

MR Pipelines in CCP4

*Garib Murshudov, Fei Long, Alexei Vagine,
Jaclyn Bibby, Olga Mayans
Daniel Rigden*

*Ronan Keegan, Martyn
Winn, Graeme Winter
Collaborative*



Background

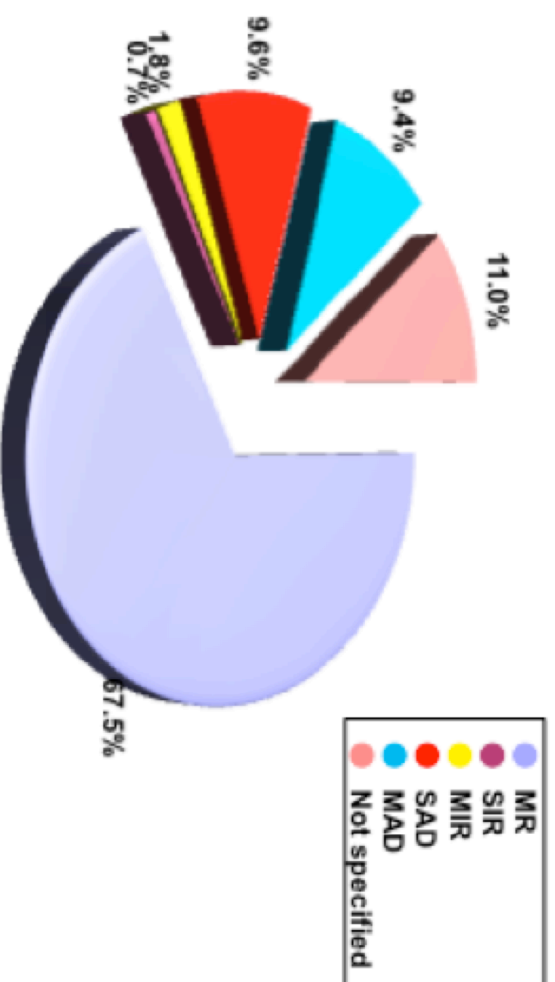


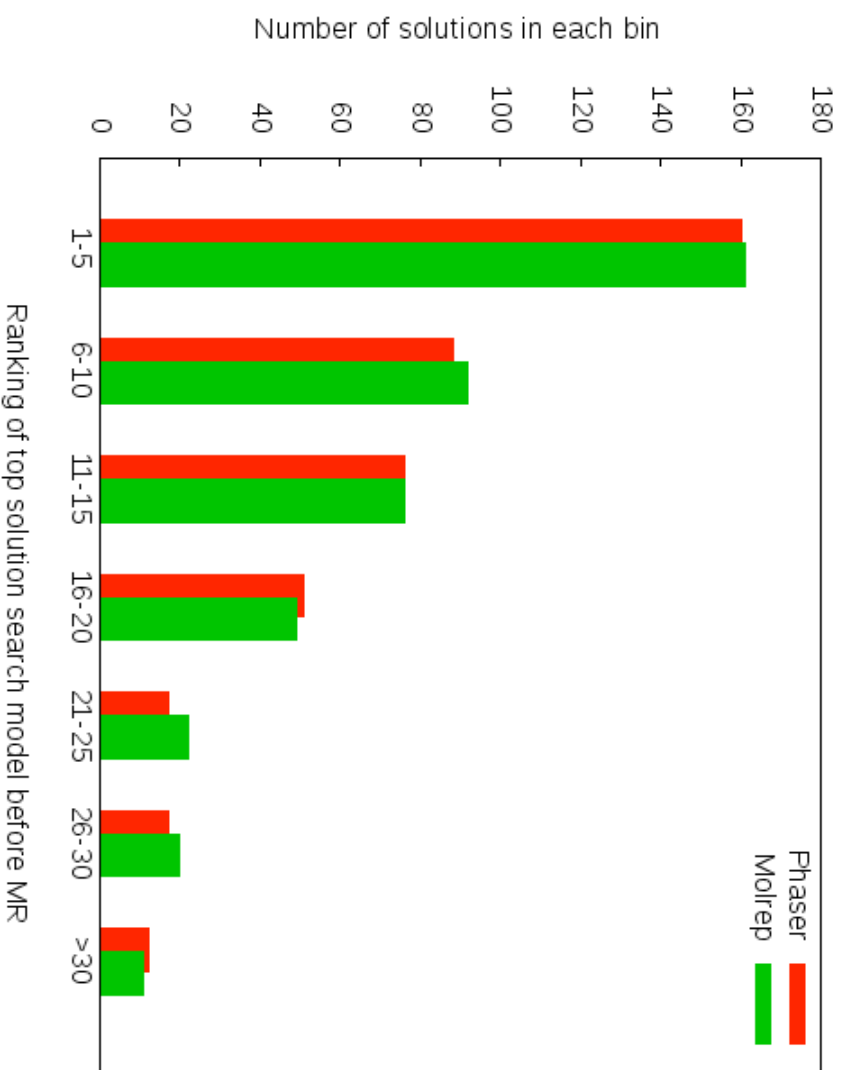
Diagram showing the percentage of structures in the PDB solved by different techniques

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

21% of structures are solved by experimental phasing



Let the computer do the work for you



Sequence homology
ranking of
best solution
(2009 pdb entries
vs. pre-2009 data)



3 pipelines, 3 routes to solution

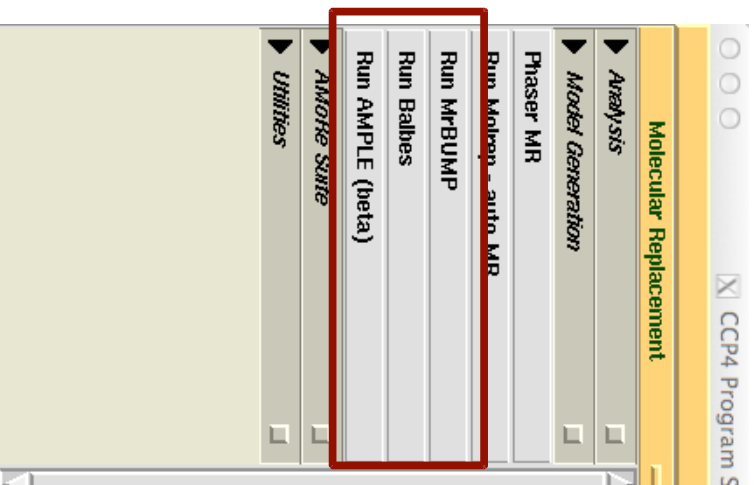
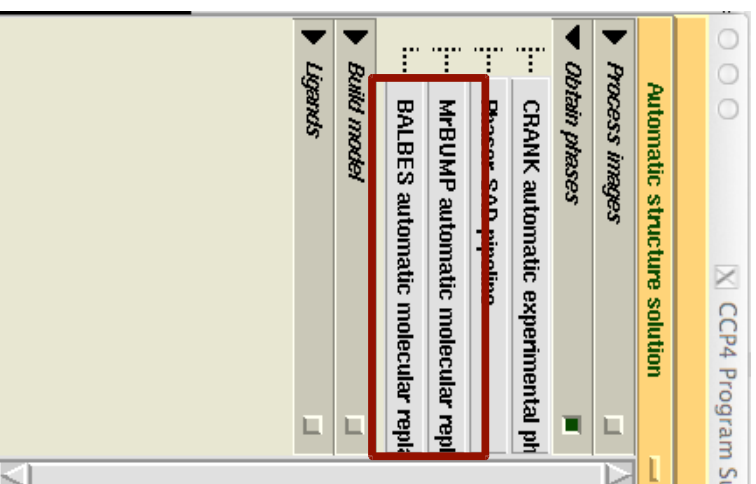
1. MrBUMP
 - Tries to mimic “best” practice of using bioinformatics tools*
1. BALBES
 - Custom database optimised for MR
1. AMPLE
 - Ab initio model generation**

*external packages clustalw, mafft, fasta

**external packages ROSETTA, QUARK, SPICKER



Using the pipelines: ccp4i



www.ysbl.york.ac.uk/balbes

THE UNIVERSITY *of* York

York Structural Biology Laboratory

University | Chemistry | YSBL

Home

CCP4

THE UNIVERSITY *of* York

York Structural Biology Laboratory

University | Chemistry | YSBL

Home (Logout) > Login > Programs > Balbes > New Balbes Run

Username: garibm

Welcome

Any problems? - !

