Handout 12
Structure refinement

Completing the structure and evaluating how good your data and model agree

Why you should refine a structure

- We have considered how atoms are located by Patterson, direct methods or specialized methods, as well as from Fourier difference maps
- The atomic positions extracted from these methods are close to the correct values, but very rarely in exactly the right place
- During a process called refinement, the starting atomic positions are optimized
  - the goal is to get an as good fit between your model and your data as possible
Possible refinement methods

- You can modify your model in order to obtain better agreement between observed and calculated Fourier maps.
- Alternatively, you can modify your model to get better agreement between observed and calculated structure factor magnitudes \(|F_o|\) and \(|F_c|\).
- For most small molecule structure solutions, the latter process is used.

Letting the computer do the work

- The refinement process can be described as a minimization process:
  - minimization of the difference between \(|F_o|\) and \(|F_c|\)
  - easily automated
- The best model is obtained by minimizing the following expression:

\[
R' = \sum_{hkl} w_{hkl} \left( \frac{|F_o| - k|F_c|}{|F_o|} \right)^2
\]

- called a least squares minimization
- \(w_{hkl}\) are weighing factors related to the quality/reliability of the data
- \(k\) is a scale factor (inverse of scale factor that would need to be applied to \(|F_o|\)
Imagine a function $F$ that is linearly dependent on a set of parameters, $x_j$:

$$F(x_1, x_2, \ldots, x_n) = p_1 x_1 + p_2 x_2 + \ldots + p_n x_n$$

Suppose we make $m$ independent measurements of $F$ for different values of $x_j$:

- we want to get the parameters $p_j$
- if $m = n$, we just have to solve a set of simultaneous equations
- if $m > n$, the system is overdetermined

Least squares minimization

Crystallography and least squares

The crystallographic function $F(hkl)$ (with the variables: atomic coordinates and thermal parameters of each atom) is not a linear function of the model parameters:

$$F(hkl) = \sum_{j=1}^{N} f_j \exp[2\pi i (hx_j + ky_j + lz_j) - \phi(hkl)]$$

The equation cannot be solved for the correct parameters in one go

Use an iterative process instead
Refinement iterations

- Our goal is to minimize
  \[ R' = \sum_{hk\ell} w_{hk\ell} \left( |F_o| - k|F_c| \right)^2 \]

- We can use an iterative least squares approach to give parameter shifts that will lead to an improved agreement between $|F_o|$ and $|F_c|$.

- The process is repeated until the suggested shifts are insignificantly small.
  - usually considered to be achieved when parameter shift $<<$ standard deviation
  - at this minimum, the derivative of $R'$ with respect to each parameter ($x_j, y_j, z_j, B_j$) should be zero

Convergence

- A refinement process is usually considered as finished when convergence is reached
  - when all parameter shifts $<<$ standard deviation

- However, the process only works well if the starting model is sufficiently good
  - many local minima in which the refinement could “get stuck”

- Other methods offer better chances to find the absolute minimum from a bad starting model
  - simulated annealing, random walk…
False minima

Errors

- The errors of all observations are included in the refinement process via the weight factors
  - weak or uncertain reflections will have less weight than strong reflections
- The least squares output will provide esds (estimated standard deviations) for all refined parameters
- Watch out for high correlation coefficients
  - correlation coefficients tell you whether two model parameters really are independent
Judging the refinement

- Statistical values are used to judge the goodness of a refinement
- The level of agreement between observed and calculated structure factors is often indicated by R factors and Goodness of Fit (GooF) values

\[
R_F = \sum_{hkl} \left( |F_o(hkl)| - k|F_e(hkl)| \right) / \sum_{hkl} |F_o(hkl)|
\]

\[
wR^2 = \left\{ \sum_{hkl} \left[ w \left( |F_o(hkl)|^2 - |F_e(hkl)|^2 \right)^2 \right] / \sum_{hkl} \left[ w |F_o(hkl)|^2 \right]^2 \right\}^{1/2}
\]

\[
GooF = S = \left\{ \sum_{hkl} \left[ w \left( |F_o(hkl)|^2 - |F_e(hkl)|^2 \right)^2 \right] / (n - p) \right\}^{1/2}
\]

- for n observations and p parameters

R factors

- For a good small molecule refinement, the final $R_F$ values are expected to be ~0.02-0.08
- Placing random atoms in a unit cell is expected to give R factors of 0.83 and 0.59 for centric and acentric space groups, respectively
- Obtaining $R_F < 0.20$ usually means that the structural model has no major errors in it
Refinement strategy

- Unit cell constants need to be refined
  - original indexing just gives approximate values

- Atomic positions obtained by Patterson searches, direct methods or other approaches should be refined
  - look at interatomic distances to judge whether refined positions make sense
  - atoms in special positions cannot move in all directions

- Atomic displacement parameters should be varied
  - can indicate wrong atomic weight: \( Z_{\text{mod}} > (\leq) Z_{\text{real}} \) leads to large (small) ADPs
  - isotropic overall temperature factors can be obtained from Wilson plots as a first approximation

Estimating the overall temperature factor

- We know that observed structure factor magnitudes are smaller than real values because of thermal motion and scaling issues

\[
K|F_o(hkl)|^2 \approx |F(hkl)|^2 = \sum f^2 e^{-2B \sin^2 \theta / \lambda^2}
\]

\[
\ln[\sum |F_o(hkl)|^2 / \sum f^2] = \ln K - 2B \sin^2 \theta / \lambda^2
\]

- for random atom placement in the unit cell

- Both K and B can be obtained from a Wilson plot
Wilson plots

It is possible to impose certain restrictions on the refinement due to some knowledge that is not inherent in the diffraction data.

- restraints and constraints

Chemical knowledge, such as bond distances and angles, can be used as a restraint in a minimization procedure.

- a restraint will make certain moves unfavorable, but will not prohibit them

Knowledge about molecular connectivity can be used.

- e.g., a rigid aromatic ring can be described by three positional and three rotational parameters instead of 4n parameters (n = number of atoms).

- this would be a rigid body constraint

When to use restraints and constraints

- Restraints and constraints can improve your data to parameter ratio by reducing the number of parameters or “adding” observations
  - very useful if your data are limited or of low quality
- They improve the convergence properties of your refinement and may allow a refinement to converge to the correct answer even if the starting model is poor
  - they make potentially disastrous parameter changes (e.g., one carbon atom moving so far that the aromatic ring will no longer be connected) unfavorable or prohibit them altogether