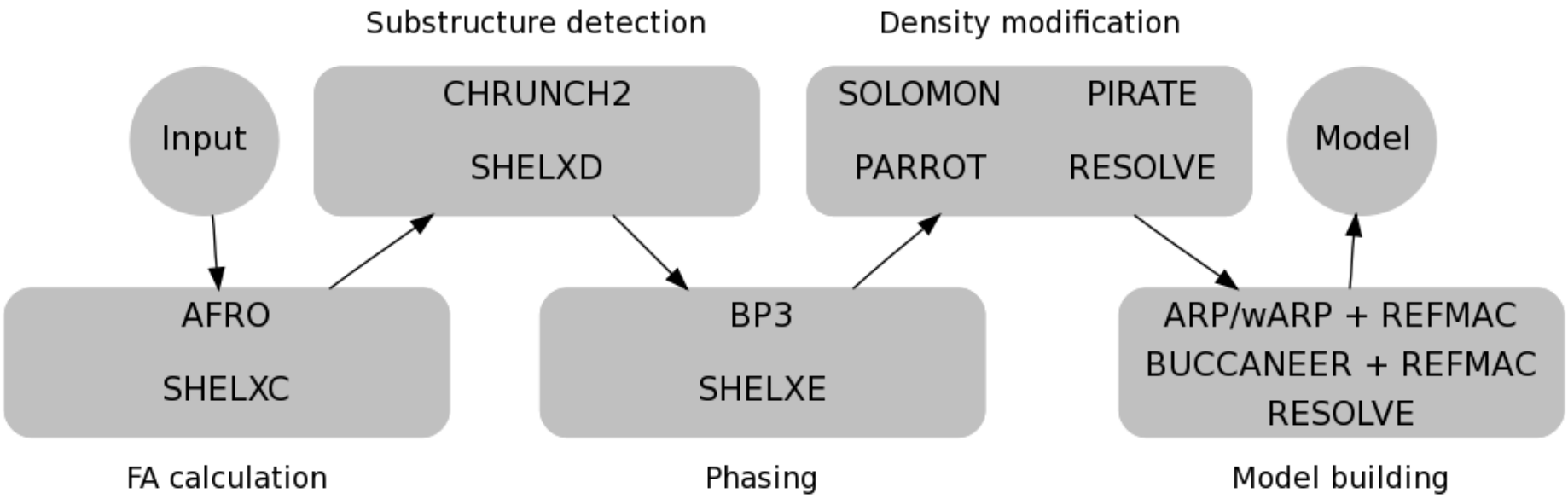
Introduction to experimental phasing and Structure solution with Crank

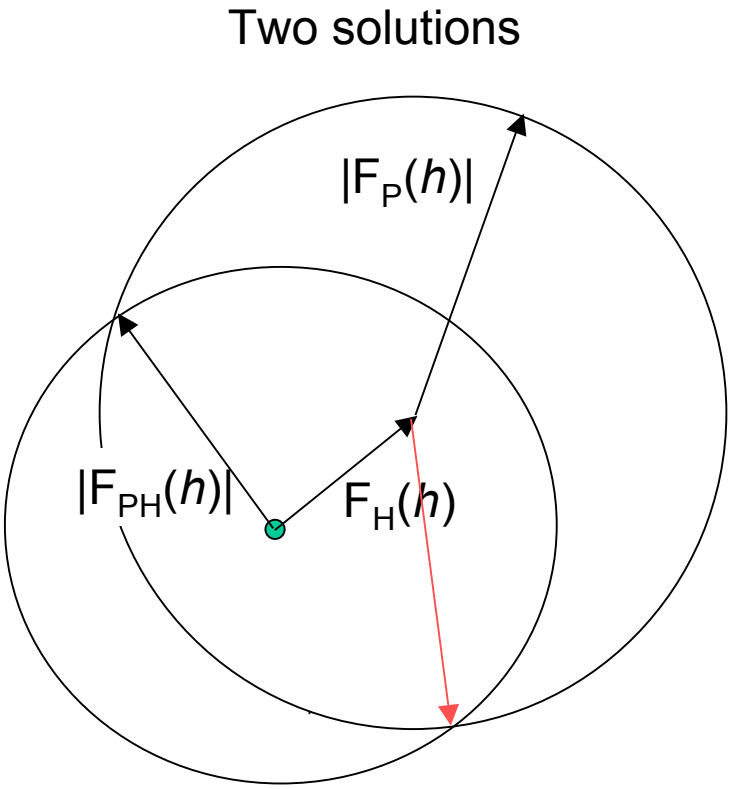
Biophysical Structural Chemistry,
Leiden University, The Netherlands

http://www.bfsc.leidenuniv.nl/software/crank/
http://www.ccp4.ac.uk/

Flow of experimental phasing and programs used within Crank.



Phase information from single isomorphous replacement



$|F_p|$ (native amplitude), $|F_{PH}|$ (derivative amplitude) and F_H (heavy atom amplitude and phase are known).

$F_p + F_H = F_{PH}$

In an ideal, error-free experiment, there are two possible phases.

