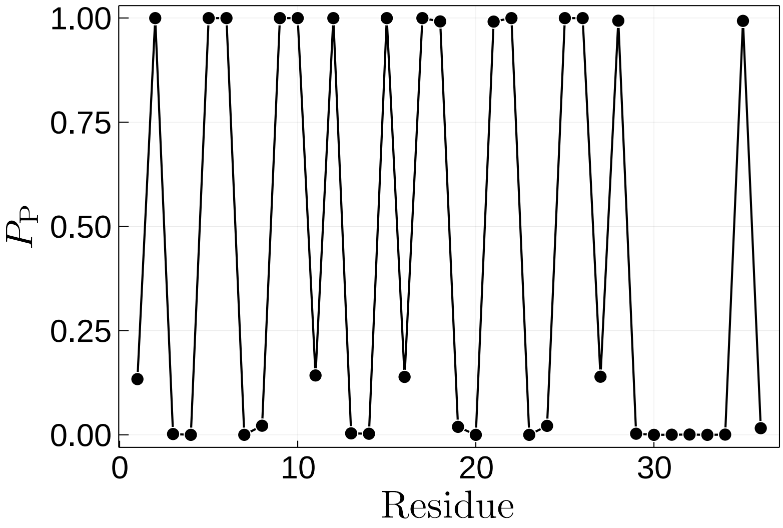

Figure 4.10: P_{P} of each residue of 6×6 MHDC with $(\epsilon_1, \epsilon_2, \epsilon_3) = (-1, 0, 0)$ and $\mu = 0.8$ (Fig. 4.2). The residue number starts from the bottom left residue of Fig. 4.2.

Chapter 5

Sequence estimation by the cavity method

We show the design results obtained by using the cavity method, an analytical method frequently used in the filed of statistical mechanics for disordered systems. The aim of this study is to check whether the cavity method yields results equivalent to those from MCMC. In this chapter, in addition to the HP model, we report the result for a realistic protein, lysozyme. The results shown below except for the result for lysozyme have been published in [81, Takahashi *et al.* 2022].

5.1 The cavity method

The cavity method is a method that extends the Bethe approximation [82, Bethe 1935] to other than two-body interactions. If the interaction network is a tree graph, the cavity method is rigorous. 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 [14, Montanari *et al.* 2004] and the prediction of contact maps [83, Weigt *et al.* 2009]. However, to the best of our knowledge, there are no studies of protein design using the cavity method.

5.2 Belief Propagation

The algorithm for marginalization of posterior probability derived by applying the cavity method is belief propagation (BP). The BP was proposed in 1982 as an algorithm to efficiently calculate the probability of an event occurring in the field of Bayesian networks, a graphical model that describes causal relationships by probabilities [84, Pearl 1982]. Although exact results can be obtained if the graph is a tree structure, it is known to give good approximations even if the graph is not a tree [85, Pearl 1988].

5.3 Applying to the posterior

Then, we show that posterior (3.12) 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 (3.12) can be expressed as follows:

$$p(\boldsymbol{\sigma}|\boldsymbol{R}) = \frac{1}{Z(\boldsymbol{R};\beta,\mu)}\Big(\prod_a \psi_a(\boldsymbol{\sigma}_a)\Big)\prod_{i=1}^{N}\phi_i(\sigma_i). \tag{5.1}$$

In Eq. (5.1), let $\boldsymbol{\sigma}_a$ be the set of residues related to the a-th contact, and $\psi_a(\boldsymbol{\sigma}_a)$ be a function of it. That is, if $\boldsymbol{\sigma}_a = \{\sigma_i, \sigma_j\}$, $\psi_a(\boldsymbol{\sigma}_a) = \exp(\beta\sigma_i\sigma_j)$. The factor $\phi_i(\sigma_i)$ is defined by $\phi_i(\sigma_i) = \exp[\beta\mu(1-\sigma_i)]$. The contact graph of the lattice protein is determined by the conformation $\boldsymbol{R}$ explicitly.

We aim to marginalize the posterior (5.1). Some isolated residues do not interact with any other residues in a lattice protein. We illustrate an example the pair of 2D lattice conformation and its contact graph in Fig. 5.1. For such isolated residues, marginalization is easy. The summation of any other residues (not isolated) cancels in the denominator and numerator of Eq. (5.1); hence the marginal posterior of Eq. (5.1) is obtained as follows:

$$p(\sigma_i|\boldsymbol{R}) = \frac{e^{\beta\mu(1-\sigma_i)}}{\sum_{\sigma_i} e^{\beta\mu(1-\sigma_i)}}. \tag{5.2}$$

For each residue σ_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.

The BP algorithm is derived by the expectation of Eq. (5.1) under the probability distribution of the system excluding the residue σ_i: $p_{\setminus i}(\boldsymbol{\sigma}_{\setminus\sigma_i}|\boldsymbol{R})$ (which is called cavity distribution). We leave the details to the Appendix, only the results are given below:

$$\tilde{\nu}^{(t)}_{a\to i}(\sigma_i) = C_{a\to i}\sum_{\sigma_j} e^{\beta\sigma_i\sigma_j}\nu^{(t)}_{j\to a}(\sigma_j), \tag{5.3}$$

$$\nu^{(t+1)}_{i\to a}(\sigma_i) = C_{i\to a}e^{\beta\mu(1-\sigma_i)}\prod_{b\in\partial_i\setminus a}\tilde{\nu}^{(t)}_{b\to i}(\sigma_i). \tag{5.4}$$

In Eqs. (5.3) and (5.4), let a and b be indices on contacts and i and j are indices on residues. The symbol ∂_i denotes the index set of contacts related to residue σ_i.

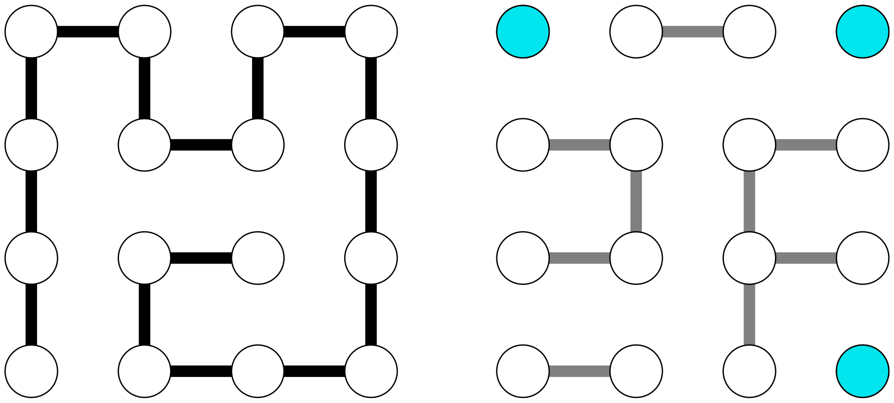

Figure 5.1: (left): A lattice protein conformation of $N = 4 \times 4$. The black line represents the protein backborn structure. (right): The grey line represents edges of the contact graph of its conformation. The colored residues are isolated residues.

$\tilde{\nu}^{(t)}_{a\to i}(\sigma_i)$ is the belief from the a-th contact to the i-th residue, $\nu^{(t+1)}_{i\to a}(\sigma_i)$ 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 $C_{a\to i}$ and $C_{i\to a}$ are the normalizing constants of each distribution function. If one properly defines $\nu^{(t=0)}_{i\to a}(\sigma_i)$ as the initial condition and computes Eqs. (5.3) and (5.4) at each step for all combinations (i, a) excluding the isolated residues, after sufficient iterations $t_{\max}$, the following belief:

$$\nu^{(t)}_i(\sigma_i) = C_i \prod_{a\in\partial_i} \tilde{\nu}^{(t-1)}_{a\to i}(\sigma_i), \tag{5.5}$$

converges to the marginal distribution $p(\sigma_i|\boldsymbol{R}) = \sum_{\boldsymbol{\sigma}\backslash\sigma_i} p(\boldsymbol{\sigma}|\boldsymbol{R})$. In Eq. (5.5) where C_i is the inverse of the normalization constant of $\nu^{(t)}_i(\sigma_i)$, we set the initial condition as a uniform distribution $\nu^{(t=0)}_{i\to a}(\sigma_i = 1) = \nu^{(t=0)}_{i\to a}(\sigma_i = 0) = 1/2$. If $\nu^{(t)}_i(\sigma_i = 1) > 1/2$ residue σ_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 [86, Kabashima and Saad 1998].

5.4 Hyper parameter optmization

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

This procedure does not mean that our design method needs to carry out MCMC for determination of μ^* before using the cavity method. Of course, the proposed design method using the cavity method also can determine μ^* in the same way. In this study, as a primary goal, we investigate whether the cavity method can achieve the same design accuracy as MCMC or not in the same conditions. Thus, we use the same value of μ^* obtained by MCMC.

The criterion of optimization of the hyper parameter μ is nontrivial. For instance, the minimization of the Bethe free energy, corresponding to the maximization of the marginal likelihood in the context of Bayesian learning, is one of promising candidates for the criterion. However, whether or not the maximization of the marginal likelihood maximizes the design accuracy is not clear. In this proposed design method, the design criterion is the maximizer of posterior marginals (MPM) according to the procedure described in the previous section. The hyper parameter μ was determined to be the value with the highest design accuracy. Thus, in this study, the criterion of optimization of the hyper parameter μ is the maximization of the design accuracy under the MPM.

5.5 Results

5.5.1 2D small conformations

Here, we test the cavity method to 2D small lattice proteins which are almost equivalent to Sec. 4.2. The difference is that in this section we do not use 3D HP model: $N = 2 \times 2 \times 3$ and $N = 3 \times 3 \times 3$.

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.

