Geometric Constraints for the Phase Problem in X-Ray Crystallography

Corinna Heldt, Alexander Bockmayr
Freie Universität Berlin, FB Mathematik und Informatik,
Arnimallee 6, 14195 Berlin, Germany
Corinna.Heldt@fu-berlin.de,
Alexander.Bockmayr@fu-berlin.de

Abstract

X-ray crystallography is one of the main methods to establish the three-dimensional structure of biological macromolecules. In an X-ray experiment, one can measure only the magnitudes of the complex Fourier coefficients of the electron density distribution under study, but not their phases. The problem of recovering the lost phases is called the phase problem. Building on earlier work by Lunin/Urzhumtsev/Bockmayr, we extend their constraint-based approach to the phase problem by adding further 0-1 linear programming constraints. These constraints describe geometric properties of proteins and increase the quality of the solutions. The approach has been implemented using SCIP and CPLEX, first computational results are presented here.

1 Introduction

Knowledge about the three-dimensional structure of biological macromolecules is an essential foundation of structural biology and biotechnology. In X-ray crystallography the arrangement of atoms within a crystal is determined from a three-dimensional representation of the electron density. From X-ray experiments one gets *diffraction data* depending on the molecular structure, i.e., the intensities of reflections of X-rays diffracted by the crystal. X-rays are scattered exclusively by the electrons in the atoms, so one is searching for a relation between the measured intensities of the beams diffracted at the object in question and the crystal structure, which can be described by the electron density distribution. The electron density represents probabilistically where electrons can be found in the molecule. With the help of diffraction data and the usage of mathematical as well as experimental methods, an electron density map can be derived. *Direct methods* use mathematical techniques to compute an electron density map from the diffraction data without any further experiments. The main problem here is the *phase problem*: experiments provide only the intensities of the X-rays diffracted in different directions and so the electron density magnitudes can be calculated, whereas the information about the phase shift is lost.

Lunin, Urzhumtsev and Bockmayr [8] proposed a 0-1 linear programming approach to direct phasing. This approach yields a set of solutions. In order to increase the quality of this solution set, we formulate some geometric properties of proteins as additional 0-1 linear programming constraints. In [3], we described the basic ideas of the 0-1 linear programming approach by Lunin, Urzhumtsev and Bockmayr [8], now we derive the new geometric constraints and present first computational results.

2 The phase problem

Every crystal consists of identical molecules, resp. complexes of molecules strictly ordered in all three dimensions. This means that we can find a parallelepiped called *unit cell* containing such a complex of molecules which builds up the whole crystal if it is repeatedly stacked together in all three dimensions. We will denote the unit cell's volume with V_{cell}. Let $\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3 \in \mathbb{R}^3$ span the unit cell. Then we can write every vector $\mathbf{r} \in \mathbb{R}^3$ in this basis, i.e., $\mathbf{r} = x_1\mathbf{b}_1 + x_2\mathbf{b}_2 + x_3\mathbf{b}_3$, where $\mathbf{x} = (x_1, x_2, x_3)^T \in [0,1]^3$ is the vector of coordinates of $\mathbf{r}$ with respect to the basis $\{\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3\}$. We are searching for the electron density distribution $\rho(\mathbf{x})$ over the crystal. Due to the crystal structure, ρ is a periodic function and therefore can be developed into a *Fourier series* [6]

$$\rho(\mathbf{x}) = \frac{1}{V_{cell}} \sum_{\mathbf{h} \in \mathbb{Z}^3} \mathbf{F}(\mathbf{h}) \exp(-2\pi i(\mathbf{h}^T\mathbf{x})), \ \mathbf{x} \in V. \tag{1}$$

The Fourier coefficients $\mathbf{F}(\mathbf{h}), \mathbf{h} \in \mathbb{Z}^3$, which are called *structure factors* in crystallography, are given by the formula

$$\mathbf{F}(\mathbf{h}) = \int_V \rho(\mathbf{x}) \exp(2\pi i(\mathbf{h}^T\mathbf{x}))d\mathbf{x}. \tag{2}$$