Runnable Programs

Login to run Balbes, Balbes

Other Options - Register

Downloads

Click on the links below

Balbes	an automata
Molrep	an automata
Refmac	a macromolecular
Jligand	a Java interface
Sfcheck	assessment
CCP4mg	an easy way
Coot	a program

Dictionary

Download the Refmac

New Balbes Run

The file formats accepted for input are **mtz** and **cif** (structure factors) and **FASTA** (sequence target). **Note:** checking the ARP/wARP checkbox will send Balbes's results to the **ARP/wARP** server (it is assumed that you agree to the ARP/wARP academic license conditions)

Structure Factors:

Sequence Target:

Instead of entering a Sequence Target file you can paste your **FASTA** sequence below:
(Note that a comment line beginning with a ';' character must precede each sequence)

Browse...

Browse...

Check Full Spacegroup:

Run ARP/wARP (on the Balbes solution):

Dissemination Level:

World

Submit

 (after clicking submit, **PLEASE WAIT** for your files to upload - this may take some time)

A decorative graphic at the bottom of the page featuring several colorful protein ribbon structures in shades of green, yellow, orange, red, and purple, arranged in a dynamic, overlapping fashion.

MrBUMP – Keegan, Winn

MrBUMP: Automated Model generation and Molecular Replacement

Job title

Program Mode: **Model search and Molecular Replacement**

SEQ in PROJ **Model search only** Browse View

MTZ in PROJ **Molecular Replacement only** Browse View

F Sigma

Free-R

Spacegroup from MTZ file: *Note that MrBUMP will assume this spacegroup for the MR stage*

MTZ out PROJECT Browse View

PDB out PROJECT Browse View

Number of molecules in the asu: ☐ Leave blank for automatic calculation

Template Search Options

User specified search models

Search Model Preparation Options

Maximum number of search results from which to generate search models for use in MR: 5

Search model types to create:

- ☐ Unmodified - No changes to search model
- ☐ PDBClip - Remove waters, hydrogens and select most probable conformations for side chains
- ☐ Molrep - Side chain truncation based on Molrep sequence alignment
- ☒ Chainsaw - Side chain truncation based on multiple alignment sequence from MaFit
- ☒ Sculptor - Side chain truncation and gap opening based on multiple alignment
- ☐ Polyalanine - Removal of all side chains

Number of Ensemble models to prepare: 1

Maximum number of prepared models to use in each Ensemble model: 5

Molecular Replacement and Refinement Options

Pre-determined Component

Enter PDB files containing already determined fixed solution for the MR search:

Edit list Add PDB file

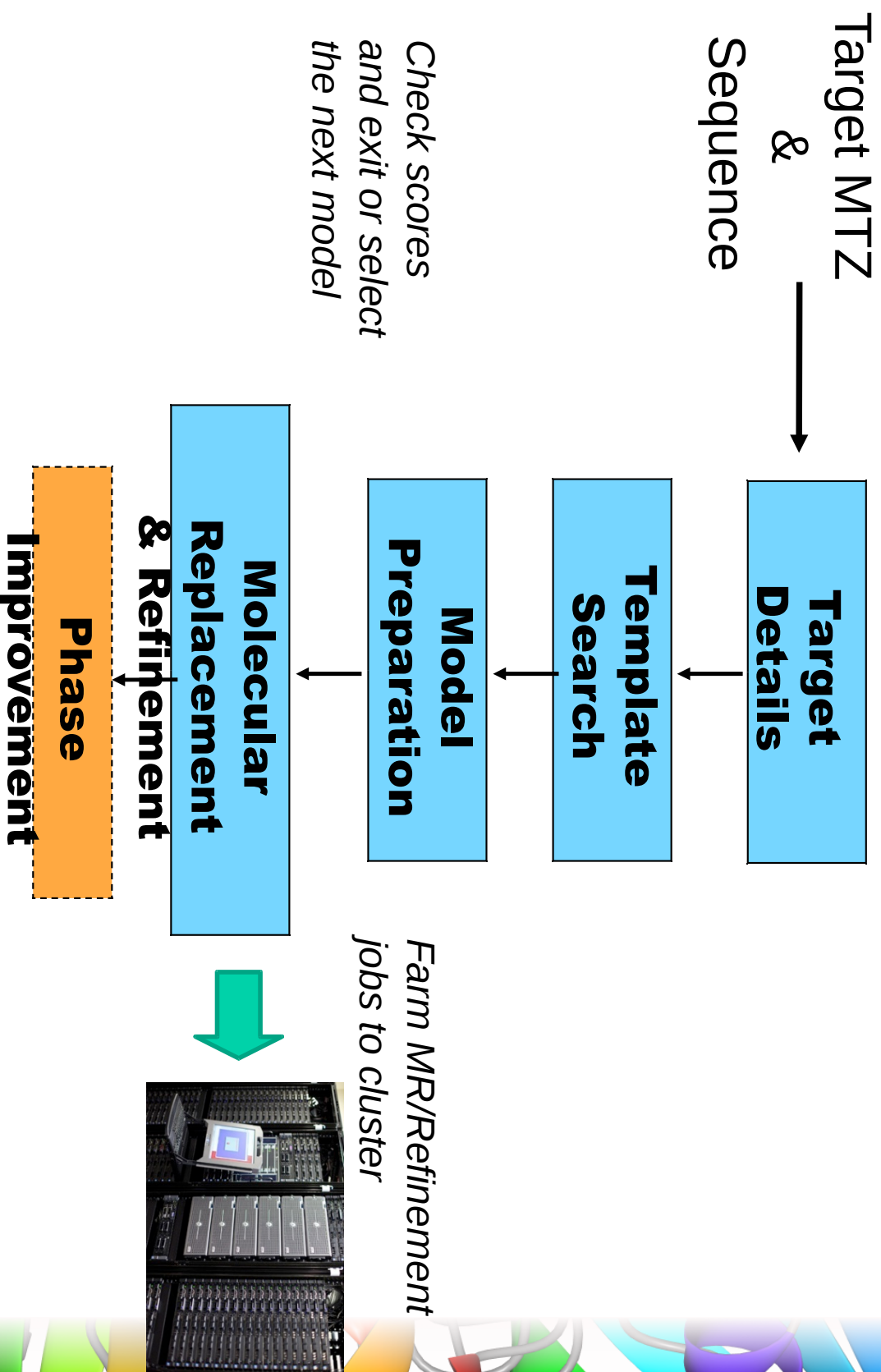
Model Building and Phase Improvement

Run Save or Restore Close

Dependency Warning!: FastA program not found (fasta34, 35 or 36) [Download from: http://www.ebi.ac.uk/fasta/](http://www.ebi.ac.uk/fasta/)



