Validation: data analysis

Pavel Afonine

phenix-online.org  lbl.gov  qrefine.com

Missoula, MT

June 27th 2024
Solving structure by crystallography

Pure sample → Crystals → Diffraction experiment → Process & check data

Solve phase problem → Initial map → Improved map → Build model

Refine model → Validate model → Publish & deposit

[Image of crystallography process, including steps and visual representations of data and models]
Validation = checking model, data and model-to-data fit are all make sense and obey to prior expectations
Validation tools: *Crystallography vs Cryo-EM*

**Exact same**
- Model

**Different**
- Data
  - Cryo-EM
  - Diffraction

**Model to data fit**
- Similar
Validation tools in Phenix

Data analysis
- **Xtriage**: Analysis of data quality and crystal defects
- **Mtriage**: Analyze quality of maps in CCP4 format

Merging statistics
- Calculates a variety of statistics for unmerged intensities, including I/σI, R-merge, R-meas, and CC1/2.

Experimental phasing

Molecular replacement

Model building

Refinement

Cryo-EM

Validation
- **Comprehensive validation (X-ray/Neutron)**: Model quality assessment, including real-space correlation and geometry inspection using MolProbity tools
- **Comprehensive validation (cryo-EM)**: Model quality assessment, including real-space correlation, for cryo-EM structures

Structure comparison
- Identify differences between multiple structures of the same protein, using multiple criteria

Calculate CC*
- Comparison of unmerged data quality with refined model, as described in Karplus & Diederichs (2012)

EMRinger
- Model validation for de novo electron microscopy structures

Ligands
Xtriage: all about your diffraction data

- Matthews coefficient probabilities
- Completeness by resolution
- Wilson plot sanity
- Detection of translational NCS (tNCS)
- Analysis of systematic absences and combination of tNCS with current space group
- Anomalous signal from measurability analysis
- Symmetry and twinning analyses
- Alternative point-group symmetry (can be detected on the basis of an R-value analyses)
Intensity statistics suggest twinning (intensities are significantly different from expected for normal data) and one or more twin operators show a significant twin fraction.

Translational NCS does not appear to be present.

Ice rings do not appear to be present.

The fraction of outliers in the data is less than 0.1%.

The data are not significantly anisotropic.

The resolution cutoff appears to be similar in all directions.

The overall completeness in low-resolution shells is at least 90%.

Overall completeness is above 90%.
Wilson B

Wilson statistics assumes atoms of the same kind are randomly distributed in the unit cell and have the same isotropic B-factors.

- Mean B and Wilson B are usually similar
  - Wilson B is dominated by strongly diffracting (lower B) atoms that contribute more to high-res reflections
    - Wilson B represents the lower end of the range of B-factors
    - Discrepancy between Wilson B and mean B is not important
Wilson plot (mean intensity vs resolution)

- The Wilson plot looks at mean intensity of diffraction by resolution, a curve which has a predictable shape.
Wilson plot (mean intensity vs resolution)

- Main reasons for deviations from expected distribution
  - Bad data (e.g., ice rings or poor data processing)
  - Macromolecule that doesn’t look like the average protein
  - Looking at only a part of the plot (e.g., low-resolution data)
Data completeness

- PDB code: 1NH2, resolution 1.9Å, showing E6-E8

2mFo-DFc , 1σ
Data completeness

Completeness by resolution:

<table>
<thead>
<tr>
<th>Resolution</th>
<th>Completeness</th>
</tr>
</thead>
<tbody>
<tr>
<td>19.9274 - 3.2441</td>
<td>0.78</td>
</tr>
<tr>
<td>3.2441 - 2.5767</td>
<td>0.99</td>
</tr>
<tr>
<td>2.5767 - 2.2515</td>
<td>1.00</td>
</tr>
<tr>
<td>2.2515 - 2.0459</td>
<td>1.00</td>
</tr>
<tr>
<td>2.0459 - 1.8993</td>
<td>0.99</td>
</tr>
</tbody>
</table>

Overall completeness in $d_{\text{min}}$-$\text{inf}$: 0.95

Systematic data incompleteness can distort maps
Non-crystallographic symmetry NCS

- Two or more molecules in the ASU related by rotation-translation
- NCS is found in about 1/3 to 1/2 of crystal structures
- Usually helps solving/refining models at medium-to-low resolution
- A special case of NCS, translational NCS (tNCS) leads to complications
Translational NCS (tNCS)

- tNCS arises when the ASU contains components that are oriented in (nearly) the same way and can be superimposed by a translation that does not correspond to any symmetry operation in the space group.