Since these are complex numbers, the structure factors can be written in the form $\mathbf{F}(\mathbf{h}) = F(\mathbf{h})\exp(i\varphi(\mathbf{h}))$, where $F(\mathbf{h}) = |\mathbf{F}(\mathbf{h})|$ is the *magnitude* and $\varphi(\mathbf{h}) \in [0, 2\pi[$ the *phase.*

The only experimental data we get in X-ray-crystallography are the reflection intensities. The intensity $I(\mathbf{h})$ of a reflection is proportional to the magnitude of the squared structure factors, with a known constant of proportionality, i.e., $C \cdot I(\mathbf{h}) = |\mathbf{F}(\mathbf{h})|^2, C \in \mathbb{R}$. Thus, all we can calculate from our experimental data are the structure factor magnitudes. The phase information is lost and must be restored by other means. This is called the *phase problem.*

3 0-1 linear programming approach

Now, the main ideas of the approach proposed in [8] are presented. Instead of calculating the electron density distribution in the whole unit cell, we will work on a grid. Using discrete Fourier transforms, we calculate electron densities at the grid points. Consider a grid $\Pi = [0, M_1 - 1] \times [0, M_2 - 1] \times [0, M_3 - 1] \subseteq \mathbb{Z}^3$, where $M = M_1M_2M_3$ is the total number of grid points. Denote by $\mathbf{M}$ the diagonal matrix $\text{diag}(M_1, M_2, M_3)$, with diagonal elements $M_1, M_2, M_3 \in \mathbb{N}$. The values of the electron density function $\rho(\mathbf{x}),\ \mathbf{x} \in V$ at the grid points are described by the *grid electron density function* $\rho_g(\mathbf{j}) = \rho(\mathbf{M}^{-1}\mathbf{j})$, $\forall \mathbf{j} \in \Pi$. We define the *grid structure factor* $\mathbf{F}_g(\mathbf{h})$ by the discrete Fourier transform

$$\mathbf{F}_g(\mathbf{h}) = \frac{1}{M} \sum_{\mathbf{j} \in \Pi} \rho_g(\mathbf{j}) \exp(2\pi i(\mathbf{h}^T\mathbf{M}^{-1}\mathbf{j})), \ \forall \mathbf{h} \in \Pi. \tag{3}$$

If we know the grid structure factors, we can restore the grid electron densities

$$\rho_g(\mathbf{j}) = \sum_{\mathbf{h} \in \Pi} \mathbf{F}_g(\mathbf{h}) \exp(-2\pi i(\mathbf{h}^T\mathbf{M}^{-1}\mathbf{j})), \ \forall \mathbf{j} \in \Pi, \tag{4}$$

using the inverse discrete Fourier transform.

In the context of direct phasing, it may be sufficient to find a binary *envelope* of the regarded molecules, i.e., a binary function representing areas where the electron density is above a certain cut-off level κ [8]. Using this idea, we may replace the unknowns $\rho_g(\mathbf{j})$ by binary variables $z_\mathbf{j} \in \{0, 1\}$, for each grid point $\mathbf{j} \in \Pi$, satisfying $z_\mathbf{j} = 0$, if $\rho(\mathbf{j}) \leq \kappa$ and $z_\mathbf{j} = 1$ otherwise.

By restricting the possible phase values $\varphi(\mathbf{h}) \in [0, 2\pi[$, $\forall \mathbf{h} \in \Pi$ to four ones, i.e., $\varphi(\mathbf{h}) \in \{\pm\frac{\pi}{4}, \pm\frac{3}{4}\pi\}$, $\forall \mathbf{h} \in \Pi$, the phase problem can be stated as a system of linear inequalities in 0-1 variables for representing the electron density values at grid points and for representing the phases. By penalizing the amount of violation, a suitable objective function can be introduced [8, 3].

