Automatic MR

Garib Murshudov
MRC-LMB, Cambridge

These slides provided by Ronan Keegan
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
MrBUMP

Ronan Keegan
(CCPC4 STFC RAL)

Martyn Winn
(STFC DL)
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

- **X-ray data (MTZ)**
- **Target sequence**
- **Template search**
- **Model preparation**
- **Molecular replacement & refinement**
- **Model building and Phase improvement**

**WEB resources**
(FASTA, PDB etc.)

**Programs available from CCP4**

**Farm MR/Refinement jobs to cluster**

**User specified models**

**Known partial structure**
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    » e.g. unpublished structures
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 enantiomorph 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

<table>
<thead>
<tr>
<th>Condition</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>final $R_{free} &lt; 0.35$ or</td>
<td>“good”</td>
</tr>
<tr>
<td>final $R_{free} &lt; 0.5$ and dropped by 20%</td>
<td></td>
</tr>
<tr>
<td>final $R_{free} &lt; 0.48$ or</td>
<td>“marginal”</td>
</tr>
<tr>
<td>final $R_{free} &lt; 0.52$ and dropped by 5%</td>
<td></td>
</tr>
<tr>
<td>otherwise</td>
<td>“poor”</td>
</tr>
</tbody>
</table>

A 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

- Search model types
  - chainsaw
  - molrep
  - plyala
  - sculptor
  - pdbclip

- MR details and files
  - mr
    - phaser
    - molrep

- Refinement details and files
  - refine
    - build

- Sequence alignment details
  - alignments
  - pdb_file
  - Model name
  - data
MrBUMP Testing Results

Model type for best solution

![Graph showing model type for best solution](image-url)
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

1. X-ray data (MTZ)
2. Target sequence
3. Template search
4. Model preparation
5. Molecular replacement & refinement
6. Molecular replacement & refinement
7. BALBES DB and associated programs
8. Possible solution
9. Model building at ARP/wARP server
10. Possible solution

search models
• 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

- 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

Domain models

链模型

Multimeric models

MARPVPEETVATERVKE

MARPVPEETVATERVKE
Search order

Sequence identity decreases

"Structure 1"

"Model 1"

"Structure 2"

"Model 2"

"Structure 3"

"Model 3"

"Multidomain model"
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 are generated if possible

Homologues from the Balbes database:

Ensemble search models:

Reference chain on the top
Ensemble models from Balbes

View Results

Please click here to refresh this page (or use the refresh page link at the bottom)

CURRENT FILE: summary.log [download this file]

BALSIE Version is 1.1.1_DB_Oct_1_2010

For the structure 1 in sequence 1

<table>
<thead>
<tr>
<th>Model</th>
<th>Chain ID</th>
<th>Similarity</th>
<th>Residues</th>
<th>Multimer?</th>
<th>Domain?</th>
<th>Monomers (expected)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td>0.857 (ENS)</td>
<td>244</td>
<td>Nonomer</td>
<td>No</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>A</td>
<td>0.846 (ENS)</td>
<td>162</td>
<td>Nonomer</td>
<td>Yes</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>A</td>
<td>0.861 (ENS)</td>
<td>72</td>
<td>Nonomer</td>
<td>Yes</td>
<td>1</td>
</tr>
</tbody>
</table>

Similarity

- 0.85 (ENS)
- 0.846 (ENS)
- 0.861 (ENS)
AMPLE

Ab initio Modelling of Proteins for molecular 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
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
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
Assessing Solutions

Results of structure rebuilding as a criterion

• SHELXE Cα 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
## Synergies

<table>
<thead>
<tr>
<th>Ab initio modelling</th>
<th>Molecular Replacement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Produces <strong>clusters</strong> of similar model structures</td>
<td>Works effectively with superposed <strong>ensembles</strong></td>
</tr>
<tr>
<td><strong>Variance</strong> within the cluster correlates with similarity to the target structure</td>
<td>Works most effectively if unreliable parts of model are <strong>truncated</strong></td>
</tr>
</tbody>
</table>
**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

<table>
<thead>
<tr>
<th>Fold</th>
<th>all-α</th>
<th>all-β</th>
<th>mixed α-β</th>
<th>all cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Success rate</td>
<td>80%</td>
<td>2%</td>
<td>37%</td>
<td>43%</td>
</tr>
</tbody>
</table>
All $\alpha$-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]
AMPLE can use Rosetta to “re-model” a particular set of templates

• Remodelling related NMR structures

• Exploiting distant homologues
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
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