Introduction to single crystal X-ray analysis XIV. Model construction and refinement, and evaluation of results

Takashi Matsumoto *

1. Introduction

Single crystal X-ray structural analysis of proteins includes these steps: expression, crystallization of protein, data collection, phase determination, model building, model fitting, and refinement of the molecular model. When the difficult steps of crystallization and phase determination are overcome, the process of structural analysis can move on to model building and refinement.

The initial electron density map after phase determination is usually noisy. Therefore, before model building, noise is reduced by density modification methods, such as "solvent flattening" and "histogram matching," resulting in an improved electron density map, which makes the model building step easier.

In the past, model building was done by making full use of a very expensive graphic work station, such as ones produced by Silicon Graphics. Additionally, the refinement calculation took more than one overnight period in most cases. However, nowadays, due to the increased speed of computers, structural analysis and model building can be carried out on a personal Windows or Mac computer, and refinement can be completed within minutes if things go well. If high-quality, high-resolution data can be obtained, most of the molecular model can be built automatically, thanks to advances in crystallographic software. However, if data resolution necessary for automatic model building cannot be obtained, it is still necessary to manually add amino acid residues one by one to build the model.

Here, construction and refinement of the model in X-ray structural analysis will be explained.

2. Density modification methods

In the past, even when initial phases were finally determined, they were often of insufficient accuracy. There was so much noise in the electron density map obtained using them that, in many cases, interpretation of it to build a model was often impossible. In these cases, to increase the accuracy of the initial phases, it was necessary to collect additional data on an isomorphous crystal substituted with a different heavy atom. However, nowadays, thanks to software development, the noise in initial electron density maps can be reduced using electron density modification methods such as "solvent flattening" and "histogram matching," improving the electron density map to the point where it is interpretable (Fig. 1). This means that there is less need for additional experiments, leading to acceleration in the structural analysis process. "Solvent flattening" and "histogram matching" used mainly in the electron density modification method will be described below. Protein crystals include very large amount of solvents. Because many of these solvent molecules remain dynamic even in crystals, the electron density in the solvent region is low and basically flat. On the other hand, in the protein region the atoms are localized so that the electron density is relatively high, and thus it is not flat. However, when the quality of the data is not good, the quality of the electron density map also decreases, resulting in the formation of peaks because various densities are included in the solvent region, which is inherently flat. In solvent flattening, the electron density in the solvent region is forcibly made flat so as to remove the peaks. As a result, the phases are improved so that the boundary between the solvent region and the protein region becomes clearer and, thus, the electron density map can be improved so that the model may be constructed with it.

Histograms of the electron density distribution of proteins whose structures have been analyzed appear almost the same. The histogram of a noisy electron density distribution shows Gaussian (normal) distribution, whereas a histogram of the electron density distribution calculated from high-quality phases gives a sharp peak having a left-right asymmetrical distribution. Histogram matching is a method where the histogram of the electron density distribution presently obtained is forcibly modified to an ideal histogram to improve the electron density distribution.

Because both methods are easy to implement and have significant positive effects on the electron density map in many instances, routine uses of them is recommended.

3. Model building

When the phases are determined and the electron density modification methods have been carried out, an interpretable electron density map can be obtained. Using the electron density map thus obtained, the molecule is constructed in accordance with the amino acid sequence of the target protein. When

* Application Laboratories, Rigaku Corporation.