We summarize our design results in Table 5.1. Table 5.1 shows the sequences generated using the cavity method and MCMC (shown in Chapter 4), where $N_{\mathrm{c}}^{(\mathrm{g})}$, $N_{\mathrm{c}}^{(\mathrm{m})}$, and $N_{\mathrm{c}}^{(\mathrm{b})}$ 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 $N_{\mathrm{c}}^{(\mathrm{g})}$ to the total number of structures N_{c}. The optimal value of chemical potential μ^* was determined as explained above, and the same value was used for the cavity method and MCMC. The value μ^* may differ for each conformation even

Table 5.1: Comparison of the cavity method and MCMC design results. The hyperparameter μ^* was calculated many times for each size in MCMC case, and the values achieve the highest success rate, SR. The same values of μ^* were used for the cavity method.

	Cavity method				MCMC				
Size	$N_c^{(g)}$	$N_c^{(m)}$	$N_c^{(b)}$	SR (%)	$N_c^{(g)}$	$N_c^{(m)}$	$N_c^{(b)}$	SR (%)	μ^*
$N = 9$	7	1	0	87.5	7	1	0	87.5	0.55
$N = 12$	29	11	0	72.5	29	11	0	72.5	0.6
$N = 16$	393	89	0	81.5	393	89	0	81.5	0.62
5×5	68	32	0	68	68	32	0	68	0.74
6×6	62	38	0	62	63	37	0	63	0.8

if the size is the same, but we use the same μ^* 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 × 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. The conformations for the $N_c^{(g)}$, $N_c^{(m)}$, and $N_c^{(b)}$ are same in the cavity method and MCMC other than the case of $N = 6 \times 6$. In the case of $N = 6 \times 6$, one conformation is designed successfully by the cavity method but MCMC failed to design it, two conformations are designed successfully by MCMC but the cavity method failed to design those two.

5.5.2 Large 2D conformations

Then, as well as Sec. 4.2.4, we show the results for the 2D $N = 32$ and $N = 50$ lattice conformation studied by [57, 58, Irbäck *et al.* 1998; Irbäck *et al.* 1999] to compare the cavity method and MCMC.

As a result, using the cavity method, we designed those two with the same sequence of [57, 58, Irbäck *et al.* 1998; Irbäck *et al.* 1999]. Fig. 5.2 shows those two designed conformations. The white balls represent H-residues, and the black balls represent P-residues.

5.5.3 Comparison of the calculation time

We also compare the calculation time of the cavity method and MCMC for the $N = 50$ (right-hand side of Fig. 5.2) using a standard PC (Apple M1 MacBook Pro with 8 GB memory). As a result, the calculation time of both cases is 5.282 seconds by MCMC and 1.433 seconds by the cavity method (BP). In the case of BP, one can

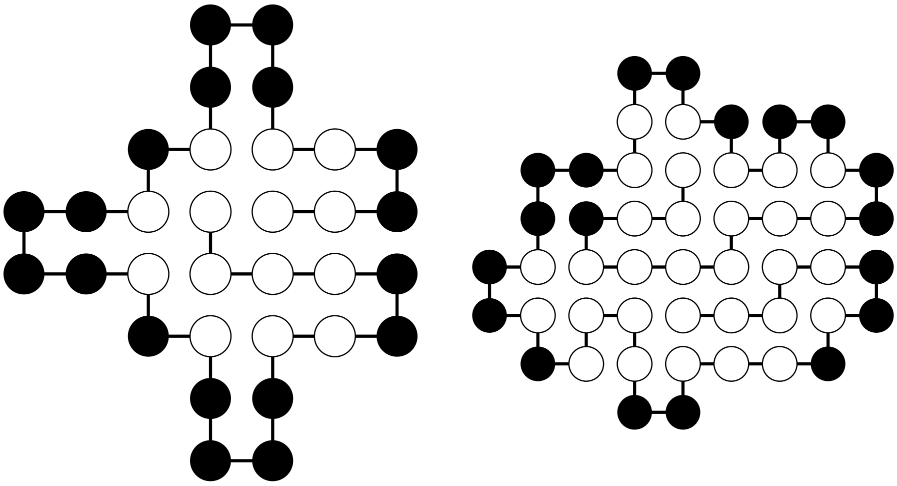

Figure 5.2: (left): Designed conformation of $N = 23$ with $\beta = 10$ and $\mu^* = 0.7$. (right): Designed conformation of $N = 50$ with $\beta = 10$ and $\mu^* = 0.85$. The white balls represent H-residues, and the black balls represent P-residues. Our present design method succeeded to design these two with the identical sequences of conventional studies [57, 58, 72, Irbäck *et al.* 1998; Irbäck *et al.* 1999; Takahashi *et al.* 2021]. Our current approach using the cavity method can be most efficient because it skips internal conformational search due to the partition function $Z_\beta(\boldsymbol{\sigma})(Z_{\beta,\mu}(\boldsymbol{\sigma}))$, and it uses the recursion of BP algorithm instead of MC sampling.

carry out the identical calculation for each residue. We, therefore, carried out the parallel computation by four threads for BP, and the result was 0.506 seconds. The cavity method is about three times faster than MCMC with no parallelization and is about ten times faster with parallelization (four threads). Parallel computation is one of the significant advantages of the cavity method.

As reported in Sec. 2.2.3, the multi sequence MC method can design 2D $N = 18$ HP model in $\mathcal{O}(10)$ CPU sec. by CPU specs in the late 1990s. Therefore, our proposed method is about 100 times faster than the multi sequence MC method, and even taking into account the evolution of CPU specifications over the past 30 years, our method is probably on the order of 10 times faster.

5.6 Redesign a real protein: Lysozyme

We also designed a real protein: lysozyme. The amino acid sequence of lysozyme is of course known and is in the data base, Protein Data Bank (PDB). Thus, we check the overlap between our estimated sequence and PDB sequence. Thus, here we carry out "redesign" of lysozyme. Lysozyme is an enzyme produced by animals. Lysozyme is composed of relatively small number of amino acid residues ($N = 129$). Therefore, lysozyme is the one of the most basic protein.

Figure 5.3: The graphics of the 3D structure of lysozyme from PDB.(PDBid: 7S2V)

We generate the contact network of lysozyme from Protein Data Bank[1] (PDB), a database for the 3D structural data of proteins and nucleic acids, and reflect it to $\Delta(r_i - r_j)$ of the model Hamiltonian. We assume that the lysozyme sequence has only H or P residue. Thus, the hydrophobic residues: A, I, L, M, F, P, Q, Y, and V (using 1 character representation) are all H residues. The hydrophilic residues: R, N, D, C, Q, E, G, H, K, S, and T are all P residues [79, Monera *et al.* 1995] as well as Sec 4.2.2.

The definition of the overlap between our estimated sequence and PDB sequence is percentage of H/P correct residues of 129 residues. The optimal chemical potential μ^* is determined by design calculations until the highest overlap.

We use C_α - C_α distance for the distance between any pair of amino acid residues. If a C_α - C_α distance is shorter than the threshold of contact, the two residues contact to each other. We use four values of threshold in order to assess the difference of effect of loops of contact network. We use 8, 7, 6, and 5Å for the contact threshold. 8Å is the most commonly used contact threshold.

Table. 5.2 shows the design result of lysozyme. The highest overlap is 0.651 (84/129) when the contact threshold is 8Å. This situation setting has not been benchmarked with other methods, but this result of H/P overlap value 0.651 is not very

[1] https://www.rcsb.org/

Table 5.2: The design result of lysozyme. The CT means the contact threshold.

CT (Å)	H/P overlap	μ^*
8	0.651	7.0
7	0.612	4.3
6	0.620	2.4
5	0.6047	0.55

high. This result means that the effect of the sparsity of the contact network is more damaging for our design method than the loop effect. μ^*s increases as the contact threshold increases.

5.7 The contact network of lysozyme

To see how loopy Lysozyme's contact network can be, here is a graphic discussion of how it looks. Fig. 5.4 through Fig. 5.7 illustrate how many loops the graph formed by the contact network of lysozyme has. Each maps an entire 3D network of contacts onto a 2D graph. They are, respectively, Fig. 5.4 when the CT is 8 Å, Fig. 5.5 when the CT is 7 Å, Fig. 5.6 when the CT is 6 Å, and Fig. 5.7 when the CT is 5 Å. As can be seen from these figures, there are many loops in the contact network when the CT is 8 Å or 7 Å, which are commonly used. However, as the CT becomes smaller, the number of loops decreases, and in the case of 5Å, the number of loops is extremely small (only one loop). However, Table 5.2 shows that the redesign accuracy increases as CT increases, i.e., as the number of loops increases. Thus, it can be seen that the smaller CT has a stronger negative effect on design accuracy.

5.8 Conclusion

In this chapter, contrary to previous computational physics studies for protein design, 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 2D lattice HP protein models, the design by the cavity method yields results almost equivalent to those from the Markov chain Monte Carlo method with lower computational cost.

We also tested our design method by the cavity method against the redesign of lysozyme, a real basic protein. The results showed a value of about 65% in the classification accuracy of H or P instead of 20 different types. We also found that the contact network of a natural protein contains many loops but that the small number of contacts has a worse effect on design accuracy than loops.

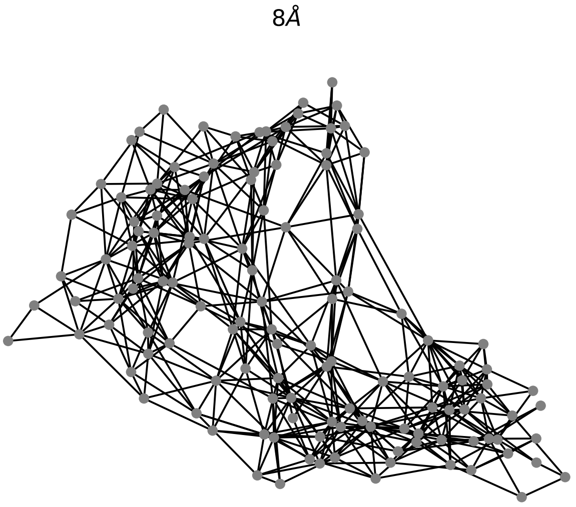

Figure 5.4: The 2D map of contact network of lysozyme when the CT is 8Å.

