

Automated Molecular Replacement with MrBUMP

Ronan Keegan

STFC Rutherford Appleton Laboratory

ronan.keegan@stfc.ac.uk

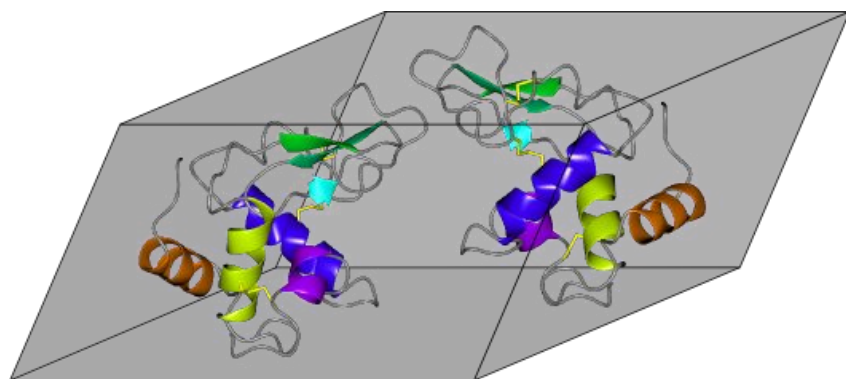


**Science & Technology
Facilities Council**

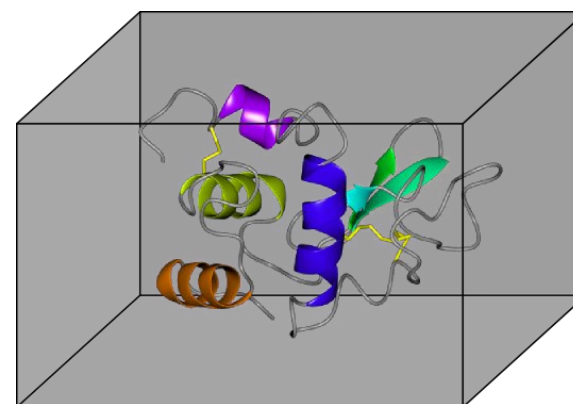


Molecular Replacement: The problem

(Slides courtesy of Gabor Bunkoczi)



Previous crystal form



Current crystal form

$$\mathcal{R}(\phi, \psi, \kappa, x, y, z)$$



Complications

Differences between target and search model:

- Deviations in the main chain
- Un-matched Loops
- Different side chains
- Different side chain conformations

Problems stemming from the experimental data:

- Incomplete data
- Anisotropy
- Twinning
- Pseudo symmetry



Automation

- Easy cases:
 - provide a one button solution
 - find the best solution and subsequently best starting point for refinement and model completion
- Difficult cases:
 - finding a solution can involve much trial and error
 - success or failure in MR can be highly sensitive to small changes in the search model
 - try many permutations for search model selection, preparation and molecular replacement
- Encapsulate complex procedures such as multiple search model creation and the exploitation of *ab initio* modelling



Ab initio modelling and MR

Deriving structure models based solely on their amino-acid sequence

Has become increasingly successful in recent years and generated models have been shown to work well as molecular replacement search models in the case of smaller proteins (up to 120 residues)

Can be a useful approach where no obvious homologue is available to use as a search model

Can also be used to re-model positioned search models from a likely but poor MR solution and improve the quality of derived phases



Cluster of aligned ab initio-generated decoy models



MR Pipelines in CCP4

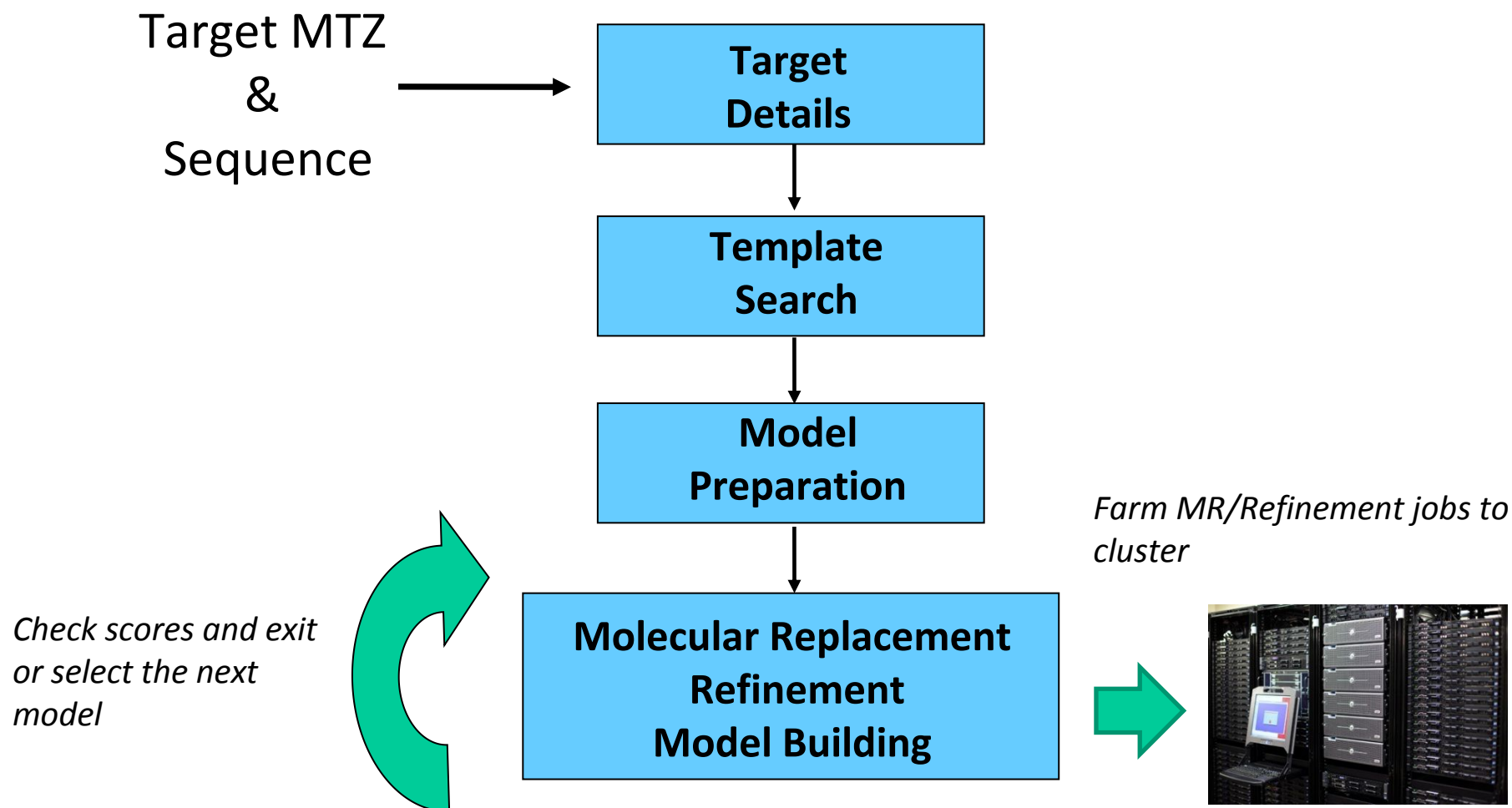
- MrBUMP & BALBES
 - Automated model search and preparation, molecular replacement, refinement and model building
- AMPLE
 - Using ab initio modeling to generate search models for molecular replacement



