Small molecules
How to identify and build them
(with ARP/wARP)
The task at hand

To find ligand density and build it!
Fitting a ligand

We have:

- electron density map (good resolution)
- + protein model
  - shows *excess density* not accounted for by protein
- *difference density map*
  - to fit the appropriate ligand into
Fitting a ligand

Challenges:

- different resolutions and data quality
- different ligand complexity / topology
- partial disorder of a ligand
- different ligands at the same site?
The protocol of a modelling task

What we know at the beginning will determine the protocol:

<table>
<thead>
<tr>
<th>Ligand identity</th>
<th>Binding site</th>
<th>1 density cluster</th>
<th>1 ligand</th>
</tr>
</thead>
<tbody>
<tr>
<td>Known</td>
<td>Known</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Unknown</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unknown</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Joana Pereira 08.11.2014
The protocol of a modelling task

What we know at the beginning will determine the protocol:

<table>
<thead>
<tr>
<th>Ligand identity</th>
<th>Known</th>
<th>Unknown</th>
</tr>
</thead>
<tbody>
<tr>
<td>Known</td>
<td></td>
<td><img src="image" alt="X" /></td>
</tr>
<tr>
<td>Unknown</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**N density clusters**

**1 ligand**
The protocol of a modelling task

What we know at the beginning will determine the protocol:

<table>
<thead>
<tr>
<th>Ligand identity</th>
<th>Known</th>
<th>Unknown</th>
</tr>
</thead>
<tbody>
<tr>
<td>Known</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unknown</td>
<td><strong>X</strong></td>
<td></td>
</tr>
</tbody>
</table>

N ligand candidates

1 density cluster
The protocol of a modelling task

What we know at the beginning will determine the protocol:

<table>
<thead>
<tr>
<th>Ligand identity</th>
<th>Binding site</th>
<th>( N ) density clusters</th>
<th>( N ) ligand candidates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Known Known</td>
<td><img src="image1" alt="Known Known" /></td>
<td><img src="image2" alt="Known Known" /></td>
<td><img src="image3" alt="Known Known" /></td>
</tr>
<tr>
<td>Known Unknown</td>
<td><img src="image4" alt="Known Unknown" /></td>
<td><img src="image5" alt="Known Unknown" /></td>
<td><img src="image6" alt="Known Unknown" /></td>
</tr>
<tr>
<td>Unknown Unknown</td>
<td><img src="image7" alt="Unknown Unknown" /></td>
<td><img src="image8" alt="Unknown Unknown" /></td>
<td><img src="image9" alt="Unknown Unknown" /></td>
</tr>
</tbody>
</table>
I am trying to solve a structure of a protein with some inhibitor. I want to know how I can put in my inhibitor in the density map of the data I got. I can see some density in the active site where the inhibitor should be...
I am not sure of how to do it.

Which case is this?

I density cluster
I ligand
Ligand fitting with ARP/wARP

Organised as a **pipeline** of core modules for **specific sub-tasks** following an intuitive approach

- **Prepare**
  - Identify ligand **binding site** and/or **ligand**
  - Sparse density map and **generate** **ligand topology**

- **Construct**
  - Construct **ensemble** of ligand models in plausible conformation to **fit sparsed density**

- **Refine**
  - Refine ligand coordinates to satisfy **geometric** constraints and **maximise** fit to density
  - Choose **best model**
Constructing the ligand

- To **address the different ligand sizes** and complexities encountered
  - to increase the **robustness** of the software

- We use 2 separate construction methods:
  1. Label swapping on the sparsed grid
  2. Metropolis search in conformation space
Constructing the ligand: graph search

- Uses the **sparse grid representation** of the electron density at the chosen binding site & topology of known ligand

### Knowledge about ligand
- Connectivity
- Distances
- Angles
- Chirality

### Knowledge about grid
- Connectivity
- Distances
- Density
- **NO IDENTITIES**
Constructing the ligand: graph search

- Label swapping
  - Ligand is expanded on the sparse density, preserving connectivities
- Every ‘dummy atom’ of the sparse grid is tried as a start point
- An exhaustive graph search is performed
- Models are scored by their fit to density and expected stereochemical features
Constructing the ligand: graph search

