Structure Determination
Summary: Crystallography in a nutshell. Lecture no. 4.
(Crystallography without tears, part 2)

Cele Abad-Zapatero
University of Illinois at Chicago
Center for Pharmaceutical Biotechnology.
Lecture no. 4: Summary
This material copyrighted by Cele Abad-Zapatero.
Crystallography:

- Description of symmetry
- Theory of Diffraction
- Structure Determination
- Refinement
Structure determination
Overview

Crystallize

Characterize Crystals
(Characterize the lattice)

Collect the Data
(gather together: h,k,l, Fo, σ)
For all the unique reflections in the crystal)

Determine phases for those reflections

Calculate electron density map
(Fourier equation or elec. Density)

Interpret the map in terms of your polypeptide chain(s) and build a model

Refine the model to optimize the fit to your Experimental values

Validate (Confirm and interpret)
Metaphor of a Diffraction Experiment as a Symphonic Concert

Wind → Orchestra stage

Audience: I, r, s; S

Crystal: h, k, l; F₀

Detector surface

'comb of wind' to change color

sound waves
Protein (macromolecular) crystals when exposed to X-ray they produce a diffraction pattern with thousands of spots.

Review of Waves, Vectors, and Complex Numbers
(The large and the small details: low and high resolution)

Wave with amplitude $F$ and phase angle $\alpha$, and the frequency of oscillation $\omega$.

Vector $F$ in complex plane with modulus $|F|$ and phase angle $\alpha$.

Stout and Jensen (1989)
The collapsing of the incoming wave on the detector
On a crystallographic experiment.

It is as if all the ‘time (or frequency)’ information of the incoming wave would be lost when the wave ‘hits’ the detector. The detector seems to absorb the wave and there is only a ‘record’ how intense the wave was:
Amplitude: $F_0(hkl)$. These are the Fo’s for all the reflections of the data set.
The structure factor equation (from atoms to $F_{hkl}$)

$$F_{hkl} = \sum_{j=1}^{n} f_j e^{2\pi i (hx_j + ky_j + lz_j)}$$

Depends on $\lambda$ and for some $\lambda$ anomalous effects.

Structure factor: contribution of all atoms to a reflection $F_{hkl}$

Contribution of all instruments to sound $S$ at seat labeled by $l$, $r$, $s$
Example for a 5 Atom Structure

For one diffraction spot as an example:-

Five atoms’ contribution to $F_h$ is shown here but many more atoms’ contributions can be vectorially added for a more complicated structure like a protein.
Protein (macromolecular) crystals when exposed to X-ray they produce a diffraction pattern with thousands of spots.

real space: x, y, z
real numbers

reciprocal space: Fhkl, phase complex numbers

real space real numbers

How to go backwards?

From measurements on a diffraction pattern to the molecules/atoms in the unit cell?

Easy if we have good phases for each of the reflections (Fhkl) of the data set.
The electron density equation

Electron density equation: electron density at \( x, y, z \) in unit cell from all reflections (all \( h, k, l \)). Summation of all the people in the audience at the concert to produce ‘density of instruments/unit volume’.

\[
\rho(x, y, z) = \frac{1}{V} \sum_{h} \sum_{k} \sum_{l} |F_{hkl}| e^{-2\pi i (hx + ky + lz - \alpha'_{hkl})}
\]
Protein structure determination (Conceptual notion 1)

- Can be compared to reconstructing a complex spatial pattern using simpler component parts.

- The parts should be added together using the same scale and origin.
Protein structure determination (Conceptual notion 2)

- The parts in your toolkit are of different sizes, varying from:
  - Large (low resolution): major features
  - Small (high resolution): finer detail.

- The parts should be added together using the same scale and origin.
Protein structure determination (Conceptual notion 3)

- The parts are not physical parts but ‘waves’ of different sizes, frequencies and phases

\[ y = A \cdot \sin(\omega t + \phi) \]

- \( A \): Amplitude of the wave
- \( \omega \): frequency of ‘waving’ (oscillation)
- \( \phi \): individual phase.
Protein structure determination
(Conceptual notion 4)

- It is like composing a symphony (or major piece of music) using the sounds (waves) of each instrument properly in phase with the rest: violin, violas, cellos, flutes..... or better:
- Reconstructing the position & movement of the instruments on a concert, based on which each and every person in the concert-hall did hear at the concert.
A mathematical insight

Given a complex function, how can you approximate its value using well-behaved analytical functions?

