Introduction to Crystallographic Refinement

Jack Tanner
Sept. 24, 2014
Outline

- The purpose of refinement
- $R_{\text{cryst}}$
- B-factors
- Geometrical (chemical) restraints
- Observations to parameters ratio
- $R_{\text{free}}$
Structural Biology Workflow

- Cloning, protein expression, and purification
- Crystallization
- X-ray diffraction data
- Interpret diffraction data
- 3-D structure
- Extract functional insights
- Interpret structure and complementary experiments
Iterative cycles of model building (MB) and refinement (REF)

phases

model & ed map

MB coot

defined map

structure factor amplitudes $|F_{obs}|$

better model

REF phenix

refined model

final model

deposit in PDB
Model building fixes problems in models

- **Experimental phasing;** initial models from autobuilding programs typically have many errors
- backbone positioned correctly, but side chain positioned wrong (e.g., wrong rotomer)
- backbone correct but side chain missing (i.e., sequence not assigned)
- sections of the polypeptide missing (e.g., loops)
- water, metal centers, cofactors, or ligands missing
- backbone fits the density poorly
- protein model built into density for ligand, metal center, or cofactor
- residues with backbone correct but sequence assignment is wrong
- polypeptide chain going in wrong direction
Model building fixes problems in models

- **Molecular replacement**
  - backbone positioned correctly, but side chain positioned wrong (e.g. wrong rotomer)
  - backbone correct but side chain missing (chainsaw search model)
  - sections of the polypeptide missing (e.g., loops)
  - water, metal centers, cofactors, or ligands missing
  - backbone fits density poorly (conformational differences from search model)

- **Ligand complexes and site-directed mutants**
  - side chain for mutated residue missing (by design)
  - side chain for active site residues missing (by design)
  - ligand is missing (by design)
  - conformational changes in active site (can be side chain and/or backbone)
The model that is output from a model building session will be consistent with the electron density map, but how well does the model agree with the experimental data ($|F_{\text{obs}}|$)?
Structure Factor Equation and the R-factor

\[ F(h,k,l) = \sum_{j=1}^{\text{atoms}} f_j \exp\{2\pi i (hx_j + ky_j + lz_j)\} \]

- Diffracted X-ray are described by complex vector \( F \).
- Given \((x_j, y_j, z_j)\) for each atom in the current model, one can generate “calculated structure factors”. The calculated SF amplitude is \(|F_c(h,k,l)|\) and can be compared to the observed amplitude, \(|F_o(h,k,l)|\).
- The R-factor (also called \( R_{\text{cryst}} \)) expresses the agreement between the model and the observed data:

\[
R_{\text{cryst}} = \frac{\sum_{h} \sum_{k} \sum_{l} \{|F_o(h,k,l)|-|F_c(h,k,l)|\}}{\sum_{h} \sum_{k} \sum_{l} |F_o(h,k,l)|}
\]
Expected values of $R_{\text{cryst}}$

$$R_{\text{cryst}} = \frac{\sum \sum \sum_{h \hspace{0.1cm} k \hspace{0.1cm} l} \{ |F_o(h,k,l)| - |F_c(h,k,l)| \} \} \sum \sum \sum_{h \hspace{0.1cm} k \hspace{0.1cm} l} |F_o(h,k,l)|$$

<table>
<thead>
<tr>
<th>Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>well-refined small molecule structure</td>
</tr>
<tr>
<td>0.1</td>
<td>well-refined protein structure with solvent and ligands</td>
</tr>
<tr>
<td>0.2</td>
<td>high-res protein structure w/o water</td>
</tr>
<tr>
<td>0.3</td>
<td>models from autobuilding</td>
</tr>
<tr>
<td>0.4</td>
<td>correct MR solution</td>
</tr>
<tr>
<td>0.5</td>
<td>random $</td>
</tr>
</tbody>
</table>
Two Purposes of Refinement

