MRC Medical Research Council

Refinement against cryo-EM maps

Garib Murshudov

MRC-LMB, Cambridge, UK

Contents

About REFMAC

Fit into EM maps

Structure based restraints

Composite map refinement

Overfitting

Effect of oversharpening

Brown et al, (2015) Tools for macromolecular model building and refinement into electron cryo-microscopy reconstructions. *Acta Cryst*. D71, 136-153

Murshudov (2016) Refinement of Atomic Structures Against cryo-EM Maps, *Methods in Enzymology*. Ed. Crowter, V579, 277-306

Useful websites

ccpem: http://www.ccpem.ac.uk/

ccpem website. In future all our developments about EM will be distributed on this website. They also develop GUI for EM tools.

coot: http://www2.mrc-lmb.cam.ac.uk/personal/pemsley/coot/

Coot website by Paul Emsley. All coot related tools are available from this website

refmac: Available from ccp4 and ccpem websites.

About REFMAC

Refmac is a program for refinement of atomic models into experimental data

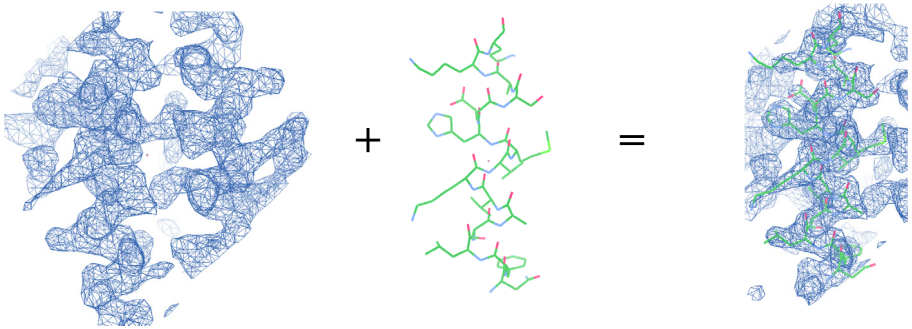
It was originally designed for Macromolecular Crystallography

It is based on some elements of Bayesian statistics: it tries to fit chemically and structurally consistent atomic models into the data.

It also can fit atomic models into cryo-EM maps.

It can do some manipulation of maps, e.g. sharpening/blurring

It is available from CCP4 and CCPEM



Data Atomic model Fit and refine

We want to fit currently available model into the data and calculate differences between them.

To do this fit properly we must use as much as possible information about model and data.

Likelihood function for cryo-EM maps

Probability distribution of "observed" structure factors given coordinates of a molecule:

$$P(F_o;F_c) = Ne^{-\frac{|F_o-F_c|^2}{\Sigma+2\sigma^2}}$$

F_o **"observed" structure factors calculated from EM map**
F_c **calculated structure factors – from coordinates**
Σ **variance of signal**
σ **variance of noise**
N **normalisation**

Major difference from crystallography: 1) variance of noise is large at high resolution; 2) complex structure factors are available

What do we know about macromolecules?

1) **Macromolecules consist of atoms that are bonded to each other in a specific way**
2) **If there are two molecules with sufficiently high sequence identity then it is likely that they will be similar to each other in 3D**
3) **It is highly likely that if there are two copies of a the same molecule they will be similar to each other (at least locally)**
4) **Oscillation of atoms close to each other in 3D cannot be dramatically different**
5) **Proteins tend to form secondary structures**
6) **DNA/RNA tend to form base-pairs, stacked bases tend to be parallel**

