Assessing The Quality Structural Models

Tianyi Shi

2020-10-27

Contents

A high-quality protein structure model explains the experimental observations well, and conforms to physical and biochemical principles. There are some parameters indicating the quality of a structure model. Some of them are used during model building and refinement, while others apply only on a refined structure.

1 The Essence of X-Ray Crystallography

The intensity at (h, k, l) in the reciprocal space, i.e. the structure factor amplitudes collected in X-ray experiments is directly related to the Fourier transform of the electron density in the real space ($I = \mathbf{F}_{hkl}^2$), the latter being calculated as:

$$\mathbf{F}_{hkl} = \int_x \int_y \int_z \rho(x, y, z) e^{2\pi i(hx+ky+lz)} \mathrm{d}x\mathrm{d}y\mathrm{d}z \tag{1}$$

In practice, since the number of atoms and reflections (hence (x, y, z) or (h, k, l) coordinates is limited, its discrete form, which can be efficiently computed, is used:[1]

$$\mathbf{F}_{hkl} = \sum_j^n f_j e^{2\pi i(hx_j+ky_j+lz_j)} \tag{2}$$

where (x_j, y_j, z_j) is the coordinates of the j-th atom in real space.

This is also known as the *structure factor equation.*

Solving the electron density from structure factors involves solving the inverse Fourier transform of the above equation, also known as the *electron density equation*, in its discrete form:

[1] Most crystallography softwares seem to use Fast Fourier Trasnform (FFT) algorithms provided by FFTW (http://www.fftw.org/) to compute discrete fourier transforms (and its inverse)

$$\rho(x,y,z) = \frac{1}{V}\sum_h \sum_k \sum_l w_{hkl} |\mathbf{F}_{\text{obs}}| e^{-2\pi i(hx+ky+lz-\alpha'_{hkl})} \tag{3}$$

where α'_{hkl} is the phase at each set of coordinates in the reciprocal space. The phases are not directly available but are necessary (actually more important than intensities) for computing $\rho(x,y,z)$ (electron density).

The phases can be solved either by phasing experiments, in which case the resolution is relatively low but the data is unbiased, or by molecular replacement (MR), in which case the resolution is relatively high but the data suffer from bias (the phases resemble the search model more than the actual structure). Thus, bias minimisation measures such as maxinum likelihood map coefficients are applied before electron density map contrstruction.

2 Assessing the Quality of A Structure During Model Building and Refinement

Solving the electron density and ultimately the atomic structure from experimental data is an iterative process, starting from a model with low resolution and imperfect phases and progressively refining the model by applying prior knowledge on the information the current model can provide so far.

In general, the first step in refinement is to construct a crude electron density map (Equation (3)) using phase angles (α') obtained either from MR or experimental phasing. This initial map, $\mathbf{F}_0$, is improved using techniques such as *density modification, solvent leveling, solvent flattening.* At some critical point in the iterative improvement of phases, the map becomes clear enough that a protein chain can be traced through it. From this point on, real-space refinement is also conducted at each round of iteration. Electron density map is obtained (details in Section 2.3), and filled with atoms, and transformed back to the reciprocal space. This process repeats until converging to a high-quality strucure. R and R_{free}, explained below, are used to assess progress in the refinement, and is also a major factor in evaluating the quality of a model.

2.1 R and R_{free}

R is a measure of the deviation of calculated intensities from models (details in Section 2.2) from the observed intensities in the diffraction pattern, defined by the following equation:

$$R = \frac{\sum ||\mathbf{F}_{\text{obs}}| - |\mathbf{F}_{\text{calc}}||}{\mathbf{F}_{\text{obs}}}$$

Since bias can easily be introduced into the R value (especially by overparameterisation, see Section 2.2) and a reduction of R value sometimes does not improve the actual quality of structure (Kleywegt and Brünger (n.d.)), a small fraction (typically around 5%[2]) of randomly selected reflections are removed from the data used for refinement. These reflections can then be used to calculate an R factor, denoted as R_{free}, whose reduction can be considered as an *unbiased estimate* of the improvement of the model.

