Automatic Molecular Replacement

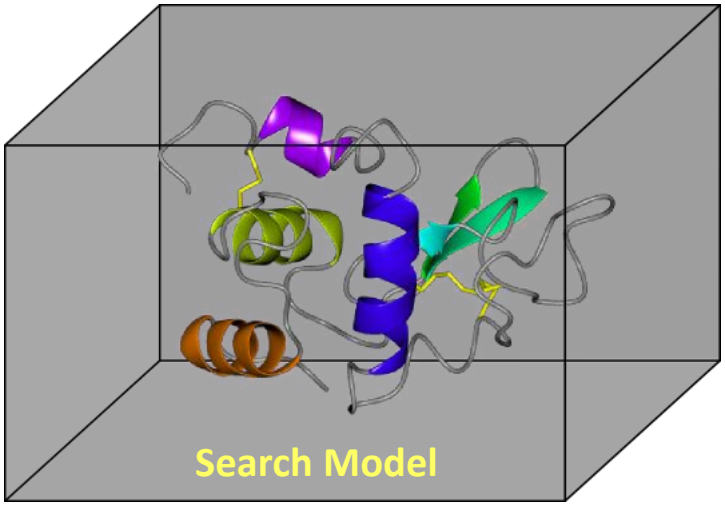
Ronan Keegan

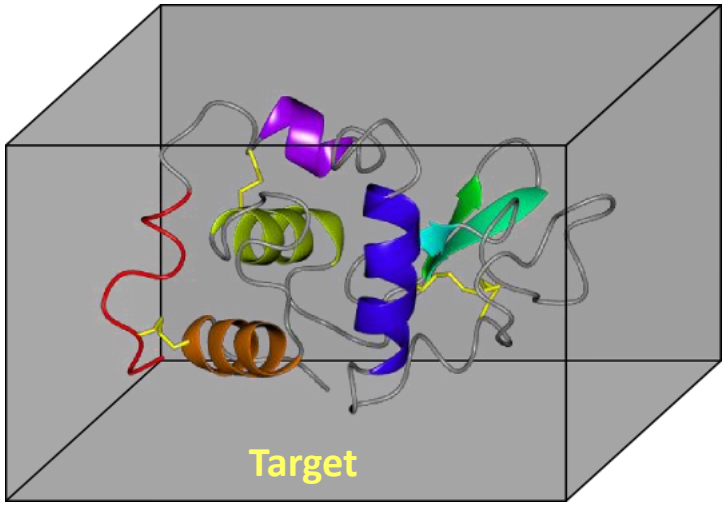
STFC Rutherford Appleton Laboratory

ronan.keegan@stfc.ac.uk

Molecular Replacement: The idea

(Slides courtesy of Gabor Bunkoczi)



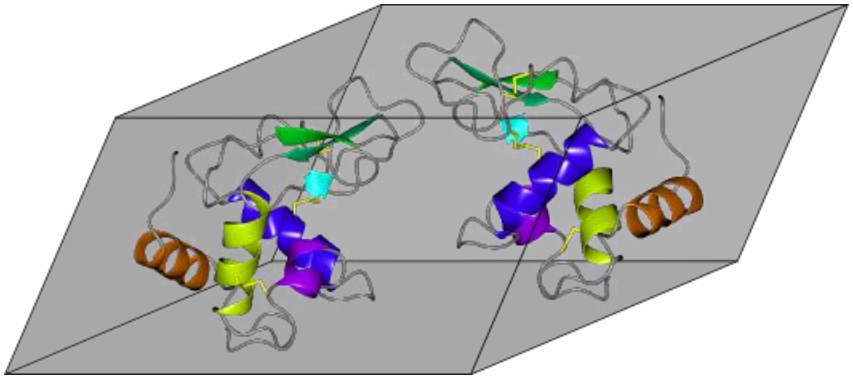


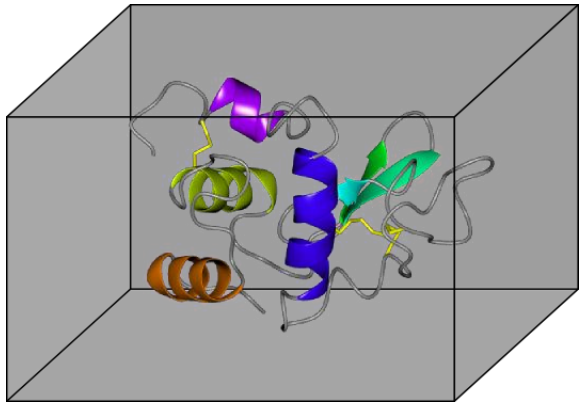
h	k	l	F	φ
0	0	1	12.6	123
0	0	2	2.1	12
0	0	3	69.9	287

h	k	l	F	φ
0	0	1	10.4	113
0	0	2	3.5	18
0	0	3	57.2	265

Molecular Replacement: The problem

(Slides courtesy of Gabor Bunkoczi)





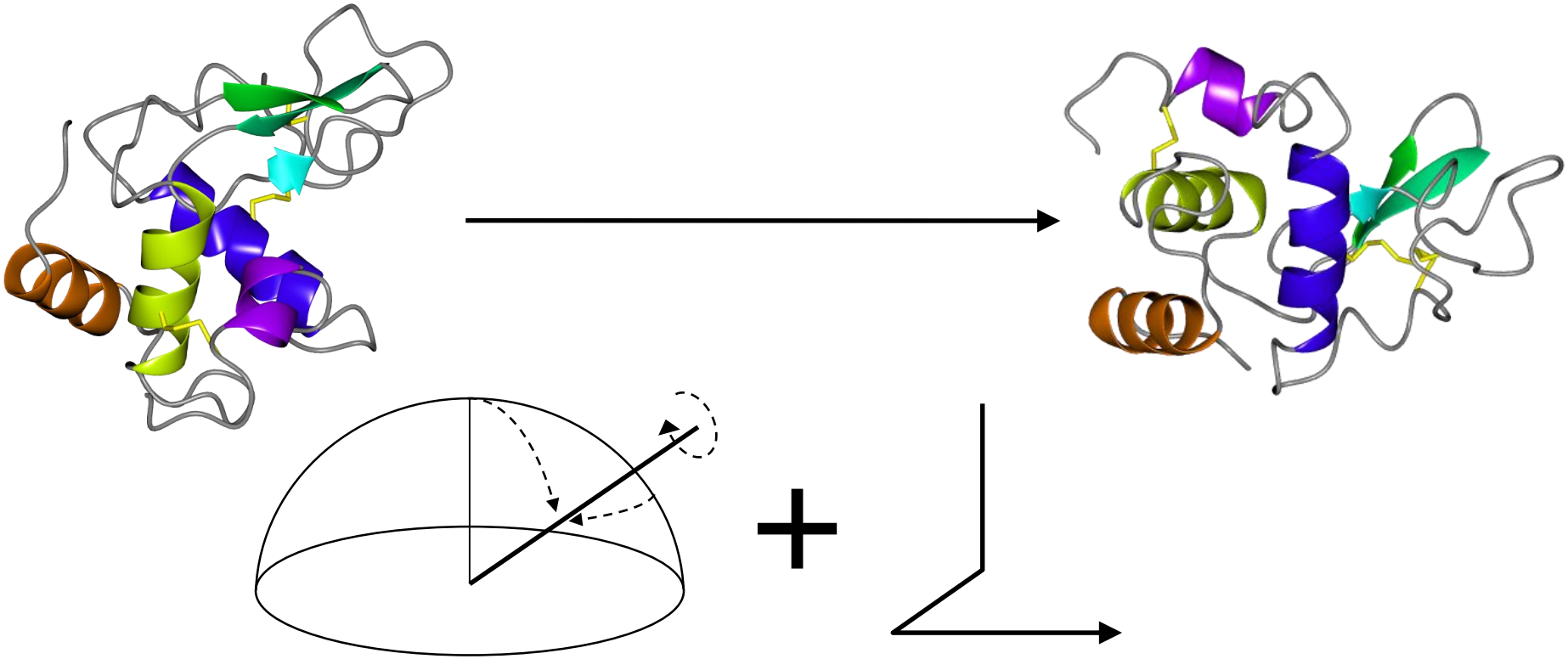
Previous crystal form

Current crystal form

$$\Re\left(\phi,\psi,\kappa,x,y,z\right)$$

Separability

(Slides courtesy of Gabor Bunkoczi)



(φ, ψ, κ) (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
  - Anistropy
  - Twinning
  - Pseudo symmetry

MR: Statistics

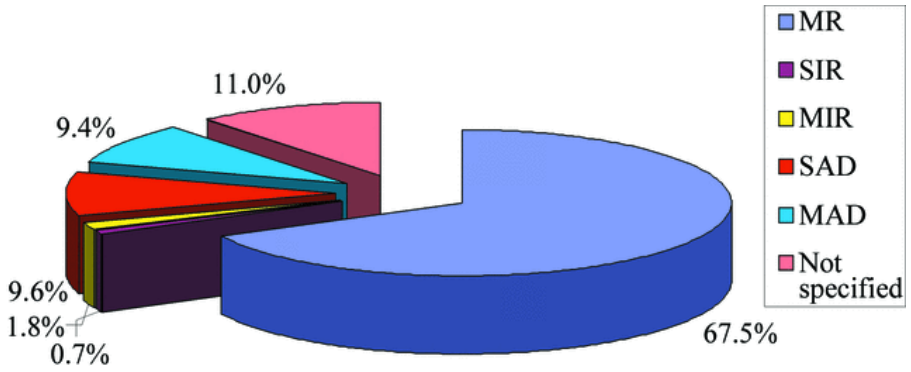


Diagram showing the percentage of structures in the PDB solved by different techniques

67.5% of structures are solved by Molecular Replacement (MR)

21% of structures are solved by experimental phasing

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

Main steps when performing MR

1. Identify suitable search model to give initial estimate of phases for unknown target structure
2. Preparation of search model for MR
3. Carry out MR using *Phaser*, *Molrep* or other program
4. Assess MR solution (e.g. MR scores, refinement scores, view density map)
   - Looks correct? – go on to model building and refinement
   - Incorrect? – return to step 1 or 2 and try to find a better or improve search model

