



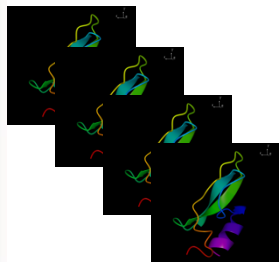
MrBUMP – Automated Molecular Replacement

***Ronan Keegan and Martyn Winn
STFC Rutherford Appleton Laboratory &
STFC Daresbury Laboratory***



**Science & Technology
Facilities Council**

The aim of MrBUMP



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

Wraps **Phaser** and/or **Molrep**.

- Also uses a variety of helper applications (e.g. Chainsaw) and bioinformatics tools (e.g. Fasta, Mafft)
- Uses 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)

Pipeline

Target MTZ
&
Sequence



Target
Details



Template
Search



Model
Preparation

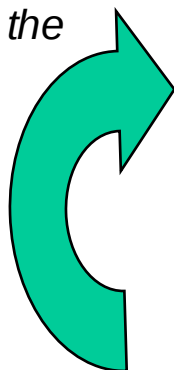


Molecular Replacement
& Refinement



Phase Improvement

*Check scores and
exit or select the
next model*



*Farm MR/Refinement jobs
to cluster*



Search for model templates

FASTA search of PDB

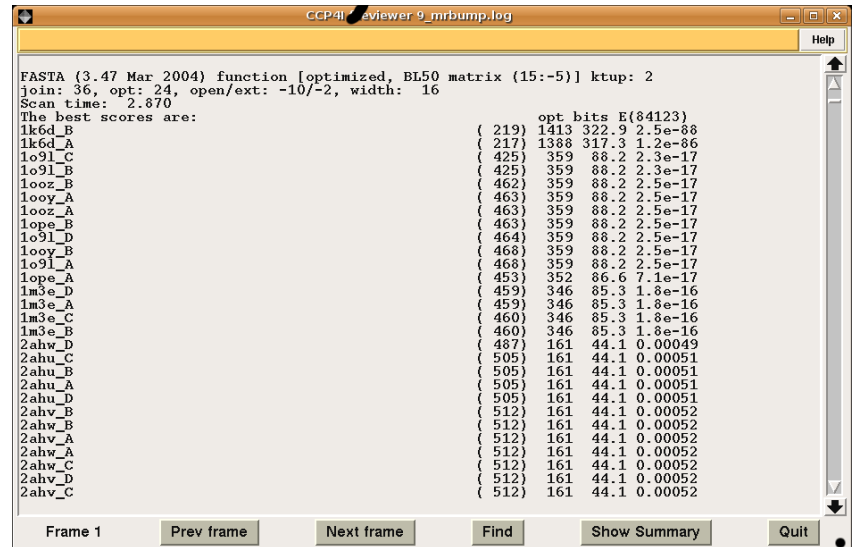
- **Sequence based search** using sequence of target structure

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. from **FFAS** or **psiBLAST** searches
- Can add local PDB files

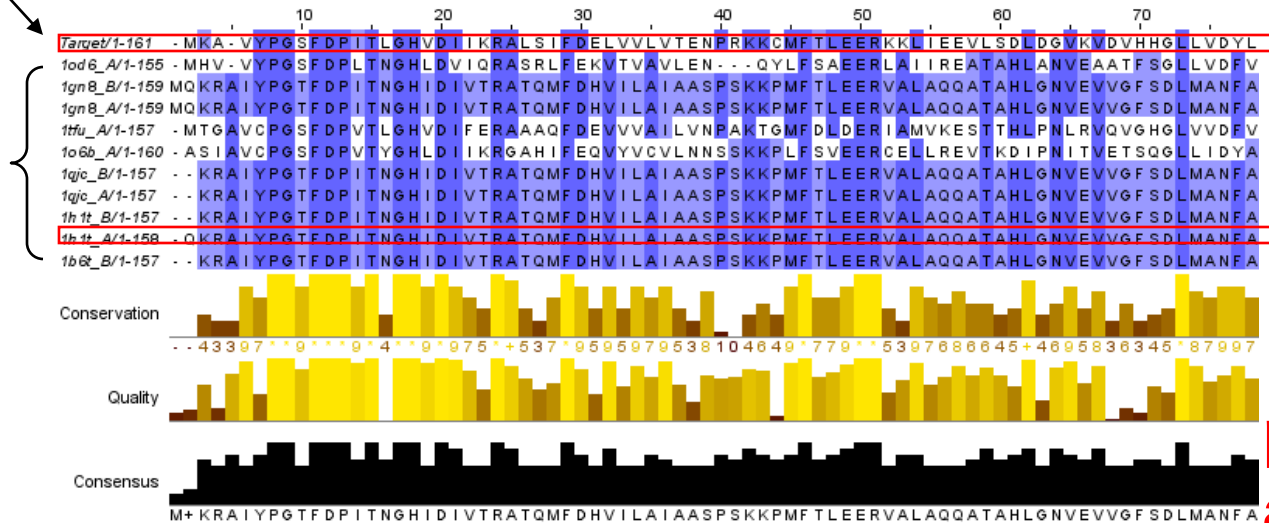


The screenshot shows a window titled "CCP4 viewer 9_mrbump.log" with a "Help" button in the top right corner. The main text area displays the output of a FASTA search performed on March 3, 2004, using an optimized BL50 matrix (15:-5) with a ktup of 2. The search parameters are: join: 36, opt: 24, open/ext: -10/-2, width: 16. The scan time was 2.870. The results show the best scores for various PDB structures, with the top hit being 1k6d_B (opt bits E(84123), score 1413.322.9, 2.5e-88). Other structures listed include 1k6d_A, 1o9l_C, 1o9l_B, 1ooz_B, 1ooz_A, 1ope_B, 1o9l_D, 1ooz_B, 1o9l_A, 1ope_A, 1m3e_D, 1m3e_A, 1m3e_C, 1m3e_B, 2ahw_D, 2ahu_C, 2ahu_B, 2ahu_A, 2ahu_D, 2ahv_B, 2ahv_A, 2ahw_C, 2ahw_D, and 2ahv_C. The window has a scroll bar on the right and buttons for "Frame 1", "Prev frame", "Next frame", "Find", "Show Summary", and "Quit" at the bottom.

```
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
1ooz_B                          ( 462) 359  88.2 2.5e-17
1ooz_A                          ( 463) 359  88.2 2.5e-17
1ope_B                          ( 463) 359  88.2 2.5e-17
1o9l_D                          ( 464) 359  88.2 2.5e-17
1ooz_B                          ( 468) 359  88.2 2.5e-17
1o9l_A                          ( 468) 359  88.2 2.5e-17
1ope_A                          ( 453) 352  86.6 7.1e-17
1m3e_D                          ( 459) 346  85.3 1.8e-16
1m3e_A                          ( 459) 346  85.3 1.8e-16
1m3e_C                          ( 460) 346  85.3 1.8e-16
1m3e_B                          ( 460) 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
2ahw_C                          ( 512) 161  44.1 0.00052
2ahw_D                          ( 512) 161  44.1 0.00052
2ahv_C                          ( 512) 161  44.1 0.00052
```

Multiple Alignment step

target



pairwise
alignment
(used in
Chainsaw)

