A Gaussian Packing Model For Phasing in Macromolecular Crystallography

Yan Yan[1] **and Gregory S. Chirikjian**[1]
[1]The Department of Mechanical Engineering, Johns Hopkins University, Baltimore, MD, USA

Abstract— *Molecular replacement (MR) is a computational method that is frequently used to obtain phase information for a unit cell packed with a macromolecule of unknown structure. The goal of MR searches is to place a homologous/similar molecule in the unit cell so as to maximize the correlation with x-ray diffraction data. MR software packages typically perform rotation and translation searches separately. This works quite well for single-domain proteins. However, for multi-domain structures and complexes, computational requirements can become prohibitive and the desired peaks can become hidden in a noisy landscape. The main contribution of our approach is that computationally expensive MR searches in continuous configuration space are replaced by a search on a relatively small discrete set of candidate packing arrangements of a multi-rigid-body model. These candidate arrangements are generated by minimizing a Gaussian-based potential function that forces the model conformations to separate from each other and not overlap within the unit cell. This is done before computing Patterson correlations rather than only performing collision checks when evaluating the feasibility of peaks. The list of feasible arrangements is short because collision-free packing requirement together with unit cell symmetry and geometry impose strong constraints. After computing Patterson correlations of the candidate arrangements, an even shorter list can be obtained using 10 candidates with highest correlations. In numerical trials, we found that a candidate from the feasible set is usually similar to the arrangement of the target structure within the unit cell. To further improve the accuracy, a Rapidly-exploring Random Tree (RRT) can be applied in the neighborhood of this packing arrangement. Our approach is demonstrated with multi-domain models in silico for 2D, with ellipses (ellipsoids in 2D) representing both the domains of the model and target structures. Configurations are defined by sets of angles between the ellipses. Our results show that an approximate configuration can be found with the mean absolute error less than 3 degrees.*

Keywords: X-ray crystallography, molecular replacement, multi-domain system, packing model, Gaussian function

1. Introduction

The field of structural biology is concerned with characterizing the shape, composition, flexibility, and motion of biological macromolecules and the complexes that they form. An ultimate goal of this field is to link these properties with the function of macromolecular structures, in the hope of better understanding biological phenomena and designing new drugs.

Here we review some of the issues involved in translating experimental data into 3D structures in the context of protein crystallography. Macromolecular X-ray crystallography (MX) has been the most used method for determining protein structures and associated complexes. It works very well for simple proteins that can be described as single rigid-bodies (called domains). This is because information about the shape of 75,000 previously solved structures in the Protein Data Bank (many of which are single-domain structures) can be used to augment new MX experimental information to gain a complete picture.

However, a challenge to MX arises in interpreting X-ray diffraction patterns for crystals composed of multi-domain systems. This is because even when a multi-domain structure has been solved previously, its overall shape may vary widely from a new version of the structure with, for example, a bound drug. In this case, a widely used computational method called the molecular replacement method (MR), which has been highly successful for single-domain proteins, becomes combinatorially intractable due to the large number of degrees of freedom in multi-domain systems. We present a new method for phasing based on geometric packing that can serve as an alternative to MR. Decades ago, the concept of building models of crystallographic unit cells to phase crystallographic data was explored in the context of small molecules [1], [2], [3]. But to our knowledge, this approach has not been pursued and is virtually unknown in the context of multi-domain macromolecular crystallography, and “phasing by packing” therefore represents a very different way of approaching the problem than MR.

The remainder of this paper is structured as follows. The mathematical aspects of the MR method for single-domain proteins is reviewed first. Then the multi-domain phase problem is formulated. Finally, we present our initial findings that diffraction patterns for multi-domain systems can be phased using our new “phasing by packing” method.

2. Essentials of Macromolecular X-Ray Crystallography (MX)

A biological macromolecule is a large collection of atomic nuclei that are stabilized through a combination of covalent bonds, hydrogen bonds, and hydrophobicity. A traditional goal in structural biology is to obtain the Cartesian coordinates of all atoms in a rigid single-domain protein.

Let $\mathbf{x}_i = (x_i, y_i, z_i)$ denote the Cartesian coordinates of the i^{th} of n atoms in a single-domain protein structure, and let $\rho_i(\mathbf{x})$ be the electron density of that atom in a reference frame centered on it. Due to thermal motions, the electron density of each of these atomic nuclei can be treated as a Gaussian distribution. The density of the whole structure is then of the form

$$f(\mathbf{x}) = \sum_{i=1}^{n} \rho_i(\mathbf{x} - \mathbf{x}_i). \tag{1}$$

The coordinates $\{\mathbf{x}_i\}$ are typically given either in a reference frame attached to a crystallographic unit cell, or to the center of mass of the protein.

MX does not provide $f(\mathbf{x})$ directly. Rather, it provides partial information about $f(\mathbf{x})$. The goal is then to computationally obtain $f(\mathbf{x})$ and fit an atomic model to it, thereby extracting the coordinates $\{\mathbf{x}_i\}$. A macromolecular crystal is composed of *unit cells* that have a discrete symmetry group, Γ. This symmetry group divides $\mathbb{R}^3$ into unit cells, $U \cong \Gamma\backslash\mathbb{R}^3$, and also describes how copies of the density $f(\mathbf{x})$ are located within the unit cell. The whole group Γ can be generated by translating unit cells and moving within the unit cell using generators $\{\gamma_1, \ldots, \gamma_m\}$. These form a subgroup of Γ, which is in turn a subgroup of the group of rigid-body motions, $SE(3)$, which will be denoted here as G.

