X-ray sociology

- more exact than NMR
- lots of Nobel prizes
- tells you where atoms are to a fraction (1/5 - 1/4) of an Å
- can work on large structures
- Hamburg is full of crystallographers
- not all proteins / nucleotides are happy to crystallise

Andrew Torda, Wintersemester 2018/2019
Summary of story

- X-ray source (synchrotron)
- Detector
- Measured reflections
- Electron density
- Coordinates (model)
Topics

- tiny bit about chemistry / crystals
- calculating electron density
- placing coordinates in density
Forming crystals

Familiar crystals – Kochsalz, sugar...
• crystal formation – small molecules
  • rigid, regular
  • soup of unordered molecules ⇌ ordered crystal, $\Delta G$ favourable

Proteins
• lots of internal mobility – not rigid
• not nice regular shapes
• soup molecules ⇌ ordered crystal, $\Delta G$ borderline
  Often can not be crystallised
Protein crystals

- small, 1 to 100 μm
- not as well ordered as small molecules

We measure an average over all molecules – consequence ..
- right hand side – the average is smeared
Diffraction

Start with 1-dimension
Think of a grid

waves
(light, sound, ..)

Spacing of peaks on the detector?
Will depend on
- distance $r_{AB}$
- wavelength $\lambda$
What we measure will be

- periodic (cosinus) in $x$
- change slowly if $\lambda$ is big

$$I = \cos\left(\frac{2\pi}{\lambda} x + \alpha\right)$$

- $\alpha$ phase – important later
- relate this to proteins

*a detector*
protein crystal and grids

- wavelengths are we talking about? X-ray $\approx 1 \text{ Å}$
- do we have grids?

Work in 3D
- makes maths more difficult
- have to rotate and add up pictures
- we have lots of pairs of atoms (AB)

X-rays interact with electrons
- hydrogen is almost invisible

What one measures
- sum over many pairs of electron clouds
Summing and Fourier transforms

Simple formula in 1D and one pair

\[ I = \cos \left( \frac{2\pi}{\lambda} x + \alpha \right) \]

- what we measure is a sum \( \sum_{\text{reflections}} \cos(... \)
- better nomenclature \( \Sigma_{hkl} - h, k, l \) are indices of detected spots

How do I break a signal into frequencies?
- fourier transform – think of pictures of media player
  - from sound in real time to picture of frequencies \( \cos(\omega) \)

Can you just take a signal and fourier transform?
- only if you know the \( \alpha \)'s
- There is phase for every reflection, but you cannot measure it
phase problem

• lots of "reflections" ($10^5$)
• you measure their size / intensity
• you cannot measure their phase $\alpha$

• if you know the $\alpha$ for each reflection
  • just calculate density

Where do the phases ($\alpha$) come from?
Ways to find phases

Can you do it directly?
• for very small molecules – yes
• protein?
  • imagine you have $10^5$ reflections
  • try just 4 possibilities in each case
  • $4^{10000}$ possibilities

Strategies – two most important
1. if you know some coordinates, substitute them into the equation – gets a good starting point (molecular replacement)
2. if you know some phases, easy to get the next ones (multiple isomorphous replacement)

+ if you have a reasonable initial guess, it can be optimised
molecular replacement

- most common – ¾ or more of structures in protein data bank
- you need to know some coordinates

- you do not know coordinates
  - you have the coordinates of something close (homologue)

- pretend these coordinates are responsible for the measurements
- substitute into formula and get an initial set of $\alpha$'s
- refine
multiple isomorphous replacement (MIR)

Remember
- for small molecules you can find phases directly (few reflections)
- if you know some phases, the next are easier

How to treat protein as if it is a small molecule?

- Heavy atoms have so many electrons they dominate the observations
- give a starting point for phases

- what are heavy atoms? Au, Pt, Hg, Br, Se, Xe

- should bind at the same position in every protein molecule
more phasing

How difficult?
- for many proteins – plenty of data but phases take years

Other methods?
- yes
resolution - meaning

• most common statistic for X-ray structures – resolution
• meaning - when do two points look like one?

