Advanced Macromolecular Structure Determination

Phasing

Tim Grüne

Dept. of Structural Chemistry, University of Göttingen

March 2011

http://shelx.uni-ac.gwdg.de

tg@shelx.uni-ac.gwdg.de
Overview

Substructure Definition and Motivation

Extracting Substructure Data from measured Data

Substructure Solution

Density Modification: applications and usage
Macromolecular Crystallography (in brief)

mosflm, xds, HKL2000, ...
xprep, shelxc, solve, ...
selxd, SnB, HySS, ...
selxe, resolve, pirate, ...
arp/warp, coot, refmac5, phenix, YOU!

Data Integration → Anomalous Differences → Substructure Solution → Phasing & Density Modification → Building & Refinement

$|F^+| - |F^-|$

< 1h → < 1h → days–weeks
The Phase Problem and Experimental Phasing
Motivation: The Phase Problem

Experiment measures: $I(hkl)$ i.e. real numbers
Motivation: The Phase Problem

Experiment measures: \( I(hkl) \) \( i.e. \) real numbers

Model Building requires: \( F(hkl) \) \( i.e. \) complex numbers:

\[
F(hkl) = |F(hkl)|e^{i\phi(hkl)}
\]
Motivation: The Phase Problem

Experiment measures: \( I(hkl) \) \( i.e. \) real numbers

Model Building requires: \( F(hkl) \) \( i.e. \) complex numbers:

\[
F(hkl) = |F(hkl)| e^{i\phi(hkl)}
\]

Connection:

\[
|F(hkl)| = \sqrt{I(hkl)}
\]

\[
\phi(hkl) = ?
\]
**Motivation: The Phase Problem**

The electron density is calculated from the structure factor $F(hkl)$:

$$
\rho(x, y, z) = \frac{1}{V_{\text{cell}}} \sum_{h,k,l} |F(h,k,l)| e^{i\phi(h,k,l)} e^{-2\pi i (hx + ky + lz)}
$$

The phase problem is one of the critical steps in macromolecular crystallography:

A diffraction experiment only delivers the amplitude $|F(hkl)|$, but not the phase $\phi(hkl)$ of the structure factor.

Without phases, a map cannot be calculated and no model can be built.
**Ab initio and Direct Methods**

Small-molecule crystallography: large number of reflections vs. small number of atoms

In small molecule crystallography the phase problem has been overcome by so-called *ab initio* methods:

A structure with not too many atoms (< 2000 non-hydrogen atoms) can be solved from a single data set — provided the resolution is better than 1.2Å (so-called “Sheldrick’s rule” [1]).
Ab initio and Direct Methods

**ab initio Methods:** phase determination directly from amplitudes, without prior knowledge of any atomic positions. Includes direct methods and the Patterson method. The most widely used of all ab initio methods are:

**Direct Methods:** phase determination using probabilistic phase relations — usually the tangent formula (Nobel prize for H. A. Hauptman and J. Karle in Chemistry, 1985).
Macromolecular Phasing

For macromolecular crystal structures, the following (main) methods to obtain initial phase information exist:

1. Anomalous Dispersion (SAD, MAD)
2. Isomorphous Replacement (SIR, MIR)
3. Molecular Replacement (MR)

This lecture concentrates on the first two methods, which can also be combined (SIRAS).
The Substructure

Central to phasing based on both anomalous dispersion and isomorphous replacement is the concept of the substructure.

The steps involve

1. Collect data set(s)
2. Create an artificial substructure data set
3. Determine the substructure coordinates
4. Calculate phases (estimates) for the original data
5. Improve phases (density modification)
What’s a Substructure?

The *Substructure* of a (crystal-) structure are the coordinates of a subset of atoms within the same unit cell.
What’s a Substructure?

The *Substructure* of a (crystal-) structure are the coordinates of a *subset* of atoms within the *same* unit cell.

It can be any part of the actual structure. In the usual sense, *substructure* refers to the *marker atoms* that are used for phasing.
What’s a Substructure?

The *Substructure* of a (crystal-) structure are the coordinates of a *subset* of atoms within the *same* unit cell.

It can be any part of the actual structure. In the usual sense, *substructure* refers to the *marker atoms* that are used for phasing.

A real crystal with the properties of the sub-structure (atom distribution vs. unit cell dimensions) cannot exist: the atoms are too far apart for a stable crystal.
Importance of the Substructure for Phasing

The substructure plays an essential part on the way towards phases for a macromolecular structure:

- The *substructure coordinates* can be determined from intensity data alone, *i.e.* without determining the phases.