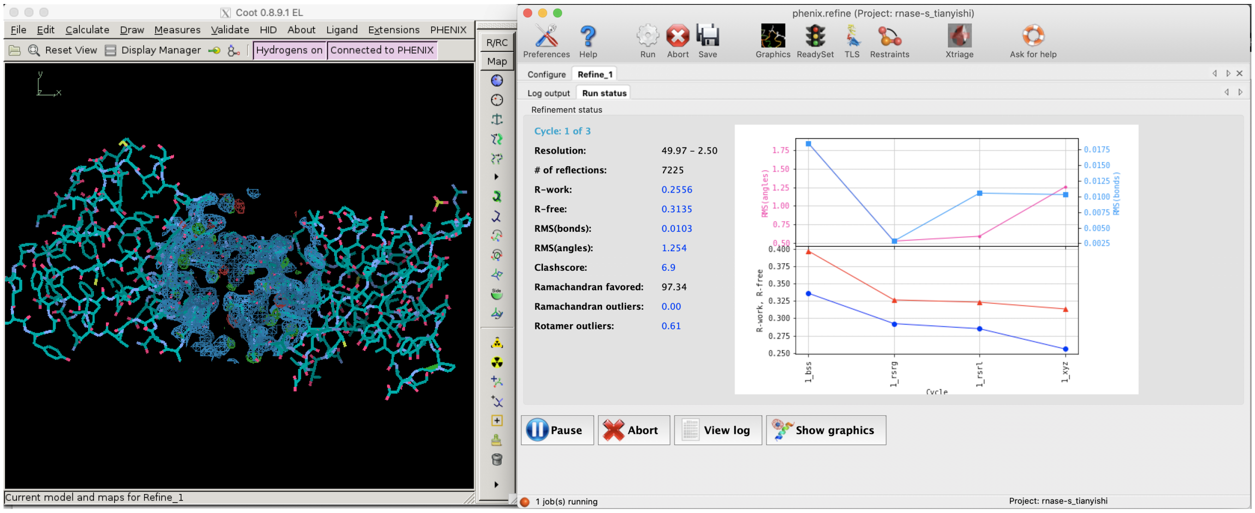
Figure 1 shows the decrease of R and R_{free} during a refinement task conducted in Phenix.

2.2 Reciprocal-Space Refinement: Refinement by least squares

Reciprocal-space refinement involves computerised attempts to improve agreement between $\mathbf{F}_{\text{obs}}$ and $\mathbf{F}_{\text{calc}}$ by *without* consideration of the maps and models. *Refinement by least squares* is the earliest successful technique and is discussed here.[3]

[2] In PDB files, this fraction is recorded in the field `FREE R VALUE TEST SET SIZE`

[3] modern refinement softwares use techniques such as simulated annealing and beyesian methods to improve the accuracy of refinement



Cycle: 1 of 3

Resolution:	49.97 – 2.50
# of reflections:	7225
R-work:	0.2556
R-free:	0.3135
RMS(bonds):	0.0103
RMS(angles):	1.254
Clashscore:	6.9
Ramachandran favored:	97.34
Ramachandran outliers:	0.00
Rotamer outliers:	0.61

Figure 1: R and R-free values decreases as refinement proceeds (left: Coot; right: phenix.refine).

The goal of refinement by least squares is, find (x_j, y_j, z_j) for all atom j whose *expected* (i.e. computed) structure factor amplitudes, $|\mathbf{F}_{\text{calc}}|$ are as close as possible to *observed* structure factor amplitudes, $|\mathbf{F}_{\text{obs}}|$. Specifically, this means minimising the function Φ:

$$\Phi = \sum_{hkl} (w_{hkl} \mid \mathbf{F}_{\text{obs}} \mid - \mid \mathbf{F}_{\text{calc}} \mid)^2_{hkl} \tag{4}$$

where w_{hkl} is the weight term that depends on the reliability of the corresponding measured intensity and $|\mathbf{F}_{\text{calc}}|$ is a variant form of Equation (2) that can include additional parameters such as B-factor (B_j) and occupancy n_j. An equation with B_j and n_j included can be written as:

$$\mathbf{F}_{\text{calc}} = \sum_{j} n_j f_j e^{2\pi i(hx_j + ky_j + lz_j) - B_j[(\sin\theta)/\lambda]^2} \tag{5}$$

Note that the equatin shows that the effect of B-factors depends on the angle of reflection $(\sin\theta)/\lambda$.