MrBUMP – Automated Molecular Replacement



MrBUMP Pipeline



Search for model templates

Phmmer/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

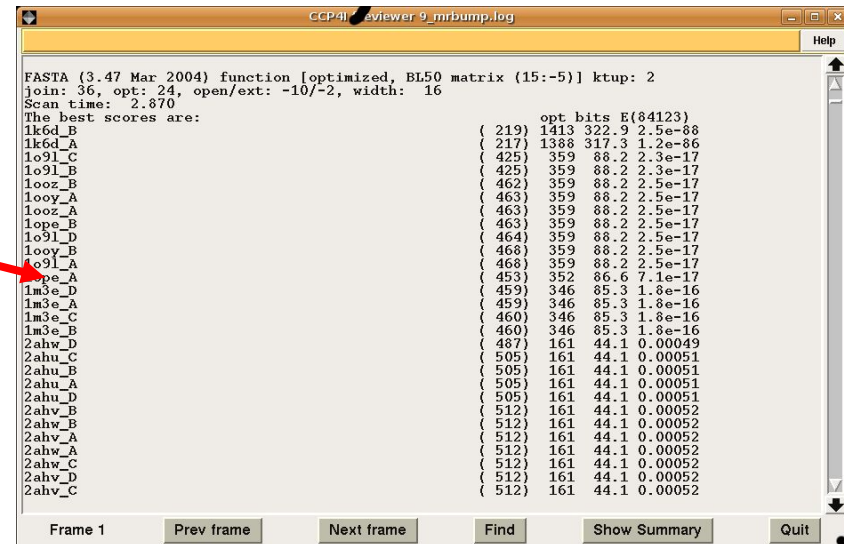
Other templates from:

Can add additional PDB id codes to the list, e.g. from external

HHPred or psiBLAST searches

New version calls HHPred directly to do search

Can also add local PDB files



The screenshot shows a window titled "CCP4 reviewer 9_mrbump.log". The text inside the window is as follows:

```
FASTA (3.47 Mar 2004) function [optimized, BL50 matrix (15:-5)] ktup: 2
join: 36, opt: 24, open/ext: -10/-2, width: 16
Scan time: 2.870
The best scores are:
opt bits E(84123)
1k6d_B (219) 1413 322.9 2.5e-88
1k6d_A (217) 1388 317.3 1.2e-86
1o9l_C (425) 359 88.2 2.3e-17
1o9l_B (425) 359 88.2 2.3e-17
1o9l_A (462) 359 88.2 2.5e-17
1o9l_D (463) 359 88.2 2.5e-17
1o9l_E (463) 359 88.2 2.5e-17
1o9l_F (463) 359 88.2 2.5e-17
1o9l_G (463) 359 88.2 2.5e-17
1o9l_H (463) 359 88.2 2.5e-17
1o9l_I (463) 359 88.2 2.5e-17
1o9l_J (463) 359 88.2 2.5e-17
1o9l_K (463) 359 88.2 2.5e-17
1o9l_L (463) 359 88.2 2.5e-17
1o9l_M (463) 359 88.2 2.5e-17
1o9l_N (463) 359 88.2 2.5e-17
1o9l_O (463) 359 88.2 2.5e-17
1o9l_P (463) 359 88.2 2.5e-17
1o9l_Q (463) 359 88.2 2.5e-17
1o9l_R (463) 359 88.2 2.5e-17
1o9l_S (463) 359 88.2 2.5e-17
1o9l_T (463) 359 88.2 2.5e-17
1o9l_U (463) 359 88.2 2.5e-17
1o9l_V (463) 359 88.2 2.5e-17
1o9l_W (463) 359 88.2 2.5e-17
1o9l_X (463) 359 88.2 2.5e-17
1o9l_Y (463) 359 88.2 2.5e-17
1o9l_Z (463) 359 88.2 2.5e-17
1m3e_D (459) 346 85.3 1.8e-16
1m3e_A (459) 346 85.3 1.8e-16
1m3e_B (459) 346 85.3 1.8e-16
1m3e_C (459) 346 85.3 1.8e-16
2ahw_D (487) 161 44.1 0.00049
2ahu_C (505) 161 44.1 0.00051
2ahu_B (505) 161 44.1 0.00051
2ahu_A (505) 161 44.1 0.00051
2ahu_D (505) 161 44.1 0.00051
2ahv_B (512) 161 44.1 0.00052
2ahv_A (512) 161 44.1 0.00052
2ahv_C (512) 161 44.1 0.00052
2ahv_D (512) 161 44.1 0.00052
2ahv_E (512) 161 44.1 0.00052
2ahv_F (512) 161 44.1 0.00052
2ahv_G (512) 161 44.1 0.00052
2ahv_H (512) 161 44.1 0.00052
2ahv_I (512) 161 44.1 0.00052
2ahv_J (512) 161 44.1 0.00052
2ahv_K (512) 161 44.1 0.00052
2ahv_L (512) 161 44.1 0.00052
2ahv_M (512) 161 44.1 0.00052
2ahv_N (512) 161 44.1 0.00052
2ahv_O (512) 161 44.1 0.00052
2ahv_P (512) 161 44.1 0.00052
2ahv_Q (512) 161 44.1 0.00052
2ahv_R (512) 161 44.1 0.00052
2ahv_S (512) 161 44.1 0.00052
2ahv_T (512) 161 44.1 0.00052
2ahv_U (512) 161 44.1 0.00052
2ahv_V (512) 161 44.1 0.00052
2ahv_W (512) 161 44.1 0.00052
2ahv_X (512) 161 44.1 0.00052
2ahv_Y (512) 161 44.1 0.00052
2ahv_Z (512) 161 44.1 0.00052
```

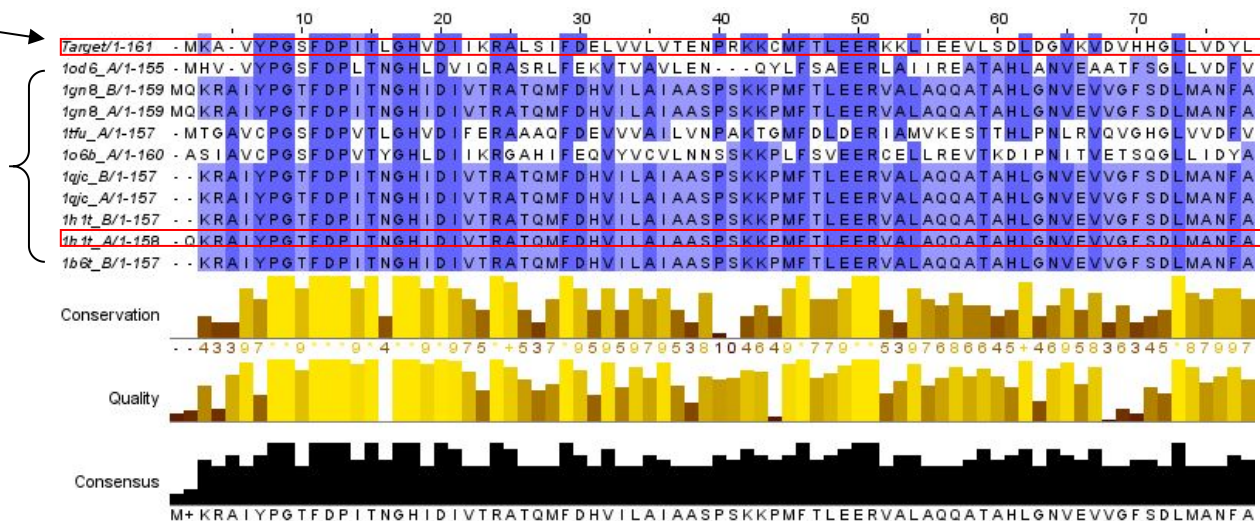


Multiple Alignment step



All model templates need to be aligned for scoring and later model preparation step

target



Jalview 2.08.1 Barton group, Dundee

pairwise alignment
(used later in Chainsaw & Sculptor)

Currently support [ClustaW](#), [MAFFT](#), [probcons](#) or [T-coffee](#) for multiple alignment

Model template scoring: $\text{score} = \text{sequence identity} \times \text{alignment quality}$

