Where do maps come from?

Kevin Cowtan

University of York
Where do maps come from?

From solving the phase problem. Either by:

1. Experimental phasing

or

2. Molecular replacement
The phase problem...
Experimental phasing

Solving the phase problem using experimental data.

- No homologous model required.
- But we need multiple sets of structure factor amplitudes.
- Solve for the *simple* difference structure.
- Differences in amplitudes arising from differences in the scattering density give us clues to the phases.

But we first need to know how to solve the phase problem directly for simple cases.
How do we solve simple structures?

Two approaches:
● The Patterson Function
● Direct Methods
The phase problem: We only measure structure factor amplitudes, not phases.

- The structure factor amplitudes give us information about the spacing between features in the map (through which reflections are strong).

- The phases give us information about where along a particular direction features are present in the map.

So we have spacing information, but are missing the position information.
The Patterson Function

http://www.ysbl.york.ac.uk/~cowtan/sfapplet/sftut7.html
The Patterson Function

• For few atoms, we can work backwards from the spacings and deduce what the structure must be.

• Guess the position of one atom, then try placing another at a spacing given by a Patterson peak, and see if we produce other plausible Patterson peaks.
  - SHELX 'crossword tables'

• Symmetry can help...
The Patterson Function

The vector between the atoms is twice the vector between the first atom and the 2-fold axis.
The Patterson Function

Harker sections:

• Special sections of the Patterson function where the peaks correspond to the distance between symmetry related atoms.

• e.g. a 2-fold along the b-axis leads to Harker peaks on the y=0 section, giving us the x and z coordinates of atoms.
The Patterson Function

The Patterson function is a type of 3-dimensional map which may be constructed in the absence of any phase information, and yet may still reveal some structural information. It is calculated according to the equation:

\[ P(u,v,w) = \sum_{hkl} |F(hkl)|^2 \exp(-2\pi i hu+kv+lw) \]

i.e. it is the Fourier transform of the intensities.

The Fourier coefficients of the Patterson may be considered as the product of the structure factors with their complex conjugates, since \(|F(hkl)|^2 = F(hkl)F(hkl)^*\). By the convolution theorem, the Patterson function must therefore be the convolution of the electron density with its inverse:

\[ P(u,v,w) = \rho(u) \otimes \rho(-u) = \sum_{xyz} \rho(x,y,z) \rho(u+x,v+y,w+z) \]

For small molecules at high resolution, the Patterson function has a peak at every position corresponding to an interatomic vector (~\(N^2\) peaks). For large structures at lower resolutions such a map is uninterpretable, but it may still be possible to identify vectors between heavy atoms, or vectors relating entire molecules in the case of translational Non-crystallographic symmetry.
Direct Methods

Methods for obtaining phase information directly from the observed structure factor magnitudes or intensities, without the need for any additional experimental data, such as a homologous structure or a heavy atom substructure.

Direct methods have traditionally been based around *triplets of strong reflections* or the *tangent formula*.

Direct methods have made the solution of small structures (up to 200 atoms) routine when data is available to atomic resolution (~2Å). More recently, dual space methods have extended these methods to larger numbers of atoms.
Direct Methods

Phase triplets

http://www.ysbl.york.ac.uk/~cowtan/sfapplet/sftut9.html
Direct Methods

For triplets of *strong* reflections whose Miller indices sum to zero, the phases will also sum to approximately zero:

$$\varphi(h) + \varphi(k) + \varphi(-h-k) \approx 0$$
Direct Methods

The Tangent Formula (Karle & Hauptman 1956)

\[
\tan(\phi_h) = \frac{\sum_{h'} |E_{h'}E_{h-h'}| \sin(\phi_{h'} + \phi_{h-h'})}{\sum_{h'} |E_{h'}E_{h-h'}| \cos(\phi_{h'} + \phi_{h-h'})}
\]

All the phase triplets combined into a single equation.
Direct Methods

Putting it together:

- Start with randomly selected phases for a few strong reflections.
- Use phase triplets or the tangent formula to estimate other phases.
- Repeat with different starting phases until a plausible solution is obtained.
What about bigger structures?

