Automated structure solution with Crank

Biophysical Structural Chemistry,
Leiden University, The Netherlands

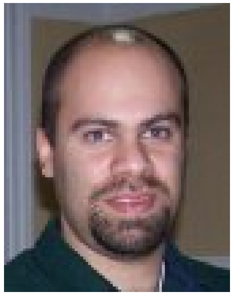
http://www.bfsc.leidenuniv.nl/software/crank/
http://www.ccp4.ac.uk/

Current developers

Pavol Skubak

Willem Jan Waterreus

Irakli Sikharulidze



Navraj S. Pannu

Jan Pieter Abrahams

RAG de Graaff

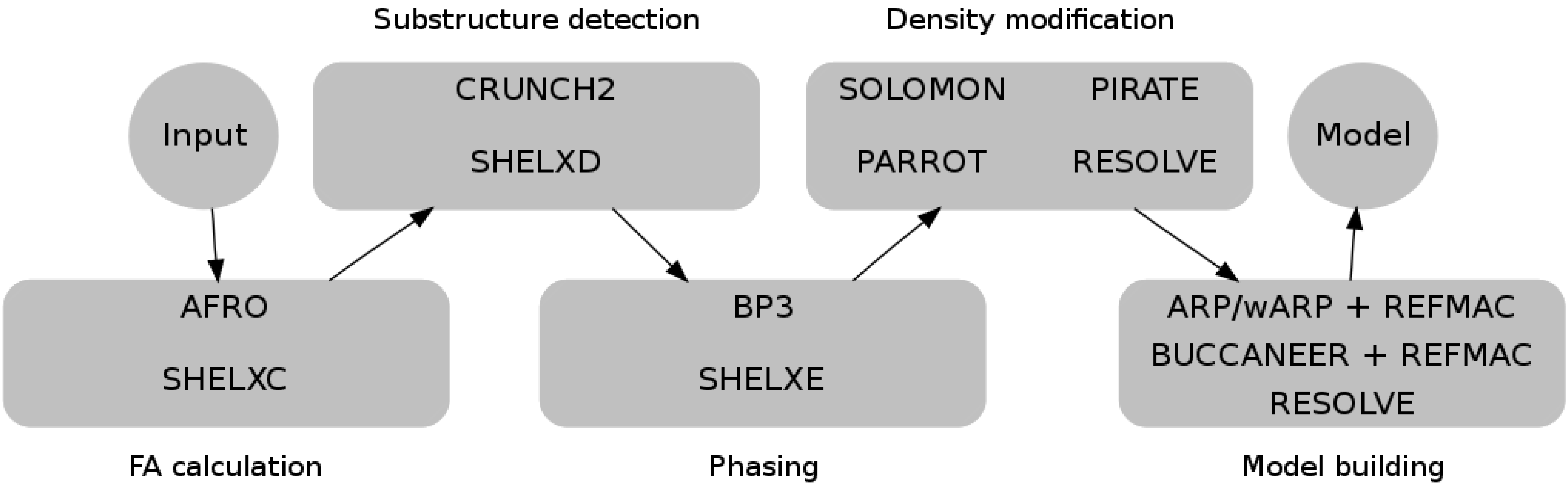
Scope of Crank version 1.4

- 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.