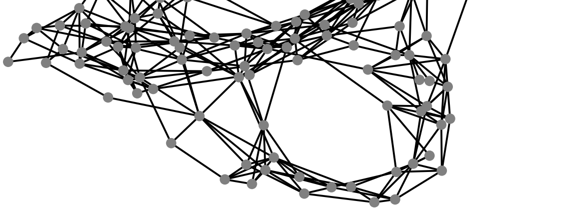

Figure 5.5: The 2D map of contact network of lysozyme when the CT is 7Å.

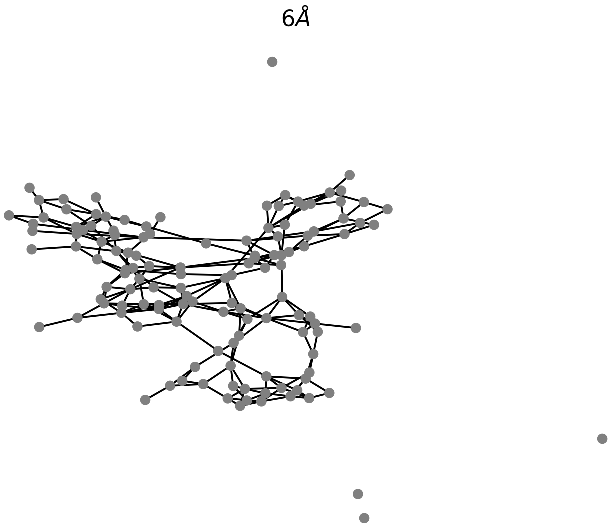

Figure 5.6: The 2D map of contact network of lysozyme when the CT is 6Å.

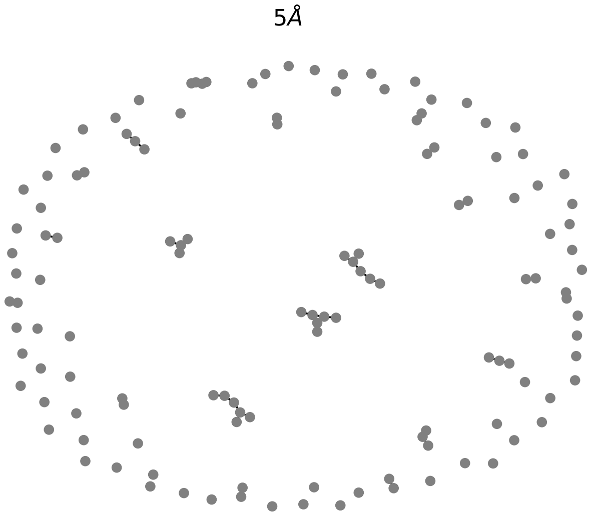

Figure 5.7: The 2D map of contact network of lysozyme when the CT is 5Å.

Chapter 6

Discussion

This chapter provides a more detailed discussion of the results from the previous chapters. The following topics are some of the most important results of this chapter: evaluation of the design methodology, validation of the prior (3.8), and so on. In the verification of the prior (3.8), we report some analyses of the necessary conditions for the prior (3.8) to hold. We also point out the possibility that our method is a protein design method on the Nishimori-line, which is the parameter region where the analytical expression of the internal energy in the analytical theory of spin glass is obtained, and that our method is an example of a Bayes optimal protein design method.

6.1 The weaknesses of our design method

First, we briefly discuss the performance evaluation of our proposed protein design method. The three main weak points of our design method are as follows:

1. Only partial structures can be designed.
2. Only one solution.
3. Cannot systematically determine μ only from information on the target structure.

For the first one, we know that not all protein structures satisfy the condition that the prior (3.8) holds, i.e., they can be designed by energy minimization. So the next task is to investigate which structures satisfy this condition. Then, to give a belief discussion, we compared the distributions of three types of interactions: H-H, H-P(P-H), and P-P between structures that our method successfully designed and structures that our method failed to design to clarify what the differences are between structures that succeed and structures that fail in our design method. Fig. 6.1 shows the average number of the three types of interactions in successful and failed structures for the design results with the 2D $N = 6 \times 6 \times 6$ cavity method.

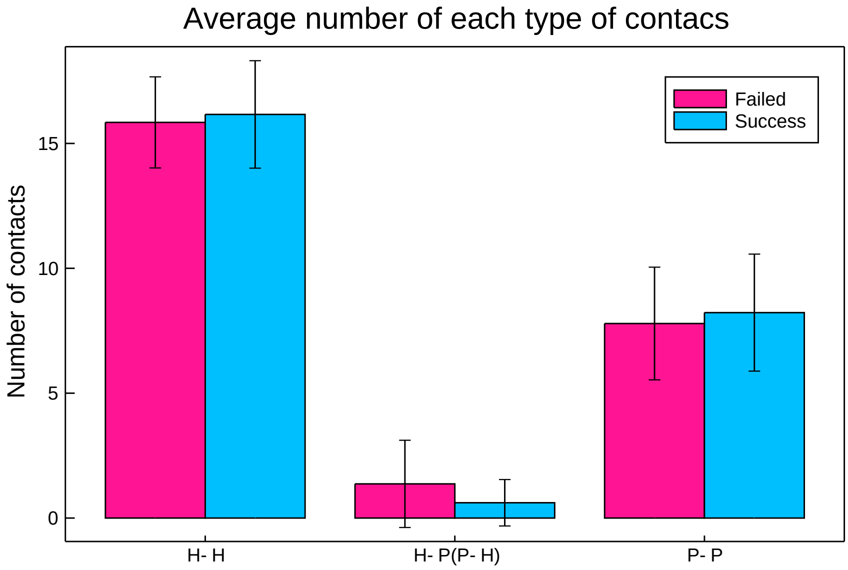

Figure 6.1: The average number of the three types of interactions in successful and failed structures for the design results with the 2D $N = 6 \times 6 \times 6$ cavity method. The each error bar is the standard deviation.

Fig. 6.1 shows that the general trend of the distribution is the same for both structures. However, the failed structures have slightly more H-P contacts, and the number of the other two interaction types is lower. Our design approach is very close to the energy minimization proposed by Shakhnovich and Gutin, as discussed in Sec. 6.2 below. Thus, under a given μ^*, we should obtain the minimum of the Hamiltonian (2.2) or (3.6) (the second term in Eq. (3.6) is structure-independent, so if the sequence $\boldsymbol{\sigma}$ is the same, the order of energies of all structure is the same whether we use Eqs. (2.2) or (3.6)). Thus, the results in Fig. 6.1 indicate that the failed structures have larger minimal frustration than those that succeed. According to the above discussion, the condition that our design method succeeds is lower minimal frustration.

For the second one we can solve the second problem by changing to simulated annealing for MCMC rather than fixing the temperature to a low level from the beginning. However, this is impossible because the cavity method only provides a unique solution. Furthermore, since there are methods to break replica symmetry in the cavity method [13, Mézard and Montanari 2009], it may be possible to output a different sequence in such a case than the one by the replica symmetric cavity method.

For the third weakness, as mentioned in Chapter 3, a solution is to find μ that satisfies the extreme value condition of the free energy $F_{\beta,\mu}(\boldsymbol{R})$. Of course, such μ may not necessarily maximize the design accuracy as it is, but satisfying the extreme value condition of the free energy $F_{\beta,\mu}(\boldsymbol{R})$ is statistically plausible and an interesting theoretical direction.

The more fundamental problem is that the design criteria for protein design methods that do not include the calculation of $Z_{\beta,\mu}(\boldsymbol{\sigma})$ are unclear. The design criterion in our method is almost maximization over the posterior (3.12), maximum a posteriori criterion (MAP criterion). The posterior (3.12) is valid for some proteins but does not seem to be a perfectly universal design criterion for all proteins. Perhaps there is a more general design criterion or minimization function that encompasses the MAP criterion for the posterior (3.12).

At the final of this section, we give a belief discussion for why Rosetta works so well. The design protocol of Rosetta is similar to our Bayesian method in the viewpoint of energy minimization without the exhaustive conformational search. Therefore, we consider Rosetta and our method identical, and the reason why Rosetta works is the same as our method. Then, we can conclude that the reason why Rosetta works is that many of the artificial protein structures designed by Rosetta are less frustrating.

6.2 The relation between our Bayesian method and the energy minimization method

Our proposed Bayesian design method omits the partition function$Z_{\beta,\mu}(\boldsymbol{\sigma})$ and minimizes the Hamiltonian (3.6). Therefore, this method is very similar to the Energy minimization (SG method) by Shakhnovich and Gutin introduced in Sec. 2.2.2. It is also similar to the method of Coluzza et al. introduced in Sec. 2.2.4.

The difference is the way of ensuring the diversity of amino acid types: SG determines the ratio of H to P (which corresponds to the determination of the magnetization constraint in a spin system)as the hyperparameter. Coluzza et al. carry out MCMC such that Eq. (2.19) is large. In contrast, our method introduces the chemical potential μ of water, like an external magnetic field in a spin system, and controls μ to ensure the number of P-residues from outside the protein.

Our method is expected to be more accurate than the SG method because it captures the property of globular proteins: hydrophilic residues are distributed on the surface and hydrophobic residues are distributed on the inside. In the method proposed by Coluzza *et al.*, the design criterion is to balance the tendency for the interaction energy between amino acid residues to be low and the tendency for each $n_a (a = 1, 2, \cdots, 20)$ to be large. This method systematically ensures the diversity of amino acid types, but we do not know if the score function of real proteins follow such a criterion. However, our method of controlling μ reflects that proteins actually express their function by interacting with water molecules. In addition, the Hamiltonian (3.6) is a more realistic Hamiltonian [65–67, Bianco *et al.* 2012; Bianco and

Franzese 2015; Bianco *et al.* 2017], and is a baseline Hamiltonian to these. Thus, the Hamiltonian (3.6) may be a more realistic score function than Coluzza *et al.*. However, as shown in Sec. 4.2.2, our method results in a two-letter array when μ is held constant. This suggests that we must find the optimal μ for each residue type. Systematically finding the optimal μ for each residue type will probably be mathematically complex.