- Used to complicate MR (Phaser now can deal with it!)
- Risk to bias OMIT map
**Translational NCS (tNCS)**

*Xtriage (Project: 1j4r)*

<table>
<thead>
<tr>
<th>Status</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Translational NCS is present at a level that may complicate refinement (one or more peaks greater than 20% of the origin)</td>
<td></td>
</tr>
<tr>
<td>The intensity statistics look normal, indicating that the data are not twinned.</td>
<td></td>
</tr>
<tr>
<td>Ice rings do not appear to be present.</td>
<td></td>
</tr>
<tr>
<td>The fraction of outliers in the data is less than 0.1%.</td>
<td></td>
</tr>
<tr>
<td>The data are not significantly anisotropic.</td>
<td></td>
</tr>
<tr>
<td>The resolution cutoff appears to be similar in all directions.</td>
<td></td>
</tr>
<tr>
<td>The overall completeness in low-resolution shells is at least 90%.</td>
<td></td>
</tr>
<tr>
<td>The completeness is 98.98%.</td>
<td></td>
</tr>
</tbody>
</table>

Please inspect all individual results closely, as it is difficult to automatically detect all issues.
Translational NCS (tNCS) and twinning

Diagnostic tests for twinning and pseudosymmetry

Using data between 10.00 to 2.21 Angstrom.

Patterson analyses

Largest Patterson peak with length larger than 15 Angstrom:

- Fractional coord.: 0.333, -0.333, -0.330
- Distance to origin: 41.406
- Height relative to origin: 62.542 %

Explanation

The p-value, the probability that a peak of the specified height or larger is found in a Patterson function of a macromolecule that does not have any translational pseudo-symmetry, is equal to 1.109e-05. p_values smaller than 0.05 might indicate weak translational pseudo symmetry, or the self vector of a large anomalous scatterer such as Hg, whereas values smaller than 1e-3 are a very strong indication for the presence of translational pseudo-symmetry. Translational pseudo-symmetry is very likely present in these data. Be aware that this will change the intensity statistics and may impact subsequent analyses, and in practice may lead to higher R-factors in refinement.

Wilson ratio and moments

Acentric reflections:

\[
\frac{\langle I^2 \rangle}{\langle I \rangle^2} = 2.430 \quad \text{(untwinned: 2.000; perfect)}
\]

\[
\frac{\langle F^2 \rangle}{\langle F \rangle^2} = 0.750 \quad \text{(untwinned: 0.785; perfect)}
\]

\[
\text{twin 1.300)} \quad \frac{\langle F^2 \rangle}{\langle F \rangle^2} = 0.885
\]
Twinning

- Twinning is a crystal growth disorder

Typically only merohedral twinning is dealt with in a meaningful way in macromolecules
Twinning

- Merohedral twining occurs when your crystal is composed of identical but rotated crystals combined together such that their lattices matching

\[
I_{OBS}(h) = \alpha_1 I(h) + \ldots + \alpha_N I(T_N h)
\]

\[
\alpha_1 + \ldots + \alpha_N = 1
\]
Twinning

- Twinning parameterization
  - **Twin law** describes orientation of different species relative to each other (rotation matrix T that transforms hkl indices of one species into the other)
  - **Twin fraction (α):** fractional contribution of each component
    - Estimated by Xtriage
    - Refined by phenix.refine

\[
I_{\text{OBS}}(h) = \alpha_1 I(h) + \ldots + \alpha_N I(T_N h)
\]

\[
\alpha_1 + \ldots + \alpha_N = 1
\]
Twinning

- tNCS can mask effects of twinning
- If both are present, intensity distributions may look like normal
  - First check for tNCS and use different test for twinning (L-test)
- If crystal is twinned, you have lost information
- Maps going to have model bias that is worse than usual
- Experimental phasing may be difficult
- False symmetry may appear
Watch for outliers

- **R-factor** in resolution bins helps to identify:
  - Problem with bulk-solvent modeling
  - Problems at high resolution
  - Artifacts (green line):

    INDE 3 5 -42 IOBS= 99999.999 SIGIOBS= 0.000