High-resolution X-ray crystal structure analysis of bovine H-protein and the treatment of diffraction data

(Bovine 由来 **H-protein** の高分解能X線結晶構造解析と
その回折強度データの取り扱い**)**

A Doctoral Thesis

By

Akifumi Higashiura

Submitted to the Graduate School of Science

Osaka University

Japan

April, 2010

Acknowledgements

This study has been carried out under the direction of Professor Atsushi Nakagawa of Institute for Protein Research, Osaka University. I would like to thank him for his incessant guidance and encouragement throughout this work.

I am deeply grateful to Dr. Koji Inaka of Maruwa Foods and Biosciences, for his kind and useful discussion, and also to Mr. Masaru Sato of Japan Aerospace Exploration Agency and Mr. Hiroaki Tanaka of Confocal Science, for their discussion and advice in high-resolution X-ray crystallography.

I would like to thank Dr. Harumi Hosaka and Mr. Makoto Matsuda, Institute for Protein Research, Osaka University for their kindly advices and initial experiments of bovine H-protein. I deeply thank Mr. Takeshi Kurakane of Institute for Protein Research, Osaka University for his useful help in the preparation, crystallization and data collection of bovine H-protein.

I wish sincere thanks to Associate Professor Mamoru Suzuki of Institute for Protein Research, Osaka University for his advices. I also express thanks to Dr. Yusuke Yamada and Dr. Masahiko Hiraki of High energy accelerator research organization for their useful help in data collection at Photon Factory.

I express my thanks to all members in Nakagawa's laboratory for kind assistances and encouragement. I also express my thanks to all of my friends for their warm encouragement.

Finally, I would like to thank my parents for their incessant understanding and encouragement.

Akifumi Higashiura

April, 2010

-Table of contents-

Abbreviations

ATP	Adenosine Triphosphate
AMP	Adenosine Monophosphate
BL	Beamline
CCD	Charge-Coupled Device
CCP4	Collaborative Computational Project, Number 4
cDNA	Deoxyribonucleic Acid
DEAE	Diethylaminoethyl
DTT	Dithiothreitol
FPLC	Fast Protein Liquid Chromatography
IPTG	Isopropyl-β-D-thiogalactopyranoside
KPB	Potassium Phosphate Buffer
MWCO	Molecular Weight Cut Off
PAGE	polyacrylamide gel electrophoresis
PDB	Protein Data Bank
PMSF	Phenylmethylsulfonyl Fluoride
PTFE	Polytetrafluoroehylene
RMS	Root Mean Square
SDS	sodium dodecyl sulfate
T.ma	*Thermotoga maritima*
THF	Tetrahydrofolate
Tris	Tris(hydroxymethyl)aminomethane
T.Th	*Thermus thermophilus*

Chapter 1
Introduction

1.1. High-resolution X-ray crystallography of macromolecules

Recently, macromolecular X-ray crystallography has been significantly advanced by high-brilliance and low-divergence synchrotron beams, high-performance and high-precision large area detectors, cryo-cooling techniques, state-of-the-art data reduction programs, and mathematical improvement of the refinement software. Improvement in the methods and techniques of protein crystallography has pushed up the resolution limit and, the quality of protein structures (Dauter *et al*., 1997; Schmidt & Lamzin, 2002; Dauter, 2003; Vrielink & Sampson, 2003; Petrova & Podjarny, 2004). As of February 2010, ~250 structures of which resolution are higher than 1.0 Å have been deposited against in the Protein Data Bank (PDB; Berman *et al*., 2000; Berman *et al*., 2002), out of ca.60,000 structures. Thus far, the protein structures that have been refined against the ultra-high resolution data (< 0.7 Å) include crambin (Jelsch *et al*., 2000), hen egg white lysozyme (Wang *et al*., 2006) and human aldose reductase (Howard *et al*., 2004), at resolutions of 0.54, 0.65 and 0.66 Å, respectively (Table 1.1). However, examples of high-resolution X-ray crystallography are still not numerous: only 92 structures (<0.15% of the total) are beyond 0.9 Å resolution (Figure 1-1).

The number of measurements (namely, diffraction data) required for structure determination at 0.9 Å resolution is about 1.4 times that of 1.0 Å. This increase in experimental data enables us to refine structural parameters more precisely and accurately. In high-resolution and well-refined structures, we can visualize multiple conformations of main and/or side chains, accurate solvent structures, and determine anisotropic temperature factors. Many hydrogen

atoms are visualized in hydrogen-omit maps generated using high-resolution data, and the coordinates of hydrogen atoms can also be refined at ultra-high resolution (<0.7 Å). Hydrogen atom positions are often important for understanding the function of enzymes (Vrielink & Sampson, 2003). As the resolution is improved, the number of X-ray diffraction data also increases. Hence, it is clear that high-resolution data have an advantage in the structural refinement of proteins. However, it is difficult to collect, scale and merge high-resolution data and to carry out refinement of structures with a large number of parameters.

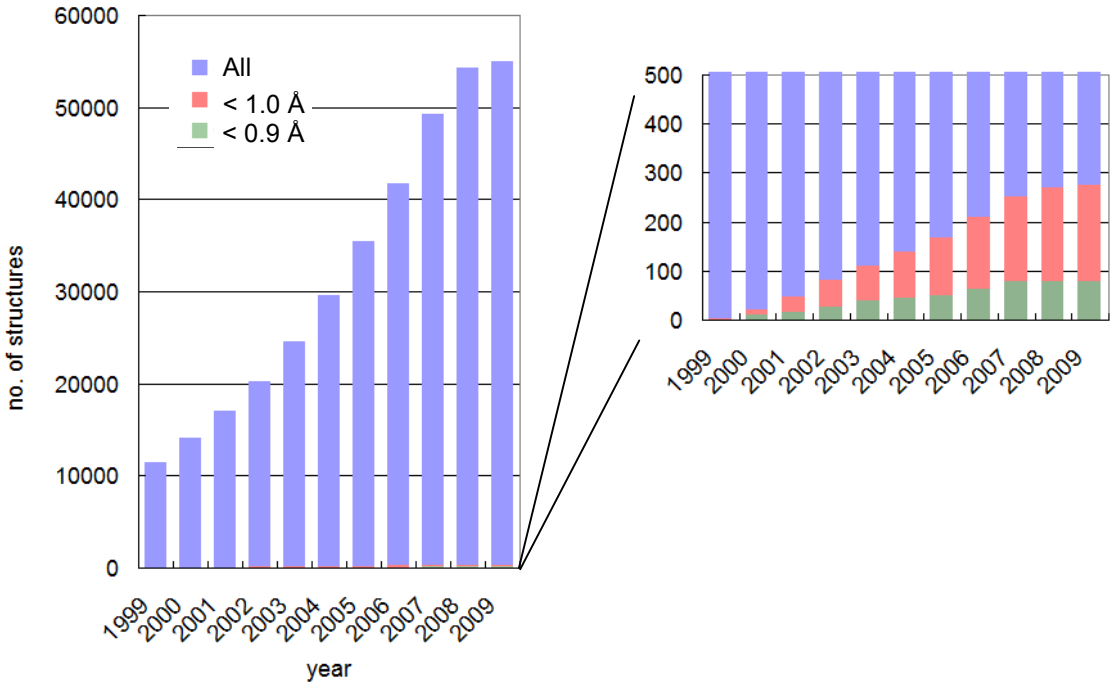


Figure 1.1 The number of X-ray crystal structures of proteins deposited in Protein Data Bank as of 2009.

The total number of structure is rapidly increasing, but the number of higher-resolution structure is not numerous.

Table 1.1 The list of top 10 highest-resolution X-ray crystal structures.

All the data were quoted from Protein Data Bank.

Protein	Resolution/Å	no. of residues	PDB ID	
Crambin	0.54	46	1ejg	Jelsch *et al.* (2000)
Antifreeze protein	0.62	64	1ucs	Ko *et al.* (2003)
Hen egg white lyzosyme	0.65	129	2vb1	Wang *et al.* (2007)
Aldose reductase	0.66	316	1us0	Howard *et al.* (2004)
Rubredoxin	0.69	54	1yk4	Bonisch *et al.*(2005)
PDZ2	0.73	76	1r6j	Kang *et al.* (2004)
Hydrophobin	0.75	70	2b97	Hakanpaa *et al.* (2006)
Amicyanin	0.75	105	2ov0	Carrell *et al.* (2007)
Serine protease	0.78	275	1gci	Kuhn *et al.* (1998)
PAK pilin	0.78	144	1x6z	Dunlop *et al.* (2005)

1.2. Bovine H-protein

The glycine cleavage system is a mitochondrial multi-enzyme system that consists of four different proteins (P, H, T and L-proteins). These proteins together catalyze the oxidative cleavage of glycine (Figure 1.3). Glycine cleavage system is widely distributed in animals, plants and bacteria. The H-protein is a monomeric protein with molecular weight of ~14 K, and plays a centric role in glycine cleavage. The lipoic acid prosthetic group covalently bonded to a specific lysine residue of the H-protein interacts with specific sites on the P, T and L-proteins (Kikuchi *et al.*, 2008). Non-ketotic hyperglycinemia, an inborn error of metabolism inherited as an autosomal recessive trait, is characterized by a massive accumulation of glycine in plasma and cerebrospinal fluids with severe neurological symptoms. It is caused by mutations in the genes

encoding the components of the glycine cleavage system. More than 85% of the patients are deficient in P-protein activity, whereas the remaining patients are deficient in T-protein activity (Tada & Kure, 1993).

The gene for bovine H-protein was isolated from bovine liver cDNA library (Fujiwara *et al.*, 1990), and the purified recombinant apo-H-protein was lipoylated ((1), (2), Figure 1.2) and activated *in vitro* by lipoyltransferase, using lipoyl-AMP as the lipoyl donor (Fujiwara *et al.*, 1992). The X-ray crystal structure of human T-protein was determined and the graphical coupling with pea H-protein was made and the complex model provides a reasonable picture of the catalytic process (Ikeda *et al.*, 2005). Bovine H-protein shows high level of homology with human H-protein, so the atomic structure of bovine H-protein may be possible to clarify the mechanism of the non-ketotic hyperglycinemia.

Lipoic acid + ATP -> Lipoyl-AMP + PPi (1)

Lipoyl-AMP + apo H-protein -> holo H-protein + AMP (2)

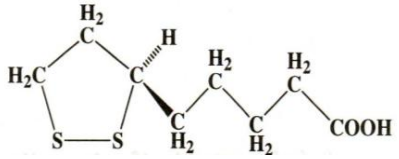


Figure1.2. The chemical structure of R-(+) lipoic acid.

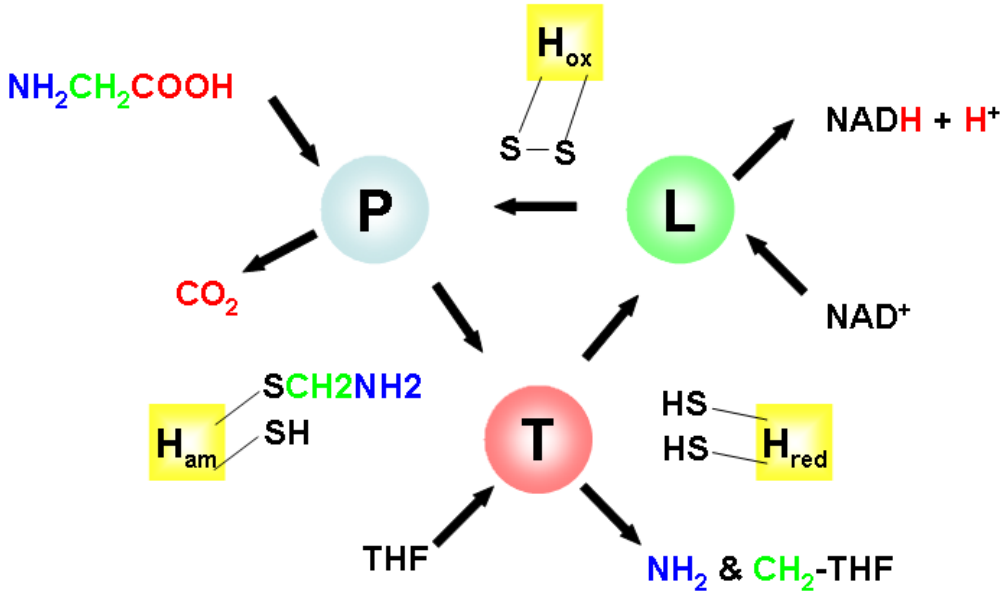


Figure1.3. Outline of the glycine cleavage system.

The figure is adapted from Douce *et al* (2001), and Nakai *et al.* (2005). The glycine cleavage system catalyzes the reversible cleavage of glycine, yielding carbon dioxide, ammonia, 5,10-methylenetetrahydrofolate, and a reduced pyridine nucleotide. It consists of four component proteins termed P-protein, H-protein, T-protein, and L-protein. P-protein catalyzes the decarboxylation of the glycine molecule concomitantly with the transfer of the residual aminomethyl group to a sulfur atom on the lipoyl group of the oxidized H-protein (H_{ox}), generating the aminomethylated H-protein (H_{am}). Next, the T-protein catalyzes the transfer of a methylene group from H_{am} to THF, resulting in the release of NH_3 and the generation of reduced H-protein (H_{red}). Finally, the dihydrolipoyl group of H_{red} is oxidized by L-protein and the lipoyl group of H_{ox} is regenerated, thereby completing the catalytic cycle.

1.3. X-ray crystal structures of other H-proteins

The first structure of H-protein from pea leaves, with a reduced lipoic acid, was determined at 2.6 Å resolution (Pares *et al.*, 1994). The structure consists of two antiparallel β-sheets forming a "sandwich" structure, one with four strands including residues 13-19, 22-28, 74-80, and 101-105 and one with three strands including residues 37-43 and two adjacent antiparallel strands (55-61 and 65-70) joined by a loop (hairpin β-motif) in which the lipoylated lysine residue is situated. The overall structure is shown in Figure 1.4. The ribbon model was generated from PDB ID:1hpc.

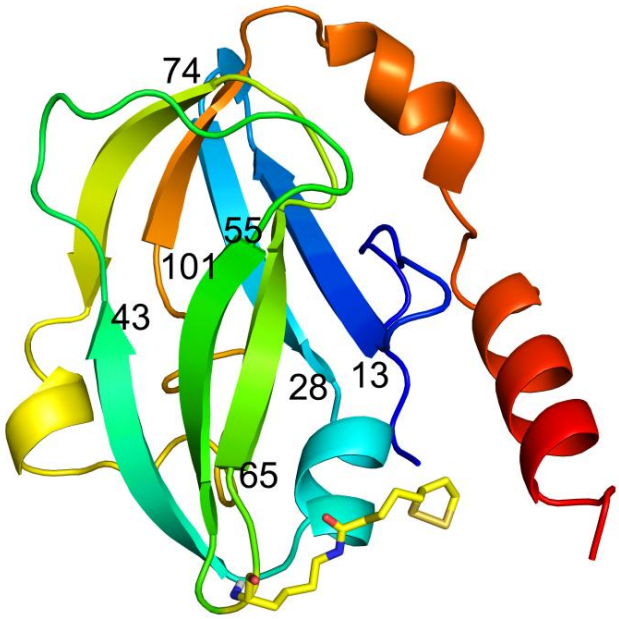


Figure 1.4 The overall structure of pea leave H-protein (PDB ID; 1hpc).

Schematic ribbon representation coloured from blue (N-terminus) to red (C-terminus). The numbers in figure mean the residue number. The lipoylated lysine residue 61 was represented with sticks. The figure was produced with the program *PyMOL* (DeLano, 2002).

Later, H-proteins from pea leaves with an oxidized lipoic acid and aminomethyl-lipoic acid were determined at 2.0 and 2.2 Å resolution, respectively (Pares *et al.*, 1995). Furthermore, the structure of *Thermus thermophilus* HB8 H-protein expressed in *E. coli* was determined at 2.5 Å resolution, and the structural similarity with pea H-protein was reported (Nakai *et al.*, 2003). The *Thermotoga maritima* structure determined at 1.65 Å resolution was deposited in Protein Data Bank (Joint Center for Structural Genomics (JCSG), 2005).

1.4. Purposes in this study

Currently, it is difficult to obtain a crystal that diffracts at high-resolution, so that the examples of high-resolution X-ray crystallography have not been well studied (Figure 1.1). However, it is clear that high-resolution data have an advantage in structural refinement of proteins and visualization of protein features not available at moderate resolution. In other words, more reliable structures are obtained from high-resolution X-ray diffraction data. For example, precise bond length and angles, multiple conformations, hydrogen atoms, and the electronic properties and charge distribution of macromolecules are directly observed from the experimental data.

As the resolution is pushed up (Figure 1.5), the amount of data increases substantially (Figure 1.6). Hence, high-resolution data make it more difficult to collect (Figure 1.7), scale and merge data, and to perform structural refinement with many structural parameters (Figure 1.8). We need more examples of high-resolution X-ray crystallography to treat high-resolution data, however so few high-resolution structures are currently available. Only 92 structures (<0.15% of the total) are beyond 0.9 Å resolution.

In this study, I attempted to establish the methodology of high-resolution X-ray crystallography of proteins, and the bovine H-protein was used for model protein for high-resolution X-ray crystallography.

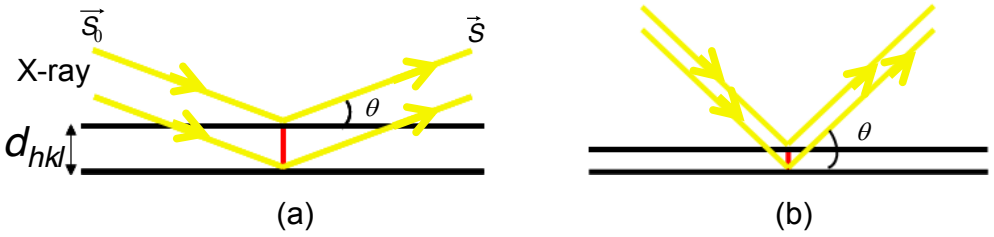


Figure 1.5 Diffractions from two lattice planes.

(a)The incident $(\vec{S}_0)$ and the reflected $(\vec{S})$ X-rays make an angle θ with the lattice planes. When the angle satisfies the Bragg law ($\lambda_{X-ray} = 2d_{hkl}\sin\theta$), the intensities of X-ray diffraction can be measured. This lattice plane spacing (d_{hkl}) is defined as a resolution. (b) The angle between the incident and the reflected X-rays are bigger to satisfy the Bragg law as the lattice plane spacing has smaller value (high-resolution).

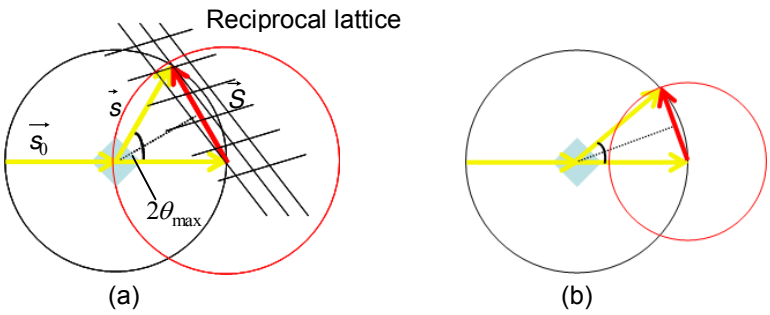
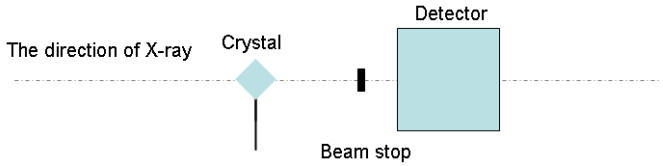
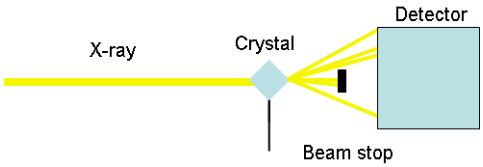


Figure 1.6 The number of observable reflections.

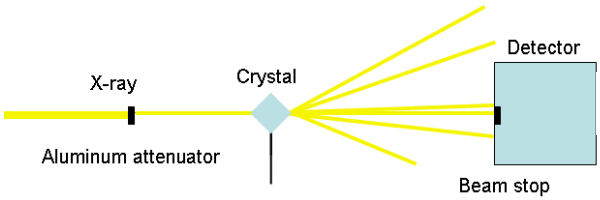
(a)The circle shows Ewald sphere, and the radius is $1/\lambda_{X-ray}$. The incident $(\vec{s}_0)$ and the reflected $(\vec{s})$ X-rays make an angle θ and $|\vec{s}_0| = |\vec{s}| = 1/\lambda_{X-ray}$. The points of the reciprocal lattice on the surface of the sphere are reflecting position, and the reflection of the highest resolution makes an angle θ_{max}. The number of reciprocal lattice points in the sphere generated from the $|\vec{S}|$ is equal to the number of observable reflections. $|\vec{S}| = 2\sin\theta/\lambda_{X-ray} = 1/d_{hkl}$ (b) In the case of small θ_{max} angle.



(a)



(b)



(c)

Figure 1.7 Data collection of X-ray diffractions from a protein crystal.

(a) The layout of experimental devices in X-ray experiment. (b) The experimental condition for collecting high-resolution X-ray diffraction. High-resolution data sets were collected with a long exposure time and a short camera distance. (c) The experimental condition for collecting low-resolution X-ray diffraction. To collect low resolution data, the X-ray was attenuated by an aluminium attenuator to avoid saturation of high-intensity diffractions. The beam stop was escaped in order to measure the intensities of the lowest-resolution reflections.

It is difficult to collect the complete data set, because of the limitation of experimental devices

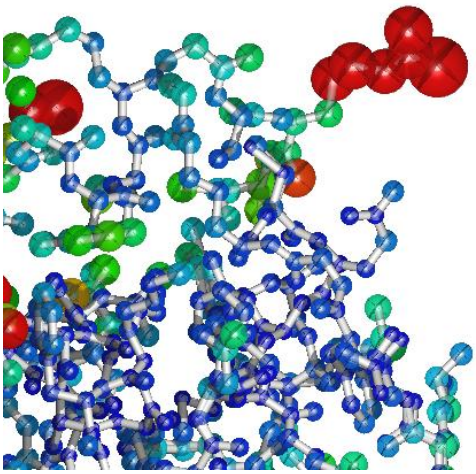


Figure 1.8 Many structural parameters "Anisotropic temperature factor".

In the X-ray crystallography, the atoms are represented by a coordinate (x,y and z) and a temperature factor (B) at moderate resolution (>~1.2 Å resolution). If the number of reflections is good enough (<~1.2 Å resolution) for determining the parameters, the temperature factors of atoms are represented using 6 parameters (U_{11}, U_{22}, U_{33}, U_{12}, U_{13}, U_{23}) for the anisotropy of electrons. These 6 parameters are called an anisotropic temperature factor. The anisotropic temperature factors are visualized as an ellipsoid coloured with red (the highest temperature factor) to blue (the lowest temperature factor).

The high quality crystals of bovine H-protein were obtained by the micro-seeding method from solution of highly purified H-protein, and these crystals enabled to start the study of high-resolution X-ray crystallography (Chapter2).

The structure of bovine H-protein was determined at 0.88 Å resolution. This is the first ultra-high-resolution structure of H-proteins. In the final model, ~40% of hydrogen atoms were visualized (Chapter 3).

To overcome the limitation of the hardware for X-ray experiments (Figure

1.6), three data sets were measured in order to obtain a complete set of high-resolution data. These three data sets were scaled and merged in several conditions, and the quality was assessed in general way (R_{merge} for measured intensities, R_{factor} for protein model and so on.). However, the advantage of the data could not be judged. To evaluate the advantage of merged data sets, hydrogen atoms were utilized as the indicator (Chapter 4).

Through the assessment of many combined data sets, the importance of low-resolution data for visualizing hydrogen atoms was apparent. (Chapter 5).