The result of an MX experiment is a diffraction pattern. This is the magnitude of the Fourier transform of the full contents of the crystallographic unit cell. Mathematically, this is written for a single-domain protein as

$$\hat{P}(g;\mathbf{k}) = \left| \mathscr{F}\left(\sum_{j=0}^{m-1} f((\gamma_j \circ g)^{-1} \cdot \mathbf{x} \right) \right|, \tag{2}$$

where $|\cdot|$ denotes the modulus of a complex number, $c = a + ib = |c|e^{i\phi}$. Our reason for using the notation $\hat{P}(g;\mathbf{k})$ will be explained shortly. Here $g \in G$ is the unknown pose of the protein that is sought, and $\circ$ is the group operation for both G and Γ. In particular, it is well-known in robotics that each rigid-body motion consists of a rotation-translation pair $g = (R, \mathbf{t})$, and the composition of any two rigid-body motions g_1 and g_2 defines the operation $\circ$:

$$g_1 \circ g_2 = (R_1, \mathbf{t}_1) \circ (R_2, \mathbf{t}_2) = (R_1 R_2, R_1 \mathbf{t}_2 + \mathbf{t}_1). \tag{3}$$

Given that $g = (R, \mathbf{t}) \in G$ is a rotation-translation pair, its action on $\mathbb{R}^3$ is defined by

$$g \cdot \mathbf{x} = R\mathbf{x} + \mathbf{t}. \tag{4}$$

Then the density of a collection of single-domain proteins in the unit cell for $j = 0, \ldots, m-1$ will be $\sum_{i=0}^{m-1} f((\gamma_i \circ g)^{-1} \cdot \mathbf{x})$.

The difficulty in extracting $f(\mathbf{x})$ from the MX data is that this measurement folds in both information about $f(\mathbf{x})$ and the symmetry group Γ, and kills the phase information, $\phi(\mathbf{k})$, without which $f(\mathbf{x})$ cannot be recovered by inverse Fourier transform. Moreover, there is an unknown $g \in G$ that describes how each symmetry-related copy of $f(\mathbf{x})$ sits in the unit cell. Single-domain MR is mostly about finding the unknown g, and most commonly this is done by dividing the search into rotational and translational parts.

The number of proteins in a unit cell, the crystallographic space group, Γ, and aspect ratios of the unit cell can be taken as known inputs in MR computations, since they are all provided by experimental observation. And from homology modeling, it is often possible to have reliable estimates of the shape of each domain in a multi-domain protein. What remains unknown are the relative positions and orientations of theses domains and the overall position and orientation of the symmetry-related copies of the proteins within the unit cell.

Once these are known, a model of the unit cell can be constructed and used as an initial phasing model that can be combined with the X-ray diffraction data. This is, in essence, the molecular replacement approach that is now more than half a century old [4], [5]. Many powerful software packages for molecular replacement include those described in [6], [7]. Typically these perform rotation searches first, followed by translation searches.

3. The Multi-Domain Molecular Replacement Method (NMR)

The molecular replacement (MR) method, originally developed in the 1960s [4], [10], [11], [12] is a computational method for phasing X-ray diffraction data for biomolecular structures. It has been integrated into crystallographic structure determination codes [6], [14]. Though MR has been wildly successful for single-domain proteins, significant issues arise when using MR for multi-domain proteins and complexes.

Currently two major computational paradigms exist for phasing of X-ray diffraction patterns of multi-domain proteins: (1) use existing software packages to obtain candidate peaks in the rotation function for individual domains separately, then solve for the translation function [13]; (2) attempt to morph multi-domain candidate models that contain their full “6N” degrees of freedom and iteratively refine those models [8]. Both methods suffer from different aspects of the “curse of dimensionality,” which we seek to circumvent using a combination of our initial results reported in [9] and new approaches based on advanced mathematical concepts that are new to the crystallography community.

Consider a multi-domain protein or complex consisting of N rigid bodies. If $f_i(\mathbf{x})$ denotes the density of the i^{th} body, then the density of the whole complex will be of the form $f(\mathbf{x}) = \sum_{i=1}^{N} f_i(g_i^{-1} \cdot \mathbf{x})$ where $g_i = (R_i, \mathbf{t}_i)$ is a rigid-body motion consisting of a rotation-translation pair and $g_i^{-1} \cdot \mathbf{x} = R_i^T(\mathbf{x} - \mathbf{t}_i)$. These motions are the unknowns in our problem.

