RESOLVE model-building

Tom Terwilliger
Los Alamos National Laboratory
RESOLVE model-building at moderate resolution

• FFT-based identification of helices and strands
• Extension with tripeptide libraries
• Probabilistic sequence alignment
• Automatic molecular assembly
Placement of helical and extended templates

- Identify locations with FFT-based convolution search
- Maximize CC of template with map
- Superimpose each fragment in corresponding library (helix,sheet) on template
- Identify longest segment in good density, score = $\langle\text{density}\rangle \times \sqrt{\text{Natoms}}$
Initial model-building – strand fragments
Chain extension by placement of tripeptide fragments

• Look-ahead scoring: find fragment that can itself be optimally extended

• C-terminal extension. Start at C-terminus of protein

• Each of 10000 fragments: superimpose CA C O on same atoms of last residue in chain (extending by 2 residues): pick best 10

• Each of best 10: extend again by 2 residues and pick best 1; score for 2-residue extension = best <density> for 4-residue extension based on this 2-residue extension

• N-terminal: same, but going in opposite direction
Chain extension
(result: many overlapping fragments)
Assembly of main-chain

• Choose highest-scoring fragment

• Test all overlapping fragments as possible extensions

• Choose one that maximizes score when put together with current fragment

• When current fragment cannot be extended: remove all overlapping fragments, choose best remaining one, and repeat
Main-chain as a series of fragments
(choosing the best fragment at each location)
Side-chain rotamer templates

- Define side-chain orientation based on N CA C of main-chain
- Up to 40 rotamers per side chain

- Create template from average calculated electron density based on all occurrences of rotamer in 637 unique proteins

- Total of 400 side-chain templates
Scoring side-chain templates at each position

- Identify side-chain orientation from N CA C of main-chain
- Get CC of template with density -> Z-score
- (Compare CC with mean, SD of all side chain density with this template)
- \( P(\text{this side-chain/rotamer is correct}) = P(\text{this side-chain/rotamer}) \times P(Z) \)
Evaluating which side-chain template is best matched by a pattern of density:
*A good match to a glycine means more than a good match to an alanine*
Side-chain template matching to identify sequence alignment to map (IF5A data)
Relative probability for each amino acid at each position
(Correct amino acids in bold)

|   | G | A | S | V | I | L | M | C | F | Y | K | R | W | H | E | D | Q | N | P | T |
| 1 |  6 |  5 |  4 | 18 | 18 |  6 |  1 |  1 |  1 |  2 |  6 |  2 |  2 |  1 |  9 |  6 |  1 |  0 |  1 |  4 |
| 2 |  4 | 11 | 14 |  5 |  37 |  5 |  2 |  0 |  2 |  0 |  0 |  2 |  3 |  0 |  0 |  1 |  2 |  0 |  0 |  0 |  6 |
| 3 | 11 | 23 |  5 |  12 |  5 |  5 |  2 |  0 |  1 |  3 |  7 |  3 |  1 |  0 |  5 |  3 |  2 |  0 |  2 |  2 |
| 4 |  7 |  9 |  6 |  16 |  8 |  5 |  2 |  0 |  1 |  3 |  8 |  4 |  1 |  0 |  7 |  6 |  2 |  0 |  3 |  4 |
| 5 | 31 |  7 |  3 |  7 |  4 |  2 |  1 |  0 |  1 |  3 |  5 |  4 |  1 |  0 |  6 |  2 |  2 |  0 | 11 |  1 |
| 6 |  1 |  3 |  3 |  41 | 14 |  8 |  0 |  0 |  0 |  0 |  2 |  1 |  0 |  0 |  2 |  4 |  0 |  0 |  1 |  9 |
| 7 |  0 |  0 |  0 |  0 |  0 |  0 |  0 |  0 |  15 | 63 |  1 |  0 |  17 |  1 |  0 |  0 |  0 |  0 |  0 |  0 |
| 8 |  2 |  3 |  6 | 23 | 10 |  6 |  2 |  1 |  0 |  1 |  4 |  3 |  0 |  0 |  5 | 16 |  1 |  0 |  1 |  6 |
| 9 |  96 |  0 |  0 |  0 |  0 |  0 |  0 |  0 |  0 |  0 |  0 |  0 |  0 |  0 |  0 |  0 |  0 |  0 |  0 |  0 |
Addition of side-chains to fixed main-chain positions
Accuracy of side-chain identification probabilities

Predicted percentage correct

Actual fraction correct
Accuracy of sequence alignment probabilities

![Graph showing accuracy of sequence alignment probabilities](image-url)
Model-building vs resolution for nearly-perfect data (IF5A)
Automated NCS identification with RESOLVE

• Expand heavy-atom sites within radius R of origin
• Make list of all pairs of sites, sorted by distance between sites d

• Choose any 3 HA sites – a triangle ABC
• Find all other sets of 3 HA sites that form the same triangle
  • If some exist (DEF) -> this might correspond to NCS
  • If none…try another set of 3 HA sites

• Testing NCS: Sites ABC match sites DEF
• Does density near ABC match (after rotation/translation) density near DEF?
### Automated NCS identification with RESOLVE

