Sunday, June 13th, 2010

Macromolecular Phasing with `shelxc/d/e`

CCP4 Workshop

APS Chicago, June 2010

Tim Grüne

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

tg@shelx.uni-ac.gwdg.de

Overview

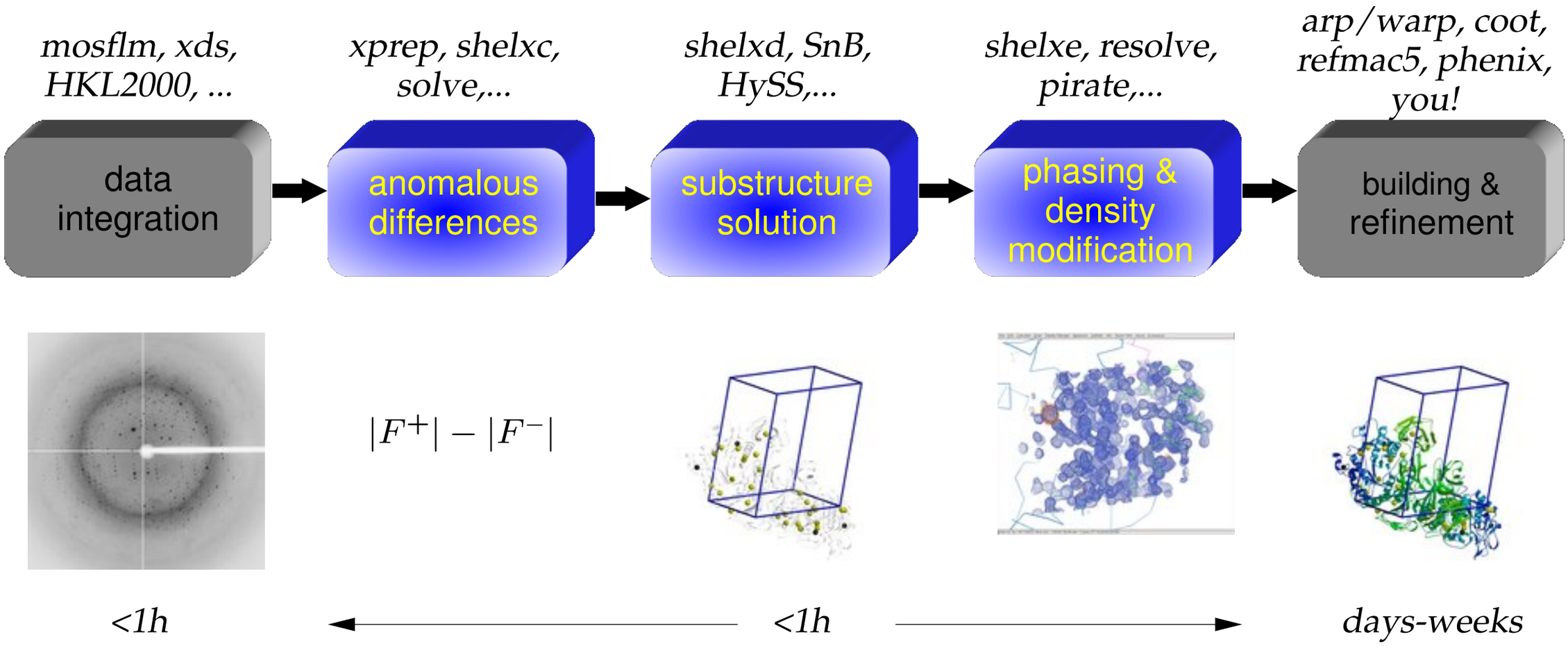
Substructure Definition and Motivation

Extracting Substructure Data from measured Data

Substructure Solution

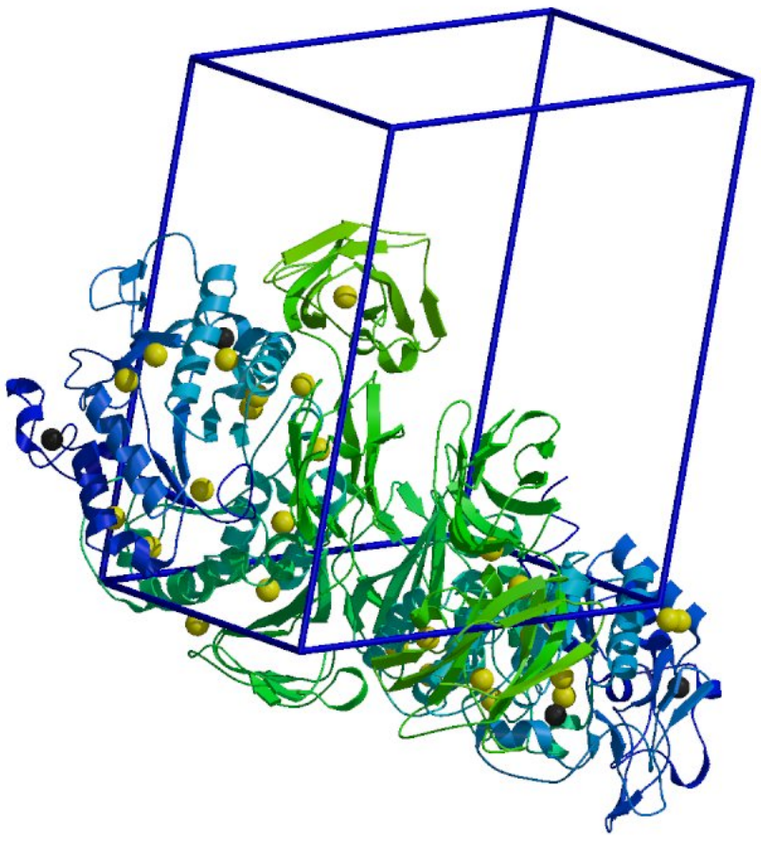
Experimental Considerations

Macromolecular Crystallography (in brief)



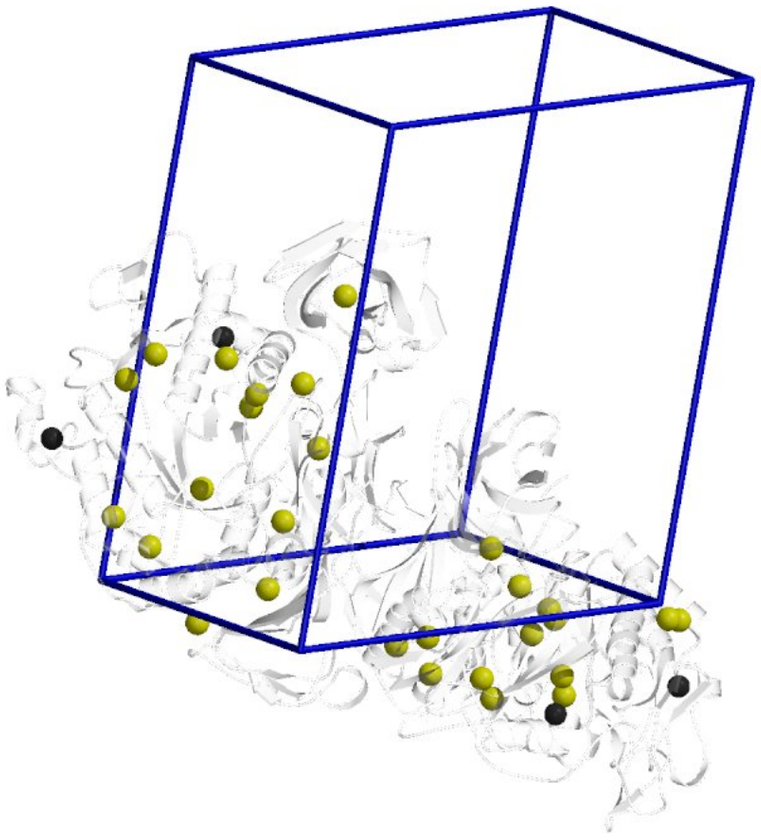
Substructure and the Phase Problem

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 most cases *substructure* refers to the *marker atoms* that are used for phasing.
A "real" substructure crystal cannot exist: the atoms are too far apart for a stable crystal.

Importance of the Substructure for Phasing

- The *substructure coordinates* can be determined by means of *direct methods* from intensities alone by means of experimental phasing experiments (isomorphous replacement, anomalous dispersion).

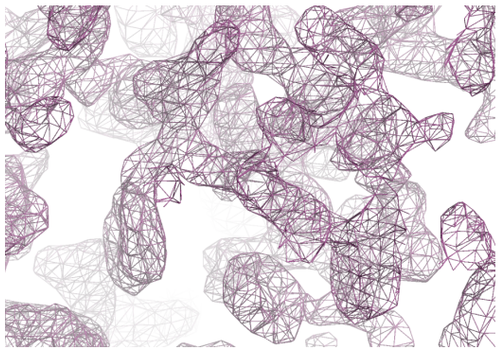
- Once the substructure coordinates are known, they can be used to overcome the phase problem, i.e. an initial map can be determined from which *model building* starts.

Motivation: The Phase Problem (1/2)

The phase problem is one of the major problems in macromolecular crystallography:

A single diffraction experiment delivers the amplitude $\|F\|$, but not the phase ϕ of the structure factors for each measured reflection (hkl).

Therefore we *cannot* directly calculate the electron density



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

Motivation: The Phase Problem (2/2)

In small molecule crystallography the phase problem has been overcome by *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.

***ab initio* Methods:** phase determination directly from amplitudes, without prior knowledge of any atomic positions, including direct methods and the Patterson method.

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

Once we know the substructure, the phases for the reflections of the "real" data can be determined — or at least estimated — to calculate an interpretable electron density map.