External (reference) structure restraints

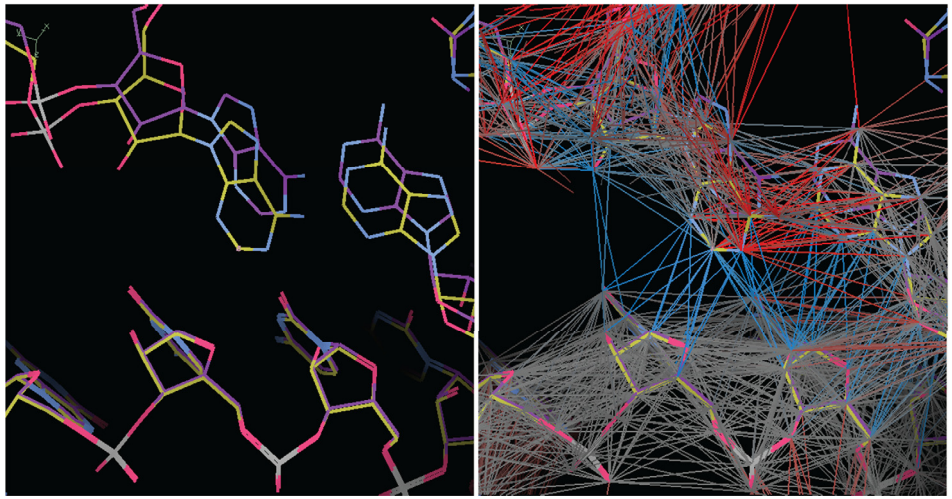
Restraints to external structures are generated by the program ProSmart:
1) Aligns structure in the presence of conformational changes. Sequence is not used
2) Generates restraints for aligned atoms
3) Identifies secondary structures (at the moment helix and strand, but the approach is general and can be extended to any motif).
4) Generates restraints for secondary structures

Note 1: ProSmart has been written by Rob Nicholls and available from him and CCP4.

Note 2: Robust estimator functions are used for restraints. I.e. if differences between target and model is very large then their contributions are down-weighted

Restraints: reference restraints

All restraints can be visualised and applied in Coot:



Yellow=target, purple=reference

Restraints to current distances (jelly-body)

The term is added to the target function:

$$\sum_{pairs} w(|d| - |d_{current}|)^2$$

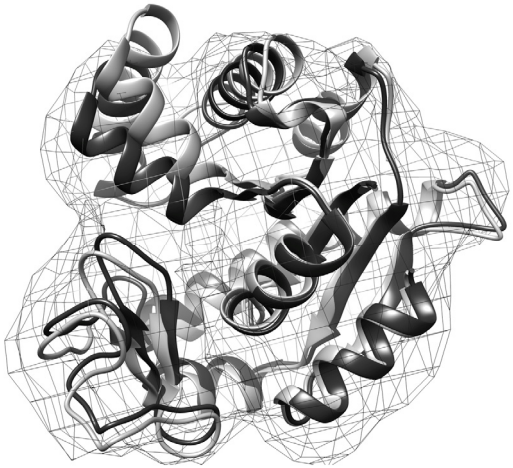
Summation is over all pairs in the same chain and within given distance (default 4.2A). $d_{current}$ is recalculated at every cycle. This function does not contribute to gradients. It only contributes to the second derivative matrix.

It is equivalent to adding springs between atom pairs. During refinement inter-atomic distances are not changed very much. If all pairs would be used and weights would be very large then it would be equivalent to rigid body refinement.

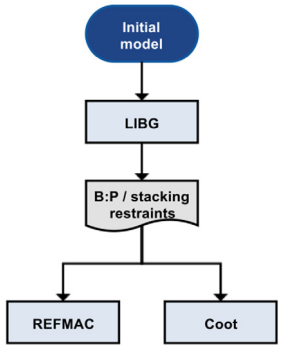
It could be called “implicit normal modes”, “soft” body or “jelly” body refinement.

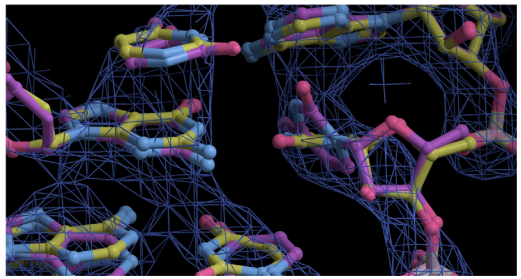
Example: 10A

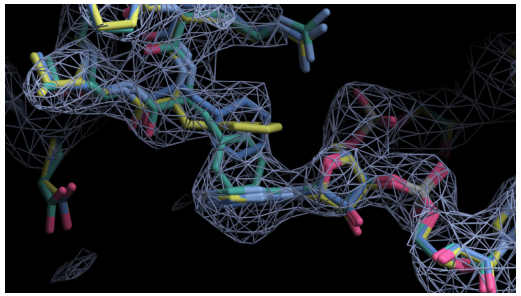
After positioning the molecule with molrep (grey model) and 150 cycle of jelly body refinement (black model)



