Refinement

Contents

Sources of model

Sources of model

1) Molecular replacement

Problems

- Wrong loops
- Wrong side chains
- Missing domains
- Slightly wrong orientation and position

Advantage

- Starting model with more or less correct sequence

Sources of model

1) Molecular replacement

Problems

- Wrong loops
- Wrong side chains
- Missing domains
- Slightly wrong orientation and position

Advantage

- Starting model with more or less correct sequence

2) Experimental phasing

Problems

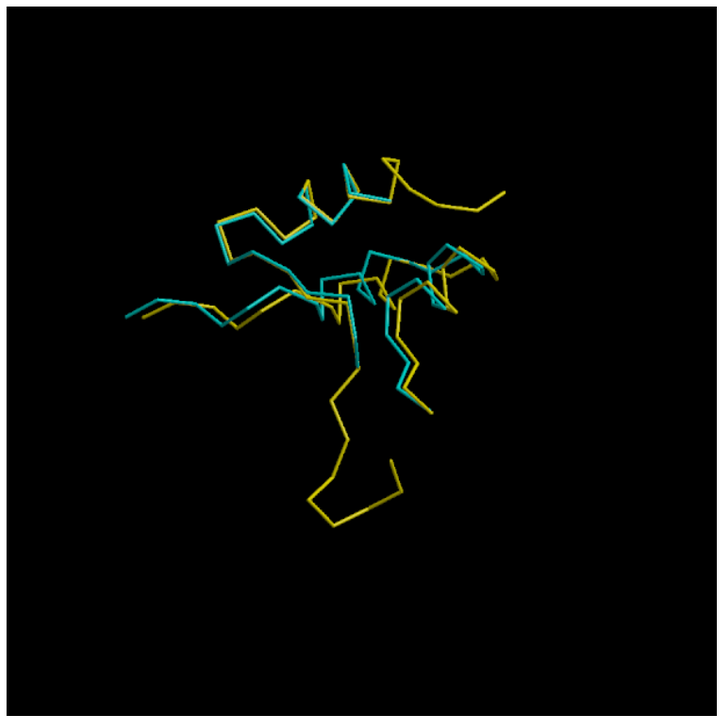
- Incomplete model
- Misfitted fragments

Advantage

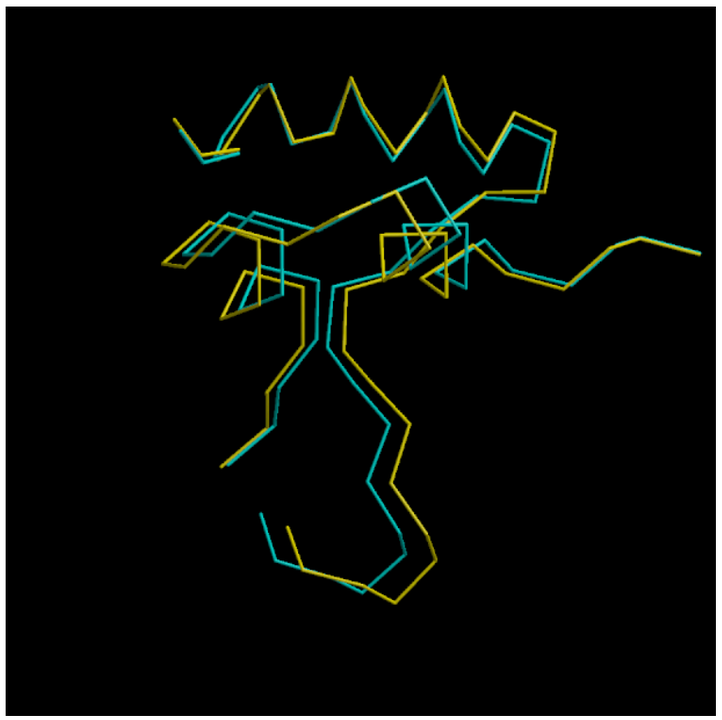
- Phase information is available

Examples of model errors

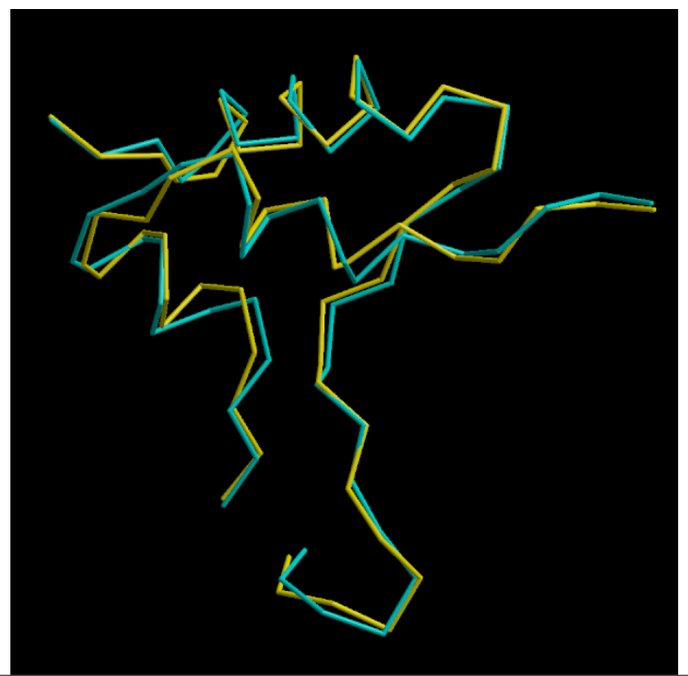
a) Absent atoms, random errors



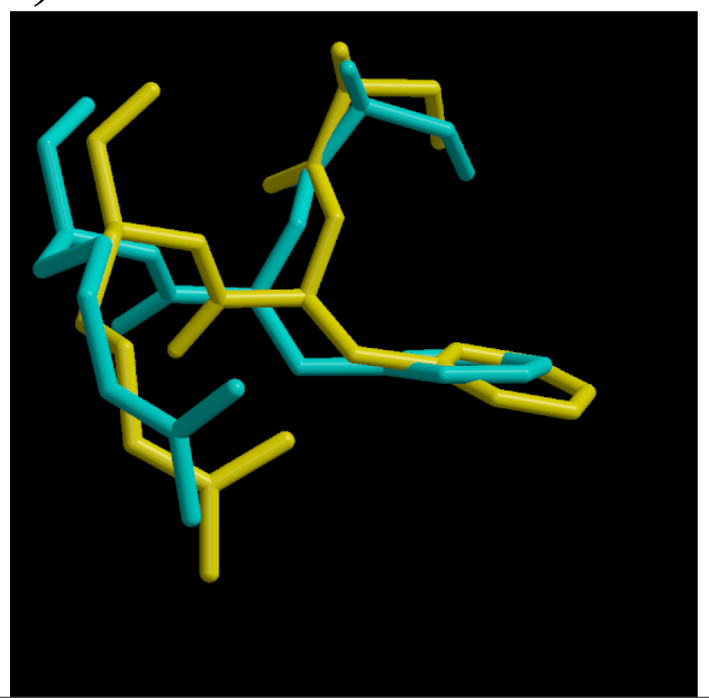
b) Rotation



c) Random errors



d) Random atoms: Close up



Target function

We want to find such model parameters that gives best agreement with the experimental data and at the same time keeps chemical integrity of the model. The total function has a form:

$$f_T(\text{model}) = f_X(\text{data},\text{model}) + w\, f_G(\text{model},\text{restraints})$$

f_T - Total function,

f_X - function that describes fit of model into data. It is minus logarithm of likelihood function.

f_G - function that describes fit of the model into chemical information (bond lengths, angles, chiralities, planarities etc)

w - is a weight

What is likelihood?

What is likelihood?

Likelihood is the probability of observing given data if the current parameter values are correct. In other words likelihood (in probabilistic sense) is agreement between observations and proposed model parameters.

What is likelihood?

Likelihood is the probability of observing given data if the current parameter values are correct. In other words likelihood (in probabilistic sense) is agreement between observations and proposed model parameters.

