Protein Structure Determination
Xray Lecture 7, 8

Anomalous Dispersion
Density modification
Refinement
Heavy atom real-space coordinates can be found from Patterson space coordinates and symmetry.

Harker peaks. \( R \cdot (x,y,z) - (x,y,z) = (u,v,w) \), for all symmetry operators \( R \).

Harker sections are where Harker peaks are found in the Patterson map.

\( F_H \) can be calculated once the coordinates of the heavy atoms are know.

Harker diagrams show how the phases may be determined given the \( F_H \) and the amplitudes before and after adding the heavy atom.
Phase probability distribution

Radii are $F_p$ and $k*F_{ph}$

Width are $\sigma_p$ and $k*\sigma_{ph}$

The red area are the places in Argand space where both $F_p$ and $F_{ph}$-$F_H$ can be
Most probable versus best phase

The degree of overlap in the amplitude rings is a measure of phase probability.

Shaded regions are possible solutions.

The “best phase” for calculating density is the probability-weighted average phase. It may have low probability, even zero!
Figure of merit

*Figure of merit* “m” is a measure of how good the phases are.

C is the “center of mass” of a ring of phase probabilities (probability is the “mass”). Assume the radius of the ring is 1. If the probabilities are sharply distributed, \( m \approx 1 \). If they are distributed widely, \( m \) is smaller.

\[
F_{\text{best}}(hkl) = m F(hkl) e^{-i\alpha_{\text{best}}}
\]

Reverse FT using figure of merit, \( m \), and best phase.

\[
\rho(r) = \sum_h F_{\text{best}}(h) e^{-2\pi h \cdot r} = \sum_h m |F(h)| e^{i\alpha_{\text{best}}} e^{-2\pi h \cdot r}
\]
In class exercise: Figure of Merit

\[ F_p = 5.00 \quad \sigma = 0.5 \]

\[ F_{PH1} = 5.50 \quad \sigma = 0.8 \quad F_{H1} = 2.23 \quad \alpha_{H1} = -63.4^\circ \]

\[ F_{PH2} = 4.50 \quad \sigma = 0.9 \quad F_{H2} = 0.50 \quad \alpha_{H2} = -164^\circ \]

1. Draw three circles separated by vectors \( F_{H1} \) and \( F_{H2} \).
2. Draw circular “error bars” of width \( 2\sigma \).
3. Draw circle plot of \( F_p \) phase probabilities.
4. By eye, estimate the centroid \( c \) of probability.
5. What is the Figure of Merit, \( m \)?
Errors (uncertainties) in amplitude lead to errors in phase calculation.

Expected phase error is expressed using the Figure of Merit.

The “Best” phase is the probability-weighted average of all possible phases.

Best phase and figure of merit are used in the reverse FT to get the electron density map. (initial)
Anomalous dispersion

**Theory**

Innermost, bound electrons scatter with a phase shift, relative to free electrons.

Think of it as delayed scatter.

Only some heavy atoms are anomalous at some wavelengths.

Phase delay $\beta$ shown for an atom with phase 0.

$\Delta F_i$ imaginary part

$\Delta F_r$ real part of anomalous scatter

$\Delta F$ real part of anomalous scatter
Friedel's Law: \( F(hkl) = F^*(-h-k-l) \)

Friedel's Law

- Friedel reflection reflects from the same Bragg planes, opposite side.
- Friedels have same amplitude, opposite sign.

...made to be broken...
Anomalous dispersion violates Friedel's Law

We call one Friedel the "plus" and the other the "minus" depending on the sign of $h$.

Phase delay is always a positive angle whether it's the + or the - reflection. So Freidel’s Law is now broken.
Selenomethionine (sMet) is methionine with the Sulfur replaced by a Selenium.

Selenium scatters anomalously at 0.98Å, and normally at 1.54Å.

In Multiwavelength Anomalous Dispersion (MAD) we collect data on one crystal, two wavelengths.

