Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

New model-fitting and model-completion programs for automated iterative nucleic acid refinement

Keitaro Yamashita,[a] Yong Zhou,[a,b] Isao Tanaka[a,b] and Min Yao[a,b]*

[a]Graduate School of Life Science, Hokkaido University, Sapporo, Hokkaido 060-0810, Japan, and [b]Faculty of Advanced Life Science, Hokkaido University, Sapporo, Hokkaido 060-0810, Japan

Correspondence e-mail:
yao@castor.sci.hokudai.ac.jp

Received 26 February 2013
Accepted 15 March 2013

In the past decade many structures of nucleic acids have been determined, which have contributed to our understanding of their biological functions. However, crystals containing nucleic acids often diffract X-rays poorly. 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 in the automated refinement system *LAFIRE*, as well as acting as independent programs. *NAFIT* performs sequential grouped fitting with empirical torsion-angle restraints and antibumping restraints including H atoms. *NABUILD* extends the model using a skeletonized map in a coarse-grained manner. It has been shown that *NAFIT* greatly improves electron-density fit and geometric quality and that iterative refinement with *NABUILD* significantly reduces the $R_{\rm free}$ factor.

1. Introduction

The number of structures of nucleic acids, including those complexed with proteins, 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 crystals containing nucleic acids do not usually diffact to high resolution, several authors have made a great deal of effort to tackle the difficulties involved. Initial model-building programs such as *phenix.autobuild* (Terwilliger *et al.*, 2008; Adams *et al.*, 2010), *ARP*/*wARP* (Hattne & Lamzin, 2008), *NUT*/*DHL*/*RSR* (Pavelcik & Schneider, 2008; Pavelcik, 2012) and *Nautilus* (Cowtan, 2012) utilize features that can be observed in lower resolution electron-density maps. Semi-automated model building can be performed with *RCrane* (Keating & 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 contains many conformational errors and unconstructed regions. Thus, repeated refinements and model rebuilding are required in subsequent steps, which makes structure analysis time-consuming.

doi:10.1107/S0907444913007191

Here, we present the new refinement tools for nucleic acids *NAFIT* and *NABUILD*. Their 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 a real-space refinement (fitting) program with empirical torsion-angle restraints and antibumping restraints including H atoms to maintain geometric quality. *NABUILD* is a chain-extension program that uses graph interpretations 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 the *LAFIRE* refinement strategy significantly reduced the $R_{\rm free}$ factor for test cases with resolutions in the range 2.1–3.12 Å.

2. Model fitting by *NAFIT*

Fitting is the process of fine-tuning 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 the density-weighted grouped local correlation coefficient (GLCC),

$$\mathrm{GLCC}_i = g_i \frac{\langle \rho_{\rm obs} Z \rangle_i}{[\langle \rho_{\rm obs}^2 \rangle_i \langle Z^2 \rangle_i]^{1/2}}, \tag{1}$$

$$g_i = \frac{1}{N} \int\limits_{\min(\rho_{\rm obs})}^{\langle \rho_{\rm obs} \rangle_i} h(\rho_{\rm obs})\, {\rm d}\rho_{\rm obs}. \tag{2}$$

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

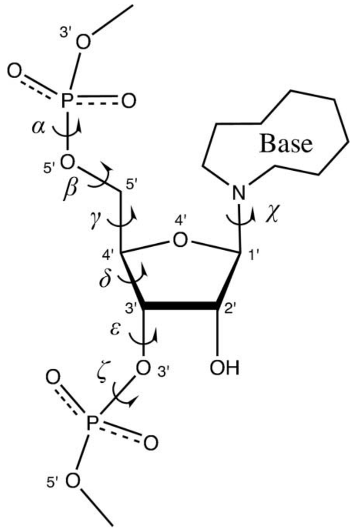


Figure 1
Definitions of nucleotide torsion angles.

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

$$E = E_{\rm geometry} + wE_{\rm map}. \tag{3}$$

Here, w is the fitting weight and

$$\begin{aligned} E_{\rm geometry} = {} & E_{\rm bond} + E_{\rm angle} + E_{\rm chiral} + E_{\rm plane} \\ & + E_{\rm nonbonded} + E_{\rm naconf1d}, \end{aligned} \tag{4}$$

$$E_{\rm map} = -\sum_i^{N_{\rm atoms}} Z_i \rho(x_i, y_i, z_i). \tag{5}$$

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 the Clipper library (Cowtan, 2002). The formulations of $E_{\rm bond}$, $E_{\rm angle}$, $E_{\rm chiral}$, $E_{\rm plane}$, $E_{\rm nonbonded}$ and their derivatives are the same as in *Coot* (SVN 4120). $E_{\rm naconf1d}$ is defined in §2.1. Minimization was performed using the conjugate-gradient method as implemented in *GSL* (Galassi *et al.*, 2009).

2.1. Conformational restraints

In *NAFIT*, $E_{\rm naconf1d}$ 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 (Fig. 1). In addition, as described above, nucleic acid structures are 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 unfavourable conformations. $E_{\rm naconf1d}$ 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).

$$E_{\rm naconf1d} = -\sum_\theta \sum_i^{N_\theta} \log p_\theta(\theta_i). \tag{6}$$

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).