Solving the minimum of Φ analytically is impractical, and instead numerical methods are used, which would lead to a minimum closest to the starting value.To prevent the refinement converging to a local minimum, it is important that the starting parameters be near the *global minimum*. Is also important *not* to include too many parameters (such as B-factor) at the initial stages of resolution when the resolution is low, as this would decrease the radius of convergence (Figure 2).

To minimise the number of parameters used during early stages of refinement (and thus to increase radius of convergence), individual (x, y, z) coordinates are actually not used. Instead, only torional angles ψ and ϕ are allowed to change, and all bond lengths and angles are fixed to their theoretical average, side chains are assumed to be in their preferred conformation, and peptide linkages are fixed to be planar. This strategy is known as *restrained* reciprocal space refinement. As refinement proceeds, more parameters, from individual (x, y, z) coordinates to isotropic B-factors and finally anisotropic factors, can be added into calculation.

2.3 Real-Space Refinement: Map Fitting

Map fitting or *model building* entails building a molecular model that fits realistically into the current electron density contour map.

To reduce the bias (towards $\mathbf{F}_{\text{calc}}$ $(\mathbf{F}_{\text{c}})$) when constructing the electron density map, Fourier syntheses of $\mathbf{F}_{\text{obs}}$ and $\mathbf{F}_{\text{calc}}$ are used. A Fourier synthesis $m\mathbf{F}_{\text{o}}$ - $n\mathbf{F}_{\text{c}}$ is calculated as:

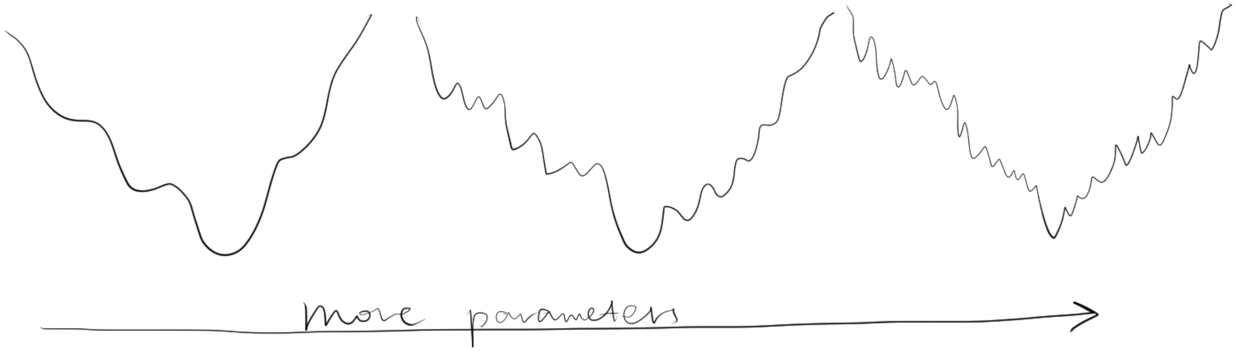


Figure 2: Adding number of parameters improves precision of refinement, but makes it more unlikely to reach the global minimum from a given point. Thus, refinement starts with a small number of parameters, and more parameters are only added after the success of previous lower-resolution refinement steps.

$$\rho(x, y, z) = \frac{1}{V}\sum_{h}\sum_{k}\sum_{l}(m|\mathbf{F}_{\mathrm{o}}| - |\mathbf{F}_{\mathrm{c}}|)e^{-2\pi i(hx+ky+lz-\alpha'_{hkl})}$$

and its corresponding electron density map is called an $m\mathbf{F}_{\mathrm{o}}$ - $n\mathbf{F}_{\mathrm{c}}$ map.

