Saturday, June 27th, 2009

The Principles of `shelxd` for Macromolecular Phasing

CCP4 Workshop

APS Chicago, June 2009

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

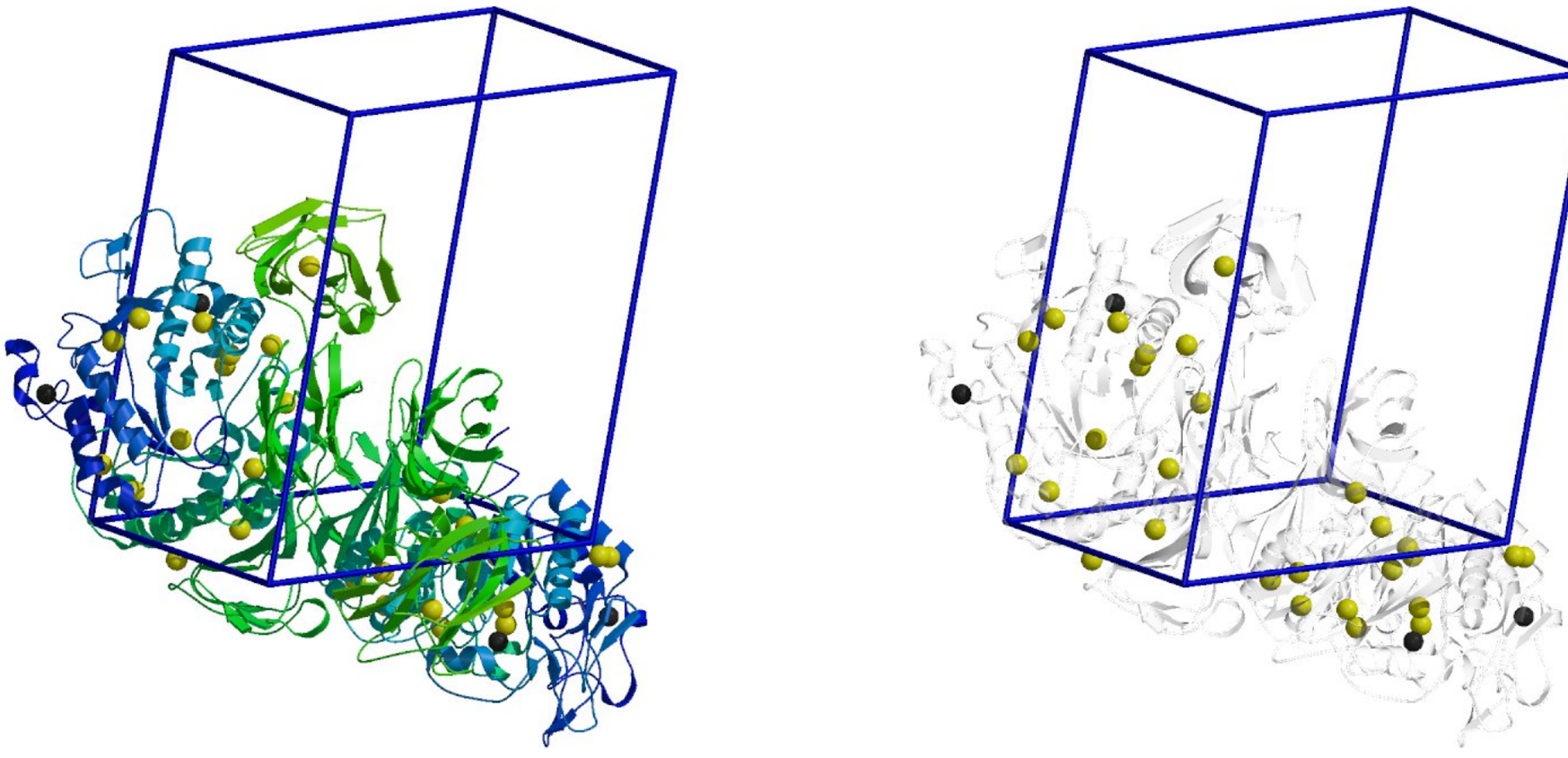
Substructure Solution

Experimental Considerations

Substructure Definition and Motivation

Substructure: What is it?

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



Substructure: Why is it so important?