6.3 Why does the cavity method work well?

The cavity method works well because the lattice protein contact graph has a graph structure suited to the Bethe approximation. In the calculation of the Bethe approximation, one assumes that the fluctuations in the effect from the next-nearest neighbor spin are negligible and are replaced by the average value. Thus, the fact that the cavity method works well means that the assumption is likely to hold. In other words, a possible reason why the cavity method works well is that the evolution of the amino acid sequence is coevolved between residues in direct contact with each other and that the effect on evolution from other residues makes little difference from residue to residue.

6.4 Biological evidences of the prior

The prior (3.8) implies that sequences enriched in polar residues are likely to be evolutionarily selected for relatively large β and μ values, owing to the effect of the second term in the Hamiltonian (3.6). This implication is consistent with the fact that organisms have many intrinsically disordered proteins (IDPs), which are characterized by a high proportion of polar residues and lack of an ordered three-dimensional structure; for example, Oates et al. estimated that the percentage of disordered residues in the human proteome is between 37% and 50% [87, Oates *et al.* 2012]. IDPs are important components of the cellular signaling machinery, allowing the same polypeptide to undertake different interactions with different partners [88, Wright and Dyson 2015]. In addition, recent studies have shown that IDPs play an important role in formation of membraneless organelles, enabling internal spatiotemporal control of complex biochemical reactions in a cell [89, Boeynaems *et al.* 2018]. These observations suggest that the physical property of the prior (3.8) is advantageous for cells to efficiently perform complex chemical reactions.

6.5 Verification of a necessary conditions for the prior

To verify totally the prior (3.8), the statistical mechanical hypothesis for protein sequence evolution, one should survey protein sequence data base to generate the probability distribution $p(\boldsymbol{\sigma})$, and proof $p(\boldsymbol{\sigma}) = Z_{\beta_p,\mu_p}(\boldsymbol{\sigma})/\Xi_{\beta_p,\mu_p}$. It is difficult to carry out, but to get collateral evidence of the hypothesis is not very difficult.

Then, we check an important necessary condition of the prior (3.8). The necessary condition of the hypothesis is that the free energy (3.6) of native amino acid sequences $\boldsymbol{\sigma}^N$ is lower than the free energy of random sequences $\boldsymbol{\sigma}^R$. To do this, using all conformational patterns of 5×5 2D lattice proteins, we calculate $F_{\beta_p,\mu_p}(\boldsymbol{\sigma}^N)$, and change in it when mutations were added to the sequence $\boldsymbol{\sigma}^N$. We use the good sequences obtained in Sec. 5.6.1 of 5×5 lattice proteins for the native sequence $\boldsymbol{\sigma}^N$.

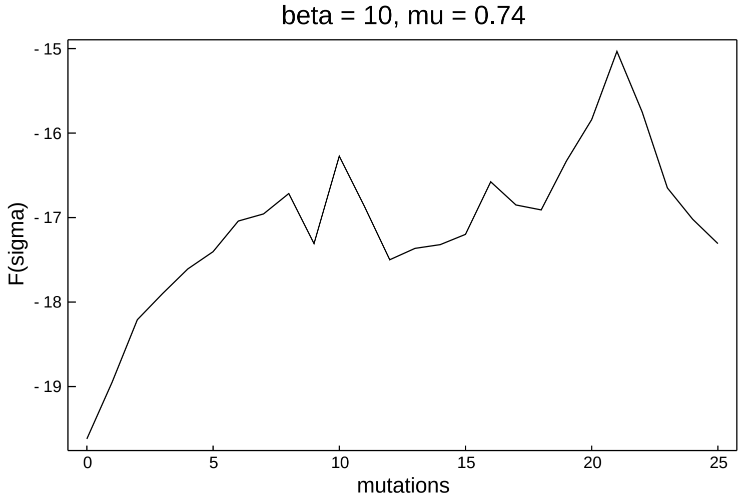

Figure 6.2: Change of $F_{\beta_p,\mu_p}(\boldsymbol{\sigma})$ with mutation in the case of $\beta_p = 10$ and $\mu_p = \mu^* = 0.74$

First, in the case of $\beta_p = 10$ and $\mu_p = \mu^* = 0.74$, the case of design calculation of this study, we calculate free energy change with mutation. In the following results, we use one 5×5 square lattice conformation successfully designed (#812) and its good sequences estimated by the cavity method. According to Fig.6.1, $F_{\beta_p,\mu_p}(\boldsymbol{\sigma})$ increases as the number of mutations of residues increases. The location of the mutation is chosen each time randomly, but once a mutation, i.e., a spin-flip, has occurred, it is not chosen. Therefore, the horizontal axis is the *Hamming distance* from the correct sequence. The result of Fig.6.1 satisfies the necessary condition mentioned earlier.

Then, we investigate above behavior of $F_{\beta_p,\mu_p}(\boldsymbol{\sigma})$ when β_p and μ_p change. The reason of this analysis is to clarify whether the necessary condition of the prior (3.8) is valid only for some special parameter values or for arbitrary parameter values.

The results is shown in Fig.6.2. We see nine parameter patterns. The rightward is lower temperature, and the upward μ_p is larger, i.e., the more hydrophilic the arrangement. The middle value is the same as Fig. 6.1. Hence, the figure is the same as Fig. 6.1. In the upper right parameter pattern, the value of $F_{\beta_p,\mu_p}(\boldsymbol{\sigma})$ is so small that the line almost disappears; Fig. 6.2 shows that, although subtle,

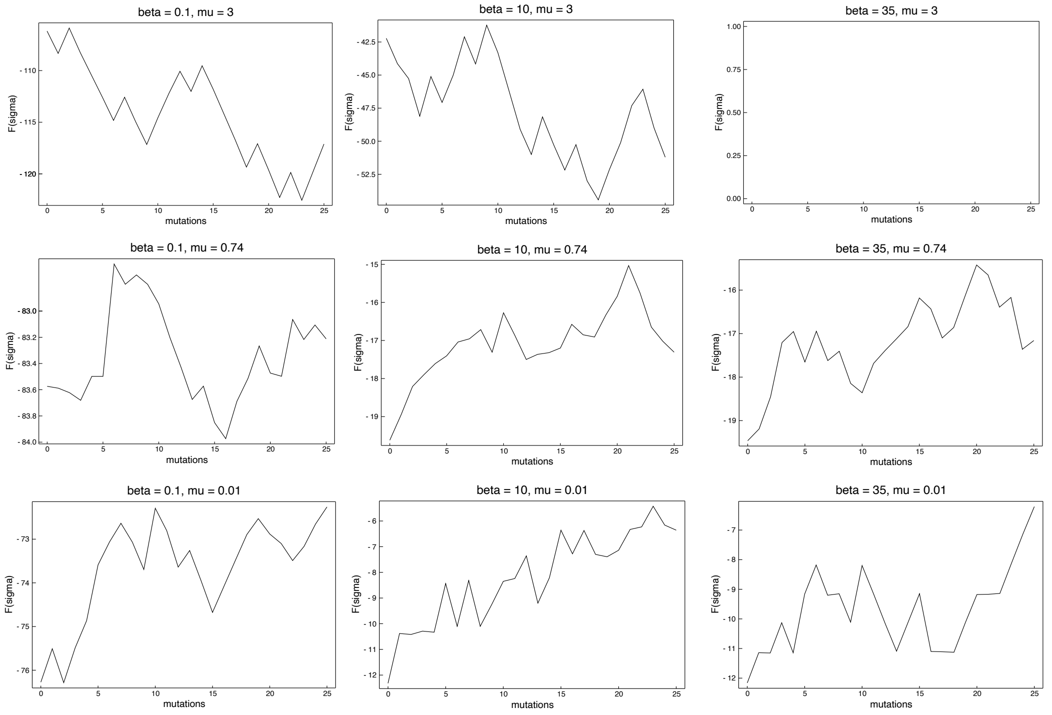

Figure 6.3: Change of $F(\boldsymbol{\sigma})$ with mutation in the case of $\beta_p = 10$ and $\mu_p = \mu^* = 0.74$

temperature appears to satisfy the necessary condition for a wider range of prior than chemical potentials, but the low-temperature case satisfies it more strongly. For chemical potentialsμ_p, if they are too large, i.e., too hydrophilic, they do not satisfy the necessary condition of the prior (3.8) at all. According to above observation, it is suggested that the necessary condition for prior is not satisfied for all parameter regions, but only for certain parameter value combinations (β_p, μ_p). We can conclude from this that if prior (3.8) holds, it is likely to hold only for some special parameter value (β_p, μ_p).

6.6 The relation between our protein design method and the Nishimori-line: Protein design as a Bayes optimal error-correcting code

Here, we point out that the derivation of the posterior (3.12) has interesting theoretical implications. This derivation of the above is consistent with the fact that in calculations for physical quantities of the spin glass $\pm J$ model, the partition function for the spin configuration cancels out on the Nishimori line. The Nishimori line is the hypersurface in the parameter space [90, Nishimori 1980] and achieves the Bayes optimality, an upper bound on accuracy in error-correcting codes [91–94, Ruján 1993; Nishimori 1993; Sourlas 1994; Iba 1999]. 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 optimal[1]. 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.

