The development of model fitting and completion program for automated iterative nucleic acid refinement

Keitaro Yamashita

A thesis submitted to Hokkaido University for the degree of Doctor of Philosophy in the Graduate School of Life Science

2013

Abstract

Over the past decade, many nucleic acid structures have been determined, which have contributed to our understanding of their biological functions. However, the crystals containing nucleic acid often poorly diffract X-rays. This makes electron density interpretation difficult and requires a great deal of expertise in crystallography and knowledge of nucleic acid structure. Here, new programs called *NAFIT* and *NABUILD* for fitting and extending nucleic acid models are presented. These programs can be used as modules for the automated refinement system, *LAFIRE*, as well as acting as independent programs. *NAFIT* performs sequential grouped fitting with empirical torsion angle restraints and anti-bumping restraints including hydrogen atoms. *NABUILD* extends the model using a skeletonized map in a coarse-grained manner. It was shown that *NAFIT* greatly improved electron density fit and geometric quality and *NABUILD* with iterative refinement significantly reduced R_{free} factor.

Contents

List of Figures

List of Tables

Acknowledgements

First of all, I would like to thank my supervisors, Prof. Isao Tanaka and Prof. Min Yao, for all of their advice and guidance over the past few years. Their guidance helped me in all the time of research and writing of this thesis. I am so grateful for the opportunity to work on this project, which has allowed me to learn so many things in crystallography.

I would like to thank Prof. Tamiki Komatsuzaki for his careful reading and much useful advice for improving this thesis.

I would also like to thank all of the laboratory members for their help and enjoyable environment. I have spent almost all the time in the laboratory since I became a member and really enjoyed the laboratory life in good atmosphere. Most notably, I thank Dr. Akiyoshi Nakamura for his much advice and hints for improving the programs. Besides that, He kindly provided Thg1-tRNA complex data and the initial model for the test of *NAFIT*.

I would finally like to thank my family for their encouragement and support.

Author's Declaration

I hereby declare that this PhD thesis entitled "The development of model fitting and completion program for automated iterative nucleic acid refinement" has been compiled by me for the degree of Doctor of Philosophy in Life Science under the supervision of Prof. Isao Tanaka and Prof. Min Yao, the Faculty of Advanced Life Science, Hokkaido University. This thesis has not been previously submitted for the award of any degree, diploma, associateship, fellowship or its equivalent to any other University or Institution.

Abbreviations

PDB	Protein Data Bank
DNA	Deoxyribonucleic acid
RNA	Ribonucleic acid
tRNA	transfer RNA
LAFIRE	Local correlation coefficient-based Automatic FItting for REfinement
GLCC	Grouped Local Correlation Coefficient
SVN	subversion (a version control system)
GSL	GNU Scientific Library
MLF	Maximum Likelihood target using amplitude (F)
MLHL	phased Maximum Likelihood target using amplitude and Hendrickson-Lattman coefficients
ADP	Atomic Displacement Parameter
NCS	Non-Crystallographic Symmetry
SAD	Single wavelength Anomalous Dispersion
MAD	Multiple wavelength Anomalous Dispersion
GUI	Graphical User Interface
r.m.s.d.	root-mean-square deviation, defined as $\sqrt{\frac{1}{N}\sum_{i=1}^{N}\delta_i^2}$, where δ_i is the distance between corresponding atoms (when comparing atoms) or $\rho_i - \langle\rho\rangle$ (when referring to map level).

Chapter 1

Introduction

The number of nucleic acid structures, including those complexed with protein, is increasing rapidly and their biological functions are being determined. Although a number of programs and methods are available for automated building and refinement of protein structures, there are relatively few such programs for nucleic acids. As the diffractivity of crystals containing nucleic acid is usually not so high (Fig. 1.1), several authors have made a great deal of effort to tackle the difficulties involved. Initial model building programs,

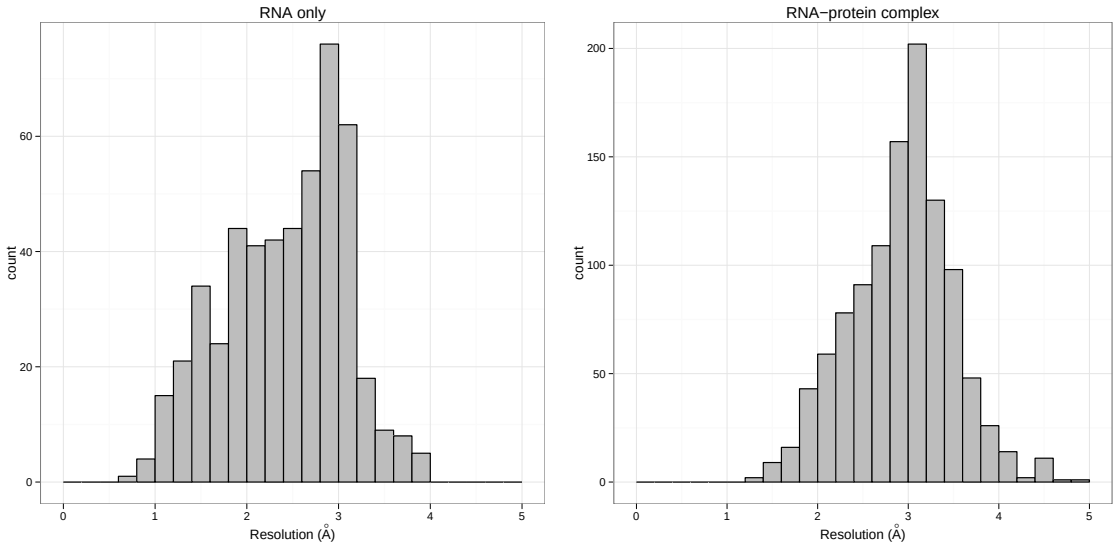


Figure 1.1: Histogram of maximum resolution of RNA and protein-RNA complex crystals. Entries with maximum resolution of more than 5 Å are omitted in the histogram. Reference: Protein Data Bank (Berman *et al.*, 2000) as of February 15, 2013.

such as *phenix.autobuild* (Terwilliger *et al.*, 2008*b*; Adams *et al.*, 2010), *ARP/wARP* (Hat-

tne and Lamzin, 2008), *NUT/DHL/RSR* (Pavelcik and Schneider, 2008; Pavelcik, 2012), and *Nautilus* (Cowtan, 2012), utilize features that can be seen in lower resolution electron density maps. Semi-automated model building can be performed with *RCrane* (Keating and Pyle, 2012). As nucleic acid structures have many rotatable bonds in the main chain, the conformation is ambiguous at lower resolution and model building is therefore error-prone. Such errors can be detected by *MolProbity* (Chen *et al.*, 2010) and corrected by *RNABC* (Wang *et al.*, 2008) and *ERRASER* (Sripakdeevong *et al.*, 2011; Chou *et al.*, 2013). A general molecular replacement technique has been proposed (Scott, 2012), which utilizes ideal A-form RNA fragments. Although these methods and programs are useful, the initial model usually has many conformational errors and unconstructed regions. Thus, in the subsequent step, the repeated refinements and model rebuilding are required, which makes the structure analysis time-consuming.

Here, we present the new refinement tools for nucleic acids, *NAFIT* and *NABUILD*. The functions are incorporated into the automated refinement program *LAFIRE* (Local correlation coefficient-based Automatic FItting for REfinement; Yao *et al.*, 2006; Zhou *et al.*, 2006), which repeats refinement, model fitting, and extension without human intervention. *NAFIT* is the real space refinement (fitting) program with empirical torsion angle restraints and anti-bumping restraints including hydrogen atoms to maintain geometric quality. Here, anti-bumping restraint is used to prevent non-bonded atom pair from approaching too close to each other. *NABUILD* is the chain extension program that uses graph interpretation constructed from a skeletonized map. A coarse-grained nucleic acid model is first built and the full atomic structure is constructed by *NAFIT*. *NAFIT* and *NABUILD* with *LAFIRE* refinement strategy significantly reduced the R_{free} factor for test cases with resolutions in the range from 2.1 to 3.12 Å.

Chapter 2

Background

2.1 Overview of macromolecular X-ray crystallography

X-ray crystallography is one of the most powerful methods to determine molecular structures in ångström order accuracy. When X-rays hit the material, X-rays interact with electrons. For crystal structure determination, we utilize the Thomson (elastic) scattering. The crystal is absolutely needed because the scattering wave from single molecule is too weak to be measured to higher angle.

When X-rays with the frequency of ν hit, the wave $\Psi(\boldsymbol{r}^*)$ scattered by electron density $\rho(\boldsymbol{r})$ of molecule can be expressed by

$$\begin{aligned} \Psi(\boldsymbol{r}^*) &= E_{\mathrm{e}} e^{2\pi i \nu t} \int_V \rho(\boldsymbol{r}) e^{2\pi i \boldsymbol{r}^* \cdot \boldsymbol{r}} d\boldsymbol{r}, & (2.1) \\ &= E_{\mathrm{e}} e^{2\pi i \nu t} \boldsymbol{F}(\boldsymbol{r}^*). & (2.2) \end{aligned}$$

Here, E_{e} is amplitude of scattered wave by single electron, and $\boldsymbol{F}(\boldsymbol{r}^*)$ is called structure factor that is a Fourier transform of $\rho(\boldsymbol{r})$. The vector $\boldsymbol{r}^* = (\boldsymbol{s}_1 - \boldsymbol{s}_0)/\lambda$ defines the scattering direction and is called reciprocal vector where $\boldsymbol{s}_0, \boldsymbol{s}_1, \lambda$ are the incident direction, the scattered direction, and the wavelength of X-rays, respectively (illustrated in Fig. 2.1). The structure factor is a complex number

$$\boldsymbol{F}(\boldsymbol{r}^*) = F(\boldsymbol{r}^*) e^{i\varphi(\boldsymbol{r}^*)} \quad (F \geq 0) \tag{2.3}$$

where F and φ are called structure amplitude and phase, respectively.

For crystal, the infinite repeat of the unit cell along the basic translation vectors $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}$, the structure factor is discretized so that $\boldsymbol{F}(\boldsymbol{r}^*)$ has value only when h, k, l are all integer

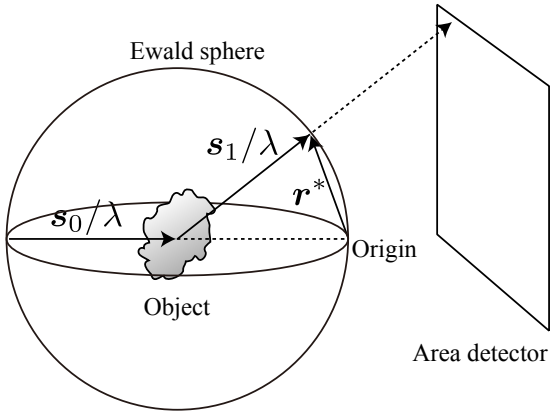


Figure 2.1: Illustration of $\boldsymbol{s}_0, \boldsymbol{s}_1, \boldsymbol{r^*}$. The sphere centered at $-\boldsymbol{s}_0/\lambda$ with the radius of $1/\lambda$ in the reciprocal space is called Ewald sphere. Diffraction occurs only for reciprocal lattice points on the surface of the Ewald sphere.

for $\boldsymbol{r^*} = h\boldsymbol{a}^* + k\boldsymbol{b}^* + l\boldsymbol{c}^*$. Here, reciprocal basis vectors $\boldsymbol{a}^*, \boldsymbol{b}^*, \boldsymbol{c}^*$ are defined as

$$\boldsymbol{a} \cdot \boldsymbol{a}^* = 1, \quad \boldsymbol{b} \cdot \boldsymbol{a}^* = 0, \quad \boldsymbol{c} \cdot \boldsymbol{a}^* = 0, \tag{2.4}$$

$$\boldsymbol{a} \cdot \boldsymbol{b}^* = 0, \quad \boldsymbol{b} \cdot \boldsymbol{b}^* = 1, \quad \boldsymbol{c} \cdot \boldsymbol{b}^* = 0, \tag{2.5}$$

$$\boldsymbol{a} \cdot \boldsymbol{c}^* = 0, \quad \boldsymbol{b} \cdot \boldsymbol{c}^* = 0, \quad \boldsymbol{c} \cdot \boldsymbol{c}^* = 1. \tag{2.6}$$

If we could obtain $\boldsymbol{F}$ from experiment, the electron density map can be calculated by inverse Fourier transform.

$$\rho(\boldsymbol{r}) = \frac{1}{V} \sum_{\boldsymbol{r^*}} \boldsymbol{F}(\boldsymbol{r^*}) e^{-2\pi i \boldsymbol{r^*} \cdot \boldsymbol{r}}. \tag{2.7}$$

By interpreting the map, the atomic model can be constructed. However, phase information $\varphi(\boldsymbol{r^*})$ cannot be directly obtained from diffraction experiment, while structure amplitude $F(\boldsymbol{r^*})$ can be obtained by measuring diffraction intensity $I(\boldsymbol{r^*})$ which is proportional to $F^2(\boldsymbol{r^*})$. This phase problem has been a central problem of crystallography. Several methods have been developed to obtain phase information. In macromolecular crystallography, there are two routinely used methods: molecular replacement (Rossmann and Blow, 1962) and heavy atom method (Green *et al.*, 1954). When substantial portion of the target protein is known and its structure is expected to be similar (say sequence identity $> 30\%$), molecular replacement is expected to work (Schwarzenbacher *et al.*, 2008). Molecular replacement determines the position and orientation of the known fragment in

the unit cell. Then the phase can be obtained as a part of structure factor calculated from the placed fragment. Heavy atom methods require artificially or naturally incorporated heavy atoms bound to the target protein. Nowadays, the most common heavy atom method is multiple wavelength anomalous dispersion (MAD; Karle, 1980; Hendrickson, 1991) or single wavelength anomalous dispersion (SAD; Hendrickson and Teeter, 1981). In all heavy atom method, substructure, the virtual crystal structure only containing the heavy atoms, is firstly determined, and then the phase is calculated according to the relationship between the structure factor and that of substructure. Density modification like solvent flattening (Wang, 1985), histogram matching (Zhang and Main, 1990), *etc.* is often required to yield the interpretable electron density map. By heavy atom method, unbiased phase information can be obtained in contrast to molecular replacement.

Once phase information is obtained, an atomic model can be constructed. Then the model is subjected to refinement, which is discussed in §2.2. As the complete model usually cannot be built at once, model (re)building and refinement need to be repeated. When refinement is finished, the structure can be published and deposited at PDB.

2.2 Macromolecular refinement

Refinement is the final process in structure determination. The goal of refinement is to obtain a molecular model which explains diffraction data well enough and is stereochemically valid. Ideally, the refined model should be complete (no unmodelled atoms), and have no errors in coordinates and displacement parameters.

As only X-ray intensities can be observed in the diffraction experiment (where no phase information obtained), only the structure amplitudes or intensities can be compared. Therefore, the observed structure amplitude $F_{\mathrm{obs}}(hkl)$ and the calculated structure amplitude of the model $F_{\mathrm{model}}(hkl)$ are usually compared. $\boldsymbol{F}_{\mathrm{model}}(hkl)$ is calculated using molecular model and bulk solvent contribution (explained later).

$$\boldsymbol{F}_{\mathrm{model}} = k(\boldsymbol{F}_{\mathrm{calc}} + \boldsymbol{F}_{\mathrm{bulk}}). \tag{2.8}$$

Here k is overall scale including isotropic and anisotropic displacement terms, and $\boldsymbol{F}_{\mathrm{calc}}$

consists of atomic contributions in the unit cell.

$$\boldsymbol{F}_{\mathrm{calc}}(\boldsymbol{r^*}) = \sum_j q_j f_j(r^*) T_j(\boldsymbol{r^*}) e^{-2\pi i \boldsymbol{r^*}\cdot\boldsymbol{r}_j}. \tag{2.9}$$

Here $\boldsymbol{r}_j$ is atomic position, q_j is occupancy of the atom, f_j is atomic form factor, and T_j represents atomic displacement explained below. The occupancy is a measure of the fraction of the atom occupying the position in the crystal. The atomic form factor is the Fourier transform of the electron density distribution of the atom, and it depends on the element. The atomic electron density distribution for each element is calculated based on spherically-symmetric approximation, and therefore, the form factor is the function of $r^*(= |\boldsymbol{r^*}|)$. The atomic displacement can be modelled isotropically or anisotropically. When isotropically modelled, T_j is expressed using B-factor (Giacovazzo, 1992).

$$T_{\mathrm{iso}}(\boldsymbol{r^*}) = e^{-B_j r^{*2}/4}, \tag{2.10}$$

$$B_j = 8\pi^2 \langle r_j'^2 \rangle. \tag{2.11}$$

Here $\langle r_j'^2 \rangle$ is the square mean shift of the atom j with respect to the position of equilibrium. When anisotropically modelled, T_j is expressed using 3×3 symmetric matrix $\boldsymbol{U}_j^*$, which defines the orientation of the thermal ellipsoid with respect to the crystallographic axes and the lengths of the three ellipsoid axes (Giacovazzo, 1992).

$$T_{\mathrm{aniso}}(\boldsymbol{h}) = e^{-2\pi^2 \boldsymbol{h}^t \boldsymbol{U}_j^* \boldsymbol{h}}. \tag{2.12}$$

Here $\boldsymbol{h} = (h\ k\ l)^t$. Note that usually $\boldsymbol{F}_{\mathrm{calc}}$ is calculated from ρ_{calc} by fast Fourier transform (FFT) instead of direct summation of eqn. (2.9) because it is much faster.

