BetaMDGP: Protein Structure Determination Algorithm Based on the Beta-complex

Jeongyeon Seo[1], Jae-Kwan Kim[2], Joonghyun Ryu[2], Carlile Lavor[3], Antonio Mucherino[4], and Deok-Soo Kim[1,2,⋆]

[1] Department of Industrial Engineering, Hanyang University
17 Haengdang-Dong, Sungdong-Ku, Seoul, 133-791, South Korea
jyseo@voronoi.hanyang.ac.kr, dskim@hanyang.ac.kr
[2] Voronoi Diagram Research Center, Hanyang University
17 Haengdang-Dong, Sungdong-Ku, Seoul, 133-791, South Korea
{jkkim,jhryu}@voronoi.hanyang.ac.kr, dskim@hanyang.ac.kr
[3] Dept. of Applied Math. (IMECC-UNICAMP), University of Campinas
13081-970, Campinas - SP, Brazil
clavor@ime.unicamp.br
[4] IRISA, University of Rennes I, France
antonio.mucherino@irisa.fr

Abstract. The molecular distance geometry problem (MDGP) is a fundamental problem in determining molecular structures from the NMR data. We present a heuristic algorithm, the BetaMDGP, which outperforms existing algorithms for solving the MDGP. The BetaMDGP algorithm is based on the beta-complex, which is a geometric construct extracted from the quasi-triangulation derived from the Voronoi diagram of atoms. Starting with an initial tetrahedron defined by the centers of four closely located atoms, the BetaMDGP determines a molecular structure by adding one shell of atoms around the currently determined substructure using the beta-complex. The proposed algorithm has been entirely implemented and tested with atomic arrangements stored in an NMR format created from PDB files. Experimental results are also provided to show the powerful capability of the proposed algorithm.

Keywords: Protein structure determination, Molecular Distance Geometry Problem, Voronoi Diagram, Quasi-triangulation, Beta-complex.

1 Introduction

One of the key challenges for understanding a protein function is understanding its structure as it is the determinant of molecular function [1]. There are two main experimental methods to determine protein structures: NMR spectroscopy [2] and X-ray crystallography [3]. Given an NMR spectroscopy file that defines the interatomic distances for some pairs of atoms, usually between the hydrogen atoms in a molecule, determination of the optimal assignment of coordinates that

⋆ Corresponding author.

M.L. Gavrilova, C.J.K. Tan (Eds.): Trans. on Comput. Sci. XXII, LNCS 8360, pp. 130–155, 2014.

satisfies the inter-distance constraints is required. The determinant of molecular structure from NMR spectroscopy is studied by solving the distance geometry problem (DGP), which is a well-known mathematical problem. The DGP is to find an embedding of a weighted undirected graph $G = (V, E, d)$ in an arbitrary dimensional space [4–6]. Each vertex $v \in V$ corresponds to a point x_v in space, and there is an edge between the two vertices if and only if their relative distance is known. The length d of an edge is its weight. Formally, the DGP is the problem of finding the location of x such that $u, v \in V$, $\forall (u, v) \in E$, and $|| x_u - x_v || = d_{u,v}$, where $d_{u,v}$ is the distance between u and v. Hence, the DGP is called a constraint satisfaction problem from a mathematical point of view because a set of coordinates must be found to satisfy the constraints. The DGP can be solved in polynomial time if the complete set of the exact distances is available [7] but is NP-hard for a general sparse set of distances even in three-dimensional space [8]. In other words, it is very difficult to correctly solve for the general setting of NMR spectroscopy in practice because it contains only a subset of the complete graph between hydrogen atoms.

We are interested in a particular class of the DGP called the molecular distance geometry problem (MDGP) arising in biology where the vertices of G represent the atom centers of a molecule. The aim of the MDGP is to identify the three-dimensional molecular conformation in three dimensional space using the Euclidean distance. The MDGP is of crucial importance for biomedical problems because a molecular function is primarily determined by its structure. While X-ray crystallography produces the absolute coordinates of the atom locations, the NMR produces the relative distance information among the atoms, usually within 5 Å [2]. Hence, the MDGP is the core problem for NMR technology.

Let x_i be the coordinate of the atom i and D the given set of the distance $d_{i,j}$ between the atom i and the atom j, $i \neq j$. The problem is to find x_i, $i = 1, ..., n$ such that

$$||x_i - x_j|| = d_{i,j}, \qquad \forall d_{i,j} \in D. \tag{1}$$

The most common approach to the MDGP is to formulate the problem as a continuous optimization problem [9–13].

$$Min. \sum_{(i,j) \in D} (||x_i - x_j||^2 - d_{i,j}^2)^2 \tag{2}$$

In real NMR files, the distances are given with the lower and the upper bounds [14]. The MDGP with the lower and the upper bounds is to find a set of positions $x_1, ..., x_n$ in the three-dimensional space such that

$$l_{i,j} \leq ||x_i - x_j|| \leq u_{i,j}, \qquad \forall d_{i,j} \in D \tag{3}$$

where l_{ij} and u_{ij} are the lower and the upper bounds on the distances, respectively.

The standard formulation by Crippen and Havel [5] is to solve the following minimization problem:

$$Min. \sum_{(i,j) \in D} p_{i,j}(x) \tag{4}$$

$$p_{i,j}(x) = Min^2\{\frac{||x_i - x_j||^2 - l_{i,j}^2}{l_{i,j}^2}, 0\} + Max^2\{\frac{||x_i - x_j||^2 - u_{i,j}^2}{u_{i,j}^2}, 0\} \quad (5)$$

Crippen and Havel applied the MDGP to protein modeling [5, 15, 7]. The MDGP has been studied by many groups: the embedding algorithm approach by Crippen and Havel [5, 15], the graph reduction approach by Hendrickson [16, 17], the approaches based on the global optimization method by Moré and Wu [9, 10] and An and Tao [11, 12], and the geometric build-up algorithm by Dong, Wu, Sit, and Yuan [18–21]. In particular, the embedding algorithm by Crippen and Havel has been adopted in NMR modeling through programs such as CNS, XPLOR, and XPLOR-NIH [22, 23]. Under certain assumptions, the problem can be formulated as a combinatorial optimization problem, called the discretizable MDGP (DMDGP) [24]. While the NP-hardness of the problem is unavoidable [24], the Branch and Prune (BP) algorithm solves the DMDGP effectively and efficiently for proteins [25]. Previous approaches usually become numerically unstable as solution process progresses because the number of constraints to determine the coordinate of a new atom gets larger.

In this regard, we propose a heuristic algorithm, called the BetaMDGP, to maintain a constant number of constraints to determine the coordinate of a new atom. The BetaMDGP algorithm to the MDGP uses the beta-complex, which is a derivative geometric construct from the Voronoi diagram of atoms and effectively provides the proximity information among atoms [26–28]. Using the beta-complex, the BetaMDGP reduces the number of distance constraints required for determining new atom coordinate and thus finds the solution very efficiently. While the previous approaches are numerically unstable, the BetaMDGP provides the more stable solution compared to the previous algorithms because the BetaMDGP keeps the number of constraints constantly during the solution process. The BetaMDGP consists of two parts. First, we determine the coordinates for the centers of four nearby atoms to define the tetrahedral seed structure to start the process. Second, the BetaMDGP adds other atoms around the boundary of this determined substructure (at the beginning, it is the seed structure) using the beta-complex. The molecular structure is determined by sufficiently repeating this second procedure. It turns out that the proposed algorithm, in its current form, determines the protein structures very effectively and efficiently compared to existing algorithms. All figures of the molecular structures are created by the BetaMol program developed by the VDRC (http://voronoi.hanyang.ac.kr) that is free to download [29].

2 Methods

The proposed algorithm is based on three geometric constructs: the Voronoi diagram of atoms, the quasi-triangulation, and the beta-complex. Consider the set $P = \{p_1, p_2, \ldots, p_n\}$ where $p_i \in P$ is a point in three-dimensional space. The ordinary Voronoi diagram VD of P is the tessellation of the space with a set of n Voronoi cells (V-cells) where the V-cell $\mathrm{VC}(p_i)$ is the set of the locations

that are closer to p_i than to the others. Consider the set $A = \{a_1, a_2, \ldots, a_n\}$ where $a_i = (p_i, r_i) \in A$ is a spherical atom with the center p_i and radius r_i in three-dimensional space. The Voronoi diagram $\mathcal{VD}$ of A is the tessellation of the space with a set of n V-cells where the $\mathrm{VC}(a_i)$ is the set of the locations that are closer to the boundary of a_i than to the boundary of any other atom. $\mathcal{VD}$ is more formally called the additively weighted Voronoi diagram in computational geometry and is different from the ordinary Voronoi diagram of points VD. $\mathcal{VD}$ can be represented as $\mathcal{VD} = (V^{\mathcal{V}}, E^{\mathcal{V}}, F^{\mathcal{V}}, C^{\mathcal{V}})$ where the Voronoi vertex (V-vertex) $v^{\mathcal{V}} \in V^{\mathcal{V}}$ corresponds to the center of the empty sphere tangent to the boundaries of four nearby atoms; the Voronoi edge (V-edge) $e^{\mathcal{V}} \in E^{\mathcal{V}}$ corresponds to the locus of the center of the empty sphere tangent to the boundaries of three nearby atoms; the Voronoi face (V-face) $f^{\mathcal{V}} \in F^{\mathcal{V}}$ corresponds to the locus of the center of the empty sphere tangent to the boundaries of two nearby atoms; the V-cell $c^{\mathcal{V}} \in C^{\mathcal{V}}$ corresponds to an atom. The topology among the V-vertices, V-edges, V-faces, and V-cells in $\mathcal{VD}$ are usually maintained in a radial-edge data structure [30]. $\mathcal{VD}$ can be computed in $O(n^3)$ time for general spherical balls in the worst case but takes $O(n)$ time for molecular atoms on average. See [31] for $\mathcal{VD}$ and see [32] for the Voronoi diagram in general.