Error-correcting codes are a problem in which the receiver receives data containing noise in the sender's message and estimates (decodes) the original message. This problem was an essential theme in the early days of statistical mechanical informatics, and spin glass theory has been applied [86, 95–98, Sourlas 1989; Kabashima and Saad 1998; Kabashima and Saad 1999; Kanter and Saad 1999, Kabashima *et al.* 2000]. The message from the sender, i.e., the true parameters, is regarded as the correct amino acid sequence. Noise is added to the correct sequence to form the 3D structure. If the receiver's estimation of the sequence from the 3D structure data is regarded as decoding, the protein design problem can be regarded as the same as the error-correcting code problem (Fig. 6.4).

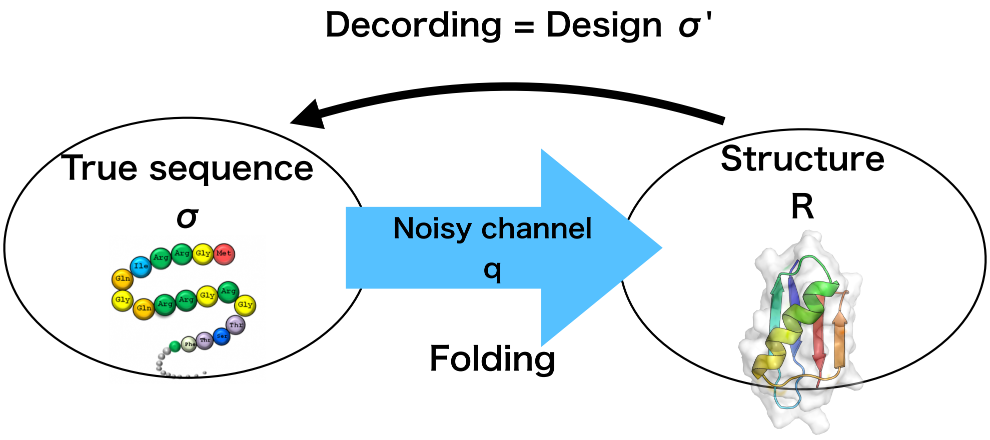

Figure 6.4: The representation of protein design as error-correcting codes.

[1] An estimation method that minimizes the expected value by the likelihood function and prior distribution of some loss function determined by the true and estimated parameters is called a Bayes optimal method. For instance, the loss function is typically the square error.

Chapter 7

Summary and future works

This last chapter give the summery and the future works of our study. The future works given below are wide-ranging. It includes both theoretical and methodological studies.

7.1 Summary

The study of this doctoral thesis addressed the inverse protein folding problem called protein design. We have emphasized the fundamental scientific importance of protein design. Statistical mechanical informatics is, in the author's opinion, one of the most appropriate theories for the fundamental scientific purpose: understanding the "design principle" of proteins.

We have applied Bayesian learning to protein design. The typical approach in statistical mechanical informatics is the Bayesian learning formulation and statistical mechanical analysis. We obtained the posterior (3.12) based on Anfinsen's dogma and the statistical mechanical hypothesis for evolution: "the amino acid sequences which have lower free energy (3.7) have remained through Darwinian evolution", originally proposed by us.

We carried out two estimation method for the posterior: MCMC and the cavity method. The latter is a kind of mean-field approximation frequently used in statistical mechanical informatics. As a result, on applying the 2D HP model, MCMC successfully finds an amino acid sequence for which the target conformation has a unique ground state. However, the performance was not as good for the 3D lattice HP models compared to the 2D models. The performance of the 3D model improves on using a 20-letter lattice proteins. The cavity method yields results almost equivalent to those, 2D small HP model from MCMC with lower computational cost. We carried out "redesign" of lysozyme using the cavity method, and achieved 65% of all estimated residues that match the true residues with 2-letter level. However, this result may not be very high.

We added some brief analysis about the prior. We find that at certain particular temperatures and chemical potentials, the positive sequence's free energy (3.7) is lower than that of the random sequences. It means that one of the necessary conditions for the prior (3.8) is proved.

Based on these results, as the tentative solution for "what is the design principle of proteins?", we propose energy minimization of the target structure with the addition of chemical potential as a candidate protein design principle. That is only valid for evolutionarily selected temperatures and chemical potentials. We also discussed that it is more likely valid for less frustrated structures.

7.2 Future works

7.2.1 Building statistical mechanical informatics of protein design

7.2.1.1 Two goals of statistical mechanical informatics

As explained in Chapter 1, the objectives of statistical mechanical informatics are twofold: performance evaluation (theoretical analysis) and method proposal. The study that proposed the statistical mechanics of compressed sensing, [99, Ganguli and Sompolinsky 2010], used a replica method to identify parameter regions for Bayesian optimal estimation, and the subsequent study proposed a specific method for compressed sensing based on such an analysis. The subsequent study proposed a specific compressed sensing method using the cavity method based on such an analysis [100, Krzakala *et al.* 2012].

However, this study has only proposed a method. Therefore, we will now promote the construction of a theory of statistical mechanics of protein design. We will construct a statistical mechanics for random graph (contact network) spin glass, assuming that the prior (3.8) holds. We also use a random chemical potential as an external field. Random field spin glass is a random graph with a fully connected system [101, Soares *et al.* 1994] and a random graph [102, Erichsen *et al.* 2021]. Therefore, based on these studies, we identify "parameter regions where the design will be successful".

Also, the chemical potential μ of water probably corresponds to the pressure P. Therefore, the parameter region in which the design described above will succeed must be consistent with the work of Bianco *et al.*, as presented at the end of Chapter 2. We will therefore discuss the consistency of our statistical mechanical informatics theory with the findings of Bianco *et al.*.

7.2.1.2 The Nishimori-surface: a relation between evolution and Bayes optimal

The Nishimori-line discussed in Sec. 6.6 and its relationship to Bayes optimal will be further elaborated. First, we will confirm whether our design method is indeed one of the Bayes optimal methods. Then, we find an equation satisfying the inverse temperature and chemical potential (β, μ) that achieves the Bayes optimal, and draw the Nishimori-"surface".

7.2.2 Improving the design method

7.2.2.1 *A priori* decision of μ^*

One of the most significant weaknesses of our method is that μ cannot be determined a priori; as discussed in Sec. 6.1, minimizing the Bayesian free energy $F_{\beta,\mu}(\boldsymbol{R})$ is an effective way to address this. Another approach is to incorporate the Hamiltonian elements related to the structure and μ. It can also contribute to the determination of the Nishimori-surface mentioned earlier. In addition, since the cavity method can be calculated independently for each residue, the chemical potential mu_i, which varies with the position of each residue, can in principle, be estimated. This could be used to predict how the surrounding water molecules will be distributed from the natural structure of the protein.

7.2.2.2 Applying the 1-step RSB cavity method

The 1-step RSB (Replica symmetry breaking) cavity method is a one-step Replica symmetry breaking method for the usual cavity method, i.e., the replica symmetric cavity method. Here, the replica is a pseudo system that appears in the replica method, an analytical method that performs double averaging of thermal and random variables. The replica symmetry assumption assumes that the value of the order parameter between any two different replicas is constant. In the cavity method, it is known that breaking the replica symmetry improves the accuracy when there are loops or long-range interactions in the interaction network [13, Mézard and Montanari 2009]. Therefore, we will apply the 1-step RSB cavity method for realistic proteins. At this time, we will also discuss the statistical significance of the results.

7.2.2.3 Graph neural network with Message passing

Belief propagation is an equivalent algorithm to the message passing algorithm. And it has been recently reported that message passing improves the ability to learn feature interactions between nodes in a graph neural network (GNN) [103, Rizvi *t al.*]. Therefore, we will apply our method to the GNN-based protein design method.

Appendix A

Derivation of the update rules of belief propagation for lattice proteins

We first consider the cavity distribution of the posterior (5.1) in the main text. Cavity distribution is the joint probability distribution of the system without residues which do not relate to σ_i. Its formula is given by

$$p_{\setminus i}(\boldsymbol{\sigma}_{\setminus \sigma_i}|\boldsymbol{R}) = \frac{\prod_{b \notin \partial_i} \psi_b(\boldsymbol{\sigma}_b) \prod_{j \neq i} \phi_j(\sigma_j)}{\sum_{\boldsymbol{\sigma}_{\setminus \sigma_i}} \prod_{b \notin \partial_i} \psi_b(\boldsymbol{\sigma}_b) \prod_{j \neq i} \phi_j(\sigma_j)}, \tag{A.1}$$

where $\psi_b(\boldsymbol{\sigma}_b)$ is given by $\psi_b(\boldsymbol{\sigma}_b) = \exp(\beta \sigma_j \sigma_k)$ if $\boldsymbol{\sigma}_b = \{\sigma_j, \sigma_k\}$. The factor $\phi_j(\sigma_j)$ is defined by $\phi_j(\sigma_j) = \exp[\beta\mu(1-\sigma_j)]$. Then, for all systems, the marginal distribution $p(\sigma_i|\boldsymbol{R}) = \sum_{\boldsymbol{\sigma}_{\setminus \sigma_i}} p(\boldsymbol{\sigma}|\boldsymbol{R})$ can be rigorously expressed using the cavity distribution (A.1) as follows:

$$p(\sigma_i|\boldsymbol{R}) = \frac{\langle \prod_{a \in \partial_i} \psi_a(\boldsymbol{\sigma}_a) \phi_i(\sigma_i) \rangle_{\setminus \sigma_i}}{\sum_{\boldsymbol{\sigma}_{\setminus \sigma_i}} \langle \prod_{a \in \partial_i} \psi_a(\boldsymbol{\sigma}_a) \phi_i(\sigma_i) \rangle_{\setminus \sigma_i}}, \tag{A.2}$$

where $\langle \cdot \rangle_{\setminus \sigma_i}$ means the expectation by the cavity distribution (A.1).