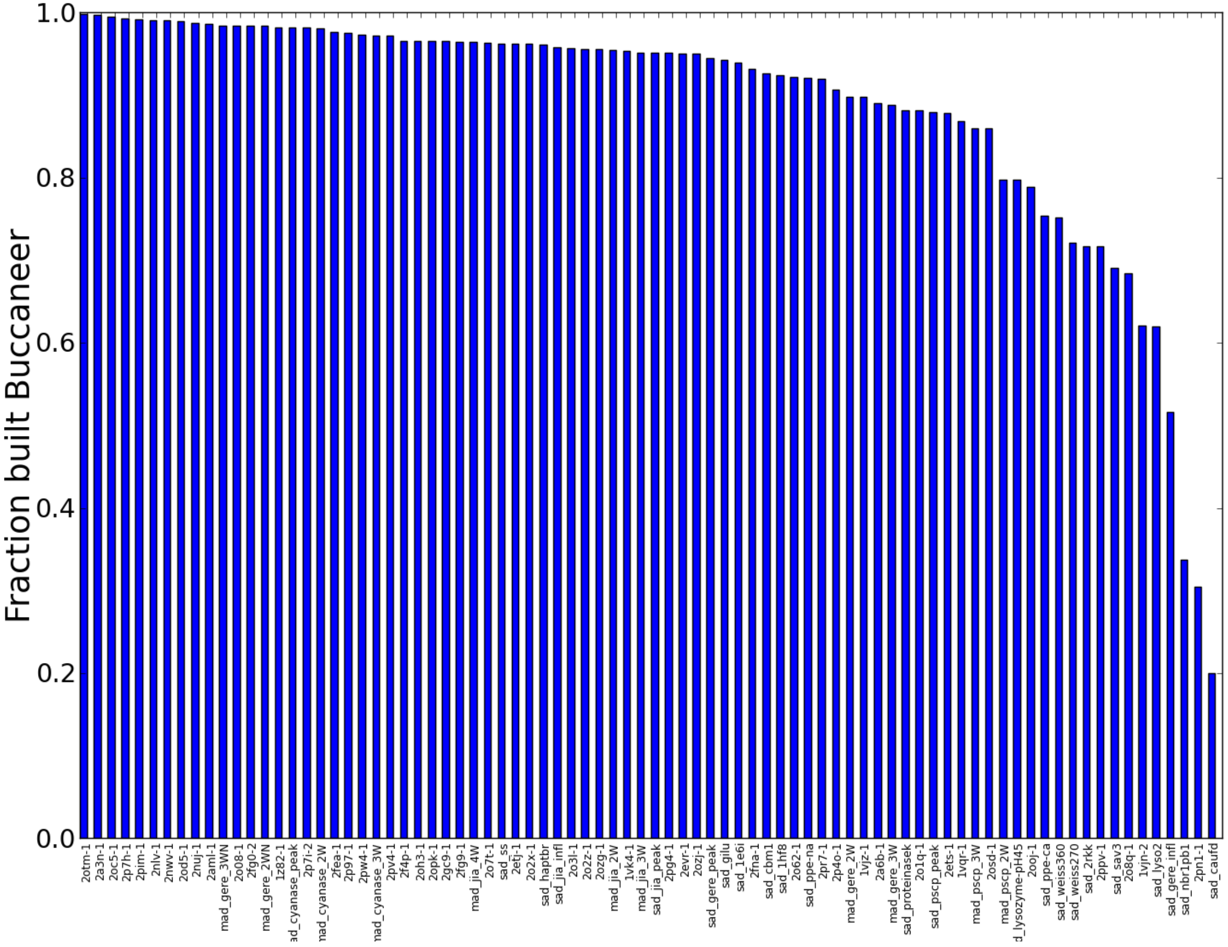
Assessing Crank 1.4's robustness

- 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.

Flow of Crank



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

$$\mathrm{E}(|F_A|;|F^+|,|F^-|) = \frac{\iiint |F_A| P(|F_A|,\alpha_A,|F_+|,\alpha_+,|F_-|,\alpha_-)\,\mathrm{d}|F_A|\,\mathrm{d}\alpha_A\,\mathrm{d}\alpha_+\,\mathrm{d}\alpha_-}{\iiint P(|F_A|,\alpha_A,|F_+|,\alpha_+,|F_-|,\alpha_-)\,\mathrm{d}|F_A|\,\mathrm{d}\alpha_A\,\mathrm{d}\alpha_+\,\mathrm{d}\alpha_-}$$

