Low Resolution Refinement with Coot and Refmac5

Tim Grüne

http://www.shelx.org

December 14^{th}, 2004

Introduction

Coot Graphical Interface by Paul Emsley (http://www.ysbl.york.ac.uk/~emsley) for model building. Includes an interface to refmac5; freely available (Gnu Public License) — but not so easy to install

Recommended reading:

- Coot tutorial: *coot-0.0.pdf* (44 pages)
- Keyboard shortcuts: *coot-keys-and-buttons.pdf* (1 page)
- Short overview: *tutorial-booklet.ps* (8 pages)

Refmac5 Macromolecular refinement program by Garib Murshudov *et al.*, integrated into the CCP4 program suite (http://www.ccp4.ac.uk)

Low Resolution Data

Low resolution data suffer from a low data to parameter ratio.

$\Rightarrow$ Model validation becomes more difficult

$\Rightarrow$ Model validation becomes more important

$\Rightarrow$ External information must be used (standard geometries, Ramachandran plot, etc.)

$\Rightarrow$ Number of R_{free} reflections $\geq$ 500 (not only 5%)

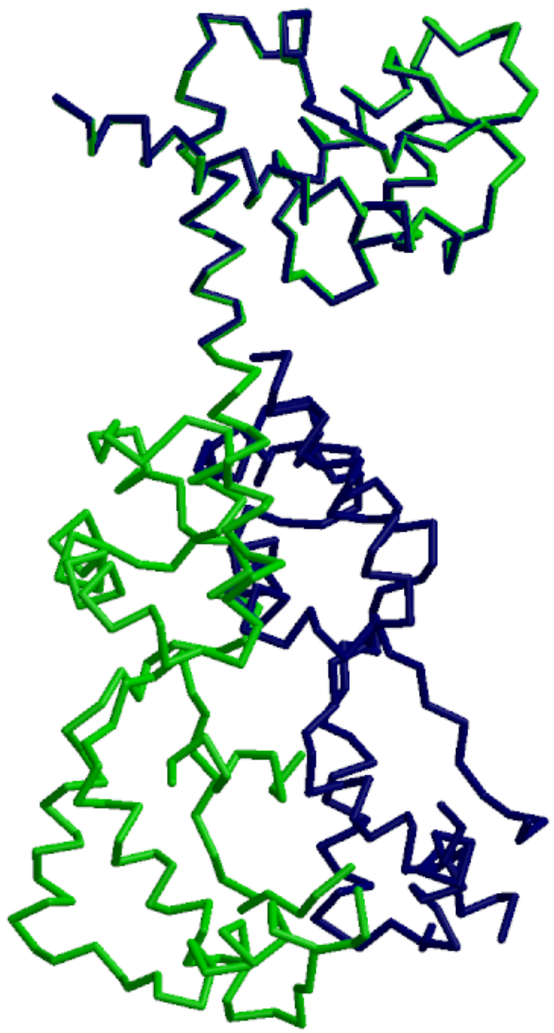
$\Rightarrow$ R_{free} reflections must not “see” the model at any stage of refinement (reflections must be flagged as soon as possible)

The Example Data

C-terminal fragment of “ISWI”, a 120 kDa nucleosome remodelling factor

1. 36 kDa, 304 residues
2. space group $P6_5$, only 70° collected
3. solvent content 60%
4. 27,162 reflections, 6277 unique ($\approx 4.3 : 1$)
5. $I/\sigma_I = 17(4.8)$, $R_{meas} = 5.7\%(32.7\%)$

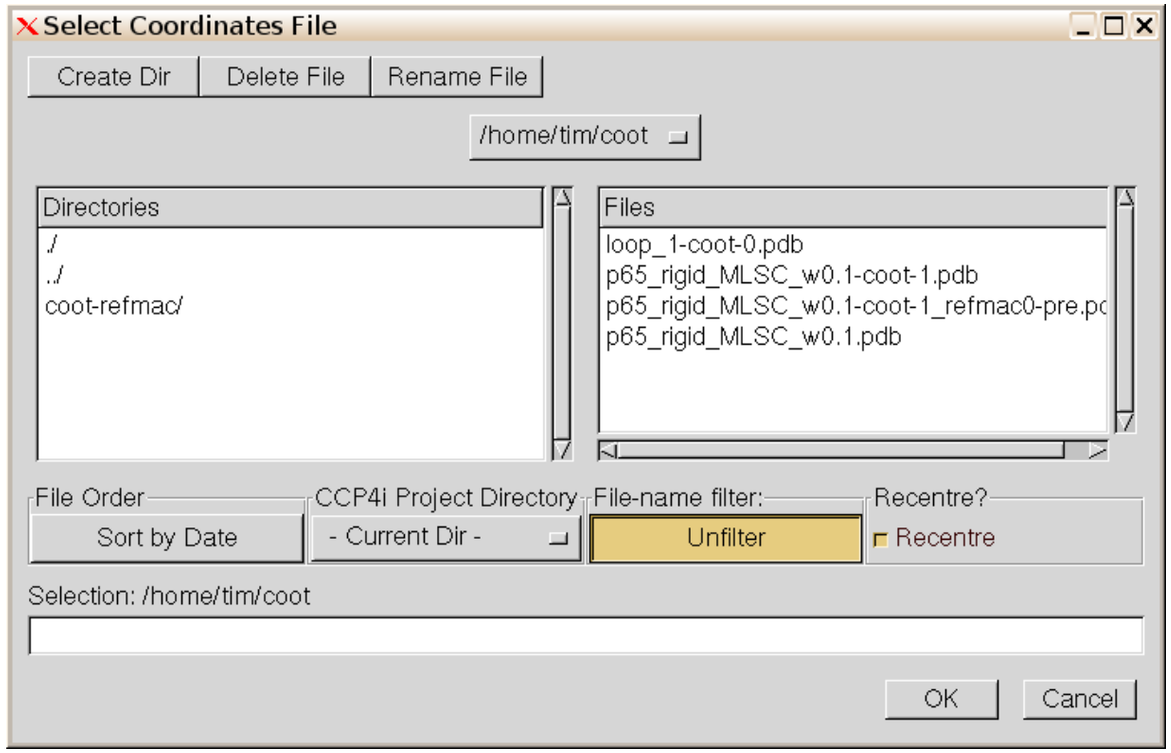
Solved by molecular replacement from the 1.9 Å structure of the same molecule in $C2$. *Phaser* placed three domains one after the other.



Indicator of correctness: Order of domains similar (but not identical) to initial structure.

