Refinement against cryo-EM data with CCP-EM using REFMAC & Coot

Tutorial written by Rob Nicholls, Tom Burnley, Colin Palmer & Garib Murshudov

Tutors: Tom Burnley, Colin Palmer, Arjen Jakobi, Agnel Joseph

EMBO Workshop, September 2021

In this tutorial we consider the refinement of a beta-galactosidase model into a 2.9 Å map from a cryo-EM reconstruction (tutorial.mrc), using a homologous model from the PDB as a starting point (homolog.pdb).

For the purposes of this tutorial, in order to make it computationally tractable, the map and box size has been reduced based on symmetry. So, whilst the biological assembly comprises four chains, only one chain is present in the provided map. The homologous model used as the starting point corresponds to chain A from the deposited model with PDB code: 5a1a.

Part 1) Map and Model Inspection

`(a) Open Coot, and load the model (homolog.pdb) and map (tutorial.mrc)`

(You can do this on the command line by typing "`coot --map tutorial.mrc --pdb homolog.pdb`")

- If you don't see the map, but you do see it listed in the Display Manager, then it could be that the contour level is too high – try decreasing the map contour level (scroll down with the mouse, or press "-" a few times).
- It is often useful to display the box that encapsulates the map:
  Draw -> Cell & Symmetry -> Show Unit Cells? -> Yes
  Then zoom in/out until you see the whole box (right click and drag, or use the "m" and "n" keys).
- Increase the map radius in order to display the whole map (e.g. to 70 Å in this case):
  Edit -> Map Parameters -> Map Radius EM.
- When zoomed out, it is often useful to display the model as a CA trace (via the Display Manager, pull down menu to the right of the Active check box).

`(b) Improve the fit of the model to the map`

The model has been translated and rotated so that it nearly fits the map, but perhaps it is not a perfect fit.

- Try using the rigid body fitting in order to optimise the overall fit of the model to the map:
  Calculate -> Modelling -> Rigid Body Fit Molecule…
- Does the model now fit the map reasonably well? If yes click "Accept" in the "Accept Refine" panel. If not then "Reject" and try again!

Note: In practice, if the position or rotation of the model is too far from the optimal solution in the map, some fitting tools may not work. In such cases, you may need to try various tools before achieving a reasonable model from which to start full-model refinement. There are a variety of tools in Coot that can help with this (e.g. Rotate Translate Molecule, Jiggle Fit "J", Morph Fit, real space refinement with self-restraints, etc.), but in this case the Rigid Body Fit should be sufficient. Other options include using Molrep in the CCP-EM interface, which attempts to obtain the optimal superposition between model and map automatically.

(c) Inspect differences between the model and map (don't spend too long on this!)

- Are there any regions of the density for which there is no model? (Validate -> Unmodelled blobs)
- Are there any regions of the model for which there is no supporting density? (Validate -> Density fit analysis). Very quickly and roughly remove three large regions (consecutive stretches of residues) for which there is no supporting evidence in the map (Right toolbar: Delete Item (Bin Icon) -> Delete Zone -> click two residues). *Remember that this model came from a different (homologous) structure, so it is no surprise that there are differences between the starting model and the map.*
- The model includes a large number of water molecules. Should these be included in the model? (Validate -> Check/Delete Waters). You won't see waters if you're still viewing the model as a CA trace, so switch back to a "Bonds" representation in the Display Manager. Click through a few of the waters. Are they supported by map density?
- How many waters are deleted by the "Check/Delete Waters" tool when you select "Delete" mode instead of "Check" mode?
- Are there any ligands in the model (click the Go To (Next) Ligand icon, which is next to the "Display Manager" and "Go To Atom" buttons at the top of the window)? You may need to select a display representation in the "Display Manager" that allows you to see the ligand (e.g. "CAs+Ligands" or "Bonds"). Does the map support the modelled ligand?

Note: if you were to fit the model into the density using Molrep, be aware that Molrep removes all HETATM records (i.e. ligands, metal ions, waters, etc.) from the model, so you would need to copy the ligand into the model before proceeding. You would do this by superposing homolog.pdb onto molrep.pdb (Calculate -> SSM Superpose), copying the ligand out of homolog.pdb (Edit -> Copy Fragment -> "//A/2001"), and then merging the copied ligand into molrep.pdb (Edit -> Merge Molecules).