To construct the empirical distribution function p_θ, we downloaded 'RNA09' coordinate files from the Richardsons' website (http://kinemage.biochem.duke.edu/databases/rnadb.php), which are nonredundant RNA structures solved at 3.0 Å resolution or better (Richardson *et al.*, 2008). The frequencies of torsion-angle values for each entry were calculated using the statistical package *R* (R Development Core Team, 2008). Frequencies below a threshold in each bin, which may be outliers, were replaced by zero. Such frequencies were then 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. 2.

2.2. Ribose-pucker correction for RNA

For RNA, ribose pucker is usually limited to either C3′-*endo* or C2′-*endo* (Richardson *et al.*, 2008). However, it is difficult to determine the correct ribose pucker at low or medium resolution, and ribose pucker has thus shown to be error-prone by the Protein Data Bank X-ray Validation Task Force (Read *et al.*, 2011). The correct ribose pucker of RNA can be deduced based on the base–phosphate perpendicular distance (Davis *et al.*, 2007; Keating & Pyle, 2010). Ribose pucker is corrected in *NAFIT*. As the radius of convergence by minimization is limited and $E_{\rm naconf1d}$ gives local minima for both C3′-*endo* and C2′-*endo*, it is necessary to construct good starting coordinates before minimization. Residues with incompatible base–phosphate perpendicular distances and δ angles are detected and are subjected to simple rebuilding as follows. The ribose atoms are replaced by those of the deduced puckering and then rotated around the C5′–C1′ axis to give an acceptable ε angle (∼208°). Finally, minimization is performed for fine-tuning of atomic coordinates. This procedure is very simple and gives reasonable coordinates with correct puckering.

2.3. Nonbonded atom interactions and H atoms

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

To maintain geometric quality, H atoms are automatically generated before the fitting procedure. They contribute to the geometric term, but Z for each H atom is set to 0, so that they do not contribute to $E_{\rm map}$ in (2). Fitting including H atoms greatly reduces interatomic clashes without deteriorating the fit to the electron-density map. For hydrogen-bonding atoms the critical distance is reduced by 0.3 Å. A fixed value of 1.8 Å is used for H-atom pairs. These constants were obtained empirically.

2.4. 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 of the current residue are selected for grouped fitting (similar to the so-called ‘sphere refinement’ in *Coot*).