Maxmimising likelihood means that we want to find such a parameter set that has the best possible agreement with the observations.

Sources of errors and likelihood

Sources of errors and likelihood

There are two source of errors.

Sources of errors and likelihood

There are two source of errors.

- Experimental errors

  Usually approximated using Gaussian distribution

Sources of errors and likelihood

There are two source of errors.

- Experimental errors

  Usually approximated using Gaussian distribution

- Model

  – Errors in positions and B values

  – Misplaced atoms

  – Unaccounted atoms

Sources of errors and likelihood

There are two source of errors.

- Experimental errors

  Usually approximated using Gaussian distribution

- Model

  – Errors in positions and B values

  – Misplaced atoms

  – Unaccounted atoms

Sources of errors and likelihood

There are two source of errors.

- Experimental errors

  Usually approximated using Gaussian distribution

- Model
  - Errors in positions and B values
  - Misplaced atoms
  - Unaccounted atoms

When building likelihood function all these errors should be taken into account.

Model errors

Distribution of complex structure factor is approximated by Gaussian (using Central limit theorem)

$$P(F;Fc) \cong ke(-(F - Fc)^2 / \Sigma) / \Sigma$$

F – true structure factor, Fc calculated structure factor. Both they are complex in general. Σ - is average of square of the differences $<|F - Fc|^2>$

When integrating out phases it becomes Rice distribution or non-central χ^2 distribution with degrees of freedom 1 (centric) and 2 (acentric)

If phase information is available they are added at this stage

Final likelihood

- Final likelihood is:

$$P(|F_o|;F_c) = \int P(|F_o|;|F|)P(|F|;F_c)$$

It can be considered as weighted sum of all probability distribution of observations with weight from another probability distribution. It is generalisation of convolution.

Here various approximations can be used. Least-squares target is one of them. REFMAC uses another approximation. Intensity based least-square as in SHELXL is third approximation

Using phases

If phase information from experimental phasing is available (or other sources) then they could be used in refinement. At early and medium stages of refinement they improve behaviour of refinement and can help to produce better electron density maps.

Usually phase probability distributions are represented using Hendrickson Lattman (HL) coefficients.

There are two ways of using phase information in refinement:

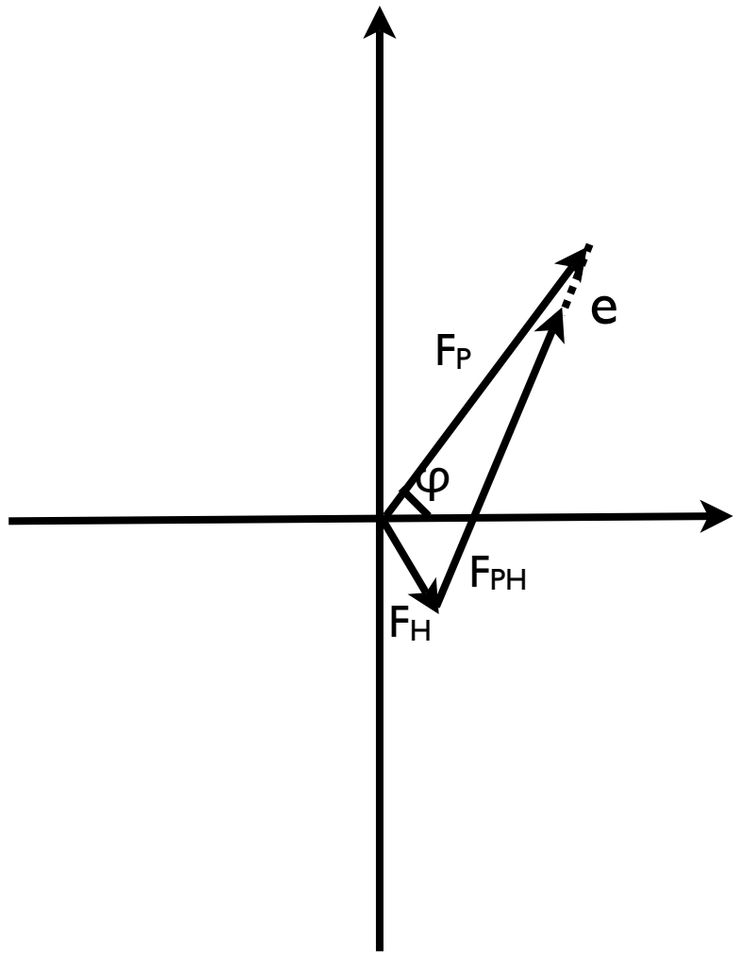
1) After experimental phasing we derive phase probability distribution and use them in refinement (using HL coefficients);

2) Use them directly, i.e. refinement of heavy atoms and protein atoms are carried out simultaneously. At the moment only SAD (available from refmac version 5.5) and SIRAS (will be available from refmac version 5.6) cases have been implemented.

Lack of isomorphism and distributions

$e = |F_{PH}| - |F_P + F_H|$

F_H is known (contribution of heavy atom). $|F_{PH}|$ is known - amplitude of structure factors of derivative, $|F_P|$ is known - amplitude of structure factors of native crystal. We do not know phases of F_P. For each value of phase we can calculate above expression. Probability distribution of phase angle (Blow and Crick) can be calculated (that is how MLPHARE calculate them. Other programs have more sophisticated approaches but essence is the same)



$P(\varphi) = N \exp(-e^2/(2*S\text{^}2))$

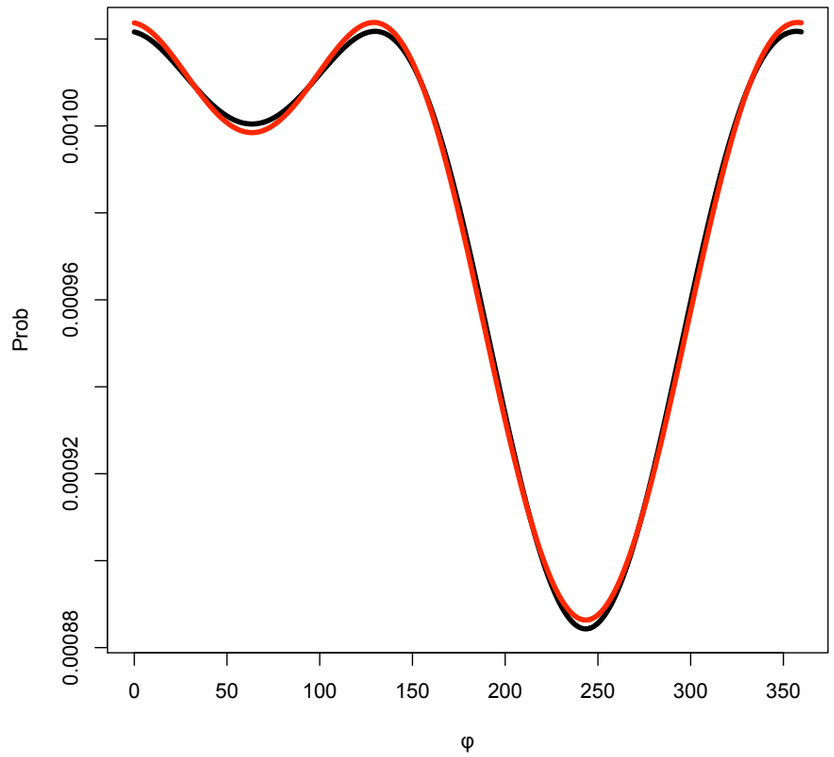
N is normalisation coefficient, S is sigma. It will give use distribution of phases. This distribution is the most important outcome of phasing program. They are used in density modification and model building.

Phase probability distributions

Phase probability distributions are usually approximated using Hendrickson Lattman (HL) coefficients. It means that for whole distribution we need only four parameters. It seems that using HL coefficients retain all features of bimodal phase probability distributions.

