MR introduction

Michail Isupov

Biosciences, University of Exeter

Outline of this talk

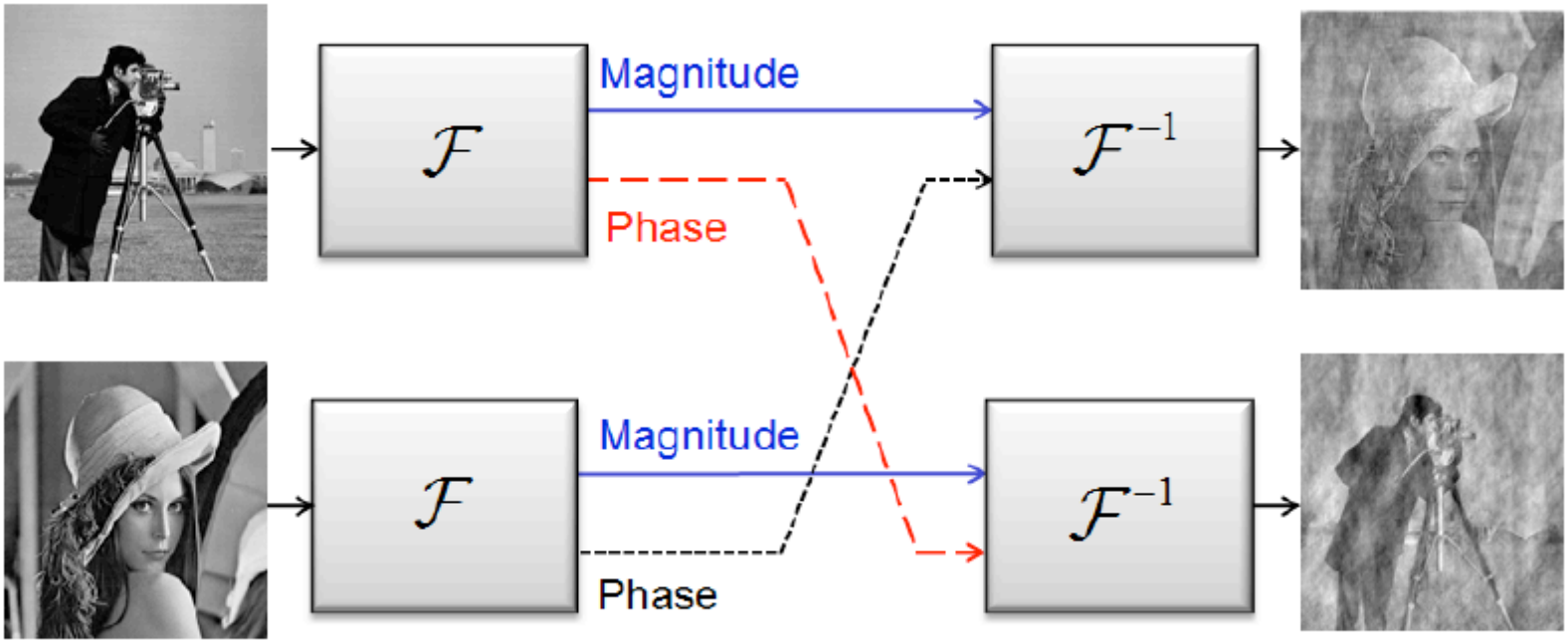
- **What is Molecular replacement**
- **Functions for Molecular Replacement in Molrep**

- **MR pipelines in CCP4: examples**

Importance of phases

It is well known that phases are very important.

This is one of the examples demonstrating it:



PHASE PROBLEM

Phases (lost in experiment) are more important for quality of electron density maps than intensities (we measure).
Initial estimation of phases can be obtained from small differences in experimental intensities or by positioning a known structure of a related macromolecule in a new cell.
Two phasing techniques are used to resolve phase problem and obtain interpretable ED maps.

1. Molecular Replacement (MR) and
2. Experimental Phasing (EP).

The phase error range is 0-180°, random phase error is 90°.
The task of a crystallographer is to obtain an initial set of phases with an average phase error below 80° and to improve these phases to an average error around 45° or better.

MOLECULAR REPLACEMENT

If a structure is known of a homologue molecule (for proteins > 30 % sequence identity) molecular replacement (MR) is the method of choice. MR positions the known structure into unknown crystal cell and this gives a set of initial phases which can be improved by structure refinement and density modification.

Some structures are solved with as low sequence identity as 15-20 %. Proportion of structures solved by MR grows with an increase in number of PDB entries.

Overall results reported in PDB (2006)

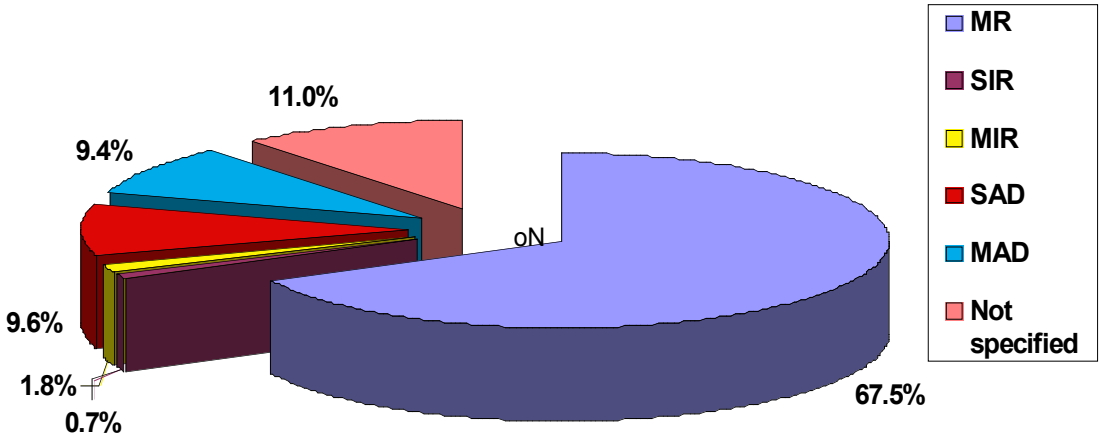
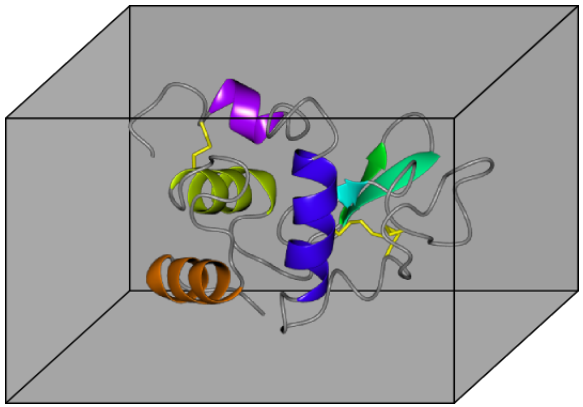


Diagram showing the percentage of structures in the PDB solved by different techniques

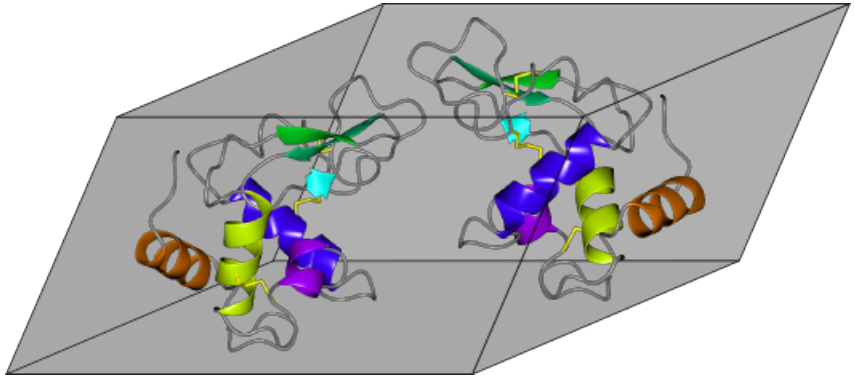
At least 80% of structures are solved by Molecular Replacement (MR) (2016)
Around 10% of structures are solved by experimental phasing