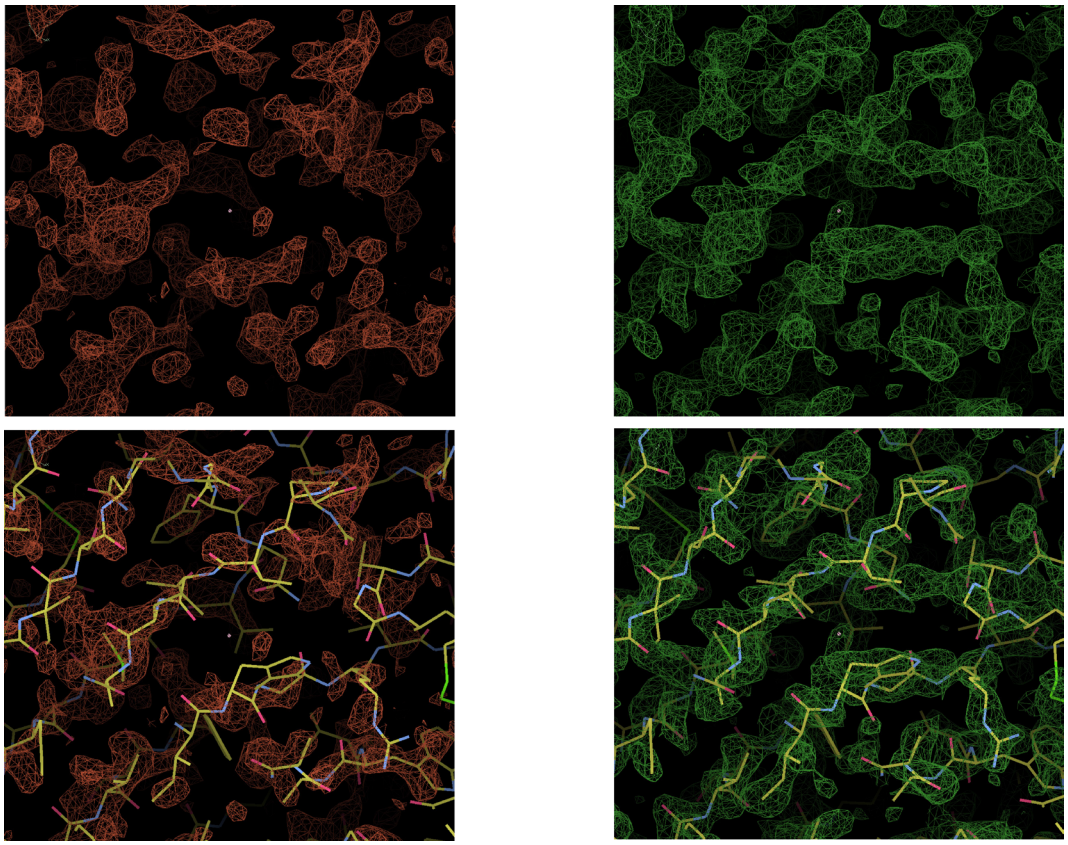

Fig. 1. Comparison of the electron density map before (upper left) and after (upper right) electron density modification; and superposition of the model structure onto each electron density map (lower left: before modification, and lower right: after modification).

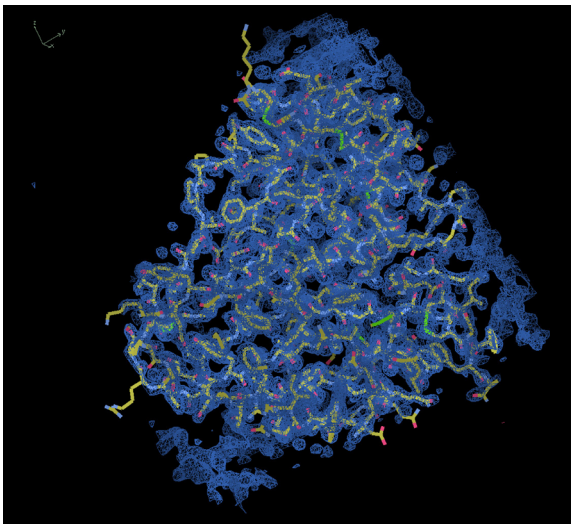

Fig. 2. Thaumatin after automatic construction with Buccaneer, wherein 80% or more of it is automatically constructed.

high-resolution electron density (about 2.0 Å or better) is obtained, interpretation of the electron density map to construct the molecule is automatically done by the programs like ARP/wARP[1], RESOLVE[2], and Buccaneer[3], [4] in accordance with the known amino acid sequence. When high-resolution and high-quality electron density is available, about 80% or more of the molecular model can be automatically constructed (Fig. 2). Especially with Buccaneer, a few successful examples of automatic construction can be seen even in an electron density map where the data are of poor quality with a lot of noise and with comparatively low resolution (about 2.5 to 3.0 Å). Below, the automatic construction steps using Buccaneer are described.

1. Candidate Cα positions are explored.
2. Fragment chain candidates are constructed by connecting Cα.
3. The fragment chain candidates are merged.
4. The N-terminal and C-terminal of the fragment obtained by the merge are correlated, thereby connecting it to a larger fragment having greater continuity.
5. Amino acid chains are assigned.
6. Correct amino acid chain is inserted and removed.
7. Removal of amino acid residues of steric hindrance.
8. Reconstruction is carried out.

In Buccaneer, the process from exploration of Cα to construction is carried out step-by-step, and the characteristics of the electron density of the main chain calculated from a known protein structure are explored from the electron density of the main chain of the target protein, thereby obtaining the electron density map of a probable main chain. By so doing, it is presumed that the success rate of automatic model construction including a turn region and a loop region is high even in a low-resolution electron density map.