Proof. We separate the posterior (5.1) in the main text into the part that includes a and the part that does not include σ_i,

$$p(\boldsymbol{\sigma}|\boldsymbol{R}) = \frac{\left(\prod_{a \in \partial_i} \psi_a(\boldsymbol{\sigma}_a) \prod_{i \neq j} \phi_i(\sigma_i)\right)\left(\prod_{b \notin \partial_i} \psi_b(\boldsymbol{\sigma}_b) \prod_{j \neq i} \phi_j(\sigma_j)\right)}{\sum_{\boldsymbol{\sigma}} \left(\prod_{a \in \partial_i} \psi_a(\boldsymbol{\sigma}_a) \prod_{i \neq j} \phi_i(\sigma_i)\right)\left(\prod_{b \notin \partial_i} \psi_b(\boldsymbol{\sigma}_b) \prod_{j \neq i} \phi_j(\sigma_j)\right)}. \tag{A.3}$$

We divide the denominator and the numerator of right hand side of Eq. (A.3) by the constant $\sum_{\boldsymbol{\sigma}\backslash_{\sigma_i}} \prod_{b\notin\partial_i} \psi_b(\boldsymbol{\sigma}_b) \prod_{j\neq i} \phi_j(\sigma_j)$. The numerator of (A.3) is then

$$\begin{aligned} &\frac{\Big(\prod_{a\in\partial_i} \psi_a(\boldsymbol{\sigma}_a) \prod_{i\neq j} \phi_i(\sigma_i)\Big)\Big(\prod_{b\notin\partial_i} \psi_b(\boldsymbol{\sigma}_b) \prod_{j\neq i} \phi_j(\sigma_j)\Big)}{\sum_{\boldsymbol{\sigma}\backslash_{\sigma_i}} \prod_{b\notin\partial_i} \psi_b(\boldsymbol{\sigma}_b) \prod_{j\neq i} \phi_j(\sigma_j)} \\ &\qquad\qquad = \prod_{a\in\partial_i} \psi_a(\boldsymbol{\sigma}_a) \prod_{i\neq j} \phi_i(\sigma_i) p_{\backslash i}(\boldsymbol{\sigma}_{\backslash\sigma_i}|\boldsymbol{R}), \end{aligned}$$

and the dominator is calculated similarly. Hence,

$$p(\boldsymbol{\sigma}|\boldsymbol{R}) = \frac{\prod_{a\in\partial_i} \psi_a(\boldsymbol{\sigma}_a) \prod_{i\neq j} \phi_i(\sigma_i) p_{\backslash i}(\boldsymbol{\sigma}_{\backslash\sigma_i}|\boldsymbol{R})}{\sum_{\boldsymbol{\sigma}} \prod_{a\in\partial_i} \psi_a(\boldsymbol{\sigma}_a) \prod_{i\neq j} \phi_i(\sigma_i) p_{\backslash i}(\boldsymbol{\sigma}_{\backslash\sigma_i}|\boldsymbol{R})}.$$

Finally, by the marginalization $p(\sigma_i|\boldsymbol{R}) = \sum_{\boldsymbol{\sigma}\backslash_{\sigma_i}} p(\boldsymbol{\sigma}|\boldsymbol{R})$, we obtain

$$\begin{aligned} p(\sigma_i|\boldsymbol{R}) &= \frac{\sum_{\boldsymbol{\sigma}\backslash_{\sigma_i}} \prod_{a\in\partial_i} \psi_a(\boldsymbol{\sigma}_a) \prod_{i\neq j} \phi_i(\sigma_i) p_{\backslash i}(\boldsymbol{\sigma}_{\backslash\sigma_i}|\boldsymbol{R})}{\sum_{\boldsymbol{\sigma}} \prod_{a\in\partial_i} \psi_a(\boldsymbol{\sigma}_a) \prod_{i\neq j} \phi_i(\sigma_i) p_{\backslash i}(\boldsymbol{\sigma}_{\backslash\sigma_i}|\boldsymbol{R})} \\ &= \frac{\sum_{\boldsymbol{\sigma}\backslash_{\sigma_i}} \prod_{a\in\partial_i} \psi_a(\boldsymbol{\sigma}_a) \prod_{i\neq j} \phi_i(\sigma_i) p_{\backslash i}(\boldsymbol{\sigma}_{\backslash\sigma_i}|\boldsymbol{R})}{\sum_{\sigma_i} \sum_{\boldsymbol{\sigma}\backslash_{\sigma_i}} \prod_{a\in\partial_i} \psi_a(\boldsymbol{\sigma}_a) \prod_{i\neq j} \phi_i(\sigma_i) p_{\backslash i}(\boldsymbol{\sigma}_{\backslash\sigma_i}|\boldsymbol{R})} \\ &= \frac{\langle \prod_{a\in\partial_i} \psi_a(\boldsymbol{\sigma}_a) \phi_i(\sigma_i) \rangle_{\backslash\sigma_i}}{\sum_{\boldsymbol{\sigma}\backslash_{\sigma_i}} \langle \prod_{a\in\partial_i} \psi_a(\boldsymbol{\sigma}_a) \phi_i(\sigma_i) \rangle_{\backslash\sigma_i}}. \end{aligned}$$

■

We define following effective potential $\psi_i^{\mathrm{eff}}(\sigma_i)$ by the numerator of right hand side of Eq. (A.2):

$$\psi_i^{\mathrm{eff}}(\sigma_i) = \sum_{\boldsymbol{\sigma}\backslash_{\sigma_i}} \Big(\prod_{a\in\partial_i} \psi_a(\boldsymbol{\sigma}_a)\Big) \phi_i(\sigma_i) p_{\backslash i}(\boldsymbol{\sigma}_{\backslash\sigma_i}|\boldsymbol{R}). \tag{A.4}$$

Additionally, for $a \in \partial_i$, we consider the system excluding $\psi_a(\boldsymbol{\sigma}_a)$, called a-cavity system. We define the marginal distribution of σ_j which is included in $\boldsymbol{\sigma}_a$ under a-cavity system by $\nu_{j\to a}(\sigma_j)$. If the contact graph of given target conformation is a tree, excluding any residue σ_i makes the contact graph divided into independent part per residues related to the contact a.

Therefore, if the contact graph of given target conformation is a tree, we can calculate (A.4) as follows:

$$\begin{aligned} \psi_i^{\mathrm{eff}}(\sigma_i) &= \phi_i(\sigma_i) \prod_{a\in\partial_i} \Big(\sum_{\boldsymbol{\sigma}_a\backslash_{\sigma_i}} \psi_a(\boldsymbol{\sigma}_a) \sum_{\boldsymbol{\sigma}\backslash_{\boldsymbol{\sigma}_a}} p_{\backslash i}(\boldsymbol{\sigma}_{\backslash\sigma_i}|\boldsymbol{R}) \Big) \\ &= \phi_i(\sigma_i) \prod_{a\in\partial_i} \Big(\sum_{\boldsymbol{\sigma}_a\backslash_{\sigma_i}} \psi_a(\boldsymbol{\sigma}_a) \prod_{j\in\partial_a\backslash_i} \nu_{j\to a}(\sigma_j) \Big) \\ &= \phi_i(\sigma_i) \prod_{a\in\partial_i} \Big(\sum_{\sigma_j} \psi_a(\boldsymbol{\sigma}_a) \nu_{j\to a}(\sigma_j) \Big). \qquad (j \in \partial_a\backslash_i) \end{aligned} \tag{A.5}$$

In Eq. (A.5), we used $\sum_{\boldsymbol{\sigma}_a \backslash \sigma_i} = \sum_{\sigma_j}$ and $\prod_{j \in \partial_a \backslash_i} \nu_{j \to a}(\sigma_j) = \nu_{j \to a}(\sigma_j)$, because each index set ∂_a has only two indices in the lattice HP model.

Then, we consider the effective potential of a-cavity system $\psi^{\text{eff}}_{i \to a}(\sigma_i)$. $\psi^{\text{eff}}_{i \to a}(\sigma_i)$ is obtained by excluding $\psi_a(\boldsymbol{\sigma}_a)$ from $\psi^{\text{eff}}_i(\sigma_i)$. Hence we obtain $\psi^{\text{eff}}_{i \to a}(\sigma_i)$ as follows:

$$\psi^{\text{eff}}_{i \to a}(\sigma_i) = \phi_i(\sigma_i) \prod_{b \in \partial_i \backslash a} \Big(\sum_{\sigma_k} \psi_b(\boldsymbol{\sigma}_b) \nu_{k \to b}(\sigma_k) \Big). \qquad (k \in \partial_b \backslash_i)$$

From the definition of $\nu_{j \to a}(\sigma_i)$ explained above, $\nu_{j \to a}(\sigma_i)$ is obtained by normalization $\psi^{\text{eff}}_{i \to a}(\sigma_i)$. Therefore, let $j = k$, we obtain following expression:

$$\nu_{j \to a}(\sigma_i) = \frac{\phi_i(\sigma_i) \prod_{b \in \partial_i \backslash a} \Big(\sum_{\sigma_j} \psi_b(\boldsymbol{\sigma}_b) \nu_{j \to b}(\sigma_j) \Big)}{\sum_{\sigma_i} \phi_i(\sigma_i) \prod_{b \in \partial_i \backslash a} \Big(\sum_{\sigma_j} \psi_b(\boldsymbol{\sigma}_b) \nu_{j \to b}(\sigma_j) \Big)}. \qquad (j \in \partial_b \backslash_i)$$

Let $\tilde{\nu}_{a \to i}(\sigma_i)$ be the distribution function derived by normalization $\sum_{\sigma_j} \psi_a(\boldsymbol{\sigma}_a) \nu_{j \to a}(\sigma_j)$, the “belief” from a to σ_i. Then, we obtain following expression of two distribution functions:

$$\begin{aligned} \tilde{\nu}_{a \to i}(\sigma_i) &= C_{a \to i} \sum_{\sigma_j} \psi_a(\boldsymbol{\sigma}_a) \nu_{j \to a}(\sigma_j), \qquad (j \in \partial_b \backslash_i) \\ \nu_{i \to a}(\sigma_i) &= C_{i \to a} \phi_i(\sigma_i) \prod_{b \in \partial_i \backslash a} \tilde{\nu}_{b \to i}(\sigma_i), \end{aligned}$$

where $C_{a \to i}$ and $C_{i \to a}$ are normalization constant of each distribution functions, respectively. When $\partial_a = \{i, j\}$, one can use $\psi_a(\boldsymbol{\sigma}_a) = e^{\beta \sigma_i \sigma_j}$. In addition, using $\phi_i(\sigma_i) = \exp[\beta \mu (1 - \sigma_i)]$, we obtain the update rules of BP: Eq. (5.3) and Eq. (5.4) in the main text.