Applications of the Voronoi diagram use the traversal on its topology structure, and the dual of the Voronoi diagram is frequently used for this purpose because it simplifies the traversal algorithms [33, 34]. The dual structure of the ordinary Voronoi diagram VD is well-known as the Delaunay triangulation which has many powerful properties primarily for it being a simplicial complex [32]. However, the dual of the Voronoi diagram of atoms $\mathcal{VD}$, now known as the quasi-triangulation $\mathcal{QT}$, was recently defined and characterized by Kim and colleagues as follows. $\mathcal{QT} = (V^{\mathcal{Q}}, E^{\mathcal{Q}}, F^{\mathcal{Q}}, C^{\mathcal{Q}})$ where $v^{\mathcal{Q}} \in V^{\mathcal{Q}}$ is mapped from $c^{\mathcal{V}} \in C^{\mathcal{V}}$; $e^{\mathcal{Q}} \in E^{\mathcal{Q}}$ is mapped from $f^{\mathcal{V}} \in F^{\mathcal{V}}$; $f^{\mathcal{Q}} \in F^{\mathcal{Q}}$ is mapped from $e^{\mathcal{V}} \in E^{\mathcal{V}}$; $c^{\mathcal{Q}} \in C^{\mathcal{Q}}$ is mapped from $v^{\mathcal{V}} \in V^{\mathcal{V}}$. Note that all the simplexes in $\mathcal{QT}$ are mapped from the simplexes in $\mathcal{VD}$ and all the mappings are one-to-one. The conversion between $\mathcal{VD}$ and $\mathcal{QT}$ can be done in $O(m)$ time in the worst case where m represents the number of simplexes in $\mathcal{QT}$. $\mathcal{QT}$ is known to have a phenomenon called an anomaly. For the details of $\mathcal{QT}$, see [35, 36, 27, 37].

The beta-complex corresponding to the real-value β is a subset of $\mathcal{QT}$ such that every simplex σ in $\mathcal{QT}$ is removed if a spherical probe of radius β can pass through σ without intersecting the atoms corresponding to it. Hence, each simplex in the beta-complex represents the proximity among some atoms within the molecular boundary. The beta-shape is defined by the region of the space bounded by the boundary of the beta-complex. Hence, the boundary of the beta-shape determines the proximity among the atoms on the molecular boundary with respect to the probe. We emphasize here that the beta-complex can be computed very efficiently from the quasi-triangulation, and its correctness is mathematically guaranteed. For the details, see [28, 26, 27].

Fig. 1 illustrates the idea of these geometric constructs in the plane. Fig. 1(a) shows a two-dimensional molecule A consisting of nine atoms. Fig. 1(b) is the Connolly surface of A corresponding to the black circular probe. Note that there

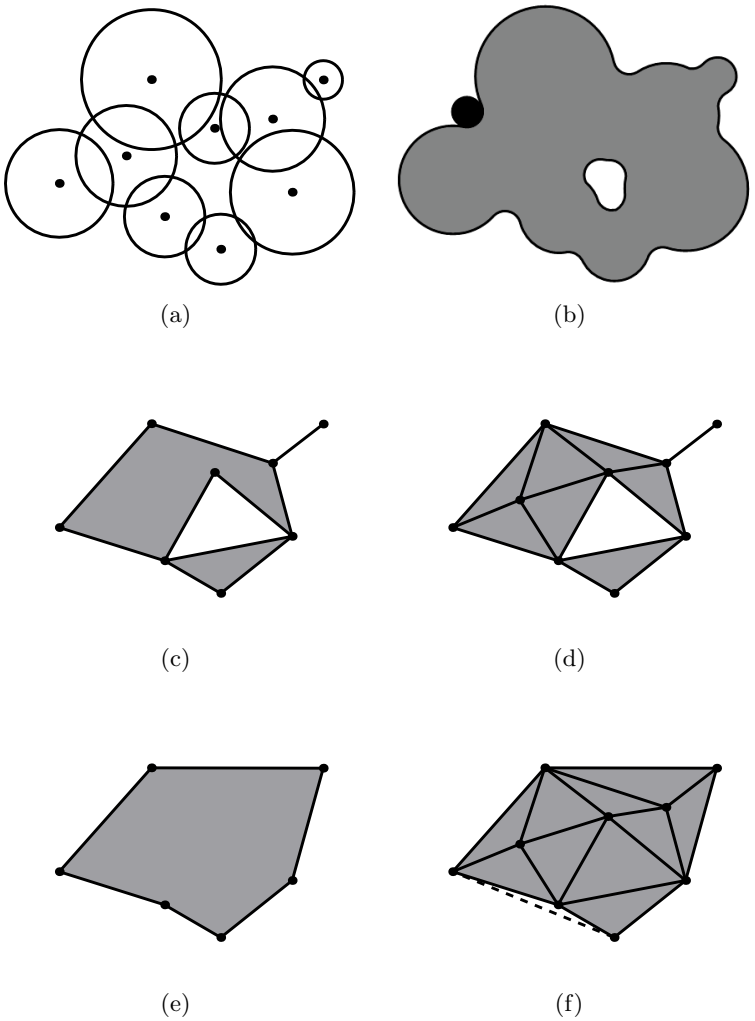


Fig. 1. Molecule, the Connolly surface, the beta-shape, and the beta-complexes in 2D. (a) A molecule (9 atoms), (b) the Connolly surface corresponding to a small probe, (c) the corresponding beta-shape, (d) the corresponding beta-complex, (e) a beta-shape corresponding to a larger probe, and (f) the corresponding beta-complex.

is an interior void. Fig. 1(c) shows the beta-shape corresponding to the Connolly surface of Fig. 1(b). The beta-shape has an interior void corresponding to the void of the Connolly surface and a dangling edge corresponding to a pair of atoms that are exposed to or touched by the probe. The boundary of the beta-shape has 8 vertices and 10 edges (7 on the exterior boundary and 3 on the interior void). Fig. 1(d) shows the corresponding beta-complex. Note that each vertex of the beta-shape and beta-complex corresponds to an atom. Fig. 1(e) and (f) show the beta-shape and the beta-complex corresponding to a larger probe, respectively. Note that both the dangling edge and the internal void have now disappeared. The dotted line segments in Fig. 1(f) together with the simplexes of the beta-complex form the quasi-triangulation of the molecule.

Based on these three constructs, the proposed BetaMDGP algorithm grows a molecular structure by adding one atom at a time that is selected by using the beta-complex for an appropriate value of the probe radius β. In the proposed algorithm, we start from tetrahedron τ consisting of four atoms which are guaranteed to be in a close neighborhood in a certain sense that will be described below. Then, we grow the structure by adding one shell of nearby atoms. Thus, we call the idea of this algorithm *"shell-growing."*

We first consider a two-dimensional example shown in Fig. 2. Suppose that Fig. 2(a) shows a true two-dimensional molecular structure that is stored in an NMR file. We first choose three nearby atoms which must form a (red-colored) seed triangle $t_0 = (a_1, a_4, a_{10})$ consisting of the centers of the three atoms a_1, a_4, and a_{10} in Fig. 2(b). The triangle t_0 can be determined by arbitrarily choosing one atom, say a_1, and two nearby atoms by looking at the distances to a_1. Let $T_0 = \{t_0\}$ and compute $\mathcal{BC}(T_0)$ whose boundary $\partial\mathcal{BC}(T_0)$ has three edges (the red chain in Fig. 2(b)). In this particular case, $\partial\mathcal{BC}(T_0)$ coincides with the boundary of the seed triangle t_0. We call $\partial\mathcal{BC}(T_0)$ the *shell* Sh_0 of T_0. Then, for each edge of Sh_0, we define another triangle by choosing another atom closest to two atoms consisting of the edge. After we determine the additional three triangles in such a fashion, say t_1, t_2, and t_3, we get Fig. 2(b). We call this operation shell-growing. Let $T_1 = \{t_0, t_1, t_2, t_3\}$. Then, we compute the beta-complex $\mathcal{BC}(T_1)$ for some value of β as shown in Fig. 2(c). Consider the red-colored $\partial\mathcal{BC}(T_1)$ the *shell* Sh_1 of T_1. Applying the shell-growing process once more by adding a new triangle for each edge $e \in Sh_1$, we get another set T_2 as shown in Fig. 2(d). Then, Fig. 2(e) shows the beta-complex $\mathcal{BC}(T_2)$ as well as T_2. Fig. 2(f) shows the last step of this model construction process to add the last atom a_7 which is under-determined. Note that a_7 can be placed at either a_7' or a_7'' without violating the distance constraint. Hence, there can be multiple solutions in the MDGP depending on the condition of the distance constraints in the input data. In such a case, however, adding another constraint on such an atom can uniquely determine the molecular structure. It is notable that such under-determined situations frequently arise in real NMR files.