In Fig. 3, a section of the electron density of lysozyme at different resolutions is shown. As the resolution

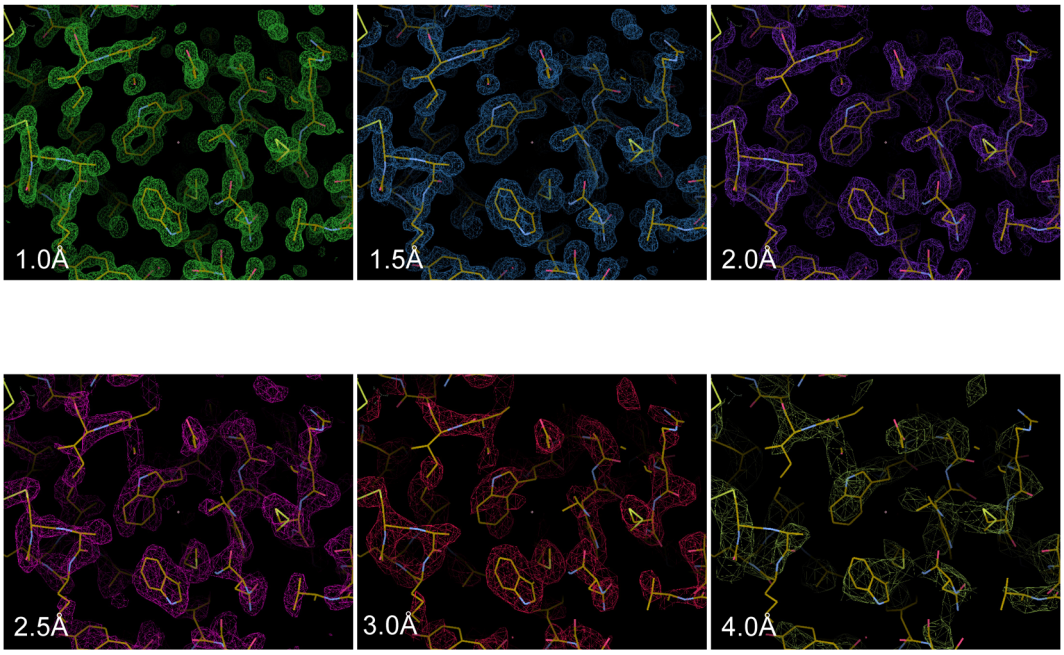

Fig. 3. Electron density of lysozyme at different resolutions.

decreases, the electron density of the side chains becomes increasingly unclear, finally leading to fragmentation even in the electron density of the main chain, indicating that the quality of the electron density is significantly decreased. Interpretation of the electron density at low resolution is difficult and, thus, it is easy to see why model construction with this type of density is difficult.

For reference, a conventional manual procedure for model building is as follows. First, as a foothold for model building, a part of the electron density of a characteristic sequence portion (a region where there are numerous amino acids having large side chains, such as tryptophan (Trp), phenylalanine (Phe), and tyrosine (Tyr)) is searched. On the basis of the characteristic electron density, the direction of the N-terminal is determined, and then an amino acid is assigned. In regions where the side chains are unclear, alanine (Ala) is temporarily assigned, followed by tracing Cα. When the side chain can be seen as refinement progresses, the temporary Ala assignments are replaced by the true amino acid one by one. Therefore, manual model construction consists of steps requiring exceptional patience. Accordingly, nowadays it is general practice that in the beginning, automatic model building is carried out to construct a fragment structure. Then, construction and bonding of the fragments are manually carried out for the rest of it.

4. Refinement

When model building is finished, the refinement step follows. When the molecular replacement method is used in phase determination, model building is not necessary; thus, the refinement can be carried out after phase determination.

Because the structure factor amplitudes (Fc) can be calculated from the molecular model, the coordinate and temperature factor are modified using the least-squares method or the like so as to make Fc consistent with the measured structure factor amplitudes (Fo). On the basis of the electron density map newly calculated using phases that are improved by the refinement, the model structure is fitted. By repeating structure refinement and model fitting, a more probable structure is constructed. This step is called refinement.

Besides the least-square method for refinement, there is also the simulated annealing (SA) method, in which the molecular dynamic method is utilized. In this method, the model at a hypothetically high temperature (about 2,500 K) escapes from any local minima by gradually lowering the temperature to room temperature, thereby converging to a more stable structure (global minimum). Because the convergence radius of the refinement in the SA method is larger than that with the least-squares method, even if the initial model structure includes significant errors, these errors can be corrected in the course of refinement; however, the calculation time is substantially longer than the least-squares method. The SA method can be used with CNS[5] or Phenix[6].

There has been a revolution, too, in software for refinement that uses low-resolution data. In the past, there has been a problem when refinement is carried out using low-resolution data (about 4 Å) because the ratio of the number of reflections to the number of refinement parameters is low. Model construction is excessively advanced (over-modelling) to lower R_{work}, thereby leading to an increase in the difference with R_{free} (for R_{work} and R_{free}, see "5. Assessment of the Results"). In order to solve this problem, in Refmac 5[7], the refinement program is provided with "jelly body refinement," the mode for low resolution (Fig.

