Introduction to Molecular Replacement, Dimple and MrBUMP

Martyn Winn, Daresbury Laboratory

Ronan Keegan, Rutherford Laboratory

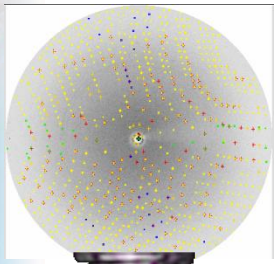
Research Complex at Harwell

BBSRC bioscience for the future

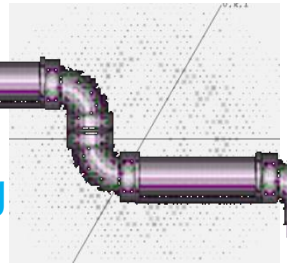
MRC Medical Research Council

Science & Technology Facilities Council

Structure Solution



Data processing



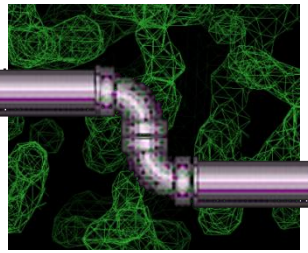
MTZ

28th

29th

Phasing

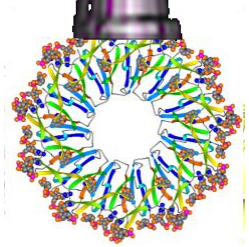
30th



map

Building

31st



CCP4

Simplest case: known structure

e.g. protein-ligand complex, mutant structure

- Known structure (.pdb file)
- Data collected and processed (.mtz file)
- Target structure with same symmetry and cell

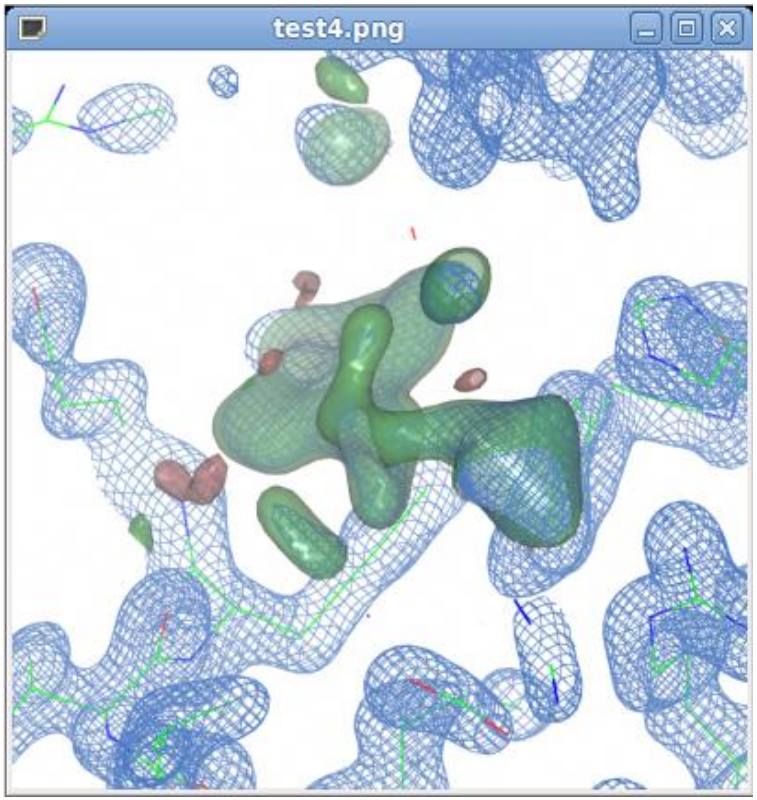
$\Rightarrow$ phase problem already solved

move directly to refinement

DIMPLE

Need to refine known structure against new data. Shows minor changes, e.g. new ligand

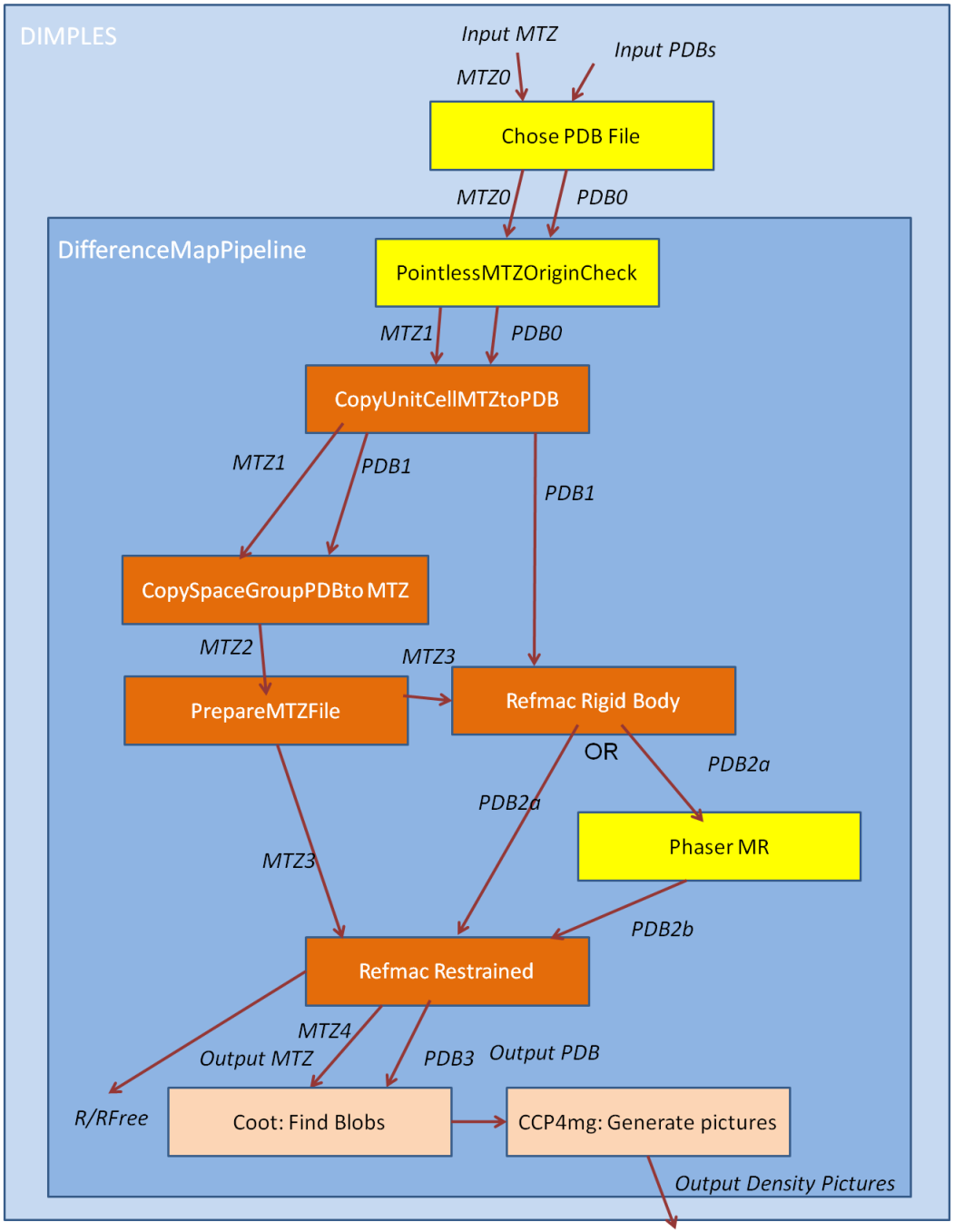
DIMPLE is simple CCP4 pipeline (to be released soon!)



An output map from the DIMPLE pipeline. The blue mesh represents all electron density. The green solid surface indicates an area of positive difference density, i.e. where there is density that is not accounted for by the structure model of the target protein.

$\Rightarrow$ possible location of the bound drug candidate.

CCP4


DIMPLES
Input MTZ
Input PDBs
MTZ0
Chose PDB File
MTZ0
PDB0
DifferenceMapPipeline
PointlessMTZOriginCheck
MTZ1
PDB0
CopyUnitCellMTZtoPDB
MTZ1
PDB1
PDB1
CopySpaceGroupPDBto MTZ
MTZ2
MTZ3
PrepareMTZFile
Refmac Rigid Body
OR
PDB2a
PDB2a
Phaser MR
MTZ3
PDB2b
Refmac Restrained
MTZ4
Output MTZ
PDB3
Output PDB
R/RFree
Coot: Find Blobs
CCP4mg: Generate pictures
Output Density Pictures

Molecular Replacement

1. What if structure not known?

   Similar one?

2. What if target cell/symmetry different?

   Position model in new cell?

