NP-Hardness Results for Protein Side-chain Packing

Tatsuya Akutsu
takutsu@ims.u-tokyo.ac.jp
Human Genome Center, Institute of Medical Science, University of Tokyo
4-6-1 Shirokanedai, Minato-ku, Tokyo 108, Japan

Abstract

This paper shows that the problem of finding a protein side-chain packing is computationally hard (NP-hard), where the problem is defined here as a combinatorial search problem using rotamer library. Although this result does not suggest a new method, it gives a justification for previous methods using such heuristics as simulated annealing, neural networks, genetic algorithms, and Gibbs sampling.

1 Introduction

The protein side-chain packing problem (*side-chain packing*, in short) is, given an amino acid sequence and spatial information on the backbone chain (the main chain), to find side-chain conformation with the minimum potential energy. More precisely, it seeks a set of χ $(\chi_1, \chi_2, \cdots)$ angles whose potential energy becomes the minimum, where positions of atoms in the main chain are fixed (see Fig. 1). This problem is important for protein structure prediction because positions of atoms in the side-chains are not determined directly by the *homology modeling* approach or the *threading* approach, and thus side-chain packing is required as the second stage of these approaches.

A variety of computational techniques have been also applied to side-chain packing. For example, simulated annealing [8], neural networks [7], genetic algorithms [11], Gibbs sampling [12] have been applied as well as other heuristic search techniques [1, 3, 9]. Although side-chain packing is a continuous search (optimization) problem, most methods treat this problem as a combinatorial search problem using *rotamer libraries* [1, 4, 10]. Recall that, in a rotamer library, a set of candidates of torsion angles $\{\chi^1, \chi^2, \cdots, \chi^k\}$ is associated with each type of amino acids, and each candidate is called a *rotamer*.

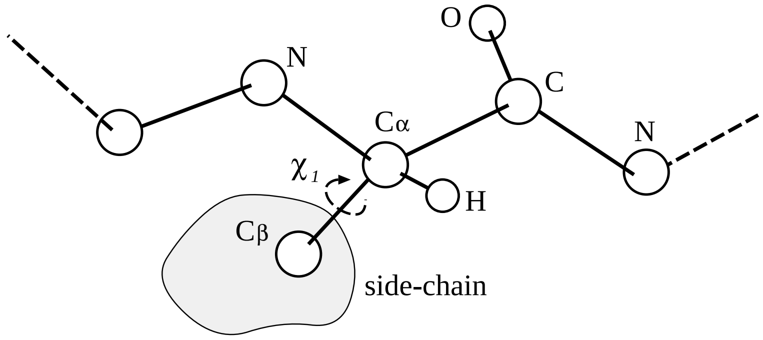

Figure 1: Side-chain packing is a problem of finding a set of torsion angles χ $(\chi_1, \chi_2, \cdots)$ which minimizes the potential energy where positions of atoms in the main chain are fixed.

In this paper, we prove that the side-chain packing problem using rotamers is NP-hard. Although this result does not suggest a new method, it is important because it gives a justification for using such heuristic methods as simulated annealing, neural networks, genetic algorithms, and Gibbs sampling. Moreover, this result gives a more concrete evidence of the difficulty of the side-chain packing problem than the arguments based on the size of search space give [3, 8]. Note that it is widely believed that NP-hard problems do not have polynomial time algorithms [5], and thus using a heuristic method is a good choice for an NP-hard problem. Moreover, we prove that *side-chain packing with perturbation*, which corresponds to a refinement procedure by MD (molecular dynamics), is also NP-hard.

2 Side-Chain Packing with Rotamers

2.1 Definition of the Problem

As described in Introduction, side-chain packing is a problem of finding side-chain conformation with the minimum potential energy. Although it is a continuous optimization problem, most methods treat this problem as a combinatorial search problem using rotamer libraries [1, 4, 10]. Since no formal definition is known, here we define the side-chain packing problem with a rotamer library as a geometric problem.

Let Σ be the set of amino acids (i.e., $|\Sigma| = 20$). For each amino acid $x \in \Sigma$, a fixed shape (3D region) $s(x)$ of the side-chain and a set of rotamers (i.e., a set of possible torsion angles) $r(x) = \{\chi_x^1, \cdots, \chi_x^{k_x}\}$ are given preliminary. Note that the axis of rotation is fixed to the line including C_α and C_β.