J.B.J. Fourier (1768-1830) French mathematician, showed that any periodic function $\psi(x)$ can be expressed as a sum of other simpler functions as:

$$\psi(x) = \frac{a_0}{2} + a_1 \cos \alpha + a_2 \cos 2\alpha + a_3 \cos 3\alpha + b_1 \sin \alpha + b_2 \sin 2\alpha + b_3 \sin 3\alpha + \ldots$$

$$\alpha = 2\pi(x/c);$$
If the phases are known, everything is simple and straightforward.

\[
\rho(x,y,z) = \sum \sum \sum |F_{hkl}| \cdot e^{-2\pi i (hx + ky + lz - \alpha)}
\]

A simple summation of terms:

\(h,k,l\): all the reflections in the data set with the corresponding phase for that reflection.
Wave with amplitude $F$ and phase angle $\alpha$, and the frequency of oscillation $\omega$.

Vector $F$ in complex plane with modulus $|F|$ and phase angle $\alpha$.

Stout and Jensen (1989)
A Simple Example of a Fourier Series

Any periodic structure can be described by a periodic function: a series of sine and cosine terms.

This applies to the simple one-dimensional, periodic step function shown here, as well as the complicated three-dimensional electron density of molecules in a crystal.

Experiment with the Fourier applet at http://www.falstad.com/fourier/
Breaking a square wave into components:

\[ s = a + b + c + d \]

- (a) \( \sin(\omega t) \)
- (b) \( \frac{1}{3}\sin(3\omega t) \)
- (c) \( \frac{1}{5}\sin(5\omega t) \)
- (d) \( \frac{1}{7}\sin(7\omega t) \)
The electron density equation

Electron density equation: electron density at \(x, y, z\) in unit cell from all reflections (all \(h, k, l\)). Summation of all the people in the audience at the concert to produce ‘density of instruments/unit volume’.

\[
\rho(x, y, z) = \frac{1}{V} \sum_h \sum_k \sum_l |F_{hkl}| e^{-2\pi i (hx + ky + lz - \alpha'_{hkl})}
\]

Periodic function: Crystal repeats the unit Cell

Fourier components of El. Density: Low, medium, high resolution terms.
Relative Importance of Intensities and Phases

Fourier transform of duck amplitudes with cat phases yields a cat!

Explore the Fourier site at http://www.ysbl.york.ac.uk/~cowtan/fourier/
Making a square wave from Wave functions

\[ s = a + b + c + d \]

(a) \( \sin (\omega t) \)  
(b) \( \frac{1}{3} \sin (3\omega t) \)  
(c) \( \frac{1}{5} \sin (5\omega t) \)  
(d) \( \frac{1}{7} \sin (7\omega t) \)

a: low resolution (overall size/shape)  
b: medium size features (arms, legs)  
c: fine detail: nose, fingers  
d: eyebrows, finger nails
In our Orchestra-Concert metaphor

Determining the structure will be like locating all the instruments (i.e. atoms) on the orchestra (i.e. asymmetric unit of our hypothetical crystal) and their motions.
MAKING-UP A CRYSTAL: The motif; i.e. The asymmetric unit
A 2-D view of our hypothetical crystal
Metaphor of a Diffraction Experiment as a Symphonic Concert

Wind → Orchestra stage

 Audience

 l, r, s; S

sound waves

'comb of wind' to change color

crystal: h, k, l; F_o,

Detector surface
A critical concept: RESOLUTION OF THE DATA

Radius of the sphere that contains all the data (max res)

Data shell

h=-1: main floor
h= 0: mezzanine
h= 1: 1st balcony
h= 2: 2nd balcony

Diffraction limit

weak spots at the edge

Stage of imaginary auditorium

shadow of the beam stop to trap the original x-ray beam
## An inspiring metaphor (parallel)

