Introduction to single crystal X-ray analysis

IX. Protein structure analysis and small molecule structure analysis

Akihito Yamano[*]

1. Introduction

The previous series have discussed single crystal X-ray analysis of small molecules. This series will discuss structure analysis of proteins using X-ray diffraction. The explanation will focus in particular on differences between structure analysis of proteins and that of small molecules.

In structure analysis of proteins or small molecules, the purpose of single crystal X-ray analysis is to determine molecular structure and the basic principle is the same. However, there are considerable differences in the method pursuing each stage of structure analysis. These differences can be attributed primarily to differences in the size of the molecules. Whereas small molecules normally have a molecular weight ranging from a few hundred to thousand daltons, while that of proteins ranges from a few thousand to million daltons. There also are large differences in the components of the molecules and their chemical degrees of freedom. The atomic species comprising small molecules, and their framework, are diverse and there are no limitations in principle, however proteins are basically made up of 20 amino acids of known structure. Furthermore, the proteins of any living organisms on the earth are all comprised of L-amino acids alone, with just a few exceptions, and thus their absolute configurations are known. Also, an α-helix, which is one distinctive structure of proteins, is always a right-handed helix when tracing from the N-terminus to the C-terminus.

Despite the large amount of structure-related information which exists beforehand as described above, it is not unusual for structure analysis of a new protein to take a few years, including the stages of expression and purification of the target molecule. Why is this troublesome task carried out? The reason is that various benefits can be obtained once the structure of a protein has been determined.

2. What is learned in structure analysis of a protein?

Although it is not clear how long it will take to reach the final goal, humanity seems to be moving toward a complete understanding of the phenomenon of life at the atomic level. What is important in understanding the phenomenon of life is to understand metabolism and biological reactions at the atomic level, and this cannot be achieved without elucidating the atomic-level structure of the proteins which are the key players in living organisms. Once a specific metabolic pathway has been determined, the reaction can be controlled by creating compounds which act on the involved proteins.

For example, if research elucidates the structure of a protein which is the key to the life cycle of a virus, then it is possible to create an anti-viral drug. The method of developing drugs based on the three-dimensional structure of the target protein, obtained typically via X-ray structure analysis, is called Structure Based Drug Design (SBDD). An example of a remarkable success in the early stages of SBDD was the development of drugs that block the HIV protease. In the final stage of replicating itself, HIV protease cuts a protein produced all at once by the host cell to create its components. Since HIV protease governs this critical reaction, when the protease is blocked, HIV cannot to replicate itself and eventually perishes, even if it succeeds infecting to a host. Today, SBDD has progressed further, and numerous drugs with higher specificity have been appeared. These drugs block a DNA–RNA reverse transcriptase specific only to retroviruses such as HIV, and the integrase which catalyzes the integration of HIV DNA into the host's DNA.

A more familiar example is the case of headache medications. Figure 1 shows the mechanism whereby a headache occurs when a person drinks excessively. The ingested alcohol is broken down into acetaldehyde by alcohol dehydrogenase. Acetaldehyde is converted to acetyl-CoA by acetaldehyde dehydrogenase, which then enters the TCA cycle. The TCA cycle is the metabolic pathway plays the central role in the glycolytic pathway, and this pathway produces ATPs, one of the main energy carriers inside the body. There is no problem when all acetaldehyde are broken down by acetaldehyde dehydrogenase, however when the breakdown cannot keep up with the speed of alcohol ingestion, a headache will occur. Acetaldehyde has the effect of relaxing the smooth muscle of blood vessels, and inflammation occurs when blood vessels expand, then bradykinin is produced. Bradykinin activates phospholipase A_2, and the activated phospholipase A_2 breaks down fatty acids, and turns them into arachidonic acid. Arachidonic acid is then converted into prostaglandin (PGE_2) by cyclooxygenase. This PGE_2 is an algesic substance resulting headache. If a molecule which inhibits phospholipase A_2 or cyclooxygenase is

[*] Application Laboratory, Rigaku Corporation.