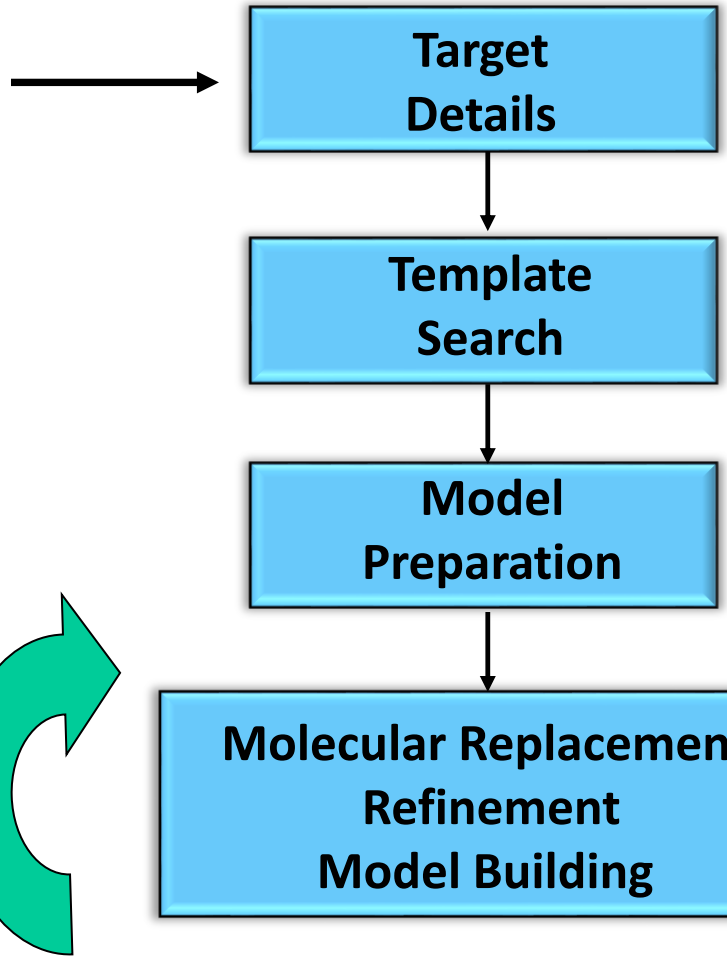
Automatic MR in CCP4

- Homologues available:
  - Find best search models and optimise their chances of success using various search model preparation procedures
  - ***MrBUMP, BALBES***
- No obvious homologous structures:
  - Generate artificial models using ab initio modelling or fragment libraries
  - ***AMPLE, ARCIMBOLDO***

MrBUMP

MrBUMP Pipeline

Target MTZ
&
Sequence



Check scores and exit or select the next model

Farm MR/Refinement jobs to cluster

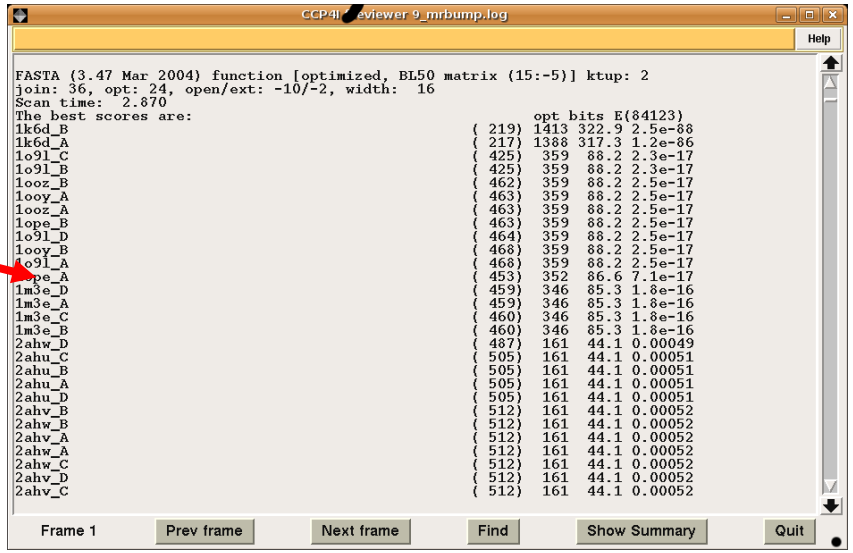
Search for model templates

Phmmer search of PDB

– Sequence based search using sequence of target structure

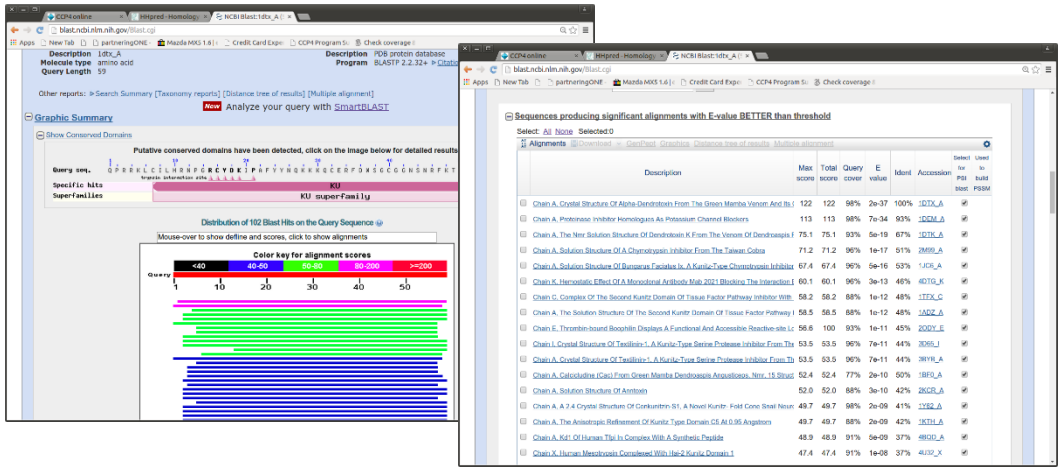
All of the resulting PDB id codes are added to a list

These structures are called model templates



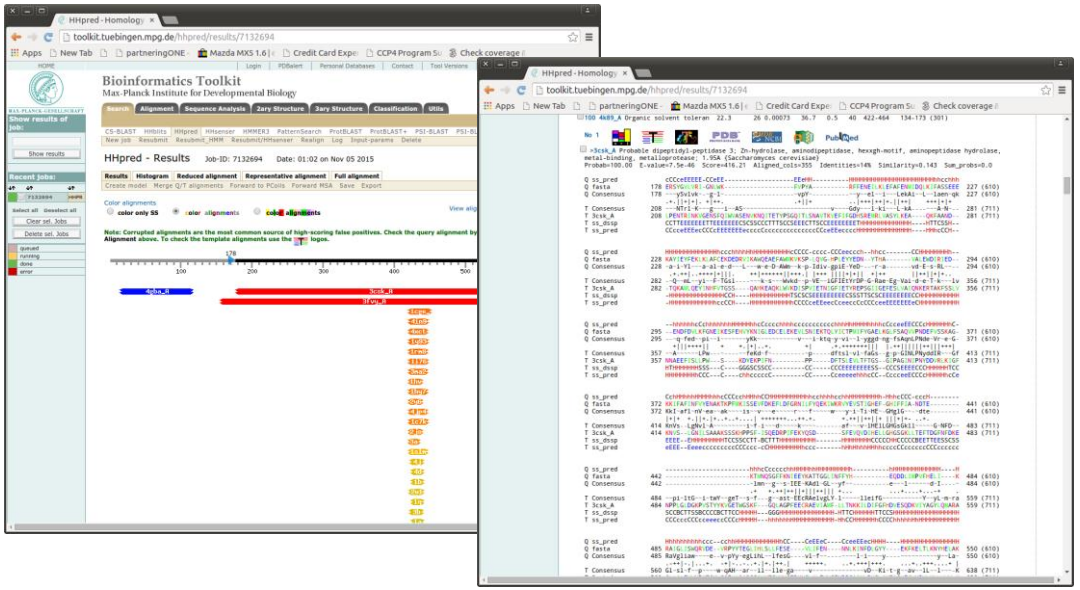
Other templates from:

- 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 (default for this online server)
- Can also add local PDB files



- Psi-Blast
  - Profile fitting method for sequence-based search
  - Quick

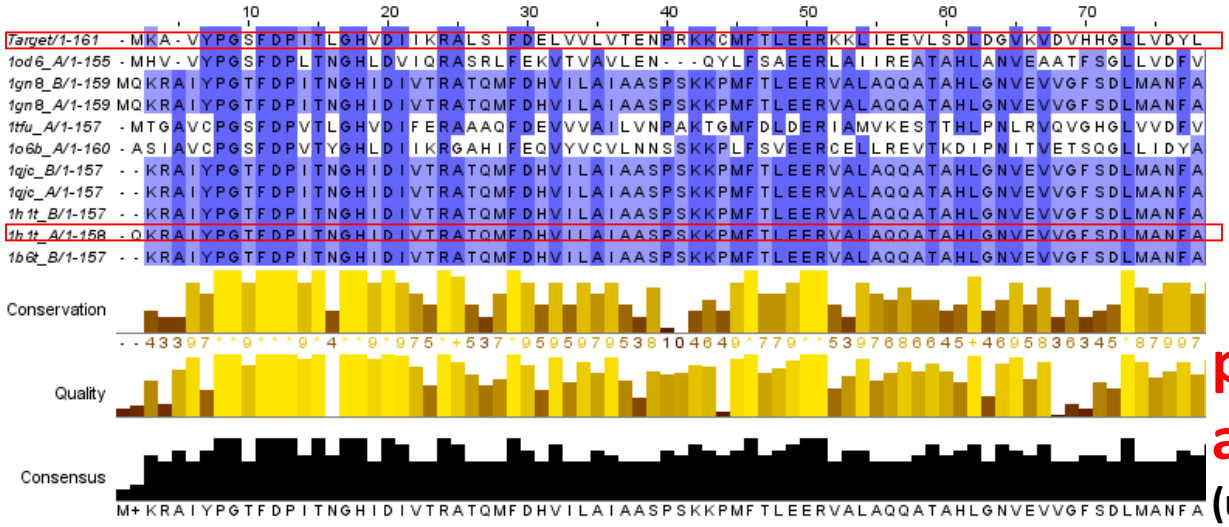
- HHPred
  - HMM method search across sequence alignment databases rather than single sequences
  - More sensitive and better for finding distant homologues