• Resolution in X-ray.. depends on how scattered the waves are

planes in a crystal
waves and resolution

- bottom path = lower path + wavelength $\lambda$
  then they reinforce each other
- difference in lengths is $n \lambda$
  $n\lambda = \overline{ABC} = 2(d \sin \theta)$
- then $d = \frac{n\lambda}{2 \sin \theta}$

Resolution consequences
- smaller wavelength $\lambda$, the better
- angle $\theta$ – you cannot do much but..
In the centre .. reflections
• did not diffract much
• due to low resolution information
• strong

Further out
• high resolution information
• signal weak
Mostly a function of crystal
- crystal not so regular and / or
- atoms mobile
  - you are seeing an average
  - there is no high resolution information
disorder – static and dynamic

static

dynamic

model for this - soon
model fitting - refinement

- you have initial electron density and initial phases
- have to fit atoms (build a model)

- what are the variables
  - \( x, y, z \) for atoms
  - \( B \)-factors (mobility) ... next slide

- Given atoms, you can calculate density
- given density, you can calculate reflections (structure factors \( F \))

Refinement
- how well do the structure factors from your model agree with measurements?
**B-factors**

A model for atom location

- Gaussian (normal) probability

Why Gaussian?

- pretend a particle moves in a harmonic well $E(x) = x^2$
- from Boltzmann relation, $p(x) \propto \exp\left(\frac{-x^2}{kT}\right)$ (formal in Sommersemester)
  - $k$ Boltzmann constant, $T$ temperature

![Graph showing energy and probability distribution](image)
**B-factors interpretation**

Probability distribution

- how likely are you see an atom at a position or
- what is the typical movement at room temperature?

\[ B = 8 \pi^2 u^2 \quad \text{where } u \text{ are fluctuations in Å} \]

\[ u = \left( \frac{B}{8\pi^2} \right)^{\frac{1}{2}} \]

- if \( B = 50 \text{ Å}^2 \), typical displacement \( \approx 0.8 \text{ Å} \)
- if \( B = 20\text{Å}^2 \), typical displacement \( \approx 0.5 \text{ Å} \)

The connection to fitting...
fitting to density

The path to this picture...
• water molecules
• atoms have different
  • sizes
  • electron clouds
• mobilities
Refinement – cost function – $R$

The cost function

- atoms $\rightarrow$ density $\rightarrow$ structure factors ($F$)
  - $F_{hkl}^{calc}$ structure factors calculated
  - $F_{hkl}^{obs}$ structure factors observed

very important... $R$ factor

$$R = \frac{\sum_{hkl} |F_{hkl}^{obs} - F_{hkl}^{calc}|}{\sum_{hkl} |F_{hkl}^{obs}|}$$
Variables in refinement

What are the variables?

- $x, y, z$ for every atom + $B$-factors

- adding ions and water to fill density
<table>
<thead>
<tr>
<th>Method</th>
<th>#par</th>
<th>#obs</th>
<th>Rw</th>
<th>Rf</th>
</tr>
</thead>
<tbody>
<tr>
<td>molecular replacement</td>
<td>3</td>
<td>775</td>
<td>44.5</td>
<td></td>
</tr>
<tr>
<td>rigid body</td>
<td>9</td>
<td>2997</td>
<td>46.4</td>
<td>47.4</td>
</tr>
<tr>
<td>first round</td>
<td>3887</td>
<td>13818</td>
<td>30.2</td>
<td>34.3</td>
</tr>
<tr>
<td>after first</td>
<td>3735</td>
<td></td>
<td>26.2</td>
<td>31.4</td>
</tr>
<tr>
<td>SHELXWAT</td>
<td>4091</td>
<td></td>
<td>20.2</td>
<td>24.6</td>
</tr>
<tr>
<td>bld + SHELXWAT</td>
<td>4231</td>
<td></td>
<td>18.7</td>
<td>22.8</td>
</tr>
<tr>
<td>include all data</td>
<td>4203</td>
<td>33993</td>
<td>19.1</td>
<td>21.1</td>
</tr>
<tr>
<td>ANIS 20</td>
<td>9453</td>
<td></td>
<td>16.0</td>
<td>19.2</td>
</tr>
<tr>
<td>Rebuild SHELXWAT</td>
<td>9557</td>
<td></td>
<td>13.4</td>
<td>15.8</td>
</tr>
<tr>
<td>rebuild</td>
<td>10481</td>
<td></td>
<td>12.3</td>
<td>15.5</td>
</tr>
<tr>
<td>rebuild</td>
<td>10819</td>
<td></td>
<td>12.1</td>
<td>15.1</td>
</tr>
<tr>
<td>rebuild</td>
<td>10838</td>
<td></td>
<td>11.6</td>
<td>14.6</td>
</tr>
<tr>
<td>rebuild</td>
<td>11494</td>
<td></td>
<td>11.3</td>
<td>14.4</td>
</tr>
<tr>
<td>rebuild</td>
<td>11576</td>
<td></td>
<td>11.0</td>
<td>14.0</td>
</tr>
<tr>
<td>rebuild</td>
<td>11774</td>
<td></td>
<td>10.7</td>
<td>13.8</td>
</tr>
</tbody>
</table>
## Refinement Progress