Jalview 2.08.1 Barton group, Dundee

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

Model template scoring: score = sequence identity X alignment quality

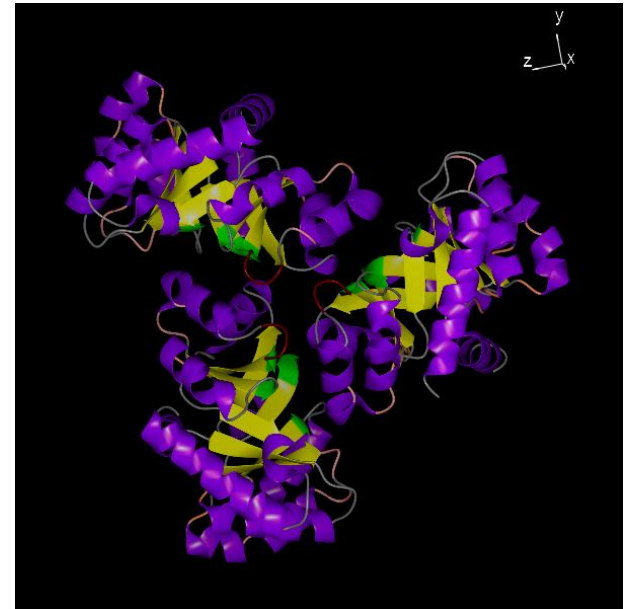
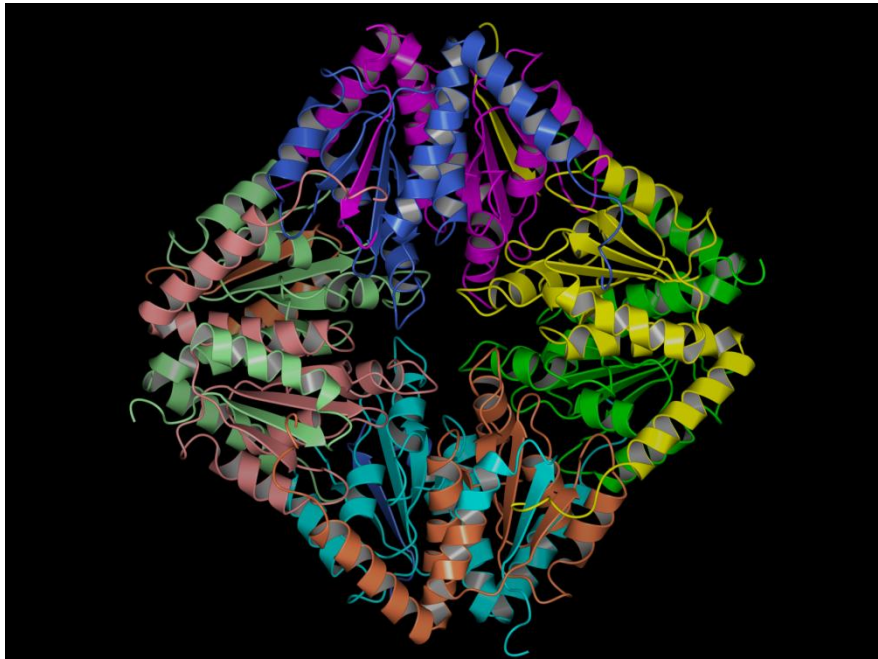
Domains

- Suitable templates for target domains may exist in isolation in 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
- 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

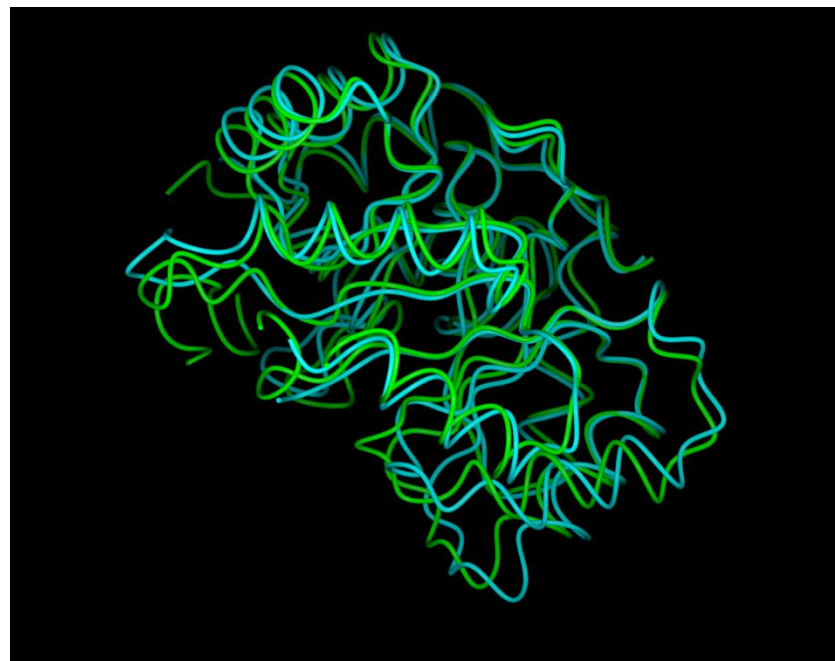
- Use template multimer as model for target multimer (currently uses PQS, will use PISA)
- Better signal-to-noise ratio than monomer, *if* assembly is correct for the target.
- Biologically relevant multimers more likely transferable



prepared with CCP4mg

Ensemble models

- Create ensembles of top search models, for use in additional runs of **Phaser**.
- Models must be sufficiently similar (MW and rmsd)
- **Molrep** can also use ensembles (not yet implemented)



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 contains a model preparation function which will align the template sequence with the target sequence and prune 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

more side
chain
truncation

deal with
deletions

Also create an ensemble models based on top models



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

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

Spacegroup from MTZ file: 'P 31 2 1'

Do MR using enantiomorphic spacegroup as well ☒

- [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%

⇒ “good”

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

⇒ “marginal”

otherwise

⇒ “poor”

conservative



Phase improvement

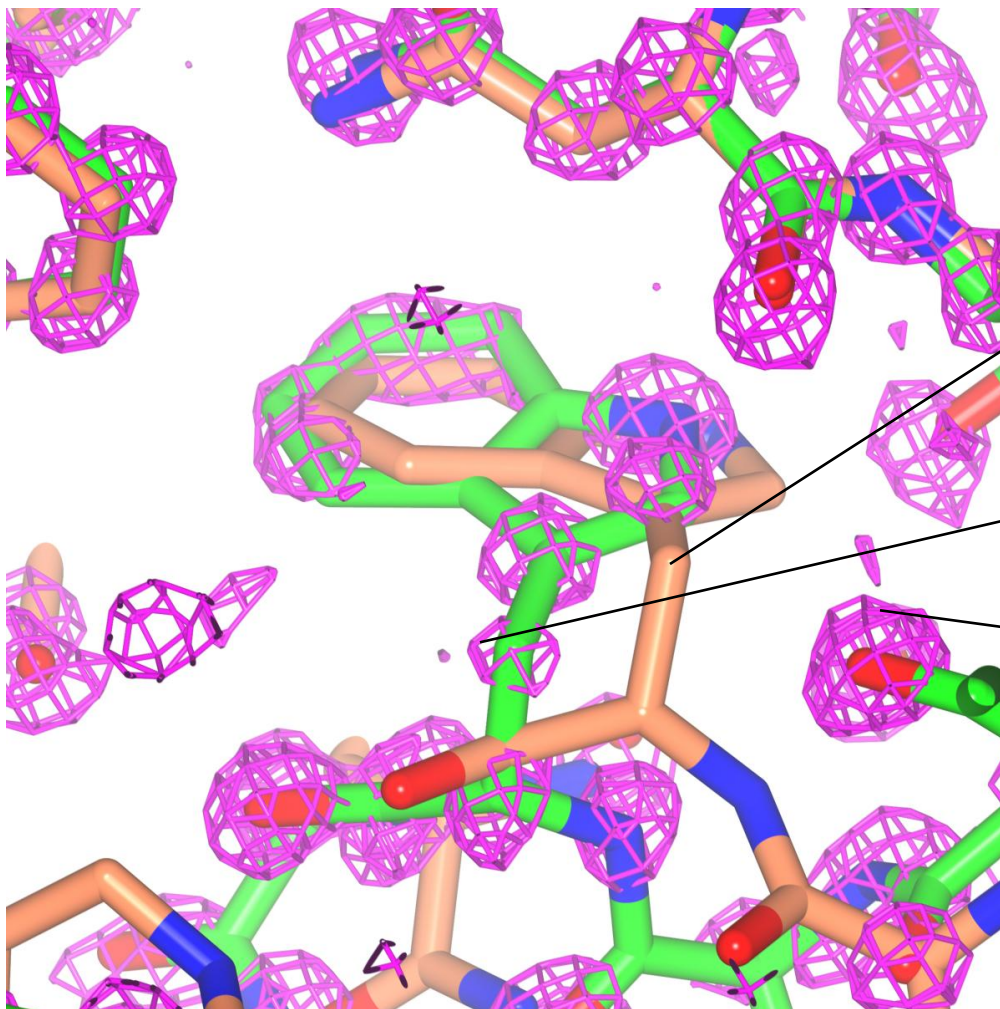
If resolution better than 1.7Å use Acorn procedure:
initial phase set from refined MR solution
artificial phase extension to 1.0Å
dynamic density modification