Finding Substructure Coordinates with Experimental Methods

Going backwards: Amplitudes from Coordinates

Inversely to the calculation of the electron density $\rho(x, y, z)$ from the structure factors $F(h, k, l)$ by the Fourier transformation (see Eq. 1), we can calculate $F(h, k, l)$ from knowing the positions (x, y, z) and types of all atoms within the unit cell:

$$F(h,k,l) \propto \sum f_j e^{2\pi i(hx+ky+lz)} \tag{2}$$

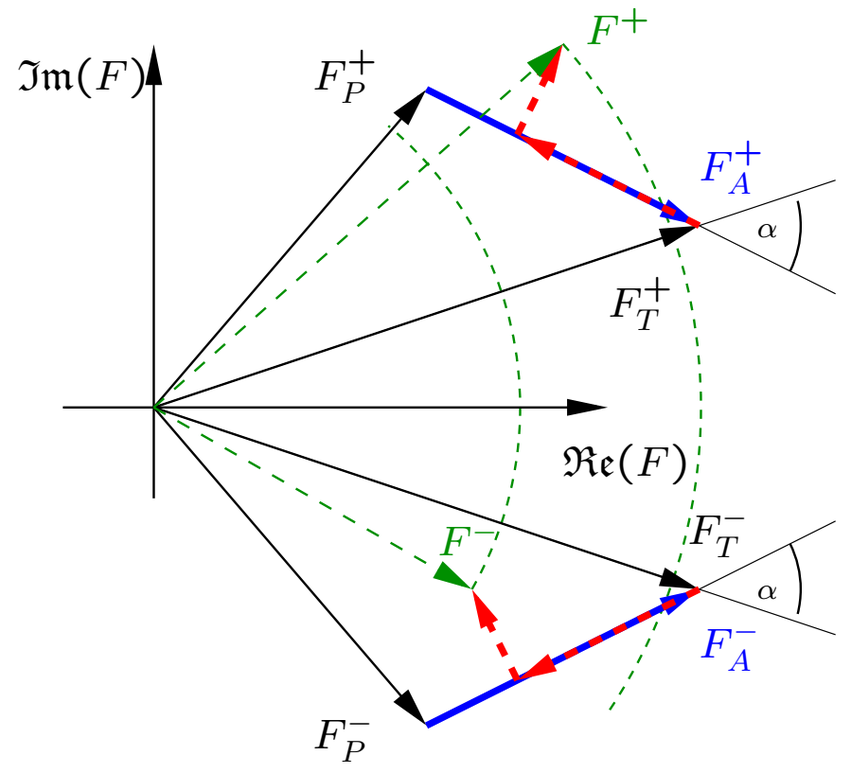
f_j is the atomic scattering factor specific to each atom type (C, N, P, *etc.*). In the presence of anomalous scattering, f_j splits into a "normal" part, only dependent on the scattering angle θ and two "anomalous" parts, only dependent on the wavelength λ:

$$f_j^{\text{anom}} = f_j(\theta) + f_j'(\lambda) + i f_j''(\lambda)$$

The MAD/SAD Phase Diagram

Since the equation for $F(h,k,l)$ (Eq. 2) is a "simple" sum, one can group it into sub-sums. The graphical representation of this "grouping" is the phase diagram on the right. In the case of SAD and MAD the following "grouping" has turned out to be useful:

$$F^{\pm} = \underbrace{\sum_{\text{non-substructure}} f_\mu e^{2\pi i(\pm\mathbf{h})\mathbf{r}_\mu}}_{F_P} + \underbrace{\sum_{\text{substructure}} f_\nu e^{2\pi i(\pm\mathbf{h})\mathbf{r}_\nu}}_{F_A} + \sum_{\text{substructure}} (f'_\nu + i f''_\nu) e^{2\pi i(\pm\mathbf{h})\mathbf{r}_\nu}$$



$$F_T^+ = F_P^+ + F_A^+, \qquad F_T^- = F_P^- + F_A^-$$

The factor if'' causes the breakdown of Friedel's law, *i.e.* $|F^+| \neq |F^-|$.

Simulating a Small Molecule Data Collection

In order to “simulate” a small-molecule experiment for the substructure coordinates, we must know $|F_A|$, the (non-anomalous) contribution of the substructure.

The experiment provides us with the amplitudes of the Bijvoet pairs $|F^+(hkl)|$ and $|F^-(hkl)|$.

The connection between

$|F^\pm|$ (measured Bijvoet pair),
$|F_A|$ (non-anomalous part of Bijvoet pair),
and the angle $\alpha = \phi_T - \phi_A$

is made by a formula derived by Karle (1980) and Hendrickson, Smith, Sheriff (1985)*:

$$|F^\pm|^2 = |F_T|^2 + a|F_A|^2 + b|F_A||F_T| \pm c|F_A||F_T| \sin\alpha \tag{3}$$

* $a = \frac{f''^2+f'^2}{f^2}$, $b = \frac{2f'}{f}$, $c = \frac{2f''}{f}$

Flotsam and Jetsam

If we subtract above equations for $|F^+|^2$ and $|F^-|^2$ from each other and use the approximation $|F^+| + |F^-| \approx 2|F_T|$, the result is

$$|F^+(hkl)| - |F^-(hkl)| \approx c|F_A(hkl)| \sin(\alpha) \tag{4}$$

This approximation holds for each reflection individually.