- 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

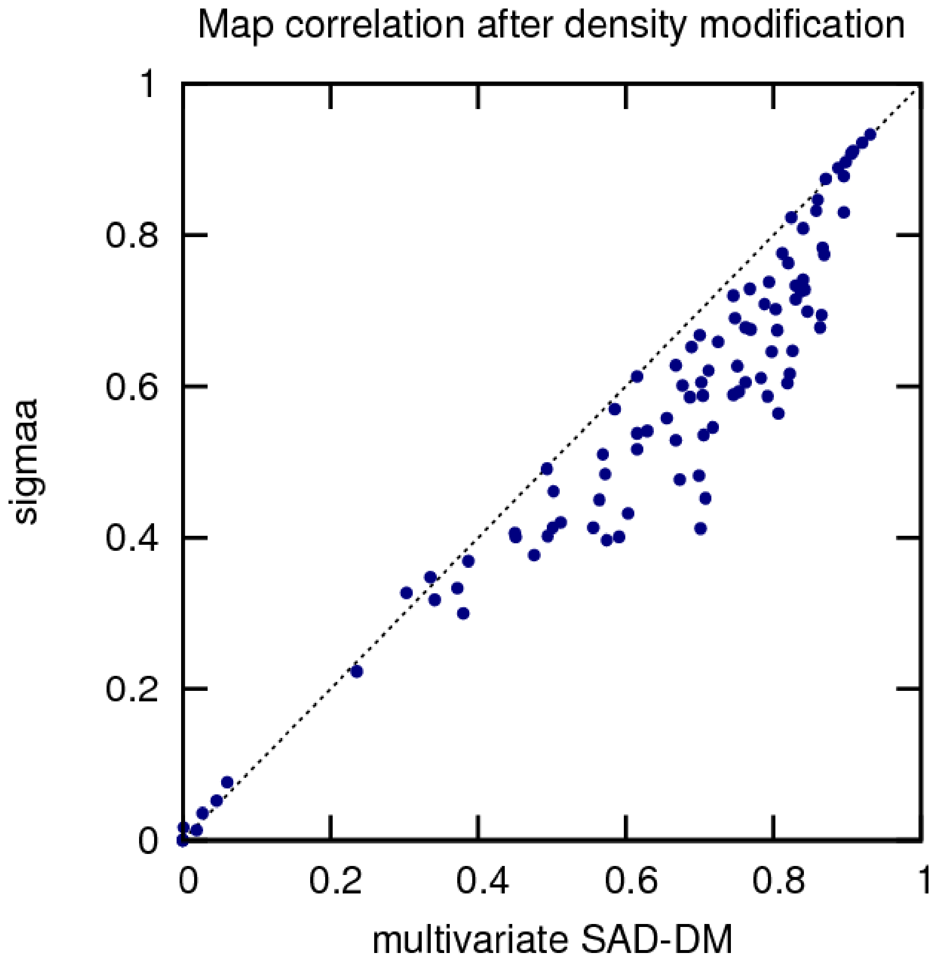
$$P(|F^+|,|F^-|;|F_c^+|,\alpha_c^+,|F_c^-|,\alpha_c^-) = \frac{\iint P(|F^+|,\alpha^+,|F^-|,\alpha^-,|F_c^+|,\alpha_c^+,|F_c^-|,\alpha_c^-)\,\mathrm{d}\alpha_+\,\mathrm{d}\alpha_-}{P(|F_c^+|,\alpha_c^+,|F_c^-|,\alpha_c^-)}$$

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

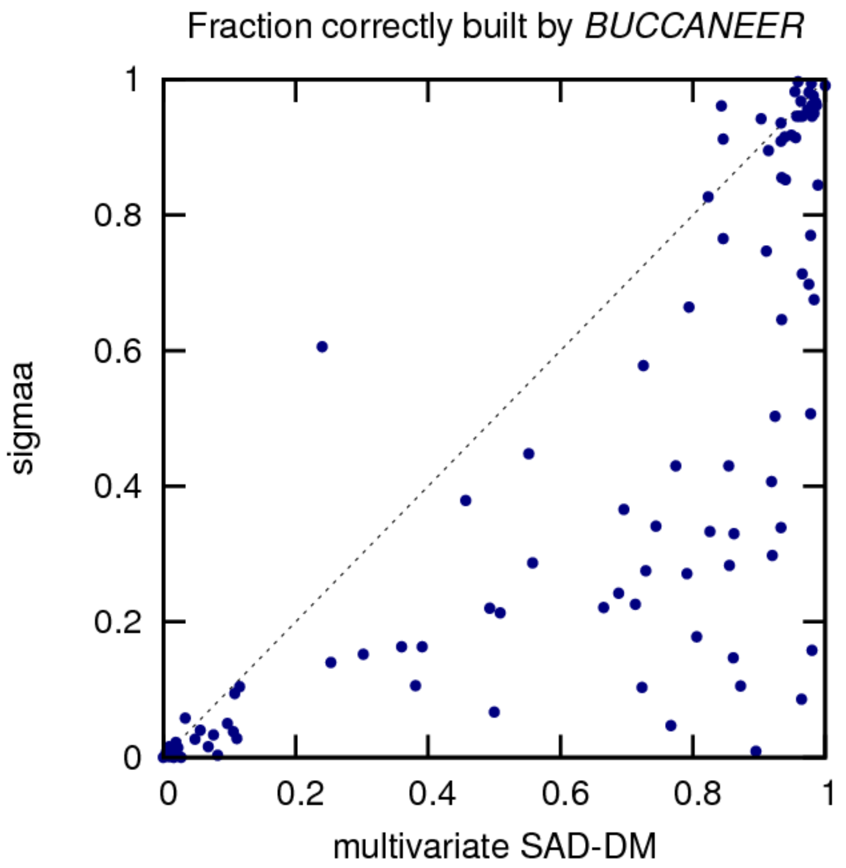
MULTICOMB: Multivariate phase combination for density modification