Solving phases: using the anomalous signal

Anomalous scattering arises due to a phase-shift of X-rays diffracted by tightly bound electrons

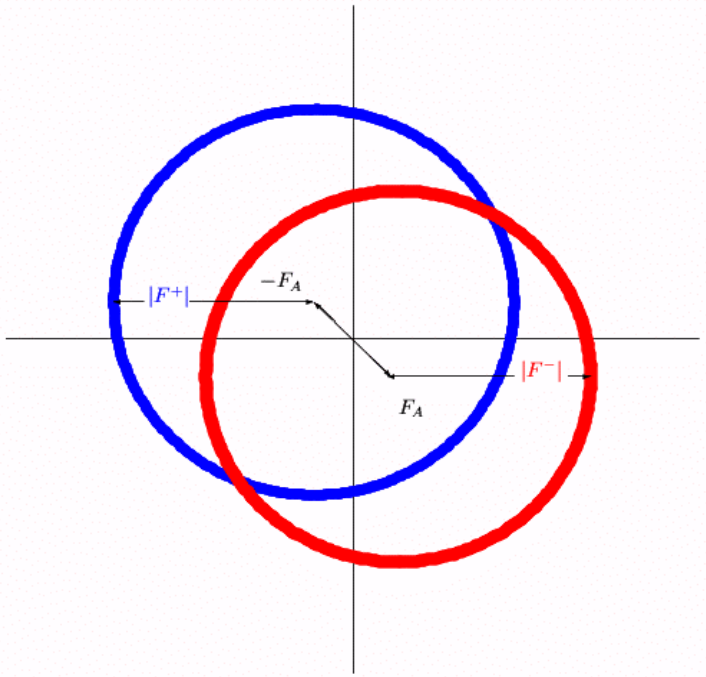
The strength of the anomalous effect depends on the wavelength of the X-rays

Anomalous scattering causes very small changes in intensity between F(+) and F(-)

Anomalous scattering causes the addition of a f' and f'' term to the atomic scattering factor.

- $\boldsymbol{F^{+}(h)} = \sum (f_j + f_j' + i\, f_j'') \exp(2\pi i\, h \cdot x_j)$
- $\boldsymbol{F^{-}(h)} = \sum (f_j + f_j' - i\, f_j'') \exp(2\pi i\, h \cdot x_j)$

Single-wavelength Anomalous Diffraction

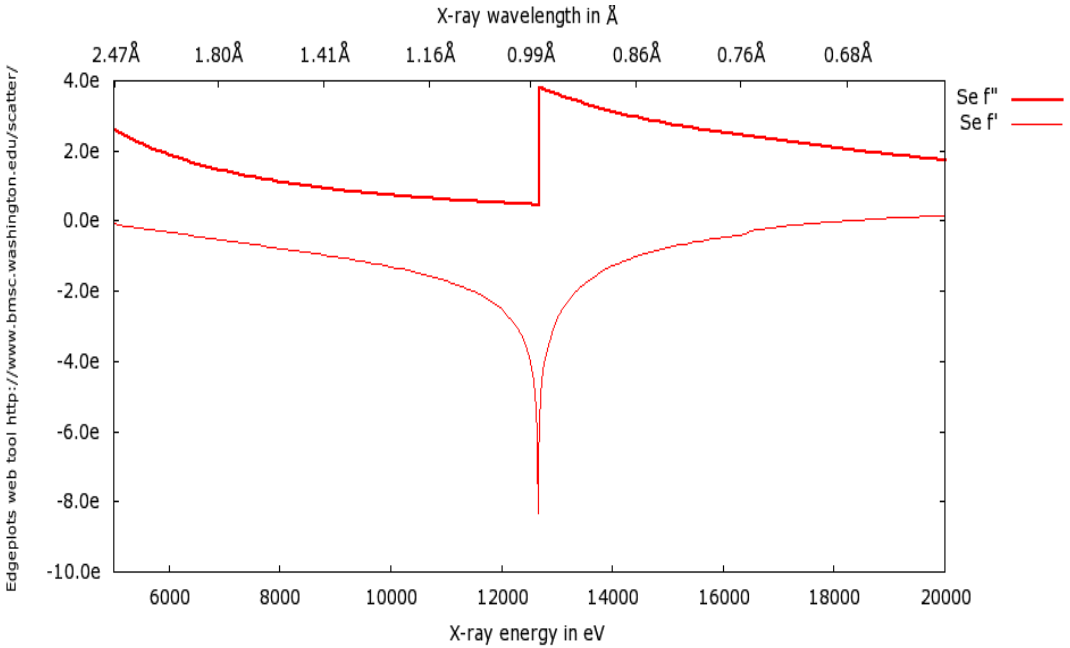


$F_A = \sum_j f''_j \exp(2\pi i h \cdot x_j)$

$F = F^+ - F_A$, $F = F^- + F_A$

Solving structures using Friedel pairs collected at one wavelength from a crystal that contains an anomalous substructure.

Effect of changing the wavelength: Anomalous scattering



- $\boldsymbol{F^{+}(h)} = \sum (f_j + f_j' + i\, f_j'') \exp(2\pi i\, h \cdot x_j)$
- $\boldsymbol{F^{-}(h)} = \sum (f_j + f_j' - i\, f_j'') \exp(2\pi i\, h \cdot x_j)$

MAD – multiple wavelength anomalous diffraction

Change wavelengths to perturb the structure factor to get different measurements.

More measurements can resolve the phase ambiguity.