Remember:

1. Our goal is an estimate for $|F_A|$, the structure factor amplitude for the substructure atoms, so that we can apply direct methods.

2. We know $|F^+| = \sqrt{I(hkl)}$ and $|F^-| = \sqrt{I(\bar{h}\bar{k}\bar{l})}$ directly from the experiment

3. c is wavelength-dependent. By using *normalised* structure factor amplitudes $|E(hkl)|$ instead of $|F(hkl)|$, shelxd becomes *wavelength independent*. Small-molecule programs usually work with normalised structure factors.

The Angle α

Up to the factor $\sin(\alpha)$ we have arrived at an expression that allows us to calculate $|F_A|$ (or rather $|E_A|$) from the difference of the Bijvoet pair $|F^+|$ and $|F^-|$.

In the case of MAD, we can further eliminate this angle because there is one of the above equation (4) for each wavelength.

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

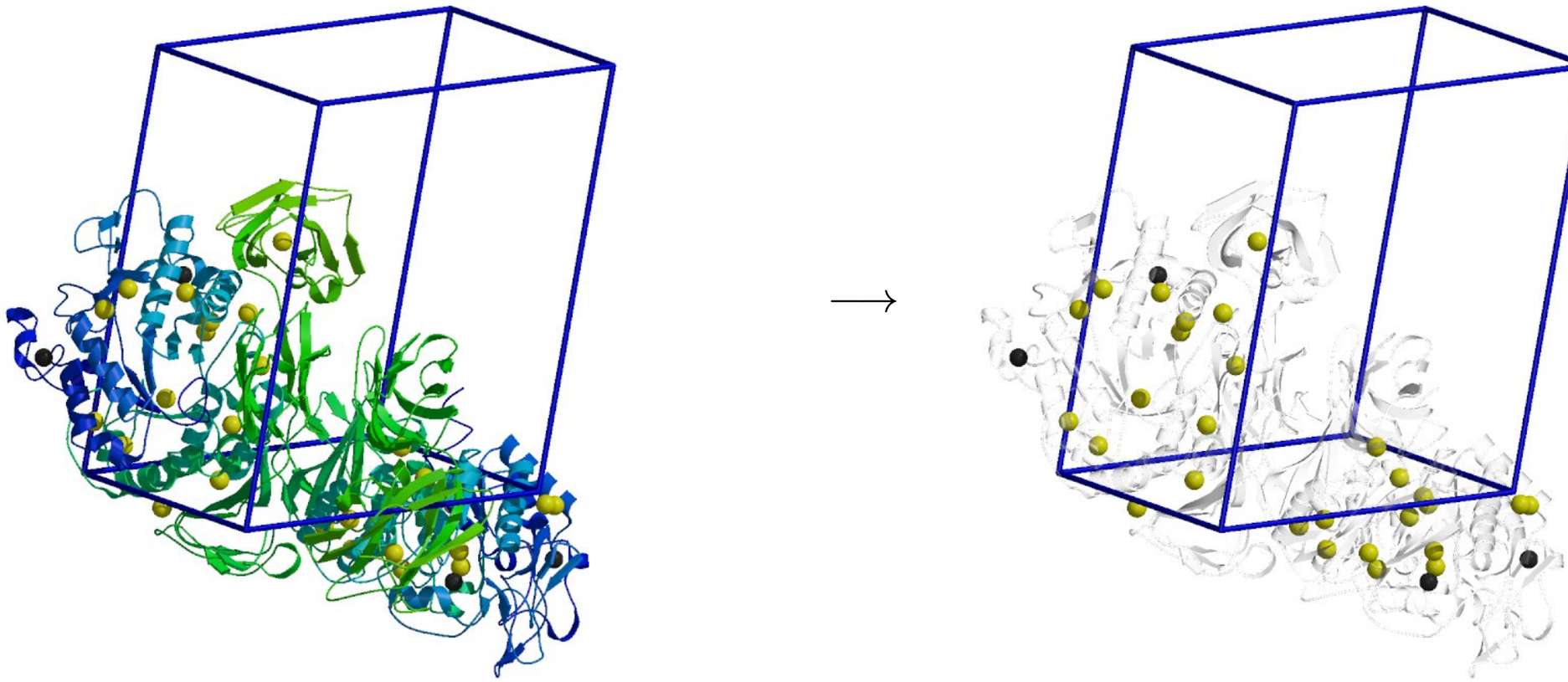
Why is this justified?

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

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**.

Direct Methods

Direct methods have been applied to solve structures with more than 2000 independent non-hydrogen atoms (1gyo). This, however, requires atomic resolution at 1.2 Å and better.

For substructure solution, anomalous data to 2.5–3 Å are usually sufficient and even 5 Å may work.

This is because the distances between atoms of the (hypothetical) substructure crystal are generally quite large, much larger than the data set resolution.

Normalised Structure Factors

Experience has shown 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

$$E(hkl)^2 = \frac{F(hkl)^2/\varepsilon}{\left\langle F(hkl)^2/\varepsilon \right\rangle} \tag{5}$$

It is calculated per resolution shell ($\approx$ 20 shells over the whole resolution range). ε is a statistical constant used for the proper treatment of centric and acentric reflections.