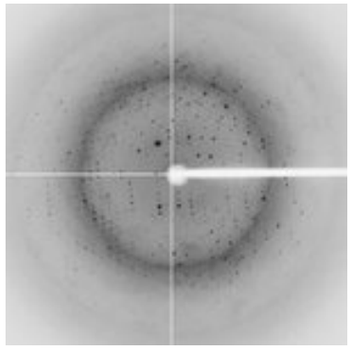
A “real” substructure crystal cannot exist: the atoms are too far apart for a stable crystal.

If we knew the coordinates of the substructure, we could solve the phase problem for the whole structure.

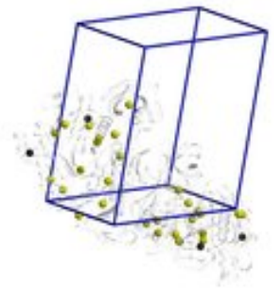
Procedure (Overview)

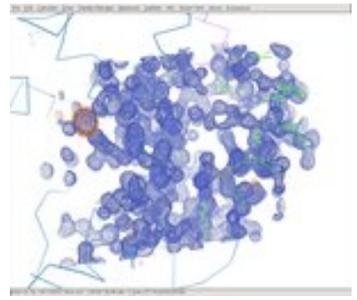
mosflm, xds, ... | *xprep, shelxc, ...* | *shelxd, SnB, ...* | *shelxe, resolve, ...* | *arp/warp, coot, you refmac5, phenix,...*

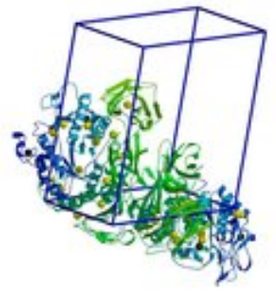
data collection processing → extract substructure data → calc. substructure coord's → all phases & interpretable map → building & refinement



h,k,l, F_A







Motivation: Benefit of “Substructure” for Macromolecules (1/2)

A major problem in macromolecular crystallography is the *phase problem*:

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

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

from the measured data.

Motivation: Benefit of “Substructure” for Macromolecules (2/2)

In small molecule crystallography there is not really a phase problem:

A structure with not too many atoms ($<$ 1000 non-hydrogen atoms) can be solved from a single data set (with so-called *ab initio* methods).

The term *direct methods* encompasses all *ab initio* methods that derive the reflection phases from the amplitudes using probabilistic phase relations — usually the tangent formula.

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

Extracting Substructure Data from measured Data

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 (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:

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

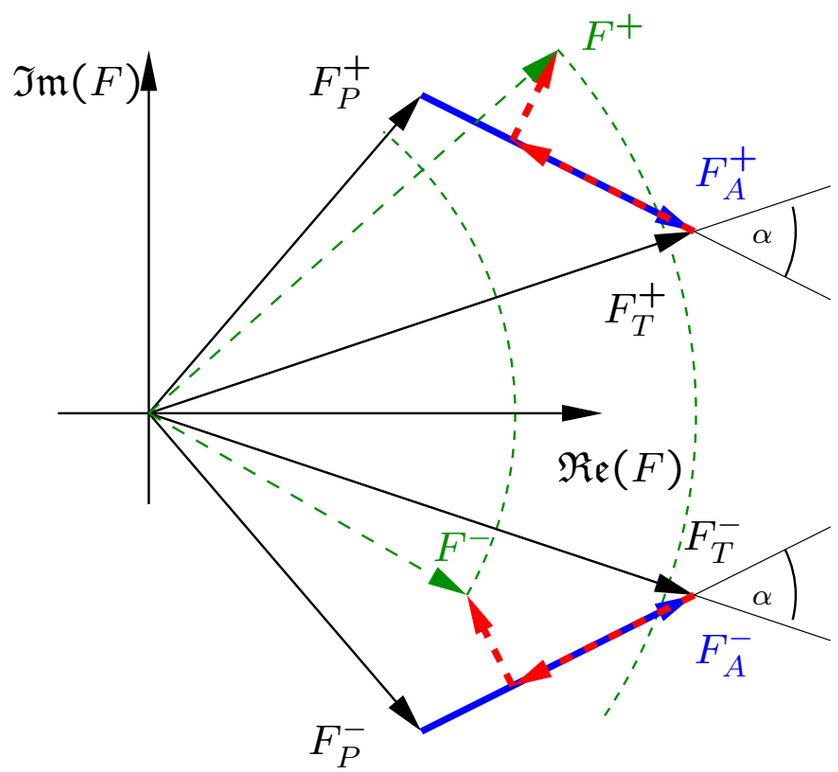
f_j is the atomic scattering factor specific to each atom type ($C, N, P,$ *etc.*). In the presence of anomalous scattering, f_j splits into a "normal" part, only dependent on the scattering angle θ and two "anomalous" parts, only dependent on the wavelength λ:

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

The Phase Diagram for SAD

Since 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 the following “grouping” has turned out to be useful:

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



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

Preparing for F_A

The reason for this rather non-intuitive splitting is a formula derived by Karle (1980) and Hendrickson, Smith, Sheriff (1985).

It puts $|F^{\pm}|$, $|F_A|$, and the angle $\alpha = \phi_T - \phi_A$ into context*:

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

Remember: The experiment provides us (only) with their amplitudes $|F^+|$ and $|F^-|$, but we would like to know F_A, the (non-anomalous) contribution of the subtructure.

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

Finally - the Substructure Amplitude

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

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

This approximation is for each reflection individually.

Remember:

1. Our goal is to know $|F_A|$, the structure factor amplitude for the substructure atoms
2. We know $|F^+|$ and $|F^-|$ directly from the experiment
3. c is just a constant

*more correctly the result should be stated as $|F(\mathbf{h})| - |F(-\mathbf{h})| \approx c|F_A(\mathbf{h})| \sin(\alpha)$, but for lazyness one often omits **h**

The Angle α

Up to the factor $\sin(\alpha)$ we have arrived at an expression that allows us to calculate $|F_A|$ 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 equations for each wavelength.

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

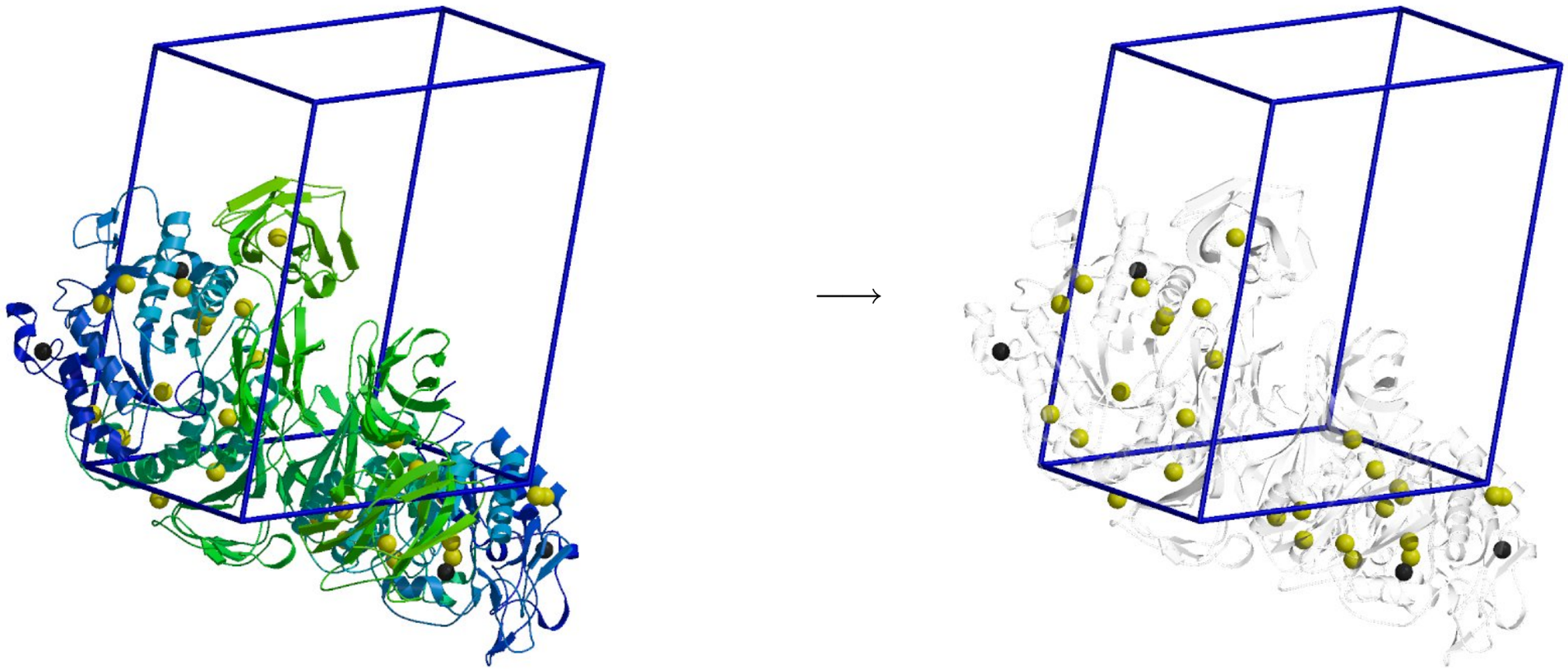
Why is this justified?