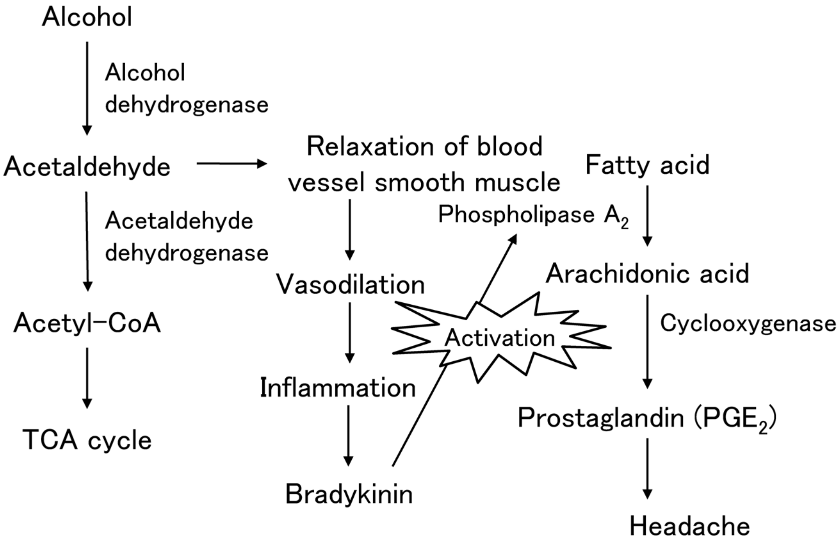

Fig. 1. Mechanism of headache occurrence after excessive ingestion of alcohol. If the activity of phospholipase A_2 or cyclooxygenase is inhibited, the pain producing substance will no longer be produced. Thus a pain-relieving efficacy can be expected.

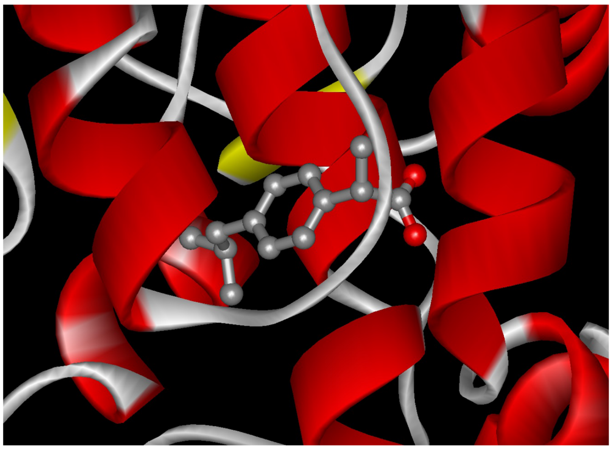

Fig. 2. Cocrystal structure of cyclooxygenase and the cyclooxygenase inhibitor ibuprofen. Some commercially available pain relievers contain ibuprofen as their effective ingredient.

designed, it can be a pain killer. In the initial stage of searching for drug candidate, scaffolds generally include compounds selected using techniques such as computer-based docking calculation referencing to compound libraries; comparison with existing drugs; and natural materials contained in medicinal herbs and the like. X-ray structure analysis is performed by bonding these compounds to target proteins, and then the structure is optimized to improve bonding strength. Figure 2 shows the cocrystal structure of cyclooxygenase and the pain-reliever ibuprofen in the PDB (Protein Data Bank). Ibuprofen is already commercially available as the pain-relieving ingredient in cold medications, and it is known to bond efficiently to the active sites of cyclooxygenase.

3. What is the difference between protein structure analysis and small molecule structure analysis?

Structure analysis of proteins and small molecules is the same in terms of the purpose, basic principle, and analysis procedure, but there are major differences in the specifics of each stage of analysis. The following section provides a review of the small molecule analysis procedure.

3.1. Procedure for small molecule structure analysis

Small molecule compounds are ordinarily prepared through chemical synthesis. Then the compound is crystallized. Crystallization is performed mainly by using a technique called the batch method. The compound is dissolved in a solvent, and then the solvent is gradually evaporated, thereby producing a supersaturated solution. Whether or not crystallization succeeds depends on the type of solvent, the concentration speed, and the temperature.