(d) After any of the above quick pre-refinement model trimming that you choose to do – i.e. removing loop regions for which there is little or no supporting map density – save the coordinates of the fitted model: File -> Save Coordinates.
Now close Coot.
Continue either with your model or with the provided file: homolog_prepared.pdb

Part 2) Refinement Preparation

(a) Before refining the model, first investigate whether the map should be blurred/sharpened for refinement

- Open the CCP-EM interface.
- Select the “MRC to MTZ” task. Provide the ‘tutorial.mrc’ map, and specify the resolution (2.9 Å). If you like, you can also provide an input model here (‘homolog_prepared.pdb’) – it will not be used in the map sharpening, but can be helpful for visualising the results in Coot afterwards.
  Now run the job.
- Once the job has finished, look at the output plots in the “Plot” and “Plot ln” tabs. Is there an argument for blurring/sharpening the map for refinement?
  We will discuss this in the tutorial session, but in general, a well-sharpened map has structure factors that gradually decay towards zero at high resolution.
- If you click on the “Coot” button at the top-left of the interface, Coot will open with the array of blurred/sharpened maps loaded, along with the model (if you provided one as an input when you ran the job).
- *If you didn’t provide a model, you might not be able to see the map because you are at the corner of the box. You can display the box as discussed earlier and manually move around until you find the map density, or you can go straight to the middle of the box using Coot’s Cryo-EM module. First activate the Cryo-EM module (Calculate -> Modules -> Cryo-EM), then in the new Cryo-EM menu, click “Go to Box Middle”.*
- Look at the sharpened and blurred maps. You might need to use the Display Manager to show and hide them, or select them in turn using the Scroll button to change the contour levels. Judging them by eye, which map do you think is most useful? Does this agree with the selection you made from the results plots earlier?

(b) Refinement with map blurring

Now try to refine the model using Refmac5, blurring the map using a B-factor of 50 Å2

- In the CCP-EM interface, select the Refmac5 task.
- Set the following:
  - Input model: the model fitted in Coot earlier or homolog_preparered.pdb
  - Input map: tutorial.mrc
  - Resolution: 2.9 Å
- Open the “Refinement options” and set the following which will blur the map by 50 $Å^2$ prior to refinement.
  - Sharpen / blur: 50.0
- Now run the job.
- Did the job fail? If so, investigate why – ensure the “Pipeline” tab is selected, and click the “Refmac refine (global)” stage that is highlighted red. Why did it fail?
- *When CCP-EM detects an error, it will show the error output from the failed program. You can use a drop-down menu to switch to the standard output log which can also give useful information to identify the problem.*
- Continue on to the next step…

(c) Refinement with ligand restraints

Before being able to refine the model, we must ensure that we have restraint dictionaries for all new/unknown ligands (some common ligands are in the CCP4 Monomer Library and dealt with automatically). Our model contains an unknown ligand “PTP”. Consequently, we must create a dictionary for this compound. The provided file “PTP.cif” is an mmCIF file that contains a description that matches the ligand in the model (see the *Note* below). Now create a dictionary for this ligand using AceDRG:

- In the CCP-EM interface, select the AceDRG task and set the following:
    - mmCIF file: PTP.cif
    - Monomer code: PTP
      This will ensure that the output dictionary will work with the PTP ligand in the model.
- Now run the job. It will take a few minutes.

Running this command will create a dictionary called “PTP_acedrg.cif”, along with coordinates for a low-energy conformer “PTP_acedrg.pdb”, in the job directory. Once the job has finished running, select the “Launcher” tab to find the location of these files – you will need this for the next step.

Note: mmCIF files for ligands that aren’t in the CCP4 Monomer Library can often be found on the PDBeChem website (https://www.ebi.ac.uk/pdbe-srv/pdbechem). If an mmCIF file were not available, we could instead have used a SMILES string corresponding to the ligand. This would be an adequate starting point for ligand dictionary generation, the problem being that the atom names would be lost. Consequently, either the atoms would need to be renamed, or alternatively the ligand could be removed from the model and coordinates with the new atomic nomenclature refitted.
Addendum: The ligand used in this tutorial, 2-phenylethyl 1-thio-beta-D-galactopyranoside, is normally given the code PTQ. Since the tutorial was written, PTQ has been added to the CCP4 monomer library and therefore the refinement would run without the need to create a custom ligand dictionary. For this tutorial the ligand has been renamed as PTP to ensure the AceDRG step is still needed.

Part 3) Refinement

`(a) Refinement with a ligand dictionary`

Now try to refine the model again, using the newly created ligand dictionary.

- In the CCP-EM interface, select the Refmac5 task.
- Provide the map, your fitted model, and specify the resolution (2.9 Å).
- Also, in the “Input ligand” field, provide the PTP_acedrg.cif file that you created in the previous step.
- Open the “Refinement options”, and type “50.0” into the “Sharpen / blur” field.
- Now run the job (this might take a few minutes, depending on computing power).