We need to take a stepwise approach:

1. Use Patterson or Direct methods to solve a substructure of easily detected atoms.
   - Either heavy atoms, or anomalous scatterers

2. Use the information from the substructure, along with multiple sets of structure factor amplitudes, to infer phases for the rest of the structure.
   - SAD, MAD, SIR, MIR, SIRAS, RIP, RIPAS etc.
Anomalous scattering from an atom occurs when an incident X-ray is absorbed and re-emitted by an atom, leading to a phase shift in the scattered X-ray.

Used in SAD/MAD phasing:

- Single anomalous dispersion (SAD)
- Multiwavelength anomalous dispersion (MAD)
SAD

\[ f(\theta, \lambda) = f_0(\theta) + f'(\lambda) + i f''(\lambda) \]

Dispersive term
Absorption term

Se K edge = 0.9795 Å
Anomalous dispersion

The change in the scattered wave can be represented by a phase shift (which is mathematically equivalent to an imaginary component to the electron density), and an attenuation. The attenuation of the atomic scattering factor $f$ is given the symbol $f'$, and the imaginary component of the atomic scattering factor $f''$. 
Anomalous dispersion

This shift, unlike the phase shift due to the difference in scattering path, is the same of both members of a Bijvoet pair of reflections (i.e. +h and -h), and leads to the Friedel opposite reflections having different intensities and phases.
F(h)
F(-h)

Opposite phase difference $\rightarrow$ opposite phase

$\varphi(-h) = -\varphi(h)$
Suppose one atom scatters anomalously...
Path differences are *opposite* for \(+h\) and \(-h\).

Anomalous phase shifts are *the same* for \(+h\) and \(-h\).
No anomalous...
No anomalous...

\[ |F(h)| = |F(-h)| \]
With anomalous...
With anomalous...

\[ |F(h)| \neq |F(-h)| \]
SAD

|F(h)|

|F(-h)|
SAD

\[ |F(h)| \quad |F(-h)| \]
Many normal scatterers...
SAD

• If we can solve the anomalous scattering substructure, we know the blue arrows in amplitude and phase.

• This means we can determine the phase of the red arrow, and thus the electron density.

• Solve the anomalous structure by using the anomalous differences as fake structure factor amplitudes, and use Patterson or Direct Methods.
Phase ambiguity...
Harker construction...

We can get the unknown red arrows from the known blue arrows and the known amplitudes. Shift the circles by the negative of the known part.
Bimodal phase probability...

Represented by Hendrickson-Lattman coefficients – A, B, C, D
MAD

- Multiwavelength anomalous dispersion.
- Collect data at multiple wavelengths, leading to different \( f', f'' \) values for each set of structure factors.
- Can resolve some of the phase ambiguity, at the cost of more crystal damage.
- Largely superseded by SAD + density modification.
MAD

Harker construction...

Multiple observations with different anomalous contributions give multiple (hopefully consistent) phase indications.
Data collection

We need to measure the differences accurately...

\[ F_+ \quad I \quad I \quad F_- \]

i.e. same crystal, similar dose
Data collection

Inverse beam:

Symmetry equivalent of opposite on one image:
Data collection

In practice:

☑️ Optimize for SAD
Isomorphous replacement

Use multiple crystals with different contents to produce multiple sets of diffraction amplitudes. These can then be used to infer phases.

- Single isomorphous replacement (SIR) – native plus a 'derivative' with additional (heavy) atoms.
- Multiple isomorphous replacement (MIR) – native plus multiple derivatives.
- Single isomorphous replacement with anomalous scattering (SIRAS) – native with no anomalous, derivative with anomalous.
Isomorphous replacement

Advantages:

- Doesn't depend on methionines for Se-Met.
- Heavy atoms can be chosen for strong scattering without tunable source.

Disadvantages:

- Need multiple crystals.
- Additional noise comparing amplitudes across crystals.
SIR

$|F_{\text{native}}|$  

$|F_{\text{derivative}}|$
Harker construction...
MIR

- Add more derivatives.
- Heavy atoms may be in the same places (with different scattering), or in different places.
Harker construction...
Where does this get us?

An initial set of phase estimates.

$\phi, FOM$

$A, B, C, D$
Where does this get us?