Where *HHPred* is used scoring and alignment come from *HHPred* search results



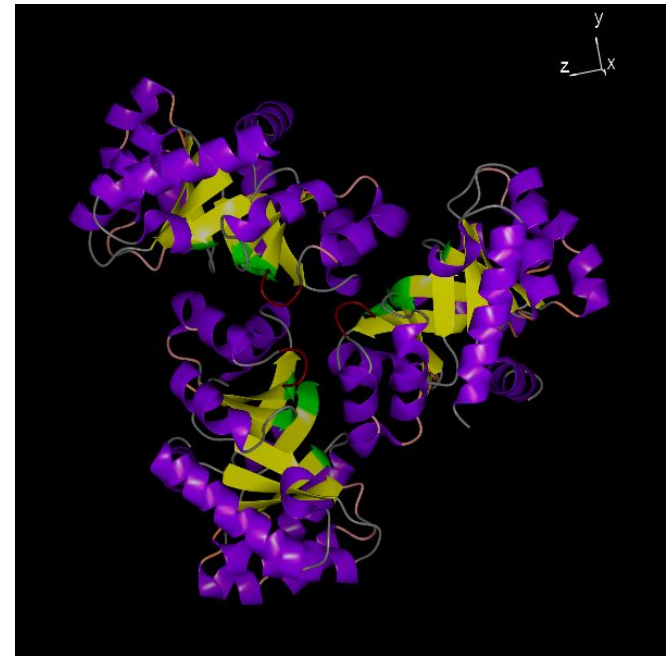
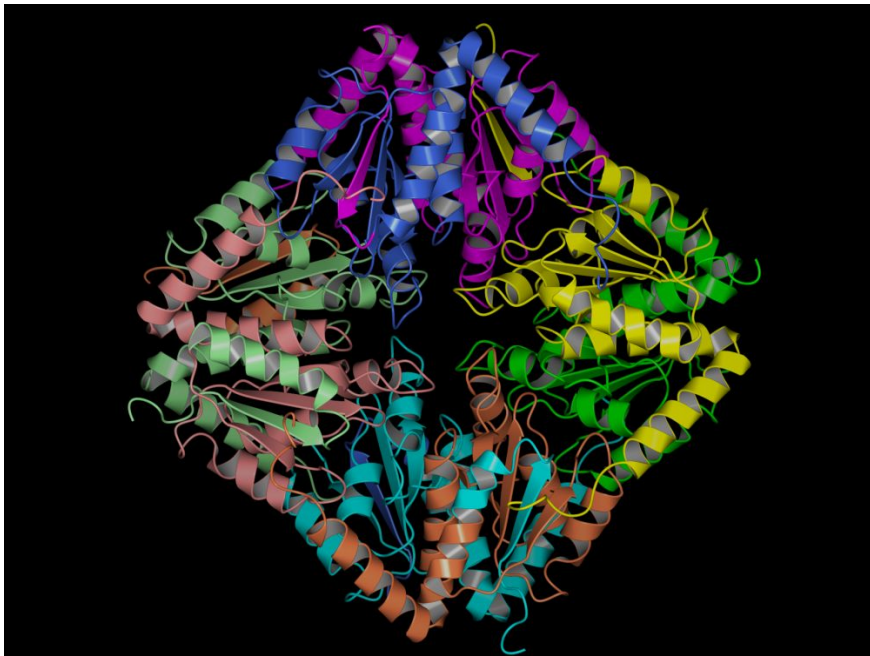
Domains

- Suitable templates for target domains may exist in isolation in the PDB, or in combination with dissimilar domains
- In case of relative domain motion, may want to solve domains separately
- SCOP database is scanned to see if domains exist for each of the PDBs in the list of templates (to be replaced with BALBES database)
- Domains are then extracted from the parent PDB structure file and added to the list of template models as additional search models for MR.



Multimers

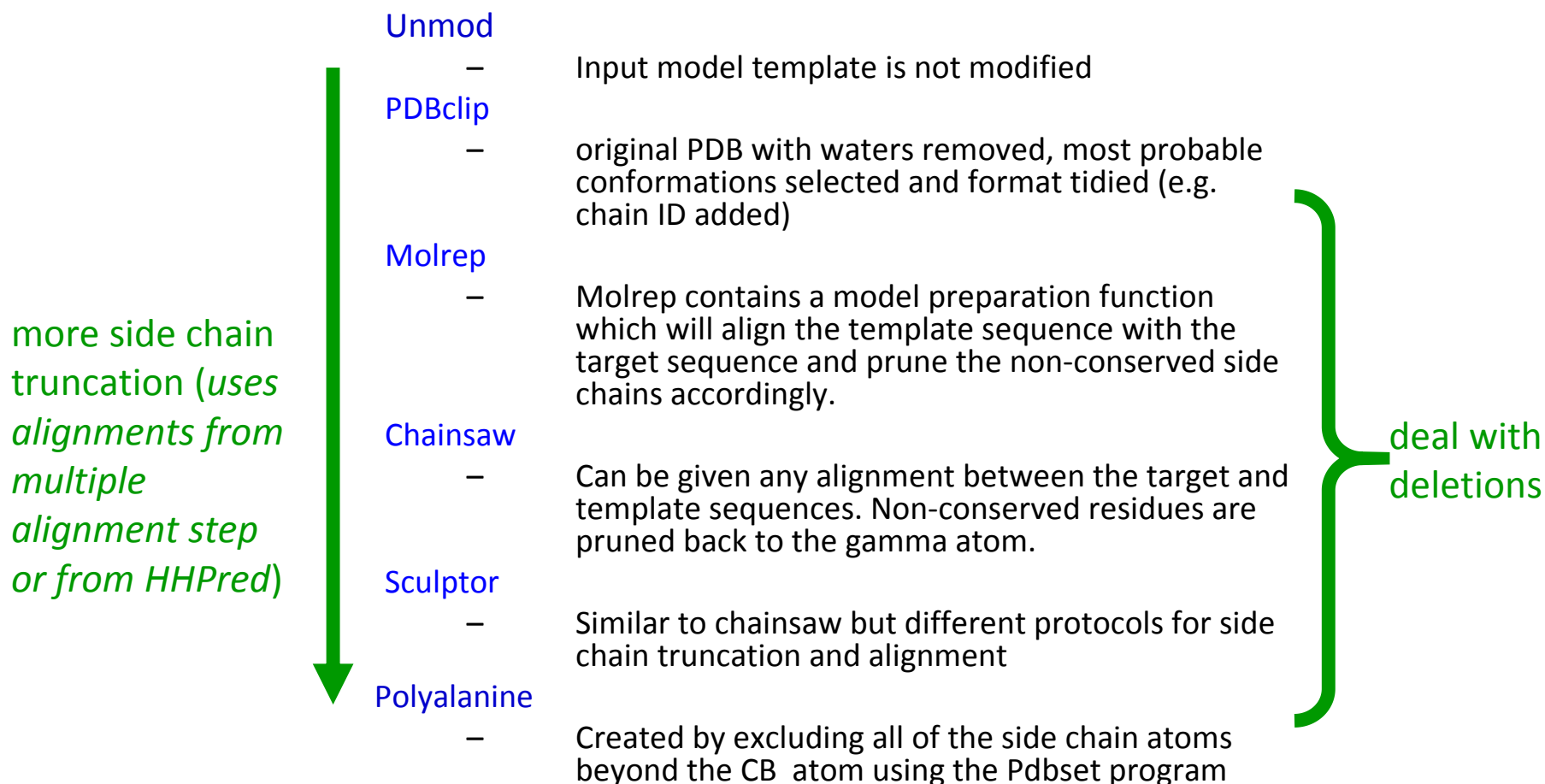
- Use template multimer as model for target multimer (currently uses PQS but will also be replaced by BALBES database)
- Better signal-to-noise ratio than monomer, *if* assembly is correct for the target.
- Biologically relevant multimers more likely transferable
- Can be determined manually using PISA