Result:

CC for medium Es good indicator of solution
Use E-maps for re-building

At lower resolutions:

Use parrot (todo!)



dUTPase from *C.jejuni*
data to 1.65Å

positioned/refined
search model

final model (1w2y)

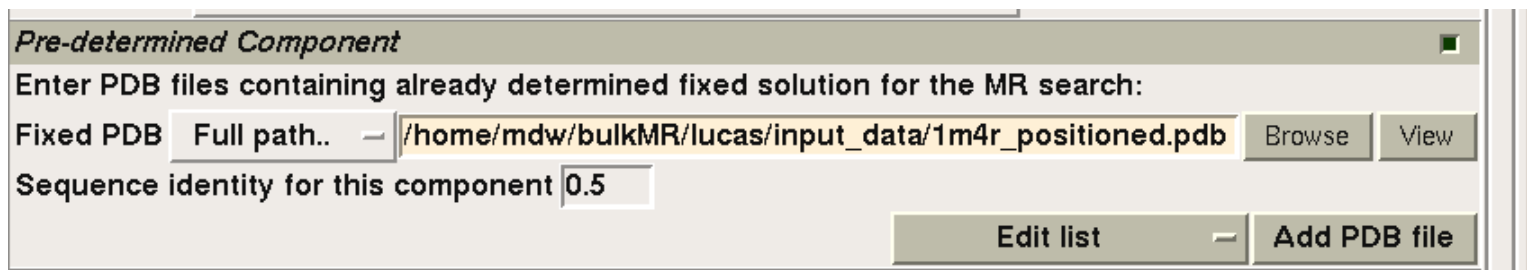
Acorn map (as
generated by
MrBUMP)

CC: 0.078 → 0.156
ARP/wARP re-builds
into Acorn map

ccp4mg

Inclusion of fixed models

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



The screenshot shows a software dialog box titled "Pre-determined Component". It contains the instruction "Enter PDB files containing already determined fixed solution for the MR search:". Below this, there is a label "Fixed PDB" followed by a text input field labeled "Full path.." containing the file path "/home/mdw/bulkMR/lucas/input_data/1m4r_positioned.pdb". To the right of the input field are "Browse" and "View" buttons. Below the input field 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 are two buttons: "Edit list" and "Add PDB file".

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

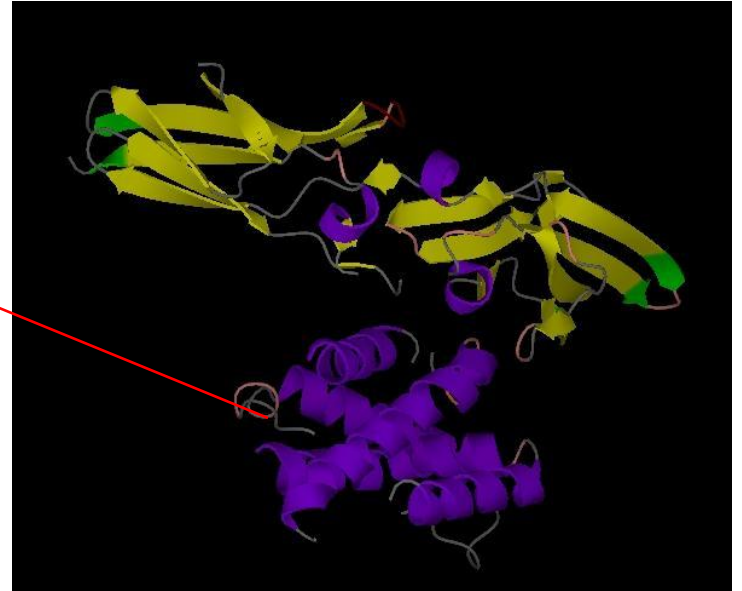
Example (with thanks to Lucas Bleicher)

1:1 complex (1 copy in a.s.u.)
data to 1.9Å in C2

Small protein (151 res) already
solved, easy to locate in complex.

Larger protein (217 res), 2
domains, nearest homologs
around 26%

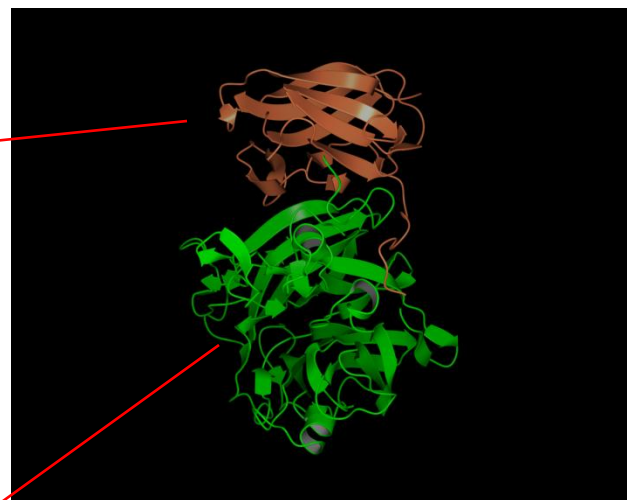
Run MrBUMP with small protein fixed.



Example (thanks to Elieen Vandermarliere)

Target is an arabinofuranosidase
Data to 1.55Å in P212121

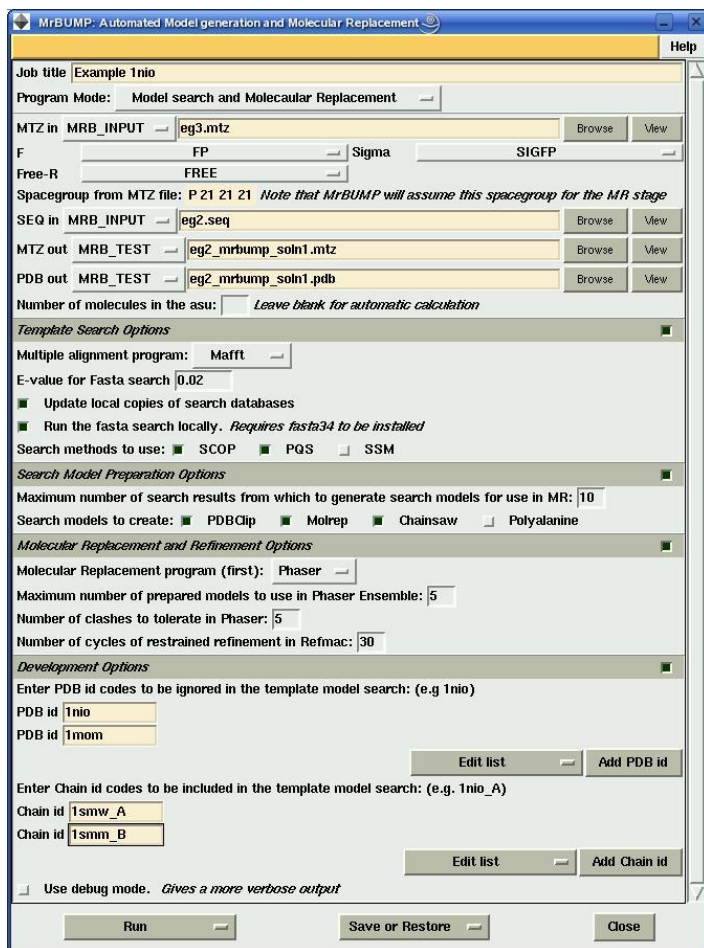
Small C domain (144 res) solved with
34% seq ident model
(1w9t_B_MOLREP best out of 4
solutions)