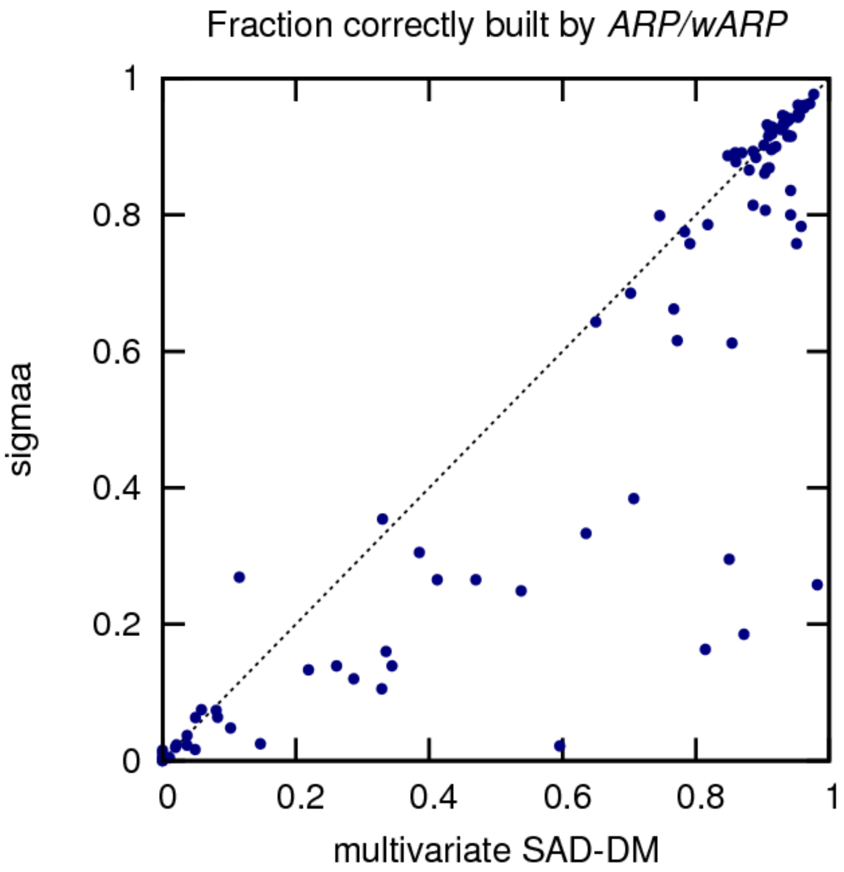
- Current density modification procedures
  - neglect the correlation between the original map and the density modified map.
  - approximate the original phase information with a 1 dimensional Hendrickson-Lattman distribution
- To overcome these shortcomings, we implemented a multivariate function which explicitly takes into account the correlation between the original, density modified and heavy atom structure factors.