Coot — Presentation of PDB files and Maps

- Supported file formats: pdb, mmCIF for coordinates; mtz, mmCIF, phs for reflections/phases; CCP4 map files

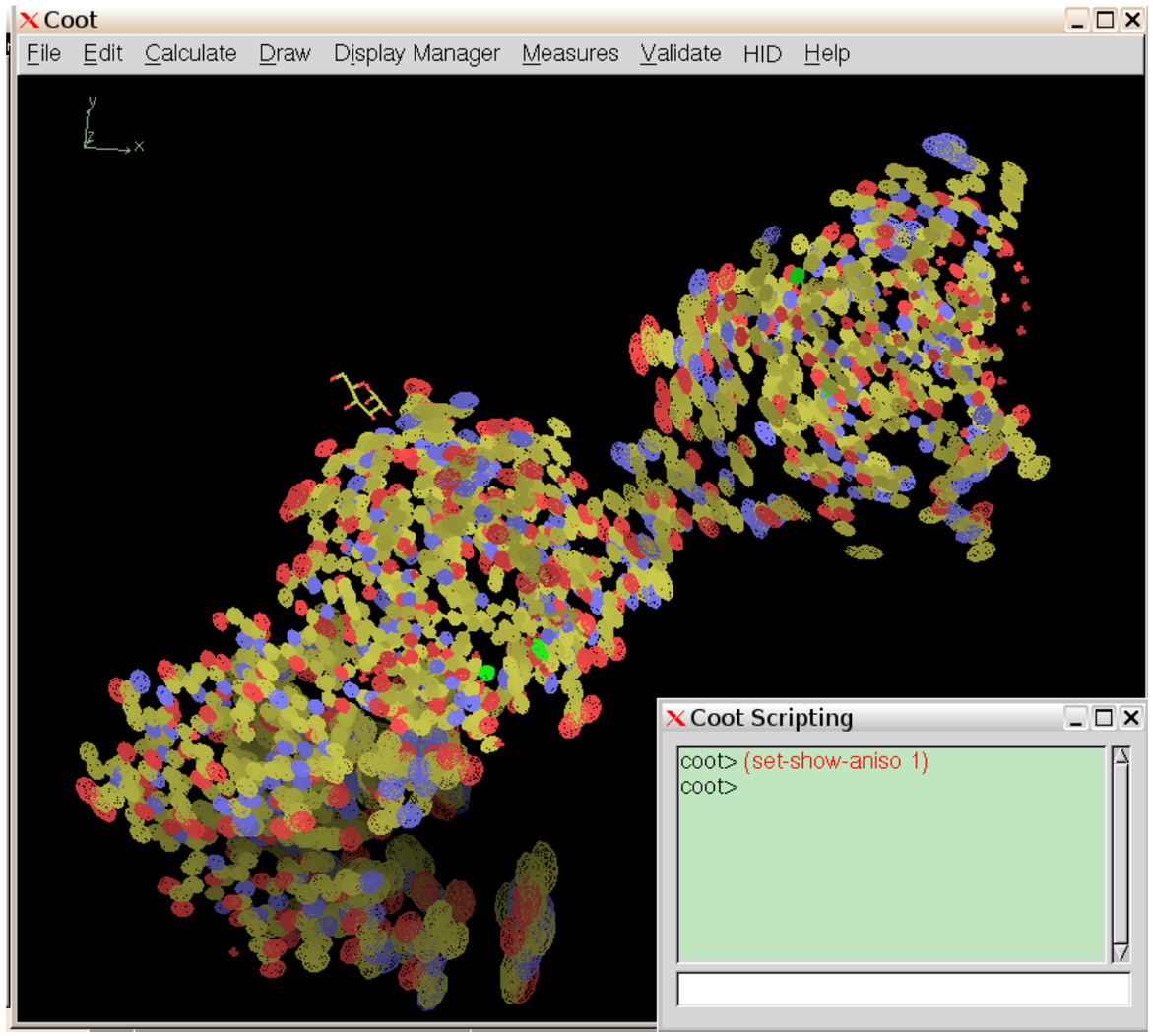


Coot — Presentation of PDB files and Maps

- Supported file formats: pdb, mmCIF for coordinates; mtz, mmCIF, phs for reflections/phases; CCP4 map files
- Calculates map by FFT from reflection files

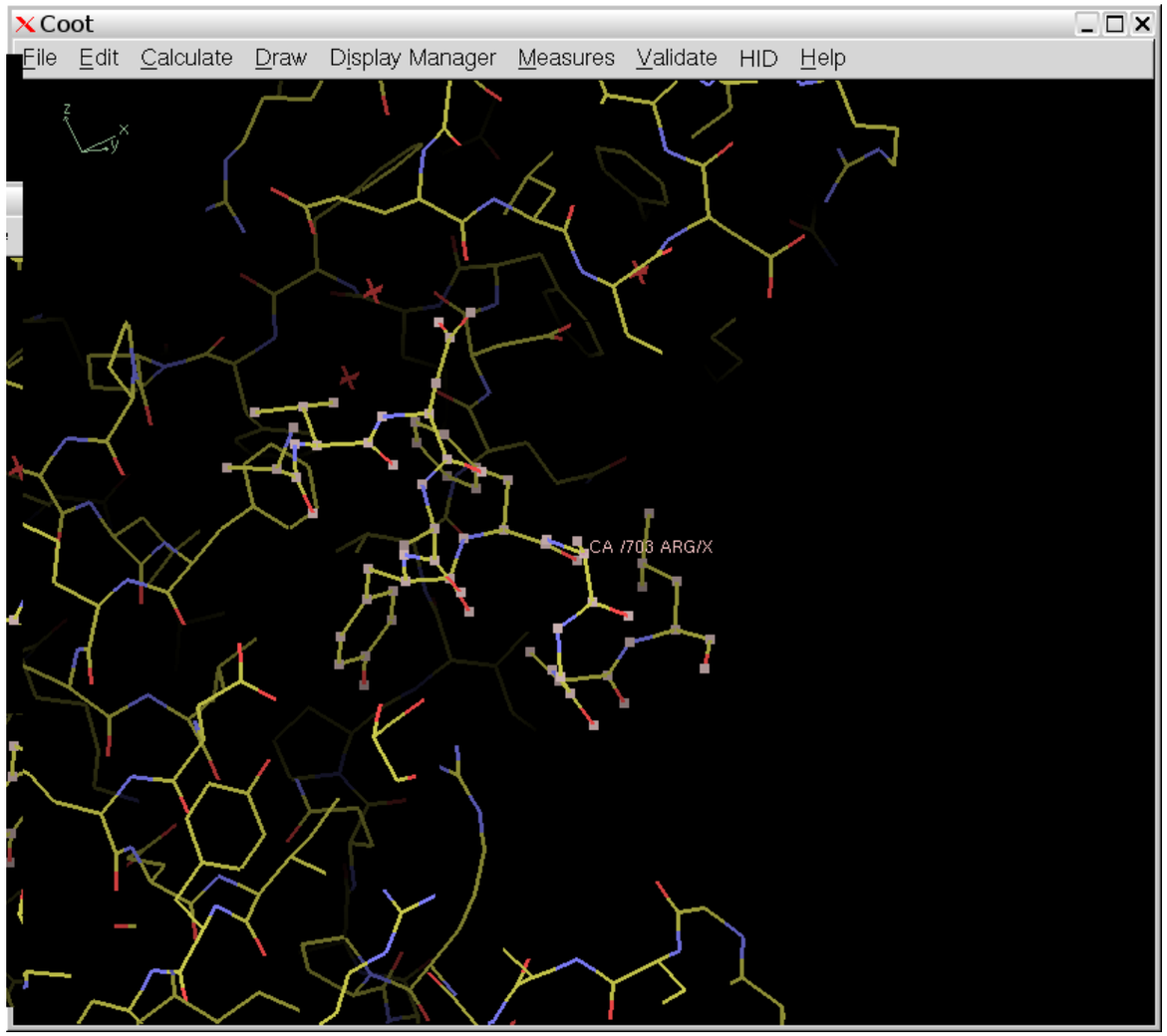
Coot — Presentation of PDB files and Maps

- Supported file formats: pdb, mmCIF for coordinates; mtz, mmCIF, phs for reflections/phases; CCP4 map files
- Calculates map by FFT from reflection files
- Reading of files straight forward through FILES menu
- Presentation of coordinates file with isotropic or anisotropic temperature factors