Adjust the parameters of the model to

(1) improve the agreement between the model and the experimental data, i.e, achieve low $R_{\text{cryst}}$

and

(2) ensure that the model makes good chemical sense (molecular geometry)
Model have parameters

model: $y = a + bx$

R-square = 0.87
Parameters of the crystallographic model

$$F(h,k,l) = \sum_{j=1}^{\text{atoms}} f_j \exp\{2\pi i (hx_j + ky_j + lz_j)\}$$

$$F(h,k,l) = \sum_{j=1}^{\text{atoms}} f_j Q_j \exp\{-B_j (\sin \theta / \lambda)^2\} \exp\{2\pi i (hx_j + ky_j + lz_j)\}$$
Parameters of the crystallographic model

- \( f_j \), atomic scattering factor; known for each element and fixed during refinement (see plot on next slide)
- \((x_j, y_j, z_j)\) for each atom; adjustable parameter
- Occupancy (\( Q_j \)) is the fraction of time atom \( j \) spends at position \((x_j, y_j, z_j)\); usually fixed at \( Q=1.0 \).
  - allowed values of 0-1; note that \( Q=0 \) means the atom is absent
  - \( Q \) and \( B \) are correlated; typically can’t refine both \( Q \) and \( B \)
  - \( Q \neq 1 \) is used for modeling multiple side chain conformations and weakly bound ligands
- \( B_j \) is related to an atom’s motion; adjustable parameter

\[
F(h,k,l) = \sum_{j=1}^{\text{atoms}} f_j Q_j \exp\{-B_j (\sin\theta/\lambda)^2\} \exp\{2\pi i (hx_j + ky_j + l z_j)\}
\]
Atomic Scattering Factor, $f_j$

- describes the scattering power of an atom
- also called the atomic form factor
- has units of electrons
- these parameters are known for every element of interest; values are coded into all refinement programs
- fixed during refinement
- see Rupp, Fig. 6-12. for plots for several elements

$\frac{f}{(\sin \Theta)/\lambda}$

scattering factor for O
More on B-factors

- also called temperature factor and Debye-Waller factor
- related to the average displacement of an atom from its mean position: 
  \[ B = 8\pi^2<(r-r_0)^2> \]
- has units of ?
- isotropic B is used for typical protein structures (d > 1.2 Å)
- typical range for protein structure is 10-100 Å²
- refinement of anisotropic B (a matrix of B-factors for each atom) is possible for atomic resolution data; anisotropic B sometimes represented by thermal ellipsoids (Bottoms et al. Prot. Sci. 2004)
- don’t confuse the ANISOU lines in typical pdb with anisotropic B-factors; these lines express the results of TLS B-factor refinement
- contribution of atom to \( F_c \) is depends on scattering angle and is thus resolution-dependent. Recall the following:
  - from SF equation: \( \exp[-B_j(sin\theta/\lambda)^2] \)
  - resolution: \( d = \lambda/2sin\theta \)
contribution of atom to $F_c$ is resolution-dependent
More on B-factors - TLS Refinement

• recognizes that motion of groups of atoms are correlated, so their B-factors will be correlated

• “group” typically refers to a protein chain or a domain

• Translation, Libration, Screw

• adds only 6 parameters per group but lowers both $R_{\text{cryst}}$ and $R_{\text{free}}$ substantially

• you must define the groups; in phenix one types...
  • chain a
  • chain a and resid 20-30
  • chain a and not resname FAD

• results expressed in the header of pdb and in ANISOU lines. Don’t confuse the ANISOU lines with anisotropic B-factors that one would see in atomic resolution structures.
The parameters of refinement are listed in the PDB file.