Phase probability distribution with HL coefficients

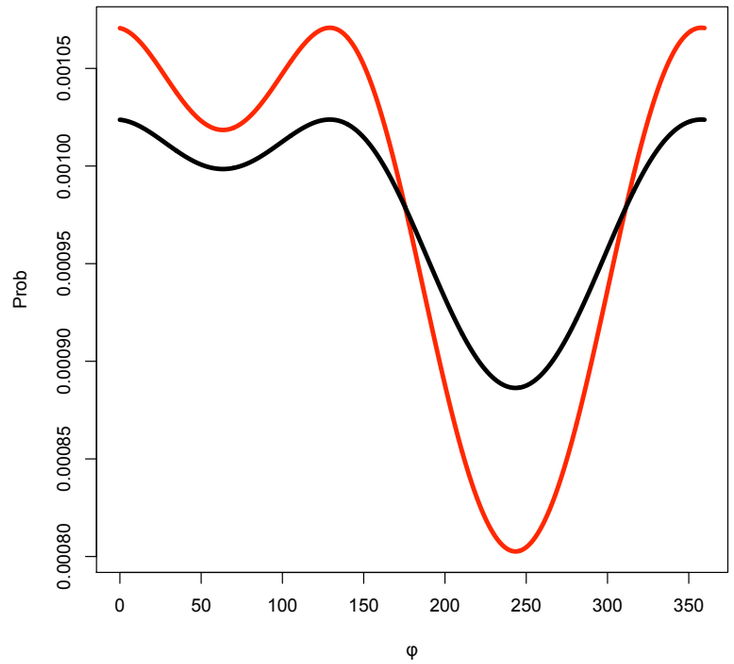
$$P(\varphi)=N\exp(A\cos(\varphi)+B\sin(\varphi)+C\cos(2\varphi)+D\sin(2\varphi))$$



Black line - actual probability distribution calculated using Blow-Crick method.
Red line - Probability distribution calculated using HL coefficients

Phase probability distributions

Phase probability distributions are very important. If they are wrong then later stages of structure solution may be affected seriously.
If two exactly same derivatives are used then the resulting probability distribution may be affected. Probability distribution may become unjustifiable sharp. For example it means that as if phase errors are small when they are not.

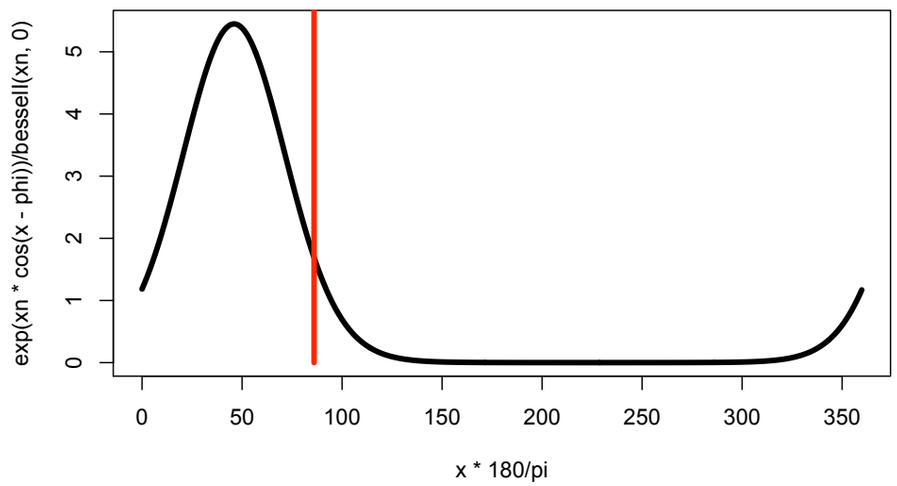


Black line - actual probability distribution
Red line - Probability distribution calculated usi
HL coefficients twice larger than actual ones

Importance of probability distribution

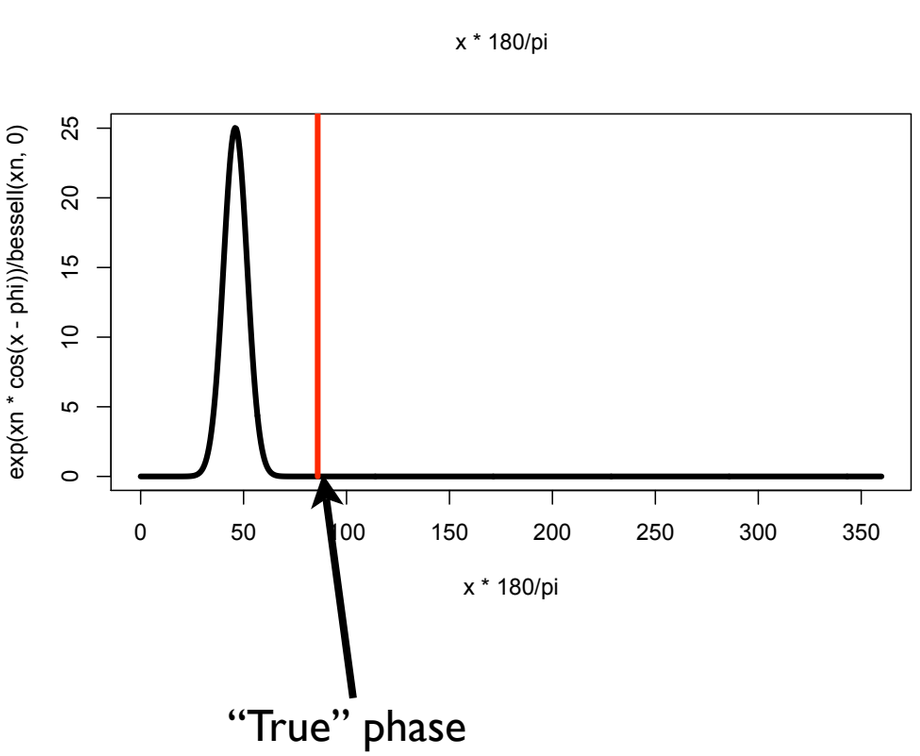
Phase probability distribution with the same centroid and different distribution

a)



If you use distribution a) then refinement will have chance to move atoms so that to achive “true” phases. If you use