At synchrotron sites, the wavelength can be changed/tunable.

Radiation damage becomes an even greater danger for MAD, as you expose the same crystal for longer periods.

Finding heavy atom sites

Need to first estimate "F_a" or "substructure factor amplitude" (a first guess is $F_a = \Delta F = |F^+ - F^-|$ for SAD).

Patterson methods:

By examining the map of interatomic distances, heavy atoms can be located. A Patterson map can be calculated without any phases.

Direct methods:

Uses mathematical methods from small molecule crystallography that exploits relation between structure factors

Hybrid (Patterson/Direct methods)

Combination of both

After we have initial phases: Density modification – solvent flattening

$\rho_{flattened}(x) = \mu(x)\, \rho(x)$

$F_{flattened}(h) = \sum_k F(k)\, G(h-k)$

Where $\mu(x)$ is a solvent mask, $\mu(x)$ is 1 in a region of protein, 0 in a region of solvent

G(h) is the Fourier transform of $\mu(x)$.

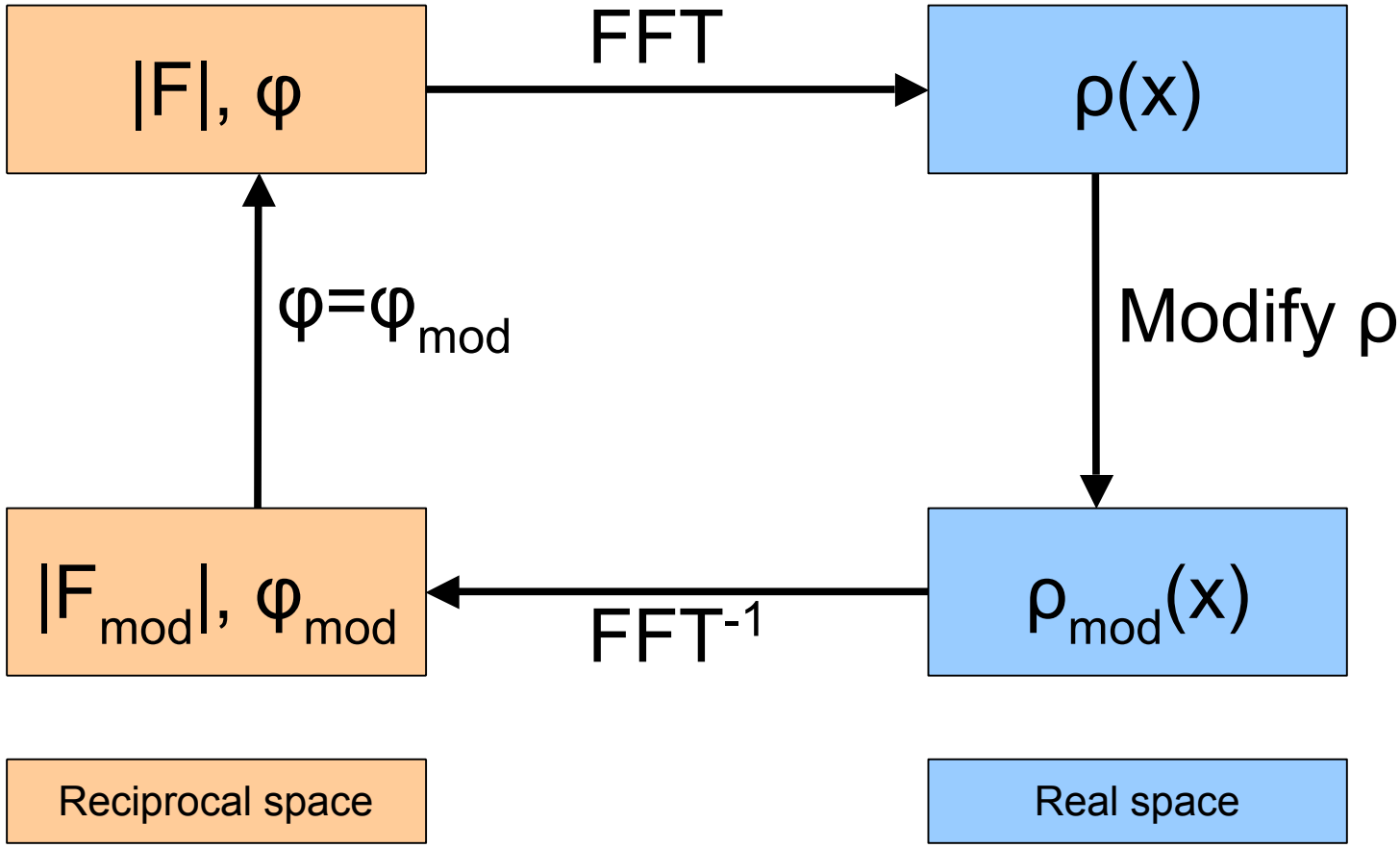
Density modification

Convolution and a solvent mask gives a relation between a structure factor of one Miller index to structure factor of other Miller indices.