Simply put, the $2\mathbf{F}_{\mathrm{o}}$ - $\mathbf{F}_{\mathrm{c}}$ map resembles a molecular surface, and a $\mathbf{F}_{\mathrm{o}}$ - $\mathbf{F}_{\mathrm{c}}$ map emphasises the error (positive density implies that the unit cell contains more electron density in this region than implied by the model ($\mathbf{F}_{\mathrm{c}}$). Near the end of refinement, the $\mathbf{F}_{\mathrm{o}}$ - $\mathbf{F}_{\mathrm{c}}$ map becomes rather empty except in problem areas, which may need to be corrected manually.

Fitting a molecular model into the electron density map depends on prior knowledge, such as average bond lengths and angles, the amino acid sequence of the protein, properties of peptide chains, etc. For example, we know that carboxyl oxygens in adjacent amino acid residues in a β-sheet point in opposite directions. Thus, once a β-sheet along with one or two carboxyl oxygen are discernible, we can make a sensible guess of the positions of all other carbonyl oxygens.

3 Assessing Quality After Automatic Model Building and Refinement

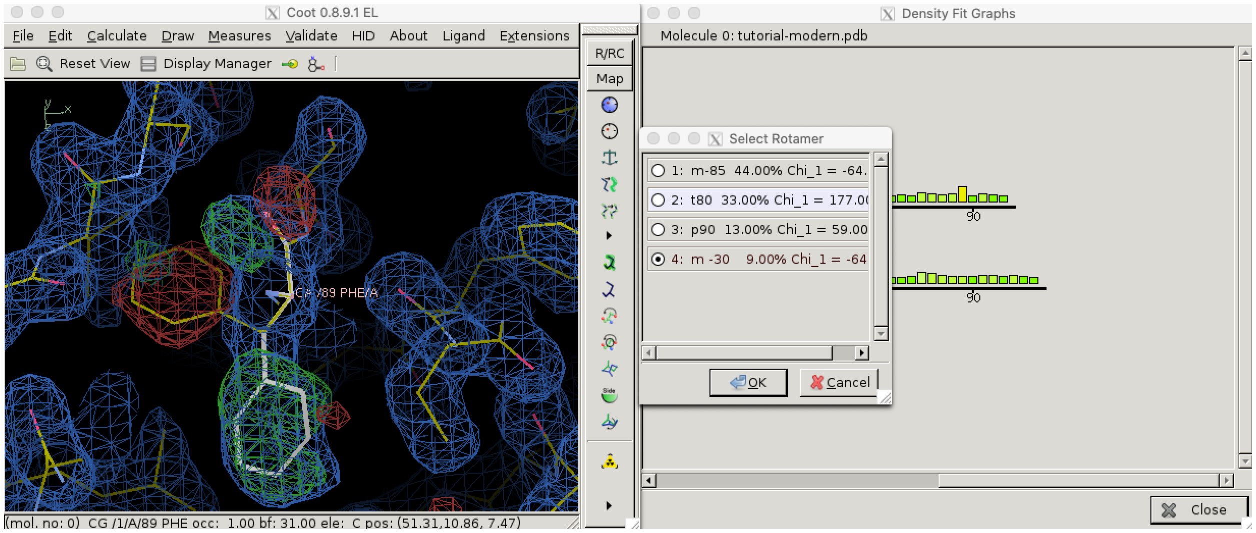
3.1 Density fit analysis and local geometry validation

Automatic model building and refinement use the decrease of R value as an indicator of progress and terminates when R is considered to be sufficiently low. This may lead to situations where the global R is favourable but local geometry can still be improved.

Local geometry validation programs, such as “Density fit analysis” in Coot (Figure 3), evaluate the model geomtry on a per-residue basis and flag outliers. These outliers can then be fixed manually. With the aid of electron density contour maps (where model atoms lie outside $2\mathbf{F}_{\mathrm{o}}$ - $\mathbf{F}_{\mathrm{c}}$ contours, the $\mathbf{F}_{\mathrm{o}}$ - $\mathbf{F}_{\mathrm{c}}$ will often show the atoms with negativel contours, with nearby positive contours pointing to correct locations for these atoms).