If m identical copies of this complex are arranged symmetrically in a unit cell by symmetry operators $\gamma_j = (Q_j, \mathbf{a}_j) \in \Gamma$ (which is the group consisting of n discrete rigid-body motions that are known a priori from the crystal symmetry and geometry), an X-ray diffraction experiment provides the magnitude (without phase) of the Fourier transform of $\sum_{j=0}^{m-1} f(\gamma_j^{-1} \cdot \mathbf{x})$. In contrast, the model density for a single domain and its symmetry mates is $\sum_{j=0}^{m-1} f_i(h_i^{-1} \circ \gamma_j^{-1} \cdot \mathbf{x})$ where h_i is the candidate position and orientation. In traditional MR, the Fourier transform of the Patterson functions, $\hat{P}(g_1, ..., g_N; \mathbf{k}) = \mathscr{F}[P(g_1, ..., g_N; \mathbf{x})]$ and $\hat{p}_i(h_i; \mathbf{k}) = \mathscr{F}[p_i(h_i; \mathbf{x})]$, that correspond to these densities and their correlation are respectively

$$\hat{P}(g_1, ..., g_N; \mathbf{k}) = \left| \sum_{j=0}^{m-1} \mathscr{F}[f(\gamma_j^{-1} \cdot \mathbf{x})] \right|, \tag{5}$$

$$\hat{p}_i(h_i; \mathbf{k}) = \left| \sum_{j=0}^{m-1} \mathscr{F}[f_i(h_i^{-1} \circ \gamma_j^{-1} \cdot \mathbf{x})] \right|, \tag{6}$$

$$c(h_i) = \int_{\mathbf{x} \in \mathscr{C}} P(g_1, ..., g_N; \mathbf{x}) p_i(h_i; \mathbf{x}) d\mathbf{x} \tag{7}$$

where the Fourier transform $\mathscr{F}$ converts a function of spatial position, $\mathbf{x}$, into a function of spatial frequency, $\mathbf{k}$. The real-space Pattersons themselves are obtained by applying the inverse Fourier transform. Of the quantities in (5)-(7), $\hat{P}(g_1, ..., g_N; \mathbf{k})$ comes from the experiment (this is the multi-domain version of (2)), and $\hat{p}_i(h_i; \mathbf{k})$ and $c(h_i)$ are computed. Here $\mathscr{C}$ is the unit cell and in MR searches the hope is that peaks in the function $c(\cdot)$ correspond to $h_i = g_i$. The difficulty is that, unlike the single domain case, in the multi-domain case P depends on many g_j's that all interact with each other. Therefore, peaks in this rotational correlation function do not necessarily correspond to good overall matches.

4. PHASING BY PACKING

Instead of running traditional MR searches on domain orientation or full conformation, we propose to construct packing models for the multi-domain systems of interest. This will generate candidate sets of motions $\{h_1, ..., h_N\}$ that can then be used to construct a *model* of $P(h_1, ..., h_N; \mathbf{x})$ *rather than* $p_i(h_i; \mathbf{x})$. If $P(h_1, ..., h_N; \mathbf{x})$ and $P(g_1, ..., g_N; \mathbf{x})$ match well to each other, then that is a much stronger indication that $h_i = g_i$ than having high correlations between $p_i(h_i; \mathbf{x})$ and $P(g_1, ..., g_N; \mathbf{x})$.

In this approach, an ellipsoid or a combination of several ellipsoids are used to approximate the convex hull of each domain of protein structures. A multi-ellipsoid-shaped model is built for a multi-domain structure and packed in space to detect feasible packing arrangements. The most important crystal packing constraint is that protein macromolecules do not collide with (or insert into) each other. With high protein-water volume ratio in crystals, they usually have to "smartly" close packed. Since the allowable motion is severely restricted, we can find a discrete candidate set to represent all the feasible packing arrangements. Noticing Gaussian functions have infinite tails, a Gaussian-based cost function (GCF) is constructed to evaluate the level of overlapping (or closeness) among ellipsoids with each ellipsoid represented by a Gaussian function or a mixture of Gaussian functions. The candidate packing arrangements can be obtained by minimizing the GCF to force the packing model to separate from each other and not overlap within the unit cell.

The shape of an ellipsoid can be captured by equidensity contours of a Gaussian function with the mean located at the ellipsoid center and the covariance matrix related to its semi-axis lengths. An arbitrarily oriented ellipsoid in $\mathbb{R}^n$ can be described as

$$(\mathbf{x} - \mu)^T R^T A R (\mathbf{x} - \mu) = 1, \tag{8}$$

where R is the rotation matrix, and $A = \mathrm{diag}[1/a_1^2, 1/a_2^2, \cdots, 1/a_n^2]$, with a_i denoting the semi-axis length of the ellipsoid. Compared with a Gaussian function in $\mathbb{R}^n$,

$$\rho(\mathbf{x}; \mu, \Sigma) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2} (\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu) \right), \tag{9}$$

we can see that when $\Sigma^{-1} = R^T A R$, the equidensity contours of the Gaussian function are ellipsoids with semi-axis lengths $k \cdot a_1$, $k \cdot a_2$, $\cdots$, $k \cdot a_n$, where $k \in \mathbb{R}_{\geq 0}$. To more accurately capture the shape of the ellipsoid with semi-axis lengths a_1, a_2, $\cdots$, a_n, we want the Gaussian function to have high and steady value inside the ellipsoid region and a quick drop outside it. We note that it is not necessary to eliminate the tail outside the ellipsoid since the interaction among the tails can help push the ellipsoids away from each other. We use a Gaussian mixture function $\psi(\mathbf{x}; \mathbf{a}, \mathbf{b})$, i.e.,