Chapter 2
Preparation and crystallization of bovine H-protein

2.1. Expression and purification

Expression and purification of bovine H-protein were basically performed as previously described (Fujiwara *et al.*, 1992). For X-ray crystallography, the expression and purification methods were partially modified.

2.1.1. Expression

The gene of bovine H-protein was ligated into the expression vector pET-3a and the plasmid was transformed into *E.Coli* BL21(DE3)pLysS carrying the gene for T7 lysozyme. The cells were grown in 300 ml of M9ZB medium (1 g of NH_4Cl, 3 g of KH_2PO_4, 6 g of Na_2HPO_4, 4 g of glucose, 1 ml of 1 M $MgSO_4$, 10 g of Bacto-trypton, and 5 g of NaCl in 1 liter of water) containing 100 μg/ml ampicillin and 34 μg/ml chloramphenicol. The culture was grown to OD_{600}=0.6 at 310 K and isopropyl-β-D-thiogalactopyranoside (IPTG) was added to a final concentration of 0.2 mM for induction of bovine H-protein expression. After incubation at 310 K for 4 hr, the culture was centrifuged at 8,000×g and the cell pellet was frozen for the purification.

2.1.2. Purification

The frozen cell pellet was thawed in lysis buffer containing 50 ml of 20 mM Tris-HCl, pH 7.5, 1 mM DTT, 40 μM PMSF and 10 μg/ml Leupeptine and sonicated 4 times for 5 minutes. The lysate was centrifuged at 48,400 g and the crude bovine H-protein was obtained in the supernatant. The supernatant was applied to a DEAE Sepharose^{TM}(GE Healthcare) column (15 ml volume)

equilibrated with the 20 mM potassium phosphate buffer (KPB), pH 7.4. The column was washed with 3 column volumes of 20 mM KPB, pH 7.4 buffer and eluted with 3 column volumes of 20 mM KPB containing 100 mM, 200 mM and 300 mM NaCl, respectively. The elution of 20 mM KPB with 200 mM NaCl was pooled and dialyzed against 20 mM KPB, pH 7.4 for 8 hr. After the dialysis the solution containing bovine H-protein were applied to a HiTrap DEAE FF^{TM}(GE Healthcare) column (5 ml volume) equilibrated with 20 mM KPB. The column was washed with 3 column volumes of same buffer and developed with a 0-500 mM NaCl gradient in 50 ml of 20 mM KPB using AKTA FPLC system (GE Healthcare). Fractions of 7 ml eluted by 20 mM KPB, pH 7.4 containing ~250 mM NaCl were concentrated with VIVASPIN 3,000 MWCO (VIVA SCIENCE) to 4 ml volume and diluted to 50 ml with 20 mM KPB, pH 7.4. The diluted solution was applied to a HiTrap Q HP^{TM}(GE Healthcare) equilibrated with the 20 mM KPB, pH 7.4 and developed with a 0-500 mM NaCl gradient in 50 ml of 20 mM KPB using AKTA FPLC system (GE Healthcare). Fractions of 7 ml eluted by 20 mM KPB, pH 7.4 containing ~250 mM NaCl were concentrated with VIVASPIN 3,000 MWCO (VIVA SCIENCE) to 4 ml volume. The concentrated solution was applied to a HiLoad^{TM} 16/60 Superdex 75pg (GE Healthcare) equilibrated and developed with the 20 mM KPB, pH 7.4 containing 150 mM NaCl, 0.5 mM DTT and 5% Glycerol (Figure 2.1.) and fractions containing bovine H-protein were pooled. The purified protein solution was concentrated to 15 mg/ml with VIVASPIN 3,000 MWCO (VIVA SCIENCE) for a cyrstallization (Figure 2.2). The protocol for purification of bovine H-protein is summarized in Figure 2.3.

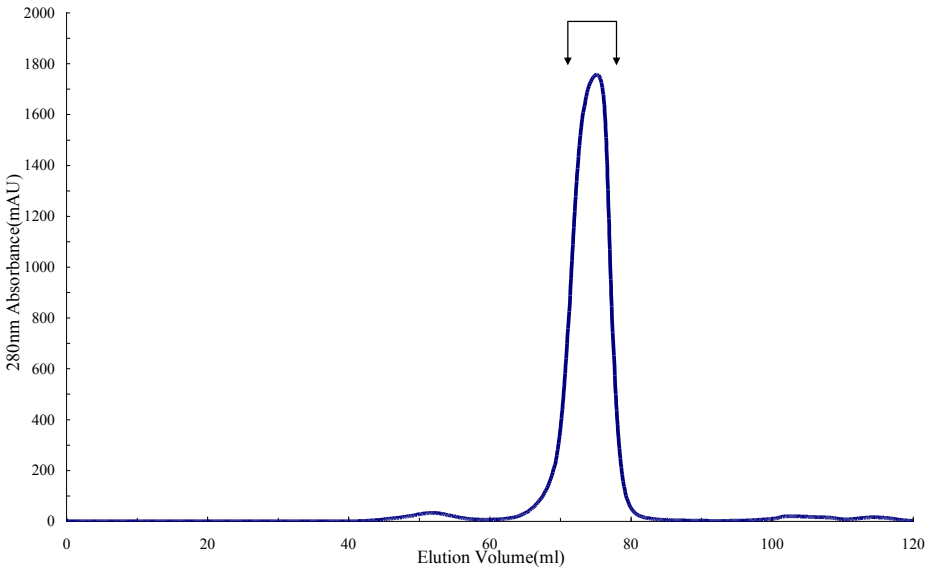


Figure 2.1 Chart of size exclusion chromatography of bovine H-protein.

Size exclusion chromatography was carried out using the column of HiLoad^{TM} 16/60 Superdex75pg (GE Healthcare). The elution profile was monitored at 280 nm. Bovine H-protein eluted in the elution volume 70~80 ml. The eluted fractions as shown in the figure were concentrated for crystallization.

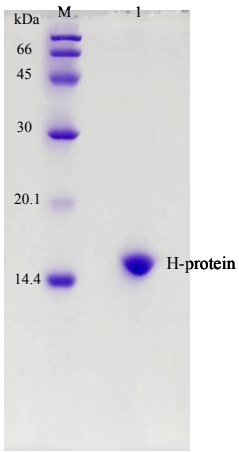


Figure 2.2 SDS-PAGE of highly purified bovine H-protein.

Purified bovine H-protein and a molecular weight marker electrophoresed in 13% SDS-poly-acrylamide gel. Lane 1 is highly purified bovine H-protein. M means the molecular weight marker for SDS-PAGE.

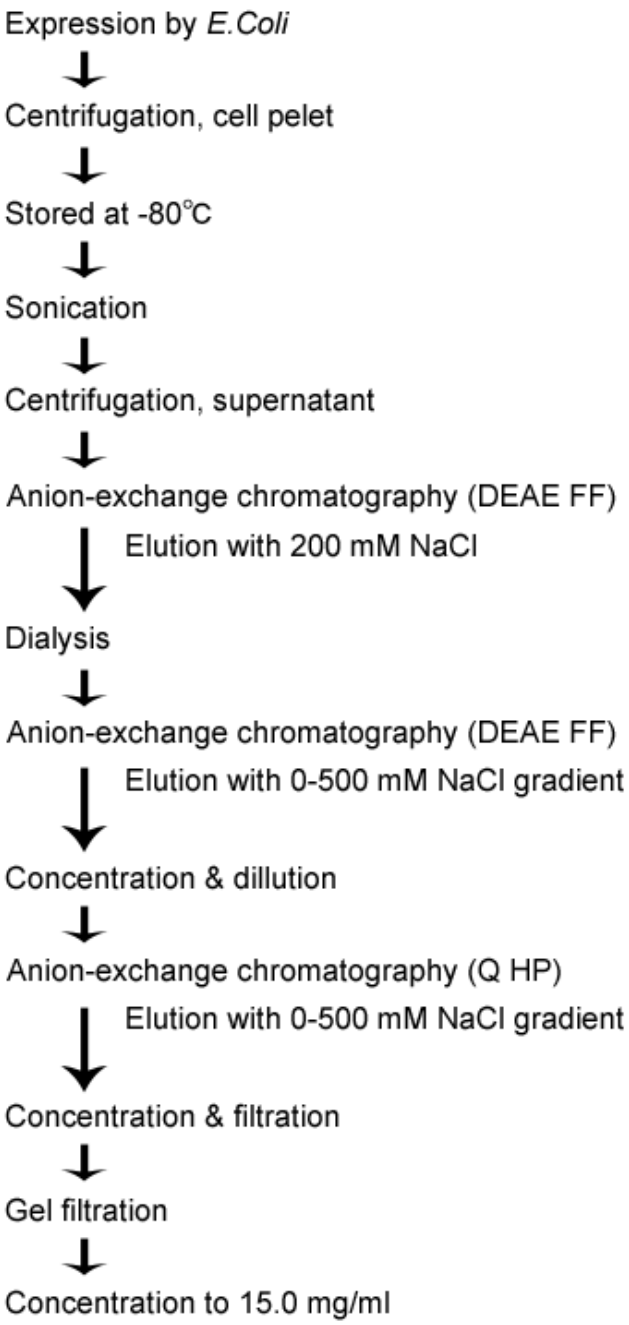


Figure 2.3 Protocol of bovine H-protein purification

2.2. Crystallization

2.2.1. Hanging drop vapour diffusion method

Crystallization was carried out by the hanging-drop vapour diffusion method (Figure 2.4). Crystals of bovine H-protein were grown in 2 μl drops containing a 1:1 (v/v) mixture of 15 mg/ml protein solution and 2.2-2.5 M ammonium sulfate in 0.1 M citrate buffer pH 2.8-3.2 at 288 K. Cluster-like crystals which were not suitable for X-ray experiments were obtained after a few days (Figure 2.5 (a)). A single crystal could not be obtained in any crystallization conditions (pH, precipitants, additives, temperatures and so on).

2.2.2. Micro seeding method

In order to obtain single crystals for X-ray experiment, the crystallization was carried out by a micro-seeding method. Micro seeds of H-protein crystals were generated using Seed Bead^{TM} (HAMPTON RESEARCH). This kit contains 'seed beads' manufactured from polytetrafluoroethylene (PTFE) for crushing crystals in 1.5 ml microcentrifuge tube. Reservoir solution of 50 μl was pipetted into the microcentrifuge tube and the cluster-like crystals were also transferred. The crystals in the tube with the bead were crushed using vortex for 90 sec, and the reservoir solution of 450 μl was added into the tube.

Single crystals for X-ray experiments were grown in 2 μl drops containing a 1:0.8:0.2 (v/v) mixture of 15 mg/ml protein solution, 2.2-2.5 M ammonium sulfate in 0.1 M citrate buffer pH 2.8-3.2, and the diluted seed solution. Single crystals suitable for X-ray experiments, with dimensions of 0.2-0.3 mm, were obtained after a few days (Figure 2.3 (b)).

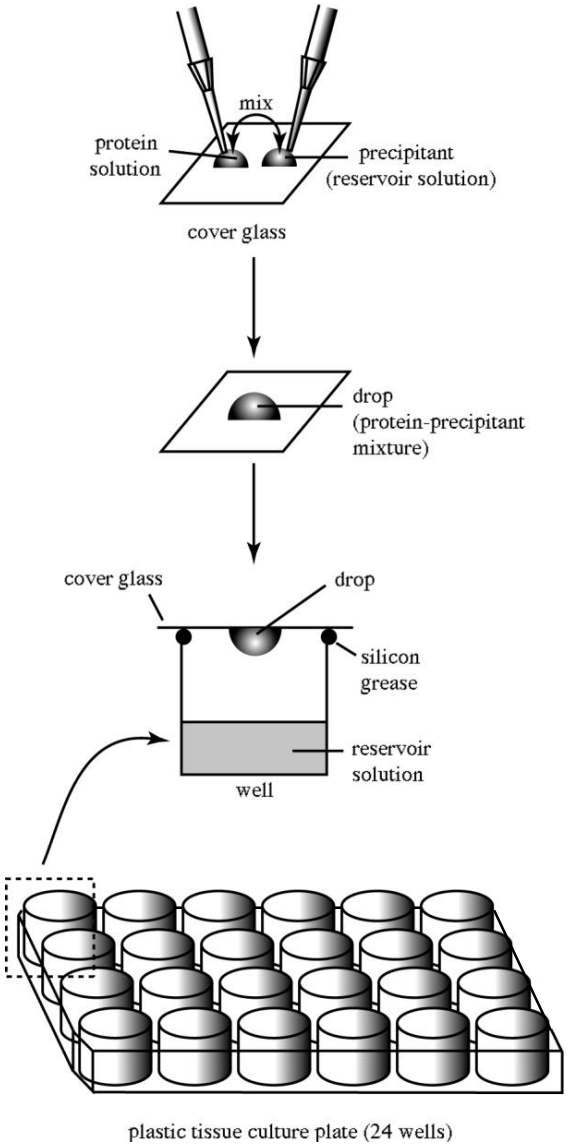
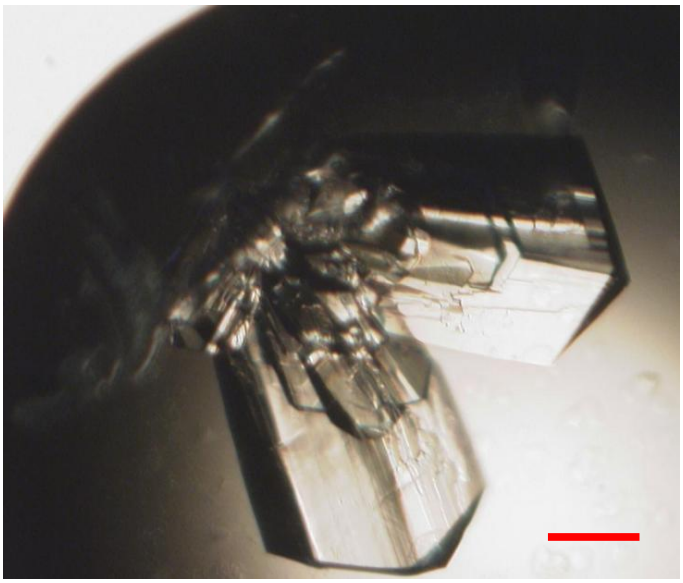
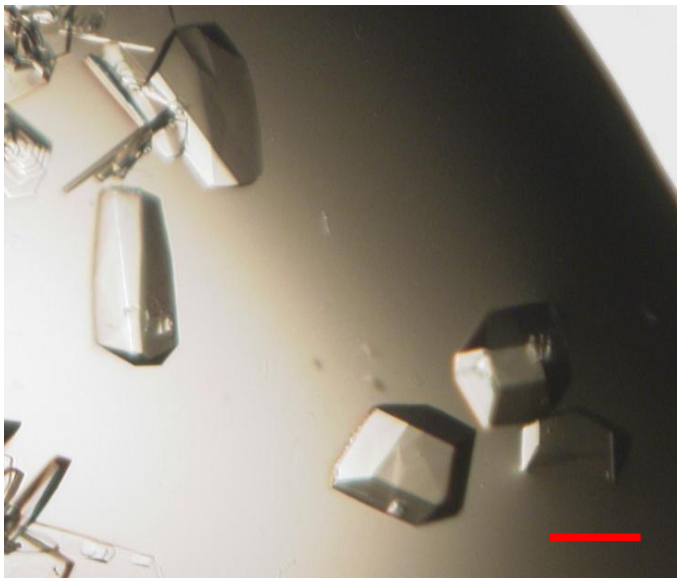


Figure 2.4 Hanging drop vapor diffusion method

The protein solution placed on cover glass was mixed with the same volume of reservoir solution. The wells containing precipitant solution were covered with them.



(a)



(b)

Figure 2.5 Crystals of bovine H-protein.

(a) Cluster-like crystals which were not suitable for X-ray experiments. (b) Single crystals for X-ray experiments. The bars represent 100 μm.

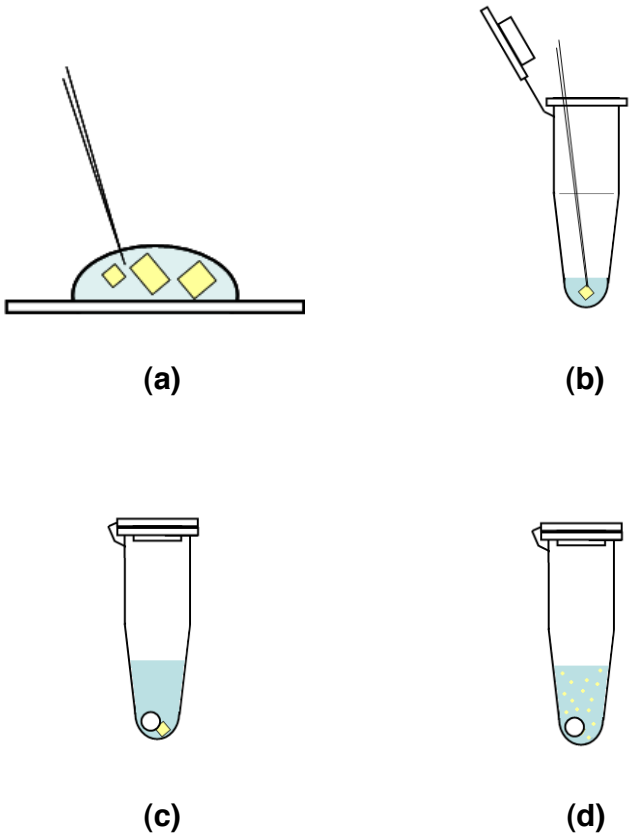


Figure 2.6 Micro-seeding method using Seed Bead™ (HAMPTON RESERACH).

(a) A crystal was picked up from the crystallization drop. (b)The picked crystal was transferred to the reservoir solution of the crystallization in microcentrifuge tube. (c) The crystal was crushed using vortex for 90 sec with the PTFE bead in the diluted reservoir solution. (d) The solution with micro seed.

Chapter 3
Data collection and processing

3.1. Data collection

Data collections were performed using synchrotron radiation from Photon Factory beamline BL-5A equipped with an ADSC Quantum315 CCD (ADSC) detector at 90K under a nitrogen vapour stream. Before freezing, each single crystal was transferred to the reservoir solution, including 30% (v/v) glycerol, using a cryo-loop. Three data sets were collected for high-, mid- and low-resolution data (Sevcik *et al.*, 1996). High-resolution data sets were collected with a long exposure time and a short camera distance, in order to measure higher Bragg angle diffractions (Figure 3.1). To collect low resolution data completely, the X-ray was attenuated by an aluminium attenuator to avoid saturation of high-intensity diffractions. Additionally, the front beam stopper was escaped in order to measure the intensities of the lowest-resolution reflections ((hkl)=(001)) (Figure 3.2). Mid-resolution data were collected in order to connect the high- and low-resolution data for data processing. All data sets were measured from a single crystal in a fixed position and the sequence of data collections was high-, mid- and low-resolution data set. The details of data collections are shown in Table 3.1.

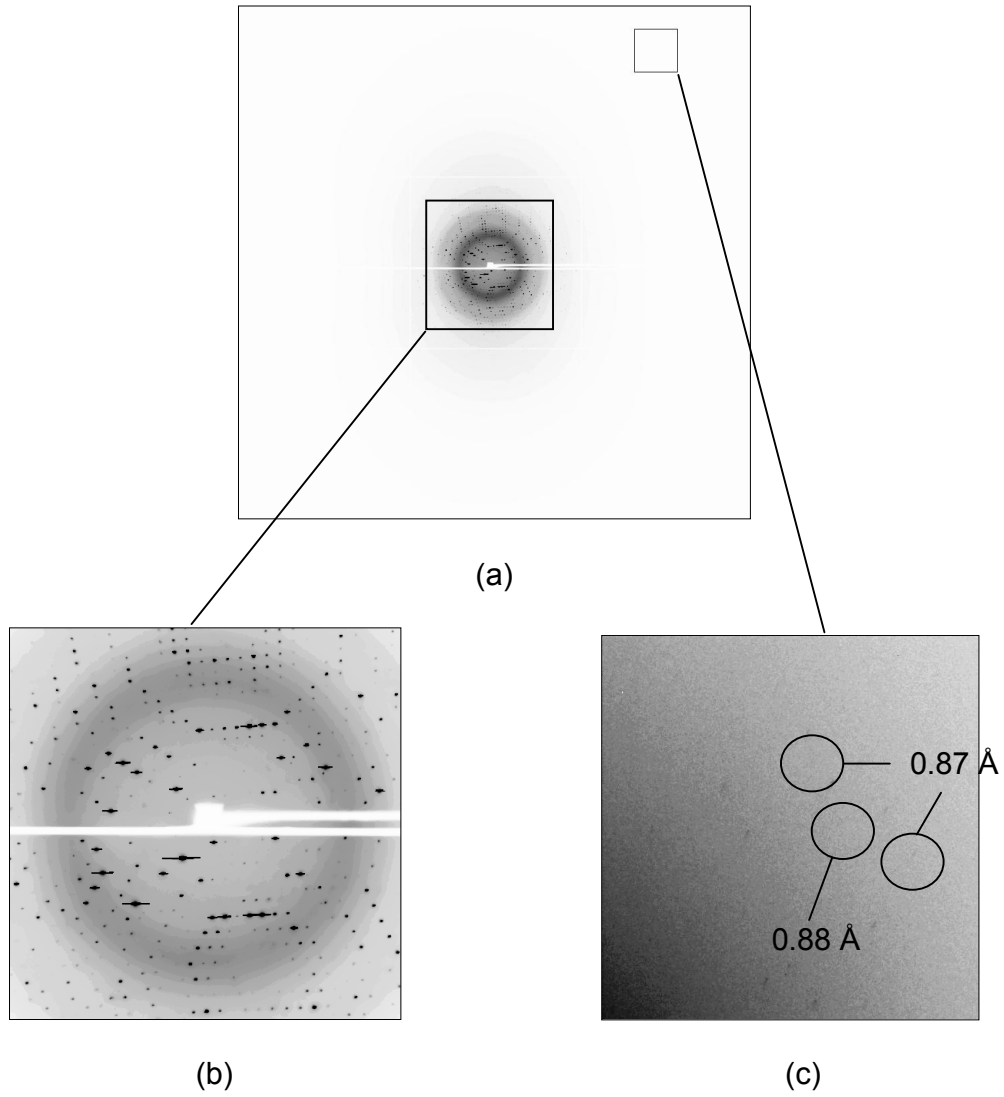


Figure 3.1 X-ray diffraction images of high-resolution data.

(a) The initial diffraction image of high-resolution data set. The resolution at the edge of the detector is 0.88 Å. (b) Zoom of Low-resolution regions. The streaks of the diffraction arise from the high-intensity exceeding the dynamic range of the detector. (c) Zoom of high-resolution regions. The maximum 0.87 Å resolution was observed.