When crystals are obtained successfully, then diffraction intensity is measured using an X-ray diffractometer. With semiconductor detectors such as the PILATUS 200K, measurement is typically completed in a few minutes to a few hours because the PILATUS 200K has a short readout time of 7 ms, and it enables true shutterless data collection.

In order to calculate electron density, it is necessary to estimate initial phases. The direct method is mostly used in small molecule structural analysis. The theory of the direct method is difficult to understand, but it has been implemented as a program, and thus executing it is extremely easy. By simply providing reflection intensity and crystallographic information, the program calculates initial phases. A wide variety of such programs are available, and can be obtained either free or for a comparatively low price. When phases have been determined, it is possible to calculate electron density. With the direct method, normalized structure factors represented as "E" are used, and thus a list of peak positions in the E-map is output when the calculation is finished. A molecular model is built assuming that atoms are present at the peak positions in the E-map. In general, precision of initial phases is relatively low, and therefore at the beginning, parts having higher electron density such as metal atoms, five-membered rings and six-membered rings are readily observed. Even if a part of the molecule is not observed, one should proceed to the least-squares refinement, and calculate a difference Fourier map taking the residual |Fo−Fc| as coefficients. At this time, the phase precision is already improved over the initial phase, and thus the remaining structure which could not be observed at the initial stage appears. Each stage of single crystal X-ray analysis can be regarded as a step of improving phases to approach the true phases.

3.2. Protein structure analysis

In protein structure analysis, the basic stages of analysis are the same as in small molecule structure analysis. However, there are major differences in the execution of each stage.

3.2.1. Sample acquisition

Small molecules are primarily prepared using chemical synthesis. Proteins, on the other hand, are produced through overexpression using genetic engineering techniques, and purified using techniques such as solid metal ion affinity chromatography, affinity chromatography using antigen-antibody reactions, and gel filtration chromatography in which molecules are separated by size. To make separation and purification easier using these liquid chromatography methods, the mainstream approach is to express as a chimeric protein linked beforehand with another protein, or as protein with histidine chain tags on the N-terminal and/or C-terminal. An expression system most suited to the purpose and conditions is chosen from various options such as *E. coli*, insect cells, and human cells.

3.2.2. Crystallization

A variety of methods have been devised for protein crystallization, ranging from conventional methods to the latest techniques using electronic elements and magnetic fields. More than enough research results have been reported on crystallization techniques themselves. However, the most important point for protein crystallization is not the crystallization technique itself; rather, it is preparation of easily crystallized protein. At the stage where protein is overexpressed, various constructs are prepared incorporating modifications at sites expected to be crucial for crystallization, or modifications for limiting relative movement between domains. Then the construct most suitable for crystallization is selected.

The most popular crystallization method is the vapor diffusion technique (Fig. 3(a)). First a precipitant is added to the same buffer solution used to dissolve the protein. A sufficient amount of reservoir solution is added to the crystallization well. The same amount of reservoir solution is added to the drop of protein solution on a cover slip, and it is turned upside down and shielded. The concentration of precipitant in the crystallization drop is nearly one-half of that of reservoir, and therefore water migrates to reservoir until it becomes at the same concentration as the reservoir solution. Concentration and super saturation occur due to this process, and crystal is obtained.

At present, progress is being made with the hardware and software needed for data collection and phase determination, and, as in the case of small molecule structure analysis, if a good crystal can be obtained, then it is possible to achieve the final structure with