Suitable search model

Similar to the target structure (this is assumed to be indicated by sequence similarity)

To represent a large fraction of the target structure (or of specific domains)

Experimental data

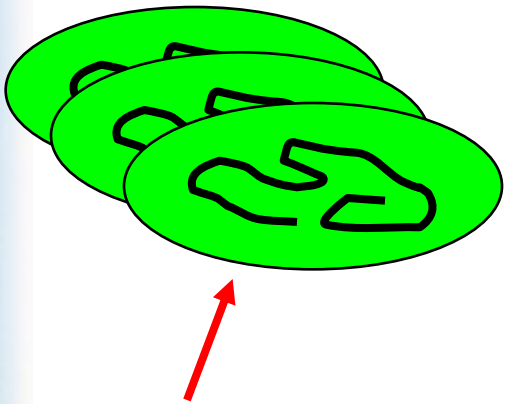
Complete at low resolution to help MR

Extend to high resolution to help refinement

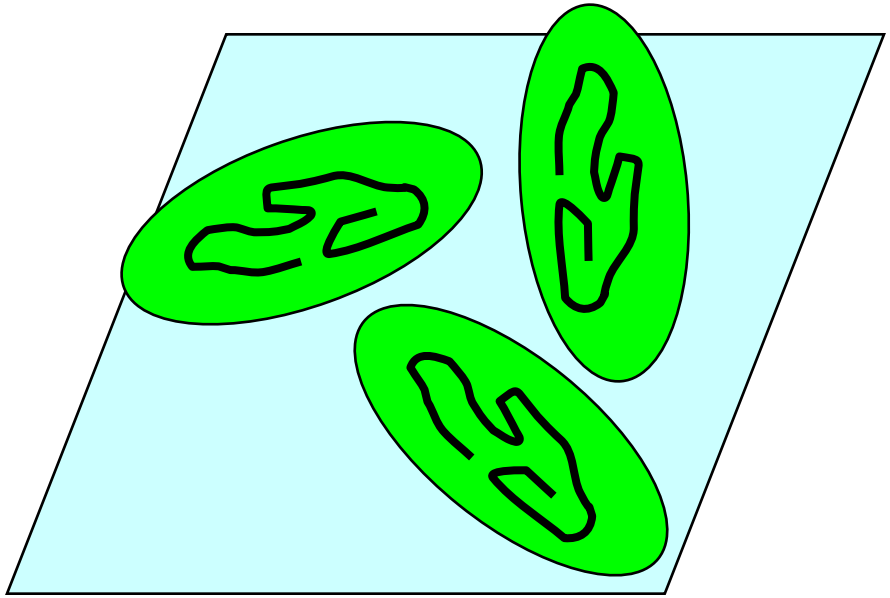
CCP4

MR: What are we trying to do?

Use experimental data to position protein search models in the crystal asymmetric unit



approximate model of protein



Gives initial phases for model correction and refinement

Molecular Replacement in CCP4

- MR Programs:
  - *Molrep*
    - automated molecular replacement given an existing template search model
  - *Phaser (MR)*
    - automated molecular replacement based on using likelihood methods to determine the best solution
- Helper programs:
  - Matthews_coef
  - Sfcheck
  - Polarrfn
  - Chainsaw

MR Pipelines in CCP4

- *MrBUMP*
  - automated template model search and preparation through to MR and refinement
  - Brute force approach – emphasis put on generating many search models
  - Can use both Phaser and Molrep for MR
- *BALBES* - Garib talk tomorrow 
  - automated molecular replacement pipeline using a customised version of the PDB database to provide search models for the MR process using Molrep
  - Support for complexes and searching across all related spacegroups

What do we know from the diffraction data?

- *Some* information on the spacegroup:
  - Cannot distinguish between enantiomorphic spacegroups from diffraction data alone, e.g. P43 vs P41
  - May be unsure about some screw axes *(are 00l, l odd really absent??)*
- The quality of the experimental intensities.
  - Complete? Saturated at low resolution? Anisotropic?
  - Are the intensity statistics reasonable? Could the crystal be twinned?
- Volume of the asymmetric unit, size of the target protein(s) $\Rightarrow$ likely number of molecules it contains (Matthews coefficient).
- Internal arrangement of molecules in asymmetric unit ...

  Non-Crystallographic Symmetry (NCS)

CCP4

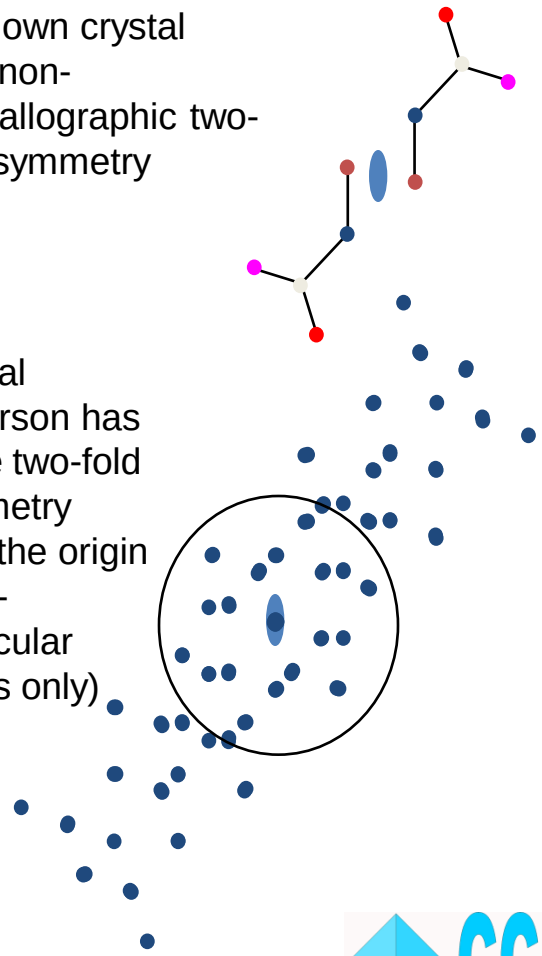
Rotational NCS

Self rotation function from data alone.

If > 1 molecule in the asymmetric unit, then the self rotation function of the Patterson on itself gives a peak at the angle corresponding to the relative rotation between the two.

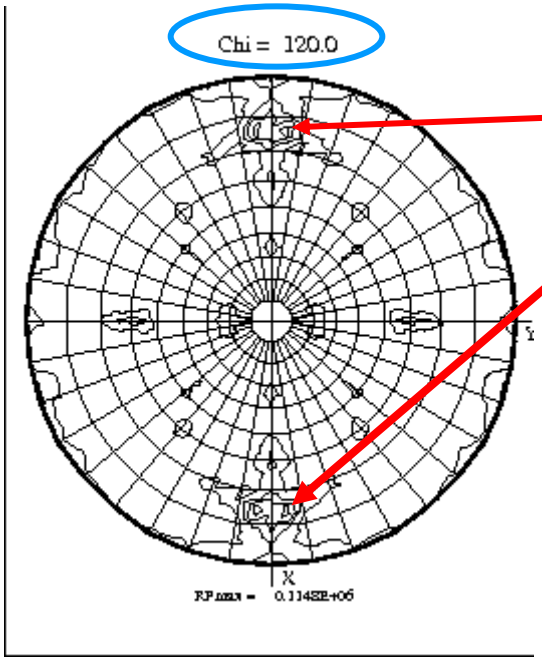
Asymmetric unit of unknown crystal with non-crystallographic two-fold symmetry

Crystal Patterson has same two-fold symmetry near the origin (intra-molecular peaks only)

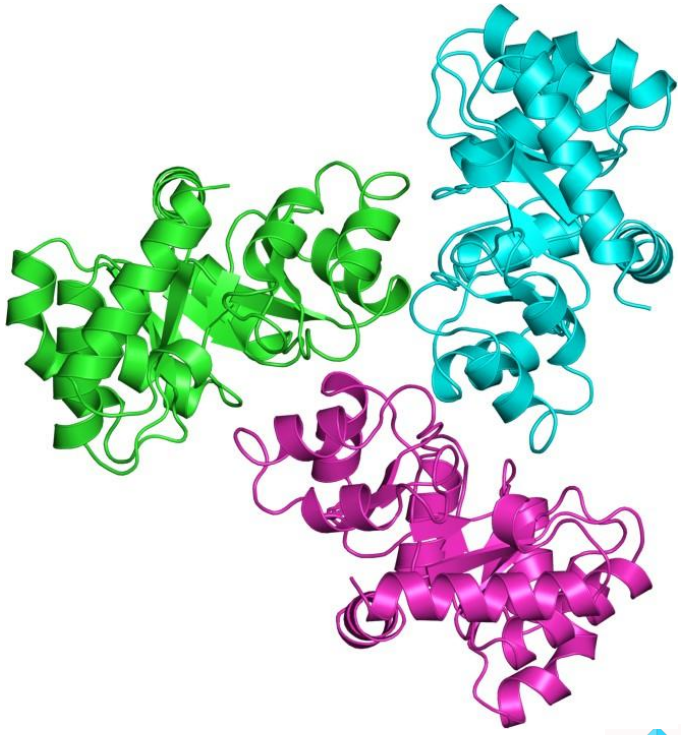


- Confirming how many copies of the structure you have in the asymmetric unit.
- Checking an MR solution
- Used in Locked Rotation Function