The denominator $\left\langle F(hkl)^2/\varepsilon \right\rangle$ as is averaged per resolution shell.

The normalised structure factor E is independent of the thermal motion (B-factor) and the spatial electron distribution around the atom.

Starting Point: The Sayre Equation

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

$$F(\mathbf{h}) = q(\sin(\theta)/\lambda) \sum_{\mathbf{h}'} F_{\mathbf{h}'} F_{\mathbf{h}-\mathbf{h}'} \tag{6}$$

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.

Tangent Formula (Karle & Hauptman, 1956)

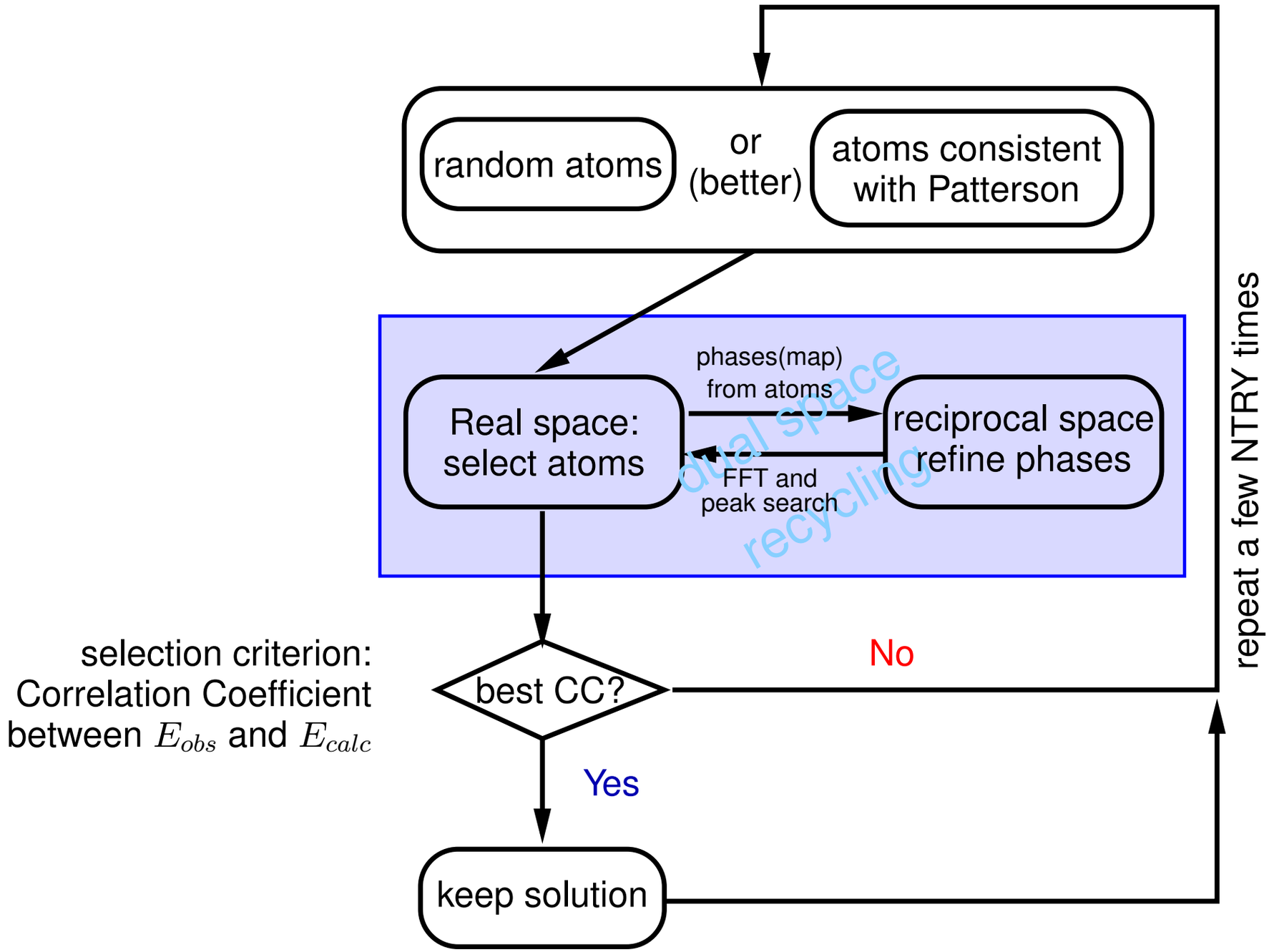
While the Sayre-equation directly is not very useful (because it requires complete data), it serves to derive the tangent formula

$$\tan(\phi_{\mathbf{h}}) \approx \frac{\sum_{\mathbf{h}'} |E_{\mathbf{h}'} E_{\mathbf{h}-\mathbf{h}'}| \sin(\phi_{\mathbf{h}'} + \phi_{\mathbf{h}-\mathbf{h}'})}{\sum_{\mathbf{h}'} |E_{\mathbf{h}'} E_{\mathbf{h}-\mathbf{h}'}| \cos(\phi_{\mathbf{h}'} + \phi_{\mathbf{h}-\mathbf{h}'})} \tag{7}$$

Direct methods and shelxd in particular do the following:

1. Assign a random phase to each reflection. They will not fulfill the tangent formula.
2. Refine the phases to improve their fit to the tangent formula
3. Calculate a map and pick the strong peaks to go back to the tangent forumla (step 2).

shelxd flow-chart



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

shelxd: The Real Space Part

shelxd provides two methods to improve the quality of the search:

PATS - Patterson Seeding the initial atoms are not chosen completely arbitrarily, but such that they obey the Patterson map of the data. This adds some “chemical” or “geometrical” information to the process and therefore helps to improve the result.