The method used to calculate the substructure emphasises strong reflections. In our case it means large differences between the Bijvoet pairs. This difference is large if the angle is close to $90°$ or $270°$, i.e. if $\sin(\alpha) \approx \pm 1$, in which case it is a fairly good approximation - good enough to solve the structure, as plenty of solved structures confirm again and again.

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 dataset**.

Direct Methods

Direct methods have been applied to solve structures with more than 2000 independent non-hydrogen atoms (1gyo). These, however, require 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 distance between atoms of the (hypothetical) substructure crystal is generally quite large, much larger than the data set resolution.

Starting Point: The Sayre Equation

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

$$F(\mathbf{h}) = q(\sin(\theta)/\lambda) \sum_{\mathbf{h}'} F_{\mathbf{h}'} F_{\mathbf{h}-\mathbf{h}'} \tag{5}$$

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 beamstop.

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

$$E(hkl)^2 = \frac{F(hkl)^2/\varepsilon}{\left\langle F(hkl)^2/\varepsilon \right\rangle} \tag{6}$$

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 electron distribution around the atom.

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

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

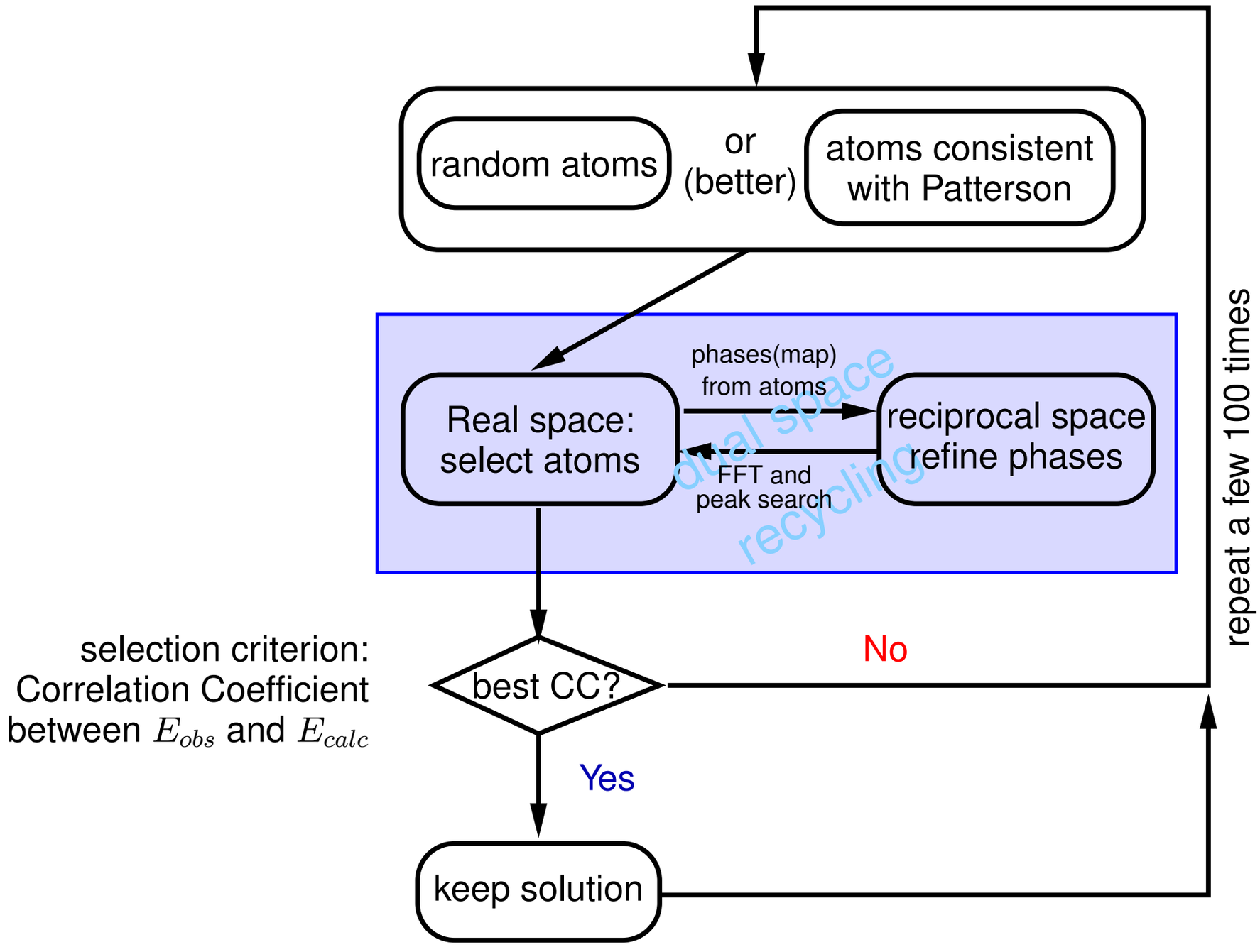
Given a set of *test* phases (one for each reflection), the tangent formula will usually not be fulfilled. By altering the test phases so that they better match the desired ideal values from Eq. 7, one can now refine these phases.