<table>
<thead>
<tr>
<th>Method</th>
<th>Data Used</th>
<th># Par</th>
<th># Obs</th>
<th>Rw</th>
<th>Rf</th>
</tr>
</thead>
<tbody>
<tr>
<td>Molecular Replacement</td>
<td>15-4.0</td>
<td>3</td>
<td>775</td>
<td>44.5</td>
<td></td>
</tr>
<tr>
<td>Rigid Body</td>
<td>10-2.5</td>
<td>9</td>
<td>2997</td>
<td>46.4</td>
<td>47.4</td>
</tr>
<tr>
<td>First Round</td>
<td>10-1.5</td>
<td>3887</td>
<td>13818</td>
<td>30.2</td>
<td>34.3</td>
</tr>
<tr>
<td>After First Round</td>
<td>10-1.5</td>
<td>3735</td>
<td></td>
<td>26.2</td>
<td>31.4</td>
</tr>
<tr>
<td>SHELXWAT</td>
<td>4091</td>
<td></td>
<td></td>
<td>20.2</td>
<td>24.6</td>
</tr>
<tr>
<td>bld + SHELXWAT</td>
<td>4231</td>
<td></td>
<td></td>
<td>18.7</td>
<td>22.8</td>
</tr>
<tr>
<td>Include All Data</td>
<td>10-1.1</td>
<td>4203</td>
<td>33993</td>
<td>19.1</td>
<td>21.7</td>
</tr>
<tr>
<td>ANIS 20</td>
<td>9453</td>
<td></td>
<td></td>
<td>16.0</td>
<td>19.2</td>
</tr>
<tr>
<td>Rebuild SHELXWAT</td>
<td>9557</td>
<td></td>
<td></td>
<td>13.4</td>
<td>15.8</td>
</tr>
<tr>
<td>Rebuild</td>
<td>10481</td>
<td></td>
<td></td>
<td>12.3</td>
<td>15.3</td>
</tr>
<tr>
<td>Rebuild</td>
<td>10819</td>
<td></td>
<td></td>
<td>12.1</td>
<td>15.1</td>
</tr>
<tr>
<td>Rebuild</td>
<td>10838</td>
<td></td>
<td></td>
<td>11.6</td>
<td>14.6</td>
</tr>
<tr>
<td>Rebuild</td>
<td>11494</td>
<td></td>
<td></td>
<td>11.3</td>
<td>14.4</td>
</tr>
<tr>
<td>Rebuild</td>
<td>11576</td>
<td></td>
<td></td>
<td>11.0</td>
<td>14.0</td>
</tr>
<tr>
<td>Rebuild</td>
<td>11774</td>
<td></td>
<td></td>
<td>10.7</td>
<td>13.8</td>
</tr>
</tbody>
</table>
**refinement progress**

<table>
<thead>
<tr>
<th>data used</th>
<th># par</th>
<th># obs</th>
<th>Rw</th>
<th>Rf</th>
</tr>
</thead>
<tbody>
<tr>
<td>molecular replacement</td>
<td>15-4.0</td>
<td>3</td>
<td>775</td>
<td>44.5</td>
</tr>
<tr>
<td>rigid body</td>
<td>10-2.5</td>
<td>9</td>
<td>2997</td>
<td>46.4</td>
</tr>
<tr>
<td>first round</td>
<td>10-1.5</td>
<td>3887</td>
<td>13818</td>
<td>30.2</td>
</tr>
<tr>
<td>after first</td>
<td></td>
<td>3735</td>
<td></td>
<td>26.2</td>
</tr>
</tbody>
</table>

- Start: Low resolution data is enough – just a few parameters
- #par and #obs
  - where is the molecule?
  - where is molecule + phases?
  - add in first atoms
    - number of parameters grows at each step
  - add in more data (#obs)
## Refinement Progress

