Automated model building with ARP/wARP

Gerrit Langer
Towards structural systems biology

$10^{25}$ atoms
Reasons and aims

Expanding the capabilities of macromolecular crystallography
What ARP/wARP 7 can do for you:

**Protein model building**
- Classic protein chain tracing at down to 2.7 Å resolution (partial tracing even at ~3.5 Å)
- Tracing helices and strands down to ~4.5 Å

**Model completion**
- Loop building
- DNA/RNA modelling
- Ligand fitting
- Solvent building

**Graphics Front End**

---

G. Langer: Automated model building with ARP/wARP

EMBL
Basic ARP essentials – in a picture


Fig. 16. A region of FDH model near Ala 20. The \((3F_o - 2F_c)\) electron density after ARP is contoured at \(1\sigma\) above the mean density. The initial model is shown in red, final model in green. Points indicate intermediate positions of atoms during refinement. Trajectories of movement of atoms which are retained from the initial model are coloured in green. Intermediate positions of the CB atom which moved out of density and was automatically removed are shown in blue. The final position of the CA atom was occupied by a new atom picked up in positive difference density, of which the trajectory is shown in red.
Basic ARP essentials – some remarks

• **In real space:**
  Atom update = removing and adding free atoms based on the current density. Free atoms are unrestrained. **Hybrid models** = a part of the atoms has a chemical assignment and the rest are free.

• **In reciprocal space:** Unrestrained refinement of free atoms and restrained refinement of chemically assigned atoms.

• The procedure *cycles* between real space and reciprocal space thereby forming a unified process of model building and refinement.
Basics: The hybrid model

- A partial protein model is used together with a free-atom model.
- Chemically assigned parts of the model provide restraints for refinement.
- The scheme of restraints and the free atoms are iteratively updated.
- The hybrid model is converging to the final model.
Basics: Creating the free atom model

Finding a parametrisation of the spatial density distribution.

Now building 1 initial model(s)

Had to go as low as 1.95 sigma, to complete atoms search.
Basics: The atom update

Removal and addition of atoms based on density at atomic centers

2Fo-Fc map

Fo-Fc map

ARP/wARP will be iterated with REFMAC5
50 refinement / model update cycles will be run in total.
Atoms will be removed below 1.0 sigma in 2mFoDFc map and added above 3.2 sigma in mFoDFc map.
Refinement – the simple approach

Refinement and model building are the two sides of structure solution: Model building determines the presence, nature and initial values of free parameters and refinement optimizes these values.

**Observed Data**
- $F_{111,\text{obs}}$
- $F_{112,\text{obs}}$
- $F_{121,\text{obs}}$
- $F_{131,\text{obs}}$
- $F_{211,\text{obs}}$
- $\cdots$
- $F_{\text{hkl,obs}}$

**Calculated Data**
- $F_{111,\text{calc}}$
- $F_{112,\text{calc}}$
- $F_{121,\text{calc}}$
- $F_{131,\text{calc}}$
- $F_{211,\text{calc}}$
- $\cdots$
- $F_{\text{hkl,calc}}$

**Model**
- $\text{N Atoms + Restraints}$
- $\text{xyzB(1)}$
- $\text{xyzB(2)}$
- $\text{xyzB(3)}$
- $\text{xyzB(4)}$
- $\cdots$
- $\text{xyzB(N)}$

**Fo-Fc**

- Calculate
- Fit through minimising difference
- Constantly updated
Atomic parameters vs. number of reflections

Parameters: Atomic XYZ, B

The theoretical number of reflections:

Reflections per atom for 50% solvent content ($V_m=2.4\,\text{Å}^3/\text{Da}$) and an average molecular weight of 14:

Example:
2.0 Å resolution - 9 reflections per atom
2.3 Å resolution - 6 reflections per atom
2.7 Å resolution - 4 reflections per atom

\[
N_{\text{refl}} = \frac{2\pi V_{AU}}{3d^3}
\]

\[
N_{\text{refl/atom}} \approx \frac{70}{d^3}
\]

or

\[
N_{\text{refl/residue}} \approx \frac{560}{d^3}
\]
Atomic parameters vs. number of reflections

Mathematically one should always have more observations than parameters to make optimisation algorithms stable!