$$\psi(\mathbf{x}; \mathbf{a}, \mathbf{b}) = \sum_{i=1}^{n} \frac{a_i}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp\left(-\frac{b_i}{2} (\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu) \right), \tag{10}$$

instead of a single Gaussian $\rho(\mathbf{x})$ to approximate an ellipsoid. In the 1D case in Fig. 1, with both variances $\sigma = 1$, we can see that compared to the single Gaussian $\frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2})$, the Gaussian mixture function with $\mathbf{a} = 0.44 \cdot [3, -1]$ and $\mathbf{b} = 1.16 \cdot [1, 3]$, i.e.,

$$\psi(\mathbf{x}; \mathbf{a}, \mathbf{b}) = \frac{1.32}{\sqrt{2\pi}} \exp(-0.58x^2) - \frac{0.44}{\sqrt{2\pi}} \exp(-1.73x^2), \tag{11}$$

has a flatter top and faster decay tails.

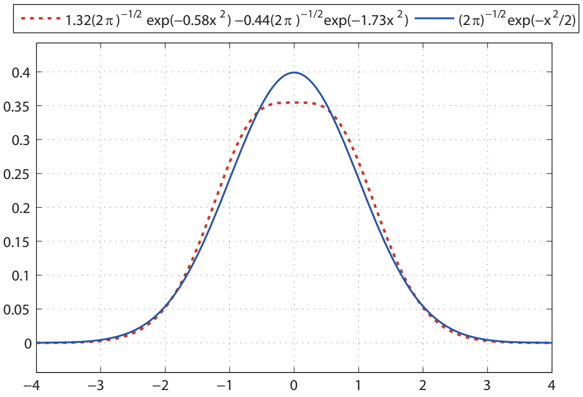

Fig. 1
THE COMPARISON BETWEEN A SINGLE GAUSSIAN WITH A MIXTURE OF GAUSSIANS.

The ellipsoid model of i^{th} domain in a multi-domain structure under a symmetry group Γ can be approximated by $\psi((h_i^{-1} \circ \gamma_j^{-1} \cdot \mathbf{x});\mathbf{a},\mathbf{b})$, where h_i is rigid-body operation of the i^{th} domain and γ_j is the symmetry operator in the symmetry group Γ. Therefore we define the GCF as

$$\text{GCF}(h_1,\cdots,h_N) \triangleq \int_{\mathbb{R}^n} \left[\sum_{j=0}^{m-1}\sum_{i=1}^{N} \psi((h_i^{-1} \circ \gamma_j^{-1} \cdot \mathbf{x}),\mathbf{a},\mathbf{b})\right]^2 d\mathbf{x}. \tag{12}$$

An advantage of Gaussian functions is that the integration of quadratic terms over $\mathbb{R}^n$ has a closed-form expression. We derived it as follows,

$$\begin{aligned}&\int_{\mathbb{R}^n} \rho_1(\mathbf{x};\mu_1,\Sigma_1)\rho_2(\mathbf{x};\mu_2;\Sigma_2)d\mathbf{x} \qquad (13)\\ =&\int_{\mathbb{R}^n}(2\pi)^{-n/2}(\det\Sigma_1)^{-1/2}\exp(-\frac{1}{2}(\mathbf{x}-\mu_1)^T\Sigma_1^{-1}(\mathbf{x}-\mu_1))\\ &(2\pi)^{-n/2}(\det\Sigma_2)^{-1/2}\exp(-\frac{1}{2}(\mathbf{x}-\mu_2)^T\Sigma_2^{-1}(\mathbf{x}-\mu_2))d\mathbf{x}\\ =&(2\pi)^{-n}(\det\Sigma_1\det\Sigma_2)^{-1/2}\int_{\mathbb{R}^n}\exp(-\frac{1}{2}(\mathbf{x}-\mu_1)^T\Sigma_1^{-1}(\mathbf{x}-\mu_1)\\ &-\frac{1}{2}(\mathbf{x}-\mu_2)^T\Sigma_2^{-1}(\mathbf{x}-\mu_2))d\mathbf{x}.\end{aligned}$$

$$\begin{aligned}\text{Since }&\int_{\mathbb{R}^n}\exp(-\frac{1}{2}\mathbf{x}^TM\mathbf{x}-m^T\mathbf{x}-C) \qquad (14)\\ &=(2\pi)^{n/2}(\det M)^{-1/2}\exp(\frac{1}{2}m^TM^{-1}m-C),\end{aligned}$$

(13) can be rewritten in a closed-form as

$$\begin{aligned}&\int_{\mathbb{R}^n}\rho_1(\mathbf{x};\mu_1,\Sigma_1)\rho_2(\mathbf{x};\mu_2;\Sigma_2)d\mathbf{x} \qquad (15)\\ =&(2\pi)^{-n}(\det\Sigma_1\det\Sigma_2\det(\Sigma_1^{-1}+\Sigma_2^{-1}))^{-1/2}\\ &\exp(\frac{1}{2}(\mu_1^T\Sigma_1^{-1}+\mu_2^T\Sigma_2^{-1})(\Sigma_1^{-1}+\Sigma_2^{-1})(\Sigma_1^{-T}\mu_1+\Sigma_2^{-T}\mu_2)\\ &-\frac{1}{2}(\mu_1^T\Sigma_1^{-1}\mu_1+\mu_2^T\Sigma_2^{-1}\mu_2)).\end{aligned}$$

The closed-form expression of the GCF can be easily derived from (15).