An initial map, using centroid map coefficients.

\[
E_{\text{map}}(h) = \text{FOM} \times |F(h)| \times \exp(i \varphi(h))
\]

For a general phase probability distribution, the least noisy map is obtained by weighting using a FOM given by the distance of the centre of mass of the phase probability distribution from the origin.
X-ray structure solution pipeline...
Density Modification

• Traditional density modification: e.g. 'dm', 'solomon', 'parrot', CNS

• Statistical density modification: e.g. 'resolve', 'pirate'
Density modification

Starting point:
- Structure factor amplitudes
- Phase estimates:
  - MR: Unimodel distribution
  - SAD: Biomodal distribution
Density modification

- Density modification is a problem in combining information:
Density modification

1. Rudimentary calculation:

\[ |F|, \varphi \]

\[
\varphi = \varphi_{\text{mod}}
\]

\[ \rho(x) \]

\[ \rho_{\text{mod}}(x) \]

Reciprocal space

Real space

FFT

FFT\(^{-1}\)

Modify \(\rho\)
Density modification

2. Phase probability distributions:

\[ |F|, P(\varphi) \xrightarrow{\text{centroid}} |F_{\text{best}}|, \varphi_{\text{best}} \xrightarrow{\text{FFT}} \rho(x) \]

\[ P(\varphi) = P_{\text{exp}}(\varphi) \times P_{\text{mod}}(\varphi) \]

\[ \text{Modify } \rho \]

\[ |F_{\text{mod}}|, \varphi_{\text{mod}} \xrightarrow{\text{FFT}^{-1}} \rho_{\text{mod}}(x) \]

Reciprocal space

Real space
Density modification

3. Bias reduction (gamma-correction):

\[ |F|, P(\phi) \xrightarrow{\text{centroid}} |F_{\text{best}}|, \phi_{\text{best}} \xrightarrow{\text{FFT}} \rho(x) \]

\[ P(\phi) = P_{\text{exp}}(\phi) \times P_{\text{mod}}(\phi) \]

\[ P_{\text{mod}}(\phi) \xrightarrow{\text{likelihood}} |F_{\text{mod}}|, \phi_{\text{mod}} \xrightarrow{\text{FFT}} \rho_{\gamma}(x) \]

Modify \( \rho \)

\( \gamma \)-correct

J.P. Abrahams
Density modification

4. Maximum Likelihood H-L:

\[ |F|, P(\phi) \rightarrow |F_{\text{best}}|, \phi_{\text{best}} \rightarrow \text{FFT} \rightarrow \rho(x) \rightarrow \text{Modify } \rho \rightarrow \rho_{\text{mod}}(x) \rightarrow \gamma\text{-correct} \rightarrow \rho_{\gamma}(x) \]
Density modification

5. Statistical density modification:

\[ |F|, P(\phi) \]

\[ \Rightarrow \]

\[ |F_{\text{best}}|, \phi_{\text{best}} \]

\[ \Rightarrow \]

\[ \rho(x) \]

\[ P(\phi) = P_{\text{exp}}(\phi) \times P_{\text{mod}}(\phi) \]

\[ \Rightarrow \]

\[ P(\rho(x)) \]

FFT

Infer

Transform distribution

RESOLVE, PIRATE

Real space

Reciprocal space
Density modification

Traditional density modification techniques:
- Solvent flattening
- Histogram matching
- Non-crystallographic symmetry (NCS) averaging
Solvent flattening
Histogram matching

A technique from image processing for modifying the protein region.

- Noise maps have Gaussian histogram.
- Well phased maps have a skewed distribution: sharper peaks and bigger gaps.

Sharpen the protein density by a transform which matches the histogram of a well phased map.

Useful at better than 4A.
Non-crystallographic symmetry

- If the molecule has internal symmetry, we can average together related regions.
- In the averaged map, the signal-noise level is improved.
- If a full density modification calculation is performed, powerful phase relationships are formed.
- With 4-fold NCS, can phase from random!
Non-crystallographic symmetry

Crystallographic:
- Aligned 2-fold
- Aligned 6-fold

Non-crystallographic:
- Unaligned 2-fold
- Aligned 5-fold
Non-crystallographic symmetry

Useful terms:
• Proper and improper NCS: (closed and open)

• Multi-domain averaging:

• Multi-crystal averaging:
Non-crystallographic symmetry

- How do you know if you have NCS?
  - Cell content analysis – how many monomers in ASU?
  - Self-rotation function.
  - Difference Pattersons (pseudo-translation only).

- How do you determine the NCS?
  - From heavy atoms.
  - From initial model building.
  - From molecular replacement.
  - From density MR (hard).

- Mask determined automatically.
Density modification using Parrot

Job title
- Estimate solvent content from sequence.
- Get NCS from heavy atoms.
- Get NCS from MR/partial model.