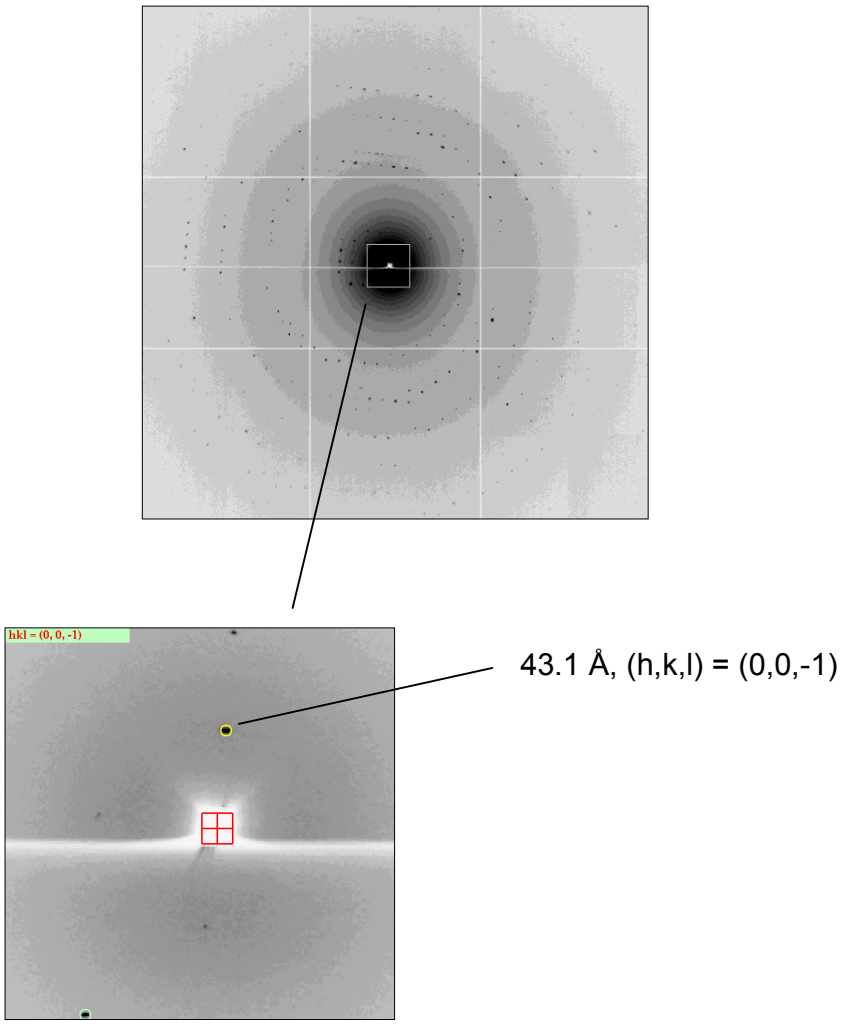


Figure 3.2 X-ray diffraction image of low-resolution data.

Resolution at the edge of the detector is 1.96 Å. The intensity from 43.1 Å diffraction was measured in the low-resolution data set.

Table 3.1 The experimental condition of data collection.

Low data set collection was performed without front beam stopper to collect very low resolution data.

Data set	High-resolution	Mid-resolution	Low-resolution
Beamline		Photon Factory BL-5A	
Detector		ADSC Q315	
Wavelength (Å)		1.0	
Camera distance (mm)	59.9	90.8	284.2
Max resolution (Å)	0.88	0.99	1.96
Oscillation range (°)	1.0	1.0	1.0
No. of images	360	180	180
X-ray exposure time (sec.)	10.0	3.0	1.0
Aluminium attenuator (mm)	0.0	0.0	0.3†

†X-ray was attenuated by 65.4%.

3.2. Scaling and merging data

High-, mid- and low-resolution data sets were measured in fixed position using a single crystal. All data sets were integrated, scaled and merged using the program *DENZO* and *SCALEPACK* as implemented in the *HKL*-2000 program package (Otwinowski & Minor, 1997). High-, mid- and low-resolution data sets were scaled and merged using high-resolution data set as the reference. The statistics of merged data sets are shown in Table 3.2. An overall *B* factor of 7.50 $Å^2$ was estimated from the Wilson plot (Figure 3.3) using the program *TRUNCATE* (Collaborative Computational Project, Number 4, 1994).

Table 3.2 Statistics of merged data

Data of highest-resolution shells are in parentheses.

Space group	*C*2
Cell dimensions (Å, °)	a = 84.41, b = 41.25, c = 43.05, β = 91.18
Resolution range (Å)	43.1 - 0.88 (0.89 - 0.88)
No. of observed reflections	1,113,257
No. of unique reflections	115,668 (3,712)
Redundancy	9.6 (4.7)
Completeness (%)	98.9 (95.4)
$\langle I \rangle / \langle \sigma(I) \rangle$	70.0 (3.72)
R_{merge} †	4.7 (38.6)
No. of reject reflections	22,060
Wilson B ($Å^2$)	7.50

† $R_{merge} = \sum_{hkl} \sum_j \left| I_j(hkl) - \overline{I(hkl)} \right| / \sum_{hkl} \sum_j \overline{I(hkl)}$, where $I_j(hkl)$ and $\overline{I(hkl)}$ are the observed intensity of measurement j and the mean intensity for reflections with index *hkl*, respectively.

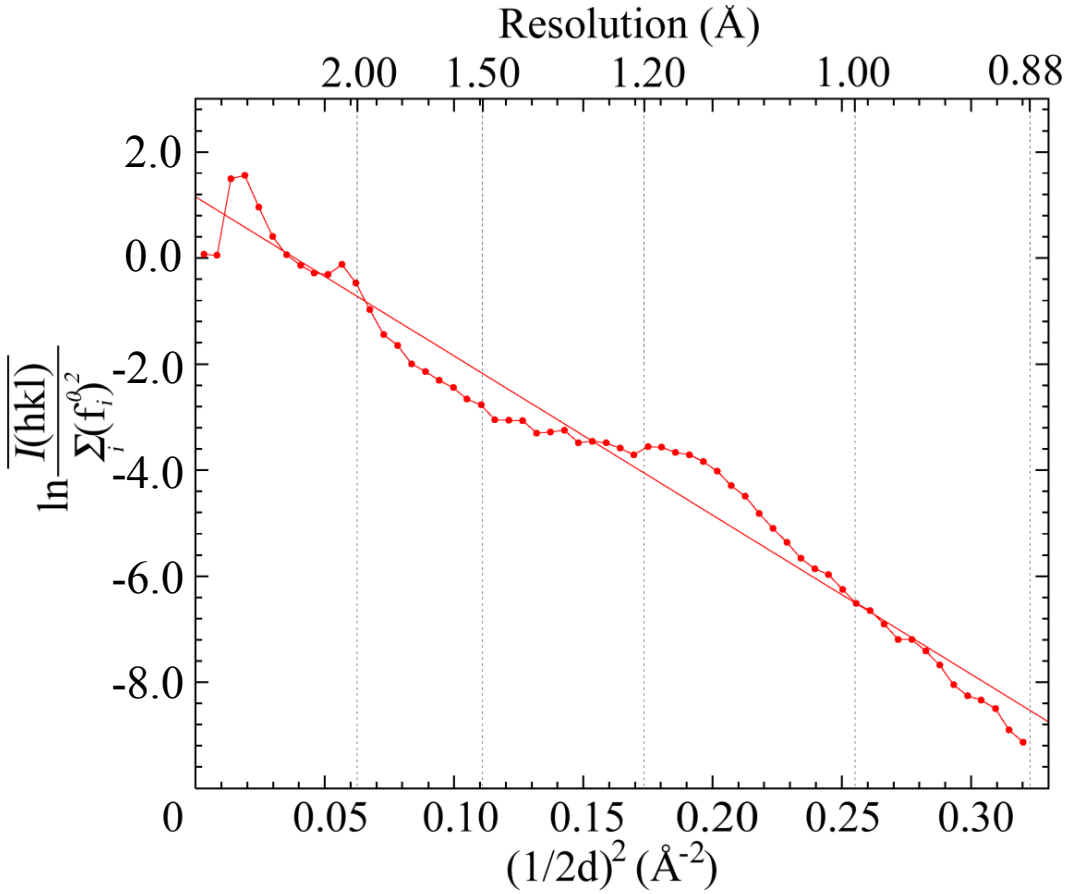


Figure 3.3 Wilson plot of merged data set.

The line represents y=2(-7.50)x+1.58, where y and x are the vertical and horizontal axis. I, f_i and d represent an intensity, atomic scattering factor and lattice plane distance. Wilson plot is defined as the equation (3) (Wilson, 1942).

$$\ln\frac{\overline{I(hkl)}}{\sum_i\left(f_i^0\right)^2} = \ln K - 2B\frac{\sin^2\theta}{\lambda^2} \qquad (3)$$

where B is temperature factor and K is scale factor.

3.3. Phase determination and structural refinement

3.3.1. Phase determination

Phase determination and structural refinement were carried out against the merged data, using the high-resolution data set as a reference batch (Table 3.2). The merged data sets were converted to the structure factor amplitude using the program *TRUNCATE* (Collaborative Computational Project, Number 4, 1994). Matthews Coefficient (Matthews, 1968) and solvent content calculated by the program *Matthews_coef* (Collaborative Computational Project, Number 4, 1994) were 2.68 $Å^3 \cdot Da^{-1}$ and 54.1%, respectively. These values suggested the presence of one molecule in the asymmetric unit (Matthews, 1968). The initial structure of H-protein was determined by molecular replacement method using the program *Molrep* (Vagin & Teplyakov, 1997) with the pea H-protein structure (PDB ID 1hpc; Pares *et al.*, 1995) as a search model. All side-chains and water molecules were removed from the search model. The *R* factor between observed and calculated structure factors of the best solution were 0.57, and its value was lower than other solutions. In addition, the contrast against other solutions was good enough.

3.3.2. Overviews of structural refinement

In order to calculate the R_{free} factor, 5% of reflections were randomly selected and excluded in the following refinement for a cross-validation analysis (Brünger, 1997). The program *ARP/wARP* (Perrakis *et al.*, 1999) was used to build missing side-chain atoms and add water molecules. The refinement process utilized the programs *phenix.refine* (Adams *et al.*, 2002) for water molecule picking, *SHELXL* (Sheldrick & Schneider, 1997) for maximum least-squares refinement and *Coot* (Emsley & Cowtan, 2004) for manual model building and adjustments. In the isotropic or anisotropic temperature

factor refinement, the coordinates or temperature factors of hydrogen atoms were not refined. In the refinement with riding hydrogen atoms, no hydrogen atoms were added to hydroxyl groups, water molecules and other solvent molecules. Finally, all occupancies of water molecules were refined. After the each of refinement step, the model was inspected and rebuilt manually using the *Coot*. In the maximum least-squares refinement using *SHELXL*, the structural parameters were refined against the intensities. The schematic of refinement process are shown in Figure 3.4.

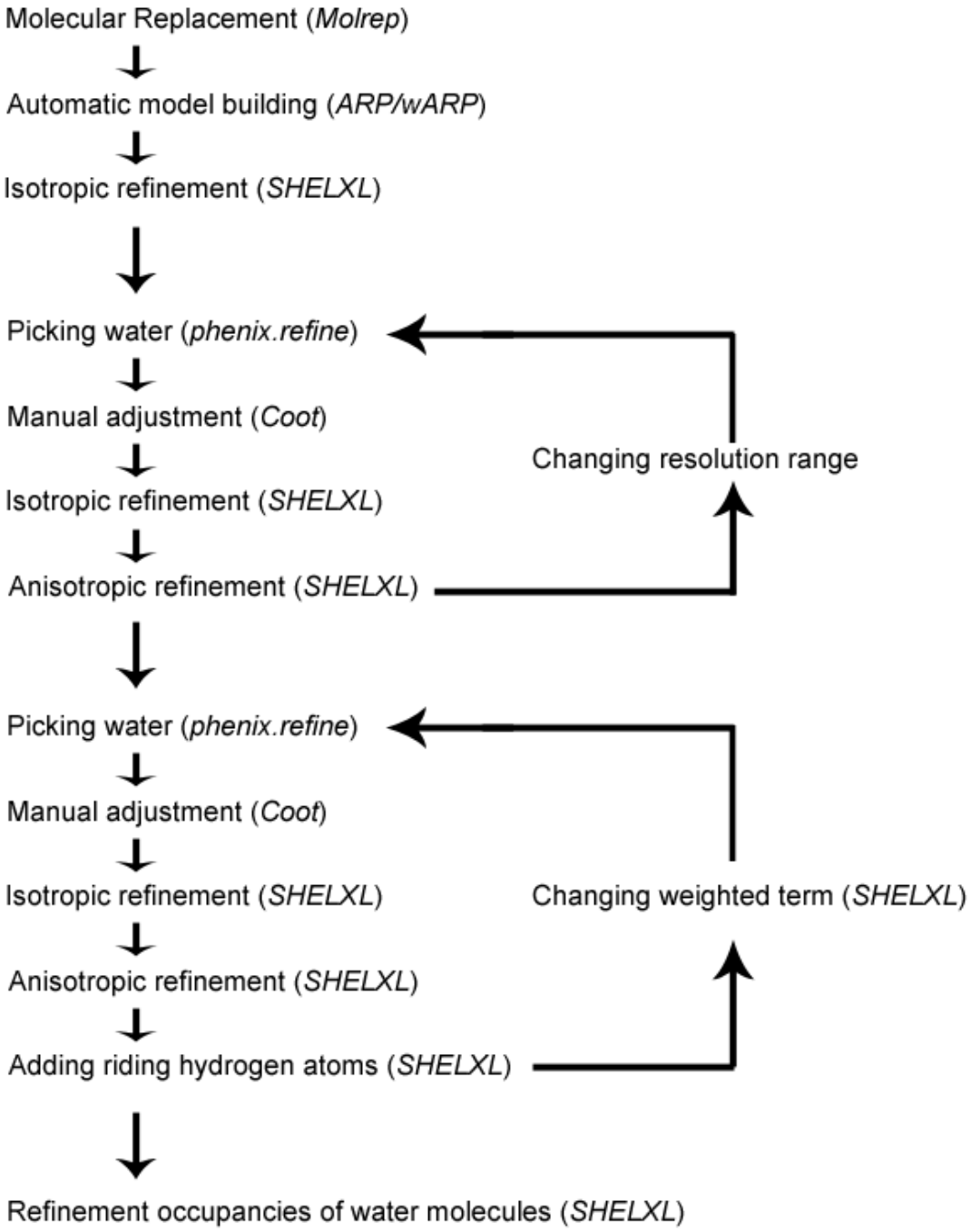


Figure 3.4 Protocol of structure analysis and refinement.

3.3.3. Details of structural refinement

The model from molecular replacement was used as an initial model. The missing side-chain atoms and water molecules were automatically built by the program *ARP/wARP*, resulting in an *R* factor and an R_{free} factor of 18.3% and 20.7%, respectively. The first step of an isotropic temperature factor refinement using the program *SHELXL* at 10.0-1.5 Å resolution converged to an *R* factor of 18.1% and an R_{free} factor of 20.8%. After the isotropic temperature factor refinement, disordered residues and multiple conformations were manually added using the program *Coot*. After the manual adjustment, water picking was carried out using *phenix.refine*, resulting in an *R* factor of 19.0% and the R_{free} of 20.2%. Isotopic refinement was carried out, and the maximum resolution was gradually extended to 1.1 Å. After further manual adjustments, the occupancies of multiple conformations and flexible side chains were determined. Anisotropic temperature refinements were carried out using the program *SHELXL* against data to 10.0-1.1 Å. After several rounds of anisotropic temperature factor refinement using *SHELXL*, *R* factor and R_{free} factor were 11.9% and 14.0%, respectively. The resolution was extended to 0.88 Å, and water picking using *phenix.refine*, isotorpic and anisotropic temperature factor refinement using *SHELXL* lowered *R* factor and R_{free} factor to 11.2% and 12.9%, respectively. Next step, water picking, isotropic and anisotropic temperature factor refinement were carried out at the resolution rage of 43.1-0.88 Å, resulting in an *R* factor and R_{free} factor of 11.4% and 13.2%. The resulting *R* factor and R_{free} factor were increased by 0.2% and 0.3% because of the low-resolution data were newly used for the refinement. Subsequently, hydrogen atoms were added to the model at the calculated positions without refinement. No hydrogen atoms were added to water and small molecules. After the refinement with riding hydrogen atoms, the *R* factor and the R_{free} factor fell by 1.1% and 1.2%, respectively. The

refinement of water molecule occupancies converged at values of 10.1% and 11.8% for the R factor and the R_{free} factor, respectively. The refinement procedure is summarized in Table 3.3.

Table 3.3 The refinement procedure.

Round / Action taken / Program†			Resolution (Å)	††R (%)	††R_{free} (%)	No. of atoms (water)
1	Molecular replacement	M	43.1-1.50	64.0	-	641 (0)
2	Model building automatically	A	43.1-1.50	18.3	20.7	1084 (154)
3	Isotropic B-factor refinement	S	10.0-1.50	18.1	20.8	1093 (133)
4	Isotropic B-factor refinement and picking water automatically	P	10.0-1.10	19.0	20.2	1254 (289)
5	Isotropic B-factor refinement after manual adjustment	S	10.0-1.10	18.1	19.8	1331 (262)
6	Anisotropic B-factor refinement against same atoms	S	10.0-1.10	12.3	14.4	1331 (262)
7	Isotropic B-factor refinement after manual adjustment	S	10.0-1.10	17.6	19.4	1352 (250)
8	Anisotropic B-factor refinement against same atoms	S	10.0-1.10	11.9	14.0	1352 (250)
9	Isotropic B-factor refinement and picking water automatically	P	10.0-0.88	17.5	18.3	1399 (297)
10	Isotropic B-factor refinement after manual adjustment	S	10.0-0.88	18.4	19.9	1438 (297)

Table 3.3 The refinement procedure. (continued)

Round	Action taken	Program	Resolution (Å)	R (%)	R_{free} (%)	No. of atoms (water)
11	Anisotropic B-factor refinement against same atoms	S	10.0-0.88	11.2	12.9	1438 (297)
12	Isotropic B-factor refinement and picking water automatically	P	43.1-0.88	17.1	18.2	1471 (291)
13	Isotropic B-factor refinement after manual adjustment	S	43.1-0.88	18.5	20.1	1444 (285)
14	Anisotropic B-factor refinement against same atoms	S	43.1-0.88	11.4	13.2	1444 (285)
15	Adding riding hydrogen atoms	S	43.1-0.88	10.3	12.0	1444 (285)
16	Changing weighted term	S	43.1-0.88	10.2	11.9	1444 (285)
17	Isotropic B-factor refinement after manual adjustment	S	43.1-0.88	18.3	19.9	1431 (274)
18	Anisotropic B-factor refinement against same atoms	S	43.1-0.88	11.2	12.9	1431 (274)
19	Adding riding hydrogen atoms	S	43.1-0.88	10.3	11.8	1431 (274)
20	Refinement of water occupancies	S	43.1-0.88	10.1	11.8	1431 (274)

† The refinement was performed with the programs *Molrep* (M), *ARP/wARP* (A), *SHELXL* (S), and *phenix.refine* (P).

†† $R = \sum_{hkl} \| F_{obs} | - | F_{calc} \| / \sum_{hkl} | F_{obs} |$, where $| F_{obs} |$ and $| F_{calc} |$ are the observed and calculated structure factor amplitude for reflections with index *hkl*, respectively. An *R* factors calculated by *SHELXL* are for reflections with $F_{obs} > 4\sigma(F_{obs})$.

3.4. Atomic structure of bovine H-protein

3.4.1. Quality assessment of the structure

The quality of the final model was checked using the program *WHAT IF* (Variend, 1990) and *MolProbity* (Davis *et al*., 2004). The statistics of the final model are reported in Table 3.4, and the Ramachandran plot (Ramachandran & Sasisekharan, 1968) are shown in Figure 3.5. The plots of *B* factors averaged over main-chain and side-chain were shown in Figure 3.6.

Table 3.4 Statistics of final model.

Resolution range (Å)	43.1-0.88
R factor (%)	11.3 (10.1)†
R_{free} factor (%)	13.2 (11.8)†
No. of protein atoms	1135
No. of solvent atoms	296
R.m.s. deviations from ideal geometry	
Bond distances (Å)	0.017
Bond angles (°)	2.15
Ramachandran plot	
Residues in most favored regions (%)	96.6
Residues in additional allowed regions (%)	3.4
Standard deviation of ω value (°)	7.19
Mean *B* factors (Å^2)	
Protein atoms (all/main chain/side chain)	11.4/9.28/13.2
Small molecules	20.5
Water molecules	28.1

†*R* factors in parentheses are for reflections with $F_o > 4\sigma(F_o)$.

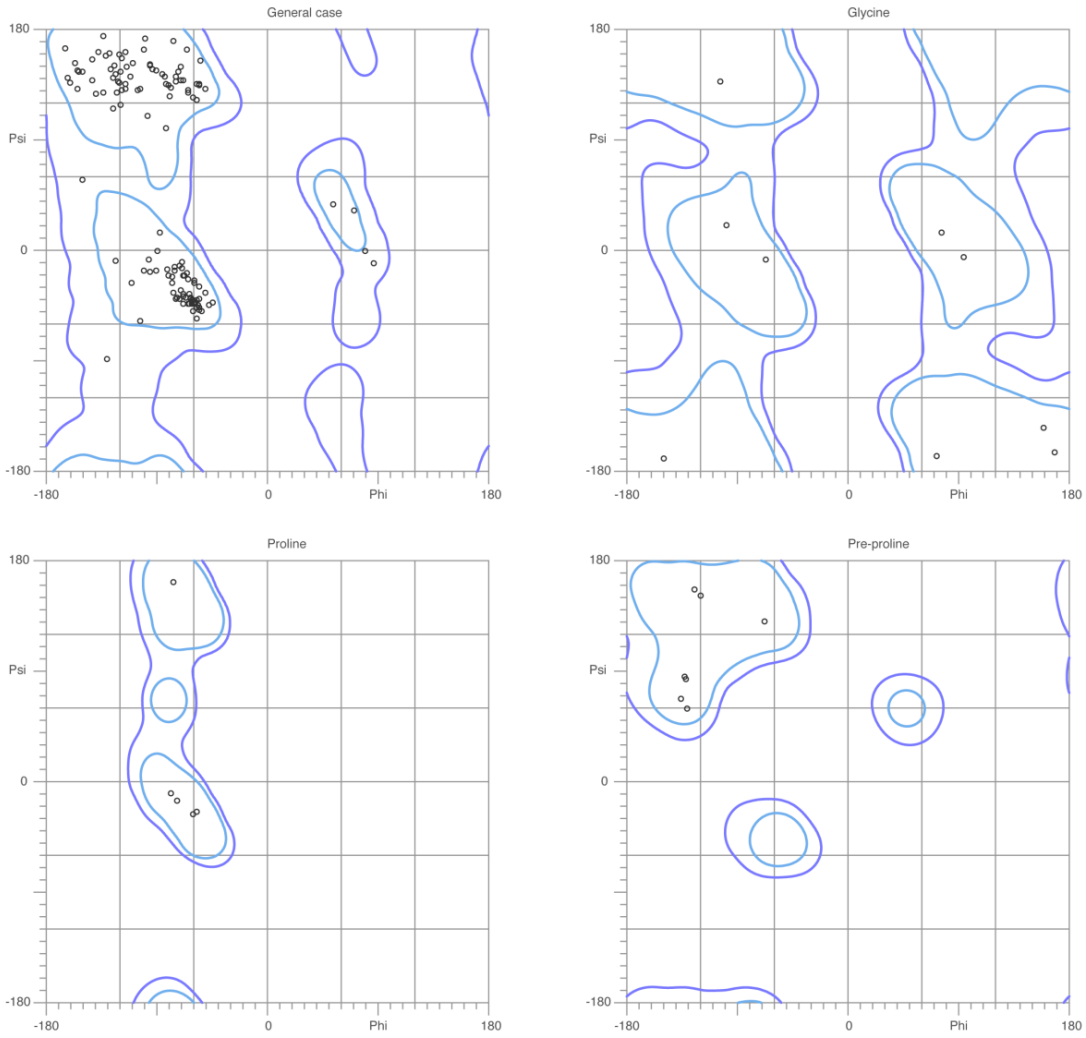


Figure 3.5 Ramachandran plot.

Phi and Psi represent the angles of main-chain C^{α}-N and C^{α}-C, respectively, and Ramachandran plot are used to validate the structural model. Regions surrounded by blue lines mean allowed region and light-blue mean favored regions (Lovell *et al.*, 2003). In this case, 96.6% of all residues were in favored regions, and 3.4% of all residues were in allowed regions. There were no outliers. This plot was generated by program *MolProbity* (Davis *et al.*, 2004).