Let $a = a_1 \ldots a_n$ be an input amino acid sequence. Along with a sequence a, the initial configuration (determined by 3D position and Euler angles) of side-chain $s(a_i)$ is given for each amino acid a_i. Note that $s(a_i)$ can take k_{a_i} configurations determined from $r(a_i)$. Let $\chi_{a_i}^j(s(a_i))$ denotes the region of the side-chain of a_i when rotamer $\chi_{a_i}^j$ is selected. Then, the problem is to decide whether or not there exists a sequence of indices $I = (I_1, \ldots, I_n)$ such that $\chi_{a_i}^{I_i}(s(a_i)) \cap \chi_{a_j}^{I_j}(s(a_j)) = \emptyset$ holds for all $i \neq j$.

Note that this definition looks strange since it is defined as a geometric decision problem although the original problem is an energy minimization problem. However, there is no problem since we consider a hardness result. Note that the original energy minimization problem is much harder than this decision problem because the potential energy will be $+\infty$ if two side-chains intersect.

Although spatial information on the main chain is not considered in this definition, the result can be modified for a case that the main chain is taken into account. In this definition, rotamers on χ_1 angle (i.e., rotations on $C_\alpha C_\beta$ bonds) are only considered, and rotamers on $\chi_2, \chi_3, \cdots$ are not considered since the effects of $\chi_2, \chi_3, \cdots$ are usually less important than that of χ_1. Although we could modify the definition and the proof so that rotamers on $\chi_2, \chi_3, \cdots$ are taken into account, it would be complicated, and thus we consider rotamers on χ_1 angle only.

2.2 Hardness

In this section, we will show a hardness result for side-chain packing. Note that, in the proof below, we construct a *virtual* protein structure, in which the shapes of side-chains are far from those of real proteins. However, we use this construction for the sake of simplicity of the presentation, and it can be modified so that structures similar to real proteins are constructed.

Theorem 1: Protein side-chain packing with a rotamer library is NP-hard.
(Proof) We use a reduction from 3SAT [5]. Recall that 3SAT is, given a collection $C = \{c_1, \ldots, c_m\}$ of clauses over a set $V = \{v_1, \ldots, v_n\}$ of variables, to decide whether or not there exists a truth assignment for V that satisfies all clauses, where each clause consists of three literals. Recall also that a truth assignment is a function from V to $\{1(\text{true}), 0(\text{false})\}$, and a clause is satisfied if at least one literal (variable or its negation) becomes true. For example, all clauses in $C = \{\{a, b, \overline{c}\}, \{a, \overline{b}, c\}, \{\overline{a}, b, c\}\}$ are satisfied by an assignment $a = 1, b = 1, c = 1$, but not satisfied by an assignment $a = 0, b = 0, c = 1$.

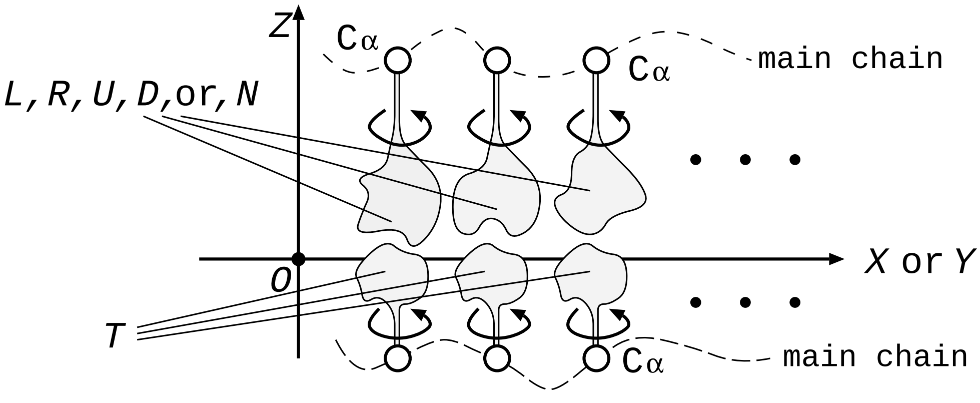

Figure 2: Placement of side-chains in the proof of Theorem 1.