Index of starting grid point

Dead branches

Start Atom

Surviving start points

No of ligand atoms placed

Joana Pereira 08.11.2014

Tim Wiegels 30.03.12

YEARS

1974–2014

EMBL
Constructing the ligand: metropolis search

- Perform a **random walk** in parameter space biased towards the **optimum of a score function**
  - Parameter space: **position, orientation** and **conformation**
  - Score function: ‘**pseudo’ map correlation**
  - Advantage: **less degrees of freedom**

Rigid groups

Rotatable bonds
Constructing the ligand: metropolis search

- Evolution of a model during a metropolis optimisation

Final result: 9000 moves + refinement

FMN to 1jqv at 2.1 Å: 0.23 Å rmsd.
Is there a simple way to determine whether ligand is bound or not by comparing the diffraction patterns between ligand-free (structure known) and ligand-soaked protein crystals? I would like to solve the ligand bound protein structure, but before I do so, **I have to find out if the ligand is actually bound and if so, where.**

Thank you very much!

---

**Which case is this?**

- N density clusters
- I ligand
Automatic binding site identification

- The difference density map from **low to high** contour threshold
Automatic binding site identification

- Capturing the dependence of the difference density map on changing the contour thresholds: **the fragmentation tree**

![Graph showing the dependence of cluster volume on contour threshold](image)
Shape features to measure similarity

An electron density map calculated from the ligand to be fit is compared to each potential density cluster:

Find the matching cluster for this ligand!

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

Decision is based on feature vectors
We are working on protein/inhibitor complex structure. However, we did find a strange density at the active site, it looks really like GSH, the natural co-enzyme of this protein. We tried to use very simple solution to get crystal then exclude the possibility of buffer molecules, but that density is always there.

I want to identify this ligand. Are there any methods to do this work?

Which case is this?

I density cluster

N ligand candidates
It's all in the shape, baby!
Shape features to measure similarity

- To **assign a likelihood to a ligand** knowing the binding site

Final ligand ranking
High throughput ligand identification

Density map (or otherwise surface of protein pocket)

Ligand

Segmentation

Projection/normalisation

Calculation of shape descriptors

Database

Ranking

Comparison

84 of the most common ligands in crystallography

C. Carolan & V. Lamzin, Acta D (July 2014)
Flexibility - Ligand Conformations!

- full ligand flexibility is considered

Currently, a database of 84 of the most common ligands in crystallography is screened.

Beware of multiple ligands, possibly covalently bound, and coordinated solvent among other things!
I have a ligand in the binding pocket and mediocre data. **The core of the ligand is well defined, however there are chains/tails of the ligand which are not**….The question here is how the chains/tails should be modelled (if at all)

Which case is this?

None of the Above
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
- Compound chosen in cocktail case

Artificial Cocktail
Partial ligand building in ArpNavigator.
Success rate

- Tested on > 20k PDB entries from EDS
- X-ray resolution limits: 1.0 to 3.0 Å
- Building the **largest fully occupied ligand**
- Correctness criterion: r.m.s.d. < 1.0 Å

For ARP/wARP 7.2

- 5…6 atoms, any map corr.
- 7…100 atoms, any map corr.
- 10…40 atoms, map corr. > 80%
- 20…40 atoms, map corr. > 80%
Success criteria: r.m.s.d < 1Å

Correct

\[ \text{rmsd} = 0.4\text{Å} \]

Correct

\[ \text{rmsd} = 1.0\text{Å} \]

Incorrect

\[ \text{rmsd} = 1.8\text{Å} \]

Incorrect

\[ \text{rmsd} = 1.5\text{Å} \]

*In yellow: the deposited ligand*
Ligand building methods

- Sparse grids
- Conformational fit
- Fine skeletons

- High-throughput ligand identification using shape descriptors

Building models which do NOT have repetitive motifs like peptides or nucleotides
The people - The power!

Collaborators

EMBL Hamburg: Gleb Bourenkov, Santosh Panjikar
MRC Laboratory, Cambridge: Garib Murshudov’s group
Rutherford Appleton Laboratory: CCP4 team

Previous Members

Ciaran Carolan, Tim Wiegels

Funding