Bibliography

[1] 笹井理生. **蛋白質の柔らかなダイナミクス**. 培風館, 2008.

[2] Christian B Anfinsen. Principles that govern the folding of protein chains. *Science*, 181(4096):223–230, 1973.

[3] John Ellis. Proteins as molecular chaperones. *Nature*, 328(6129):378–379, 1987.

[4] Ivan Coluzza. Computational protein design: a review. *Journal of Physics: Condensed Matter*, 29(14):143001, 2017.

[5] Simona Cocco, Christoph Feinauer, Matteo Figliuzzi, Rémi Monasson, and Martin Weigt. Inverse statistical physics of protein sequences: a key issues review. *Reports on Progress in Physics*, 81(3):032601, 2018.

[6] Arthur G Street and Stephen L Mayo. Computational protein design. *Structure*, 7(5):R105–R109, 1999.

[7] Shaun M Lippow and Bruce Tidor. Progress in computational protein design. *Current opinion in biotechnology*, 18(4):305–311, 2007.

[8] Ilan Samish, Christopher M MacDermaid, Jose Manuel Perez-Aguilar, and Jeffery G Saven. Theoretical and computational protein design. *Annual review of physical chemistry*, 62:129–149, 2011.

[9] Ilan Samish. *Computational protein design.* Springer, 2017.

[10] Marc Mézard, Giorgio Parisi, and Miguel Angel Virasoro. *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*, volume 9. World Scientific Publishing Company, 1987.

[11] Hidetoshi Nishimori. *Statistical physics of spin glasses and information processing: an introduction.* Number 111. Clarendon Press, 2001.

[12] 樺島祥介. 情報統計力学とは何かー情報学でも more is different ー. **人工知能学会誌**, 22(3), 2007.

[13] Marc Mezard and Andrea Montanari. *Information, physics, and computation.* Oxford University Press, 2009.

[14] Andrea Montanari, M Müller, and Marc Mézard. Phase diagram of random heteropolymers. *Physical review letters*, 92(18):185509, 2004.

[15] Daniel J Amit, Hanoch Gutfreund, and Haim Sompolinsky. Spin-glass models of neural networks. *Physical Review A*, 32(2):1007, 1985.

[16] ACC Coolen. Statistical mechanics of recurrent neural networks i—statics. In *Handbook of biological physics*, volume 4, pages 553–618. Elsevier, 2001.

[17] John Hertz, Anders Krogh, and Richard G Palmer. *Introduction to the theory of neural computation*. CRC Press, 2018.

[18] Joseph D Bryngelson and Peter G Wolynes. Spin glasses and the statistical mechanics of protein folding. *Proceedings of the National Academy of sciences*, 84(21):7524–7528, 1987.

[19] T Garel and H Orland. Mean-field model for protein folding. *EPL (Europhysics Letters)*, 6(4):307, 1988.

[20] Richard A Goldstein, Zaida A Luthey-Schulten, and Peter G Wolynes. Optimal protein-folding codes from spin-glass theory. *Proceedings of the National Academy of Sciences*, 89(11):4918–4922, 1992.

[21] AM Gutin and EI Shakhnovich. Ground state of random copolymers and the discrete random energy model. *The Journal of chemical physics*, 98(10):8174–8177, 1993.

[22] 白木賢太郎. **相分離生物学**. 東京化学同人, 2019.

[23] Keqin Chen and Frances H Arnold. Enzyme engineering for nonaqueous solvents: random mutagenesis to enhance activity of subtilisin e in polar organic media. *Bio/Technology*, 9(11):1073–1077, 1991.

[24] Keqin Chen and Frances H Arnold. Tuning the activity of an enzyme for unusual environments: sequential random mutagenesis of subtilisin e for catalysis in dimethylformamide. *Proceedings of the National Academy of Sciences*, 90(12):5618–5622, 1993.

[25] Michael S Packer and David R Liu. Methods for the directed evolution of proteins. *Nature Reviews Genetics*, 16(7):379–394, 2015.

[26] Stephen F Parmley and George P Smith. Antibody-selectable filamentous fd phage vectors: affinity purification of target genes. *Gene*, 73(2):305–318, 1988.

[27] Julia Koehler Leman, Brian D Weitzner, Steven M Lewis, Jared Adolf-Bryfogle, Nawsad Alam, Rebecca F Alford, Melanie Aprahamian, David Baker, Kyle A

Barlow, Patrick Barth, et al. Macromolecular modeling and design in rosetta: recent methods and frameworks. *Nature methods*, 17(7):665–680, 2020.

[28] Yang Hsia, Jacob B Bale, Shane Gonen, Dan Shi, William Sheffler, Kimberly K Fong, Una Nattermann, Chunfu Xu, Po-Ssu Huang, Rashmi Ravichandran, et al. Design of a hyperstable 60-subunit protein icosahedron. *Nature*, 535(7610):136–139, 2016.

[29] Jacob B Bale, Shane Gonen, Yuxi Liu, William Sheffler, Daniel Ellis, Chantz Thomas, Duilio Cascio, Todd O Yeates, Tamir Gonen, Neil P King, et al. Accurate design of megadalton-scale two-component icosahedral protein complexes. *Science*, 353(6297):389–394, 2016.

[30] Gabriel L Butterfield, Marc J Lajoie, Heather H Gustafson, Drew L Sellers, Una Nattermann, Daniel Ellis, Jacob B Bale, Sharon Ke, Garreck H Lenz, Angelica Yehdego, et al. Evolution of a designed protein assembly encapsulating its own rna genome. *Nature*, 552(7685):415–420, 2017.

[31] Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.

[32] Yufeng Liu and Haipeng Gong. Using the unfolded state as the reference state improves the performance of statistical potentials. *Biophysical journal*, 103(9):1950–1959, 2012.

[33] Wenze Ding, Kenta Nakai, and Haipeng Gong. Protein design via deep learning. *Briefings in bioinformatics*, 23(3):bbac102, 2022.

[34] Zhixiu Li, Yuedong Yang, Eshel Faraggi, Jian Zhan, and Yaoqi Zhou. Direct prediction of profiles of sequences compatible with a protein structure by neural networks with fragment-based local and energy-based nonlocal profiles. *Proteins: Structure, Function, and Bioinformatics*, 82(10):2565–2573, 2014.

[35] James O'Connell, Zhixiu Li, Jack Hanson, Rhys Heffernan, James Lyons, Kuldip Paliwal, Abdollah Dehzangi, Yuedong Yang, and Yaoqi Zhou. Spin2: Predicting sequence profiles from protein structures using deep neural networks. *Proteins: Structure, Function, and Bioinformatics*, 86(6):629–633, 2018.

[36] Jingxue Wang, Huali Cao, John ZH Zhang, and Yifei Qi. Computational protein design with deep learning neural networks. *Scientific reports*, 8(1):1–9, 2018.

[37] Sheng Chen, Zhe Sun, Lihua Lin, Zifeng Liu, Xun Liu, Yutian Chong, Yutong Lu, Huiying Zhao, and Yuedong Yang. To improve protein sequence profile prediction through image captioning on pairwise residue distance map. *Journal of chemical information and modeling*, 60(1):391–399, 2019.

[38] Yuan Zhang, Yang Chen, Chenran Wang, Chun-Chao Lo, Xiuwen Liu, Wei Wu, and Jinfeng Zhang. Prodconn: Protein design using a convolutional neural network. *Proteins: Structure, Function, and Bioinformatics*, 88(7):819–829, 2020.

[39] Yifei Qi and John ZH Zhang. Densecpd: improving the accuracy of neural-network-based computational protein sequence design with densenet. *Journal of chemical information and modeling*, 60(3):1245–1252, 2020.

[40] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[41] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. *Advances in neural information processing systems*, 29, 2016.

[42] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[43] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

[44] John Ingraham, Vikas Garg, Regina Barzilay, and Tommi Jaakkola. Generative models for graph-based protein design. *Advances in neural information processing systems*, 32, 2019.

[45] Bowen Jing, Stephan Eismann, Patricia Suriana, Raphael JL Townshend, and Ron Dror. Learning from protein structure with geometric vector perceptrons. *arXiv preprint arXiv:2009.01411*, 2020.

[46] Alexey Strokach, David Becerra, Carles Corbi-Verge, Albert Perez-Riba, and Philip M Kim. Fast and flexible protein design using deep graph neural networks. *Cell systems*, 11(4):402–411, 2020.

[47] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

[48] Kit Fun Lau and Ken A Dill. A lattice statistical mechanics model of the conformational and sequence spaces of proteins. *Macromolecules*, 22(10):3986–3997, 1989.

[49] Marek Cieplak and Jayanth R Banavar. Energy landscape and dynamics of proteins: an exact analysis of a simplified lattice model. *Physical Review E*, 88(4):040702, 2013.

[50] Erik Van Dijk, Patrick Varilly, Tuomas PJ Knowles, Daan Frenkel, and Sanne Abeln. Consistent treatment of hydrophobicity in protein lattice models accounts for cold denaturation. *Physical review letters*, 116(7):078101, 2016.