<table>
<thead>
<tr>
<th>Diffraction Experiment</th>
<th>Symphonic concert</th>
</tr>
</thead>
<tbody>
<tr>
<td>X-rays</td>
<td>wind behind stage</td>
</tr>
<tr>
<td>Crystal</td>
<td>stage with musical instruments repeated by symmetry</td>
</tr>
<tr>
<td>Crystal rotation</td>
<td>stage rotation</td>
</tr>
<tr>
<td>Detector</td>
<td>audience</td>
</tr>
<tr>
<td>Diffraction spot at h,k,l</td>
<td>person at position l, r, s in the audience</td>
</tr>
<tr>
<td>Measure I (Intensity of spot)</td>
<td>Measure Intensity. of sound ($S$)</td>
</tr>
<tr>
<td>$I(h,k,l)$ (modulus)</td>
<td>$S(l,r,s)$ (modulus)</td>
</tr>
</tbody>
</table>
Structure determination
Overview

Crystallize

Characterize Crystals
(Characterize the lattice)

Collect the Data
(gather together: h, k, l, Fo, σ)
For all the unique reflections in the crystal)

Determine phases for those reflections

Calculate electron density map
(Fourier equation or elec. Density)

Interpret the map in terms of your polypeptide chain(s) and build a model

Refine the model to optimize the fit to your Experimental values

Validate (Confirm)
Refinement: Fo vs. Fc’s

\[ F_{hkl} = \sum_{j=1}^{n} f_j e^{2\pi i (hx_j + ky_j + lz_j)} \]

Structure factor: contribution of all atoms to reflection \( F_{hkl} \)

Contribution of all instruments to sound \( S \) at seat labeled by \( l, r, s \)

The structure factors calculated from all the atoms in your structure (\( Fc(hkl) \)) are compared with the ones measured from your diffraction experiment \( Fo(hkl) \) to see how they agree:

\[ R = \frac{\sum (hkl) |Fo| - \sum (hkl) |Fc|}{\sum (hkl) |Fo|} \]

Often given as percent
Refinement

Permits modeled atoms to move in order to minimize the $|F_o - F_c|$ differences.

All structures are refined to some degree before publication. The success of refinement depends upon the resolution of the data: a low resolution structure (e.g., 4 Å) will not refine well, whereas a high resolution structure (e.g., better than 3 Å) has many more observed data points ($F_o$s), and should converge very nicely. The higher the resolution, the more confidence in the correctness of the structure. This is due to the degree of overdeterminacy of the calculation: the ratio of the number of observations to the number of variables.

In refinement, the changes in atomic positions must obey the restraints of idealized bond lengths and angles (see Restraints).

Residual: The “$R$-factor” is the overall indication of how well the theoretical data calculated from the model ($F_c$) agree with the observed data ($F_o$):

$$R = \frac{\sum |F_o - F_c|}{\sum F_o} \times 100$$

The $R$-factor is usually presented as a percent (as above). The values to expect are as follows:

<table>
<thead>
<tr>
<th>Model Status</th>
<th>$R$-factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random, totally incorrect structure</td>
<td>~59%</td>
</tr>
<tr>
<td>Unrefined structure just solved by MIR or MR</td>
<td>40–55%</td>
</tr>
<tr>
<td>Excellently refined and rebuilt high resolution structure</td>
<td>&lt;20%</td>
</tr>
<tr>
<td>Perfect structure and perfect data</td>
<td>0%</td>
</tr>
</tbody>
</table>

As a coincidental but general rule, the $R$-factor should be approximately 10 times the resolution limit. (i.e., a 1.9 Å structure should have an $R$-factor of ~19% or better).
<table>
<thead>
<tr>
<th>Structure determination</th>
<th>Find position/struc.stage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overview (crystal)</td>
<td>Overview (stage instrum)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Crystallize</th>
<th>Assemble musicians at concert</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Hall. No audience, no detector</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Characterize crystals</th>
<th>Characterize the auditorium</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Characterize the lattice)</td>
<td>seat capacity &amp; spacing betw. seats</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Collect the Data</th>
<th>End of concert ushers collect</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Gather together: $h,k,l, F_o, \sigma$)</td>
<td>$l, r, s, S$ for all attendees concert</td>
</tr>
<tr>
<td>for all the unique reflections in the crystal</td>
<td>for all the unique reflections in the crystal</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Determine phases for those reflections</th>
<th>Determ. Phases of sound waves</th>
</tr>
</thead>
<tbody>
<tr>
<td>need to find phase of $F_o$ vector</td>
<td>to your neighbor’s (vectors)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Calculate electron density map</th>
<th>Add up all the terms together</th>
</tr>
</thead>
<tbody>
<tr>
<td>including the phase of each ref.</td>
<td>including the phase for each seat</td>
</tr>
</tbody>
</table>

| Interpret the map in terms of | Interpret the result in terms of |
| polypeptide chain(s) and build a model | orchestra composition and loc. |

| Refine the model to optimize the fit to your experimental values | adjust the position, details to fit the Intensity you hear |

| Validate (Confirm) | Make sure everything seems OK |
Solving a structure means: ‘Phasing’ Techniques (crystal)