Coot — Presentation of PDB files and Maps

- Supported file formats: pdb, mmCIF for coordinates; mtz, mmCIF, phs for reflections/phases; CCP4 map files
- Calculates map by FFT from reflection files
- Reading of files straight forward through FILES menu
- Presentation of coordinates file with isotropic or anisotropic temperature factors
- Marking of atoms with zero occupancy by grey box (very useful for files downloaded from the PDB)

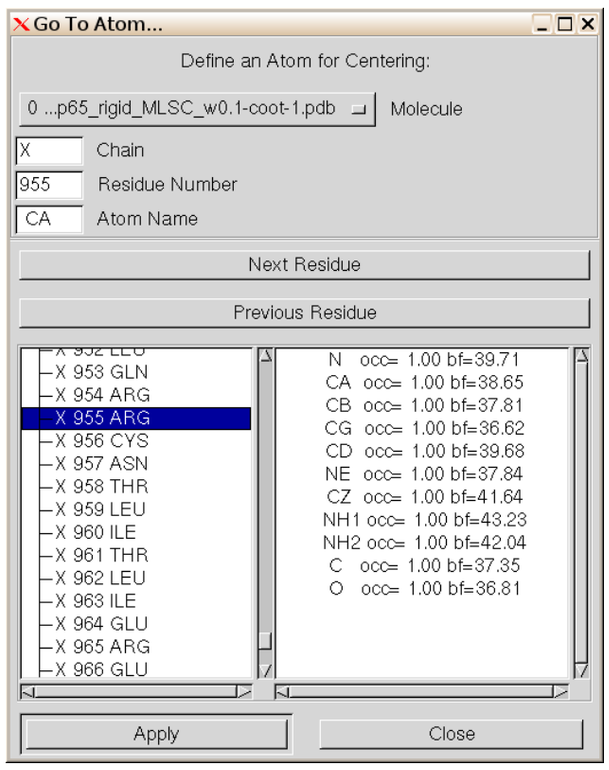


Coot — Navigation

Mouse 1. middle mouse button: centring on atom

2. Ctrl-key and left mouse button: dragging (translation) of map

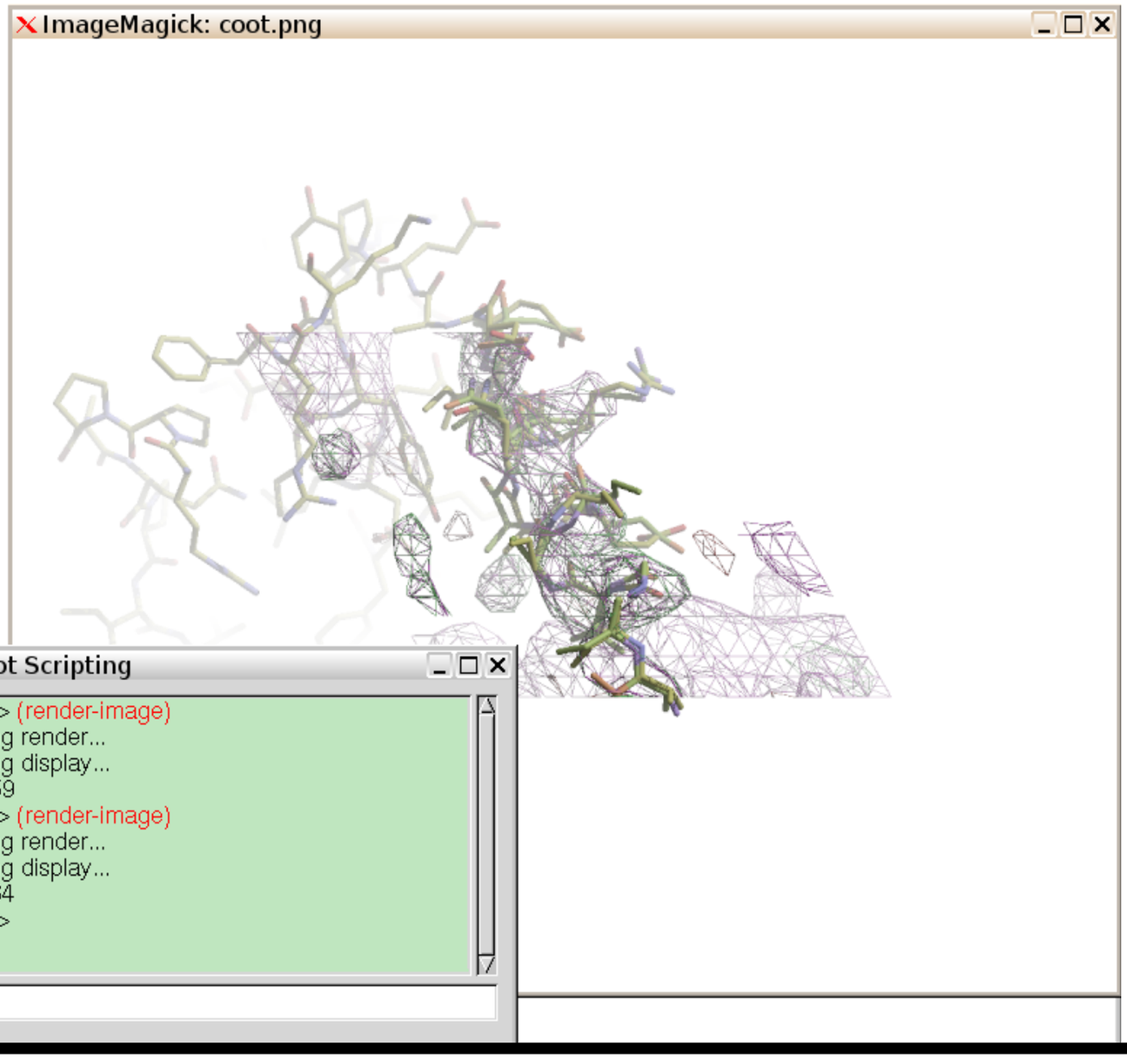
Go-to-atom-menu (from “Draw–menu”) direct naming of target atom/residue; menu also displays information about temperature values.



The space bar centres on the next C_α of the currently selected model.

Coot — Scripting

Coot understand the scripting language scheme (python should also be supported). Scheme is a variant of lisp, but following the examples from tutorial and manual, it allows variable configuration of Coot. The user can read in whole scripts or use the scripting menu to execute single commands, e.g. in order to render the current view:



Detecting Features in low Resolution Maps

Automated building is often not an option with low resolution data.

With no model available, good starting points for model building are the secondary structure elements of proteins, α–helices and β–sheets.