Suppose that ∂T denotes the boundary of the union of the underlying space taken by each triangle $t \in T$. Then, $\partial\mathcal{BC}(T)$ may or may not be identical to ∂T. For example, Fig. 2(c) shows that ∂T_1 has six vertices but $\partial\mathcal{BC}(T_1)$ has five

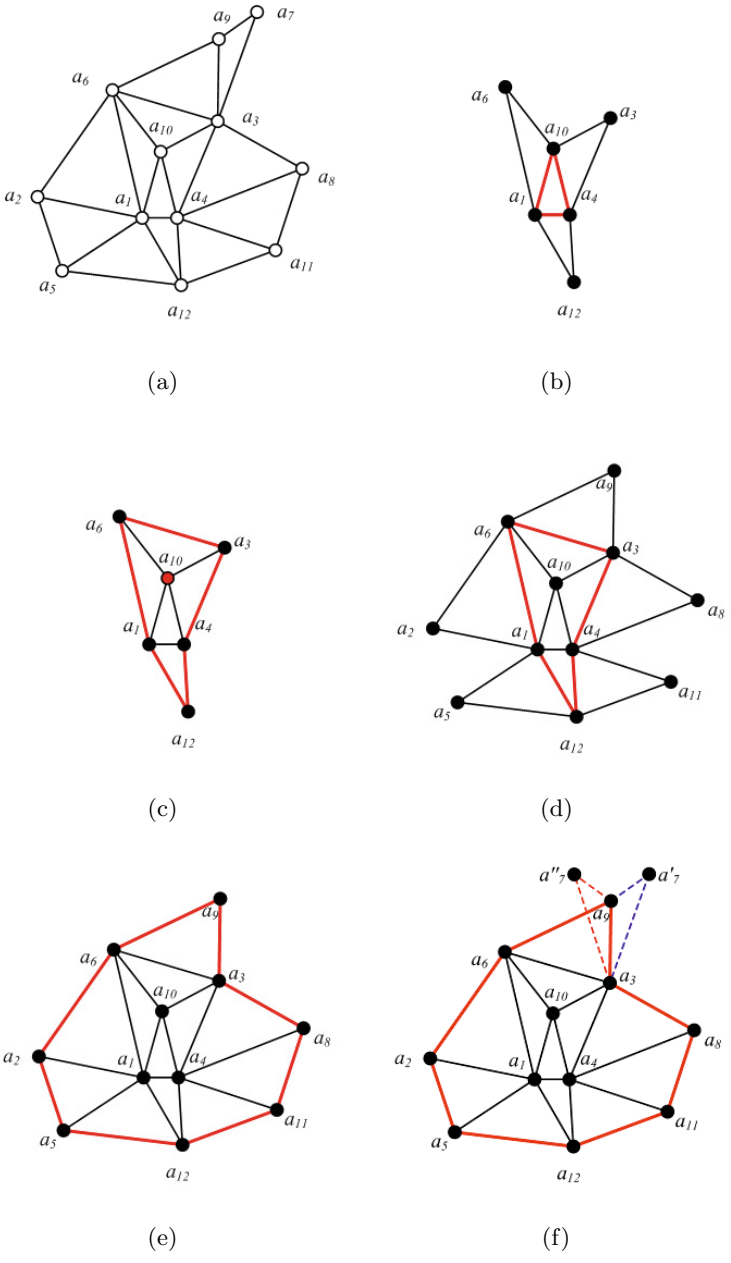


Fig. 2. The idea of the BetaMDGP algorithm in two-dimensional space. (a) The true structure to determine; (b) $T_0 = \{t_0\}$ (t_0 is the (red-colored) seed triangle) and $T_1 = \{t_0, t_1, t_2, t_3\}$; (c) $\mathcal{BC}(T_1)$ and (red-colored) $\partial\mathcal{BC}(T_1)$; (d) T_2; (e) $\mathcal{BC}(T_2)$; and (f) a multiple solution case in the MDGP.

vertices only: a_{10} does not appear on $\partial\mathcal{BC}(T_1)$. Therefore, when we compute T_2, we can safely ignore a_{10} from further consideration. In Fig. 2(e), if we compute $\partial\mathcal{BC}(T_2)$, we can now ignore three atoms from further consideration (i.e., a_1, a_4, and a_{10}). This reduction can contribute to the solution quality because it simplifies the solution process by removing the conflicting constraints as much as possible. It also contributes to algorithmic efficiency. The accumulation of the round-off error does not occur in the BetaMDGP, and thus the numerical stability is also improved. The three-dimensional MDGP can be similarly solved using the three-dimensional beta-complex. According to our experiment, the number of omitted atoms for the three-dimensional MDGP is significant.

Now, we consider a three-dimensional example of the BetaMDGP. See Fig. 3. Suppose that Fig. 3(a) shows the true three-dimensional molecular structure stored in an NMR file. We start the process with a seed tetrahedron $\tau_0 = (a_1, a_4, a_5, a_{11})$ consisting of the centers of four closely located atoms a_1, a_4, a_5, and a_{11}. The BetaMDGP algorithms grows τ_0 (the gray tetrahedron in Fig. 3(b)) by adding one shell of atoms around the current $T = \{\tau_0\}$ as follows: i) Compute the beta-complex of the current T for the appropriate β-value, ii) find the set ΔT of the new tetrahedron added to T for the faces on the boundary of the beta-complex $\partial\mathcal{BC}(T)$, and iii) $T = T \cup \Delta T$. Repeating this procedure a sufficient number of times correctly determines the structure of a molecule from the NMR data. In this paper, we use $\beta = 1.4$Å, which corresponds to the radius of the probe for a water molecule. The gray tetrahedron in Fig. 3(b) shows the seed tetrahedron τ_0. Let $T_0 = \{\tau_0\}$. The beta complex $\mathcal{BC}(T_0)$ is identical to τ_0. Then, for each face of τ_0, we define another tetrahedron by choosing the other atom closest to the vertices of the face. After we determine the additional four tetrahedron, say τ_1, τ_2, τ_3, and τ_4, we get Fig. 3(c) as the *shell-growing*. Let $T_1 = \{\tau_0, \tau_1, \tau_2, \tau_3, \tau_4\}$. Then, we compute the beta-complex $\mathcal{BC}(T_1)$ of T_1 for some value of β, as shown in Fig. 3(d). Consider $\partial\mathcal{BC}(T_1)$, i.e. the *shell* Sh_1 of T_1. Applying the shell-growing process once more using the faces on $\partial\mathcal{BC}(T_1)$, we get another set T_2 as the new tetrahedron added to T_1 for each face $f \in \partial\mathcal{BC}(T_1)$ as shown in Fig. 3(e). $\partial\mathcal{BC}(T_2)$ becomes Sh_2 as shown in Fig 3(f).

Suppose that ∂T denotes the boundary of the union of the underlying space taken by each tetrahedron $\tau_i \in T$. Then, like its two-dimensional counterpart, $\partial\mathcal{BC}(T)$ may or may not be identical to ∂T. For example, Fig. 3(d) shows that ∂T_1 has eight vertices but $\partial\mathcal{BC}(T_1)$ has seven vertices only: a_1 does not appear on $\partial\mathcal{BC}(T_1)$. Therefore, when we compute T_2, we can ignore a_1 from further consideration.

The following algorithm briefly describes the three-dimensional BetaMDGP algorithm. The input of the BetaMDGP algorithm is an atom set A where $a_i = (p_i, r_i) \in A$ is an atom with the unknown center p_i (but its radius r_i is known) and the distance set D where its element $d_{i,j} < \rho_{cutoff}$ is the inter-atomic distance between a_i and a_j. We used the usual cutoff distance 5 Å in order to simulate the NMR data. The output of the BetaMDGP algorithm is the atom set $\tilde{A}$ where $\tilde{a_i} = (\tilde{p_i}, r_i) \in \tilde{A}$ is an atom with the known coordinate of the center $\tilde{p_i}$. Step 1 determines the seed tetrahedron with the coordinates of the constituting

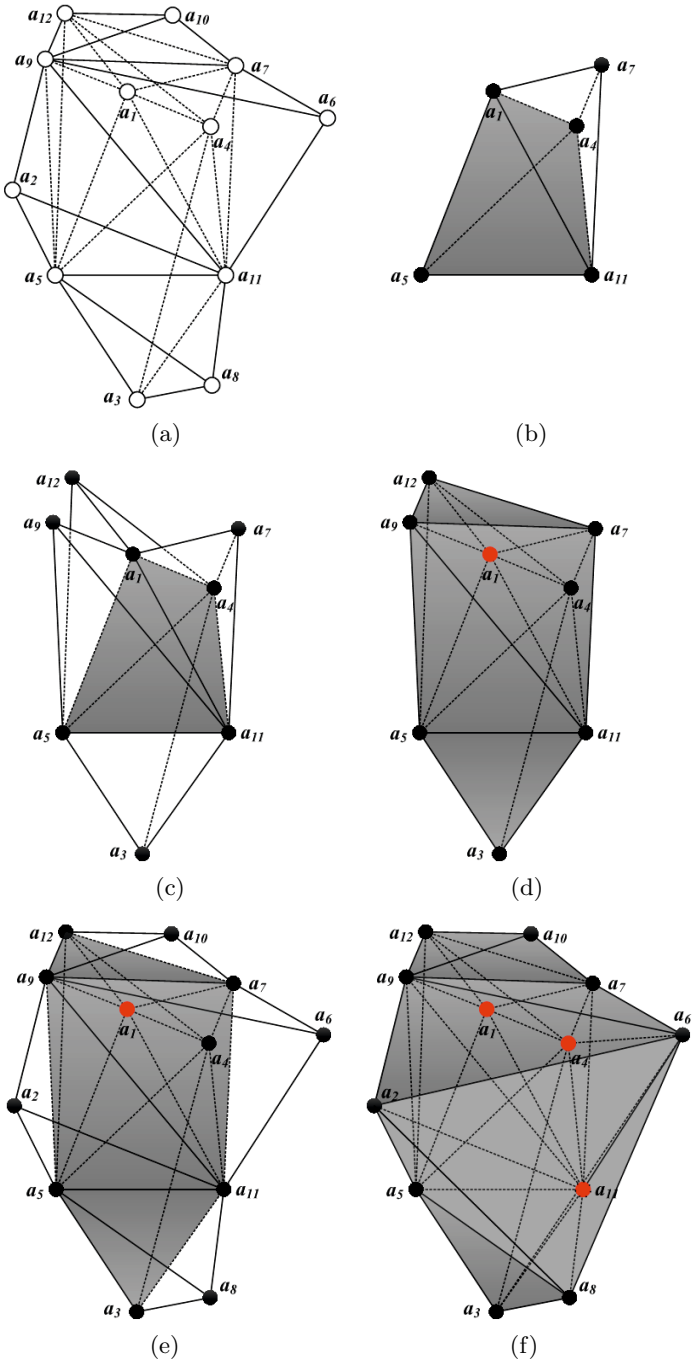


Fig. 3. The idea of the BetaMDGP algorithm in the three-dimensional space. (a) The true structure to determine; (b) $T_0 = \{\tau_0\}$ (τ_0 is the (gray-colored) seed tetrahedron) and τ_1 added to T_0; (c) $T_1 = \{\tau_0, \tau_1, \tau_2, \tau_3, \tau_4\}$; (d) $\mathcal{BC}(T_1)$; (e) T_2; and (f) $\mathcal{BC}(T_2)$

atoms. Step 2 performs the sell-growing procedure to determine as many atoms as possible. A newly determined atom a_p has three distances from the three atoms of f_p on $\partial\mathcal{BC}(\tilde{A})$. The average distance among a_p to the three atoms of f_p is $dist(a_p, f_p)$. However, the shell-growing procedure may not be able to exhaust all the atoms because there can be some atoms (which are called under-determined) where each does not have four distances from the atoms on $\partial\mathcal{BC}(\tilde{A})$. If such an under-determined atom exists, we choose an arbitrary location as long as it does not violate both its distance constraints and the well-packed molecular structure property.