Mr BUMP pipeline



MrBUMP – search for model templates

FASTA search of PDB

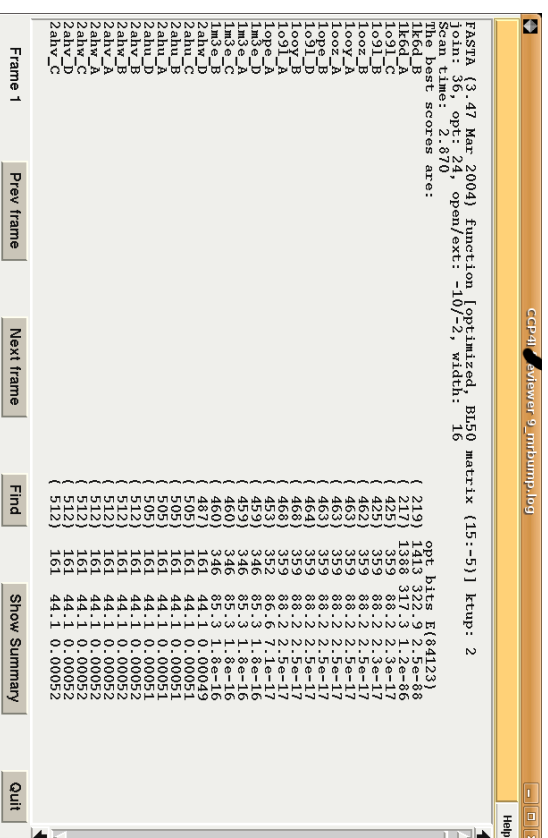
- ## Sequence based search using sequence of target structure

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

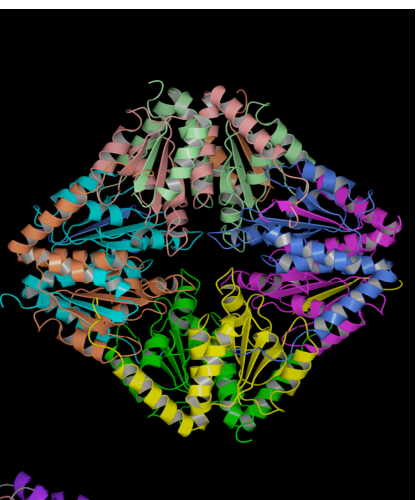


MrBUMP – domains and multimers



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



Use template multimer as model for target multimer

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



MrBUMP – search models

Search models can be prepared in five 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.

Sculptor

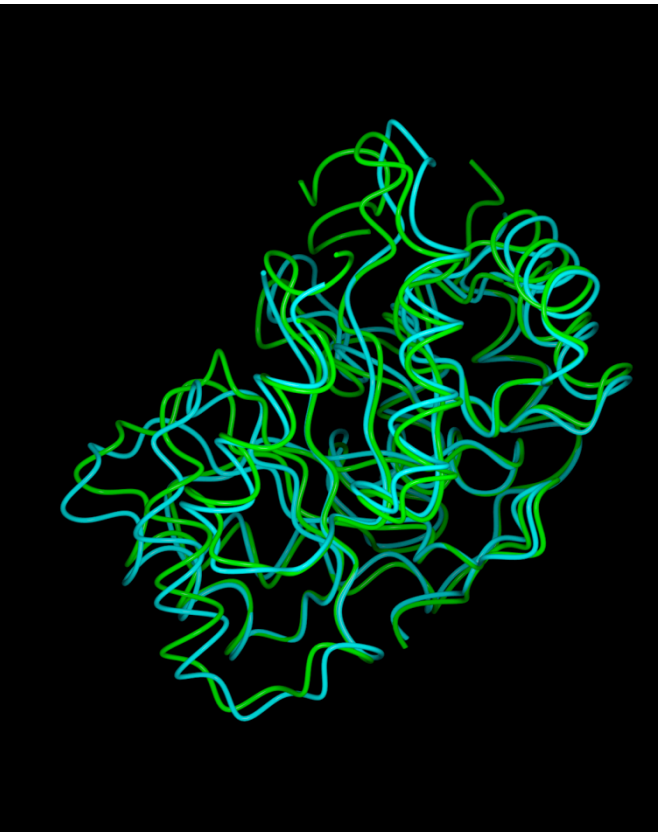
- Similar to chainsaw but different protocols for side chain truncation and alignment

Polyalanine

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



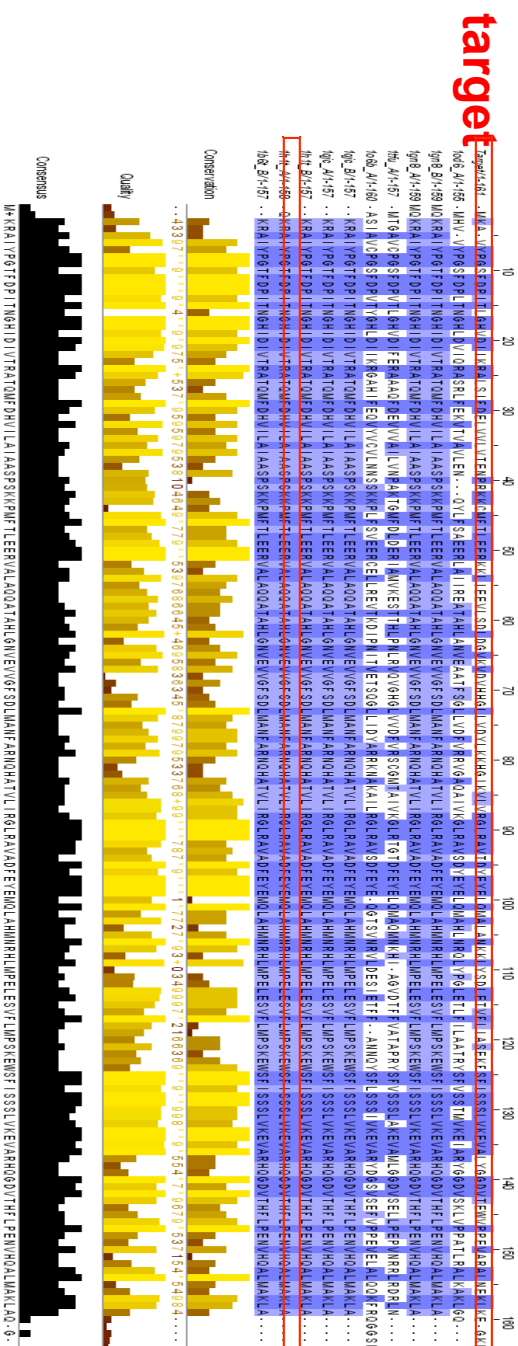
MrBUMP - ensembles



- Create ensembles of top search models, for use in additional runs of [Phaser](#).
- Models must be sufficiently similar (MW and rmsd)
- Future versions will incorporate phaser.ensemble to provide more ensemble search models



MrBUMP – multiple alignment



Pairwise alignment
Used in chainsaw and sculptor

Jalview 2.08.1 Barton group, Dundee

Model template scoring: score = sequence identity X alignment quality



MrBUMP – MR and Rigid Body Refinement

Running MR

For each search model, MR done with **Molrep** or **Phaser** or both.