- Once the substructure coordinates are known, they can be used to overcome the phase problem, *i.e.* an initial map can be determined from which *model building* starts.
Common Trouble and Pitfalls

Typical problems:

- The substructure can be determined, but still the phases for the whole macromolecule cannot be determined.

- Especially at low resolution: Difficult to tell whether or not the substructure and thus the electron density map are correct.

- The latter problem can sometimes be overcome by complementary information (e.g. MR-SAD)
What about Sheldrick’s Rule?

The Sheldrick’s Rule says that direct methods only work when the data resolution is 1.2 Å or better.

Macromolecules *seldomly* diffract to this resolution.

So why can we still use direct methods to determine the substructure?
Sheldrick’s Rule

Sheldrick’s Rule refers the real structures and the 1.2 Å-limit is about the distance where single atoms can be resolved in the data. The substructure is an artificial crystal and atom distances within the substructure are well above the usual diffraction limits of macromolecules, even for 4 Å data.
Substructure Contribution to Data

The electron density $\rho(x, y, z)$ can be calculated from the (complex) structure factors $F(hkl)$. Inversely the structure factor amplitudes can be calculated once the atom content of the unit cell is known:

$$F(hkl) = \sum_{\text{in unit cell}} \text{atoms } j f_j(\theta_{hkl}) e^{-8\pi^2 U_i \sin^2 \theta_{hkl} \lambda^2} e^{2\pi i (hx_j + ky_j + lz_j)}$$  \hspace{1cm} (2)

- $f_j$ atomic scattering factor: depends on atom type and scattering angle $\theta$. Values tabulated in [2], Tab. 6.1.1.1
- $U_j$ atom displacement parameter, ADP: reflects vibrational motion of atoms
- $2\pi(hx_j + ky_j + lz_j)$ phase shift: relative distance from origin
- $(x_j, y_j, z_j)$: fractional coordinates of $j^{th}$ atom
A Short Break: Fractional Coordinates

Coordinates in a PDB-file are *orthogonal coordinates* and what we are most used to.

*E.g.* the coordinates (4, 7, 10):

4 Å along $x$-axis, 7 Å along $y$-axis, and 10 Å along $z$-axis. The axes are *orthogonal* to each other.
A Short Break: Fractional Coordinates

Fractional coordinates describe a position with respect to the unit cell vectors,

\[ \vec{p} = x\vec{a} + y\vec{b} + z\vec{c} \]

For a point in the unit cell: \(0 \leq x, y, z \leq 1\), hence the name fractional coordinates.

Fractional coordinates simplify many computations and are also used in the Fourier transformations to calculate electron density from structure factors and vice versa.
Back to Substructure Contribution to Data

Example: Composition of two different reflections \((hkl)\) and \((h'k'l')\) of a 5-atoms crystal.

The contribution of each atom to **the intensity** \((e.g.\) the 5\(^{th}\): difference between red and cyan vector\) varies from spot to spot. It cannot be deduced directly from the measured intensities.
Extraction of Substructure Contribution

Experimental determination (i.e. other than by molecular replacement) of the contribution of the substructure to the measured intensities is achieved by

- **Anomalous Dispersion**: Deviation from *Friedel's Law* and comparison of
  \[ |F(hklt)| \text{ with } |F(\bar{h}\bar{k}\bar{l})| \]

- **Isomorphous replacement**: Changing the intensities without changing the unit cell and comparison of
  \[ |F_{\text{normal}}| \text{ with } |F_{\text{altered}}| \]

The alteration is usually achieved by the introduction of some heavy atom type into the crystal, either by soaking or by co-crystallisation.
Anomalous Scattering

The atomic scattering factors $f_j(\theta)$ describe the reaction of an atom’s electrons to X-rays.

They are different for each atom type (C, N, P, ...), but normally they are real numbers and do not vary significantly when changing the wavelength $\lambda$.

If, for one atom type, the wavelength $\lambda$ matches the transition energy of one of the electron shells, anomalous scattering occurs.

This behaviour can be described by splitting $f_j(\theta)$ into three parts:

$$f_j^{\text{anom}} = f_j(\theta) + f'_j(\lambda) + if''_j(\lambda)$$

Near the transition energy, the latter two, $f'$ and $f''$, vary strongly with the wavelength.
Anomalous Scattering

Since the equation for \( F(h, k, l) \) (Eq. 2) is a “simple” sum, one can group it into sub-sums. In the case of SAD and MAD the following “grouping” has turned out to be useful, with its phase diagram on the right (ADP \( e^{-8\pi^2 U_j \sin^2 \theta_{hkl}} \lambda^2 \) omitted for clarity):
Anomalous Scattering