Result is 3 data sets from one crystal

Datasets: $F_{\lambda_1}$, $F_{\lambda_1}^{+\lambda_2}$, $F_{\lambda_1}^{-\lambda_2}$
MAD vector math

$\lambda_1$ is a wavelength where Se does not absorb, normal diffraction.

$\lambda_2$ is a wavelength where Se does absorb, anomalous $\Delta F$ is added.

\[
\begin{align*}
F_{-\lambda_1}(h k l) & \quad \text{Friedel mates collected at non-anomalous wavelength (equivalent)} \\
F_{+\lambda_1}(h k l) & \\
F_{-\lambda_2}(h k l) & \quad \text{Friedel mates collected anomalous wavelength (not equivalent)} \\
F_{+\lambda_2}(h k l)
\end{align*}
\]
After adding anomalous, vector addition is not the mirror image.
Flip $F^{-\lambda_2}$ to $F^{-\lambda_2}^*$

\[
F^{+\lambda_2} = F^{+\lambda_1} + \Delta F^+
\]
\[
F^{-\lambda_2} = F^{-\lambda_1} + \Delta F^{-*}
\]

now we’re solving for the same vector
switch order of addition (offset circles)

\[ F_{-\lambda_2}^- = \Delta F^- + F_{+\lambda_1}^+ \]
\[ F_{+\lambda_2} = \Delta F^+ + F_{+\lambda_1} \]
complete Harker diagram for MAD

$\lambda_1(h \ k \ l)$

$\lambda_2(h \ k \ l)$

$\Delta F_*$

$F_{\lambda_1}(h \ k \ l)$

$F_{\lambda_2}(h \ k \ l)$

$F_{\lambda_2}(h \ k \ l)^*$

Argand
Exercise 5: Solve a MAD Harker

Use compass, ruler, graph, colored pencils. Do the vector addition.

measured intensities

\[ F_{\lambda_1} = 12.0 \quad \alpha_{\lambda_1} = \text{???} \]
\[ F^{-\lambda_2} = 12.0 \]
\[ F^{+\lambda_2} = 15.0 \]

calculated heavy atom structure factors

\[ \Delta F_{H,r} = 3.0 \quad \alpha_{Hr} = 120^\circ \]
\[ \Delta F_{H,i} = 2.0 \]

ANSWER: \( \alpha_{\lambda_1} = 58^\circ \)
Solving MAD using math

F^-_{\lambda_2}(h k l)^* = F^+_{\lambda_1}(h k l) + \Delta F^-(h k l)

= |F^-_{\lambda_2}(h k l)|e^{-i\alpha_2} = |F^+_{\lambda_1}(h k l)|e^{i\alpha_1} + \Delta F^-(h k l)e^{i(\alpha_h-\beta_h)}

|F^-_{\lambda_2}(h k l)|\cos(\alpha_2^-) = |F^+_{\lambda_1}(h k l)|\cos(\alpha_1) + |\Delta F^-(h k l)|\cos(\alpha_h-\beta_h)

-|F^-_{\lambda_2}(h k l)|\sin(\alpha_2^-) = |F^+_{\lambda_1}(h k l)|\sin(\alpha_1) + |\Delta F^-(h k l)|\sin(\alpha_h-\beta_h)

F^+_{\lambda_2}(h k l) = F^+_{\lambda_1}(h k l) + \Delta F^+(h k l)

= |F^+_{\lambda_2}(h k l)|e^{i\alpha_2^+} = |F^+_{\lambda_1}(h k l)|e^{i\alpha_1} + \Delta F^+(h k l)e^{i(\alpha_h+\beta_h)}

|F^-_{\lambda_2}(h k l)|\cos(\alpha_2^+) = |F^+_{\lambda_1}(h k l)|\cos(\alpha_1) + |\Delta F^-(h k l)|\cos(\alpha_h+\beta_h)

|F^-_{\lambda_2}(h k l)|\sin(\alpha_2^+) = |F^+_{\lambda_1}(h k l)|\sin(\alpha_1) + |\Delta F^-(h k l)|\sin(\alpha_h+\beta_h)

Four equations, Three unknowns.