The main procedures of generating candidate phasing models by packing can be described by a flowchart in Fig. 2. In the first step, we discretize the configuration space by a coarse grid, and find the configuration with the smallest GCF value inside each "configuration cell" defined by the grid. The collision-free ones of these configurations form the candidate set of packing arrangements. This discrete candidate set reduces the whole configuration space to a much shorter list. We note that with a closed-form expression, minimizing the GCF is less computationally expensive compared to calculating $c(h_i)$ in traditional MR searches (see (7)).

In the next step, we use a Fourier-based cost function (FCF), where

$$\begin{aligned}&\text{FCF}(h_1,\ldots,h_N) \qquad (16)\\ &=\left[\int_{\mathbf{k}\in\Omega}\left(\hat{P}(g_1,\ldots,g_N;\mathbf{k})-\hat{P}(h_1,\ldots,h_N;\mathbf{k})\right)^2d\mathbf{k}\right]^{\frac{1}{2}},\end{aligned}$$

to sort these collision-free configurations from low to high. In our simulation, the function $f_i(\mathbf{x})$ defined in Sec. 3 are chosen to be the set indicator function for the ellipsoid representing body i. Then $\hat{P}(g_1,\ldots,g_N;\mathbf{k})$ and $\hat{P}(h_1,\ldots,h_N;\mathbf{k})$ are defined in (5) and (6), respectively.

Minimizing $\text{FCF}(h_1,\ldots,h_N)$ is similar to finding peaks in $c(h_i)$ except that we use a multi-domain model rather than a single-domain one. After the sorting, we keep 10 configurations with lowest FCF as a candidate list. These candidates indicate high correlations with the target structure. The FCF has the rugged surface of the configuration space, so to further improve the accuracy, a stochastic sampling method—Rapidly-exploring random tree (RRT) algorithm [15] is used to minimize the FCF around the "best candidate". The best candidate can be first chosen as the one with the lowest FCF in the set. If its FCF cannot be reduced below a threshold value C after running the RRT, we switch the best candidate to the one with the next lowest FCF.

5. EXPERIMENTAL EXAMPLE

In this section, the approach to phasing by using packing models is demonstrated in a 2D planar case, with ellipses representing both the domains of the model and target structures. All the angular parameters of the target structure are treated as being unknown, and the only priori information that we have is the magnitude of the Fourier transform of the electron density function $\hat{P}(g_1,\ldots,g_N;\mathbf{k})$. Our goal is to find the closest model configuration $\{h_1,\ldots,h_N\}$ with respect to the target structure $\{g_1,\ldots,g_N\}$. To illustrate our approach, a multi-ellipse-shaped "rabbit" with one "face" and two "ears" is constructed as a packing model for a 3-domain structure in P1 symmetry. Since translations have no impact on the packing result in P1 symmetry, the rabbit

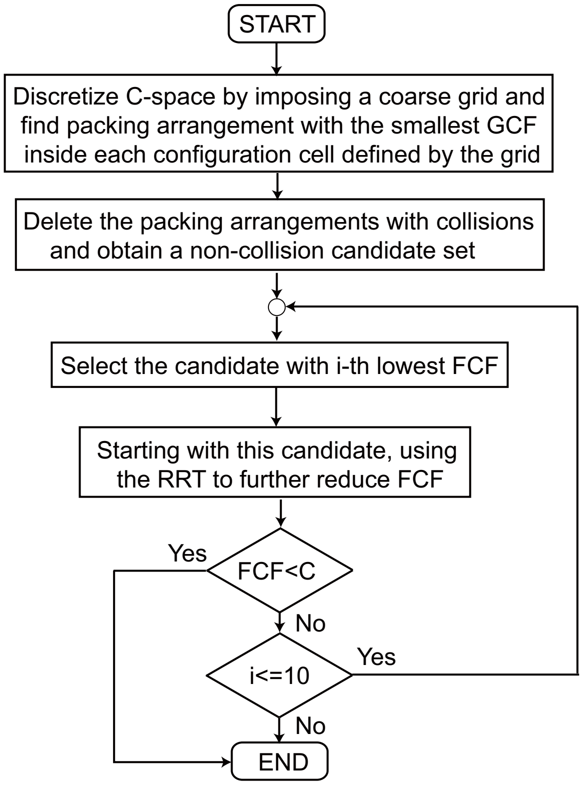
START
Discretize C-space by imposing a coarse grid and find packing arrangement with the smallest GCF inside each configuration cell defined by the grid
Delete the packing arrangements with collisions and obtain a non-collision candidate set
Select the candidate with i-th lowest FCF
Starting with this candidate, using the RRT to further reduce FCF
FCF<C
Yes
No
i<=10
Yes
No
END

Fig. 2

FLOWCHART OF FINDING CANDIDATE PHASING MODELS BY PACKING.

model has 3 DOF— the rotations of the face, θ_1 and two ears, θ_2 and θ_3 (see the dimensions and ranges of motion in Table 1).