Algorithm. `Three-dimensional BetaMDGP`
Input:
$A = \{a_1, a_2, \ldots, a_n\}$ where $a_i = (p_i, r_i) \in A$ is an atom with the unknown center p_i and the known radius r_i of a particular type in the NMR file
$D = \{d_{i,j} | d_{i,j}$ the distance between a_i and a_j, $d_{i,j} < \rho_{cutoff}\}$
Output:
$\tilde{A} = \{\tilde{a_1}, \tilde{a_2}, \ldots, \tilde{a_n}\}$ where $\tilde{a_i} = (\tilde{p_i}, r_i)$ with the known coordinate $\tilde{p_i}$
Step 1. Initialization:
Step 1.1. Make a seed tetrahedron τ_0 with four nearby atoms in A.
Step 1.2. Insert the four atoms of τ_0 to $\tilde{A}$.
Step 1.3. Determine the coordinates of the four atoms in $\tilde{A}$.
Step 1.4. $A \leftarrow A - \tilde{A}$
Step 2. Shell-growing: While $A \neq \emptyset$,
Step 2.1. Compute the beta-complex $\mathcal{BC}$ of $\tilde{A}$.
Step 2.2. Find the set $\mathcal{F}_\beta$ of the faces on $\partial\mathcal{BC}(\tilde{A})$.
Step 2.3. While $\mathcal{F}_\beta \neq \emptyset$,
- Get a face $f_p \in \mathcal{F}_\beta$, $\mathcal{F}_\beta \leftarrow \mathcal{F}_\beta - \{f_p\}$.
- Get an atom $a_p \in A$ which has three distances from the three atoms of f_p, $dist(a_p, f_p)$ is the shortest from f_p.
- $\tilde{A} \leftarrow \tilde{A} + \{a_p\}$ and determine the coordinate of a_p.
- $A \leftarrow A - \{a_p\}$
- If $A = \emptyset$, terminate the shell-growing process.

End-while.
Step 2.4. If such a_p does not exist,
- Go to Step 3.

End-if.
End-while.
Step 3. Marginal process: While $A \neq \emptyset$,
- Get an atom $a_i \in A$.
- $\tilde{A} \leftarrow \tilde{A} + \{a_i\}$ and determine the coordinate of a_i by using the distance related a_i.
- $A \leftarrow A - \{a_i\}$

End-while.
Step 4. Terminate.

3 Results

The proposed BetaMDGP algorithm has been validated through implementation and testing with data obtained from the PDB files. The input files contain the atomic pairs whose inter-atomic distances fall within the usual cutoff distance of 5Å. The experiment of the BetaMDGP algorithm shows extremely good solution quality in that the recovered structures are very close to the original PDB structures from geometric measures points of view such as the RMSD between the equivalent atoms in both the original PDB models and the reconstructed models, the existence and distribution of the interior voids, and the distribution of the covalent bond lengths. All the experimental results were visualized using the BetaMol program [29]. Note that all the reconstructed structures are displayed after it was superposed with the original PDB model using the structure superposition program, the BetaSuperpose [38]. We tested the BetaMDGP algorithm with three types of NMR data created from the PDB files: i) data without an interval (all atom types), ii) data with an interval (all atom types), and iii) data with hydrogen atoms with an interval. The computational environment is as follows: Intel Core2 Duo E6550 CPU and 4 GB memory on a Windows 7 Ultimate platform.

As the first test, we created NMR files according to the inter-atomic distances within a 5Å cutoff radius. In other words, we computed all the pairwise inter-atomic distances for all the atoms in each PDB file and output the atom pairs with an inter-atomic distance shorter than 5Å into an NMR file. See Fig. 4. The red structures in Fig. 4(a) and (d) show the true structures of 2lt8 (558 atoms) and 1xba (2068 atoms) in the PDB after we removed all the hetero atoms and water molecules. Note that 2lt8 in Fig. 4(a) was determined by NMR spectroscopy and therefore it is one (to be specific, the first one) of the ensembles. 1xba in Fig. 4(d) is from the X-ray crystallography. The blue structure in Fig. 4(b) shows the reconstructed structure by the BetaMDGP algorithm using the input file from the 2lt8. From the visual inspection, we can see that both Fig. 4(a) and (b) are very similar. Fig. 4(c) shows the ribbon models of both the structures after they were superposed. This figure shows that the backbones are almost identical. Fig. 4(e) and (f) are the reconstructed structure and the ribbon models for the 1xba model, respectively. The reconstruction for 1xba also has a similar shape as its original PDB model.

We also checked the interior structures by computing the voids. A void is a cavity in a molecular interior that is accessible to some molecule and is important for understanding the molecular characteristics. Fig. 5(a) shows the distribution of the interior voids of the PDB structure 1xba. The dark red color denotes the voids where a spherical probe with the radius 1.4Å (corresponding to a water molecule) can be placed. Fig. 5(b) shows the same information for the reconstructed structure. Note the similarity of the void distribution for the water molecules. Fig. 5(c) and (d) show the distribution of the voids corresponding to a probe with the radius 1.0Å. Both the original structures and the reconstructed structures are remarkably similar!

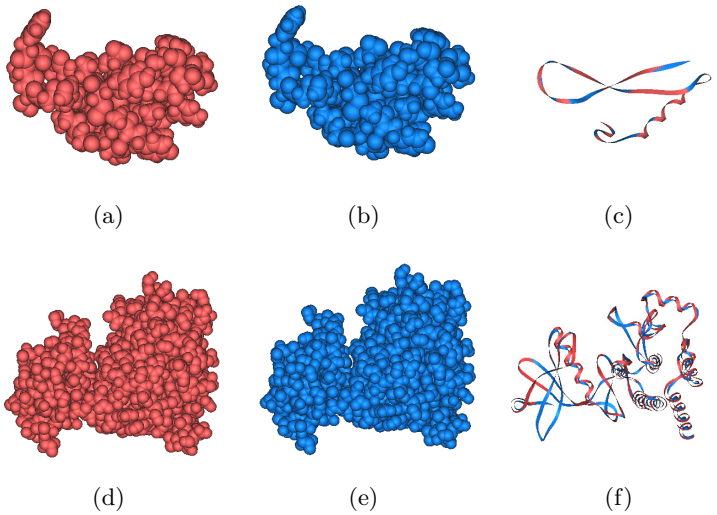


Fig. 4. Structure comparison. (a) and (d) the PDB structures 2lt8 (558 atoms) and 1xba (2068 atoms), respectively; (b) and (e) the reconstructed structures by the BetaMDGP; and (c) and (f) the ribbon models for the backbones of the true (red) and reconstructed (blue) structures after they were superposed.

For validation of the solution quality of the proposed algorithm, visual inspection is of course insufficient. We statistically checked the solution quality as well. First, we checked the root mean squared deviation (RMSD) between the PDB models and the reconstructed models as follows. Let $dist_i^{pdb2beta}$ be the distance between an atom a_i in PDB and its reconstructed atom by using the BetaMDGP algorithm. The RMSD for n atoms is given as

$$\text{RMSD} = \sqrt{\frac{1}{n}\sum_{1}^{n}(dist_i^{pdb2beta})^2}. \tag{6}$$

Table 1 shows the statistics of the RMSDs and the computation time. Column A is the PDB ID of the original PDB models used in the first test and the number of atoms is in column B. Columns C and D are the number of residues and ensembles, respectively. Note that the 1xba model determined from X-ray crystallography has no ensemble. Column E shows the statistics of the RMSDs between the original PDB models and its reconstructions. Note that the 2lt8 and 2jwu models were determined by NMR spectroscopy. Hence, we reported the average value of the RMSDs (E3) for each ensemble after the reconstructions of all the ensemble instances were separately computed by the BetaMDGP program. Similarly, we reported the minimum (E1), the maximum (E2), and the standard deviation (E4) value of the RMSDs. The average RMSDs (E3) for 2lt8, 2jwu,

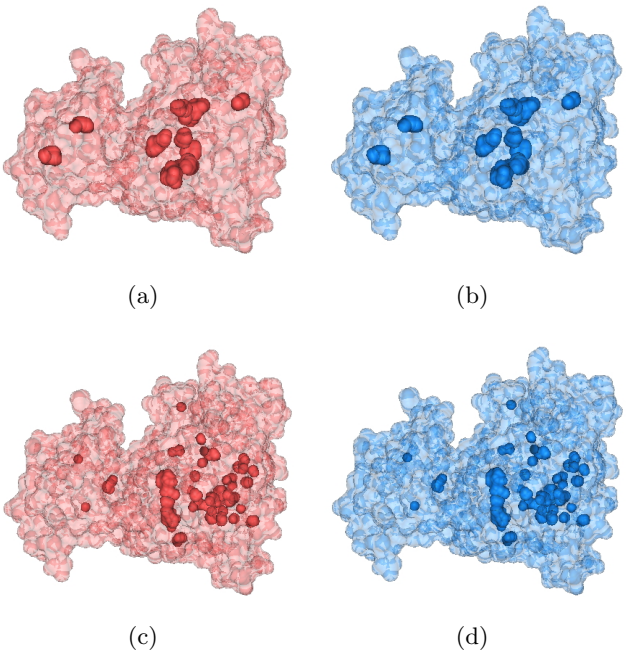


Fig. 5. Void distributions of 1xba. (a) and (c) the interior voids for the PDB structures and (b) and (d) the interior voids for the reconstructed structures ((a) and (b): $\beta = 1.4$, (c) and (d): $\beta = 1.0$).

