Automatic MR

Garib Murshudov

MRC-LMB, Cambridge

These slides provided by Ronan Keegan

CCP4

Science & Technology Facilities Council

MR pipelines in CCP4

MrBUMP - Ronan Keegan, Martin Winn

BALBES – Fei Long, Alexei Vagin, Garib Murshudov

Ample - Ronan Keegan, Martyn Winn, Jaclyn Bibby, Jens Thomas, Olga Mayans and Daniel Rigden

CCP4

Science & Technology Facilities Council

MrBUMP

Ronan Keegan
(*CCP4 STFC RAL*)

Martyn Winn
(*STFC DL*)

CCP4

Science & Technology Facilities Council

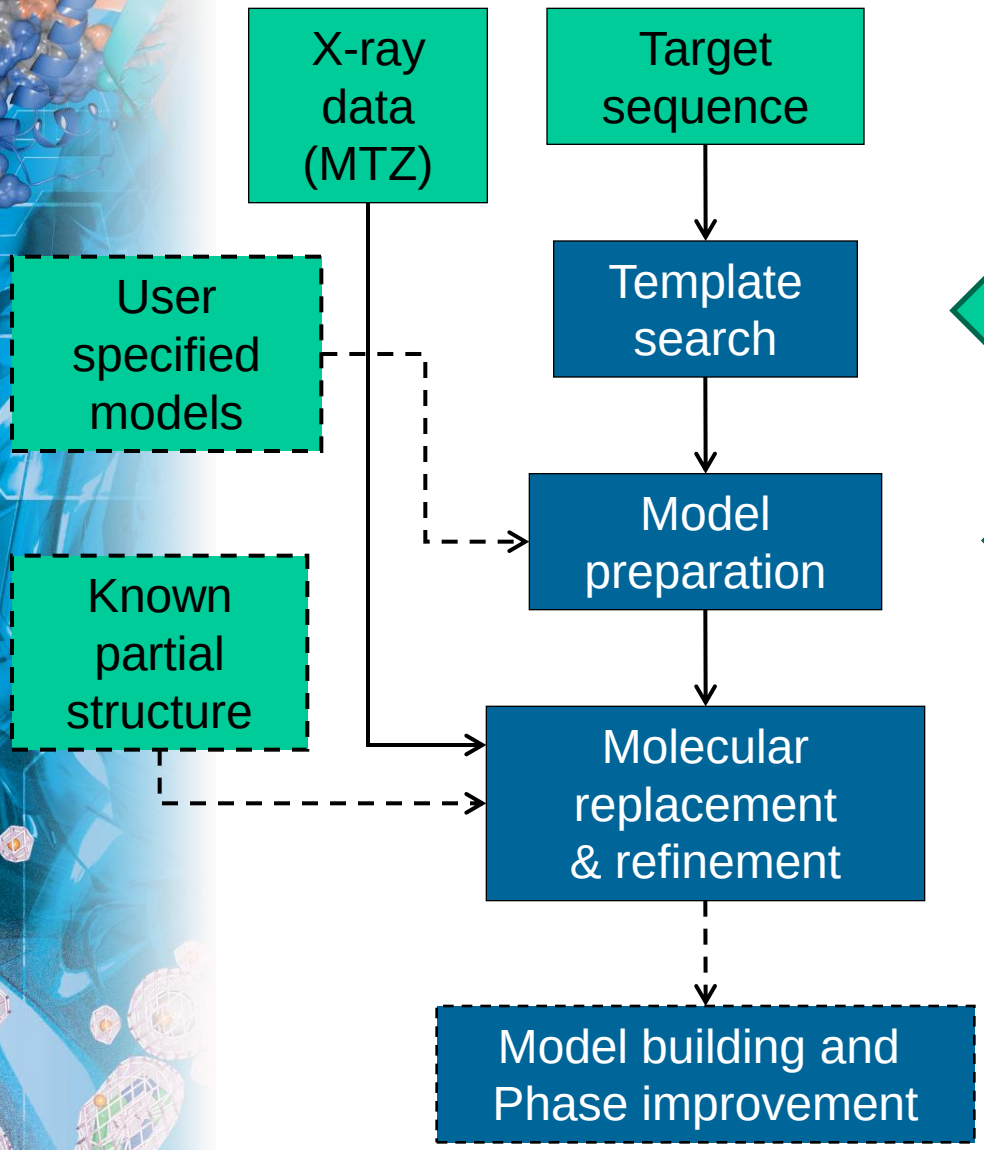
MrBUMP features

- Particular emphasis on generating a variety of search models
- Uses a variety of bioinformatics tools and helper applications
- Uses on-line databases

- Favourable cases: gives “one-button” solution
- Complicated cases: lead generation

(suggests likely search models for manual investigation)

MrBUMP Pipeline

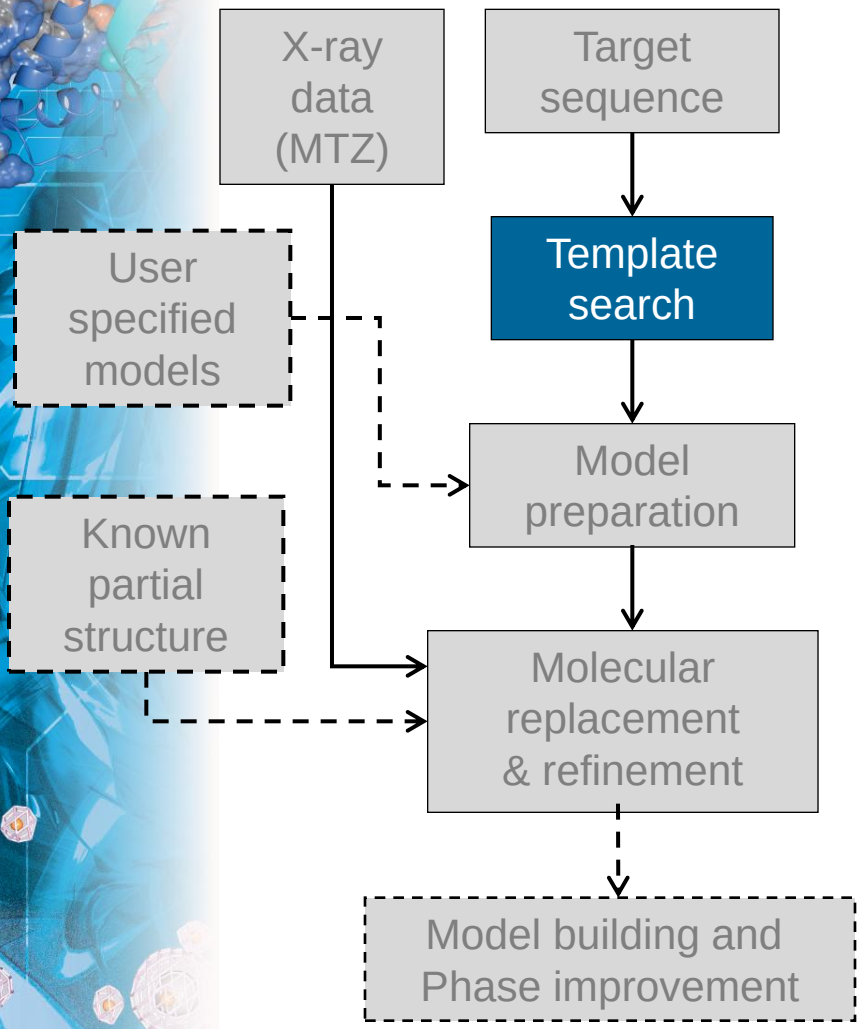


WEB resources (FASTA, PDB etc.)

Programs available from CCP4

Farm MR/Refinement jobs to cluster

MrBUMP Pipeline



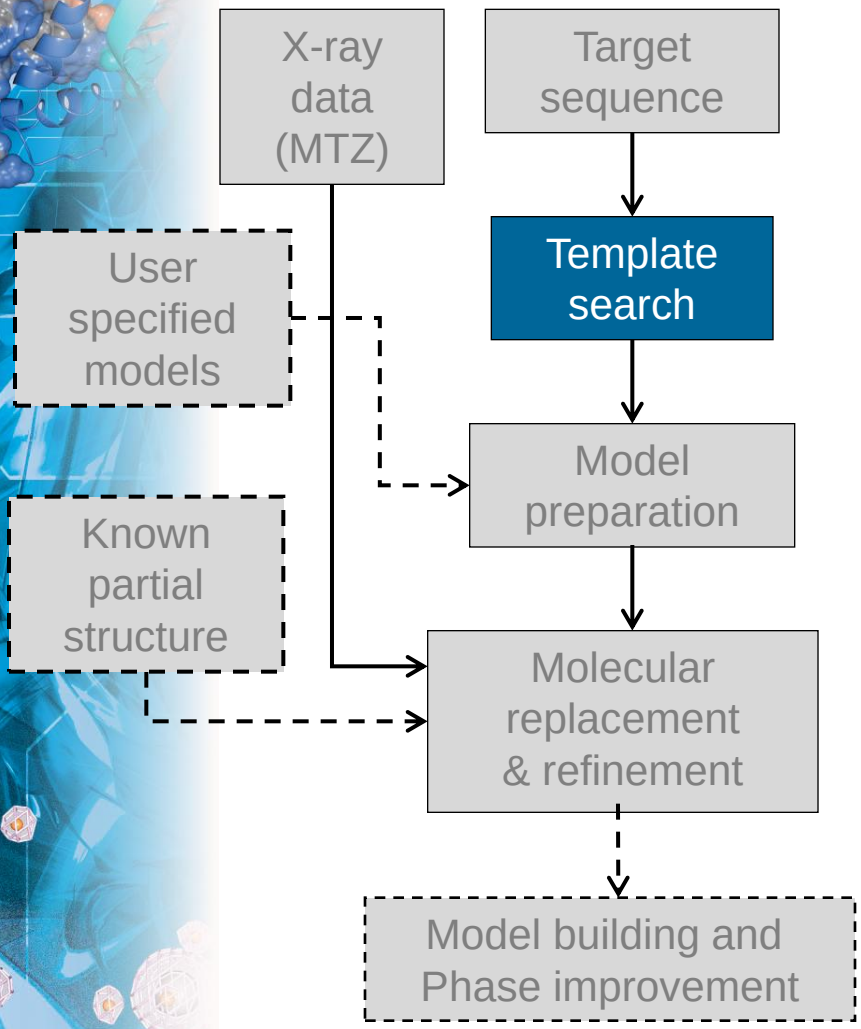
Initial list of model templates

- **Sequence based search using target sequence**
  - FASTA search of the PDB

Model template scoring

- **Based on multiple sequence alignment**
  - ClustalW, MAFFT coming with CCP4
  - Probcons, T-coffee optionally installed