For the Gaussian mixture function in this 2D planar case, we use the same ratios of a_1, a_2 and b_1, b_2 as the 1D case in (11), i.e., $a = m_a \cdot [3,-1]$ and $b = m_b \cdot [1,3]$. m_b^*—the optimal value of m_b, is chosen to "stretch or shrink" the Gaussian mixture function so that it can " best" represent the defined ellipse. After that, m_a^* is calculated to normalize the Gaussian mixture function with m_b^*. More specifically, we define m_b^* as

$$m_b^* = \arg\max_{m_b} |S_{\text{cand.}}(m_b)|, \tag{17}$$

where $S_{\text{cand.}}$ represents the non-collision candidate set, generated by obtaining the packing arrangements with the smallest GCF value inside each configuration box defined by the grid, and deleting the collision ones afterwards. $|S_{\text{cand.}}(m_b)|$ denotes the number of non-collision candidates in this set. With the optimal m_b, the GCF forces the packing models to separate from each other to the greatest extent, and the size of the non-collision candidate set is therefore maximized. Fig. 3 shows the size of the non-collision candidate set $|S_{\text{cand.}}(m_b)|$ with different m_b values under 3 different defined grids (in 30-, 40- and 60-degree increments). We can see when $m_b = 0.2$, $|S_{\text{cand.}}(m_b)|$ has the highest value, and the peak is independent of how we define the grid. In the experiment, we use the 30-degree grid, and 48 non-collision candidates can be found. With $m_b = 0.2$, we compare the contours of the Gaussian mixture function with the rabbit shape in Fig. 4, and we can see that it fits the shape of the rabbit model well. Also in Fig. 5, we compare collision checking results with GCF values in the θ_1-θ_2 plane with fixed θ_3=-90 degrees. It is shown that all non-collision configurations are located in the low GCF value regions, which demonstrates that by minimizing the GCF, the ellipses are less likely to have overlapping.

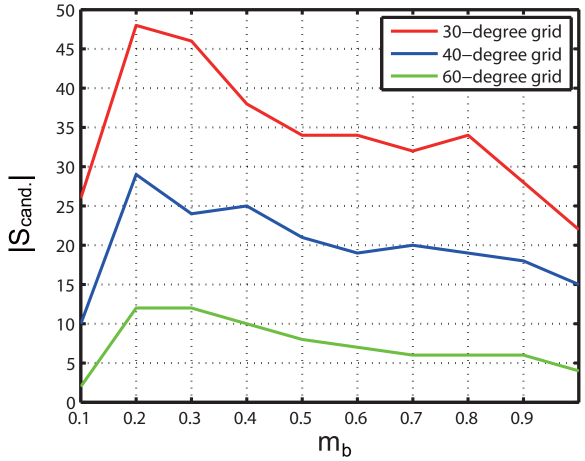

Fig. 3

THE SIZE OF THE NON-COLLISION CANDIDATE SET WITH DIFFERENT m_b VALUES UNDER 3 DIFFERENT DEFINED GRIDS (IN 30-, 40- AND 60-DEGREE INCREMENTS, RESPECTIVELY).

An example of packing results with the target structure randomly sampled in space is illustrated in Fig. 6. After generating the candidate set by minimizing the GCF, and sorting these candidates by the FCF from high to low, the best candidate in the set (Candidate 1 in Fig. 7) shows 1.50, 17.81 and 10.97 degrees of the error in θ_1, θ_2 and θ_3, respectively. After running the RRT around this candidate, these errors are further reduced to only 0.79, 2.14 and 0.19 degrees respectively, less than 1.2 % of the total rotation range. Table 2 shows 10 different numerical trials and the mean absolute errors (MAE), mean$\{\Delta\theta_1, \Delta\theta_2, \Delta\theta_3\}$, are all below 3 degrees.

6. CONCLUSIONS

Macromolecular crystallography has been the traditional workhorse for determining structural models in the field of biophysics. Within macromolecular crystallography, the molecular replacement method has been a highly successful

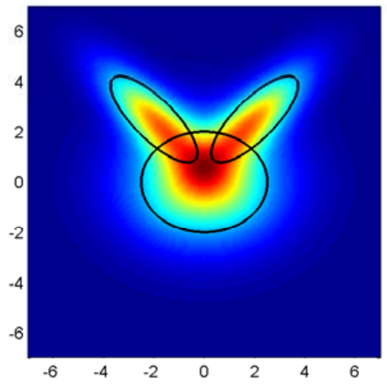

Fig. 4

THE COMPARISON OF THE RABBIT SHAPE WITH THE CONTOURS OF THE GAUSSIAN MIXTURE FUNCTION ($m_b = 0.2$).