Table 1. Statistics of the RMSDs and the computation times from the BetaMDGP

PDB ID	#atoms	#res.	#ensem.	RMSD (Å) (E)				time(sec)
				min.	max.	avg.	stdev.	
(A)	(B)	(C)	(D)	(E1)	(E2)	(E3)	(E4)	(F)
2lt8	558	43	20	0.008	0.104	0.030	0.025	9.51
2jwu	922	56	20	0.001	0.208	0.017	0.046	7.62
1xba	2068	334	·	·	·	0.041	·	104.68

and 1xba were 0.030, 0.017, and 0.041Å, respectively. They are all tiny. The computation took 9.51, 7.62, and 104.68 sec, respectively (F). It currently seems relatively high because our current implementation of the Voronoi diagram and beta-complex algorithms are not optimally tuned for the MDGP problem. We expect this problem will be remedied in our future version with an expected computation reduction of tenfold or more.

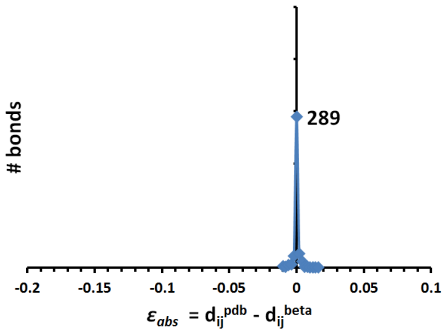
We also checked the distributions of the covalent bond lengths in both the PDB and the reconstructed structures. Let d_{ij}^{pdb} and d_{ij}^{beta} be the length between

the covalent bonded a_i and a_j in the PDB and the reconstructed model, respectively. Let $\epsilon_{abs} = d_{ij}^{pdb} - d_{ij}^{beta}$ be the absolute error. Fig. 6 shows the distribution of the absolute error ϵ_{abs} for the three examples. Note that all three graphs show that the distributions are extremely focused with a mean value of zero. We note that the volumes of the voids can also be computed and compared from a statistical point of view. From this test, we conclude that the proposed BetaMDGP algorithm reconstructs the original PDB structure effectively and efficiently.

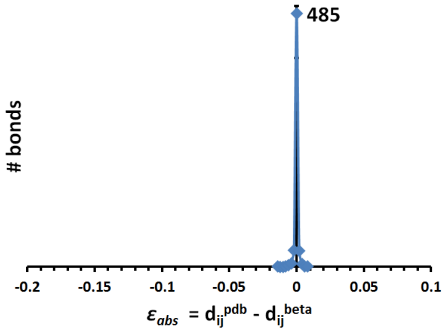
For the second test, we experimented with data consisting of the interval distances. We computed the inter-atomic distances from all the ensembles of the PDB model whose structures were determined from NMR spectroscopy. We created the lower and the upper bounds of the interval of each edge by the minimum value and the maximum value of the inter-atomic distances of the edges, respectively. Then, the medium value of interval shorter than 5Å was used as the input data for the test.

The red structures in Fig. 7(a) and (d) show the true structures of 2jmy (281 atoms, 19 models in the ensembles) and 2jwu (922 atoms, 20 models in the ensembles) in the PDB whose structures were determined from NMR spectroscopy. The red structures in Fig. 7 are one (to be specific, the model with the minimum RMSD after the superposition with the reconstruction) of the ensembles. We compared the reconstructed structure by the BetaMDGP with all the original ensemble structures. The blue structure in Fig. 7(b) shows the reconstructed structure by the BetaMDGP algorithm using the input files of the 2jmy. Fig. 7(c) shows the ribbon models for the superposed backbones of the original structure in Fig. 7(a) and the reconstructed structure in Fig. 7(b). Note that they are very close. Fig. 7(d), (e), and (f) are for the 2jwu model. Table 2 shows a summary of these experimental results. Column D denotes the number of models in the ensembles in the original PDB models. Column E shows the minimum of RMSD (E1), the maximum of RMSD (E2), the average of RMSD (E3), and the standard deviation of RMSD (E4) between each of the ensembles of the original model and the reconstructed model using the interval distance. The minimum RMSD (E1) between the reconstructed structure and 19 ensembles of 2jmy is 2.21, and the average RMSD (E3) is 2.34 Å. These RMSDs are obviously larger than the RMSD for the experiment with the data without an interval. This may be because we used the medium value of the interval distance as the input to the BetaMDGP program. From the second test, we also conclude that the BetaMDGP algorithm reconstructs the PDB structures at a fairly sufficient level of accuracy and efficiency.

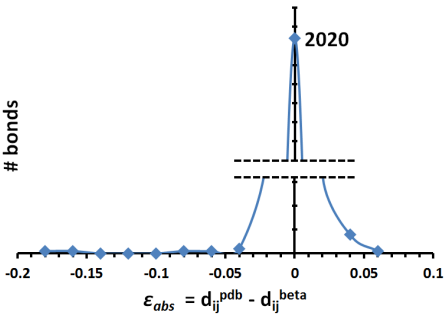
For the third test, we experimented the BetaMDGP algorithm for the input data consisting of only the hydrogen atoms with intervals for each distance-defined pair. We computed the inter-atomic distances between only the hydrogen atoms from all the ensembles of the PDB file. Then, we created the lower and the upper bounds of the interval by the minimum value and the maximum value of the inter-atomic distances, respectively. The medium value of interval shorter than 5Å is used as the input file.



(a)



(b)



(c)

Fig. 6. Difference in the covalent bond lengths between the original and the reconstructed structure. PDB ID: (a) 2lt8 (558 atoms), (b) 2jwu (922 atoms), and (c) 1xba (2068 atoms).

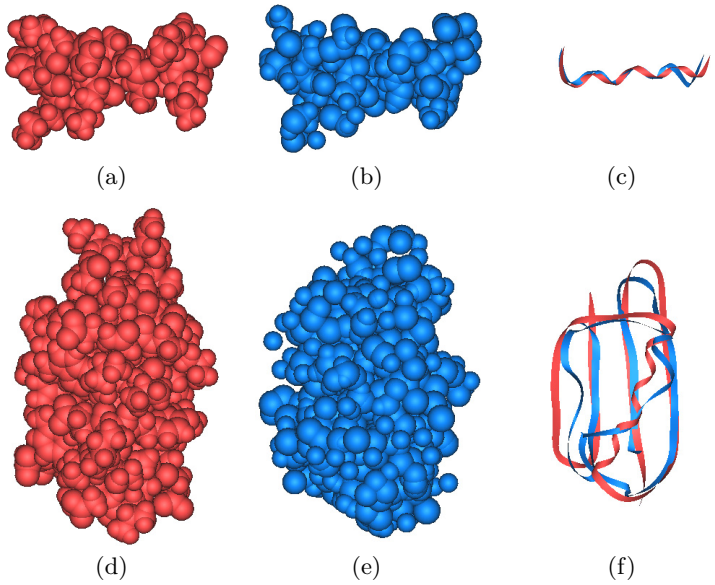


Fig. 7. Structure comparison. (a) and (d) the original protein structures 2jmy and 2jwu, respectively; (b) and (e) the reconstructed structures by the BetaMDGP; and (c) and (f) the ribbon models of the original (red) and reconstructed (blue) structures after they were superposed.

Table 2. Summary of the RMSDs and the computation times from BetaMDGP with intervals

PDB ID	#atoms	#res.	#ensem.	RMSD (Å) (E)				time(sec)
				min.	max.	avg.	stdev.	
(A)	(B)	(C)	(D)	(E1)	(E2)	(E3)	(E4)	(F)
2jmy	281	15	19	2.21	2.59	2.34	0.10	1.972
2jwu	922	56	20	4.36	4.55	4.48	0.04	11.812

Table 3 shows a summary of this experimental result. In this experiment, we used the input distance of the two types: i) the medium value of the interval distance (Row I) and ii) the random value of each interval distance (Row II). The data in Row II are the results of 500 experiments. Column B is the number of hydrogen atoms and column E shows the statistics of the RMSDs between all the ensembles of the original model and the reconstructed model. The minimum RMSD (E1) of Row II (by the random choice) is significantly smaller than the minimum RMSD of Row I (by the medium choice). This implies that we may get better reconstruction if the distribution of the distances for each atom pair can be used. In this experiment, we used the RMSD as the measure of quality for the

reconstructed models. However, we believe that this may not be an appropriate measure for the reconstructed model from the test data with intervals produced from the PDB ensemble. The development of an appropriate measure is an issue for further study.