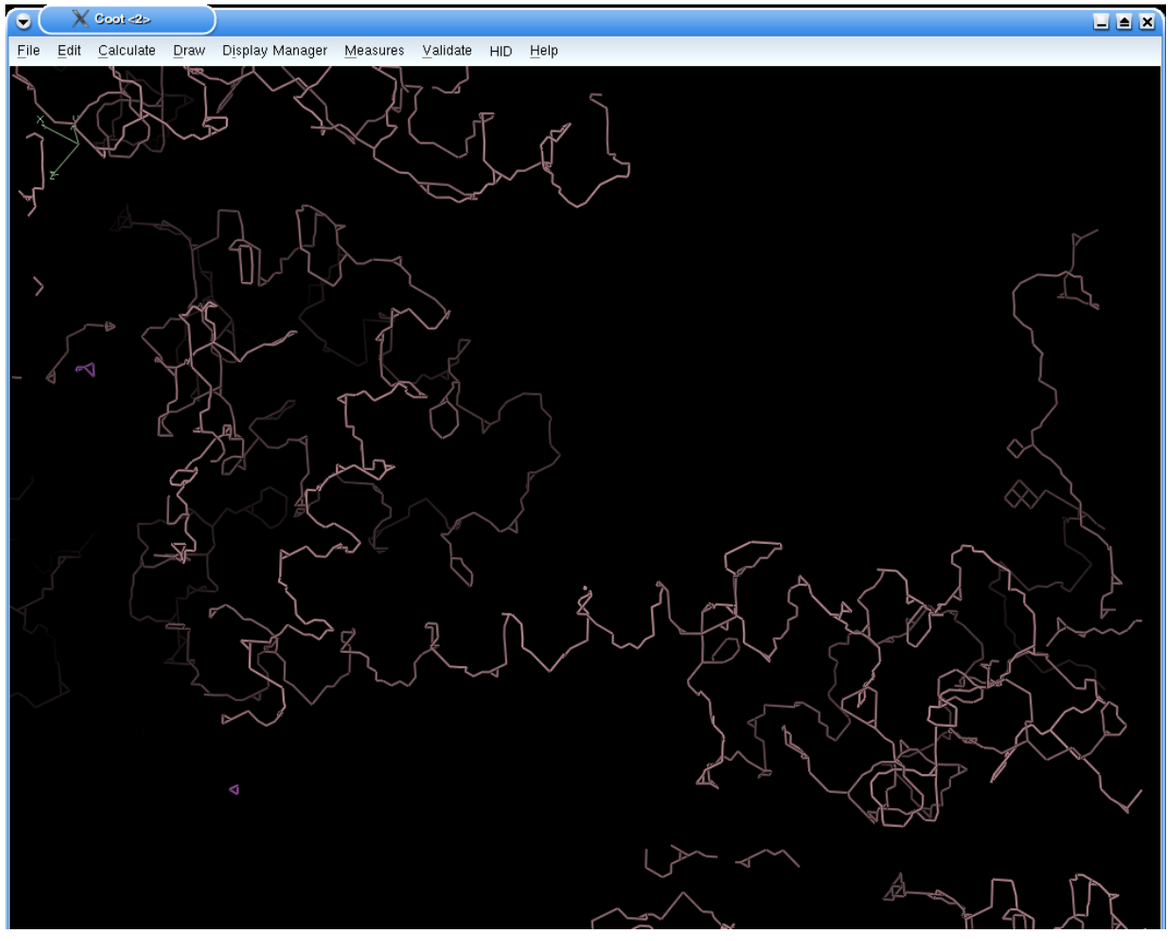
With Coot, secondary structure elements can be placed by combination of skeletonisation and baton building.

Skeletonisation finds a connected path through the electron density that crosses maxima of electron density (algorithm by J. Greer (1974))

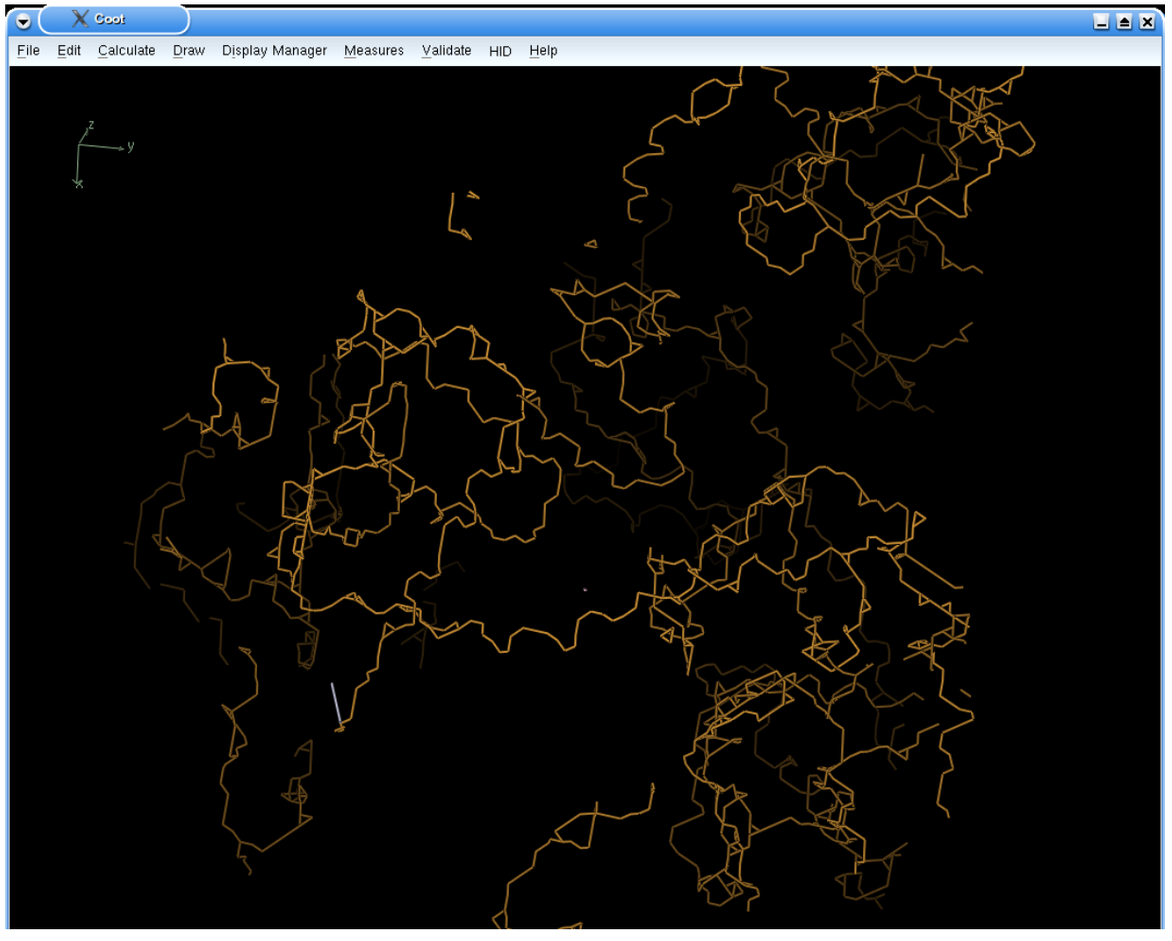
Baton represent two subsequent C_α atoms of a peptide chain, i.e., they are 3.8Å apart.

