Recent developments in Crank

Navraj S. Pannu
Leiden University, The Netherlands

Current developers

Pavol Skubak

Ruben Zubac

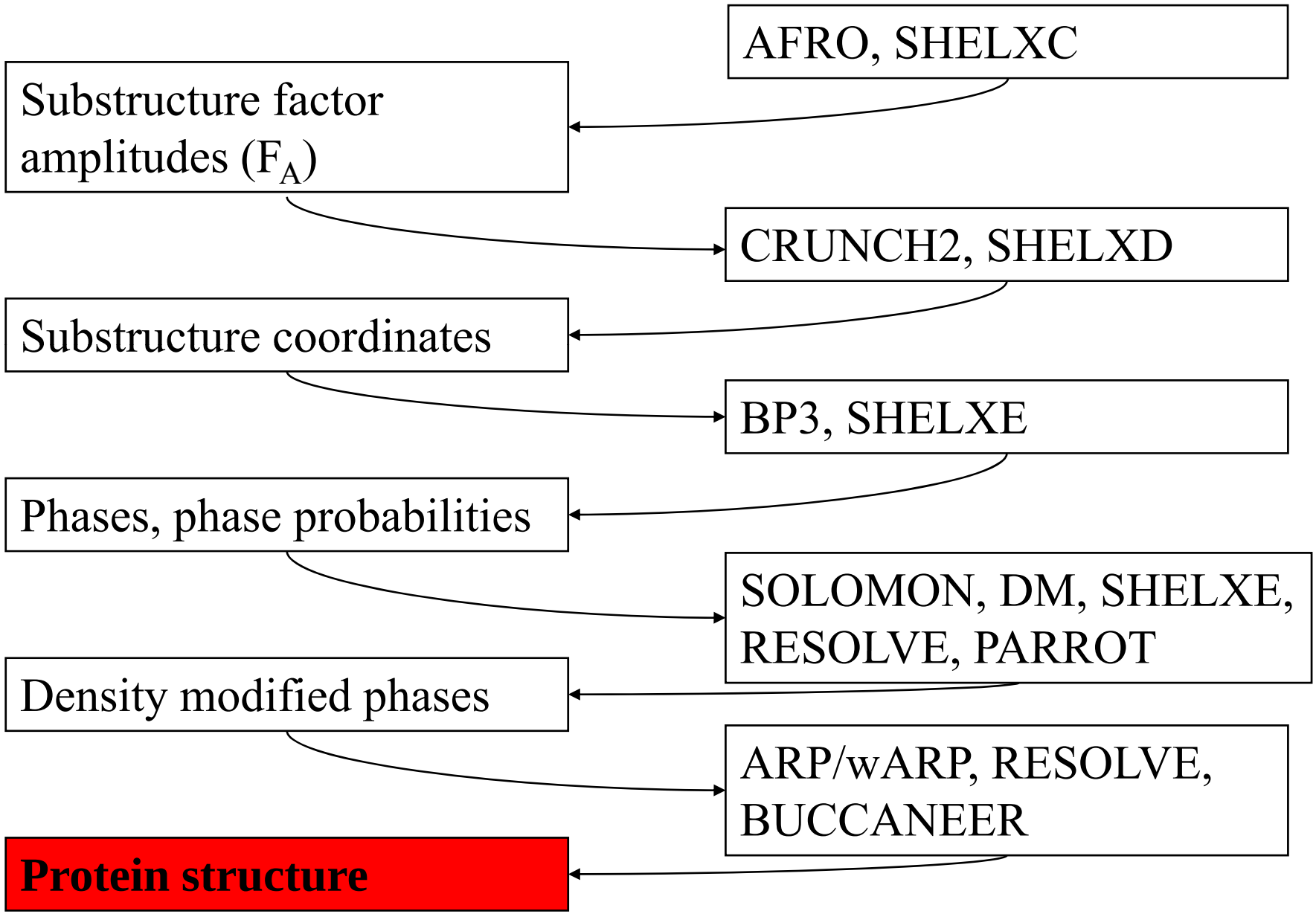
Irakli Sikharulidze

Jan Pieter Abrahams

RAG de Graaff

Willem-Jan Waterreus

Flow of CRANK



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..

Output from substructure determination

- If substructure coordinates are found, usually all positions are determined accurately.