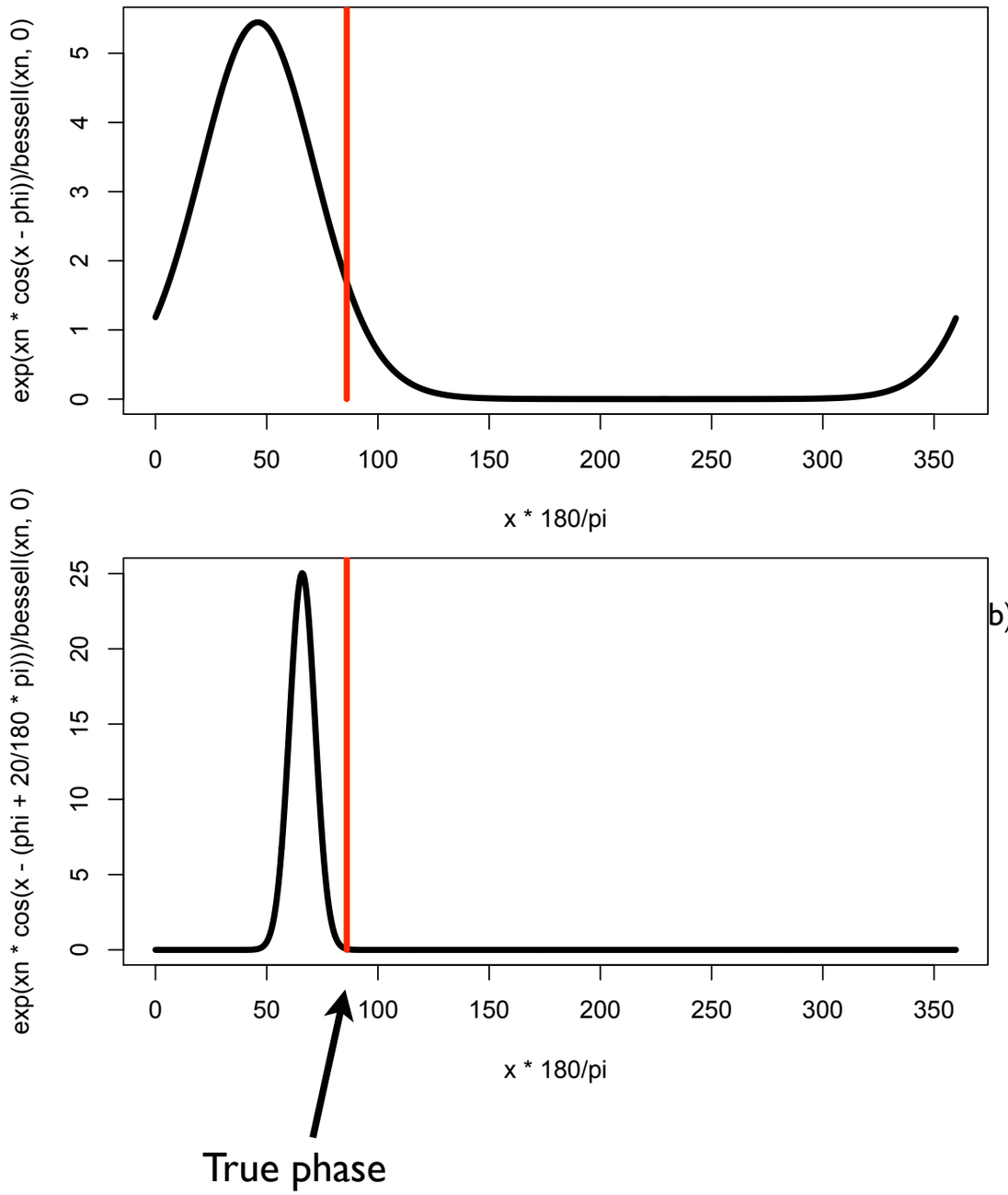
b)



Importance of phase probability distributions

Sometimes it may be more important to derive more accurate probability distribution than more accurate centroid phase (or best phase)

a) has less accurate phase but “true” phase have reasonable probability; b) has more accurate phases but “true” phase has vanishing probability. For map calculation b) might be better but for refinement a) is preferable



Density modification programs produces improved phases and they should be used for initial model building. However as a rule they give very optimistic phase probability distribution.

Shortcomings of using ABCD directly

- Dependent on where you obtained your Hendrickson-Lattman coefficients
- Assumes that your prior phase information is independent from your model phases!

Using phases directly

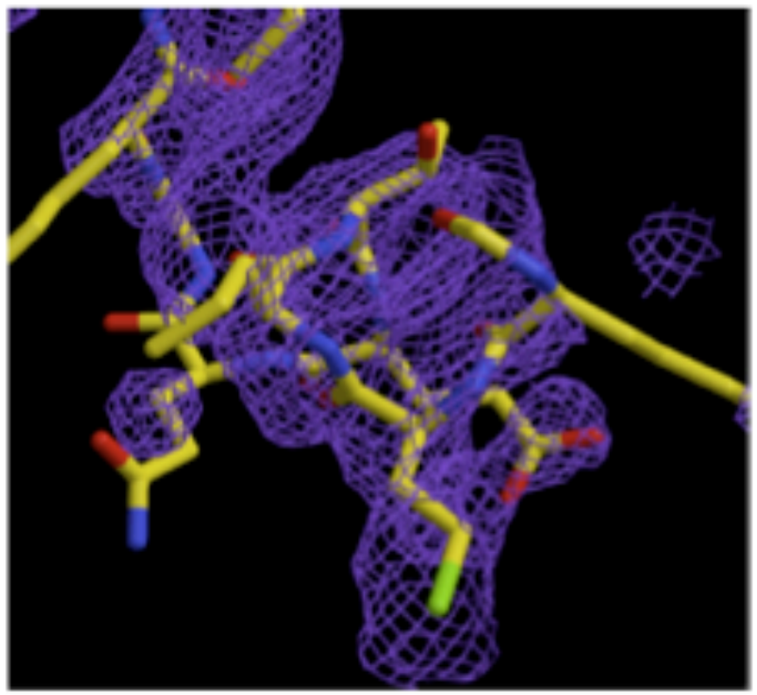
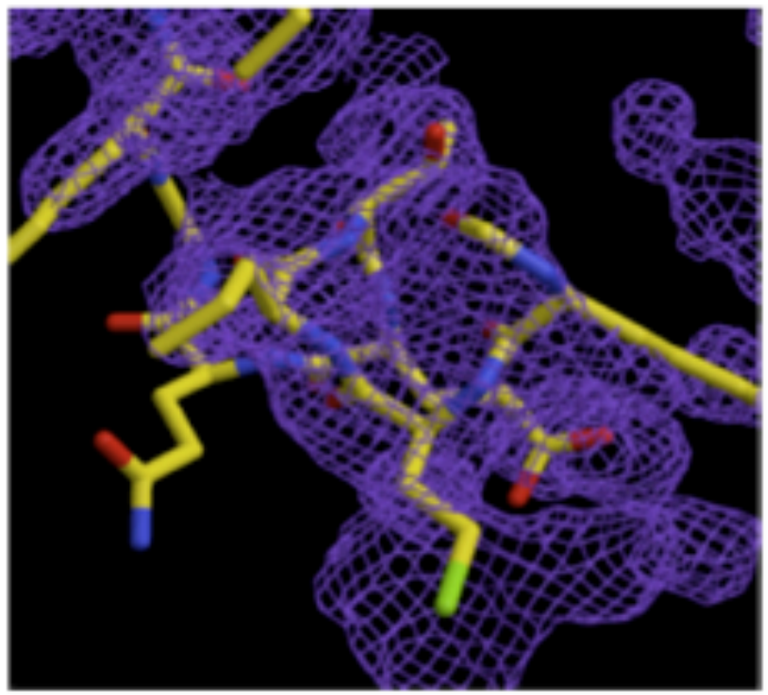
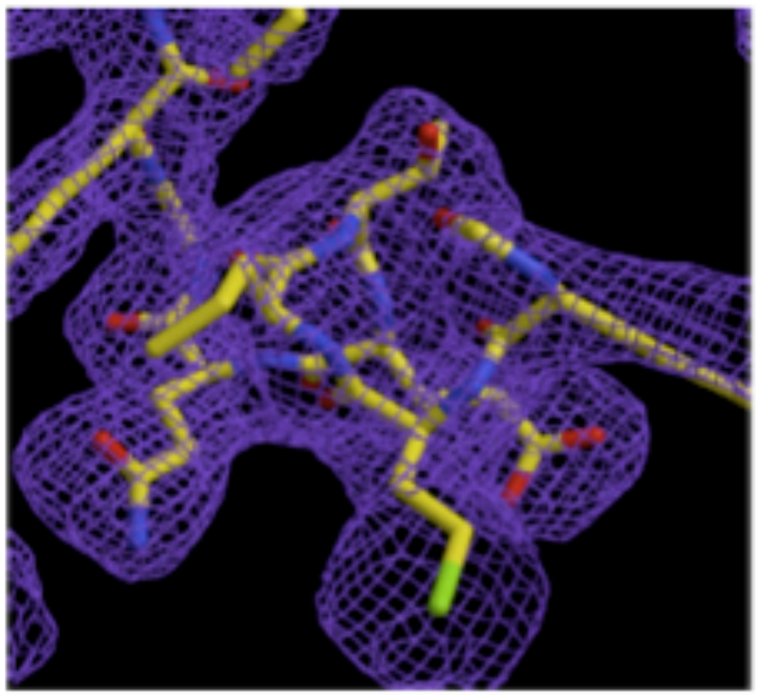
To use all information available in the data (including phase information) it is necessary to derive joint probability distribution of all observations given model parameters. In general it is a tricky problem (computationally very expensive). Pavol Skubak from Leiden has implemented two cases

1) SAD using joint probability distribution of observed Bijvoet pairs

2) SIRAS using joint probability distribution of single derivative with anomalous scatterers and structure factors of native crystals.

