NMR vs X-ray, precision, certainty

Main methods
• X-ray crystallography and NMR

Others
• cryo-electron microscopy (cryo EM)
• small-angle X-ray scattering (SAXS)
• neutron diffraction

• Dominated by proteins, but most comments apply to nucleotides
Techniques for structures

Mostly X-ray

- NMR
- cryo EM
- neutron diffraction
- ...

X-ray
Structure solving techniques

X-ray  89 %
NMR   9 %
cryo-EM 2 %
nobel prize 2017

Can you combine methods?
• X-ray + NMR - rare
• X-ray + cryo-EM – more common
• low and high-resolution X-ray – sometimes

Why focus on X-ray and NMR?
• emphasis in this course on atomistic detail
• still most important
cryo-EM and SAXS

Why will I not speak about cryo-EM?

• fashionable, but look at resolution

• distance between two residues \((C_i^\alpha, C_{i+1}^\alpha) = 3.8 \text{ Å}\)

• cryo-EM cannot tell which residue is which

• getting better every year

• not quite atomic detail

SAXS – even less detail

SAXS = small angle X-ray scattering
Genauigkeit

Why do I care about accuracy?

- What is a bond length? (1.07, 1.54, 1.32 .. Å) easy
- How does the energy change as I move an atom?
- I want to understand protein-ligand binding
  - where is my ligand?
  - with which residues does it interact?
  - can I predict the effect of a mutation / substitution?

A line from the protein data bank

| ATOM | 41  | N   | ASP A  | 3   | 35.790 | 11.466 | -9.466 | 1.00 | 16.15 |

\[ x = 35.790, \ y = 11.466, \ z = \ldots \quad \text{total fantasy (10^{-13} m)} \]
Error definitions

- Accuracy, precision – not good words
- Certainty / uncertainty / confidence – more in statistics
- Resolution – nice word – defined later

What do I mean by error? $x \pm y$
- should say 90% confidence, one $\sigma$, 75 % quartiles, ..

How do I interpret this?
- I imagine a Gaussian (normal) distribution
accuracy / precision

basically bad words

- do not use Wikipedia + Übersetzung schwierig

small error
- systematic error
- large random error

repetitions do not help
- repetitions increase certainty

if you know it, you would correct for it
- usually modelled with a gaussian
Systematic Errors – are they relevant?

Hopefully not too much

• X-ray – very small effects – perhaps present in old structures
  • not all programs use exactly the same references for bond lengths / angles
• NMR – distances
  • if you only use upper bounds are you changing the distributions?
  • error in calibrating NOE → distance conversion
    • all distances will be too large / small
• NMR – calculation of structures
  • some methods produce more compact structures

Should not be much of a problem in modern data
Why is the Gaussian distribution sacred?

Random numbers (noise, errors)
• take uniform random numbers from 0 to 1
• add a few dozen together and get the sum
• repeat many times
• the sums are normal (Gaussian) distributed around $\frac{1}{2}$

If I have a process which is genuinely random
• best modelled with a Gaussian

Are errors always Gaussian? No – more later
• Errors from your growth estimations/spectrometer?
  • No, but probably a good starting point

Atomic coordinates?
• this lecture
measured data → resolution → electron density → coordinate error and other error

X-ray - fitting structure to data
Resolution

Do we know the error in X-ray coordinates?
• no

Do we know the resolution?
• yes
  • property of crystal and reflections one measures

What does resolution $r_{res}$ mean?
• distance $r_{ij}$ between two points $i, j$
• If $r_{ij} < r_{res}$
  I cannot resolve two points – they look like one object
Resolution

1.2 \times 10^5 structures

75\% of data

1\frac{3}{4} - 2\frac{1}{2} \text{ Å}

are they getting better?
all X-ray structures

structures since 2016
X-ray resolution

Cannot say if they get better
• old structures only get updated if resolution improves
• new (big) complexes are solved that could not be before (low resolution)

If I have 1.5 Å resolution are my coordinates only known to 1.5 Å? No
• I have many reflections – many estimates of position
• I add much chemical information (bond lengths, angles)

What is the error really (simulated data)?
For resolution near 1.5 to 2.0 Å

- I have errors around 0.2 to 0.3 Å

Mobility

• We have uncertainty – from resolution, incomplete data
• we also have mobility

• no matter how good the data is
  the positions of atoms are not fixed
B-Factors

2ei5
all atoms

2ei5
backbone
B-factors

red – blue / mobile less mobile
• surface more mobile / core fixed

Formal meaning

\[ B = 8 \pi^2 u^2 \]

say \( u \) is the average displacement

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

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

units? \( \text{Å}^2 \)

• there are different kinds of \( B \)-factors
Types of $B$-factors

How reliable / meaningful?

- the less certain the coordinates, the larger the $B$-factor (part of fitting – automatic – not done by hand)
- different programs give different values

Is $8\pi^2 u^2$ OK?

- in one dimension?

ask me where Gaussian form comes from - not for Klausur – harmonic model
Anisotropic $B$-factors

How does an atom in a protein move?

- the middle of a protein is not very symmetric

- we could better describe mobility with more numbers

Big problem

- more numbers needs better, high-resolution data rather rare
### $B$-factors one will meet

<table>
<thead>
<tr>
<th></th>
<th>data necessary</th>
<th>number of parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>every atom anisotrophic</td>
<td>lots</td>
<td>lots</td>
</tr>
<tr>
<td>every atom</td>
<td>normal</td>
<td></td>
</tr>
<tr>
<td>per-residue averaged</td>
<td>poor data</td>
<td>few</td>
</tr>
</tbody>
</table>
very mobile atoms

- $B$-factors: physical meaning for mobility of individual atoms

What else does one see?