Since the equation for $F(h, k, l)$ (Eq. 2) is a “simple” sum, one can group it into sub-sums. In the case of SAD and MAD the following “grouping” has turned out to be useful, with its phase diagram on the right (ADP $e^{-8\pi^2 U_j \sin^2 \theta_{hkl} \lambda^2}$ omitted for clarity):

$$F(hkl) = \sum_{\text{non-substructure}} f_\mu e^{2\pi i hr_\mu}$$

$$+ \sum_{\text{normal}} f_\nu e^{2\pi i hr_\nu}$$

$$+ \sum_{\text{anomalous}} f'_\nu e^{2\pi i hr_\nu}$$

$$+ i \sum_{\text{anomalous}} f''_\nu e^{2\pi i hr_\nu}$$

Normal presentation, because $f'$ usually negative

$$\text{Im}(F)$$

$$\text{Re}(F)$$
Breakdown of Friedel’s Law

Now compare $F(hkl)$ with $F(\overline{h}\overline{k}\overline{l})$ in the presence of anomalous scattering:

$$F(\overline{h}\overline{k}\overline{l}) = \sum_{\text{substructure}} f_\mu e^{2\pi i(-h)r_\mu} + \sum_{\text{normal}} f_\nu e^{2\pi i(-h)r_\nu} + \sum_{\text{anomalous}} f'_\nu e^{2\pi i(-h)r_\nu} + i \sum_{\text{anomalous}} f''_\nu e^{2\pi i(-h)r_\nu}$$

Friedel’s Law is valid for the $f_\mu$, $f_\nu$, and $f'_\nu$ parts:
Breakdown of Friedel’s Law

Now compare $F(hkl)$ with $F(\bar{h}\bar{k}\bar{l})$ in the presence of anomalous scattering:

\[
F(\bar{h}\bar{k}\bar{l}) = \sum_{\text{non-substructure}} f_\mu e^{2\pi i (\bar{h}r_\mu)} + \sum_{\text{substructure}} f_\nu e^{2\pi i (\bar{h}r_\nu)} + \sum_{\text{normal}} f_\nu e^{2\pi i (-h)r_\nu} + \sum_{\text{anomalous}} f_\nu' e^{2\pi i (-h)r_\nu} + i \sum_{\text{anomalous}} f_\nu'' e^{2\pi i (-h)r_\nu}
\]

The complex contribution of $if_\nu''$ violates Friedel’s Law:

\[
|F^+| \neq |F^-|
\]
How all this helps

Karle (1980) and Hendrickson, Smith, Sheriff (1985) published the following formula:

\[
|F^+|^2 = |F_T|^2 + a|F_A|^2 + b|F_A||F_T| + c|F_A||F_T| \sin \alpha
\]

\[
|F^-|^2 = |F_T|^2 + a|F_A|^2 + b|F_A||F_T| - c|F_A||F_T| \sin \alpha
\]

with \(F_T = F_P + F_A\) the non-anomalous contribution of structure and substructure.

**NB:** \(\alpha(hkl) = \alpha(\bar{h}\bar{k}\bar{l})\), \(|F_A(hkl)| = |F_A(\bar{h}\bar{k}\bar{l})|\), and \(|F_T(hkl)| = |F_T(\bar{h}\bar{k}\bar{l})|\)

\[\ast a = f'' + f'^2, \quad b = 2f', \quad c = 2f''\]
Combining Theory and Experiment

The diffraction experiment measures the Bijvoet pairs $|F^+|$ and $|F^-|$ for many reflections.

In order to “simulate” a small-molecule experiment for the substructure, we must know $|F_A|$.

With the approximation $\frac{1}{2}(|F^+| + |F^-|) \approx |F_T|$ the difference between the two equations above yields

$$|F^+| - |F^-| \approx c|F_A| \sin \alpha$$
Status Quo – A Summary

- Our experiment measures $|F^+| = \sqrt{I(hkl)}$ and $|F^-| = \sqrt{I(\bar{h}\bar{k}\bar{l})}$.

- We are looking for (an estimate) of $|F_A|$ for all measured reflections.
  - These $|F_A|$ mimic a small-molecule data set from the substructure alone.
  - The small-molecule data set can be solved with direct methods, even at moderate resolution.

- With the help of Karle and Hendrickson, Smith, Sheriff, we already derived

$$|F^+| - |F^-| \approx c|F_A| \sin \alpha$$

We are nearly there!
The Factor $c$

Before we know $|F_A|$, we must “get rid of” $c$ and $\sin \alpha$.

$$c = \frac{2f''(\lambda)}{f(\theta(hkl))}$$

can be calculated for every reflection provided $f''(\lambda)$.