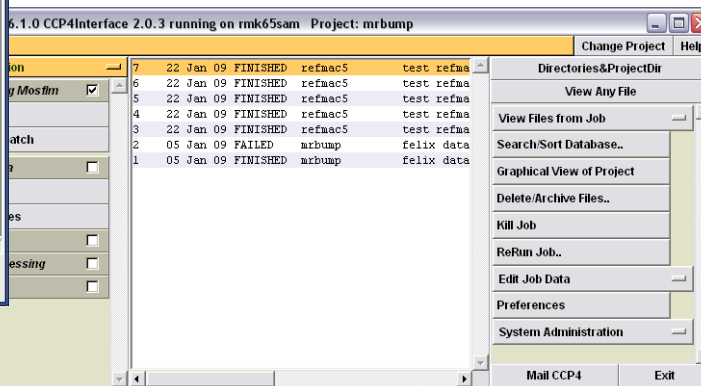
With C domain solution fixed, large N domain (345 res) solved with
28% seq ident model
(1gyh_C_CHNSAW best out of 7 solutions)

Acorn: CC increases from 0.04 to 0.18
ARP/wARP then builds 457/493 residues to R/Rfree
0.185/0.225

MrBUMP in CCP4i

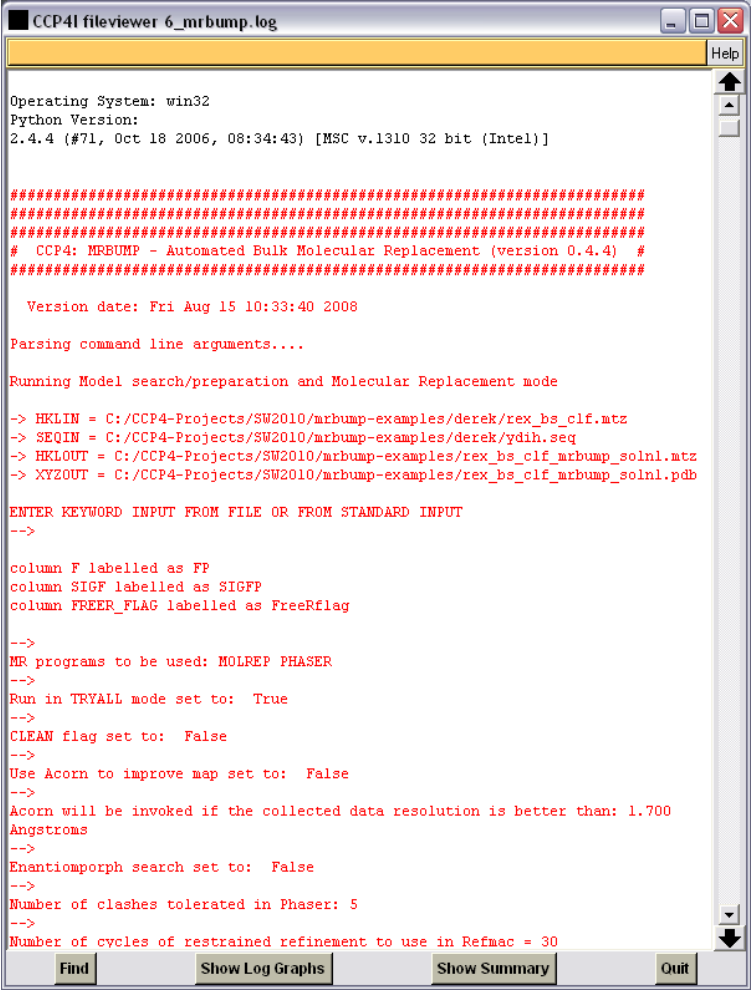


- MrBUMP included in CCP4 6.1 series
- Runs on Linux, OSX and Windows.
- Comes with CCP4 GUI .
- Can also be run from the command line with keyword input
- Tutorials available



MrBUMP Output

- Log file gives summary of models tried and results of MR
- May get several putative solutions
- Ease of subsequent model rebuilding, 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
-->

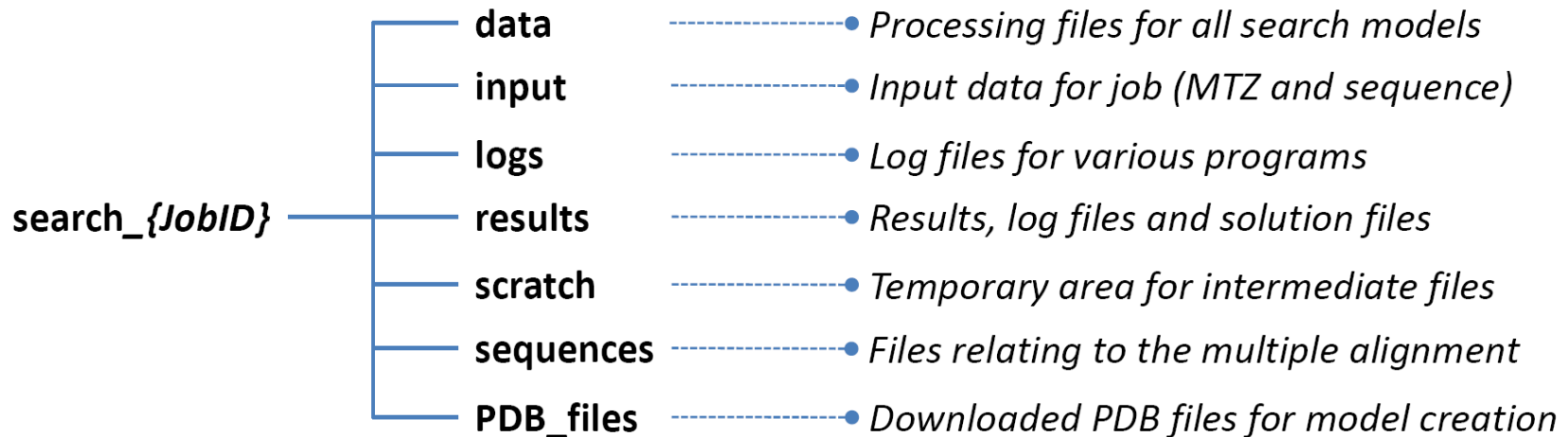
column F labelled as FP
column SIGF labelled as SIGFP
column FREER_FLAG labelled as FreeRflag