Comparison of sigmaa vs. multivariate SAD function



Results of model building: sigmaa vs. multivariate SAD





β-correction method: bias reduction in density modification

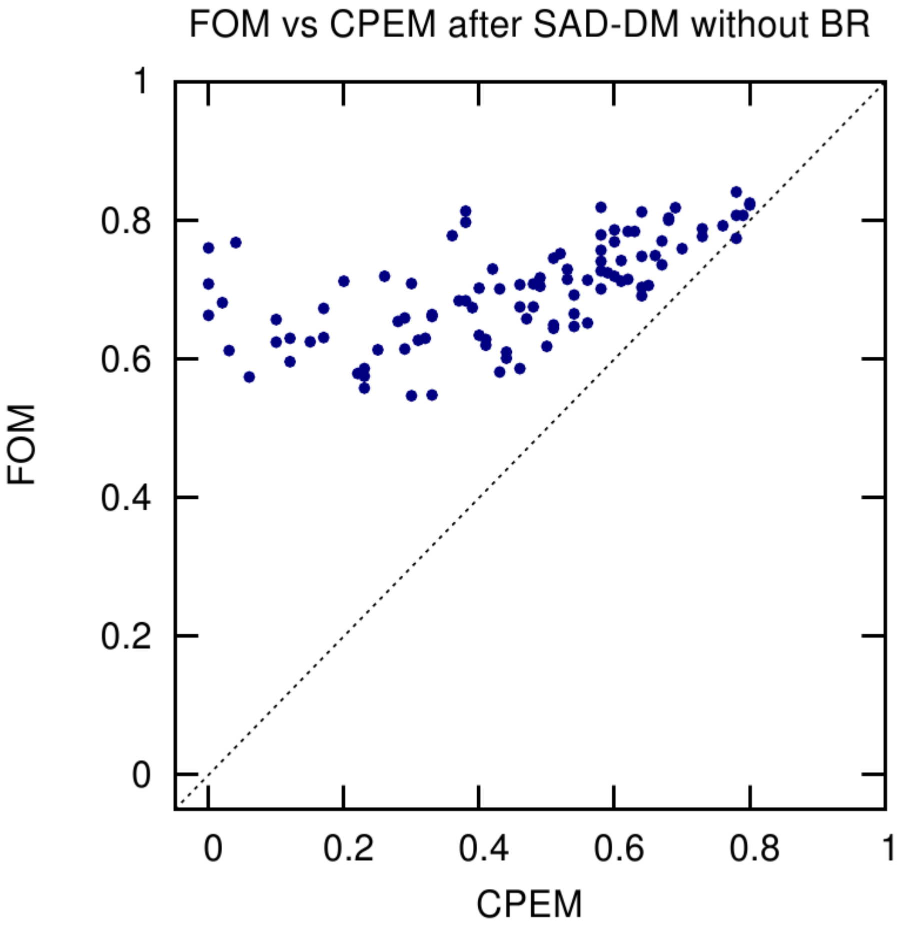
- Density modified map is obtained from experimental map leading to artificially high correlations between the observed and modified amplitudes.

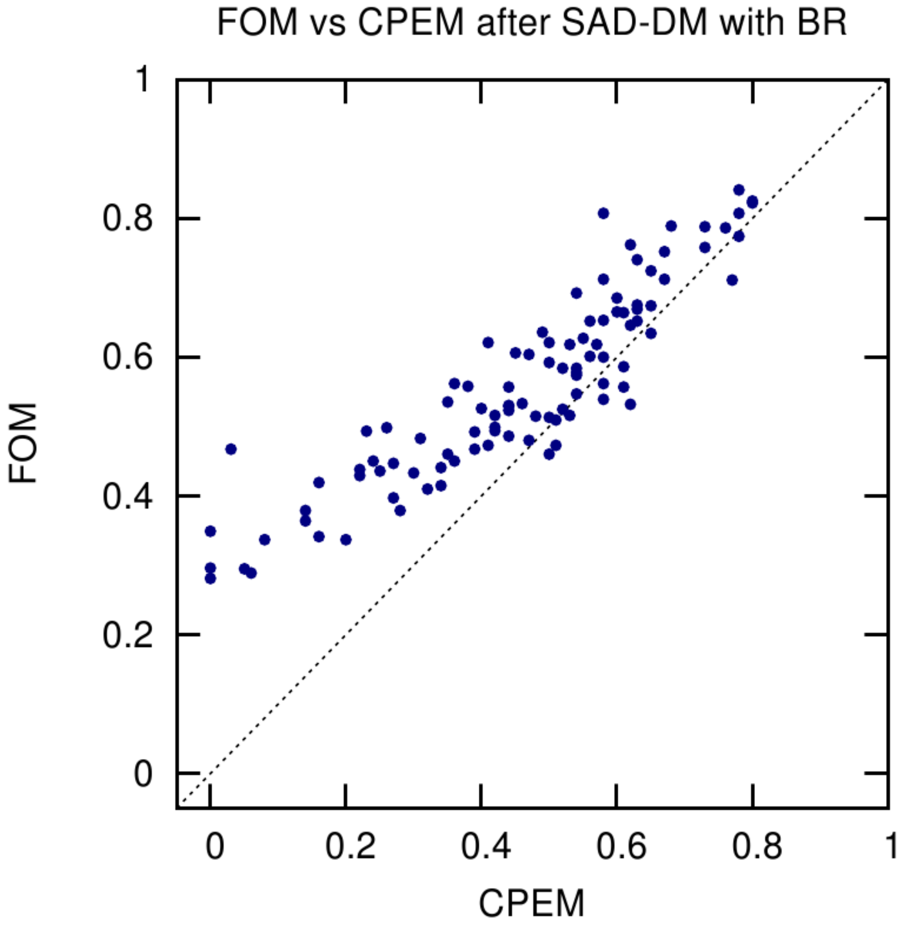
$$\langle F_o F_c \rangle = \beta D \langle |F_o||F_c| \rangle$$