During a MAD experiment, $f''(\lambda)$ is usually measured by a fluorescence scan.

To avoid the dependency on $f''(\lambda)$, the program *shelxd* [3] uses *normalised structure factor amplitudes* $|E(hkl)|$ instead of $|F(hkl)|$, which are very common for small-molecule programs.
The Angle $\alpha$

In the case of MAD, multi-wavelength anomalous dispersion, we can calculate $\sin \alpha$ because there is one equation for each wavelength.
The Angle $\alpha$

In the case of MAD, multi-wavelength anomalous dispersion, we can calculate $\sin \alpha$ because there is one equation for each wavelength.

In the case of SAD, the program \texttt{shelxd} approximates $|F_A| \sin(\alpha) \approx |F_A|$.

Why is this justified?

Bijvoet pairs with a strong anomalous difference ($|F^+| - |F^-|$) have greater impact in direct methods. The difference is large, however, when $\alpha$ is close to $90^\circ$ or $270^\circ$, \textit{i.e.} when $\sin(\alpha) \approx \pm 1$. This coarse approximation has proven good enough to solve hundreds or thousands of structures with \texttt{shelxd}.
Fluorescence Scan

In order to find the wavelength with the strongest response of the anomalous scatterer, the values of $f'$ and $f''$ are determined from a fluorescence scan.

In order to get the strongest contrast in the different data sets, MAD experiments collect data at

1. maximum for $f''$ (peak wavelength)
2. minimum for $f'$ (inflection point)
SIR/MIR

Single isomorphous replacement requires two data sets:

1. Macromolecule in absence of heavy metal (Au, Pt, Hg, U, . . .), called the native data set.
   Measures $|F_P|$.
2. Macromolecule in presence of heavy metal, called the derivative data set.
   Measures $|F|$.

$|F_A|$ can then be estimated from the difference

$$|F| - |F_P|$$
SAD and MAD vs. SIR and MIR

**SAD/ MAD**
- $|F_A|$ from $|F^+|$ vs. $|F^-|$ (one data set)
- Signal improved by wavelength selection
- Requires tunable wavelength (synchrotron)
- "multi": one crystal, different wavelengths

**SIR/ MIR**
- $|F_A|$ from native vs. derivative (two data sets)
- Signal improved by heavy atom (large atomic number $Z$)
- Independent from wavelength (inhouse data)
- "multi": several derivatives = several crystals
Isomorphous replacement is hardly used for phasing anymore:

- Incorporation of heavy atom into crystal often alters the unit cell ⇒ non-isomorphism between native and derivative ruins usability of data sets
- Heavy atom in structure ⇒ go to synchrotron and try MAD instead of SIR/MIR
However, one often has a high-resolution data set from a native crystal and one (lower resolution) data set with anomalous signal.

They can be combined into a SIRAS experiment:

- Anomalous data set: SAD
- Anomalous data set as derivative vs. native: SIR

These are two independent sources of phase information and usually work better than either alone.
\( |F_A| \): Substructure Solution with Direct Methods
Direct Methods

Having figured out the values $|F_A|$ from our measured data we are actually pretending having collected a data set from a crystal with exactly the same (large) unit cell as our actual macromolecule but with only very few atoms inside.

We artificially created a small molecule data set.
Normalised Structure Factors

Experience shows that direct methods produce better results if, instead of the normal structure factor $F(hkl)$, the normalised structure factor is used.

The normalised structure factor is calculated as $E(hkl)^2 = \frac{F(hkl)^2/\varepsilon}{\langle F(hkl)^2/\varepsilon \rangle}$

$\varepsilon$ is a statistical constant used for the proper treatment of centric and acentric reflections; the denominator $\langle F(hkl)^2/\varepsilon \rangle$ as is averaged per resolution shell. It is calculated per resolution shell (≈ 20 shells over the whole resolution range).

This reduces the strong fall-off with scattering angle $\theta$. 
Starting Point for Direct Methods: The Sayre Equation

In 1952, Sayre published what now has become known as the Sayre-Equation

\[ F(h) = q \left( \frac{\sin(\theta)}{\lambda} \right) \sum_{h'} F_{h'} F_{h-h'} \]

This equation is exact for an “equal-atom-structure” (like the substructure generally is).