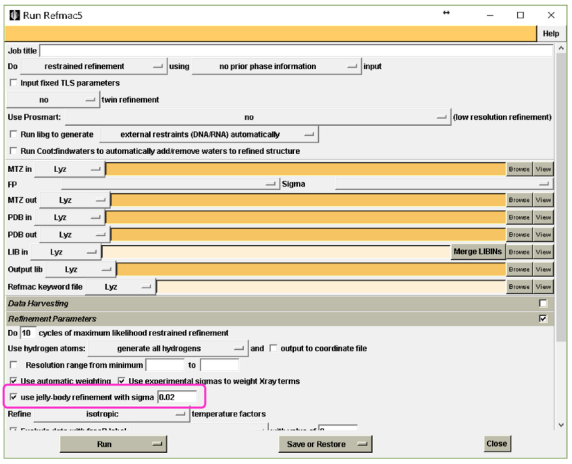

Fig. 4. Refinement mode for low-resolution, "jelly body refinement", of Refmac5.

4). With "jelly body" refinement, more restrictions/restraints are used than for ordinary refinement to avoid the divergence of the values of R_{work} and R_{free}.

5. Evaluation of the results

For assessment of the refinement results, in general, R_{work} and R_{free} are used. R_{work} is an indicator to evaluate the validity of the model structure. R_{free} is the R value calculated using only the test reflections not used in the refinement (5% of the independent reflection is used in most cases). R_{free} indicates whether or not the molecular model is biased by overfitting in the course of the refinement. When the refinement is properly carried out, both R_{work} and R_{free} decrease, with R_{free} having a slightly higher value (higher than R_{work} by about 5% in many cases). If the model is wrongly assigned to the electron density derived from noise, or overfitting is made due to low resolution data, R_{free} does not decrease but only R_{work} decreases, leading to significant divergence between R_{work} and R_{free} (10% or more). If a situation like this occurs, the entire structure needs to be precisely studied in order to correct mistakes in the model.

6. Conclusion

Thanks to the increased speed of PCs, it is now possible to perform the model building and refinement steps of structural analysis even with a laptop computer. Further, because of software advances, by using the electron density modification method, a sufficiently interpretable electron density map can be obtained. In addition, by effectively utilizing automatic model construction, a significant reduction in the time consumed for the refinement can also be realized. This means that competitors studying a similar protein as the target are also enjoy the increased speed of the structural analysis, and the number of competitors is increasing at the same time. Therefore, competition appears to be much fiercer than before. The key to winning this competition depends on collecting high-resolution and high-quality data. For this, if the measurement can be conveniently done with a familiar in-house system, waiting times and travel costs can be significantly reduced. In the system mounted with a high-speed and highly sensitive Hybrid Photon Counting detector represented by HyPix-6000HE (Rigaku) and PILATUS (DECTRIS) (Fig. 5), measurement time is decreased and the protein size and quality of the protein crystal that can be dealt with is dramatically improved.

Fig. 5. Hybrid Photon Counting detectors: HyPix-6000HE (left) and PILATUS 200K (right).

To obtain high-quality data, a high-quality crystal is absolutely necessary. Thus it is important to crystallize a high-purity protein from a solution suitable for crystallization. For in-solution assessment of whether a protein is suitable for crystallization, BioSAXS-2000, which was introduced in the 11th part of this basic course on single crystal X-ray structural analysis, is suitable.

References

(1) G. Langer, S. X. Cohen, V. S. Lamzin and A. Perrakis: *Nature Protocols.*, **3** (2008), 1171−1179.
(2) T. C. Terwilliger: *Acta Cryst.*, **D59** (2003), 38−44.
(3) K. Cowtan: *Acta Cryst.*, **D62** (2006), 1002−1011.
(4) K. Cowtan: *Acta Cryst.*, **D64** (2008), 83−89.
(5) A. T. Brunger, P. D. Adams, G. M. Clore, P. Gros, R. W. Grosse-Kunstleve, J.-S. Jiang, J. Kuszewski, N. Nilges, N. S. Pannu, R. J. Read, L. M. Rice, T. Simonson and G. L. Warren: *Acta Cryst.*, **D54** (1998), 905−921.
(6) P. D. Adams, P. V. Afonine, G. Bunkoczi, V. B. Chen, I. W. Davis, N. Echols, J. J. Headd, L. W. Hung, G. J. Kapral, R. W. Grosse-Kunstleve, A. J. McCoy, N. W. Moriarty, R. Oeffner, R. J. Read, D. C. Richardson, J. S. Richardson, T. C. Terwilliger and P. H. Zwart: *Acta Cryst.*, **D66** (2010), 213−221.
(7) G. N. Murshudov, A. A. Vagin and E. J. Dodson: *Acta Cryst.*, **D53** (1997), 240−255.