Basepair and parallel plane restraints

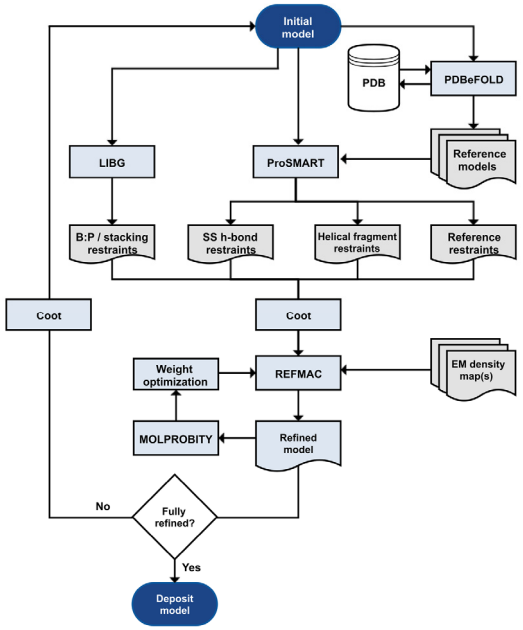






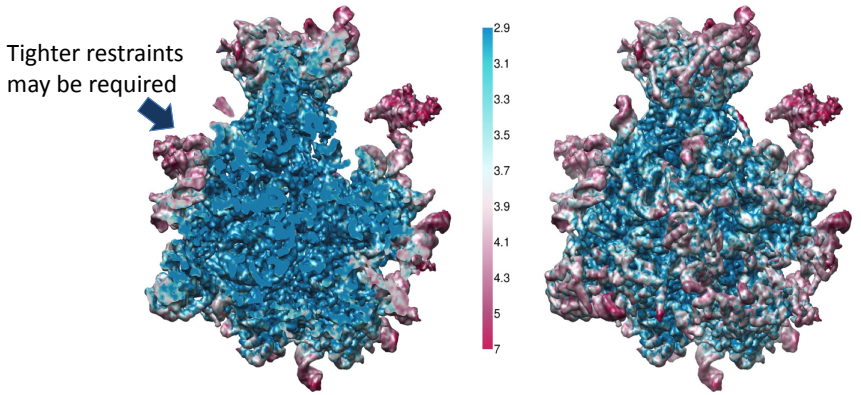
- Base-pair, parallelization and pucker restraints for nucleic acids
- Also suitable for X-ray data

Refinement protocol



- REFMAC can refine models against EM maps
- Input can be multiple or composite maps
- External restraints can be applied for different regions to account for variation in local resolution
- Electron scattering factors are used
- Symmetry restraints can be applied

Restraints can be tuned to local resolution



Electron scattering factor

One way of using electron scattering factor is through Mott-Bethe formula.

For individual atom:

$$f_e(s) = o\frac{me^2}{2h^2}\frac{Z_n - (1+c)f_X(s)}{|s|^2}$$

Individual atoms (with screening and when o=1, c=0):

$$f_e(s) = \frac{me^2}{2h^2}\frac{Z_n - f_X(s)}{|s|^2 + \lambda^2}$$

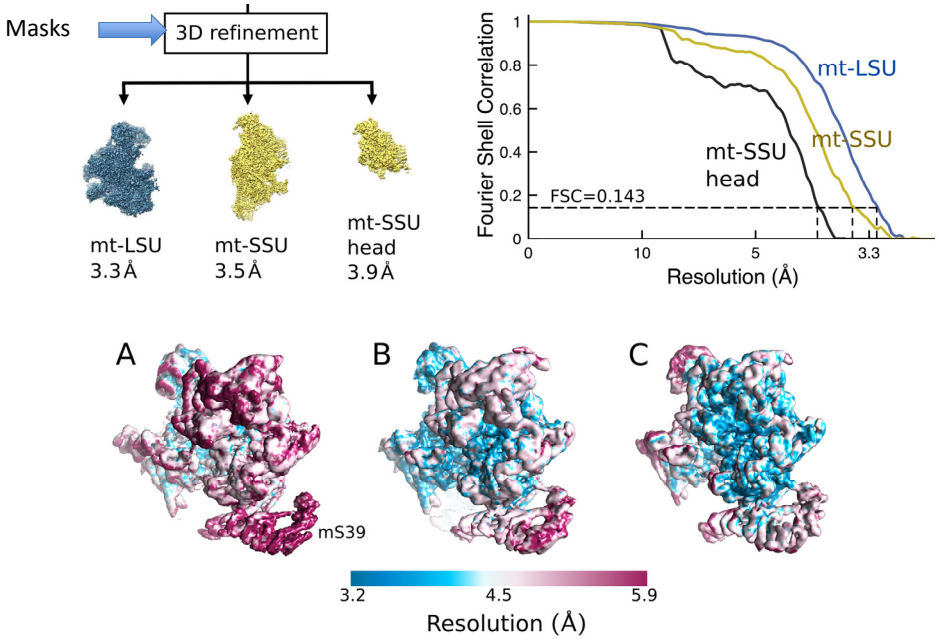
For total Fourier coefficient:

$$F_e = \frac{me^2}{2h^2}\frac{F_Z - F_X}{|s|^2} = \frac{me^2}{2h^2}\frac{F_{Z-X}}{|s|^2}$$

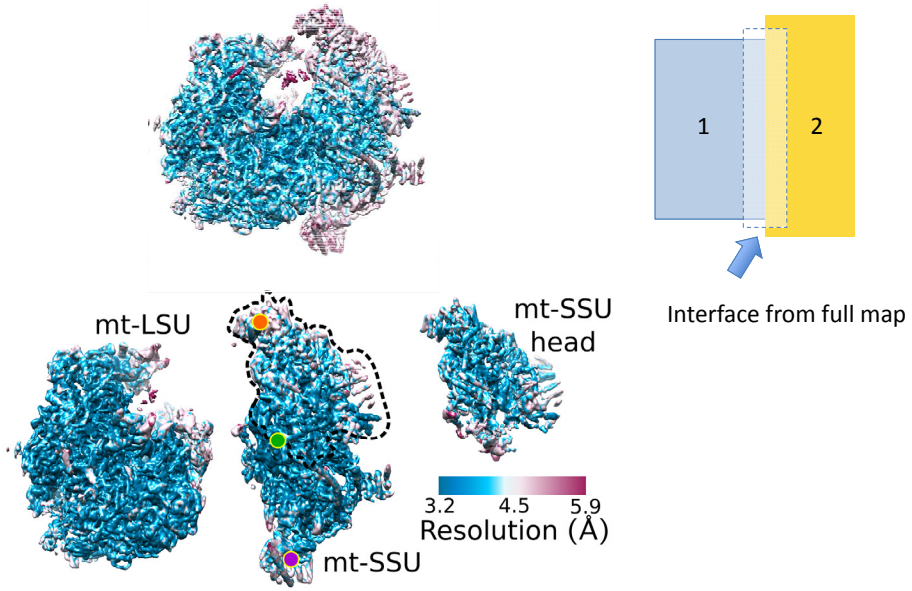
$$F_{Z-X} = \sum (Z_j - f_{j,X}(s))e^{-B_j|s|^2/2}$$

Refinement against composite maps

Masking improves local resolution

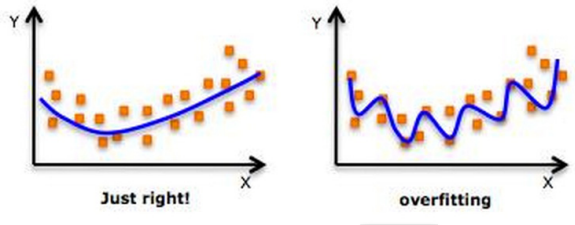


Composite map refinement



Overfitting

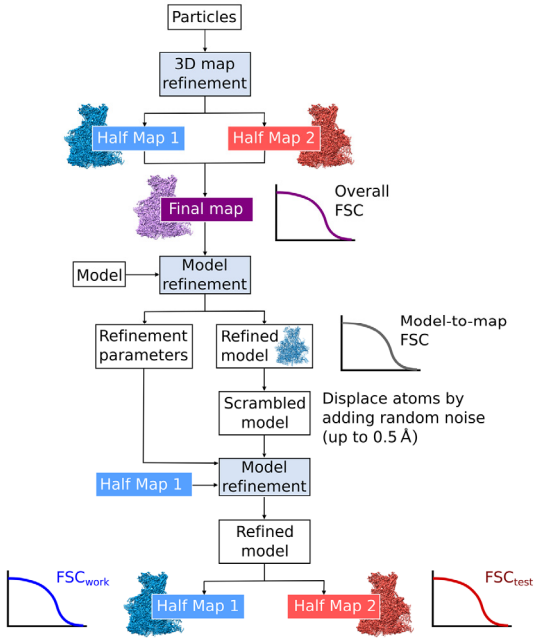
Overfitting



What leads to overfitting?