Map Skeletonisation

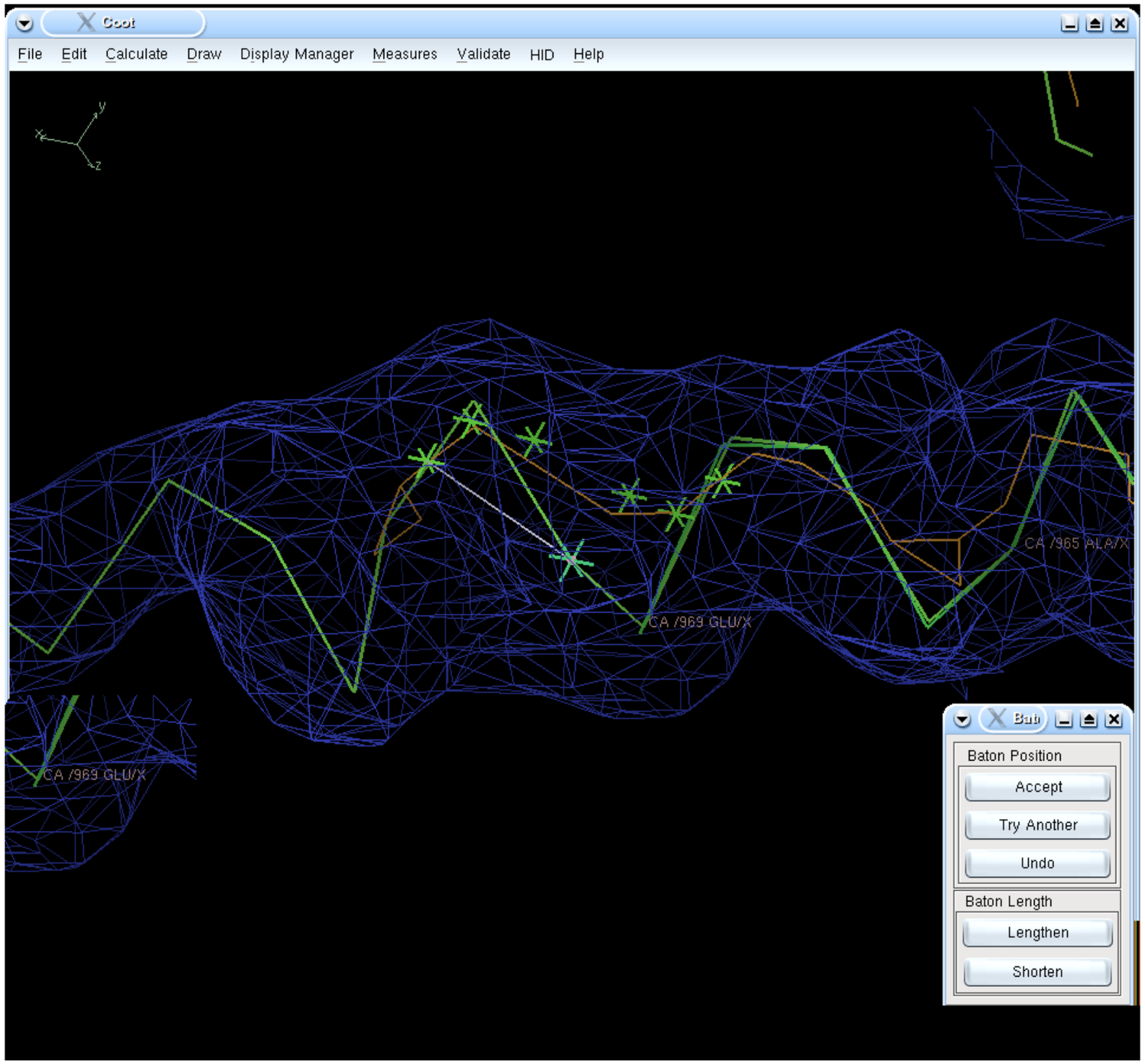
Example skeleton at 2.0Å resolution: Skeleton passes through C_α atoms, Side chains become visible