In general, the resulting 0-1 linear program for solving the phase problem does not have a unique optimal solution, but a set of different optimal solutions. In order to reduce the number of those and at the same time increase the quality of the remaining ones, additional constraints can be added.

4 Additional constraints

In the electron density distribution of a protein, no peaks of very high or very low electron density occur, if an appropriate resolution is used. This means, in the grid electron density distribution, no isolated points of high electron density surrounded by low electron density values as well as no isolated points of low electron density surrounded by high electron density values are expected to occur.

Definition 1 (Neighbour relation)**.** *Two grid points* $\mathbf{j}_1 \in \Pi$ *and* $\mathbf{j}_2 \in \Pi$ *are neighbours, denoted by* $\mathbf{j}_1 \mathfrak{n} \mathbf{j}_2$*, if and only if* $\parallel \mathbf{j}_1 - \mathbf{j}_2 \parallel_2 = 1$*.*

Definition 2 (Isolated point)**.** *A binary grid point* $z_{\mathbf{j}} \in \Pi$ *is called isolated if and only if* $z_{\mathbf{j}} = 0 \Rightarrow z_{\mathbf{i}} = 1, \ \forall \ \mathbf{i}\mathfrak{n}\mathbf{j}$ *and* $z_{\mathbf{j}} = 1 \Rightarrow z_{\mathbf{i}} = 0, \ \forall \ \mathbf{i}\mathfrak{n}\mathbf{j}$*.*

Every interior grid point has six neighbours, thus the condition $-5 \le z_{\mathbf{j}} - \sum\limits_{\mathbf{i}\mathfrak{n}\mathbf{j}} z_{\mathbf{i}} \le 0$, for all $\mathbf{j} \in \Pi$ states the exclusion of isolated interior grid points.

5 Connectivity

At low resolution and a high enough cut-off level κ, the high-level region $\Omega_\kappa \overset{def}{=} \{\mathbf{j} : \rho(\mathbf{j}) > \kappa\}$ is expected to consist of a small number of connected components, which should be equal to the number of molecules inside the unit cell [9]. At lower cut-off level these components merge into fewer regions. So it is possible to give an upper bound for the number of molecules in advance.

We define a graph representing properties of the binary grid electron density maps. Let $G_\Pi = (V_\Pi, E_\Pi)$ be an undirected graph with $M = M_1 \cdot M_2 \cdot M_3$ vertices denoted by $v_{\mathbf{j}}, \ \mathbf{j} \in \Pi$. Vertices $v_{\mathbf{j}} \in V_\Pi$ and $v_{\mathbf{i}} \in V_\Pi$ with $\mathbf{j}$ and $\mathbf{i}$ being neighbours are connected by edges, i.e., $E_\Pi = \{e = (v_{\mathbf{j}}, v_{\mathbf{i}}) \mid \mathbf{j}\mathfrak{n}\mathbf{i}\}$. Let $V_\Pi^* \subseteq V_\Pi$ be the set of vertices with a corresponding electron density above the cut-off level, i.e., the set of vertices satisfying $V_\Pi^* = \{v_{\mathbf{j}} \mid z_{\mathbf{j}} = 1, \ \mathbf{j} \in \Pi\}$. With $E_\Pi^* \subseteq E_\Pi$ we denote the set of edges in the subgraph $G_\Pi^* = (V_\Pi^*, E_\Pi^*)$ induced by V_Π^*.

The binary grid electron density distribution contains $K \in \mathbb{N}$ components, if and only if the corresponding graph $G_\Pi^* = (V_\Pi^*, E_\Pi^*)$ contains K connected components. Figure 2 shows the graph representing the binary grid electron density distribution. Black filled vertices represent grid electron density values above the cut-off level, neighboured black vertices are connected by solid edges.

