MR Pipelines in CCP4

Collaborative

Garib Murshudov, Fei Long, Alexei Vagin, Jaclyn Bibby, Olika Mayans, Daniel Rigden, Martyn Winn, Graeme Winter, Ronan Keegan, Martyn Daniel Rigden

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

21% of structures are solved by Experimental Phasing (EP).

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

- MrBUMP tries to mimic “best” practice of using external packages clustalw, mafft, fasta
- BALBES custom database optimised for MR
- AMPLE Ab initio model generation
- MIRBUMP bioinformatics tools

*external packages ROSETTA, QUARK, SPICKER
**external packages clustalw, mafft, fasta
Using the pipelines: ccp4i
MRBUMP – Keegan, Winn
Mr BUMP pipeline

Molecular Replacement Replacement

Target MTZ & Sequence

Template Preparation

Model Search

Details Target

Phase Improvement

Farm MR/Refinement jobs to cluster

the next model and exit or select

Check scores

Sequence Target MTZ

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

Sequence based search using sequence of target structure

Other templates from:

- SSM search using top hit from the FASTA search
- Can add additional PDB id codes to the list, e.g. from:
  - SSM or psiBLAST searches
  - Can add local PDB files

These structures are called model templates
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 – domains and multimers
MrBUMP – search models

Search models can be prepared in five ways:

1. **PDBclip**
   - Original PDB with waters removed, most probable conformations selected and format tidied (e.g., chain ID added).

2. **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.

3. **Chainsaw**
   - Similar to Chainsaw but different protocols for side chain truncation and alignment.
   - Chainsaw can be given any alignment between the template and the target sequence, non-conserved residues are pruned back to the gamma atom.

4. **Sculptor**
   - Similar to Chainsaw but different protocols for side chain truncation and alignment.

5. **Polyalanine**
   - Molrep contains a model preparation function which will align the template sequence with the target sequence and prune the non-conserved side chains according to excluding all of the side chain atoms beyond the CB atom using the Pdbset program.

Search models can be prepared in five ways:
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.ensembler to provide more ensemble search models.

Create ensembles of top.
Model template scoring:

score = sequence identity \times alignment quality

MrBUMP – multiple alignment

Jalview 2.0.8 Barton group, Dundee

MrBUMP – multiple alignment

Pairwise alignment

and sculptor

chainsaw

used in

alignments
MrBUMP – MR and Rigid Body Refinement

Running MR

For each search model, MR done with Molrep or Phaser or both. Molrep and Phaser programs run mostly with defaults.

 outcomes include:

<table>
<thead>
<tr>
<th>Best</th>
<th>&quot;good&quot;</th>
<th>Bad</th>
<th>&quot;poor&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Initial Rfree &gt; 0.52 and dropped by 5%</td>
<td>Final Rfree &lt; 0.35 or Final Rfree &lt; 0.5 and dropped by 20%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Initial Rfree &gt; 0.48 or Final Rfree &lt; 0.48</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>otherwise</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Refinement for restrained refinement: The resulting models from molecular replacement are passed to Refmac.

Rigid body scoring: Allow Molrep / Phaser to set resolution limits and weights.

Weights:
- initial Rfree < 0.5 and final Rfree > 0.48
- initial Rfree < 0.35 and final Rfree > 0.5
- otherwise

The change in the Rfree value during refinement is used as a rough estimate of how good the resulting model is.
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
- SHELEXE

NEW: Model building programs can be invoked post-MR to give a better assessment of whether or not the MR solution is correct.
Prepositioned models/complexes

MrBUMP will now accept one or more positioned models.

Phase extension

If resolution better than 1.7 Å use Acorn procedure

Artificial phase extension to 1.0 Å

Initial phase set from refined MR solution

These are included as fixed models in all MR jobs.

MrBUMP - other
- schematic

- database

- Manager

- Inputs

- Outputs

- programs
β
α
λ
β
ε
σ

Organises

Data Resolution for molecular replacement

Twinning

Pseudo translation

Sequence

Estimates number of molecules in the asymmetric unit

Finds template structures with their domain and multimer

"Corrects" template molecules using sequence alignment

Protocols

Runs various protocols with molecular replacement and

Refinement and makes decisions accordingly

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

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

Chains. The internal database has around 35000 unique entries organized according to:

- database
Model selection

Input sequence

Database of domains

Full chain model

Best multi-domain model
All models are corrected by sequence alignment and by accessible surface area.

**model preparation**
Example

Reference chain on the top

Ensemble search models:

Homologues from the Balbels database:

PolyGeo - ensemble models
Final decisions are made based on R-factors after refinement.

Since we have similar structures we can use them in refinement.

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

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

Space group

AMPLE – Bidy, Keegan
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
AMPLé – Monte Carlo sampling

Initial fragment assembly stage requires relatively modest computing power. Refinement stage can require supercomputing resources.

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

Side chains added to selected predictions. Refinement under a more realistic physics-based force field.

1000's of "Decoys" assembled of fragments from PDB structures.
AMPLE - synergies

ab initio modelling

Molecular Replacement

Produces clusters of similar structures superposed ensembles

Works effectively with approximating the target
AMPLE – the pipeline

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

Cluster into groups according to RMS deviation.

Generate large number of trial models.

Select templates from largest cluster.

THESEUS

SPICKER

ROSETTA

Cluster Centroid
AMPLE – pipeline in more detail

Refinement and model building

MR with MRBUMP (both Phaser and Molrep)

(x2)

(x3)

Add all side chains
Add reliable side chains
Leave as polyclanline

Recluster truncated models selecting the largest cluster (if necessary 2, 3A)

(x6 per target)

(x6 per target)

Truncate the most variable residues using six thresholds

Variance scores for each residue

ML Align with Theseus Gliding

Top cluster models from Rosetta or Gaurk

Spickerer

Quark models

Rosetta models

GD1

AMP LE – 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

Thanks to authors of all underlying programs

Alexei Vagin, University of York (retired)
Ronan Keegan, RAL
Marvin Wim, Daresbury Laboratory
Jacky Bibby, Daniel Rigden, Olga Mayans, University of Liverpool
Carib Mursudov, Fei Long, MRC-LMB