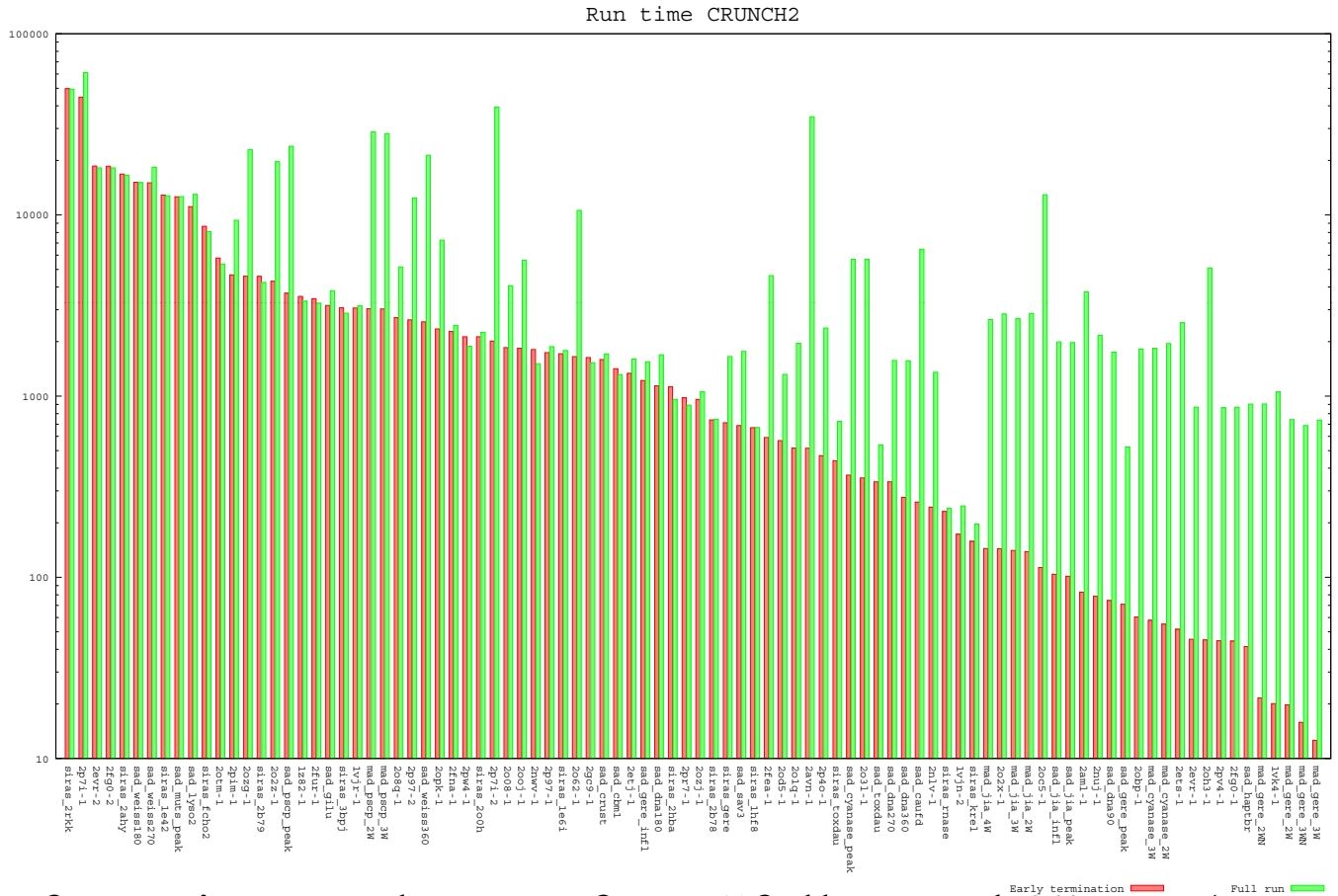
- Indicators of a correct solution:
  - CCweak > 30% in SHELXD
  - FOM > 1.0 in CRUNCH2

  (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.

Speeding up substructure detection



- Plot of run-time vs data set for a “full crunch2” run (green) and a crunch2 with early termination (red).

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.

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

- Most 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 $\boldsymbol{P(|F^{+}|, |F^{-}|; |F_c^{+}|, |F_c^{-}|)}$

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)

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 automatically built 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?