PDB file refined without TLS

<table>
<thead>
<tr>
<th>ATOM</th>
<th>1</th>
<th>N</th>
<th>LEU A</th>
<th>5</th>
<th>5.815</th>
<th>18.670</th>
<th>26.851</th>
<th>1.00</th>
<th>31.21</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATOM</td>
<td>2</td>
<td>CA</td>
<td>LEU A</td>
<td>5</td>
<td>6.758</td>
<td>17.790</td>
<td>27.543</td>
<td>1.00</td>
<td>31.43</td>
<td>C</td>
</tr>
<tr>
<td>ATOM</td>
<td>3</td>
<td>C</td>
<td>LEU A</td>
<td>5</td>
<td>6.164</td>
<td>16.930</td>
<td>28.675</td>
<td>1.00</td>
<td>31.23</td>
<td>C</td>
</tr>
<tr>
<td>ATOM</td>
<td>4</td>
<td>O</td>
<td>LEU A</td>
<td>5</td>
<td>6.752</td>
<td>15.910</td>
<td>29.047</td>
<td>1.00</td>
<td>31.21</td>
<td>O</td>
</tr>
<tr>
<td>ATOM</td>
<td>5</td>
<td>CB</td>
<td>LEU A</td>
<td>5</td>
<td>7.984</td>
<td>18.572</td>
<td>28.044</td>
<td>1.00</td>
<td>32.17</td>
<td>C</td>
</tr>
<tr>
<td>ATOM</td>
<td>6</td>
<td>CG</td>
<td>LEU A</td>
<td>5</td>
<td>8.014</td>
<td>19.392</td>
<td>29.337</td>
<td>1.00</td>
<td>32.48</td>
<td>C</td>
</tr>
<tr>
<td>ATOM</td>
<td>7</td>
<td>CD1</td>
<td>LEU A</td>
<td>5</td>
<td>8.234</td>
<td>18.512</td>
<td>30.562</td>
<td>1.00</td>
<td>33.15</td>
<td>C</td>
</tr>
<tr>
<td>ATOM</td>
<td>8</td>
<td>CD2</td>
<td>LEU A</td>
<td>5</td>
<td>9.138</td>
<td>20.424</td>
<td>29.226</td>
<td>1.00</td>
<td>34.31</td>
<td>C</td>
</tr>
</tbody>
</table>

PDB file refined with TLS

<table>
<thead>
<tr>
<th>ATOM</th>
<th>624</th>
<th>N</th>
<th>THR A</th>
<th>89</th>
<th>23.227</th>
<th>38.757</th>
<th>21.801</th>
<th>1.00</th>
<th>22.07</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANISOU</td>
<td>624</td>
<td>N</td>
<td>THR A</td>
<td>89</td>
<td>2611</td>
<td>2104</td>
<td>3673</td>
<td>-494</td>
<td>-551</td>
<td>585</td>
</tr>
<tr>
<td>ATOM</td>
<td>625</td>
<td>CA</td>
<td>THR A</td>
<td>89</td>
<td>22.871</td>
<td>37.786</td>
<td>20.769</td>
<td>1.00</td>
<td>20.49</td>
<td>C</td>
</tr>
<tr>
<td>ANISOU</td>
<td>625</td>
<td>CA</td>
<td>THR A</td>
<td>89</td>
<td>2397</td>
<td>1990</td>
<td>3400</td>
<td>-508</td>
<td>-496</td>
<td>626</td>
</tr>
<tr>
<td>ATOM</td>
<td>626</td>
<td>CB</td>
<td>THR A</td>
<td>89</td>
<td>24.108</td>
<td>37.212</td>
<td>20.035</td>
<td>1.00</td>
<td>22.73</td>
<td>C</td>
</tr>
<tr>
<td>ANISOU</td>
<td>626</td>
<td>CB</td>
<td>THR A</td>
<td>89</td>
<td>2667</td>
<td>2311</td>
<td>3658</td>
<td>-565</td>
<td>-461</td>
<td>693</td>
</tr>
</tbody>
</table>

Total number of parameters in the model is 5 times the number of atoms. But, typically Q is fixed, so the number of refined parameters is only 4 times the number of atoms.
Recall the Two Purposes of Refinement

Adjust the parameters of the model to

(1) improve the agreement between the model and the experimental data, i.e., achieve low $R_{\text{cryst}}$

and

(2) ensure that the model makes good chemical sense (molecular geometry)
Criterion 2: model must make good chemical sense

- For a small protein and very high resolution data (0.7 Å), a model that satisfies (1) will automatically satisfy (2). This is called unrestrained refinement.

- For a protein with $d > 1$ Å resolution, there is not enough information in the diffraction data to uniquely determine all the parameters of the model. More information is needed for stable refinement. This extra information takes the form of stereochemical restraints.