WEED - Random omit maps after real space refinement, 30% of the peak positions in the map are left out from phase calculation and subsequent phase refinement. This is similar to omit maps in model building and refinement and improves the results. The idea behind this is to reduce “model bias” or “over-emphasis” from strong contributors.

PATS is the default for macromolecules. **WEED** is only recommended when the Patterson map cannot be interpreted, which is the case mainly in two cases:

1. too many atoms in the substructure.

2. the Laue group is one of the higher cubic ones.

shelxd Usage

shelxd is started from the command line:

```
#> shelxd name
```

Input		Output	
name.hkl	Anomalous differences $\|\|F^+\| - \|F^-\|\|$ with initial estimates for α	name.res	Substructure coordinates
name.ins	File with Instructions	name.lst	Log file (same as Terminal output)

The word name can only contain letters and numbers. Period '.', and spaces ' ' cannot be used within name.

The .ins-file contains the instructions for shelxd. It is easiest to prepare it the shelxc, xprep, or the GUI hkl2map (which uses shelxc), see tutorial.

Density Modification and Auto-building with shelxe

shelxe — the “eierlegende Wollmilchsau”

shelxe was originally devised as a quick check whether or not the collected data are sufficient for structure solution.

The “standard” phase information for shelxe comes from the substructure coordinates found by shelxd (or sharp, or solve/ HySS, or . . .).

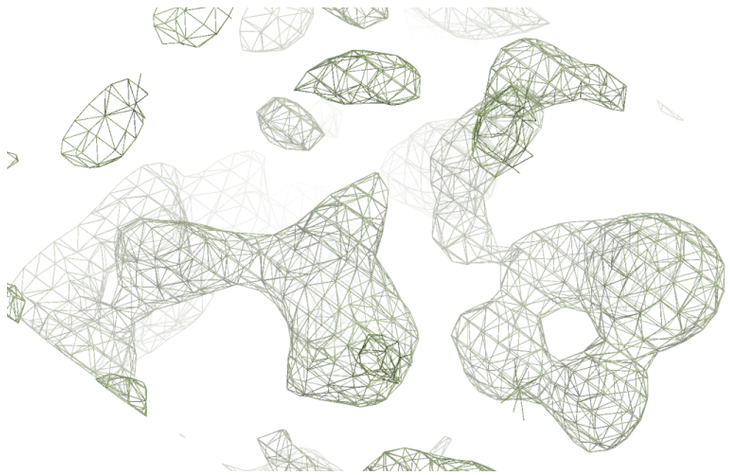
Its methods have proven more powerful than just being a “quick check” and the program has evolved since, including the Free Lunch Algorithm and Backbone Auto-building

shelxe can also combine a (poor) MR-solution with SAD phases — MR-SAD (*Autorickshaw*, www.embl-hamburg.de/auto-rickshaw)

shelxe can also be used to extend an MR-solution when you get stuck — without anomalous data (*Arcimboldo*, http://chango.ibmb.csic.es/ARCIMBOLDO).

Density Modification with shelxe

Input: Phase information of low quality, *e.g.* substructure coordinates from shelxd