We introduce 0-1 variables $e_{\mathbf{j_1},\mathbf{j_2}}$ for $\mathbf{j_1}, \mathbf{j_2} \in \Pi$ with $\mathbf{j_1} \ \mathfrak{n} \ \mathbf{j_2}$. These variables should take the value 1, if the corresponding edge connects two neighbouring nodes $\mathbf{j_1}, \mathbf{j_2} \in \Pi$ with $z_{\mathbf{j_1}} = z_{\mathbf{j_2}} = 1$, and 0 otherwise. The constraint $-1 \le 2e_{\mathbf{j_1}\mathbf{j_2}} - z_{\mathbf{j_1}} - z_{\mathbf{j_2}} \le 0$, for all $\mathbf{j}, \mathbf{j_1}, \mathbf{j_2} \in \Pi$ with $\mathbf{j_1} \ \mathfrak{n} \ \mathbf{j_2}$, ensures this condition.

Now, a 0-1 linear programming approach will be presented to model that a binary grid electron density distribution satisfies the '*K-component-constraint*', i.e., it contains at most

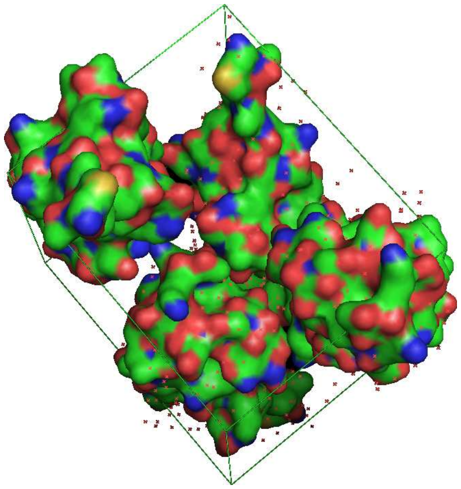

Figure 1: Unit cell of Protein G

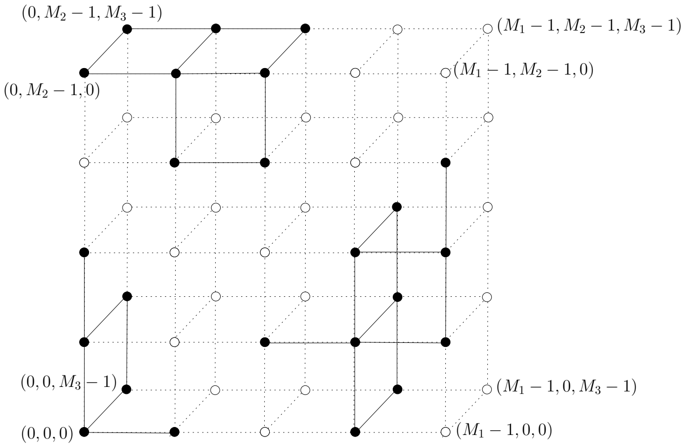

Figure 2: The graph $G^*_\Pi = (V^*_\Pi, E^*_\Pi)$

$K \in \mathbb{N}$ components. For any subset $\emptyset \neq T \subsetneq \Pi$ we introduce a binary variable u_T indicating whether T contains grid points $\mathbf{j} \in \Pi$ where the variable $z_\mathbf{j}$ takes the value 1.

$$u_T \overset{def}{=} \begin{cases} 1, & \text{if } \sum\limits_{\mathbf{j} \in T} z_\mathbf{j} \geq 1 \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

If this is the case for more than K disjoint subsets, there have to be edges connecting some of these components, otherwise the 'K-component-constraint' would be violated.