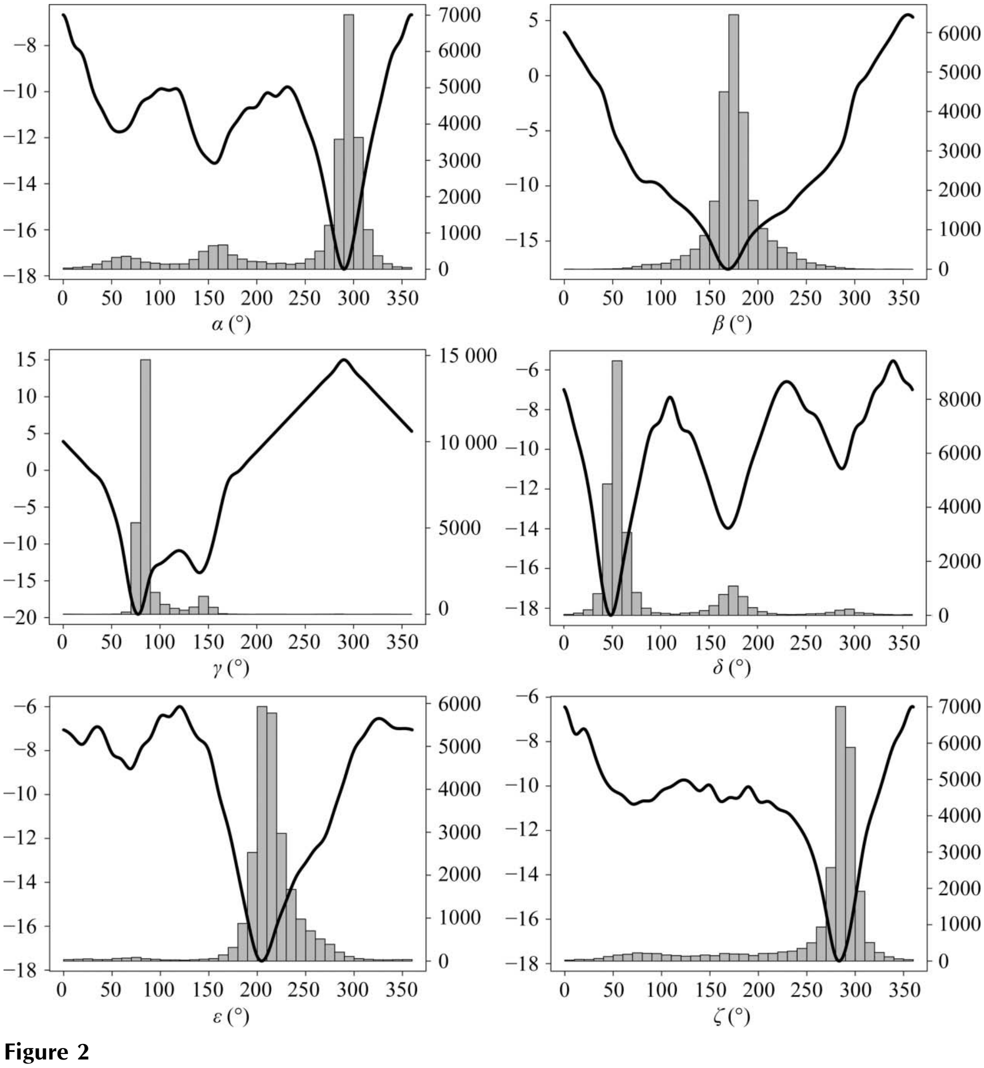


Figure 2
One-dimensional conformational restraint functions (left axis, heavy lines) and frequency (right axis, histograms) for RNA. The restraint for the χ angle is not shown.

(ii) Atoms that are not selected are fixed and take part in nonbonded interaction (including other chains).

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

(iv) H atoms are generated.

(v) The target function E is minimized.

Using this sequential grouped fitting, minimization is expected to stably converge owing to the limited number of variables, and neighbouring residues can be fitted into the electron-density map with resolution of steric clashes.

3. Chain extension by *NABUILD*

For model completion by *NABUILD*, two building modes are available: (i) gap building, which connects residues already built, and (ii) terminal building, which starts building from the termini. *NABUILD* finds the main-chain path extracted from the skeletonized electron-density map. Map skeletonization is performed after extracting 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 blobs, and in such cases models cannot be built owing 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. 3). 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 the points P, S and B as representative points of the 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 P-atom coordinate, the centre of the sugar-ring atoms (C1′, C2′, C3′, C4′ and O4′) and the centre of the base atoms, respectively. The $\mathbf{B}$ direction is also defined as the vector normal to the base plane. In building, the points PSB and the direction $\mathbf{B}$ are determined and atomic coordinates are then calculated.

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

(i) The missing residue ranges are determined by comparing the PDB file and the sequence file.

(ii) A $2mF_o - DF_c$ map is calculated with a grid spacing of about 0.5 Å.

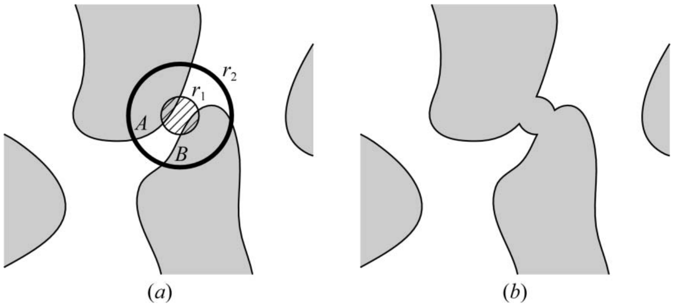


Figure 3
Gap filling before skeletonization. The grey regions represent electron-density blobs. (*a*) For each grid point, blobs are clustered within r_2 from the point (A and B). (*b*) If the points in the sphere with radius r_1 belong to more than one cluster, a pseudo-blob is inserted.

(iii) Based on the initial model, positions near the existing B positions are recorded along with $\mathbf{B}$ directions as possible B positions and $\mathbf{B}$ directions because they may form stacking interactions or base pairs.

(iv) The density map is skeletonized using the 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. 3).

(v) Skeleton points around residues that have already been built are removed, with the exception of sugar atoms at the 3′ terminus and the 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 ten candidates for each range (§3.1).

(vii) Candidates are merged into chains (§3.2).

3.1. Building on the 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 of the current terminal phosphate group is constructed, where N is the number of residues to be built. For gap building, a graph covering the skeleton points within min(1.70N + 13.3, 50) Å of the midpoint of the terminal phosphate groups is constructed. 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 $\langle$GLCC$\rangle$ is calculated.

3.1.1. 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*a*). For terminal building, the end point is not defined. The paths are found by a breadth-first search, keeping up to 1000 intermediate paths. If the number of intermediate paths exceeds 1000, those with lower scores are removed.

The path score is defined based on the P positions as $\langle Z_P\rangle - \langle p\rangle$, where p is the planarity score of the blob belonging to P and Z_P is the electron-density Z score at position P calculated from the mean and standard deviation of the density values at path points near position P. The 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 an 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 λ_1, λ_2 and λ_3 are eigenvalues given by PCA ($\lambda_1 \leq \lambda_2 \leq \lambda_3$). The discriminant is learned from real data and is expected to be positive for a planar blob ($\lambda_1/\lambda_3 << 1$, $\lambda_1/\lambda_2 << 1$) and negative for a spherical blob ($\lambda_1 \simeq \lambda_2 \simeq \lambda_3$). The currently used discriminant is $-18.82\lambda_1/\lambda_3 + 1.2694\lambda_1/\lambda_2 + 6.3484$ and the map level is dynamically adjusted so that the volume of the blob is close to the expected value ($\sim$14 Å^3).