$$\beta = \frac{cov(|F_o^{free}|, |F_c^{free}|)}{cov(|F_o^{work}|, |F_c^{work}|)}$$

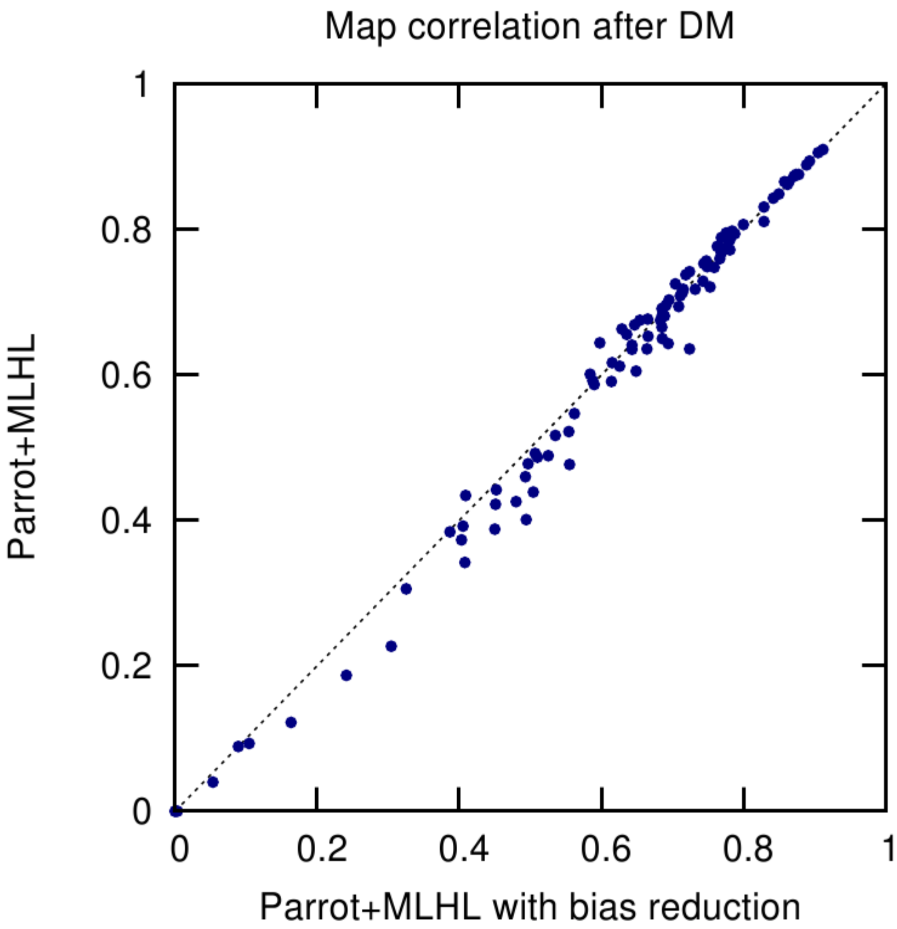
- β correction is applied to the Luzzati error parameter to reduce bias of modified data.