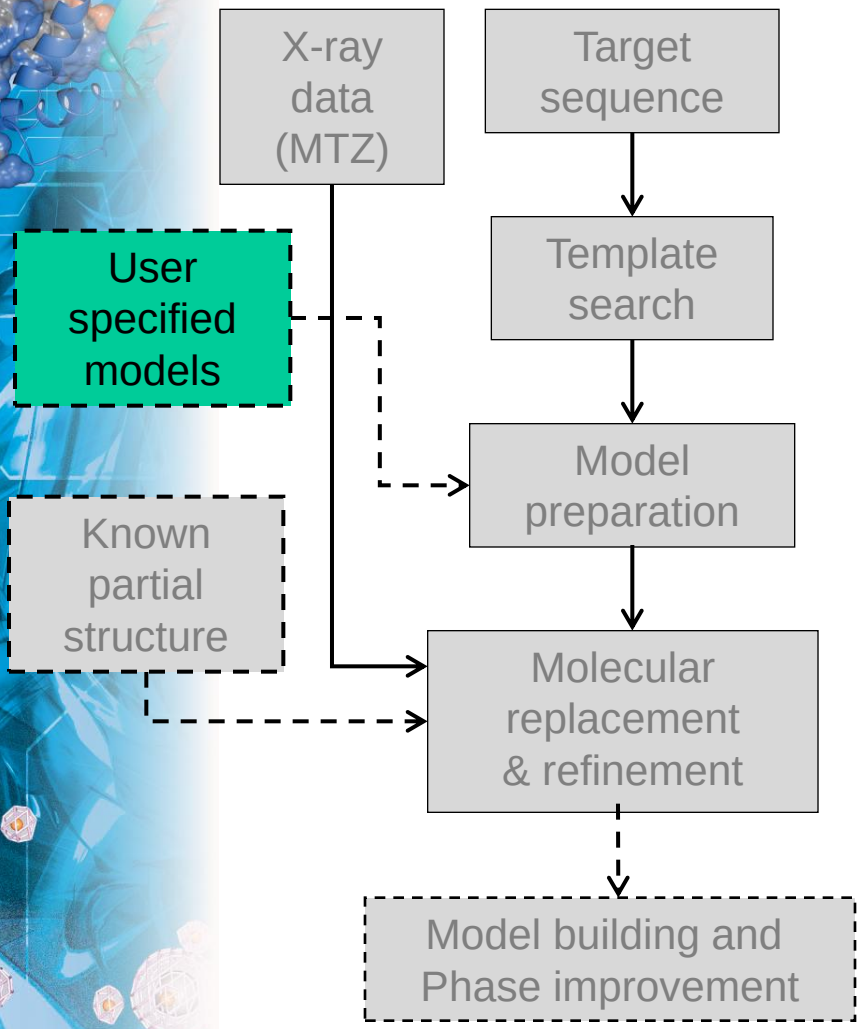
MrBUMP Pipeline



Additional model templates

- **Monomers**
  - Search PDBeFold (SSM) for top hits from the initial FASTA search
- **Domains if possible**
  - Domain definitions from SCOP for each initial template
- **Homo-multimers if possible**
  - Multimer definitions manually generated using PISA

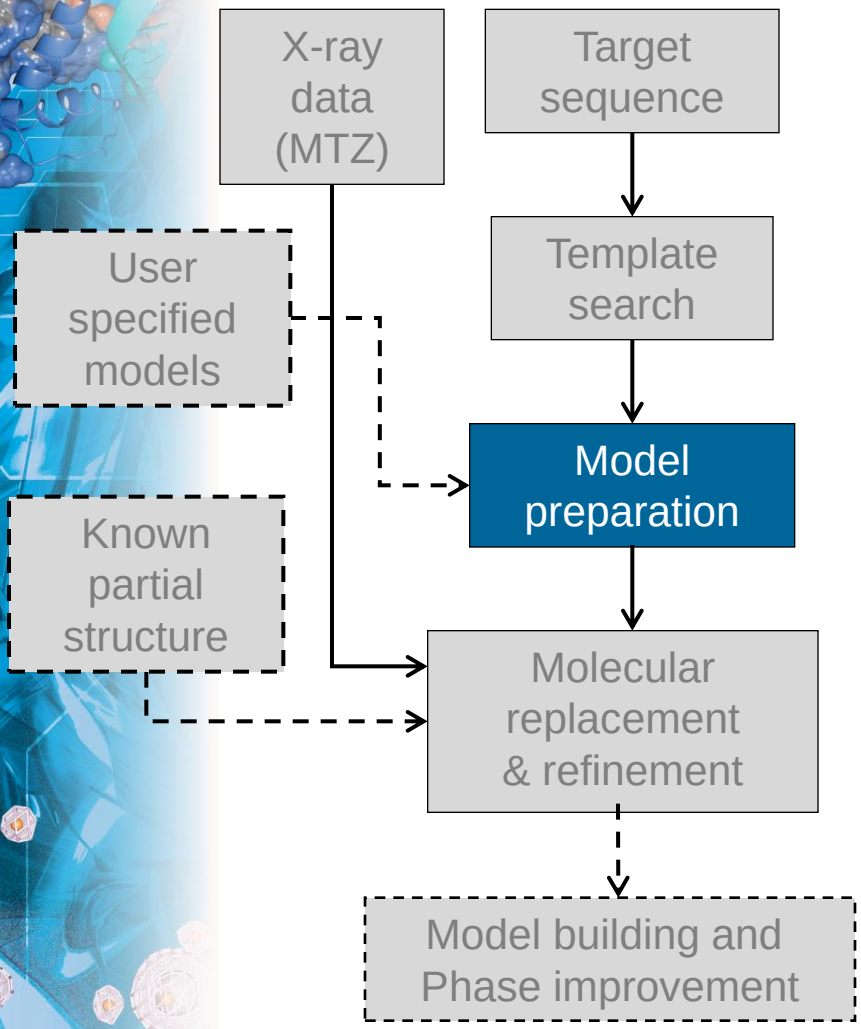
MrBUMP Pipeline



Additional model templates

- **Monomers**
  - PDB codes and chain IDs provided by a user
    - » e.g. from FFAS or psiBLAST searches
  - PDB files provided by a user
    - » e.g. unpublished structures

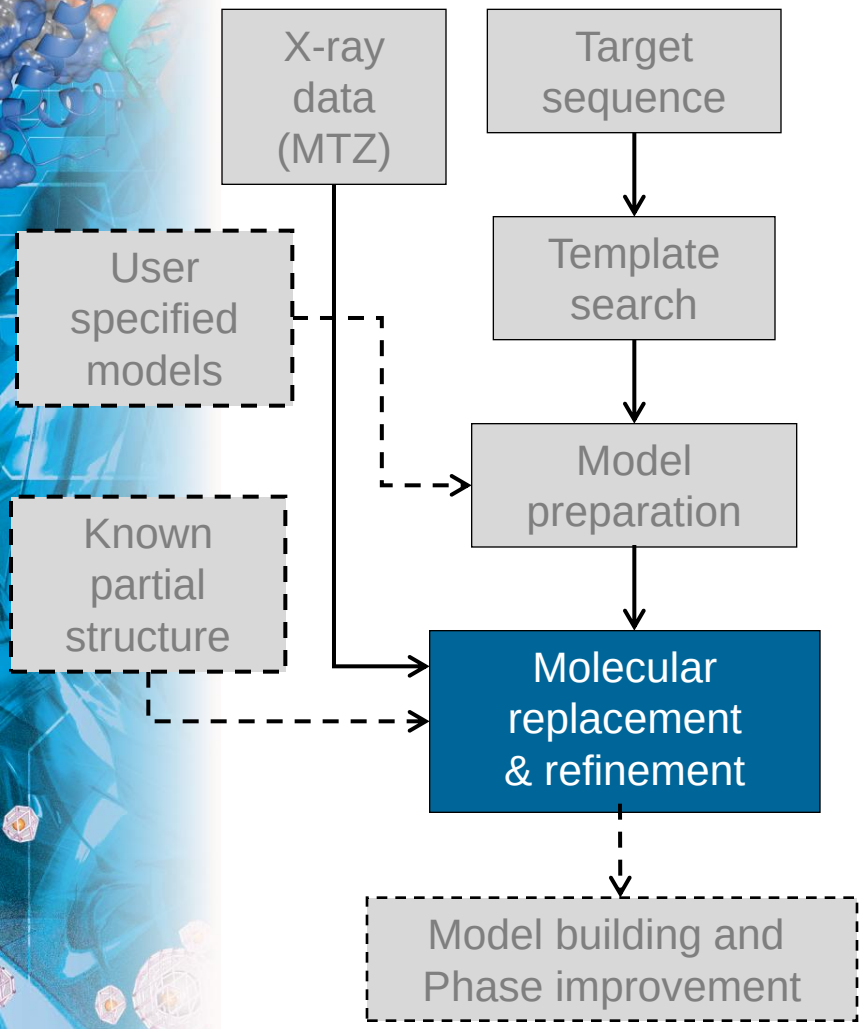
MrBUMP Pipeline



Search Model Preparation

- **Several options for each template**
  - PDBclip (unmodified)
  - Molrep
  - Chainsaw
  - Sculptor
  - polyalanine
- **Ensembles if requested**
  - generated using Superpose from top score search models

MrBUMP Pipeline



Molecular Replacement

- **Molrep or Phaser or both for each search model**
- **Test for enantiomorphic space groups if requested by a user**

Restrained Refinement

- **Using Refmac for each MR result**
  - Refinement results are used for scoring MR solutions