Macromolecular crystals include large amounts of solvent molecules. Most of them are disordered in the crystal, and therefore, they are invisible in the electron density map (Fig. 2.2). However, as they have small and flat electron density, the bulk solvent density ρ_{bulk} largely contributes to low resolution structure factors. Therefore, it is important to include $\boldsymbol{F}_{\mathrm{bulk}}$, the Fourier transform of ρ_{bulk}, in the structure factor calculation. $\boldsymbol{F}_{\mathrm{bulk}}$ is calculated by Babinet's principle (Moews and Kretsinger, 1975) or solvent mask based method (Phillips, 1980; Jiang and Brünger, 1994; Afonine *et al.*, 2005). The latter method is more sophisticated and largely used recently. It defines the bulk solvent region by excluding modeled atoms, and then finds the parameters which describe the bulk solvent

density, say $k_{\rm sol}$ and $B_{\rm sol}$ in the following equation.

$$\boldsymbol{F}_{\rm bulk} = k_{\rm sol} e^{-B_{\rm sol} r^{*2}/4} \mathcal{F}[m(\boldsymbol{r})], \tag{2.13}$$

$$m(\boldsymbol{r}) = \begin{cases} 1 & (\boldsymbol{r} \in \text{bulk solvent region}) \\ 0 & (\boldsymbol{r} \in \text{protein region}) \end{cases}. \tag{2.14}$$

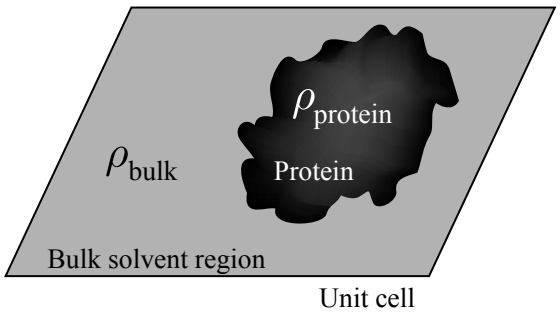


Figure 2.2: Bulk solvent model. Bulk solvent region can be explicitly defined by masking the protein region.

The degree of refinement is evaluated by reliability factors, $R_{\rm work}$ and $R_{\rm free}$.

$$R_{\rm work} = \frac{\sum_{hkl \notin T} |F_{\rm obs}(hkl) - F_{\rm model}(hkl)|}{\sum_{hkl \notin T} F_{\rm obs}(hkl)}, \tag{2.15}$$

$$R_{\rm free} = \frac{\sum_{hkl \in T} |F_{\rm obs}(hkl) - F_{\rm model}(hkl)|}{\sum_{hkl \in T} F_{\rm obs}(hkl)}. \tag{2.16}$$

Here, T is called test set. Test set is a randomly chosen subset (normally $5 \sim 10\%$) of hkl in resolution sphere. In refinement, as none of reflections belonging to T are used for minimization, degree of refinement can be objectively evaluated by $R_{\rm free}$ (Brünger, 1992; Kleywegt and Brünger, 1996). Correctly refined model should have small $R_{\rm work}$ and $R_{\rm free}$ factors, and difference of the factors should be moderately small. Obviously, R-factors should achieve zero for perfect model and perfect data, while the random (thus completely wrong) structure gives $R = 0.586$ for acentrics (Wilson, 1950). Achievable values mainly depend on the resolution.

Refinement programs perform minimization of the target function E with respect to atomic coordinates and/or displacement parameters *etc.*

$$E = wE_{\rm xray} + E_{\rm geometry}. \tag{2.17}$$

Here, E_{xray} is an "X-ray" term which represents how diffraction data are explained by the model, E_{geometry} is a geometric term to restrain stereochemical parameters of the model to ideal or favoured values, and w is the weight to balance E_{xray} and E_{geometry}. Normally there are two options for E_{xray}; least square (LS) target and maximum likelihood (ML) target. For macromolecular refinement, the ML target is normally preferred, especially when incompleteness or error of the model is severe, which will be explained in §2.2.1. E_{geometry} basically includes terms of bond length, bond angle, torsion angle, planarity, chirality (chiral volume), and non-bonded atom contact. Optionally, restraints of non-crystallographic symmetry (NCS), Ramachandran favorability[1)] (Kleywegt and Jones, 1996), reference model similarity[2)] (Headd *et al.*, 2012; Nicholls *et al.*, 2012; Smart *et al.*, 2012), *etc.* can be used mainly depending on resolution or stage of refinement. Especially at lower resolution, tighter restraints are essential, but they could introduce a bias into the model. The weight w should be properly adjusted.

2.2.1 Refinement target

The simplest refinement target for E_{xray} is a least square target (Konnert, 1976).

$$Q_{\mathrm{LS}} = \sum_{hkl} \frac{1}{\sigma^2(F_{\mathrm{obs}})} \Big(F_{\mathrm{obs}}(hkl) - F_{\mathrm{model}}(hkl) \Big)^2, \tag{2.18}$$

where, $\sigma^2(F_{\mathrm{obs}})$ is the estimated variance of F_{obs}. However, least square target is inappropriate for macromolecular refinement in most cases. For example, in macromolecular refinement, the model is often incomplete and refinement is used to yield better phase which shows unmodelled portions more clearly in real space. By least square target, the model cannot be properly improved in such case because unincorporated contributions of missing atoms cannot be compensated by any means. Maximum likelihood target overcomes the problems of least square target by taking model incompleteness and errors into account.

Likelihood is defined as the conditional probability of F_{obs} given F_{model}. To maximize

[1)] The Ramachandran restraint is used to prevent peptide torsion angles φ, ψ from having unfavored values.

[2)] The reference model restraint is used to keep the conformation similar to the reference model.

the likelihood, the maximum likelihood target for refinement is defined as

$$Q_{\mathrm{ML}} = -\log \prod_{hkl} p(F_{\mathrm{obs}}; F_{\mathrm{model}}) = -\sum_{hkl} \log p(F_{\mathrm{obs}}; F_{\mathrm{model}}). \tag{2.19}$$

As there are a number of forms for maximum likelihood refinement target (Pannu and Read, 1996; Adams *et al.*, 1997; Murshudov *et al.*, 1997; Lunin *et al.*, 2002; Blanc *et al.*, 2004), here the form used in *PHENIX* package (Adams *et al.*, 2010) is taken for an example.

With the assumption that atoms in the model have random independent errors in positions and displacement parameters, and missing atoms are independently and uniformly distributed in the unit cell, the probability $p(\boldsymbol{F}_{\mathrm{obs}}; \boldsymbol{F}_{\mathrm{model}})$ for acentric reflections becomes two dimensional Gaussian centered on $\alpha \boldsymbol{F}_{\mathrm{model}}$ due to the central limit theorem (Fig. 2.3).

$$p(\boldsymbol{F}_{\mathrm{obs}}; \boldsymbol{F}_{\mathrm{model}}) = \frac{1}{\pi\varepsilon\beta} \exp\left(-|\boldsymbol{F}_{\mathrm{obs}} - \alpha \boldsymbol{F}_{\mathrm{model}}|^2/\varepsilon\beta\right). \tag{2.20}$$

For centric reflections[3)], the distribution is one dimensional.

$$p(\boldsymbol{F}_{\mathrm{obs}}; \boldsymbol{F}_{\mathrm{model}}) = \frac{1}{\sqrt{2\pi\varepsilon\beta}} \exp\left(-|\boldsymbol{F}_{\mathrm{obs}} - \alpha \boldsymbol{F}_{\mathrm{model}}|^2/2\varepsilon\beta\right). \tag{2.21}$$

Here, α reflects the error in atomic coordinates, and β reflects both atomic coordinate errors and the amount of scattering density of missing (unmodelled) atoms. α and β are considered as constant inside thin resolution shells and estimated by optimization of likelihood function with respect to α and β for each resolution shell using reflections belonging to test set (Lunin and Skovoroda, 1995).

As we do not have any information of phase φ, phase is eliminated by integrating out φ to obtain $p(F_{\mathrm{obs}}; F_{\mathrm{model}})$, the conditional probability of F_{obs} given F_{model}. For acentrics,

$$p(F_{\mathrm{obs}}; F_{\mathrm{model}}) = F_{\mathrm{obs}} \int_0^{2\pi} p(\boldsymbol{F}_{\mathrm{obs}}; \boldsymbol{F}_{\mathrm{model}}) d\varphi, \tag{2.22}$$

$$= \frac{2F_{\mathrm{obs}}}{\varepsilon\beta} \exp\left(-\frac{F_{\mathrm{obs}}^2 + \alpha^2 F_{\mathrm{model}}^2}{\varepsilon\beta}\right) I_0\left(\frac{2\alpha F_{\mathrm{obs}} F_{\mathrm{model}}}{\varepsilon\beta}\right). \tag{2.23}$$

Here F_{obs} is multiplied as Jacobian to transform $p(\boldsymbol{F}_{\mathrm{obs}}; \boldsymbol{F}_{\mathrm{model}})$ to polar coordinates, and $I_0(x)$ is the modified Bessel function of the zeroth order. For centrics,

$$p(F_{\mathrm{obs}}; F_{\mathrm{model}}) = \int_0^{2\pi} p(\boldsymbol{F}_{\mathrm{obs}}; \boldsymbol{F}_{\mathrm{model}}) d\varphi, \tag{2.24}$$

[3)] centric reflections are the phase restricted reflections. If there exists $\boldsymbol{R}$ such that $\boldsymbol{h}^t \boldsymbol{R} = -\boldsymbol{h}^t$, the phase of reflection $\boldsymbol{h}$ is restricted to $\pi \boldsymbol{h} \boldsymbol{T} \pm n\pi$ ($n \in \boldsymbol{Z}$) where $(\boldsymbol{R}, \boldsymbol{T})$ is a symmetry operator ($\boldsymbol{R}$ is rotation matrix and $\boldsymbol{T}$ is translation vector) belonging to the space group to transform atomic fractional coordinate $\boldsymbol{x}$ to $\boldsymbol{R}\boldsymbol{x} + \boldsymbol{T}$.

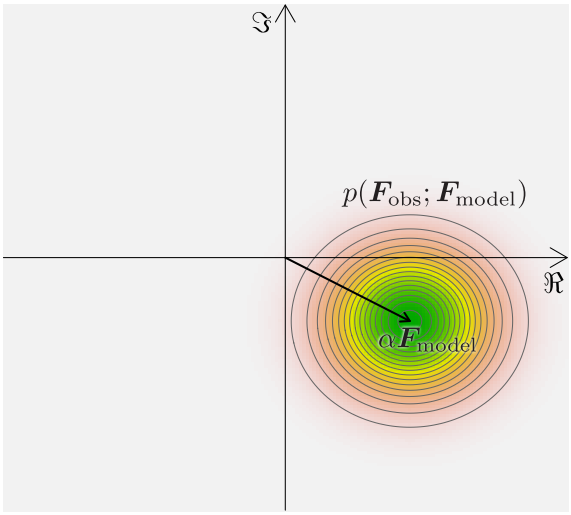


Figure 2.3: An example of $p(\boldsymbol{F}_{\mathrm{obs}}; \boldsymbol{F}_{\mathrm{model}})$ for acentrics on the complex plane. A 2-dimensional Gaussian centered on $\alpha \boldsymbol{F}_{\mathrm{model}}$ is drawn as a contour plot.

$$= \sqrt{\frac{2}{\pi\varepsilon\beta}} \exp\left(-\frac{F_{\mathrm{obs}}^2 + \alpha^2 F_{\mathrm{model}}^2}{2\varepsilon\beta}\right) \cosh\left(\frac{\alpha F_{\mathrm{obs}} F_{\mathrm{model}}}{\varepsilon\beta}\right). \quad (2.25)$$

The negative logarithm of product of $p(F_{\mathrm{obs}}; F_{\mathrm{model}})$ is called maximum likelihood target using amplitude (MLF) and is minimized by gradient-driven optimization method.

In case the phase is determined experimentally *i.e.* by heavy atom method, prior phase probability is available, which can be used for refinement. The phase probability distribution is represented by Hendrickson-Lattman (HL) coefficients (Hendrickson and Lattman, 1970). HL coefficients (A, B, C, D) encode a bimodal phase probability function $P(\varphi)$.

$$P(\varphi) = N \exp\left(A \cos\varphi + B \sin\varphi + C \cos 2\varphi + D \sin 2\varphi\right). \quad (2.26)$$

Here N is a normalization factor such that $\int_0^{2\pi} P(\varphi) d\varphi = 1$. When $C = D = 0$, $P(\varphi)$ is unimodal. The prior phase probability distribution encoded by HL coefficients can be incorporated by replacing eqn. (2.22) with

$$p(F_{\mathrm{obs}}; F_{\mathrm{model}}) = F_{\mathrm{obs}} \int_0^{2\pi} P(\varphi) p(\boldsymbol{F}_{\mathrm{obs}}; \boldsymbol{F}_{\mathrm{model}}) d\varphi. \quad (2.27)$$

The negative logarithm of product of eqn. (2.27) is called phased maximum likelihood target using HL coefficients (MLHL) that was shown to be beneficial if available (Pannu *et al.*, 1998).

2.2.2 Real space refinement

Models often are incomplete and have irremovable errors by minimization of the refinement target. Moreover, as accurate phase information cannot be obtained from diffraction experiment, the accurate electron density map cannot be obtained at once. The phase information is updated and improved in the course of refinement. By using improved phase, models can be fixed and missing atoms can be added. Therefore, refinement is an iterative process.

As shown by Read (1986), for acentrics, the true structure factor $\boldsymbol{F}$ and $\boldsymbol{F}_{\mathrm{model}}$ are related by

$$m|\boldsymbol{F}_{\mathrm{obs}}|e^{i\varphi_{\mathrm{model}}} = \frac{\boldsymbol{F}}{2} + \frac{D\boldsymbol{F}_{\mathrm{model}}}{2} + \text{noise terms}, \tag{2.28}$$

where D is called Luzzati D-factor which is equivalent to α in eqn. (2.20), and m is called figure of merit (FOM) defined by

$$m = \langle \cos(\varphi - \varphi_{\mathrm{model}}) \rangle = \int_0^{2\pi} P(\varphi - \varphi_{\mathrm{model}}) \cos(\varphi - \varphi_{\mathrm{model}}) d\varphi. \tag{2.29}$$

Therefore $\boldsymbol{F}$ can be approximated by

$$\boldsymbol{F} \approx (2mF_{\mathrm{obs}} - DF_{\mathrm{model}})e^{i\varphi_{\mathrm{model}}}. \tag{2.30}$$

We call the electron density map calculated from eqn. (2.30) "$2mF_{\mathrm{o}} - DF_{\mathrm{c}}$ map" (or simply $2F_{\mathrm{o}} - F_{\mathrm{c}}$ map) and routinely compare the model and this map. Another routinely used map is "$mF_{\mathrm{o}} - DF_{\mathrm{c}}$ map" (or simply $F_{\mathrm{o}} - F_{\mathrm{c}}$ map), which effectively shows the difference electron density between the true model and the current model. Note that these maps are less biased, but their accuracy depends on the accuracy and completeness of the current model. Several methods have been proposed to overcome the severe model bias: multi-start simulated annealing (Hodel *et al.*, 1992; Rice *et al.*, 1998), prime-and-switch phasing (Terwilliger, 2004), iterative-build OMIT maps (Terwilliger *et al.*, 2008*a*), averaged kick maps (Pražnikar *et al.*, 2009), *etc.*

If experimental phase information is used in the form of prior phase probability distribution in refinement, the phase combined $2mF_{\mathrm{o}} - DF_{\mathrm{c}}$ map can be calculated. By the refinement programs, instead of eqn. (2.30)

$$2m_{\mathrm{comb}}F_{\mathrm{obs}}e^{i\varphi_{\mathrm{comb}}} - DF_{\mathrm{model}}e^{i\varphi_{\mathrm{model}}} \tag{2.31}$$

is computed by default if MLHL is used. Here, m_{comb} and φ_{comb} are obtained by combining the prior phase probability distributions with the probability distribution of φ_{model}. Unimodal phase distribution of φ_{model} is constructed with estimated FOM.

