Model Completion and Ligand Building with ARP/wARP

Victor Lamzin

EMBL Hamburg
ARP/wARP

Protein model building (to ~3.5 Å)
Helices/Strands (to ~4.5 Å)

DNA/RNA building (to ~3.5 Å)

Ligand building

ArpNavigator
Modelling a Protein Chain
Methods for Building Protein Structure

<table>
<thead>
<tr>
<th></th>
<th>Peptides</th>
<th>Di-peptides</th>
<th>Short fragments</th>
<th>Long chains</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-</td>
<td>✓</td>
<td>-</td>
<td>✓</td>
</tr>
<tr>
<td>Di-peptides</td>
<td></td>
<td></td>
<td>-</td>
<td>✓</td>
</tr>
<tr>
<td>Short fragments</td>
<td></td>
<td></td>
<td>-</td>
<td>✓</td>
</tr>
<tr>
<td>Long chains</td>
<td></td>
<td></td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
From Global to Local Shape Features

\[ f = [f_1, f_2, \ldots, f_n] \]

- Peptide
- Inv. Peptide
- Noise
Dependence on the Data Resolution

Calculated map of protein G

3 Å (a)  4 Å (b)  5 Å (c)  6 Å (d)  8 Å (e)
Dependence on the Data Resolution

1.5 Å

3.0 Å

3.8 Å

6-7 Å
Iterative Protein Model Building with ARP/wARP
Iterative Protein Model Building with ARP/wARP

<table>
<thead>
<tr>
<th>Resolution</th>
<th>Estimated fraction of automatically built protein structure (7/2010)</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 2.0Å</td>
<td>Over 90%</td>
</tr>
<tr>
<td>2.3Å</td>
<td>84%</td>
</tr>
<tr>
<td>2.6Å</td>
<td>80%</td>
</tr>
<tr>
<td>3.0Å</td>
<td>74%</td>
</tr>
<tr>
<td>3.5Å</td>
<td>65%</td>
</tr>
</tbody>
</table>
Fitting in Poorly Ordered Loops

Chain fragments are built and docked to sequence
Auto-NCS Detection

Initially built fragments → Identified NCS matches → Extended fragments

4 protein test structures, resolution 3.0 – 3.2 Å, NCS order 2 – 5
Observed improvement in all 4 cases
Average model completeness increased by 3 – 10%
Average number of chain fragments decreased by up to 20%
Short helix/strand fragments (3 to 5 Cα candidates) are built.

Longer traces are formed of which the best (in red) are kept.

Traces are clustered.

Assemblies are averaged.
Modelling Secondary Structure – Best Case Scenario

Helices

Strands
Modelling Secondary Structure

Helices for a 600-residue protein can be built in a few seconds on a modern iMac.
Tracing RNA/DNA Chains
Tracing RNA/DNA Chains
Building RNA/DNA Models


- The 30S ribosomal subunit
  - Resolution: 3.05 Å
  - Experimental phases (MCC 0.73)
- Auto-building with ARP/wARP
  - Modelled 1,302 out of 1,513 nucleotides (86%) with backbone r.m.s.d. to deposited structure: 0.7 Å
  - Located 1,121 nucleobases with 0.6 Å accuracy in location, and 12° in orientation
- Backbone fragmented in 75 chains
- Built in around 6h (cf. several months of manual work)
Crystallographic Ligand Building in a Nut-Shell

electron density map (good resolution)
Crystallographic Ligand Building in a Nut-Shell
Ligand Building Methods in ARP/wARP

Sparse grids

Conformational fit

Fine skeletons
Implementing Novel Shape Descriptors

1) Surface to volume ratio
2) Bounding box limits
3) Moments of inertia
4) Rotation match score
5) Eigenvalues
6) Distance histogram
7) Geodesic distance histogram

3\textsuperscript{rd} order moment invariants

<table>
<thead>
<tr>
<th></th>
<th>( / )</th>
<th>( | )</th>
<th>( + )</th>
<th>( \square )</th>
<th>( \bigcirc )</th>
<th>( \complement )</th>
<th>( | )</th>
<th>( \bigtriangleup )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \mu_{00} )</td>
<td>16</td>
<td>32</td>
<td>32</td>
<td>60</td>
<td>32</td>
<td>46</td>
<td>31</td>
<td>45</td>
</tr>
<tr>
<td>( \mu_{20} )</td>
<td>170</td>
<td>200</td>
<td>340</td>
<td>2255</td>
<td>900</td>
<td>1228</td>
<td>775</td>
<td>1130</td>
</tr>
<tr>
<td>( \mu_{30} )</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3049</td>
<td>3596</td>
<td>2850</td>
</tr>
</tbody>
</table>
Zernike Reconstruction of a Ligand

Orders shown:
6, 10, 12, 15 and 18
Success Rate for Ligand Building

~4,000 structures from the PDB (ESD) which are:

Resolution of the X-ray data: 3.0 Å or higher
Ligand size 20 to 40 atoms

Of these 2,800 ligands have RS MCC of 0.8 and higher

‘Correctly built’ corresponds to a NN RMSD of < 1.0 Å
Success Rate of Ligand Building

In yellow: the deposited ligand
Partial disorder, partially occupied ligands

- Deposited in PDB
- Built automatically with ARP/wARP when the full ligand is given

SAM in 1v2x at 1.5Å

Artificial cocktail:

Compound chosen in cocktail case
Cocktail Screening

Final ligand ranking
High-Throughput Ligand Identification

Density map → Segmentation → Projection/normalisation → Calculation of shape descriptors e.g. Zernike moments

Ligand → Ranking → Comparison → Database
Known structures allow us to identify particular patterns associated with ligand binding.

ATP in 1b38

<table>
<thead>
<tr>
<th>Amino Acid</th>
<th>Binding to Amino Acids</th>
</tr>
</thead>
<tbody>
<tr>
<td>GLY</td>
<td></td>
</tr>
<tr>
<td>ARG</td>
<td></td>
</tr>
<tr>
<td>ASP</td>
<td></td>
</tr>
<tr>
<td>GLU</td>
<td></td>
</tr>
<tr>
<td>LYS</td>
<td></td>
</tr>
<tr>
<td>ILE</td>
<td></td>
</tr>
<tr>
<td>PHE</td>
<td></td>
</tr>
<tr>
<td>TYR</td>
<td></td>
</tr>
<tr>
<td>THR</td>
<td></td>
</tr>
<tr>
<td>SER</td>
<td></td>
</tr>
<tr>
<td>LEU</td>
<td></td>
</tr>
<tr>
<td>HIS</td>
<td></td>
</tr>
<tr>
<td>VAL</td>
<td></td>
</tr>
<tr>
<td>ASN</td>
<td></td>
</tr>
<tr>
<td>MET</td>
<td></td>
</tr>
<tr>
<td>ALA</td>
<td></td>
</tr>
<tr>
<td>GLN</td>
<td></td>
</tr>
<tr>
<td>PRO</td>
<td></td>
</tr>
<tr>
<td>TRP</td>
<td></td>
</tr>
<tr>
<td>CYS</td>
<td></td>
</tr>
</tbody>
</table>

PDBemotif
Acknowledgements

EMBL Hamburg

Ciaran Carolan (ligands)
Johan Hattne (nucleic acids)
Gerrit Langer (helices/strands, arpnavigator, ligands)
Tim Wiegels (auto-NCS)

NKI Amsterdam

Krista Joosten (loops)