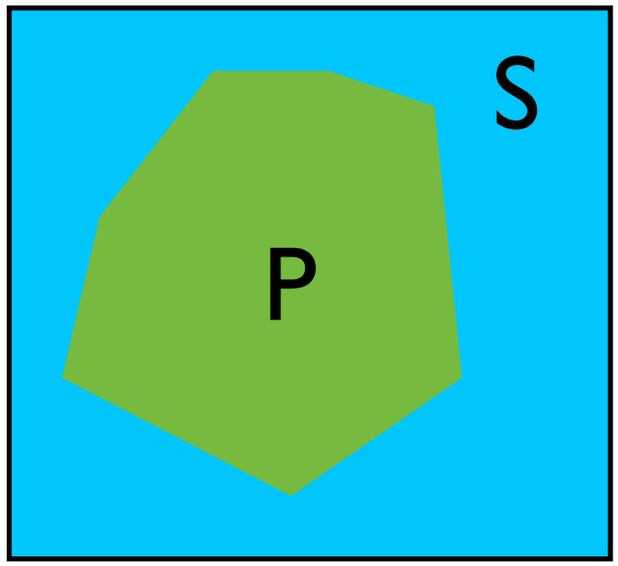
Test cases - example

- transhydrogenase: data to 2.5 Å, SAD on 16 Se, 364 residues in AU
- ARP/wARP model building with different functions for refinement

NO PHASES	INDIRECT (MLHL)	DIRECT (SAD)
		
151 residues built	244 residues built	347 residues built

Bulk solvent
Method 1: Babinet's bulk solvent correction

At low resolution electron density is flat. Only difference is that solvent has lower density than protein. If we would increase solvent just enough to make its density equal to that of protein then we would have flat density (constant). Fourier transformation of constant is zero (apart from F000). So contribution from solvent can be calculated using that of protein. And it means that total structure factor can calculated using contribution from protein only



$\rho_s+\rho_p=\rho_T$ <==> $F_s+F_p=F_T$

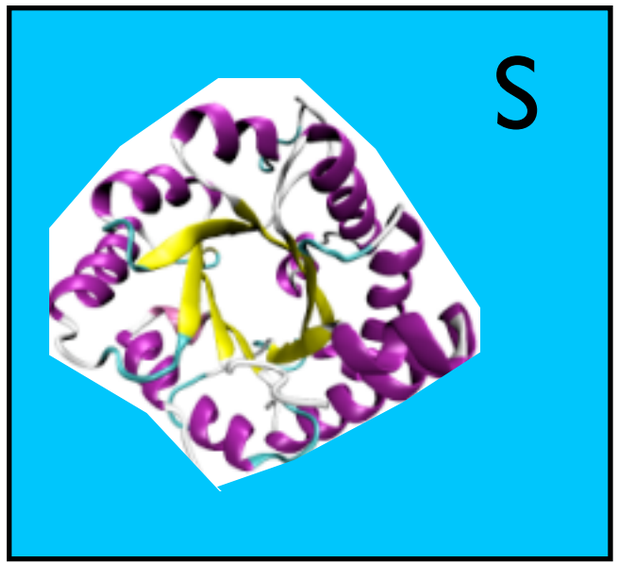
$\rho_s+k\rho_p=c$ <==> $F_s+kF_p=0$

$F_s=-kF_p$ ==> $F_T=F_p-kF_p=(1-k)F_p$

k is usually taken as $k_b \exp(-B_b s^2)$. k_b must be less than 1. k_b and B_b are adjustable parameters

Bulk solvent
Method 2: Mask based bulk solvent correction

Total structure factor is the sum of protein contribution and solvent contribution. Solvent region is flat. Protein contribution is calculated as usual. The region occupied by protein atoms is masked out. The remaining part of the cell is filled with constant values and corresponding structure factors are calculated. Finally total structure factor is calculated using



$$F_T = F_p + k_s F_s$$

k_s is adjustable parameter.

Mask based bulk solvent is a standard in all refinement programs. In refmac it is default.

Parameters

Parameters

Usual parameters

Parameters

Usual parameters

1) Positions x,y,z

Parameters

Usual parameters

1) Positions x,y,z
2) B values – isotropic or anisotropic

Parameters

Usual parameters

1) Positions x,y,z
2) B values – isotropic or anisotropic
3) Occupancy

Parameters

Usual parameters

1) Positions x,y,z
2) B values – isotropic or anisotropic
3) Occupancy

Parameters

Usual parameters

1) Positions x,y,z
2) B values – isotropic or anisotropic
3) Occupancy

Derived parameters

Parameters

Usual parameters

1) Positions x,y,z
2) B values – isotropic or anisotropic
3) Occupancy

Derived parameters

4) Rigid body positional
   - After molecular replacement
   - Isomorphous crystal (liganded, unliganded, different data)

Parameters

Usual parameters

1) Positions x,y,z
2) B values – isotropic or anisotropic
3) Occupancy

Derived parameters

4) Rigid body positional
   - After molecular replacement
   - Isomorphous crystal (liganded, unliganded, different data)
5) **Rigid body of B values – TLS**
   - Useful at the medium and final stages
   - At low resolution when full anisotropy is impossible

Parameters

Usual parameters

1) Positions x,y,z
2) B values – isotropic or anisotropic
3) Occupancy

Derived parameters

4) Rigid body positional
   - After molecular replacement
   - Isomorphous crystal (liganded, unliganded, different data)
5) **Rigid body of B values – TLS**
   - Useful at the medium and final stages
   - At low resolution when full anisotropy is impossible
6) Torsion angles

TLS

TLS

ADPs are an important component of a macromolecule
Proper parameterization
Biological significance

Displacements are likely anisotropic, but rarely we have the luxury of refinining individual aniso-U. Instead iso-B are used.

TLS parameterization allows an intermediate description [Schomaker & Trueblood (1968) Acta Cryst. B24, 63-76].