By looking at $2mF_{\mathrm{o}}-DF_{\mathrm{c}}$ and $mF_{\mathrm{o}}-DF_{\mathrm{c}}$ maps, model errors can be corrected (fitting) and models can be built into unmodelled density blobs, using computer graphics software like *Coot* (Emsley and Cowtan, 2004; Emsley *et al.*, 2010). It often requires experienced crystallographer's assessment and knowledge of stereochemistry. After this real space refinement, reciprocal space refinement is performed again, and then more accurate phase information is obtained. This dual space recycling should be repeated until refinement converges.

Appendix: Derivation of $2mF_{\mathrm{o}} - DF_{\mathrm{c}}$ map coefficient

Let $\boldsymbol{F}_N$ and $\boldsymbol{F}_P^C$ be the true structure factor and the structure factor of partial model with coordinate error, respectively (Fig. 2.4). $\boldsymbol{\delta}$ is defined by $\boldsymbol{\delta} = \boldsymbol{F}_N - D\boldsymbol{F}_P^C$. By cosine low,

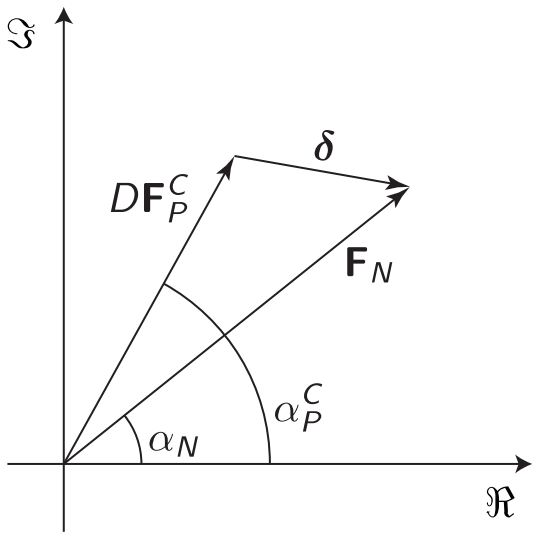


Figure 2.4: Relationship between $\boldsymbol{F}_N$ and $\boldsymbol{F}_P^C$.

$$\delta^2 = F_N^2 + D^2 F_P^{C\,2} - 2\cos(\alpha_N - \alpha_P^C) D F_P^C F_N. \tag{2.32}$$

Note that the expected amplitude and phase of $\boldsymbol{\delta}$ are uncorrelated with those of $\boldsymbol{F}_P^C$. By taking the expected values of both sides,

$$\langle \delta^2 \rangle = F_N^2 + D^2 {F_P^C}^2 - 2mDF_P^C F_N. \tag{2.33}$$

By using the relationship

$$F_N^2 = \boldsymbol{F}_N \boldsymbol{F}_N^* = \boldsymbol{F}_N (D\boldsymbol{F}_P^{C*} + \boldsymbol{\delta}^*), \tag{2.34}$$

the equation can be rearranged:

$$\langle \delta^2 \rangle = \boldsymbol{F}_N (D\boldsymbol{F}_P^{C*} + \boldsymbol{\delta}^*) + D^2 {F_P^C}^2 - 2mDF_P^C F_N. \tag{2.35}$$

By isolating mF_N, we get

$$mF_N = \frac{\boldsymbol{F}_N}{2DF_P^C}(D\boldsymbol{F}_P^{C*} + \boldsymbol{\delta}^*) + \frac{D}{2}F_P^C - \frac{\langle \delta^2 \rangle}{2DF_P^C}. \tag{2.36}$$

Mutiplying both sides by $e^{i\alpha_P^C}$, we get

$$\begin{aligned} mF_N e^{i\alpha_P^C} &= \frac{\boldsymbol{F}_N}{2D\boldsymbol{F}_P^{C*}}(D\boldsymbol{F}_P^{C*} + \boldsymbol{\delta}^*) + \frac{D}{2}\boldsymbol{F}_P^C - \frac{\langle \delta^2 \rangle}{2D\boldsymbol{F}_P^{C*}}, & (2.37) \\ &= \frac{\boldsymbol{F}_N}{2} + \frac{D}{2}\boldsymbol{F}_P^C + \frac{\boldsymbol{\delta}^*}{2}e^{2i\alpha_P^C} + \frac{\delta^2 - \langle \delta^2 \rangle}{2D\boldsymbol{F}_P^{C*}}. & (2.38) \end{aligned}$$

The Fourier transform of the last two terms will be a background noise (Main, 1979).

2.3 *LAFIRE*: automatic refinement system

As described in §2.2, refinement is an iterative process and often requires interventions of experienced crystallographers. It makes the structure analysis time-consuming. *LAFIRE* (Local correlation coefficient-based Automatic FItting for REfinement; Yao *et al.*, 2006; Zhou *et al.*, 2006) has been developed to automate this iterative refinement process. *LAFIRE* consists of three parts: (i) partial model building, (ii) model modification (fitting) including evaluation of the current model, and (iii) a process control system that includes interfaces with refinement programs. For refinement, users can choose which program is used from *CNS* 1.2 or 1.3 (Brünger *et al.*, 1998; Brunger, 2007), *REFMAC*5 (Murshudov *et al.*, 2011), *phenix.refine* (Afonine *et al.*, 2012), or *autoBUSTER* (Bricogne *et al.*, 2011). Note that *autoBUSTER* reports ${R^{\text{xpct}}}_{\text{work}}$ and ${R^{\text{xpct}}}_{\text{free}}$ factors instead of conventional R

factors, which are calculated using F_{xpct} instead of F_{model} (Blanc *et al.*, 2004). F_{xpct} is an expected structure amplitude based on the probability distribution $p(F_{\mathrm{obs}}; F_{\mathrm{model}})$. The refinement target can be either maximum likelihood using amplitude (MLF) or phased maximum likelihood using amplitude and Hendrickson-Lattman coefficients (MLHL). Besides *REFMAC*5, *LAFIRE* requires *CCP4* (Collaborative Computational Project, Number 4, 1994) programs mainly for handling reflection data.

Automatic refinement is performed based on the refinement strategy of *LAFIRE*, which is shown schematically in Fig. 2.5. First, the sequence of initial model is checked against

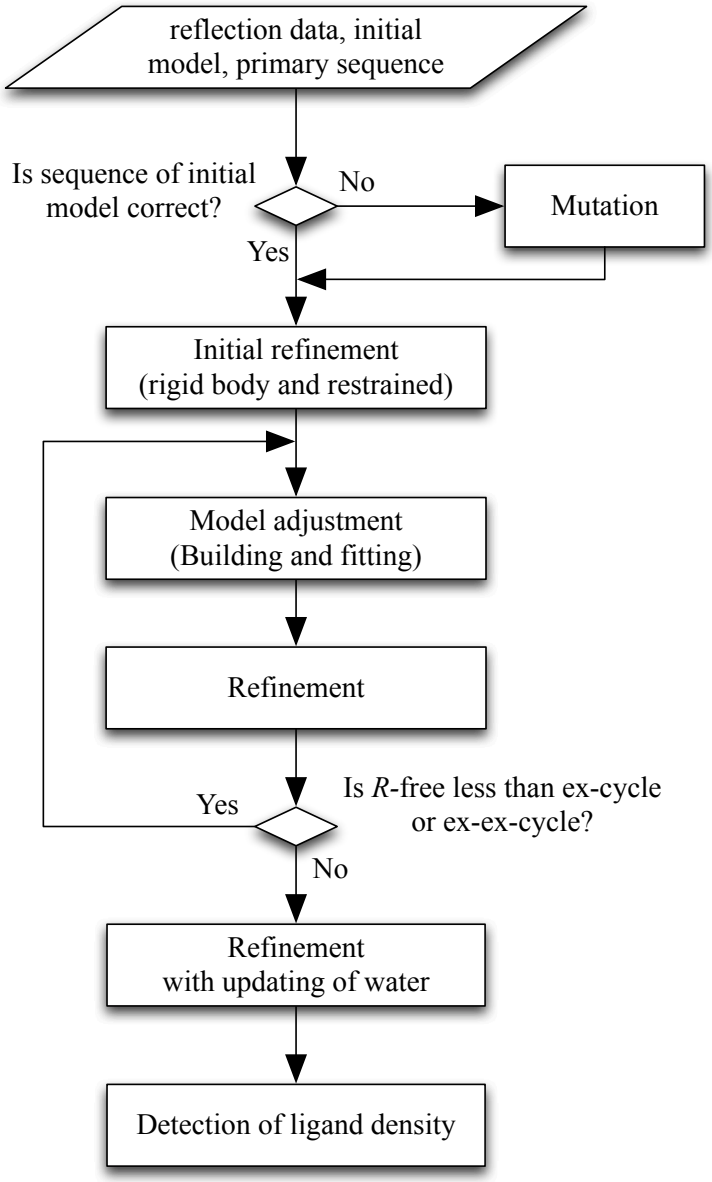


Figure 2.5: Schematic representation of the refinement strategy of *LAFIRE*.

the primary sequence given by user. Incorrectly assigned residues are mutated. Next, the sequence-corrected initial model is subjected to initial refinement starting from rigid body refinement. Rigid body domains are defined for each chain. Following the full atomic restrained refinement, fitting and building are performed based on the $2mF_o - DF_c$ map given by the refinement program. The fitted and extended model is then refined and the cycle of model adjustment and refinement procedures is iterated while R_{free} is less than ex-cycle (the previous cycle) or ex-ex-cycle (two cycles before). Using the model with lowest R_{free}, refinement with updating of water sites is started and iterated. Finally, the density blobs of possible ligands are identified.

LAFIRE has a graphical user interface (GUI) implemented with PyQt4. The GUI can be used for preparation of input files, parameters, and selection of molecule type to be fitted or extended. R_{free} and R_{work} factors and r.m.s.d. from ideal bond length values of each cycle are displayed while running *LAFIRE*. When the job is over, a summary of the refinement is displayed and *PyMOL* (Schrödinger, LLC, 2010) and *Coot* (Emsley *et al.*, 2010) are ready to begin visual inspection of the result.

In this study, *NAFIT* and *NABUILD*, the model fitting and completion programs for nucleic acid, are incorporated into the model adjustment procedure after running the refinement program.

2.4 Nucleic acid structure

Nucleic acids are biological molecules which play important roles such as accumulation of genetic information in organisms. Nucleotide, a monomer of nucleic acid, consists of base (nucleobase), sugar (β-D-ribose for RNA or β-D-2′-deoxyribose for DNA), and phosphate (Fig. 2.6). Nucleotide varies in base (Fig. 2.7). Primary bases are adenine (A), guanine (G), cytosine (C), thymine (T) for DNA. For RNA, uridine (U) is used instead of thymine. Primary bases are classified into purine derivatives (A, G) and pyrimidine derivatives (C, U, T). For purine derivatives, N_9 atom and $C_{1'}$ atom in sugar forms β-glycoside bond. For pyrimidine derivatives, N_1 forms β-glycoside bond with $C_{1'}$. Nomenclature of nucleic acid atoms and torsion angles follows the recommendations by IUPAC-IUB joint commission on biochemical nomenclature (1983).

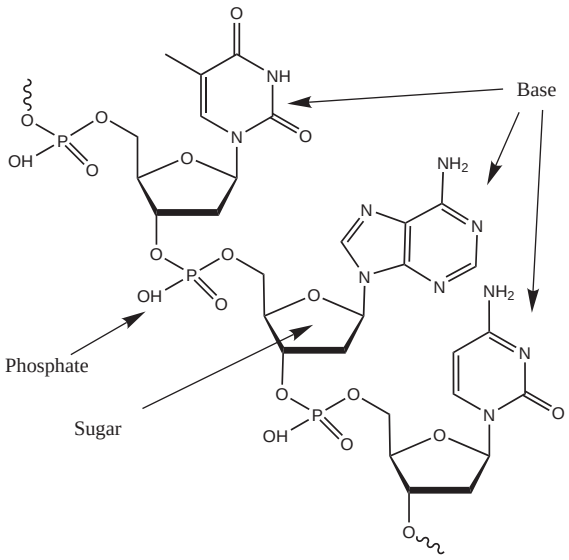


Figure 2.6: Nucleic acid structure (DNA).

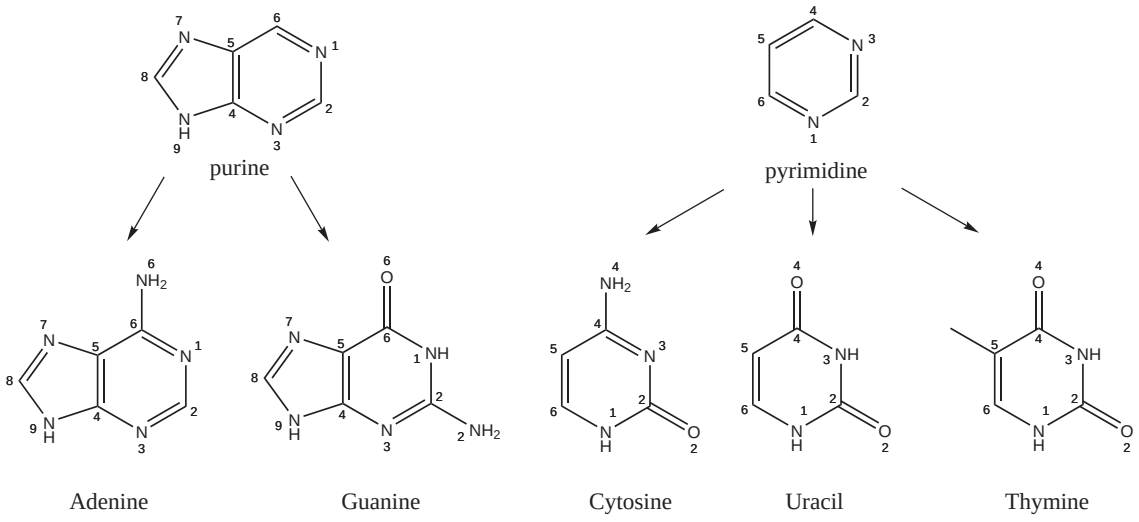


Figure 2.7: Structure of nucleobases.

In nucleic acid, $O_{5'}$ in sugar is phosphorylated and forms phosphodiester bond with previous nucleotide (Fig. 2.8). Names of chemically equivalent two oxygen atoms in phosphate group, O_1 and O_2 are defined clockwise when viewed $O_{5'}$ along the direction $O_{3'} \rightarrow P$.

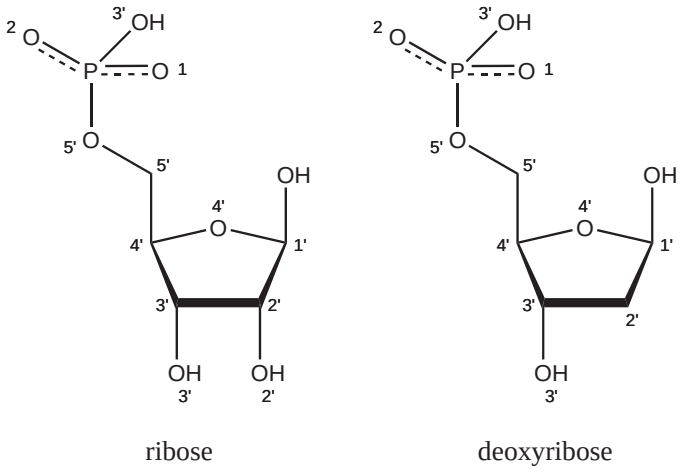


Figure 2.8: Nomenclature of sugar and phosphate atoms.

2.4.1 Definitions of torsion angles

Three dimensional molecular structures are defined by bond lengths, bond angles, and rotations of atoms along bonds. For bonded atoms A–B–C–D, the rotation around B–C bond is described by torsion angle. The torsion angle is defined as an angle between A–B bond and C–D bond when viewed from B→C direction.

In nucleic acid, all bonds in sugar and phosphate, and glycoside bond are rotatable. Torsion angle nomenclature is described in Table 2.1 and Fig. 2.9.