MR programs run mostly with defaults

Allow Molrep / Phaser to set resolution limits and weights

Rigid body scoring

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”



Mr BUMP – model building

- **New:** 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



MrBUMP - other

Prepositioned models/complexes

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

Pre-determined Component

Enter PDB files containing already determined fixed solution for the MR search:

Fixed PDB

Sequence identity for this component

Phase extension

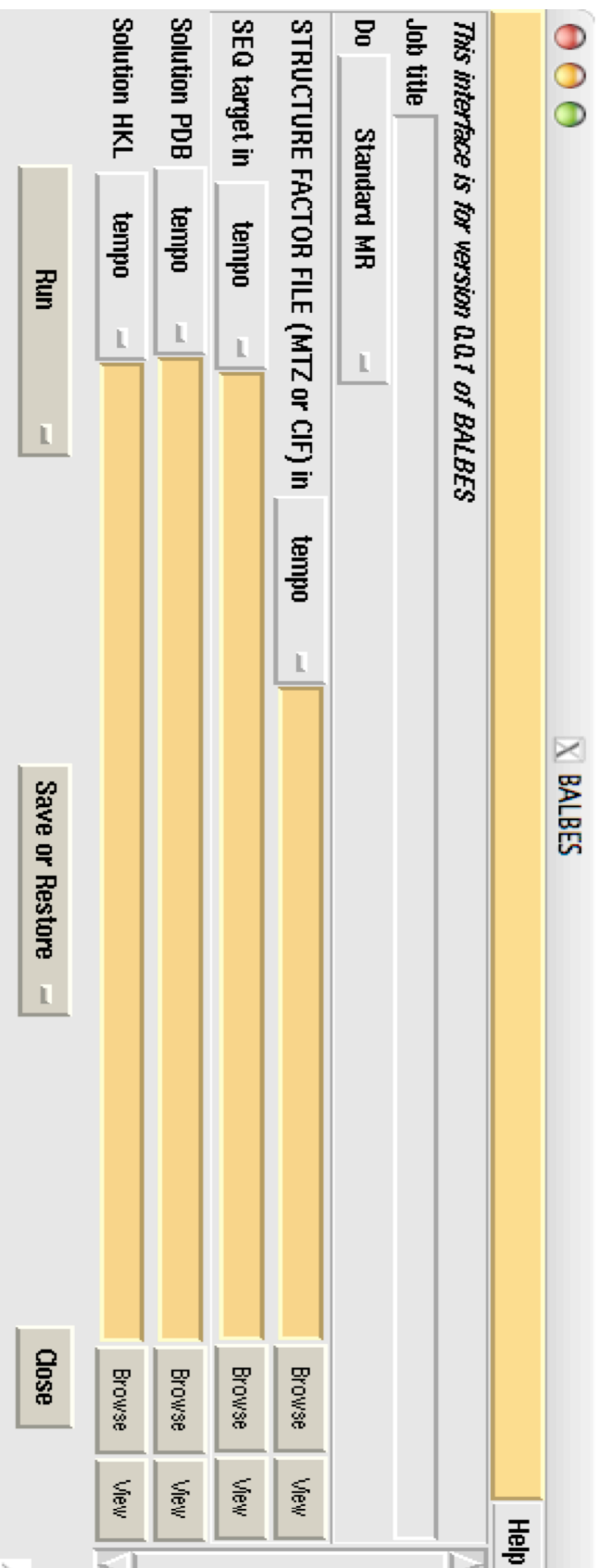
If resolution better than 1.7 Å use Acorn procedure

Initial phase set from refined MR solution

Artificial phase extension to 1.0 Å



Bαλβεσ — Murshudov, Long, Vagine

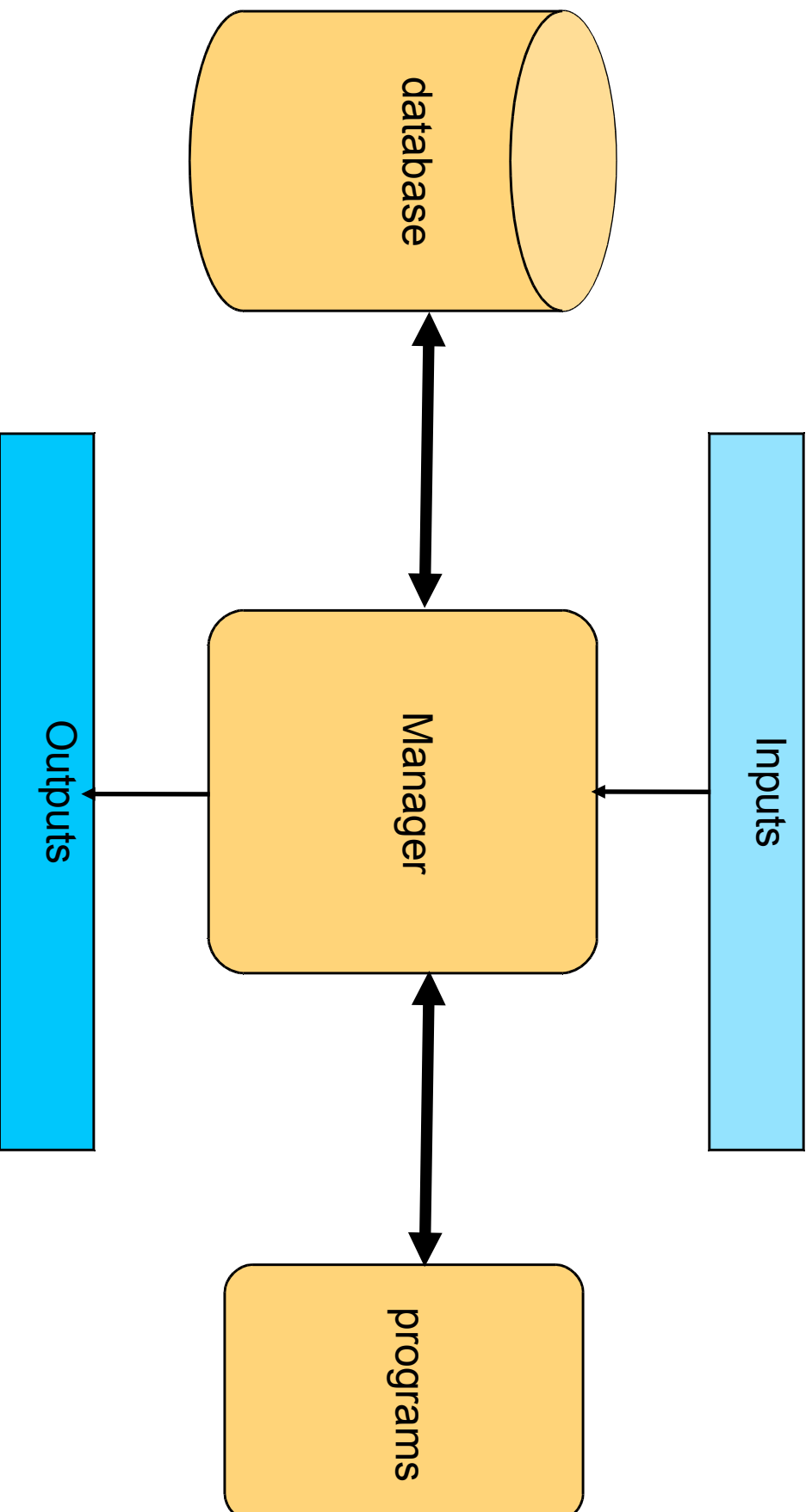


% balbes -f *structure_factors_file* -s *sequence_file* -o *output_directory*

- f required
- s required
- o optional



$\beta\alpha\beta\epsilon\sigma$ - schematic



$\beta\alpha\lambda\beta\epsilon\sigma$ - manager

Organises

Data

Resolution for molecular replacement

Data completeness and other properties

Twinning

Pseudo translation

Sequence

Finds template structures with their domain and multimer organisations

Estimates number of molecules in the asymmetric unit

“Corrects” template molecules using sequence alignment

Protocols

Runs various protocols with molecular replacement and refinement and makes decisions accordingly



$\beta\alpha\lambda\beta\epsilon\sigma$ - database

Organised according to:

Chains . The internal database has around 35000 unique entries selected from more than 51,000 present in the PDB.