<table>
<thead>
<tr>
<th>Method</th>
<th>Data Used</th>
<th># Par</th>
<th># Obs</th>
<th>$R_w$</th>
<th>$R_f$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Molecular Replacement</td>
<td>15-4.0</td>
<td>3</td>
<td>775</td>
<td>44.5</td>
<td></td>
</tr>
<tr>
<td>Rigid Body</td>
<td>10-2.5</td>
<td>9</td>
<td>2997</td>
<td>46.4</td>
<td>47.4</td>
</tr>
<tr>
<td>First Round</td>
<td>10-1.5</td>
<td>3887</td>
<td>13818</td>
<td>30.2</td>
<td>34.3</td>
</tr>
<tr>
<td>After First</td>
<td>10-1.5</td>
<td>3735</td>
<td></td>
<td>26.2</td>
<td>31.4</td>
</tr>
<tr>
<td>SHELXWAT</td>
<td></td>
<td></td>
<td></td>
<td>20.2</td>
<td>24.6</td>
</tr>
<tr>
<td>bld + SHELXWAT</td>
<td></td>
<td>4091</td>
<td></td>
<td>18.7</td>
<td>22.8</td>
</tr>
<tr>
<td>Include All Data</td>
<td>10-1.1</td>
<td>4203</td>
<td>33993</td>
<td>19.1</td>
<td>21.7</td>
</tr>
<tr>
<td>ANIS 20</td>
<td></td>
<td>9453</td>
<td></td>
<td>16.0</td>
<td>19.2</td>
</tr>
<tr>
<td>Rebuild SHELXWAT</td>
<td></td>
<td>9557</td>
<td></td>
<td>13.4</td>
<td>15.8</td>
</tr>
<tr>
<td>Rebuild</td>
<td></td>
<td></td>
<td></td>
<td>12.3</td>
<td></td>
</tr>
<tr>
<td>Rebuild</td>
<td></td>
<td>10481</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rebuild</td>
<td></td>
<td>10819</td>
<td></td>
<td>12.1</td>
<td></td>
</tr>
<tr>
<td>Rebuild</td>
<td></td>
<td>10838</td>
<td></td>
<td>11.6</td>
<td>14.6</td>
</tr>
<tr>
<td>Rebuild</td>
<td></td>
<td>11494</td>
<td></td>
<td>11.3</td>
<td>14.4</td>
</tr>
<tr>
<td>Rebuild</td>
<td></td>
<td>11576</td>
<td></td>
<td>11.0</td>
<td>14.0</td>
</tr>
<tr>
<td>Rebuild</td>
<td></td>
<td>11774</td>
<td></td>
<td>10.7</td>
<td>13.8</td>
</tr>
</tbody>
</table>

*Note: $R$ of bit more than 20% is typical.*
What is happening in refinement

Similar problem to NMR refinement
- use a minimizer
  - move $x$, $y$, $z$ and $B$-factors
    - until agree with experimental data ($F_{hkl}^{\text{obs}}$)
    - maintain known chemistry – bond lengths angles

Different to NMR
- more variables
  - $B$-factors
    - you can add water and ions to fill density
- more data
- search for a single solution (not many possibilities)
Overfitting?
- you give me data
- true model is a line \( y = ax + m \)
  - two parameters
- I fit to a polynomial
  \[ y = ax^4 + bx^3 + cx^2 + dx + m \]
- apparently better fit
  - will not predict correctly

How would you detect this?
- I have half a dozen points
- do fitting on five – see how good the fit on the sixth point is

In general – fit on 90 % the data and test on remaining 10 %
## Refinement Progress