1. Insufficient data (low resolution, partial occupancy)
2. Ignoring data (cutting by resolution)
3. Sub-optimal parameterisation
4. Bad weights
5. Excess of imagination

Validation of overfitting

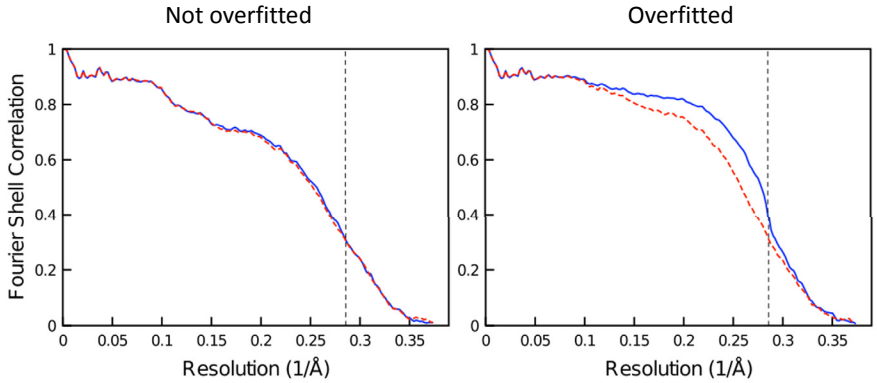


Reference map sharpening

Cross validation requires maps to be on the same sharpening level. Refmac can sharpen given map to a reference map. Usually reference map is taken as full reconstruction map.

Fourier transformation (structure factor) from reference map is calculated and then average values of modules of structure factors in resolution shells are calculated. For given map average values of modules of Fourier coefficients in resolution bins scaled to reference maps coefficients

Validation of overfitting



FSC_{work} = model refined against half map 1, compared to half map 1
FSC_{free} = model refined against half map 1, compared to half map 2

Monitoring fit to density: $FSC_{average}$

- Measure of fit to density (analogous to crystallographic Rfactor)
- Used to follow the progress of refinement
- $FSC_{average}$ avoids a dependence on weight
- FSC is calculated over resolution shells
- If shells are sufficiently narrow the weights are roughly the same within each shell

$$R_f = \frac{\sum_{\mathbf{h}} w_{\mathbf{h}} ||\mathbf{F}_{1\mathbf{h}}| - |\mathbf{F}_{2\mathbf{h}}||}{\sum_{\mathbf{h}} w_{\mathbf{h}} |\mathbf{F}_{1\mathbf{h}}|} \quad e^{-Bs^2/4}$$

$$FSC_{average} = \frac{\sum_{i=1}^{N_{shell}} N_i FSC_i}{\sum_{i=1}^{N_{shell}} N_i}$$

Effect of oversharpening

A report in Science:

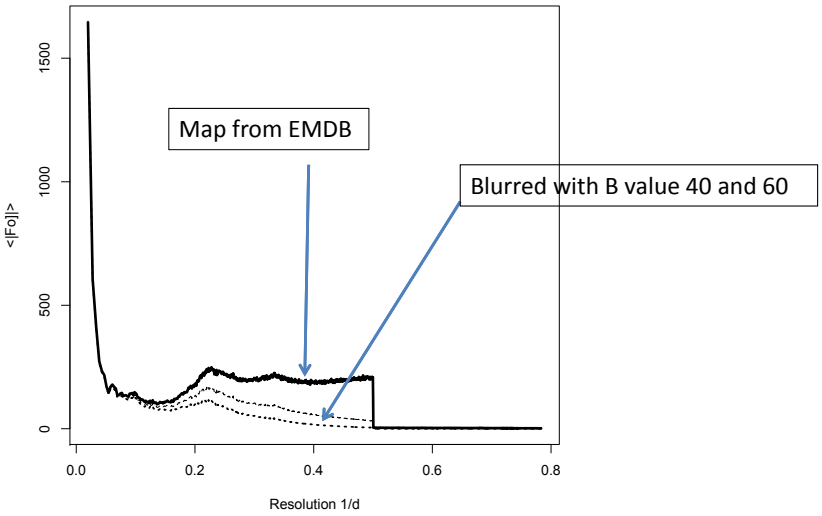
Bartesaghi A, Merk A, Banerjee S, Matthies D, Wu X, Milne J, Subramaniam S "2.2 A resolution cryo-EM structure of beta-galactosidase in complex with a cell-permeant inhibitor" SCIENCE (2015)