Search Model Preparation

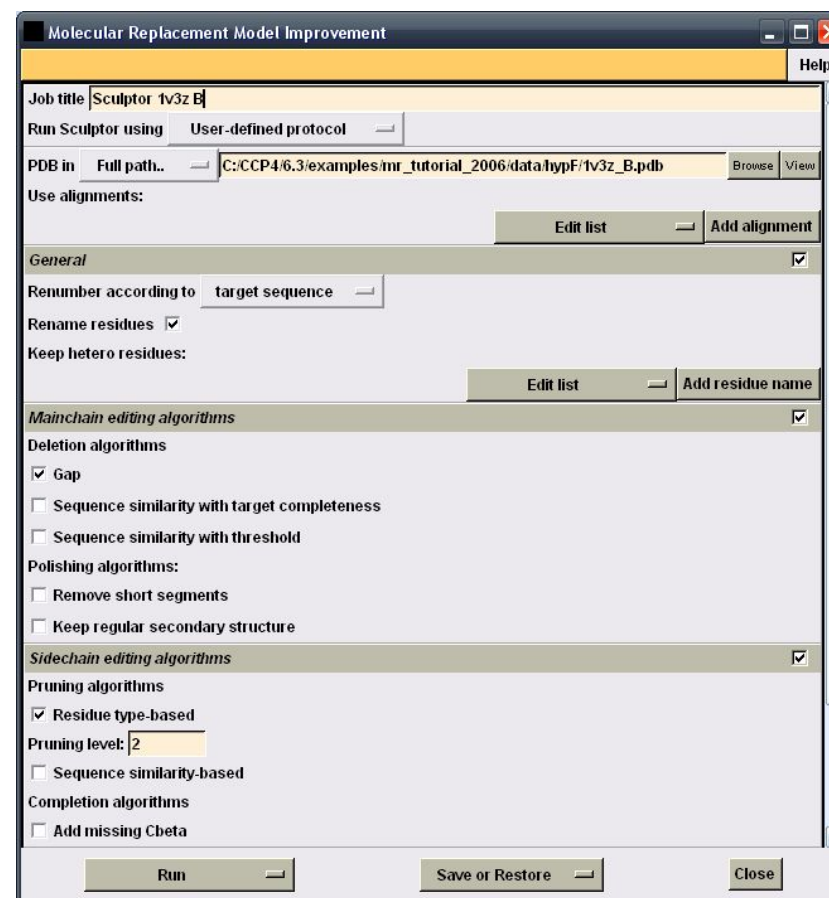


Search models are prepared in six different ways:



Phaser.Sculptor

- Multi-protocol search model preparation
- Allows for fine-grain control of the parameters used to decide on search model truncation
- Takes into account
 - Surface accessibility
 - Weighted sequence alignment



Ensemble models

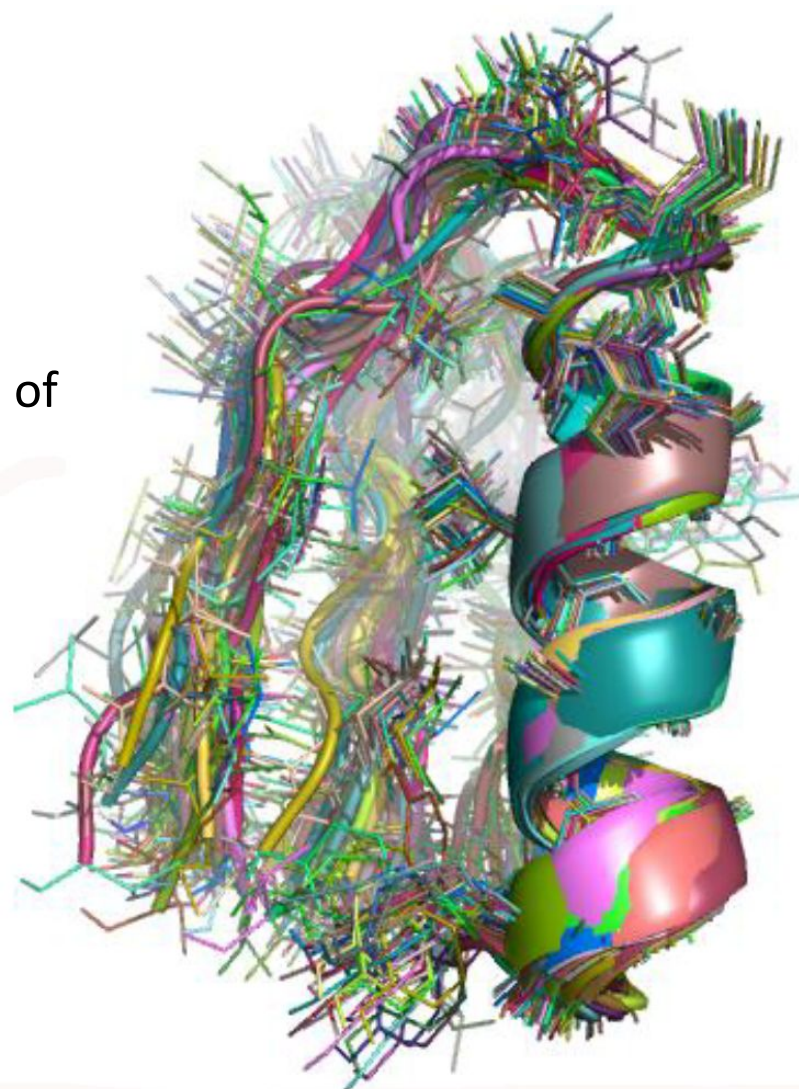
Create ensembles of top search models, for use in additional runs of [Phaser](#) and [Molrep](#)

MR programs will weight up regions of low alignment variance and down weight regions with more variability

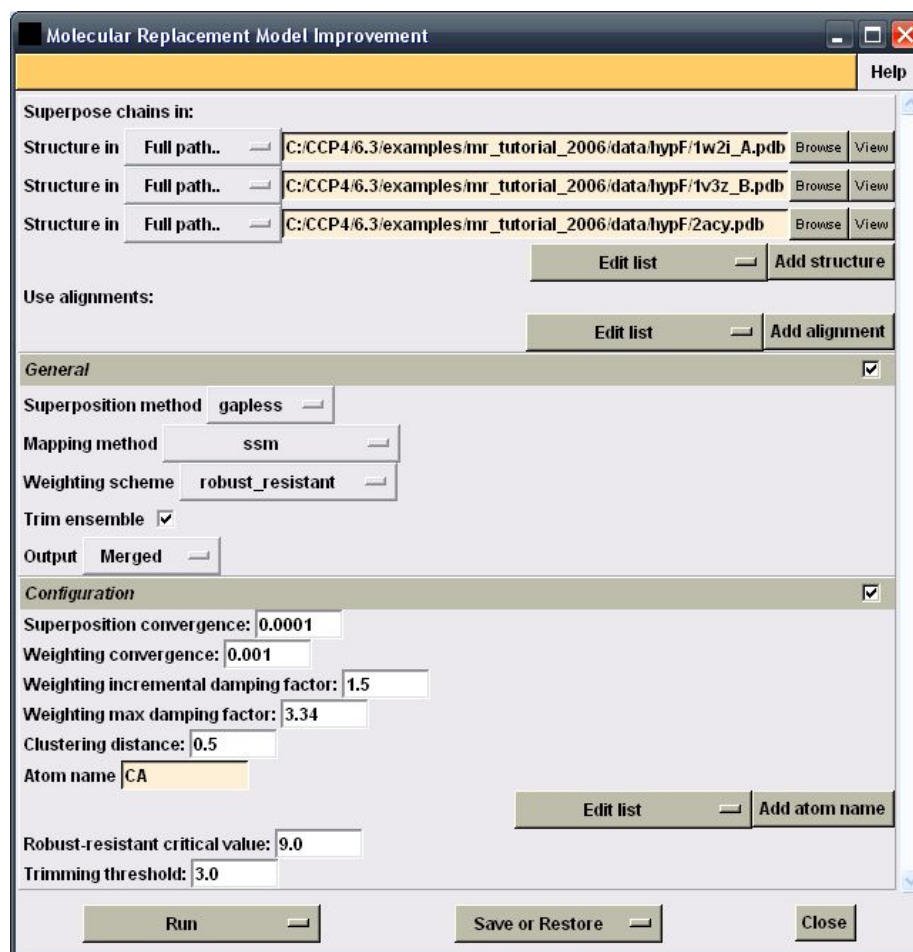
Can give solutions in cases where individual models fail

Models must be sufficiently similar (MW and rmsd)

Use Gesamt or Superpose to align structures



Phaser.Ensembler



- Generate ensembles from set of similar homologues
- Several protocols:
 - Gapped or non-gapped
 - Different alignment methods including manual input of alignment
 - Trimming based on RMSD threshold