MR Problem

Known crystal structure



New crystal structure

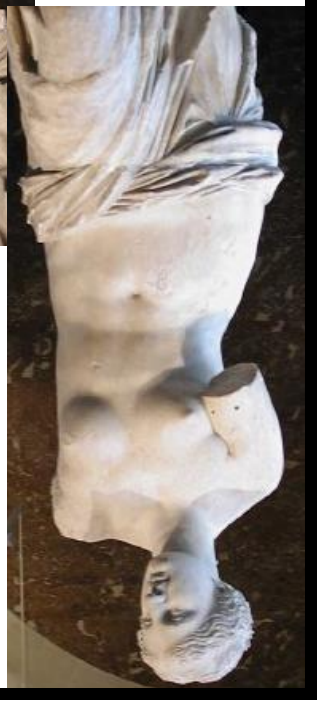


Given:
- Crystal structure of a homologue
- New X-ray data

Find:
- The new crystal structure

model corrected MR solution → Refinement → result
Rebuilding

H_2O



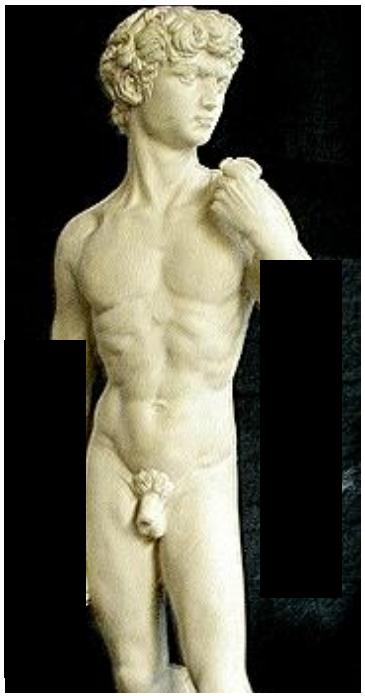
H_2O

Model

Corrected model

MR solution

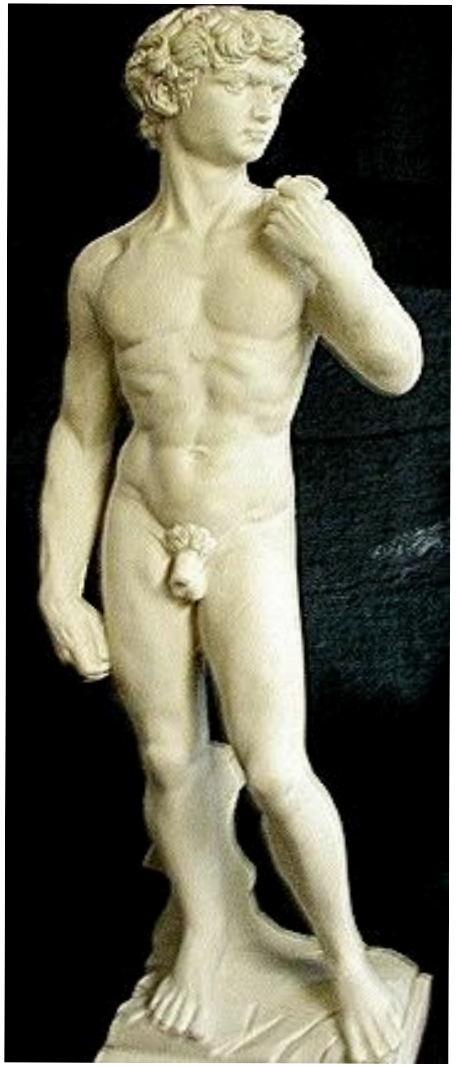
after refinement



Model

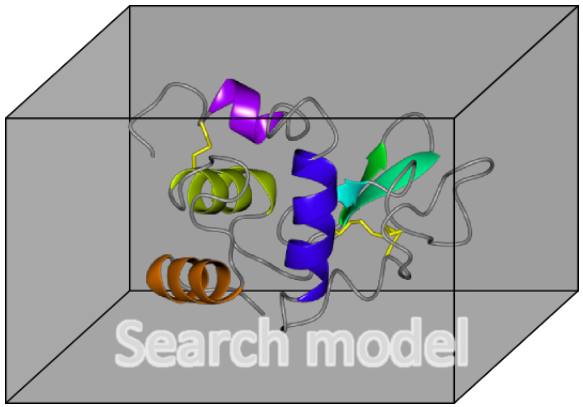
Corrected model

MR solution after refinement

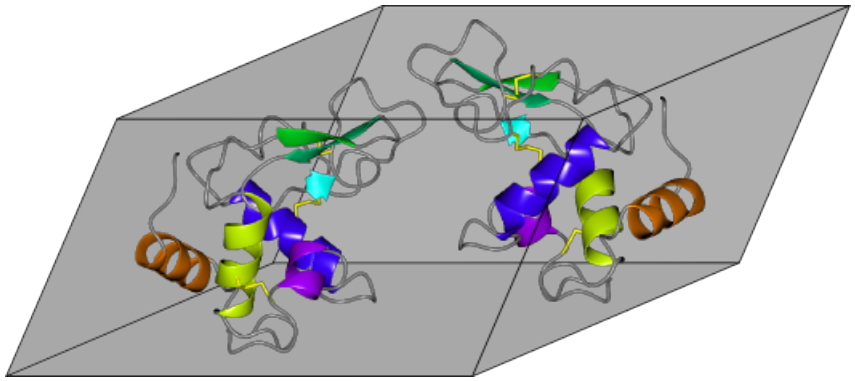


MR Technique

Known crystal structure



New crystal structure



Method:

- 6×N - dimensional global optimisation
  - one 6-d search for each molecule in the AU
    - >> split further to orientation + translation searches = 3 + 3
    - >> fast search step using FFT

Required:

- Scoring
  - the match between the data and an (incomplete) crystal model
  - ideally: the highest score = correct solution

Molecular Replacement in CCP4

MR Programs

- AMoRe
- Molrep
- Phaser

MR Pipelines

- MrBUMP
- Balbes
- Morda
- AMPLE
- Arcimboldo
- Simbad

CCP4 programs for model preparation

Single model correction:

- Chainsaw
- Molrep
- Sculptor

Preparation of ensemble models – fitting models:

- Lsqkab
- ProSMART
- SSM (also in Coot)
- Gesamt

Automated preparation of ensemble models:

- Ensembler
- MrBUMP-CCP4MG

Molrep

Alexey Vagin

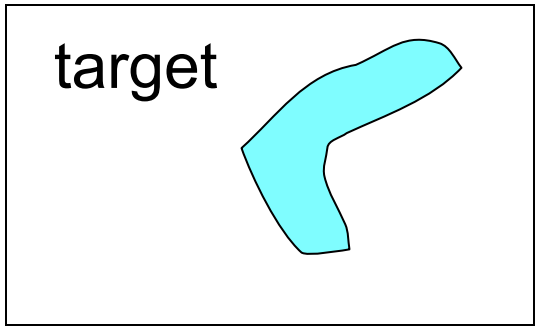
YSBL University of York

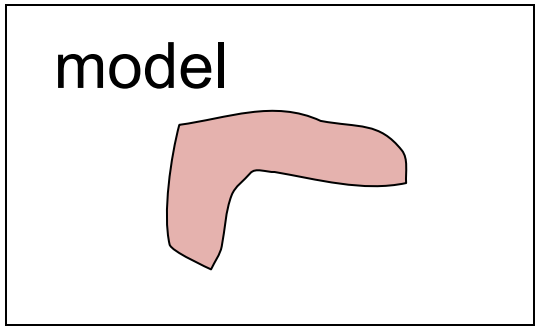
Default protocol

```
molrep -f data.mtz -m model.pdb -mx fixed.pdb -s target.seq
```

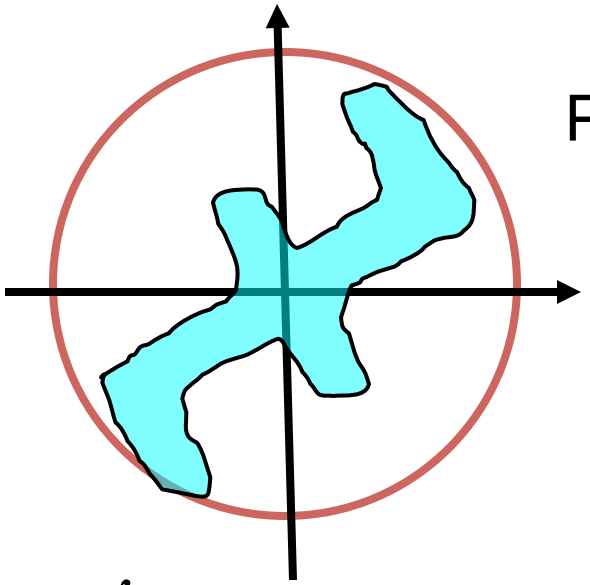
- model correction based on sequence and structure information
- defines the number of molecules per AU
- anisotropic correction of the data
- weighting the data according to model completeness and similarity
- check for pseudotranslation and use it if present
- 30+ peaks in Cross RF for use in TF (accounts for close peaks)
- applied packing function
- make use of partial structure (fixed model)

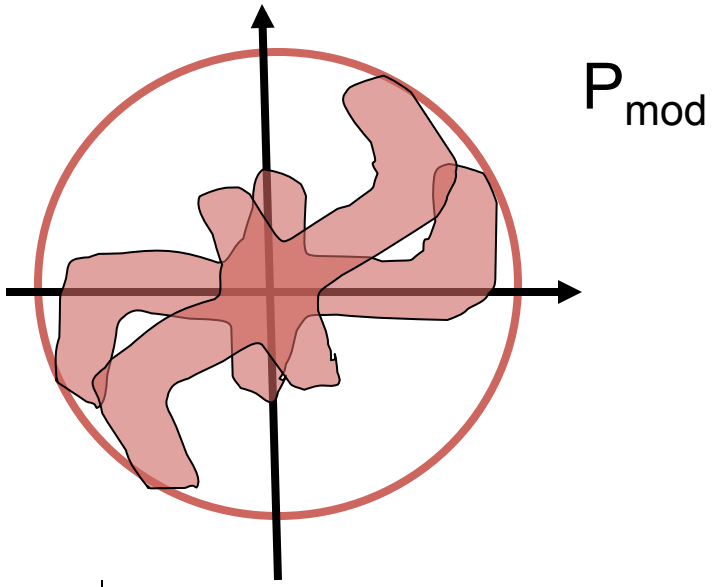
Cross Rotation Function





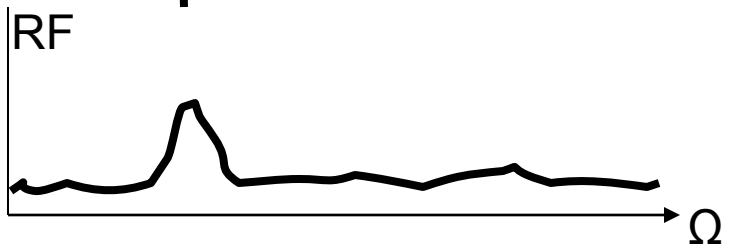
Patterson



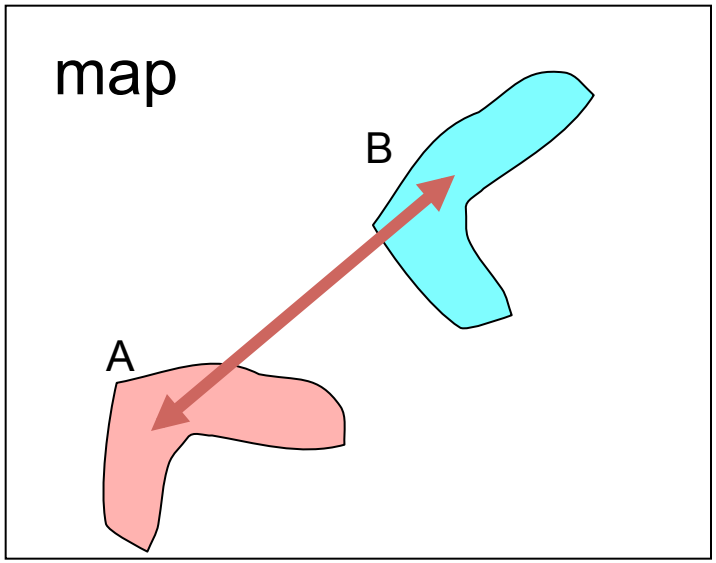


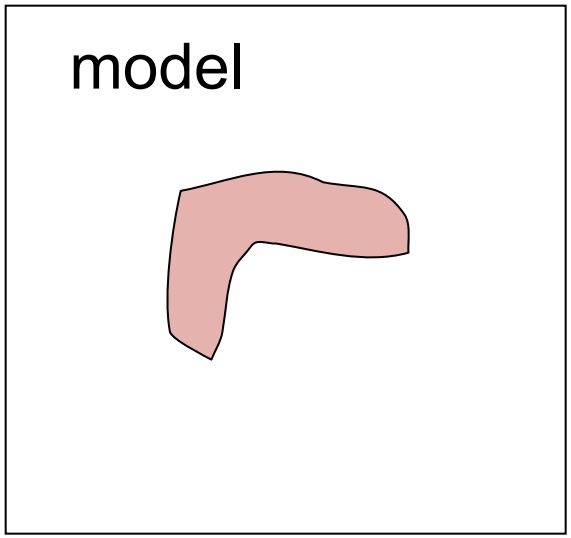
$$RF(\Omega) = \int P_{obs}(r)\ \Re_{\Omega}\{ P_{mod}(r)\}\ dr$$