Data for (unsolved) work structure:
- Work SEQ in: PROJECT
- Work MTZ in: PROJECT
- FP: SIGFP
- HLA: HLB
- HLC: HLD
- Use Free-R flag: [ ]
- Use map coefficients: [ ]
- Use PHI/FOM instead of HL coefficients: [ ]

Results for work structure:
- Work MTZ out: PROJECT
- Output column label prefix: parrot

Options
- Number of cycles of phase improvement to run: 3

Optional parameters

Run
Save or Restore
Close
Parrot: simple vs NCS averaged

Map correlations compared with and without NCS averaging.
Model Building

Model building software:

- Proteins:
  - Buccaneer
  - ARP/wARP
  - Phenix autobuild

- Nucleic acids:
  - Nautilus/Coot
  - ARP/wARP
  - Phenix autobuild
Buccaneer: Method

Compare simulated map and known model to obtain likelihood target, then search for this target in the unknown map.

Reference structure:

Work structure:
Buccaneer: Method

- Compile statistics for reference map in 4A sphere about Ca => LLK target.
- Use mean/variance.
Buccaneer

Use a likelihood function based on conserved density features.
The same likelihood function is used several times. This makes the program very simple (<3000 lines), and the whole calculation works over a range of resolutions.

**Finding, growing:** Look for C-alpha environment

(4.0A sphere about Cα)

**Sequencing:** Look for C-beta environment

(5.5A sphere about Cβ)

ALA  CYS  HIS  MET  THR  ...  x20
Buccaneer

10 stages:

- Find candidate C-alpha positions
- Grow them into chain fragments
- Join and merge the fragments, resolving branches
- Link nearby N and C terminii (if possible)
- Sequence the chains (i.e. dock sequence)
- Correct insertions/deletions
- Filter based on poor density
- NCS Rebuild to complete NCS copies of chains
- Prune any remaining clashing chains
- Rebuild side chains
Buccaneer

Case Study:

A difficult loop in a 2.9A map, calculated using real data from the JCSG.
Find candidate C-alpha positions
Grow into chain fragments
Join and merge chain fragments
Sequence the chains
Correct insertions/deletions
Prune any remaining clashing chains
Comparison to the final model
Buccaneer: Results

Model completeness not very dependent on resolution:
Buccaneer: Results

Model completeness dependent on initial phases:
Buccaneer

Chain tracing/refinement using Buccaneer/Refmac

Job title

Data for (unsolved) work structure: (Note: perform phase improvement/density modification first)

Specify an initial model to be extended.

Work SEQ in PROJECT

Work MTZ in PROJECT

FP

HLA

HLC

Free R flag

Use Free-R flag: Use map coefficients: Use PHI/FOM instead of HL coefficients:

Work PDB out PROJECT buccaneer.pdb

Options

Number of cycles of building/refinement to run: 3

Buccaneer parameters

Refmac parameters

Run

Save or Restore

Close
Buccaneer

What you need to do afterwards:

• Tidy up with Coot.
  – Or ARP/wARP when resolution is good.
  – Buccaneer+ARP/wARP better+faster than ARP/wARP.

• Typical Coot steps:
  – Connect up any broken chains.
  – Use density fit and rotamer analysis to check rotamers.
  – Check Ramachandran, molprobity, etc.
  – Add waters, ligands, check un-modeled blobs..
  – Re-refine, examine difference maps.
Buccaneer: Summary

A simple, (i.e. MTZ and sequence), very fast method of model building which is robust against resolution.

User reports for structures down to 3.8A when phasing is good.

Results can be further improved by iterating with refinement in refmac (and in future, density modification).

Proven on real world problems.

Use it when resolution is poor or you are in a hurry. If resolution is good and phases are poor, then ARP/wARP may do better. Best approach: Run both!
Nautilus:

● A new tool for nucleic acid model building

● Automated (CCP4i) or interactive (Coot)

● Starting from:
  – Experimental phasing
  – Molecular replacement
  – Protein complexes
Good afternoon Kevin. Welcome to Coot
Acknowledgements

Help:

- JCSG data archive: www.jcsg.org
- Garib Murshudov, Raj Pannu, Pavol Skubak
- Eleanor Dodson, Paul Emsley, Randy Read, Clemens Vonrhein

Funding:

- The Royal Society, BBSRC