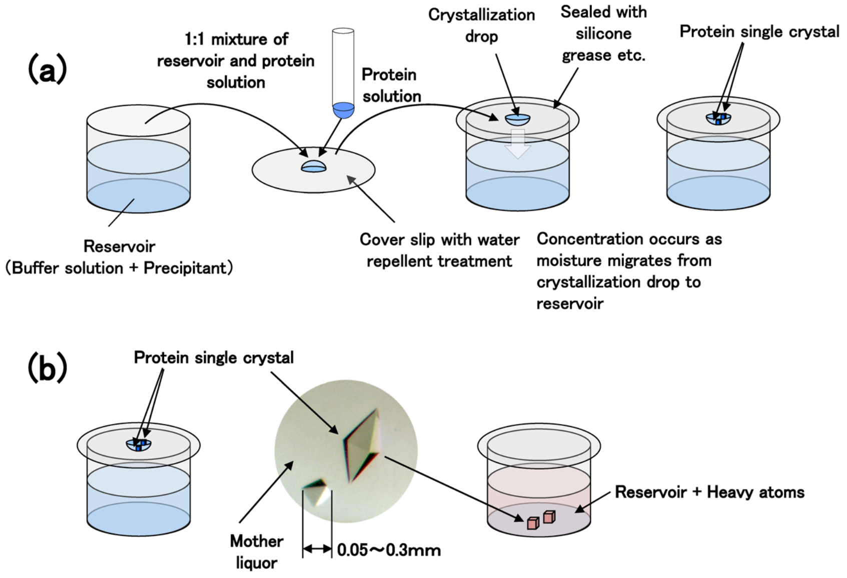
(a)
1:1 mixture of reservoir and protein solution
Protein solution
Crystallization drop
Sealed with silicone grease etc.
Protein single crystal
Reservoir (Buffer solution + Precipitant)
Cover slip with water repellent treatment
Concentration occurs as moisture migrates from crystallization drop to reservoir
(b)
Protein single crystal
Reservoir + Heavy atoms
Mother liquor
0.05～0.3mm

Fig. 3. (a) Set-up for crystallization using the hanging drop vapor diffusion method. (b) Heavy atom soaking method.

a far greater probability than before. Therefore, it is no exaggeration to say that the major determinant of success in protein structure analysis is how biochemical experiments are carried out.

3.2.3. Data collection

In small molecule structure analysis, data collection is often done at low temperature for reasons such as preventing evaporation of crystallization solvent, suppressing thermal vibration of atoms, and suppressing damage to crystals. In protein structure analysis, on the other hand, measurement is almost always performed at a low temperature around 100 K. This is because protein crystals contain typically 50% solvent, that is water, and they are also more easily damaged by X-rays. A cryoprotectant must be added to perform low-temperature measurement. The problem is mother liquor for crystallization picked up with a crystal by a loop. The mother liquor is buffer solution whose main component is water, therefore when it is frozen as it is, it forms hexagonal ice and swells to degrade the crystal unless it is frozen extremely quickly. Initially, it was thought that crystals are damaged during freezing because the water inside protein crystals freezes and forms ice, however water molecules inside crystals are stabilized by the protein itself, and therefore ice formation due to freezing is expected to be limited.

Ordinary freezing is performed with a cold nitrogen gas stream of a low-temperature system. The low-temperature gas stream is first blocked with a card or similar item, and after setting the crystal in place, the card is quickly removed, thereby increasing the freezing speed. This is called the flash freezing method. There is also a method in which the loop used to scoop up the crystal is quickly dipped into low-temperature liquid such as liquid nitrogen contained in a Dewar. Compared with gas, a liquid medium is known to have better heat conduction, but if freezing is not done with a great care, it may have the opposite effect to what was intended. The most common cause of failure is the nitrogen gas layer present on top of liquid nitrogen. Since the gas layer has a comparatively mild temperature gradient, it impedes rapid freezing of the crystal. This can be avoided in various ways, such as filling the Dewar up to the rim with liquid nitrogen, or blowing away the gas layer using some methods(1).

3.2.4. Phase determination

In small molecule structure analysis, initial phases are determined through calculation, almost without exception. In protein structure analysis, on the other hand, it is not practical yet to determine phases using the direct method, but usually determined experimentally. A typical technique is the multiple isomorphous replacement (MIR) method. In the MIR method, the phase of the original (native) protein is derived by using the slight discrepancy in phase which arises when heavy atoms are introduced. Theoretically, it is possible to calculate the phase of the native crystal if there are a total of three sets of data: for the native crystal, and a minimum of two different types of heavy atom derivatives.