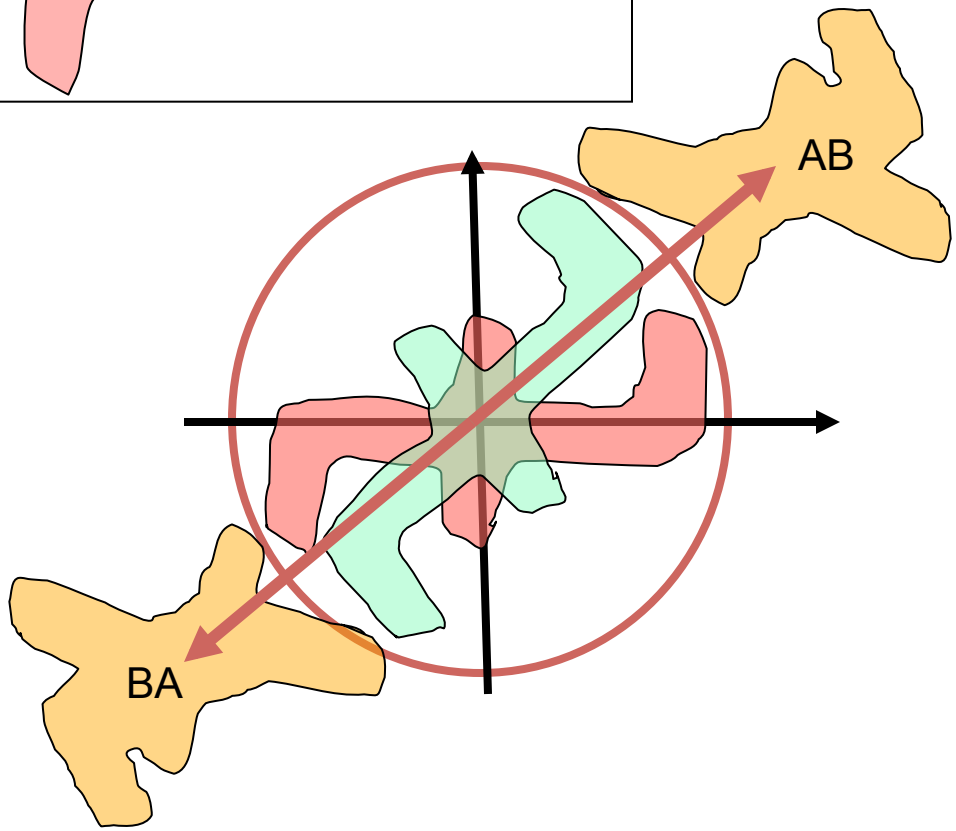
$\Re_{\Omega}$ - rotation operator

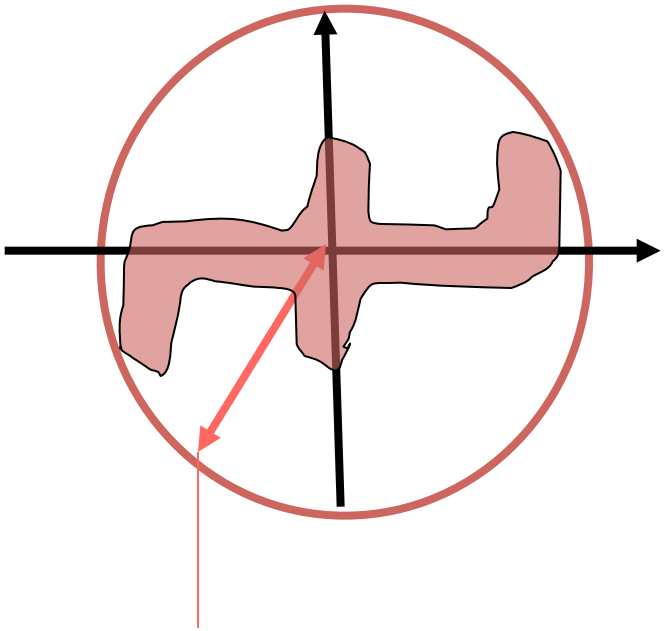


Optimal radius of integration

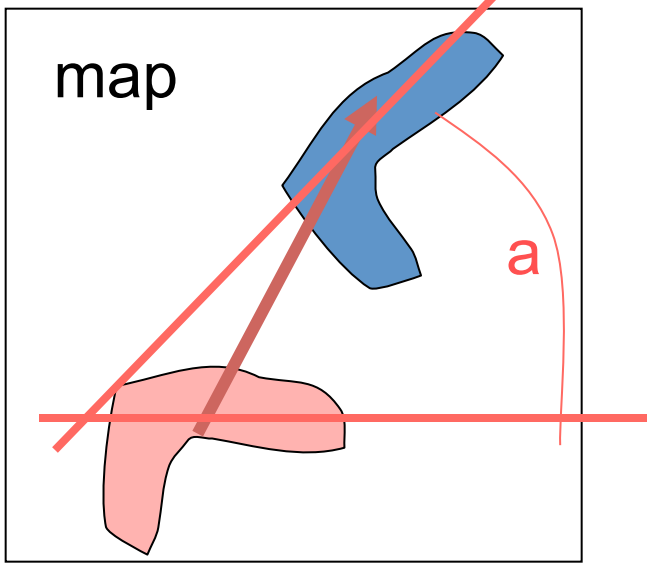




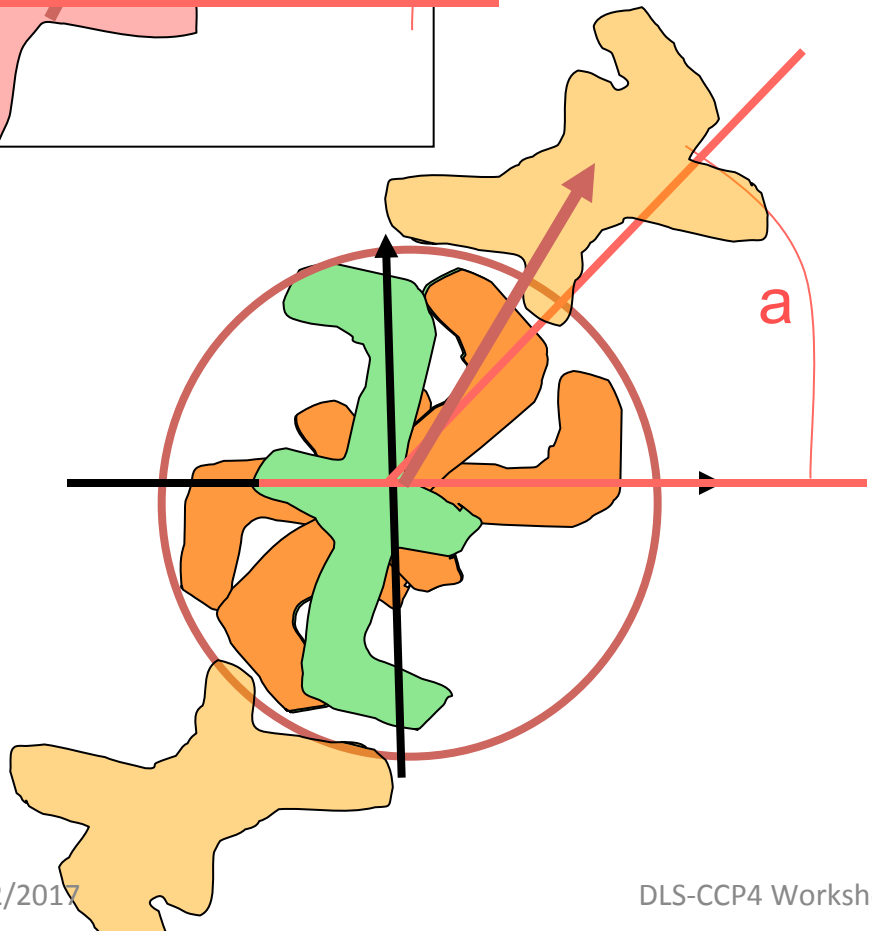


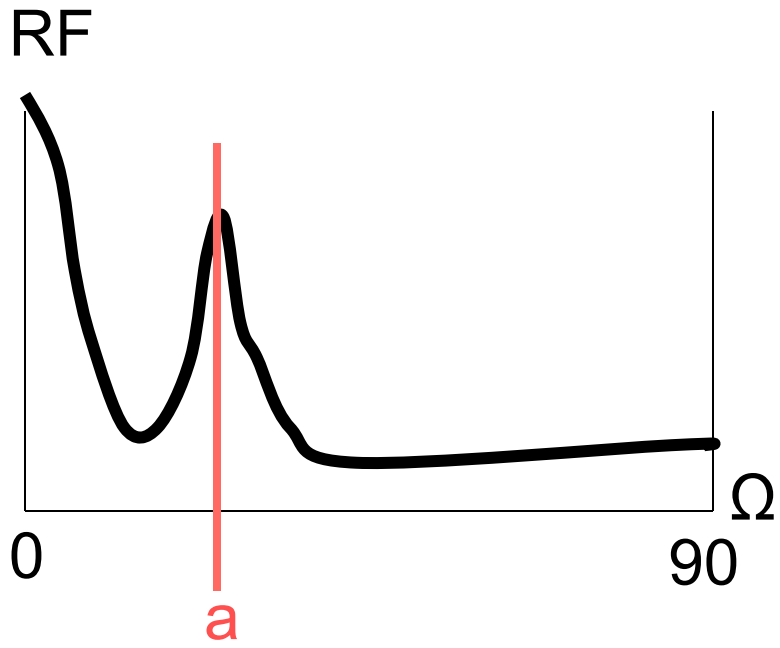


Self Rotation Function

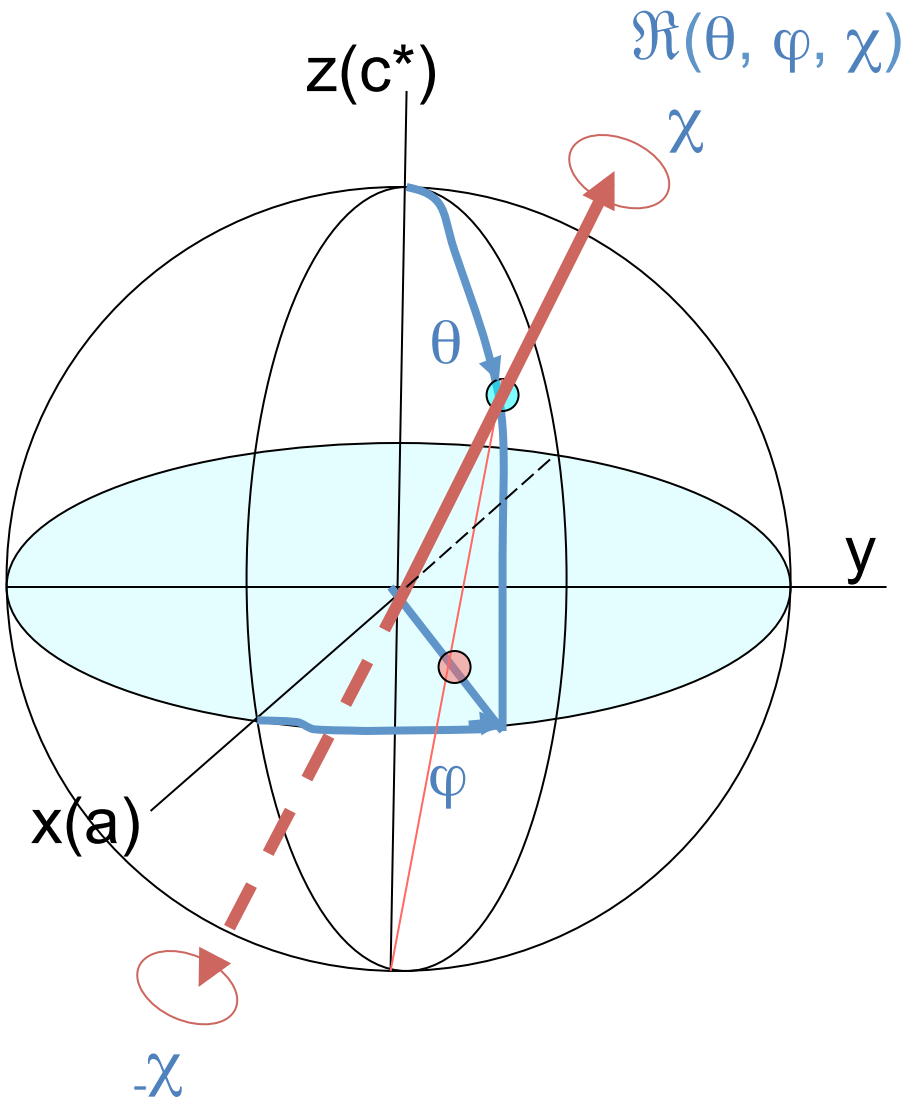


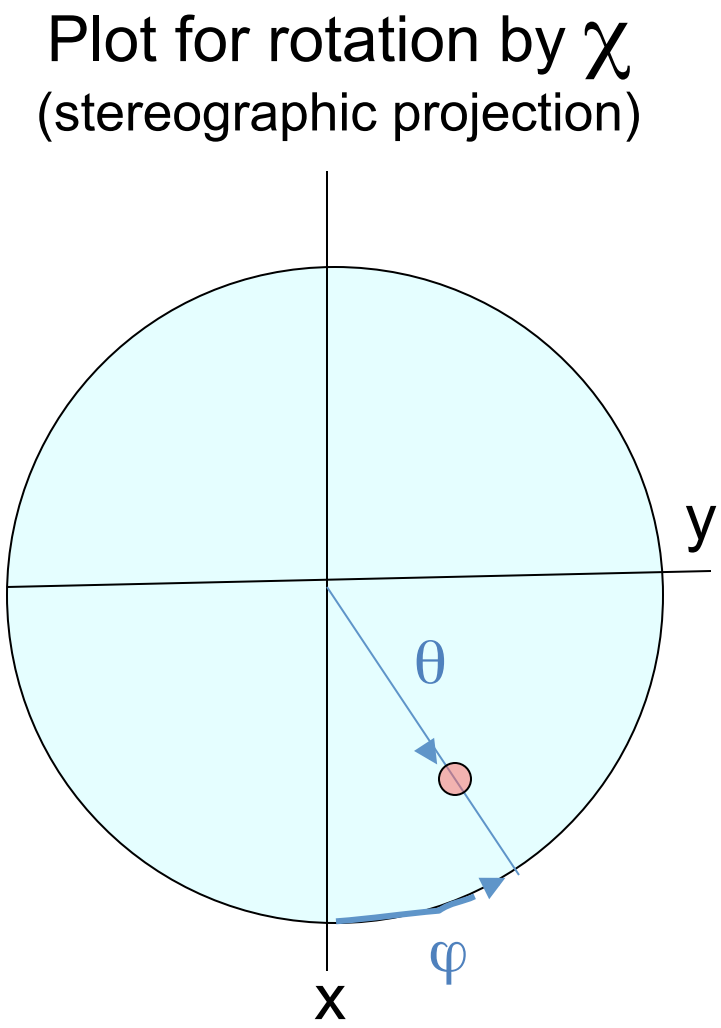
$$RF(\Omega) = \int P_{obs}(r)\ \ \mathfrak{R}_{\Omega}\{ P_{obs}(r)\}\ \ dr$$





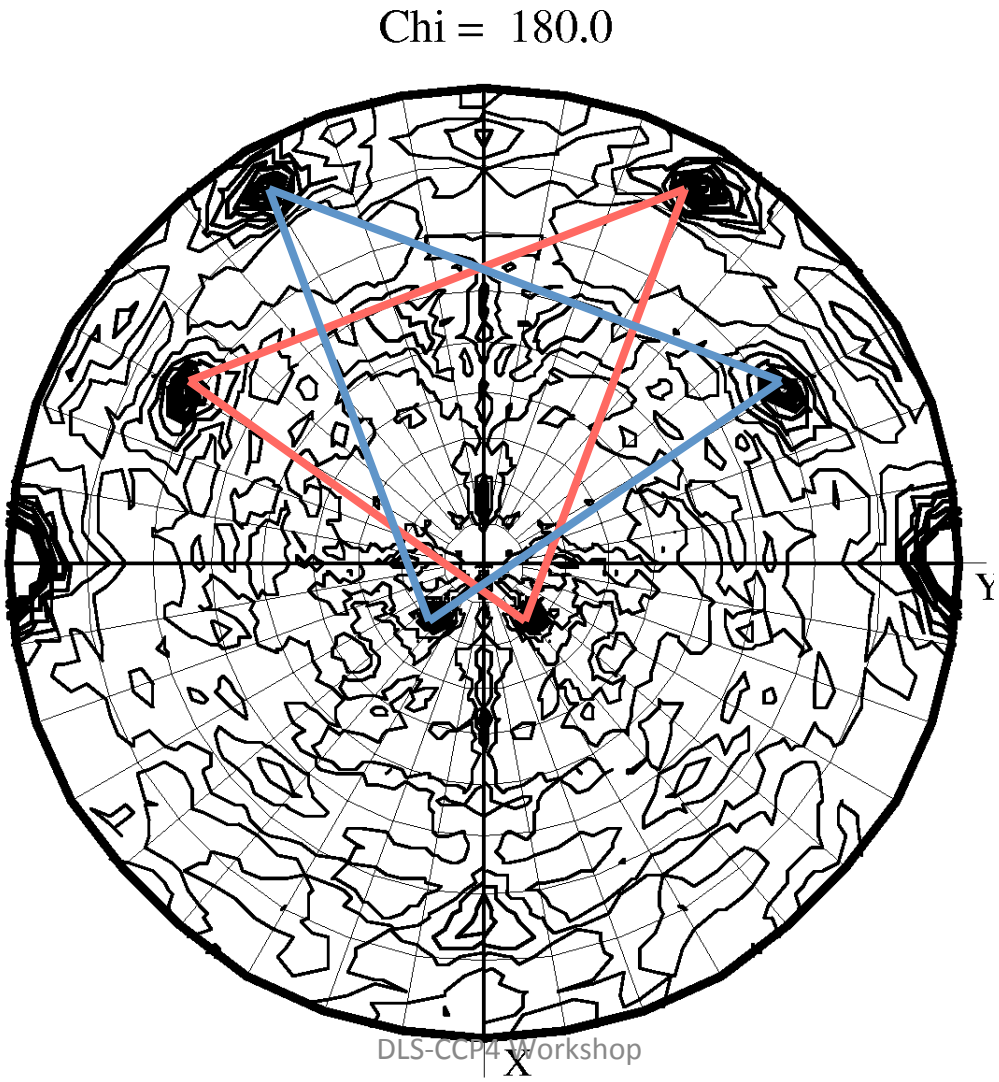
Self Rotation Function

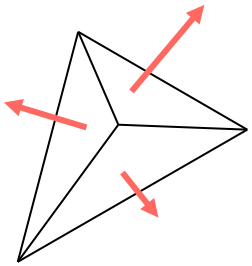


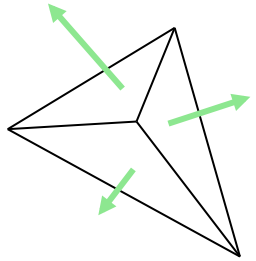


Self Rotation Function

Space group $P2_1$ one tetramer







Translation function

$$TF(\mathbf{t}) = \iiint P^{\text{obs}}(\mathbf{r}) \times P^{\text{calc}}_{\text{cross}}(\mathbf{t},\mathbf{r})\,\mathrm{d}\mathbf{r}^3$$

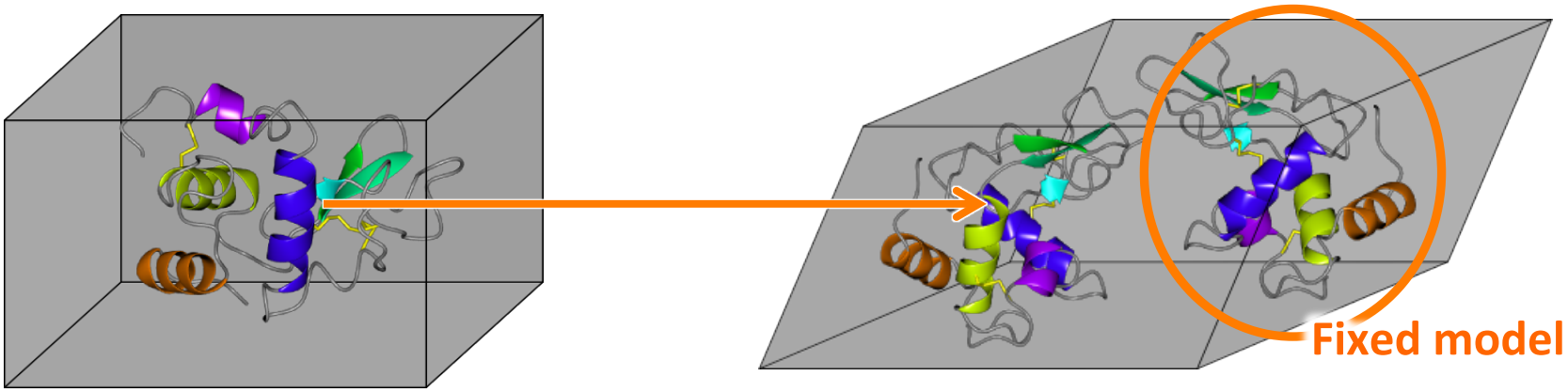
- Conventional (Patterson based) TF:
  - absence of any phases

$$TF(\mathbf{t}) = \iiint \rho^{\text{obs}}(\mathbf{r}) \times \rho^{\text{calc}}(\mathbf{t},\mathbf{r})\,\mathrm{d}\mathbf{r}^3$$