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 Phmmer, MAFFT, for multiple alignment

Model template scoring: score = sequence identity X alignment quality

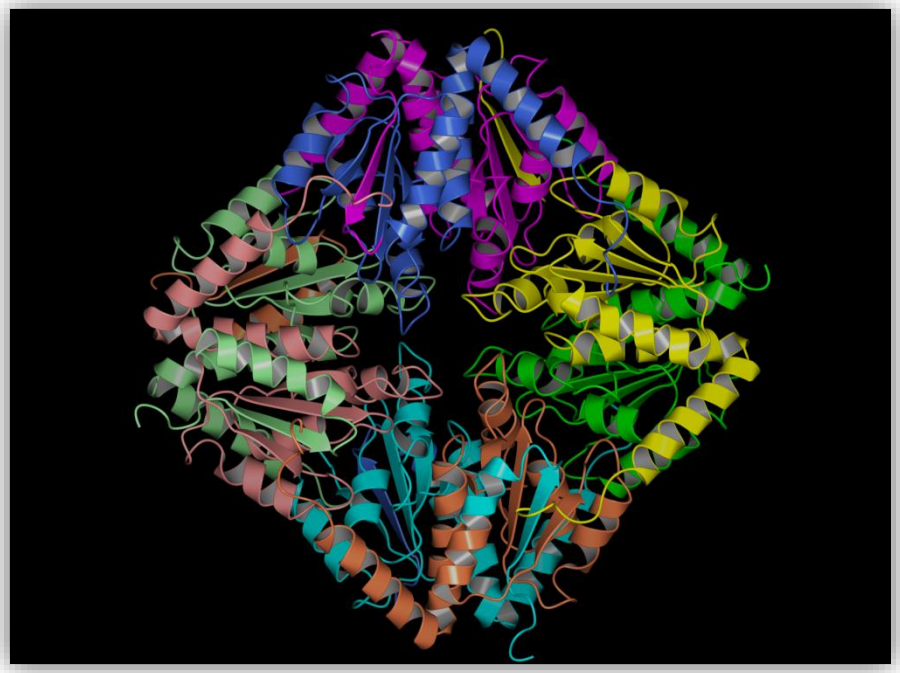
Where *HHPred* is used scoring and alignment come from *HHPred* search results

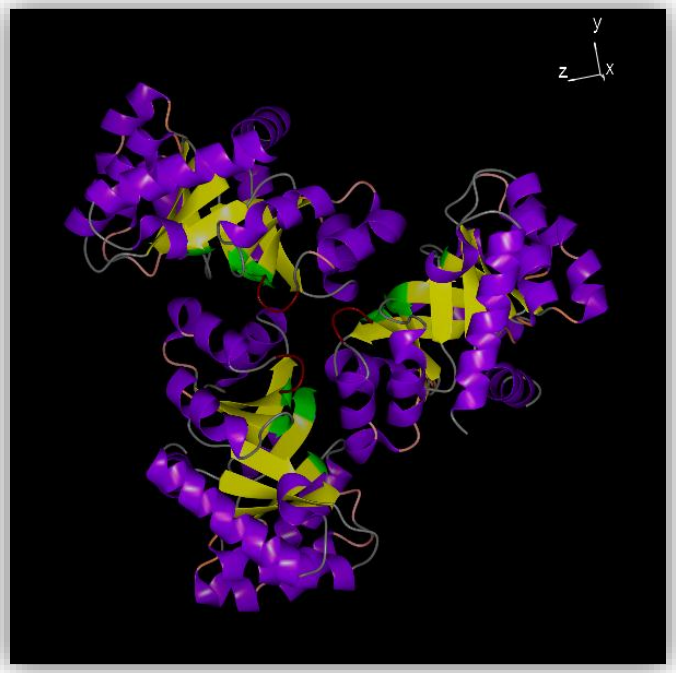
Domains

- 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
- 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:

more side chain truncation (*uses alignments from multiple alignment step or from HHPred*)

Unmod
– Input model template is not modified

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.

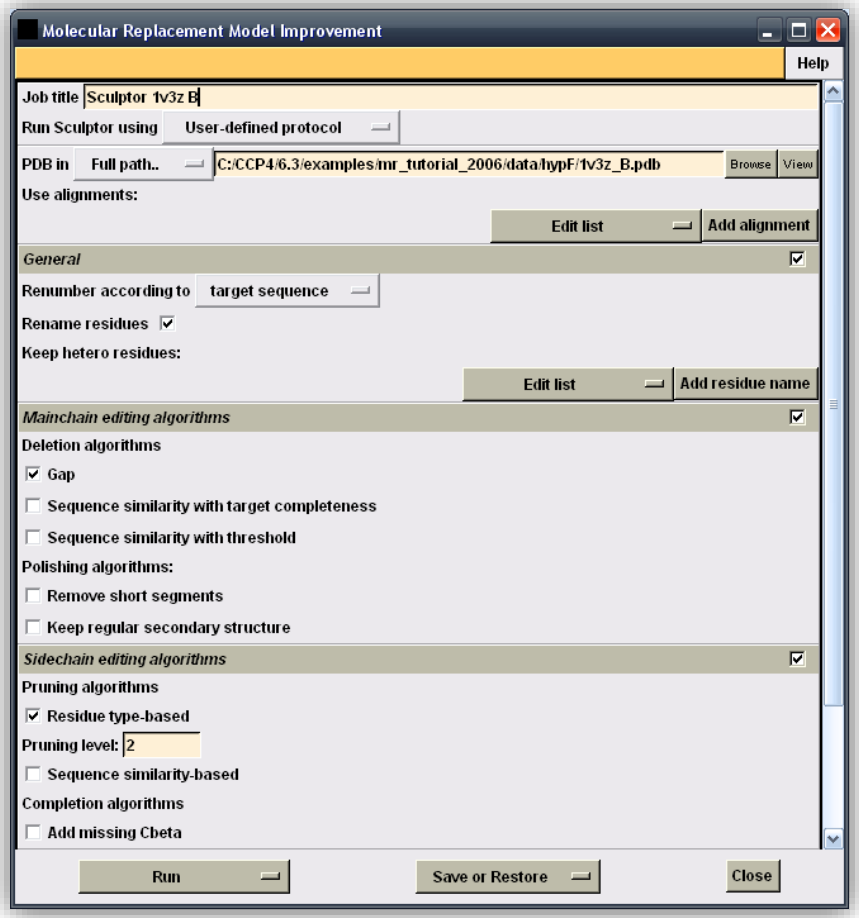
Sculptor
– Similar to chainsaw but different protocols for side chain truncation and alignment

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

deal with deletions

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

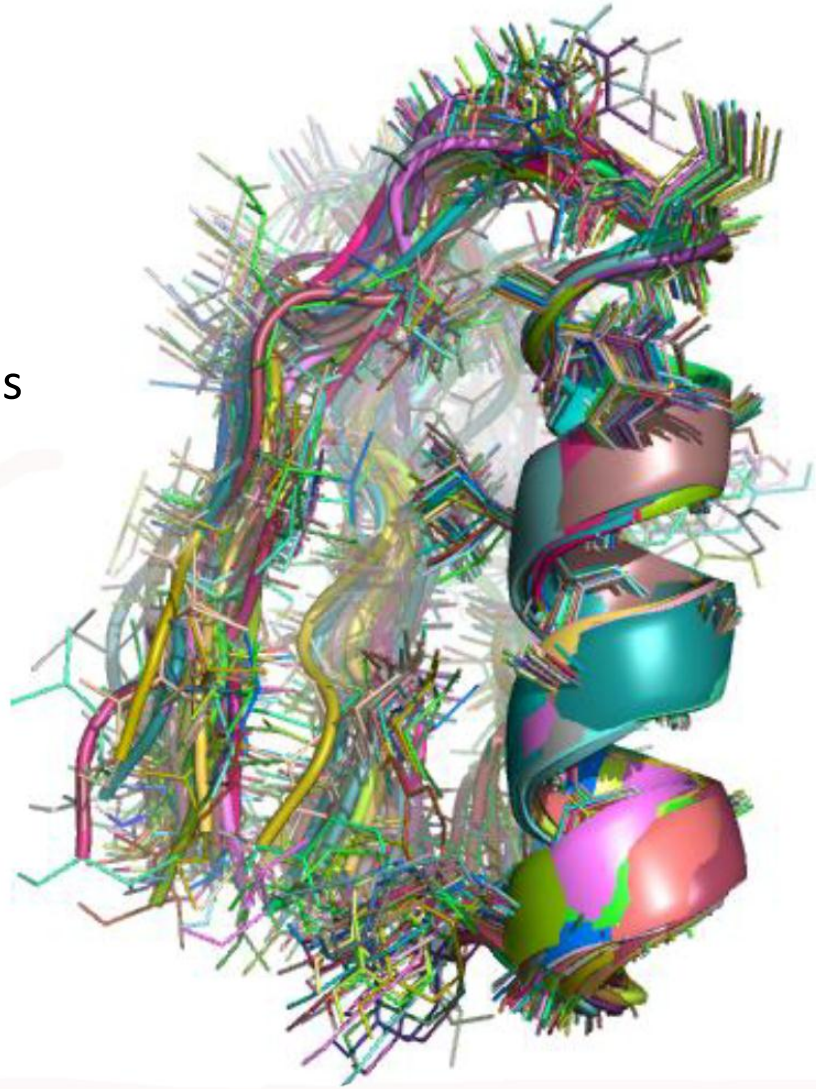
- Create ensembles of top search models, for use in additional runs of Phaser and Molrep