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

Allow Molrep / Phaser to set resolution limits and weights

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 Refmac



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%



“good”

final Rfree < 0.48 or
final Rfree < 0.52 and dropped by 5%



“marginal”

otherwise



“poor”

conservative



Model Building

Model building programs can be invoked post-MR to give a better assessment of whether or not the MR solution is correct

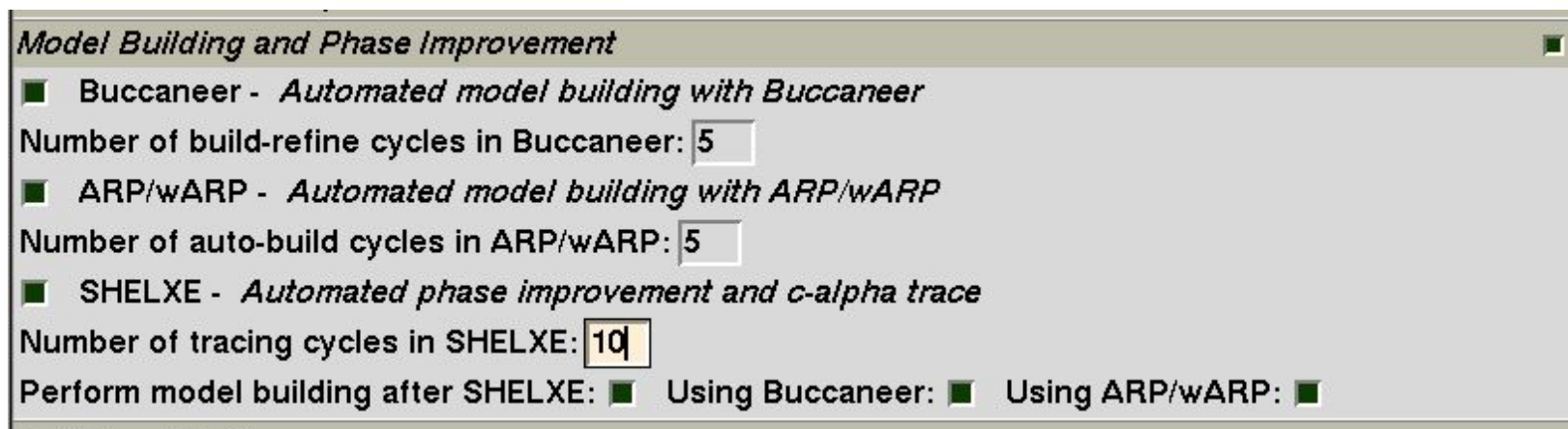
Options:

Buccaneer

ARP/wARP

SHELXE

SHELXE can also be combined with Buccaneer/ARPw/ARP to bring model closer to completion



Assessing Solutions with SHELXE

- Attempt to rebuild the structures
 - SHELXE: partial CC of >25% & average fragment length of 10 or more
 - Further confirmation provided through building with ARP/wARP and Buccaneer.

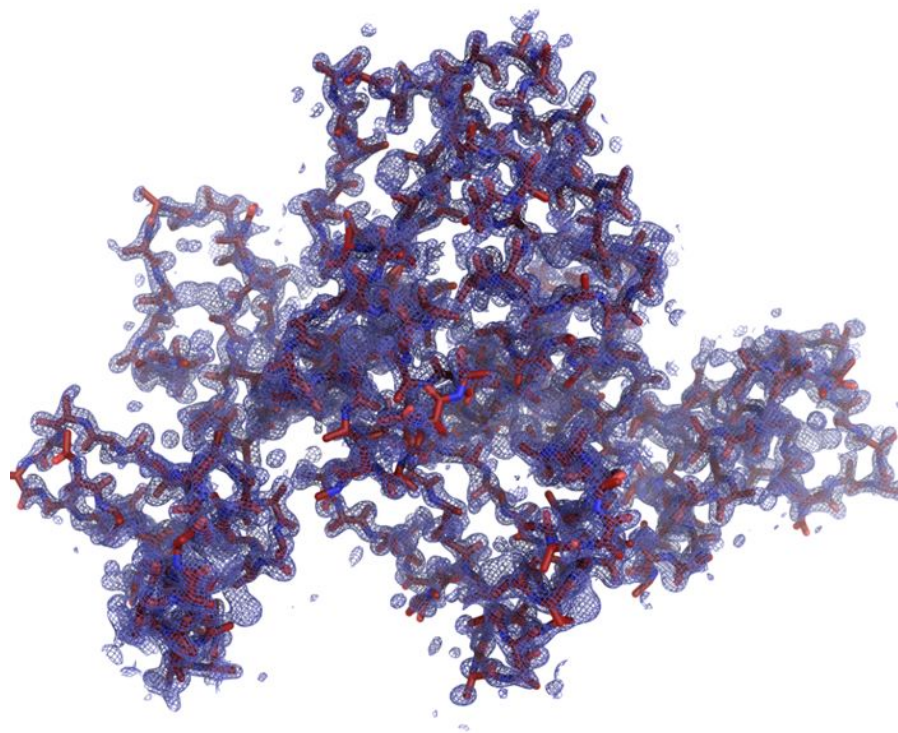
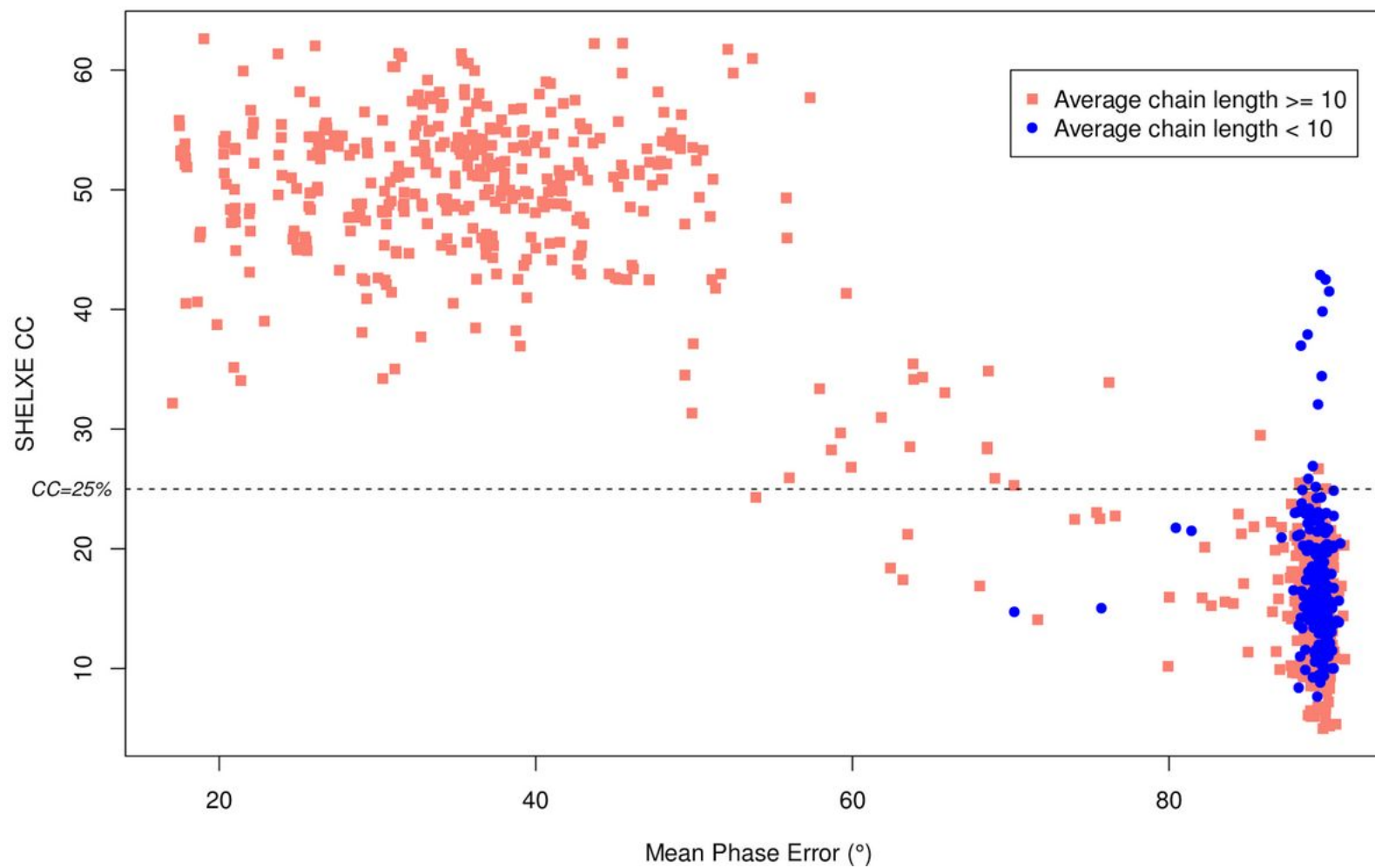