- Phased TF:
  - (poor) experimental phases
  - phases from partial model
  - Fitting into EM reconstruction

t - vector of translation

Fixed partial model in Patterson search



Almost the same equation as for a single molecule search,

$$TF(\mathbf{t}) = \sum_{\mathbf{h}} I_{\mathbf{h}}^{\text{obs}} \times \left| F_{\mathbf{h}}^{\text{fixed}} + F_{\mathbf{h}}^{\text{calc}}(\mathbf{t}) \right|^2$$

Translation function

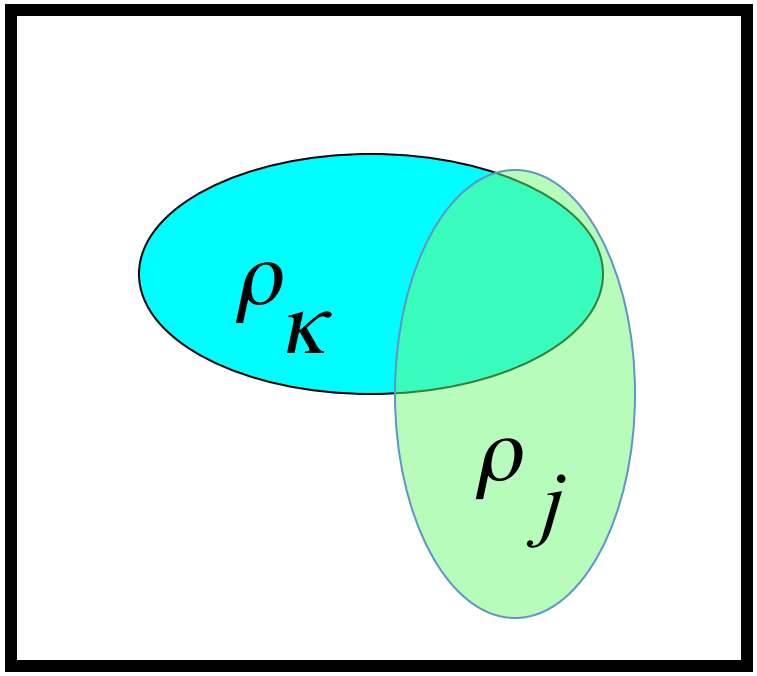
Patterson based TF

- Space group P1
  - » First copy: no translation function needed
  - » All subsequent copies: 3-d translation function
- Polar space groups (P2, $P2_1$, C2, P3, $P3_1$, $P3_2$, …, $P6_5$)
  - » First copy: translation function is 2-dimensional
  - » All subsequent copies: 3-d translation function
- Other space groups
  - » All copies: 3-d translation function

Phased TF

- » All copies: 3-d translation function

Fast Packing Function



Estimation of overlap:

$$\int \rho_k(\mathbf{r},\mathbf{s})\,\rho_j(\mathbf{r},\mathbf{s})\,d\mathbf{r}$$

Packing function:

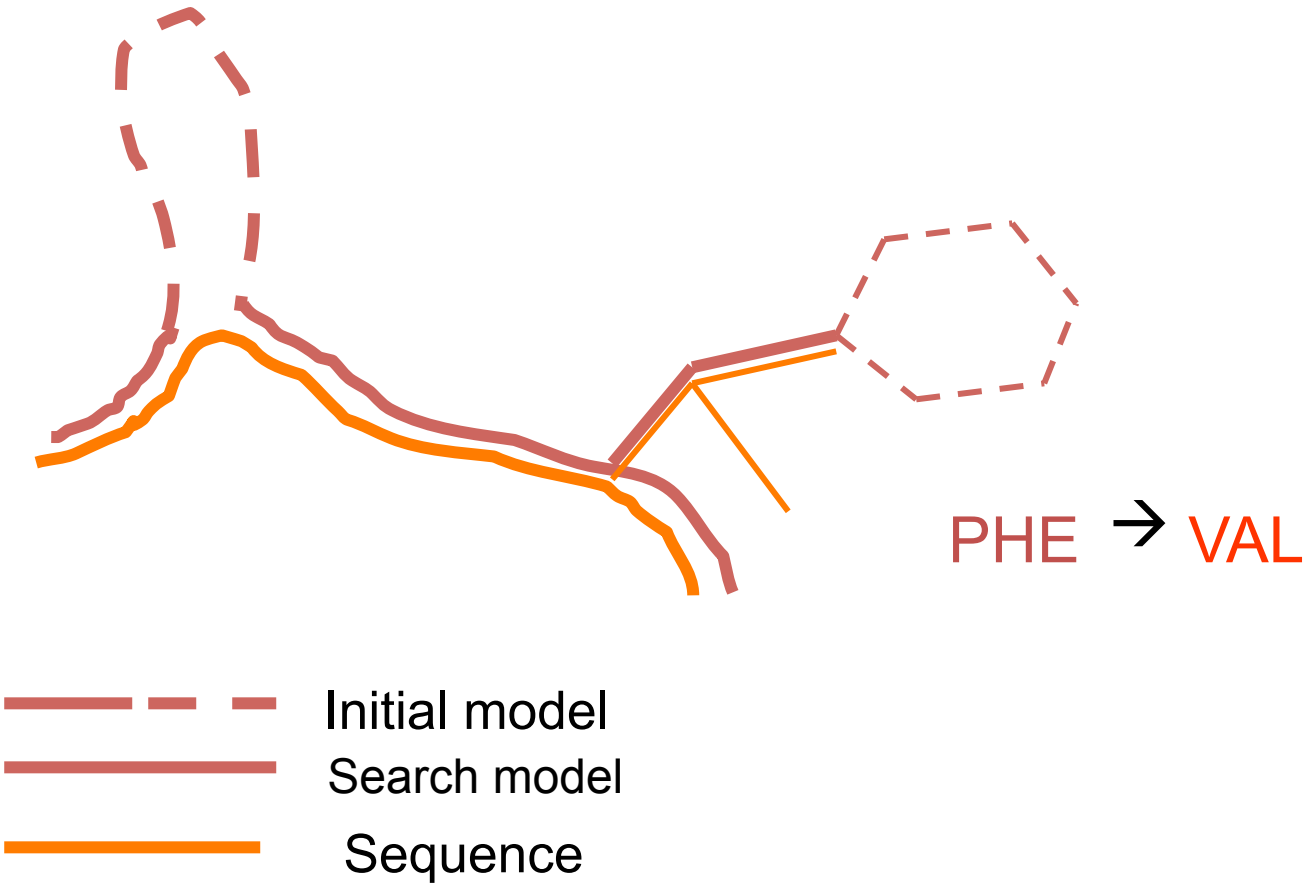
$$P(\mathbf{s}) = \mathbf{1} - \sum_k \sum_{j \neq k} \int \rho_k(\mathbf{r},\mathbf{s})\,\rho_j(\mathbf{r},\mathbf{s})\,d\mathbf{r}$$

Model modification in MOLREP

```
molrep -f data.mtz -m model.pdb -s target.seq
```

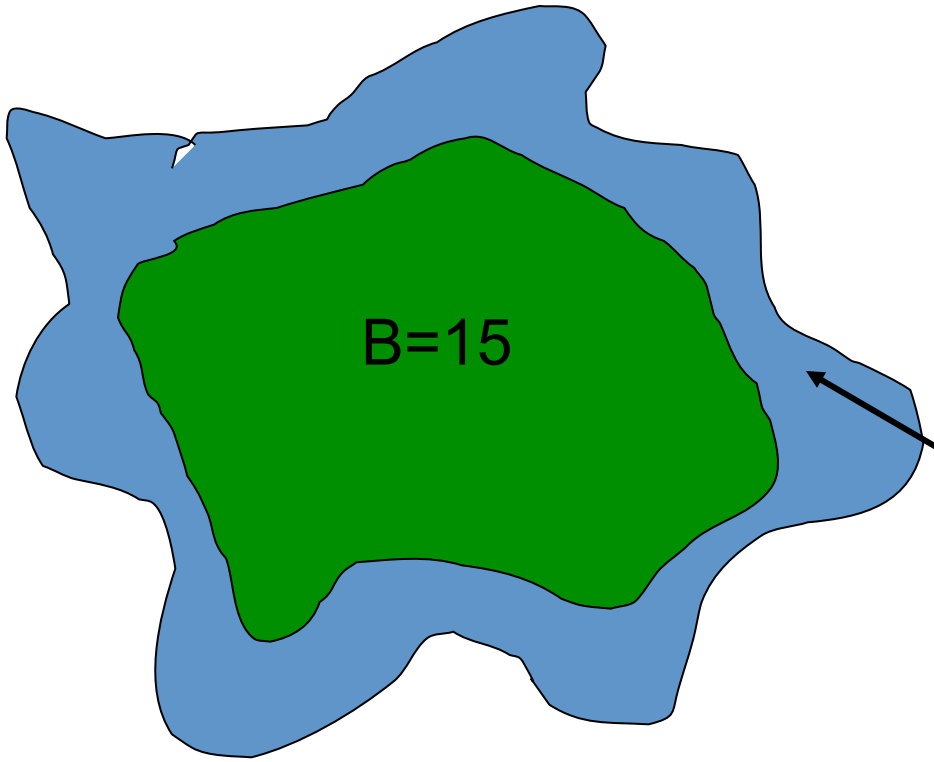
- Performs model correction:
  - Identifies secondary structure in the model
  - Aligns target and model sequences
    - » no deletions or insertions in α-helixes or β-strands
  - Retains aligned residues
  - Retains "aligned" atoms in aligned residues
- Adds B-factor to residues exposed to solvent
- Uses sequence identity to down-weight high resolution data

Automatic correction of the model using sequence alignment



Model improvement

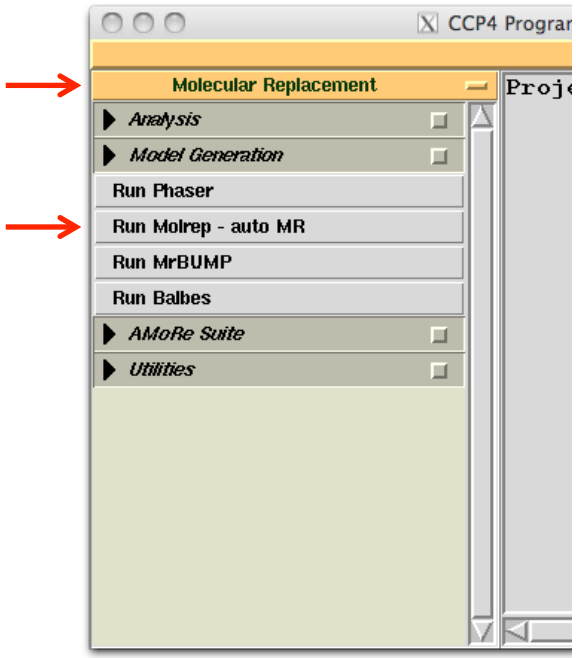
Set atomic B values according to accessible surface area

