Phasing by Molecular Replacement

Bio5325

Spring 2006

Molecular Replacement

Definition: Using phases from a known structure as the initial estimates to phase an unknown protein structure.

We are guessing/hoping that the unknown resembles the known model protein.

Procedure: position/orient known protein in unit cell of unknown to best match experimental diffraction data. Improve model by adding missing pieces and refining atomic parameters to better agree with experiment.

What can go wrong? Errors in the MR model are perfectly correlated with calculated starting phases = **model bias** that is hard to detect and correct.

In contrast, building a model into experimentally phased (heavy atom method) electron density results in errors that are uncorrelated with the starting phases. Improvements in the protein geometry and it's fit to the density will result in a model with more accurate phases that can be combined with experimental phases to increase the accuracy of the electron density.

Conclusion: experimental phases are always preferable to MR phases.

Molecular Replacement

How similar must the unknown/known proteins be in order for MR to succeed?

- Inaccurately placed atoms become more evident at high resolution. At low resolution, the MR model may be reasonably accurate even though it fails to recapitulate high resolution features of the unknown protein.
- Missing atoms contribute equally to error across all resolution. This missing information contributes to noise that obscures the "signal" of a correctly placed/oriented MR model.
- A reasonable MR model might result in a R_{cryst} = 0.45-0.48 prior to model refinement (recall that fully refined models typically have a R_{cryst} = 0.20-0.26). A successful model typically includes an accurate representation of >70% of the unknown structure.

Model Refinement

The initial molecular replacement solution is refined against the experimental data (F_{hkl}'s) to improve model accuracy. New features of the unknown protein (additional side chains, missing segments) will appear in the electron density if the phases are improving. This is the same principle as the difference Fourier used to find "missing" heavy atoms in the isomorphous replacement method.

Full atom refinement may fail if initial model is rough (inaccurate, poorly placed). In this case, the model is far from the true minimum and small random changes in atomic positions sampled during model refinement do not sample the correct solution.

Rigid body refinement of the initial MR solution may provide a more accurate starting point for full atom refinement. Rigid body refinement consists of 3 translational and 3 rotational parameters. We're treating the model as one rigid object. The model can be further divided into domains that are refined as independent bodies (can be linked by "springs" = geometric constraints).

Model Refinement

During model refinement, we are comparing $|F_{obs}|$ (containing experimental errors, contributions from solvent scattering) to $|F_{calc}|$ (Fourier amplitudes of the "perfect protein" in a vacuum).

Solvent scattering/contrast is most evident at low resolution (F_{obs} ~12-9Å), whereas model inaccuracy (F_{calc}) is increasingly evident when comparing higher resolution terms.

Can add a "solvent mask" term to F_{calc}'s to improve agreement with F_{obs} at low resolution. This improves scaling of F_{calc} to F_{obs}.

Placing the MR Model in the Unit Cell of the Unknown Protein

Goal: superimpose each domain of the MR model protein onto homologous domains of the unknown protein.

Test: all possible orientations/positions of the protein in the unit cell.

Target function: calculate agreement between F_{obs} and F_{calc} as model is rotated/translated. Use simple difference $|F_{obs} - F_{calc}|$ or correlation function between observed and calculated (MR model) structure factor amplitudes.

Placing the MR Model in the Unit Cell of the Unknown Protein

Practical: usually need to break the problem into 2 steps. Rotation function (Patterson based vector superposition) sets orientation, followed by a translation function to position the model in the unit cell (recall that the Patterson function superimposes interatomic vectors on a single origin, so translations are lost).

Big problem: more than 1 protein molecule in asymmetric unit of unknown crystal. Too many combinations to test all orientations/positions of multiple molecules in a global search. Modeling this unit cell with a single protein MR model may result in too many “missing atoms” and failure to identify the correct solution.

Placing the MR Model in the Unit Cell of the Unknown Protein

How finely must all possible orientations/translations be sampled?

F_{calc} and F_{obs} must be correlated in the highest resolution shell that is sampled by MR calculations.

At 4 Å resolution, a 1 Å error in atomic coordinates causes a ¼ wave (90 deg.) error in the phases!

For a globular protein having a ~10 Å radius, a rotational error of 5 deg. would correspond to 1 Å in placement of atoms on and around the protein's outer surface.