Testing Crank's robustness

- Plot of fraction built versus data set by just varying one model building program (red) versus another (green).

SAD function in model refinement

- Current 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) (Pannu *et al.*, 1998)

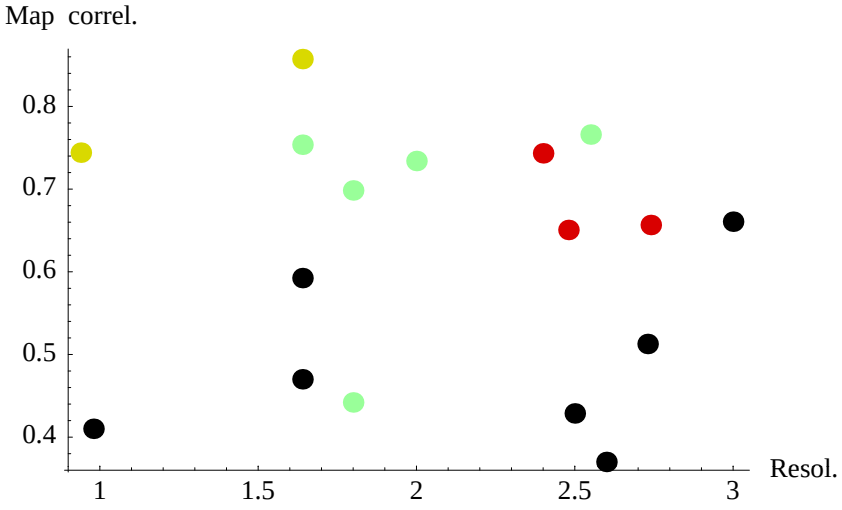
Shortcomings of MLHL

- 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).

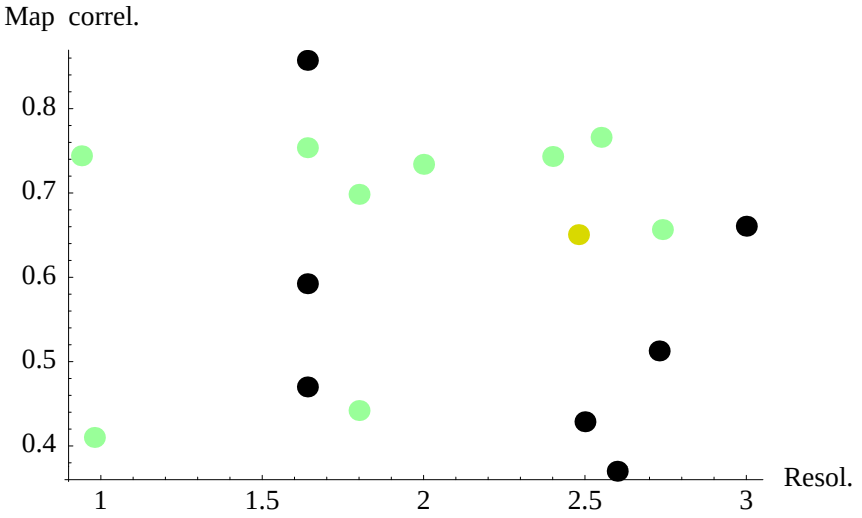
Tests of SAD function in refinement

- The SAD function was tested on many real SAD data sets (various anomalous signals and resolution ranges) against MLHL and RICE function in automated model building with iterative refinement in ARP/wARP + REFMAC.
- Input to ARP/wARP an REFMAC created with CRANK using CRUNCH2 or SHELXD, BP3 and DM.
- Skubak *et al.* (2004,2005) Acta Cryst D.