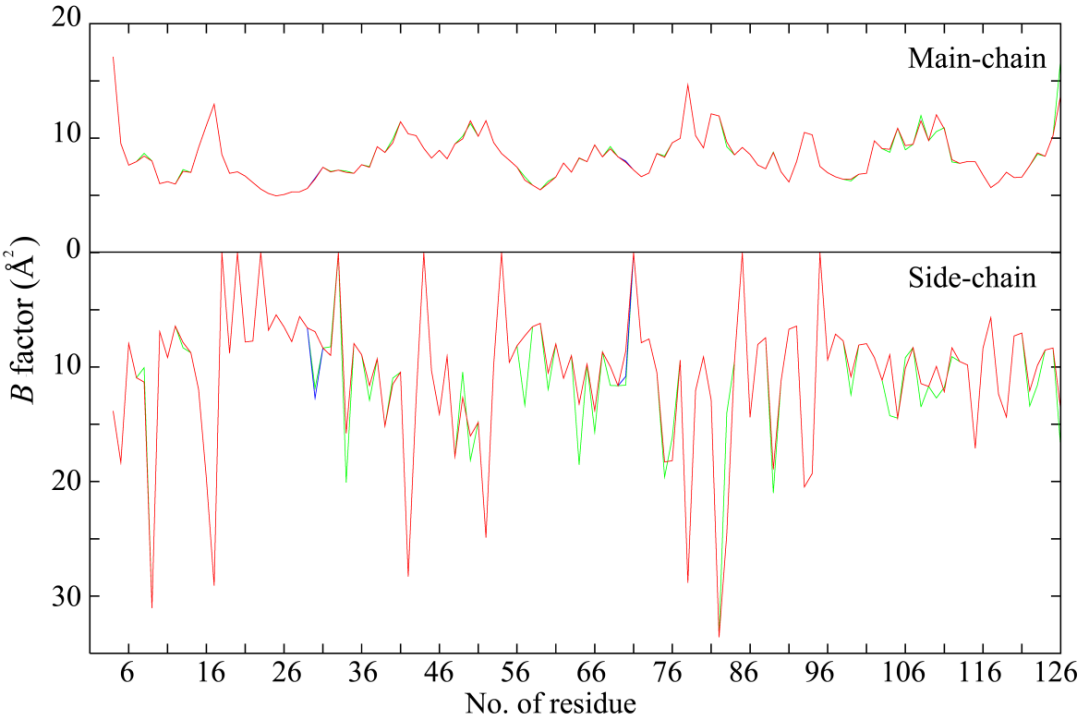


Figure 3.6 Plots of averaged *B* factors.

B factors averaged over main-chain and side-chain atoms for final model. Red line, green line, and blue line represent conformation A, conformation B, and conformation C, respectively. All *B* factors of main-chains and side-chains are smaller than 20 and 35 Å^2. There are no multiple conformations which has extremely high *B* factors.

3.4.2. Overall structure of bovine H-protein

The overall structure of bovine H-protein is shown in Figure 3.7. This structure mainly consists of two antiparallel β-sheets, and helices at the C-terminus joined to the main domain by a flexible linker. The two antiparallel β-sheets form a β–sandwich. One sheet is composed of four strands comprising of residues 10-15, 18-23, 70-75, and 97-101, and the other is composed of three strands comprising of residues 34-39, 51-57 and 61-65. There are also two short β-strands (4-5 and 45-46). The long disordered helix consisting of residues 104-109 (3_{10}-helix) and 113-125 (α-helix) is positioned at the C-terminus. Two other short α-helices (residues 25-31, 85-88) and a 3_{10}-helix (residues 77-79) are positioned in the loops that join the β-sheets. Although lysine 59 in one of these loops is the site of lipoylation, the residue has no lipoyl prosthetic group in this structure. The secondary structures were assigned by *DSSP* (Kabsch, W. & C. Sander, 1983), and shown in Figure 3.8.

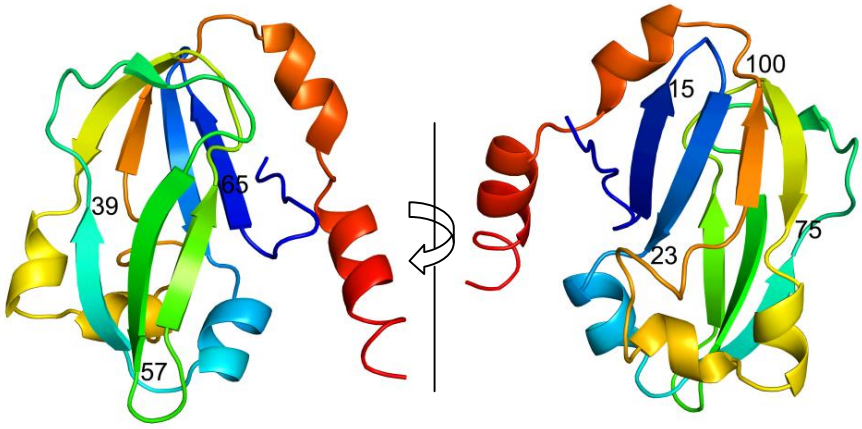


Figure 3.7 The overall structure of bovine H-protein.

Schematic ribbon representation coloured from blue (N-terminus) to red (C-terminus). The numbers in figure mean the residue number. The figure was produced with the program *PyMOL* (DeLano, 2002).

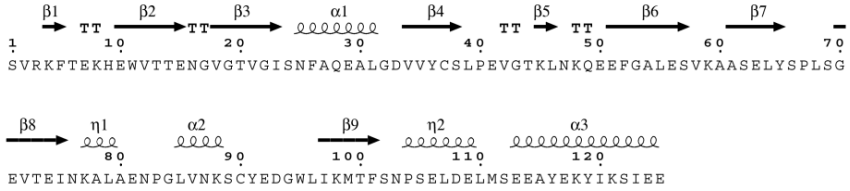


Figure 3.8 Secondary structure assignment of bovine H-protein.

α, β, η and TT represent α-helix, β-strand, 3_{10}-helix and turn, respectively. The secondary structures were assigned by *DSSP* and represented by *EsPript* (Gouet, P., *et al.,* 2003, http://espript.ibcp.fr/ESPript/ESPript/index.php).

3.4.3. Comparisons between bovine and other H-proteins

H-proteins in the glycine cleavage systems are widely distributed in many species, and the X-ray crystal structures of H-proteins from pea and bacteria were deposited in Protein Data Bank (PDB ID 1dxm, 1hpc, 1htp, 1onl and 1zko). The structural similarity between pea and *T. thermophilus HB8* H-protein has been previously reported (Nakai *et al.*, 2003). Bovine H-protein shares 49% amino acid sequence identity with pea, 43% with *T. thermophilus HB8* and 50% with *T. maritima* proteins (Figure 3.8). As expected, the secondary structures of the bovine and other H-proteins are almost the same, and the r.m.s. deviations for C^{α} atoms were 0.575-1.14 Å (Table 3.5 and Figure 3.10). Differences are observed mainly in the C-terminal helices, flexible loops and around residue 59 of bovine H-protein. These results show that the secondary structures are highly conserved in H-proteins of glycine cleavage systems.

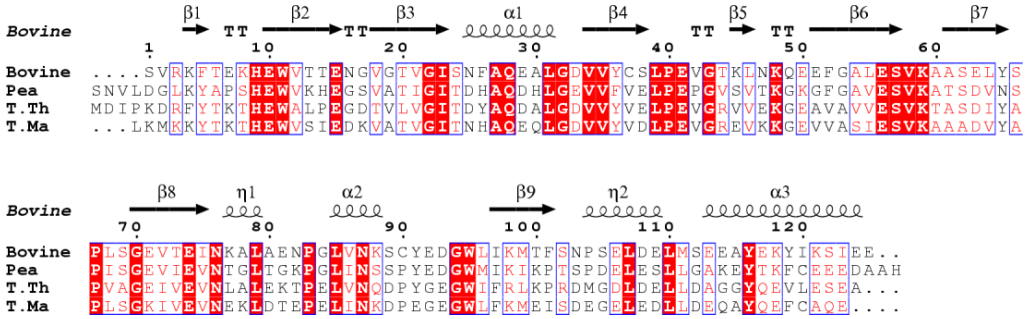


Figure 3.9 Alignment of amino-acid sequences of H-proteins.

Tt and Tm represent *T.thermohpilus* and *T.maritima*, respectively. α, β, η and TT represent α-helix, β-strand, 3_{10}-helix and turn, respectively. The secondary structures were assigned by *DSSP* and represented by *EsPript* (Gouet, P., *et al.,* 2003, http://espript.ibcp.fr/ESPript/ESPript/index.php)

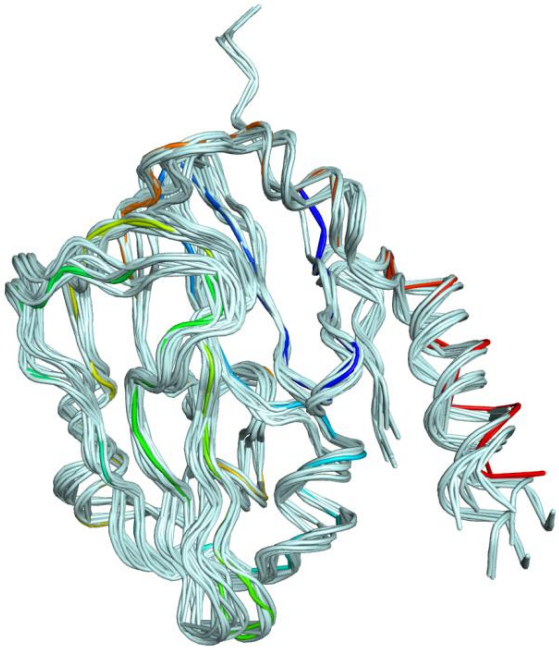


Figure 3.10 Superimposition of bovine H-protein and other H-proteins.

Bovine H-protein is coloured from blue to red and other H-proteins are coloured cyan. PDB codes of other H-protein are 1hpc, 1htp, 1dxm, 1onl and 1zko. The structures were superimposed by least-square fitting of main-chain atoms.

Table 3.5 Maximum and mean differences and r.m.s. deviations between bovine and other H-proteins.

Tt and Tm represent *T.thermohpilus* and *T.maritima*, respectively.

Reference range of bovine H-protein / range of residues	Species	PDB code (Chain ID) / range of residues	Mean diff. (Å)	R.m.s. dev. (Å)	Max diff. (Å)
3-125	Pea	1dxm(A) /7-129	0.729	0.864	3.245
3-125	Pea	1dxm(B) /7-129	0.700	0.828	2.833
3-125	Pea	1hpc(A) /7-129	0.697	0.824	3.176
3-125	Pea	1hpc(B) /7-129	0.658	0.796	2.817
3-125	pea	1htp /7-129	0.623	0.730	2.797
3-125	Tt	1onl(A) /7-129	0.746	0.903	4.085
3-125	Tt	1onl(B) /7-129	0.887	1.142	5.826
3-125	Tt	1onl(C) /7-129	0.827	1.070	6.141
3-122	Tm	1zko(A) /4-123	0.466	0.591	4.493
3-122	Tm	1zko(B) /4-123	0.455	0.575	4.635

3.5. Features of high-resolution structure of bovine H-protein

3.5.1. Crystal packing

Crystals diffracting to atomic-resolution tend to have lower solvent contents (35-40%) and lower symmetry (Schmidt & Lamzin, 2002; Bönisch *et al.*, 2005). However, crystals of bovine H-protein had a solvent content of 54.8% (Matthews coefficient V_M = 2.72 Å^3 Da^{-1}). The high quality of the crystals despite high solvent content could be rationalized in terms of crystal packing. The flexible loops that join β-strands and the long helices at the C-terminus, which was separated from the core are strongly held by other symmetry-related molecules (Figure 3.11).

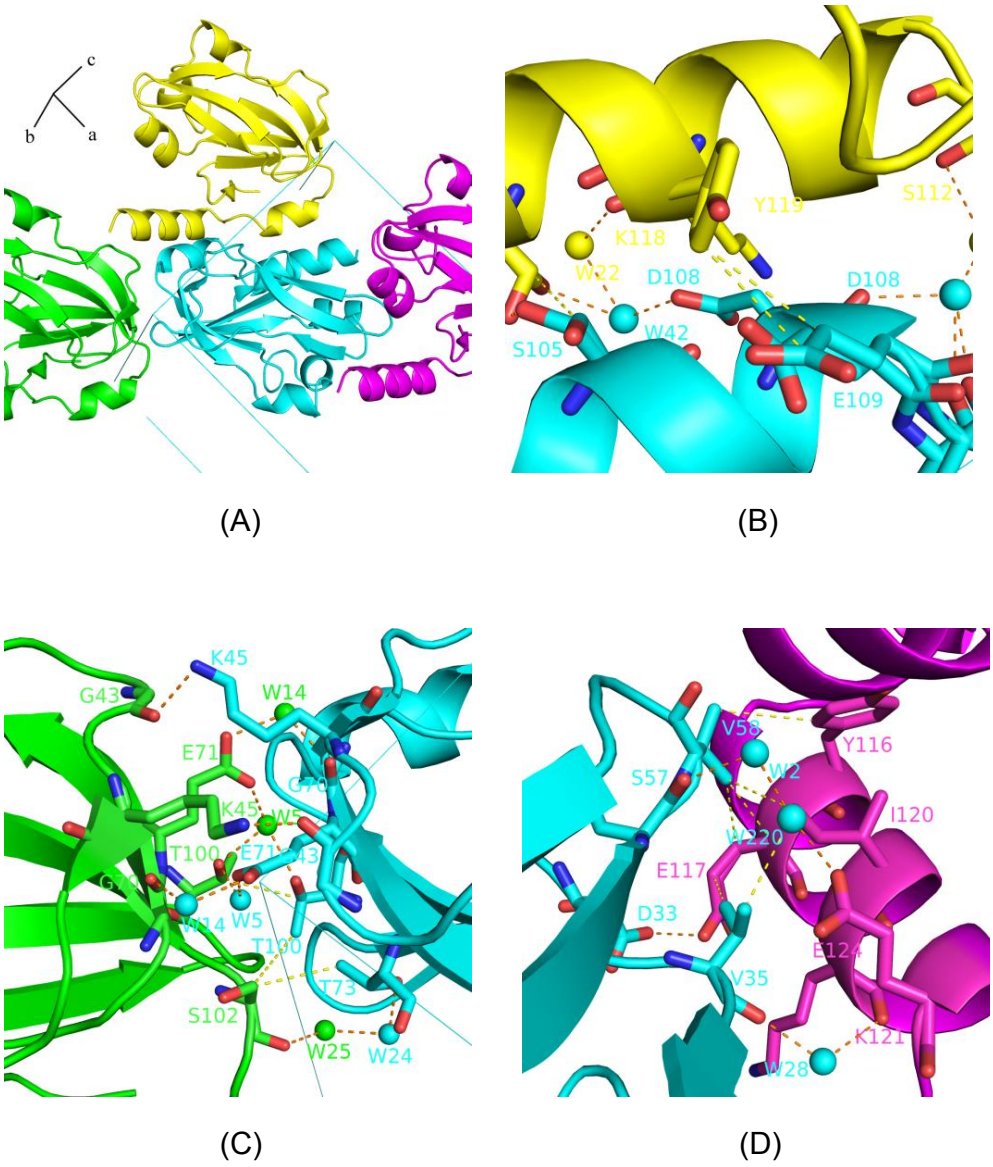
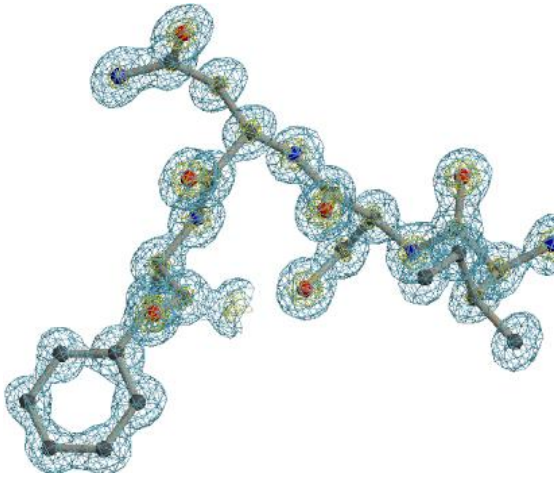


Figure 3.11 Crystal packing of the bovine H-protein.

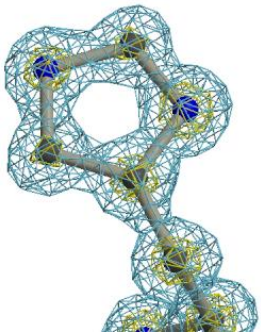
(A) Green, yellow and cyan molecules show (-x, y, -z), (-x, y, -z+1) and (-x+1/2, y+1/2, -z+1), respectively. (B), (C) and (D) show the regions of intermolecular interactions. Hydrogen bonds and hydrophobic interactions represented by red and yellow dotted lines. The figures were produced with the program *PyMOL* (DeLano, 2002).

3.5.2. Electron density map

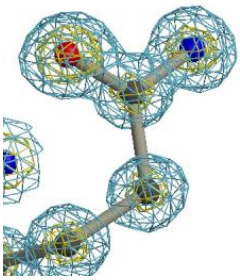
The final electron density map showed clear electron density of non-hydrogen atoms. The identities of atoms could be clearly assigned on the basis of electron density in many regions of the protein, as well as for solvent atoms (Figure 3.12). However, N-terminus region could not be modeled because of poor electron density.



(a)



(b)



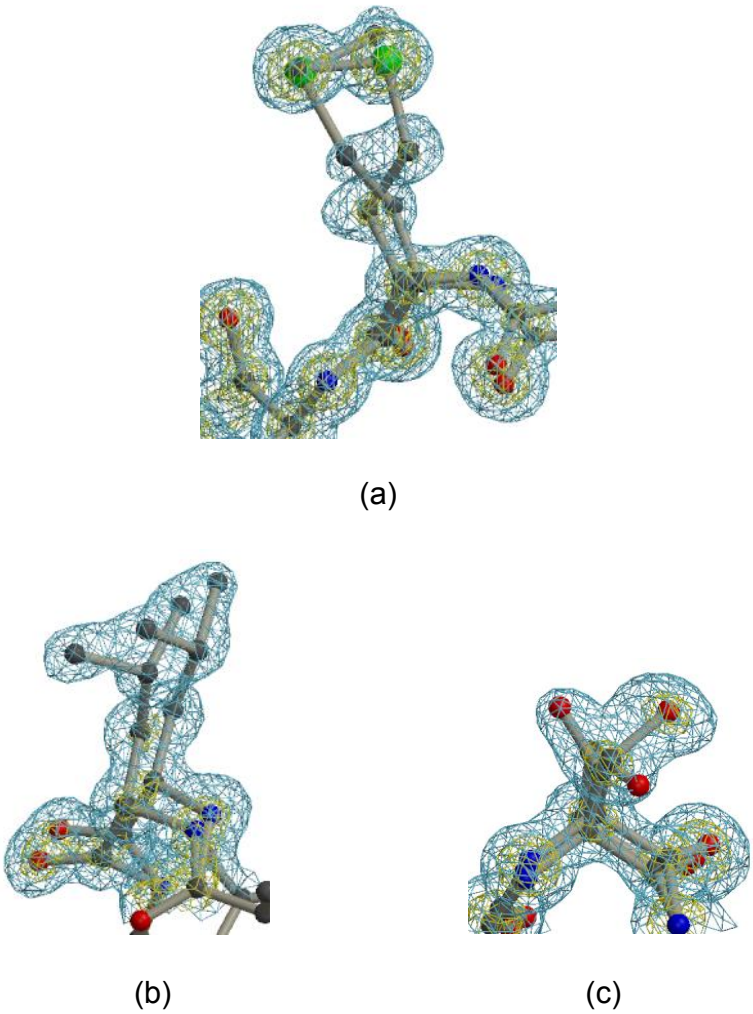
(c)

Figure 3.12 Electron density maps.

σ_A-weighted $2F_o$-F_c electron density map contoured at 1.00 e/Å ^{3} (blue) and 3.0 e/Å^3 (gold). (a) ILE 22-SER 23-ASN 24-PHE 25 (b) HIS 9 (c) GLN 28. The figures were produced with the program *POVscript+* (Fenn *et al.*,2003).

3.5.2. Multiple conformations

The final model consists of 1135 protein atoms, 251 water molecules, 2 sulfate ions, and a glycerol molecule. Multiple conformations are observed for 29 residues in the final model, and the examples of them are shown in Figure 3.13. The distribution of multiple conformations is summarized in Figure 3.14 and Figure 3.15.



(a)

(b) (c)

Figure 3.13 Electron density maps around multiple conformations.

σ_A-weighted $2F_o$-F_c electron density maps contoured at 0.50 e/Å ^{3} (blue) and 2.0 e/Å^3 (gold), (a) Met 111 (b) Leu 107 (C) Ser 123. The figures were produced with the program *POVscript+* (Fenn *et al.*,2003).

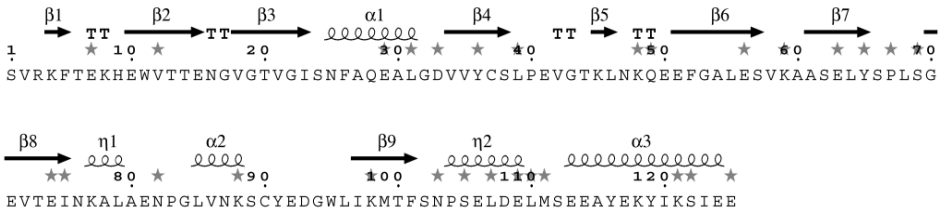


Figure 3.14 The distribution of multiple conformations.

Asterisks represent the residues assigned as a multiple conformation. α, β, η and TT represent α-helix, β-strand, 3_{10}-helix and turn, respectively.

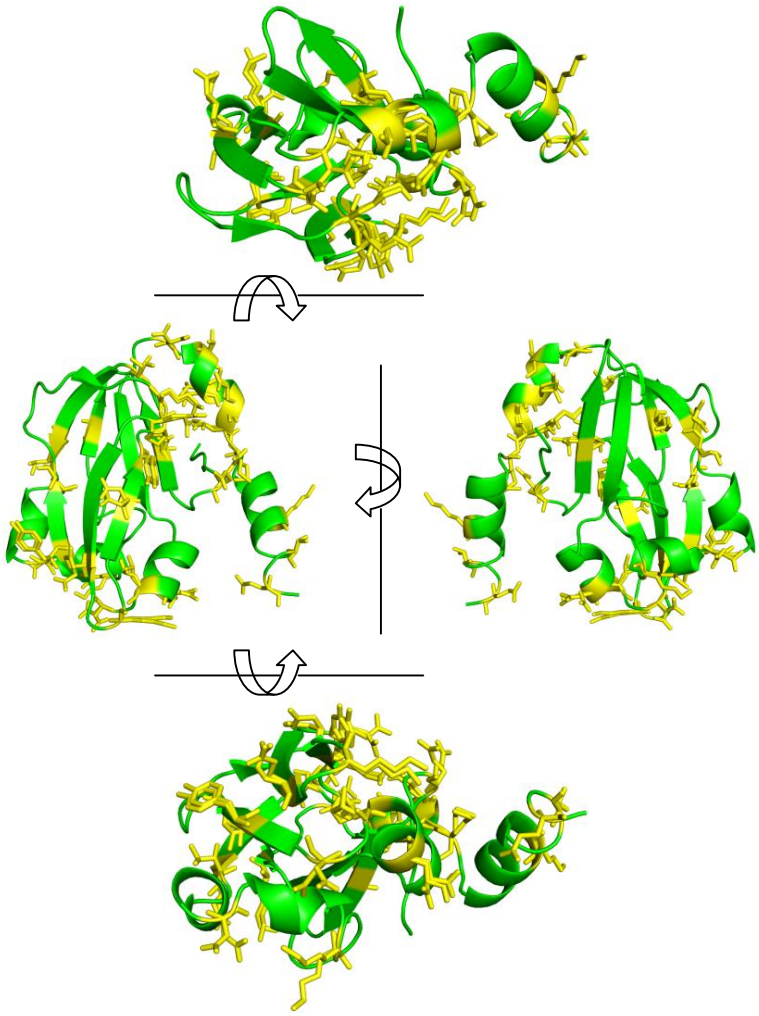
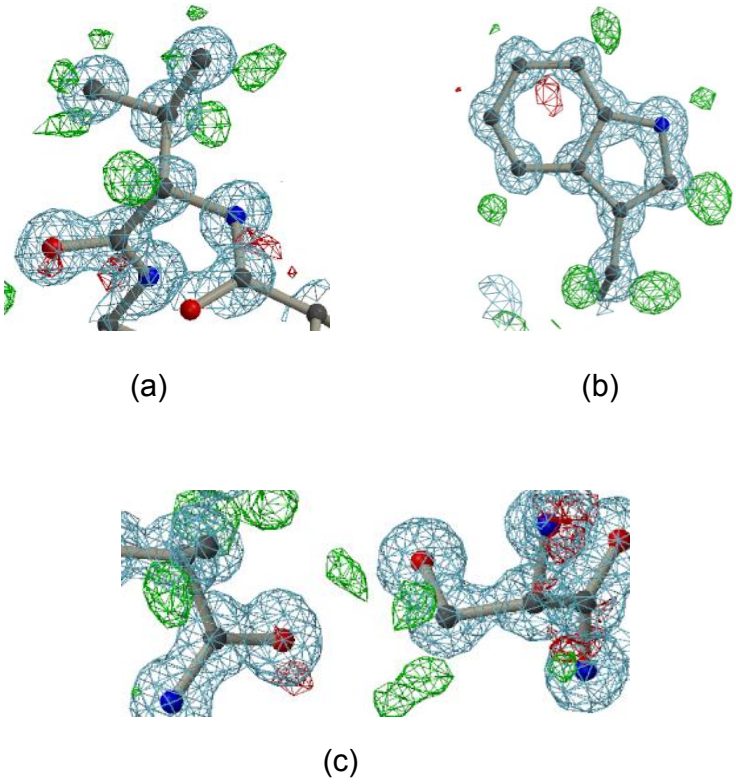


Figure 3.15 The distribution of multiple conformations.