Self rotation function for 1vlw

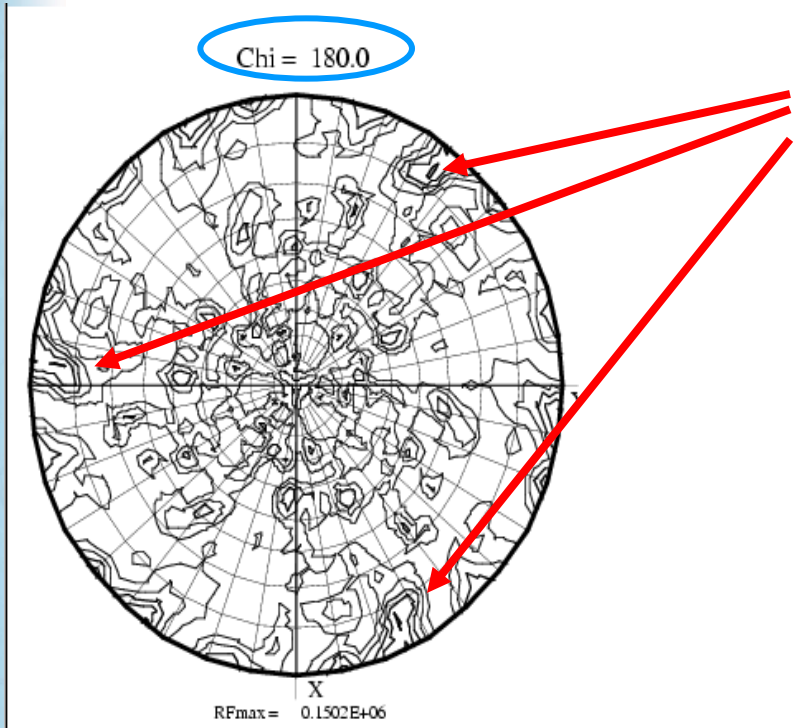


2 symmetry-related 3-folds

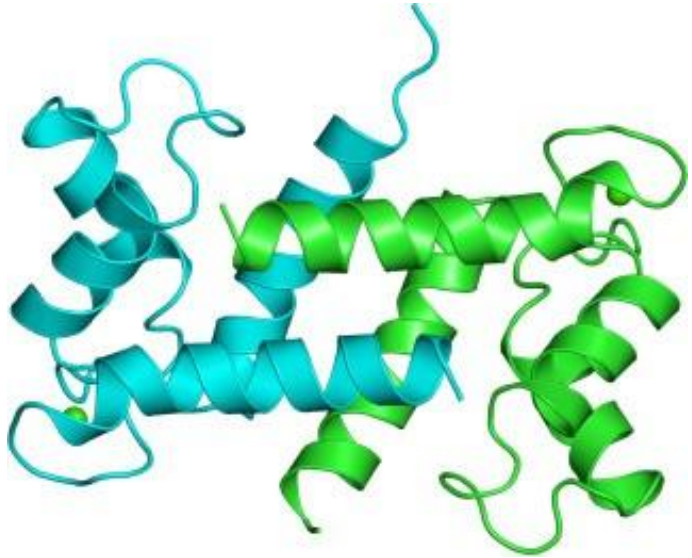


C 2 2 21

Self Rotation Function for S100



3 symmetry related 2-folds



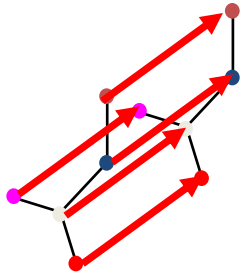
H 3

Translational NCS

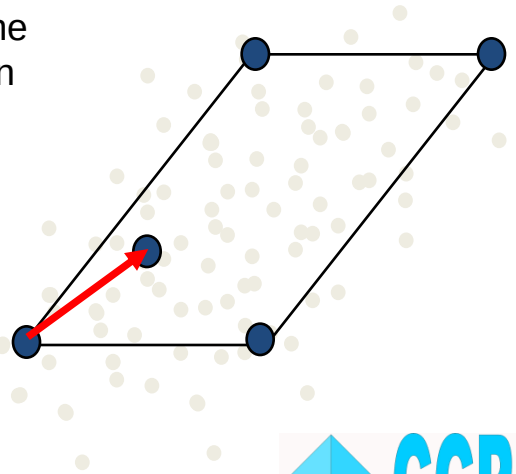
If the asymmetric unit contains two molecules related by a translation, then the native Patterson will have a large peak at the position representing this translation.

NB Patterson always has peak at origin !

Asymmetric unit of unknown crystal structure with non-crystallographic translation.

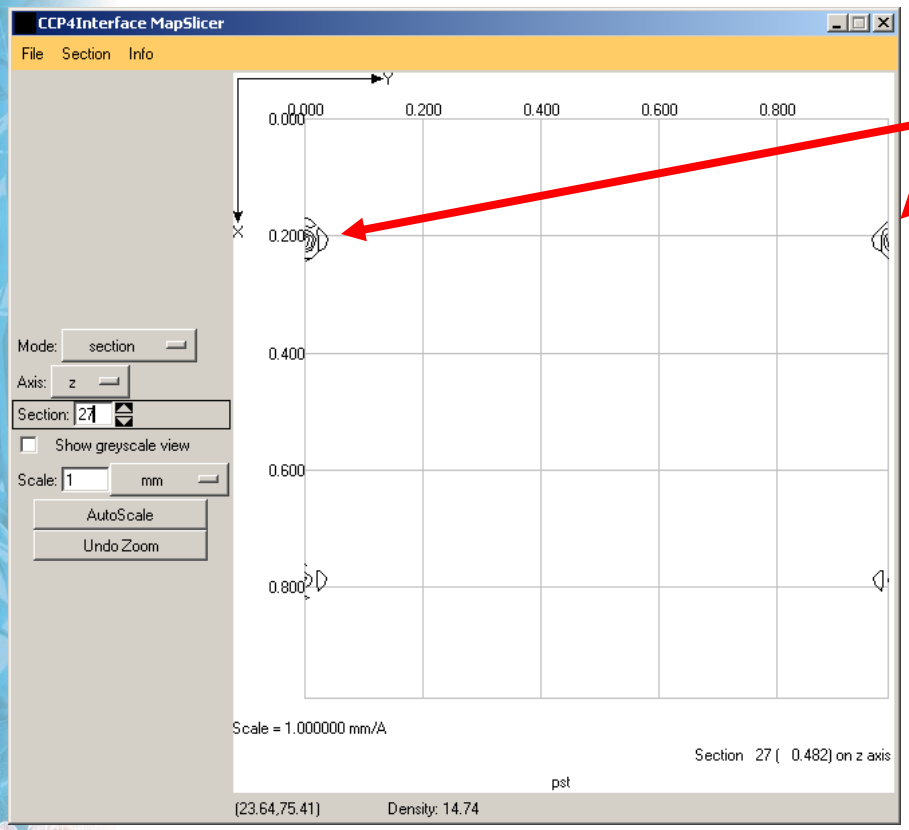


Crystal Patterson has origin sized peak at the translation vector.



Beware, Patterson peaks can also arise from internal regularities, e.g. helices or DNA

Native Patterson for pst



Translational NCS vector
(0.209, 0.000, 0.487)

Peak is 0.286 the height of origin peak.
Molrep looks for peaks higher than 0.125 of origin peak.

Translational NCS

Non-crystallographic translations introduce awkward structure factor correlations, and can make structures difficult to refine.

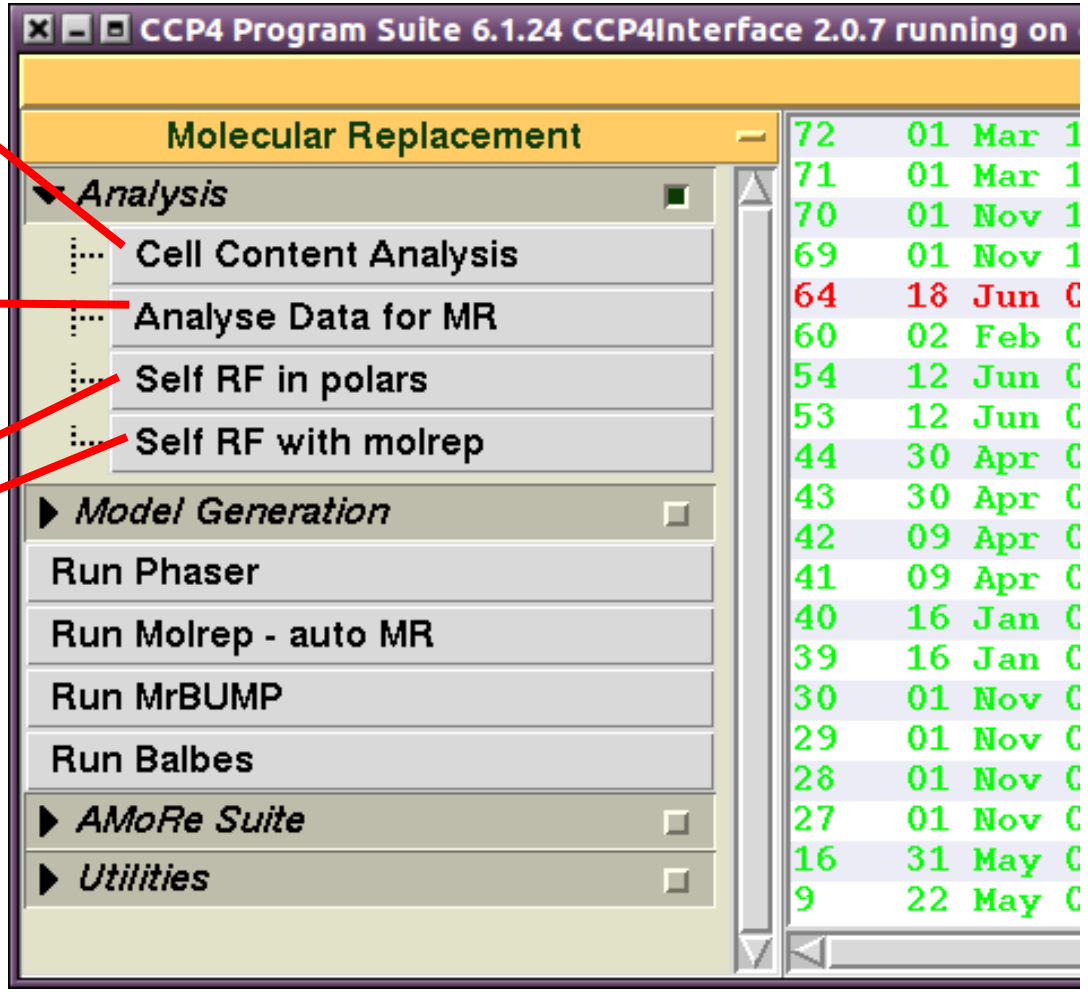
Molrep can use NCS translation vector to generate dimer for use in translation search.