- If we refined solely to satisfy criterion (1) we would end up with a very low $R$-factor, but the stereochemistry of the protein would be unacceptable.

- If we refined solely to satisfy criterion (2) we would end up with nearly perfect stereochemistry, but the $R$-factor would be unacceptably high. Proper refinement strikes a sensible balance between the two opposing goals.
Think of refinement as an energy minimization problem

\[ V = V_{\text{x-ray}} + \text{Weight} \times V_{\text{geom}} \]

Pseudo-potential energy (unphysical energy)

\[ (|F_o| - k_{\text{scale}} \times |F_c|)^2 \]

Geometry "Force field"

\[ R_{\text{cryst}} \]

geometry
More on $V_{\text{geom}}$

• Why do we need this term?
  • there is not enough information in our measured data ($F_o$) to uniquely determine the (x,y,z) and B of all the atoms in the structure. The refinement is said to be “underdetermined”.

• $V_{\text{geom}}$ restrains your structure during refinement to have the geometrical qualities we typically associate with proteins
  • such as...
  • correct chirality: L-amino acids, not D
  • planar aromatic rings
  • bond angles consistent with hybridization (e.g., CA sp$^3$, C sp$^2$)
  • bond lengths consistent with known atomic bonding radii (e.g., 1.54 Å for C-C, 1.23 Å for C=O)
  • phi and psi consistent with Ramachandran plot
  • atoms have finite size and cannot get too close to each other
Observations/Parameters Ratio

• need \( n/p > 10 \) to consider unrestrained refinement.
• Some examples of \( n/p \) omitting restraints (\( p=4 \times \) number of atoms)
  • rPV, 1.05 Å, 132,869 unique obs., 2828 atoms, \( n/p = 11.7 \)
  • rP4/NMN, 1.35 Å: 66,003 unique obs., 2207 atoms, \( n/p = 7.5 \)
  • rP4/2'AMP, 1.9 Å, 24,575 unique obs., 2085 atoms, \( n/p = 2.9 \)
  • BjPutA 2.1 Å, 172,549 unique obs., 15,636 atoms, \( n/p = 2.7 \)
• The geometrical restraints serve as additional “observations” increasing the observation/parameter ratio, \( n/p \)
• Calculation of \( n/p \) including restraints is complicated because the geometrical restraints are not independent from each other. Rupp BMC has a discussion of this subtlety.
Energy minimization goes downhill always, but simulated annealing can go over pseudo-energy barriers.

Simulated annealing has a larger radius of convergence than minimization.

Protein Conformation
Mathematical Form of $V_{\text{geom}}$

$$V_{\text{geom}} = V_{\text{bonded}} + V_{\text{nonbonded}}$$
Molecular Mechanics Force Field

\[ V = \text{bonded energy} + \text{nonbonded energy} \]

\[ V_{\text{bonded}} = \sum_{\text{bonds}} k_B(r - r_0)^2 + \sum_{\text{angles}} k_\theta(\theta - \theta_0)^2 + \sum_{\text{dihedrals}} k_\phi(1 + \cos(n_\phi + \delta)) \]

\[ V_{\text{nonbond}} = \sum_{i=1}^{N} \sum_{j=i+1}^{N} 4\varepsilon_{ij} \left[ \left( \frac{\sigma_{ij}}{R_{ij}} \right)^{12} - \left( \frac{\sigma_{ij}}{R_{ij}} \right)^{6} \right] + \frac{Z_i Z_j q^2}{4\pi\varepsilon_0 \varepsilon_r R_{ij}} \]

\( N = \text{number of atoms} \)
\( \varepsilon_r = \text{relative permittivity or dielectric constant (80 for water)} \)
\( R_{ij} = \text{distance between atoms i and j} \)
\( Z_i = \text{partial charge of atom i, (-1 to +1)} \)

parameters in red must be determined for each "atom type" using spectroscopy, \textit{ab initio} calculations and other methods.
Bond Energy

Dihedral Energy

van der Waals
Lennard - Jones 6-12 Potential

Electrostatics - Coulomb Potential
van der Waals
Lennard - Jones 6-12 Potential
Geometry statistics to check after each round