Through constraints we reduce the number of free parameters. $d [\text{Å}]$

Through restraints we add the information about the known distribution of their values.
Autotracing: Recognising peptides

Dummy atoms with a distance close to that of \( \text{C}_\alpha - \text{C}_\alpha \) pairs are chosen for peptide recognition.

- Look for \( \text{C}_\alpha - \text{C}_\alpha \) distances (atoms 2.8 - 4.8 Å apart)
- Placing the peptide plane
  - Rotate into best density
- Real-Space-Refine the peptide into the density
Autotracing: Recognising peptides

The task:
Separating true peptides from false ones.

peptide
good fit

peptide
but poor fit

non-peptide
but good fit
Autotracing: Recognising peptides

A peptide density shape is described by \( \sim 1,000 \) parameters, which are then reduced to 1 parameter:

\[
Y = \sum_{i=1,N} w_i (p_{\text{obs},i} - p_{\text{template},i})^2
\]

1.9Å resolution

Target density
Weight function
Similarity searched for
Autotracing: From peptides to di-peptides

Restrict possible di-peptide conformations.
Autotracing: The branching problem

Some $\text{C}_\alpha$-$\text{C}_\alpha$ connections (peptides) are guessed correctly and some are not. The average connectivity is generally higher than 1 leading to an enormous amount of trace routes!

A depth-first search algorithm performs a limited graph search for the sake of execution speed.
Sequence docking
Side chain fitting

• Complete side chains are built in real space at once.
• Discrete best rotamer search.
• Torsional angle refinement.
The ‘classic’ workflow at a glance
A typical ARP/wARP run

<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>

![Graph showing Chains, Residues / 10, and R factor (%) over Cycle](graph.png)
Building loops – the idea

- Missing portions in the middle of a protein model shall be obtained

- Can we fill these gaps by using the knowledge contained in the partial model?
  1) Number of missing residues - Loop length
  2) Sequence info of partial model - Loop sequence
  3) Structure of partial model - Loop restraints
  4) Positions of known fragments - Loop constraints (Anchors)

\[
P(\varphi_2, \theta_1 | \varphi_1, \theta_0) = \frac{P(\varphi_1, \varphi_2, \theta_0, \theta_1)}{P(\varphi_1, \theta_0)}
\]
• **Hypothesis:**
  1) Angles and dihedral angles of the Cα’s in tetra-peptides can be used to suggest the positions of a 5th Cα
  2) Iterate this extension to generate possible chain traces to bridge the gap
Boosting the ARP/wARP performance: NCS

The exploitation of NCS properties of protein molecules helps to increase the completeness of the built model.

For more details listen to Tim’s talk tomorrow!
Conditional Restraints
Conditional Restraints

Often the fraction of the protein built can be increased.
Secondary structure modeling

- Secondary structure is abundant in proteins and has well conserved stereochemical features.
- Helices can be easily spotted even at lower resolution and be used as anchor fragments for chain tracing when the classic ARP/wARP approach works less well.
- On average 43% helices and 22% β-strands.

Calculated map of protein G at a) 3Å, b) 4Å, c) 5Å, d) 6Å, e) 8Å resolution
Secondary structure modeling

Short helix/strand fragments (3 to 5 Cα candidates) are recognised by their geometry: distances and angles.

Short fragments overlap to form long traces.

The amount of combinations is huge. Only the unique and best traces are kept to limit complexity.