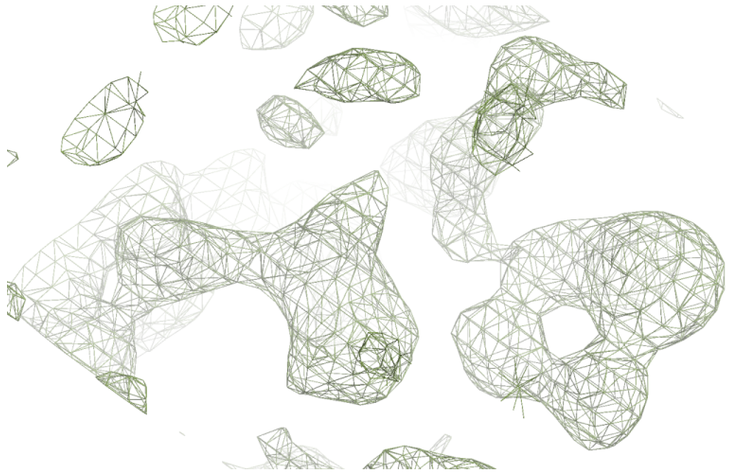


MAD *w/o DM*

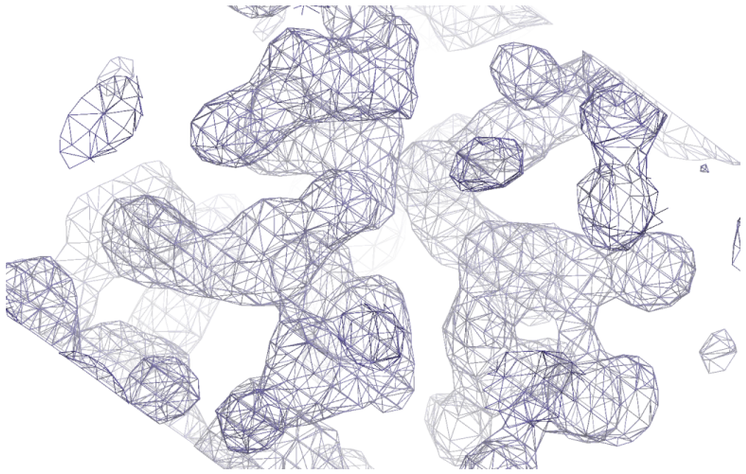
Density Modification with shelxe

Input: Phase information of low quality, *e.g.* substructure coordinates from shelxd

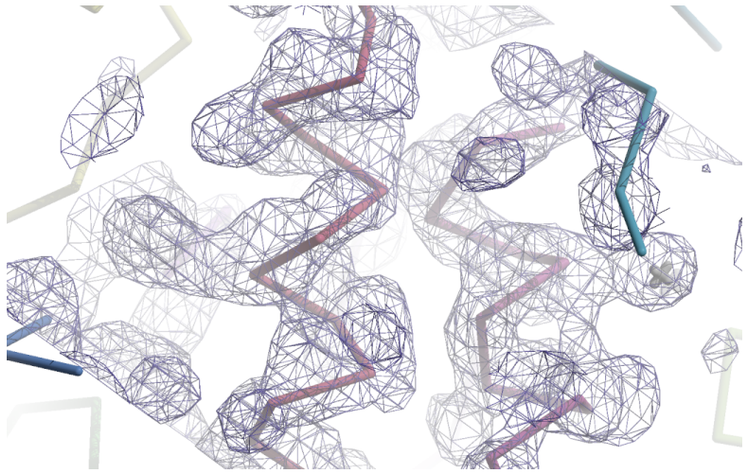
Output: An interpretable electron density map, including a Poly-Ala trace (if applicable).



MAD *w/o DM*



MAD *with DM*



DM with Poly-ALA model

shelxe Usage

Unlike shelxc and shelxd, shelxe does not use an input file with instructions but takes all options from the command line or has reasonable defaults, *e. g.*

```
shelxe tpp phase -h -s0.6 -a
```

Input		Output	
tpp.hkl	"native" intensities	tpp.phs	electron density map
phase.res	substructure coordinates (from shelxd)	tpp.hat	improved substructure coordinates (*cf.* tpp.res, shelxd)
phase.hkl	source of α-estimates (from shelxd)	tpp.pdb	Poly-ALA model (-a)
		tpp.lst	log-File

Input and Output names cannot be changed, they are fixed with the input names (here: tpp and phase).

shelxe Usage

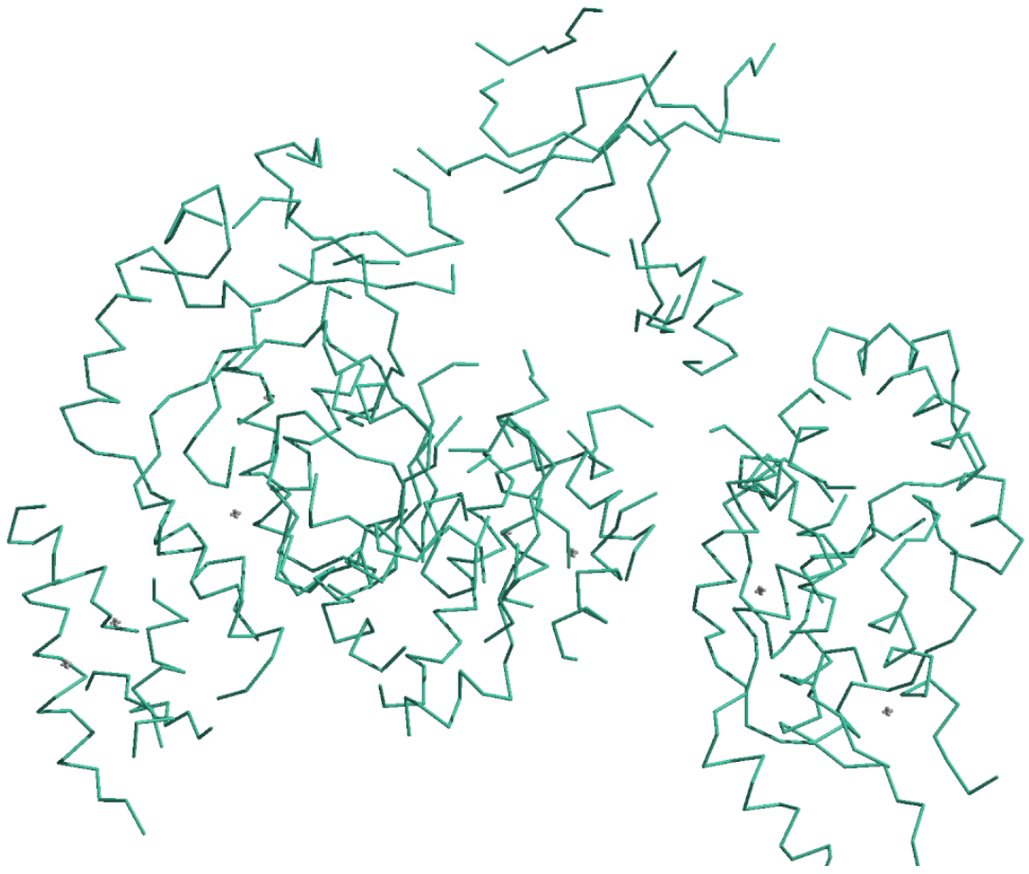
The inverted coordinates of the substructure found by shelxd ($(x, y, z) \longrightarrow (-x, -y, -z)$) is a solution for shelxd as much as the solution put out by shelxd. The program has no way to distinguish between these two.