Data analysis before MR

Matthews coefficient
Number copies in a.s.u.

Native Patterson
(translational NCS)

Self RF
(rotational NCS)



Molecular Replacement

Traditionally split MR search into 2 steps (cf. EPMR, Queen of Spades, etc.):

1. Determine orientation of search model
   - cross rotation function (CRF)
2. Determine position of search model for given orientation
   - translation function (TF)
   - check for clashes between symmetry related models

Look at this in context of Molrep and Phaser

Selection of Search Models

- Search PDB for homologous proteins
- Don't always just take the first
- Look at search model first and edit

Discussed more under MrBUMP and tomorrow.

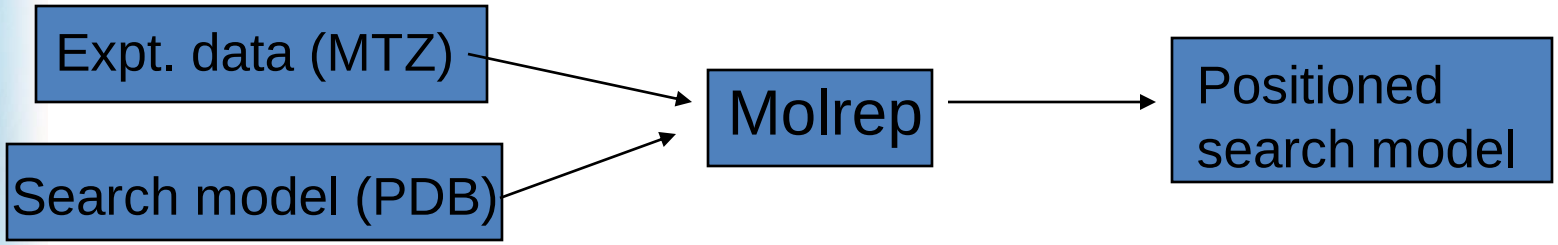
Molrep

Alexei Vagin
University of York
http://www.ysbl.york.ac.uk/~alexei/molrep.html

CCP4

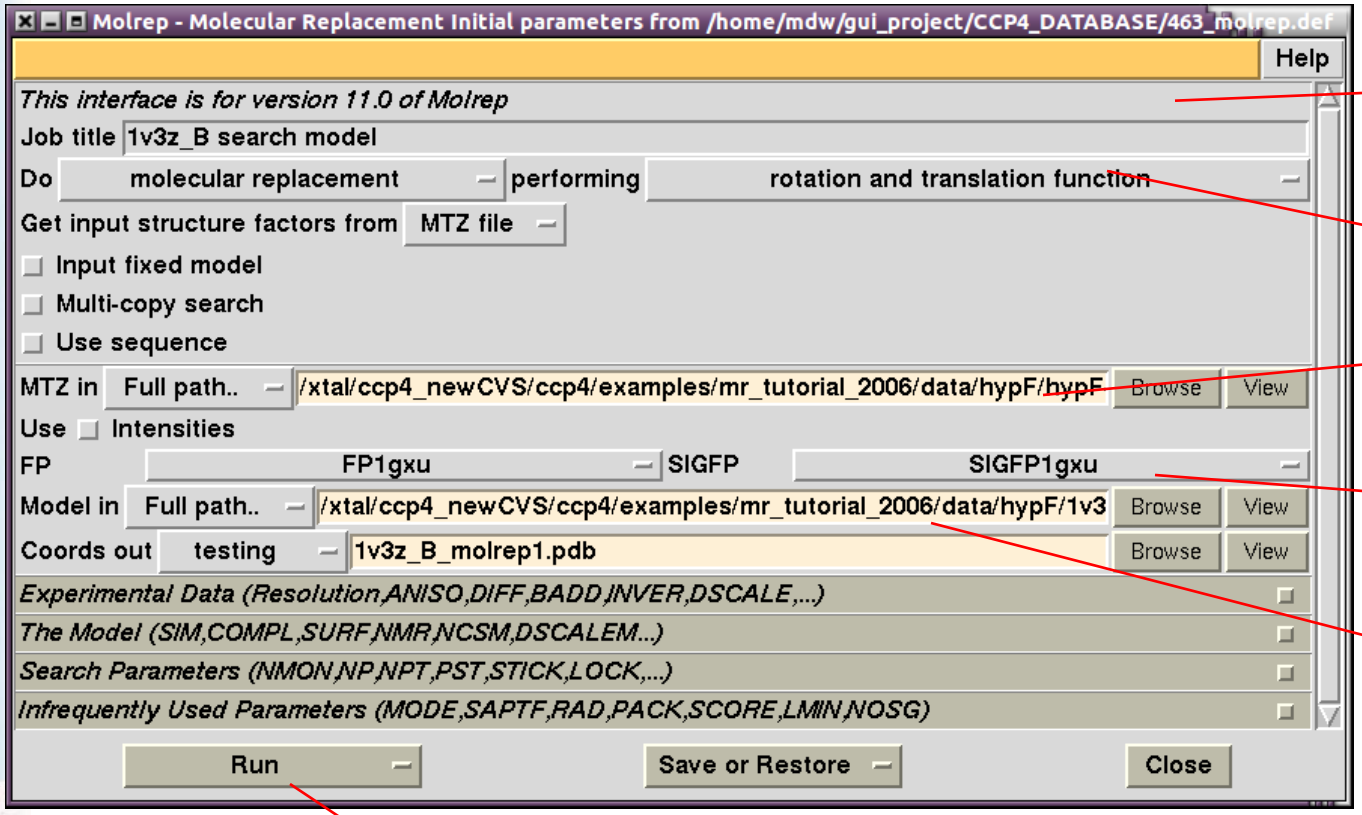
Molrep: overview of functionality

Performs complete MR in single step:



- Individual steps for more difficult cases: CRF, TF, rigid-body
- Self RF, locked CRF
- Multi-copy search: dyad search, multi-monomer
- Phased TF, spherically-averaged phased TF
- Improve search model
- Other search models: electron density map, EM map, NMR models (inc. ensembles of homologous proteins)
- Fit model in electron density map / EM map

MR for straightforward case via GUI



Other Molrep parameters

SG ALL

Check all compatible spacegroups

High resolution limit

Absolute cut-off (RESMAX)
Default estimated

Low resolution cut-off

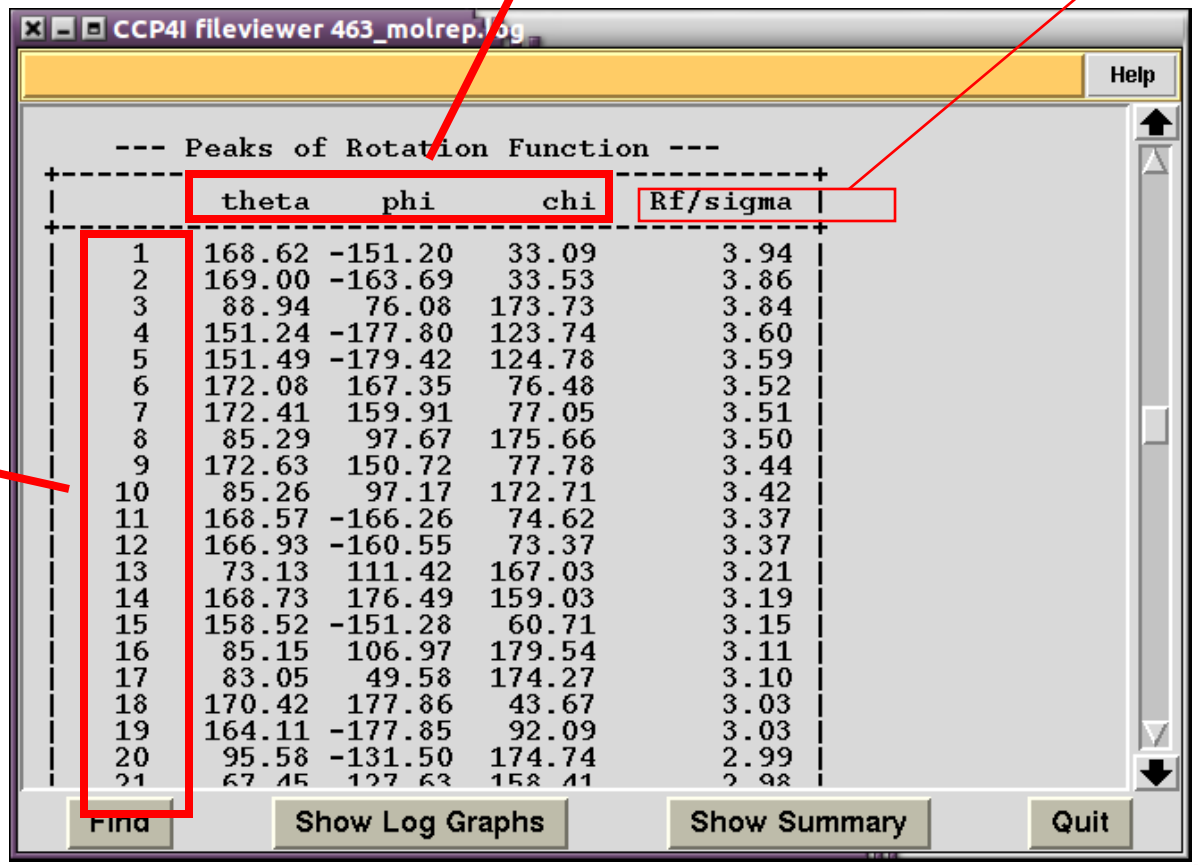
Molrep uses soft cut-off, B_{off} (BOFF)
From minimum resolution (RESMIN) or size of search model (COMPL)