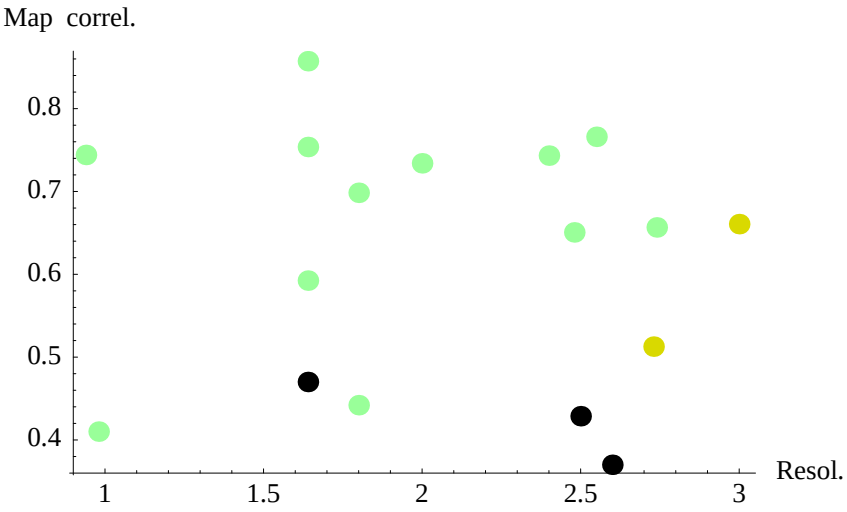
Results



Rice function



MLHL function



SAD function

Green: 80 – 100% built

Yellow: 50 – 80% built

Red: 20 – 50% built

Black: 0 - 20% built

Cases in the ShelxC/D/E pipeline with differences between SAD and RICE in wARP*+ Refmac

	Resol. (Å)	Anom. atoms	Experiment	Time# (min)	Residues RICE/SAD/FINAL
MutS	3.0	46 Se	SAD (peak)	156.2	493/1093/1600
subtilisin	1.77	3 Ca, S	SAD	55.65	6/259/275
thioesterase	2.5	8 Se	SAD (infl)	86.7	300/542/572
gere	2.75	12 Se	MAD(p/i)	28.81	43/110/444
cyanase	2.41	40 Se	MAD (p/i)	69.44	71/669/1560
thioesterase I	1.81	20 Br	SAD(peak)	65.46	35/431/462

*10 wARP cycles.

#Total Crank time for SAD target in Refmac.

Remarks on current Crank usage

- With a sufficient anomalous signal and resolution, structures can be solved automatically.
- When structures can not, first determine which step has failed (ie. Was a substructure found? Do I know the number of monomers? Etc.) Crank attempts to make re-running steps easier.

Future developments

- Improved density modification using data directly.
- Using SIRAS data directly in model refinement and multivariate SIRAS phasing.
- Joint ligand refinement.
- Multivariate F_A.

Developments in phase combination/density modification

- “Traditional” density modification attempts to combine a density modified map with experimental phases.
- Rather than assume independence, we attempt to model the correlation between the modified map and the original phases directly.
- We have developed a phase combination algorithm to consider the modified structure factors, SAD data and SAD heavy atom substructures together in a multivariate fashion.

Preliminary results

	After phasing	Current Solomon	Multivariate
ss	65.98/0.412	56.95/0.571	46.61/0.735
subtilisin	65.67/0.392	58.17/0.496	57.30/0.509
gerE	70.98/0.403	66.49/0.457	56.25/0.670
T4 bact.	73.78/0.365	67.72/0.420	63.55/0.457
rnase	77.02/0.180	74.46/0.176	67.89/0.312

FA estimation

- First step in solving a structure by SAD/MAD or SIRAS is to determine F_A values.

- F_A is the structure factor amplitude corresponding to the substructure to input to direct methods programs (*i.e.* SHELXD or CRUNCH2)

Current F_A estimation

- F_A is currently estimated by $| |F^+| - |F^-| |$ 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.

Multivariate SAD equation

$$E(|F_A|,|F^+|,|F^-|) =$$

$$\int\int\int\int |F_A|\, P(|F_A|, \alpha_A,|\, |F^+|, \alpha^+,|F^-|, \alpha^-)\, d|F_A|\, d\alpha_A\, d\alpha^+ d\alpha^-$$

- 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 equation has been implemented – which reduces to Giacovazzo's equation if Bijvoet phases are equal.