3.1.2. Finding *PSB*. The first P position is defined using the terminal residue. To build from the 3′ terminus, the path point 2.0–4.3 Å away from S of the 3′ terminus with the maximum electron density is defined as P. To build from the 5′ terminus, the P position is redefined by the path point within 2.0 Å 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 following conditions.

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

(ii) The linear distance P_i–P_{i+1} is between 4.5 and 7 Å.

(iii) The angle $\angle(P_{i-1}$–P_i–$P_{i+1})$ is larger than 37.8°.

These constants were defined by analyzing the RNA09 data.

The 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*) of 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*b*).

3.1.3. *PSB* to atomic coordinates. Based on the *PSB* positions and the **B** direction, atomic coordinates are calculated sequentially (Fig. 4*c*). Firstly, the phosphate group is placed at position P, with O5′ directed toward S. Next, the centre of the base atoms is placed at B, the base plane is oriented along the **B** direction and the N atom forming a glycosidic bond (N9 for purine and N1 for pyrimidine) is directed toward S. The centre of the sugar atoms is then placed at S and oriented by aligning the normal vector of the sugar plane to $\vec{SB} + \vec{SP}$, and C5′ is directed towards P. Translation and rotation of sugar atoms are optimized by the downhill simplex method to satisfy geometric conditions.

The placed atoms are fitted to the electron density together with the adjacent residue (Fig. 4*d*). In this fitting, the *PSB* positions are used as target positions for 'mean position restraint'. The mean position restraint is designed for restraining a group of atoms $\mathbf{x}_1, \ldots, \mathbf{x}_N$ to a given position $\mathbf{a}$ (P, S or B). The equation added to the target function of fitting is defined as

$$E_{\rm mpos} = \frac{1}{\sigma^2}\left(\frac{\sum_i^N \mathbf{x}_i}{N} - \mathbf{a}\right)^2. \qquad (7)$$

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

3.2. Merging candidates

When candidates are collected for each missing residue range, a single candidate should be selected in each range and integrated into a single chain without severe steric clashes. Firstly, 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

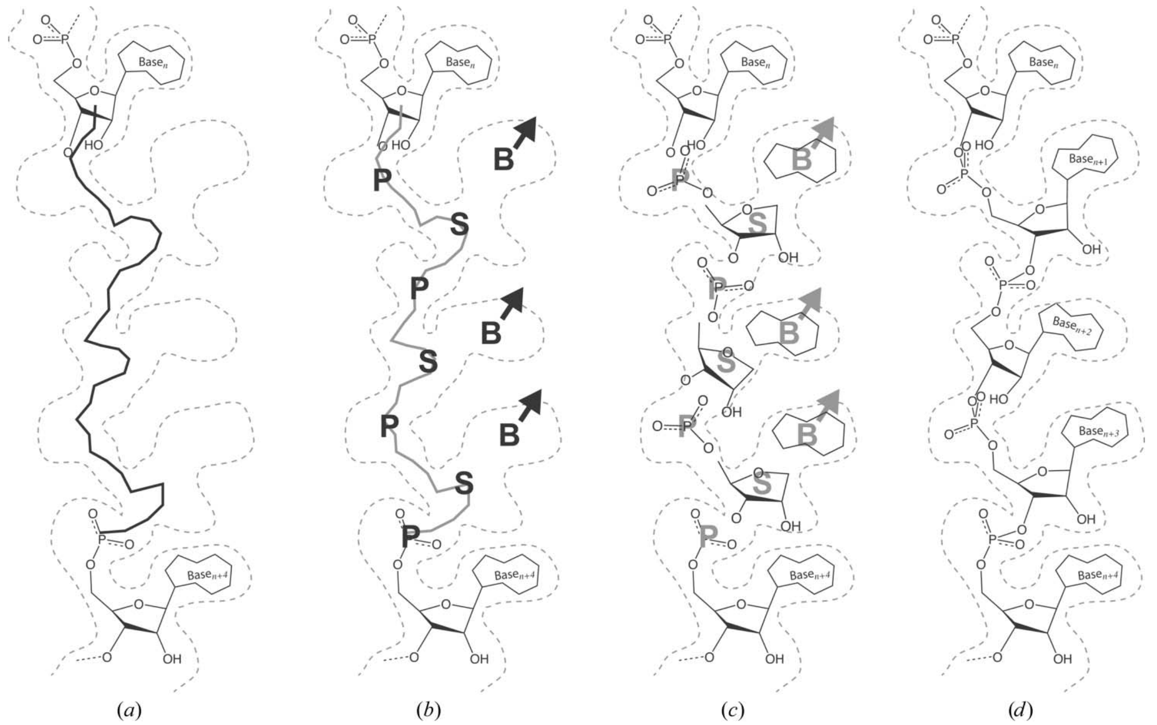


Figure 4
NABUILD procedure. (*a*) The path from the 3′ sugar to the 5′ phosphate is extracted. (*b*) The positions P, S and B and the **B** direction are determined. (*c*) Atoms of each part are roughly positioned and oriented. (*d*) Atomic coordinates are refined by fitting with mean position restraints.