The root mean square deviations of bonds and angles from their ideal values are listed in the PDB file.

grep "Final" *pdb

REMARK Final: r_work = 0.1874 r_free = 0.2209 bonds = 0.007 angles = 1.022

phenix.refine tends to produce tight geometry. Jablonski et al. Acta D. 2007 suggest rmsd bonds of 0.015-0.020 for structures refined to $R = 0.15 - 0.20$, but this is debatable.
Geometry statistics

- Other metrics are calculated by the user
  - Ramachandran plot outliers (coot, procheck)
  - MolProbity web server (comprehensive validation)
  - PDB validation server
  - PROCHECK (procheck pdbfile resolution)

- Adopt a “Clean As You Go Policy” in model building and refinement by validating early and often; validation always points out problems in the structure.
Cross-validation and $R_{\text{cryst}}$

- $R_{\text{cryst}}$ tends to decrease as more parameters (e.g. atoms) are added, which can lead to overfitting. $R_{\text{cryst}}$ is “biased”.

- $R_{\text{free}}$ is an R-factor calculated from a set of reflections that is set aside and NOT used in refinement. Developed by Axel T. Brunger (Nature 1992). Since these reflections were not included in refinement, $R_{\text{free}}$ is unbiased.

- $R_{\text{free}}$ is calculated from the test set (usually 5 % of reflections chosen randomly)

- $R_{\text{cryst}}$ is calculated from the working set (other 95 % of data)

- Seems strange to throw away data, but omitting a small percentage of reflections - chosen randomly - does not affect maps. Rupp, BMC, Fig. 9-10 shows this well.

References
International Tables F, chapter 18.2
Why is $R_{\text{cryst}}$ biased?

- $R_{\text{cryst}}$ tends to decrease as more parameters (e.g. atoms) are added, which can lead to overfitting. $R_{\text{cryst}}$ is “biased”.

$$R_{\text{cryst}} = \frac{\sum \sum \sum | |F_o(h,k,l)| - |F_c(h,k,l)| |}{\sum \sum \sum |F_o(h,k,l)|}$$
$R_{\text{cryst}}$ is biased by refinement because of its similarity to the refinement target function.

$$R_{\text{cryst}} = \frac{\sum \sum \sum_{h,k,l} \{ |F_o(h,k,l)| - |F_c(h,k,l)| \}}{\sum \sum \sum_{h,k,l} |F_o(h,k,l)|}$$

Target function for refinement:

$$V = V_{\text{x-ray}} + \text{Weight} \cdot V_{\text{geom}}$$

(geometry term)
Overfitting: adding more parameters to the model lowers $R_{cryst}$ but does not increase the information content of the model.
$R_{\text{free}}$ is a good indicator of phase error and the information content of the model

$R_{\text{cryst}}$

$R_{\text{free}}$

phase error

Fig. 4 of Brunger, Nature 1992
performed refinements starting from different models model

<table>
<thead>
<tr>
<th></th>
<th>2365 randomly placed atoms</th>
<th>protein without chemical restraints</th>
<th>protein with chemical restraints</th>
<th>add 314 good waters</th>
<th>add 1850 randomly placed waters</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R_{\text{cryst}}$</td>
<td>0.16</td>
<td>0.14</td>
<td>0.21</td>
<td>0.15</td>
<td>0.13</td>
</tr>
<tr>
<td>$R_{\text{free}}$</td>
<td>0.54</td>
<td>0.43</td>
<td>0.25</td>
<td>0.21</td>
<td>0.35</td>
</tr>
<tr>
<td>phase error</td>
<td>90°</td>
<td>48°</td>
<td>35°</td>
<td>34°</td>
<td>42°</td>
</tr>
</tbody>
</table>

Table 3 from Brunger, Nature 1992
Case Study of Problematic Data: LmUGM

- space group P3(1)21, a = b = 82.3 Å, c=129.4 Å
- 2.1 Å resolution
- 3247 atoms (No. parameters = 4 x 3247 = 12988)
- No obvious crystal pathologies such as twinning
- Advanced model has high R and R-free
  - $r_{work} = 0.321$ $r_{free} = 0.370$ $bonds = 0.003$ $angles = 0.731$