There are a number of methods for preparing heavy atom derivative crystals, but the classical method involves soaking the native crystal in heavy atom solution. This is called the soaking method (Fig. 3(b)). At first glance this looks easy, but there are many parameters to be explored such as the heavy atom type, concentration, and soaking time. When conditions are good, heavy atoms bond at specific sites of the protein molecule while crystallinity is maintained. To determine whether or not heavy atoms have been

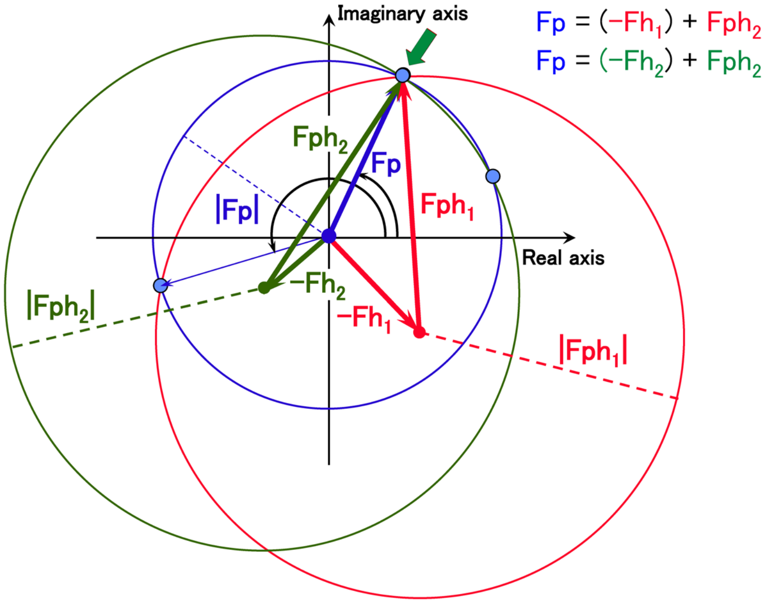

Fig. 4. Principle of phase determination using the multiple isomorphous replacement method. Blue, red and green are obtained, respectively, from the native crystal, first heavy atom derivative, and second heavy atom derivative.

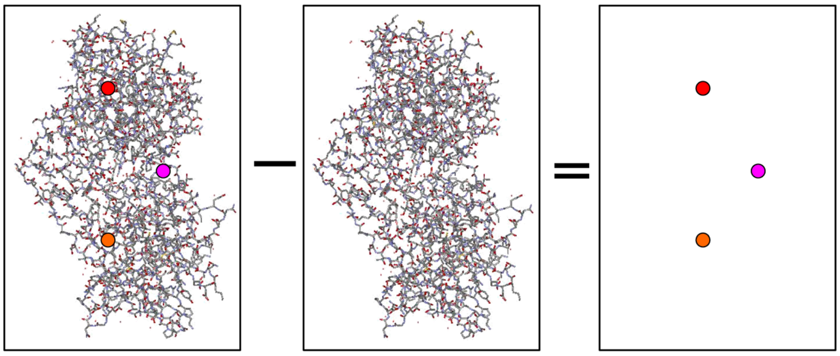

Fig. 5. Principle of determining structure of heavy atoms using the direct method. When the native protein structure is subtracted from that of the heavy atom derivative, only the structure of heavy atoms remains. The structure of the heavy atoms alone is comparatively simple, therefore it is possible to identify the positions of heavy atoms using the direct method.