Torsion angle	Atoms
α	$_{(i-1)}O_{3'}$ - P - $O_{5'}$ - $C_{5'}$
β	P - $O_{5'}$ - $C_{5'}$ - $C_{4'}$
γ	$O_{5'}$ - $C_{5'}$ - $C_{4'}$ - $C_{3'}$
δ	$C_{5'}$ - $C_{4'}$ - $C_{3'}$ - $O_{3'}$
ε	$C_{4'}$ - $C_{3'}$ - $O_{3'}$ - $P_{(i+1)}$
ζ	$C_{3'}$ - $O_{3'}$ - $P_{(i+1)}$ - $O_{5'(i+1)}$
χ	$O_{4'}$ - $C_{1'}$ - N_1 - C_2 (Py)
	$O_{4'}$ - $C_{1'}$ - N_9 - C_4 (Pu)
ν_0	$C_{4'}$ - $O_{4'}$ - $C_{1'}$ - $C_{2'}$
ν_1	$O_{4'}$ - $C_{1'}$ - $C_{2'}$ - $C_{3'}$
ν_2	$C_{1'}$ - $C_{2'}$ - $C_{3'}$ - $C_{4'}$
ν_3	$C_{2'}$ - $C_{3'}$ - $C_{4'}$ - $O_{4'}$
ν_4	$C_{3'}$ - $C_{4'}$ - $O_{4'}$ - $C_{1'}$

Table 2.1: Torsion angles in nucleotide.

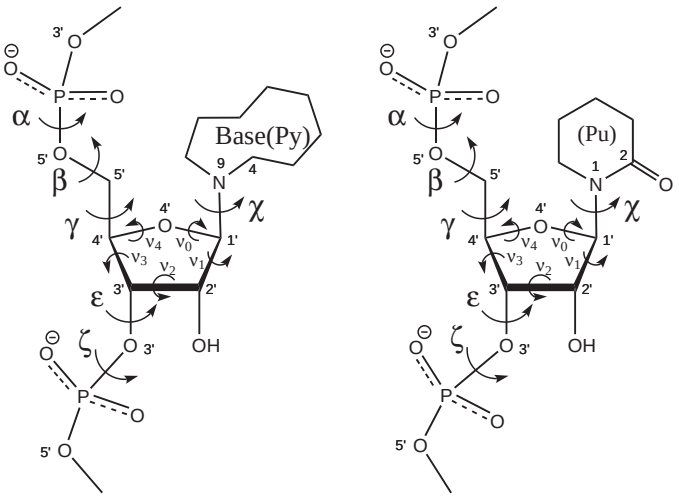


Figure 2.9: Torsion angles in nucleotide.

Chapter 3

Model fitting by *NAFIT*

Fitting is a fine-tuning process of a given model to improve the fit to the electron density and the model quality. The fit to the electron density map is evaluated by density-weighted grouped local correlation coefficient (GLCC).

$$\mathrm{GLCC}_i = g_i \frac{\sum_{j\in i} \rho_{\mathrm{obs}j} Z_j}{\left[\sum_{j\in i} \rho_{\mathrm{obs}j}^2 \sum_{j\in i} Z_j^2\right]^{1/2}}, \tag{3.1}$$

$$g_i = \frac{1}{N} \int_{\min(\rho_{\mathrm{obs}})}^{\langle\rho_{\mathrm{obs}}\rangle_i} h(\rho_{\mathrm{obs}}) d\rho_{\mathrm{obs}}. \tag{3.2}$$

Here, g_i is a weighting factor that accounts for the quality of the density map of the group i, $h(\rho_{\mathrm{obs}})$ is the number of grid points that have a density value of ρ_{obs}, $\min(\rho_{\mathrm{obs}})$ is the minimum density value of the map, and N is the number of grid points in the map. The average value $\langle\rho_{\mathrm{obs}}\rangle_i$ is calculated from atoms in the group i. $h(\rho_{\mathrm{obs}})$ is constructed in the asymmetric unit. The formulation is almost the same as described (Yao *et al.*, 2006), but Z_j (atomic number of atom j) is used instead of ρ_{calc} to reduce computational complexity. GLCC is used for evaluation, and is not used in fitting because a less computationally expensive function is used instead. Geometric quality includes bond lengths, bond angles, torsion angles, chiral volumes, planes, and steric clashes.

The fitting function is designed based on real space refinement of *Coot* (Emsley and Cowtan, 2004) with the additional features discussed later. Fitting is performed by minimization of the target function E.

$$E = E_{\mathrm{geometry}} + wE_{\mathrm{map}}. \tag{3.3}$$

Here, w is the fitting weight, and

$$E_{\text{geometry}} = E_{\text{bond}} + E_{\text{angle}} + E_{\text{chiral}} + E_{\text{plane}} + E_{\text{nonbonded}} + E_{\text{naconf1d}}, \quad (3.4)$$

$$E_{\text{map}} = -\sum_{i}^{N_{\text{atoms}}} Z_i \rho(x_i, y_i, z_i). \quad (3.5)$$

Here, Z is the atomic number and $\rho(x, y, z)$ is the density at the point (x, y, z), which is calculated by cubic interpolation of a given electron density map using *Clipper* library (Cowtan, 2002). Formulations of $E_{\text{bond}}, E_{\text{angle}}, E_{\text{chiral}}, E_{\text{plane}}, E_{\text{nonbonded}}$ and their derivatives are the same as in *Coot* (revision 4120 in SVN). E_{naconf1d} is defined in §3.1. Minimization is performed using conjugate gradient method implemented in *GSL* (Galassi *et al.*, 2009).

3.1 Conformational restraints

In *NAFIT*, E_{naconf1d} is introduced to yield a reasonable nucleic acid conformation. In contrast to peptides, which have two rotatable bonds in the main chain, the nucleotide main chain is more flexible with six rotatable bonds (§2.4.1). In addition, as described above, nucleic acid structure is often solved at medium or low resolution. This makes it more difficult to interpret the electron density map and determination of the accurate conformation is difficult (Wang *et al.*, 2008). Therefore, an appropriate restraint is needed for fitting nucleic acid structures at medium or low resolution to avoid unfavorable conformation. E_{naconf1d} is introduced for this purpose. This involves a one-dimensional conformational restraint based on an empirical distribution function. A similar function is used in the RNA structure prediction program *FARFAR* (Das *et al.*, 2010).

$$E_{\text{naconf1d}} = -\sum_{\theta} \sum_{i}^{N_\theta} \log p_\theta(\theta_i). \quad (3.6)$$

Here, p_θ is the frequency of torsion angle θ ($\theta = \alpha, \beta, \gamma, \delta, \varepsilon, \zeta, \chi$). p_θ and its derivative $\partial p_\theta / \partial \theta$ are calculated continuously using cubic spline interpolation with periodic boundary conditions by *GSL* (Galassi *et al.*, 2009). The conformational restraint could be extended to multidimension for correlated angles such as δ-ε.

To construct the empirical distribution function p_θ, we downloaded "RNA09" coordinate files from Richardson's website (http://kinemage.biochem.duke.edu/databases/rnadb.php), which are non-redundant RNA structures solved at 3.0 Å resolution or better (Richardson

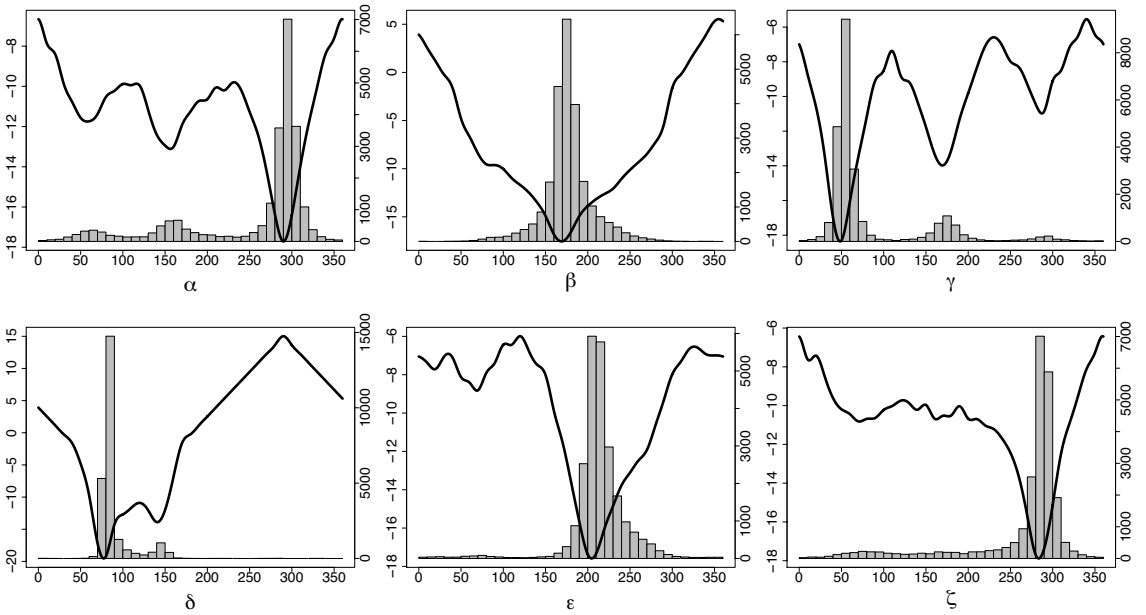


Figure 3.1: One-dimensional conformational restraint functions (left axis, heavy lines) and frequency (right axis, histogram) for RNA. The restraint for χ angle is not shown.

et al., 2008). The frequencies of torsion angle values for each entry were calculated with the statistical package *R* (R Development Core Team, 2008). Frequencies below a threshold in each bin, which may be outliers, were replaced with zero. Then, such frequencies were replaced by small values so that a gradient toward the nearest positive frequency was made as in Ramachandran restraint construction in *Coot* Emsley *et al.* (2010). The constructed restraint functions are shown as heavy lines in Fig. 3.1.

3.2 Ribose pucker correction for RNA

For RNA, ribose pucker is usually limited to either $C_{3'}$-endo or $C_{2'}$-endo (Richardson *et al.*, 2008). However, it is difficult to determine correct ribose pucker at low or medium resolution, and therefore ribose pucker was shown to be error-prone by Protein Data Bank X-ray Validation Task Force (Read *et al.*, 2011). Correct ribose pucker of RNA can be deduced based on base-phosphate perpendicular distance (Davis *et al.*, 2007; Keating and Pyle, 2010). Ribose pucker is corrected in *NAFIT*. As the radius of convergence by minimization is limited and E_{naconf1d} gives local minima for both $C_{3'}$-endo and $C_{2'}$-endo, it is necessary to construct good starting coordinates before minimization. Residues with in-

compatible base-phosphate perpendicular distance and δ angle are detected and subjected to simple rebuilding as follows. The ribose atoms are replaced with those of deduced puckering, then rotated around the axis $C_{5'}$– $C_{1'}$ to give an acceptable ε angle ($\sim 208°$). Finally, minimization is performed for fine-tuning of atomic coordinates. This procedure is very simple and gives reasonable coordinates with correct puckering.

3.3 Non-bonded atom interactions and hydrogen atoms

$E_{\text{nonbonded}}$ is a repulsive term for non-bonded atom pairs. Non-bonded interactions are considered to the fourth and more distant atoms from each moving atom with the exception of atoms in rings. The interactions with non-moving atoms whose positions are not refined in the minimization are also considered to avoid steric clashes. The critical distance is defined as the sum of van der Waals radii multiplied by 0.9, which gives a reasonable compromise between geometric quality and density fit. The atoms whose distance is larger than the critical distance have no contribution to $E_{\text{nonbonded}}$.

To maintain geometric quality, hydrogen atoms are automatically generated before the fitting procedure. They contribute to the geometric term, but Z for each hydrogen is set to 0 so that they do not contribute to E_{map} in equation (3.3). Fitting including hydrogen atoms greatly reduces interatomic clashes without deteriorating fit to the electron density map. For hydrogen bonding atoms, the critical distance is reduced by 0.3 Å. The fixed value of 1.8 Å is used for hydrogen atom pairs. These constants are obtained empirically.

3.4 Sequential grouped fitting

NAFIT performs fitting on a chain-by-chain basis. The following procedure is repeated starting from the first residue to the end residue:

i. Residues within 5 Å from the current residue are selected for grouped fitting (like the so-called "sphere refinement" in *Coot*).

ii. Atoms not selected are fixed and take part in non-bonded interaction (including other chains).

iii. If base pair information is provided, pseudo bonds of base pairs are generated.

iv. Hydrogen atoms are generated.

v. The target function E is minimized.

By this sequential grouped fitting, minimization is expected to be stably converged due to limited number of variables and neighboring residues can be fitted into the electron density map with resolution of steric clashes.

Chapter 4

Chain extension by *NABUILD*

For model completion by *NABUILD*, two building modes are available: (i) gap-building connects residues already built and (ii) terminal-building starts building from the termini. *NABUILD* finds the main chain path extracted from the skeletonized electron density map. Map skeletonization is performed after extracting the density blobs with each density value above 0.8 r.m.s.d. However, the main chain cannot always be traced on the connected density blob, and in such cases models cannot be built due to the gap in density. To overcome this problem, an artificial density blob is inserted to connect blobs that lie within a certain distance before skeletonization (Fig. 4.1). This blob insertion increases false main chain paths but they can be easily removed by scoring.

If the main chain path is identified, a coarse-grained nucleic acid model is constructed. We define P, S, and B points as representative points for three parts of the nucleotide. The set of P, S, and B is called PSB. For a given nucleotide, P, S, and B are defined as the phosphorous atom coordinate, center of sugar ring atoms ($C_{1'}$, $C_{2'}$, $C_{3'}$, $C_{4'}$, and $O_{4'}$), and center of base atoms, respectively. The B direction is also defined as the vector normal to the base plane. In building, PSB points and B direction are determined and then atomic coordinates are calculated.

The procedure for chain completion by *NABUILD* is as follows.

i. The missing residue ranges are identified by comparing PDB file and sequence file.

ii. $2mF_{\mathrm{o}} - DF_{\mathrm{c}}$ map is calculated with a grid spacing of about 0.5 Å.

iii. Based on the initial model, positions near the existing B positions are recorded along

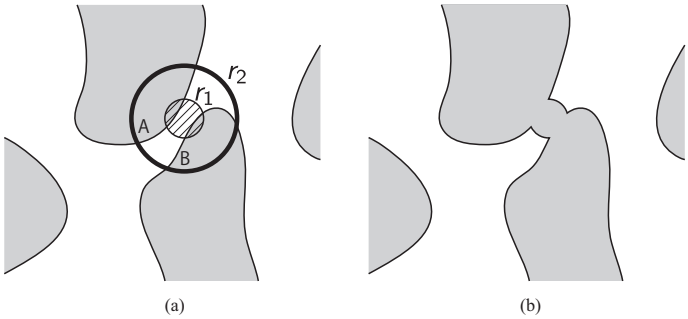


Figure 4.1: Gap filling before skeletonization. Gray regions represent electron density blobs. For each grid point, blobs are clustered within r_2 from the point (A and B in (a)). If the points in the sphere with radius r_1 belong to more than one cluster, a pseudo-blob is inserted (b).

with B directions as possible B positions and directions because they may form stacking interactions or base pairs (Fig. 4.2).

iv. The density map is skeletonized (the ridges connecting density peaks are traced) using *Clipper* library (Cowtan, 2002) after binarization with a threshold of 0.8 r.m.s.d. and gap filling with $r_1 = 1.0$ Å, $r_2 = 2.0$ Å (Fig. 4.1). These constants are obtained in a few trials and there should be room for improvement.

v. Skeleton points around residues already built are removed with the exception of sugar atoms at the $3'$ terminus and phosphate group at the $5'$ terminus. Skeleton points around protein molecules are also removed.

vi. Chain extensions for the residue ranges are performed to collect up to 10 candidates for each range (§4.1).

vii. Candidates are merged into chains (§4.2).