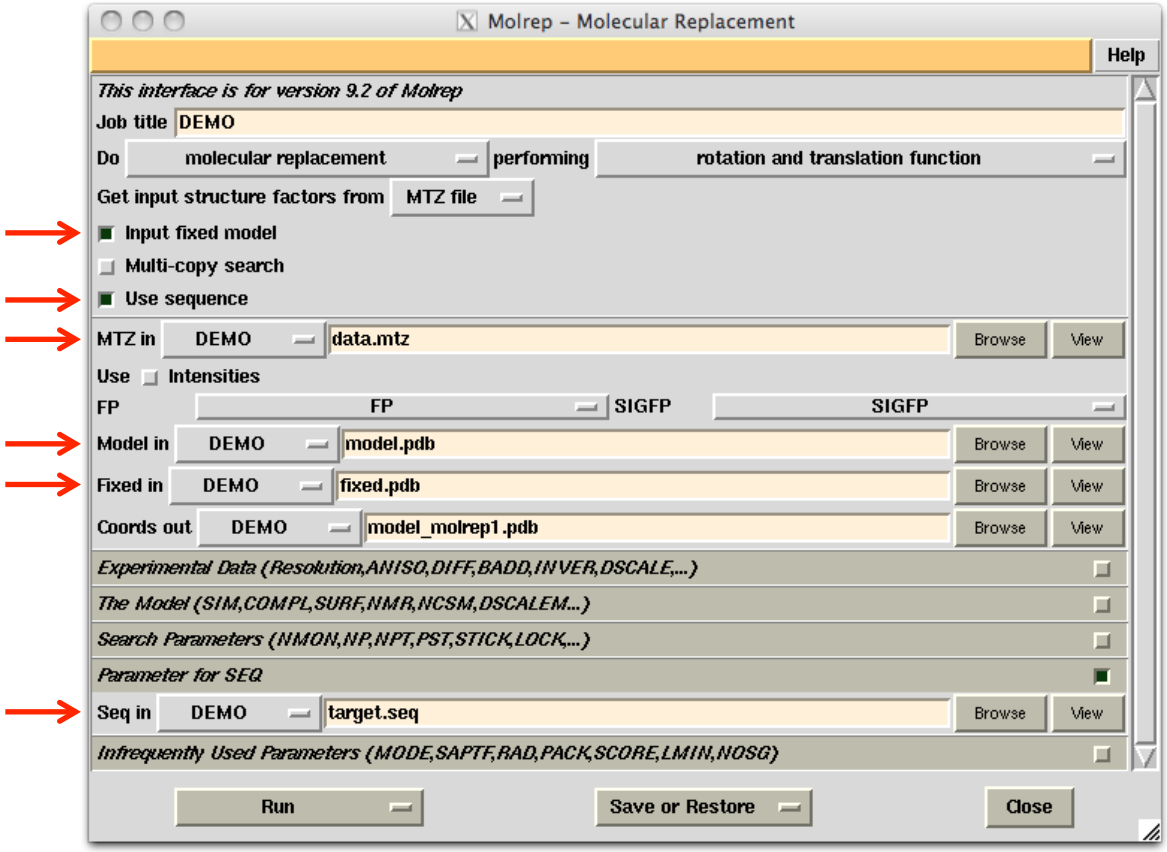


$B\ (A^2) = 15\ (A^2) + 10\ SA\ (A^2)$

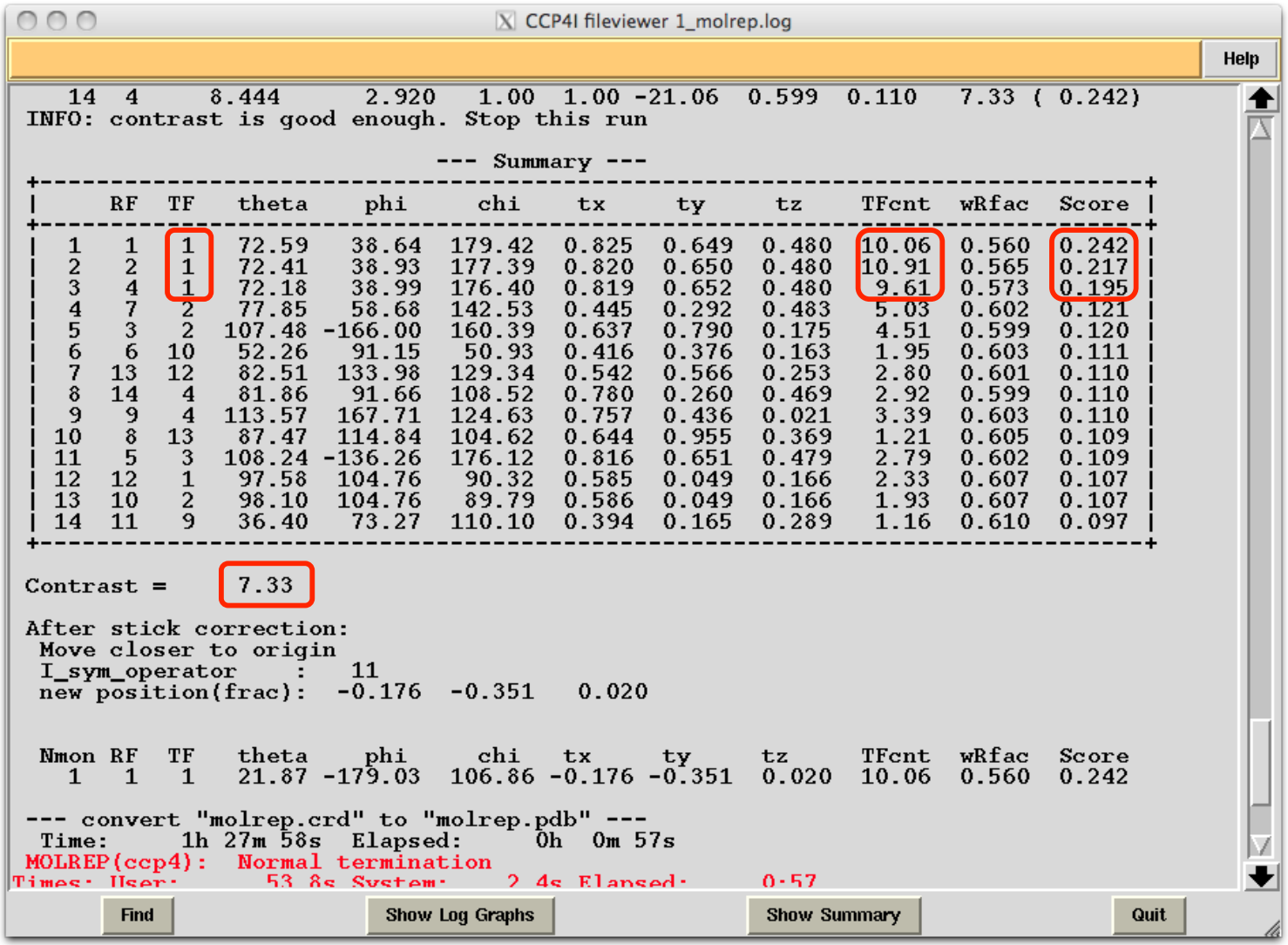
SA – surface area

Molrep in CCP4I

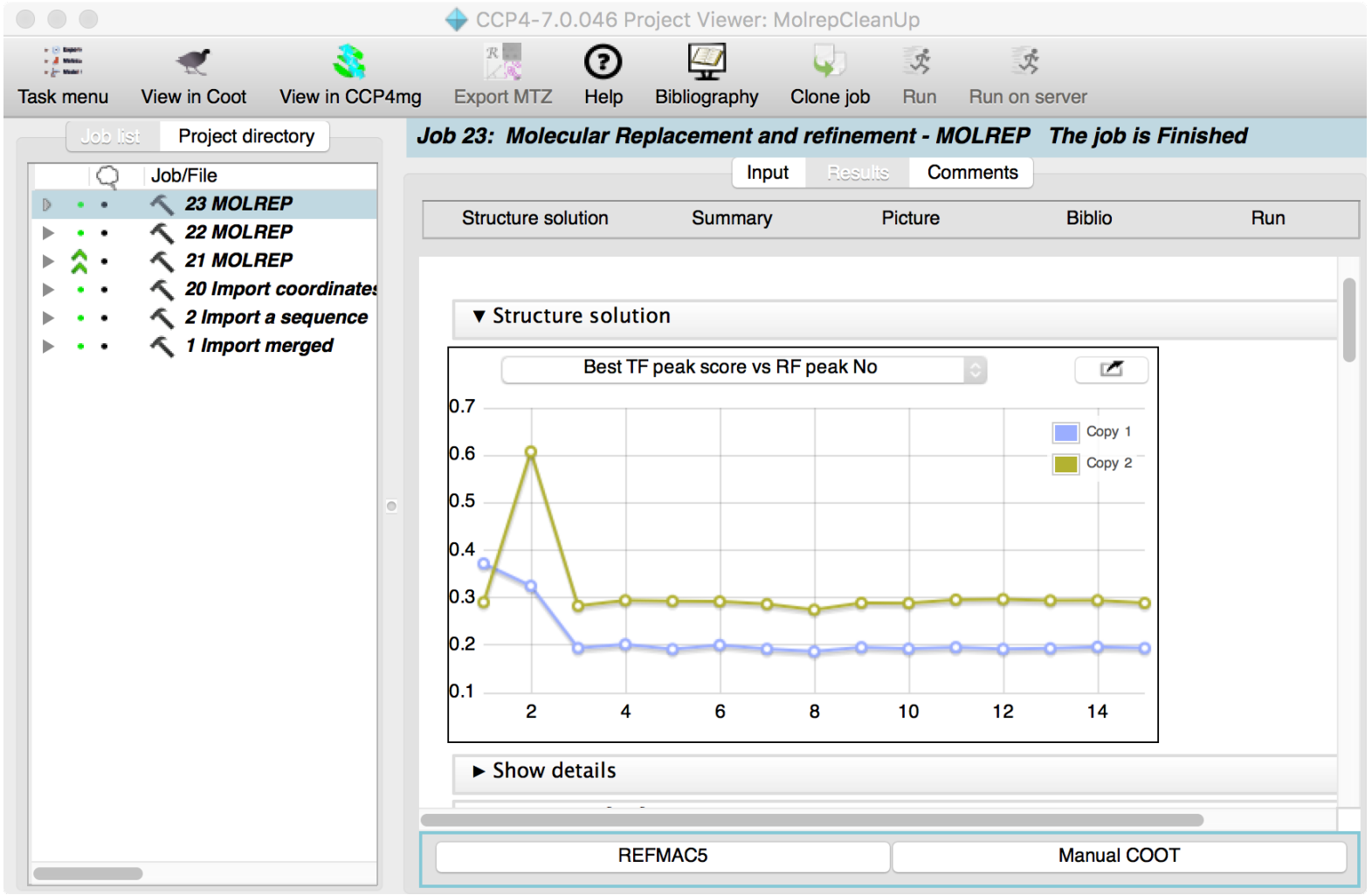




Log-file


CCP4I fileviewer 1_molrep.log
Help
14 4 8.444 2.920 1.00 1.00 -21.06 0.599 0.110 7.33 (0.242)
INFO: contrast is good enough. Stop this run
--- Summary ---
| RF TF theta phi chi tx ty tz TFcnt wRfac Score |
| 1 1 1 72.59 38.64 179.42 0.825 0.649 0.480 10.06 0.560 0.242 |
| 2 2 1 72.41 38.93 177.39 0.820 0.650 0.480 10.91 0.565 0.217 |
| 3 4 1 72.18 38.99 176.40 0.819 0.652 0.480 9.61 0.573 0.195 |
| 4 7 2 77.85 58.68 142.53 0.445 0.292 0.483 5.03 0.602 0.121 |
| 5 3 2 107.48 -166.00 160.39 0.637 0.790 0.175 4.51 0.599 0.120 |
| 6 6 10 52.26 91.15 50.93 0.416 0.376 0.163 1.95 0.603 0.111 |
| 7 13 12 82.51 133.98 129.34 0.542 0.566 0.253 2.80 0.601 0.110 |
| 8 14 4 81.86 91.66 108.52 0.780 0.260 0.469 2.92 0.599 0.110 |
| 9 9 4 113.57 167.71 124.63 0.757 0.436 0.021 3.39 0.603 0.110 |
| 10 8 13 87.47 114.84 104.62 0.644 0.955 0.369 1.21 0.605 0.109 |
| 11 5 3 108.24 -136.26 176.12 0.816 0.651 0.479 2.79 0.602 0.109 |
| 12 12 1 97.58 104.76 90.32 0.585 0.049 0.166 2.33 0.607 0.107 |
| 13 10 2 98.10 104.76 89.79 0.586 0.049 0.166 1.93 0.607 0.107 |
| 14 11 9 36.40 73.27 110.10 0.394 0.165 0.289 1.16 0.610 0.097 |
Contrast = 7.33
After stick correction:
Move closer to origin
I_sym_operator : 11
new position(frac): -0.176 -0.351 0.020
Nmon RF TF theta phi chi tx ty tz TFcnt wRfac Score
1 1 1 21.87 -179.03 106.86 -0.176 -0.351 0.020 10.06 0.560 0.242
--- convert "molrep.crd" to "molrep.pdb" ---
Time: 1h 27m 58s Elapsed: 0h 0m 57s
MOLREP(ccp4): Normal termination
Find
Show Log Graphs
Show Summary
Quit

CCP4I2


CCP4-7.0.046 Project Viewer: MolrepCleanUp
Task menu
View in Coot
View in CCP4mg
Export MTZ
Help
Bibliography
Clone job
Run
Run on server
Job list
Project directory
Job/File
23 MOLREP
22 MOLREP
21 MOLREP
20 Import coordinates
2 Import a sequence
1 Import merged
Job 23: Molecular Replacement and refinement - MOLREP The job is Finished
Input
Results
Comments
Structure solution
Summary
Picture
Biblio
Run
▼ Structure solution
Best TF peak score vs RF peak No
Copy 1
Copy 2
0.7
0.6
0.5
0.4
0.3
0.2
0.1
2
4
6
8
10
12
14
▶ Show details
REFMAC5
Manual COOT

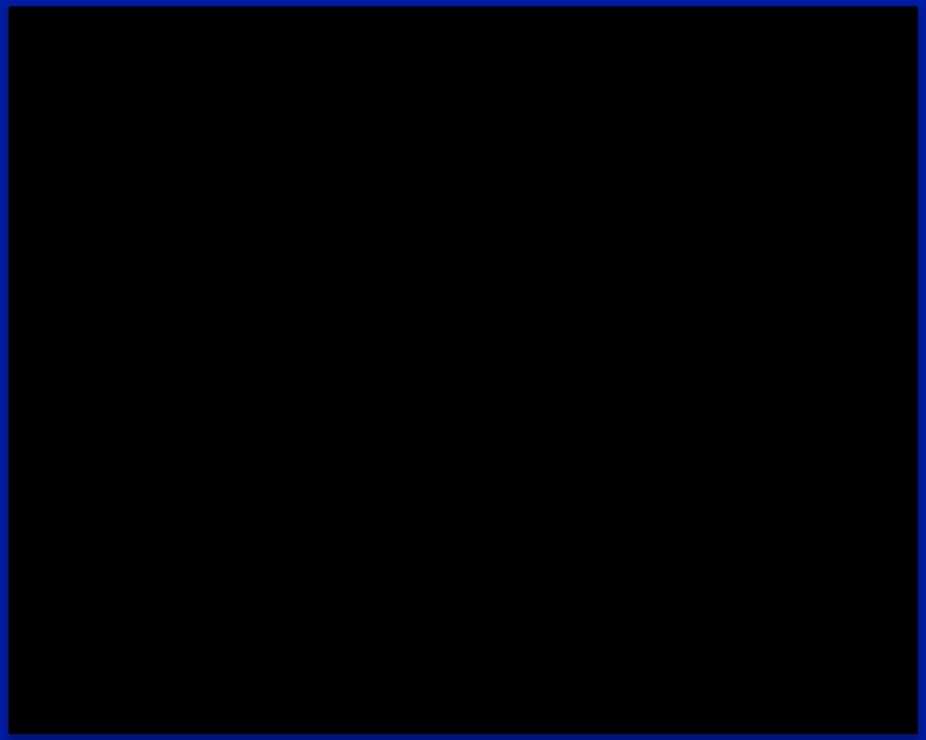
Structure solution: User's viewpoint

Structural biologists often present difficult cases solved by a combination of techniques and unusual approaches to combat problems and pathologies such as twinning, wrong SG/origin assignment, OD-structure, disorder of large parts of the structure etc.