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

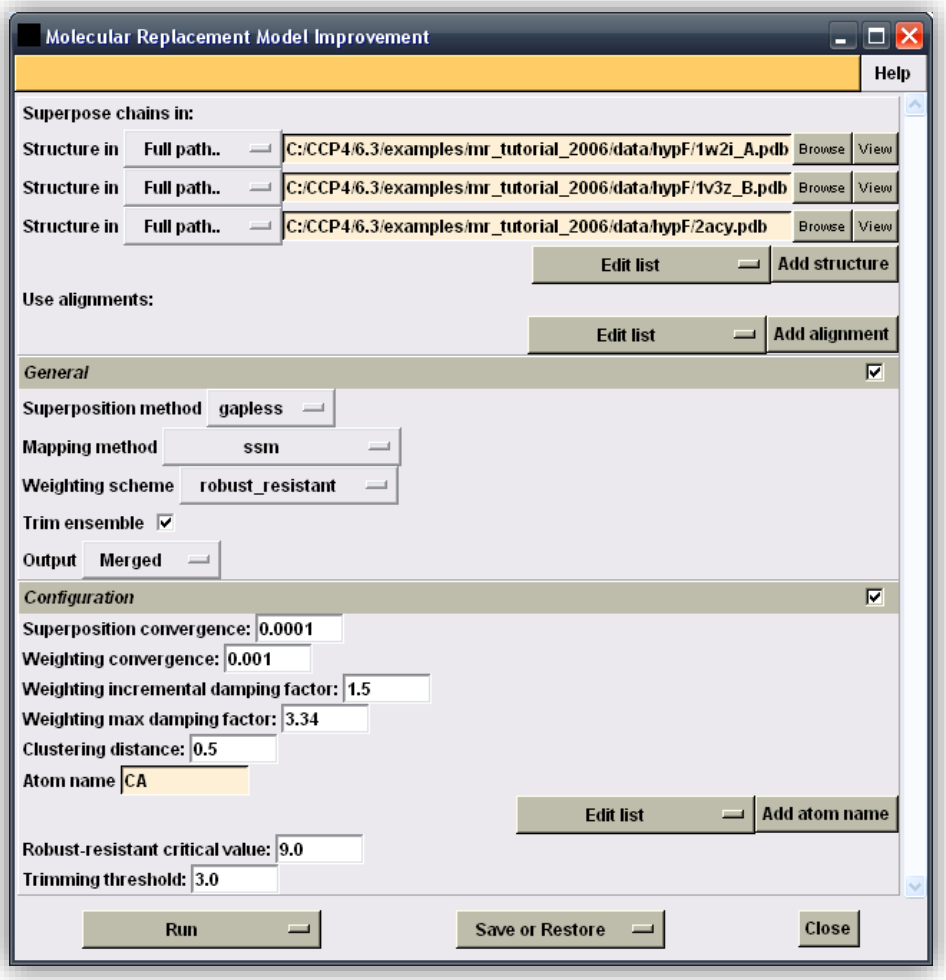
- Can give solutions in cases where individual models fail

- Models must be sufficiently similar (MW and rmsd)

- Use Gesamt or Superpose to align structures



Phaser.Ensembler

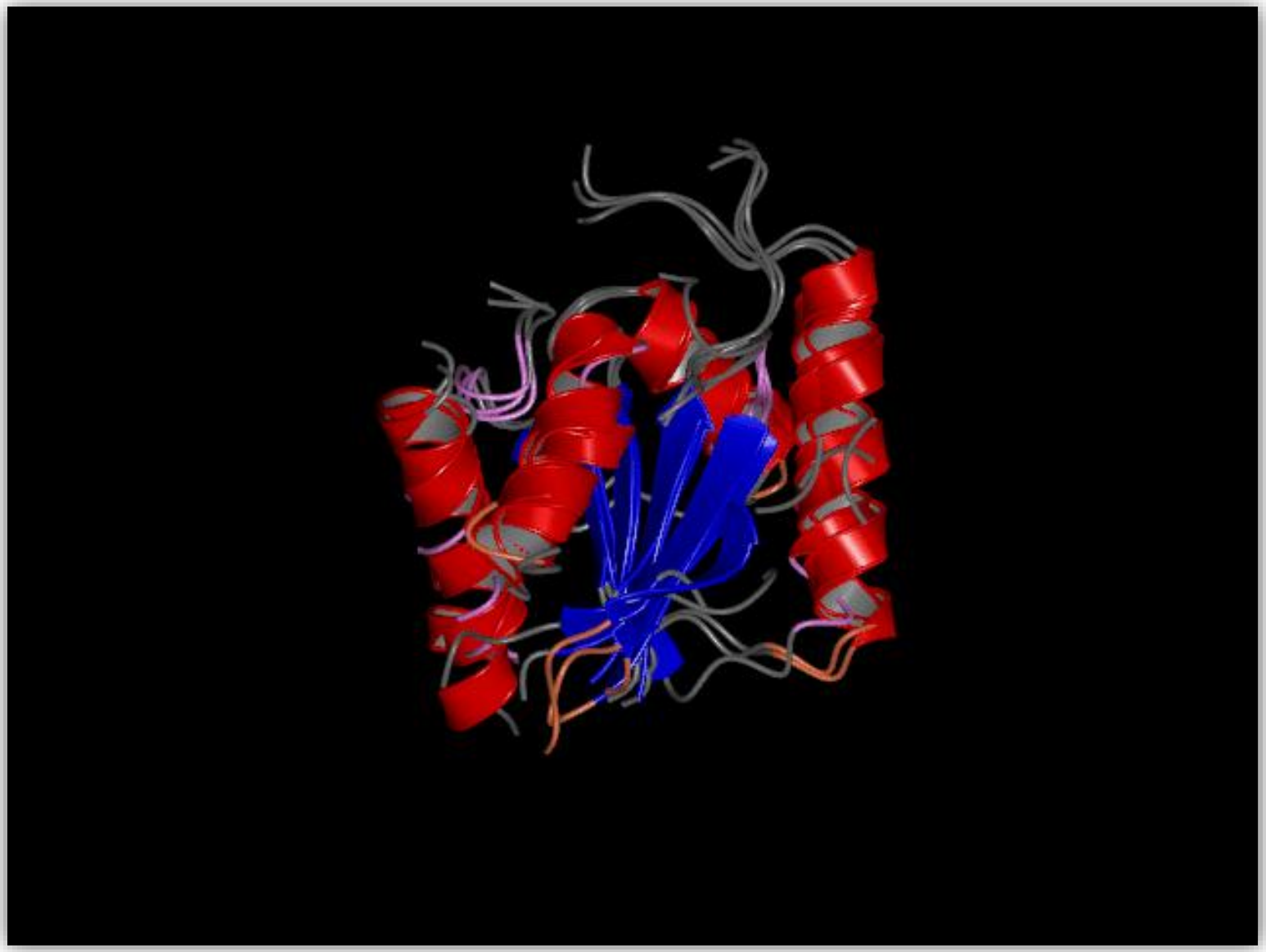


- Generate ensembles from set of similar homologues
- Several protocols:
  - Gapped or non-gapped
  - Different alignment methods including manual input of alignment
  - Trimming based on RMSD threshold

Cluster/Truncate in MrBUMP

- Plan to implement AMPLE method in MrBUMP in near future
- Homologues structurally aligned using *Gesamt* to form initial ensemble
- Test cases: 1E59, 1EBB, 2QNI, 3DCY, 1DKM, 1QWO
- 1UJB, 2A6P and 3C7T as distant homologues in ensemble

	Structures in superposition					
PDB	1UJB/2A6P/3C7T	1UJB/2A6P	2A6P/3C7T	1UJB/3C7T	**MrBUMP**	**MrBUMP** no 1UJB
1E59	0/57	-	-	-	0/10	-
1EBB	11/57	0/57	1/63	14/63	3/10	0/7
2QNI	34/57	14/57	19/63	27/63	0/10	-
3DCY	45/57	41/57	40/63	45/63	2/10	0/7
1DKM	0/57	-	-	-	0/10	-
1QWO	0/57	-	-	-	0/10	-



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
  $\Rightarrow$ 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
- **LLG** scores given in table
  - > 120 – clear solution
  - > 60 < 120 – most likely a solution
  - < 60 – worthy of investigation

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"

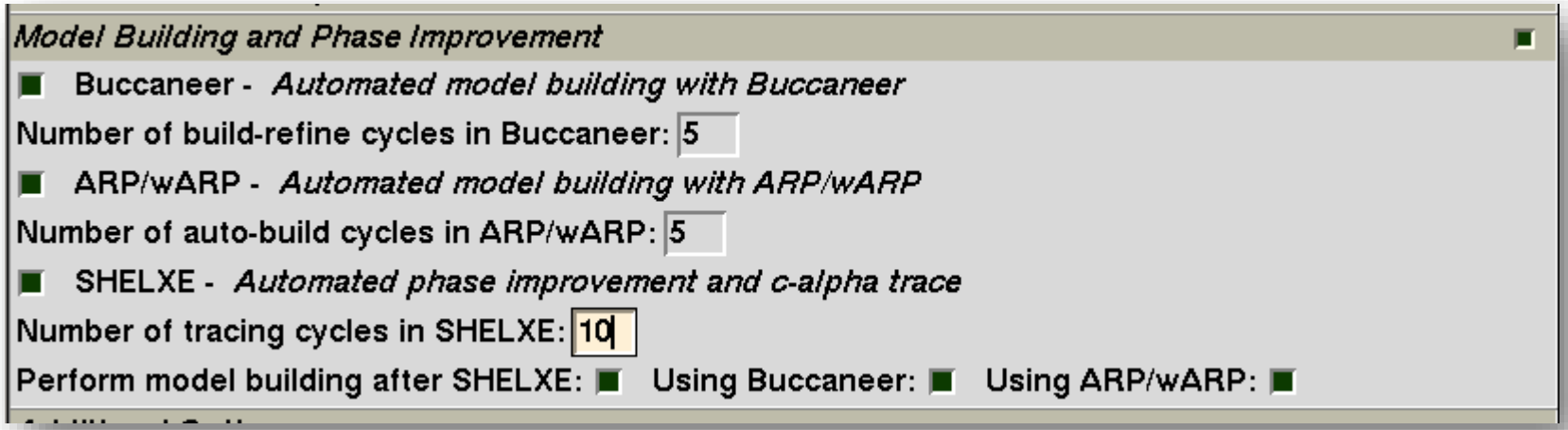
final Rfree < 0.48 or
final Rfree < 0.52 and dropped by 5% ⇒ "marginal"

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

- More stringent method of testing MR solution: 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.