Table 3. Statistics of the RMSDs and computation times from BetaMDGP with intervals (only hydrogen atoms)

	PDB ID	#atoms	#res.	#ensem.	RMSD (Å) (E)				time(sec)
					min.	max.	avg.	stdev.	
	(A)	(B)	(C)	(D)	(E1)	(E2)	(E3)	(E4)	(F)
I	2jmy	153	15	19	7.06	7.50	7.28	0.12	1.45
	2jwu	467	56	20	8.43	8.58	8.50	0.04	3.46
II	2jmy	153	15	19	3.69	13.41	6.85	1.08	1.33
	2jwu	467	56	20	5.64	16.37	9.42	1.48	3.21

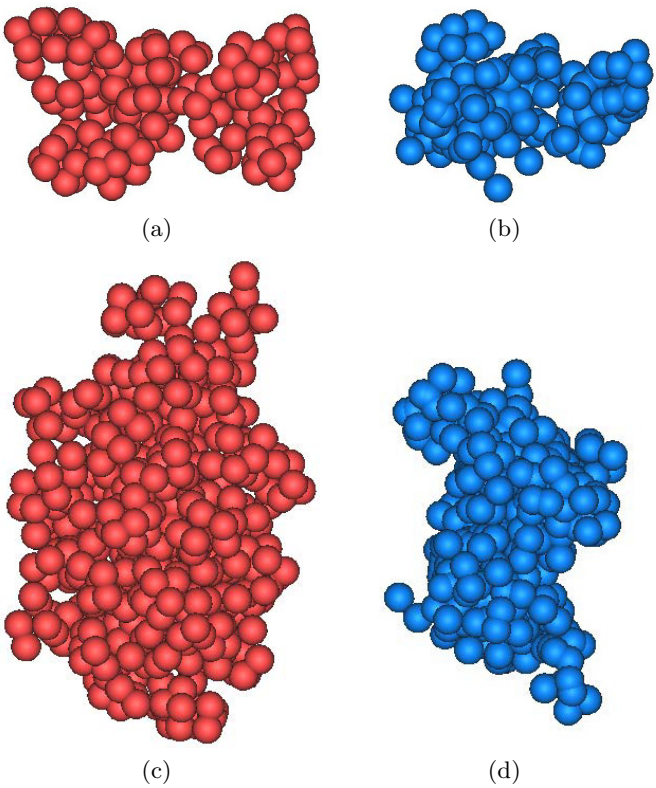


Fig. 8. Structure comparison. (a) and (c) the hydrogen atoms of the original protein structures of 2jmy and 2jwu; respectively and (b) and (d) the reconstructed structures by the BetaMDGP.

Fig. 8 shows the result of the experiments in Row II. The red model in Fig. 8(a) shows the true structures of 2jmy (153 hydrogen atoms) in the PDB. The blue model in Fig. 8(b) is the reconstructed model for the hydrogen atoms of 2jmy. The atoms in Fig. 8(b) are more closely positioned than the true structure in Fig. 8(a). Fig. 8(c) and (d) are for the 2jwu model (467 hydrogen atoms).

4 Discussions

The BetaMDGP algorithm was compared with other popular methods such as DGSOL and BP (Branch-and-prune) that we were able to benchmark. Other programs, for example, the geometric build-up algorithm were not available. First, the DGSOL is a program for solving MDGP based on the global continuation method with Gaussian smoothing of a merit function that only depends on the sparse distance data [9, 10] and can be freely downloaded from the DGSOL web site (http://www.mcs.anl.gov/more/dgsol/). In the current release, the DGSOL uses a variable-metric limited-memory code to trace the minimizers and can determine protein structures up to 200 atoms [10]. In this experiment, we also used the test data set obtained from the DGSOL site. Among the three available fragments consisting of 50, 100, and 200 atoms with a 1gpv structure (1840 atoms in total) from the PDB (The DGSOL provides only this one structure on the web site), we chose to test the biggest fragment with 200 atoms. The DGSOL determines the lower and the upper bounds of the distance intervals as follows. If $d_{i,j} = ||x_i - x_j||$ is the distance within the 5Å cutoff distance between atoms i and j, then the lower bound $l_{i,j} = d_{i,j}(1 - \varepsilon)$ and the upper bound $u_{i,j} = d_{i,j}(1 + \varepsilon)$ for some epsilon with $0 < \varepsilon < 1$. We ran both the BetaMDGP and the DGSOL using the various input data with intervals generated with different ε values.

Table 4 shows the test results of the BetaMDGP and the DGSOL. Columns A and B show the number of atoms and edges (i.e. the number of atom pairs with known distances in the input data), respectively. Column C shows the different ε values used for interval generation. Recall Eq. (5). Column D2 is what the DGSOL produced by Eq. (5) which describes how much the reconstructed structure satisfies the distance constraints with intervals. Column D1 shows the statistics using Eq. (5) from the reconstructed structure by the BetaMDGP. While the value of both the algorithms are tiny, the BetaMDGP values are smaller. Columns E1 and E2 show the RMSDs, defined by Eq. (6), from both BetaMDGP and DGSOL, respectively. Note that the RMSD of the BetaMDGP is significantly smaller than that of the DGSOL. The computation times in columns F1 and F2 show that the computation times from the BetaMDGP are significantly faster than the DGSOL.

Fig. 9 visually illustrates the experimental result of the case $\varepsilon = 0.16$ in Table 4. The red model in Fig. 9(a) is the segment of the original PDB model 1gpv that we extracted. It corresponds to the segment consisting of 200 atoms in the input file defined by the DGSOL website. Hence, this is the target structure that we want to reconstruct. The blue one in Fig. 9(b) and the green one in Fig. 9(c) are the reconstructions by the BetaMDGP and the DGSOL, respectively. Observe that the reconstruction by the BetaMDGP is very close to the original

Table 4. Comparison of the BetaMDGP and the DGSOL. The fragment used for the test has 200 atoms from the PDB structure 1gpv (1840 atoms in total). The parameter ε is used for the interval generation.

PDB ID: 1gpv			$p_{i,j}(x)$ (D)		RMSD (Å) (E)		time (sec) (F)	
#atoms (A)	#edges (B)	ε (C)	BetaMDGP (D1)	DGSOL (D2)	BetaMDGP (E1)	DGSOL (E2)	BetaMDGP (F1)	DGSOL (F2)
200	3300	0.04	0.001	0.007	0.004	5.395	1.273	22.256
200	3300	0.08	0.000	0.097	0.039	2.528	1.304	23.510
200	3300	0.12	0.000	0.000	0.085	5.055	1.394	22.097
200	3300	0.16	0.000	0.000	0.013	2.470	1.381	25.079

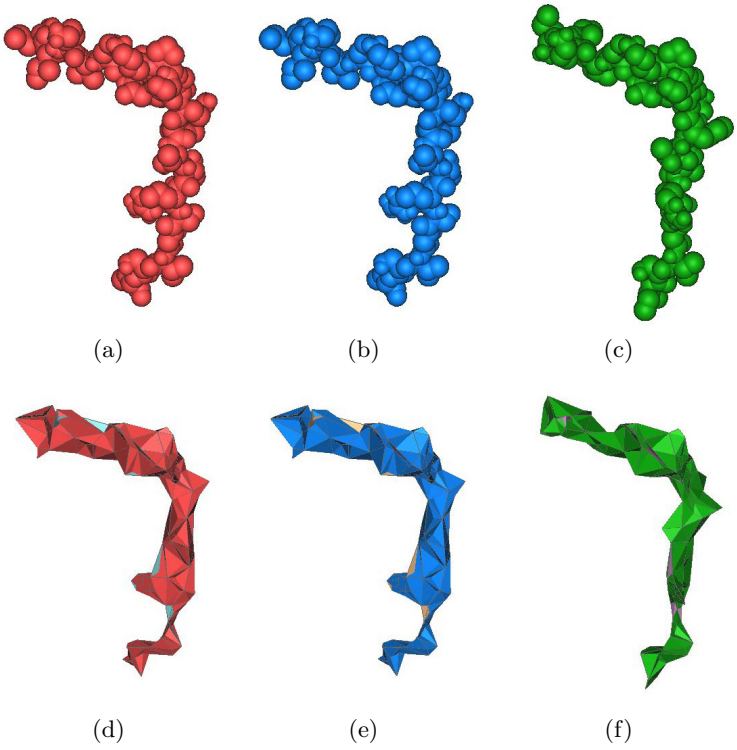


Fig. 9. Comparison of the structures reconstructed from the BetaMDGP and DGSOL against the original structure from PDB (1gpv). (a) the original protein structures (PDB code:fragment of 1gpv); (b) and (c) the reconstructed structures by the BetaMDGP and DGSOL, respectively; and (d), (e), and (f) the corresponding beta-shapes (β=1.4Å).

PDB model, whereas the one by the DGSOL is significantly different from the original model (The atom in the vertical column of the original PDB model does not exist). Fig. 9(d), (e), and (f) are the beta-shapes of the structures in Fig. 9(a), (b), and (c) where $\beta = 1.4$ Å respectively. The beta-shapes clearly show the reconstruction quality.

Second, the BetaMDGP was also compared with the MD-jeep program which implemented the Branch and Prune (BP) algorithm [25]. The MD-jeep (version 0.1) and test problems can be freely downloaded from http://www.antoniomucherino.it/en/mdjeep.php. The MD-jeep reconstructs the backbone structures through the formulation of a combinatorial optimization problem and uses a branch and prune strategy. To be compatible with the input data to the MD-jeep, we also generated an input file to the BetaMDGP from the same PDB files according to the rules used for the MD-jeep as follows. Given a PDB file, we first extracted N, C_α, and C atoms with their coordinates on a backbone. Then, the pairwise distances falling within 5 Å were computed in order to simulate NMR data as an input file to the BetaMDGP algorithm. We verified the identity of the MD-jeep input files downloaded from its web site and the generated BetaMDGP input files using the number of atoms and the interatomic distances. Running both the BetaMDGP and the MD-jeep algorithms produces their reconstructions which obviously contain only N, C_α, and C atoms, missing O and C_β (i.e., the first atom on each side-chain). Among the possible solutions found by MD-jeep, we used a solution that MD-jeep provided. We remark that the solution quality is likely to be improved if all MD-jeep solutions are used. Fig. 10(a), (b), and (c) visually show the ribbon models of the backbone of the original PDB structure (red), that of the BetaMDGP reconstruction (blue), and that of the MD-jeep reconstruction (green), respectively. Each reconstructed structure is displayed after it is superposed with the original one from the PDB file. Note that the BetaMDGP reconstructions are closer to the original models.

Given a backbone structure with known amino acid sequence information, it is possible to recover the entire protein structure by solving the side-chain prediction problem, abbreviated as the SCP-problem, which predicts the optimal conformation of the side-chains of all the amino acids in a protein. The general approach to the SCP-problem is to use a rotamer library which is derived by statistically clustering the observed side-chain conformations of known protein structures in the PDB [39–42]. The optimality is defined by the minimum potential energy of the protein structure determined by the conformation of all the side-chains where the energy is given by a forcefield. The SCP-problem is known as NP-hard [43–45] and is useful for flexible protein-ligand docking [46, 47] and homology modeling [48–50].