Example skeleton at 3.4Å resolution: Skeleton follows the main chain



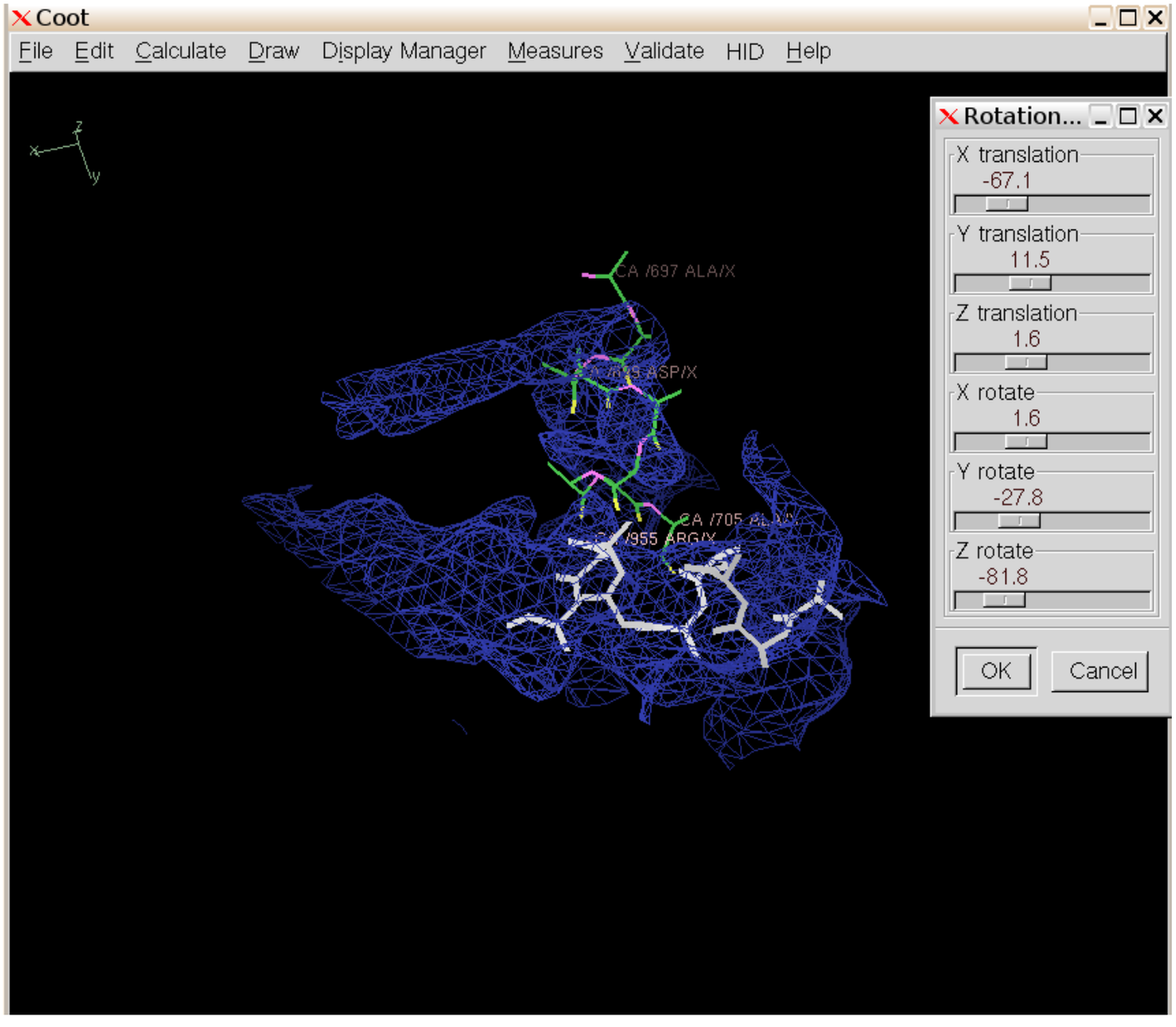
Baton Building



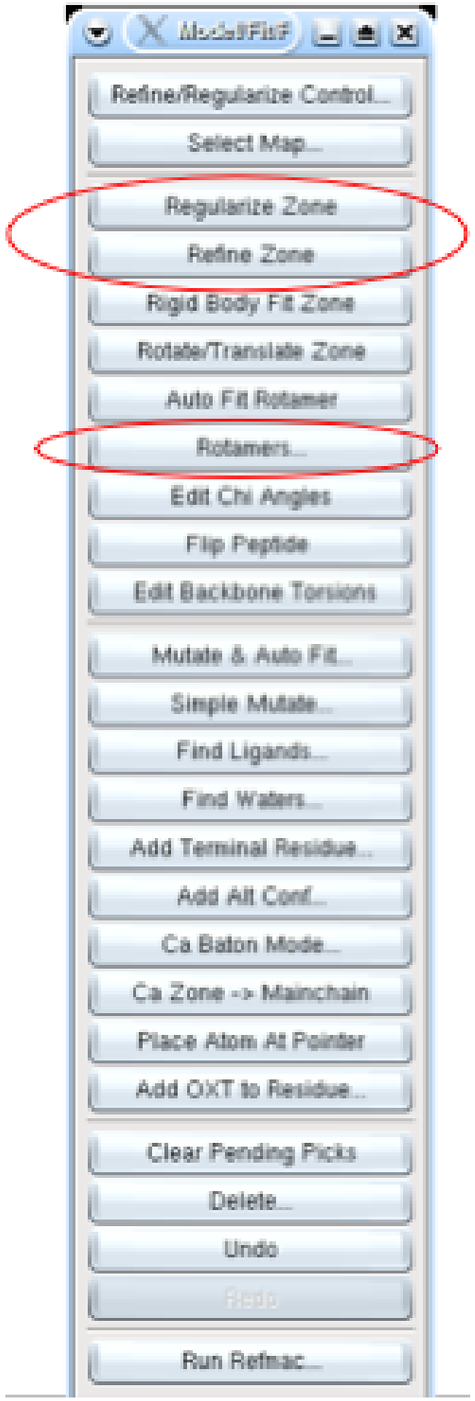
The smarter way

At low resolution, it is probably wiser to manually place a template helix or strand into the density.

1. Extract a helix from an pdb–file with appropriate length (10–20 residues)
2. Truncate side–chains to alanines
3. place and orient template into target density
4. mutate side chains to correct residue types (where possible, e.g. Trypto-phane)



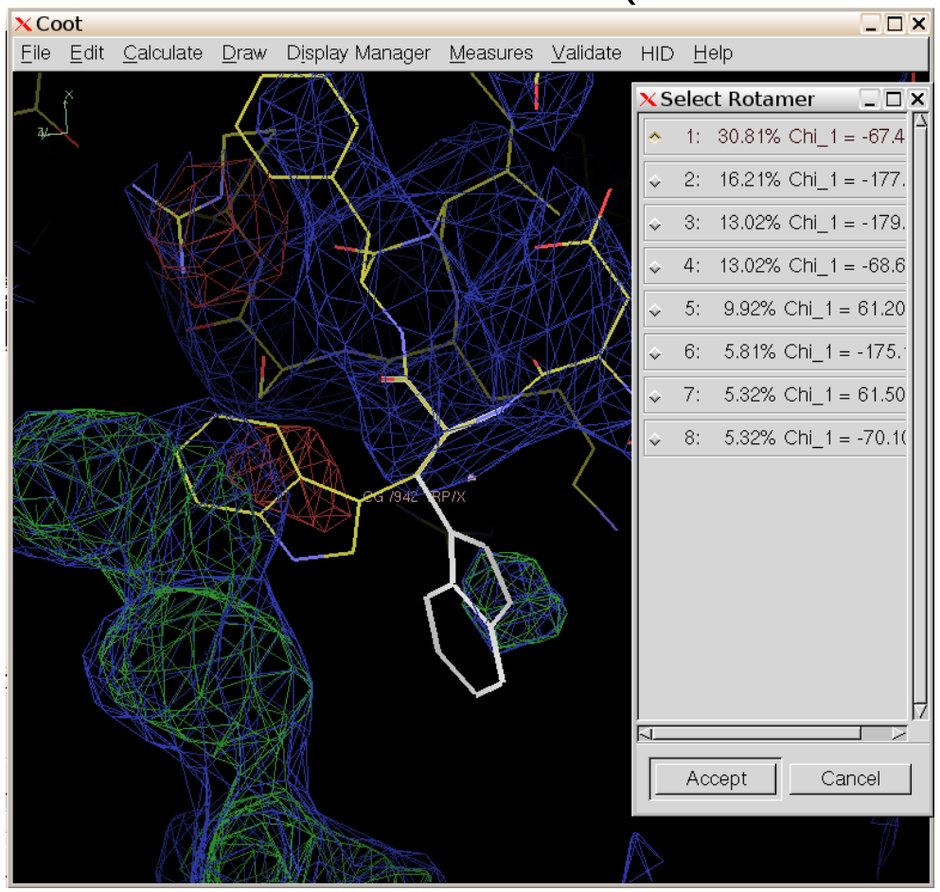
Local Refinement and Regularisation



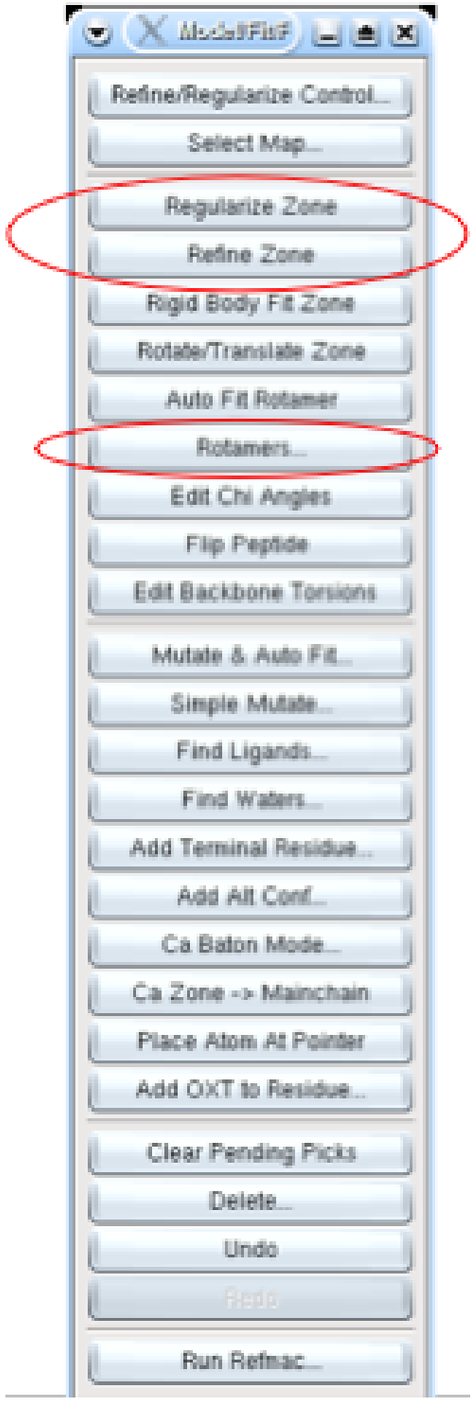
The “Model/Fit/Refinement” menu (accessible via “Calculate” show Coot’s options for model building and refinement).