<table>
<thead>
<tr>
<th>Structure</th>
<th>Number of sites found by SOLVE</th>
<th>NCS</th>
<th>NCS (found from heavy-atom sites)</th>
<th>NCS (electron-density map)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NDP Kinase</td>
<td>9</td>
<td>3-fold</td>
<td>3-fold</td>
<td>3-fold</td>
</tr>
<tr>
<td>Hypothetical</td>
<td>16</td>
<td>2-fold</td>
<td>2-fold</td>
<td>2-fold</td>
</tr>
<tr>
<td>Red Fluorescent Protein</td>
<td>26</td>
<td>4 copies</td>
<td>4 copies</td>
<td>4 copies</td>
</tr>
<tr>
<td>AEP Transaminase</td>
<td>66</td>
<td>6 copies</td>
<td>6 copies</td>
<td>6 copies</td>
</tr>
<tr>
<td>Formate dehydrogenase</td>
<td>12</td>
<td>2-fold</td>
<td>2-fold*</td>
<td>2-fold</td>
</tr>
<tr>
<td>Gene 5 protein</td>
<td>2</td>
<td>None</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>Armadillo repeat from β-catenin</td>
<td>15</td>
<td>None</td>
<td>2 copies</td>
<td>None</td>
</tr>
<tr>
<td>Dehalogenase</td>
<td>13</td>
<td>None</td>
<td>3 copies</td>
<td>None</td>
</tr>
<tr>
<td>Initiation Factor 5A</td>
<td>4</td>
<td>None</td>
<td>None</td>
<td>None</td>
</tr>
</tbody>
</table>
Molecular assembly in RESOLVE

List all chains assigned to sequence (anywhere in space)

A possible arrangement consists of:

• Each chain assigned to a molecule
• Each chain assigned to a symmetry-related position

Score a possible arrangement based on:

• Plausibility of gap distances between position of C of residue i and N of residue j
• RMS distance of chains from molecular center
• RMSD of NCS symmetry for corresponding atoms

• Try a reasonable starting arrangement (each chain assigned to the center of an NCS copy)
• Adjust by moving chains and groups of chains randomly from one symmetry-related position to another. Choose based on score.
Molecular assembly in RESOLVE

Summary of molecular assembly results (NDP-kinase)

NCS copies: 3

Molecule: 1  Chain: 1  Score for molecular location: 0.83

<table>
<thead>
<tr>
<th>Frag</th>
<th>Start</th>
<th>End</th>
<th>N</th>
<th>Overlap</th>
<th>Link</th>
<th>Mol</th>
<th>NCS</th>
<th>NCS</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>17</td>
<td>64</td>
<td>48</td>
<td>0</td>
<td>6.6</td>
<td>4.5</td>
<td>0.7</td>
<td>31.0</td>
<td>51.0</td>
</tr>
<tr>
<td>2</td>
<td>69</td>
<td>74</td>
<td>6</td>
<td>0</td>
<td>24.5</td>
<td>19.6</td>
<td>0.5</td>
<td>3.0</td>
<td>3.7</td>
</tr>
<tr>
<td>3</td>
<td>115</td>
<td>137</td>
<td>23</td>
<td>0</td>
<td>14.4</td>
<td>5.2</td>
<td>0.8</td>
<td>20.5</td>
<td>22.7</td>
</tr>
<tr>
<td>4</td>
<td>166</td>
<td>186</td>
<td>21</td>
<td>0</td>
<td>5.2</td>
<td></td>
<td>0.6</td>
<td>9.5</td>
<td>22.4</td>
</tr>
</tbody>
</table>

Residues placed for this molecule: 98

Total residues placed: 309 of 588 or 52%
Residues built without side chains: 65
Total residues built: 374 or 63%

Total score for this arrangement: 314.4
Initial automated structure solution, density modification, NCS-identification, and model-building

<table>
<thead>
<tr>
<th>Structure</th>
<th>Res. (Å)</th>
<th>% of main-chain built</th>
<th>% of side chains built</th>
</tr>
</thead>
<tbody>
<tr>
<td>Granulocyte stimulating factor (Rozwarski et al., 1996)</td>
<td>3.5</td>
<td>50%</td>
<td>0%</td>
</tr>
<tr>
<td>β-catenin (Huber et al., 1997)</td>
<td>2.7</td>
<td>81%</td>
<td>62%</td>
</tr>
<tr>
<td>Gene 5 protein</td>
<td>2.6</td>
<td>61%</td>
<td>11%</td>
</tr>
<tr>
<td>NDP Kinase (Pédelacq et al, 2002)</td>
<td>2.6</td>
<td>56%</td>
<td>37%</td>
</tr>
<tr>
<td>Hypothetical (P. aerophilum ORF, NCBI accession number AAL64711)</td>
<td>2.6</td>
<td>79%</td>
<td>75%</td>
</tr>
<tr>
<td>2-Aminoethylphosphonate (AEP) Transaminase (Chen et al., 2000)</td>
<td>2.6</td>
<td>85%</td>
<td>81%</td>
</tr>
<tr>
<td>Red Fluorescent Protein (Yarbrough et al, 2001)</td>
<td>2.5</td>
<td>88%</td>
<td>88%</td>
</tr>
<tr>
<td>Initiation factor 5A (Peat et al., 1998)</td>
<td>2.1</td>
<td>84%</td>
<td>84%</td>
</tr>
</tbody>
</table>
The PHENIX Project

Lawrence Berkeley Laboratory
Paul Adams, Ralf Grosse-Kunstleve, Pavel Afonine, Nat Echols, Nigel Moriarty, Jeff Headd, Nicholas Sauter, Peter Zwart

Los Alamos National Laboratory
Tom Terwilliger, Li-Wei Hung

Cambridge University
Randy Read, Airlie McCoy, Gabor Bunkoczi, Rob Oeffner

Duke University
Jane & David Richardson, Vincent Chen, Chris Williams, Bryan Arendall, Laura Murray

An NIH/NIGMS funded Program Project