4.1 Building on extracted path

Building is performed over a limited radius to avoid excessive computation. For terminal building, a graph covering the skeleton points within $\min(3.10N + 17.1, 50)$ Å from the current terminal phosphate group is constructed, where N is the number of residues to be

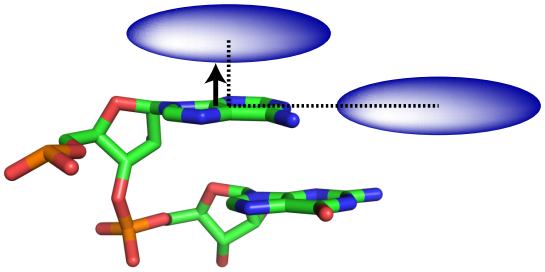


Figure 4.2: Recording positions close to existing B positions, along with B directions. If base existed in the shaded region, it could be expected that its direction was nearly the same as that of existing base in the neighborhood.

built. For gap building, a graph covering the skeleton points within $\min(1.70N + 13.3, 50)$ Å from the midpoint of the terminal phosphate groups is constructed. The relational expressions between N and the radius was obtained by linear regression analysis of maximum distance on N using the large ribosomal subunit structure in 2.2 Å resolution (Schmeing *et al.*, 2005, PDB code: 1vq8). If building from the $3'$ terminus fails, building from the $5'$ terminus is performed. Following path extraction and determination of atomic coordinates, the candidates are fitted and $\langle \mathrm{GLCC} \rangle$ is calculated.

4.1.1 Path extraction

The skeleton points in the defined space are extracted and a graph is constructed where the vertices are each skeleton point and the edge weight is the distance. Skeleton points nearest to the start and end points are determined. Possible paths from the graph that run from start to end, and the P positions are found (Fig. 4.3 (a)). For terminal building, the end point is not defined. The paths are found by breadth-first search[1)] keeping up to 1000 intermediate paths. If number of intermediate paths exceeds 1000, those with lower scores are removed.

Path score is defined based on P positions as $\langle Z_{\mathrm{P}} \rangle - \langle p \rangle$, where p is planarity score of the blob belonging to P and Z_{P} is electron density Z score at P position calculated

[1)]Breadth-first search (BFS) is a commonly known algorithm in graph theory, which systematically searches all vertices. BFS begins at a starting vertex and inspects the neighboring vertices. Then, for each of those neighbors, it inspects their neighbors unvisited. This procedure is repeated until all reachable vertices are inspected.

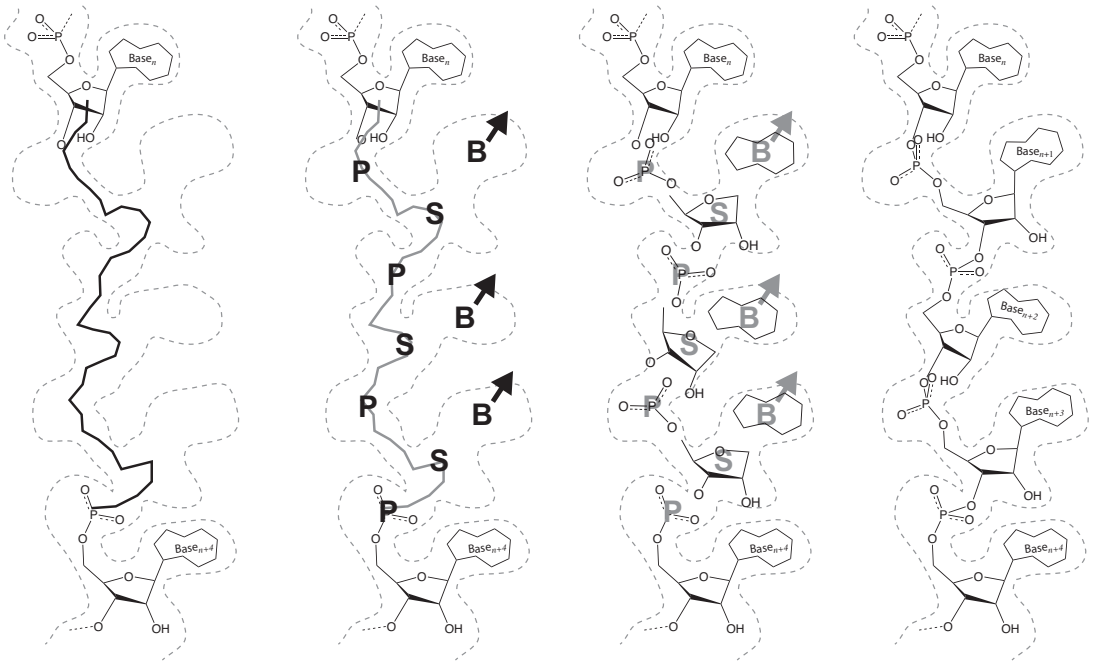


Figure 4.3: *NABUILD* procedure. (a) The path from the 3′ sugar to 5′ phosphate is extracted. (b) Positions of P, S, and B, and B direction are determined. (c) Atoms of each part are roughly positioned and oriented. (d) Atomic coordinates are refined by fitting with mean position restraints.

from the mean and standard deviation of the density values at path points near the P position. Planarity score p is introduced to discriminate P from B. To calculate p, principal component analysis (PCA) (Pearson, 1901) is performed using the grid point coordinates of electron density map above a certain level within 2.2 Å from P. p is defined by Fisher's linear discriminant (Fisher, 1936) as a linear combination of λ_1/λ_3 and λ_1/λ_2, where $\lambda_{1,2,3}$ are eigen values given by PCA ($\lambda_1 \leq \lambda_2 \leq \lambda_3$). The discriminant is learned from real data and expected to be positive for a planar blob ($\lambda_1/\lambda_3 \ll 1, \lambda_1/\lambda_2 \ll 1$) and negative for a spherical blob ($\lambda_1 \simeq \lambda_2 \simeq \lambda_3$). The currently used discriminant is $-18.823\lambda_1/\lambda_3 + 1.2694\lambda_1/\lambda_2 + 6.3484$ and the map level is dynamically adjusted so that the volume of the blob is close to the expected value (~ 14 Å^3).

4.1.2 Finding PSB

The first P position is defined using the terminal residue. To build from the 3′ terminus, the path point from 2.0 to 4.3 Å (empirically determined numbers) away from S of the 3′ terminus with the maximum electron density is defined as P. To building from the 5′ terminus, the P position is redefined by the path point within 2.0 Å (empirically determined number) from the original P with the maximum density. The next P positions are defined sequentially. The $(i+1)$th P will be determined from the path point with the maximum density that satisfies the conditions.

i. The distance along the path between P_i and P_{i+1} is from 8 to 12 Å.

ii. The linear distance $\overline{\mathrm{P}_i\mathrm{P}_{i+1}}$ is from 4.5 to 7 Å.

iii. The angle $\angle \mathrm{P}_{i-1}\mathrm{P}_i\mathrm{P}_{i+1}$ is larger than 37.8°.

These constants were defined by analyzing RNA09 data (the minimum or the maximum values were taken after manual inspection to exclude outliers).

Top 100 paths are subjected to PSB position calculation. S_i is determined as the midpoint of P_i and P_{i+1} along the path. B_i is determined based on the electron density from geometrically allowed points. Similar PSB candidates with r.m.s.d.(PSB) less than 1.0 Å are removed. For each PSB remaining, the B direction is calculated based on the shape of the density blob and prerecorded possible positions and normal vectors (Fig. 4.3 (b)).

4.1.3 PSB to atomic coordinates

Based on the PSB positions and the B direction, atomic coordinates are calculated sequentially (Fig. 4.3 (c)). First, the phosphate group is placed at the P position, with $O_{5'}$ directed toward S. Next, the center of the base atoms is placed at B, the base plane is oriented along B direction, and the nitrogen atom forming a glycosidic bond (N_9 for purine and N_1 for pyrimidine) is directed toward S. Then, the center of the sugar atoms is placed at S and oriented by aligning the normal vector of the sugar plane to $\overrightarrow{\mathrm{SB}} + \overrightarrow{\mathrm{SP}}$, and $C_{5'}$ is directed toward P. Translation and rotation of sugar atoms are optimized by the downhill simplex method to satisfy geometric conditions.

The placed atoms are fitted against the electron density together with the adjacent residue. In this fitting, the PSB positions are used as target positions for "mean position restraint" (Fig. 4.3 (d)). The mean position restraint is designed for restraining group of atoms $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ to a given position $\boldsymbol{a}$ (P, S, or B). The equation added to the target function of fitting is defined as

$$E_{\mathrm{mpos}} = \frac{1}{\sigma^2} \left(\frac{\sum_i^N \boldsymbol{x}_i}{N} - \boldsymbol{a} \right)^2 . \tag{4.1}$$

Orientation of base atoms should be compared to the flipped one because both orientations could be fitted almost equally to the electron density of typical resolution. To determine the most likely base orientation, GLCC and χ angle are compared between the original and flipped coordinates. If the difference of GLCC is less than 0.05, the orientation is determined based on the χ frequency, otherwise that with the larger GLCC is chosen.

4.2 Merging candidates

When candidates for each missing residue range are collected, a single candidate should be selected in each range and integrated into a single chain without severe steric clashes. First, the top candidates for each residue range are checked to determine whether they clash with each other. Base atoms are excluded from this calculation because base clashing can be resolved after merging. If more than one atomic contact within 2 Å exists, the candidates are marked as clashing pair. The candidates that are not involved in clashing are accepted. For clashing candidate pairs, a non-clashing candidate pair is searched among

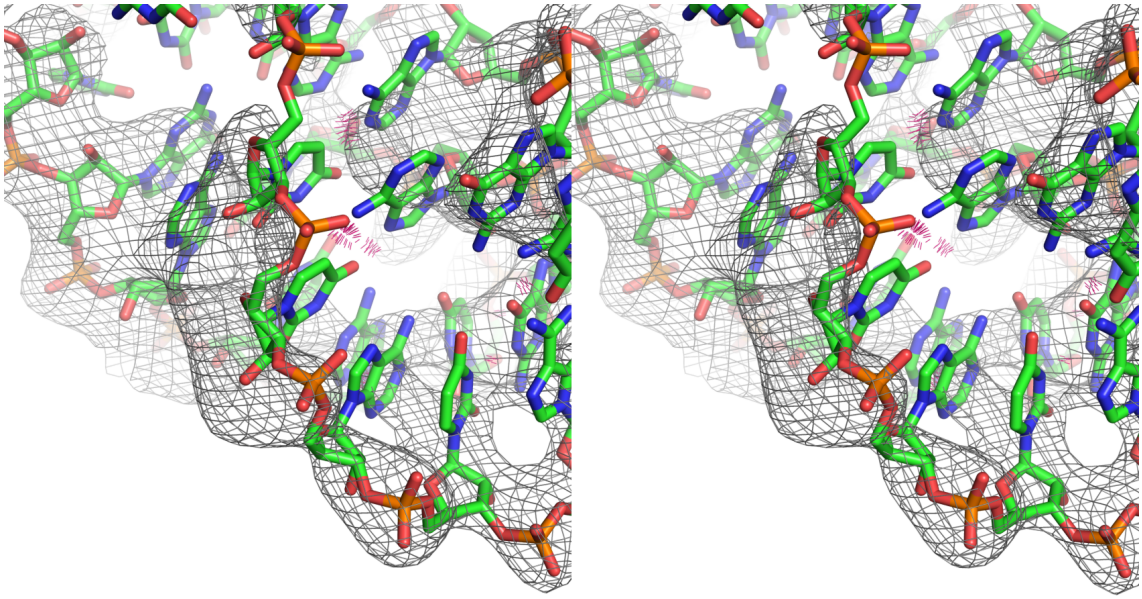
the candidates. All non-clashing candidates are integrated into a single chain. If clashing of base atoms is found, fitting is performed for the residues and those within 3.5 Å. In this fitting, phosphorus atoms are fixed.

Chapter 5

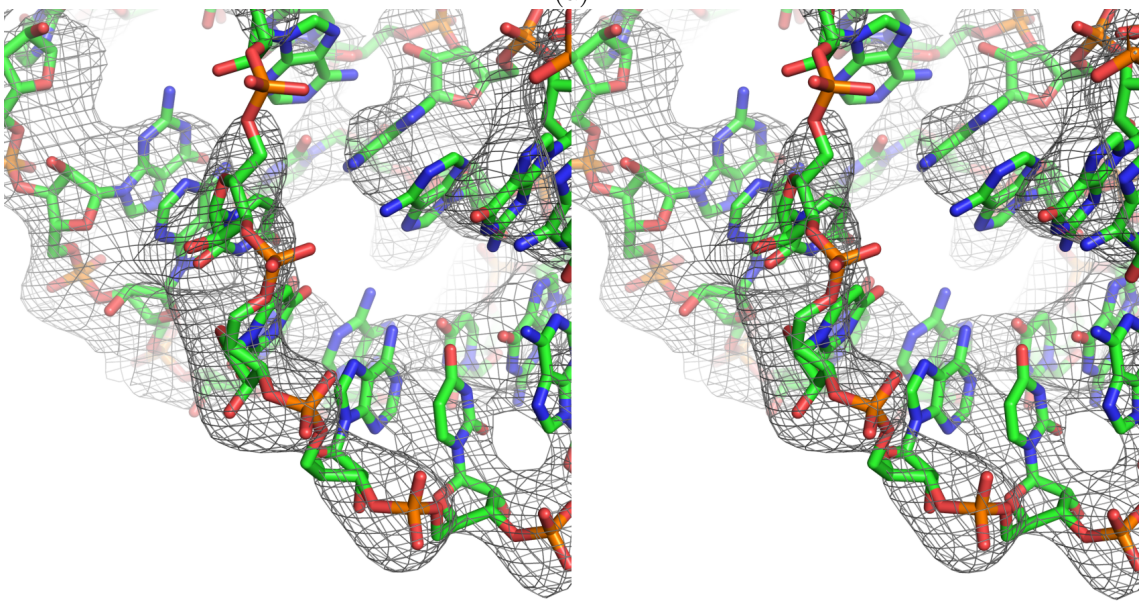
Test cases

5.1 Fitting by *NAFIT* to a low resolution map

We demonstrated the performance of *NAFIT* with $tRNA^{His}$ protein complex data at very low resolution (4.2 Å). The crystal structure of Thg1-$tRNA^{His}$ complex was solved by the molecular replacement method using the Thg1 structure. The electron density of tRNA appeared in both $2mF_o - DF_c$ and $mF_o - DF_c$ maps, and the $tRNA^{Phe}$ (Shi and Moore, 2000, PDB code: 1ehz) model was manually placed. Two tRNA chains of 75 nucleotides were present in the asymmetric unit. The tRNA sequence was then mutated according to $tRNA^{His}$. As some residues of the tRNA model showed a poor fit to the electron density map, they were roughly adjusted using *Coot*. Then, we used *NAFIT* to improve the manually adjusted model. The resultant model showed a better fit to the electron density map and a better geometric quality (Fig. 5.1). A fixed value of 2 was used for fitting weight w, which was determined in a number of trials. We performed individual sites and grouped ADP (one group per chain) refinement by *phenix.refine* (ver. dev-1218) with automatic secondary structure restraints and torsion angle NCS restraints (Headd *et al.*, 2012) to show how *NAFIT* enhanced model quality. We also performed the same refinement by *autoBUSTER* (ver. 2.10) with automatic NCS restraints. As a result, the refined model with *NAFIT* showed better geometric quality and explained diffraction data well compared to the refined model without *NAFIT* (Table 5.1).



(a)



(b)

Figure 5.1: Stereo drawing of the *NAFIT* result for Thg1-tRNA complex at 4.2 Å resolution. Target $2mF_o - DF_c$ electron density is shown only around the displayed molecule at the level of 1.0 r.m.s.d. Bad overlaps ($\geq$ 0.4 Å) are displayed as red spikes given by the *probe* command in *MolProbity* (Word *et al.*, 1999). (a) Before fitting. (b) After fitting, steric clashes are resolved and the model shows a better fit to the electron density map.

Table 5.1: Fitting of tRNA to the map of Thg1-tRNA complex at 4.2 Å resolution. This table shows that the fitting strategy with anti-bumping restraints including hydrogen atoms and naconf1d restraints greatly improved geometric quality and density fit at very low resolution. *phenix.refine* and *autoBUSTER* were used to refine and evaluate the initial model[†] and the improved model[‡]. Note that *autoBUSTER* reports R^{xpct} factors (see §2.3).

		⟨GLCC⟩*		Outliers***			
	R_{work}, R_{free}	chain1, chain2	clash score**	pucker	bond	angle	suite
(initial[†])		0.587, 0.374	36.1	18	26	30	0
NAFIT (none)		0.701, 0.575	81.4	4	0	0	0
NAFIT (+naconf1d)		0.680, 0.500	65.1	3	0	2	0
NAFIT (+H)		0.698, 0.565	7.93	9	0	0	0
NAFIT[‡] (+H+naconf1d)		0.670, 0.496	4.17	3	0	1	0
phenix.refine	0.3420, 0.4177	0.583, 0.363	23.4	6	0	0	0
NAFIT[‡]+*phenix.refine*	0.3370, 0.3671	0.614, 0.391	15.9	4	0	1	0
autoBUSTER	0.3722, 0.4054	0.531, 0.369	0.834	18	0	0	0
NAFIT[‡]+*autoBUSTER*	0.3256, 0.3578	0.663, 0.479	1.67	17	0	0	0

*Averaged GLCC is calculated for each chain. GLCC of chain2 is still poor even in the final model as the electron density around the molecule is poor.