Theorem 1 ([7])**.** *A binary grid electron density distribution* $z^* \in \{0,1\}^{M_1 \times M_2 \times M_3}$ *contains* *at most K components if it satisfies the following constraints:*

$$-1 \leq 2e_{\mathbf{j_1 j_2}} - z_{\mathbf{j_1}} - z_{\mathbf{j_2}} \leq 0, \tag{6}$$

$$\frac{1}{|T_i|}\sum_{\mathbf{j}\in T_i} z_{\mathbf{j}} \leq u_{T_i} \leq \sum_{\mathbf{j}\in T_i} z_{\mathbf{j}}, \tag{7}$$

$$\sum_{i=1}^{K+1} u_{T_i} - K \leq \sum_{(\mathbf{j_1},\mathbf{j_2})\in\delta(T_1,\ldots,T_{K+1})} e_{\mathbf{j_1 j_2}} \tag{8}$$

$$u_{T_i},\ z_{\mathbf{j}}, e_{\mathbf{j_1 j_2}} \in \{0,1\}, \tag{9}$$

$$\forall \emptyset \neq T_1,\ldots,T_{K+1} \subsetneq \Pi,\ \bigcup_{i=1}^{K+1} T_i = \Pi,\ T_i \cap T_j = \emptyset,\ \forall i \neq j,\ i,j \in \{1,\ldots K+1\},$$

$$\forall \mathbf{j}, \mathbf{j_1}, \mathbf{j_2} \in \Pi,\ \textit{with}\ \mathbf{j_1}\ \mathfrak{n}\ \mathbf{j_2}.$$

Here $\delta(T_1,\ldots,T_{K+1})$ denotes the set of all edges connecting two different components T_i, T_j, with $i \neq j \in \{1,\ldots K+1\}$.

The number of constraints in (8) grows exponentially in the number of nodes. Using a separation algorithm within a branch-and-cut framework [7], only certain violated inequalities will be added to the formulation.

In the constraint programming literature, global constraints for restricting the number of connected components have been studied in [5].

6 Computational results

In order to evaluate the approach, real protein data from the Protein Data Bank [1] was taken. For the implementation, we used SCIP Version 1.2.0 [2] together with CPLEX 11.0 [4] as IP-solver. SCIP can solve mixed-integer as well as constraint integer programming problems. The running time to calculate a solution on a 6 × 6 × 6-grid (216 independent grid points) on a i686 with 4 processors, a 3GHz CPU and 3GB RAM was about 10 minutes CPU time without additional constraints, and about 50 minutes CPU time with all constraints added. In the latter case, about 900 search nodes and 23MB of memory were needed, without the additional constraints 250 search nodes and 28MB of memory.

Once a set of solutions has been calculated, we evaluate the quality of those solutions. Using the minimal molecular volume that has been defined in the solution process to specify the number of non-zero grid values, the grid electron density distribution of the original protein is binarised. The distance $D(z_{exact}, z^i_{calc})$ between the resulting binary electron density z_{exact} and the calculated ones z^i_{calc}, $i \in \{1,\ldots,N\}$, where $N \in \mathbb{N}$ is the number of computed solutions, is defined by

$$D(z_{exact}, z^i_{calc}) \overset{def}{=} \sum_{\mathbf{j}\in\Pi} \left| z_{exact}(\mathbf{j}) - z^i_{calc}(\mathbf{j}) \right|. \tag{10}$$

The smaller the distance value, the better the quality of the considered solution. The smallest distance reached in the test run is $D_{min} \overset{def}{=} \min_{i=1,\ldots,N} D(z_{exact}, z^i_{calc})$.

As the exact solution normally is not known in advance, we use a method to get an average solution from the set of computed solutions. One possibility to calculate such an average solution for a set of $N \in \mathbb{N}$ solutions is the following one:

$$z_{av}(\mathbf{j}) \overset{def}{=} \frac{1}{N} \sum_{i=1}^{N} z^i_{calc}(\mathbf{j}), \ \forall \mathbf{j} \in \Pi, \quad D_{av} \overset{def}{=} D(z_{exact}, z_{av}). \tag{11}$$

Obviously, in general z_{av} is not a binary function. Using the defined molecular volume value, it can be binarised and compared to the exact solution.