Yellow residues represent the multiple conformations. The figures were produced with the program *PyMOL* (DeLano, 2002).

3.5.3. Hydrogen atoms

Hydrogen atom has only one electron, so that its electron density is relatively poorer than other atoms in the electron density map determined by X-ray crystallography. However, in high-resolution X-ray diffraction data the hydrogen atoms are visualized as a peak of hydrogen-omit electron density map. In this study, refinements were performed against the data of 0.88 Å resolution, and ~40% of hydrogen atoms were visualized (Figure 3.16), hydrogen atoms bonded to atoms of multiple conformations, waters and other solvent molecules were excluded from the counting.



(a) (b)

(c)

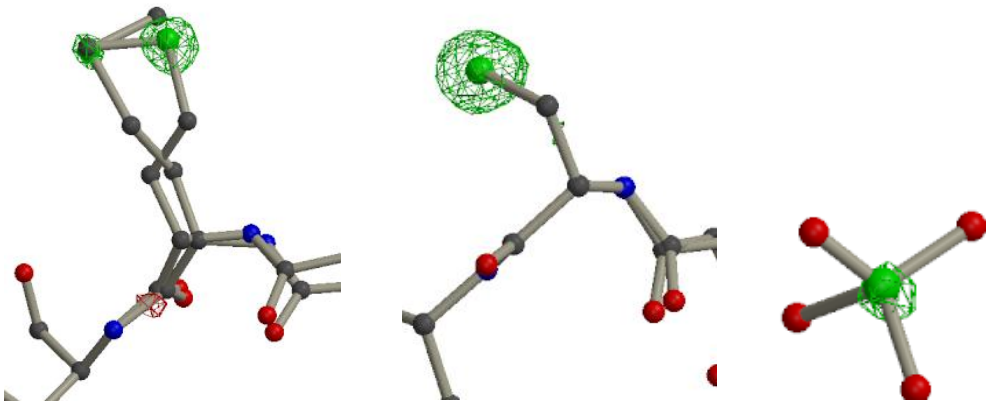
Figure 3.16 Electron density maps around regions finely observed hydrogen atoms. σ_A-weighted $2F_o$-F_c (blue) and F_o-F_c (green or red) electron density countoured at 1.0 e/Å^3 and $\pm$0.18 e/Å^3 (a) VAL 58 (b) TRP 11 (C) SER 31, green meshes represents the electron density of hydrogen atoms. The figures were produced with the program *POVscript+* (Fenn *et al.*,2003).

3.5.4. Anomalous scattering from sulfate atoms

Anomalous scattering occurs if the electrons in an atom cannot be regarded as free electrons. The atomic scattering factor, f is expressed as a sum of the normal scattering factor, f^0, and anomalous scattering term, f' and f''.

$$f = f^0 + f' + if''$$

The imaginary component of anomalous scattering term, f'' is proportional to the atomic absorption coefficient of the atom at X-ray wavelength. The K-absorption edge for sulfate atom is at 5.016 Å, so that the anomalous scattering for sulfur atoms is very weak (f''=0.24 e) at the experimental wavelength of 1.0 Å. Nevertheless, the anomalous electron density of sulfate atoms were clearly visualized (Figure 3.17), this results showed that the structural refinement were precisely finished.



(a) (b) (c)

Figure 3.17 Anomalous peaks of sulfate atoms.

Anomalous difference electron density map countoured at +0.03 e/Å^3 (green or red), the data collections were performed using 1.0Å wavelength X-ray. (f''=0.24 e), (a) Met 111 (b) Cys 90 (c) SO_4 5002. The figures were produced with the program *POVscript+* (Fenn *et al.*,2003).

Chapter 4
Optimization of merging procedure

4.1. Quality of high-, mid- and low-resolution data set

Three data sets were collected at resolution cut-offs of 0.88 Å (high-resolution data set), 0.99 Å (mid-resolution data set) and 1.96 Å (low-resolution data set). The details of data collections are shown in Table 3.2. Each data was integrated, scaled and merged using the program *DENZO* and *SCALEPACK* as implemented in the *HKL*-2000 program package. The statistics of each data set are given in Table 4.1. The overall R_{merge} values on intensities for high-, mid- and low-resolution data were 4.7, 3.2 and 4.4%, and the completeness were 95.6, 97.9 and 99.5%, respectively. The resolution limit (0.88 Å) of the high-resolution data was determined based on the R_{merge} (~40%) of the shell; $\langle I \rangle / \langle \sigma(I) \rangle$ values of the highest resolution shells were larger than 1.5 in all data sets. Figure 4.1(a) shows the completeness as a function of resolution for the high-, mid- and low-resolution data sets. The high- and mid-resolution data sets were nearly complete at resolution greater than ~2.0 Å. However, the completeness was not satisfactory at lower resolution (> ~15 Å), probably due to saturation of intensities and the experimental conditions. The completeness of the low-resolution data set shows that the reflections of lowest resolution could be measured without a front-beam stop between the crystal and the detector. Figure 4.1.(b) shows the $\langle I \rangle / \langle \sigma(I) \rangle$ as a function of resolution for high-, mid- and low-resolution data sets. The $\langle I \rangle / \langle \sigma(I) \rangle$ values of three data sets were not equivalent at the same resolution range, because of the differences among the experimental conditions for each data set. Figure 4.1(c) shows the R_{merge} as a function of resolution for high-, mid- and low-resolution data sets. The insufficiency of diffraction intensities from short X-ray exposure time increased in

the R_{merge} of mid-resolution data set at ~1.0 Å resolution. Similarly, the R_{merge} for the low-resolution data set was higher than the R_{merge} for the high-resolution data set at ~2.0 Å resolution. Wilson plots for the three data sets are shown in Figure 4.1(d) The large differences observed in the absolute scale and *B*-factors of the three data sets reflect the conditions under which the data were recorded. To obtain a complete data set at overall resolution (43.1-0.88 Å), these three data sets had to be merged and scaled carefully.

Table 4.1 Scaling statistics of high-, mid-, and low-resolution data sets.

Data of highest-resolution shells are in parentheses.

Data set	High-resolution	Mid-resolution	Low-resolution
Space group		*C*2	
Cell dimensions (Å, °)	*a* = 84.41	*a* = 84.61	*a* = 84.41
	b = 41.25	*b* = 41.34	*b* = 41.30
	c = 43.05	*c* = 43.16	*c* = 43.14
	β = 91.18	β = 91.10	β = 91.14
Resolution range (Å)	14.9-0.88	20.4-1.00	43.1-1.96
	(0.89-0.88)	(1.00-0.99)	(1.98-1.96)
No. of observed reflections	690,629	277,583	38,945
No. of unique reflections	111,864 (3,714)	81,155 (1,903)	10,767 (295)
Redundancy	6.2 (3.7)	3.4 (2.7)	3.6 (3.0)
Completeness (%)	95.6 (95.4)	97.9 (69.3)	99.5 (86.3)
$\langle I \rangle / \langle \sigma(I) \rangle$	47.4 (2.64)	34.3 (1.66)	49.2 (25.2)
R_{merge} (%)	4.7 (42.6)	3.2 (47.6)	4.4 (6.1)
No. of reject reflections	12091	127	233
Wilson B (Å^2)	6.64	7.92	13.8

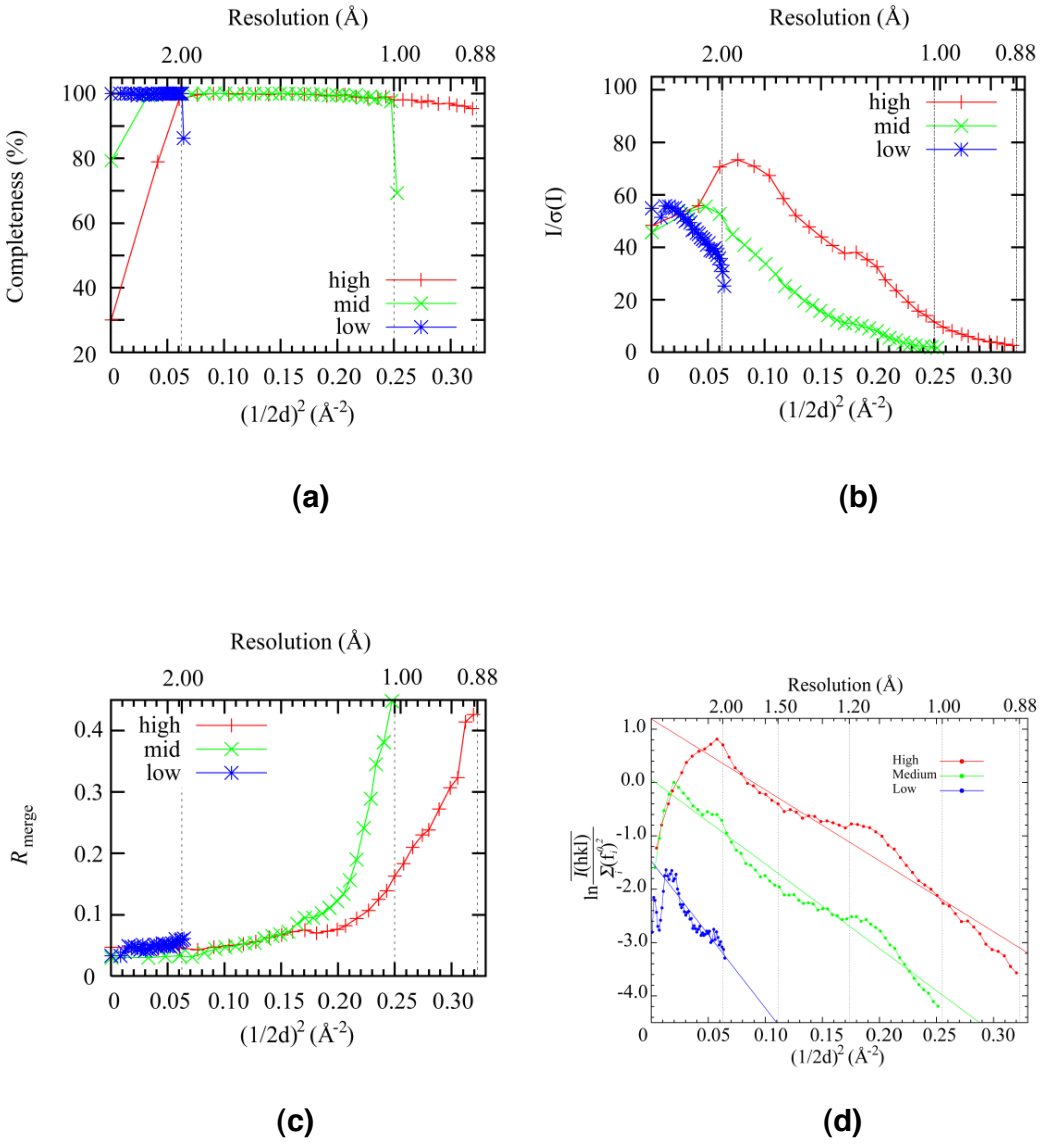


Figure 4.1 Scaling statistics of each data set.

(a) Completeness, (b) <*I*>/<$\sigma(I)$>, and (c) R_{merge} as a function of resolution.

d represents the lattice plane distance (Å). (d)Wilson plots. The lines represent; Red: y=2(-6.64)x+1.12, Green: y=2(-7.92)x+0.0554, Blue: y=2(-13.8)x-1.48, where y and x are the vertical and horizontal axis, respectively. I, f_i and d represent an intensity, atomic scattering factor and lattice plane distance

4.2. Assessment methods of scaled data

It was shown that the values of $<I>/<\sigma(I)>$ of high-, mid- and low-resolution data sets were different and did not overlap (Figure 4.1(b)). In addition, Wilson plots for the three data sets are shown in Figure 4.1(d). The large differences observed in the absolute scale and *B*-factors of the three data sets. To obtain the complete data set at overall resolution, the high-, mid- and low-resolution data sets were attempted to merge using *SCALEPACK* in several merging trials. In order to assess the results of merged data, methods of "Re-refinement" and "Counting hydrogen atoms" were utilized.

4.2.1. Re-refinement

In order to assess the quality of merged data correctly, random-coordinate noise at a maximum of 0.1 Å was added to the final model using the program *PDBSET* (Collaborative Computational Project, Number 4, 1994). When the input file for *SHELXL* refinement was generated by *SHELXPRO,* all anisotropic *B*-factors were converted to isotropic *B*-factors, and the partial occupancies of water molecules were replaced to unity. The re-refinement was performed by *SHELXL* using an input file derived from the coordinate without any biases. The first step was 30 cycles of an isotropic refinement against the data up to maximum resolution, followed by 30 cycles of an anisotropic refinement against the same data. Subsequently, the refinement with riding hydrogen atoms and determination of water molecule occupancies were 30 cycles, respectively. The *R* factors of each resolution shell were calculated using *SHELXPRO* as a function of 30 resolution shells. The re-refinement protocol is shown in Figure 4.2.

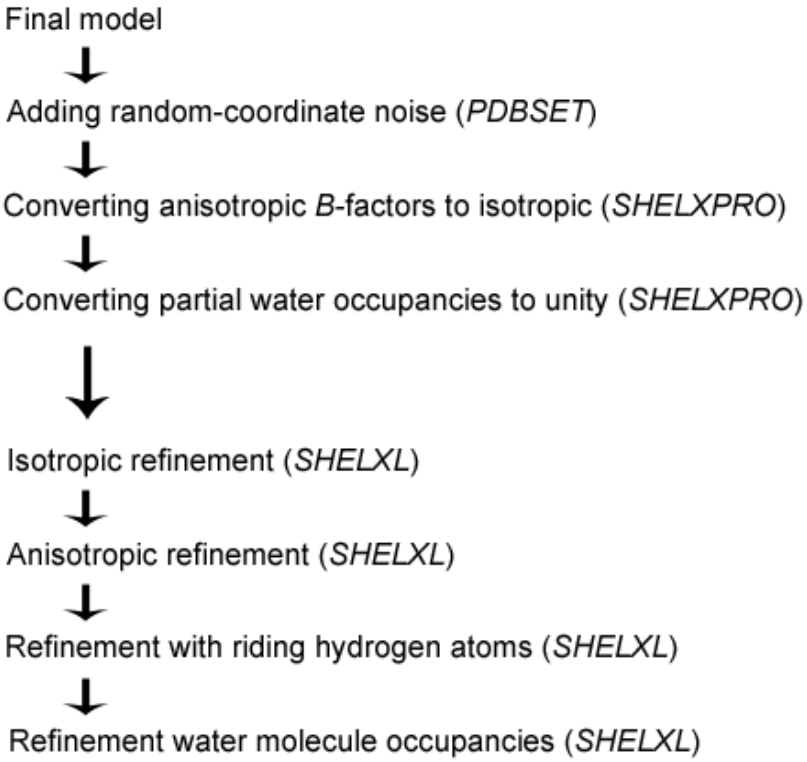


Figure 4.2 Re-refinement protocol

4.2.2. Counting hydrogen atoms with significant electron density

In the refinement process, hydrogen atoms were added and subjected to refinement as a riding model. In order to avoid bias that might result from the riding model treatment of hydrogen positions (Wang *et al.*, 2007), random-coordinate noise with a maximum of 0.1 Å displacement was added to the final model using the program *PDBSET.* All anisotropic *B*-factors were converted to isotropic *B*-factors by *SHELXPRO*. The re-refinement was performed by *SHELXL*, using the input file derived from the coordinates with random errors. The first step was 30 cycles of an isotropic refinement against the data to maximum resolution, followed by 30 cycles of an anisotropic refinement against the same data. The structure factors of F_o-F_c and the phase were exported by *SHELXPRO* and the exported data were converted to the MTZ format using the program *F2MTZ* (Collaborative Computational Project, Number 4 1994). The σ_A-weighted F_o-F_c difference Fourier electron density maps, sampled at 864, 432, 432 grid positions along the unit cell edges were generated. Thus, the size of each grid was smaller than $0.1\times0.1\times0.1$ Å. The number of electrons appearing at the calculated positions of atoms were determined using the program *MAPMAN* (Uppsala Software Factory, RAVE package) with the 'PEEK VALUE' command including an 'INTERPOLATE' option. Hydrogen atoms bonded to protein atoms with partial occupancies, water molecules and small molecules were ignored to avoid miscounts. The threshold for determining hydrogen atoms was 0.16 electrons. This threshold agrees with the result of a manual counting of the hydrogen atoms. The schematic of re-refinement protocol are shown in Figure 4.3.

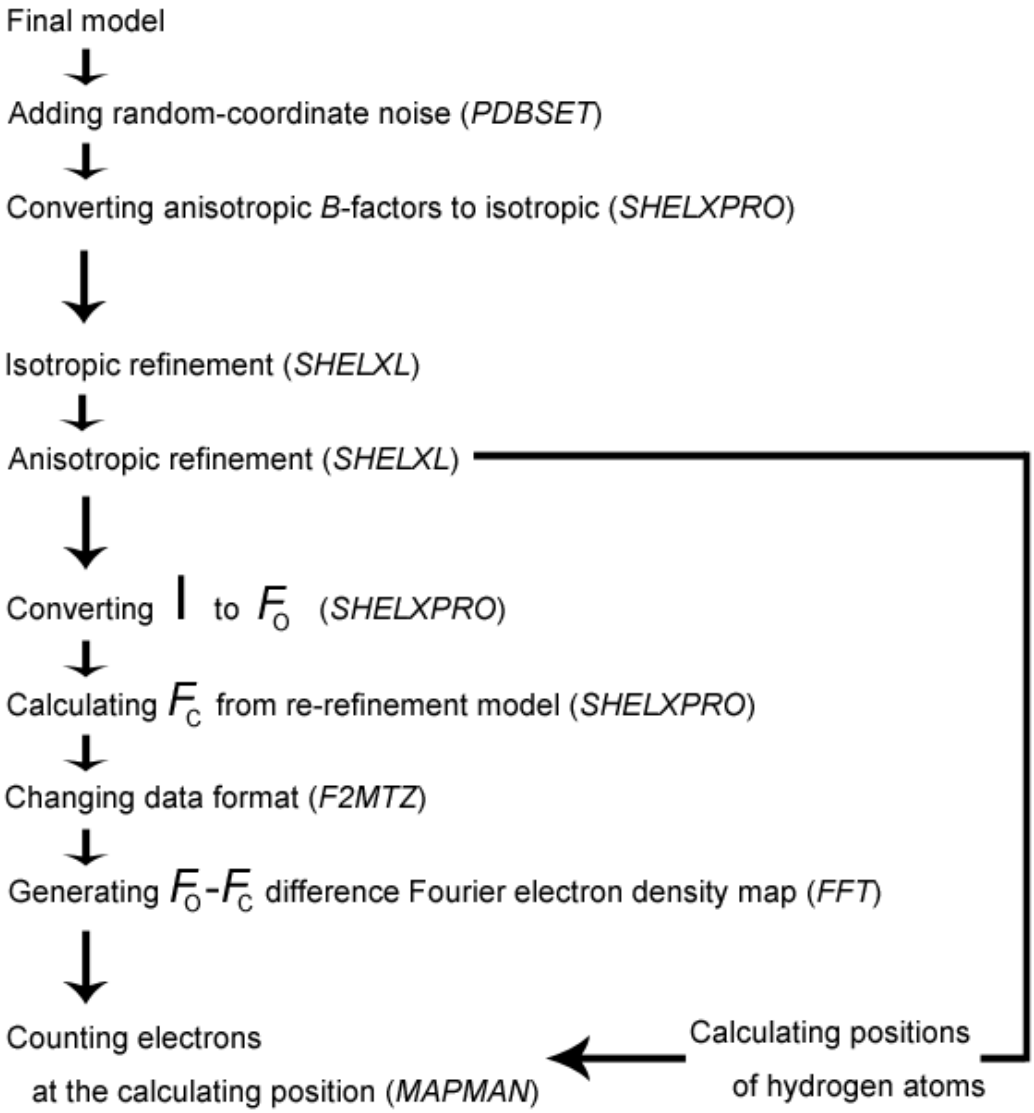


Figure 4.3 Protocols of counting hydrogen atoms.

I, F_o and F_c represents the intensity, observed and calculated structure factor, respectively.

4.3. Reference data for scaling and merging

To obtain the complete data set, the high-, mid- and low-resolution data sets were merged using *SCALEPACK* with the 'REFERENCE BATCH' option. The high-, mid- and low-resolution data sets were individually scaled and merged, and used as the reference.

4.3.1. Quality of merged data set

The scaling and merging statistics are given in Figure 4.4 and Table 4.2. The overall completeness was similar for the three data sets. The overall R_{merge} values on intensities were 4.7, 4.5 and 4.6% for the data obtained using high-, mid- and low-resolution data as the reference data. When the mid-resolution data was used as the reference data in merging, the R_{merge} value was slightly better than the others. On the other hand, R_{merge} values in the highest resolution shells were 38.6, 44.5 and 54.0% for the high-, mid- and low-resolution reference data, respectively. The overall $<I>/<\sigma(I)>$ values were 70.0, 70.8, 82.4 and the values in the highest-resolution shell were 3.72, 2.45, 2.60, for the high-, mid- and low-resolution reference data, respectively. The merged data set obtained using the high-resolution data as reference showed the highest $<I>/<\sigma(I)>$ value in the highest-resolution shell, but the overall value was the lowest. In contrast, the data set obtained using the low-resolution data as reference showed lowest value in the highest-resolution shell and the highest value in overall. The data set obtained using the mid-resolution data as reference had the lowest value in the outer shell. Wilson *B* values of merged data obtained using high-, mid- and low-resolution data as reference were significantly different (Table 4.2), and Wilson plots of merged data sets showed significant differences at the high-resolution range (Figure 4.4(d)). It could not be judged by these statistics which reference was the best with respect to the final

merged data.