**_MolProbity_ clash score is defined as the number of bad overlaps per 1000 atoms (Word *et al.*, 1999) and calculated by *phenix.clashscore* after removing protein atoms.

***RNA model validation is given by *phenix.rna_validate*. Bond/angle outliers are counted if deviation from the ideal value is larger than 4σ. Pucker outliers are ε outlier or incompatible δ and base-phosphate perpendicular distance (Davis *et al.*, 2007; Keating and Pyle, 2010). Suite outliers are given by *suitename* (Richardson *et al.*, 2008).

5.2 Building by *NABUILD* and iterative refinement

The test of *NABUILD* was performed along with iterative refinement because it is expected that improved phase given by refinement programs may enable us to build residues that cannot be built in the previous map.

5.2.1 Sample model preparation

We demonstrated the performance of *NABUILD* with eight data sets obtained from PDB (Table 5.2 and Fig. 5.2). For each PDB entry, we downloaded the coordinates files and the reciprocal space data including structure amplitudes, test flags for R_{free} calculation, and experimental phases if present. Experimental phase information is available for 2zm5, 3d0u, 3bwp, and 3sd1 in the form of Hendrickson-Lattman coefficients. For this test, we truncated all except those residues with the canonical A-form RNA conformation from the model deposited in PDB because such residues could be modelled relatively easily, *e.g.*, by searching for double-stranded helices. A-form RNA residues were identified as "1a" class by the *suitename* program (ver. 0.3.070628, Richardson *et al.*, 2008). All non-1a residues and isolated residues were removed to generate an initial model. Protein chains were added to the initial model if present in the model deposited in PDB. Other components, including ligands, ions, and water, were removed.

5.2.2 Protocol

The tests were performed using *LAFIRE* GUI. The pruned model as a starting model and the mtz file were provided. The refinement program was set to *phenix.refine* (ver. dev-1218) with MLF or MLHL (if available with the exception of 3d0u for which the phase quality was low) target and torsion angle NCS restraints if more than one copy existed in the asymmetric unit. In the model adjustment procedure, only nucleic acid chains were fitted and extended, and other polymer chains were fixed to avoid steric clashes with nucleic acids if present. Before *NABUILD*, the residues with GLCC lower than 0.8 built by *NABUILD* in previous cycles are removed. This criterion is determined empirically. The refinement scheme was based on *LAFIRE*'s refinement strategy, which is described in §2.3.

Table 5.2: Example data for the test of *NABUILD* and iterative refinement.

PDB code (chain)	$d_{\min}$ (Å)	space group	protein residues	RNA residues
1n78 (C, D)	2.10	$C222_1$	936	75, 75
3u56	2.10	$P2_12_12_1$	228	80
3sd1	2.27	$P2_12_12_1$	0	89
3dj2	2.50	$P2_12_12_1$	0	174
2zm5 (C, D)	2.55	$P2_12_12_1$	611	74, 69
3d0u	2.70	$P3_2$	0	161
3cul (C, D)	2.75	$C2$	183	92, 92
3bwp	3.12	$P2_12_12_1$	0	356

1n78 is $tRNA^{Glu}$ complexed with GluRS (Sekine *et al.*, 2003). 3u56 is an 80-nt 23S RNA complexed with mutant TthL1 (released in 2012 by Tishchenko *et al.*). 3sd1 is a tetrahydrofolate riboswitch, which was solved by the Ir-SAD method (Trausch *et al.*, 2011). 3dj2 is a lysine riboswitch (Serganov *et al.*, 2008). 2zm5 is $tRNA^{Phe}$ complexed with MiaA, which was solved by the Se-SAD method (Chimnaronk *et al.*, 2009). 3d0u is a lysine riboswitch, which was solved by the Ir-SAD method (Garst *et al.*, 2008). 3cul is an aminoacyl-tRNA synthetase ribozyme (Xiao *et al.*, 2008). 3bwp is a group II intron, which was solved by the MAD method with a combination of two derivatives (Toor *et al.*, 2008).

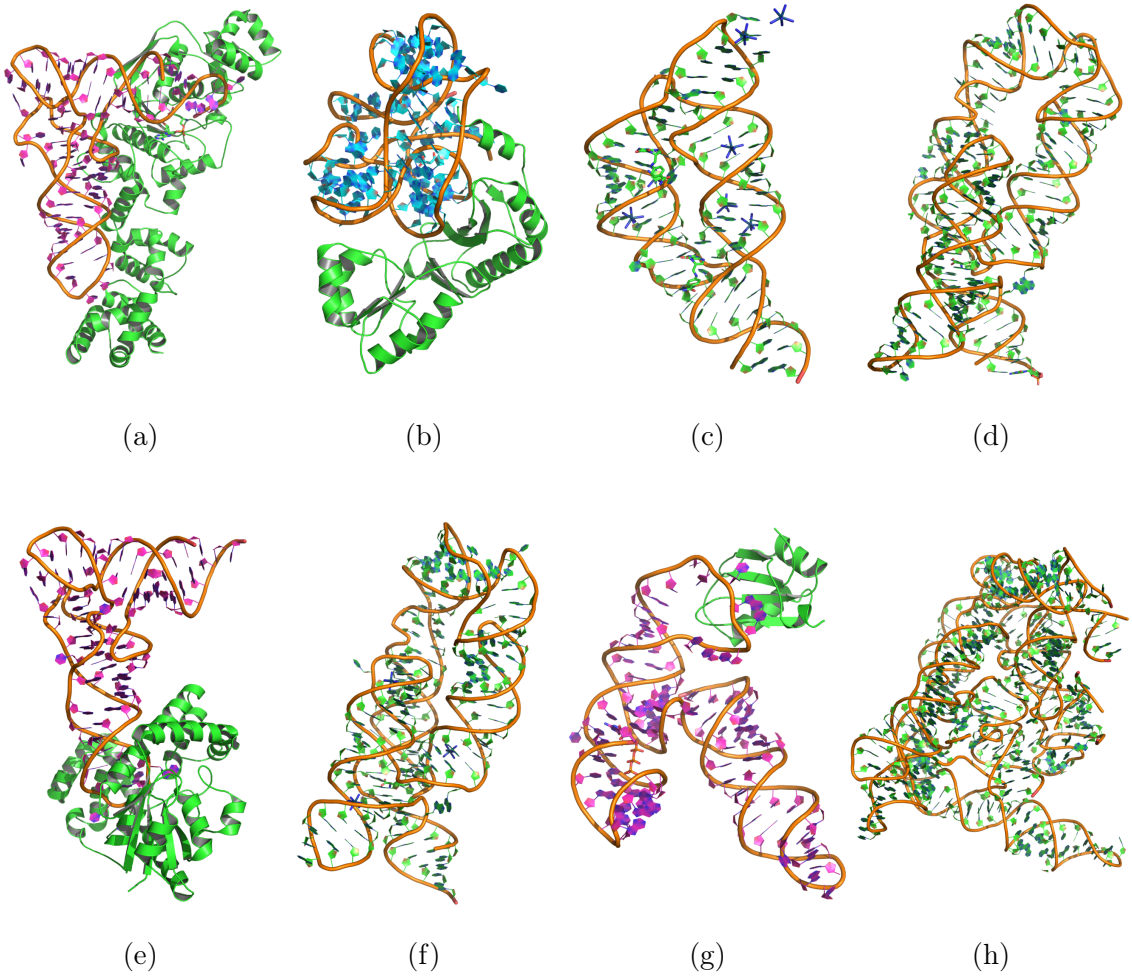


Figure 5.2: 3D structures of examples for the test of *NABUILD* and iterative refinement. (a) 1n78 (chain A&C), (b) 3u56, (c) 3sd1, (d) 3dj2, (e) 2zm5 (chain A&C), (f) 3d0u, (g) 3cul (chain A&C), (h) 3bwp. Chain IDs are not denoted if only single chain is present.

5.2.3 Results

The results of iterative refinement by *LAFIRE* with *NAFIT* and *NABUILD* are summarized in Table 5.3. Detailed results are shown in Fig. 5.3.

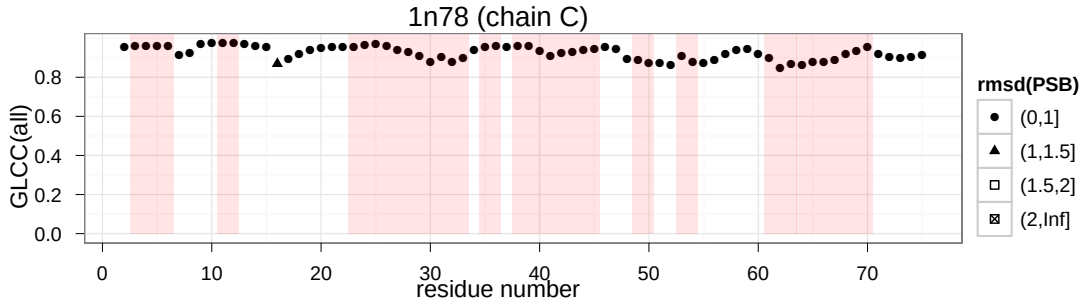
In all cases, significantly lower R_{free} values were obtained than the starting models. In some cases, especially when the resolution was high, model building and refinement were nearly completed. The refinement was always satisfactory even though the model was not completed. For the 2zm5 test case, the C chain was almost perfectly built (Fig. 5.4), while the D chain was incomplete due to the very poor electron density. Although manual intervention was still needed, it was greatly reduced.

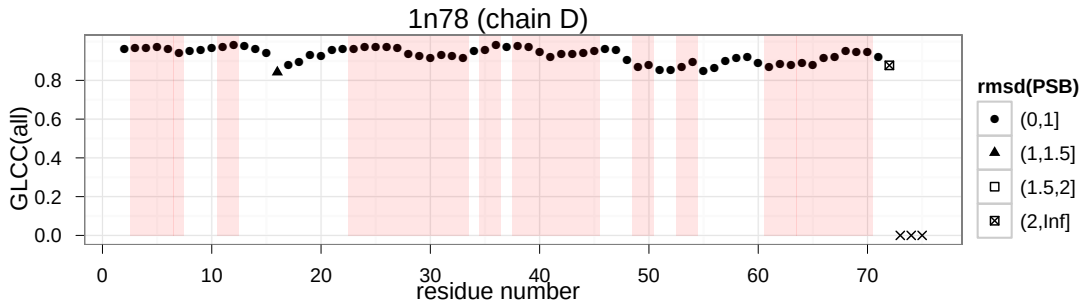
Table 5.3: Results of *NABUILD* and iterative refinement. If the r.m.s.d. value of P, S, and B positions between the built model and the PDB model is less than 1.5 Å in each residue, the residue is counted as "correctly built".

	number of residues							
PDB code (chain)	needed to be built	correctly built*	wrongly built*	unbuilt	R_{free} initial	R_{free} final	R_{free} of PDB model†	refinement target
1n78 (C)	33	33 (33)	0 (0)	0	0.3085	0.2386	0.2520	MLF
1n78 (D)	32	28 (29)	1 (0)	3				
3u56	32	30 (32)	2 (0)	0	0.3436	0.2286	0.2321	MLF
3sd1	29	28 (28)	1 (1)	0	0.4689	0.3047	0.2634	MLHL
3dj2	49	42 (43)	3 (2)	4	0.3868	0.2575	0.2626	MLF
2zm5 (C)	47	45 (46)	2 (1)	0	0.3870	0.2864	0.2637	MLHL
2zm5 (D)	41	9 (9)	10 (10)	22				
3d0u	55	52 (53)	3 (2)	0	0.3550	0.2084	0.2077	MLF
3cul (C)	32	23 (25)	9 (7)	0	0.4099	0.2874	0.2797	MLF
3cul (D)	25	22 (23)	3 (2)	0				
3bwp	165	91 (97)	28 (22)	46	0.4552	0.3374	0.2873	MLHL

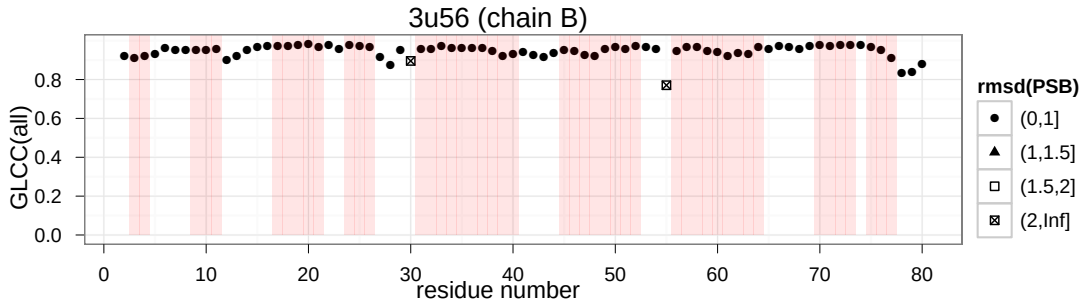
*Values in parentheses represent the number of correctly/wrongly built main chain residues.

† Evaluated by *phenix.maps* with bulk solvent correction and anisotropic scaling using the model deposited in PDB. Note that the final model given by *LAFIRE* does not include ligands and ions.

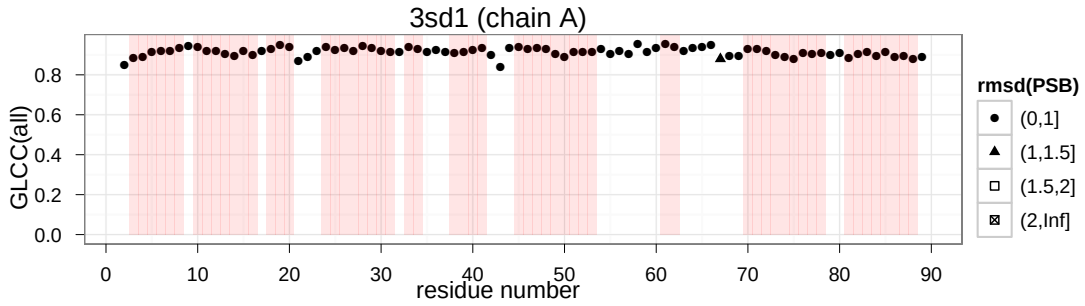




(a)

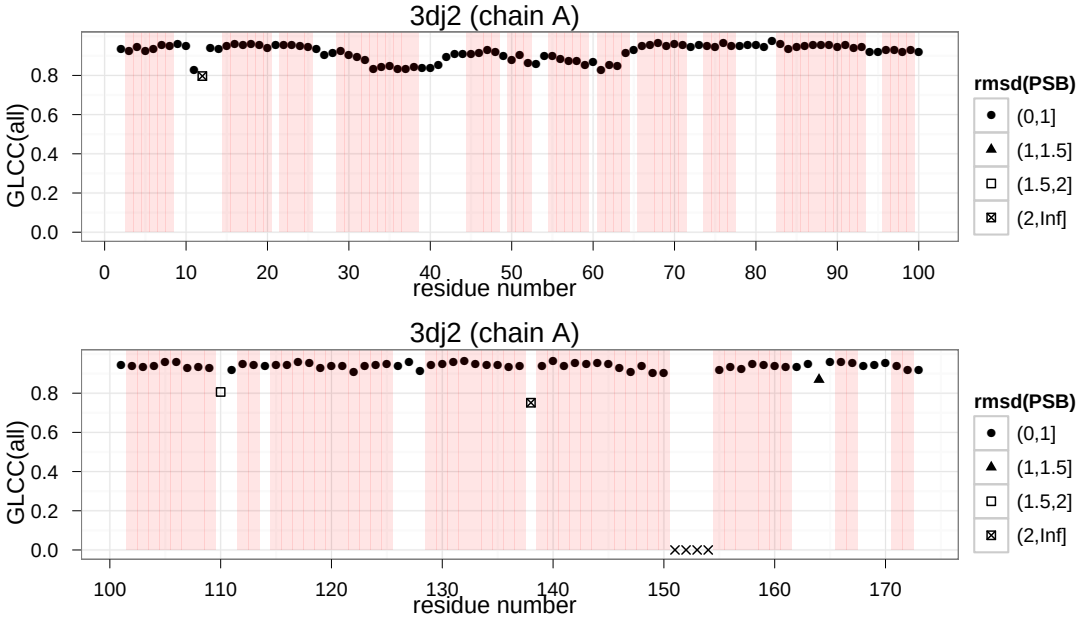


(b)