**Knowns:** all three amplitudes $|F|$l, phase and amplitude of real part $\alpha_h$, phase and amplitude of imaginary part $\beta_h$

**Unknowns:** phases $\alpha_1, \alpha_2^-, \alpha_2^+$
Phase probability distribution

\[ F_{\lambda_1}(h k l) - F_{\lambda_2}(h k l)* \]

Argand
F's that are both Syms AND Friedels are centric reflections

Example: $F(0 \ k \ l)$ and $F(0 \ -k \ -l)$ when $a$ is a 2-fold are centric

Draw any set of Bragg planes parallel to the 2-fold. Project the density onto a line. Notice: The projected density is centrosymmetric.

Phase can only be 0 or 180°.
Why Centric reflections are so useful

• All **centric** reflections have real phase = 0° or 180°

Therefore for **centric reflections**:

\[ |F_{ph}| = |F_p| \pm |F_h| \]

*is exact*. Not approximate.

Plus, no phase error for centric reflections.

± = + if the phase of Fp and Fh are both 0, or both 180,
± = – otherwise.

If \( |F_{ph}| < |F_p| \) and the phase of Fh is -180°, what is the phase of Fp?
180°? or -180°?
What if If \( |F_{ph}| > |F_p| \)?
Summary: Heavy atom phasing methods

**SIR** = single isomorphous replacement, without anomalous.

Fourier transform uses: Figure-of-merit weighted amplitudes, alpha-best phases and centric reflections.

**MIR** = multiple isomorphous replacement, without anomalous. Same Fourier terms, but Figure-of-merit is generally better than for SIR.
Summary: Heavy atom phasing methods

**SAD** = single wavelength anomalous dispersion. Phases from F+ and F- from one crystal.

Fourier transform uses: Figure-of-merit weighted amplitudes, alpha-best phases and centric reflections.

**MAD** = multi-wavelength anomalous dispersion. Phases from three datasets from one crystal at 2 wavelengths (or more). F+, F- at anomalous wavelength, and F at non-anomalous wavelength.
Friedel mates are reflections from the same set of Bragg planes, opposite sides.

Friedels have identical amplitudes, phases with inverted sign (normally).

Anomalous dispersion is when Friedels have non-identical amplitudes.

Some heavy atoms have anomalous at certain wavelengths.

Se-Met is used for MAD phasing. (multiwavelength)

In a MAD Harker, we have 1 circle for the non-anomalous wavelength, 2 circles for the anomalous wavelength, F+ and F-*(complex conjugate of F-).
Is the initial map good enough?

(1) The map is calculated using $\alpha_{\text{best}}$.

(2) The map is contoured and displayed using \{InsightII, MIDAS, XtalView, FRODO, O, ...\}

(3) A “trace” is attempted.
Model building

$e^-$ density cages (1 $\sigma$ contours) displayed using InsightII
Information used to build the first model:

**Density**

\[
\rho(r) = \sum_{h} F_{\text{best}}(h) e^{-2\pi h \cdot r} = \sum_{h} m |F(h)| e^{i\alpha_{\text{best}} e^{-2\pi h \cdot r}}
\]

**Sequence and Stereochemistry**

Models are built initially by identifying characteristic sidechains (by their shape) then tracing forward and backward along the backbone density until all amino acids are in place.

Alpha-carbons can be placed by hand, and numbered, then an automated program will add the other atoms (MaxSprout).
Class exercise:

Tracing an electron density map

sequence: AGDLLEHEIFGMPPAGGA

Can you locate the density above in the sequence?
R-factor: How good is the model?

Calculate $F_{\text{calc}}$'s based on the model.