- Using isomorphous differences. Soak heavy atoms (Hg, Pt, etc.) and compare the intensities of the reflections to calculate the phases using the position of the heavy atoms.

- Molecular replacement (faster, simpler if similar structure exists). Critical for SBDD.
- Refinement
Solving a structure means;-
‘Phasing’ Techniques (music stage)

• Using isomorphous ‘concerts’. That is to say. Ask the musicians to repeat the concert including some additional ‘HEAVY’ instruments: i.e. Tuba; gather the data etc. calculate where the Tuba was and synchronize all the waves.

• Molecular replacement (faster, simpler if similar structure exists): i.e. you might know the orchestra (or similar) from other concert.

• Refinement
Molecular Replacement harnesses a known 3D structure

- Molecular replacement is the placement of a known protein structure into a different new crystal form: you already know which object is present in the crystal. VERY COMMON in SBDD.

- Molecular replacement uses a homology 3D model ie where the amino acid sequence identity with a known 3D structure is >40% and places it into the new crystal unit cell. This is the starting point for further final refinement (next step).

- Championed by Michael G. Rossmann in the 60’s. Pioneer the method even though the programs (computers) of the time were not the best. Completely routine nowadays. Takes minutes to ‘solve’ a structure.
Method of Isomorphous replacement

Max Perutz

John Kendrew

To show the applicability of this method for the determination of the first two protein structures: hemoglobin and myoglobin was the crowning achievement of these two scientists: i.e. birth of protein crystallography

HOW DID THEY DO IT?
Addition of one Hg

Same unit cell - *isomorphous*

Intensity changed, position of spot unchanged

Two ‘precession photos’ of spots, slightly displaced for clarity, of the same protein one with Hg bound the other without.
Figure 1.14. Microdensitometer trace across the same row of X-ray reflexions from two different crystals of hemoglobin, one without and the other with mercury atoms attached to the two reactive cysteines. Note the changes in intensity of equivalent reflexions.
Ways to estimate the phase

If we can change each reflection in a known way, and measure the amplitude $|F|$ with and without the change, then we can work out the phase.

We can change the structure factor for each reflection in two ways:

- add a few (heavy) atoms (*isomorphous replacement*)
- change the scattering of some atoms by changing the wavelength (*anomalous scattering*)

$$F(h) = \sum_j f_j \exp(2\pi i h \cdot r_j)$$
Pig 2 Zn insulin
Pb derivative
Diffraction pattern

Figure courtesy of E J Dodson
The Harker Construction for SIR

Follow these steps to graphically solve the phase problem:

1) Draw a circle of radius $|FP|$ centered at the origin.
2) Draw $-FH$ from the origin (both $|FH|$ and $\alpha_H$ are known).
3) Draw a circle of radius $|FPH|$ at the end of the $-FH$ vector.

... the phase solution is at an intersection of the circles!

Major problem with SIR: Which intersection?

There is a twofold ambiguity in the phase solution. The ambiguity is a function of $FP - FH$ ... worst when they are perpendicular, and best (single solution) when they are collinear.
Phase probability for one reflection in a SIR experiment. $F_{\text{best}}$ is the centroid of the distribution. The map calculated with $|F_{\text{best}}|e^{i\alpha_{\text{best}}}$ (or $m|FP|e^{i\alpha_{\text{best}}}$, where $m$ is the figure of merit, $\langle \cos \Delta \alpha \rangle$) has the least error. In this example, $m = 0.29$ implies a $73^\circ$ error in the phase angle.
(a) A 2.6 Å SIR electron-density map with the final $\alpha$-carbon trace of the structure superimposed.
\[ \rho_x = \frac{1}{V} \sum m|FP|\exp(i\alpha_{\text{best}})\exp(-2\pi i \mathbf{h} \cdot \mathbf{x}). \]
(b) A close-up of the map with all atoms of the final structure superimposed.
Note that the map is NOT-interpretable.