Scoring MR solutions

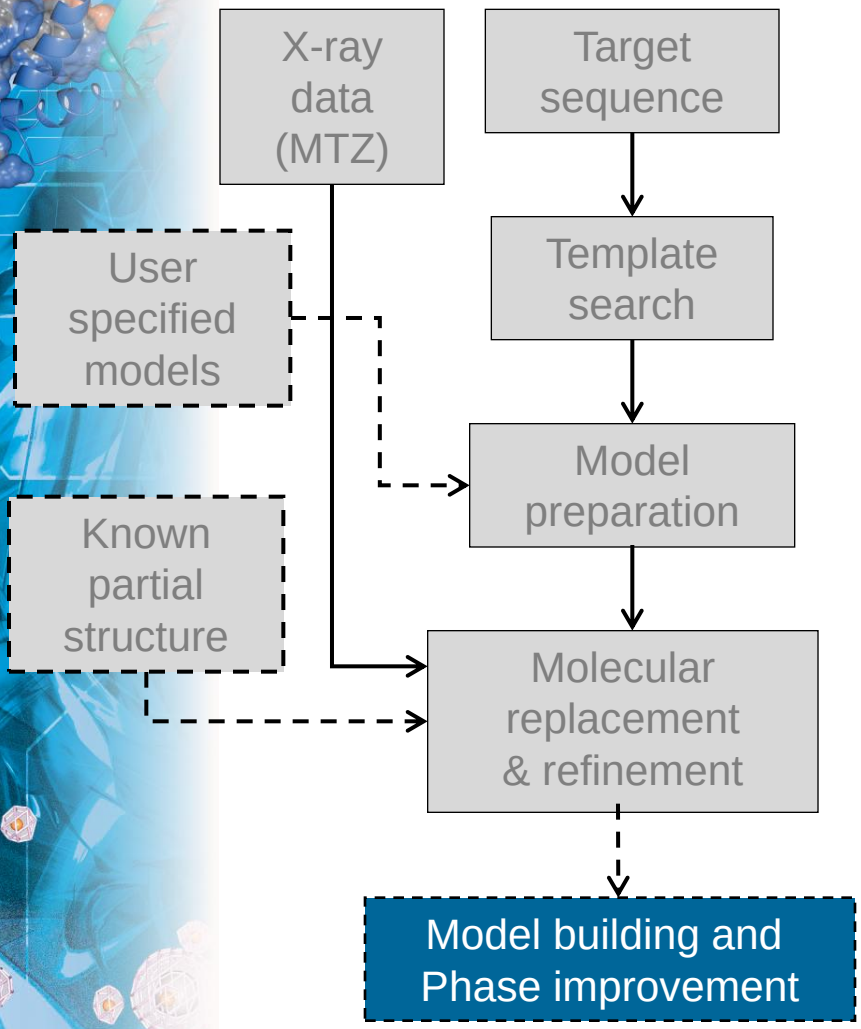
final Rfree < 0.35 or
final Rfree < 0.5 and dropped by 20% → “good”

final Rfree < 0.48 or
final Rfree < 0.52 and dropped by 5% → “marginal”

otherwise → “poor”

conservative approach,
a "poor" solution can still be a solution

MrBUMP Pipeline



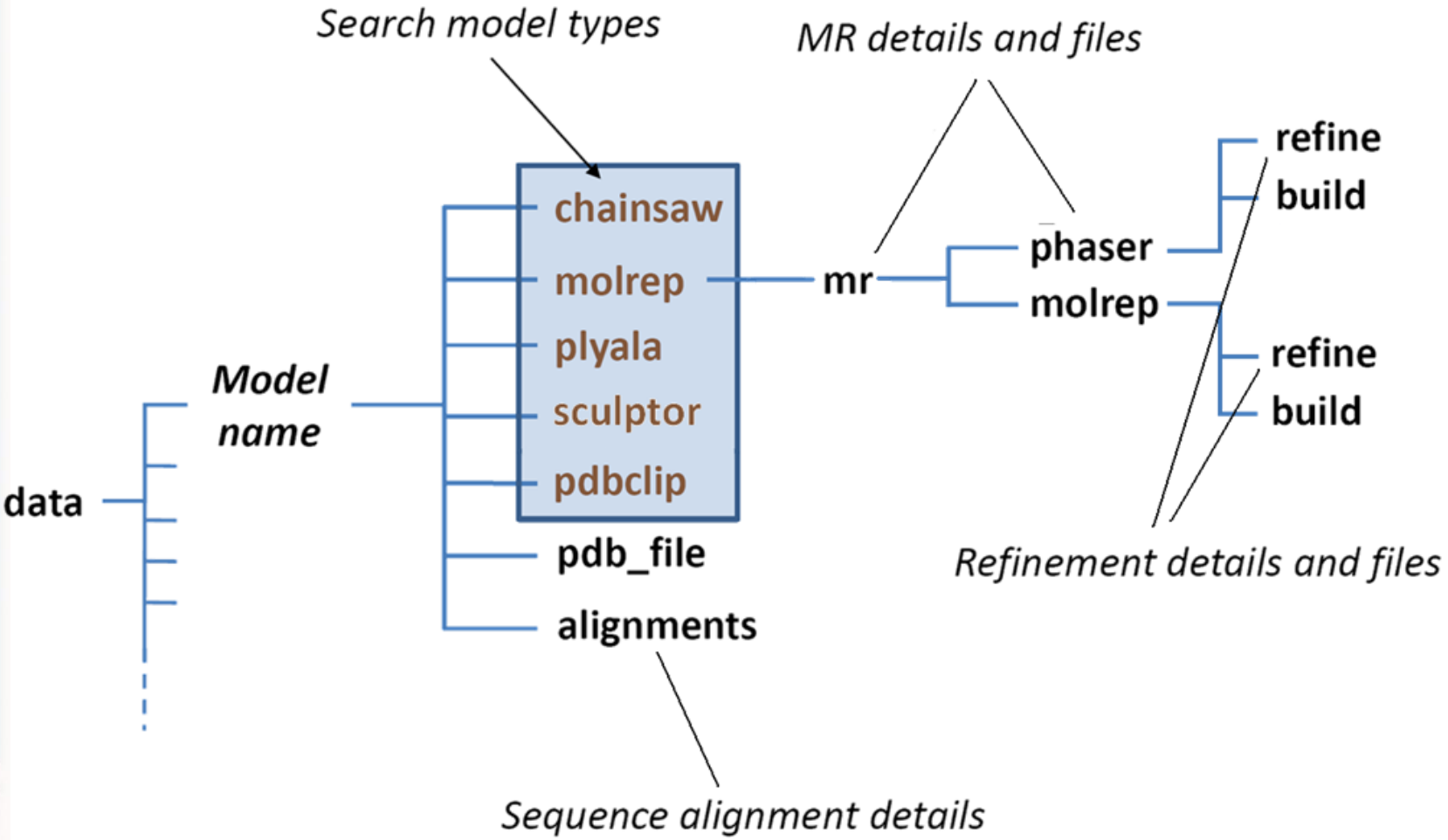
Model building

- **Buccaneer, ARP/wARP or SHELXE if requested**
- **Model completeness is used for scoring MR solutions**

Phase improvement

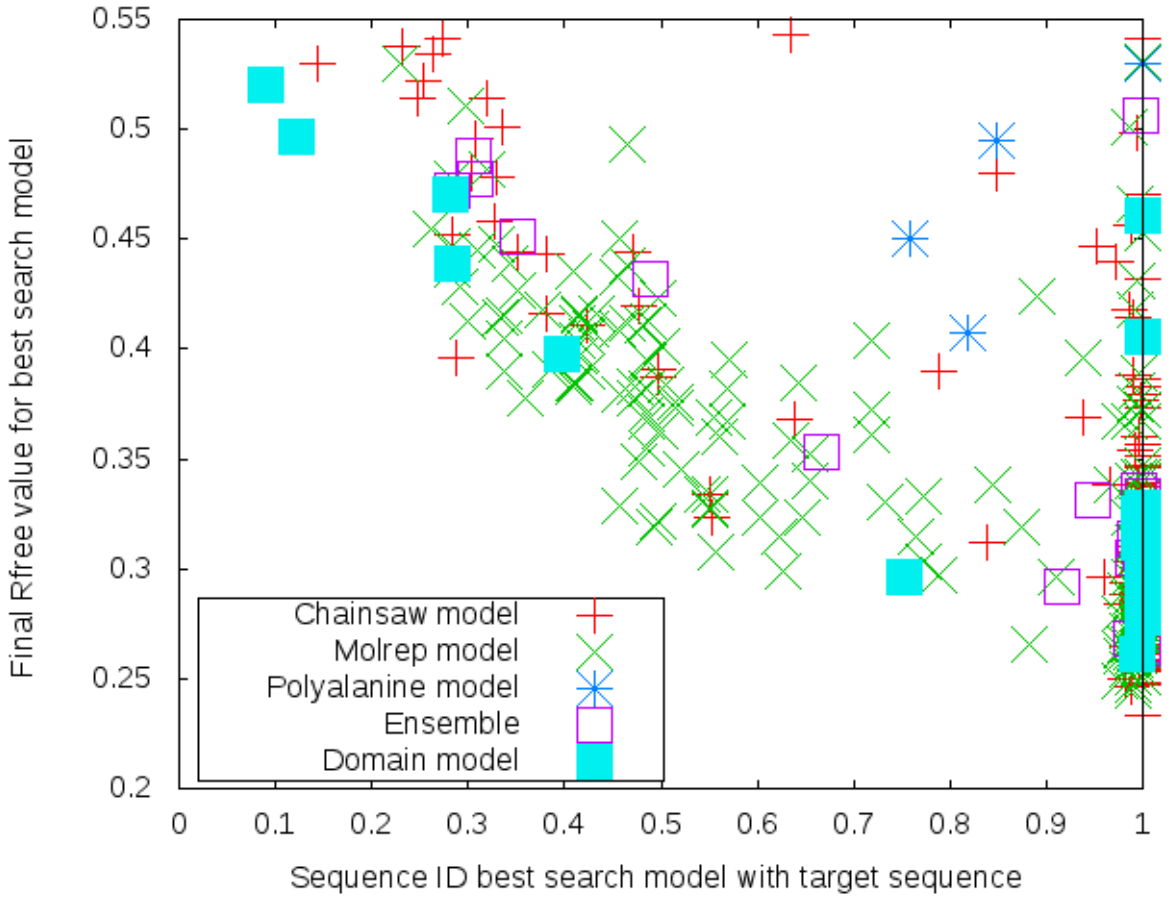
- **Acorn**
  if resolution is better than 1.7Å
- **CC for medium Es is used for scoring MR solutions**