For shelxe, however, this matters. For this reason shelxe must always be run twice. The inversion is taken care of automatically by shelxe by simply adding the option -i to the command line options. Any output file will have _i appended to its name, so no file will be overwritten.

The first run without -i is referred to as the “original hand”, the second run with -i as the “inverted hand”.

`shelxe` — Indicators of Success

How to tell whether the solution from `shelxe` is correct, and which hand?



The easiest way to see if `shelxe` found the correct solution is available *via* the Poly-ALA chain it builds. It should not be too fragmented and should look like a protein.

This can be controlled with `coot` which reads the `.res` (substructure from `shelxd`) and `.hat` (improved substructure from `shelxe`) files, the `.phs`-file (electron density map) and the `.pdb`-file (Poly-ALA model).

A correct solution has long chains in few fragments.

shelxe — Indicators of Success

The *shelxe-log file* reports the

```
CC for partial structure against native data =  27.31 %
```

A value above 30% is a strong indicator for successful structure solution, values between 20% and 30% are worth further investigation like *recycling of heavy atom positions* (see *e.g.* tutorial).

```
Pseudo-free CC = 66.96 %
```

above 65% can also be considered a success.

These figures stem from the solution of Tripeptidyl-peptidase I (A. Pal, 2009), 2×571 a.a., Se-MAD at 2.35Å. shelxe traced 662 a.a. in 5 chains.

“Extensions” of shelxe

shelxe and Arcimboldo

Arcimboldo combines Molecular Replacement with Density Modification:

- Search small fragments (helices of 14 residues) with phaser
- Extend large number of possible solutions with auto-building in shelxe

shelxe and Arcimboldo

Arcimboldo combines Molecular Replacement with Density Modification:

- Search small fragments (helices of 14 residues) with phaser
- Extend large number of possible solutions with auto-building in shelxe

Arcimboldo is like *ab initio* phasing at 2Å resolution.

- Computationally very expensive
- Use Cluster *via* web-interface http://chango.ibmb.csic.es/ARCIMBOLDO

Learning and Using shelxc/d/e with hkl2map

When you are scared by the command line input or the crude numbers output by shelxc/d/e, a good way of getting familiar with the triad is the GUI hkl2map by Thomas Schneider, http://webapps.embl-hamburg.de/hkl2map.

hkl2map can be used for SAD, SIR, SIRAS, and MAD.

hkl2map shows several graphics which facilitate interpring the shelx output.

Please see the tutorial for more information.

Summary

shelxd and shelxe are fairly robust programs. The most sensitive parameter to shelxd is the high resolution cut-off:

Do not include (anomalous) noise in you substructure noise.

shelxe is even less sensitive. Before or without auto-tracing, the *solvent content* (-s) had to be chosen correctly.

In borderline-cases, consider recycling of the refined heavy atom coordinates.

The best preparation for anomalous phasing: Make sure you carefully collect your data.

The licensed program xprep is the advanced version of shelxc. It applies an improved scaling of anomalous data and there are cases where using xprep instead of shelxc is crucial to structure solution (unfortunately).

References and Further Reading

References from the lecture and further reading:

Drenth, J. (2007). *Principles of Protein X-Ray Crystallography*. Springer.

Pal, A. *et al.* (2009). Structure of Tripeptidyl-peptidase I Provides Insight into the Molecular Basis of Late Infantile Neuronal Ceroid Lipofuscinosis . *Journal of Biological Chemistry*, 284:3976–3984.

Rodríguez, D. D. e. (2009). Crystallographic *ab initio* protein structure solution below atomic resolution. *Nature Methods*, 6(9).

Rupp, B. (2009). *Biomolecular Crystallography: Principles, Practice, and Application to Structural Biology*. Garland Science.

Sheldrick, G. e. (2001). *International Tables for Crystallography*, volume F, chapter 16.1: *Ab initio* phasing. Kluwer Academic Publishers.