Figure courtesy of Andrea Thorn

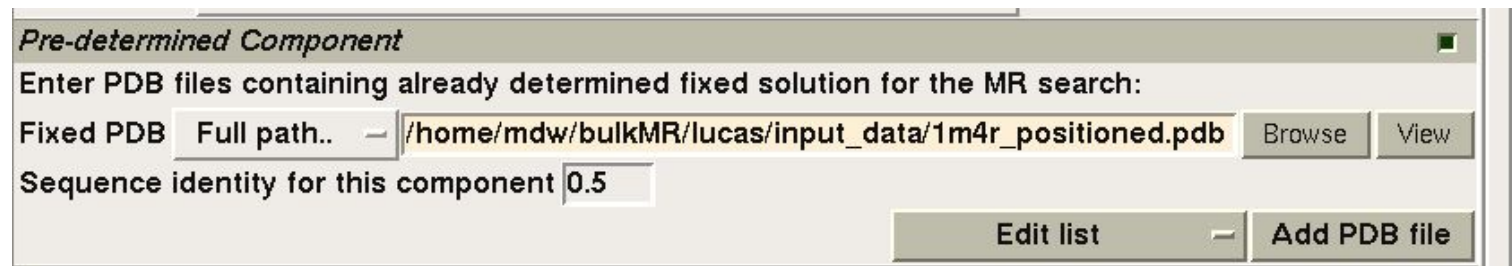


SHELXE CC vs. Mean Phase Error



Inclusion of fixed models

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

A screenshot of a software dialog box titled "Pre-determined Component". The dialog has a light gray background and a title bar. Below the title, there is a text label "Enter PDB files containing already determined fixed solution for the MR search:". Below this, there is a section for "Fixed PDB" with a "Full path.." label, a text input field containing the path "/home/mdw/bulkMR/lucas/input_data/1m4r_positioned.pdb", and two buttons labeled "Browse" and "View". Below the path field, there is a label "Sequence identity for this component" followed by a text input field containing the value "0.5". At the bottom right of the dialog, there are two buttons labeled "Edit list" and "Add PDB file".

Thus, solve complexes through consecutive runs of MrBUMP.
Automation of this in progress

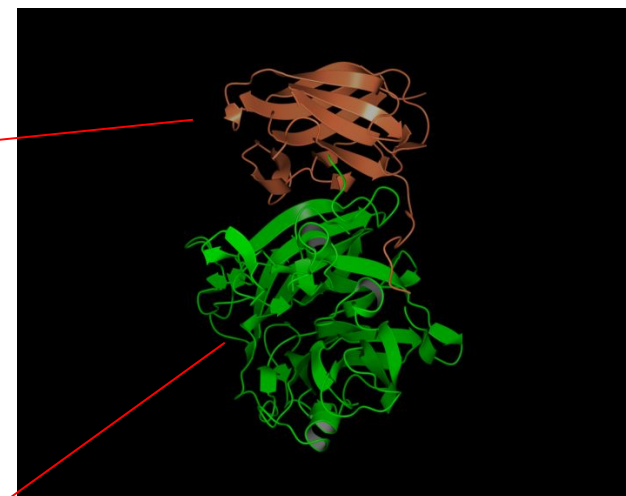


Example (thanks to Elie Vandermarliere)

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)

ARP/wARP then builds 457/493 residues to R/Rfree
0.185/0.225



MrBUMP output...

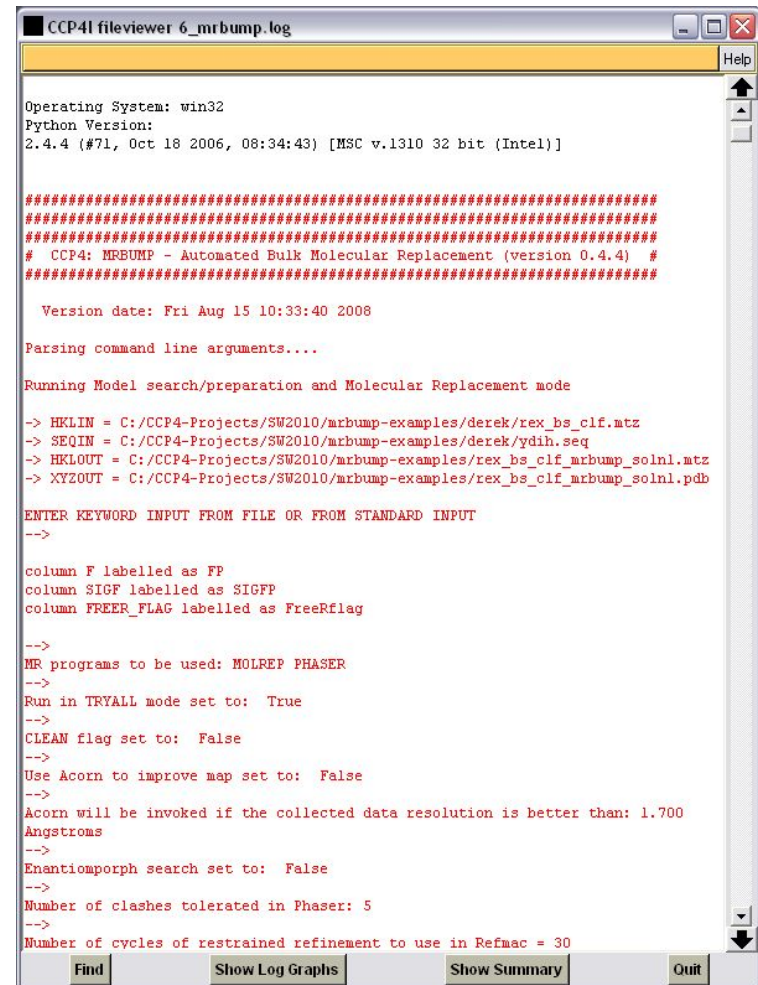
Log file gives summary of models tried
and results of MR

May get several putative solutions

Ease of subsequent model re-building,
model completion may depend on
choice of solution