MrBUMP data directory

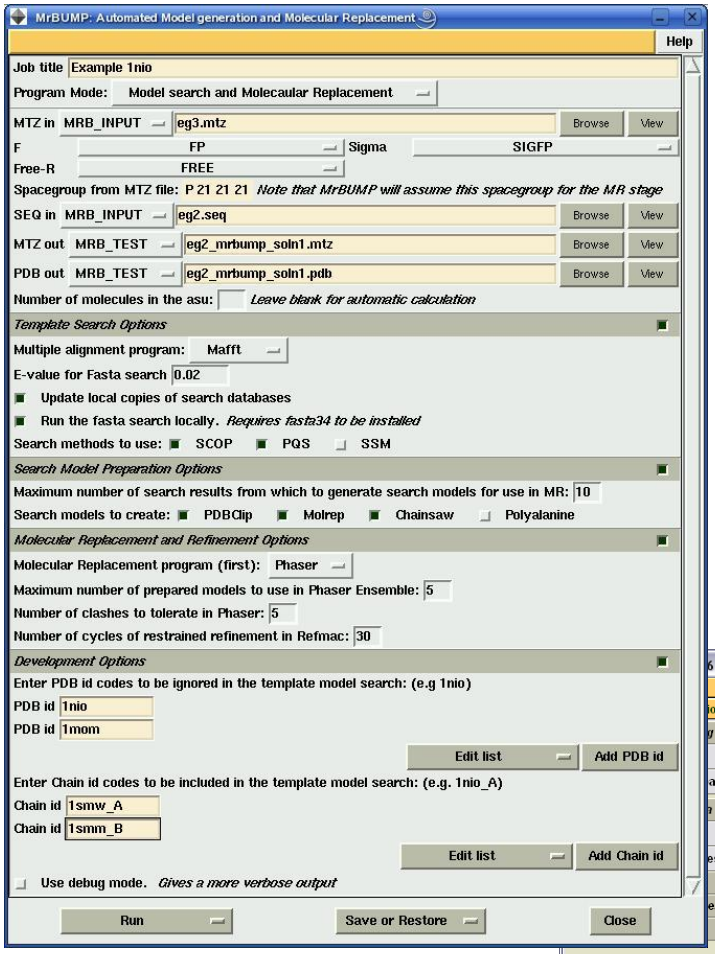


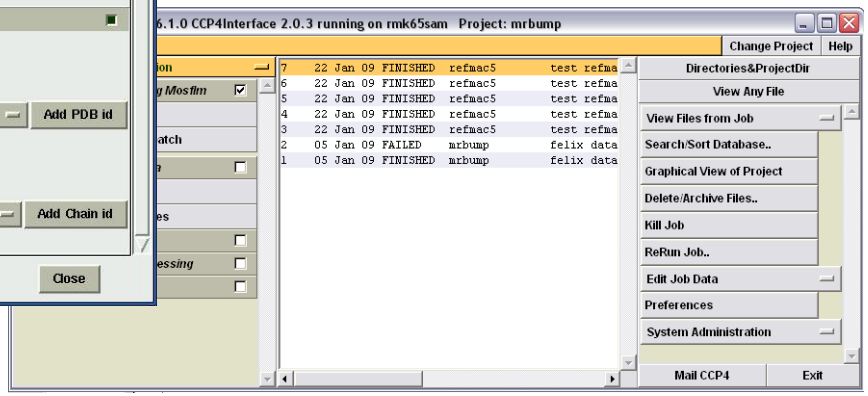
MrBUMP Testing Results

Model type for best solution



MrBUMP in CCP4i





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

Balbes

Fei Long
Garib Murshudov

MRC LMB Cambridge

Alexey Vagin

YSBL University of York

Balbes features

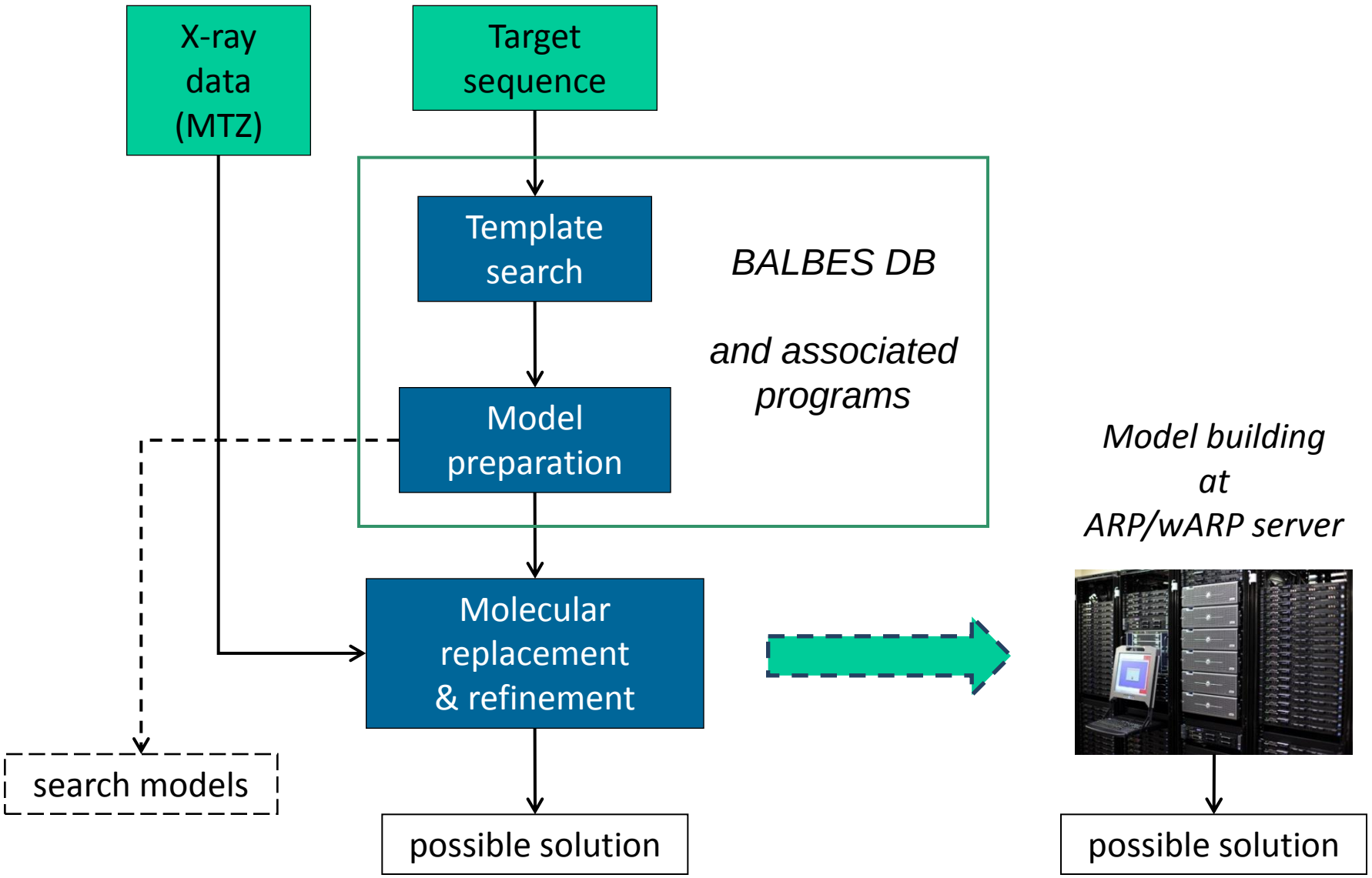
Input
- mtz-file
- sequence

Output
- the best solution
- search models for manual investigation

- Balbes has a database containing preprocessed data from the PDB

- Structure solution using Molrep and Refmac
  - Uses the search in the density for a second, third etc. component

BALBES Pipeline

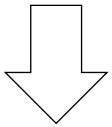


Database

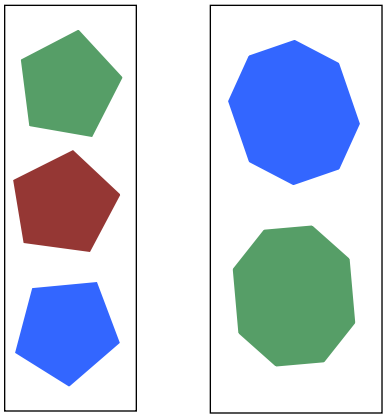
- Non-redundant chains from the PDB
  - The PDB entries of better resolution are preferred
- More than 30000 domain definitions
  - flexible parts removed
  - hierarchically organized according to 3D similarity
- Multimers
  - generated using PISA definition
- Relations
  - domain entry points to original PDB ID, chain ID and residues
  - multimer entry to original PDB ID, chain IDs