All entries in the PDB are analysed according to their identity. Only non-redundant sets of structures are stored.

Domains. The DB contains 35000 domain definitions. Loops and other flexible parts are removed from the domain definitions.

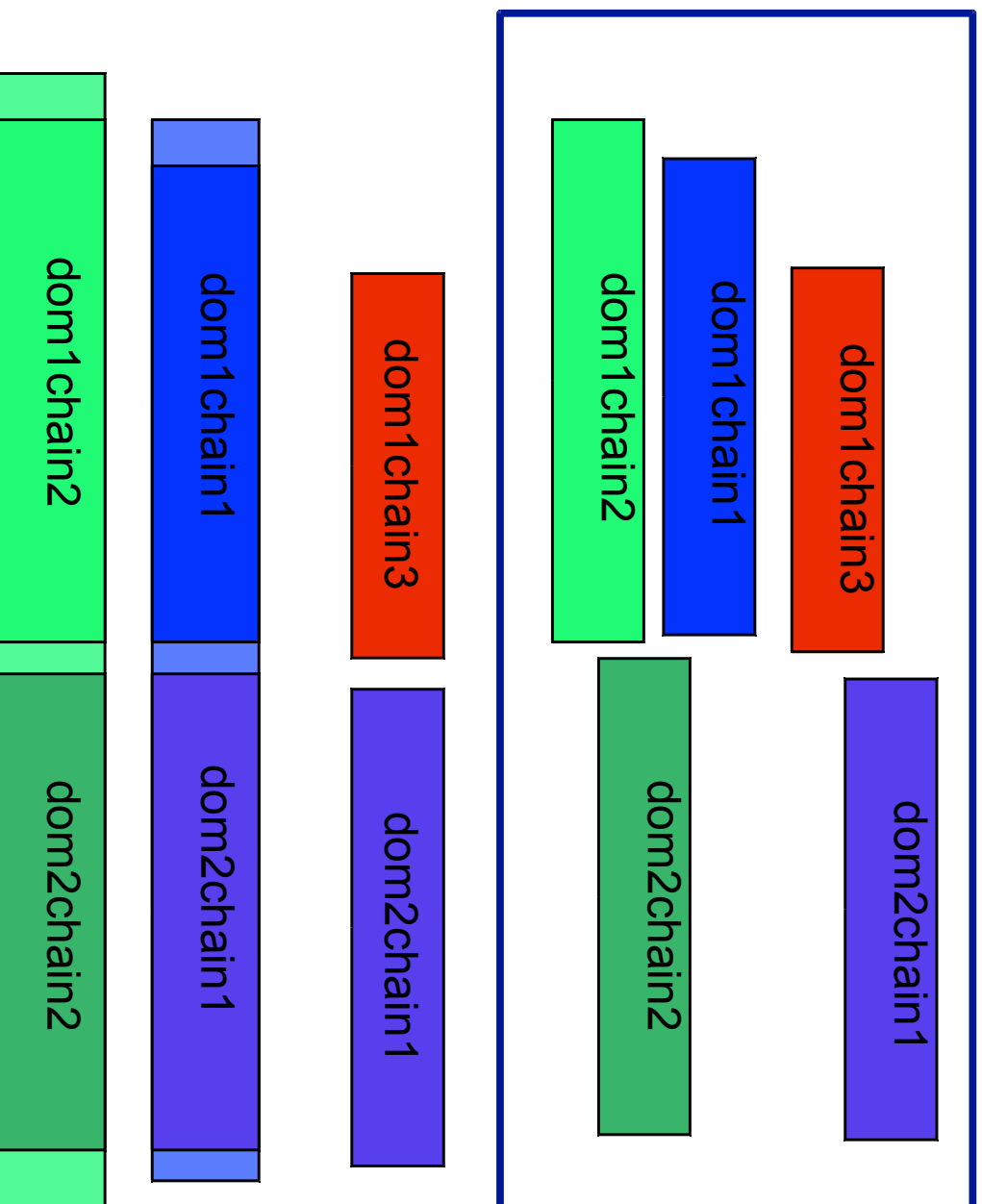
Multimers of structures (using PISA)

Hierarchy is organized according to sequence identity and 3D similarity (rmsd over Ca atoms).