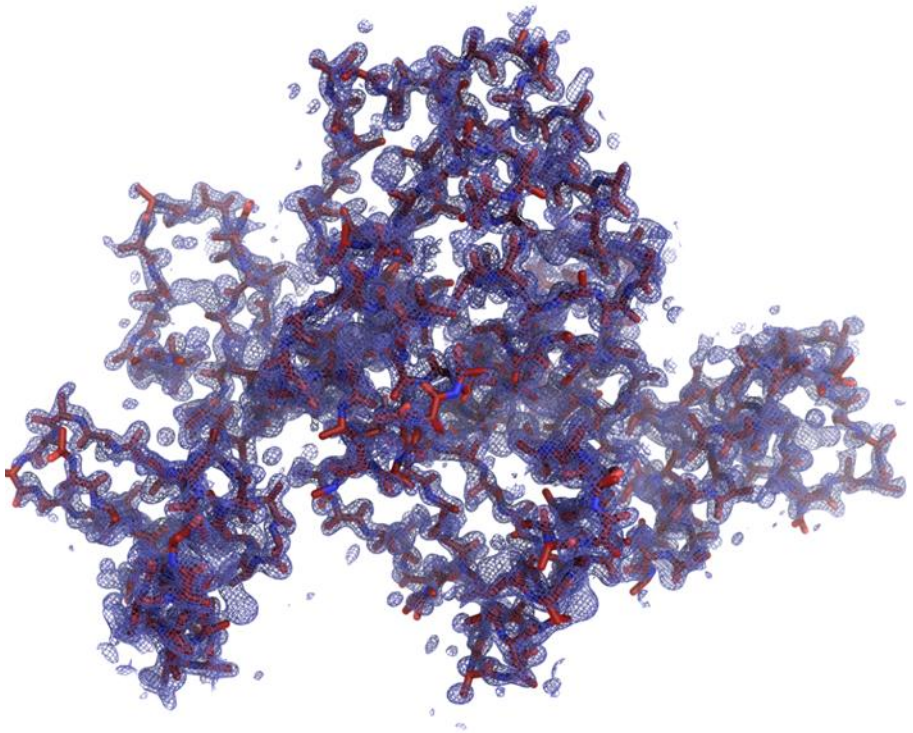
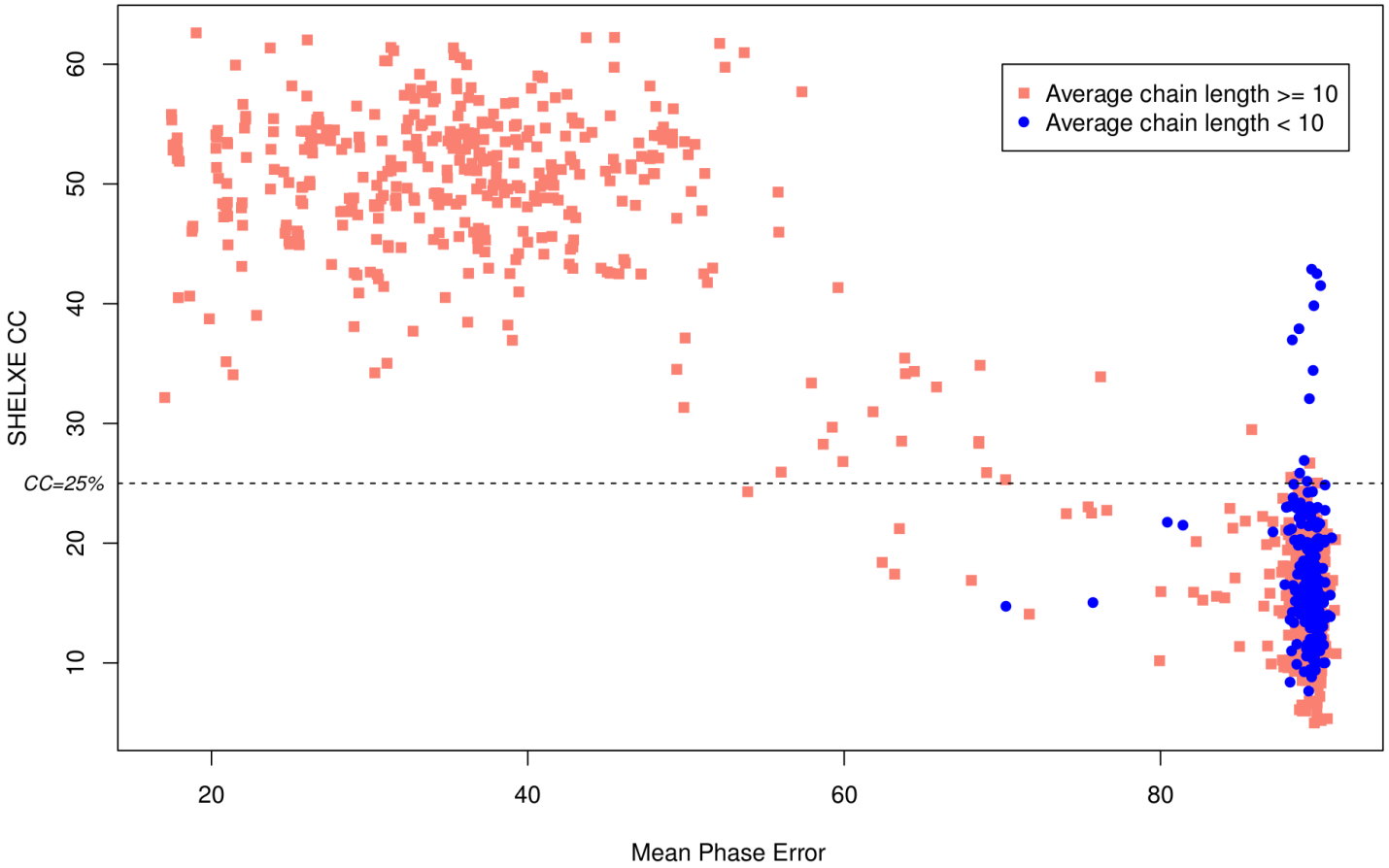


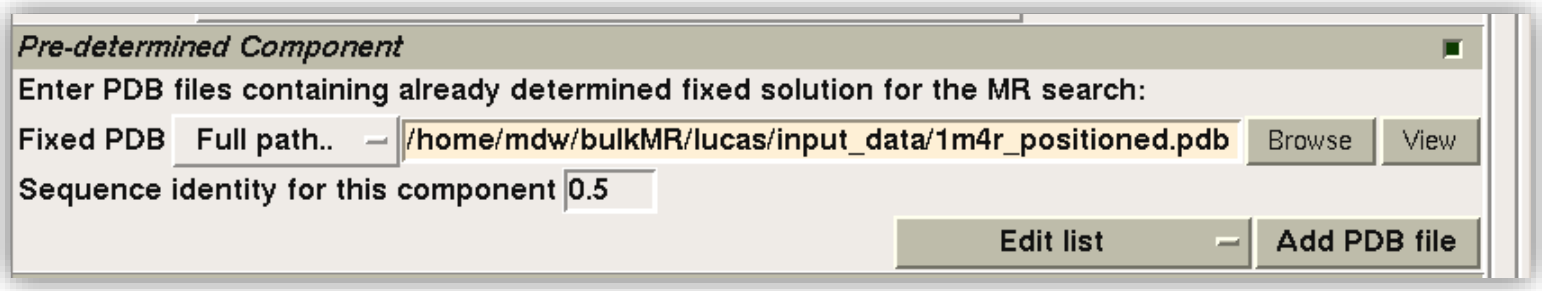
Figure courtesy of Andrea Thorn



- Criteria for success:
  - CC of better than 25%
  - Average fragment chain length of 10 or more
- Result for > 1000 AMPLE jobs

Inclusion of fixed models

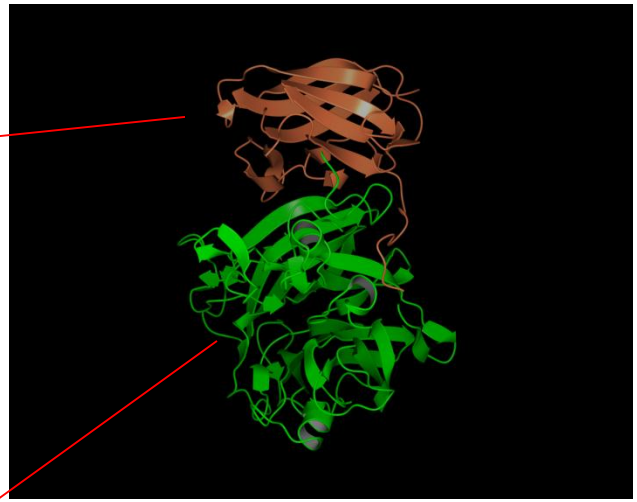
- MrBUMP will accept one or more positioned models.
- These are included as fixed models in all MR jobs.



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

Example (thanks to Elien Vandermarliere)

Target is an arabinofuranosidase
Data to 1.55Å in P212121



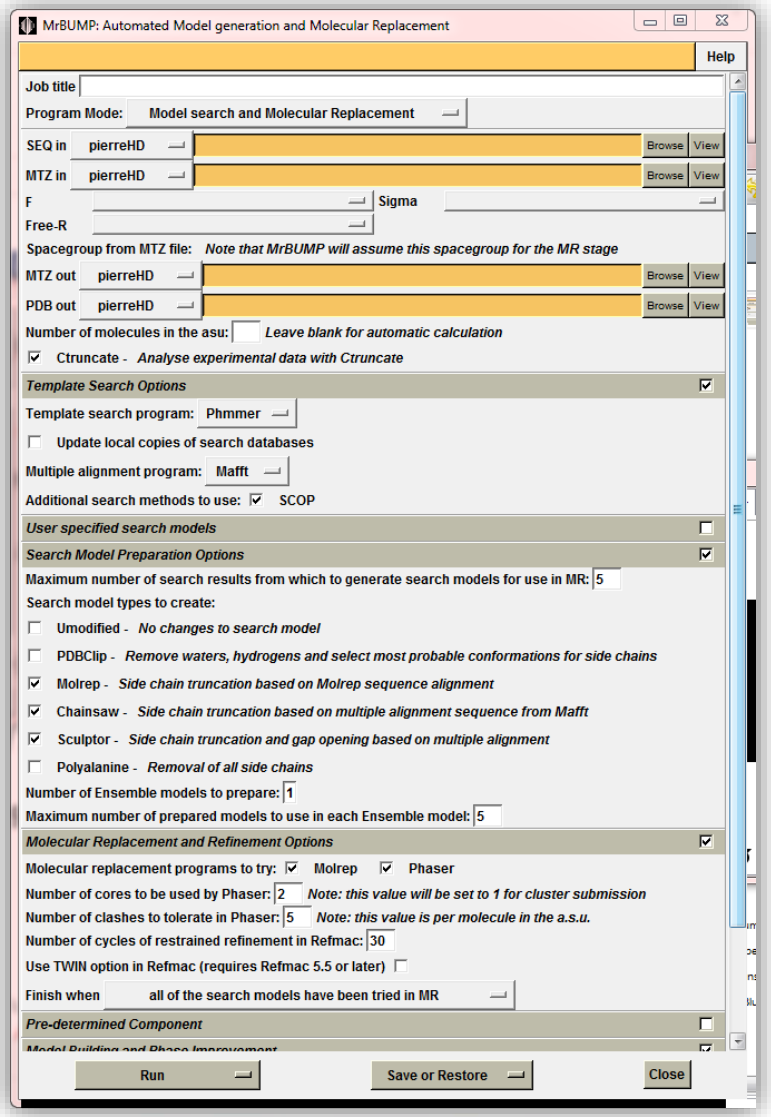
Small C domain (144 res) solved with 34% seq ident model
(1w9t_B_MOLREP best out of 4 solutions)