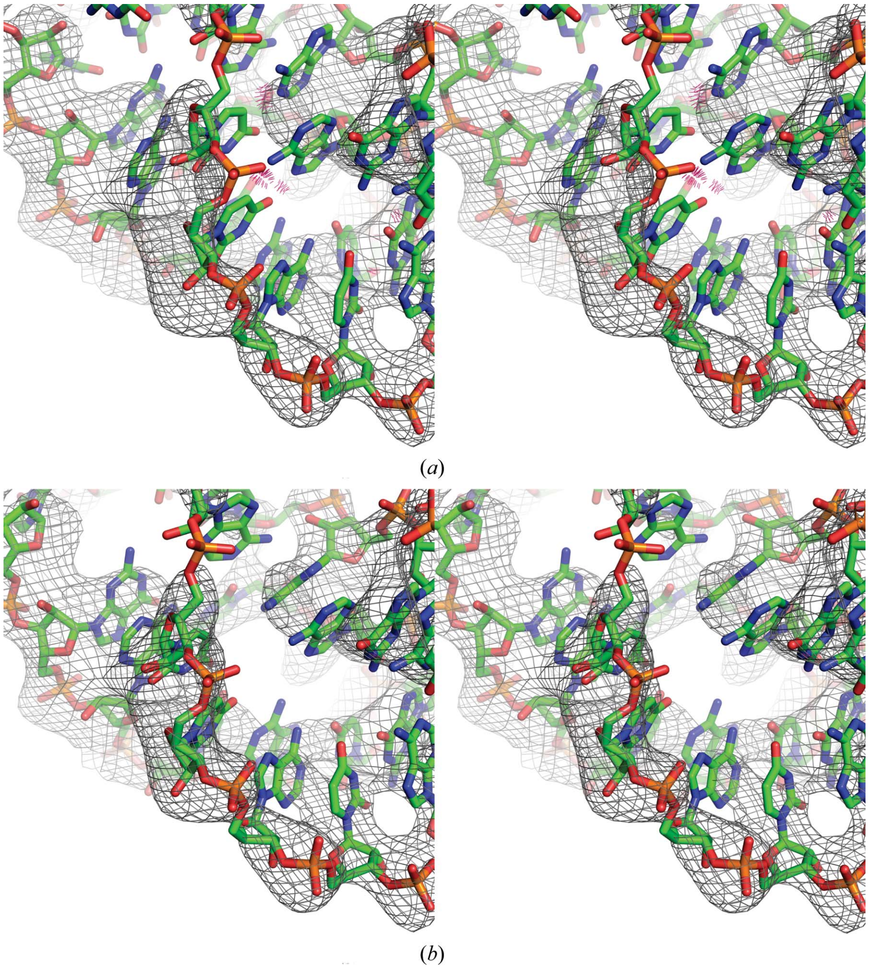
exists within 2 Å, the candidates are marked as a clashing pair. Candidates that are not involved in clashing are accepted. For clashing candidate pairs, a nonclashing candidate pair is searched for among the candidates. All nonclashing candidates are integrated into a single chain. If clashing of base atoms is found, fitting is performed for these residues and for those within 3.5 Å. In this fitting, P atoms are fixed.

Table 1
Fitting of tRNA to the map of the Thg1–tRNA complex at 4.2 Å resolution.

This table shows that the fitting strategy with antibumping restraints including H atoms and naconf1d restraints greatly improved the geometric quality and density fit at very low resolution. *phenix.refine* was used to refine and evaluate the initial model and the improved model (*NAFIT**).

		⟨GLCC⟩†		Outliers‡			
	$R_{\rm work}$, $R_{\rm free}$	Chain 1, chain 2	Clashscore§	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

† Averaged GLCC is calculated for each chain. GLCC is still poor for chain 2 even in the final model as the electron density around the molecule is poor. ‡ RNA model validation is given by *phenix.rna_validate*. Bond/angle outliers are counted if the deviation from the ideal value is larger than 4σ. Pucker outliers are ε outliers or incompatible δ and base–phosphate perpendicular distance (Davis *et al.*, 2007; Keating & Pyle, 2010). Suite outliers are given by *suitename* (Richardson *et al.*, 2008). § The *MolProbity* clashscore is defined as the number of bad overlaps per 1000 atoms (Word *et al.*, 1999) and is calculated by *phenix.clashscore* after removing protein atoms.


(*a*)

(*b*)

Figure 5
Stereo drawing of the *NAFIT* result for the Thg1–tRNA complex at 4.2 Å resolution. Target $2mF_{\rm o} - DF_{\rm c}$ electron density is shown only around the displayed molecule at the 1.0 r.m.s.d. level. Bad overlaps (≥0.4 Å) given by the *probe* command in *MolProbity* (Word *et al.*, 1999) are displayed as red spikes. (*a*) Before fitting. (*b*) After fitting; steric clashes are resolved and the model shows a better fit to the electron-density map.