Worth checking “poor” solutions

Top solution available from ccp4i



```
CCP4i fileviewer 6_mrbump.log
Help
Operating System: win32
Python Version:
2.4.4 (#71, Oct 18 2006, 08:34:43) [MSC v.1310 32 bit (Intel)]

#####
# CCP4: MRBUMP - Automated Bulk Molecular Replacement (version 0.4.4) #
#####

Version date: Fri Aug 15 10:33:40 2008

Parsing command line arguments...

Running Model search/preparation and Molecular Replacement mode

-> HKLIN = C:/CCP4-Projects/SW2010/mrbump-examples/derek/rex_bs_clf.mtz
-> SEQIN = C:/CCP4-Projects/SW2010/mrbump-examples/derek/ydih.seq
-> HKLOUT = C:/CCP4-Projects/SW2010/mrbump-examples/rex_bs_clf_mrbump_soln1.mtz
-> XYZOUT = C:/CCP4-Projects/SW2010/mrbump-examples/rex_bs_clf_mrbump_soln1.pdb

ENTER KEYWORD INPUT FROM FILE OR FROM STANDARD INPUT
-->

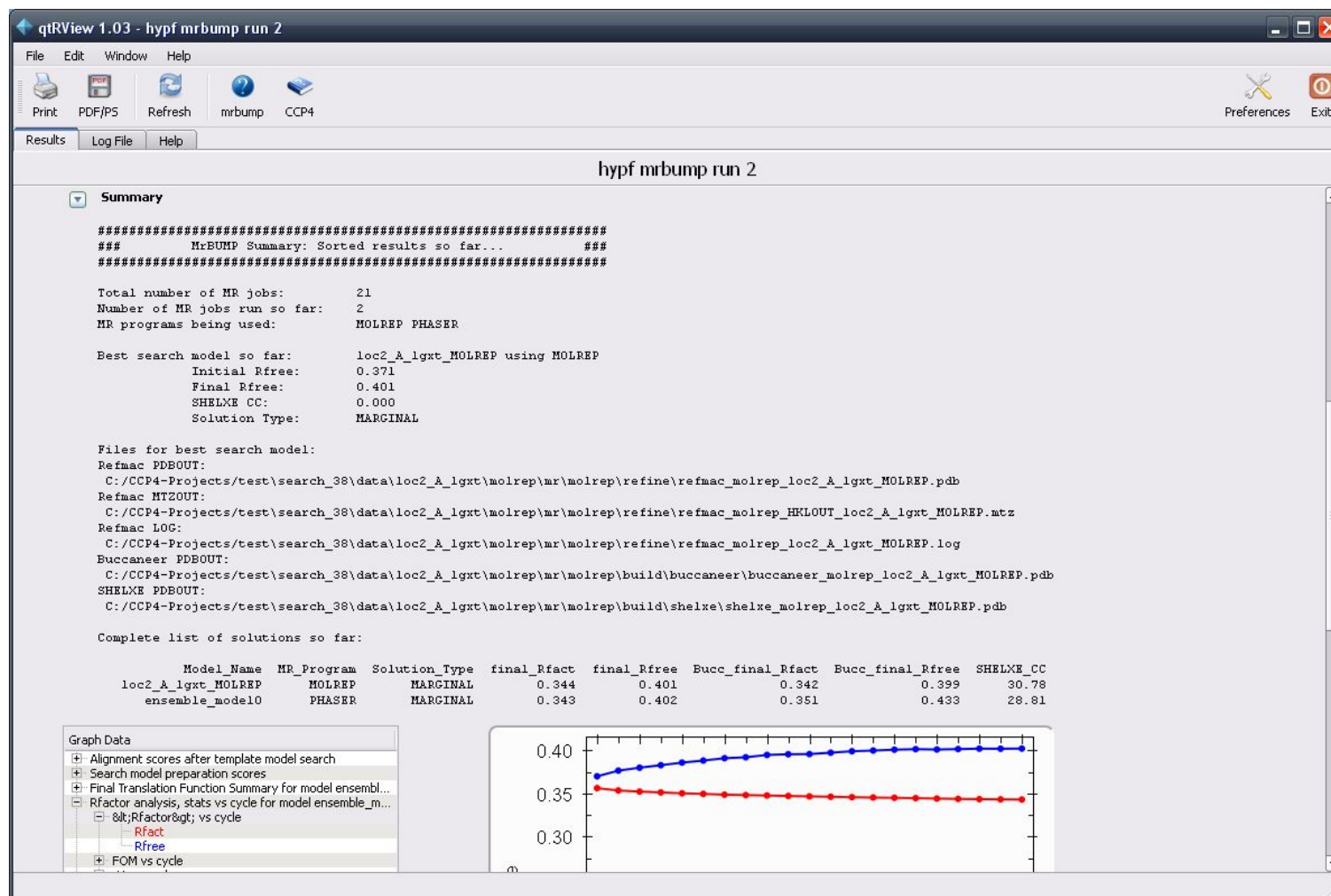
column F labelled as FP
column SIGF labelled as SIGFP
column FREER_FLAG labelled as FreeRflag

-->
MR programs to be used: MOLREP PHASER
-->
Run in TRYALL mode set to: True
-->
CLEAN flag set to: False
-->
Use Acorn to improve map set to: False
-->
Acorn will be invoked if the collected data resolution is better than: 1.700
Angstroms
-->
Enantiomorph search set to: False
-->
Number of clashes tolerated in Phaser: 5
-->
Number of cycles of restrained refinement to use in Refmac = 30

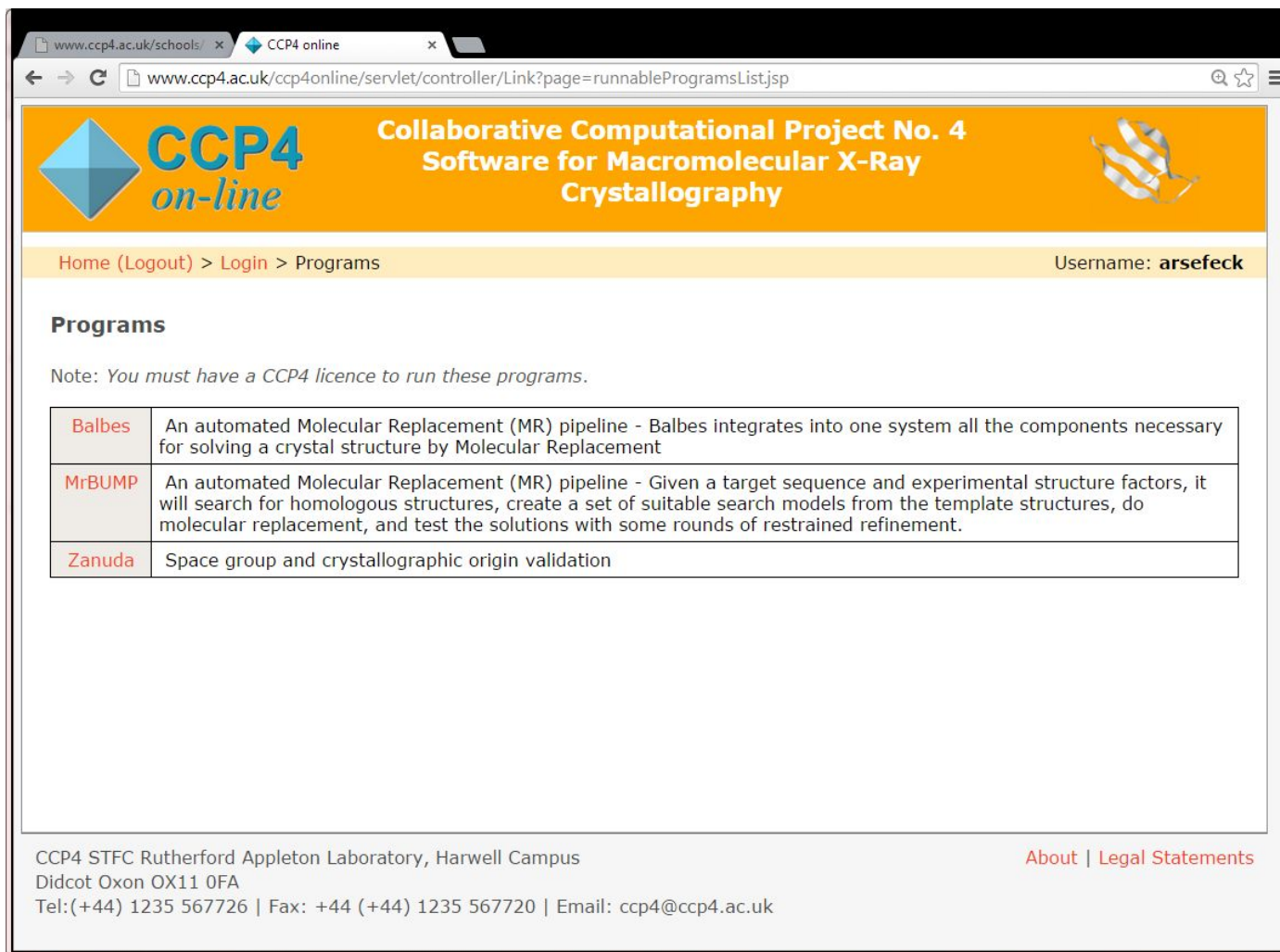
Find Show Log Graphs Show Summary Quit
```



MrBUMP output - QTRview



CCP4 Online



The screenshot shows a web browser window with the address bar displaying `www.ccp4.ac.uk/ccp4online/servlet/controller/Link?page=runnableProgramsList.jsp`. The page has an orange header with the CCP4 on-line logo, the text "Collaborative Computational Project No. 4 Software for Macromolecular X-Ray Crystallography", and a ribbon diagram icon. Below the header is a navigation bar with links "Home (Logout) > Login > Programs" and a username "Username: arsefeck". The main content area is titled "Programs" and includes a note: "Note: You must have a CCP4 licence to run these programs." Below this is a table listing three programs: Balbes, MrBUMP, and Zanuda. The footer contains contact information for CCP4 STFC Rutherford Appleton Laboratory, Harwell Campus, Didcot Oxon OX11 0FA, including telephone, fax, and email addresses, as well as links for "About" and "Legal Statements".