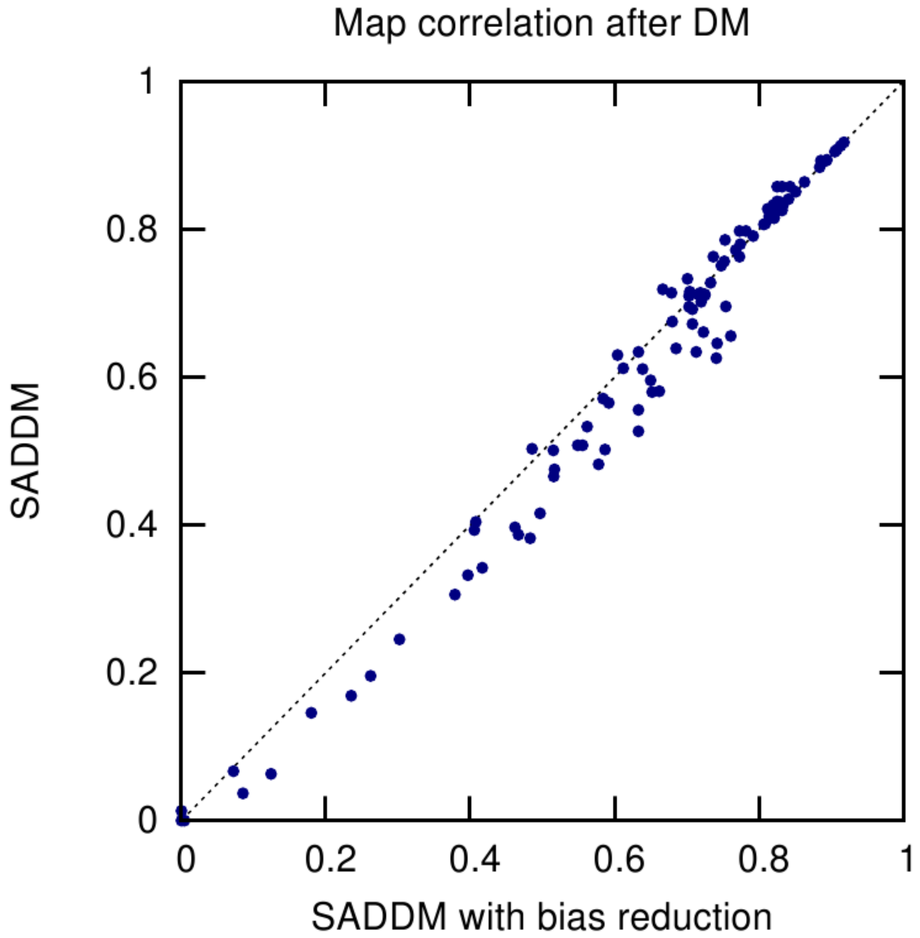
FOM and phase error after DM with/without bias reduction


FOM vs CPEM after SAD-DM without BR


FOM vs CPEM after SAD-DM with BR

Map correlations after DM with/without bias reduction





SAD and SIRAS functions in model refinement

- Previous functions in REFMAC:
  - No prior phase information (Rice function) (Murshudov *et al.*,1997), (Bricogne and Irwin, 1996), (Pannu and Read, 1996)
  - Prior phase information used indirectly in the form of Hendrickson-Lattman coefficients (MLHL function) (Pannu *et al.*, 1998)

Features of MLHL function

- Dependent on where you obtained your Hendrickson-Lattman coefficients.
- Assumes that your prior phase information is independent from your model phases!
- ***Benefit***: General approach for all experiments (MAD, SAD, MIRAS).

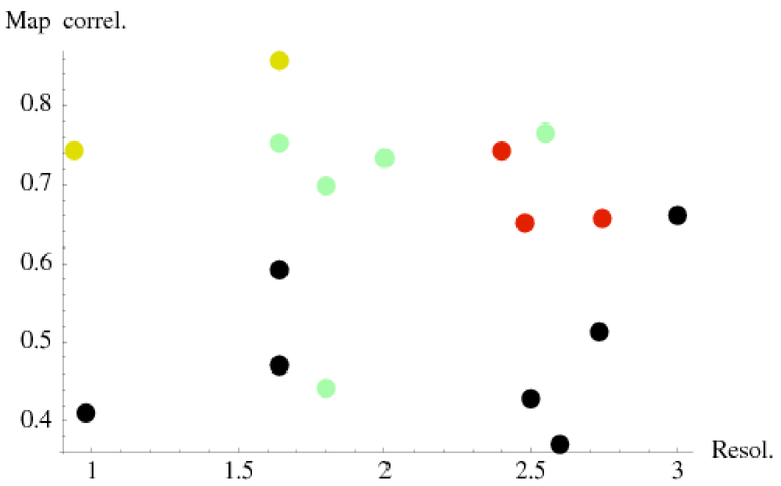
Multivariate SIRAS function for phasing and model refinement