Compute R-factor

$$R = \frac{\sum_{h} |F_{\text{obs}}(h)| - |F_{\text{calc}}(h)|}{\sum_{h} |F_{\text{obs}}(h)|}$$

Depending on the space group, an R-factor of ~55% would be attainable by scaled random data.

The R-factor must be < ~50%.

Note: It is possible to get a high R-factor for a correct model. What kind of mistake would do this?
What can you do if the phases are not good enough?

1. Collect more heavy atom derivative data
2. Try density modification techniques.

Density modification:

initial phases → Fo’s and (new) phases → Map → Modified map → Fc’s and new phases
Density modification techniques

1. Solvent Flattening

(1) Draw envelope around protein part
(2) Flatten. Set solvent \( \rho \) to \( \langle \rho \rangle \).
(3) Forward transform (from density, to get \( \alpha_{\text{calc}} \))
(4) Reverse transform (using \( F_{\text{obs}} \) and \( \alpha_{\text{calc}} \))
2. Skeletonization

Make the density look like a chain

(1) Calculate map.
(2) Skeletonize it (draw ridge lines)
(3) Prune skeleton so that it is “protein-like”
(4) Back transform the skeleton to get new phases.

Protein-like means: (a) no cycles, (b) no islands
3. Non-crystallographic symmetry

NCS transformation is: \( M r_1 + v = r_2 \)

(1) To find \( M \), use Self Rotation function.

(2) To find \( v \), use Patterson Correlation function.

(3) Calculate map (\( \rho \)) using \( F_{obs} \) and \( \alpha_{calc} \).

(4) Set \( \rho'(r_1) = \rho'(r_2) = <\rho(r_1) + \rho(r_2) > \)

(4) Back transform to get new \( \alpha_{calc} \).

Average NCS-equivalent density

repeat 3-5 till converged
A survey of electron density maps

The early (low-resolution) days:

Before computers, maps were contoured on stacked pieces of plexiglass. A “Richards box” was used to build the model.
At 4-6Å resolution, alpha helices look like sausages.
Medium resolution

~3 Å data is good enough to see the backbone with space in between.
The program BONES traces the density automatically, if the phases are good.
BONES models need to be manually connected and sidechains attached. MaxSPROUT converts a fully connected trace to an all-atom model.
discontinuous density can be resolved by model building and refinement.
Contouring at two density cutoffs sometimes helps
Holes in rings are a good thing

Seeing a hole in a tyrosine or phenylalanine ring is universally accepted as proof of good phases. You need at least 2Å data.
Can you see in stereo?

Try this at home. In 3D, the density is much easier to trace.
New rendering programs

“CONSCRIPT: A program for generating electron density isosurfaces for presentation in protein crystallography.” M. C. Lawrence, P. D. Bourke
Great map: holes in rings
Superior map: Atomicity

Rarely is the data this good. 2 holes in Trp. All atoms separated.
Only small molecule structures look this good.

Atoms are separated down to several contours. Proteins are never this well-ordered. But this is what the density really looks like.
Refinement

• The *gradient* of the R-factor with respect to each atomic position may be calculated.

• Each atom is moved down-hill along the gradient.

• "Restraints" may be imposed.

\[
\frac{dR}{dv_i} = \frac{\sum_h \left| F_{obs}(h) \right| - \left| F_{calc}(h) \right|}{\sum_h \left| F_{obs}(h) \right|}
\]

Coordinates of atom \(i\)

Try working out the derivative of R with respect to the x-coordinate of atom \(i\). Hint: use the cosine part of each F.
$|F_o - F_c|$ w/respect to any coordinate of any atom is a cosine function.

$$\frac{dR}{dx_{99}} = \text{sum of } \frac{d|F_o - F_c|}{dx_{99}} \text{ over all } |F_o - F_c|.$$
What is a restraint?

A restraint is a function of the coordinates that is lowest when the coordinates are “ideal”, and which increases as the coordinates become less ideal.