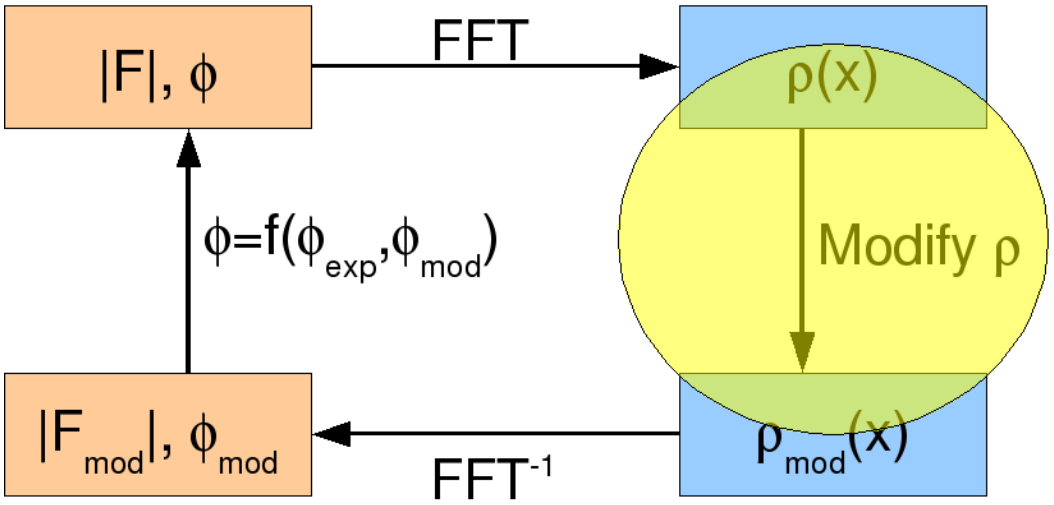
By iterating, and imposing our solvent mask, we can improve phases and our electron density map.