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

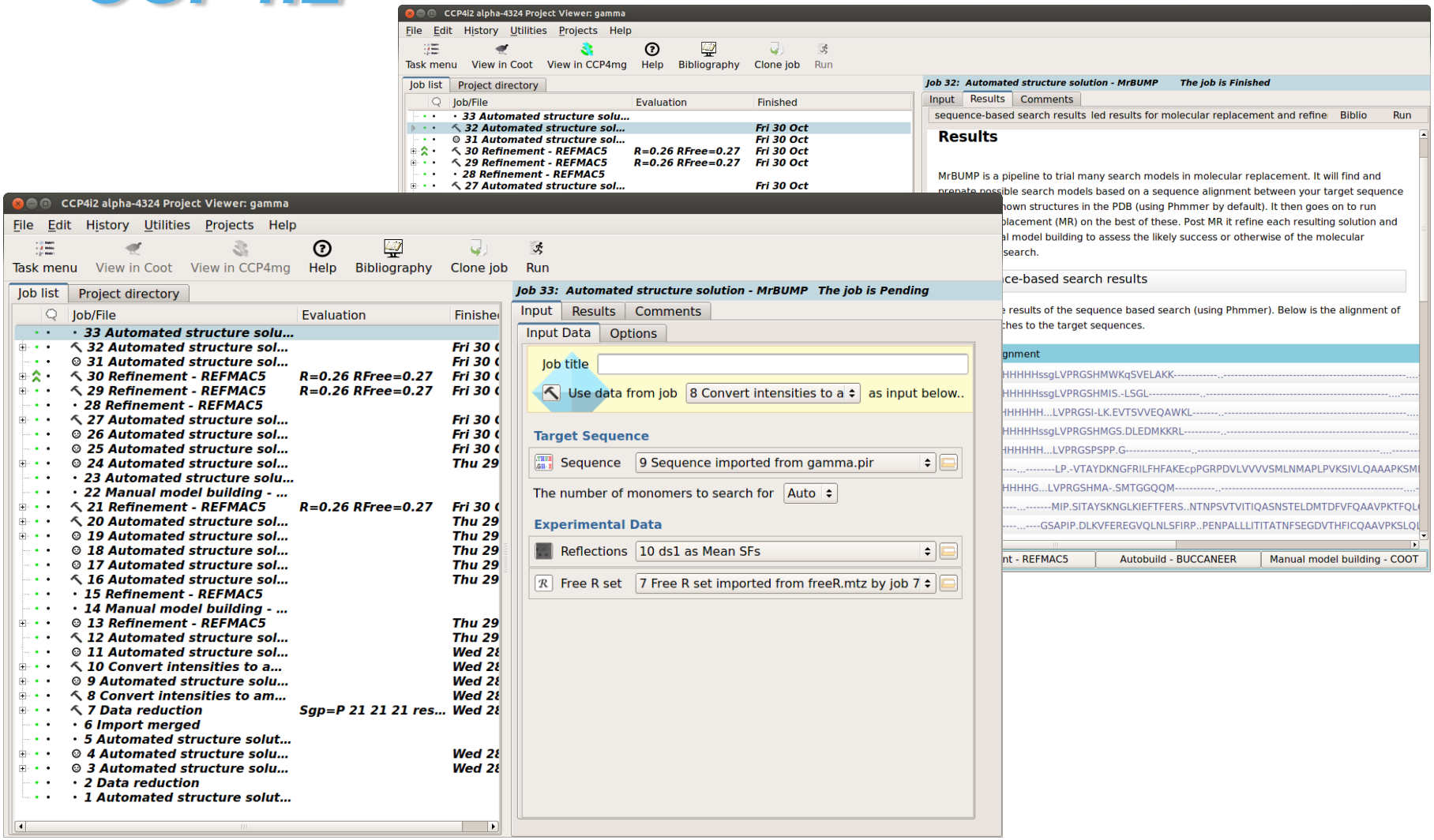
ARP/wARP then builds 457/493 residues to R/Rfree 0.185/0.225

Using MrBUMP

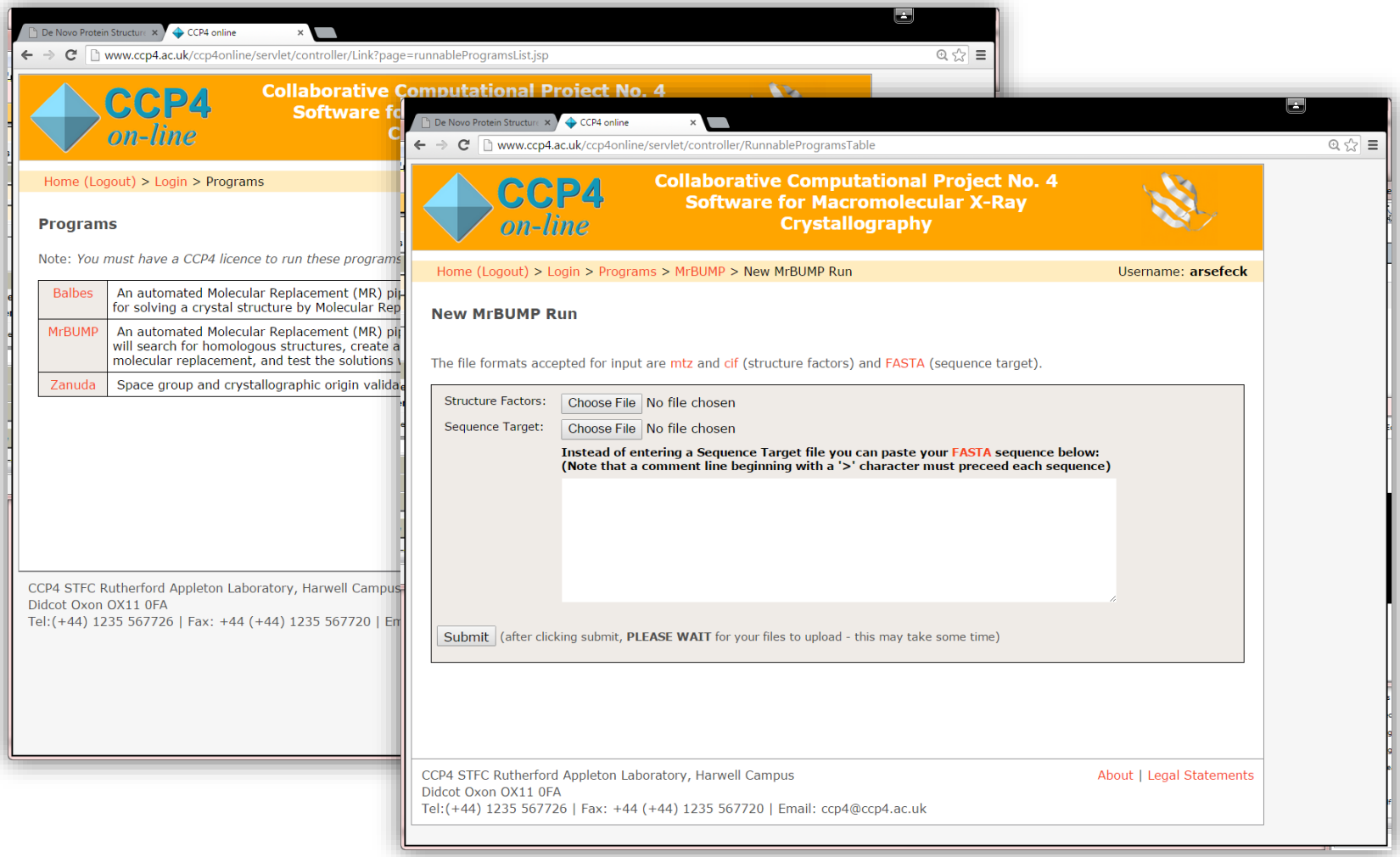
- CCP4i GUI
- Searching for a model through to model building
- All steps can be adjusted
- Different options for model search, alignment, molecular replacement and model building
- Sensible default options
- Can take several hours to run