Stereochemical restraints

- bond lengths
- bond angles
- torsion angles

also...
planar groups
B’s
Calculated phases, observed amplitudes = hybrid F's

- $F_c$'s are calculated from the atomic coordinates
- A new electron density map calculated from the $F_c$'s would only reproduce the model. (of course!)
- Instead we use the observed amplitudes $|F_\text{obs}|$, and the model phases, $\alpha_\text{calc}$.

Hybrid back transform:

$$\rho(\mathbf{r}) = \sum_{\mathbf{h}} \left| F_{\text{obs}}(\mathbf{h}) \right| e^{-i\left(2\pi(\mathbf{h} \cdot \mathbf{r}) + \alpha_{\text{calc}}(\mathbf{h})\right)}$$

Hybrid maps show places where the current model is wrong and needs to be changed.
Difference map: $F_0 - F_c$ amplitudes

The $F_0$ “native” map $\rho(F_0)$ differs from the $F_c$ map $\rho(F_c)$ in places where the model is wrong. So we take the difference. In the difference map:

- Missing atoms? $\Rightarrow \rho(F_0 - F_c) > 0.0$
- Wrongly placed atoms? $\Rightarrow \rho(F_0 - F_c) < 0.0$
- Correctly modeled atoms? $\Rightarrow \rho(F_0 - F_c) = 0.0$

Q: Subtracting densities (real space) is the same as subtracting amplitudes (reciprocal space) and transforming. T or F?
The $F_o$ map plus the difference map is

$F_o$ where the differences are zero (the atoms are correct)

Less than $F_o$ where the model has wrong atoms.

Greater than $F_o$ where the model is missing atoms.

$F_o + (F_o - F_c) = 2F_o - F_c$
The “free R-factor”: cross-validation

The free R-factor is the test set residual, calculated the same as the R-factor, but on the “test set”.

Free R-factor asks “how well does your model predict the data it hasn’t seen?”

\[ R_{\text{free}} = \frac{\sum_{h \in T} |F_{\text{obs}}(h)| - k|F_{\text{calc}}(h)|}{\sum_{h \in T} |F_{\text{obs}}(h)|} \]

Note: the only difference is which \(hkl\) are used to calculate.
Why cross-validate?

Because if you have more parameters than data, you can over-fit the data.

R-factor = 0.000??
Fitting and overfitting

Fit is correct if *additional data*, not used in fitting the curve, fall on the curve.

Low residual in the “*test set*” justifies the fit.

residual is not zero now.
cross-validation

=Measuring the residual on data (the “test set”) that were not used to create the model.

The residual on test data is likely to be small if

\[
\frac{\text{data}}{\text{parameters}} \quad \text{is large.}
\]
parameters versus data

Example:

Papain crystal structure has 25,000 reflections.

Papain has 2000 non-H atoms,

times 4 parameters each \((x, y, z, B)\),

equals 8000 parameters

data/parameters = 25,000/8000 ≈ 3  \(<--  this~is~too~small!\)
**restraints count as “data”**

Bond lengths, angles, etc. are “measurements” that must be fit by the model. The true “residual” should include deviations from ideal bond lengths, angles, etc.

In practice, residual in restraints (e.g. deviations from ideal bond lengths, angles) is very low. This means that restraints are essentially “constraints”.

- bond lengths
- bond angles
- van der Waals
- planar groups
- torsion angles
constraints reduce the number of parameters

Bond lengths, angles, and planar groups may be fixed to their ideal values during refinement ("Torsion angle refinement").

Using constraints, Ser has 3 parameters, Phe 4, and Arg 6.

There are an average 3.5 torsion angles per residue.

Papain has ~700 torsion angle parameters.

\[ \therefore \text{data/parameter}=\frac{25,000}{700} \approx 35 \]
radius of convergence concept

...=How far away from the truth can it be, and still find the truth?

radius of convergence depends on *data* & *method*.