Particularly important for low resolution structures:

Rotamers . . . library of side–chain conformations (R. Dunbrack et al. (1997))



Local Refinement and Regularisation

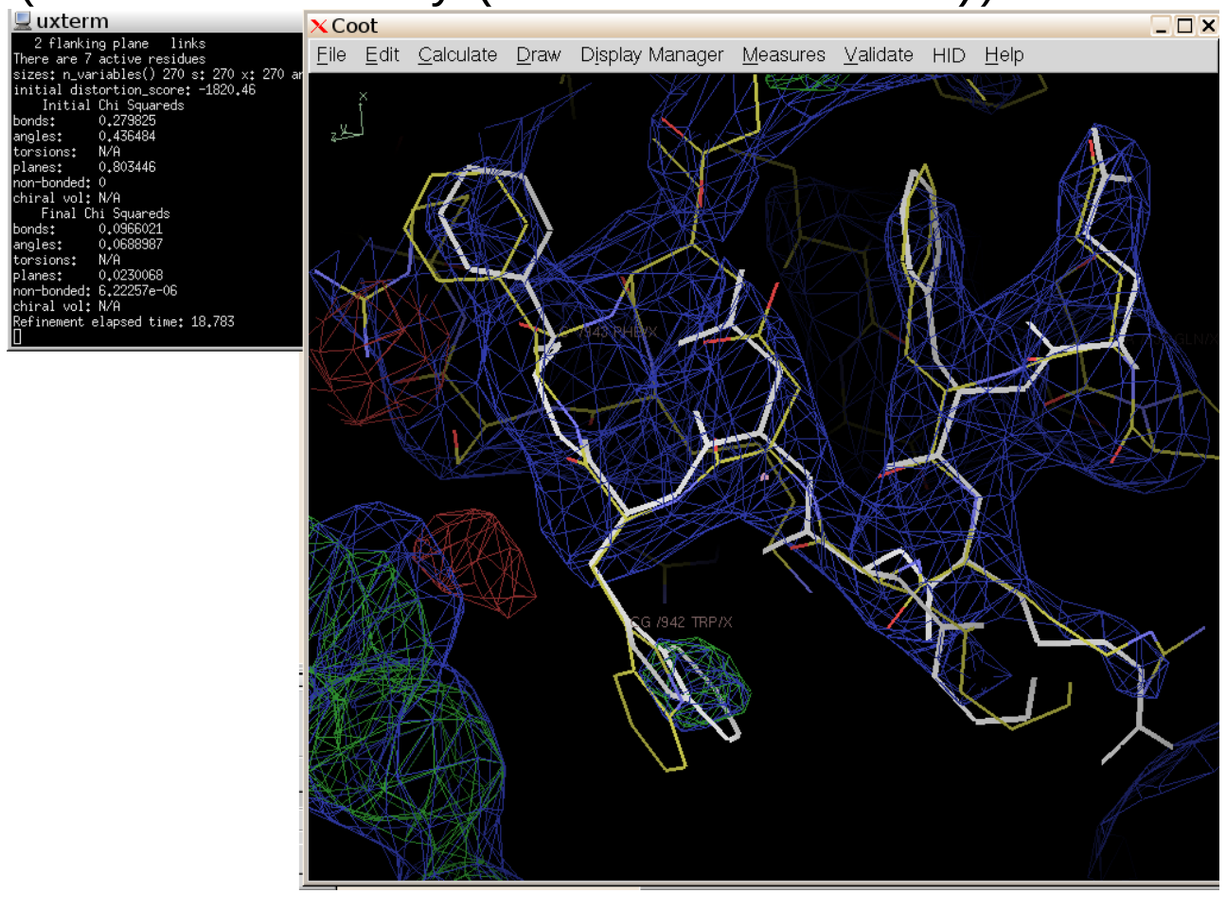


The “Model/Fit/Refinement” menu (accessible via “Calculate” show Coot’s options for model building and refinement).

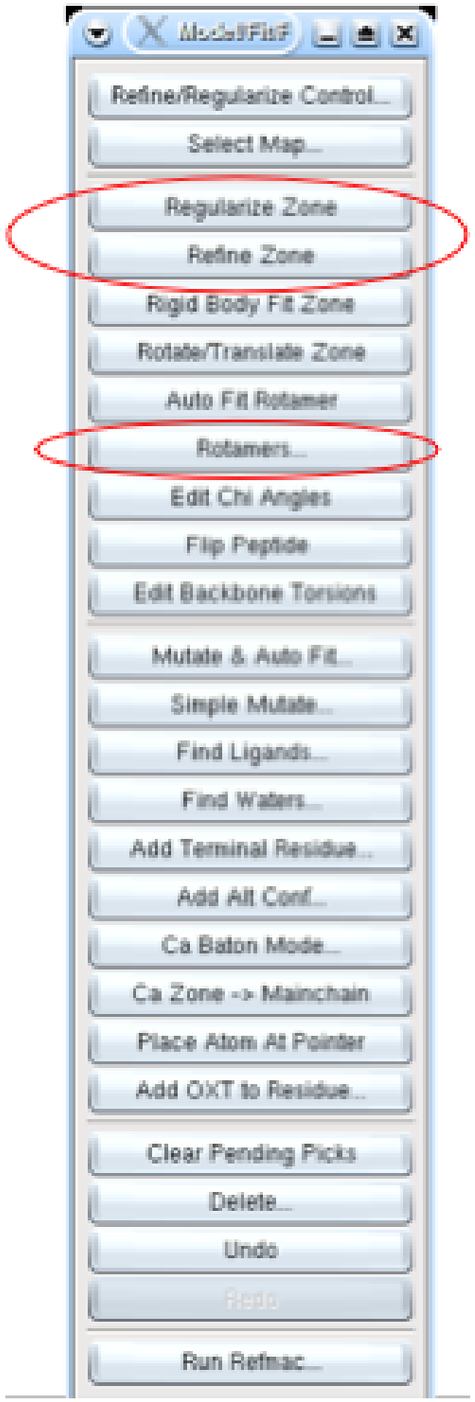
Particularly important for low resolution structures:

Rotamers . . . library of side–chain conformations (R. Dunbrack et al. (1997))

Refine Zone real space refinement of selected zone to map; takes geometry terms into account (ration controlled by `(set-matrix 20.0)`)