-->
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
```

Output files

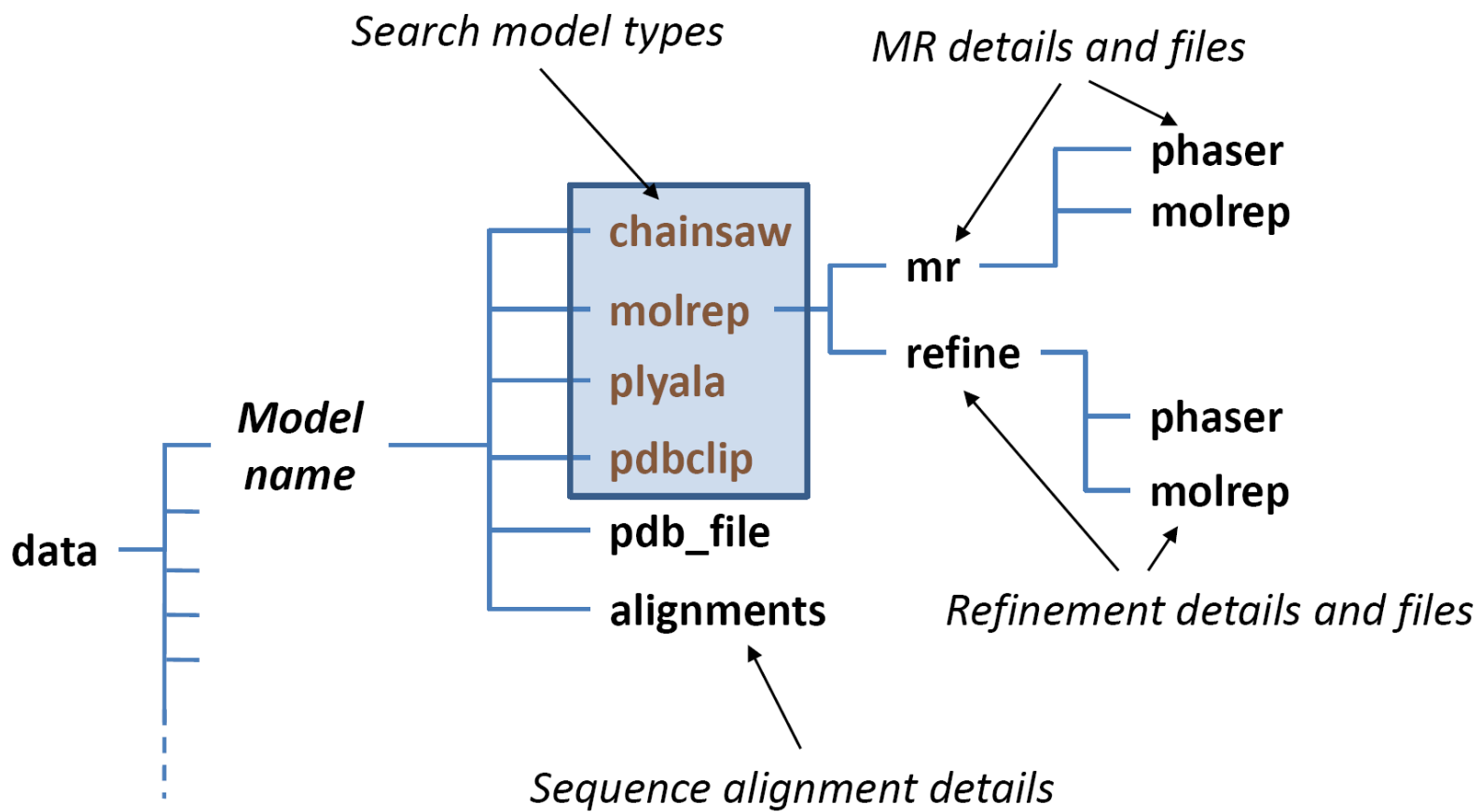
- Detailed results located in:
 - `<ccp4i project directory>/search_<job number>`





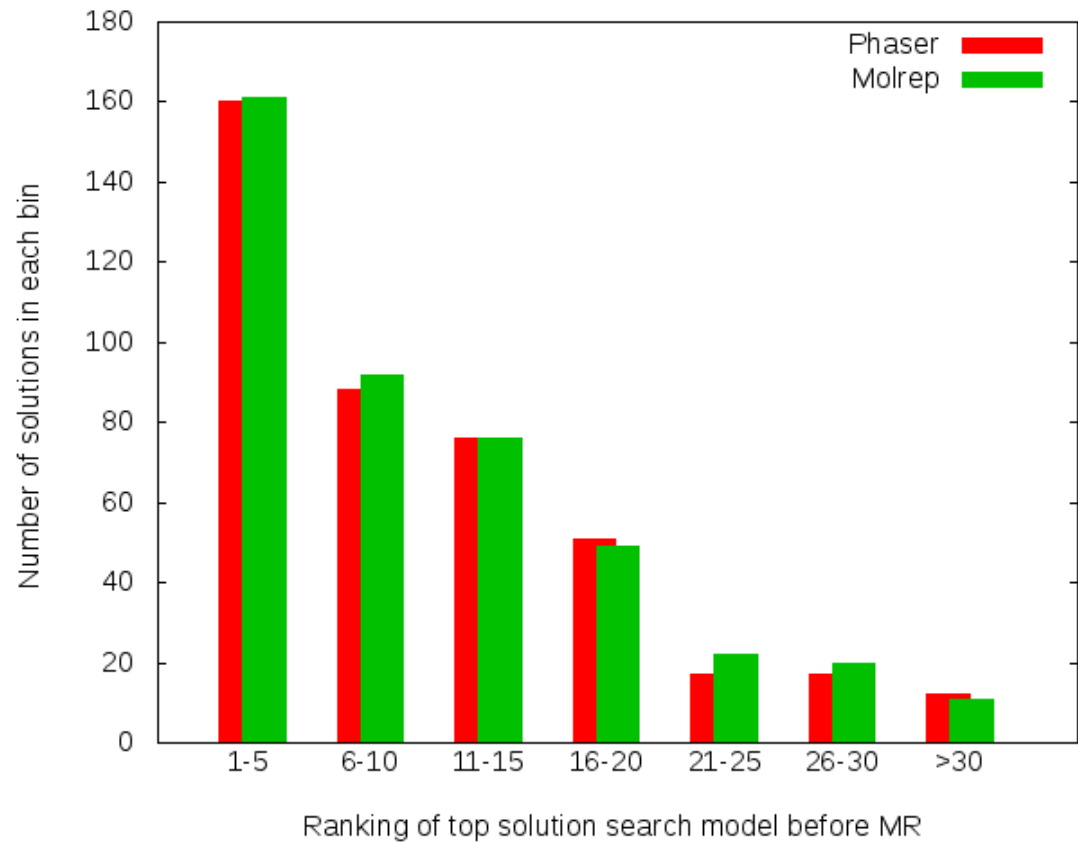
Output files

- 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 projectdirectory>/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**".



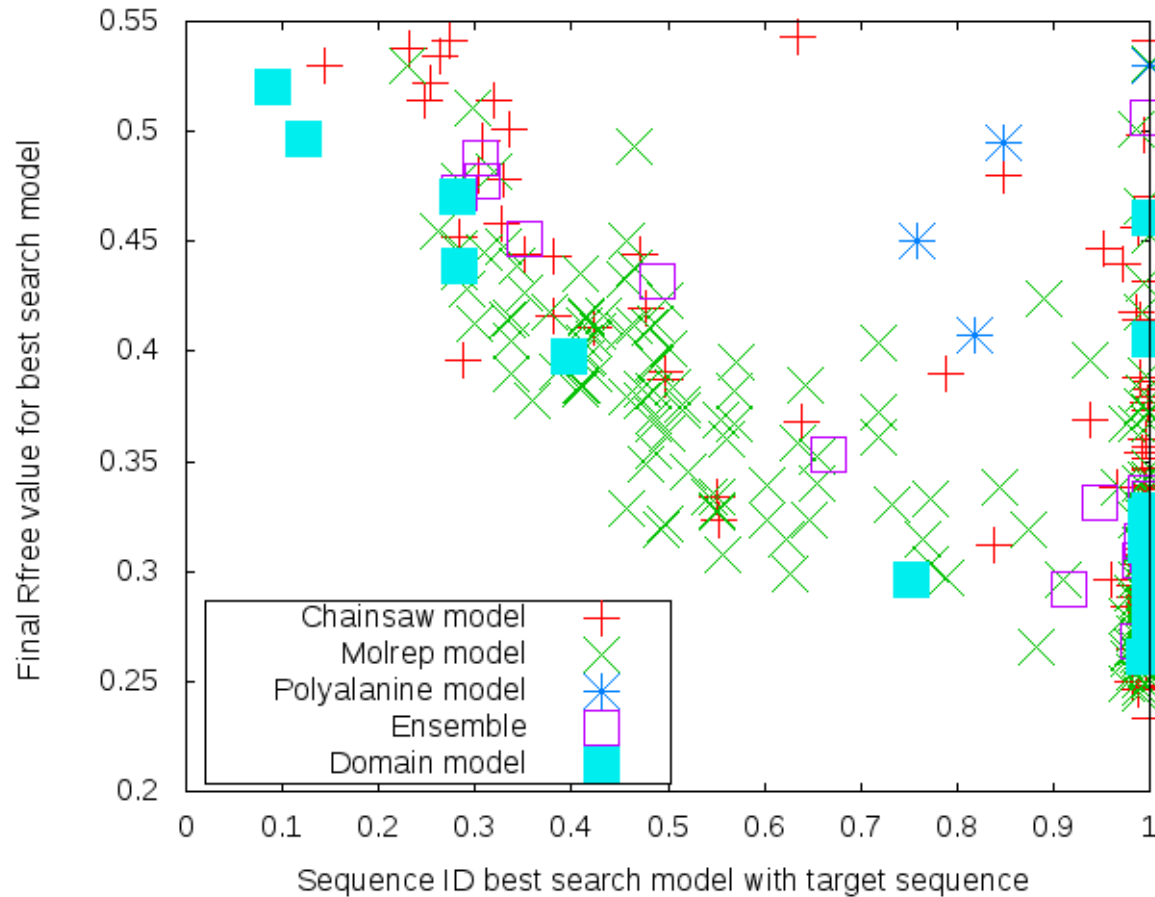
MrBUMP Testing Results

- Sample set from 2009 PDB depositions processed
- Best MR solutions based against initial scoring of the search model used



MrBUMP Testing Results

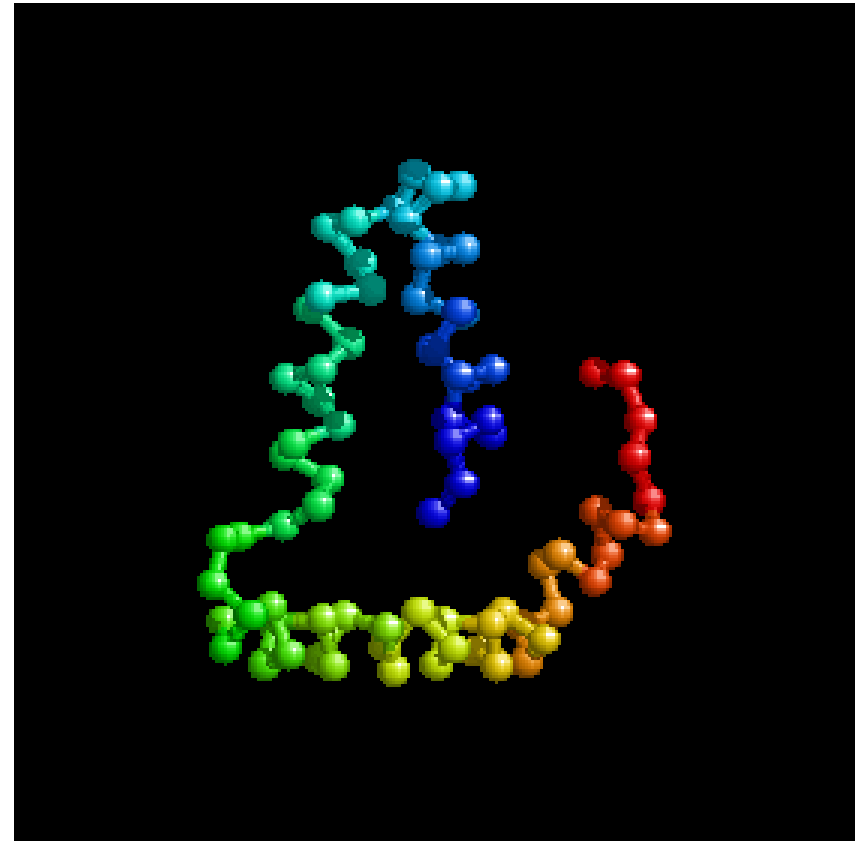
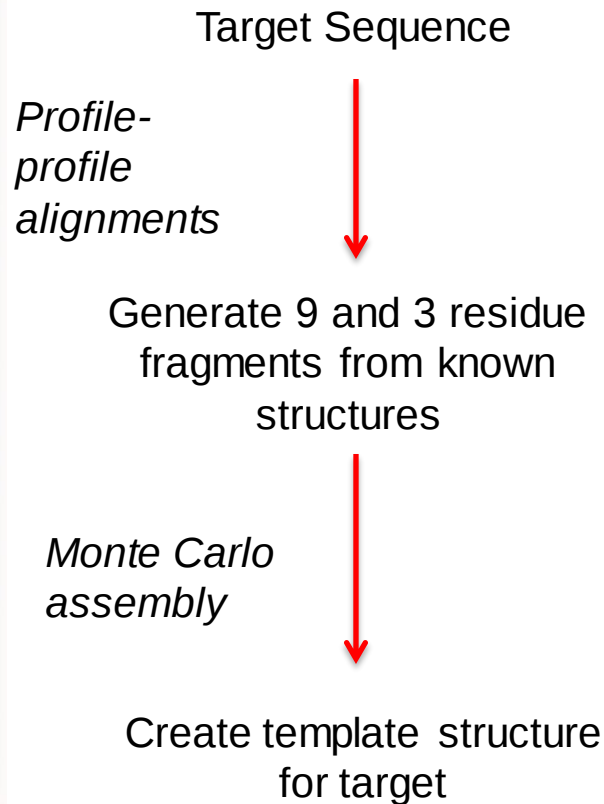
- Model type for best solution



AMPLE – ab initio modelling for Molecular Replacement

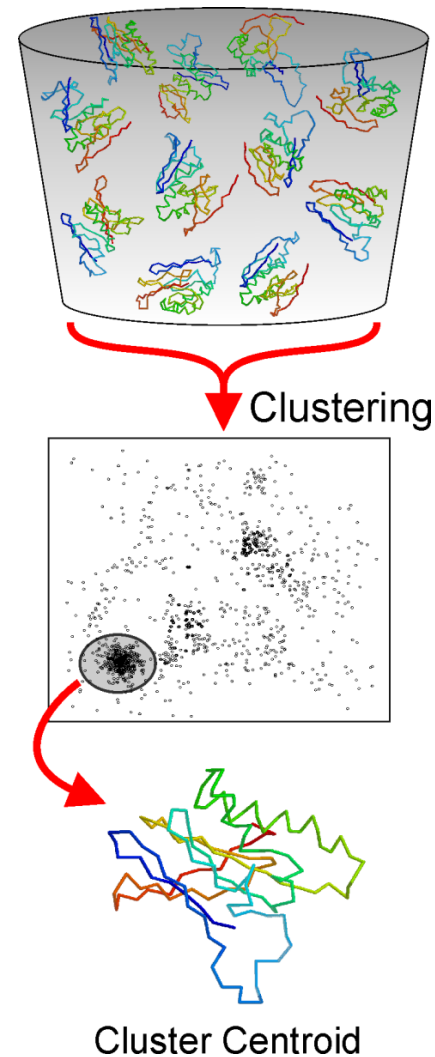
- *ab initio* aka *de novo*, template-free
- Emerging, 1997 onwards
- Protein sequence alone encodes structure...
- ∴ In principle, can address *any* protein
- CPU-intensive
- Limited to max 100-120 residues (but this is typical domain size!)
- Examples: Rosetta, Quark

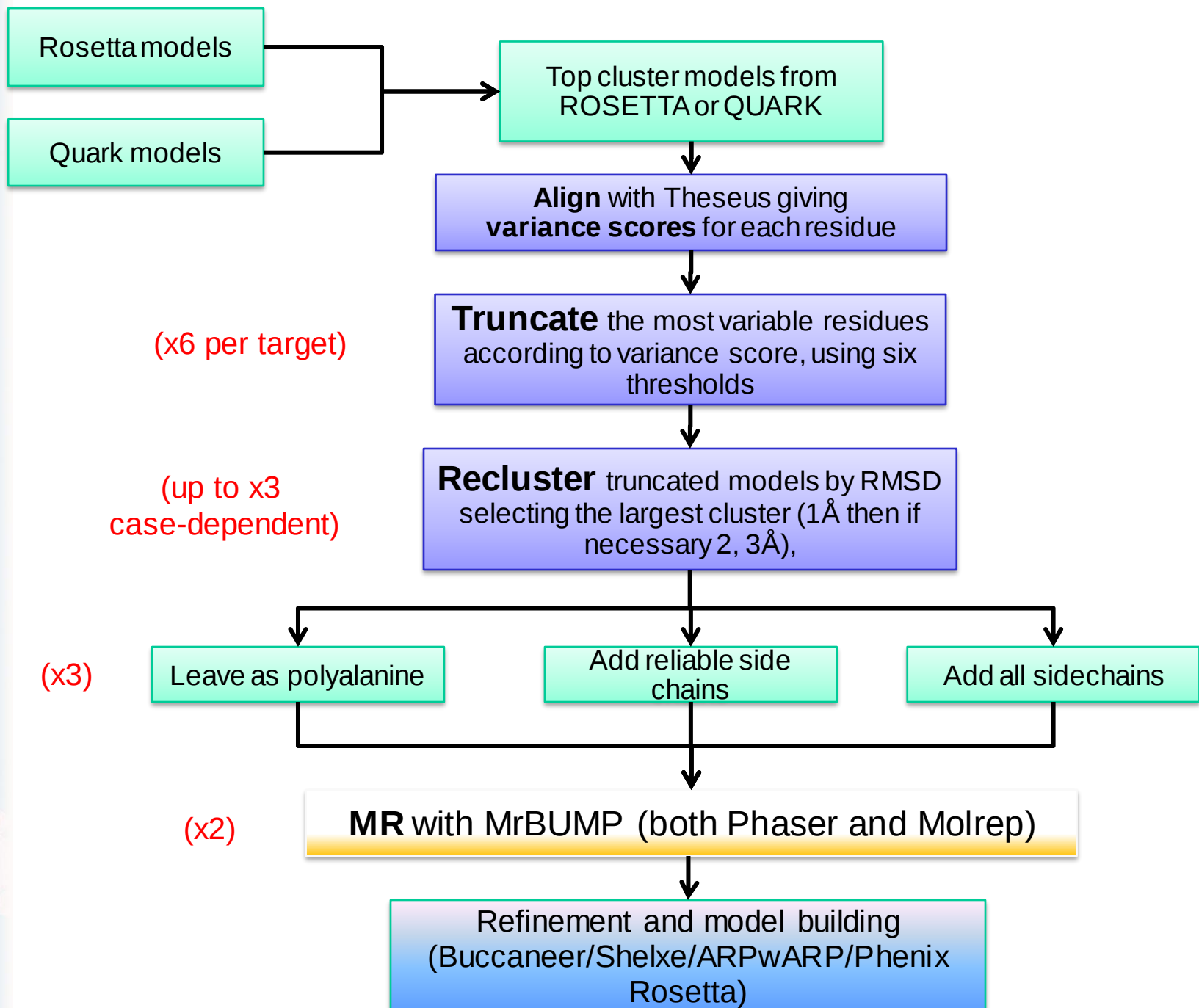
Ab initio modelling – fragment assembly: Rosetta