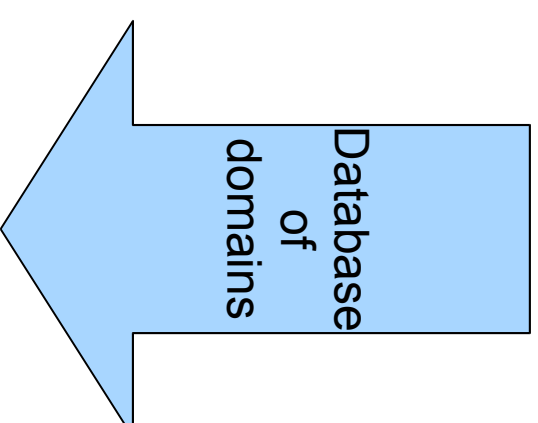


$\beta\alpha\lambda\beta\epsilon\sigma$ – model selection

DVSGTVCLSALPPEATDTLNLIASDGPFPYSQDGWVFQ



Input sequence



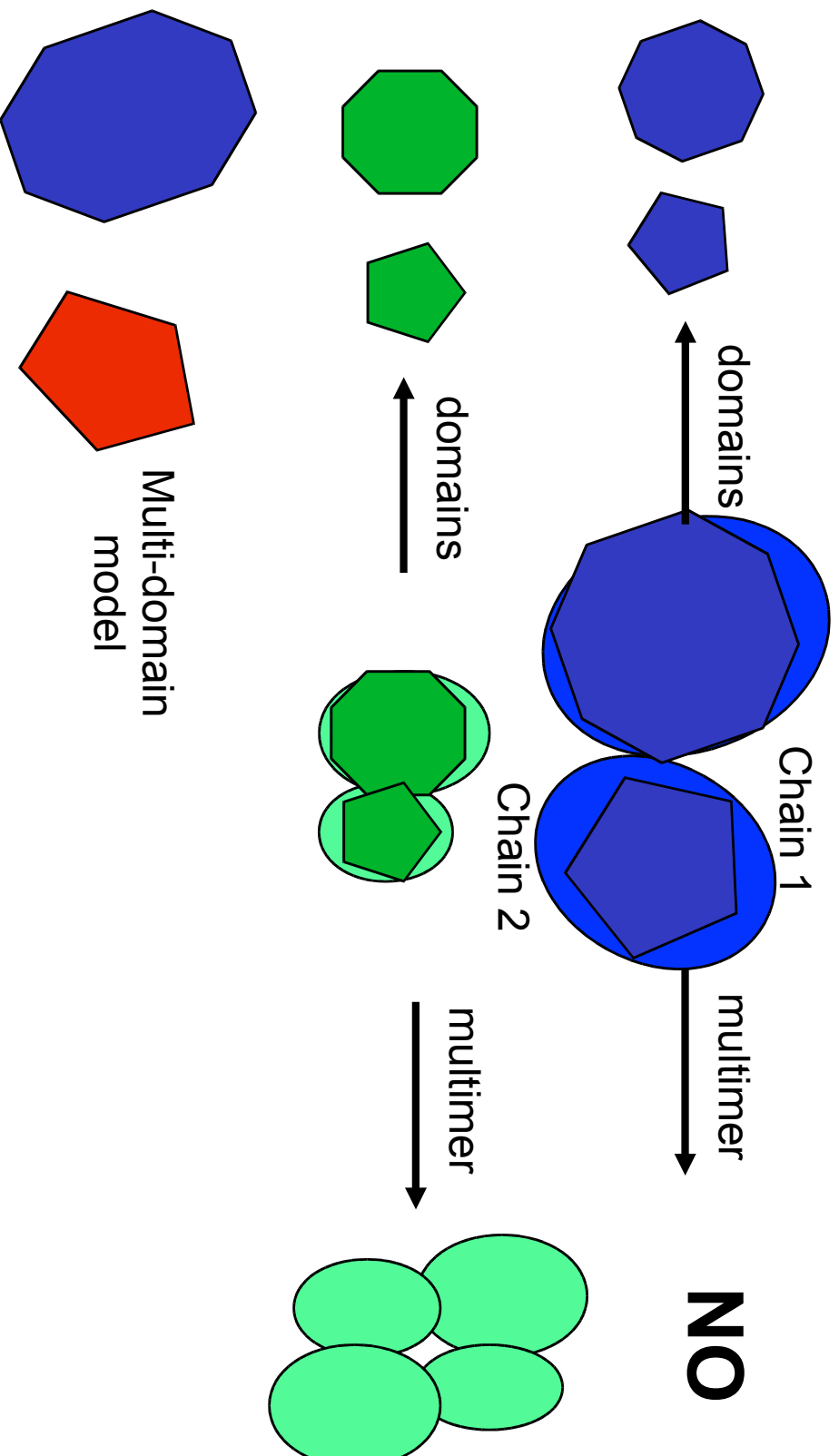
Best multi-domain model

Full chain models



$\beta\alpha\beta\beta\epsilon\sigma$ – model preparation

All models are corrected by sequence alignment and by accessible surface area

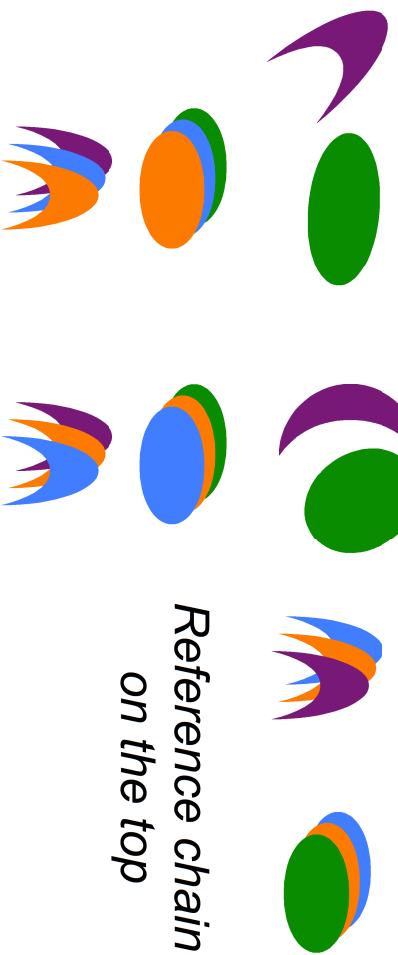


$\beta\alpha\lambda\beta\epsilon\sigma$ – ensemble models

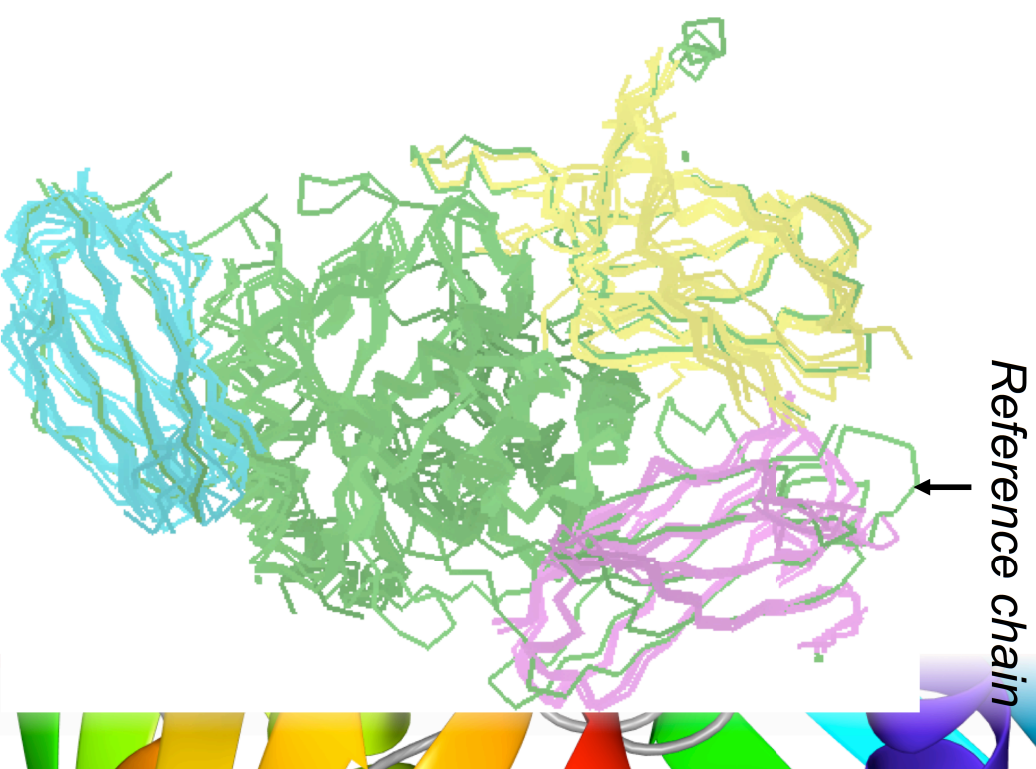
Homologues from the Balbes database:



Ensemble search models:



Example:



$\beta\alpha\lambda\beta\epsilon\sigma$ – refinement

R-values

Final decisions are made based on R-factors after refinement.
Since we have similar structures we can use them in refinement.

Jelly body

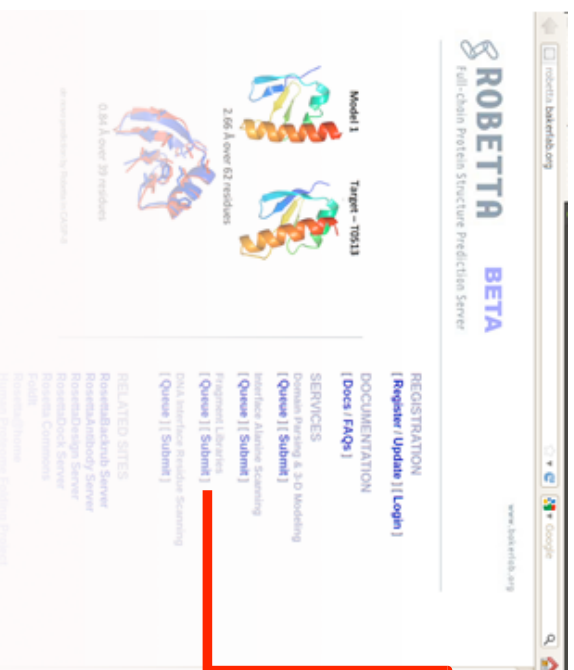
In refinement stage “jelly-body” refinement is used. It seems to increase success rate, especially for multidomain cases.