CCP4i2

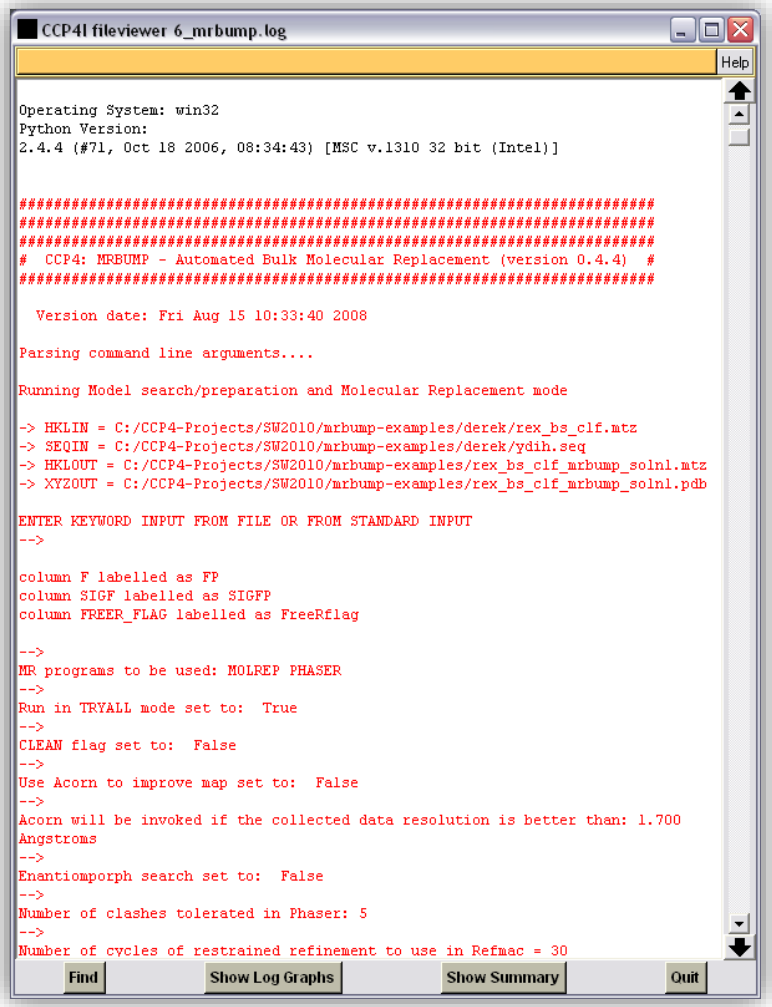

CCP4i2 alpha-4324 Project Viewer: gamma
File Edit History Utilities Projects Help
Task menu
View in Coot
View in CCP4mg
Help
Bibliography
Clone job
Run
Job list
Project directory
Job/File
Evaluation
Finished
Job 33: Automated structure solution - MrBUMP The job is Pending
Input
Results
Comments
Input Data
Options
Job title
Use data from job
8 Convert intensities to a
as input below..
Target Sequence
Sequence
9 Sequence imported from gamma.pir
The number of monomers to search for
Auto
Experimental Data
Reflections
10 ds1 as Mean SFs
Free R set
7 Free R set imported from freeR.mtz by job 7
Job 32: Automated structure solution - MrBUMP The job is Finished
Results
MrBUMP is a pipeline to trial many search models in molecular replacement.
Autobuild - BUCCANEER
Manual model building - COOT

MrBUMP Online

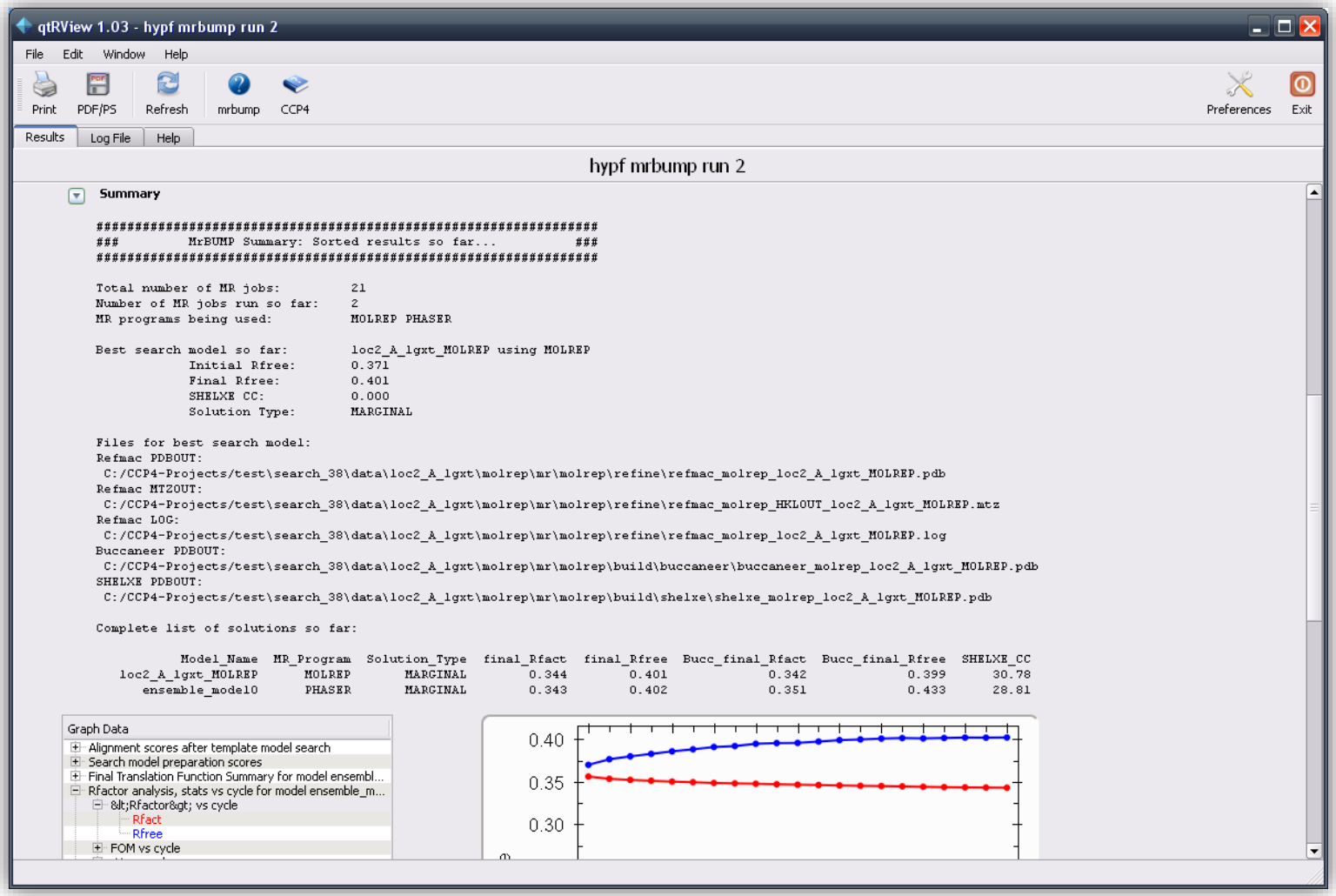


MrBUMP output…

- 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



MrBUMP output - QTRview



BALBES

Balbes – Murshudov, Long, Vagin

- Alternative approach to the problem
- Customised database of models designed specifically for use in molecular replacement
- Redundant protein chains have been removed and domain and multimer definitions are pre-determined for all chains in the database
- Focus is on deriving set of well-crafted search models for the use in MR

βαλβεσ

The Balbes Database

- Length > 15 amino-acid residues
- Resolution better than 3.5 Angstroms
- Basic entry is a macromolecular subunit = Domains
- Redundancy has been removed:
  - If 2 entries have sequence ID > 80 % and C-alpha RMSD < 1 Å the one refined against higher resolution is retained
- Multimeric forms are determined using PISA and construction details stored
- 50000 domain entries (*derived from 80000 X-ray structures in PDB*)

Target Analysis

X-ray Data Analysis

- What resolution range to use for molecular replacement
- Data completeness and other properties
- Twinning – information presented to user
- Pseudo translation – can be accounted for in MR step

Sequence Analysis

- Finds template structures with their domain and multimer definitions
- Estimates number of molecules in the asymmetric unit
- “Corrects” template molecules using sequence alignment

Search Protocols

- Ensemble models are constructed and favoured
- Multimers:
  - Start by searching with the largest unit possible – use Oligomer if available, then whole chains, then domains
- Complexes:
  - Step-by-step approach
    1. Search for largest domain is done using conventional MR
    2. Refinement of Solution
    3. Search for next largest domain is done in difference map of previous refinement
    4. Refinement of new model
    5. Repeat steps 3 and 4 for further domains if required
- Space Groups:
  - Search all related space groups (performed in parallel on web server)

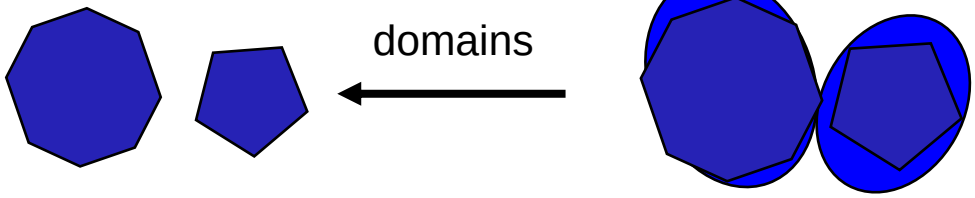
Models generated from a single sequence

Domain models

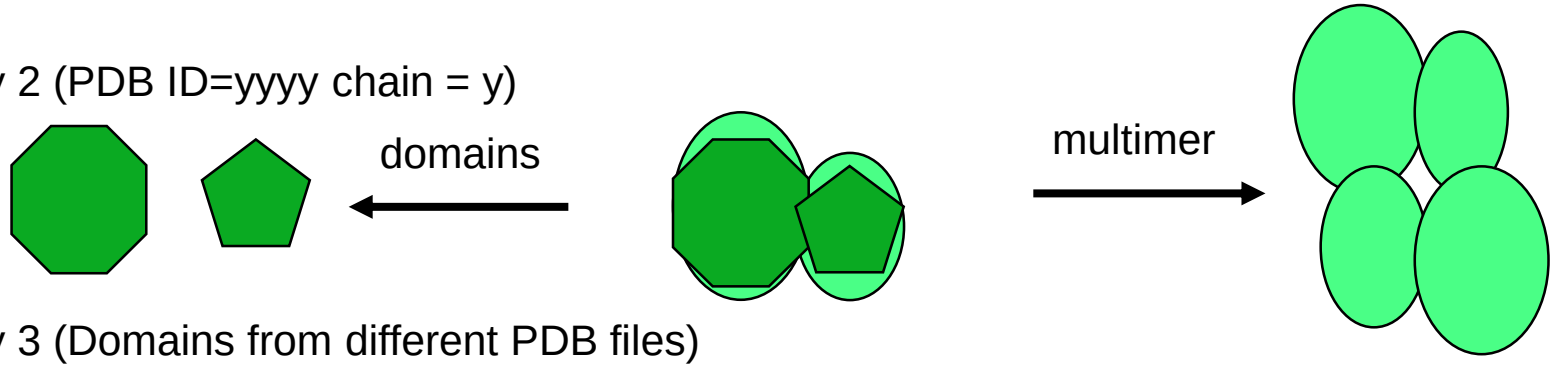
Full Chain models

Oligomeric models

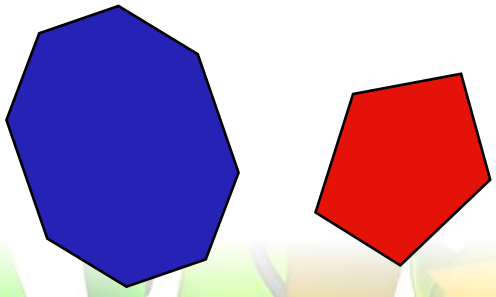
Family 1 (PDB ID=xxxx chain = x)