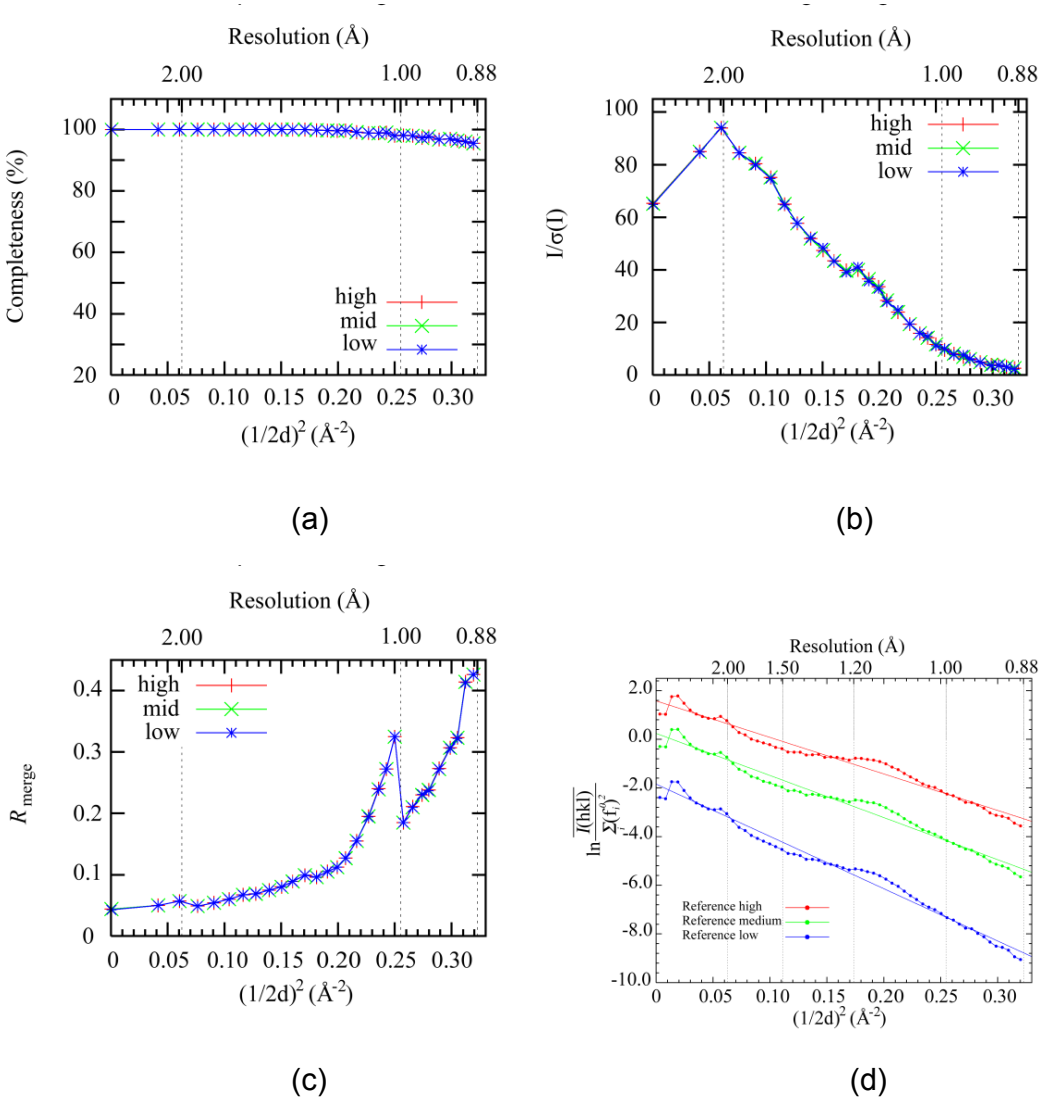


Figure 4.4. Merging statistics of each reference batch.

(a) Completeness, (b) $\langle I \rangle / \langle \sigma(I) \rangle$, and (c) R_{merge} as a function of resolution. d represents the lattice plane distance (Å). (d) Wilson plots. The lines represent; Red: y=2(-7.50)x+1.58, Green: y=2(-8.65)x+0.241, Blue: y=2(-10.7)x+1.84, where y and x are the vertical and horizontal axis, respectively. I, f_i and d represent an intensity, atomic scattering factor and lattice plane spacing (Å).

Table 4.2. Merging data using high-, mid- or low-resolution data as a reference batch.

Data of highest-resolution shells are in parentheses.

Reference data set	High-resolution	Mid-resolution	Low-resolution
Space group		*C*2	
Cell dimensions (Å, °)	*a* = 84.41	*a* = 84.41	*a* = 84.41
	b = 41.25	*b* = 41.34	*b* = 41.31
	c = 43.05	*c* = 43.16	*c* = 43.14
	β = 91.18	β = 91.18	β = 91.14
Resolution range (Å)		43.1 - 0.88 (0.89 - 0.88)	
No. of observed reflections	1,113,257	1,082,860	1,025,526
No. of unique reflections	115,668 (3,712)	116,414 (3,683)	116,058 (3,690)
Redundancy	9.6 (4.7)	9.3 (3.4)	8.8 (3.6)
Completeness (%)	98.9 (95.4)	98.8 (95.0)	98.9 (95.8)
$\langle I \rangle / \langle \sigma(I) \rangle$	70.0 (3.72)	70.8 (2.45)	82.4 (2.60)
R_{merge}	4.7 (38.6)	4.5 (44.5)	4.6 (54.0)
No. of reject reflections	22,060	25,148	21,843
Wilson *B* ($Å^2$)	7.50	8.65	10.7

4.3.2. *R* factors after re-refinement

Using the final model with random errors, the re-refinements were performed against the merged data with the reference from high-, mid- or low-resolution data set. The R/R_{free} factors were 10.0/11.8, 9.9/11.9 and 9.9/12.1 for reflections with $F_o > 4\sigma(F_o)$, and were 11.1/13.2, 10.9/13.1 and 11.2/13.6 for all reflections from the merged data with references from high-, mid- and low-resolution data, respectively. Figure 4.5 shows the R_{free} factor as a function of resolution for each merged data set. The *R* factors for the merged data with the reference from low-resolution data were much higher than for the merged data with references from high- and mid-resolution data at a resolution range of 1.07-0.88 Å. The shortage of overlap reflections between high- (or mid-) and low-resolution data increased in the *R* factors. These results show that differences in the reference batch particularly affected the *R* factors at high resolution. In conclusion, these results suggest that reflections in the high-resolution range are affected by the reference intensities in merging; furthermore, merged data with a reference from high-resolution data produces better data for structural refinement.

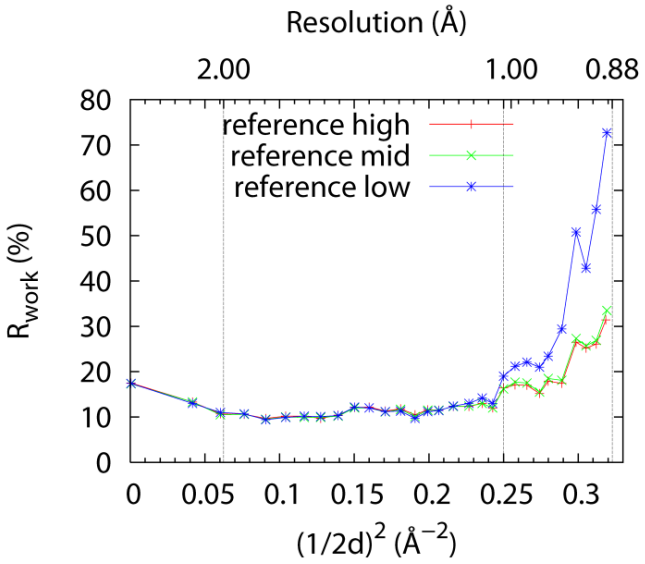


Figure 4.5. R_{free} factors after re-refinement.

R_{free} factors after re-refinement against the data using references from high-, mid- and low-resolution image as a function of resolution. d represents the lattice plane spacing (Å).

4.3.3 Counting hydrogen atoms

In ultra-high-resolution X-ray crystallography, significant electron density for a large number of hydrogen atoms can be observed in σ_A-weighted F_o-F_c hydrogen-omit maps. To assess the quality of merged data sets obtained using the high-, mid- and low-resolution data sets as a reference, the numbers of visualized hydrogen atoms were counted in σ_A-weighted F_o-F_c hydrogen-omit maps. Results are given in Table 4.3, and parts of maps are shown in Figure 4.6. The numbers of visualized hydrogen atoms are 279, 226 and 170 for the high-, mid- and low-resolution reference, respectively; the number for the high-resolution reference is about 1.6 times larger than the number for low-resolution reference. The reliabilities of structures using each data set are validated by the R values in Table 4.3.

4.3.4 Applying negative *B*-factors

Significant differences in the number of visualized hydrogen atoms and the Wilson *B* values were observed among data sets obtained using different data as the reference data for merging. To clarify the effect of Wilson *B* values in the visualization of hydrogen atoms, we modified the merged data obtained using the low- or mid-resolution data as the reference data as follows.

$$F_{\mathrm{B-map}} = \exp\left(-B\sin^2\theta/\lambda^2\right)\times\left(mF_0 - DF_c\right)$$

where $F_{\mathrm{B\text{-}map}}$ is the structure factor for σ_A-weighted F_o-F_c map with a negative *B* factor and mF_o-DF_c is the coefficient for calculation of the standard σ_A-weighted F_o-F_c map. The negative *B* factors increased the number of visualized hydrogen atoms in the low- and mid-resolution reference data sets (Table 4.3). In addition, Figures 4.6(D) and (E) show that the quality of maps with negative *B* values is almost equal to the map from the high-resolution reference (Figure 4.6(A)).

These results showed that the visualization of hydrogen atoms were sensitive to the *B* values in hydrogen-omit maps. In general, the overall *B* values were affected by radiation damage. In our data collection, radiation damage was obvious, especially in the high-resolution data set, and the scaling *B* values were increased by 1.4 Å^2 in the final frame of high-resolution data sets relative to the initial frame. To consider the effect of the radiation damage and the selection of reference data set separately, high-, mid- and low-resolution data sets were merged using typical images as references. The initial and the final batches of the high-resolution data set, and the initial batch of the mid- and low-resolution data sets, were chosen for the reference batch; the resulting Wilson *B* values of the merged data were 7.0, 8.4, 8.5 and 11.5 Å^2, respectively (Figure 4.7). These images were collected at the same position of the crystal. Compared with the total exposure time of high- or mid-resolution data sets, the total exposure time of the low-resolution data set was very short; therefore, the increase in *B* factor

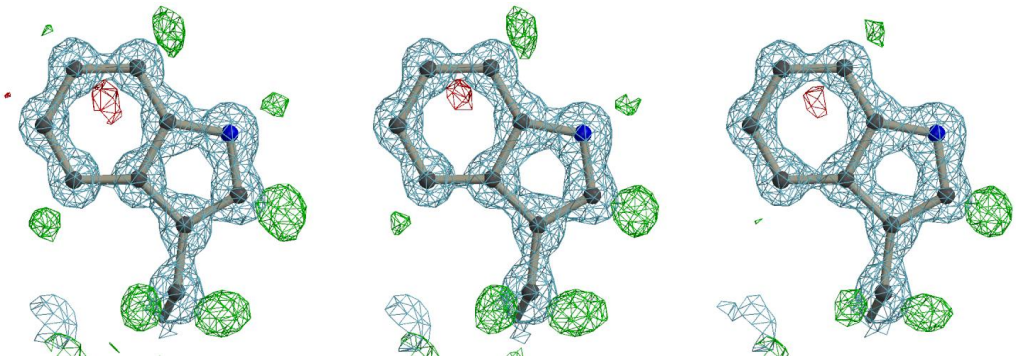
must be smaller than that for the high- or mid-resolution data set. Nevertheless, the Wilson *B* factor of the merged data sets using low-resolution data as a reference was higher than the others. In the merging process, the choice of a reference data set has to be completely irrelevant, so the higher Wilson *B* factor of the low-resolution reference data might be due to the limitation of the scaling algorithm. These results showed the choice of reference data set was important.

In addition, when we applied larger negative *B* values, hydrogen densities were emphasized but noise also increased significantly (Figure 4.8).

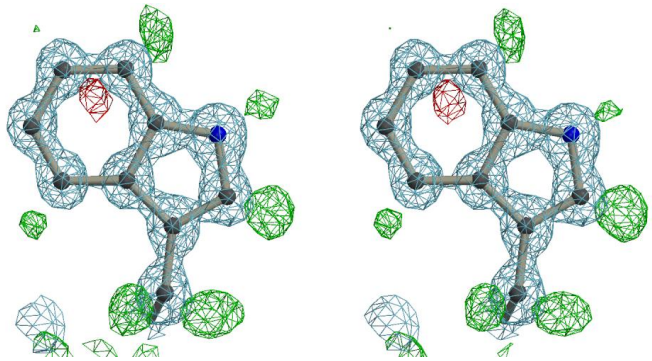
Table 4.1. *R* values against re-refinement models and the number of visualized hydrogen atoms for data sets with a reference.

†*R* factors in parentheses are for reflections with $F_o > 4\sigma(F_o)$.

Data set	R_{work}/R_{free} after re-refinement (%)	Applied B factor (Å^2)	No. of visualizing hydrogen atoms (%)	RMS deviation for H-omit map (e/Å^3)
high	12.2(11.2)/14.4(13.0)	0.0	279 (41.5)	0.0682
mid	12.1(11.1)/14.4(13.2)	0.0	226 (33.6)	0.0648
low	12.4(11.2)/14.5(13.5)	0.0	170 (25.3)	0.0615
mid (-B)	-	-1.15	264 (39.3)	0.0675
low (-B)	-	-3.20	268 (39.9)	0.0680



(A) (B) (C)



(D) (E)

Table 4.6 Electron density maps calculated using each reference data.

σ_A-weighted $2F_o$-F_c (blue) and σ_A-weighted F_o-F_c (positive: green or negative: red) electron density maps contoured at 1.00 e/Å^3 and ±0.15 e/Å^3, respectively. The green and red represent the positive and negative peaks. All figures show the map around residue Trp11. (A) Reference high-resolution data set. (B) Reference mid-resolution data set. (C) Reference low-resolution data set. (D) Reference mid-resolution data set with application of B=-1.15 Å^2. (E) Reference mid-resolution data set with application of B=-3.20 Å^2. The figures were produced with the program *POVScript+* (Fenn *et al.*, 2003).

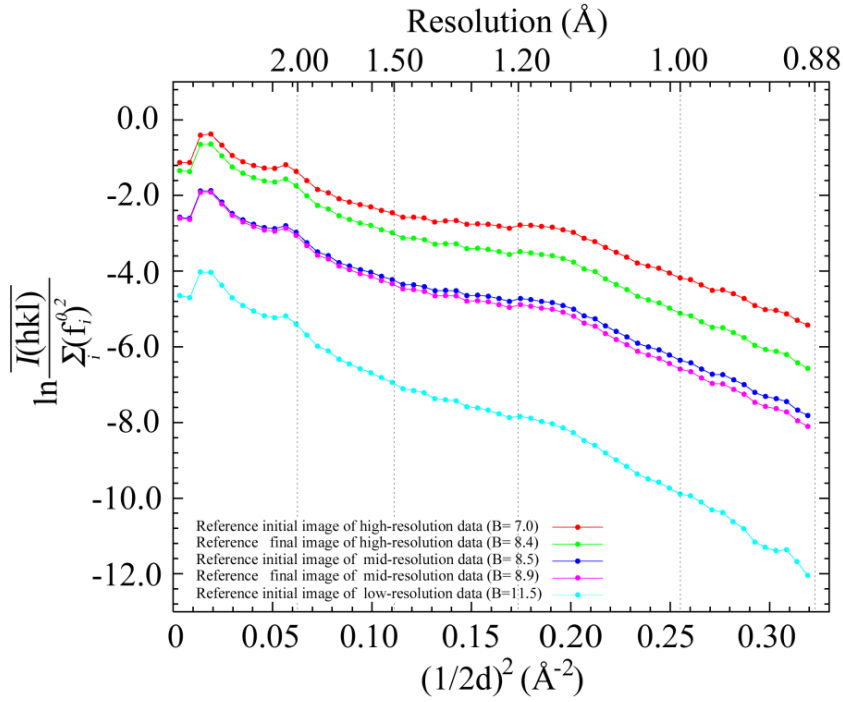


Figure 4.7 The Wilson plots of merged data set using a typical image as a reference.

The *B* means Wilson *B* factor. The lines represent; Red: y=2(-7.0)x-0.18, Green: y=2(-8.4)x-0.81, Blue: y=2(-8.5)x-2.0, Cyan: y=2(-8.9)x-2.1, Light blue: y=2(-11.5)x-4.1 where y and x are the vertical and horizontal axis, respectively. I, f_i and d represent an intensity, atomic scattering factor and lattice plane spacing (Å).

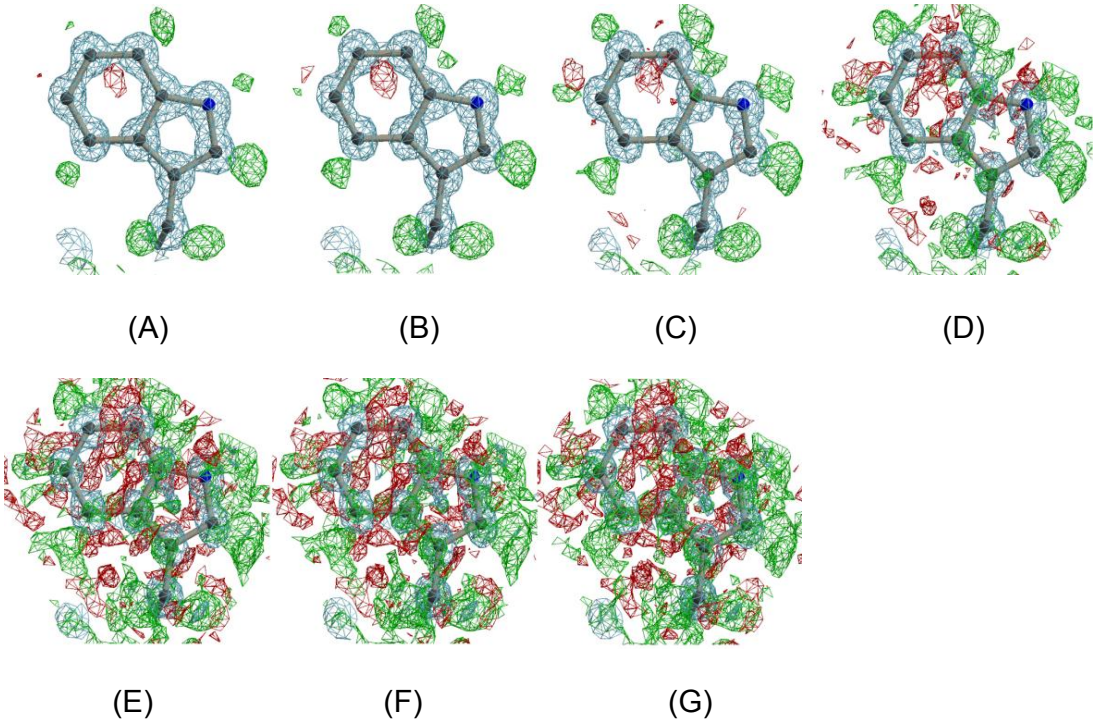


Figure 4.8 Electron density maps which are applied further negative B values.

σ_A-weighted $2F_o$-F_c (blue) and σ_A-weighted F_o-F_c (green or red) electron density maps contoured at 1.00 e/Å^3 and 0.15e/Å^3, respectively. The green and red represent the positive and negative peaks. All figures show the map around residue Trp11. (A), (B), (C), (D) and (E) Reference the high-resolution data set with application of B=0.0, -1.0, -2.0, -3.0, -4.0, -5.0, -6.0 and -7.50 Å^2, respectively. (F) Reference the mid-resolution data with application of B=-8.65 Å^2. (G) Reference the low-resolution data set with the application of B=-10.7 Å^2. The applied *B* values of (E), (F) and (G) are equal to each Wilson *B* value.

4.4. Truncating data set

Figure 4.4(a) shows that the completeness of merged data was high enough for the structural refinement at the overall resolution. In contrast, Figure 4.4(b) and Figure 4.4(c) demonstrate that the curves of R_{merge} and $<I>/<\sigma(I)>$ are not smooth at the resolution edge of low- and mid-resolution data sets. In order to obtain higher-quality merged data, the resolution of mid- and low-resolution data sets were truncated after the integration.

4.4.1. Truncating mid-resolution data set

The integration data from low-resolution data set were truncated using an in-house-written program at around the resolution limit of 1.0 Å. The edges of highest resolution of the mid-resolution data set were cut to 1.0, 1.2, 1.4, 1.6 and 1.8 Å, respectively, and the merging was carried out using *SCALEPACK* with high-resolution data as a reference.

The completeness of merged data with truncated mid-resolution reflections as a function of resolution are shown in Figure 4.9(a). The values of completeness for all cases were good enough at overall resolution. The curves of $I/\sigma(I)$ varied as the resolution edge of mid-resolution data were truncated (Figure 4.9(b)). In the case of the resolution cutoff of 1.0 and 1.2 Å were almost same at overall resolution, and the curve was smooth at the mid-resolution (~1.3 Å). However, in the other cases, the values were smaller, and the curve was not smooth at the each cutoff resolution. The curve of R_{merge} were especially changed (Figure 4.9 (c)). In the case of resolution cutoff of 1.4 Å, the curve was more smooth than other cases at mid-resolution.

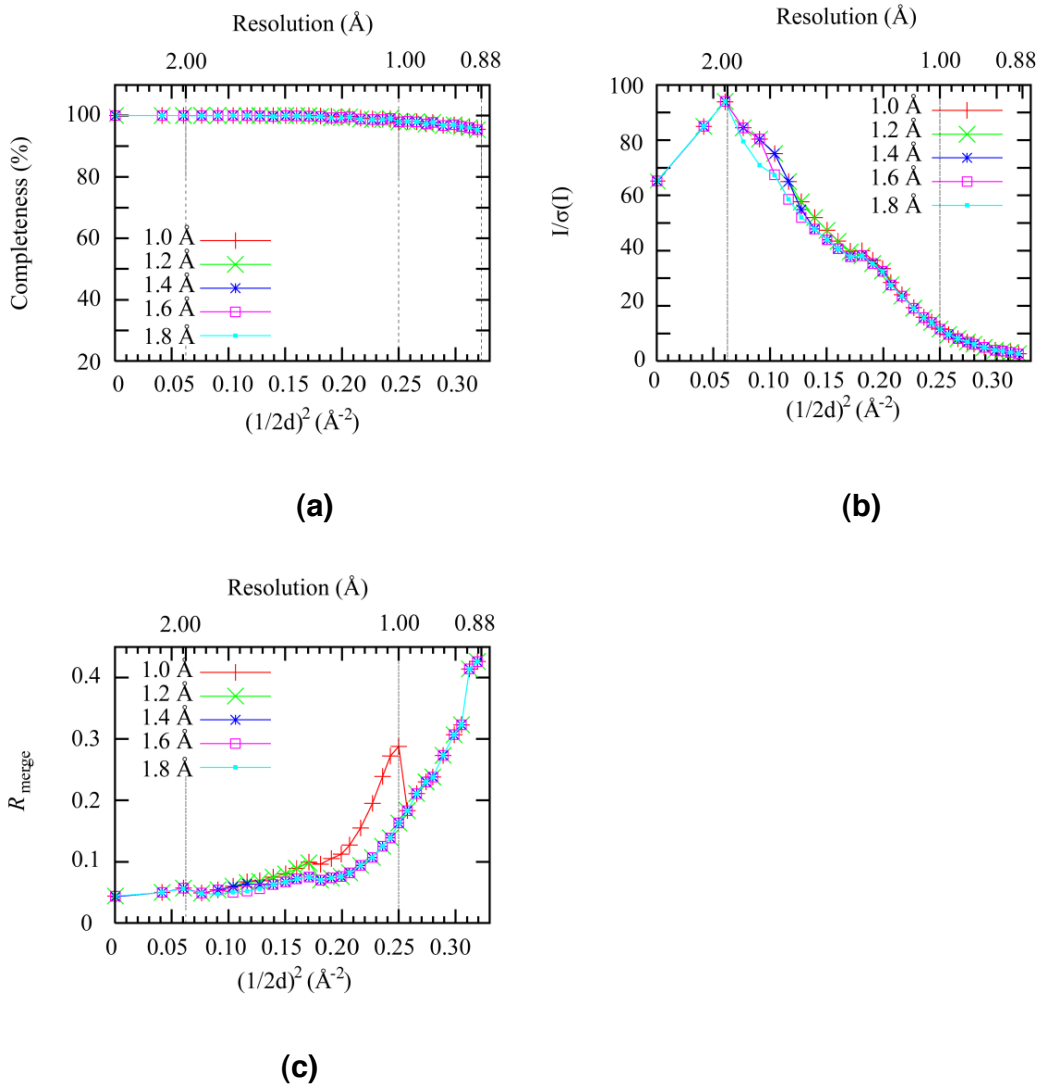


Figure 4.9 Merging statistics with truncated mid-resolution data set.