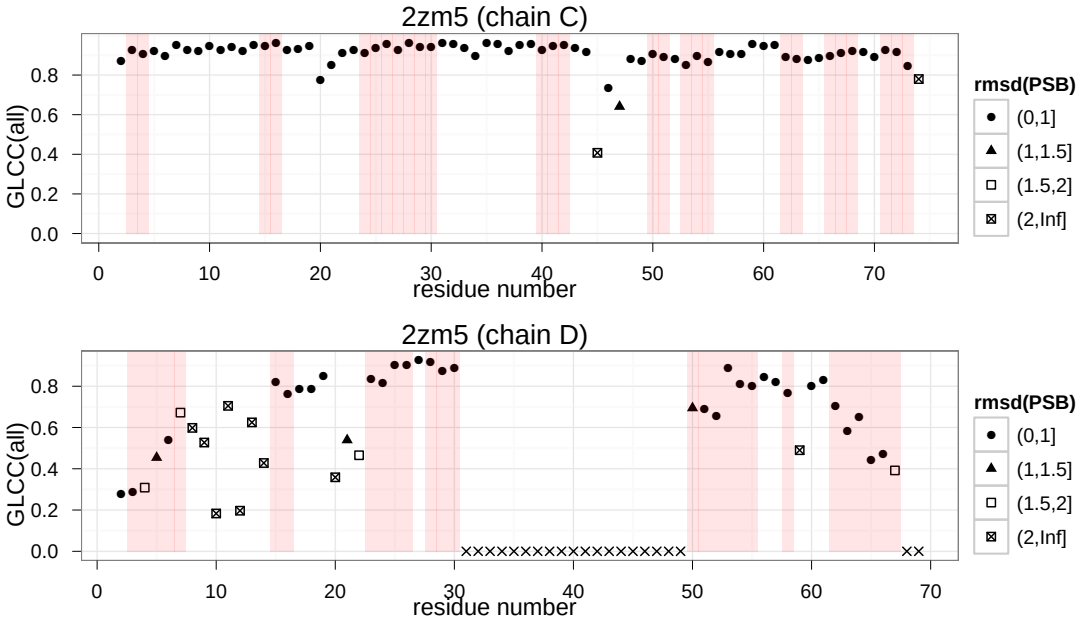


(c)

Figure 5.3: Detailed results of *NABUILD* and iterative refinement. GLCC was calculated using the final $2mF_{\mathrm{o}} - DF_{\mathrm{c}}$ map. Shaded regions represent the residues which are already included in the starting models. The symbols represent r.m.s.d.(PSB) between built residue and correct residue of PDB model. $(a, b]$ represents left-open, right-closed interval ($a <$ r.m.s.d. $\leq b$). '$\times$' marks on the base line represent the unbuilt residues. (a) 1n78. (b) 3u56. (c) 3sd1.

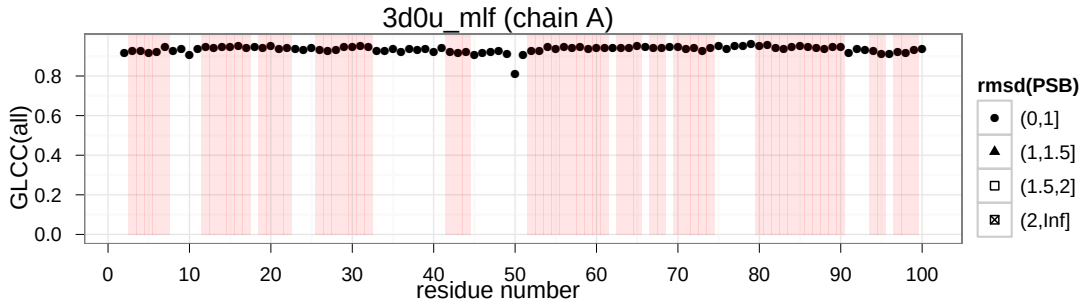


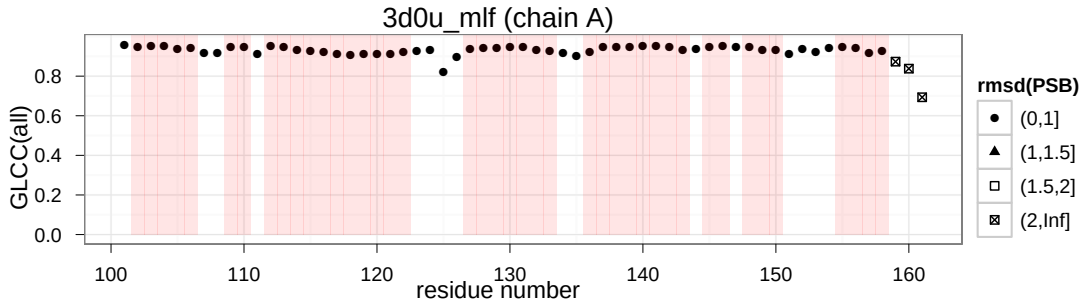
(d)



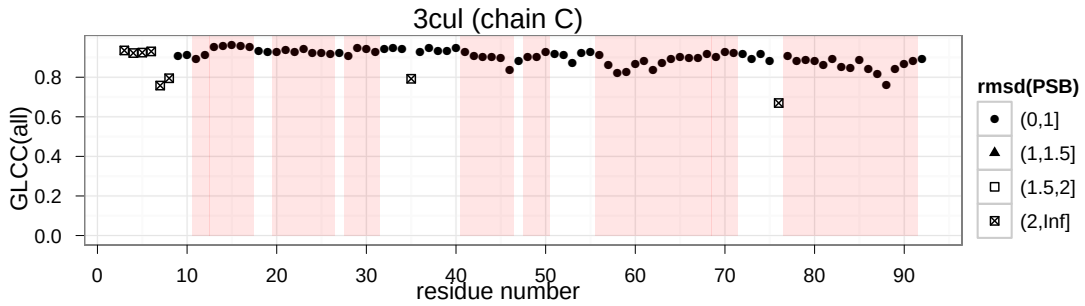
(e)

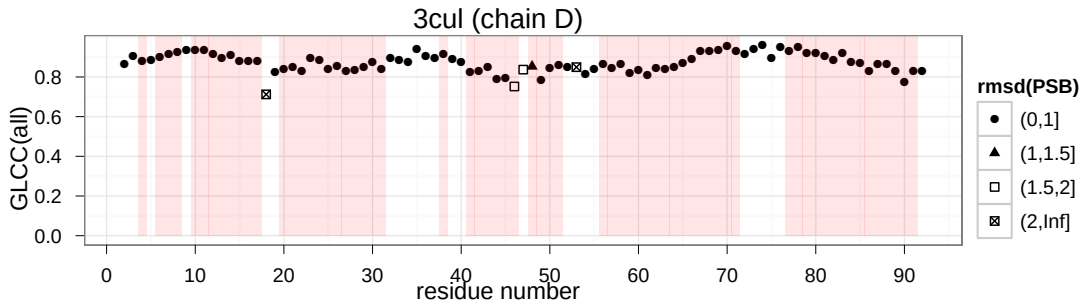
Figure 5.3 (*cont.*): (d) 3dj2. (e) 2zm5.





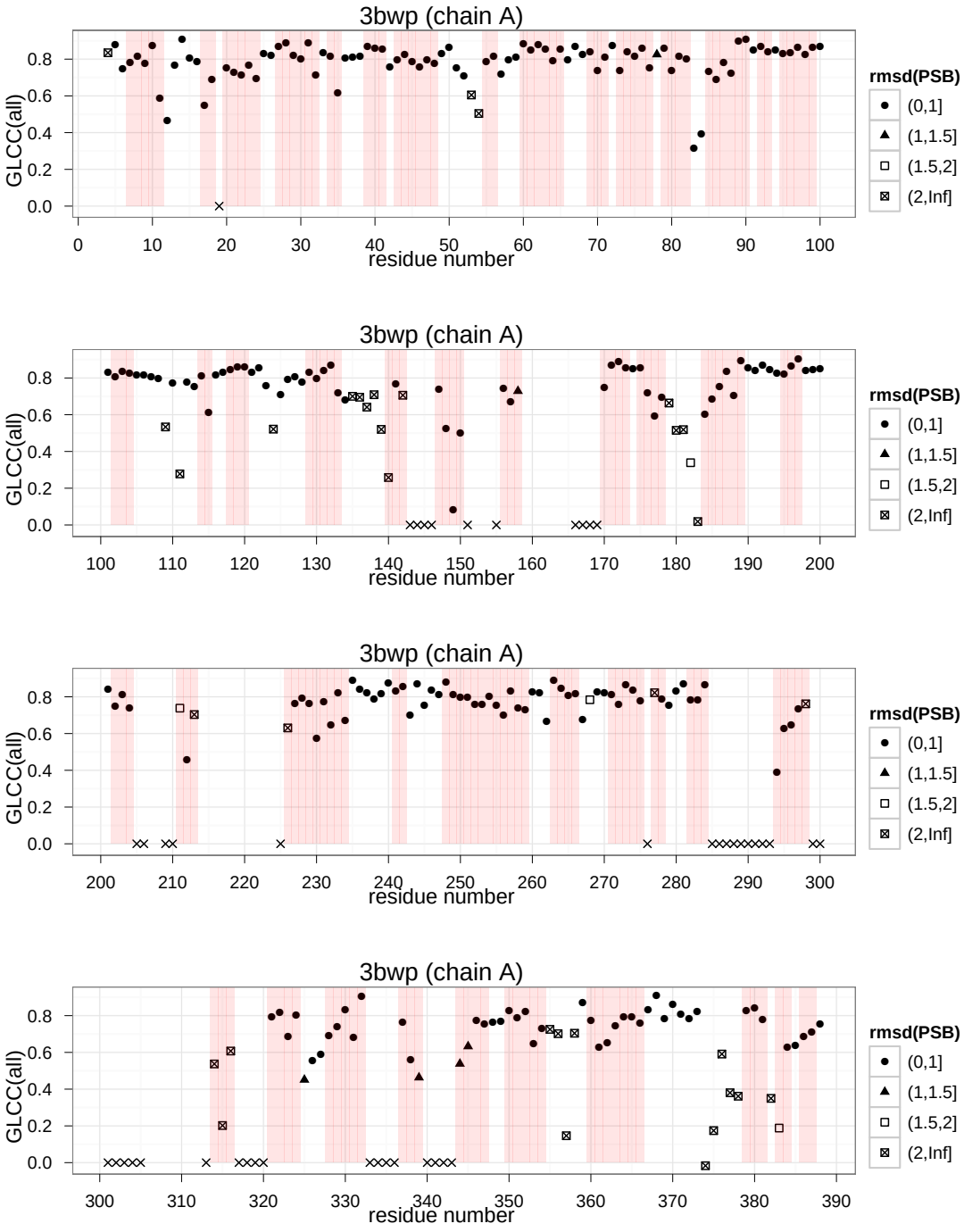
(f)





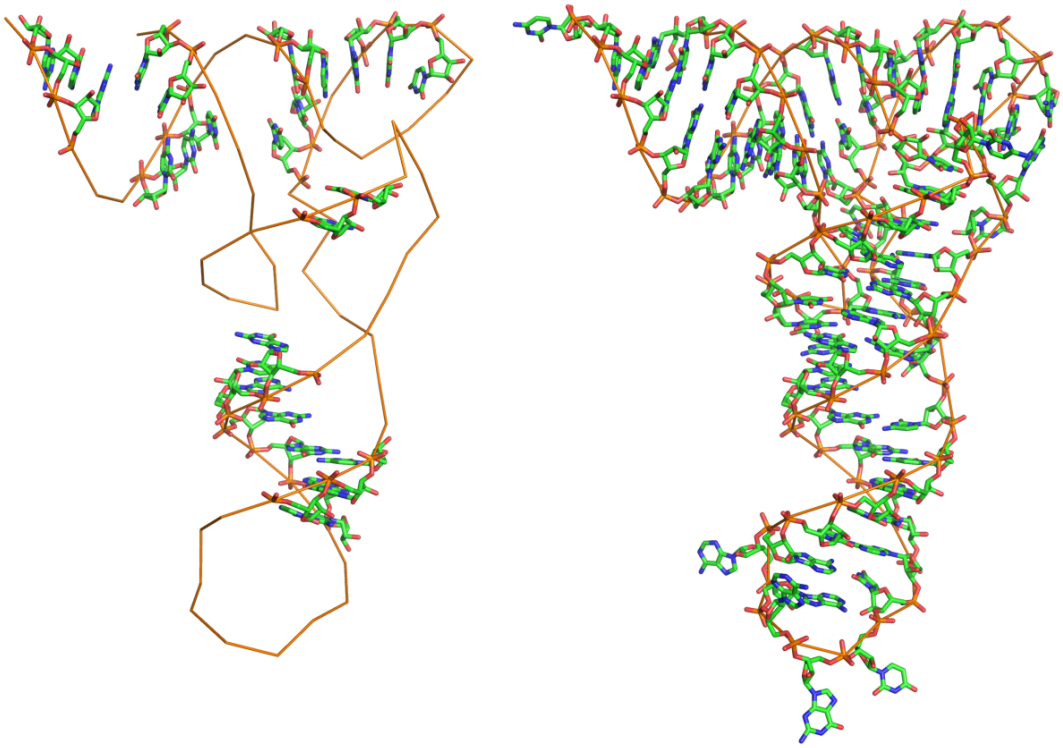
(g)

Figure 5.3 (*cont.*): (f) 3d0u. (g) 3cul.



(h)

Figure 5.3 (*cont.*): (h) 3bwp.



(a) (b)

Figure 5.4: Result for 2zm5 (C chain). The chain trace of the model deposited in PDB is shown as orange lines. Protein atoms are not shown. (a) The starting model was prepared by truncating non-A-form residues. (b) After running *LAFIRE*, C chain was built almost perfectly.

Chapter 6

Conclusion and Outlook

LAFIRE was developed to simulate the refinement process performed by experienced crystallographers (Yao *et al.*, 2006). In the present study, new fitting and extension programs for nucleic acids were developed and incorporated into *LAFIRE*.

NAFIT fits the nucleic acid model into the electron density map with much better geometric quality even at low resolution, as demonstrated for Thg1-tRNA complex. As fitting is essentially minimization by the conjugate gradient method, the convergence radius is limited and therefore the results strongly depend on the initial coordinates. Future development will include a function to generate better initial coordinates before minimization, such as ribose pucker correction in §3.2. Of course, as not all outliers are necessarily wrong, decision making should depend on careful inspection of the electron density map as well as torsion angle preference.

NABUILD extends the nucleic acid model by path extraction from the electron density map. It works reasonably well at medium resolution (2.5 Å) and is tolerable at around 3.0 Å. It depends on the phase accuracy. To build as many residues as possible, *NABUILD* imposes weak preconditions: (i) electron density map of the main chain must be connected at some level (however, gaps of less than 1.0 Å are allowed); (ii) the electron density value of P must be larger than S; (iii) only the geometric relationship among the observed PSB positions is allowed. In fact, the wrongly built or unbuilt residues in the test cases almost always did not fulfill conditions (i) or (ii). An algorithm to overcome these problems is currently under development.

NAFIT and *NABUILD* focus on RNA structure. Although they may also work well

for DNA, some parameters, such as preferred conformation, are not optimal for DNA structures. Parameters specialized for DNA will be included in the next release.

Bibliography

Adams, P. D., Afonine, P. V., Bunkóczi, G., Chen, V. B., Davis, I. W., Echols, N., Headd, J. J., Hung, L.-W., Kapral, G. J., Grosse-Kunstleve, R. W., McCoy, A. J., Moriarty, N. W., Oeffner, R., Read, R. J., Richardson, D. C., Richardson, J. S., Terwilliger, T. C. and Zwart, P. H. (2010). *PHENIX*: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallographica Section D*, **66**(2), 213–221.

Adams, P. D., Pannu, N. S., Read, R. J. and Brünger, A. T. (1997). Cross-validated maximum likelihood enhances crystallographic simulated annealing refinement. *Proceedings of the National Academy of Sciences*, **94**(10), 5018–5023.

Afonine, P. V., Grosse-Kunstleve, R. W. and Adams, P. D. (2005). A robust bulk-solvent correction and anisotropic scaling procedure. *Acta Crystallographica Section D*, **61**(7), 850–855.