Multiple Isomorphous Replacement (MIR)

Because of the phase ambiguity, SIR is usually not good enough to solve the phase problem. The best approach? Use multiple heavy atom derivatives! Their combination should yield an unambiguous phase solution.

The Harker construction for two derivatives: Two or more derivatives should give a unique solution, with only one intersection of all circles.
MIR (cont)

The phase probability for each reflection of the multiple derivatives can be plotted linearly:

The optimum phase is calculated for every $hkl$, and then all the structure factors and phases are combined in a Fourier summation to calculate an electron density map.

Practical problems with MIR:
1) Harsh chemistry: heavy atom reagents often damage protein crystals.
2) Heavy atom reagents often enhance radiation sensitivity.
3) Derivatives are not always isomorphous.
4) Substitution is often incomplete, giving a weak FH signal.
5) The FH signal is low, ~10% of FP signal, and may not be detectable.
6) The electron density near the site of a heavy atom in an MIR map is usually un-interpretable.

MIR jargon: lack of closure errors, most probable vs. 'best' phases, figure of merit, phasing power, etc.
Altering the scattering properties of certain atoms in the protein

This use of multiple wavelengths (at least 2 or more) is called Mult. Anomalous Diff.

- **Selenomethionine** protein production is now a reasonably straightforward technique of protein production.
- One in 57 amino acids, on average, is methionine.
- Intensity changes with wavelength are small but, being all on one crystal, viable to measure.
- Very small errors of non-isomorphism.
MAD with Seleno-methionine

• Avoids the search for isomorphous heavy atom derivatives, which is/was a trial and error process.

• Thus higher throughput of protein crystal structure determination is achievable; even at a genome numbers scale>>> ie the field of structural genomics has arrived!
Seleno-methionine hydroxymethylbilane synthase fluorescence scan measured at SRS 9.5
Seleno-methionine hydroxymethylbilane synthase has 5 Se-met residues. The protein MW is 34 kDa; now known to be near the average MW of proteins based on genome sequencing results.

MAD and modified electron density maps

Figure courtesy of E J Dodson
The end result is an electron density map that can be interpreted.
Molecular Replacement harnesses a known 3D structure

- Molecular replacement is the placement of a known protein structure into a different ie new crystal form: you already know which object is present in the crystal. VERY COMMON in SBDD.

- Molecular replacement uses a homology 3D model ie where the amino acid sequence identity with a known 3D structure is >40% and places it into the new crystal unit cell. This is the starting point for further final refinement (next step).
**Six parameter Search (Rotation and Translation of the known object)**

**VERY IMPORTANT IN SBDD SINCE YOU SEEK STRUCTURES OF THE SAME TARGET WITH MANY DIFFERENT LIGANDS.**

If we can find the rotation and translation that puts the model in the correct position in the crystal cell, THEN we can calculate approx. phases.

Figure courtesy of Prof Eleanor Dodson
When symmetry is present, we only have to find one rotation and translation operator; the other one is given by the symmetry.
This is a summary of the key Protein Crystallography Concepts in a ‘nutshell’.

Electron density interpretation

Refinement

Analysis of a Structural Paper.

Questions?
The collapsing of the incoming wave on the detector on a crystallographic experiment.

X-rays or sound waves

It is as if all the ‘time (or frequency)’ information of the incoming wave would be lost when the wave ‘hits’ the detector. The detector seems to absorb the wave and there is only a ‘record’ how intense the wave was:

Amplitude: $F_0(hkl)$. These are the Fo’s for all the reflections of the data set.
Refinement: \( F_0 \) vs. \( F_c \)’s

\[
F_{hkl} = \sum_{j=1}^{n} f_j e^{2\pi i (hx_j + ky_j + lz_j)}
\]

Structure factor: contribution of all atoms to reflection \( F_{hkl} \)

Contribution of all instruments to sound \( S \) at seat labeled by \( l, r, s \)

The structure factors calculated from all the atoms in your structure (\( F_c(hkl) \)) are compared with the ones measured from your diffraction experiment \( F_0(hkl) \) to see how they agree:

\[
R = \frac{\sum_{hkl}|F_0|-\sum_{hkl}|F_c|}{\sum_{hkl}|F_0|}
\]