These examples are highly educational.
The fact that well over half of the structures can be (are being) solved by 'a black box' MR or EP pipelines is often underreported.

A Black Box crystallography pipeline

Data, sequence



Refined Model

User does not have to know what is inside

Practical approach to phasing : always start with pipelines.

CCP4 suite contains a number of pipelines both for MR and EP phasing. These are available from CCP4 online and two interfaces – ccp4i and ccp4i2.

Time spent on solving the structure without trying pipelines is wasted unless a learning element is involved. Even for learners it may be more useful to analyze the solution of a finished structure (post-mortem) than waste time trying to unsuccessfully solve a challenging case.

I have tried a number of phasing pipelines in the three CCP4 interfaces.

CCP4 Phasing pipelines.

MR Pipelines

MRBump	CCP4 online, CCP4i, CCP4i2 (ca 150 publications in WoS)
BALBES	CCP4 online, CCP4i (400 publications)
MORDA	CCP4 online, CCP4i, CCP4i2

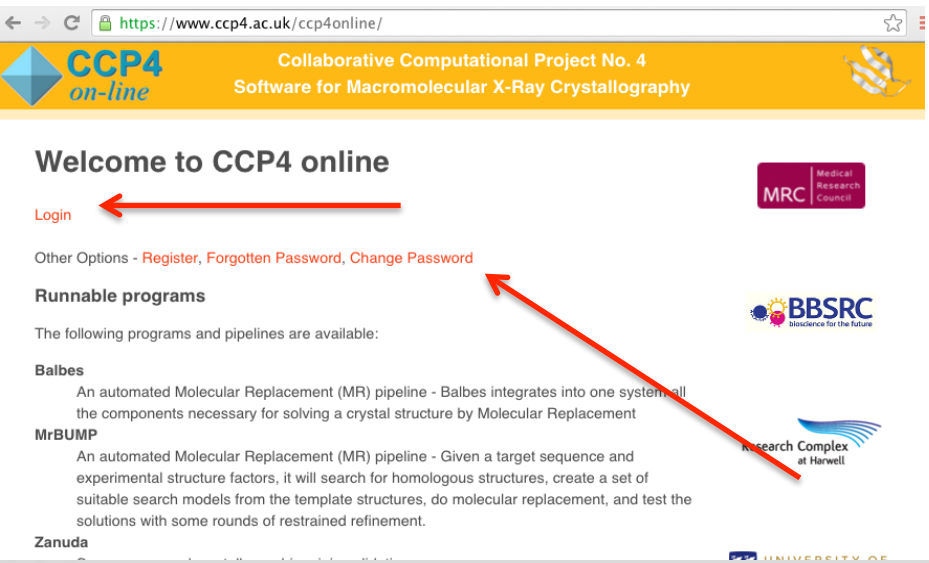
Experimental phasing pipelines

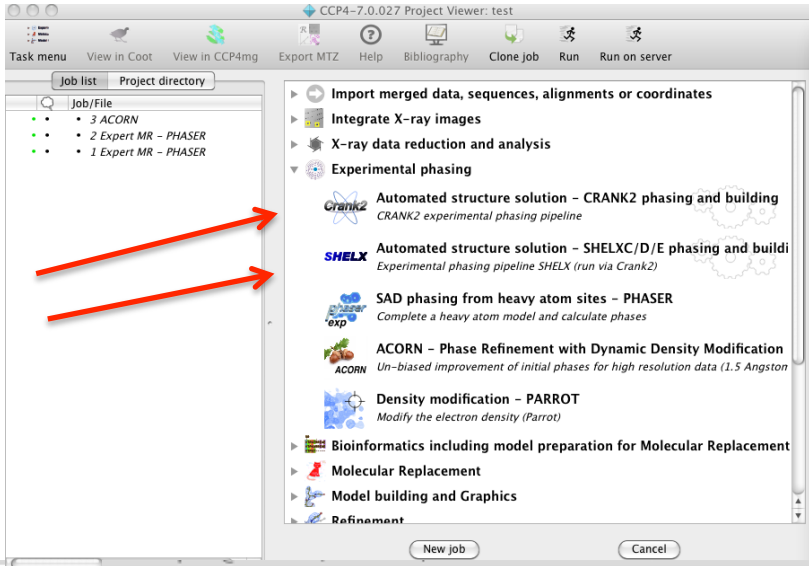
CRANK2	CCP4 online, CCP4i (SAD), CCP4i2
SHELX (via Crank2)	CCP4 online, CCP4i2

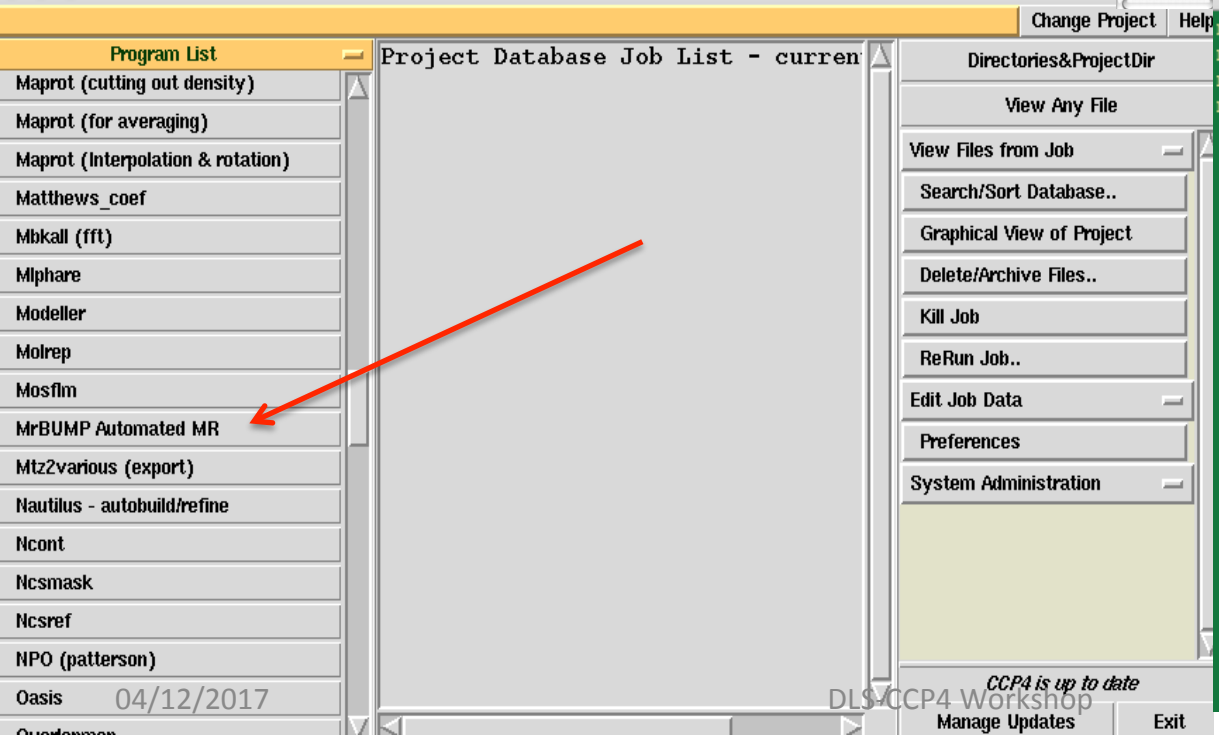
Ab-initio/exhaustive phasing pipelines for difficult cases (not discussed in this talk)

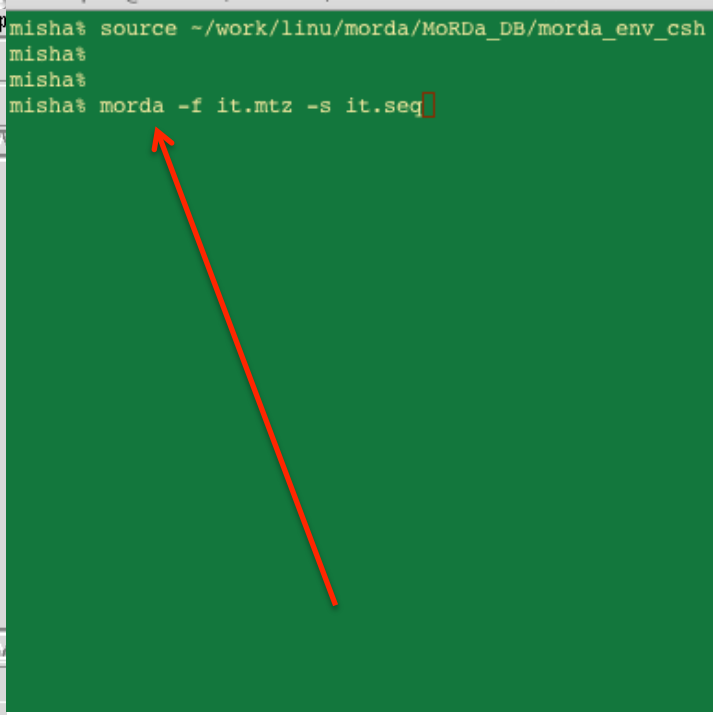
AMPLE	CCP4 online, CCP4i, CCP4i2
ARCHIMBOLDO	CCP4i
SIMBAD	CCP4 online, CCP4i

CCP4 interfaces









MR pipelines in CCP4.

- Around 80% of new PDB X-ray entries are reported to be solved by MR, less than 10 % by experimental phasing.
- Therefore once data for a project become available there is a very good chance that the protein structure can be solved by straight MR.
- A putative MR solution needs to be validated/finalised by the downstream refinement, this may require data/model manipulations which may be not easy/trivial for a beginner structural biologist.
- MR phasing pipelines refine putative solutions for several models in all candidate space groups and earmark the best of them .

MR pipelines.

- MR Phasing pipelines approach structure solution by using several models (including oligomers and separate domains) and model assemblies.

- The search parameters are optimized on many cases, therefore these pipelines are very efficient and sometimes it is challenging even to reproduce the pipeline results by running stand-alone MR programs.

MR pipeline test.

- To test existing MR pipelines in CCP4, data and sequences of 5 unpublished (not present in PDB) medium–sized structures with better than 2 Å resolution were submitted to each of CCP4 MR pipelines (MrBump, BALBES, MORDA) on CCP4 online server. Three of these cases have medium difficulty and two are easy.

- All three pipelines have found the solution and built the correct model for each project, as can be verified by final R-factors of the refined/partially refined models.

- Presented are important points of these structure determinations.