4. *LAFIRE* updates and incorporation of *NAFIT* and *NABUILD*

LAFIRE consists of three parts: partial model building, model modification (fitting) including evaluation of the current model and a process-control system that includes interfaces with refinement programs. For refinement, users can choose which program is used from *CNS* v.1.2 or v.1.3 (Brünger *et al.*, 1998; Brunger, 2007), *REFMAC*5 (Murshudov *et al.*, 2011), *phenix.refine* (Afonine *et al.*, 2012) and *autoBUSTER* (Bricogne *et al.*, 2011). The refinement target can be either maximum likelihood using amplitudes (MLF) or phased maximum likelihood using amplitudes and Hendrickson–Lattman coefficients (MLHL). Initially, rigid-body refinement is performed, in which the rigid-body domain is defined for each chain. Next, following full atomic restrained refinement, fitting and building are performed based on the $2mF_{\rm o} - DF_{\rm c}$ map given by the refinement program. The *LAFIRE* processes are controlled by monitoring the $R_{\rm free}$ factor and are repeated until there is no further improvement. Finally, refinement with updating of water sites is performed.

NAFIT and *NABUILD* are incorporated into the model-adjustment procedure after running the refinement program. Before *NABUILD*, residues with a GLCC lower than 0.8 built by *NABUILD* in previous cycles are removed. This criterion was determined empirically. During *NAFIT* and *NABUILD*, protein atoms are fixed to avoid steric clashes with nucleic acids.

LAFIRE has a graphical user interface (GUI) implemented in PyQt4. The GUI can be used for the preparation of input files and parameters and for the selection of the molecule type to be fitted or extended. The $R_{\rm free}$ and $R_{\rm work}$

Table 2
Example data for the test of *NABUILD* and iterative refinement.

1n78 is tRNA^{Glu} complexed with GluRS (Sekine *et al.*, 2003). 3u56 is an 80 nt 23S RNA complexed with mutant TthL1 (S. V. Tishchenko, E. Y. Nikonova, O. S. Kostareva, A. G. Gabdulkhakov, A. V. Sarskikh, W. Piendl, S. V. Nikonov, M. B. Garber & N. A. Nevskaya, unpublished work). 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 using a combination of two derivatives (Toor *et al.*, 2008).

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	*C*2	183	92, 92
3bwp	3.12	$P2_12_12_1$	0	356

factors and the 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; http://www.pymol.org) and *Coot* (Emsley *et al.*, 2010) are ready to begin visual inspection of the result.

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 the 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} (PDB entry 1ehz; Shi & Moore, 2000) 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 poor fit to the electron-density map, they were roughly adjusted using *Coot*. We then used *NAFIT* to improve the manually adjusted model. The resultant model showed a better fit to the electron-density map and better geometric quality (Fig. 5). A fixed value of 2 was used for the fitting weight *w*, which was determined in a number of trials. We performed individual-site and grouped ADP (one group per chain) refinement using *phenix.refine* (v.dev-1218) with automatic secondary-structure restraints and torsion-angle NCS restraints (Headd *et al.*, 2012) to show how *NAFIT* enhanced the model quality. As a result, the model refined with *NAFIT* showed better geometric quality and explained the diffraction data well compared with the model refined without *NAFIT* (Table 1).

5.2. Building by *NABUILD* and iterative refinement

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

5.2.1. Sample model preparation. We demonstrated the

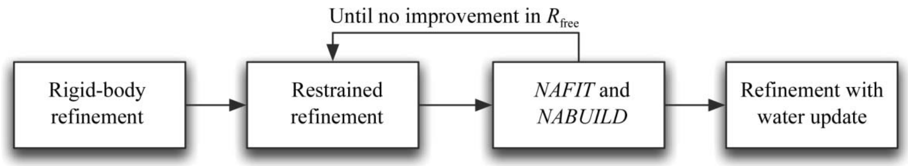


Figure 6
Simplified representation of the refinement protocol. *NAFIT* and *NABUILD* always use the latest electron-density map given by the refinement program.