Space group

Future version will use more extensive search of space groups and decision on space group will be made after refinement.



AMPLE – Biddy, Keegan



AMPLE (beta) - Ab initio modelling for Molecular Replacement Initial parameters from /home/rolo/

Job title 2cmp ab initio model MR Ample User Guide Help

Input Files

SEQ in	ample	2cmp_1asia	Browse	View
MTZ in	ample	2cmp.mtz	Browse	View
F	FP	Sigma	SIGFP	
Free-R	FREE			
Number of models to generate:	500			
Number of Processors	1			

Fragment Files

To generate fragment files for your sequence you need to use the Robetta Fragment library server

Click here to go to online Robetta Server (registration required)

3mers (aaXXXX_03_05.200_v1_3) /home/rmk65/CCP4/Projects/ample Browse

9mers (aaXXXX_09_05.200_v1_3) /home/rmk65/CCP4/Projects/ample Browse

Rosetta Installation

Rosetta Installation Directory /home/rmk65/opt/src/rosetta3.4 Browse

Run Save or Restore Close

Documentation available
from <http://ccp4wiki.org>



AMPLE – *ab initio* structure prediction

Ab initio (or *de novo*) structure prediction is the prediction of a target structure fold based purely on its sequence information.

Methods have greatly improved in recent years with the aid of the CASP experiments
(Critical Assessment of Protein structure prediction).

Some examples are:

Rosetta

I-TASSER

QUARK



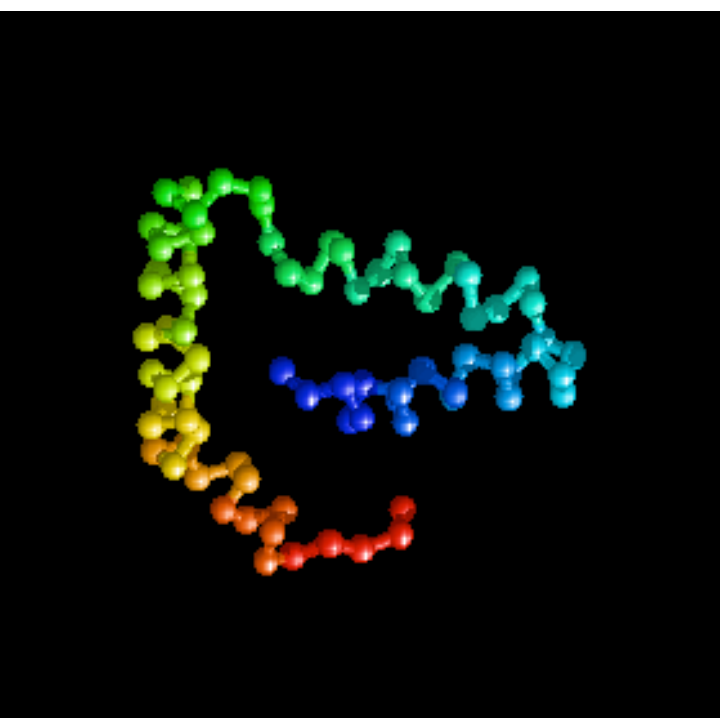
AMPLE – Monte Carlo sampling

1000's of “Decoys” assembled of fragments from PDB structures.

Decoys are clustered and centroid representatives of largest cluster are considered candidate fold predictions.

Side chains added to selected decoys.

Refinement under a more realistic physics-based force field.



Initial fragment assembly stage requires relatively modest computing power.

Refinement stage can require supercomputing resources.



AMPLE - synergies

ab initio modelling

Produces **clusters** of similar model structures

Molecular Replacement

Works effectively with superposed **ensembles** approximating the target



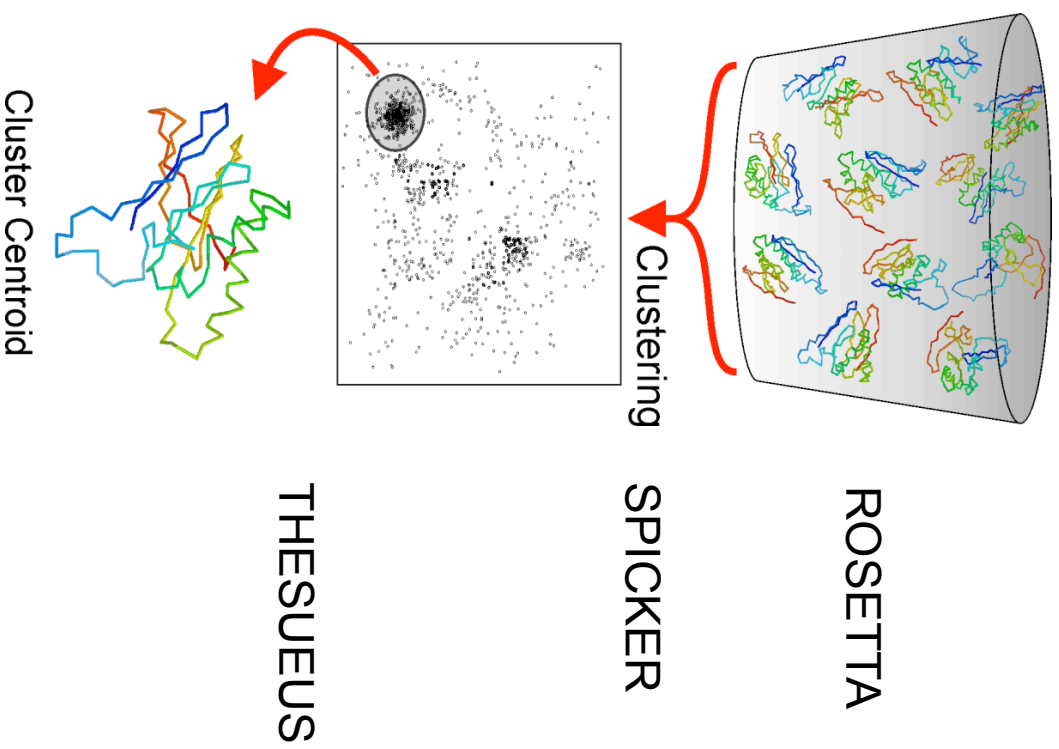
AMPLE – the pipeline

Designed to make ab initio modelling for use in MR easily accessible to users.