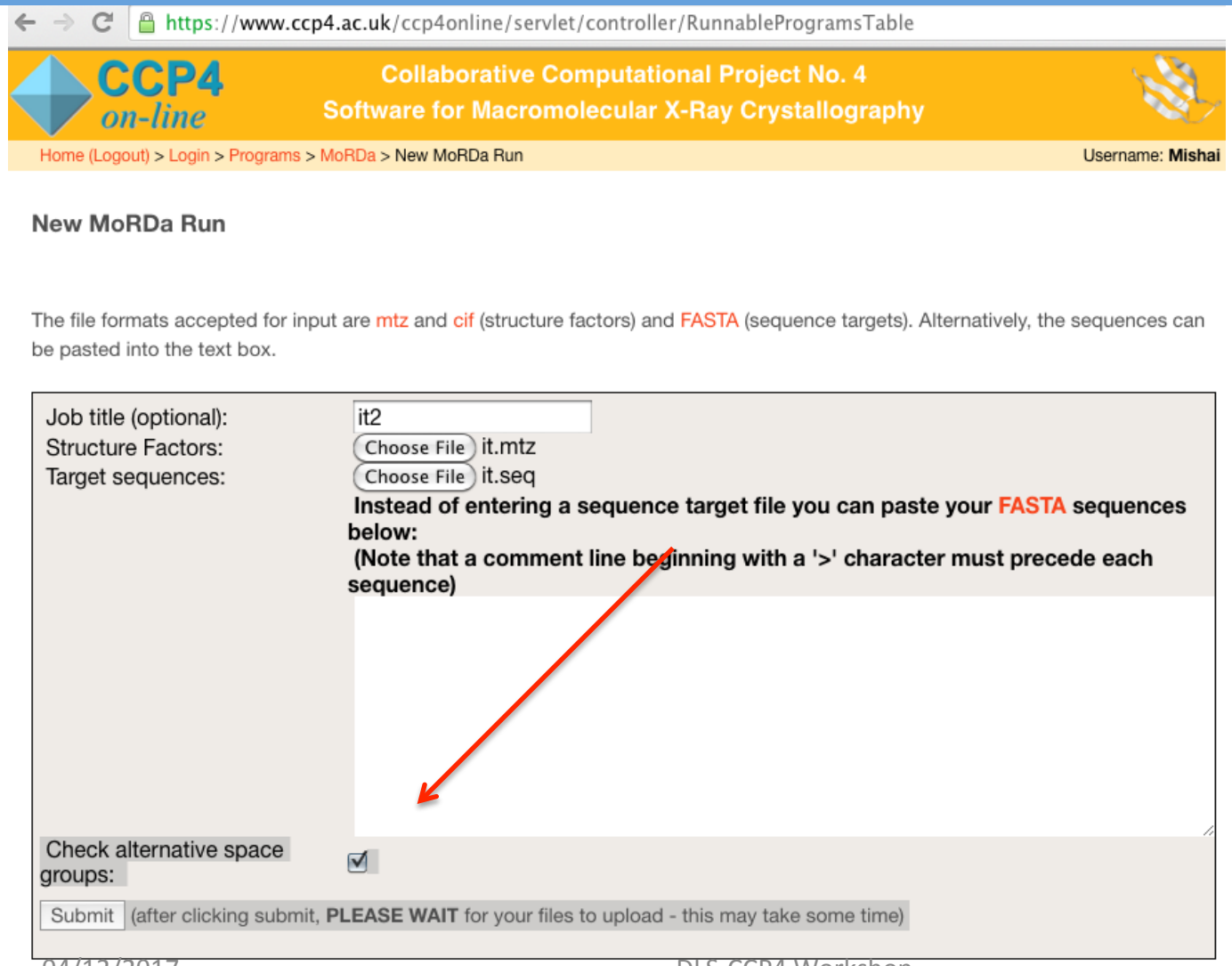
MR pipeline trial proteins.

Project	Resolution Å	Highest identity model %	Space group	Unit cell a,b,c Å, α, β, γ °	Monomers in a.u.
Serine hydroxy-methyl transferase *Thermoanaerobacter*	1.6	70	$P2_1$	49.7, 108.2, 73.5 90, 90.3, 90	2 x 413 aa
Metagenomic epoxide hydrolase Sample Tomsk55	1.6	31	$C222_1$	41.2, 84.2, 157.5, 90, 90, 90	1 x 297 aa
Metagenomic epoxide hydrolase Sample China65	1.4	34	C2	163.9, 46.2, 73.9, 90, 106.9, 90	2 x 293 aa
Sugar transaminase *Archaeoglobus*	1.3	53	$P2_12_12_1$	70.2, 105.9, 111.2 90, 90, 90	2 x 371 aa
Sugar transamina-se*Thermoanaerobacter*	1.5	39	$P3_221$	76.4, 76.4, 134.4 90, 90, 120	1 x 375 aa

Treatment of space group ambiguity in CCP4 phasing pipelines.

- Space group assignment is done at the data integration/scaling stage (POINTLESS).
- Enantiomorph ambiguity arises when systematic absences used for the space group assignment, are the same for two distinct space groups, e.g for both P43 and P41 general condition for reflection of type 00*l* is *l=4n, i.e.* both space groups are possible and can only be resolved by phasing.
- Unfavourable orientation of the crystal used for data collection may result in the intensities not being measured along crystallographic axis, e.g. if the reflections 0*k*0 in an orthorhombic space group are not measured, the axis y can be either screw or rotational.
- Pseudo-symmetry may result in pseudo-absences in the diffraction pattern, affecting assignment of a correct space group.
- Straightforward solution is to try phasing in all possible space groups.

Do not forget to tick the box if space group ambiguity is expected. Not present in MrBump online – run jobs for all candidate space groups.



Enantiomorph choice (sugar transaminase)

MoRDa

Please cite the following paper, if you used a solution from MoRDa:
A. Vagin, A. Lebedev, *Acta Cryst.* (2015). A71, s19

MoRDa homepage contains more details and instructions for local installation.

PROCESS **8644678102** HAS ENDED

Check alternative space groups: ☐

Report | Prep | Details

- Job Details
- Input Data
- Best Solution

No	DB code	ens	ident	Nres	Nexp	Nfound	Q	correct	space group
1	3dr4A	+	0.386	363	1	1	0.480		P 31 2 1

PROCESS **7506573332** HAS ENDED

Check alternative space groups: ☑

Report | Prep | Details

- Job Details
- Input Data
- Best Solution

No	DB code	ens	ident	Nres	Nexp	Nfound	Q	correct	space group
1	3dr4A	+	0.386	363	1	1	0.830	***	P 32 2 1

The best solution for the project epoxide hydrolase Tomsk55 by MORDA is based on the PDB entry which is not present in not-renewed BALBES database.

```
#-------------------------------------------------------------------------------#
#                          A structure is suggested by BALBES                   #
#                       Its probability to be a solution is 99.0%               #
#-------------------------------------------------------------------------------#
The solution model index is sq1st3m1
|-------------------------------------------------------------------------------|
| ITS PDB FILE       |                 results/refmac_final_result.pdb          |
|-------------------------------------------------------------------------------|
| ITS MTZ FILE       |                 results/refmac_final_result.mtz          |
|-------------------------------------------------------------------------------|
| R_ini/R_fin        |  0.5460/0.4300  |  Rfree_ini/Rfree_fin  |  0.5450/0.4570 |
|-------------------------------------------------------------------------------|
| ITS Q FACTOR       |                          0.630                           |
|-------------------------------------------------------------------------------|
```

MoRDa homepage contains more details and instructions for local installation.

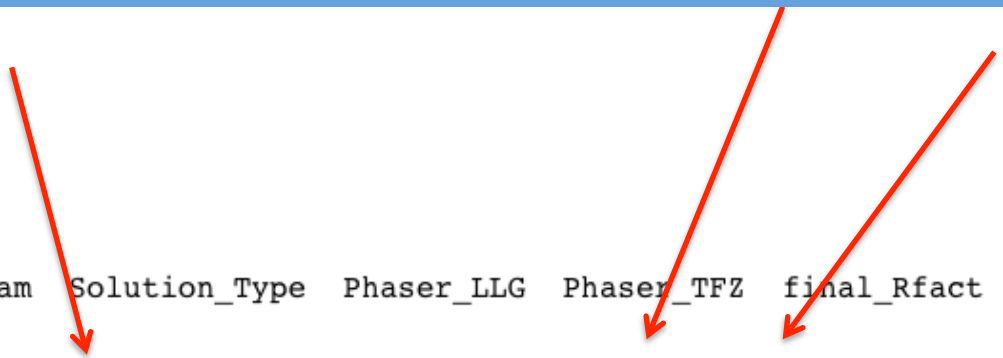
PROCESS **1265910401** HAS ENDED

Report | Prep | Details

Searches for sequence 1

No	DB code	ens	ident	Nres	Nexp	Nfound	Z score	initial R-Rfree	final R-Rfree	Q	correct
1	4inzA	+	0.311	286	1	1	11.687	0.562-0.568	0.369-0.402	0.736	***
2	4inzA_1	+	0.379	195	1	1	11.485	0.562-0.568	0.407-0.443	0.667	***
3	2e3jA	+	0.280	293	1	1	7.740	0.562-0.576	0.448-0.470	0.611	**
4	2e3jA_1	+	0.333	204	1	1	8.262	0.560-0.565	0.450-0.476	0.597	**

MrBump appears to be misleadingly pessimistic in evaluating possible solutions for the project epoxide hydrolase Tomsk55 (31% identity).



$$

Model	Model_Name	MR_Program	Solution_Type	Phaser_LLG	Phaser_TFZ	final_Rfact	final_Rfree	Bucc_final
$$								
1	4inz_A_CHNSAW	MOLREP	MARGINAL	-- --	0.405	0.427	0.258	0.291 -- --	42.790	55
2	4inz_A_SCLPTR	MOLREP	MARGINAL	-- --	0.405	0.429	0.261	0.293 -- --	42.870	34
3	4inz_A_CHNSAW	PHASER	MARGINAL	215.000	14.900	0.406	0.430	0.291 0.313 -- --	43.000	34
4	4inz_A_MOLREP	MOLREP	MARGINAL	-- --	0.407	0.431	0.302	0.335 -- --	42.160	33
5	4inz_A_SCLPTR	PHASER	MARGINAL	215.000	14.900	0.406	0.432	0.286 0.320 -- --	42.460	38
6	4inz_A_MOLREP	PHASER	MARGINAL	181.000	13.500	0.411	0.432	0.268 0.304 -- --	42.420	38
7	4opm_A_CHNSAW	PHASER	MARGINAL	55.000	7.300	0.481	0.513	0.452 0.490 -- --	42.140	33
8	4lxh_A_SCLPTR	PHASER	POOR	42.000	6.800	0.505	0.524	0.504 0.531 -- --	N/A	N/A

Comparison of MR pipelines in CCP4

- The three CCP4 MR pipelines are consistently solving medium difficulty cases and are likely to be mutually complementary in more difficult projects.
- Balbes and Morda have their own internal databases, this speeds up calculations, MrBump relies on external resources.
- Own database may be a weakness, since regular update is required. Balbes database is few years behind. Databases used by MrBump are renewed externally except for the sequence database.
- Parallelisation made the online version of Morda the preferred option.
- To solve a new project or to check whether yours is a challenging one, the data and sequence need to be input in all CCP4 MR pipelines.

How to approach a new project.

1. Submit data to all CCP4 MR phasing pipelines. If a solution is found, proceed to finalize refinement/validation.
2. Not conclusive - try density modification/model rebuilding of putative solutions (Shelxe, Buccaneer, Arp/Warp), especially if FreeR is below 45-48%.
3. Try to address space group/origin mis-assignment (Zanuda).
4. If Q-factors of putative MR solutions are low and FreeR values are well over 50% it is worth trying MR pipelines using data collected on similar crystals which may happen to be better (more ordered, not-twinned, have less radiation damage etc).
5. Consider experimental phasing, crystallisation of separate domains, try stand-alone MR programs (Phaser, Molrep, Amore), ab-initio phasing pipelines or seek help of expert crystallographers.

Acknowledgements

- Alexei Vagin CCP4
- Garib Murshudov LMB
- Andrey Lebedev CCP4
- Ronan Keegan CCP4
- Jennifer Littlechild Exeter