It requires, however, complete data including \( F(000) \), which is hidden by the beam stop, so \textit{per se} the Sayre-equation is not very useful.
Tangent Formula (Karle & Hauptman, 1956)

The Sayre-equation serves to derive the tangent formula

\[
\tan(\phi_h) \approx \frac{\sum_{h'} |E_{h'}E_{h-h'}| \sin(\phi_{h'} + \phi_{h-h'})}{\sum_{h'} |E_{h'}E_{h-h'}| \cos(\phi_{h'} + \phi_{h-h'})}
\]

which relates three reflections \( h, h', \) and \( h - h'. \)

H. A. Karle & J. Hauptman were awarded the Nobel prize in Chemistry in 1985 for their work on the tangent formula.
Solving the Substructure

Direct methods (and in particular *shelx*) do the following:

1. Assign a random phase to each reflection. They will not fulfil the tangent formula.
2. Refine the phases using $|F_A|$ (or rather $|E_A|$) to improve their fit to the tangent formula.
3. Calculate a map, pick the strong peaks (as putative substructure).
4. Calculate phases from the peak coordinates and return to step 2.
Solving the Substructure

Steps 1-4 are repeated *many times*, each time with a new set of random phases.

Every such attempt is evaluated and the best attempt is kept as solution for the substructure.

The substructure is thus found *by chance*. It works even if the data quality is only medium — but may require more trials.

The cycling between steps 2-3, i.e. phase refinement in reciprocal space vs. peak picking in direct space, is called *dual space recycling* [4, section 16.1]
Dual-Space Recycling

Real space:
select atoms

reciprocal space
refine phases

random atoms or (better) atoms consistent with Patterson

phases(map)
from atoms

FFT and peak search

repeat a few 100–1000 times

selection criterion:
Correlation Coefficient between $E_{\text{obs}}$ and $E_{\text{calc}}$

best CC?

No

Yes

keep solution
Phasing the Rest

Once the coordinates of the substructure are found, they serve to calculate the phases for the whole structure, *e.g.* with the Harker Construction. With those phases, an *initial* electron density map for the new structure can be calculated:

\[
\begin{align*}
I(hkl) & \rightarrow |F_A| & (x, y, z)_A & \rightarrow \phi(hkl)_A & \int f\mu & \rightarrow \phi(hkl) \\
\end{align*}
\]

\[
\rho(x, y, z)
\]
Density Modification
Density Modification

Phases from experimental methods are *only coarse approximations* of the real phases.

For SAD or SIR even worse than for MAD or MIR.

Phases need to be further improved before the electron density map has got a quality which allows a model to be built.

This is the task of *density modification*. 
How to Improve Phases

Density modification ("DM") programs improve the experimental phases by adding external knowledge to the data.

Roughly speaking, there are two types of density modification:

- Classical DM
- Statistical DM
Classical D. M.: Solvent Flattening

Models from small-molecule crystals usually describe every atom present in the crystal.

In macromolecular crystals, about 40-70 % of the crystal volume consists of disordered solvent molecules (water, salts, PEG, . . .)

The electron density of disordered molecules is flat, i.e. unobserved.
Solvent Flattening: Procedure

1. Calculate map from experimental phases and observed data
2. Detect solvent region
3. Modify electron density in solvent region, e.g. set to zero or a constant value.
4. Back-transform electron density into phases and amplitudes
   → Improved phases from more reasonable map
5. Make amplitudes consistent with observed data
   → Data consistency, avoid “overfitting”
Classical D. M.: Other Methods

Other than solvent flattening (and solvent flipping), other classical methods include

- Histogram matching: Make the density distribution more similar to what is expected for a macromolecular crystal
- Exploiting non-crystallographic symmetry (NCS): With more than one copy per molecule in the asymmetric unit, the density around and at each molecule should be similar to each other
- Model building
Classical D. M.: Model Building

The most extreme “chemical knowledge” about the phases lies in a model itself.

Therefore, model building, even a partial model, is a very powerful way of density modification.

Often, model building is a separate step (e.g. in the programs buccaneer (K. Cowtan), and arp/warp (V. Lamzin/ A. Perrakis)).

shelxe (G. Sheldrick - poly-ALA only, beta-test version) combines model building and density modification.
Programs for “classical” DM

- DM (K. Cowtan)
- shelxe (G. Sheldrick)
- solomon (J.-P. Abrahams)
- CNS (A. Brunger)
Statistical D. M.

Statistical density modification was first implemented in the program “resolve” (T. Terwilliger) and is now also available via “pirate” (K. Cowtan) (see [5] for an overview).

Statistical density modification is supposed to be less biased towards the experimentally determined phases, and therefore more likely to converge towards the “real” phases.
Pipelines for Structure Solution

There is not *the* one program that works best. If one fails to find a solution, another one might still succeed.

Some pipelines that combine various programs for various tasks (substructure solution, density modification, automated model building), *e.g.* Autorickshaw [6] or Crank [7].
References