Phaser
Molecular replacement
Gábor Bunkóczi

CCP4 Fukuoka School,
30 October 2012
Accounting for model error

Position errors

Calculate from sequence identity (Chothia & Lesk)

Model incompleteness

Calculate from unit cell composition
Likelihood function

$F_c$ – model structure factor
Dependent on orientation and position of model

$D$ – Luzzati-factor ($0 \leq D \leq 1$)
Dependent on model errors

$\sigma_\Delta$ – uncertainty
Dependent on model errors and incompleteness
Reflection likelihood

$$DF_c(R_1, r_1)$$

$$F_{obs}$$
Reflection likelihood

\[ DF_c(R_2, r_2) F_{obs} \]
Reflection likelihood

\[ F_{\text{obs}} \]

\[ DF_c(R_3, r_3) \]
Reflection amplitude likelihood
Rotation and translation functions

Identical mathematical forms!

Translation function:
Reflection amplitude likelihood (for all reflections)

Rotation function:
Reflection amplitude likelihood with $\sigma_\Delta$ is increased to account for unknown translation
Molecular replacement workflow

Rotation search → Translation search → Refinement → Packing test
Identifying solutions
Multiple component searches

Sequential approach

Molecular Replacement

Component 1

Component 2

Signal increase for subsequent components!
Ordering the search
Ordering the search
Search order determination

65% of structure
35% identical model

35% of structure
50% identical model

Which model is easier to find?
Search order determination

A better model explains more of the data

Better model if data resolution is low

65% of structure
35% identical model

Better model if data resolution is high

35% of structure
50% identical model
Executing sequential searches

1\textsuperscript{st} component

2\textsuperscript{nd} component
Sculptor: model improvement

- **main chain**
  Delete residues from the model that are not present in the target (e.g. extra domains)

- **side chain**
  - Prune side chain back to $C_\gamma$ if residues are different
  - Add $C_\beta$ to GLY that is not GLY in the target

- **B-factors**
  - weight up segments where sequence is highly conserved (sequence similarity)
  - weight down regions with structural flexibility (solvent accessible surface area)
Sequence similarity

weighting

matrix

LGHQH--GS----RYVM

FRKCTNNASKYLCYL

window window
B-factor weighting

Sequence similarity (residue level)

Accessible surface area (atom level)
Solve tetragonal lysozyme structure

Tetragonal lysozyme model
• Original B-factors: LLG = 806.9
• Modified B-factors: LLG = 592.02

Monoclinic lysozyme model
• Original B-factors: LLG = 328.3
• Modified B-factors: LLG = 330.6
Sculptor: model improvement

Calculations with CLUSTALW alignments

Model
Default

Model
SS mainchain

Model
Default
+ASA B-factors

Model
Default
+sequence similarity B-factors

+sequence similarity B-factors

+ ASA B-factors

sequence identity

0.15-0.20
0.20-0.25
0.25-0.30
0.30-0.35
0.35-0.40
0.40-0.54

number of models

unsolved
multi-model
single model
## Protocols

**Testcase:** 1HJ9, 1 molecule/ASU, 32% identical

<table>
<thead>
<tr>
<th>1.   2B9L_8.pdb</th>
<th>TFZ=7.3</th>
<th>LLG=41.731</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.   2B9L_10.pdb</td>
<td>TFZ=7.1</td>
<td>LLG=40.661</td>
</tr>
<tr>
<td>3.   2B9L_12.pdb</td>
<td>TFZ=7.6</td>
<td>LLG=40.644</td>
</tr>
<tr>
<td>4.   2B9L_7.pdb</td>
<td>TFZ=6.8</td>
<td>LLG=38.864</td>
</tr>
<tr>
<td>5.   2B9L_11.pdb</td>
<td>TFZ=7.3</td>
<td>LLG=38.582</td>
</tr>
<tr>
<td>6.   2B9L_9.pdb</td>
<td>TFZ=6.7</td>
<td>LLG=38.419</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.   2B9L_4.pdb</td>
<td>TFZ=6.2</td>
<td>LLG=26.755</td>
</tr>
<tr>
<td>11.  2B9L_1.pdb</td>
<td>TFZ=5.2</td>
<td>LLG=23.545</td>
</tr>
</tbody>
</table>
Ensembler: multiple superposition
Ensembler: workflow

• Find protein chains (discard solvent)
• Align residues
  – SSM
  – MUSCLE
  – Alignments
• Match atoms (by name)
• Superpose
  – gapless: use atoms present in all structures
  – gapped: use sites present in at least two structures
Ensembler: weighting