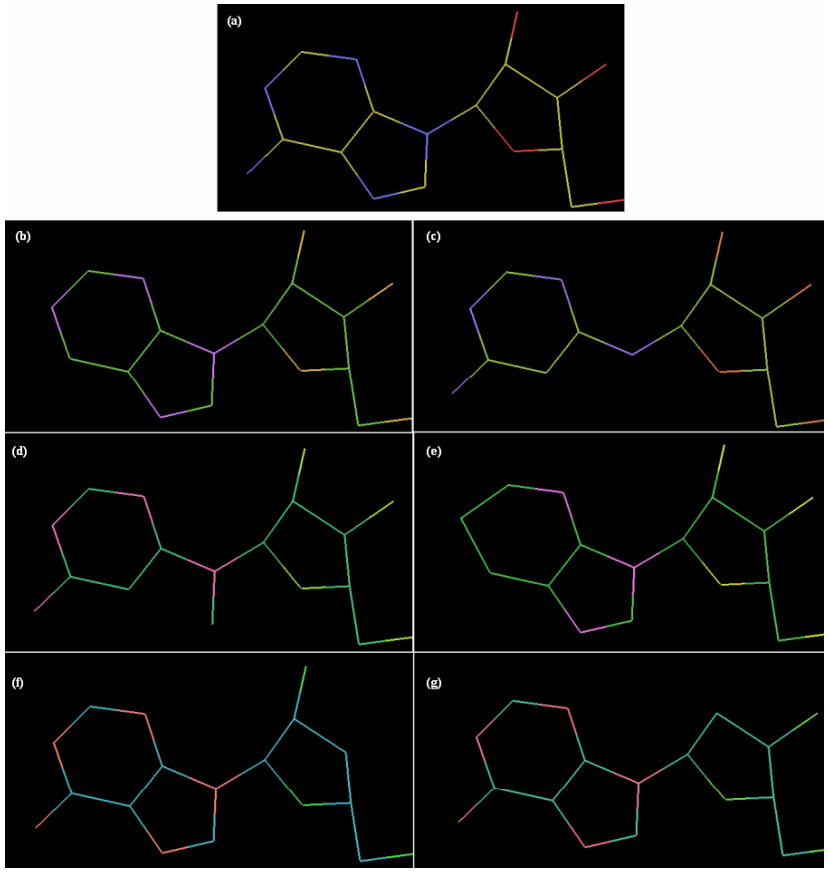
Test cases:
Correlations with final calculated E's

	Reso	Anom Atom	f″	Corr ΔE	Corr E_{multi}
Ferrodoxin	0.94	Fe	1.25	0.252	0.338
Thioesterase	2.5	Se	5.3	0.529	0.549
Lyso 180	1.64	S	0.56	0.324	0.348
Lyso 135	1.64	S	0.56	0.262	0.319
DNA 360	1.5	P	0.43	0.517	0.540
DNA 90	1.5	P	0.43	0.422	0.478

Joint refinement of structure, structure-ligand complexes

- Use multivariate function to refine together structure and structure-ligand complexes
- ***Motivation***: refining (isomorphic) data sets together could highlight the difference between the structure and provide better maps of the ligand.

Test case of ligand cocktail in ARP/wARP



(a) Correct ligand (CNA)
(b) PUI
(c) HJK
(d) YUH
(e) PUK
(f) ZIP
(g) ZIK

Results

Function	Input Ligand	1st choice	2nd choice
Current	none	PUI	CNA
Multivariate	none	CNA	PUI
Current	CNA	PUI	CNA
Multivariate	CNA	CNA	PUI
Current	PUI	PUI	CNA
Multivariate	PUI	CNA	PUI

Recall: CNA is correct ligand, PUI is incorrect

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 may be obtained by deriving a multivariate function for SIRAS modelling the correlation amongst data sets.

Results

Correctly built residues

Phasing function	Refinement function	2o0h	2b78	2b79
univariate	Rice	24	51	139
univariate	MLHL	34	96	197
univariate	Multi. SIRAS	53	345	235
multivariate	Rice	68	106	190
multivariate	MLHL	128	341	201
multivariate	Multi. SIRAS	310	356	238

Future developments

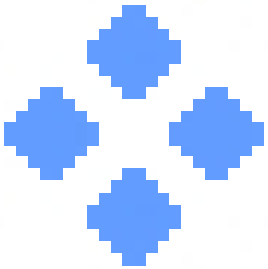
- MAD is NOT MIR – a multivariate likelihood MAD function in phasing and model refinement.
- Multi-crystal support in Refmac for dealing with (severe) non-isomorphism in SIRAS and with radiation damage.

Availability

- Crank works under Linux, MAC OS, Windows and is free software.
- Crank is available in CCP4 version 6.1.1 or from http://www.bfsc.leidenuniv.nl/software/crank/
- ***Please*** use version 1.1 or higher!
- Crank 1.3.0 (to be released in CCP4 6.1.2) has annotated logfiles for displaying results.

Acknowledgements

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



Cyttron