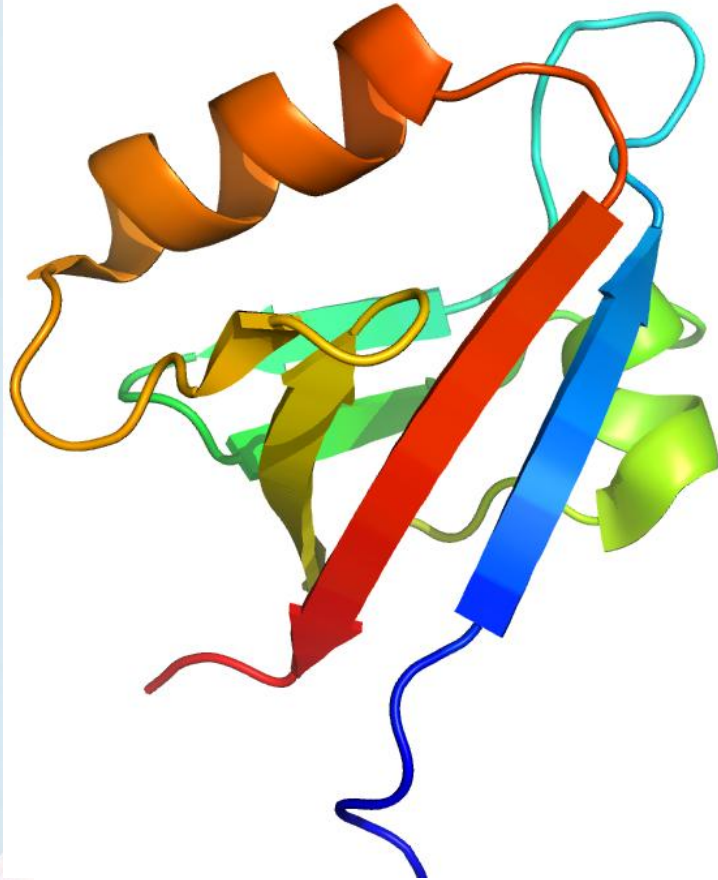
AMPLE Pipeline

- Designed to make ab initio modelling for use in MR easily accessible to users
- Generate large number of trial models
- Cluster into groups according to RMS deviation
- Select template from largest cluster

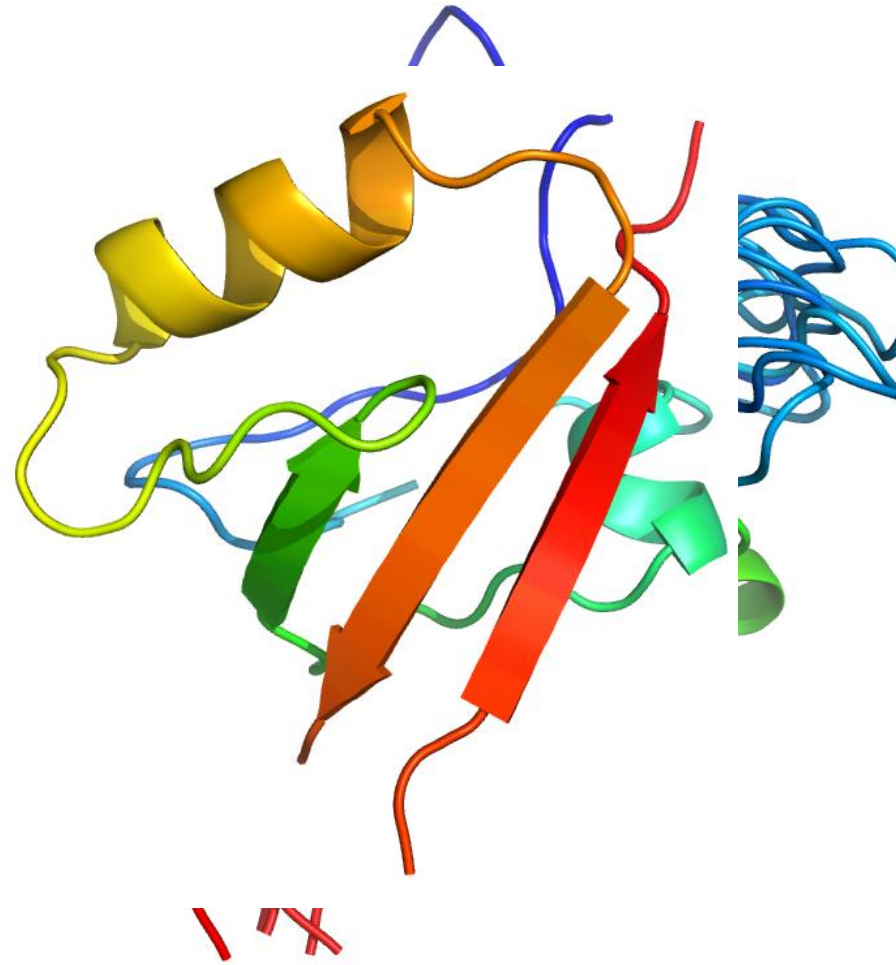




Rosetta success (1r6j, a PDZ domain)

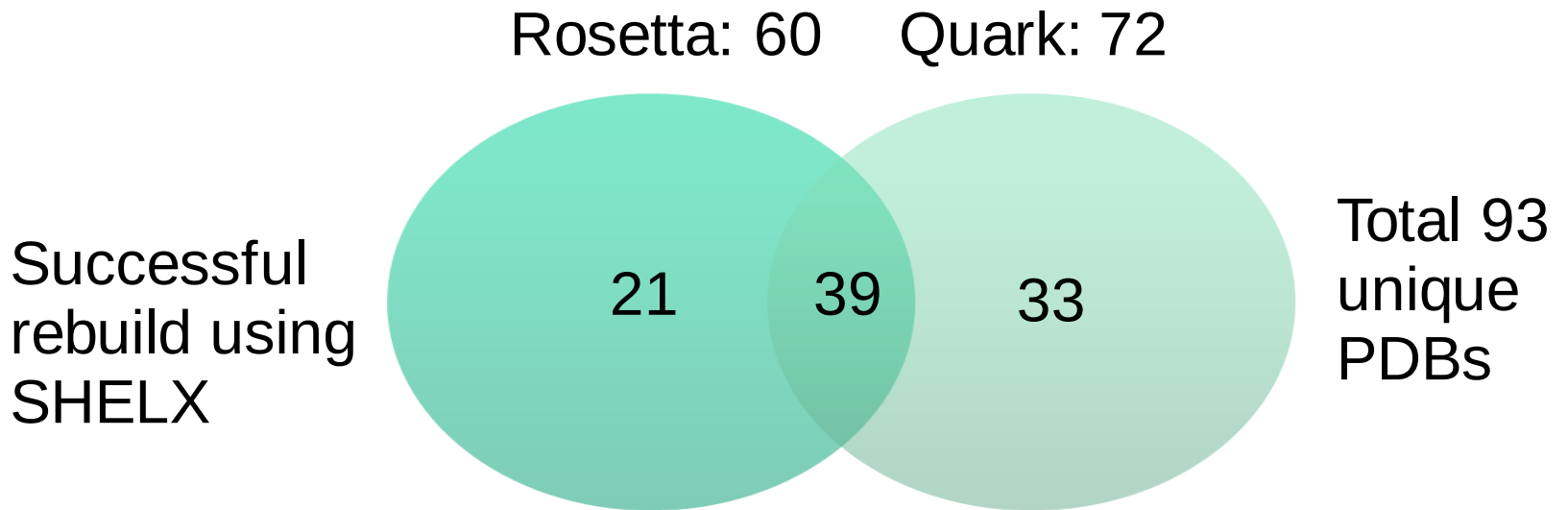


crystal structure



unanchored Rosetta rebuilt
cluster

How many are successful after MR?



Out of 296 cases, 93 are successful.

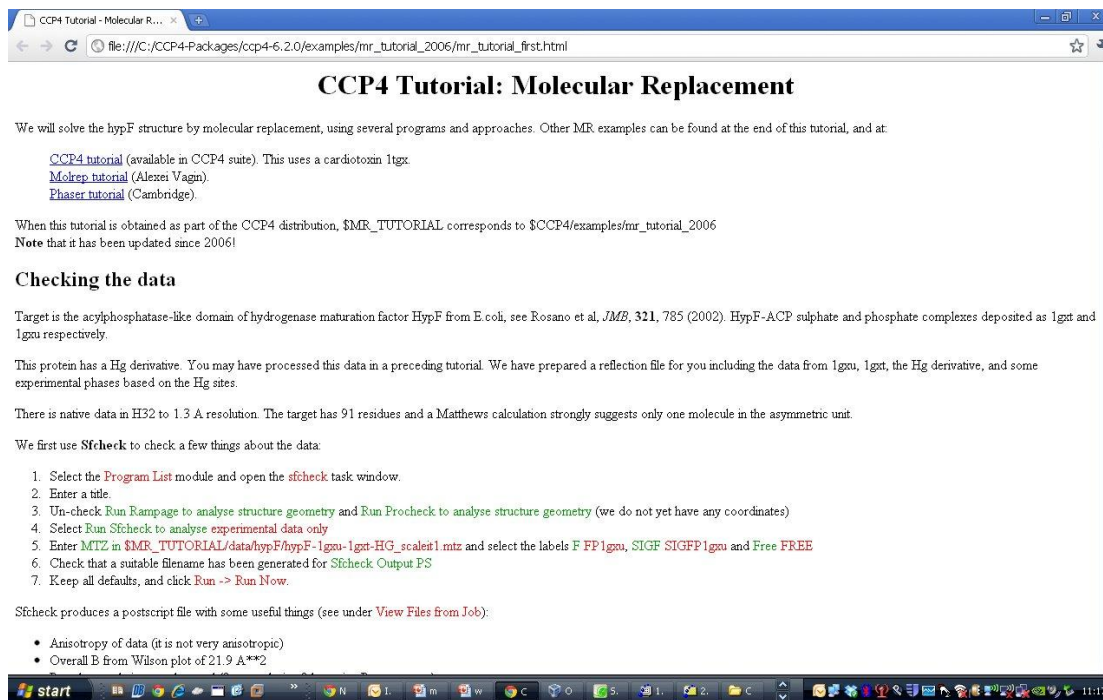


Availability

- To be released with next CCP4 release
- Still in development
- Installed on CCP4 cluster in UK and available to anyone in OIST workshop

MR/MrBUMP Tutorial

- Handout in folder or
- In the CCP4 installation under:
 - `$CCP4/examples/mr_tutorial_2006`
 - 2 HTML pages - beginners and advanced



CCP4 Tutorial - Molecular Replacement

We will solve the hypF structure by molecular replacement, using several programs and approaches. Other MR examples can be found at the end of this tutorial, and at:

- [CCP4 tutorial](#) (available in CCP4 suite). This uses a cardiotoxin 1ltx.