From an instance of 3SAT, we construct an instance of side-chain packing. In the construction, we use 6 kinds of side-chains (residues): L (left), R (right), U (up), D (down), N (neutral), T (truth assignment). For each grid point (i,j) $(1 \leq i \leq n,\ 1 \leq j \leq m)$ on XY-plane, we put two side-chains as in Fig. 2. Although we do not take care of the main chain, we consider a situation that Cα atoms are placed at $(i,j,-1)$ and $(i,j,2)$. Note that the axis of rotation of each side-chain is the line including $(i,j,0)$ and parallel to Z-axis. The construction consists of two parts: *truth assignment* part and *satisfaction testing* part.

First we describe the truth assignment part (see Fig. 3(A)). For each variable v_i, we construct m side-chains $T_{i,1}, T_{i,2}, \ldots, T_{i,m}$, where the shape of $T_{i,j}$ is defined by

$$\begin{aligned} s(T_{i,j}) &= \{(x,y,0)|\ i-0.4 < x < i+0.4, j-0.6 < y < j\} \cup \\ &\quad \{(x,y,0)|\ i-0.5 < x < i-0.4, j < y < j+0.6\} \cup \\ &\quad \{(x,y,0)|\ i+0.4 < x < i+0.5, j < y < j+0.6\}. \end{aligned}$$

For each $T_{i,j}$, we associate the rotamer set $\{\chi_0, \chi_\pi\}$, where χ_θ means the rotation by θ radian on the line including (i,j) and parallel to Z-axis. Note that, once χ_0 is selected for some $T_{i,j}$, χ_0 must be selected for all $T_{i,k}$ with $1 \leq k \leq m$. Otherwise, some pair $(T_{i,k}, T_{i,k+1})$ would intersect. Thus, we consider the following correspondence:

$$\chi_0 \text{ is selected for } T_{i,j} \iff v_i \text{ is 1 (true).}$$

Next we describe the satisfaction testing part (see Fig. 3(B)). For this part, 5 kinds of side-chains L, R, U, D, N are used. Let $\alpha_{i,j}$, $\beta_{i,j}$ and $\gamma_{i,j}$ be the regions defined by

$$\begin{aligned} \alpha_{i,j} &= \{(x,y,1)\ |\ i-1 < x < i, j-0.5 < y < j+0.5\}, \\ \beta_{i,j} &= \{(x,y,1)\ |\ i < x < i+1, j-0.5 < y < j+0.5\}, \\ \gamma_{i,j} &= \{(i,y,1)\ |\ j-0.5 < y < j\}. \end{aligned}$$

Then, shapes of $L_{i,j}$, $R_{i,j}$, $U_{i,j}$, $D_{i,j}$ and $N_{i,j}$ are defined by

$$\begin{aligned} &s(L_{i,j}) = \alpha_{i,j} \cup \{(i,j+0.4,0)\}, \quad s(R_{i,j}) = \beta_{i,j} \cup \{(i,j+0.4,0)\}, \quad s(N_{i,j}) = \alpha_{i,j}, \\ &s(U_{i,j}) = \gamma_{i,j} \cup \{(i,j+0.4,0)\}, \quad s(D_{i,j}) = \gamma_{i,j} \cup \{(i,j-0.4,0)\}, \end{aligned}$$

where the rotamer set $\{\chi_0, \chi_\pi\}$ is associated with each of $L_{i,j}$, $R_{i,j}$ and $N_{i,j}$, and the rotamer set $\{\chi_0, \chi_{2\pi/3}, \chi_{-2\pi/3}\}$ is associated with each of $U_{i,j}$ and $D_{i,j}$. Thus, for example, $\chi_\pi(s(L_{i,j})) = \beta_{i,j} \cup \{(i,j-0.4,0)\}$.