Model preparation

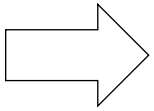
MARPVPEETVATERVKE

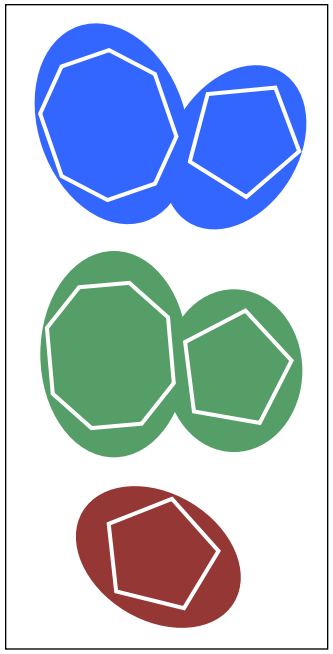


MARPVPEETVATERVKE

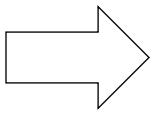


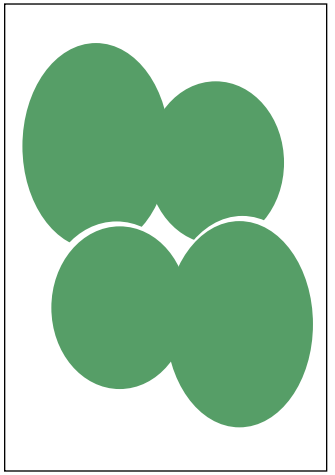
Domain models





Chain models

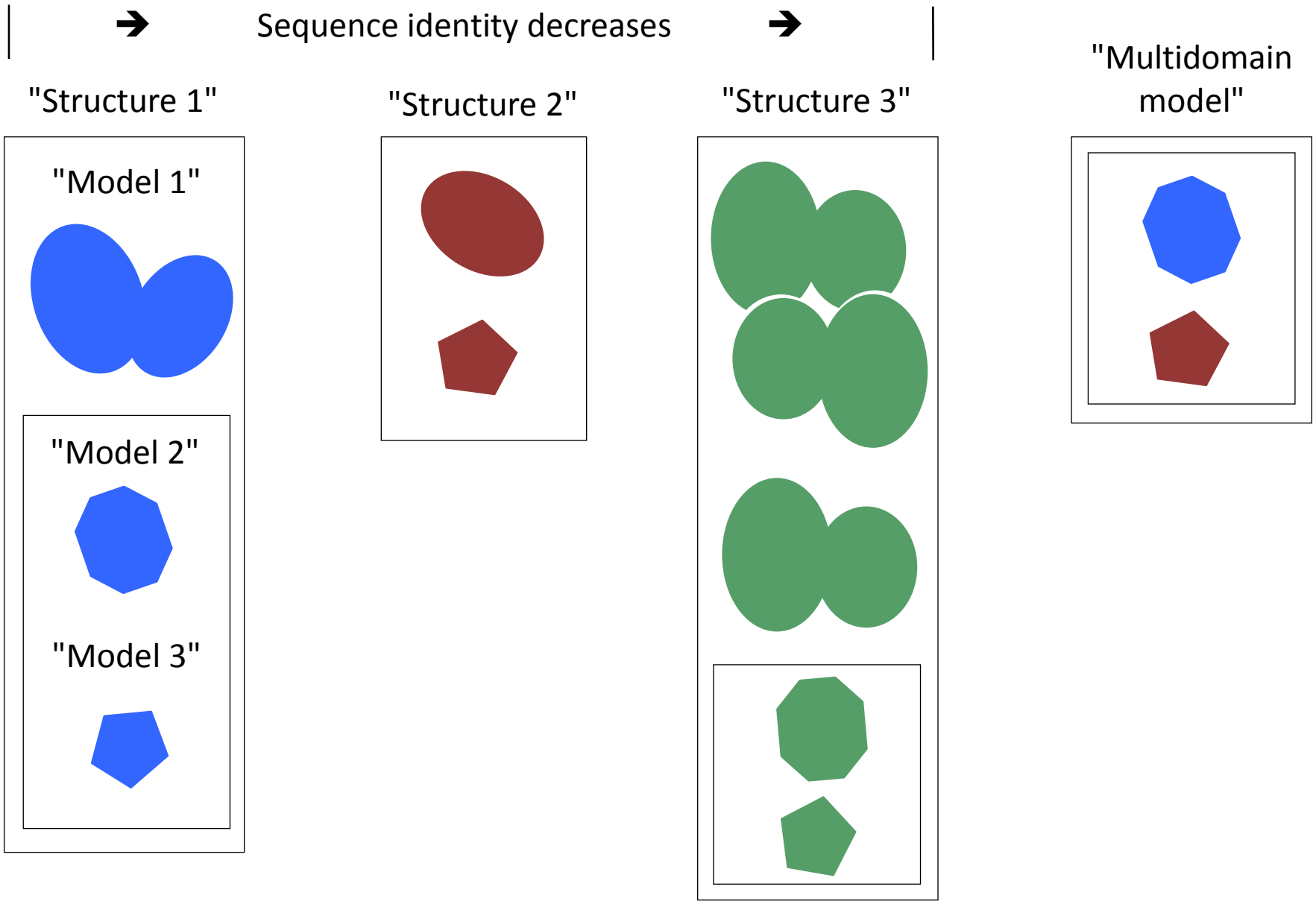




Multimeric models

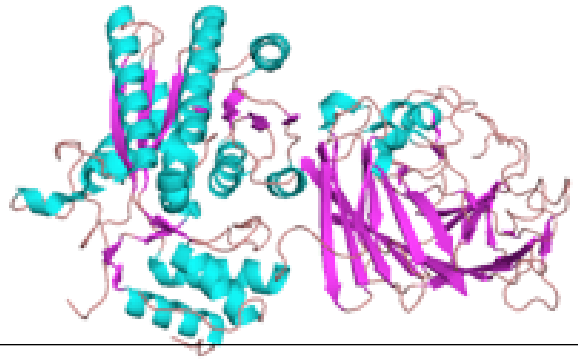
- All models are corrected
  - by sequence alignment
  - by accessible surface area
- Several input sequences mean protein-protein complex
  - hetero-multimeric models will be generated if possible

Search order



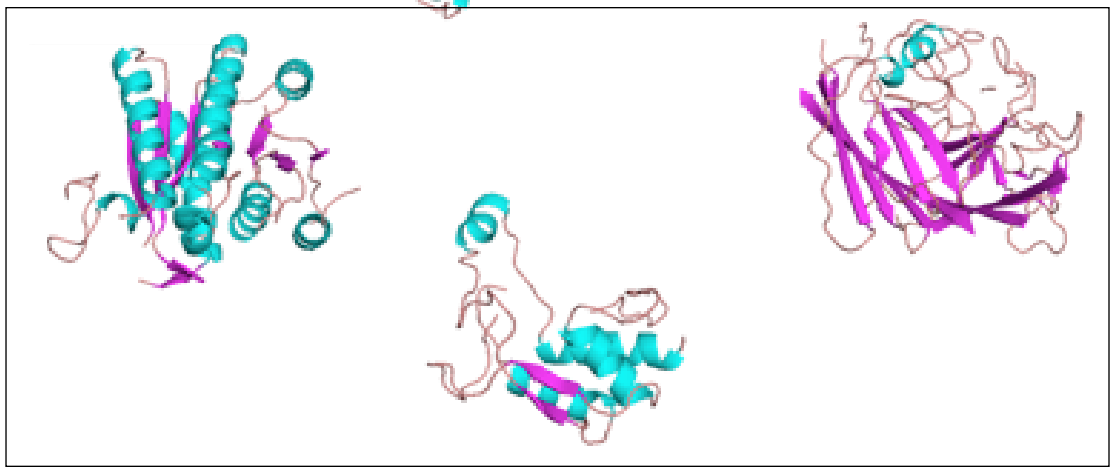
Multi-domain model (test case)

Target



1z45 - isomerase

Multi-domain model



1yga - domain of isomerase (51%)

1udc - two domains of isomerase (49%)

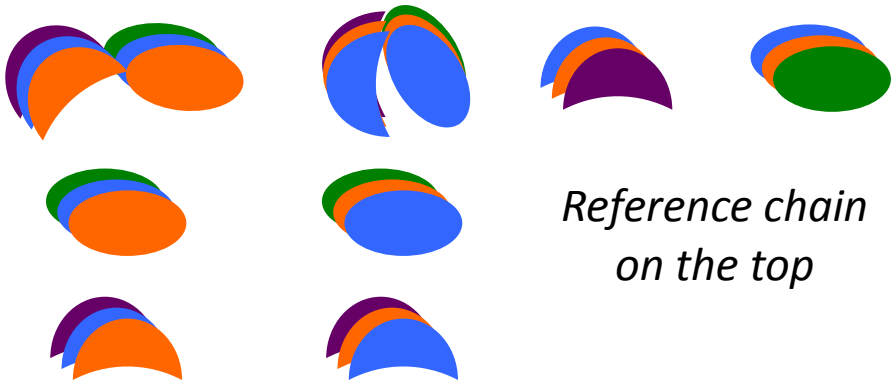
1ek6 - two domains of isomerase (55%)

Ensemble models

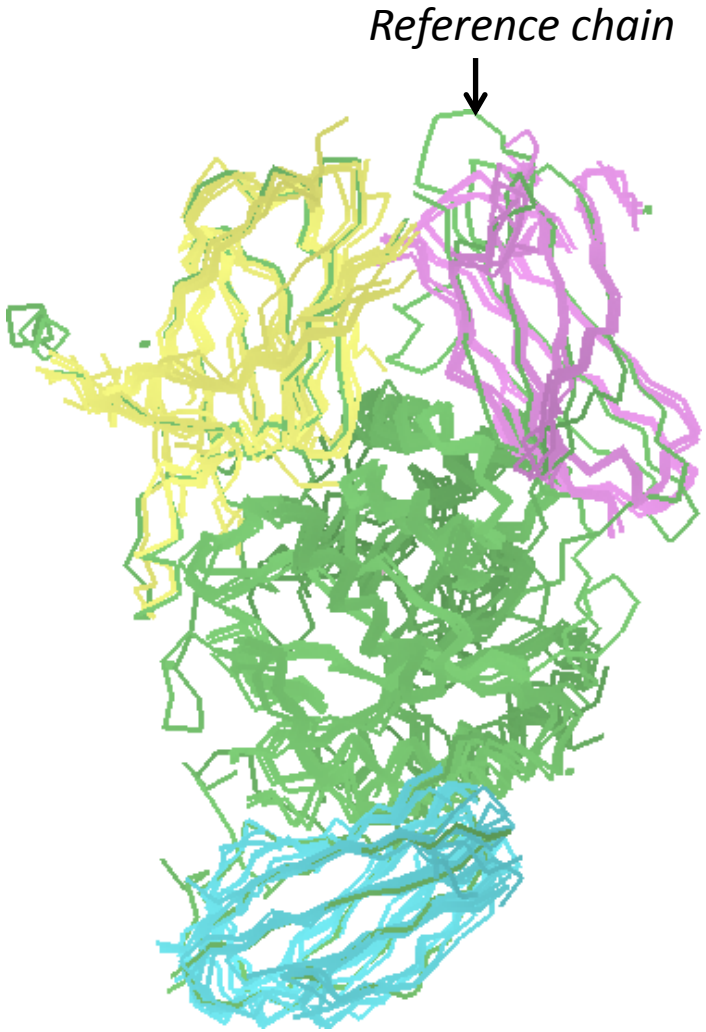
Ensemble models are generated if possible

Homologues from the Balbes database:

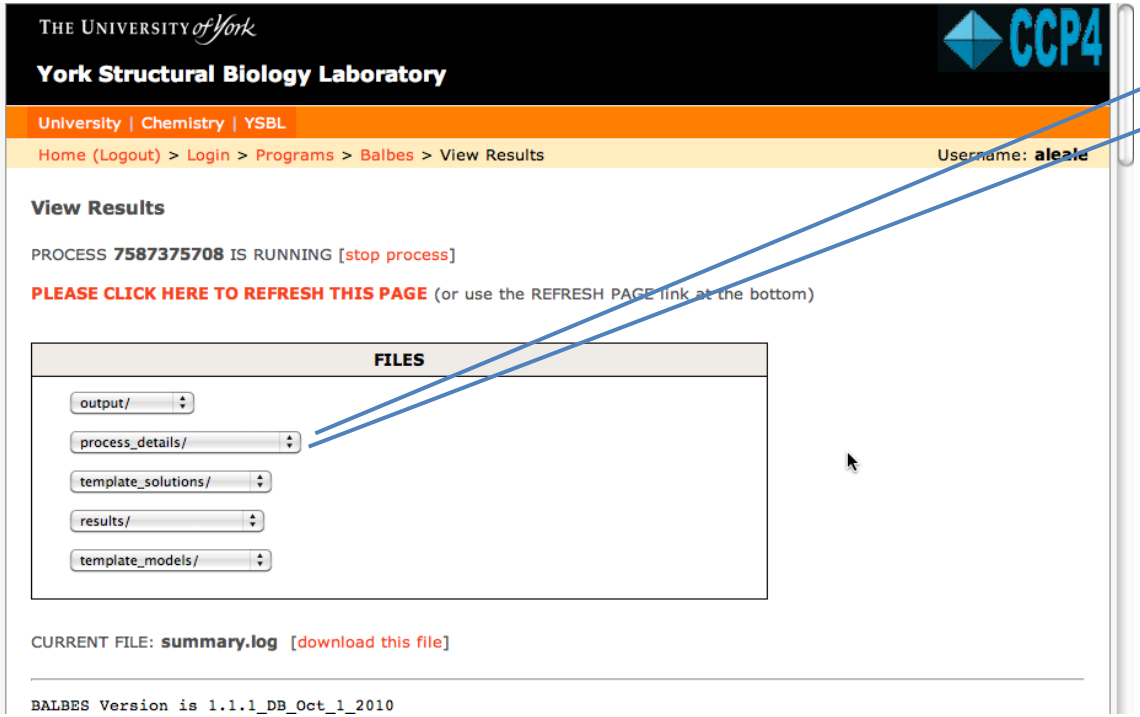
Ensemble search models:

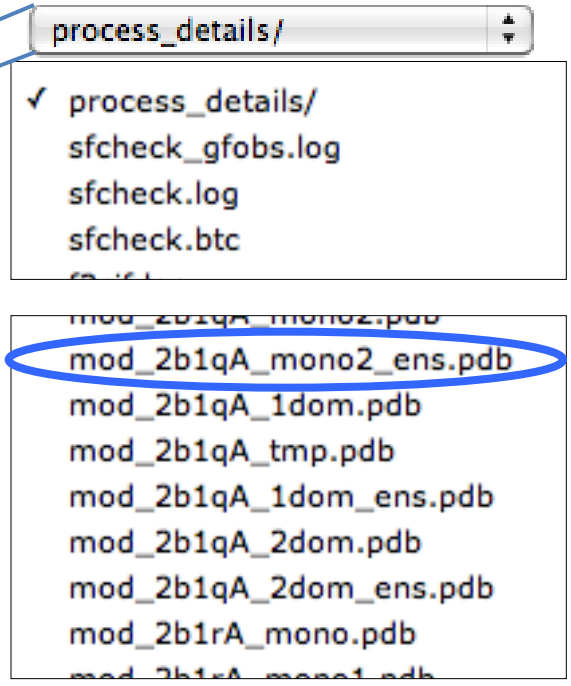


Reference chain on the top



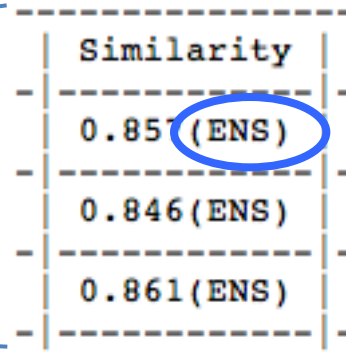
Ensemble models from Balbes





For the structure 1 in sequence 1

PDB_CODE	2b1qA					
NO. of MODEL	3					
Model	Chain ID	Similarity	Residues	Multimer?	Domain?	Monomers(expected)
1	A	0.857(ENS)	244	Monomer	No	1
2	A	0.846(ENS)	162	Monomer	Yes	1
3	A	0.861(ENS)	72	Monomer	Yes	1



AMPLE

***A**b initio* **M**odelling of **P**roteins for mo**LE**cular replacement

- Uses Rosetta to generate *ab initio* models
- Several new structures has been already solved

Jaclyn Bibby, Jens Thomas, Olga Mayans and Daniel Rigden

Institute of Integrative Biology

Ronan Keegan and Martyn Winn

Collaborative Computational Project 4 (CCP4)

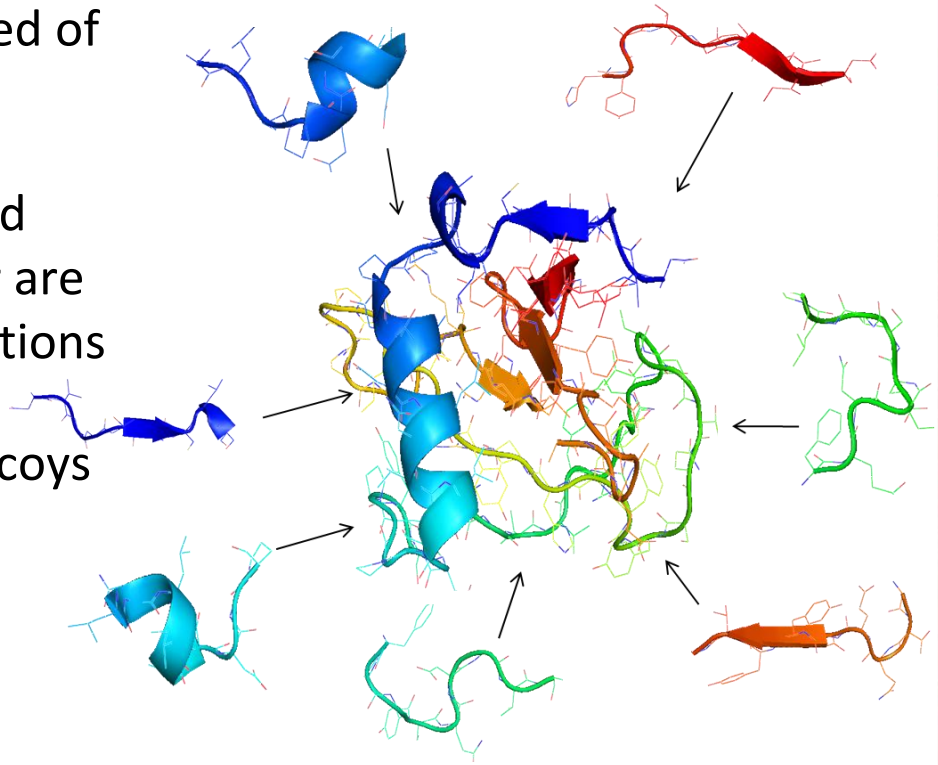
Ab initio structure prediction

Modest computing power is needed:

1. 1000’s of “Decoys” assembled of fragments from PDB structures

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

3. Side chains added to selected decoys



CCP4

Ab initio structure prediction

Modest computing power is needed:

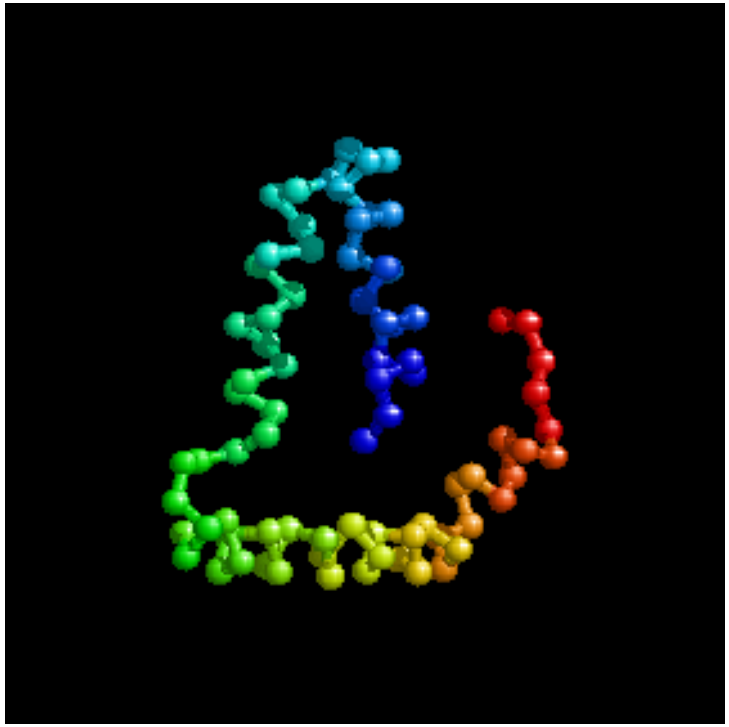
1. 1000’s of “Decoys” assembled of fragments from PDB structures

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

3. Side chains added to selected decoys

Supercomputing resources are needed:

4. Force field refinement



CCP4

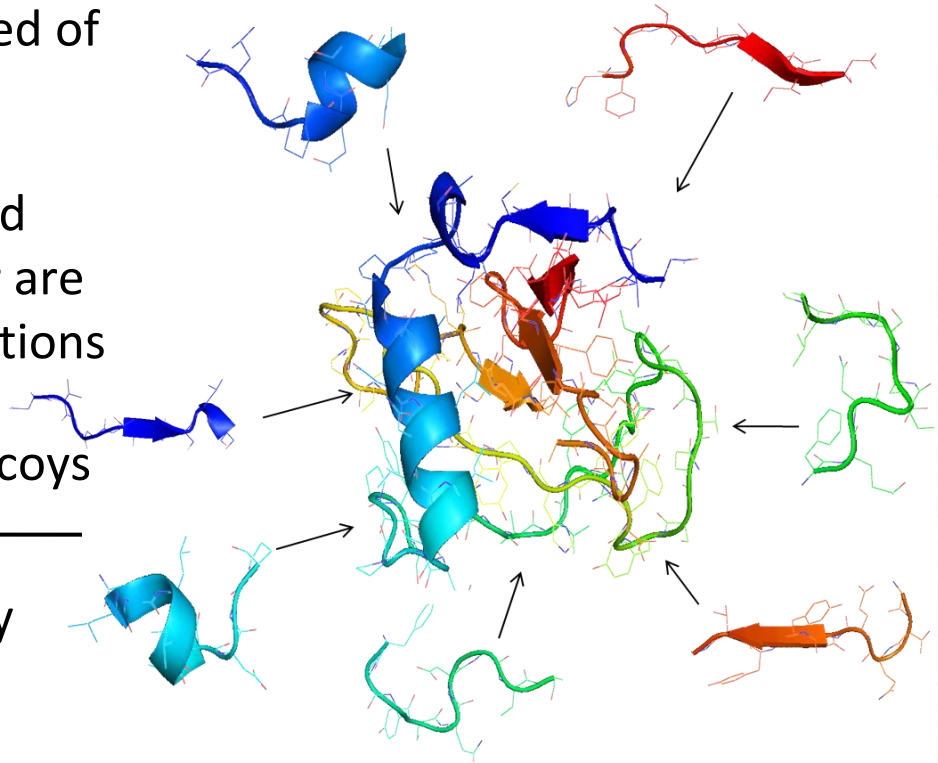
AMPLE

Modest computing power is needed:

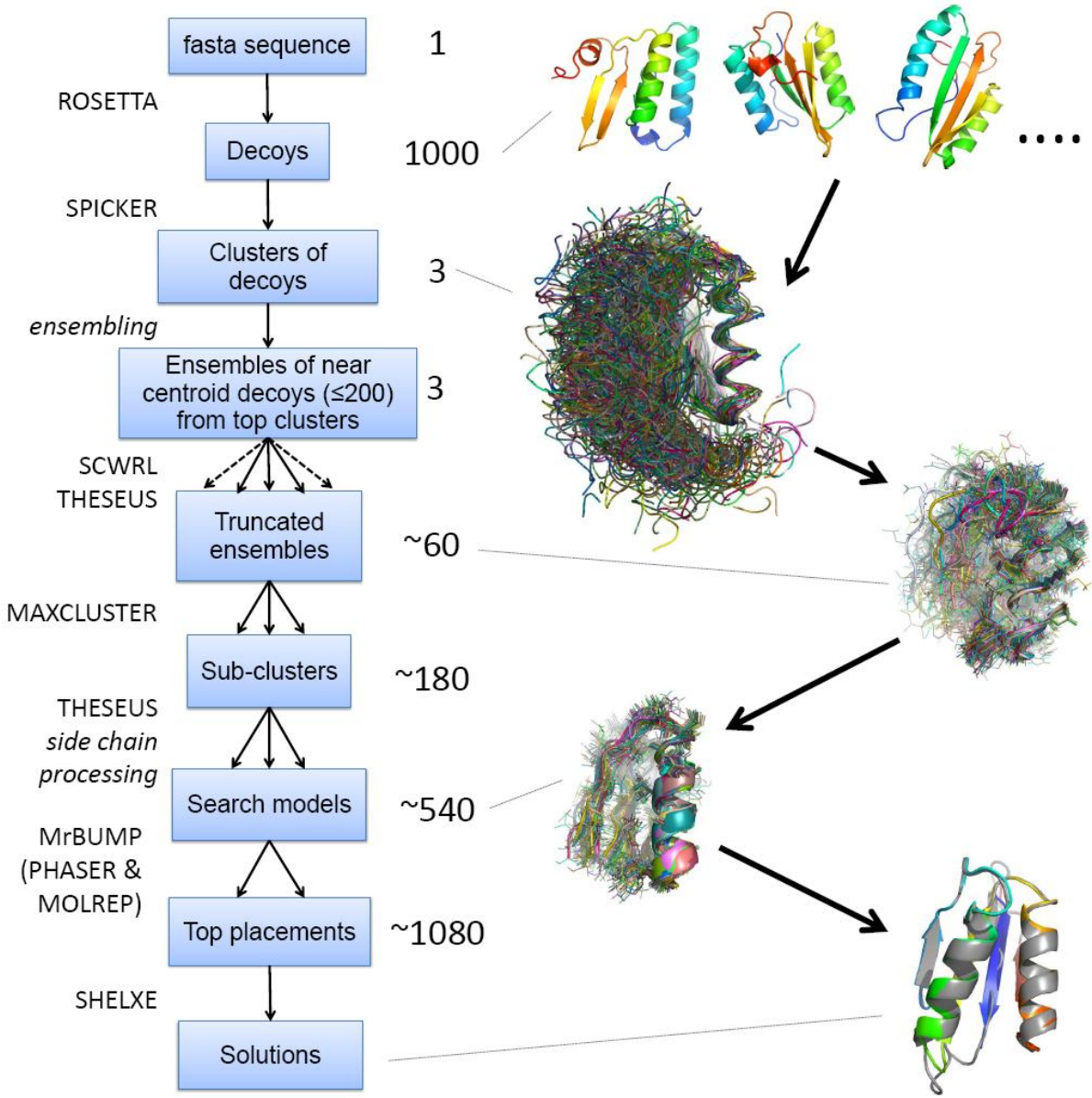
1. 1000's of "Decoys" assembled of fragments from PDB structures
2. Decoys are clustered and centroid representatives of largest cluster are considered candidate fold predictions
3. Side chains added to selected decoys

---

4. Ensembles generated from decoy models are used for Molecular Replacement



CCP4


fasta sequence
1
ROSETTA
Decoys
1000
SPICKER
Clusters of decoys
3
ensembling
Ensembles of near centroid decoys (≤200) from top clusters
3
SCWRL
THESEUS
Truncated ensembles
~60
MAXCLUSTER
Sub-clusters
~180
THESEUS
side chain processing
Search models
~540
MrBUMP
(PHASER &
MOLREP)
Top placements
~1080
SHELXE
Solutions

Assessing Solutions

Results of structure rebuilding as a criterion

- SHELXE C^{α} tracing
  - partial CC of >25%
  - average fragment length of 10 or more
- Further rebuilding including side chains
  - ARP/wARP
  - Buccaneer.

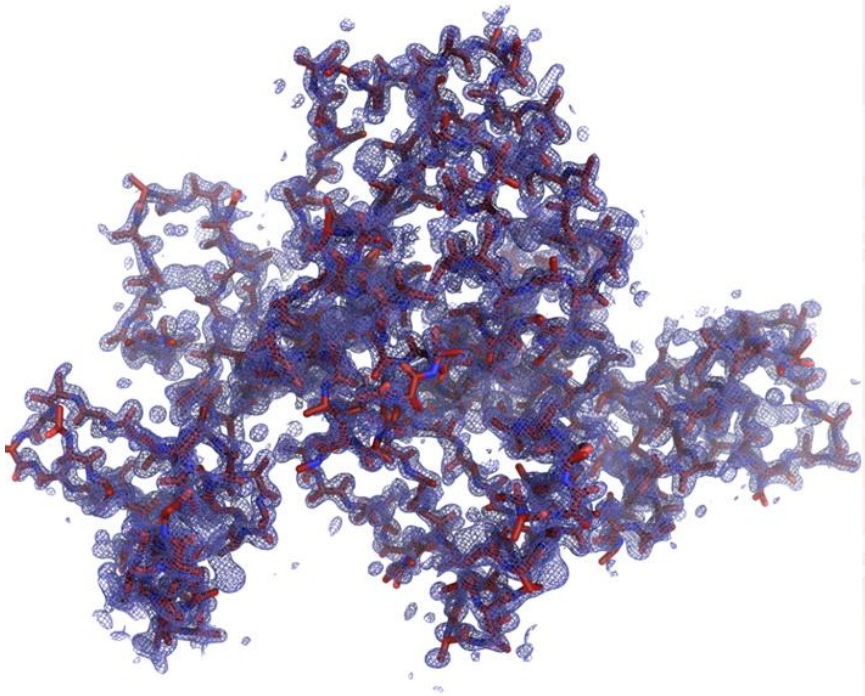
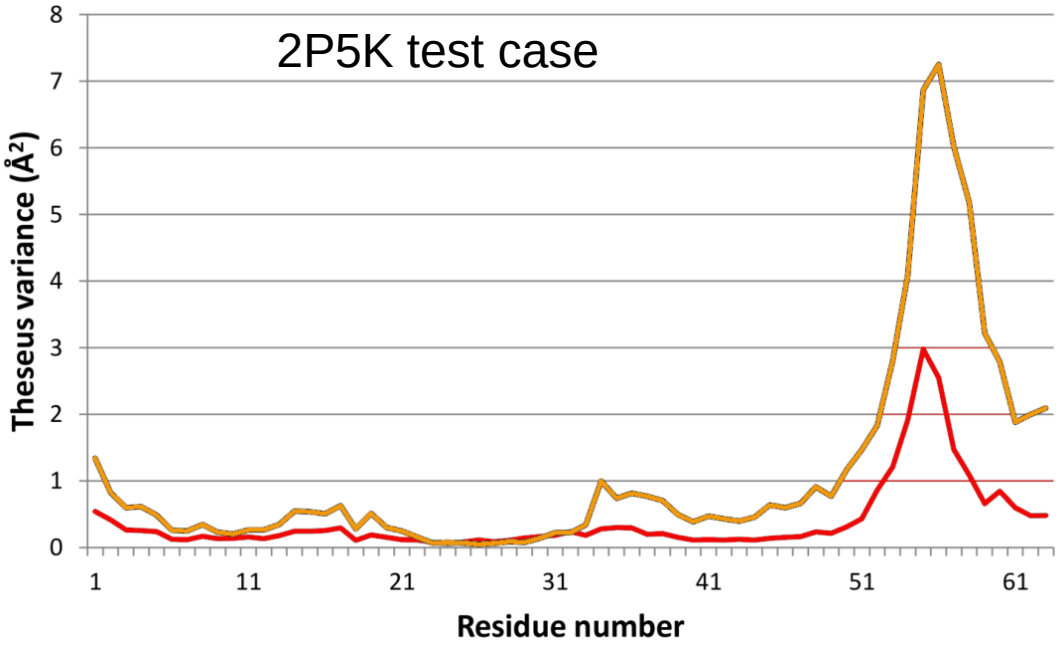