High resolution cut-off

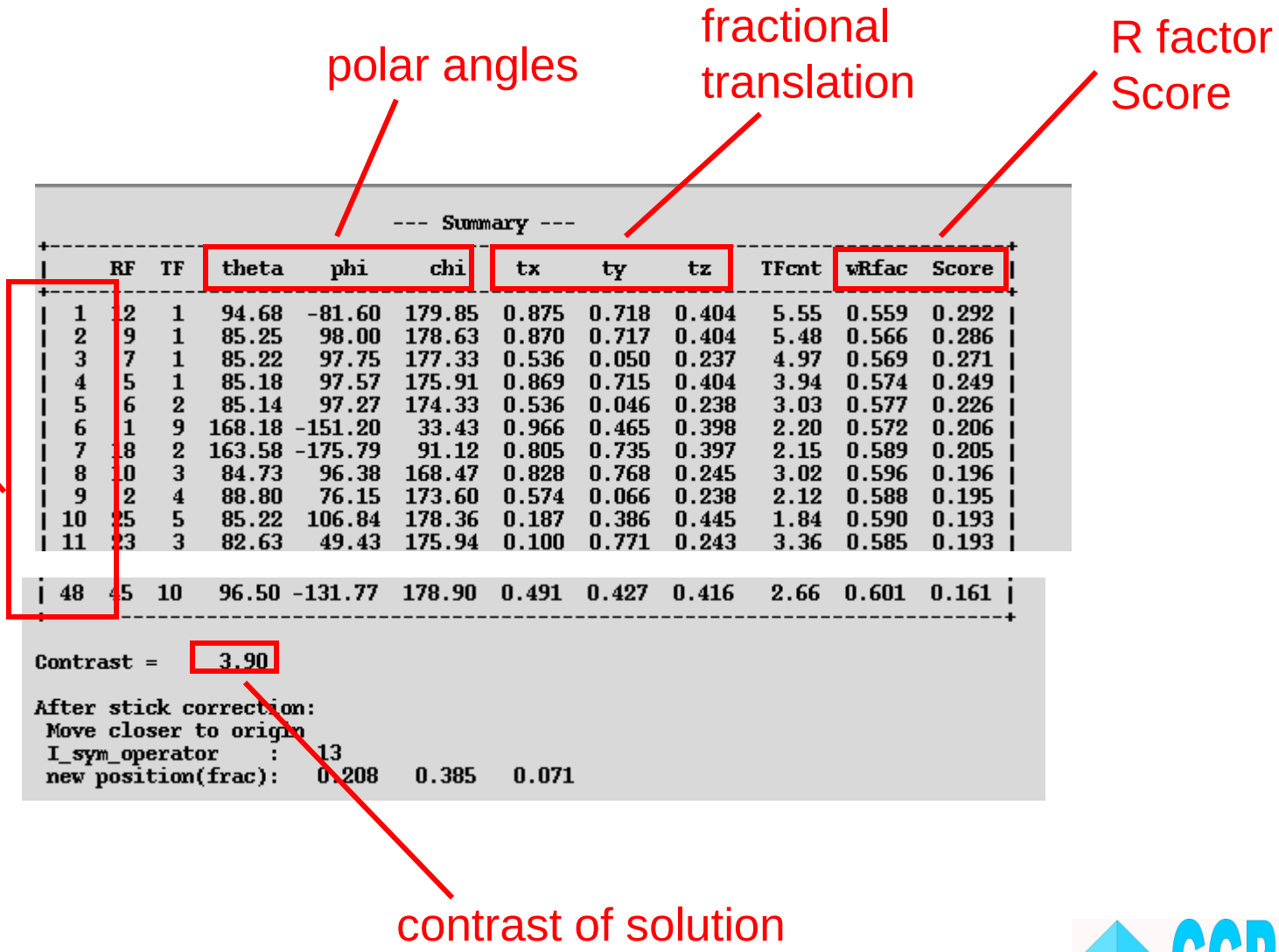
Molrep uses soft cut-off, B_{add} (BADD)
Default from sequence similarity (SIM)

$$|F|_{new} = |F|_{input} * \exp(-B_{add} * s^2) * (1 - \exp(-B_{off} * s^2))$$

Cross Rotation Function



Translation Function



Identification of solutions

Packing Function (computed from overlap of electron densities, PF = 1 means no overlap) integrated into TF search

$\Rightarrow$ downweights solutions with overlapping molecules.

SCORE = product Correlation Coefficient of intensities and maximal value of Packing Function

CONTRAST = ratio of top score to mean score:

>3.0 - definitely solution
<3.0 and > 2.0 - solution
<2.0 and > 1.5 - maybe solution
<1.5 and > 1.3 - maybe not solution, but program accepts it
<1.3 - probably not solution

Finding more than one copy in the asu

By default, Molrep will estimate number of copies to find.
Override with **NMON** keyword

"Multi-copy search" options

CRF
TF for first copy
Fix first copy
TF for second copy
Fix second copy
TF for third copy
:
:

Solving complexes

- Choose first component (largest, highest similarity)
- Solve for first component (probably need to specify **NMON** explicitly)
- New Molrep job
  - **Model in** - second component
  - **Fixed in** - positioned first component
- Repeat for all other components

CCP4

Phaser

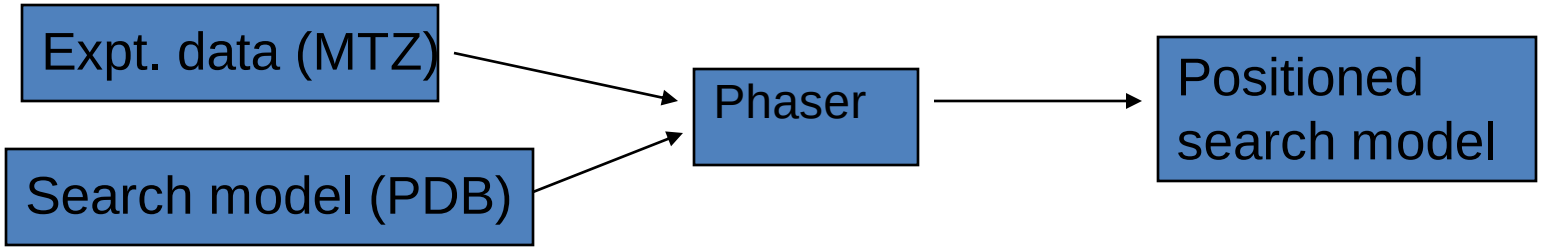
Randy Read, Airlie McCoy, Gabor Bunkoczi
University of Cambridge

http://www.phaser.cimr.cam.ac.uk/

CCP4

Overview

Performs complete MR in single step:



Use "**MODE MR_AUTO**" or "**automated search**" in the GUI

- anisotropy correction
- fast rotation function
- fast translation function
- packing
- refinement and phasing

loop over models

More functionality ...