Afonine, P. V., Grosse-Kunstleve, R. W., Echols, N., Headd, J. J., Moriarty, N. W., Mustyakimov, M., Terwilliger, T. C., Urzhumtsev, A., Zwart, P. H. and Adams, P. D. (2012). Towards automated crystallographic structure refinement with *phenix.refine*. *Acta Crystallographica Section D*, **68**(4), 352–367.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. and Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research*, **28**(1), 235–242.

Blanc, E., Roversi, P., Vonrhein, C., Flensburg, C., Lea, S. M. and Bricogne, G. (2004). Refinement of severely incomplete structures with maximum likelihood in *BUSTER–TNT*. *Acta Crystallographica Section D*, **60**(12 Part 1), 2210–2221.

Bricogne, G., Blanc, E., Brandl, M., Flensburg, C., Keller, P., Paciorek, W., Roversi, P., Sharff, A., Smart, O., Vonrhein, C. and Womack, T. (2011). *BUSTER version 2.10.0*. Cambridge, United Kingdom: Global Phasing Ltd.

Brünger, A. T. (1992). Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature*, **355**(6359), 472–475.

Brunger, A. T. (2007). Version 1.2 of the Crystallography and NMR system. *Nature Protocols*, **2**(11), 2728–2733.

Brünger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J.-S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. and Warren, G. L. (1998). *Crystallography & NMR System*: A New Software Suite for Macromolecular Structure Determination. *Acta Crystallographica Section D*, **54**(5), 905–921.

Chen, V. B., Arendall, III, W. B., Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., Murray, L. W., Richardson, J. S. and Richardson, D. C. (2010). *MolProbity*: all-atom structure validation for macromolecular crystallography. *Acta Crystallographica Section D*, **66**(1), 12–21.

Chimnaronk, S., Forouhar, F., Sakai, J., Yao, M., Tron, C. M., Atta, M., Fontecave, M., Hunt, J. F. and Tanaka, I. (2009). Snapshots of Dynamics in Synthesizing N^6-Isopentenyladenosine at the tRNA Anticodon. *Biochemistry*, **48**(23), 5057–5065.

Chou, F.-C., Sripakdeevong, P., Dibrov, S. M., Hermann, T. and Das, R. (2013). Correcting pervasive errors in RNA crystallography through enumerative structure prediction. *Nature Methods*, **10**(1), 74–76.

Collaborative Computational Project, Number 4 (1994). The *CCP*4 suite: programs for protein crystallography. *Acta Crystallographica Section D*, **50**(5), 760–763.

Cowtan, K. (2002). The Clipper Project. *Joint CCP4 and ESF-EACBM Newsletter on Protein Crystallography*, **40**.

Cowtan, K. (2012). Nautilus software for automated nucleic acid building. *CCP4 Newsletter*, **48**.

Das, R., Karanicolas, J. and Baker, D. (2010). Atomic accuracy in predicting and designing noncanonical RNA structure. *Nature Methods*, **7**(4), 291–294.

Davis, I. W., Leaver-Fay, A., Chen, V. B., Block, J. N., Kapral, G. J., Wang, X., Murray, L. W., Bryan Arendall, W., I., Snoeyink, J., Richardson, J. S. and Richardson, D. C. (2007). MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Research*, pp. W375–W383.

Emsley, P. and Cowtan, K. (2004). *Coot*: model-building tools for molecular graphics. *Acta Crystallographica Section D*, **60**(12), 2126–2132.

Emsley, P., Lohkamp, B., Scott, W. G. and Cowtan, K. (2010). Features and development of *Coot*. *Acta Crystallographica Section D*, **66**(4), 486–501.

Fisher, R. A. (1936). THE USE OF MULTIPLE MEASUREMENTS IN TAXONOMIC PROBLEMS. *Annals of Eugenics*, **7**(2), 179–188.

Galassi, M., Davies, J., Theiler, J., Gough, B., Jungman, G., Alken, P., Booth, M. and Rossi, F. (2009). *Gnu Scientific Library: Reference Manual*. Network Theory Ltd., 3rd ed.

Garst, A. D., Héroux, A., Rambo, R. P. and Batey, R. T. (2008). Crystal Structure of the Lysine Riboswitch Regulatory mRNA Element. *Journal of Biological Chemistry*, **283**(33), 22347–22351.

Giacovazzo, C. (ed.) (1992). *Fundamentals of Crystallography*. IUCr/Oxford University Press.

Green, D. W., Ingram, V. M. and Perutz, M. F. (1954). The Structure of Haemoglobin. IV. Sign Determination by the Isomorphous Replacement Method. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, **225**(1162), 287–307.

Hattne, J. and Lamzin, V. S. (2008). Pattern-recognition-based detection of planar objects in three-dimensional electron-density maps. *Acta Crystallographica Section D*, **64**(8), 834–842.

Headd, J. J., Echols, N., Afonine, P. V., Grosse-Kunstleve, R. W., Chen, V. B., Moriarty, N. W., Richardson, D. C., Richardson, J. S. and Adams, P. D. (2012). Use of knowledge-based restraints in *phenix.refine* to improve macromolecular refinement at low resolution. *Acta Crystallographica Section D*, **68**(4), 381–390.

Hendrickson, W. (1991). Determination of macromolecular structures from anomalous diffraction of synchrotron radiation. *Science*, **254**(5028), 51–58.

Hendrickson, W. A. and Lattman, E. E. (1970). Representation of phase probability distributions for simplified combination of independent phase information. *Acta Crystallographica Section B*, **26**(2), 136–143.

Hendrickson, W. A. and Teeter, M. M. (1981). Structure of the hydrophobic protein crambin determined directly from the anomalous scattering of sulphur. *Nature*, **290**(5802), 107–113.

Hodel, A., Kim, S.-H. and Brünger, A. T. (1992). Model bias in macromolecular crystal structures. *Acta Crystallographica Section A*, **48**(6), 851–858.

IUPAC-IUB joint commission on biochemical nomenclature (1983). Abbreviations and Symbols for the Description of Conformations of Polynucleotide Chains. *Pure and Applied Chemistry*, **55**(8), 1273–1280.

Jiang, J.-S. and Brünger, A. T. (1994). Protein Hydration Observed by X-ray Diffraction: Solvation Properties of Penicillopepsin and Neuraminidase Crystal Structures. *Journal of Molecular Biology*, **243**(1), 100–115.

Karle, J. (1980). Some developments in anomalous dispersion for the structural investigation of macromolecular systems in biology. *International Journal of Quantum Chemistry*, **18**(S7), 357–367.

Keating, K. S. and Pyle, A. M. (2010). Semiautomated model building for RNA crystallography using a directed rotameric approach. *Proceedings of the National Academy of Sciences*, **107**(18), 8177–8182.

Keating, K. S. and Pyle, A. M. (2012). *RCrane*: semi-automated RNA model building. *Acta Crystallographica Section D*, **68**(8), 985–995.

Kleywegt, G. J. and Brünger, A. T. (1996). Checking your imagination: applications of the free R value. *Structure*, **4**(8), 897–904.

Kleywegt, G. J. and Jones, T. A. (1996). Phi/Psi-chology: Ramachandran revisited. *Structure*, **4**(12), 1395–1400.

Konnert, J. H. (1976). A restrained-parameter structure-factor least-squares refinement procedure for large asymmetric units. *Acta Crystallographica Section A*, **32**(4), 614–617.

Lunin, V. Y., Afonine, P. V. and Urzhumtsev, A. G. (2002). Likelihood-based refinement. I. Irremovable model errors. *Acta Crystallographica Section A*, **58**(3), 270–282.

Lunin, V. Y. and Skovoroda, T. P. (1995). *R*-free likelihood-based estimates of errors for phases calculated from atomic models. *Acta Crystallographica Section A*, **51**(6), 880–887.

Main, P. (1979). A theoretical comparison of the β,γ' and $2F_o$–F_c syntheses. *Acta Crystallographica Section A*, **35**(5), 779–785.

Moews, P. C. and Kretsinger, R. H. (1975). Refinement of the structure of carp muscle calcium-binding parvalbumin by model building and difference fourier analysis. *Journal of Molecular Biology*, **91**(2), 201–225.

Murshudov, G. N., Skubák, P., Lebedev, A. A., Pannu, N. S., Steiner, R. A., Nicholls, R. A., Winn, M. D., Long, F. and Vagin, A. A. (2011). *REFMAC*5 for the refinement of macromolecular crystal structures. *Acta Crystallographica Section D*, **67**(4), 355–367.

Murshudov, G. N., Vagin, A. A. and Dodson, E. J. (1997). Refinement of Macromolecular Structures by the Maximum-Likelihood Method. *Acta Crystallographica Section D*, **53**(3), 240–255.

Nicholls, R. A., Long, F. and Murshudov, G. N. (2012). Low-resolution refinement tools in *REFMAC*5. *Acta Crystallographica Section D*, **68**(4), 404–417.

Pannu, N. S., Murshudov, G. N., Dodson, E. J. and Read, R. J. (1998). Incorporation of Prior Phase Information Strengthens Maximum-Likelihood Structure Refinement. *Acta Crystallographica Section D*, **54**(6 Part 2), 1285–1294.

Pannu, N. S. and Read, R. J. (1996). Improved Structure Refinement Through Maximum Likelihood. *Acta Crystallographica Section A*, **52**(5), 659–668.

Pavelcik, F. (2012). Application of constrained real-space refinement of flexible molecular fragments to automatic model building of RNA structures. *Journal of Applied Crystallography*, **45**(2), 309–315.

Pavelcik, F. and Schneider, B. (2008). Building of RNA and DNA double helices into electron density. *Acta Crystallographica Section D*, **64**(6), 620–626.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, **2**(11), 559–572.

Phillips, S. E. (1980). Structure and refinement of oxymyoglobin at 1.6 Å resolution. *Journal of Molecular Biology*, **142**(4), 531–554.

Pražnikar, J., Afonine, P. V., Gunčar, G., Adams, P. D. and Turk, D. (2009). Averaged kick maps: less noise, more signal...and probably less bias. *Acta Crystallographica Section D*, **65**(9), 921–931.

R Development Core Team, (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Read, R. J. (1986). Improved Fourier coefficients for maps using phases from partial structures with errors. *Acta Crystallographica Section A*, **42**(3), 140–149.

Read, R. J., Adams, P. D., Arendall, W. B., Brunger, A. T., Emsley, P., Joosten, R. P., Kleywegt, G. J., Krissinel, E. B., Lütteke, T., Otwinowski, Z., Perrakis, A., Richardson, J. S., Sheffler, W. H., Smith, J. L., Tickle, I. J., Vriend, G. and Zwart, P. H. (2011). A New Generation of Crystallographic Validation Tools for the Protein Data Bank. *Structure*, **19**(10), 1395–1412.

Rice, L. M., Shamoo, Y. and Brünger, A. T. (1998). Phase Improvement by Multi-Start Simulated Annealing Refinement and Structure-Factor Averaging. *Journal of Applied Crystallography*, **31**(5), 798–805.

Richardson, J. S., Schneider, B., Murray, L. W., Kapral, G. J., Immormino, R. M., Headd, J. J., Richardson, D. C., Ham, D., Hershkovits, E., Williams, L. D., Keating, K. S., Pyle, A. M., Micallef, D., Westbrook, J. and Berman, H. M. (2008). RNA backbone: Consensus all-angle conformers and modular string nomenclature (an RNA Ontology Consortium contribution). *RNA*, **14**(3), 465–481.

Rossmann, M. G. and Blow, D. M. (1962). The detection of sub-units within the crystallographic asymmetric unit. *Acta Crystallographica*, **15**(1), 24–31.

Schmeing, T. M., Huang, K. S., Kitchen, D. E., Strobel, S. A. and Steitz, T. A. (2005). Structural Insights into the Roles of Water and the 2'Hydroxyl of the P Site tRNA in the Peptidyl Transferase Reaction. *Molecular Cell*, **20**(3), 437–448.

Schrödinger, LLC (2010). The PyMOL Molecular Graphics System, Version 1.3r1.

Schwarzenbacher, R., Godzik, A. and Jaroszewski, L. (2008). The JCSG MR pipeline: optimized alignments, multiple models and parallel searches. *Acta Crystallographica Section D*, **64**(1), 133–140.

Scott, W. G. (2012). Challenges and surprises that arise with nucleic acids during model building and refinement. *Acta Crystallographica Section D*, **68**(4), 441–445.

Sekine, S., Nureki, O., Dubois, D. Y., Bernier, S., Chênevert, R., Lapointe, J., Vassylyev, D. G. and Yokoyama, S. (2003). ATP binding by glutamyl-tRNA synthetase is switched to the productive mode by tRNA binding. *EMBO Journal*, **22**(3), 676–688.

Serganov, A., Huang, L. and Patel, D. J. (2008). Structural insights into amino acid binding and gene control by a lysine riboswitch. *Nature*, **455**(7217), 1263–1267.

Shi, H. and Moore, P. B. (2000). The crystal structure of yeast phenylalanine tRNA at 1.93 A resolution: a classic structure revisited. *RNA*, **6**(8), 1091–1105.

Smart, O. S., Womack, T. O., Flensburg, C., Keller, P., Paciorek, W., Sharff, A., Vonrhein, C. and Bricogne, G. (2012). Exploiting structure similarity in refinement: automated NCS and target-structure restraints in *BUSTER*. *Acta Crystallographica Section D*, **68**(4), 368–380.

Sripakdeevong, P., Kladwang, W. and Das, R. (2011). An enumerative stepwise ansatz enables atomic-accuracy RNA loop modeling. *Proceedings of the National Academy of Sciences*, **108**(51), 20573–20578.

Terwilliger, T. C. (2004). Using prime-and-switch phasing to reduce model bias in molecular replacement. *Acta Crystallographica Section D*, **60**(12 Part 1), 2144–2149.

Terwilliger, T. C., Grosse-Kunstleve, R. W., Afonine, P. V., Moriarty, N. W., Adams, P. D., Read, R. J., Zwart, P. H. and Hung, L.-W. (2008*a*). Iterative-build OMIT maps: map improvement by iterative model building and refinement without model bias. *Acta Crystallographica Section D*, **64**(5), 515–524.

Terwilliger, T. C., Grosse-Kunstleve, R. W., Afonine, P. V., Moriarty, N. W., Zwart, P. H., Hung, L.-W., Read, R. J. and Adams, P. D. (2008*b*). Iterative model building, structure refinement and density modification with the *PHENIX AutoBuild* wizard. *Acta Crystallographica Section D*, **64**(1), 61–69.

Toor, N., Keating, K. S., Taylor, S. D. and Pyle, A. M. (2008). Crystal Structure of a Self-Spliced Group II Intron. *Science*, **320**(5872), 77–82.

Trausch, J. J., Ceres, P., Reyes, F. E. and Batey, R. T. (2011). The Structure of a Tetrahydrofolate-Sensing Riboswitch Reveals Two Ligand Binding Sites in a Single Aptamer. *Structure*, **19**(10), 1413–1423.

Wang, B.-C. (1985). *Resolution of phase ambiguity in macromolecular crystallography*, vol. 115, pp. 90–112. Academic Press.
`http://www.sciencedirect.com/science/article/pii/0076687985150093`

Wang, X., Kapral, G., Murray, L., Richardson, D., Richardson, J. and Snoeyink, J. (2008). RNABC: forward kinematics to reduce all-atom steric clashes in RNA backbone. *Journal of Mathematical Biology*, **56**(1), 253–278.

Wilson, A. J. C. (1950). Largest likely values for the reliability index. *Acta Crystallographica*, **3**(5), 397–398.

Word, J., Lovell, S. C., LaBean, T. H., Taylor, H. C., Zalis, M. E., Presley, B. K., Richardson, J. S. and Richardson, D. C. (1999). Visualizing and quantifying molecular goodness-of-fit: small-probe contact dots with explicit hydrogen atoms. *Journal of Molecular Biology*, **285**(4), 1711–1733.

Xiao, H., Murakami, H., Suga, H. and Ferré-D'Amaré, A. R. (2008). Structural basis of specific tRNA aminoacylation by a small in vitro selected ribozyme. *Nature*, **454**(7202), 358–361.

Yao, M., Zhou, Y. and Tanaka, I. (2006). *LAFIRE*: software for automating the refinement process of protein-structure analysis. *Acta Crystallographica Section D*, **62**(2), 189–196.

Zhang, K. Y. J. and Main, P. (1990). Histogram matching as a new density modification technique for phase refinement and extension of protein molecules. *Acta Crystallographica Section A*, **46**(1), 41–46.

Zhou, Y., Yao, M. and Tanaka, I. (2006). New algorithm for protein model building: extending a partial model in a map segment. *Journal of Applied Crystallography*, **39**(1), 57–63.