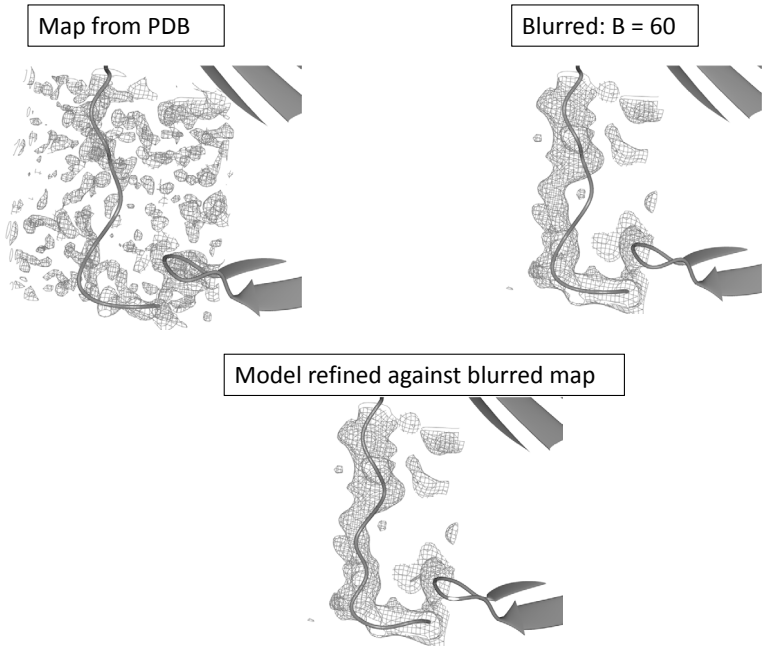
Claim 2.2A maps. Deposited map does not look like to be at 2.2A.

It is the case of over-sharpening. If we have time we could discuss ways of avoiding over-sharpening.

Map quality is better that deposited map.