Figure courtesy of Andrea Thorn

Variance and Truncation

- Strong correlation between
  - deviation from the true structure
  - variability within decoy cluster



- Recipe for model generation:

  truncation at different ***Theseus*** variance levels

CCP4

Synergies

Ab initio modelling	Molecular Replacement
Produces **clusters** of similar model structures	Works effectively with superposed **ensembles**
Variance within the cluster **correlates with similarity to the target structure**	Works most effectively if unreliable parts of model are **truncated**

Ab initio tests

Test set of 295 small proteins from the PDB (40-120 residues)

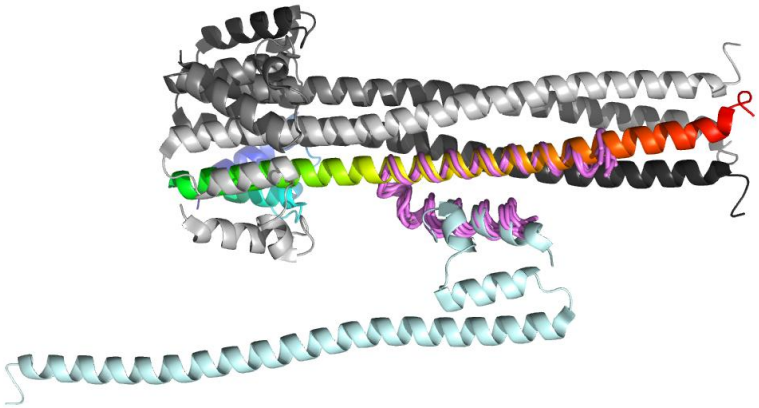
- Resolution of 2.2 Å or better
- Single molecule in the asymmetric unit
- Fragment generation step: homologues were excluded
- 1000 decoys generated for each case using Rosetta

Fold	all-α	all-β	mixed α-β	all cases
Success rate	80%	2%	37%	43%

All α-helical proteins are possible *ab-initio* targets

Coiled-coil proteins

- Difficult to solve in MR even with good homologues
- Initial tests: 80% success rate
- Some novel structures have also been solved



Transmembrane Proteins

- Very difficult to work with/crystallise
- 18 test cases of 1.45 - 2.5A resolution
- 7 clear and 5 possible successes
- 223 residue structure (3GD8) could be largest ever solved *ab initio*.

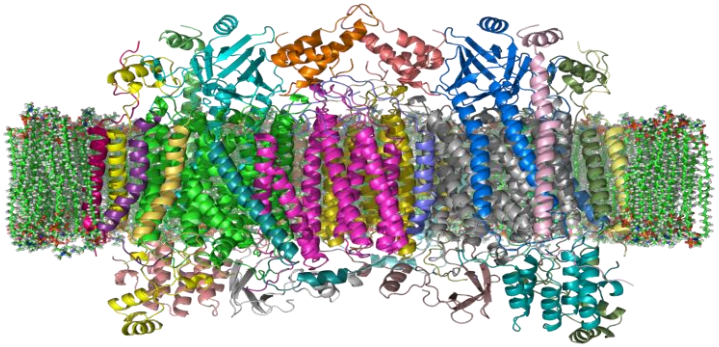
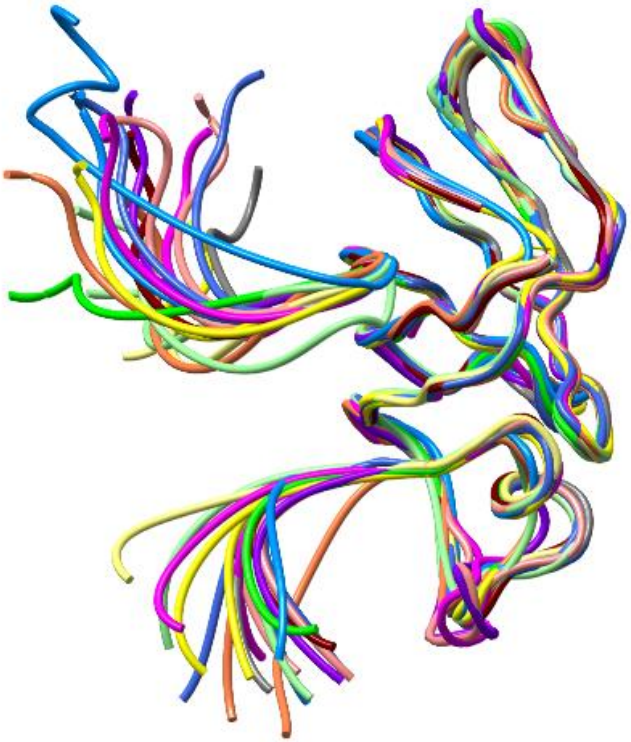


Image: http://en.wikiversity.org/wiki/File:Cytochrome_C_Oxidase_1OCC_in_Membrane_2.png

CCP4

General targets

AMPLE can use Rosetta to “re-model” a particular set of templates



- Remodelling related NMR structures
- Exploiting distant homologues

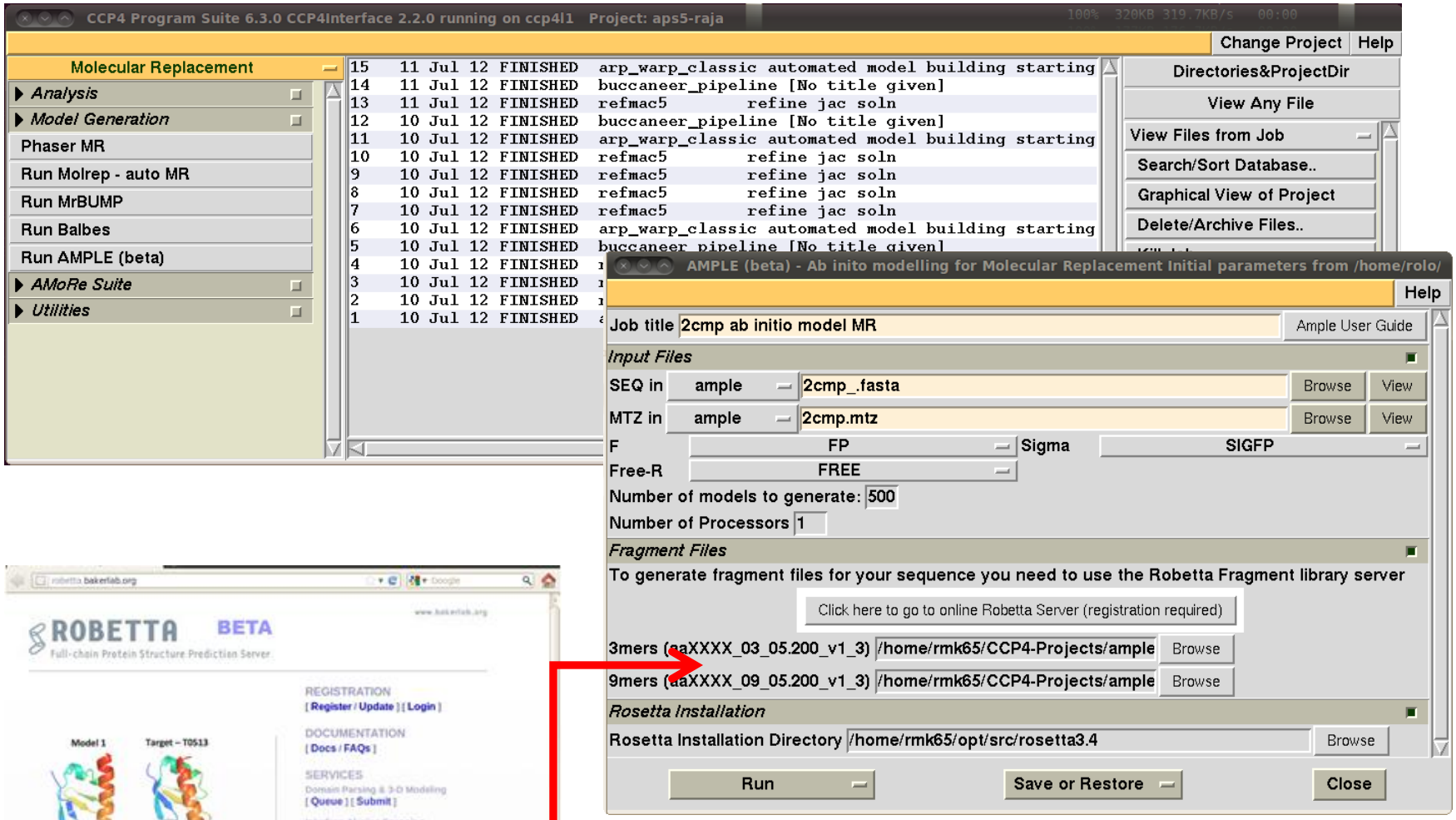
CCP4

AMPLE usage

- Complete run:2 CPU days
- A parallelised version on a cluster: 1 hour

Distributed as a part of CCP4 suite

- 6.3.0: beta version
- 6.4.0 (upcoming and installed here): improved and more robust version
- Currently required non-CCP4 packages:
  - Rosetta, SHELXE, Theseus, SPICKER, Maxcluster



Documentation available from

http://ccp4wiki.org

CCP4

Acknowledgements

AmoRe

Jorge Navaza — IBS Grenoble

Molrep

Alexey Vagin — University of York

Phaser

Randy Read — University of Cambridge

Airlie McCoy — University of Cambridge

Gabor Bunkozci — University of Cambridge

Acknowledgements

Balbes

Alexey Vagin	University of York
Fei Long	MRC LMB
Garib Murshudov	MRC LMB

MrBUMP

Ronan Keegan	STFC RAL
Martyn Winn	STFC DL

Ample

Ronan Keegan	STFC RAL
Jaclyn Bibby	University of Liverpool
Daniel Rigden	University of Liverpool
Olga Mayans	University of Liverpool