Family 2 (PDB ID=yyyy chain = y)



Family 3 (Domains from different PDB files)

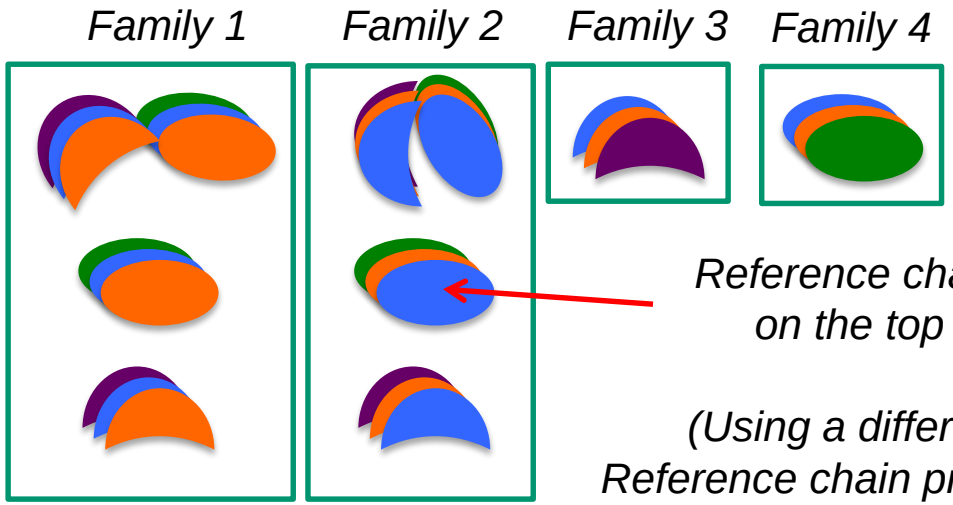


Multi-domain model

Ensemble Models

Homologues from the Balbes database:

Ensemble search models:



Reference chain on the top

(Using a different Reference chain produces a different search model)

- Example:

Refinement

Restrained refinement including "jelly-body" restraints

In refinement stage "jelly-body" refinement is used. It seems to increase success rate, especially for multidomain cases

R-values

Initial and final Rfactor and Rfree are presented to the user in the log file and final decisions are made based on a quality factor derived from these.

ARP/wARP

Optionally, users can select that the results from BALBES are forwarded to the ARP/wARP server for model building

Balbes Online

What to do if automation doesn't work?

- Step-by-step search if dealing with a complex or there are multiple copies in the asymmetric unit
  - Note: if translational NCS is present search for 2 or more copies in a step
- Try all space groups from Laue class e.g. P222, P212121, P22121, etc..
  - *Phaser* and *Molrep* support this
- Check for domains and possible hinge-motion between domains in search models
  - Break model into domains and search separately
- If multiple homologues exist with similar folds create an ensemble model
- Data problems
  - Have you processed the data correctly?
  - Try to assess whether or not you have twinning, if so you may have lower symmetry
  - Are there signs of?

Acknowledgements

- Andrey Lebedev, CCP4
- Martyn Winn, Daresbury Laboratory
- Gabor Bunkoczi, Phaser group
- BALBES: Fei Long, Garib Murshudov
- Jens Thomas, Jaclyn Bibby, Daniel Rigden, University of Liverpool
- Andrea Thorn, Tim Gruene & George Sheldrick (SHELX)
- Refmac: Garib Murshudov, LMB-MRC Cambridge
- Molrep: Alexey Vagin
- Phaser: Randy Read, Airlie McCoy
- Gesamt: Eugene Krissinel
- Thanks to authors of all underlying programs
- Funding: BBSRC, STFC
- Support from CCP4 & the Research Complex at Harwell

Acorn: Phase improvement

If resolution better than 1.7Å use Acorn procedure:

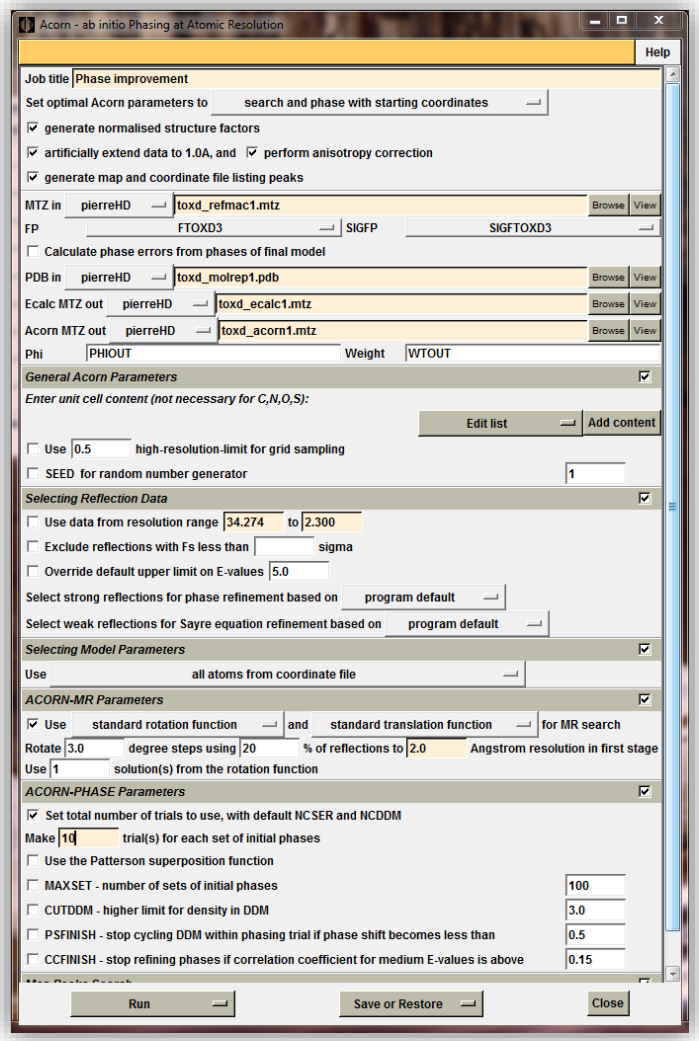
- initial phase set from refined MR solution
- artificial phase extension to 1.0Å
- dynamic density modification

Result:

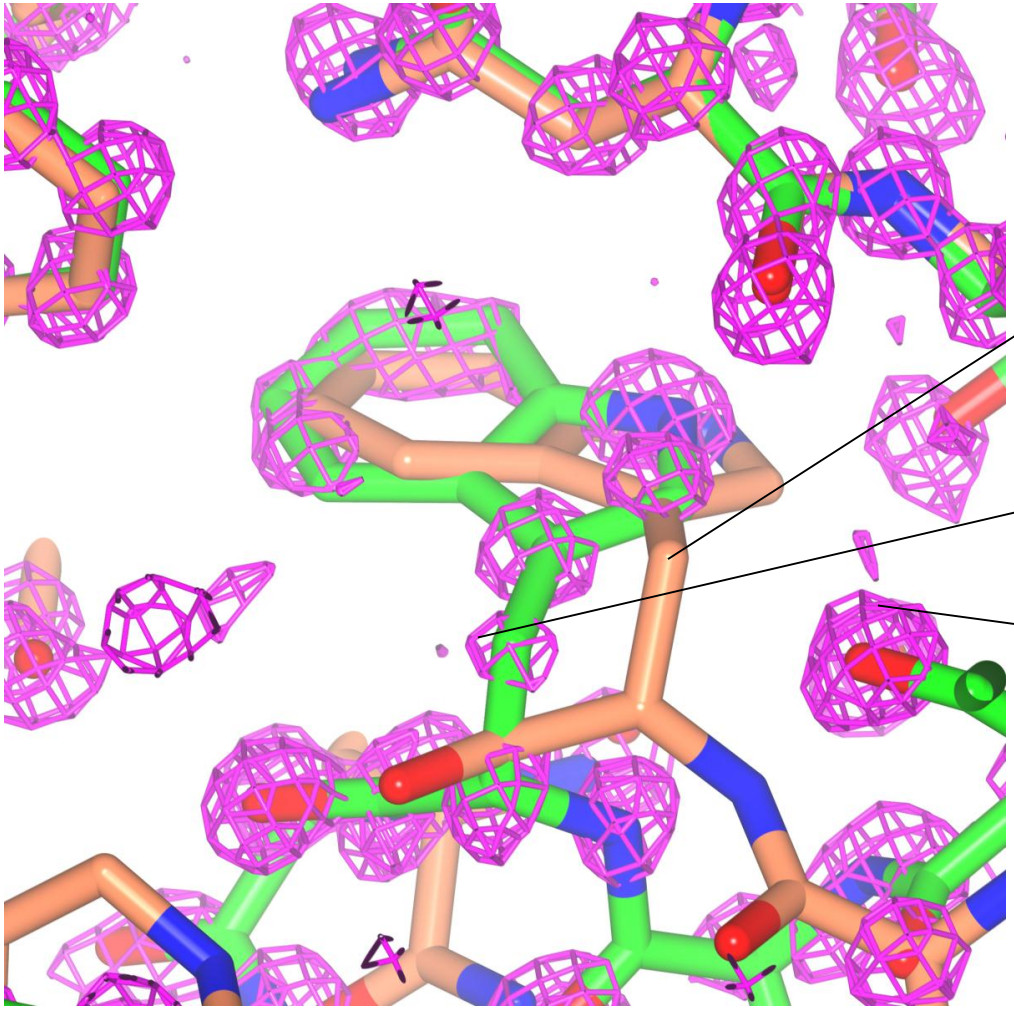
- CC for medium Es good indicator of solution
- Use E-maps for re-building

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