Local Refinement and Regularisation



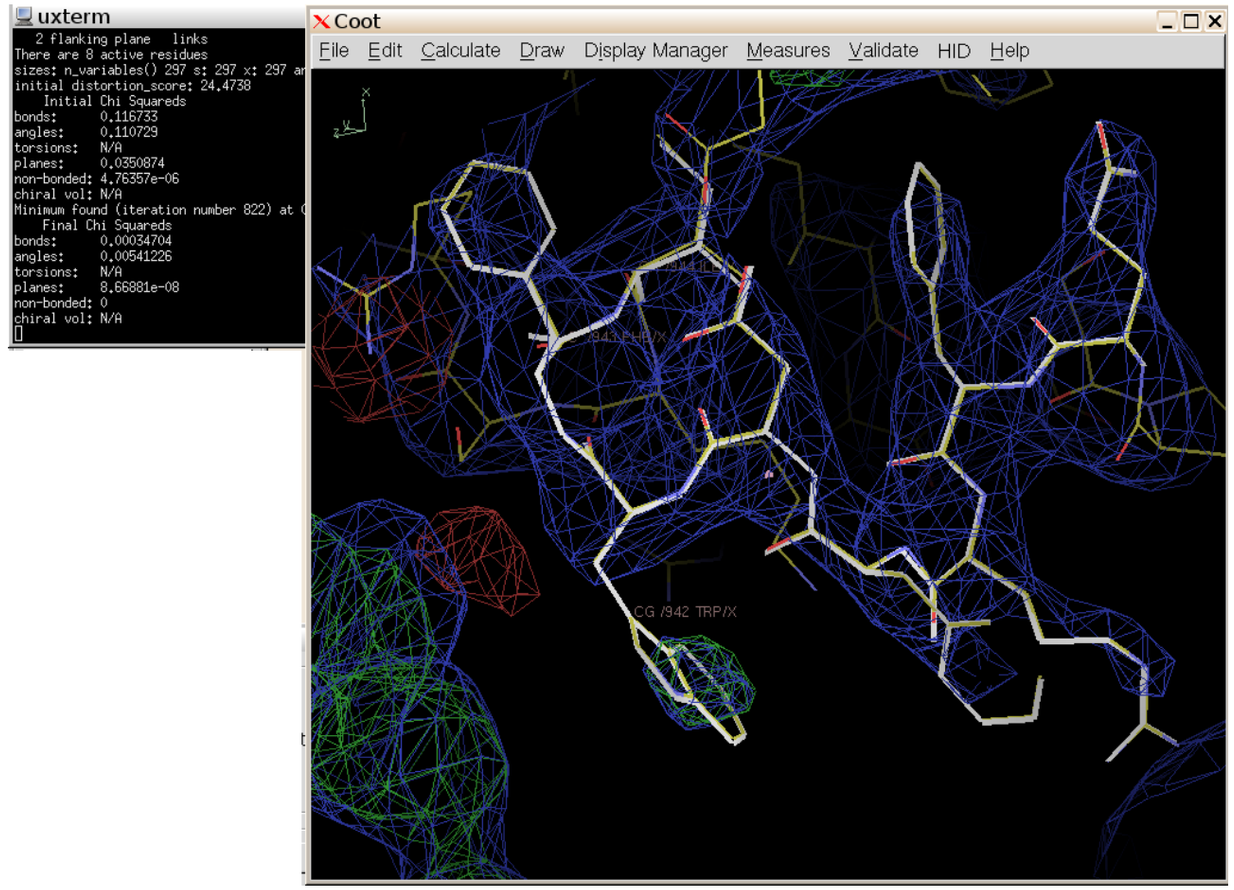
The “Model/Fit/Refinement” menu (accessible via “Calculate” show Coot’s options for model building and refinement).

Particularly important for low resolution structures:

Rotamers . . . library of side–chain conformations (R. Dunbrack et al. (1997))

Refine Zone real space refinement of selected zone to map; takes geometry terms into account (ration controlled by `(set-matrix 20.0)`)

Regularize Zone adjust the selected zone ($\leq$ 20 residues) to conform to geometric ideal values



Running Refmac5

Refmac5 can be run in three different ways:

1. directly from Coot via the “Model/Fit/Refinement” menu. This uses the default settings plus the specified pdb–file and mtz–file; default settings can be overwritten by creating a file called `refmac-extra-para` which contains extra refmac5–commands

2. via a script from the command line

3. via the CCP4–interface

The interface is comfortable and sufficient for normal cases. A script allows more control than the interface and is closer to the documentation of Refmac5 — helpful for critical (low resolution) cases!

Starting point: Script from GUI (“Run & View Com File”)

Finding the correct Parameter Settings for Refmac5

The first critical value for Refmac5 is the weighting term which relates reflection data and geometry restraints (bond lengths, angle, etc.).

The default value of 0.3 is too high for low resolution data and leads to heavy distortions of the protein:

3.4Å; weight = 0.3

Ncyc	Rfact	Rfree	FOM	rmsBOND	rmsANGLE
0	0.437	0.440	0.667	0.066	1.994
1	0.309	0.372	0.680	0.125	7.253
2	0.288	0.365	0.698	0.108	6.733
3	0.286	0.361	0.691	0.109	6.918
4	0.283	0.360	0.699	0.088	6.221
5	0.281	0.357	0.697	0.082	6.012
6	0.283	0.361	0.699	0.082	6.044
7	0.284	0.359	0.696	0.084	6.107
8	0.285	0.364	0.696	0.082	6.039
9	0.285	0.365	0.692	0.082	6.012
10	0.288	0.369	0.692	0.083	6.119

3.4Å; weight = 0.03

Ncyc	Rfact	Rfree	FOM	rmsBOND	rmsANGLE
0	0.437	0.440	0.667	0.066	1.994
1	0.346	0.378	0.687	0.050	1.838
2	0.341	0.381	0.692	0.026	1.387
3	0.340	0.382	0.691	0.011	1.111
4	0.339	0.385	0.689	0.010	1.077
5	0.340	0.385	0.690	0.009	1.077
6	0.341	0.387	0.689	0.009	1.091
7	0.342	0.389	0.688	0.009	1.071
8	0.344	0.392	0.685	0.008	1.071
9	0.346	0.395	0.684	0.008	1.071
10	0.348	0.398	0.684	0.009	1.081

An even better check for the stability of refinement is provided by procheck, also part of the CCP4–suite.

Which Parameters to Refine

The example fragment of ISWI has about 2,200 atoms to refine.

Refinement of coordinates and B–factor leads to more than 8,000 parameters, compared to 6,000 reflections;

Refmac5 allows to refine an overall B–factor, i.e., only 6,600 parameters and therefore a more stable refinement (REFI BREF OVER instead of REFI BREF ISO).

If B–values are refined to very high values (> 120Å^2), overall B–factor refinement should certainly be considered (test runs with different values).

How to tell good from bad Parameter Settings

There are three guides to help decide whether a change in parameter setting for Refmac5 was sensible or not:

1. R and R_{free} values. For low resolution data it is important to make sure that the R_{free} does not divert too much from the R–value.

2. Ideal geometries: Procheck (or Whatcheck) should often be run to check the sanity of the model. Consider Ramachandran plot, ideal bond lengths and angles.

3. The map: When testing a range of parameters for Refmac5, e.g. overall B–factor, look at the electron density map whether features appear or disappear.

Often R and R_{free} reach a minimum after a couple of cycles of refinement and then rise again. Fix NCYCLE to the number of cycles with the best values!

Keywords BINS and DAMP

DAMP regulates the shifts of scales per cycle (defaults: 0.5/0.5). For low resolution data and a bad model (beginning of refinement), these values should be lowered to stabilise refinement.

BINS For maximum likelihood methods, E–values must be calculated from structure factor amplitudes by separating the resolution range into bins (default: 20). For low resolution data, this might result in too few reflections per bin, i.e. meaningless E–values; in such a case, BINS should be set to a lower value (Bins 10)

The SCALE Keyword

A very critical parameter for Refmac5 is the SCALE keyword. From the manual:

> "We are not really sure how best to handle scaling. If you have problems please get in touch."

SCALE modifies the scaling of observed and calculated structure factor amplitudes and influences the dealing of bulk solvent (scattering from material between protein molecules in the crystal).

Sub–keywords of SCALE that are particularly important for low resolution data:

TYPE The CCP4–GUI defaults to SIMPLE, but it is generally better to use BULK, i.e. bulk solvent scaling based on Babinet's principle

BAVERAGE The overall B–factor is estimated from the Wilson plot which is only reliable for data below 4Å.

Judging from the authors' quote, there is not rule of thumb for setting these parameters. The are best being tested in a range of runs.

Acknowledgement

I am grateful to Airlie McCoy for solving the molecular replacement problem with a pre–release of *phaser*. I had tried MolRep, AmoRe, Qs, and Beast without success.

Peter Brick and Stephen Curry educated me to being careful when building models and interpreting data.

Dante Neculai encouraged me installing Coot on our network (which still took me 2 days . . .)