Density modification

1. Rudimentary calculation:



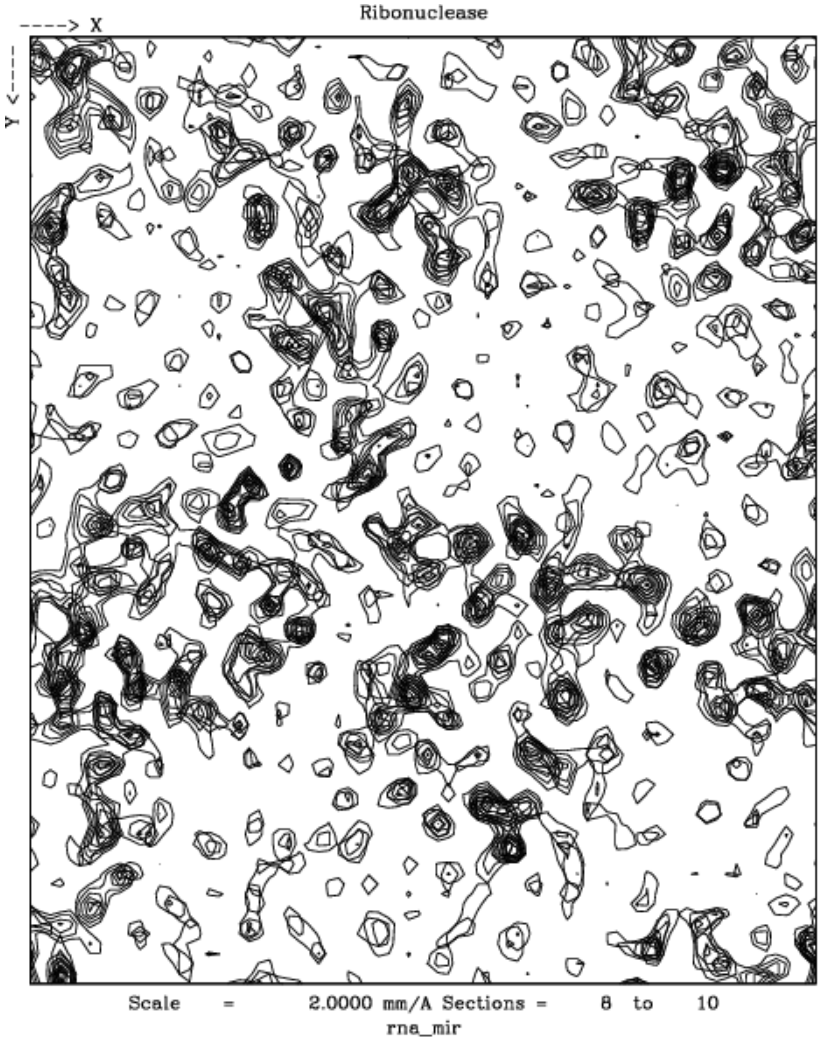
Density modification

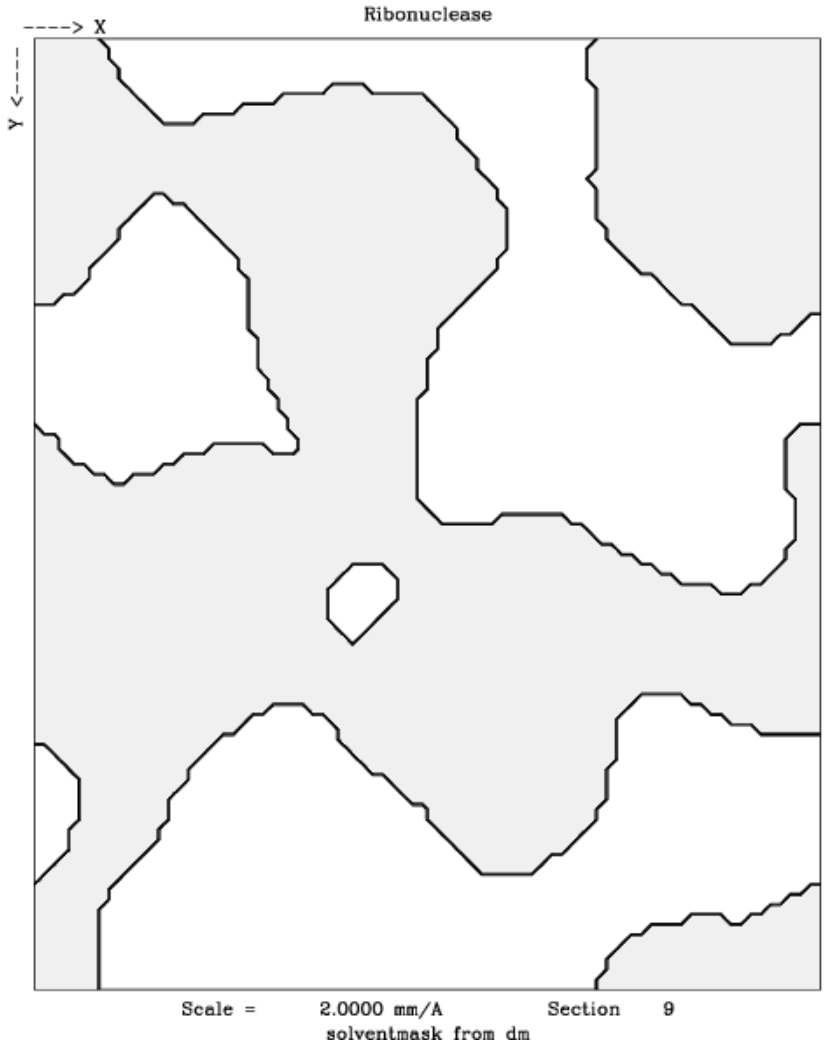


Traditional density modification techniques:

- Solvent flattening
- Histogram matching
- Non-crystallographic symmetry (NCS) averaging

Solvent flattening





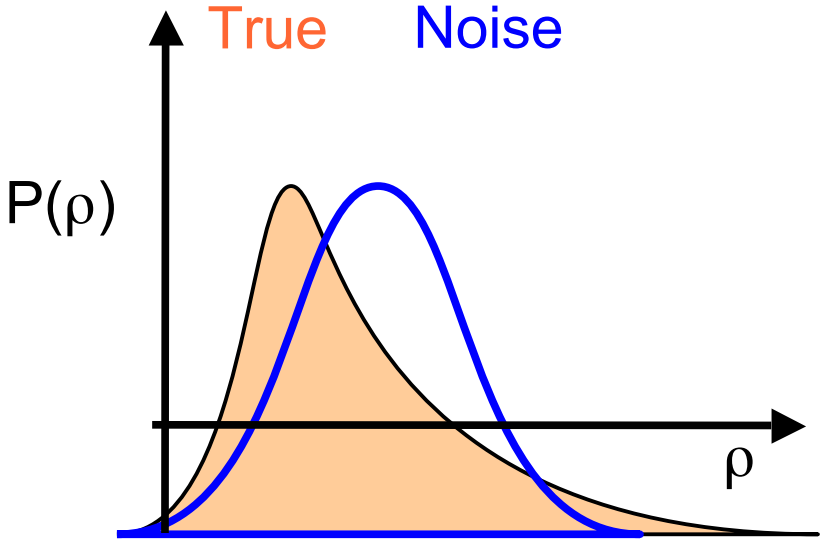
Histogram matching

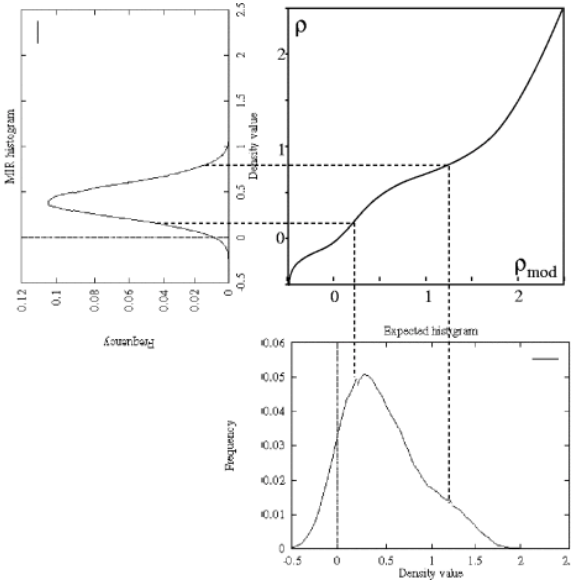
A technique from image processing for modifying the protein region.

- Noise maps have Gaussian histogram.
- Well phased maps have a skewed distribution: sharper peaks and bigger gaps.

Sharpen the protein density by a transform which matches the histogram of a well phased map.