The sum of densities at Cα positions is used as a score.
Secondary structure modeling

Center points are used to measure distances between helix or strand traces. This is used for clustering and the construction of trace ensembles. Ensembles get averaged and results refined in real space.
Secondary structure modeling: Example

1o14 at 3.2Å – good phases, 60% of the protein traced in 1min

helices: 88% complete, strands: 66% complete, rmsd of Cα atoms to reference structure: 0.9Å
Secondary structure

Tracing completeness

RMS deviation
Tracing DNA/RNA: Recognising planar objects

- Adenine
- Cytosine
- Guanine
- Thymine
- Tryptophan
- Phenylalanine
- Histidine
- Tyrosine

Other planar or near-planar objects

- Planar side chain
- Peptide bond
- Puckered ribose
- Puckered proline

- Arg, Asn, Asp, Gln, and Glu
Tracing DNA/RNA: Employing local density features

\[ \mathbf{f} = [f_1, f_2, \ldots, f_n] \]

- Plane
- Phosphate
- Noise

Normalisation, Interpolation

Feature calculation

e.g. 3\textsuperscript{rd} order moment invariants
Tracing DNA/RNA: Parallels
Tracing DNA/RNA: Example

1j5e: Wimberly et al. (2000)
Nature 407, 327–339

- The 30S ribosomal subunit
  - Resolution: 3.05 Å
  - 1,513 nucleotides (red)
  - 20 ribosomal proteins, 2,540 residues in total (green)
- Autobuilding using experimental phases
- After 3h: Modelled 1,302 out of 1,513 nucleotides (86%)
- Located 1,121 nucleobases
  - Backbone r.m.s.d. to deposited structure: 0.7 Å
ARP/wARP and molecular graphics

This should give a more intuitive access to functions such as: Building ligands, tracing helices/strands and model solvent.
ARP/wARP and molecular graphics

Some few ways of manipulating ligands have already been implemented.
Computational web services

Submit a remote job at the Hamburg Cluster

- Submit the job for remote execution at the Hamburg cluster

Your Email address: [text input]

Job data can be archived and made available to any software developer that requests them.

Run

Save or Restore

Close

www.embl-hamburg.de/ARPwARP/remote-http.html

www.embl-hamburg.de/Auto-Rickshaw/

βαλβεσ
Computational web services

ARP/wARP web service

1. Opening Page
2. Step 1
   - Your email address: gerrit@embl-hamburg.de
   - Run ARP/wARP starting from: experimental phases
   - MTZ file: /proj/ARP_Example_Data/Tracing/PSP/psp.mtz

3. Step 2
   - Proceed to step two

4. Confirmation Page
Downloading ARP/wARP

http://www.embl-hamburg.de/ARP/

ARP/wARP is a software suite for improvement and objective interpretation of crystallographic electron density maps and automatic construction and refinement of macromolecular models.

ARP/wARP 7.1 download and license

Download User Guide (pdf)
View User Guide (html)
Frequently Asked Questions and Software Patches

ARP/wARP protein model building via the web

Send E-mail to the ARP/wARP list (subscribers only):
Get ARP/wARP List Archive
Subscribe to ARP/wARP list
Unsubscribe from ARP/wARP list

Relevant literature:

People and Acknowledgements

Developers

EMBL Hamburg: Ciaran Carolan, Philipp Heuser, Victor Lamzin, Gerrit Langer, Tim Wiegels

NKI Amsterdam: Krista Joosten, Robbie Joosten, Tassos Perrakis

Collaborators

EMBL Hamburg: Gleb Bourenkov, Santosh Panjikar

York University: Garib Murshudov’s group

Daresbury Laboratory: CCP4 team

Former members

Richard Morris, Peter Zwart, Francisco Fernandez, Matheos Kakaris, Olga Kirillova, Wijnand Mooij, Diederick De Vries, Marouane Ben Jelloul, Johan Hattne, Tilo Strutz, Guillaume Evrard, Serge Cohen, Venkat Parthasarathy, Babu Pothineni, Helene Doerksen

Funding