- Robust-resistant weighting to superpose structurally similar regions
- Less sensitive to alignment errors

unweighted

weighted
Ensembler: trimming

- based on unweighted positional rmsd
- superposed positions only
Molecular replacement pipeline
Goals

- Solve easy cases quickly
- Solve difficult cases
- Simple user interface
  Minimal input: sequence, data
- Sufficient customisation
- Efficient use of computational resources
  - Workstations and clusters
  - Scale with search complexity
- Optimise for large number of models
Molecular replacement

\[(\phi, \psi, \kappa) \rightarrow (x, y, z) \rightarrow \text{Model} \rightarrow \text{Solution} \]

After a correct solution with a high Z-score has been found.
Superposition

\[ (\phi, \psi, \kappa) + (x, y, z) \]

superposition with alternative model (SSM)

Find best
Amalgamation

Incomplete structure \( \rightarrow \) TFZ $\geq 7.0$ \( \rightarrow \) Assembled structure

TFZ $\geq 7.0$
Assembly recognition

• identify (local) point group symmetry
• use full assembly in subsequent search
  better signal
• refine as set of individual molecules
  correct for local differences
## Assembly recognition

**Testcase:** Shiga-like toxin, 1 model, 4 pentamers/ASU

<table>
<thead>
<tr>
<th>Index</th>
<th>Model</th>
<th>TFZ</th>
<th>LLG</th>
<th>ΔLLG</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>monomer</td>
<td>5.3</td>
<td>43.1</td>
<td>43.1</td>
</tr>
<tr>
<td>2</td>
<td>monomer</td>
<td>11.9</td>
<td>154.7</td>
<td>111.6</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>monomer</td>
<td>22.3</td>
<td>2005.9</td>
<td>293.3</td>
</tr>
<tr>
<td>11</td>
<td>5*monomer</td>
<td><strong>42.6</strong></td>
<td><strong>3889.3</strong></td>
<td><strong>1883.4</strong></td>
</tr>
<tr>
<td>16</td>
<td>monomer</td>
<td>33.5</td>
<td>4545.4</td>
<td>656.1</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>monomer</td>
<td>38.1</td>
<td>6557.1</td>
<td>673.9</td>
</tr>
<tr>
<td>20</td>
<td>monomer</td>
<td>9.2</td>
<td>7322.9</td>
<td>765.8</td>
</tr>
</tbody>
</table>
Assembly completion

Incomplete assembly → Predicted position → Completed assembly

Known assembly → calculate TFZ and LLG
Assembly completion

Incomplete assembly

Predicted position (approximately correct)

Completed assembly
calculate TFZ and LLG

Known assembly
Local search

Large noise peaks
Fully automated solution

- 750 residues, 1 chain/asan
- 2.95 Å, I 4₁ 2 2
- Solved by SIRAS
- 82 HHpred hits

3LN6
Fully automated solution

Input

component
{
    sequence=3ln6.seq

    homology
    {
        file_name=3ln6.hhr
    }
}

After 5 minutes (50 CPUs)

Significant hit:
3NZT_9, TFZ=10.0, LLG=80

Model is only partial!
Sequence-based assembly

MIIDRLQLRSHSHLPILQATFGLERESLRIHQPTQRVAQTPHPKTLGSRNYHPYIQTQDSEPQLE

1. hit: 

2. hit: 

\[ \begin{array}{c}
\text{negligible overlap}
\end{array} \]

N. hit: 

\[ \begin{array}{c}
\text{negligible model}
\end{array} \]
Fully automated solution

After one day (50 CPUs)

Solution:
3NZT_9, LLG=80
1UC8_11, TFZ=6.4, LLG=143

Processed:
82 templates
845 models
...
1 000 000+ translation peaks
Fully automated solution

Solution

After AutoBuild

431 built, 92 placed
R=0.4058/0.4740
Acknowledgements

**PHENIX:**
- Paul Adams
- Nat Echols

**CIMR:**
- Stephen Graham
- Nikol Simecek

**Industrial Consortium:**
- Joel Bard
- Herb Klei
- Bob Nolte
- Steven Sheriff

**CCP4 team**
Acknowledgements

Python-based Hierarchical Environment for Integrated Xtallography

Computational Crystallography Initiative / LBNL
Paul Adams, Ralf Grosse-Kunstleve, Pavel Afonine, Nathaniel Echols, Jeffrey Headd, Nigel Moriarty, Nicholas Sauter, Peter Zwart

Los Alamos National Lab
Tom Terwilliger, Li-Wei Hung

UNIVERSITY OF CAMBRIDGE
Randy Read, Airlie McCoy, Robert Oeffner

DUKE University
Jane and Dave Richardson, Vincent Chen