- Currently in BP3 and SHARP, anomalous information is added for SIRAS and MAD by multiplying by a Gaussian term of Bijvoet differences (Thus, assuming independence with isomorphism term.)
- This isomorphic term also assumes uncorrelated errors.
- Better results are obtained by deriving a multivariate function for SIRAS modeling the correlation amongst data sets (Skubak et al. (2009) Acta D).

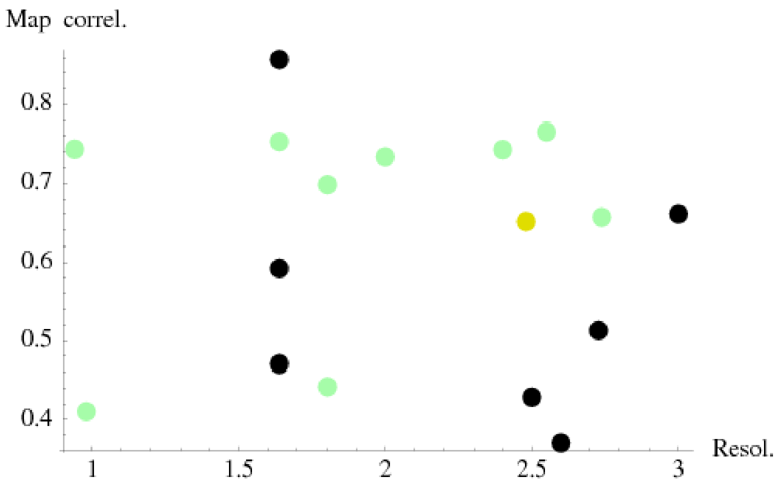
Tests of SAD and SIRAS functions in refinement

- The functions were tested on many real data sets (various phasing signals and resolution ranges) in ARP/wARP + REFMAC.
- Input created by CRANK using CRUNCH2 or SHELXD, BP3 and DM or SOLOMON.
- Skubak *et al.* (2004,2005,2009) Acta D.

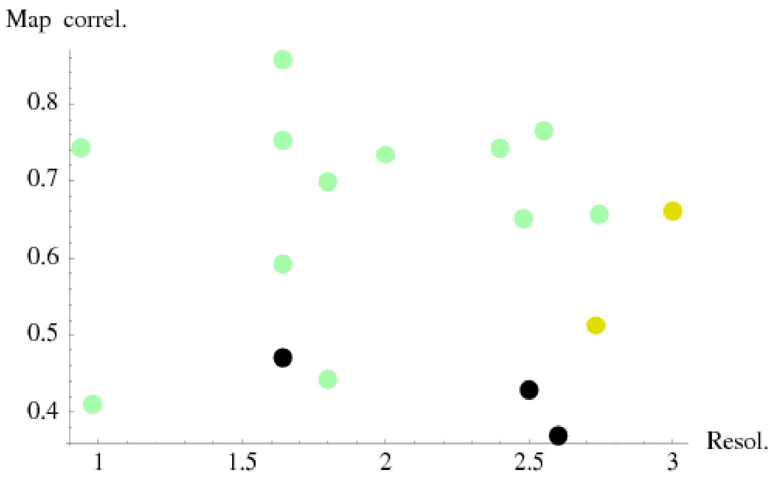
Results from SAD function



Rice function



MLHL function



SAD function

Green: 80 – 100% built

Yellow: 50 – 80% built

Red: 20 – 50% built

Black: 0 - 20% built

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)

Is my automatically build model correct?

- General comments for ARP/wARP, Buccaneer, and Resolve:
  - What fraction of residues have been built?
  - How long is the longest peptide built?
  - What fraction of amino acids built have sequence docked?

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/

NWO
Nederlandse Organisatie voor
Wetenschappelijk Onderzoek

Cyttron

CCP4