(a) Completeness, (b) $\langle I \rangle / \langle \sigma(I) \rangle$, and (c) R_{merge} as a function of resolution. d represents the lattice plane spacing (Å).

4.4.2. Truncating low-resolution data set

The integration data from low-resolution data set were truncated using an in-house-written program at around the resolution limit of 2.0 Å. The edges of highest resolution of the low-resolution data set were cut to 2.0, 2.2, 2.4 and 2.6 Å, respectively, and the merging was carried out using *SCALEPACK* using high-resolution data as a reference.

The completeness of merged data with truncated low-resolution reflections as a function of resolution are shown in Figure 4.10(a). The values of completeness for all cases were good enough at overall resolution. The value of $I/\sigma(I)$ for resolution cutoff 2.0 Å was bigger than other cases, and the curve for cutoff 2.6 Å was not smooth at ~2.6 Å resolution (Figure 4.10(b)). The curve of R_{merge} as a function of resolution were given in Figure 4.10(c). In the case of resolution cutoff of 2.4 Å, the curve was more smooth than other cases at low-resolution.

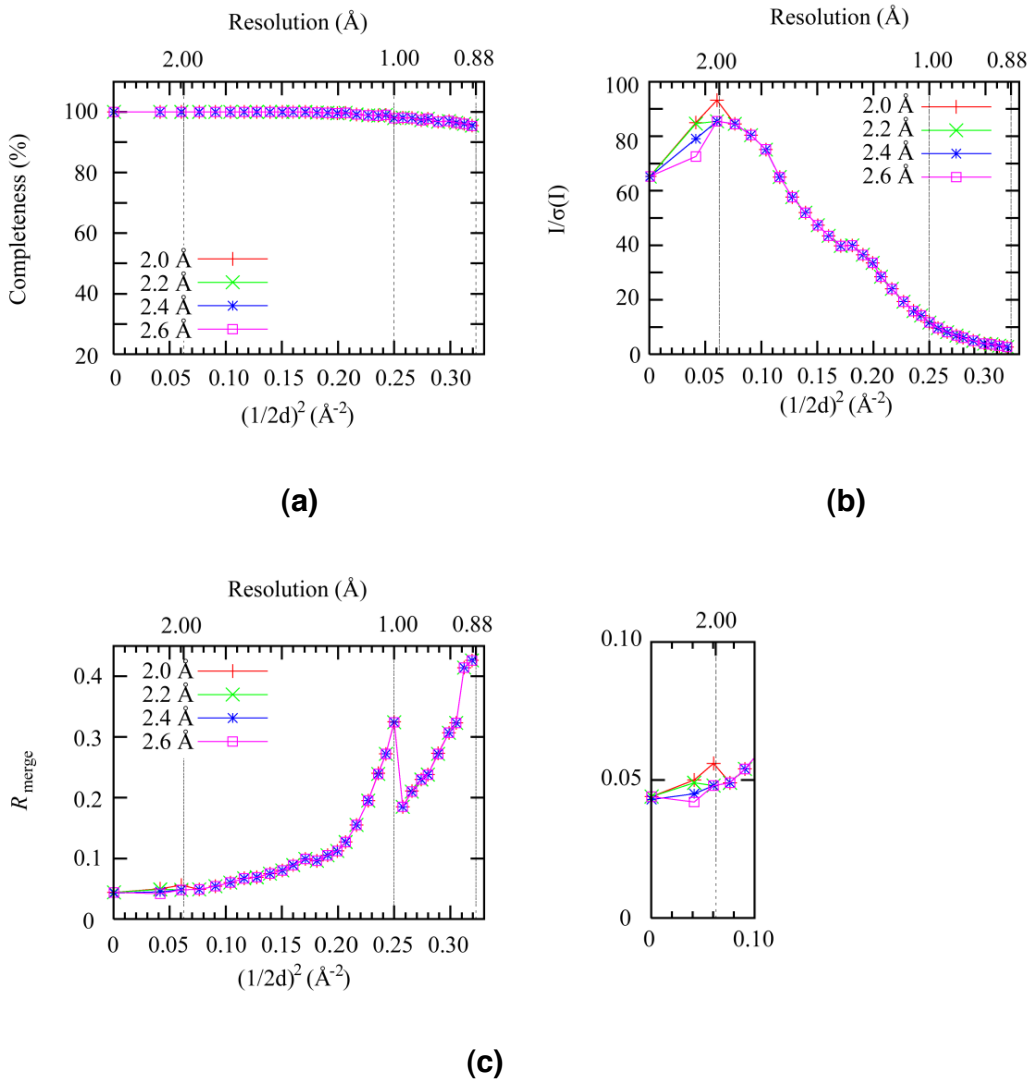


Figure 4.10 Merging statistics with truncated low-resolution data set.

(a) Completeness, (b) $\langle I \rangle / \langle \sigma(I) \rangle$, and (c) R_{merge} as a function of resolution. d represents the lattice plane spacing (Å).

4.4.3. Truncating mid- and low-resolution data set

Four different cases of merged data were tested from the results of 4.4.1 and 4.4.2. The merging was carried out using *SCALEPACK* using high-resolution data as a reference and the integration data were truncated using an in-house-written program.

4.4.3.1. Quality of merged data set

Case 1 was the control case, in which the resolution limits of mid- and low-resolution data sets were not changed. The ranges of mid- and low-resolution data sets were 20.4-0.99 Å and 43.1-1.96 Å resolution, respectively. In Case 2, the edges of highest resolution of the mid- and low-resolution data sets were cut to 1.4 and 2.4 Å resolution, respectively. In the merging with these limits of resolution, the curve of the R_{merge} as a function of resolution was smooth at the edge of resolution of truncated mid- or low-resolution data sets (Figure 4.11(c)). In Case 3, the highest resolutions of the mid- and low-resolution data sets were truncated to 1.2 and 2.0 Å, respectively, before merging the data. In this case, the curve of $\langle I \rangle / \langle \sigma(I) \rangle$ as a function of resolution was smooth at high resolution, and the value of $\langle I \rangle / \langle \sigma(I) \rangle$ was bigger than the other three cases at ~2.0 Å resolution (Figure 4.11(b)). In Case 4, the mid- and low-resolution data sets were cut to 1.8 and 2.6 Å resolution, respectively. This range of resolution is significantly truncated. In this case, the curve of R_{merge} was smooth (Figure 4.11(c)) and the $\langle I \rangle / \langle \sigma(I) \rangle$ was smaller than in the other three cases (Figure 4.11(b)).

The values of completeness for Cases 2, 3 and 4 were slightly reduced (Figure 4.11(a)). The statistics of merged data in the four different cases are summarized in Table 4.4. The number of observed and unique reflections, and the redundancy, were decreased in the truncated data (Case 2, 3 and 4). The

R_{merge} of Case 2 and 4 was improved in the average value.

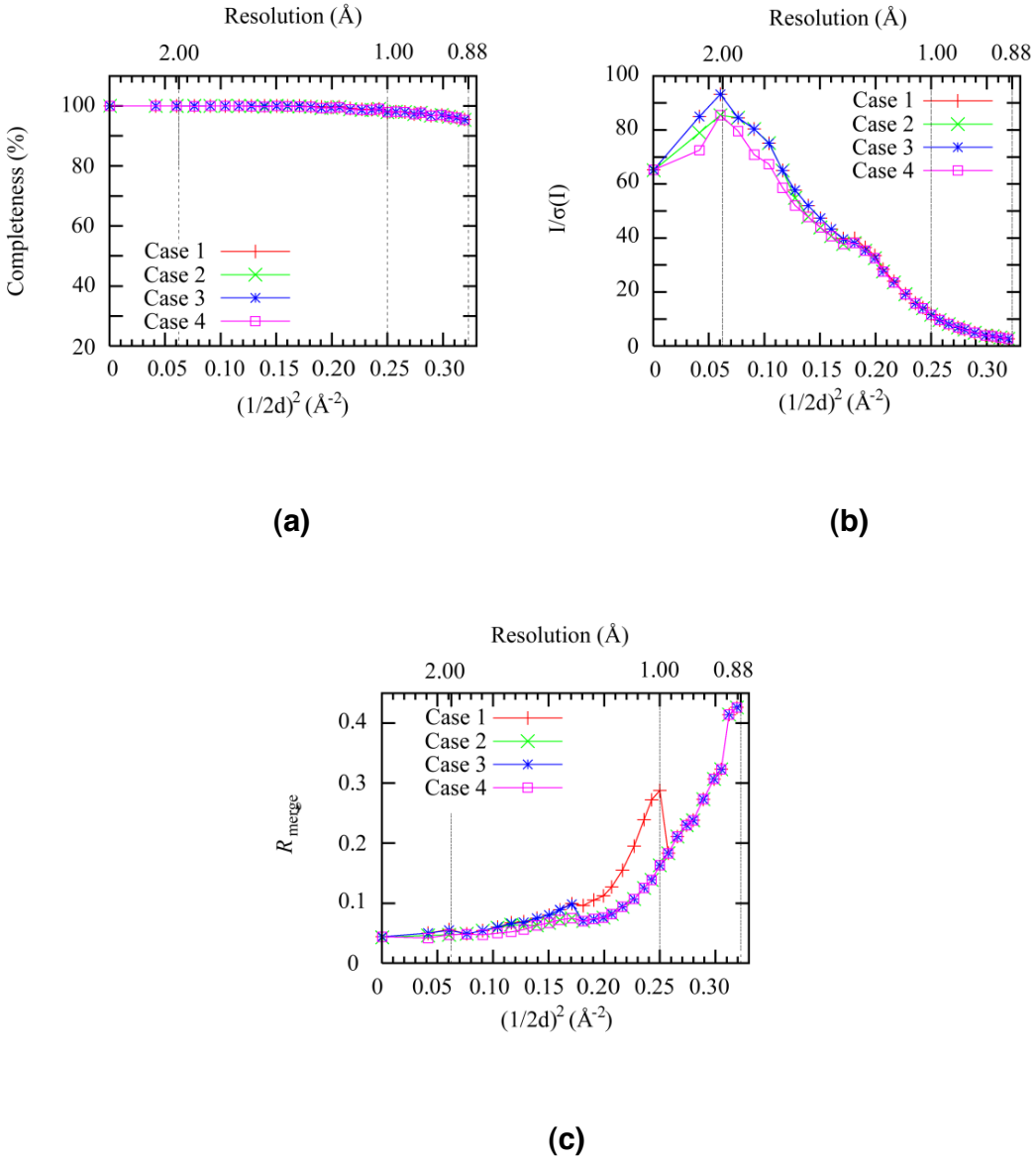


Figure 4.11 Merging statistics for each case.

(a) Completeness, (b) $\langle I \rangle / \langle \sigma(I) \rangle$, and (c) R_{merge} as a function of resolution. d represents the lattice plane distance (Å).

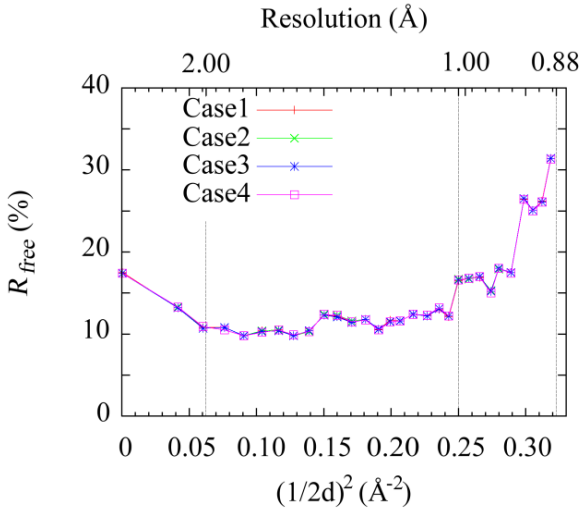
Table 4.4. Merging statistics of each case.

Data of highest-resolution shells are in parentheses.

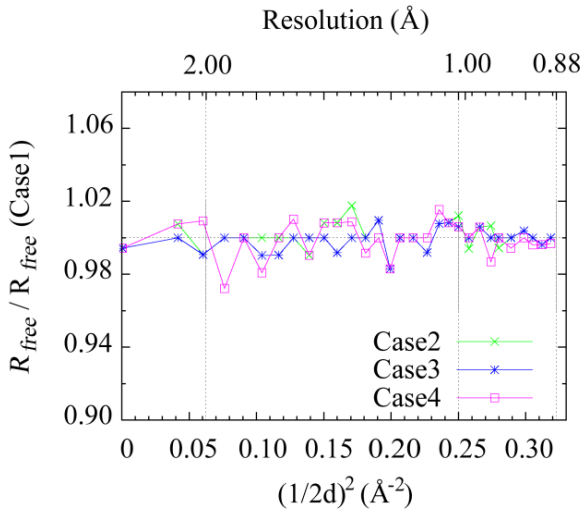
Data set	Case 1 all data	Case 2 based on R_{merge}	Case 3 based on $\langle I \rangle / \langle \sigma(I) \rangle$	Case 4 over truncating
Space group		*C*2		
Cell dimensions (Å, ˚)		a = 84.58, b = 41.29, c = 43.12, β = 91.18		
Resolution range (Å)		43.1 - 0.88 (0.89 - 0.88)		
No. of observed reflections	1,107,549	920,306	995,128	849,487
No. of unique reflections	115,668 (3,712)	115,595 (3,712)	115,615 (3,712)	115,574 (3712)
Redundancy	9.6 (4.7)	8.0 (4.7)	8.6 (4.7)	7.4 (4.7)
Completeness (%)	98.9 (95.4)	98.8 (95.4)	98.8 (95.4)	98.8 (95.4)
$\langle I \rangle / \langle \sigma(I) \rangle$	70.0 (3.72)	68.7 (3.72)	69.9 (3.72)	67.1 (3.72)
R_{merge}	4.7 (38.6)	4.5 (38.6)	4.7 (38.6)	4.4 (38.6)
No. of reject reflections	21,974	20,639	21,790	16,326

4.4.3.2. *R* factors after re-refinement

Using the final model with errors, the re-refinements were performed against the merged data in the four different cases. The resulting values for R/Rfree factors were 10.1/11.9, 10.1/11.9, 10.1/11.9 and 10.1/11.8 for reflections with $Fo > 4\sigma(Fo)$ and were 11.2/13.2, 11.2/13.2, 11.4/13.2 and 11.2/13.2 for all reflections from merged data in the Cases 1-4. Figure 4.12 (a) shows the R factors as a function of resolution for the merged data and curves were almost the same at overall resolution. To evaluate the details of the result, the R factors of Cases 2, 3 or 4 divided by the R factor of Case 1 as a function of resolution are plotted in Figure 4.12 (b). The plot of Cases 2 and 4 were slightly improved at high (< ~1.0 Å) or low resolution (>~2.0 Å) but increased in the mid-resolution (2.0-1.0 Å).



(a)



(b)

Figure 4.12 Rfactor after re-refinement.

(a) R_{work} after re-refinement against the data in each case as a function of resolution. (b) Values of *R* factor were normalized by the *R* factor of Case 1. d represents the lattice plane distance (Å).

4.4.3.3. Counting hydrogen atoms

Structural refinement against a high-resolution data enables us to observe the hydrogen atoms as F_o-F_c electron density in the hydrogen-omit map. To assess the re-refinement coordinates, the hydrogen atoms were counted in the F_o-F_c hydrogen-omit map. The number of observed hydrogen atoms was summarized in Table 4.5. The number of hydrogen atoms in Case 2 was slightly bigger than in the other cases. This result might show that the improvement of R_{merge} at high or low resolution is related to the number of visualized hydrogen atoms.

Table 4.5 Visualizing hydrogen atoms in the F_o-F_c electron density map in each case.

	Number of visualizing hydrogen atoms		
	Resolution range (Å)	Total / (%)	RMS deviation for H-omit map (e/Å^3)
Case 1 (All data)	43.1-0.88	279/ 41.5	0.06831
Case 2 (data based on R_{merge})	43.1-0.88	278 / 41.4	0.06877
Case 3 (data based on $<I>/<\sigma(I)>$)	43.1-0.88	280 / 41.7	0.06839
Case 4 (Over truncated data)	43.1-0.88	277 / 41.2	0.06813

Chapter 5
Combined data sets

5.1. Comparison of combined data

To estimate the advantages of merging high-, mid- and low-resolution data sets, we generated several merged data sets: the high-, mid- and low-resolution data set (HML, 43.1-0.88 Å resolution), the high- and mid-resolution data set (HM, 20.4-0.88 Å resolution), and the high- and low-resolution data set (HL, 43.1-0.88 Å resolution).

5.1.1. Statistics of combined data

The statistics of the HML are given in Table 4.2 (reference High-resolution column), and statistics for HM and HL are given in Table 6. The statistics of the high-resolution data set (H, 13.5-0.88 Å resolution) is summarized in Table 5.1. The completeness values in the lowest resolution shell (43.1-2.73 Å) were 100.0% for HML, 85.0% for HM, 100.0% for HL and 30.0% for H data, respectively. The maximum resolution is the same in all data sets. The merging data were obtained using high-resolution data set as a reference.

Table.5.1 Statistics of merged data sets.

Data of highest-resolution shells are in parentheses.

Data set	†HM	HL
Space group	*C*2	
Cell dimensions (Å, °)	a = 84.61, b = 41.34, c = 43.05, β = 91.18	a = 84.41, b = 41.31, c = 43.14, β = 91.14
Resolution range (Å)	20.3 - 0.88 (0.89 - 0.88)	43.1 - 0.88 (0.89 - 0.88)
No. of observed reflections	1,077,943	839,325
No. of unique reflections	115,067 (3,714)	115,574 (3,712)
Redundancy	9.4 (4.7)	7.3 (4.7)
Completeness (%)	98.3 (95.5)	98.8 (95.4)
$<I>/<\sigma(I)>$	60.8 (3.73)	60.8 (3.72)
R_{merge} (%)	4.4 (46.6)	4.4 (38.6)
No. of reject reflections	17254	15246
Wilson B (Å^2)	7.39	7.49

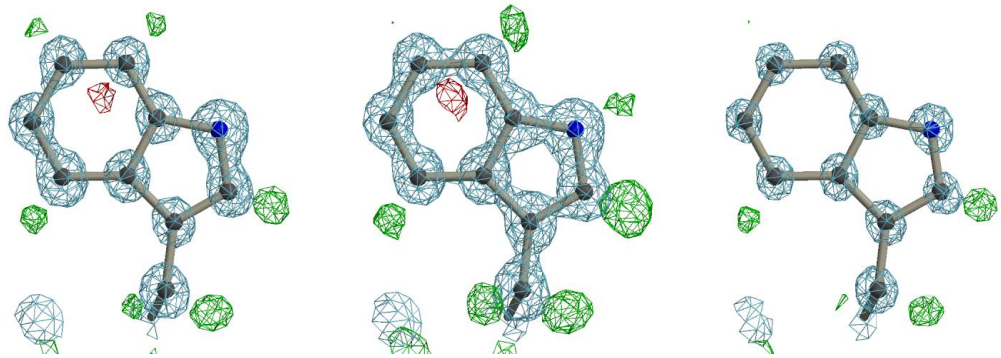
†H, M and L represent High-, Mid- and Low-resolution data sets, respectively.

5.1.2. Counting hydrogen atoms

For each of the combined data sets, the hydrogen atoms were counted in the σ_A-weighted F_o-F_c hydrogen-omit maps. The numbers of visualized hydrogen atoms and the resulting values for R and R_{free} factors of the re-refinements are summarized in Table 5.2. The number of visualized hydrogen atoms was largest in the HML data and second largest in the HL map. In addition, the number of visualized hydrogen atoms in the HML map was larger than in the HM map, and the number in the HL map is larger than in the H map. These results, with and without the low-resolution data set, suggest that low-resolution data are important for locating hydrogen atoms in σ_A-weighted F_o-F_c hydrogen-omit maps (Figure 4.6(A) and Figure, 5.1(A), (B) and (C)). The completeness values for each combined data set, as a function of resolution, are given in Figure 5.2. It is clear that the incompleteness of low-resolution reflections causes a reduction of visualized hydrogen atoms.

To compare the effect of redundancy for the visualization of hydrogen atoms, the reflections that were not present in the H data set were excluded from the HML merged data. The quality of the reflections of the 'H from HML' data was higher than for the H data due to higher redundancy in the former. In addition, the quality of the 'H from HML' data was the same as for the HML data, whereas the completeness of the 'H from HML' data was lower than that of the HML data at low resolution. Similarly, we also generated the ‘HM from HML’ and ‘HL from HML’ data. Using HM, HL and H from the HML data, the number of visualized hydrogen atoms was counted and the results are summarized in Table 7. The number of visualized hydrogen atoms in high-redundancy data was slightly larger than or nearly equal to the number in the map of low-redundancy data. Consequently, the high redundancy of merged data slightly improved the electron density of hydrogen atoms, but at low resolution the contribution was

much smaller than that of completeness.



(a) (b) (c)

Figure 5.1 Electron density maps generated from combined data sets

σ_A-weighted $2F_o$-F_c (blue) and σ_A-weighted F_o-F_c (positive: green or negative: red) electron density maps. The green and red represent the positive and negative peaks. All figures show the map around residue Trp11. σ_A-weighted $2F_o$-F_c (blue) maps contoured at 1.00 e/Å ^{3}. σ_A-weighted F_o-F_c maps of (A)HM, (B)HL and (C)H contoured at ±0.15 e/Å ^{3}. The figures were produced with the program *POVScript+* (Fenn *et al.*, 2003).

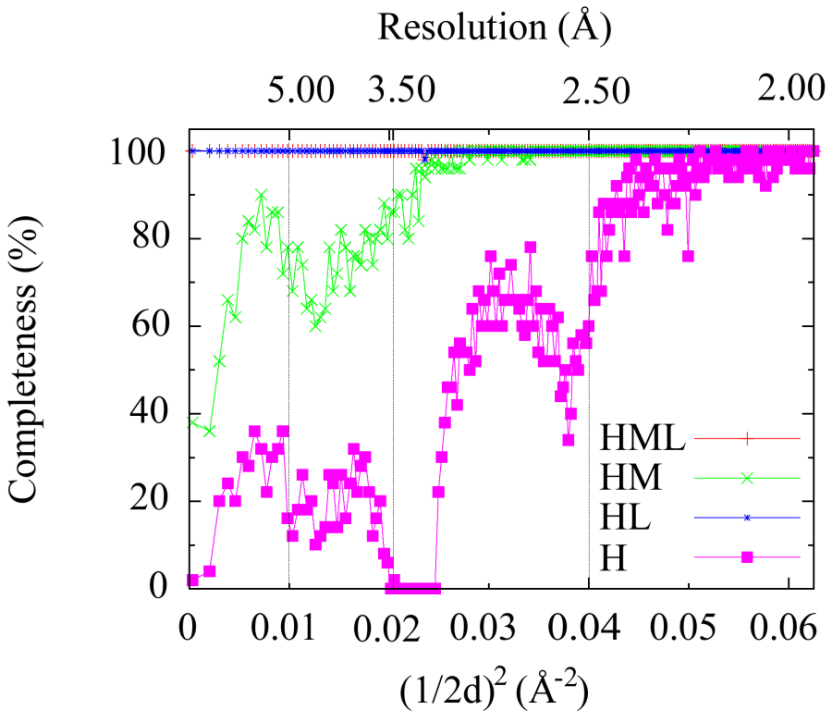


Figure 5.3 Completeness of each combined data set.