We generated the two types of missing atoms, O and C_β, with their coordinates. Then, we ran the BetaSCP program, also developed by the authors group [51], to get the entire protein structure of the backbones produced by both the BetaMDGP and the MD-jeep. Fig. 11(a) shows the structure of the original PDB files (red). Fig. 11(b) and (c) show the structures recovered by running the

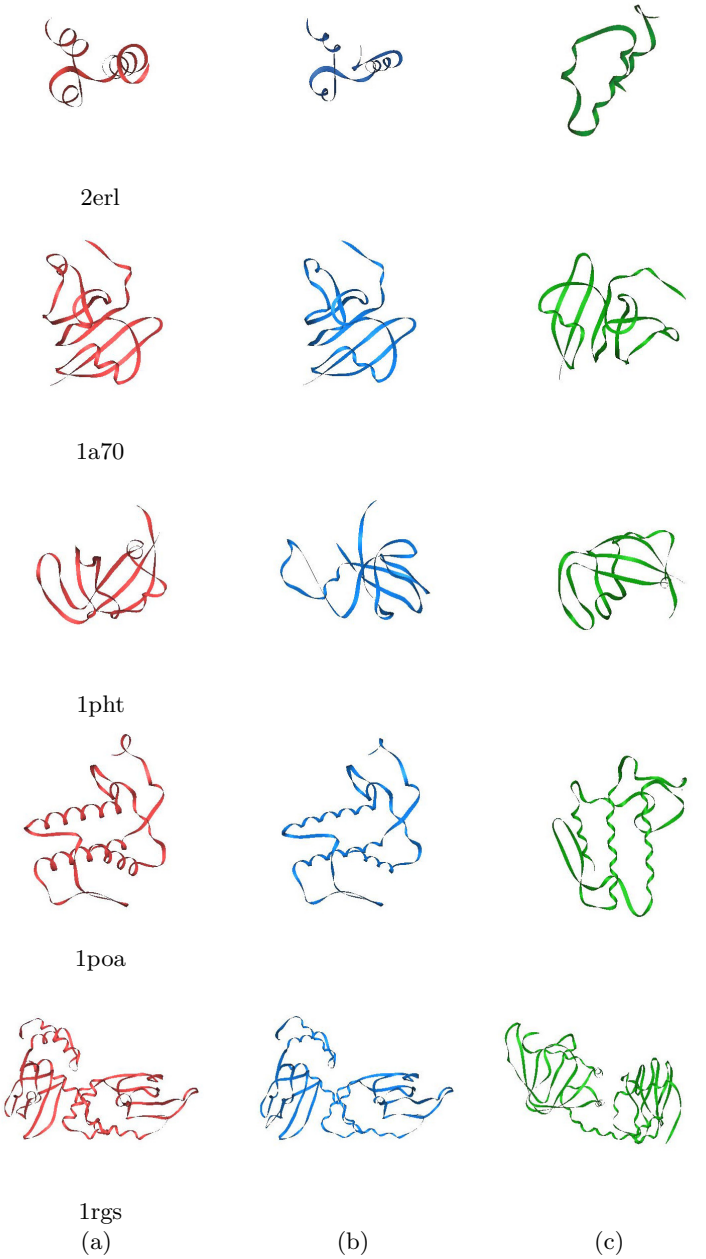


Fig. 10. Structure comparison with the ribbon models. (a) backbone of the original protein structures and (b) and (c) the reconstructed backbone structures by the BetaMDGP and the MD-jeep, respectively.

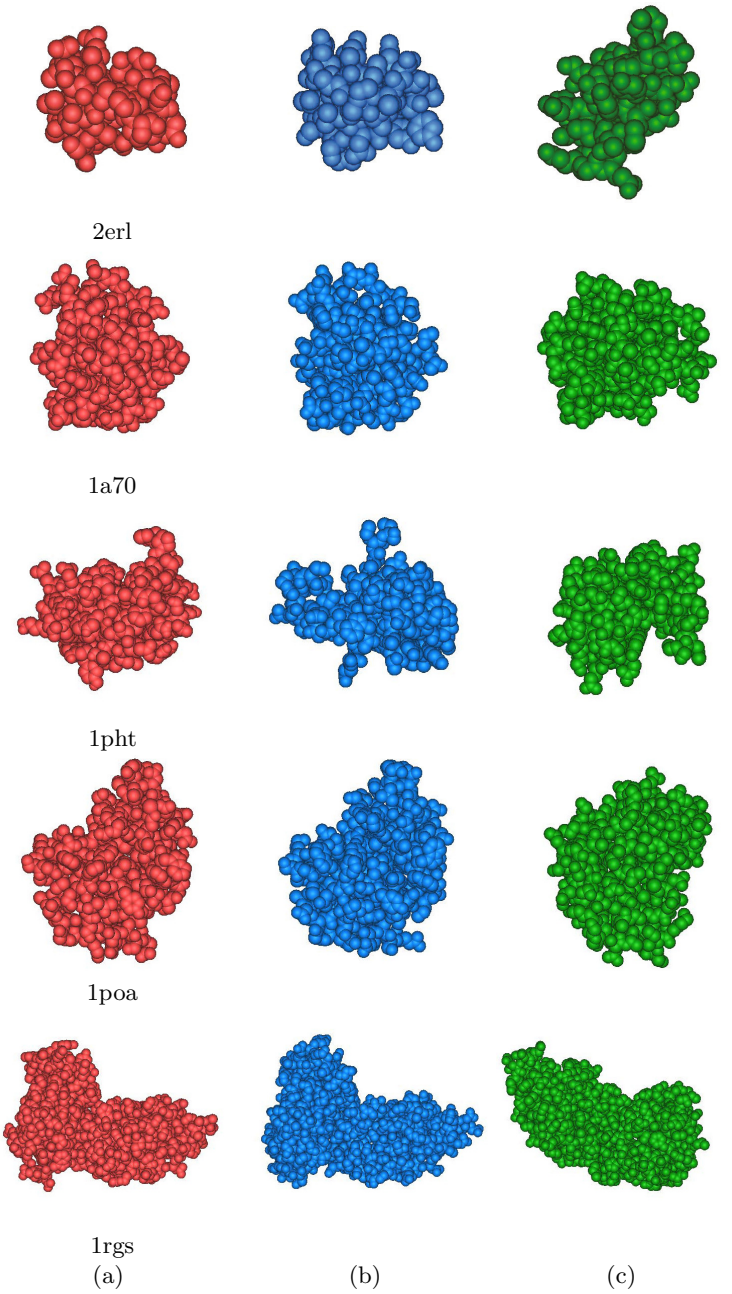


Fig. 11. Structure comparison. (a) the original protein structures and (b) and (c) the reconstructed structures by the BetaSCP program on the backbones from the BetaMDGP and the MD-jeep, respectively.

BetaSCP program on the backbones from the BetaMDGP (blue) and the MD-jeep (green), respectively. Observe that the result of the BetaMDGP is significantly better than that of the MD-jeep.

The visual result in Fig. 11 is statistically analyzed in Table 5. Row I corresponds to the BetaMDGP and Row II corresponds to the MD-jeep. Column B1 is the number of atoms of the original models and column B2 is the number of atoms in the backbones, both from the original PDB structures. Column D1 is the RMSD between the original backbone structure and the structure reconstructed by the BetaMDGP algorithm. Column D3 is the MD-jeep counterpart for column D1. Note that the backbone structures produced by the BetaMDGP are mostly better than those by the MD-jeep (with the exception 1pht). Column D2 is the RMSD between the entire original PDB structure and the reconstructed structure with the recovered side-chains. Column D4 is the MD-jeep counterpart for column D2. Note that the BetaMDGP solutions are mostly better than those of the MD-jeep. Columns E1 and E4 are the computation times for the reconstruction by the BetaMDGP and the MD-jeep, respectively. Columns E2 and E5 are the computation times for running the BetaSCP program. Columns E3 and E6 are the total computation times for solving the SCP problem after the BetaMDGP and the MD-jeep, respectively.

Table 5. Experimental statistics of the protein structures whose side-chains are recovered by the BetaSCP program on the backbones reconstructed by the BetaMDGP and MD-jeep (Row I: the BetaMDGP result; Row II: the MD-jeep result)

	PDB ID (A)	#atoms (B)		#residues (C)	RMSD (Å) (D)		time (sec) (E)		
		PDB (B1)	Backbone (B2)		BetaMDGP (D1)	Recon Entire Struct (D2)	BetaMDGP (E1)	BetaSCP (E2)	E1+E2 (E3)
I	2erl	566	120	40	0.75	0.88	4.81	1.22	6.03
	1a70	732	291	97	0.54	1.48	7.81	4.41	12.22
	1pht	810	249	83	2.16	2.95	6.14	5.16	11.30
	1poa	913	354	118	0.27	1.12	24.02	5.94	29.96
	1rgs	2015	792	264	0.67	1.98	35.46	21.34	56.80
					MD-jeep (D3)	Recon Entire Struct (D4)	MD-jeep (E4)	BetaSCP (E5)	E1+E2 (E6)
II	2erl	566	120	40	1.78	2.62	0.00	1.15	1.15
	1a70	732	291	97	2.10	3.01	0.01	4.29	4.30
	1pht	810	249	83	2.11	2.51	0.01	4.85	4.86
	1poa	913	354	118	2.22	2.83	0.01	5.64	5.65
	1rgs	2015	792	264	3.47	7.28	0.08	19.40	19.48

5 Conclusions

We proposed a new approach, the BetaMDGP, to the MDGP problem based on the beta-complex, which is a geometric construct derived from the Voronoi di-

agram of atoms. From experiments using simulated NMR files, the BetaMDGP reconstructs the original structures with surprising similarity except for the input data with interval distances. However, there are three main issues to be resolved in the future. First, the BetaMDGP algorithm needs to consider the interval distances as the current algorithm considers only the medium value of an interval. Second, we need to improve the BetaMDGP algorithm by considering the under-determined condition. The real NMR data may be more sparse than our test data. These NMR data cause a situation where the molecular structure cannot be determined by using only the input data. Therefore, we have to consider additional information to solve the under-determined condition. For example, we can consider additional information such as the chemistry information and produce the input distance from the under-determined atom using triangular inequality. Finally, we remark that the BetaMDGP algorithm needs improved computational efficiency and the convergence of the BetaMDGP algorithm with an optimization method such as the BP algorithm is likely to improve solution quality.