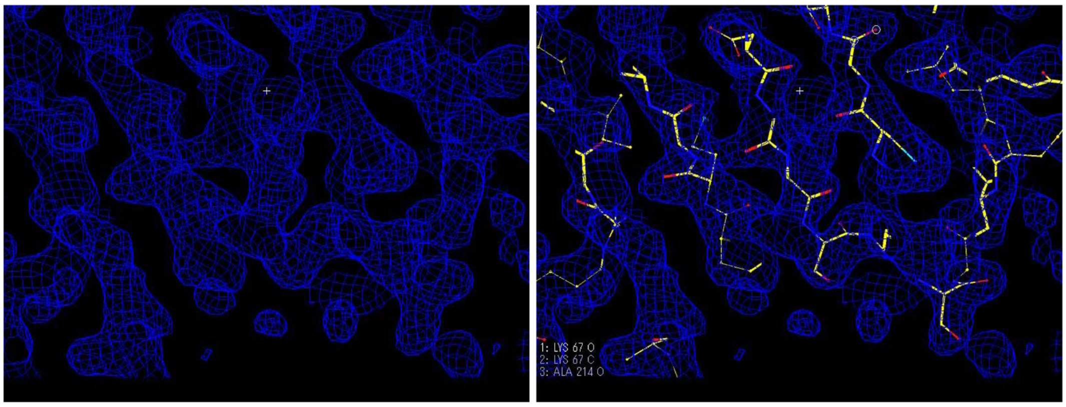

Fig. 6. Left: Electron density diagram after solvent flattening. Right: Electron density with molecular model created using automatic model building program.

introduced successfully, there is no other way but actually measuring data, and therefore a large amount of work is necessary to obtain even one derivative.

Figure 4 shows the principle of phase determination using the MIR method. The structure factors can be thought of as waves, and thus it is convenient to think on the Gaussian plane. Information from the native crystal is indicated in blue. The reflection intensity of the native crystal can be determined through data collection, but the phase cannot be measured directly. This corresponds to a situation where the lengths of the structure factor vectors are known, but their directions are unknown. It is here that heavy atom derivatives come into play.

First, there is a need to find the structure factors of heavy atoms alone. If the isomorphism of the heavy atom derivative is sufficiently high, then the residual subtracting the intensity of the native crystal |Fp| from that of heavy atom derivative $|Fph_1|$ can be regarded as the intensity of heavy atoms alone. In real space, this indicates that it is enough to determine the structure of the heavy atoms only (Fig. 5). The structure of the heavy atoms alone is simple, therefore it should be possible to determine the structure somehow. Previously, positions of the heavy atoms were determined using the Patterson method, but today it is typically done using the direct method.

Once the positions of the heavy atoms are determined, then it is possible to calculate the portion of the structure factors attributable to the heavy atoms. Since structure factors are expressed as vectors, they can be written into the diagram shown as Fh_1 in Fig. 4. However, only the reflection intensity is known for the protein part of the heavy atom derivative, therefore a circle is drawn with radius equal to the amplitude of the structure factor whose center is the end point of the structure factor of the heavy atom (red circle in Fig. 4). At the intersections of the red and blue circles from the derivate and native crystal, the vector relations are satisfied for the heavy atom derivative, native crystal, and heavy atom only. However, there are still two intersection points, hence two possibilities for the phase angle remain. Thus a second heavy atom derivative is prepared, and the same procedure is repeated. For the second derivative, the positions of the heavy atom is determined and the portion of the structure factor attributable to the heavy atom is calculated (Fh_2 in Fig. 4). When a circle is drawn with radius equal to the amplitude of the second heavy atom derivative centered at the end point of Fh_2, then in this case again there are two intersections, however only one of them is consistent with one of the

Table 1. Summary of small molecule structure analysis and protein structure analysis.

	Small molecule structure analysis	Protein structure analysis
Sample preparation	Chemical synthesis.	Biochemical expression and refinement using genetic engineering.
Crystallization	Batch method dissolving in organic solvent and concentrating, or diffusing poor solvent.	Vapor diffusion method. Concentration occurs by adding precipitant to protein dissolved in buffer solution with optimal pH for the individual proteins.
Data gathering	Mo or Cu radiation is used depending on the magnitude of absorption or necessity of absolute structure determination. There are cases where the oscillation angle during measurement is large, and cases where it is small.	No need to determine absolute structure. Cu radiation is always used with the laboratory system to raise the intensity and to resolve neighboring reflections originated from long lattices. The oscillation angle is normally 1° or less.
Phase determination	Calculation by computer using statistical techniques. Direct method.	In general, the direct method cannot be used. Determination is done experimentally by the MIR method etc.
Model building	Building is carried out based on expected structure, chemical knowledge, bonding distance and electron density.	Components are limited to 20 types of amino acids, and thus the basic approach is to fit amino acid chains. There are multiple automatic building programs.
Structure refinement	At present, this is always done using the least-squares method. Shelxl is the de facto standard.	The post popular technique is the SA method. The least-squares method is also used. Refmac, Shelxl, etc.
Evaluation/Utilization of results	Utilization of molecular structure, absolute structure and crystal packing, etc.	Primary the molecular structure. It is used for drug development and understanding biological reactions.