Schematic decomposition of ADPs

$$U = U_{cryst} + U_{TLS} + U_{int} + U_{atom}$$

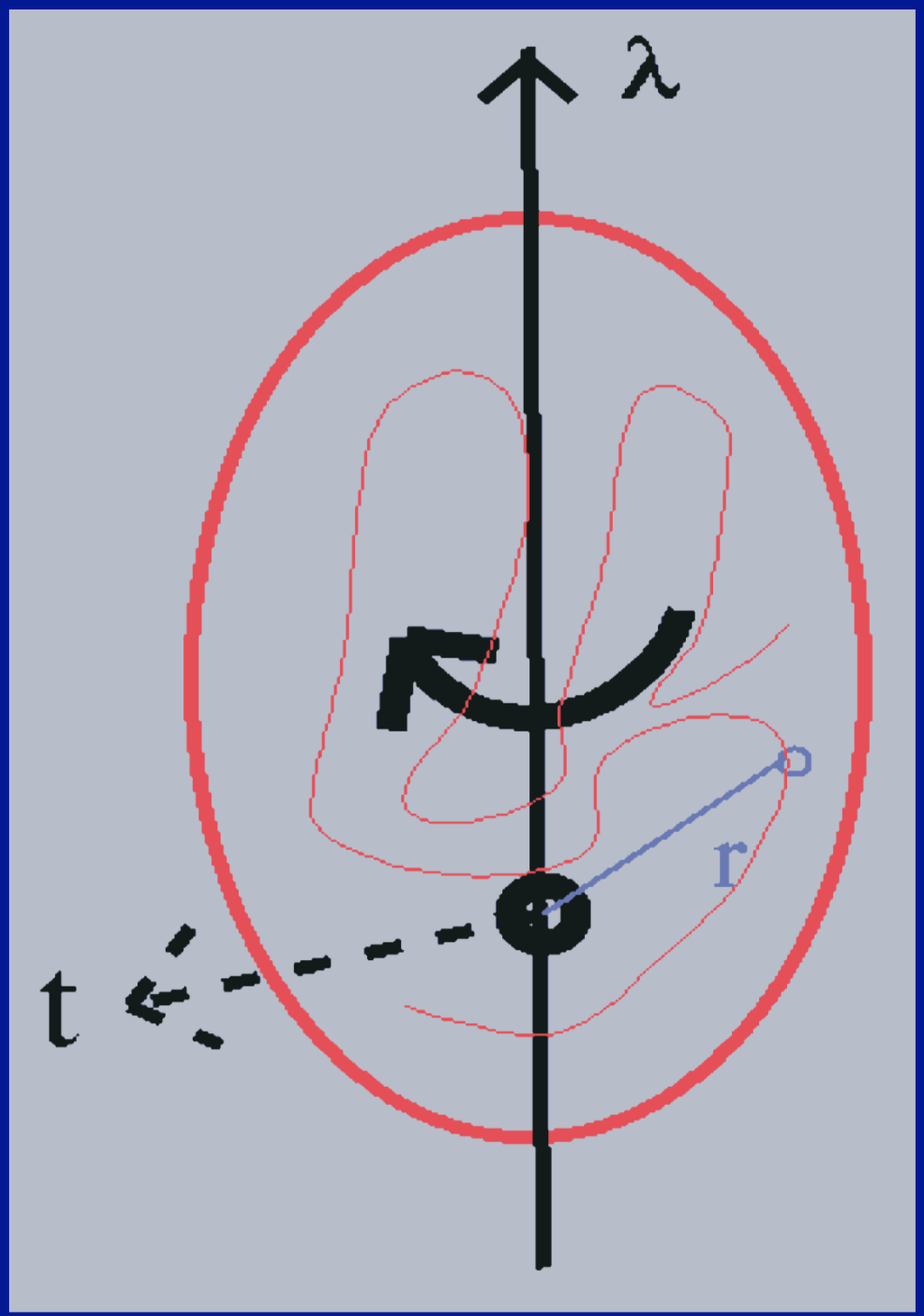
U_{cryst} : overall anisotropy of the crystal

U_{TLS} : TLS motions of pseudo-rigidy bodies

U_{int} : collective torsional librations or internal normal modes

U_{atom} : individual atomic motions

Rigid-body motion



General displacement of a rigid-body point can be described as a rotation along an axis passing through a fixed point together with a translation of that fixed point.

$$\underline{u} = \underline{t} + D\underline{r}$$

for small librations

$$\underline{u} \approx \underline{t} + \underline{\lambda} \times \underline{r}$$

D = rotation matrix
λ = vector along the rotation axis of magnitude equal to the angle of rotation

TLS parameters

Dyad product:

$$\underline{\mathbf{u}}\underline{\mathbf{u}}^{T} = \underline{\mathbf{t}}\underline{\mathbf{t}}^{T} + \underline{\mathbf{t}}\underline{\boldsymbol{\lambda}}^{T}\times\underline{\mathbf{r}}^{T} - \underline{\mathbf{r}}\times\underline{\boldsymbol{\lambda}}\underline{\mathbf{t}}^{T} - \underline{\mathbf{r}}\times\underline{\boldsymbol{\lambda}}\underline{\boldsymbol{\lambda}}^{T}\times\underline{\mathbf{r}}^{T}$$

ADPs are the time and space average

$$\mathbf{U}_{TLS} = \langle\underline{\mathbf{u}}\underline{\mathbf{u}}^{T}\rangle = \mathbf{T} + \mathbf{S}^{T}\times\underline{\mathbf{r}}^{T} - \underline{\mathbf{r}}\times\mathbf{S} - \underline{\mathbf{r}}\times\mathbf{L}\times\underline{\mathbf{r}}^{T}$$

$\mathbf{T} = \langle\underline{\mathbf{t}}\underline{\mathbf{t}}^{T}\rangle$ – 6 parameters, **TRANSLATION**

$\mathbf{L} = \langle\underline{\boldsymbol{\lambda}}\underline{\boldsymbol{\lambda}}^{T}\rangle$ – 6 parameters, **LIBRATION**

$\mathbf{S} = \langle\underline{\boldsymbol{\lambda}}\underline{\mathbf{t}}^{T}\rangle$ – 8 parameters, **SCREW-ROTATION**

Example GAPDH

Glyceraldehyde-3-phosphate dehydrogenase from Sulfolobus solfataricus [Isupov et al., (1999) J. Mol. Biol., 291, 651-60].

340 amino acids per chain

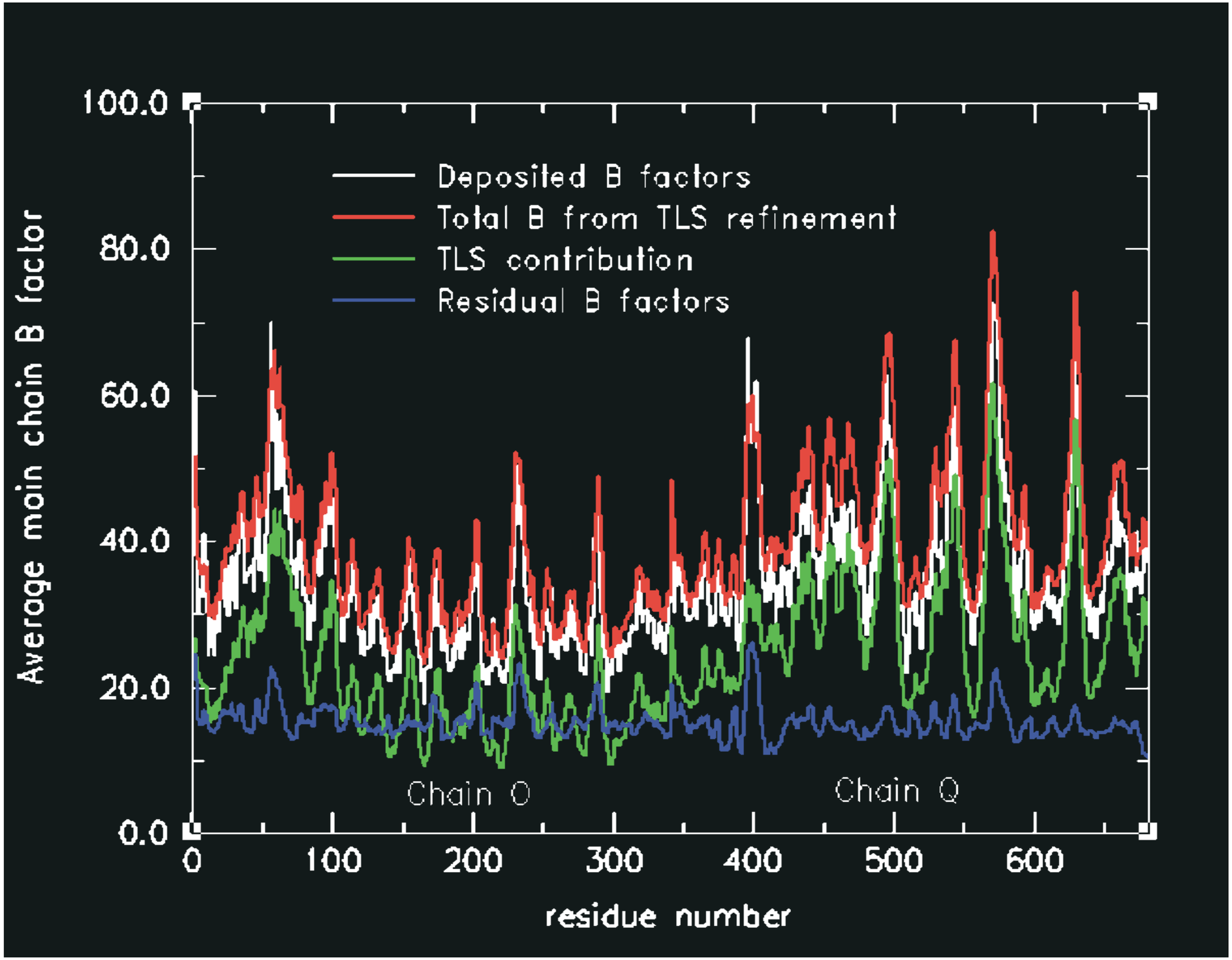
2 chains in asymmetric unit (O and Q), each molecule has NAD-binding and catalytic domains