[51] Christian Holzgräfe, Anders Irbäck, and Carl Troein. Mutation-induced fold switching among lattice proteins. *The Journal of chemical physics*, 135(19):11B611, 2011.

[52] Guangjie Shi, Thomas Vogel, Thomas Wüst, Ying Wai Li, and David P Landau. Effect of single-site mutations on hydrophobic-polar lattice proteins. *Physical Review E*, 90(3):033307, 2014.

[53] Shi-Jie Chen and Ken A Dill. Rna folding energy landscapes. *Proceedings of the National Academy of Sciences*, 97(2):646–651, 2000.

[54] Eugene I Shakhnovich and AM Gutin. A new approach to the design of stable proteins. *Protein Engineering, Design and Selection*, 6(8):793–800, 1993.

[55] Eugene I Shakhnovich and Alexander M Gutin. Engineering of stable and fast-folding sequences of model proteins. *Proceedings of the National Academy of Sciences*, 90(15):7195–7199, 1993.

[56] Flavio Seno, Michele Vendruscolo, Amos Maritan, and Jayanth R Banavar. Optimal protein design procedure. *Physical review letters*, 77(9):1901, 1996.

[57] Anders Irbäck, Carsten Peterson, Frank Potthast, and Erik Sandelin. Monte carlo procedure for protein design. *Physical Review E*, 58(5):R5249, 1998.

[58] Anders Irbäck, Carsten Peterson, Frank Potthast, and Erik Sandelin. Design of sequences with good folding properties in coarse-grained protein models. *Structure*, 7(3):347–360, 1999.

[59] Yukito Iba, Kei Tokita, and Macoto Kikuchi. Design equation: A novel approach to heteropolymer design. *Journal of the Physical Society of Japan*, 67(11):3985–3990, 1998.

[60] Kei Tokita, Macoto Kikuchi, and Yukito Iba. Dynamical equation approach to protein design. *Progress of Theoretical Physics Supplement*, 138:378–383, 2000.

[61] Yihua Wang, Baohan Wang, Yun Liu, Weizu Chen, and Cunxin Wang. A generalized approach for protein design based on the relative entropy. *Chinese Science Bulletin*, 49(5):426–431, 2004.

[62] Xiong Jiao, Baohan Wang, Jiguo Su, Weizu Chen, and Cunxin Wang. Protein design based on the relative entropy. *Physical Review E*, 73(6):061903, 2006.

[63] I Coluzza, HG Muller, and D Frenkel. Designing refoldable model molecules. *Physical Review E*, 68(4):046703, 2003.

[64] Ivan Coluzza. A coarse-grained approach to protein design: learning from design to understand folding. *PloS one*, 6(7):e20853, 2011.

[65] Valentino Bianco, Svilen Iskrov, and Giancarlo Franzese. Understanding the role of hydrogen bonds in water dynamics and protein stability. *Journal of Biological Physics*, 38(1):27–48, 2012.

[66] Valentino Bianco and Giancarlo Franzese. Contribution of water to pressure and cold denaturation of proteins. *Physical review letters*, 115(10):108101, 2015.

[67] Valentino Bianco, Giancarlo Franzese, Christoph Dellago, and Ivan Coluzza. Role of water in the selection of stable proteins at ambient and extreme thermodynamic conditions. *Physical Review X*, 7(2):021047, 2017.

[68] Karl Friston, James Kilner, and Lee Harrison. A free energy principle for the brain. *Journal of physiology-Paris*, 100(1-3):70–87, 2006.

[69] Tetsuya J Kobayashi. Implementation of dynamic bayesian decision making by intracellular kinetics. *Physical review letters*, 104(22):228104, 2010.

[70] Masahiro Kinoshita. Importance of translational entropy of water in biological self-assembly processes like protein folding. *International journal of molecular sciences*, 10(3):1064–1080, 2009.

[71] Tomonari Sumi and Hiroshi Imamura. Water-mediated interactions destabilize proteins. *Protein Science*, 30(10):2132–2143, 2021.

[72] Tomoei Takahashi, George Chikenji, and Kei Tokita. Lattice protein design using bayesian learning. *Physical Review E*, 104(1):014404, 2021.

[73] Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, (6):721–741, 1984.

[74] Kaizhi Yue and Ken A Dill. Forces of tertiary structural organization in globular proteins. *Proceedings of the National Academy of Sciences*, 92(1):146–150, 1995.

[75] Anders Irbäck and Carl Troein. Enumerating designing sequences in the hp model. *Journal of Biological Physics*, 28(1):1–15, 2002.

[76] Hao Li, Robert Helling, Chao Tang, and Ned Wingreen. Emergence of preferred structures in a simple model of protein folding. *Science*, 273(5275):666–669, 1996.

[77] Vijay S Pande, Alexander Yu Grosberg, and Toyoichi Tanaka. Heteropolymer freezing and design: towards physical models of protein folding. *Reviews of Modern Physics*, 72(1):259, 2000.

[78] Sanzo Miyazawa and Robert L Jernigan. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules*, 18(3):534–552, 1985.

[79] Oscar D Monera, Terrance J Sereda, Nian E Zhou, Cyril M Kay, and Robert S Hodges. Relationship of sidechain hydrophobicity and α-helical propensity on the stability of the single-stranded amphipathic α-helix. *Journal of peptide science: an official publication of the European Peptide Society*, 1(5):319–329, 1995.

[80] Hao Li, Chao Tang, and Ned S Wingreen. Designability of protein structures: A lattice-model study using the miyazawa-jernigan matrix. *Proteins: Structure, Function, and Bioinformatics*, 49(3):403–412, 2002.

[81] Tomoei Takahashi, George Chikenji, and Kei Tokita. The cavity method to protein design problem. *Journal of Statistical Mechanics: Theory and Experiment*, 2022(10):103403, 2022.

[82] Hans A Bethe. Statistical theory of superlattices. *Proceedings of the Royal Society of London. Series A-Mathematical and Physical Sciences*, 150(871):552–575, 1935.

[83] Martin Weigt, Robert A White, Hendrik Szurmant, James A Hoch, and Terence Hwa. Identification of direct residue contacts in protein–protein interaction by message passing. *Proceedings of the National Academy of Sciences*, 106(1):67–72, 2009.

[84] Judea Pearl. Belief propagation in hierarchical inference structure. 1982.

[85] Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan kaufmann, 1988.

[86] Yoshiyuki Kabashima and David Saad. Belief propagation vs. tap for decoding corrupted messages. *EPL (Europhysics Letters)*, 44(5):668, 1998.

[87] Matt E Oates, Pedro Romero, Takashi Ishida, Mohamed Ghalwash, Marcin J Mizianty, Bin Xue, Zsuzsanna Dosztanyi, Vladimir N Uversky, Zoran Obradovic, Lukasz Kurgan, et al. D2p2: database of disordered protein predictions. *Nucleic acids research*, 41(D1):D508–D516, 2012.

[88] Peter E Wright and H Jane Dyson. Intrinsically disordered proteins in cellular signalling and regulation. *Nature reviews Molecular cell biology*, 16(1):18–29, 2015.

[89] Steven Boeynaems, Simon Alberti, Nicolas L Fawzi, Tanja Mittag, Magdalini Polymenidou, Frederic Rousseau, Joost Schymkowitz, James Shorter, Benjamin Wolozin, Ludo Van Den Bosch, et al. Protein phase separation: a new phase in cell biology. *Trends in cell biology*, 28(6):420–435, 2018.

[90] Hidetoshi Nishimori. Exact results and critical properties of the ising model with competing interactions. *Journal of Physics C: Solid State Physics*, 13(21):4071, 1980.

[91] Pál Ruján. Finite temperature error-correcting codes. *Physical review letters*, 70(19):2968, 1993.

[92] Hidetoshi Nishimori. Optimum decoding temperature for error-correcting codes. *Journal of the Physical Society of Japan*, 62(9):2973–2975, 1993.

[93] N Sourlas. Spin glasses, error-correcting codes and finite-temperature decoding. *EPL (Europhysics Letters)*, 25(3):159, 1994.

[94] Yukito Iba. The nishimori line and bayesian statistics. *Journal of Physics A: Mathematical and General*, 32(21):3875, 1999.

[95] Nicolas Sourlas. Spin-glass models as error-correcting codes. *Nature*, 339(6227):693–695, 1989.

[96] Yoshiyuki Kabashima and David Saad. Statistical mechanics of error-correcting codes. *EPL (Europhysics Letters)*, 45(1):97, 1999.

[97] Ido Kanter and David Saad. Error-correcting codes that nearly saturate shannon's bound. *Physical Review Letters*, 83(13):2660, 1999.

[98] Yoshiyuki Kabashima, Tatsuto Murayama, and David Saad. Typical performance of gallager-type error-correcting codes. *Physical Review Letters*, 84(6):1355, 2000.

[99] Surya Ganguli and Haim Sompolinsky. Statistical mechanics of compressed sensing. *Physical review letters*, 104(18):188701, 2010.

[100] Florent Krzakala, Marc Mézard, François Sausset, YF Sun, and Lenka Zdeborová. Statistical-physics-based reconstruction in compressed sensing. *Physical Review X*, 2(2):021005, 2012.

[101] RF Soares, FD Nobre, and JRL De Almeida. Effects of a gaussian random field in the sherrington-kirkpatrick spin glass. *Physical Review B*, 50(9):6151, 1994.

[102] R Erichsen Jr, A Silveira, and SG Magalhaes. Ising spin glass in a random network with a gaussian random field. *Physical Review E*, 103(2):022133, 2021.

[103] Syed A Rizvi, Nhi Nguyen, Haoran Lyu, Ben Christensen, Josue Ortega Caro, Emanuele Zappala, Maria Brbic, Rahul M Dhodapkar, and DV Dijk. Ampnet: Attention as message passing for graph neural networks. *arXiv preprint arXiv:2210.09475*, 2022.