3.1.1 Structural parameters: bond length, bond angle, chirality, planarity, and dihedral angles

During late stages of refinement, restraints on structura paramaters such as bond lengths and angles are release such that thay are allowed to change freely as long as R values could be made to decrease. This



```
○ 1: m-85  44.00% Chi_1 = -64.
○ 2: t80  33.00% Chi_1 = 177.0(
○ 3: p90  13.00% Chi_1 = 59.00
● 4: m -30   9.00% Chi_1 = -64
```

```
(mol. no: 0)  CG /1/A/89 PHE occ:  1.00 bf: 31.00 ele:  C pos: (51.31,10.86, 7.47)
```

Figure 3: Use the 'Density fit analysis' function to evaluate model geometry on a per-residue basis and plot a histogram that shows outliers, then use the 'Rotamers' tool to fix a side chain that's pointing the wrong way.

may lead to unrealistic models with improbable structural parameters. Thus, they need to be validated. By convention, this is done by calculating the root-mean-square deviations (RMSD) of these parameters in the model from the corresponding set of values based opon the geometry of small organic molecules. These values are recorded in PDB files in `REMARK 3` fields:

```
REMARK   3  DEVIATIONS FROM IDEAL VALUES.
REMARK   3                   RMSD          COUNT
REMARK   3   BOND      :  0.003           1366
REMARK   3   ANGLE     :  0.675           1846
REMARK   3   CHIRALITY :  0.050            186
REMARK   3   PLANARITY :  0.005            246
REMARK   3   DIHEDRAL  : 15.473            459
```

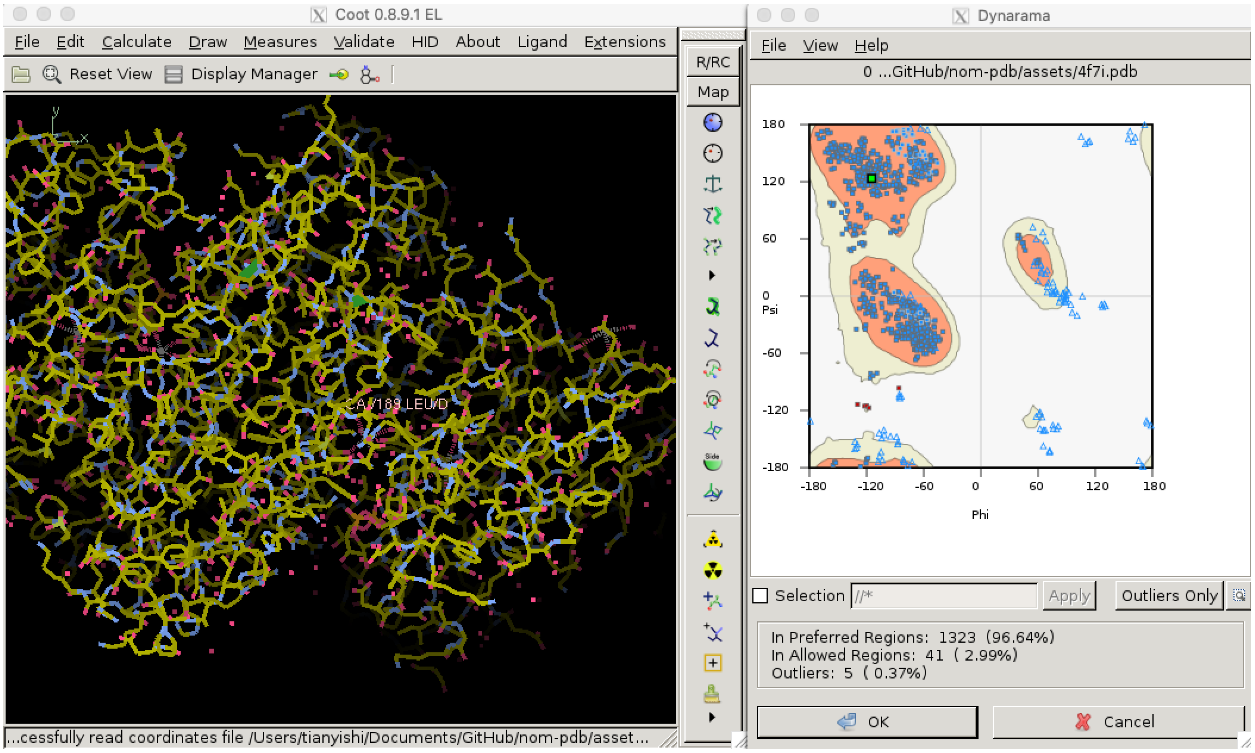
Because bond angles, lengths, chirality and planarity are restrained in early stages of refinement, they are less likely to deviate much from ideal values. Dihedral angles, which is allowed to change throughout the refinement, produce much larger RMSD and are worth closer examination.