$P4_12_12$, data to 2.05Å

Refinement GAPDH

Model	TLS	R	R_{free}
ani/rB	0	22.9	29.5
ani/rB	1	21.3	26.8
ani/rB	4	21.1	26.5
ani/20	0	29.5	35.2
ani/20	1	25.1	29.4
ani/20	4	24.4	28.8

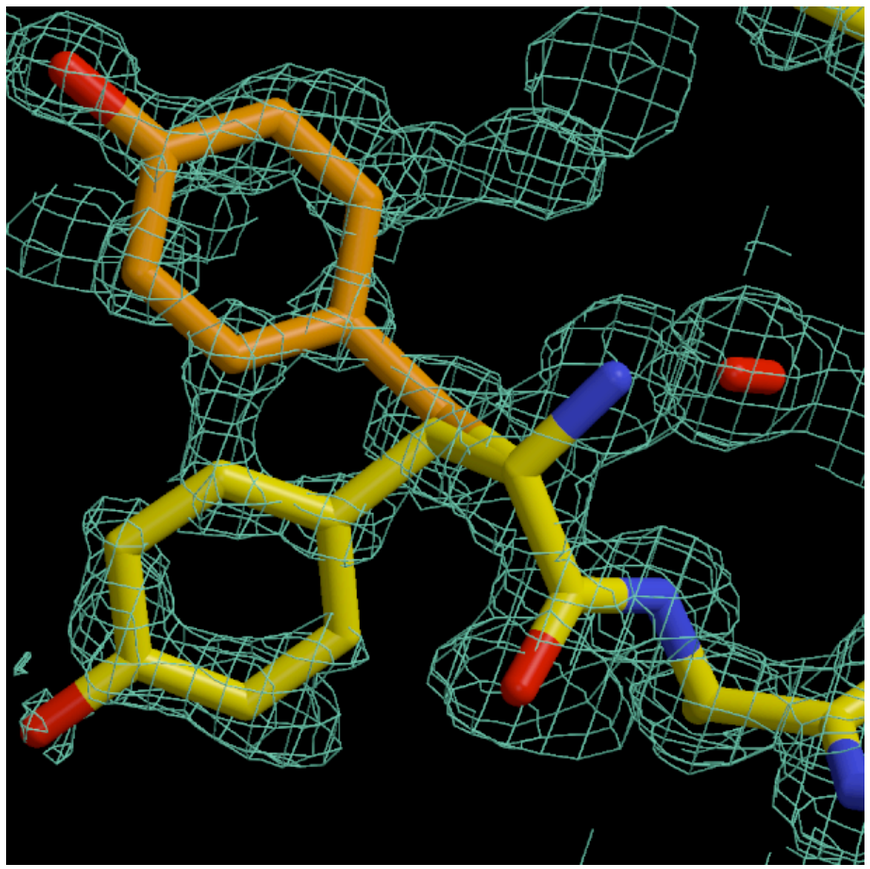
ani = anisotropic scaling
rB = TLS refinement starting from refined Bs; 20 = TLS refinement starting from Bs fixed to 20 $Å^2$


Deposited B factors
Total B from TLS refinement
TLS contribution
Residual B factors
Average main chain B factor
100.0
80.0
60.0
40.0
20.0
0.0
0
100
200
300
400
500
600
Chain O
Chain Q
residue number

Some fine tuning

Alternative conformations

Example from 0.88Å catalase structure:Two conformations of Tyrosine. Ring is clearly in two conformation. To refine it properly CB also needs to be split. It helps adding hydrogen atom on CB and improves restraints in anisotropic U values