<table>
<thead>
<tr>
<th>Refinement Strategy</th>
<th>No. params</th>
<th>R-cryst</th>
<th>R-free</th>
<th>deltaR</th>
</tr>
</thead>
<tbody>
<tr>
<td>starting model</td>
<td>12988</td>
<td>0.321</td>
<td>0.370</td>
<td>0.049</td>
</tr>
<tr>
<td>add 113 HOH with automated water picking</td>
<td>13440</td>
<td>0.287</td>
<td>0.372</td>
<td>0.085</td>
</tr>
<tr>
<td>add 253 HOH with automated water picking</td>
<td>14000</td>
<td>0.264</td>
<td>0.348</td>
<td>0.084</td>
</tr>
<tr>
<td>anisotropic B-factors, 391 HOH</td>
<td>32742</td>
<td>0.237</td>
<td>0.376</td>
<td>0.139</td>
</tr>
</tbody>
</table>
Applications of $R_{\text{free}}$

- Global check on structure quality
  - $R_{\text{free}} > 0.4$ near end of refinement indicates a bad problem with the model or the data (twinning, translational pseudosymmetry)
  
  - CRABP II structure intentionally traced backwards has $R_{\text{cryst}} = 0.21$ and $R_{\text{cryst}} = 0.62$
  
  - $R_{\text{free}} - R_{\text{cryst}} < 0.05$ (Figure 2 of Kleywegt and Brunger)
  
  - Large differences indicate problems with model and/or data

- Determine the best parameters and options for refinement
  - Optimal weights (Figure 3 of Brunger 1992)
  - Whether of non-crystallographic symmetry restraints should be used
  - Whether TLS should be used
  - Whether anisotropic B-factors refinement should be used (high resolution)
  - Estimate the expected number of water molecules to be included in model

References
Checking your imagination: applications of the free R value, Structure, 1996, Kleywegt and Brunger.
International Tables F, chapter 18.2
Practical considerations

• Selection of test set can be biased when strong NCS present
• Be careful when selecting test set for a new, isomorphous diffraction data set.
  • This occurs for mutants and ligand complexes that crystallize in same space group and unit cell as wild-type protein
  • for the new data set, use the test set that was used in the previous structure refinement...
  • ...or run simulated annealing to partially minimize phase memory

References
Checking your imagination: applications of the free R value, Structure, 1996, Kleywegt and Brunger.
International Tables F, chapter 18.2
Case Study of Problematic Data: LmUGM

- space group P3(1)21, a = b = 82.3 Å, c=129.4 Å
- 2.1 Å resolution
- No obvious crystal pathologies such as twinning
- Advanced model has high R and R-free
  - $r_{\text{work}} = 0.321$ $r_{\text{free}} = 0.370$ bonds = 0.003 angles = 0.731

<table>
<thead>
<tr>
<th>Refinement Strategy</th>
<th>R-cryst</th>
<th>R-free</th>
<th>deltaR</th>
</tr>
</thead>
<tbody>
<tr>
<td>starting model</td>
<td>0.321</td>
<td>0.370</td>
<td>0.049</td>
</tr>
<tr>
<td>add 113 HOH with automated water picking</td>
<td>0.287</td>
<td>0.372</td>
<td>0.085</td>
</tr>
<tr>
<td>add 253 HOH with automated water picking</td>
<td>0.264</td>
<td>0.348</td>
<td>0.084</td>
</tr>
<tr>
<td>refine occupancies (Q) of all atoms</td>
<td>0.293</td>
<td>0.399</td>
<td>0.106</td>
</tr>
<tr>
<td>refine Q and add 168 HOH</td>
<td>0.269</td>
<td>0.375</td>
<td>0.106</td>
</tr>
<tr>
<td>anisotropic B-factors, 391 HOH</td>
<td>0.237</td>
<td>0.376</td>
<td>0.139</td>
</tr>
<tr>
<td>anisotropic B-factors, 370 HOH, refine all Q</td>
<td>0.239</td>
<td>0.425</td>
<td>0.186</td>
</tr>
</tbody>
</table>