Program	Description
Balbes	An automated Molecular Replacement (MR) pipeline - Balbes integrates into one system all the components necessary for solving a crystal structure by Molecular Replacement
MrBUMP	An automated Molecular Replacement (MR) pipeline - Given a target sequence and experimental structure factors, it will search for homologous structures, create a set of suitable search models from the template structures, do molecular replacement, and test the solutions with some rounds of restrained refinement.
Zanuda	Space group and crystallographic origin validation

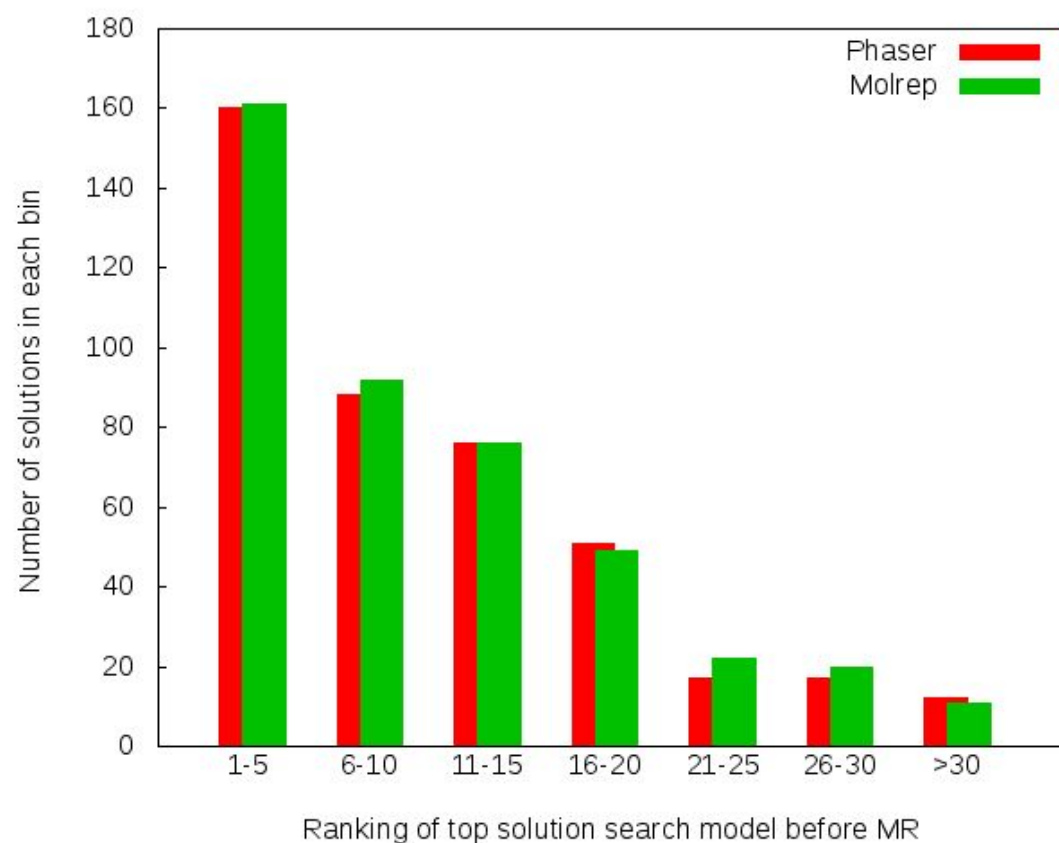


Benefit of processing many models

Sample set from 2009 PDB
depositions processed

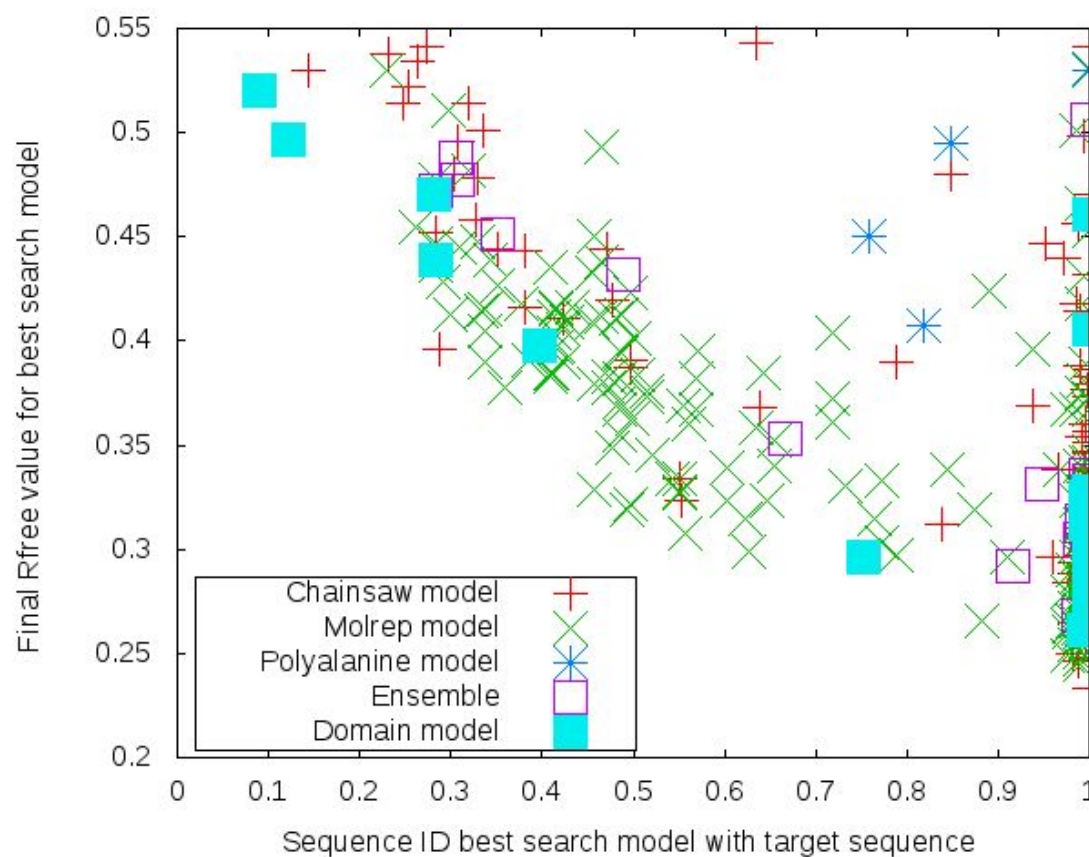
Before MR step all search
models are scored based
on sequence alignment

Best MR solutions based
against initial scoring of
the search model used
(results are for
successful MR jobs only)



MrBUMP Testing Results

- Model type for best solution (successful MR test cases)



Acknowledgements

- Jaclyn Bibby, Daniel Rigden, Jens Thomas, University of Liverpool
- Olga Mayans, University of Liverpool
- Martyn Winn, Daresbury Laboratory
- Andrea Thorn, Tim Gruene & George Sheldrick (SHELX)
- Developers of Rosetta and Quark
- Refmac: Garib Mushudov, LMB-MRC Cambridge
- Molrep: Alexei Vagin & Andrey Lebedev
- Phaser: Randy Read, Airlie McCoy & Gabor Bunkozci
- Thanks to authors of all underlying programs
- Funding:
 - BBSRC, CCP4
- Support from CCP4 & the Research Complex at Harwell