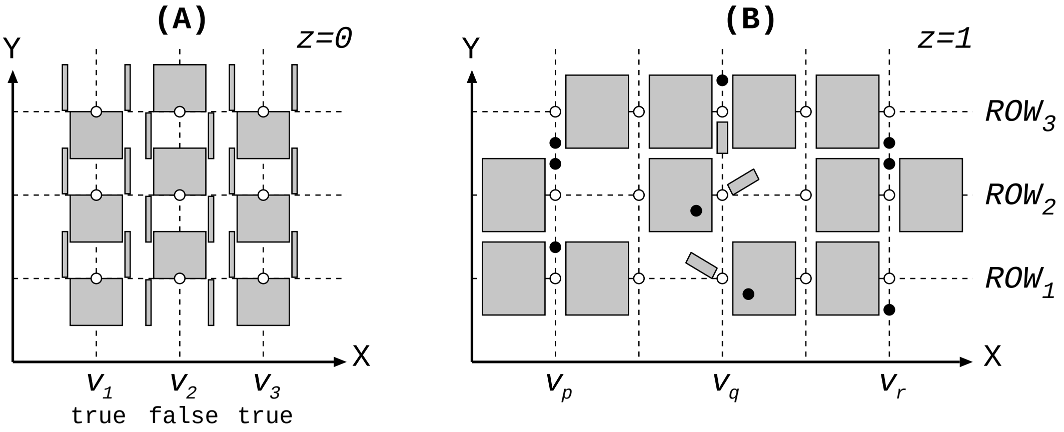

Figure 3: Illustration of (A) truth assignment part and (B) satisfaction testing part, where white circles denote grid points (axes of rotations). Fig. (A) corresponds to an assignment of $v_1v_2v_3 = 101$. In Fig. (B), ROW_1, ROW_2 and ROW_3 correspond to $v_pv_qv_r = 100$, 101 and 010, respectively, where black circles lie on the plane of $z = 0$. Note that χ_0 is selected for v_p in ROW_1 while χ_π is selected for v_r in ROW_1 because $x_{p,1} = L_{p,1}$ and $x_{r,1} = R_{r,1}$.

From each clause c_k, we construct side-chains $x_{i,k}$ $(1 \leq i \leq n)$. Let v_p, v_q, v_r be variables appearing in clause c_k, where $p < q < r$. Then, $x_{p,k}$, $x_{q,k}$ and $x_{r,k}$ are defined by

$$x_{p,k} = \begin{cases} L_{p,k}, & \text{if } v_p \in c_k, \\ R_{p,k}, & \text{if } \overline{v_p} \in c_k, \end{cases} \quad x_{q,k} = \begin{cases} U_{q,k}, & \text{if } v_q \in c_k, \\ D_{q,k}, & \text{if } \overline{v_q} \in c_k, \end{cases} \quad x_{r,k} = \begin{cases} R_{r,k}, & \text{if } v_r \in c_k, \\ L_{r,k}, & \text{if } \overline{v_r} \in c_k, \end{cases}$$

where $\overline{v}$ denotes the negation of v. For $i \notin \{p, q, r\}$, we let $x_{i,k} = N_{i,k}$.

Here, we consider the case of $c_k = \{v_p, v_q, v_r\}$ (i.e., all literals are positive). Note that, in this case, $x_{p,k} = L_{p,k}$, $x_{q,k} = U_{q,k}$ and $x_{r,k} = R_{r,k}$. Then, we can see the following properties must hold in order to avoid collisions among side-chains:

(i) If χ_0 is selected for $T_{p,k}$ (resp. $T_{r,k}$), χ_0 must be selected for $L_{p,k}$ (resp. $R_{r,k}$). Otherwise, χ_π must be selected for $L_{p,k}$ (resp. $R_{r,k}$).

(ii) If χ_0 is selected for $T_{q,k}$, χ_0 must be selected for $U_{q,k}$. Otherwise, either $\chi_{2\pi/3}$ or $\chi_{-2\pi/3}$ must be selected for $U_{q,k}$.

(iii) If $\chi_{-2\pi/3}$ is selected for $U_{q,k}$, χ_0 must be selected for $L_{p,k}$ and $N_{i,k}$ with $i < q$.

(iv) If $\chi_{2\pi/3}$ is selected for $U_{q,k}$, χ_0 must be selected for $R_{r,k}$ and χ_π must be selected for $N_{i,k}$ with $i > q$.

From (i) and (ii), we can see that selecting χ_0 for $L_{p,k}$ (resp. $U_{q,k}$, $R_{r,k}$) corresponds to an assignment of $v_p = 1$ (resp. $v_q = 1$, $v_r = 1$). From (iii) and (iv), we can see that χ_0 must be selected for at least one of $L_{p,k}$, $U_{q,k}$ and $R_{r,k}$ in order to avoid collisions. Thus, c_k is satisfiable iff. (if and only if) $x_{i,k}$'s $(1 \leq i \leq n)$ do not intersect.