Another possibility would be to choose the solution with a minimum distance from all other solutions. For every solution, the distances to all others are summed up, the solution for which this sum is minimal is chosen as reference solution z_{ref}:

$$D_{sum}(i) \overset{def}{=} \sum_{j=1}^{N} \sum_{\mathbf{j} \in \Pi} \left| z^i_{calc}(\mathbf{j}) - z^j_{calc}(\mathbf{j}) \right|, \ \forall i \in \{1, \ldots, N\}, \tag{12}$$

$$z_{ref} \overset{def}{=} z^i_{calc}, \text{ with } D_{sum}(i) = \min_{j=1,\ldots,N} D_{sum}(j), \quad D_{ref} \overset{def}{=} D(z_{exact}, z_{ref}). \tag{13}$$

In the table below some test results on $6 \times 6 \times 6$-grids are shown, based on the data for Protein G [1]. In order to get reasonable running times, a small grid size was chosen. For real applications it would be desirable to handle bigger grid sizes. A covering of 30% is forced, the original binary electron density distribution then consists of 1 component. The 70 best solutions (with respect to the objective function specified in [3]) were considered. Only 28 of them also consisted of 1 component, 49 of them consisted of at most 2. The maximum number of components in one of these 70 solutions was 9.

Constraints	# sol	p_{min}	p_{av}	p_{ref}
none	70	72%	56%	54%
iso	67	72%	62%	54%
connected (2)	49	72%	66%	63%
connected (1)	28	72%	74%	65%
iso, connected (2)	49	72%	69%	68%
iso, connected (1)	28	72%	74%	70%

In the first column, the used additional constraints are specified: either only the constraint excluding isolated points (iso), or the constraint excluding isolated points and the ‘K-component-constraint’ (connected). In brackets the maximum number of components allowed is specified. The second column shows the number of solutions from the original solution set satisfying these constraints.

In the other columns, the percentage of correct solution values is given for the different distance measures, i.e.,

$$p_{min} = \frac{|\Pi| - D_{min}}{|\Pi|}, \ p_{av} = \frac{|\Pi| - D_{av}}{|\Pi|}, \ p_{ref} = \frac{|\Pi| - D_{ref}}{|\Pi|}. \tag{14}$$

Obviously, the values of p_{av} as well as p_{ref} increase by adding stricter constraints, showing the increasing quality of the regarded solutions.

7 Conclusions and further work

Based on the 0-1 linear programming approach to model the phase problem presented at WCB 2008 [3], we derived a way to model additional 0-1 linear programming constraints representing geometric properties of proteins. First results show that adding those to the original 0-1 program results in a higher quality of the set of solutions. Now, this approach will be tested on more data and also on bigger grids. Concerning future work, one could think of better ways to create a solution from the resulting solution set or of including further constraints.

References

[1] Protein Data Bank, 2010. `http://www.rcsb.org`.

[2] T. Achterberg. *Constraint Integer Programming*. PhD Thesis, TU Berlin, July 2007.

[3] C. Brinkmann and A. Bockmayr. A constraint-based approach to the phase problem in X-ray crystallography. In *WCB08 - Workshop on Constraint Based Methods for Bioinformatics, Paris*, 2008.

[4] CPLEX. *CPLEX*. ILOG, 2010. `http://www.ilog.com/products/cplex/`.

[5] G. Dooms. *The CP(Graph) Computation Domain in Constraint Programming*. PhD Thesis, Université Catholique de Louvain, Sept 2006.

[6] J. Drenth. *Principles of protein X-ray crystallography*. Springer-Verlag, 1994.

[7] C. Heldt. Constraint based methods for phasing in crystallography. PhD Thesis, FU Berlin, in preparation.

[8] V. Y. Lunin, A. Urzhumtsev, and A. Bockmayr. Direct phasing by binary integer programming. *Acta Crystallogr A*, 58(Pt 3):283–291, May 2002.

[9] N. Lunina, V. Y. Lunin, and A. Urzhumtsev. Connectivity-based ab initio phasing: from low resolution to a secondary structure. *Acta Crystallogr D Biol Crystallogr*, 59(Pt 10):1702–1715, Oct 2003.