- [Molrep tutorial](#) (Alexei Vagin).
- [Phaser tutorial](#) (Cambridge).

When this tutorial is obtained as part of the CCP4 distribution, \$MR_TUTORIAL corresponds to \$CCP4/examples/mr_tutorial_2006
Note that it has been updated since 2006!

Checking the data

Target is the acylphosphatase-like domain of hydrogenase maturation factor HypF from E.coli, see Rosano et al, *JMB*, 321, 785 (2002). HypF-ACP sulphate and phosphate complexes deposited as 1gxv and 1gxu respectively.

This protein has a Hg derivative. You may have processed this data in a preceding tutorial. We have prepared a reflection file for you including the data from 1gxu, 1gxt, the Hg derivative, and some experimental phases based on the Hg sites.

There is native data in H32 to 1.3 Å resolution. The target has 91 residues and a Matthews calculation strongly suggests only one molecule in the asymmetric unit.

We first use Sfccheck to check a few things about the data:

1. Select the **Program List** module and open the **sfccheck** task window.
2. Enter a title.
3. Un-check **Run Rampage** to analyse structure geometry and **Run Procheck** to analyse structure geometry (we do not yet have any coordinates)
4. Select **Run Sfccheck** to analyse experimental data only
5. Enter **MTZ** in **\$MR_TUTORIAL/data/hypF/hypF-1gxu-1gxt-HG_scale1.mtz** and select the labels **F FP1gxu**, **SIGF SIGFP1gxu** and **Free FREE**
6. Check that a suitable filename has been generated for **Sfccheck Output PS**
7. Keep all defaults, and click **Run** -> **Run Now**.

Sfccheck produces a postscript file with some useful things (see under **View Files from Job**):

- Anisotropy of data (it is not very anisotropic)
- Overall B from Wilson plot of 21.9 Å²

Acknowledgements

- Martyn Winn CCP4 @ Daresbury
- Thanks to authors of all underlying programs and services
(see references in MrBUMP log file)
- Other suggestions from:
 - Vincent Fazio, CSIRO, Australia
 - Dave Meredith, Graeme Winter, Daresbury Laboratory.
 - Eugene Krissinel, EBI, Cambridge.
 - Eleanor Dodson, YSBL, York University
 - Geoff Barton, Charlie Bond, University of Dundee
 - Randy Read, Airlie McCoy, Gabor Bunkozci, Cambridge
- AMPLE:
 - Jaclyn Bibby, Daniel Rigden, Olga Mayans, University of Liverpool
- Funding:
 - BBSRC (e-HTPX, CCP4)

<http://www.ccp4.ac.uk/MrBUMP>