Solving A Synthetic Protein Structure by Iterated Projections

Ling Jin[1] and Xianrui Cheng[1]

[1]Department of Biological Science, University of Southern California, Los Angeles, CA 90089, USA

Abstract

X-ray crystallography is a well-established method to solve protein structures. X-rays diffracted by the protein crystal carry two pieces of information that together allow faithful reconstruction of the crystal structure: intensity and phase. However, only the intensity information can be collected in a diffraction experiment, making the solution of the structure a major challenge. Here we demonstrate accurate recovery of a protein structure using only intensity information. We generated artificial X-ray diffraction data of a hypothetical protein and successfully recovered its structure by applying an iterated projection algorithm. Our work privides a proof-of-principle example of protein structure reconstruction from partial information, and shows the promise of this algorithm in solving more complicated real protein structures.

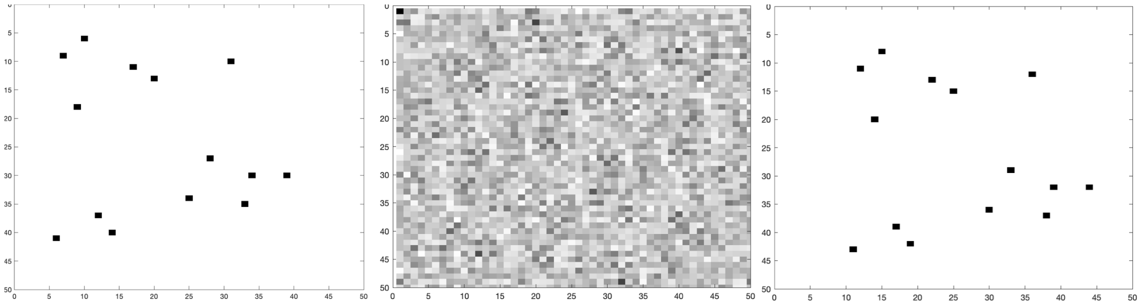
1 Introduction

The structure of a protein helps understand its function, for example how it can interact with the other molecules. When a protein is in a crystalline state, its atoms are arranged periodically in the lattice. In the following context, the lattice will be referred as the real space. Crystal atoms diffract a beam of X-ray and the scattered X-ray will be collected as diffraction pattern in the screen which is called the reciprocal space as a mathematical space constructed in a periodical way on the real space. Therefore X-ray crystallography is an effective method to examine the structure of a protein. Since the X-ray diffraction pattern is the Fourier transform of the scattering object, the reciprocal space is also called the Fourier space. If the angle and intensity of these diffracted beams in the reciprocal space can be precisely measured, one can restore the three-dimensional image of the electron density in the real space through the Inverse Fourier Transform. Then, the electron density map describing all the density of the electron can be used to determine the average position of the atoms in the crystal which is also referred to atom density map.

However, in actual experiments, only the intensity in the reciprocal space can be collected, while the phase-angle information is lost[1]. The difference-map algorithm is an iteration-based method which has a long history of being used in phase retrieval. In this algorithm, the iterative map enforces two constraints by repeated projections in the real and reciprocal spaces by projections, recovering the true density map upon computational convergence[2][3]. Here, we apply this algorithm to synthetic data and show that the implementation can fully restore the atom density map with.

2 Results

The implementation of the difference map algorithm under synthesized two-dimensional condition can completely restore the accurate atom density map from only the intensity information. The artificial X-ray diffraction data of a hypothetical protein is generated by Fourier Transform of a synthesized atom density map. By taking only the absolute value of the Fast Fourier Transform data, we simulated the situation where only the diffraction intensity can be collected in actual experiments (figure 1a and b). The algorithm completely recovers the atomic density in two-dimensional crystal(figure 1c). Since the atoms in protein crystals are arranged periodically, the atomic density restoration map constructed by the algorithm tends to shift or flip, but the relative position between atoms is preserved(figure 1a and c). Therefore the intensity information completely restores the atomic density map of the initially randomly generated protein.



(a) (b) (c)

Figure 1: **A restoring process of a two-dimensional density.** (a) Atomic density map of 14 atoms randomly synthesized in a 50x50 grid unit canvas. (b) The intensity map of the synthesized atom after Fourier transform, which is served as the starting point for restoring the atomic density map. (c) Atomic density map restored after iteration. After panning, it shows the complete restoration of the synthetic atom density map.

The difference map algorithm is used to find the intersection that satisfies the two constraints as the two projections converge. One constraint is the Fourier transform of the density in real space is aligned with the measured intensity in the reciprocal space. The other constraint is that the density of atoms in real space should be positive.The iteration of projections is monitored by the norm difference (ε), which measures the convergence of the intensity constraints and atomicity constraints. When the norm difference become zero, the intersection point will be considered as the solution. As the number of iterations increases, the gap between the two constraints narrows and ε value will eventually drop to close to zero(figure 2). At this point, the two constraint sets can be considered to converge to a fixed point as the point of intersection. This projection at the fixed point satisfies these two constraints at the same time, so it is considered to be an estimate of the true atom density map.