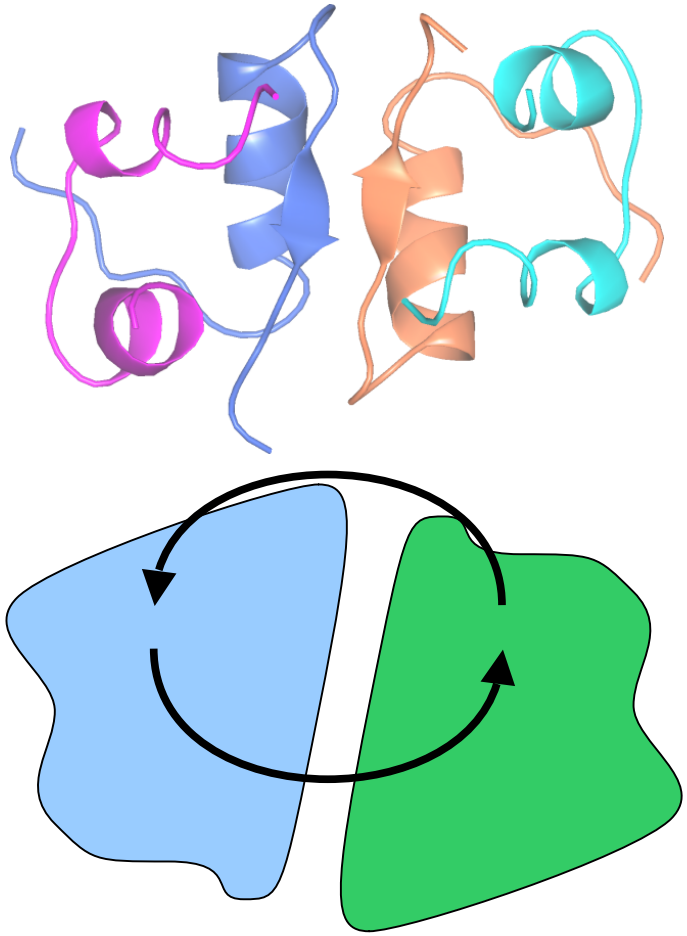
Useful at better than 4A.



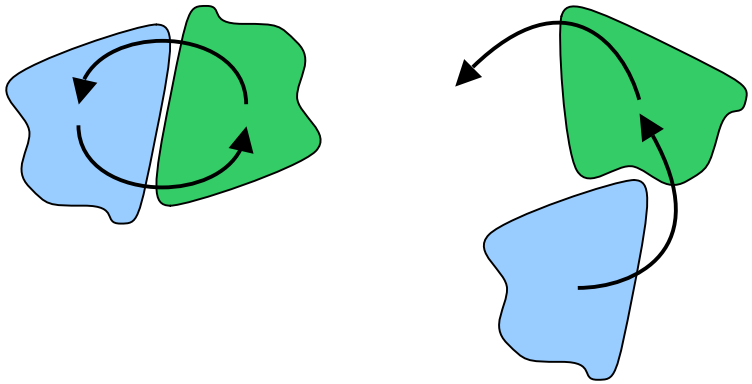


Non-crystallographic symmetry

- If the molecule has internal symmetry, we can average together related regions.
- In the averaged map, the signal-noise level is improved.
- If a full density modification calculation is performed, powerful phase relationships are formed.
- With 4-fold NCS, can phase from random!

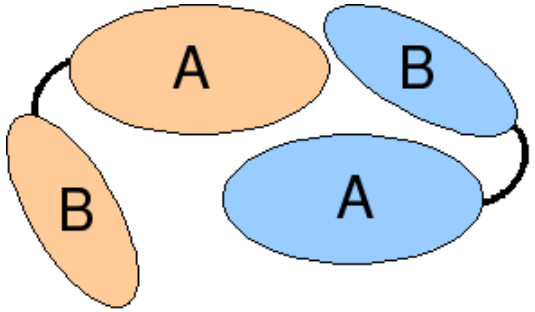


Non-crystallographic symmetry



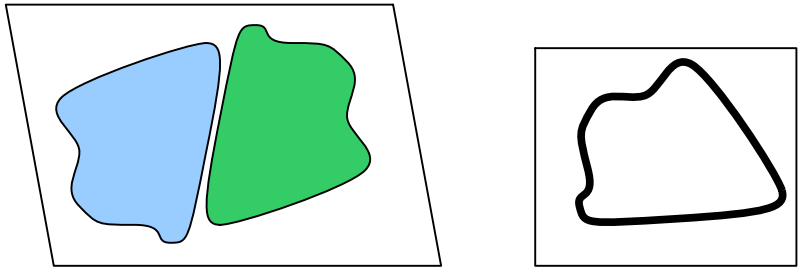
Useful terms:

- Proper and improper NCS: (closed and open)



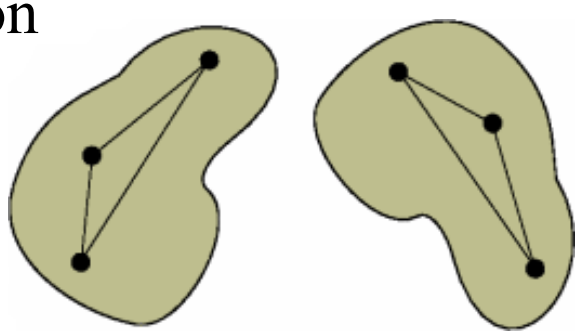
- Multi-domain averaging:

- Multi-crystal averaging:



Non-crystallographic symmetry

- How do you know if you have NCS?
  - Cell content analysis – how many monomers in ASU?
  - Self-rotation function.
  - Difference Pattersons (pseudo-translation only).
- How do you determine the NCS?
  - **From heavy atoms.**
  - From initial model building.
  - From molecular replacement.
  - *From density MR (hard).*
- Mask determined automatically.