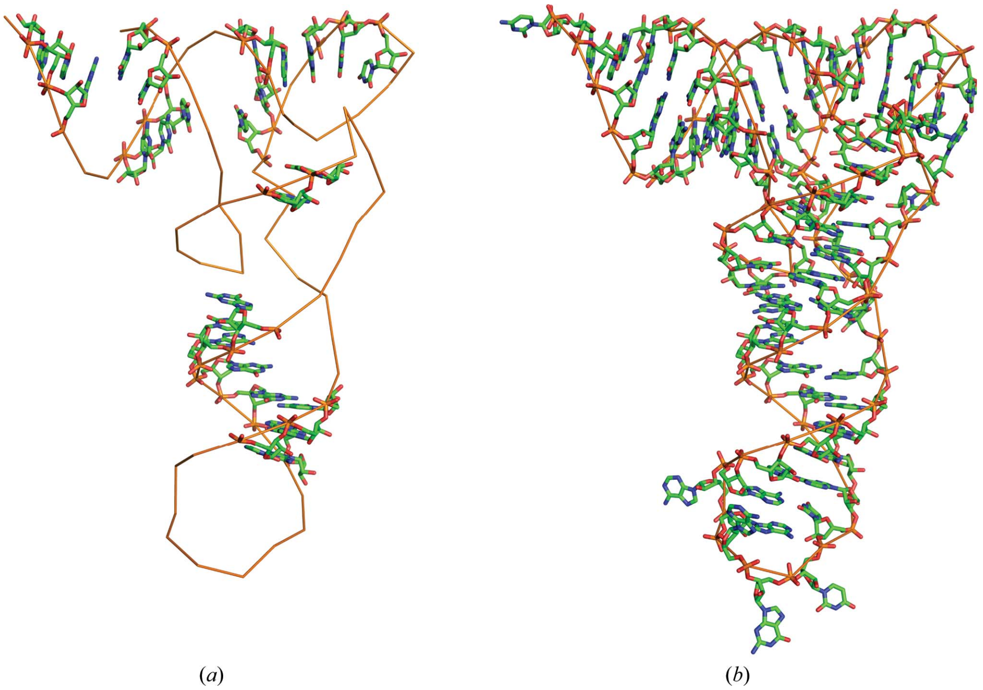


Figure 7
Result for PDB entry 2zm5 (*C* chain). The chain trace of the model deposited in the PDB is shown as orange lines. Protein atoms are not shown. (*a*) The starting model was prepared by truncating non-A-form residues. (*b*) After running *LAFIRE*, the *C* chain was built almost perfectly.

Table 3
Results of *NABUILD* and iterative refinement.

If the r.m.s.d. values of the *P*, *S* and *B* positions between the built model and the PDB model are less than 1.5 Å in each residue, the residue is counted as correctly built.

	No. of residues							
PDB code (chain)	Needing to be built	Correctly built†	Incorrectly built†	Unbuilt	$R_{\rm free}$, initial	$R_{\rm free}$, final	$R_{\rm free}$ 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/incorrectly built main-chain residues. ‡ Evaluated by *phenix.maps* with bulk-solvent correction and anisotropic scaling using the model deposited in the PDB. Note that the final model given by *LAFIRE* does not include ligands and ions.

performance of *NABUILD* with eight data sets (Table 2). Experimental phase information is available for PDB entries 2zm5, 3d0u, 3bwp and 3sd1 in the form of Hendrickson–Lattman coefficients. For this test, we truncated all residues except those with the canonical A-form RNA conformation from the model deposited in the PDB because such residues could be modelled relatively easily, for example by searching for double-stranded helices. A-form RNA residues were identified as class 1a by the *suitename* program (v.0.3.070628; Richardson *et al.*, 2008). All non-class 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 the PDB. Other components, including ligands, ions and water, were removed.

5.2.2. Protocol. The tests were performed using the *LAFIRE* GUI. The pruned model as a starting model and the mtz file were provided. The refinement program was set to *phenix.refine* (v.dev-1218) with the MLF or MLHL (if available; the exception was 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, while other polymer chains were fixed if present. The refinement scheme was based on the *LAFIRE* refinement strategy, which is shown in brief in Fig. 6.

5.2.3. Results. The results of iterative refinement by *LAFIRE* with *NAFIT* and *NABUILD* are summarized in Table 3. In all cases, significantly lower $R_{\rm free}$ values were obtained than those of 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. 7), while the *D* chain was incomplete owing to very poor electron density. Although manual intervention was still needed, it was greatly reduced.

6. Discussion

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 have been 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 the Thg1–tRNA complex. As fitting is essentially minimization using the conjugate-gradient method, the convergence radius is limited and therefore the results depend strongly on the initial coordinates. Future development will include a function to generate better initial coordinates before minimization, such as ribose-pucker correction (§2.2). 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 tolerably at around 3.0 Å. It depends on the phase accuracy. To build as many residues as possible, *NABUILD* imposes weak preconditions: (i) the 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 that of *S* and (iii) only the geometric relationships among observed *PSB* positions are allowed. In fact, residues that were incorrectly built or unbuilt in the test cases almost always did not fulfill conditions (i) or (ii). An algorithm to overcome these problems is currently in development.

NAFIT and *NABUILD* focus on RNA structure. Although they may also work well for DNA, some parameters, such as the preferred conformation, are not optimal for DNA structures. Parameters specialized for DNA will be included in the next release.

7. Availability

The features described in this paper are available in *LAFIRE* 4.0, which can be downloaded from our website at http://altair.sci.hokudai.ac.jp/g6/Research/Lafire_English.html. Precompiled versions for Linux (CentOS 6) and Mac OS X are available. The source code can be obtained upon request. *LAFIRE* requires *CCP*4 programs. To use *CNS*, *phenix.refine* or *autoBUSTER* for refinement, these programs should be installed. For help, contact lafire@castor.sci.hokudai.ac.jp.

We thank Dr Akiyoshi Nakamura for providing the Thg1-tRNA complex data and the initial model for the test of *NAFIT*. This work was supported by the Targeted Proteins Research Program (TPRP) from the Ministry of Education, Culture, Sports, Science and Technology, Japan.

References

Adams, P. D. *et al.* (2010). *Acta Cryst.* D**66**, 213–221.