<table>
<thead>
<tr>
<th>Method</th>
<th>Data Used</th>
<th># Par</th>
<th># Obs</th>
<th>Rw</th>
<th>Rf</th>
</tr>
</thead>
<tbody>
<tr>
<td>Molecular Replacement</td>
<td>15-4.0</td>
<td>3</td>
<td>775</td>
<td>44.5</td>
<td></td>
</tr>
<tr>
<td>Rigid Body</td>
<td>10-2.5</td>
<td>9</td>
<td>2997</td>
<td>46.4</td>
<td>47.4</td>
</tr>
<tr>
<td>First Round</td>
<td>10-1.5</td>
<td>3887</td>
<td>13818</td>
<td>30.2</td>
<td>34.3</td>
</tr>
<tr>
<td>After First</td>
<td>3735</td>
<td></td>
<td></td>
<td>26.2</td>
<td>31.4</td>
</tr>
<tr>
<td>SHELXWAT</td>
<td></td>
<td>4091</td>
<td></td>
<td>20.2</td>
<td>24.6</td>
</tr>
<tr>
<td>bld + SHELXWAT</td>
<td></td>
<td>4231</td>
<td></td>
<td>18.7</td>
<td>22.8</td>
</tr>
<tr>
<td>Include All Data</td>
<td>10-1.1</td>
<td>4203</td>
<td>33993</td>
<td>19.1</td>
<td>21.7</td>
</tr>
<tr>
<td>ANIS 20</td>
<td></td>
<td>9453</td>
<td></td>
<td>16.0</td>
<td>19.2</td>
</tr>
<tr>
<td>Rebuild SHELXWAT</td>
<td></td>
<td>9557</td>
<td></td>
<td>13.4</td>
<td>15.8</td>
</tr>
<tr>
<td>Rebuild</td>
<td></td>
<td>10481</td>
<td></td>
<td>12.3</td>
<td>15.3</td>
</tr>
<tr>
<td>Rebuild</td>
<td></td>
<td>10819</td>
<td></td>
<td>12.1</td>
<td>15.1</td>
</tr>
<tr>
<td>Rebuild</td>
<td></td>
<td>10838</td>
<td></td>
<td>11.6</td>
<td>14.6</td>
</tr>
<tr>
<td>Rebuild</td>
<td></td>
<td>11494</td>
<td></td>
<td>11.3</td>
<td>14.4</td>
</tr>
<tr>
<td>Rebuild</td>
<td></td>
<td>11576</td>
<td></td>
<td>11.0</td>
<td>14.0</td>
</tr>
<tr>
<td>Rebuild</td>
<td></td>
<td>11774</td>
<td></td>
<td>10.7</td>
<td>13.8</td>
</tr>
</tbody>
</table>
Crystal data – also has noise
- you can always fit to noise
  - add water, ions, move atoms
How to detect?
- fit on 90 % of data
- calculate $R$ on remaining data

\[
R_{\text{free}} = \frac{\sum_{hkl}|F_{hkl}^{\text{obs}} - F_{hkl}^{\text{calc}}|}{\sum_{hkl}|F_{hkl}^{\text{obs}}|}
\]

- but $F_{hkl}^{\text{obs}}$ is from 10% of data not used in fitting
- typical values in protein data bank – 20 to 30 %
Practical meaning of resolution

- formally – limit when two points become one point
- practical

best  < 1 Å  see H electrons
      1.2 Å  separate atoms
      2 Å  sidechains

worst  > 4 Å  overall shape
resolution in PDB

1.2 × 10^5 structures

75% of data
1 3/4 - 2 1/2 Å
Errors, uncertainty

Errors ? – Many
• good data - you cannot tell O from N
• bad data – you may slip by one amino acid
• spectacular errors
  • trace chain backwards, join wrong secondary structure units

Uncertainty
• if I have 2 Å data, what is my uncertainty ?
  Much smaller
  1. averaging over observations
  2. known bonds

Estimate of uncertainty..
coordinate error - synthetic data

Resolution near 1.5 to 2.0 Å
• errors 0.2 to 0.3 Å


two different estimates of coordinate error – not important for us
mobility – more important

2\(\pi\) all atoms

2\(\pi\) backbone

Real uncertainty
• not experimental error

• .. mobility

when mobility is high...
missing atoms

Clear pieces of structure missing

Look at $B$-factors
<table>
<thead>
<tr>
<th></th>
<th>NMR</th>
<th>X-ray</th>
</tr>
</thead>
<tbody>
<tr>
<td>certainty</td>
<td>spread amongst 50 models</td>
<td>B-factors / Gaussian model</td>
</tr>
<tr>
<td>resolution</td>
<td>no meaning</td>
<td>well defined</td>
</tr>
<tr>
<td>size</td>
<td>rarely &gt; 200 residues</td>
<td>big</td>
</tr>
</tbody>
</table>
not for discussion – make sure the ideas are clear

* Forming crystals is a question of energy differences
* Simple refraction, one dimension
* generalising to 3D not discussed
* from reflections to density via Fourier Transform
* phasing methods
* high and low resolution reflections
* static vs dynamic disorder
* fitting, overfitting
* R and R free
* B-factors and missing atoms