For the other cases (i.e., cases of $c_k \neq \{v_p, v_q, v_r\}$), similar properties hold and thus c_k is satisfiable iff. $x_{i,k}$'s $(1 \leq i \leq n)$ do not intersect.

Therefore, there exists an assignment of rotamers to side-chains which does not cause collisions among side-chains iff. there exists a truth assignment satisfying all clauses in 3SAT.

Since the above construction can be done in polynomial time, the theorem holds. □

Note that, in the above proof, rotamer flip may propagate to the very distant rotamers. But, such a phenomenon may occur in a real protein because there is a case that substitution of one residue drastically changes 3D conformation of a protein.

3 Side-Chain Packing with Perturbation

Although the conformation of the main chain is fixed in the above, better side-chain packing may be obtained if the conformation of the main chain and side-chains is perturbed. Indeed, in many methods based on homology modeling, protein conformations are refined by perturbing both the main chain and side-chains using MD (molecular dynamics) [4, 9]. However, MD programs do not always find global optimals. Thus, we consider here a problem of protein side-chain packing with perturbation.

As in Section 2.2, we ignore the main chain and consider side-chains only, and we do not consider the potential energy but consider whether or not collisions are caused. Moreover, we do not consider rotamers and we assume that the initial configuration of side-chains is fixed, in which side-chains may intersect. Then, we define *side-chain packing with perturbation* as a geometric problem: given real numbers ϵ and δ, shapes of side-chains and an initial configuration, to find a configuration such that no two side-chains intersect, where each side-chain can be translated by at most ϵ and can be rotated by $\theta \in [-\delta, \delta]$ on the centroid. Although a lot of studies have been done for geometric packing problems [2], we do not know a result which can be directly applied to this problem.

Theorem 2: Protein side-chain packing with perturabation is NP-hard.
(Proof) We use a reduction from PARTITION [5]. Recall that PARTITION is, given a set of integers $S = \{s_1, \ldots, s_n\}$, to decide whether or not there exists a subset $S' \subseteq S$ such that $\sum_{s_i \in S'} s_i = \sum_{s_j \in S - S'} s_j$. In this proof, we consider a special case that S' contains exactly one of s_{2i-1}, s_{2i} for all i since this case remains NP-complete [5].

From such an instance of PARTITION, we construct an instance of protein packing with perturbation. For the simplicity, we consider a two-dimensional case where only translations within ϵ are allowed. However, the proof can be modified for a three-dimensional case with translations and (small) rotations.

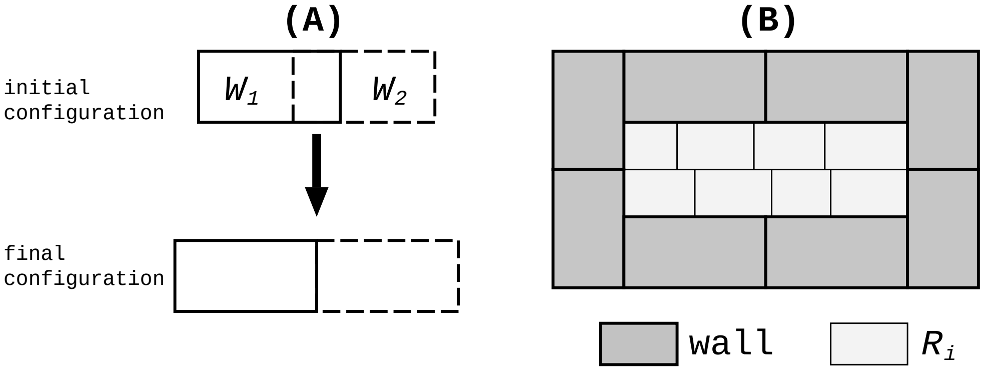

Figure 4: Constructions used in the proof of Theorem 2: (A) construction of a wall, (B) final configuration of the rectangles.