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. & Adams, P. D. (2012). *Acta Cryst.* D**68**, 352–367.

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

Brunger, A. T. (2007). *Nature Protoc.* **2**, 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. & Warren, G. L. (1998). *Acta Cryst.* D**54**, 905–921.

Chen, V. B., Arendall, W. B., Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., Murray, L. W., Richardson, J. S. & Richardson, D. C. (2010). *Acta Cryst.* D**66**, 12–21.

Chimnaronk, S., Forouhar, F., Sakai, J., Yao, M., Tron, C. M., Atta, M., Fontecave, M., Hunt, J. F. & Tanaka, I. (2009). *Biochemistry*, **48**, 5057–5065.

Chou, F. C., Sripakdeevong, P., Dibrov, S. M., Hermann, T. & Das, R. (2013). *Nature Methods*, **10**, 74–76.

Cowtan, K. (2002). *CCP4 Newsl. Protein Crystallogr.* **40**, contribution 5.

Cowtan, K. (2012). *CCP4 Newsl. Protein Crystallogr.* **48**, contribution 6.

Das, R., Karanicolas, J. & Baker, D. (2010). *Nature Methods*, **7**, 291–294.

Davis, I. W., Leaver-Fay, A., Chen, V. B., Block, J. N., Kapral, G. J., Wang, X., Murray, L. W., Arendall, W. B., Snoeyink, J., Richardson, J. S. & Richardson, D. C. (2007). *Nucleic Acids Res.* **35**, W375–W383.

Emsley, P. & Cowtan, K. (2004). *Acta Cryst.* D**60**, 2126–2132.

Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. (2010). *Acta Cryst.* D**66**, 486–501.

Fisher, R. A. (1936). *Ann. Eugen.* **7**, 179–188.

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

Garst, A. D., Héroux, A., Rambo, R. P. & Batey, R. T. (2008). *J. Biol. Chem.* **283**, 22347–22351.

Hattne, J. & Lamzin, V. S. (2008). *Acta Cryst.* D**64**, 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. & Adams, P. D. (2012). *Acta Cryst.* D**68**, 381–390.

Keating, K. S. & Pyle, A. M. (2010). *Proc. Natl Acad. Sci. USA*, **107**, 8177–8182.

Keating, K. S. & Pyle, A. M. (2012). *Acta Cryst.* D**68**, 985–995.

Murshudov, G. N., Skubák, P., Lebedev, A. A., Pannu, N. S., Steiner, R. A., Nicholls, R. A., Winn, M. D., Long, F. & Vagin, A. A. (2011). *Acta Cryst.* D**67**, 355–367.

Pavelcik, F. (2012). *J. Appl. Cryst.* **45**, 309–315.

Pavelcik, F. & Schneider, B. (2008). *Acta Cryst.* D**64**, 620–626.

Pearson, K. (1901). *Philos. Mag.* **2**, 559–572.

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

Read, R. J. *et al.* (2011). *Structure*, **19**, 1395–1412.

Richardson, J. S. *et al.* (2008). *RNA*, **14**, 465–481.

Scott, W. G. (2012). *Acta Cryst.* D**68**, 441–445.

Sekine, S., Nureki, O., Dubois, D. Y., Bernier, S., Chênevert, R., Lapointe, J., Vassylyev, D. G. & Yokoyama, S. (2003). *EMBO J.* **22**, 676–688.

Serganov, A., Huang, L. & Patel, D. J. (2008). *Nature (London)*, **455**, 1263–1267.

Shi, H. & Moore, P. B. (2000). *RNA*, **6**, 1091–1105.

Sripakdeevong, P., Kladwang, W. & Das, R. (2011). *Proc. Natl Acad. Sci. USA*, **108**, 20573–20578.

Terwilliger, T. C., Grosse-Kunstleve, R. W., Afonine, P. V., Moriarty, N. W., Zwart, P. H., Hung, L.-W., Read, R. J. & Adams, P. D. (2008). *Acta Cryst.* D**64**, 61–69.

Toor, N., Keating, K. S., Taylor, S. D. & Pyle, A. M. (2008). *Science*, **320**, 77–82.

Trausch, J. J., Ceres, P., Reyes, F. E. & Batey, R. T. (2011). *Structure*, **19**, 1413–1423.

Wang, X., Kapral, G., Murray, L., Richardson, D., Richardson, J. & Snoeyink, J. (2008). *J. Math. Biol.* **56**, 253–278.

Word, J. M., Lovell, S. C., LaBean, T. H., Taylor, H. C., Zalis, M. E., Presley, B. K., Richardson, J. S. & Richardson, D. C. (1999). *J. Mol. Biol.* **285**, 1711–1733.

Xiao, H., Murakami, H., Suga, H. & Ferré-D'Amaré, A. R. (2008). *Nature (London)*, **454**, 358–361.

Yao, M., Zhou, Y. & Tanaka, I. (2006). *Acta Cryst.* D**62**, 189–196.

Zhou, Y., Yao, M. & Tanaka, I. (2006). *J. Appl. Cryst.* **39**, 57–63.