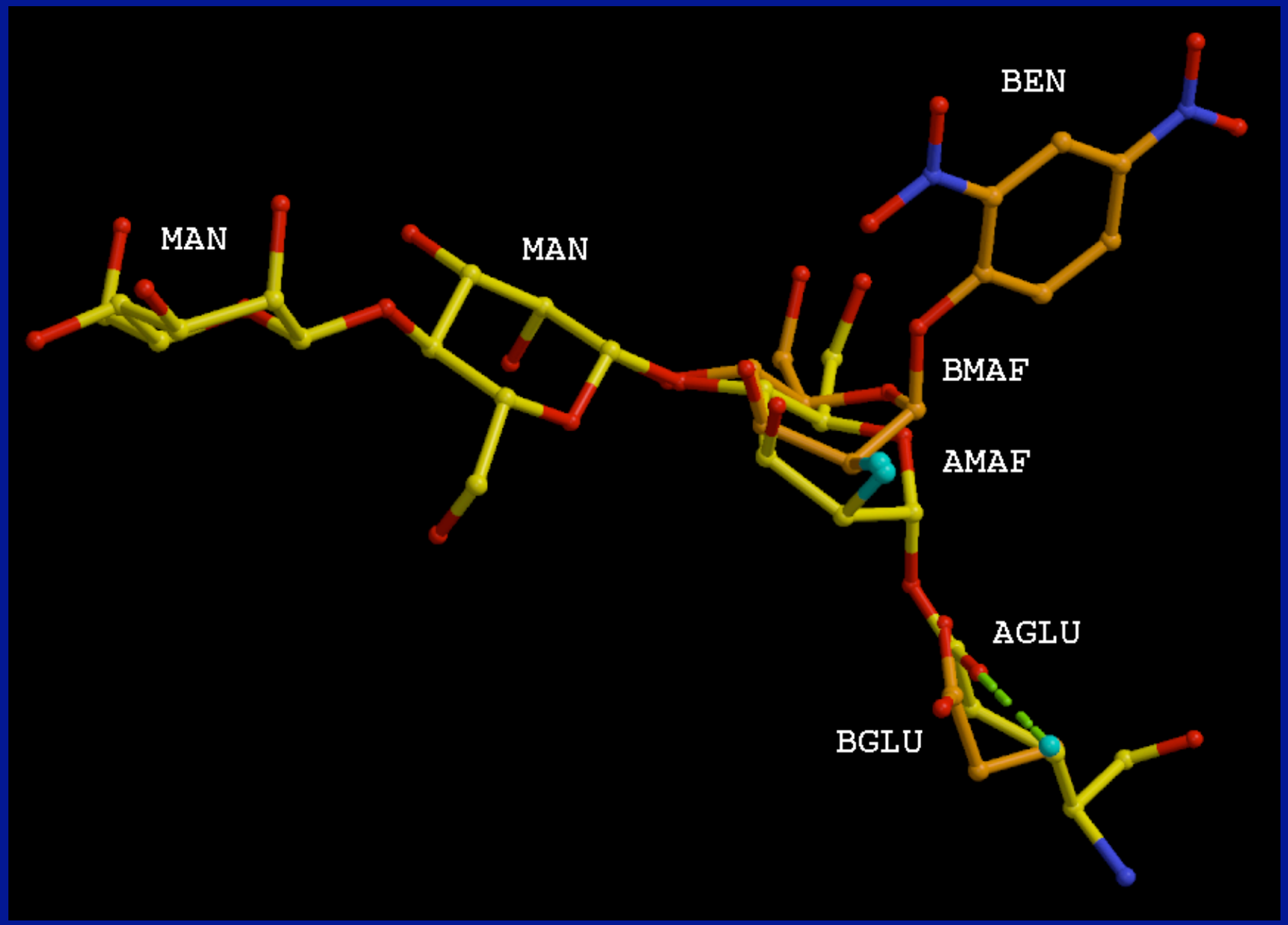


Alternative conformation: Example in pdb file

```
ATOM    977  N   GLU A  67     -11.870   9.060   4.949  1.00 12.89      N
ATOM    978  CA  GLU A  67     -12.166  10.353   4.354  1.00 14.00      C
ATOM    980  CB AGLU A  67     -13.562  10.341   3.738  0.50 14.81      C
ATOM    981  CB BGLU A  67     -13.526  10.285   3.654  0.50 14.35      C
ATOM    986  CG AGLU A  67     -13.701   9.400   2.573  0.50 16.32      C
ATOM    987  CG BGLU A  67     -13.876  11.476   2.777  0.50 14.00      C
ATOM    992  CD AGLU A  67     -15.128   9.179   2.134  0.50 17.17      C
ATOM    993  CD BGLU A  67     -15.237  11.332   2.110  0.50 15.68      C
ATOM    994  OE1AGLU A  67     -15.742  10.153   1.644  0.50 20.31      O
ATOM    995  OE1BGLU A  67     -15.598  12.213   1.307  0.50 16.68      O
ATOM    996  OE2BGLU A  67     -15.944  10.342   2.389  0.50 18.94      O
ATOM    997  OE2AGLU A  67     -15.610   8.027   2.235  0.50 21.30      O
ATOM    998  C   GLU A  67     -12.110  11.473   5.386  1.00 13.40      C
ATOM    999  O   GLU A  67     -11.543  12.528   5.110  1.00 12.98      O
```

Covalent link between residues in double conformation

Fluro-modified sugar MAF is in two conformation. One of them is bound to GLU and another one is bound to ligand BEN



Alternative conformation of links: how to handle

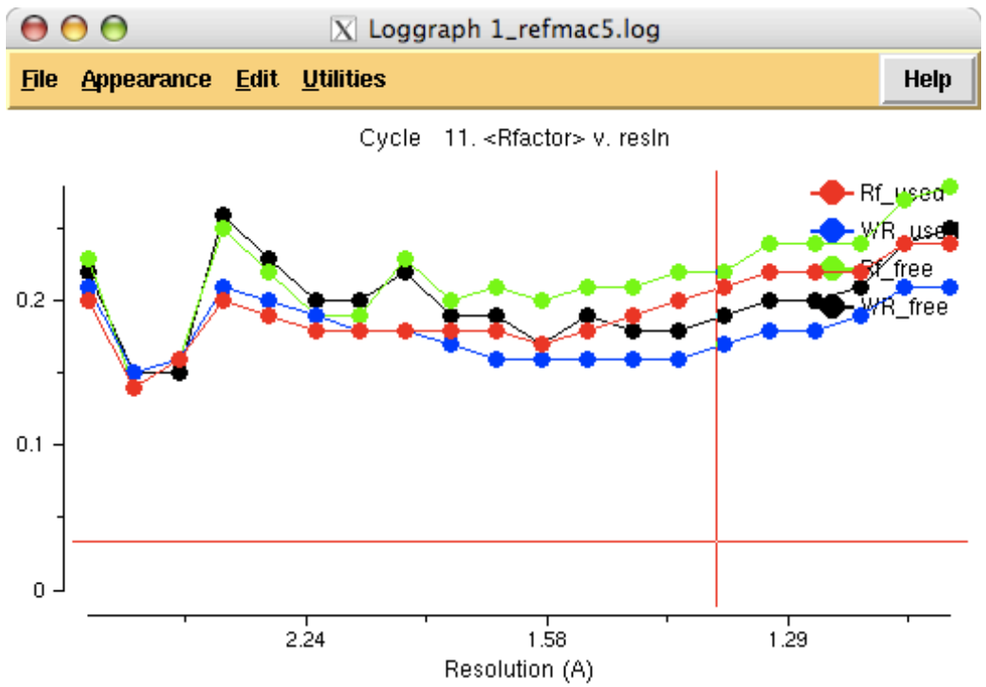
Description

Description of link(s) should be added to the library. When residues make link then each component is usually modified. Description of Link should contain it also

PDB

```
LINK         C6   BBEN B    1                   O1   BMAF S    2    BEN-MAF
LINK         OE2  AGLU A  320                   C1   AMAF S    2    GLU-MAF
```

Behaviour of R/Rfree, average Fobs vs resolution should be reasonable. If there is a bump or it has an irregular behaviour then either something is wrong with your data or refinement.



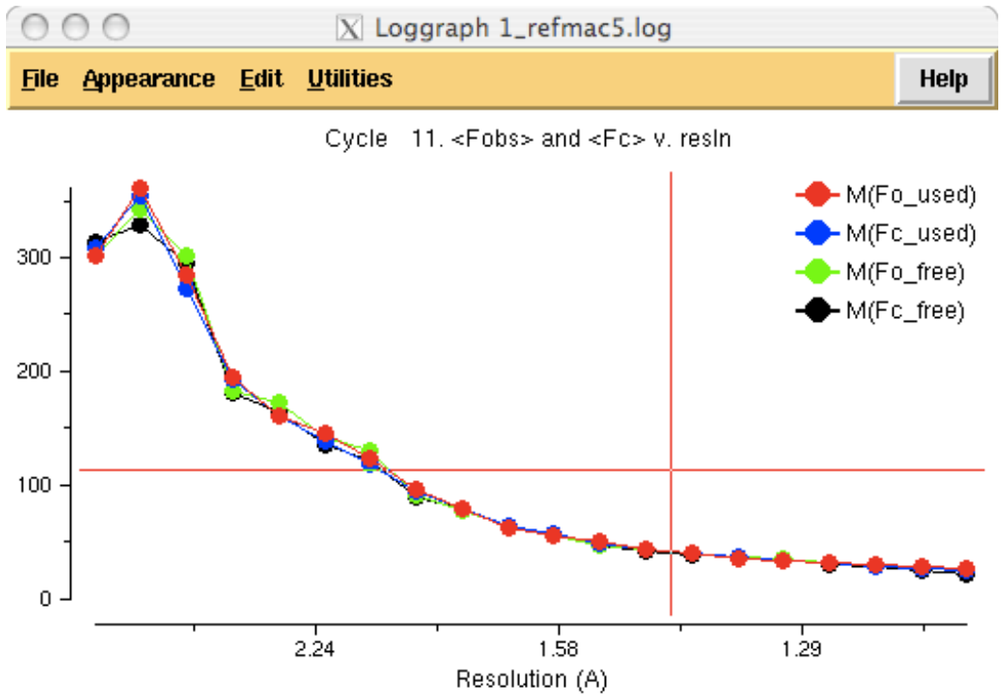
1.36,0.034

Tables in File

Cycle 10. Rfactor analysis, F distribution v resln
Cycle 10. Fom(<cos(DelPhi)>-acentric, centric, overall v resln
Cycle 11. Rfactor analysis, F distribution v resln
Cycle 11. Fom(<cos(DelPhi)>-acentric, centric, overall v resln
Rfactor analysis, stats vs cycle

Graphs in Selected Table

Cycle 11. <Rfactor> v. resln
Cycle 11. <Fobs> and <Fc> v. resln
Cycle 11. % observed v. resln



1.43,113.9

Tables in File

Cycle 10. Rfactor analysis, F distribution v resln
Cycle 10. Fom(<cos(DelPhi)>-acentric, centric, overall v resln
Cycle 11. Rfactor analysis, F distribution v resln
Cycle 11. Fom(<cos(DelPhi)>-acentric, centric, overall v resln
Rfactor analysis, stats vs cycle

Graphs in Selected Table

Cycle 11. <Rfactor> v. resln
Cycle 11. <Fobs> and <Fc> v. resln
Cycle 11. % observed v. resln

What and when

- Rigid body: At early stages - after molecular replacement or when refining against data from isomorphous crystals
- TLS - at medium and end stages of refinement at resolutions up to 1.7-1.6A (roughly)
- Anisotropic - At higher resolution towards the end of refinement
- Adding hydrogens - Higher than 2A but they could be added always
- Phased refinement - at early and medium stages of refinement
- SAD - at all stages(?)
- Twin - always
- Ligands - as soon as you see them
- What else?