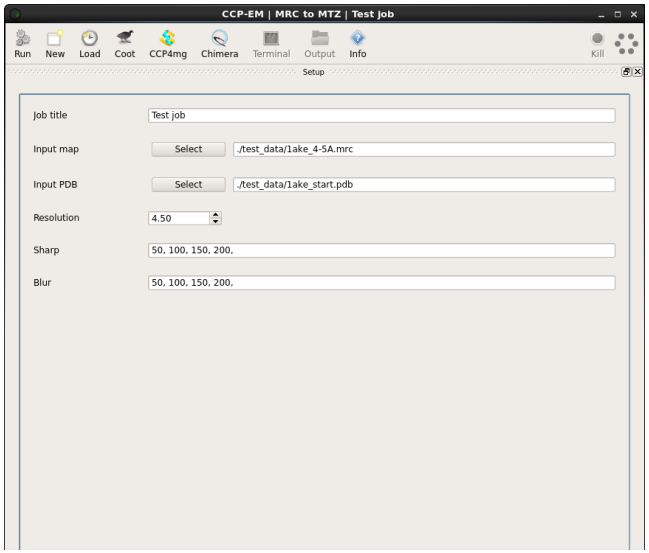
<|F|> vs resolution



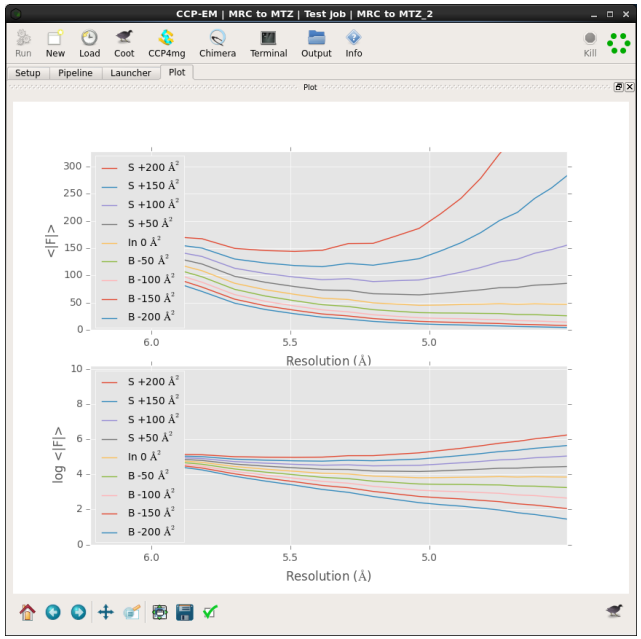


Refinement using ccp-em

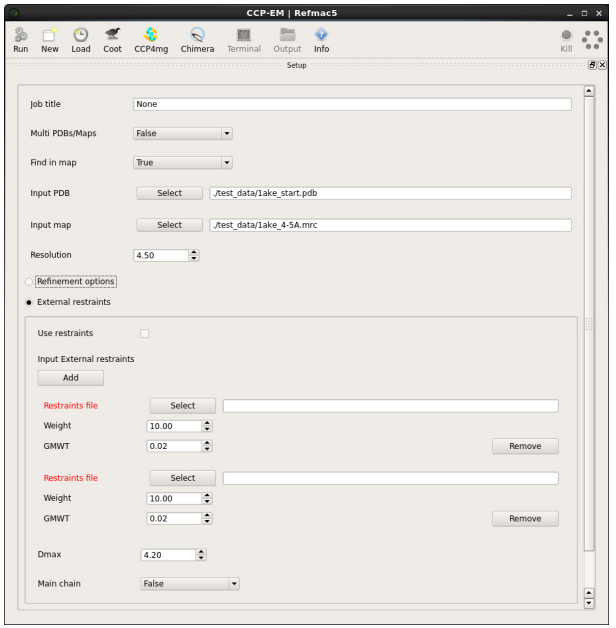
Map sharpening blurring using ccp--em



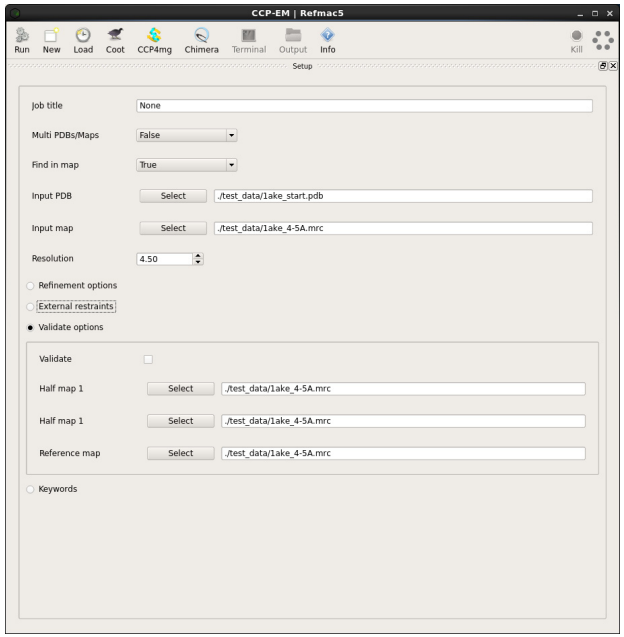
Power spectrum vs resolution



Simple GUI for search in the map



Ccp-em GUI for refinement



Some future plans

1. Calculate map noise variance and significance of differences
2. Calculate Fourier coefficient variances and use them in refinement
3. Local resolution tuned restraints
4. In very distant future: refine against images of particles

Conclusions

Refmac can be used for atomic model refinement against cryo-EM maps

Cross-validation is a problem: refinement against half data maps might be useful

External reference structure restraints help to stabilise refinement against limited and noisy data

Jelly body refinement is useful when starting local conformation is correct

Oversharpening can obscure features of the map. Multiple maps with different sharpening levels should be used for model building.

Some of the features of refmac is available from ccp-em interface

Acknowledgements

MRC-LMB

Fei Long
Rob Nicholls
Oleg Kovalevskiy
Paul Emsley
Michal Tykach
Alan Brown
Many, many EM users

CCP-EM

Tom Burnley
Martyn Winn

And people of LMB, especially those in graphics room
People of CCP4
Users of our programs