Scope of Crank version 1.4

- Crank is for SAD, MAD, MAD+native and SIRAS.
- It requires minimal input, but is highly configurable.
- User friendly gui/pipelines for our latest developments in substructure detection, phasing, density modification and model building & refinement as well as plugins to externally developed programs.

Current Crank developers

Willem-Jan Waterreus

Pavol Skubak

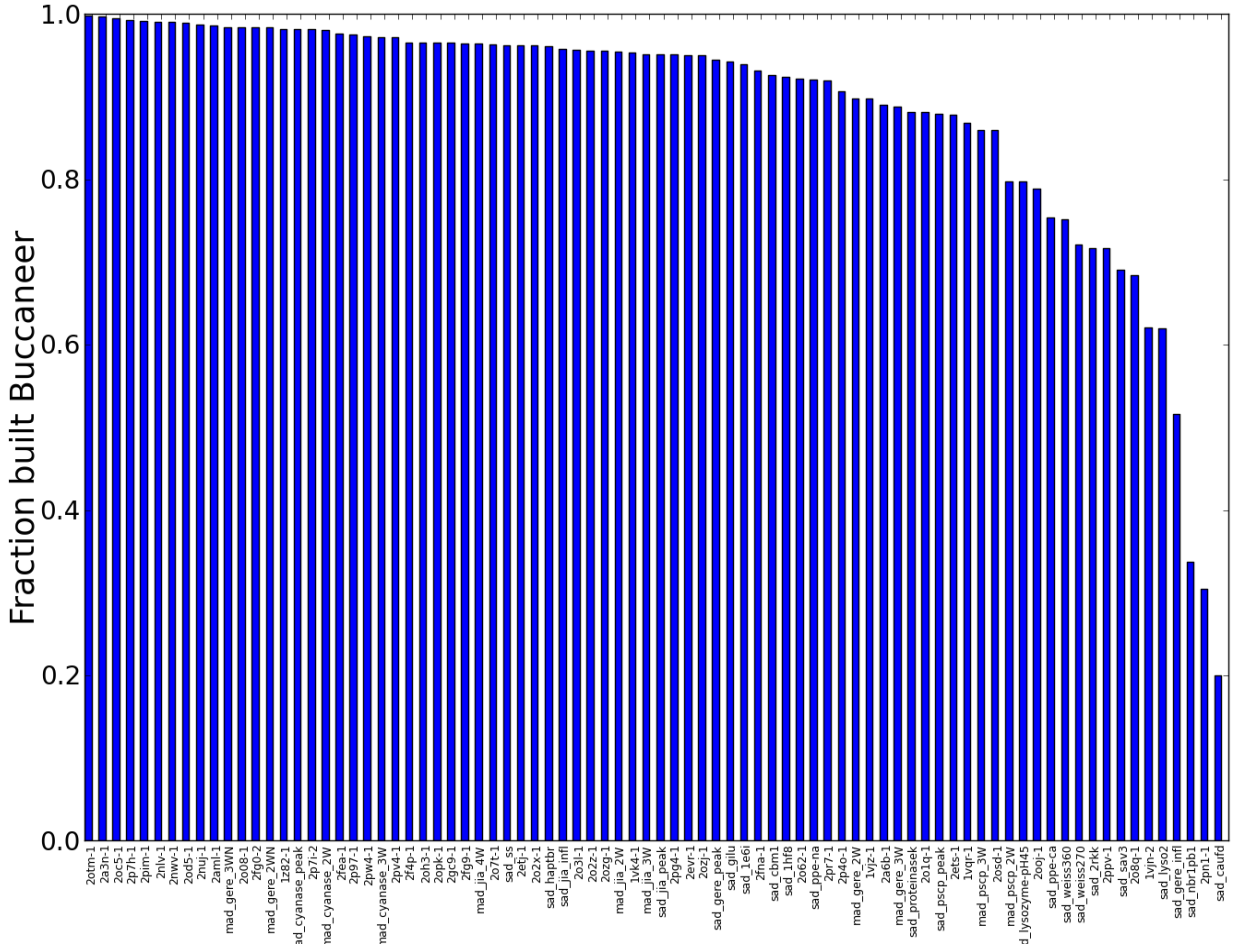
Jan Pieter Abrahams

RAG de Graaff

Assessing Crank 1.4's robustness

- A test system has been built of over 100 MAD, SAD, SIRAS data sets with a range of phasing quality and resolution.
- Over 10% are not solvable by data sets authors.

Crank pipeline in default mode: afro/crunch2/bp3/solomon/buccaneer



A challenging problem solved by default: GerE with SAD data

- GerE data set is distributed with CCP4 and originally solved by MAD + native.
- 2.7 Angstrom SAD peak data with 12 seleniums
- Could not be solved with earlier Crank versions.
- Crank version 1.3.x builds 70% by default.
- Crank version 1.4 builds 93% and builds over 70% of SAD data from inflection point.