Thus, candidate rotational orientations must be sampled in 5 deg. increments to obtain a correct solution with <90 deg. phase error for peripheral atoms of the MR model.

It would be computationally (too) expensive to do this fine rotational sampling simultaneously with all possible translations (in <1 Å increments). Full rotation/translation searches (simultaneously) are only practical if we're searching a small region of space that we know contains the correct solution.

The Rotation Function

The Patterson function is a map of all interatomic vectors in the crystal.

A spherical region of the Patterson centered on the origin includes short interatomic vectors, and excludes longer vectors relating atoms in different molecules in the crystal.

The large origin peak (self vectors) of the Patterson function can be subtracted to improve contrast in remaining regions=better signal to noise ratio.

Idea: if 2 structures have some domain in common, then at some resolution, their spherically-cut, origin-subtracted Pattersons maps should have a subset of vectors in common when the structures are properly oriented. (parameters to be optimized are underlined)

The Rotation Function

The Patterson function:

$$P(uvw) = \frac{1}{V}\sum_h\sum_k\sum_l \left|F^2_{hkl}\right| e^{-2\pi i(hu+kv+lw)}$$

Modified Patterson coefficients $|F^2_{hkl}|$ for subtraction of origin peak:

Sharpened coefficients: $E(h) = \dfrac{F(h)}{\langle F(h)\rangle}$ <F(h)> is avg. for resolution shell

Subtraction of origin: $\left(E(h) - 1.0\right) \bullet \langle F(h)\rangle$

(spike at origin = constant)

The Rotation Function

Self-rotation function: both copies of spherically-cut Patterson function come from unknown crystal. The idea is to see if there are multiple NCS-related copies of a protein inside unit cell (largest peaks are caused by crystallographic symmetry).

Cross-rotation function: sample Patterson of known model in different orientations against Patterson of unknown crystal in an attempt to find corresponding orientation of search model.

- Put 1 copy of atomic model into empty box at least 2x the size of the model, in order to avoid overlap with models in neighboring boxes of the "crystal."
- Fourier transform of model => F_{calc} => square to obtain I_{calc} => Fourier inverse => $P_{calc}(u)$ => spherical cut => rotate model and repeat.

The Rotation Function

Sampling of (α,β,γ) during RF depends on size of molecule and resolution. It is common to work at 10-4 Å resolution to determine global orientation without requiring extremely fine sampling of (α,β,γ).

Peaks of Patterson function are about 2x wider than Fourier peaks, making RF solutions inaccurate. Can refine RF solutions by Patterson Correlation (PC) refinement (see Brunger et al.)

A “correct” RF is sometimes distinguished by its high value, but it is common for correct solution to be further down the list of candidate solutions.

Customary to evaluate several RF solutions in subsequent calculations.

High crystallographic symmetry makes the RF noisy because single molecule used as search object represents smaller fraction of total interatomic vectors in unknown crystal.

Translation Function

Assume that we have a list of candidate RF solutions including the correct answer.

For each candidate RF, apply translations to generate every possible position of search molecule (on appropriately fine grid):

- Generate neighboring molecules by applying crystal symmetry.
- Fourier transform the ensemble (calculate F_{calc}).
- Evaluate the TF:

$$\mathrm{TF} = \sum_{hkl} I(h)_{obs}\, I(h)_{calc}$$

- This sum, calculated over all (hkl)s in the resolution range, minimizes the least squares residual between I_{obs} and I_{calc}.

Translation Function

Packing function: a problem with the TF is that F_{calc} becomes erroneously large with overlap of molecules in unit cell.

Parseval's theorum says that $\sum \left| F^2 \right| \propto \sum \rho^2$.

For 2 blocks of density = 1, $\sum \rho^2 = \sum 1$, which is constant no matter where the molecules slide around in the unit cell, except...

When molecules completely overlap, half as many grid points are occupied by 2x density whose square is 4. Thus, $\sum \rho^2$ may be up to 2x too large. Overlapping, symmetry related molecules (illegal) inflate values of ($I_{obs}I_{calc}$) that are summed in calculating the TF.

Solution: look for peaks in the TF that don't violate packing considerations, e.g., those corresponding to maximal volume occupied by protein (no overlap of symmetry-related molecules).