Acknowledgments. This work was supported by a National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No.2011-0020410). C. Lavor and A. Mucherino were supported by the Brazilian research agencies FAPESP and CNPq. The authors thank Prof. G. Crippen for suggesting an important direction of this study with real NMR files and Dr. S.-W. Chi at KRIBB for the help with the NMR file format.

References

1. Donald, B.R.: Algorithms in Structural Molecular Biology. The MIT Press (2011)
2. Cavanagh, J., Fairbrother, W.J., Palmer III, A.G., Rance, M., Skelton, N.J.: Protein NMR spectroscopy: principles and practice. Academic Press (2006)
3. Jan, D.: Principals of protein X-ray crystallography. Springer (2006)
4. Blumenthal, L.M.: Theory and Applications of Distance Geometry. Oxford Clarendon Press (1953)
5. Crippen, G., Havel, T.: Distance Geometry and Molecular Conformation. John Wiley & Sons, New York (1988)
6. Liberti, L., Lavor, C., Maculan, N., Mucherino, A.: Euclidean distance geometry and applications. SIAM Review 56 (article in press, 2014)
7. Havel, T.F.: Distance Geometry, vol. 4. John Wiley & Sons (1995)
8. Saxe, J.: Embeddability of weighted graphs in k-space is strongly np-hard. In: Proceedings of 17th Allerton Conference in Communications Control and Computing, pp. 480–489 (1979)
9. Moré, J.J., Wu, Z.: Global continuation for distance geometry problems. SIAM Journal of Optimization 7, 814–836 (1997)
10. Moré, J.J., Wu, Z.: Distance geometry optimization for protein strucutures. Journal of Global Optimization 15(3), 219–234 (1999)
11. An, L.T.H.: Solving large scale molecular distance geometry problems by a smoothing technique via the gaussian transform and d.c. programming. Journal of Global Optimization 27, 375–397 (2003)
12. An, L.T.H., Tao, P.D.: Large-scale molecular optimization from distance matrices by a d.c. optimization approach. SIAM Journal of Optimization 14(1), 77–114 (2003)

13. Liberti, L., Lavor, C., Mucherino, A., Maculan, N.: Molecular distance geometry methods: from continuous to discrete. International Transactions in Operational Research 18, 33–51 (2010)
14. Wüthrich, K.: NMR in Structural Biology. World Scientific, New York (1995)
15. Havel, T.: An evaluation of computational strategies for use in the determination of protein structure from distance constraints obtained by nuclear magnetic resonance. Progress in Biophysics and Molecular Biology 56(1), 43–78 (1991)
16. Hendrickson, B.: The Molecular Problem: Determining Conformation from Pairwise Distances. PhD thesis, Cornell University (1991)
17. Hendrickson, B.: The molecule problem: Exploiting structure in global optimization. SIAM Journal of Optimization 5, 835–857 (1995)
18. Dong, Q., Wu, Z.: A geometric build-up algorithm for solving the molecular distance geometry problem with sparse distance data. Journal of Global Optimization 26(3), 321–333 (2003)
19. Wu, D., Wu, Z.: An updated geometric build-up algorithm for solving the molecular distance geometry problems with sparse distance data. Journal of Global Optimization 37(4), 661–673 (2007)
20. Sit, A., Wu, Z., Yuan, Y.: A geometric buildup algorithm for the solution of the distance geometry problem using least-squares approximation. Bulletin of Mathematical Biology 71(8), 1914–1933 (2009)
21. Sit, A., Wu, Z.: Solving a generalized distance geometry problem for protein structure determination. Bulletin of Mathematical Biology, 1–28 (2011)
22. Brünger, A.T., Adams, P.D., Clore, G.M., DeLano, W.L., Gros, P., Grosse-Kunstleve, R., Jiang, J.S., Kuszewski, J., Nilges, M., Pannu, N.S., Read, R.J., Rice, L.M., Simonson, T., Warren, G.L.: Crystallography & nmr system: A new software suite for macromolecular structure determination. Acta Crystallographica Section D-Biological Crystallography D54, 905–921 (1998)
23. Schwieters, C.D., Kuszewski, J.J., Tjandra, N., Clore, G.M.: The xplor-nih nmr molecular structure determination package. Journal of Magnetic Resonance 160, 65–73 (2003)
24. Lavor, C., Liberti, L., Maculan, N., Mucherino, A.: The discretizable molecular distance geometry problem. Computational Optimization and Applications 52, 115–146 (2012)
25. Liberti, L., Lavor, C., Maculan, N.: A branch-and-prune algorithm for the molecular distance geometry problem. International Transactions in Operational Research 15, 1–17 (2008)
26. Kim, D.S., Cho, Y., Sugihara, K., Ryu, J., Kim, D.: Three-dimensional beta-shapes and beta-complexes via quasi-triangulation. Computer-Aided Design 42(10), 911–929 (2010)
27. Kim, D.S., Kim, J.K., Cho, Y., Kim, C.M.: Querying simplexes in quasi-triangulation. Computer-Aided Design 44(2), 85–98 (2012)
28. Kim, D.S., Seo, J., Kim, D., Ryu, J., Cho, C.H.: Three-dimensional beta shapes. Computer-Aided Design 38(11), 1179–1191 (2006)
29. Cho, Y., Kim, J.K., Ryu, J., Won, C.I., Kim, C.M., Kim, D., Kim, D.S.: BetaMol: a molecular modeling, analysis and visualization software based on the beta-complex and the quasi-triangulation. Journal of Advanced Mechanical Design, Systems, and Manufacturing 6(3), 389–403 (2012)
30. Cho, Y., Kim, D., Kim, D.S.: Topology representation for the Voronoi diagram of 3D spheres. International Journal of CAD/CAM 5(1), 59–68 (2005), `http://www.ijcc.org`
31. Kim, D.S., Cho, Y., Kim, D.: Euclidean Voronoi diagram of 3D balls and its computation via tracing edges. Computer-Aided Design 37(13), 1412–1424 (2005)

32. Okabe, A., Boots, B., Sugihara, K., Chiu, S.N.: Spatial Tessellations: Concepts and Applications of Voronoi Diagrams, 2nd edn. John Wiley & Sons, Chichester (1999)
33. Munkres, J.R.: Elements of Algebraic Topology. Perseus Press (1984)
34. Boissonnat, J.D., Yvinec, M.: Algorithmic Geometry. Cambridge University Press, Cambridge (1998)
35. Kim, D.S., Kim, D., Cho, Y., Sugihara, K.: Quasi-triangulation and interworld data structure in three dimensions. Computer-Aided Design 38(7), 808–819 (2006)
36. Kim, D.S., Cho, Y., Sugihara, K.: Quasi-worlds and quasi-operators on quasi-triangulations. Computer-Aided Design 42(10), 874–888 (2010)
37. Kim, D.S., Cho, Y., Ryu, J., Kim, J.K., Kim, D.: Anomalies in quasi-triangulations and beta-complexes of spherical atoms in molecules. Computer-Aided Design 45(1), 35–52 (2013)
38. Kim, J.K., Kim, D.S.: Betasuperposer: Superposition of protein surfaces using beta-shapes. Journal of Biomolecular Structure & Dynamics 30(6), 684–700 (2012)
39. Dunbrack Jr., R.L.: Rotamer libraries in the 21st century. Current Opinion in Structural Biology 12(4), 431–440 (2002)
40. Dunbrack Jr., R.L., Karplus, M.: Backbone-dependent rotamer library for proteins. Journal of Molecular Biology 230(2), 543–574 (1993)
41. Dunbrack Jr., R.L., Karplus, M.: Conformational analysis of the backbone-dependent rotamer preferences of protein sidechains. Journal of Molecular Biology 1(5), 334–340 (1994)
42. Kono, H.: Rotamer libraries for molecular modeling and design of proteins. In: Park, S.J., Cochran, J.R. (eds.) Protein Engineering and Design (2009)
43. Chazelle, B., Kingsford, C., Singh, M.: The inapproximability of side-chain positioning. Technical report, Princeton University (2004)
44. Fung, H., Rao, S., Floudas, C., Prokopyev, O., Pardalos, P., Rendl, F.: Computational comparison studies of quadratic assignment like formulations for the In silico sequence selection problem in De Novo protein design. Journal of Combinatorial Optimization 10(1), 41–60 (2005)
45. Pierce, N.A., Winfree, E.: Protein design is NP-hard. Protein Engineering 15(10), 779–782 (2002)
46. Althaus, E., Kohlbacher, O., Lenhof, H.P., Müller, P.: A combinatorial approach to protein docking with flexible side-chains. In: RECOMB 2000 Proceedings of the Fourth Annual International Conference on Computational Molecular Biology, pp. 15–24 (2000)
47. Althaus, E., Kohlbacher, O., Lenhof, H.P., Müller, P.: A combinatorial approach to protein docking with flexible side chains. Journal of Computational Biology 9(4), 597–612 (2002)
48. Lee, C., Subbiah, S.: Prediction of protein side-chain conformation by packing optimization. Journal of Molecular Biology 217(2), 373–388 (1991)
49. Tuffery, P., Etchebest, C., Hazout, S., Lavery, R.: A new approach to the rapid determination of protein side chain conformations. Journal of Biomolecular Structure & Dynamics 8(6), 1267–1289 (1991)
50. Leach, A.R.: Molecular Modelling: Principles and Applications. Prentice Hall (2001)
51. Ryu, J., Kim, D.S.: Protein structure optimization by side-chain positioning via beta-complex. Journal of Global Optimization (2012), doi: 10.1007/s10898-012-9886-3