two intersections obtained with the first derivative (thick arrow in Fig. 4). This makes it possible to determine which of the two phase angle possibilities is correct, and thus to determine the phase of the protein crystal.

3.2.5. Molecular model building

If phases can be calculated, so can the electron density. The electron density calculated with the initial phases usually has unrealistic natures such as negative electron density. In order to eliminate these regions, a procedure called solvent flattening is performed in the solvent region. Phases can be improved by carrying out a Fourier transform of the electron density map modified in real space via solvent flattening.

The electron density map after solvent flattening is shown in Fig. 6. When high-quality electron density map to this extent has been obtained, fitting of amino acid chains is performed while looking at the shape of the electron density. Good clues are provided by amino acid side chains with distinctive shape such as tryptophan, tyrosine, and phenylalanine. Even with a novel protein, it is common for the amino acid sequence to be known prior to structure analysis. Therefore, if there are some sites where the allocation is certain, the remaining fitting can be proceeded according to the amino acid sequence. At present, multiple model building software packages are available (Buccaneer(2), ARP/wARP(3), Solve/Resolve(4),(5)), and these build a molecular model automatically referring to the given amino acid sequence. The advent of molecular model building programs has dramatically reduced the time needed for protein structure analysis. Protein may be better suited to automation because the number of building components is limited to 20 amino acids.

3.2.6. Structure refinement

When the molecular model is finished, the next step is structure refinement. The simulated annealing (SA) method is a characteristic technique for protein structure analysis. In the SA method, energy is first minimized, and then the molecule temperature is virtually increased by connecting a heat bath. After that, gradual cooling is performed by separating the heat bath. This method has a larger convergence radius than the least-squares method, and it is thought that it converges on the correct structure by overcoming higher energy barriers. CNS/CNX is a typical program for the SA method. The least-squares method is also used to refine protein structure. The CCP4 basic software package for protein structure analysis includes software for the least-squares method called Refmac(6). SHELXL (the standard refinement program for small molecule structure analysis) is also has a capability of protein structure refinement. With either the SA method or least-squares method, the number of reflections to parameter ratio is lower in protein structure analysis, and hence refinement is executed while providing constraints such as amino acid structure retains its shape.

In protein structure analysis, water molecules behaving as part of the protein are occasionally found. There also are water molecules creating a hydrogen bond network covering the molecular surface and contributing to solubilization. Normally, water in the first layer is localized, and this too is incorporated into the molecular model. The final crystallographic R value is generally in the range 10–30%.

4. Conclusion

Table 1 summarizes each stage of protein structure analysis and small molecule structure analysis. The steps are common, but the way those steps are executed are markedly different. Difficulties in small molecule structure analysis are almost always attributable to crystallographic problems besides obtaining crystals. The most common are cases where analysis is performed without noticing twinning. On the other hand, the difficulty of protein structure analysis is primarily due to the difficulty of experiments.

References

(1) M. Warkentin, V. Berejnov, N. S. Husseini and R. E. Thorne: *J. Appl. Crystallogr.*, **39** (2006), 805–811.

(2) M. D. Winn *et al.*: *Acta Cryst.* **D67** (2011), 235–242.

(3) G. Langer, S. X. Cohen, V. S. Lamzin and A. Perrakis: *Nature Protocols*, **3** (2008), 1171–1179.

(4) T. C. Terwilliger: *Acta Cryst.*, **D59** (2003), 38–44.

(5) T. C. Terwilliger: *Acta Cryst.*, **D59** (2003), 45–49.

(6) G. N. Murshudov, A. A. Vagin and E. J. Dodson: *Acta Cryst.*, **D53** (1997), 240–255.