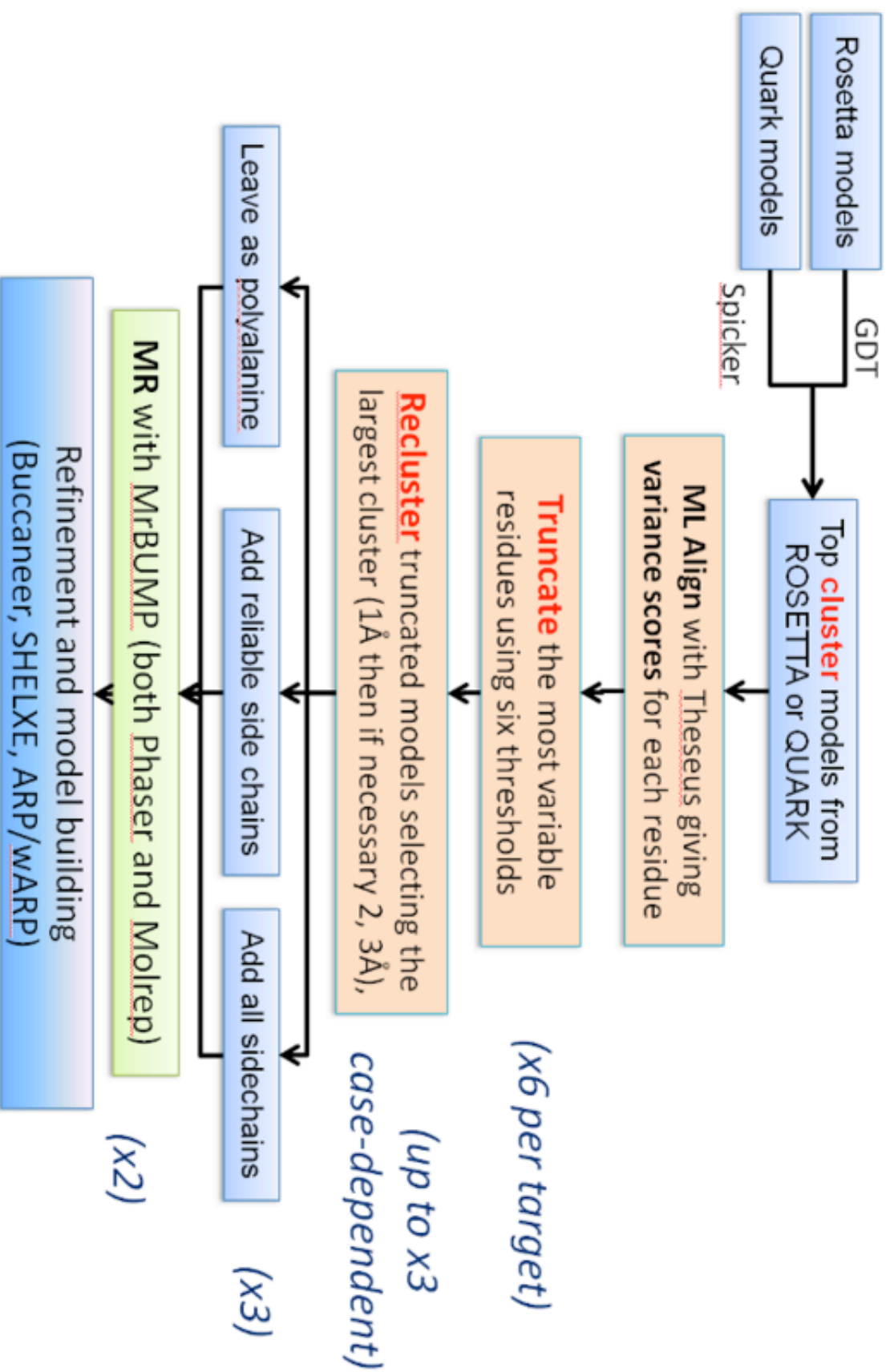
Generate large number of trial models.

Cluster into groups according to RMS deviation.

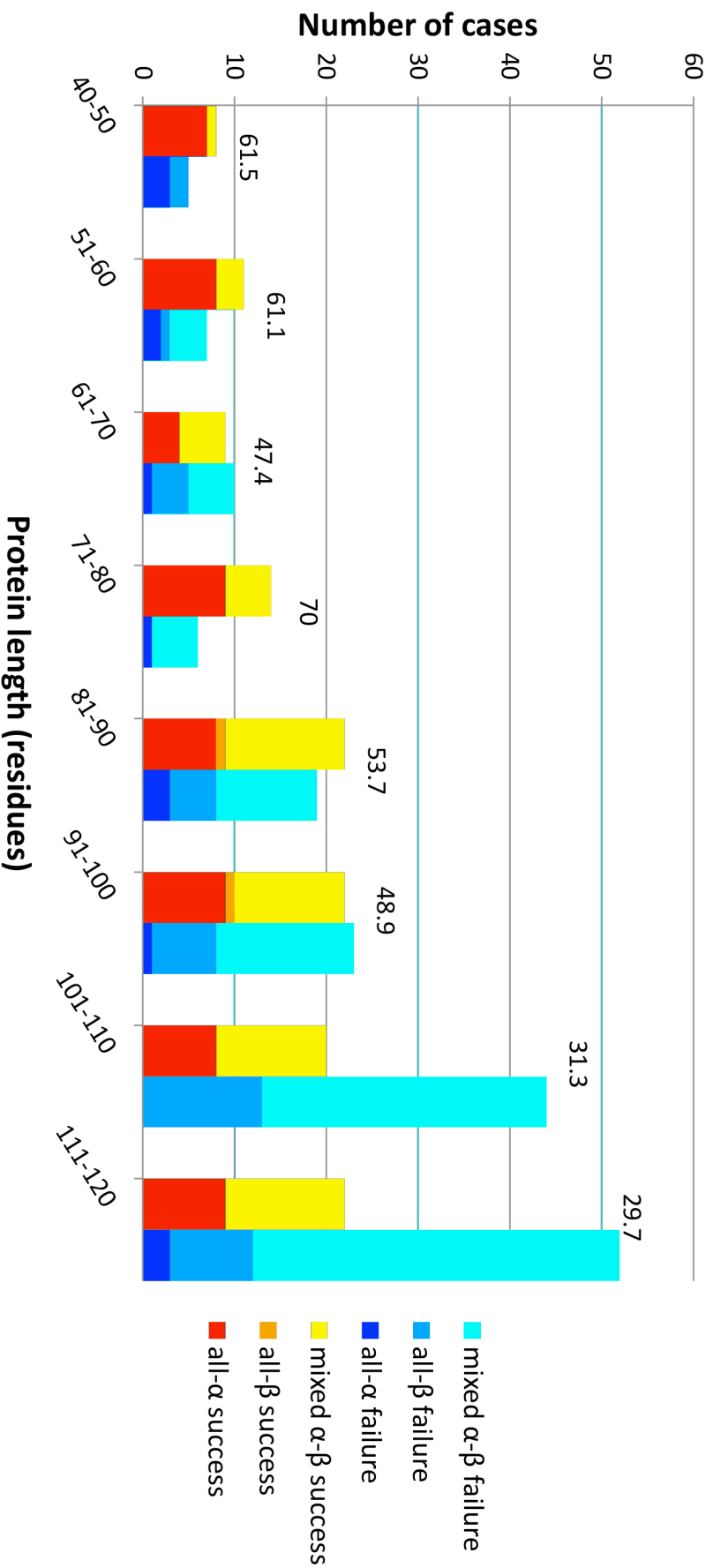
Select templates from largest cluster.



AMPLE – pipeline in more detail



AMPLE – effectiveness based on secondary structure type



Overall success rate:

all- α = 80%; **all- β** = 2%; **mixed α - β** = 37%



Acknowledgements

- Garib Murshudov, Fei Long, MRC-LMB
- Jaclyn Bibby, Daniel Rigden, Olga Mayans, University of Liverpool
- Martyn Winn, Daresbury Laboratory
- Ronan Keegan, RAL
- Alexei Vagine, University of York (retired)
- Thanks to authors of all underlying programs