More data = fewer false (local) minima

Better method = one that can overcome local minima
Molecular dynamics w/ Xray refinement increases the radius of convergence

MD samples conformational space while maintaining good geometry (low residual in restraints).

\[ E = (\text{residual of restraints}) + (R\text{-factor}) \]

\[ \frac{dE}{dx_i} \] is calculated for each atom \( i \), then we move \( i \) downhill.

**Random vectors** added, proportional to temperature \( T \).

The *simulated annealing* MD method:

(1) start the simulation “hot”

(2) “cool” slowly, trapping structure in lowest minimum.

“X-plor” Axel Brünger *et al*
Phase bias, and how to fix it.

The model biases the phases.

The effect of phase bias is local to the errors.

To correct a part of the model, we must first remove that part.

An “OMIT MAP” is calculated. The phases for an omit map are derived from a partial model, where some small part has been omitted.
This residue has been removed before calculating $F_c$.

$2F_o - F_c$ density $= F_o + (F_o - F_c) =$ The native map plus the difference map.
Difference maps

Two inhibitor peptides bound to thrombin. The inhibitors were omitted from the $F_c$ calculation. (stereo images)

The final model is submitted to the PDB

Structure Explorer - 1AAJ

Summary Information

Title: not available
Compound: Amicyanin (Apo Form)
Authors: R. C. E. Durley, L. Chen, L. W. Lim, F. S. Mathews
Exp. Method: X-ray Diffraction
Classification: Electron Transport
Source: Paracoccus Denitrificans
Primary Citation: Durley, R., Chen, L., Lim, L. W., Mathews, F. S., Davidson, V. L.: Crystal structure analysis of amicyanin and apoamicyanin from Paracoccus denitrificans at 2.0 Å and 1.8 Å resolution. Protein Sci 2 pp. 739 (1993)

Deposition Date: 09-Apr-1992
Release Date: 31-Oct-1993

Resolution [Å]: 1.80
Space Group: P 21
Unit Cell: a 28.95 b 56.54 c 27.55
angles [°]: alpha 90.00 beta 96.38 gamma 90.00

Polymer Chains: 1AAJ
Atoms: 905
HET groups: HOH

R-Value: 0.155
Residues: 105

Other data commonly reported: total unique reflections, completeness, free R-factor
“The model” means the atomic coordinates.
The first model is built from the electron density by shape recognition.
If the electron density is not interpretable, density modification procedures may be used (solvent flattening, skeletonization)
The R-factor measures the difference between observed and calculated F’s.
R-free is the R-factor calculated on a test set, not used for refinement.
Fourier coefficients for difference maps are Fo-Fc
2Fo-Fc maps are corrected Fo maps.
Refinement may be automatic (least squares) or manual (Coot).
Omit maps may be used to escape from phase bias.
From crystal to data

Indexed film

Intensity, $I_p(hkl) = F^2$

native data: $F_p$

I is always a relative term
- Bigger crystal, higher $I$
- Better crystal, higher $I$
- Longer exposure, higher $I$
- More intense Xrays, higher $I$
From data to **Patterson map**

- **native data**: $F_p$
- **heavy atom data**: $F_{ph}$

Find the *best scale factor*, $w$

Calculate $F_{diff} = w*|F_{ph}| - |F_p|$  

Fourier transform

**difference Patterson map**
From data to phases

native data: $F_p$

heavy atom data: $F_{ph}$

Calculate difference Patterson

Find heavy atom peaks on Harker sections

Solve for heavy atom positions using symmetry

Calculate heavy atom vectors

Estimate phases
From data to model

Collect **native** data: $F_p$

Collect **heavy atom** data: $F_{ph}$

Estimate phases

Calculate $\rho$

Is the map traceable? yes

Trace the map

Refine

density modification?

no
Exercise 6 -- Install software

• Install Phenix

• Install Coot (...if it doesn't come with Phenix)

• On Monday Oct 2 bring computer, start up Phenix and follow along using the files linked to the course web page.