- All steps can be run separately
- Search over spacegroups (**SGALTERNATIVE)**:
  MTZ spacegroup and enantiomorph e.g. P41 vs P43
  All spacegroups in MTZ point-group
  Selected spacegroups
- Ensemble models (see later)
- Brute RF and TF - slow and accurate
- Normal mode analysis
  Generates models (.pdb) perturbed along normal modes (e.g. to account for domain movements)

MR for straightforward case

mode

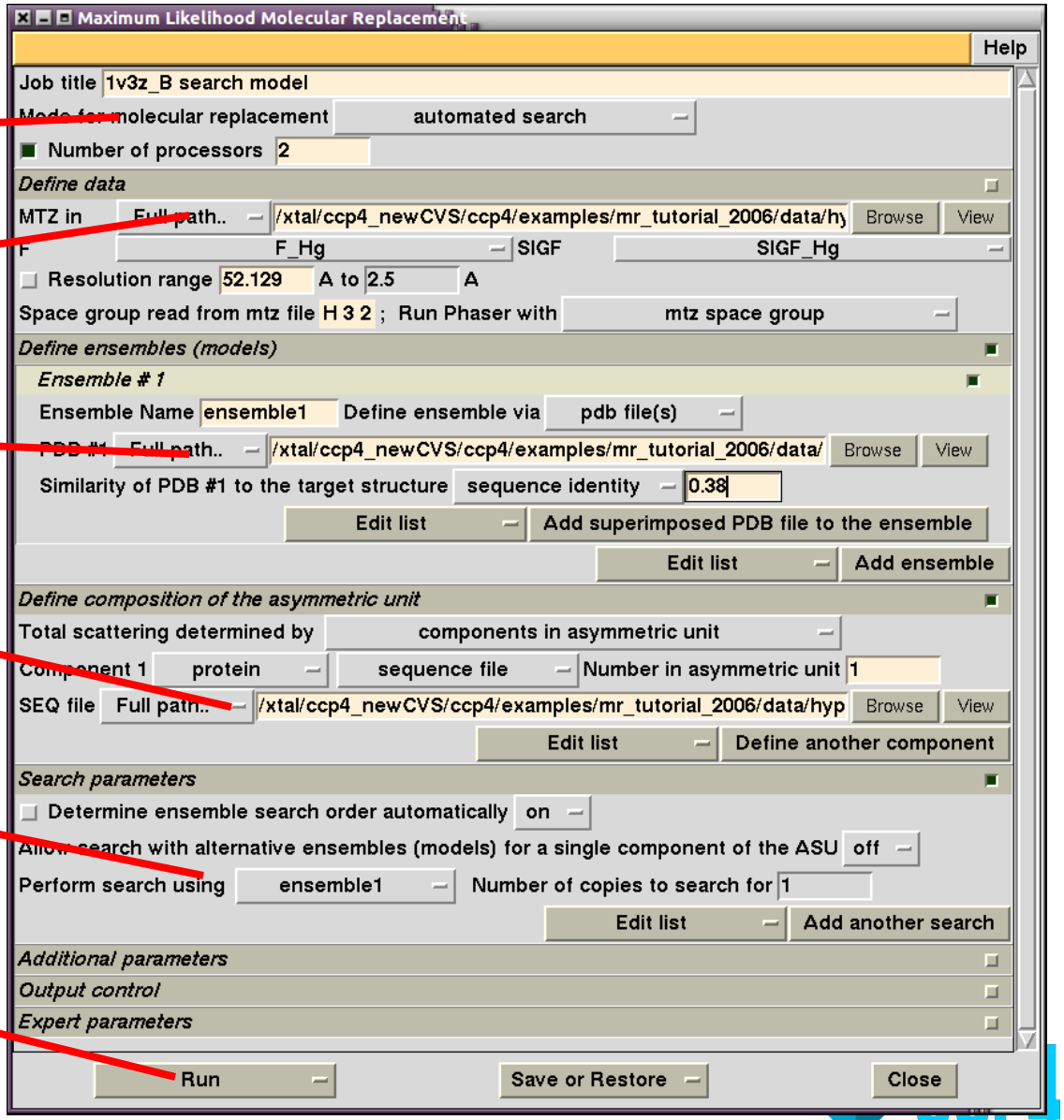
MTZ file

search model

target details

specify search

RUN IT!



FRF

Euler angles (CCP4)

Top LLG and Z-scores for FRF

CCP4I fileviewer 20_phaser_MR.log

Help

```
Top Peaks With Clustering From Rescoring of FRF #1
--------------------------------------------------
#        Rank of the peak after clustering search points
(#)      Rank of the peak before clustering search points
FSS      Fast Search Score
Z-Score  Number of standard deviations of FFS above the mean
Split    Distance (in degrees or Angstroms) to top peak
#Group   Number of points clustered in peak
raw/top  FSS of raw (unclustered) peak and top peak in cluster

You requested peaks over 75% of top (i.e. 0.75*(top-mean)+mean)
There were 6 sites over 75% of top
The sites over 75% are:
#      (#)     Euler1 Euler2 Euler3   LLG   Z-score Split #Group   raw/top
1      1        30.3    8.5  235.9  +17.29    4.63    0.0     43 100.0/100.0
               150.3    8.5  235.9
               270.3    8.5  235.9
                89.7  171.5   55.9
               209.7  171.5   55.9
               329.7  171.5   55.9
2      16       38.0    8.0  244.0  +16.68    4.50   15.7     29  86.6/ 91.1
               158.0    8.0  244.0
               278.0    8.0  244.0
                82.0  172.0   64.0
               202.0  172.0   64.0
               322.0  172.0   64.0
3      47      108.4    5.8  223.6  +15.78    4.32   54.3     30  82.5/ 86.0
               228.4    5.8  223.6
               348.4    5.8  223.6
                11.6  174.2   43.6
               131.6  174.2   43.6
               251.6  174.2   43.6
4      20       59.5    7.3  333.2  +15.68    4.30   13.3     62  85.9/ 87.1
               179.5    7.3  333.2
               299.5    7.3  333.2
                60.5  172.7  153.2
               180.5  172.7  153.2
               300.5  172.7  153.2
5      73       39.8   12.4  343.3  +13.40    3.84   17.5      1  78.9/ 78.9
               159.8   12.4  343.3
               279.8   12.4  343.3
                80.2  167.6  163.3
               200.2  167.6  163.3
               320.2  167.6  163.3
6      209      22.6   46.6  198.9  +11.82    3.53   57.8      1  67.5/ 67.5
               142.6   46.6  198.9
               262.6   46.6  198.9
                97.4  133.4   18.9
               217.4  133.4   18.9
               337.4  133.4   18.9

Fast Rotation Function Table
----------------------------
      Top     (Z)       Second     (Z)       Third     (Z)
  17.29 ( 4.63)       16.68 ( 4.50)       15.78 ( 4.32)
```

Find | Show Summary | Quit

FTF

fractional translation

FRF solution number

Top LLG and Z-scores for FRF

CCP4I fileviewer 20_phaser_MR.log

Help

```
Top Peaks With Clustering From Rescoring of SET #1 TRIAL #1
-----------------------------------------------------------
#       Rank of the peak after clustering search points
(#)     Rank of the peak before clustering search points
FSS     Fast Search Score
Z-Score Number of standard deviations of FFS above the mean
Split   Distance (in degrees or Angstroms) to top peak
#Group  Number of points clustered in peak
raw/top FSS of raw (unclustered) peak and top peak in cluster

You requested peaks over 75% of top (i.e. 0.75*(top-mean)+mean)
There were 57 sites over 75% of top
The sites over 75% are:
#     (#)     Frac X Frac Y Frac Z   LLG    Z-score Split #Group    raw/top
1     1       0.692  0.691  0.206  +25.09    4.62     0      2   56.60/ 56.60
2     14      0.509  0.989  0.037  +24.74    4.58     29     3   53.10/ 54.87
3     3       0.502  0.981  0.771  +23.07    4.37     12     2   55.89/ 56.07
4     9       0.695  0.695  0.258  +23.06    4.37     8      1   54.10/ 54.10
5     19      0.475  0.844  0.775  +22.62    4.31     16     2   52.43/ 52.43
6     6       0.571  0.450  0.132  +21.80    4.21     17     3   55.18/ 55.18
```

Find | Show Summary | Quit

CCP4I fileviewer 20_phaser_MR.log

Help

```
-----------------
TABLES OF RESULTS
-----------------

   Fast Translation Function Table: Space Group R 3 2
   --------------------------------------------------
   #SET #TRIAL      Top     (Z)     Second     (Z)     Third     (Z)    Ensemble
      1      1   25.09 ( 4.62)   24.74 ( 4.58)   23.07 ( 4.37)   ensemble1
      1      2   45.14 ( 7.13)       -     -         -     -     ensemble1
      1      3   23.91 ( 5.00)   21.21 ( 4.63)   20.14 ( 4.49)   ensemble1
      1      4   23.41 ( 4.53)   23.25 ( 4.51)   19.77 ( 4.08)   ensemble1
      1      5   21.86 ( 4.75)   21.46 ( 4.70)   20.37 ( 4.56)   ensemble1
      1      6   21.01 ( 4.78)   19.57 ( 4.61)   19.52 ( 4.60)   ensemble1
   ---- ------
```

Find | Show Summary | Quit

CCP4

Packing

Phaser does packing check after FTF

Clashes = Cα atoms closer than 3Å

Default number of clashes now a percentage of total Cα atoms

```
---------------
PACKING FUNCTION
---------------

   There is 1 solution to pack
   Packing analysis
   0% 100%
   |==| DONE


   Packing Table: Space Group H 3 2
   --------------------------------
   Solutions accepted if number of clashes <= 5% of CA atoms
   i.e. total number of clashes <= 4 AND each ensemble <= 5%
   #    #Clashes #    Accepted Annotation
   1     2       1    YES    RFZ=4.4 TFZ=7.0 PAK=2

   1 accepted of 1 solutions
```

If the model is RNA or DNA, phosphate (P) and carbon atoms (C3* and C4*) in the phosphate backbone, and nitrogen atoms in the bases are taken as the marker atoms for clashes.