3.1.2 Torsional Angles and Ramachandran Plot

Torsional (dihedral) angles ψ and ϕ are show much more variation than bond lengths and angles, but only a subset of all possible (ϕ, ψ) pairs are allowed so that adjacent amino acid side chains do not clash. Validation of torsional angle is achieved via a lookup table, where the keys are (ϕ, ψ) pairs and values are scores. A (ϕ, ψ) is considered preferred or allowed if its score is within certain thresholds. Otherwise, it is considered an outlier. Due to glycine's small size and proline's cyclic structure, the preferred/allowed regions of their torsinal angle pairs are defined differently, for example in Phenix/cctbx[4]. Torsional angle validation is often visualised with a Ramachandran plot, as shown in Figure 4.[5]

[4] In Phenix/cctbx, a lookup table (actually implemented as an array for better performance) is defined for each of the following cases: glycine, cis-proline, trans-proline, pre-proline, isoleucine/valine, and all other amino acid residues: `mmtbx/validation/ramachandran/rama8000_tables.h`

[5] Recently I've been developing a PDB file parser. It is now able to parse coordinate information reliably, which allows Ramanchandran plot analysis: https://github.com/TianyiShi2001/protein



...cessfully read coordinates file /Users/tianyishi/Documents/GitHub/nom-pdb/asset...

```
In Preferred Regions:  1323  (96.64%)
In Allowed Regions:  41  ( 2.99%)
Outliers:  5  ( 0.37%)
```

Figure 4: Validating dihedral angles with Ramachandran plot in Coot

3.1.3 Undefined regions

Often, some regions (often the terminal regions) of the protein are highly disordered and produce no detectable diffraction at high resolution and are thus invisible in the electron-density maps. It is not uncommon to omit these residues in the model. In some cases, however, exact coordinates are given for these atoms, along with very high *B*-factors (above 50 or 60 angstroms), and one must be careful not to over-interprete these regions.

3.1.4 Unexplained Density

Empty electron density may remain after all known contents of the unit cell have been located. This can either due to an artifect of missing Fourier terms, reagents used in purification or crystallisation, or it could be due to previously unknown ligands (cofactors, inhibitors or allosteric effectors).

3.1.5 Distortions due to crystal packing

Although flexibility of atoms is usually reflected by the *B*-factor, in certain situations dynamic regions of a protein molecule can be rigidly fixed in a specific conformation as a result of crystal packing interactions. It should be declared in the PDB header that the protein is affected by crystallisation. One can also check whether these interactions are likely to occur by displaying all neighbouring, symmetry-related molecules in the crystal structure and examine if any intermolecular interactions are present that are a result of crystal packing.

References

Blow, David. 2002. *Outline of Crystallography for Biologists*. Oxford University Press.

Kleywegt, G J, and A T Brünger. n.d. "Checking Your Imagination: Applications of the Free R Value." *Structure* 4 (8): 897–904. https://doi.org/10.1016/s0969-2126(96)00097-4.

Rhodes, Gale. 2006. *Crystallography Made Crystal Clear Clear*. Academic Press.

Rupp, Bernhard. 2010. *Biomolecular Crystallography: Principles, Practice, and Application to Structural Biology*. Garland Science.