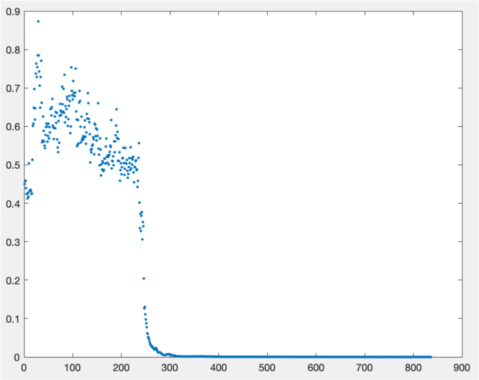


Figure 2: **A typical scatter of plot constraint differences ε evolving with the number of iterations.** Each dot represents a normalized difference between density constraint map and atomicity constraints map. As the number of iterations increases, the difference between the two constraints set gradually decreases until it approaches zero.

3 Discussion

Here we recovered crystal structure using only the intensity information. The difference map algorithm correctly solved a two-dimensional protein crystal structure from the synthetic data. This method may be extended to solve actual protein atom density map from the experimental data through the same principle.

However, the actual diffraction data is not as error-free as the synthesized data. Experimental and statistical errors in intensity measurement will lead to uncertainty and inaccuracy. In addition, the electron density map in the actual experiment will not be just a few solid black dots on the white canvas like our simulated image. In fact, an atom is not represented by a pixel, but will appear in several adjacent pixels in the form of an electron cloud with Gaussian density distribution.

The optimal value of the key parameter β has to be selected carefully in a real experiment to reach the highest accuracy and effectiveness of the algorithm. With perfect synthetic data, as long as it is in the interval $(0.4 \sim 0.8)$, no matter what value β takes, the atom density map will be fully restored. However, with imperfect data, the parameter β has to be optimized to achieve the best performance of the algorithm.

In conclusion, understanding the shape of a protein is essential for understanding protein's function, designing therapeutic drugs, and discovering diseases caused by misfolded proteins. Our current implementation can be used in crystallography to fully restore the atom density map based on synthetic intensity data. The algorithm optimized for real diffraction data might be able to solve the structure of various proteins.

4 Methods

This implementation uses the difference map algorithm to analyze the synthesized diffraction data. The magnitude of the diffraction data simulates the intensity in the reciprocal space collected by X-ray crystallography under real conditions. In the simulatated cases, the diffraction data is generated by Fast Fourier transform of a synthetic two-dimensional atom density map with a certain number of artificial atoms. Each atom is simulated by a single pixel with a density value of one, while other pixels maintain a density value of zero, indicating that there are no atoms.

The intensity projection $\Pi_F(\rho)$ optimizes the atom density ρ on the real space grid. Since the valid density has the property that the intensity in the reciprocal space should keep the correct magnitude while the distance $\|\Pi_F(\rho) - \rho\|$ in the real space is minimized. Since Euclidean distance is preserved before and after Fourier Transform, achieving the distance-minimizing property in reciprocal space gives back the projection with minimized distance in real space. In order to keep the magnitude matching the measured intensity, each point of the projection has to lay on the circle corresponding to the measured magnitude F_q. The projection in the reciprocal space can be accomplished by redirecting the intensity $\widetilde{\rho_q}$ while keeps its magnitude:

$$\left[\widetilde{\Pi}_F(\widetilde{\rho})\right]_q = \frac{F_q}{|\widetilde{\rho_q}|}\widetilde{\rho_q}. \tag{1}$$

The intensity projection in real space is expressed by the Inverse Fourier Transform of its projection in the reciprocal space:

$$\Pi_F = \mathcal{F}^{-1}\widetilde{\Pi}_F\mathcal{F}. \tag{2}$$

By applying Fast Fourier Transform, the maximum computational cost of projecting is $MlogM$, where M is the number of grid points in real or reciprocal space.

The atomic support projection $\Pi_s(\rho)$ uses a known number of compact subsets of grid points to present atoms. Under the premise that the number of atoms N is already known, atomic support is defined as the first N pixels with relatively larger positive local-maxima's and their neighbors' density values in the real space. A local maximum refers to a larger density value than any of its eight neighboring grid points under two-dimensional conditions. Since the density of an atom is positive, only the positive values are copied onto the projection with the same grid location while other grid units are set to zero:

$$[\Pi_s(\rho)]_r = \begin{cases} \rho_r & \text{if } \rho_r > 0 \\ 0 & \text{if } otherwise. \end{cases} \tag{3}$$

Difference map obtains the solution density ρ_{sol} that satisfy both intensity constraints and atomicity constraints:

$$D : \rho \mapsto \rho + \beta(\Pi_1 \circ f_2 - \Pi_2 \circ f_1)(\rho), \tag{4}$$

where $\Pi_1 = \Pi_S$ and $\Pi_2 = \Pi_F$. Each map f_i is expressed as

$$f_i = (1 + \gamma_i)\Pi_i - \gamma_i \ (i = 1, 2). \tag{5}$$

$$\gamma_1 = -1/\beta \tag{6}$$

$$\gamma_2 = 1/\beta. \tag{7}$$

During the course of iteration, as the two sets of projections converge to their intersection, the norm of the difference ε between two projections becomes smaller and smaller:

$$\varepsilon_n = \|(\Pi_1 \circ f_2 - \Pi_2 \circ f_1)(\rho(n))\| \tag{8}$$

The solution ρ_{sol} is achieved when the difference is small enough to be close to zero.

5 Code availability

Source code for the algorithm is available at https://github.com/changeless/Two-dimensional-Synthetic-Crystal-Phase-Problem

References

[1] Stokes, A. R. *Nature* (1948).

[2] Elser, V. *Acta Crystallographica Section A* (2003).

[3] Elser, V., Rankenburg, I., and Thibault, P. *PNAS* (2007).