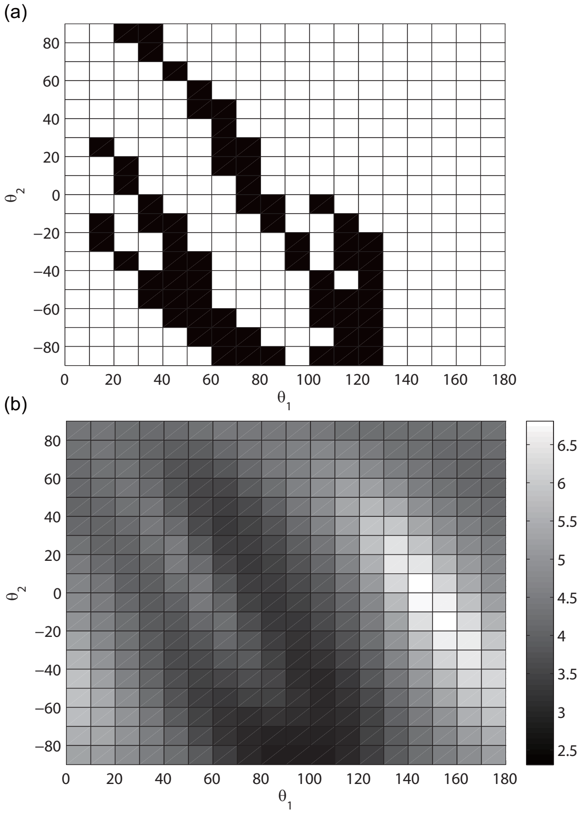

Fig. 5

THE COMPARISON OF (A) COLLISION CHECKING RESULTS WITH (B) GCF VALUES (WITH $m_b = 0.2$) IN THE θ_1-θ_2 PLANE (WITH θ_3=-90 DEGREES). IN (A), BLACK PIXELS REPRESENT THE NON-COLLISION CONFIGURATIONS AND WHITE ONES ARE COLLISION FREE. IN (B), THE PIXELS WITH DARKER COLORS REPRESENT THE CONFIGURATIONS WITH LOWER GCF VALUES, AND VISE VERSA.

Table 1

THE DIMENSIONS AND RANGES OF MOTION OF THE RABBIT PACKING MODEL

Dimensions	size of the unit cell	9 × 6.75
	semi-axis lengths of the face	2; 2.5
	semi-axis lengths of the ears	2.3; 0.92
Range of rotation	face: θ_1 (deg)	0 ∼ 180
	ears: θ_2, θ_3 (deg)	-90 ∼ 90

method for providing phasing models to combine with experimental information to obtain protein structures. In this paper we demonstrate that an alternative to molecular replacement, called "phasing by packing" is promising for multi-rigid-domain structures. Numerical results illustrate the potential of this method.

7. ACKNOWLEDGMENTS

We acknowledge NIH Grant R01GM075310 for the support of this work and Dr. E. Lattman for useful discussions.

References

[1] Hendrickson, W. A. and Ward, K. B., "A Packing Function for Delimiting the Allowable Locations of Crystallized Macromolecules," *Acta Cryst. A* 32:778-780 (1976).

[2] Williams, D.E., "Crystal Packing of Molecules," *Science* 147(3658):605-606 (1965).

[3] Damiani, A., Giglio, E., Liquori,A.M. , Mazzarell, L, "Calculation of Crystal Packing: A Novel Approach to the Phase Problem," *Nature* 215:1161-1162 (1967).

[4] Rossmann, M.G., Blow, D.M., "The Detection of Sub-Units within the Crystallographic Asymmetric Unit," *Acta Cryst.* 15:24-31 (1962).

[5] Rossmann, M.G., "Molecular replacement - historical background," *Acta Cryst. D*57:1360-1366 (2001).

[6] Navaza, J., "AMoRe: an Automated Package for Molecular Replacement," *Acta Cryst. A*50:157-163 (1994).

[7] Collaborative Computational Project Number 4, "The CCP4 suite: programs for protein crystallography," *Acta Cryst. D*50:760-766 (1994) http://www.ccp4.ac.uk/

[8] Jamrog, D.C., Zhang, Y., Phillips Jr., G.N., "SOMoRe: a multi-dimensional search and optimization approach to molecular replacement," *Acta Cryst. D*59:304-314 (2003).

[9] Jeong, J., Lattman, E., Chirikjian, G.S., "A Method for Finding Candidate Conformations for Molecular Replacement Using Relative Rotation Between Domains of a Known Structure," *Acta Cryst. D* D62, pp. 398-409, 2006.

[10] Crowther, R.A., and Blow, D.M. (1967). A method of positioning a known molecule in an unknown crystal structure. *Acta Cryst* **23**, 544.

[11] Crowther, R.A. (1972). The fast rotation function. In *The Molecular Replacement Method*, M.G. Rossmann, ed. New York: Gordon and Breach Science Publishers, 173-178.

[12] Lattman, E., Love, W.E., "A Rotational Search Procedure for Detecting a Known Molecule in a Crystal," Acta Cryst. B26, 1854-1857, 1970.

[13] Lattman, E.E., "Use of the Rotation and Translation Functions," Methods Enzymol. 115, 55-77, 1985

[14] Vagin, A., Teplyakov, A.,"MOLREP: an Automated Program for Molecular Replacement," J. Applied Cryst., 30, 1022-1025 1997

[15] LaValle, S. M. *Planning Algorithms*, Cambridge University Press, Cambridge, U.K., 2006.