Solution Files

.sol file produced at end of job

- Contains summary of all solutions
- Each solution contains rotations and usually translations -

3DIM vs 6DIM

- One line per model located
- .sol file can be read back into Phaser in later jobs

Z-score	Have I solved it?
less than 5	no
5 - 6	unlikely
6 - 7	possibly
7 - 8	probably
more than 8	definitely

RFZ = RF Z-score

TFZ = TF Z-score

Ensemble models

Phaser refers to search models as "ensembles"

Often, ensemble contains single model, as in traditional MR

But Phaser can use an ensemble of > 1 models, which may work better than any single model

Models in an ensemble must be superposed prior to use in Phaser - use e.g. Superpose in CCP4

N.B. Phaser will complain if:

- MW of models in ensemble are too different
- RMS between models is too large

(In Molrep, construct ensemble as pseudo-NMR PDB file)

CCP4

Finding more than 1 copy in the asu

Specify > 1 in **Composition of the asymmetric unit**
(keyword COMPOSITION ... NUMBER)
Specify > 1 in **Number of copies to search for**
(keyword SEARCH ... NUMBER)

Phaser will issue warnings if these numbers are wrong.

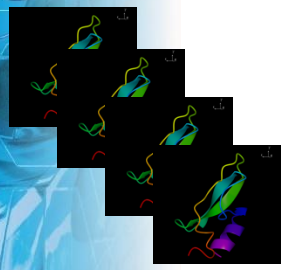
CRF
TF for first copy
Fix first copy (possibly multiple sets)
CRF for second copy
TF for second copy
Fix second copy (possibly multiple sets)
:
:

Complexes

As before, but:

- Define > 1 type of component
  - **Define composition of the asymmetric unit**
    - **Define another component**
- Define > 1 ensemble
  - **Define ensembles (models)**
    - **Add ensemble**
- Specify all searches
  - **Search parameters**
    - **Add another search**

MrBUMP



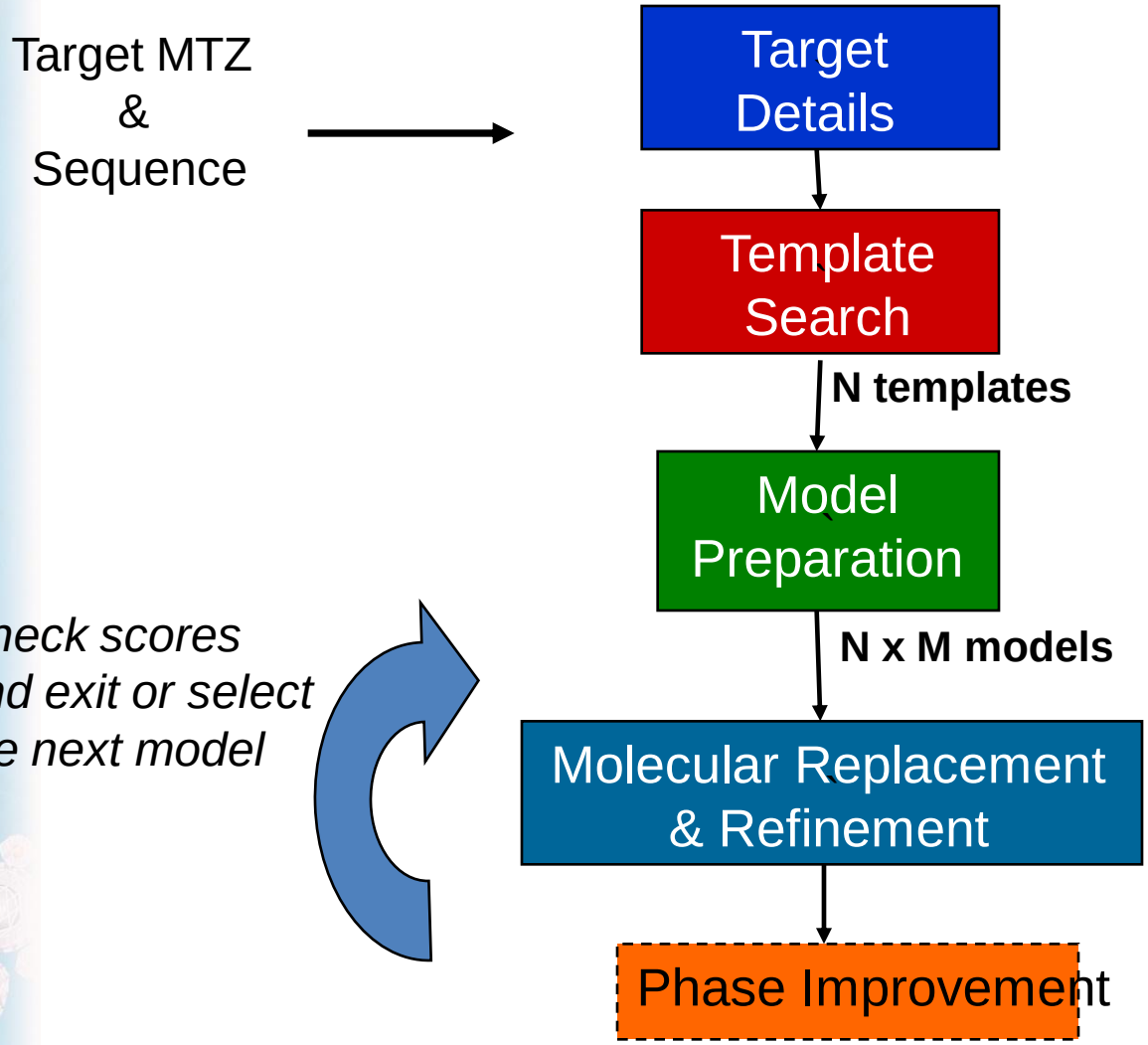
- An automation framework for Molecular Replacement.
- Particular emphasis on generating a variety of search models.

- Wraps **Phaser** and/or **Molrep**.
- Uses a variety of helper applications (e.g. Chainsaw) and bioinformatics tools (e.g. Fasta, Mafft) to generate search models
- Uses up-to-date on-line databases (e.g. PDB, Scop)

- In favourable cases, gives “one-button” solution
- In Complicated Cases, will suggest likely search models for manual investigation (lead generation)

CCP4

The Pipeline



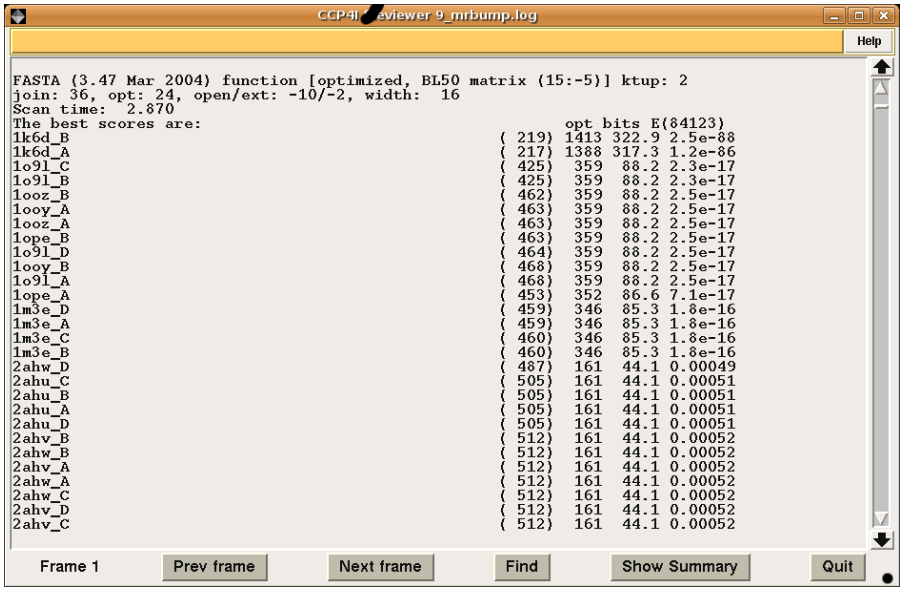
Search for model templates

FASTA search of PDB

– Sequence based search using sequence of target structure

All of the resulting PDB id codes are added to a list

These structures are called model templates



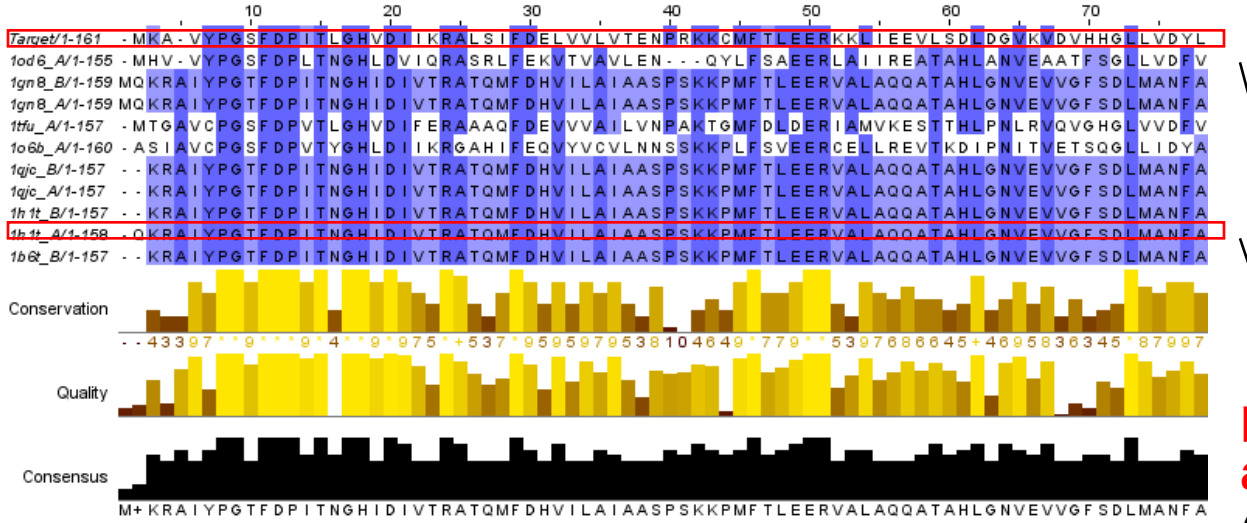
Other templates from:

- SSM search using top hit from the FASTA search
- Can add additional PDB id codes to the list, e.g. identified from FFAS or psiBLAST searches
- Can add local PDB files

Multiple Alignment step

target

model templates

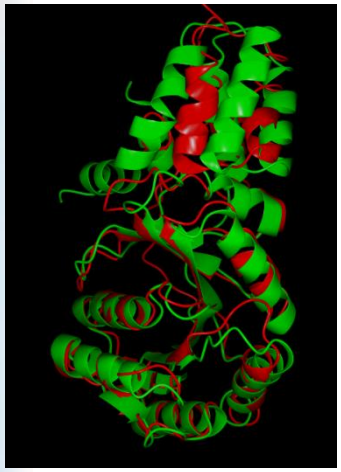


Jalview 2.08.1 Barton group, Dundee

pairwise alignment (used in Chainsaw)

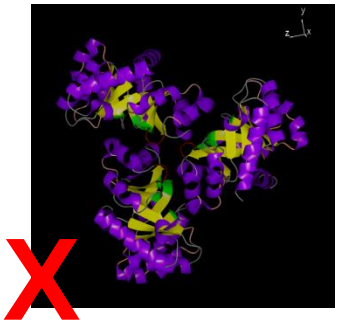
currently support ClustalW, MAFFT, probcons or T-coffee for multiple alignment

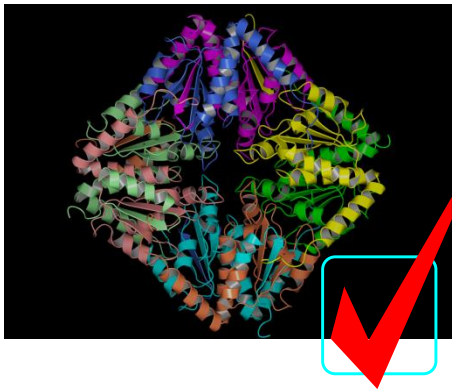
Model template scoring: score = sequence identity X alignment quality



Domains
e.g. if relative domain motion

SCOP



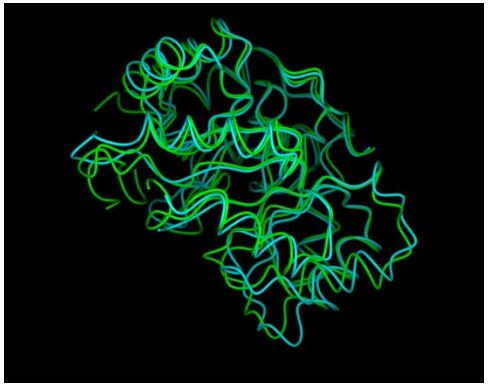


PQS/PISA

template chains

Multimers
Better signal-to-noise ratio than monomer, ***if*** assembly is correct for the target.

superpose



Ensembles
Create ensembles of search models, for use in additional run of Phaser.
Need to be similar in MW and rmsd

Search Model Preparation

Search models prepared in four ways:

PDBclip

- original PDB with waters removed, most probable conformations selected and format tidied (e.g. chain ID added)

Molrep

- Molrep model preparation function which aligns the template sequence with the target sequence and prunes the non-conserved side chains accordingly.

Chainsaw

- Can be given any alignment between the target and template sequences. Non-conserved residues are pruned back to the gamma atom.

Polyalanine

- Created by excluding all of the side chain atoms beyond the CB atom using the Pdbset program

deal with deletions

more side chain truncation

Molecular Replacement step

Running MR

- For each search model, MR done with **Molrep** or **Phaser** or both.
- MR programs run mostly with defaults
- MrBUMP provides LABIN columns, MW of target, sequence identity of search model, number of copies to search for, number of clashes tolerated

MR output

- MR scores and un-refined models available for later inspection
  ⇒ assess quality of solution, extent of model bias
- MrBUMP doesn't use MR scores, but checks for output file with positioned model, and passes to refinement step

CCP4

Testing enantiomorphic spacegroups

- 11 pairs of enantiomorphic spacegroups containing screw axes of opposite handedness, e.g. $P4_1$ and $P4_3$)
- usually both need to be tested in MR
- correct spacegroup indicated by TF and packing

- MrBUMP can test both in Molrep and/or Phaser.
- For each search model, best MR results used to fix spacegroup for subsequent steps.
- Discrimination good for good search model + correct MR solution

Restrained Refinement step

- The resulting models from molecular replacement are passed to **Refmac** for restrained refinement.
- The change in the Rfree value during refinement is used as rough estimate of how good the resulting model is.

final Rfree < 0.35 or final Rfree < 0.5 and dropped by 20%	⇒	"success"
final Rfree < 0.48 or final Rfree < 0.55 and dropped by 5%	⇒	"marginal"
otherwise	⇒	"failure"

conservative .....

Phase improvement

If resolution better than 1.7Å use Acorn procedure:

initial phase set from refined MR solution

artificial phase extension to 1.0Å

dynamic density modification

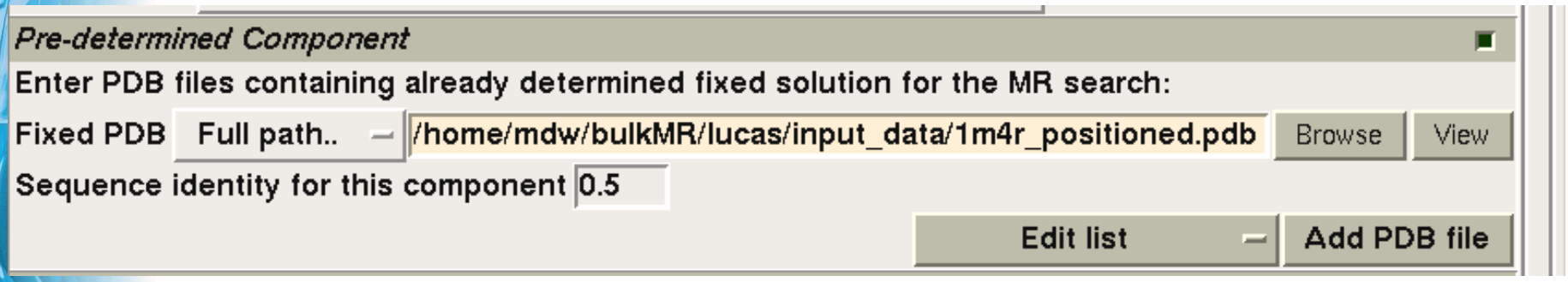
Result:

CC for medium Es good indicator of solution

Use E-maps for re-building

Inclusion of fixed models

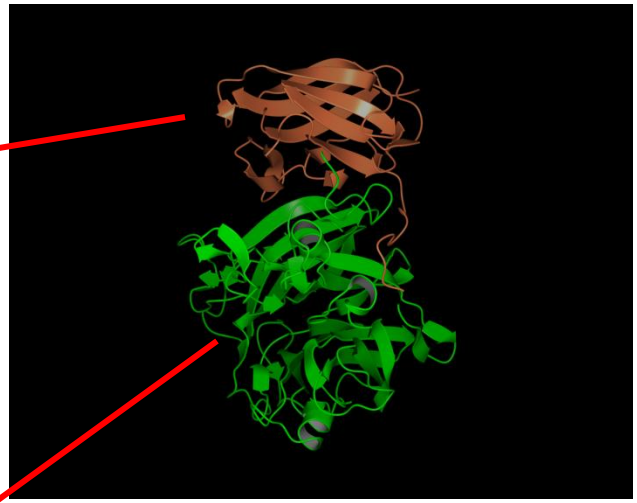
- MrBUMP will accept one or more positioned models.
- These are included as fixed models in all MR jobs.



Thus, solve complexes through consecutive runs of MrBUMP.

Example (thanks to Elien Vandermarliere)

Target is an arabinofuranosidase
Data to 1.55Å in P212121



Small C domain (144 res) solved with 34% seq ident model
(1w9t_B_MOLREP best out of 4 solutions)

With C domain solution fixed, large N domain (345 res) solved with 28% seq ident model
(1gyh_C_CHNSAW best out of 7 solutions)

Not yet solved!
Acorn: CC increases from 0.04 to 0.18
ARP/wARP then builds 457/493 residues to R/Rfree 0.185/0.225

MrBUMP output

Log file gives summary of models tried and results of MR

- May get several putative solutions
- *Ease of subsequent model re-building, model completion may depend on choice of solution*
- Worth checking “failed” solutions

Top solution available from ccp4i

Detailed results located in: <ccp4i project directory>/search_<job number>

In this directory, there are a number of subdirectories, including:

data

Contains the data files and log files from all jobs run. The directory hierarchy is of the form <template>/<search model>/<pipeline step>
e.g. <ccp4i project directory>/search_55/data/loc0_A/chainsaw/mr

results

Results from the successful search model are placed into subdirectory "solution". Other results are placed into subdirectory "marginal_solns".