Current F_A estimation

- F_A is currently estimated by $\Delta F = | |F^+| - |F^-| |$ for SAD data.
- Direct method programs are very sensitive to F_A values.
- Improving estimates can improve hit rates of direct methods and solve things that can not previously been solved.

AFRO: Multivariate SAD equation for F_A estimation

- Giacovazzo previously proposed multivariate F_A estimation, with an implementation assuming Bijvoet phases are equal.
- An equation can be obtained without the equal phase assumption requiring only one numerical integration.
- The multivariate F_A calculation leads to more substructures determined (by default) in data sets shown over ΔF.

CRUNCH2:
A program for substructure detection.

- Algebraic approach based on rank reduction of Karle/Hauptman matrices.
- Considers a higher order collection of reflections over triplets/tangent formula.
- de Graaff *et al.* (2001) *Acta Cryst.* D57, 1857-1862..

Important parameters in substructure detection

- The number of cycles run.
- The number of atoms to search for.
  - Should be within 10-20% of actual number
  - A first guess uses a probabilistic Matthew's coefficient
- The resolution cut-off:
  - For MAD, look at signed anomalous difference correlation.
  - For SAD, a first guess is 0.5 + high resolution limit.

Output from substructure determination

- If substructure coordinates are found, usually all positions are determined accurately.
- Indicators of a correct solution:
  - CCweak > 30% in SHELXD
  - FOM > 1.0 in CRUNCH2

(both are conservative criteria for a correct solution)

Validating substructure detection

- A substructure is assumed to be solved if it is over a statistical threshold defined by the detection program (ie. CCweak > 30% or CRUNCH2 FOM > 1.0)
- ***Problem***: Often, a substructure is correct, but the threshold is *not* reached.
- ***Solution***: Run Bp3 in “Check” mode, to verify if a solution is complete/correct.

BP3: Heavy atom refinement

- Can be used for SAD, MAD, S/MIR(AS).
- Refines atomic and error parameters.
- Outputs FOM, HL coefficients, PHIB to an MTZ file in original and inverted hand.
- Two “modes” of operation: normal and PHASe (fast phasing).
- Output from Bp3 should be input to a density modification program.

SAD functions in heavy atom refinement before BP3

- Earlier heavy atom refinement programs use a Gaussian (or least squares) function in Bijvoet differences ($\Delta F = |F^+|- |F^-|$) (North, 1965), (Matthews, 1966).
- The calculated Bijvoet difference is determined based on a assumed value of F and α and the heavy atom structure factor model.

Deriving a likelihood function suitable for a SAD experiment

- Include effect of model and measurement errors and correlation between observed and calculated Bijvoet pairs.
- Required joint probability distribution is

$$P(|F^+|,|F^-|;|F_c^+|,|F_c^-|) = \frac{\int\int P(|F^+|,\alpha^+|F^-|,\alpha^-,|F_c^+|,|F_c^-|)d\alpha^+ d\alpha^-}{P(|F_c^+|,|F_c^-|)}$$

- Would be suitable for substructure phasing, phase combination in density modification and model building + refinement and all combinations!

Improving the map

- Adjusting solvent content can improve the map after density modification. (Since the number of monomers is usually not known beforehand, neither is the solvent content.)
- If BP3 was run in fast mode, or SHELXE was run, a better map may result if BP3 is run in “default” mode.
- Use NCS averaging (see Crank/dm/Buccaneer demo on ccp4wiki.org).

Is my map good enough?

- Statistics from substructure phasing:
  - Look at FOM from BP3.
  - For SAD, look at Luzzati parameters.
  - Refined occupancies.
- Statistics from density modification:
  - Compare the “contrast” from hand and enantiomorph (output of solomon or shelxe).
- Does it look like a protein? (model visualization)

Conclusions/Remarks

- With a sufficient anomalous signal and resolution, structures can be solved automatically.
- When structures can not, first determine which step has failed: Crank attempts to make re-running steps easier.

Future developments

- MAD is NOT MIR – a multivariate likelihood MAD function in phasing and model refinement.
- A two-wavelength MAD function has been implemented (Sikharulidze and Pannu, in preparation) in phasing and F_A calculation and showing initial promising results.
- Multivariate functions allowing information from phasing, density modification and model building/refinement to be combined and thus no longer separating steps.

Availability & Documentation

- Crank works under Linux, MAC OS, Windows and is free software.
- Crank is available in CCP4 version 6.1.x
- ***Please* use version 1.3 or higher!**
- Crank wiki page is available:
  - http://ccp4wiki.org/
  - tested on undergraduates with no previous knowledge of crystallography/phasing

Acknowledgements

- All dataset contributors (JCSG, Z. Dauter, M.Weiss, C.Mueller-Dieckmann)
- Garib Murshudov, Kevin Cowtan, George Sheldrick, Victor Lamzin, Charles Ballard, Francois Remacle, Peter Briggs, Norman Stein, Martyn Winn
- http://www.bfsc.leidenuniv.nl/software/crank/

Nederlandse Organisatie voor Wetenschappelijk Onderzoek

Cyttron

CCP4