- Missing atoms?

- There is not enough electron density seen to place an atom

- Interpretation: the atoms are very mobile
- Usually only in loops, N- and C-termini
X-ray summary

- resolution is well-defined
- coordinate error is less well-defined
- resolution might be $1 \frac{1}{2} \text{Å}$, but coordinate error is much smaller
- mobility puts a lower limit on uncertainty

How does this compare with NMR?
How are NMR structures calculated?

- measure NOEs between H's – convert to distances
  - maybe some angles, chemical shifts, residual dipolar couplings
- distances $\rightarrow$ coordinates (distance geometry)

Distance information is

- not so accurate – often only upper bounds
- limited to short (< 5 Å) distances
- there are many sets of coordinates that fit the data

Solve the distance geometry problem 100 times
- send the best 20 or 30 structures to data bank
NMR coordinate error

- purple – what you see when you open the file
- green – 20 more “models”
more drastic example

- blue first model
- gold/brown : all models
- most of the coordinates not really determined
- angle of first helix not really known
Meaning of models

Interpretation

• Each of the models in the data file agrees with the experimental information
• All of the models are reasonable solutions

Can we take the average?

• what would the average look like?

What do average coordinates generally look like?
You cannot average coordinates

coordinates with normal bond lengths / angles

averaged (A, B) coordinates
silly bonds, angles
Using NMR coordinates

- average may have little meaning
- pick a model of your choice?
  - if the models are good – OK
  - if the models are very different you have a problem

Can one talk about accuracy/certainty?
- If you think the models cover the allowed space
- what is the average distance compared to average coordinates? (root mean square)

- What does one expect?
• take coordinates
• generate possible distances
• delete randomly
• calculate structures / compare to known coordinates

Certainty with NMR

- take set of solutions (20 to 50)
- fit to each other or average
- for each site (maybe C\(^\alpha\) ) calculate root means square difference
- gives estimate at each site of spread

- maybe average over all sites – gives very rough idea of certainty

- Gaussian distribution assumption ? Weak

- compare some features of NMR and X-ray..
Is NMR terrible?

Uncertainty is

- bigger than with X-ray
- less well estimated

There are problems with crystallography

- many proteins never crystallise
- some are difficult to phase
- a synchrotron is much more expensive than an NMR spectrometer
Distribution of errors

• I say I have 2 Å resolution or 2 Å difference between structures or 0.2 Å uncertainty – what does it mean?
• simple / classic error analysis
• if \( y = x_1 - x_2 \) with errors \( y = (x_1 \pm \epsilon_1) - (x_2 \pm \epsilon_2) \)
  • final error is \( \epsilon = \left( (\epsilon_1)^2 + (\epsilon_2)^2 \right)^{\frac{1}{2}} \)
• imagine \( \frac{1}{4} \) Å error on C and N
• final error on \( d_{CN} \)

\[
\epsilon_{d_{CN}} = \left( \left( \frac{1}{4} \right)^2 + \left( \frac{1}{4} \right)^2 \right)^{\frac{1}{2}} = \left( \frac{1}{8} \right)^{\frac{1}{2}} \approx 0.35 \text{Å}
\]

silly. I know that CN bond length is 1.32 Å
What have I done wrong?
• Intuitive – some distances are known and fixed

• Formal statistical – rule only applies to independent errors
  • bonded C and N coordinates are highly correlated

• Does simple error analysis ever apply?

Here yes
but probably not so interesting

Here yes
and probably important
uncertainty is more complicated

Mobility is not evenly distributed

• X-ray B-factors
  • very uneven
  • surface is most mobile
  • long sidechains are very mobile

• NMR
  • uncertainty also reflects mobility
Mistakes – not random, not systematic

X-ray
- usually in fitting atoms into density
- trace chain backwards
- asn and gln – N and O have the same electron density

NMR
- misassignment of peaks
- finding errors?

18.10.2018
finding errors

- a structure is solved again and looks different
- a structure is solved under slightly different conditions
- a very homologous structure is solved

- properties of structures
- are all bond lengths / angles OK?
sizing of chains

NMR
lots of peptides
not many > 200

X-ray
<table>
<thead>
<tr>
<th></th>
<th>X-ray</th>
<th>NMR</th>
<th>cryo-EM</th>
<th>SAXS</th>
</tr>
</thead>
<tbody>
<tr>
<td>resolution</td>
<td>1½ - 2½ Å</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>certainty</td>
<td>&lt; 1 Å</td>
<td>from &lt; Å</td>
<td>&gt; few Å</td>
<td>blobs</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>like for X-ray</td>
</tr>
<tr>
<td>cost</td>
<td>$$$$</td>
<td>$$$</td>
<td>$$$</td>
<td>$$$</td>
</tr>
<tr>
<td>you have protein..</td>
<td>easier if</td>
<td>less</td>
<td>less reliance on</td>
<td></td>
</tr>
<tr>
<td>how difficult is</td>
<td>similar to</td>
<td>known</td>
<td>known</td>
<td></td>
</tr>
<tr>
<td>structure?</td>
<td>known structure</td>
<td>structure</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>