[16] Chirikjian, G.S., Zhou, S., "Metrics on Motion and Deformation of Solid Models," *ASME J. Mechanical Design*, Vol. 120, No. 2, June, 1998, pp. 252-261.

Table 2

10 Numerical trials.

	Target			Best	Cand.		After	RRT		Final	errors	
Trial	θ_1	θ_2	θ_3	θ_1	θ_2	θ_3	θ_1	θ_2	θ_3	e_1	e_2	e_3
1	100.82	-72.21	-3.03	100.32	-90.00	-14.00	101.61	-74.35	-3.22	0.79	2.14	0.19
2	42.29	64.37	-69.25	43.07	60.00	60.00	42.96	64.29	-67.17	0.67	0.08	2.08
3	136.67	-68.70	-67.33	120.00	-39.37	-79.98	135.58	-67.54	-68.39	1.09	1.16	1.06
4	114.21	-63.46	-51.42	120.00	-81.03	-60.00	116.43	-64.71	-49.70	2.22	1.25	1.72
5	54.83	-49.51	-70.41	61.85	-60.00	-60.00	55.83	-50.37	-69.37	1.00	0.86	1.04
6	159.67	47.67	-2.65	173.97	26.77	14.89	160.75	44.52	0.43	1.08	3.15	3.08
7	101.63	-67.65	12.06	114.08	-88.32	31.05	103.72	-70.20	13.08	2.09	2.55	1.02
8	113.89	-73.69	30.76	120.00	-90.00	38.95	112.72	-74.80	29.70	1.17	1.11	1.06
9	66.41	27.29	-76.94	60.00	38.95	-90.00	63.20	30.78	-78.49	3.21	3.49	1.55
10	97.19	-1.59	-86.46	120.00	-39.37	-79.98	100.02	-2.89	-83.53	2.83	1.30	2.93

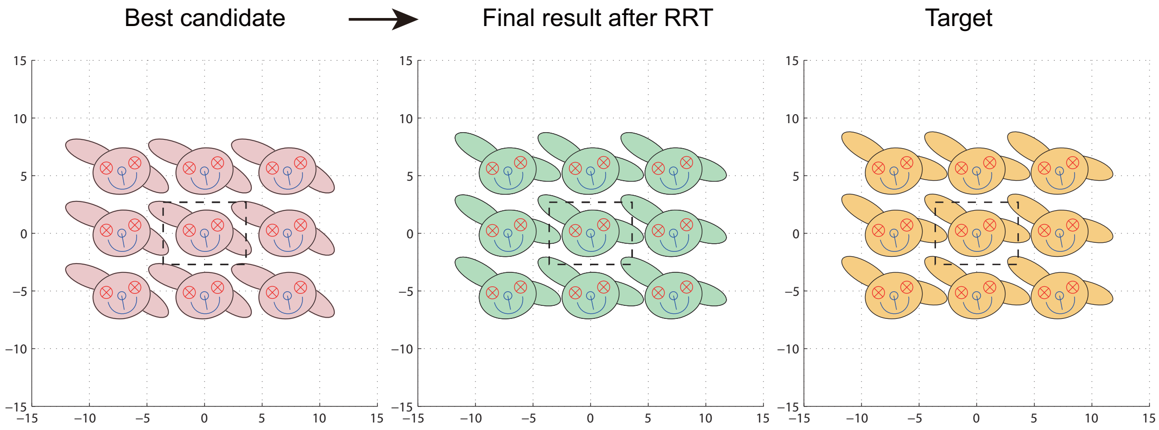

Fig. 6

An example of packing results with the target structure randomly sampled in the space.

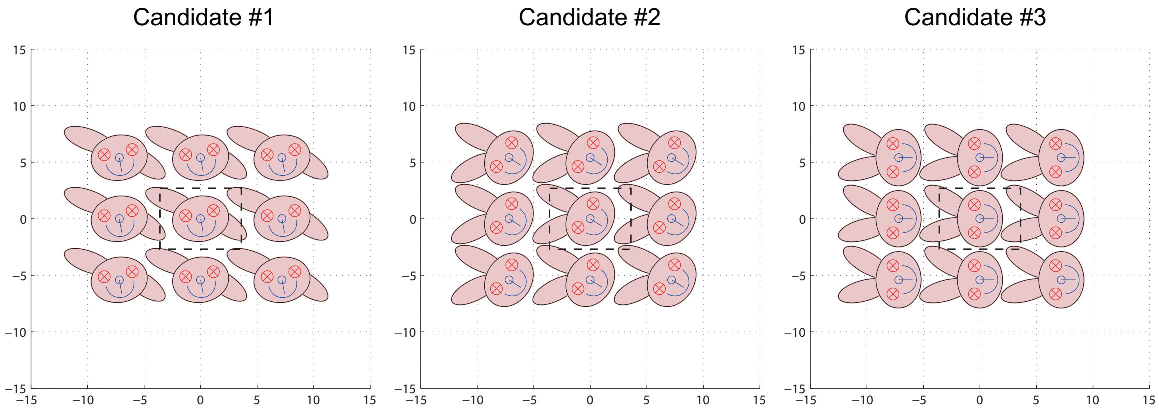

Fig. 7

3 candidate packing arrangements for the example in Fig. 6.