Completeness is plotted as a function of resolution at low-resolution. d represents the lattice plane spacing (Å). H, M and L represent High-, Mid- and Low-resolution data sets, respectively.

5.1.3. Applying negative *B*-factors

Figures 5.3(a), (b) and (c) show the σ_A-weighted maps which were calculated with negative *B*-factors. The number of visualized hydrogen atoms was comparable to that for the HML data, but the map quality was poorer than those of HML. In Figures 5.3(d), (e) and (f), the σ_A-weighted F_o-F_c maps contoured at the 2.1σ level were drawn to compare with the HML maps. Although the number of visualized hydrogen atoms was almost the same, the negative peaks were more in the σ_A-weighted F_o- F_c maps of HM, HL and H data.

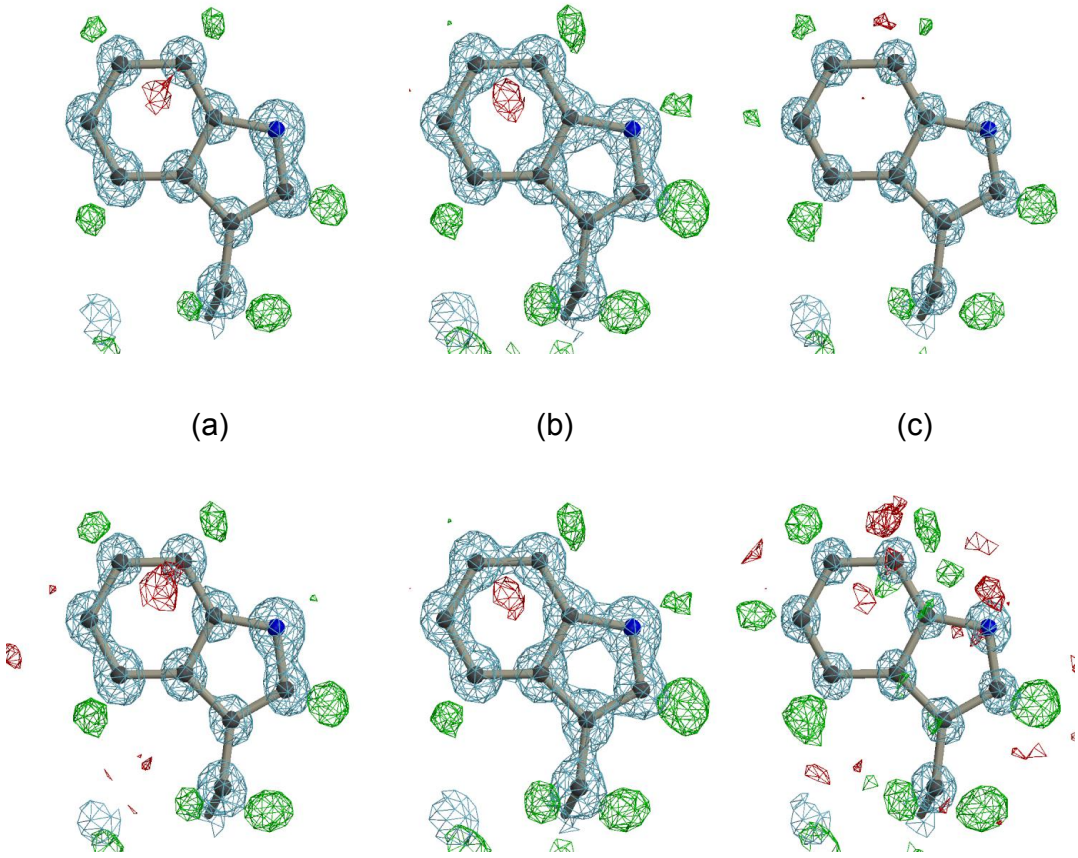


Figure 5.3 Electron density maps generated from combined data sets applied with negative B-factors

σ_A-weighted $2F_o$-F_c (blue) and σ_A-weighted F_o-F_c (positive: green or negative: red) electron density maps. The green and red represent the positive and negative peaks. All figures show the map around residue Trp11. σ_A-weighted F_o-F_c maps of (a) HM, (b)HL and (c)H are generated with increasing of B=-0.5, -0.1 and -1.8 Å^2, and contoured at ±0.15 e/Å^3 . The σ_A-weighted F_o-F_c maps of (d)HM, (e)HL and (f)H were contoured at 2.1σ with applying of the same negative *B*-factors. The figures were produced with the program *POVScript+* (Fenn *et al.*, 2003).

Table 5.2.1 *R* values against re-refinement models and the numbers of visualized hydrogen atoms for combined data sets.

Data set	R_{work}/R_{free} after re-refinement (%)	Applied *B* factor (Å^2)	No. of visualized hydrogen atoms (%)	RMS deviation for H-omit map (e/Å^3)
HML	12.2(11.2)/14.4(13.0)	0.0	279 (41.5)	0.0682
HM	12.1(11.0)/14.7(13.3)	0.0	251 (37.4)	0.0581
HL	12.3(11.2)/14.3(13.0)	0.0	276 (41.1)	0.0683
H	11.9(10.3)/14.0(12.1)	0.0	133 (19.8)	0.0420

[†]H, M and L represent High-, Mid- and Low-resolution data sets, respectively.

Table 5.2.2 *R* values against re-refinement models and the numbers of visualized hydrogen atoms for combined data sets.

Data set	R_{work}/R_{free} after re-refinement (%)	Applied *B* factor (Å^2)	No. of visualized hydrogen atoms (%)	RMS deviation for H-omit map (e/Å^3)
HM from HML	12.1(11.0)/14.7(13.3)	0.0	248 (36.9)	0.0584
HL from HML	12.2(11.2)/14.3(13.0)	0.0	278 (41.4)	0.0682
H from HML	12.0(10.6)/14.1(12.4)	0.0	137 (20.4)	0.0427
HM (-B)	-	-0.8	279 (41.5)	0.0605
HL (-B)	-	-0.1	279 (41.5)	0.0686
H (-B)	-	-2.4	278 (41.4)	0.0511

[†]H, M and L represent High-, Mid- and Low-resolution data sets, respectively.

If reflections that are not present in the HM data set are excluded from the HML merged data, the generated data set is called 'HM from HML'.

5.2. Effects of removing selected reflections

To assess the contribution of low-resolution reflections to visualization of hydrogen atoms, the reflections from high- or low-resolution, as well as randomly selected reflections, were removed from the HML data; σ_A-weighted F_o-F_c hydrogen-omit maps were calculated using these removed data with the phases and structure factors from the original data. The number of visualized hydrogen atoms, as a function of the percentage of reflections removed is plotted in Figure 5.2.

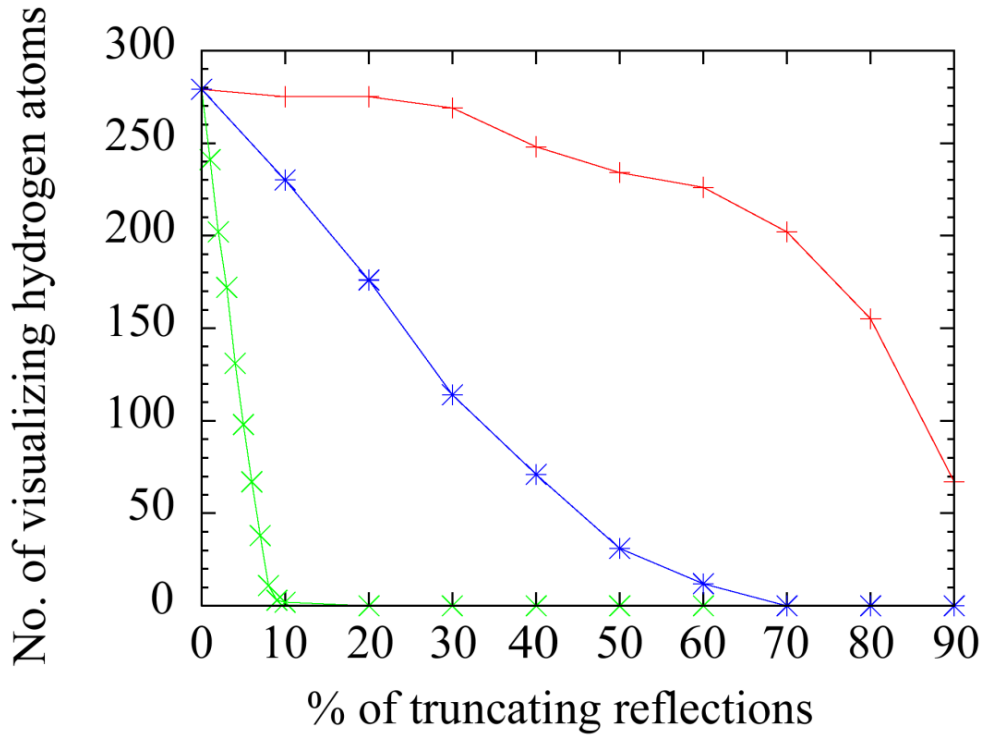


Figure 5.2 Truncation of reflections.

Red and green curves show the truncation from high and low resolution reflections, respectively. Blue curve shows the random truncation.

Removing reflections from the high-resolution data reduced the number of visualized hydrogen atoms, and only half the number were observed upon truncation of reflections by 20%. Truncation of reflections by 80, 50, and 10% from high-resolution resulted in a resolution cut-off of 1.5, 1.1, and 0.91 Å, respectively. The number of visualized hydrogen atoms in the 0.91 Å map was almost the same as the number in the 0.88 Å map. When reflections were removed randomly, the number of visualized hydrogen atoms was reduced in proportion to the percentage of the removed reflections. The resolution ranges for the randomly truncated data were ~43.1-0.88 Å. Removal 10% of the reflections decreased the number of visualized hydrogen atoms by 13%. When the reflections were removed from the low-resolution, the number of visualized hydrogen atoms dramatically reduced, and no hydrogen atoms were observed when reflections were removed even by as little as ~10%. Removal ~4% of the reflections halved the number of visualized hydrogen atoms. Removal of 4% of low-resolution reflections is equivalent to removing data below 2.9 Å resolution, and removal of 10% of low-resolution reflections is equivalent to removing data below 1.9 Å resolution.

Chapter 6
Conclusions

We collected X-ray diffraction data on a crystal of the bovine H-protein using synchrotron radiation, and determined the structure at 0.88 Å resolution. This is the highest-resolution structure of a H-protein from any source. Our comparisons of the atomic structure of bovine and other H-proteins demonstrate that the secondary structures of H-proteins are highly conserved.

To overcome the limitations of intensity measurements and the limited dynamic range of the CCD detector used, we measured three data sets in order to obtain a complete set of data under different experimental conditions. All of the high-, mid- and low-resolution data sets were of very high quality. The three data sets could be merged using any of them as a reference. The statistics of the merged data sets were similar, except for $<I>/<\sigma(I)>$ and the Wilson B values. To assess the quality of merged data sets obtained using high-, mid- and low-resolution data as reference, the number of visualized hydrogen atoms were counted in σ_A-weighted F_o-F_c hydrogen-omit maps. The results showed that the merged data set using the high-resolution data as reference was better than the other two sets. This effect is probably a result of radiation damage and the scaling algorithm. Thus, it is important to choose the high-resolution data set as the reference.

The advantage of merging three data sets was investigated using seven merged data sets (HML, HM, HL, H, HM from HML, HL from HML and H from HML). The qualities of the merged data set were assessed by comparing R factors against the re-refinement models and by counting hydrogen atoms with significant electron density, and the results suggested that the low-resolution reflections contributed significantly to visualization of hydrogen atoms. In addition, mid-resolution reflections improved redundancy and completeness,

and the number of visualized hydrogen atoms slightly increased. In summary, merging of the three data sets was effective in determining structure of high-quality as well as for locating hydrogen atoms.

When reflections from high and low resolutions were removed from the list of reflections, the number of visualized hydrogen atoms was reduced. In particular, the truncation of low-resolution reflections strongly affected the visualization of hydrogen atoms. These observations suggest that low-resolution reflections are critical in the visualization of hydrogen atoms, even in high-resolution crystal structures.

Establishment of the methodology in high-resolution X-ray crystallography will enable to determine the more precisely parameters of many protein structures including hydrogen atoms. As the number of more reliable structures in the ultra-high-resolution will be increased, the charge density studies for proteins will be accelerated.

References

Adams, P.D., Grosse-Kunstleve, R.W., Hung, L.-W., Ioerger, T. R., McCoy, A. J., Moriarty, N. W., Read, R. J., Sacchettini, J. C., Sauter, N. K. & Terwilliger, T. C. (2002) *Acta Cryst.* D**58**, 1948-1954. PHENIX: building new software for automated crystallographic structure determination

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissing, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucleic Acids Res.* **28**, 235-242. The Protein Data Bank.

Berman, H. M., Bettistuz, T., Bhat, T. N., Bluhm, W. F., Bourne, P. E., Burkhardt, K., Feng, Z., Gilliland, G. L., Iype, L., Jain, S., Fagan, P., Marvin, J., Padilla, D., Ravichandran, V., Scheider, B., Thanki, N., Weissig, H., Westbrook, J. D., & Zardecki, C. (2002). *Acta Cryst*. **D58**, 899-907. The Protein Data Bank

Bönisch, H., Schmidt, C.L., Bianco, P., Ladenstein, R., (2005). *Acta Cryst*. **D61**, 990-1004. Ultrahigh-resolution study on Pyrococcus abyssi rubredoxin. I. 0.69 Å X-ray structure of mutant W4L/R5S.

Dauter, Z. (2003). *Methods Enzymol.* **368**, 288-337.

Dauter, Z., Lamzin, V. S. & Willson, K. S. (1997). *Curr. Opin. Struct. Biol.* **7**, 681-688. The benefits of atomic resolution.

Davis, I. W., Murray, L. W., Richardson, J.S. & Richardson, D. C. (2004). *Nucleic Acids Res.* **32**, W615-W619. MOLPROBITY: structure validation and all-atom contact analysis for nucleic acids and their complexes.

DeLano, W.L., (2002). in *DeLano Scientific, Palo Alto, CA, USA.* The *PyMOL* Molecular Graphics System

Douce, R., Bourguignon, J., Neuburger, M., & Rébeillé, F. (2001). *Trends plant Sci.* **6**, 167-176. The glycine decarboxylase system: a fascinating complex.

Dunlop, K.V., Irvin, R.T., & Hazes, B. (2005). *Acta Cryst.* D**61**, 80-87. Pros and cons of cryocrystallography: should we also collect a room-temperature data set?

Emsley, P. & Cowtan, K. (2004). *Acta Cryst.* D**60**, 2126-2132. Coot: model-building tools for molecular graphics

Fenn, T.D., D. Ringe, and G.A. Petsko, (2003). *J. Appl. Cryst.* **36**, 944-947. *POVScript+:* a program for model and data visualization using persistence of vision ray-tracing.

Fujiwara, K., Okamura-Ikeda, K., & Motokawa, Y. (1990). *J. Biol. Chem*. **265**, 17463-17467. cDNA sequence, in vitro synthesis, and intramitochondrial lipoylation of H-protein of the glycine cleavage system.

Fujiwara, K., Okamura-Ikeda, K., & Motokawa, Y. (1992). *J. Biol. Chem*. **267**, 20011-20016. Expression of mature bovine H-protein of glycine cleavage system in *Escherichia coli* and *in vitro* lipoylation of the apoform.

Gouet, P., Robert, X., & Courcelle, E. (2003). Nucleic Acids Res. 31, 3320-3323. ESPript/ENDscript: Extracting and rendering sequence and 3D information from atomic structures of proteins.

Hakanpää, J., Linder, M., Popov, A., Schmidt, A., & Rouvinen, J. (2006) *Acta Cryst.* D**62**, 356-367. Hydrophobin HFBII in detail: ultrahigh-resolution structure at 0.75 A.

Howard, E.I., Sanishvili, R., Cachau, R.E., Mitschler, A., Chevrier, B., Barth, P., Lamour, V., Van Zandt, M., Sibley, E., Bon, C., Moras, D., Schneider, T.R., Joachimiak, A., & Podjarny, A. (2004). Proteins. **55**, 792-804. Ultrahigh resolution drug design I: Details of interactions in human aldose reductase-inhibitor complex at 0.66 Å.

Jelsch, C., Teeter, M. M., Lamzin, V., Pichon-Pesme, V., Blessing, R. H., & Lecomte, C. (2000). *Proc. Natl. Acad. Sci. USA,* **97**, 3171-3176. Accurate protein crystallography at ultra-high resolution: Valence electron distribution in crambin.

Kabsch, W. and C. Sander, (1983). *Biopolymers,* **22**(12), 2577-637. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features.

Kang, B.S., Devedjiev, Y., Derewenda, U., & Derewenda, Z.S. (2004). *J. Mol. Biol.* **338(3)**, 483-493. The PDZ2 domain of syntenin at ultra-high resolution: bridging the gap between macromolecular and small molecule crystallography.

Kikuchi, G., Motokawa, Y., Yoshida, T. and Hiraga, K., *Proc. Jpn. Acad. Ser. B Phys. Biol. Sci.*, **84**(7), 246-63. Glycine cleavage system: reaction mechanism, physiological significance, and hyperglycinemia.

Ko, T.P., Robinson, H., Gao, Y.G., Cheng, C.H., Devries, A.L., & Wang, A.H. (2003). *Biophys. J.*, **84**, 1228-1237. The refined crystal structure of an eel pout type III antifreeze protein RD1 at 0.62-Å resolution reveals structural microheterogeneity of protein and solvation.

Kuhn, P., Knapp, M., Soltis, SM., Ganshaw, G., Thoene, M., & Bott. R. (1998). *Biochemistry.* **37**, 13446-13452. The 0.78 Å structure of a serine protease: Bacillus lentus subtilisin.

Lovell, S. C., Davis, I. W., Arendall, W. B., de Bakker, P. I., Word, J. M., Prisant, M. G., Richardson, J. S. and Richardson, D. C. (2003). *Proteins*, **50**(3), p. 437-50. Structure validation by Calpha geometry: phi,psi and Cbeta deviation.

Matthews, B.W. (1968) *J. Mol. Biol.* **33**, 491-197. Solvent content of protein crystals.

Murshudov, G.N., Vagin, A.A., & Dodson, E.J. (1997). *Acta Cryst.* D**53**, 240-255. Refinement of Macromolecular Structures by the Maximum-Likelihood Method.

Nakai, T., Ishijima, J., Masui, R., Kuramitsu, S., & Kamiya, N. (2003). *Acta Cryst*. D**59**, 1610-1618. Structure of *Thermus thermophilus* HB8 H-protein of the glycine-cleavage system, resolved by a six-dimensional molecular-replacement method.

Nakai, T., Nakagawa, N., Maoka, N., Masui, R., Kuramitsu, S., & Kamiya., N. (2005) *EMBO J.* **24**, 1523-1536. Structure of P-protein of the glycine cleavage system: implications for nonketotic hyperglycinemia

Okamura-Ikeda, K., Hosaka, H., Yoshimura, M., Yamashita, E., Toma, S., Nakagawa, A., Fujiwara, K., Motokawa, Y. and Taniguchi, H. (2005) *J. Mol. Biol.* **351**, 1146-1159. Crystal structure of human T-protein of glycine cleavage system at 2.0 A resolution and its implication for understanding non-ketoic hyperglycinemia

Otwinowski, Z. & Minor, M. (1997). *Methods Enzymol.* **276**, 307-326. Processing of X-ray diffraction data collected in oscillation mode.

Petrova, T. & Podjarny, A. (2004). *Rep. Prog. Phys.* **67**, 1565-1605. Very high resolution X ray structures of biological macromolecules.

Pares, S., Cohen- Addad, C., Sieker, L., Neuburger, M., & Douce, R. (1994). *Proc. Natl. Acad. Sci. USA,* **91**, 4850-4853. X-ray structure determination at 2.6-Å resolution of a lipoatecontaining protein: the H-protein of the glycine decarboxylase complex from pea leaves.

Pares, S., Cohen- Addad, C., Sieker, L., Neuburger, M., & Douce, R. (1995). *Acta Cryst.* D**51**, 1041-1051. Refined structure at 2 and 2.2 Å resolution of two forms of the H-protein, a lipoamide-containing protein of the glycine decarboxylase complex.

Perrakis, A., Morris, R., & Lamzin, V. S. (1999). *Nat. Struct. Biol.* **6**, 458-463. Automated protein model building combined with iterative structure refinement.
Ramachandran & Sasisekharan (1968). *Adv. Protein Chem.* **23**, 283-438. Conformation of polypeptides and proteins.

Sevcik, J., Dauter, Z., Lamzin, V.S. & Wilson, S. (1996). *Acta Cryst*. D**52**, 327-344. Ribonuclease from *streptomyces aureofaciens* at atomic resolution.

Sheldrick, G. M. & Schneider, T. R. (1997). *Methods Enzymol.* **277**, 319-343. SHELXL: High-resolution refinement.

Shmidt, A., & Lamzin, V. S. (2002). *Curr. Opin. Struct. Biol.* **12**, 698-703. *Veni, vidi-, vici*-atomic resolution unravelling the mysteries of protein function.

Tada, K. and Kure, S. (1993). *J. Inher. Metab. Dis.* **16**, 691-703. Non-ketoic hyperglycinemia: molecular lesion, diagnosis and pathophysiology.

Vagin, A. & Teplyakov, A. (1997). *J. Appl. Cryst.* **30**, 1022-1025. MOLREP: an Automated Program for Molecular Replacement.

Vrielink, A. & Sampson, N., (2003). *Curr. Opin. Struct. Biol.* **6**, 709-715. Sub-Angstrom resolution enzyme X-ray structures: is seeing believing?

Vriend, G. (1990). *J. Mol. Graph.* **8**, 52-56. WHAT IF: A molecular modeling and drug design program.

Wang, J., Dauter, M., Alkire, R., Joachimiak, A., & Dauter, Z. (2007) *Acta Cryst*. D**63**, 1254-1268. Triclinic lysozyme at 0.65 Å resolution.

Weiss, M. S., (2001). *J. Appl. Cryst.*, **34**, 130-135. Global indicators of X-ray data quality.

List of publications

Directly related

Higashiura, A., Kurakane, T., Matsuda, M., Suzuki, M., Inaka, K., Sato, M., Kobayashi, T., Tanaka, T., Tanaka, H., Fujiwara, K., & Nakagawa, A. *Acta Cryst*. **D (In press)**. High-resolution X-ray crystallography of bovine H-protein determined at 0.88Å resolution

Not directly related

Wang, C.-Y., Miyazaki, N., Yamashita, T., Higashiura, A., Nakagawa, A., Li, T.-C., Takeda, N., Xing, L., Hjalmarsson, E., Friberg, C., Liou, D.-M., Sung, Y.-J., Tsukihara, T., Matsuura, Y., Miyamura, T. & Cheng, R.H. (2008). *Acta Cryst*. F64, 318-22. Crystallization and preliminary X-ray diffraction analysis of recombinant hepatitis E virus-like particle

Miyazaki, N., Uehara-Ichiki, T., Li, X., Bergman, L., Higashiura, A., Omura, T., & Cheng, R.H. (2008). *J. Virol.* **82(22)**, 11344-11353. Structural evolution of Reoviridae revealed by Oryzavirus in aquiring the second capsid shell