Nothing comes from Nothing

The tangent formula does not provide us with an initial set of phases for the substructure. It only allows to refine phases.

shelxd (and e.g. the program SnB) solves this problem by a statistical approach which is called dual-space recycling.

shelxd flow-chart



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

shelxd: The Real Space Part

shelxd provides two alternative 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. There are mainly two cases for which **WEED** is recommended instead, both of which lead to an uninterpretable Patterson map with too dense peaks to distinguish:

1. too many atoms in the substructure to interprete the Patterson map.

2. the Laue group is one of the higher cubic ones, also resulting in too many atoms to interprete the Patterson map.

Density Modification

Knowing the positions of the substructure atoms we can now calculate their structure factor amplitudes $|F_A|$ ***and*** their phases ϕ_A. If we knew α we would know $\phi_T = \phi_A + \alpha$ and hence could calculate a (starting) electron density map together with the amplitudes F_T.

For MAD, SIRAS, MIR we can calculate α. For SAD and SIR, however, approximations are necessary - again ....

From the phase diagram (slide 10), one can figure out, that if $|F^+| \gg |F^-|$, then $\alpha \approx 90°$ and $\alpha \approx 270°$ if $|F^-| \gg |F^+|$. In the SIR case the angles are $0°$ and $180°$ respectively.

These approximations are poor, usually too poor for model building. They are, however, good enough for density modification which improves these estimates for ϕ_T in several cycles of imposing chemical or statistical information on the reflection.

This is yet another story, though ....

Experimental Considerations

data multiplicity and independence E.g. when collecting more than 360° with 1° framewidth, shift the goniometer by 0.5° degree backwards after the first round to get “fresh”, i.e. statistically independent, data.

proper resolution limit during integration If the resolution limit during data integration is set too far into the noise region, also the data quality at lower resolution will suffer from it. Try to estimate how far your crystal diffracts and only integrate up to that resolution.

proper resolution limit for shelxd shelxd is sensitive to the resolution cut-off. Check (e.g. with xprep) to what resolution an anomalous signal could be detected and try several resolution limits around that value.

References

- Drenth, J. (2007), *Principles of Protein X-Ray Crystallography*, Springer
- G.M. Sheldrick, H.A. Hauptman, C.M. Weeks, R. Miller and I. Usón (2001), Intern. Tables for Crystallography F, *Section 16.1: Ab initio phasing*

Thank you for your attention!