First we construct a *wall* (see Fig. 4(A)), Let W_1 and W_2 be rectangles of width $2L$ and height L, where we assume that edges are not included in it. We denote the position of rectangle by the position of its centroid. Initially, W_1 and W_2 are placed at $(-L + \epsilon, 0)$ and $(L - \epsilon, 0)$ respectively, where we assume $L > 2\epsilon$. Then, W_1 and W_2 must move to $(-L, 0)$ and $(L, 0)$ in order to avoid the collision.

Using four such pairs, we construct a wall by which rectangular region $\{(x,y)|0 \leq x \leq B, -A \leq y \leq A\}$ is surrounded (see Fig. 4(B)), where $A = \sum_{i=1}^{n} s_i$ and $B = (n+1)A/2$.

From each s_i, we construct a rectangle R_i of width $A + s_i$ and height A (see Fig. 4(B)). In the initial configuration, both R_{2i-1} and R_{2i} are placed at $((i-(1/2))A, 0)$. Finally, we set $\epsilon = A$.

Then, it is easy to see that there exists a configuration which does not cause collisions iff. there exists a subset S' for PARTITION. Since the above construction can be done in polynomial time, the theorem holds. □

Although rectangles are used in the above proof, rectangles can be replaced with more molecular-like shapes. However, we do not know whether or not they can be replaced by spheres.

4 Concluding Remarks

In this paper, we have proved NP-hardness results for side-chain packing with rotamers and side-chain packing with perturbation. Note that, in these proofs, we did not use a model of real proteins but used simplified artificial models although our models could be modified in some extent, where a model includes shapes of residues and a rotamer library. Since our models are different from a model of real proteins, there may exist a polynomial time algorithm for a real model. However, if such an algorithm exists, it must be specialized to the real model (otherwise it can be applied to the problems defined in this paper). Therefore, it seems difficult to develop a polynomial time algorithm for each problem even if a real protein model is used. Of course, robust proofs (i.e., proofs which can be applied to wide variety of models) [6] are more desirable and should be studied.

Acknowledgments

This work was supported in part by a Grant-in-Aid "Genome Science" for Scientific Research on Priority Areas from The Ministry of Education, Science, Sports and Culture in Japan.

References

[1] Bower, M. J., Cohen, F. E., Dunbrack, R. L., "Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: a new homology modeling tool," *J. Mol. Biol.*, 267:1268–1282, 1997.

[2] Daniels, K. and Milenkovic, V. J., "Multiple translational containment. Part I: an approximate algorithm," *Algorithmica*, 19:148–182, 1997.

[3] Desmet, J., De Maeyer, M. Hazes, B. and Lasters, I., "The dead-end elimination theorem and its use in protein side-chain positioning," *Nature*, 356:539–542, 1992.

[4] Dunbrack, R. L. and Karplus, M., "Backbone-dependent rotamer library for proteins application to side-chain prediction," *J. Mol. Biol.*, 230:543–574, 1993.

[5] Garey, M. R. and Johnson, D. S., *Computers and Intractability: A Guide to the Theory of NP-completeness*, Freeman, NY, 1979.

[6] Hart, W. E. and Istrail, S., "Robust proofs of NP-hardness for protein folding: general lattices and energy potentials," *J. Computational Biology*, 4:1–22, 1997.

[7] Hwang, J. K. and Liao, W. F., "Side-chain prediction by neural networks and simulated annealing optimization,“ *Protein Eng.*, 8:363–370, 1995.

[8] Lee, C. and Subbiah, S., "Prediction of protein side-chain conformation by packing optimization," *J. Mol. Biol.*, 217:373–388, 1991.

[9] Levitt, M., "Accurate modeling of protein conformation by automatic segment matching," *J. Mol. Biol.*, 226:507–533, 1992.

[10] Ponder, J. W. and Richards, F. M., "Tertiary templates for proteins: use of packing criteria in the enumeration of allowed sequences for different structural classes," *J. Mol. Biol.*, 193:775–791, 1987.

[11] Tuffery, P., Etchebest, C., Hazout, S. and Lavery, R., "A new approach to the rapid determination of protein side-chain conformations," *J. Biomol. Struct. Dynam.*, 8:1267–1289, 1991.

[12] Vasquez, M., "An evaluation of discrete and continuum search techniques for conformational analysis of side-chains in proteins," *Biopolymers*, 36:53–70, 1995.