`(b) Inspect the refinement statistics`

- Once the job has finished, click on the “Results” tab. Look at the refinement statistics table, as well as the graphs.
- Were 20 cycles sufficient, i.e. does refinement seem to have converged?

- Is there evidence that refinement has improved the model, or made it worse? Consider statistics representing fit to the data (FSC average), as well as geometric quality (Rms bond/angle/chiral).
- What is the major contributing factor to the improvement of refinement statistics between Start and Finish? Is it surprising that the refinement statistics have improved? (Hint: consider the refinement protocol, particularly the treatment of atomic B-factors – look at the commands passed to Refmac5, which are listed at the top of the "Refmac refine (global)" logfile).

`(c) Visual inspection`

Click on the "Coot" button at the top-left of the interface. Coot will open with the map and both the model before and after refinement loaded and displayed. The model before refinement will be coloured yellow, and the one after refinement green. If you cannot see the map then change the contour level (e.g. using the scroll wheel), and increase the map radius (e.g. to 70 Å).

- Inspect the models. Can you see any evidence of changes in the model, or improvements to local model quality?
- Zoom out so that you can see the whole model(s). Open the Display Manager. Hide the map, and change the representation of both models to "CAs + Ligands". Now repeatedly toggle the display of one of the models on and off. What differences can you perceive between the models? What does this tell you about the differences between the underlying macromolecular structures?
- Change the representation of both models to "Colour by B-factors - CAs". Now display/undisplay each of the two models in turn, and consider the colouring of the two models. What does this tell you about the differences between the biological assemblies of the two models (i.e. the structure under refinement versus the model of the homologous macromolecule)? Does this reflect anything about the relative resolutions of the maps underlying the two models?
- Compare the Ramachandran plots corresponding to the two models (*Validate -> Ramachandran Plot*). Are any differences due to overall trends or individual residues, and are they substantial or minor?

Part 4) Different Refinement Protocols

`(a) Refinement with modified parameters`

Now let's see if we can improve the model by adjusting refinement parameters. We want to improve the fit-to-data (as judged by the FSC), without overly negatively affecting the geometry (agreement with prior knowledge). The weight used during refinement can be found directly under the "Refinement Statistics" table on the Results page – make a note of what this value was for the previous refinement run (at the time of writing the automatic weight should be around 0.015-0.02). In order to loosen the geometry / improve fit-to-density, we need to increase this value to increase the weight given to the map restraints with respect to the geometry restraints.

- Clone the previous Refmac5 refinement job in CCP-EM (double click on the job, and then select "Clone" from the top left of the window).

- In “Refinement options”, untick “Auto weight” and specify a value that is 5–10x higher than the auto weight from the previous run (e.g. 0.1). Ensure that the “Sharpen / blur” field is still set to “50.0”.
- Run the job.
- Once it has finished, compare the refinement statistics from this job and the original one.
- Click on the “Coot” button in the upper-left portion of the window. Look at the Ramachandran plot from this refinement run, and compare with the original.

Which of the models do you prefer – the one with auto-weight, or the one with the higher manual weight?

(b) Refinement with modified parameters – attempt two

Increasing the weight (e.g. to 0.1) meant that the model suffered from overfitting – the model sank into the density (FSC improved), but ended up with excessively distorted geometry. Now let’s try a lower weight in order to improve the geometry (at the expense of fit to density):

- Clone the previous Refmac5 refinement job in CCP-EM.
- Ensure that “Auto weight” is unticked in “Refinement options”, and specify a manual weight that is 2–5x lower than the auto weight (e.g. 0.01).
- Ensure that the “Sharpen / blur” field is still set to “50.0”.
- Run the job.
- Once it has finished, compare the refinement statistics from this job with lower weight, the previous job with higher weight, and the original one with automatic weighting.
- Click on the “Coot” button in the upper-left portion of the window. Look at the Ramachandran plot from this refinement run, and compare with the other two jobs.

Which of the models do you prefer – the one with auto-weight, the one with a higher weight, or the one with a lower weight?

(c) Refinement with modified parameters – attempt three – using external ProSMART restraints

When refining the model, we’ve been using jelly-body restraints, which are enabled by default in the CCP-EM interface. These are very useful in helping to stabilise refinement, but can’t help us to improve the model. Another option is to use restraints from ProSMART in order to inject prior information from a high-resolution homologue during refinement, which will help the model to retain a conformation that is more consistent with prior observations. Specifically, these restraints will ensure that the local interatomic distances within the model do not stray too far from that in the homologous model:

- From the main CCP-EM screen, select “ProSMART” from the task list on the left.
- Ensure that the “Alignment mode” is set to “Reference model” (this is the default).
- Provide your PDB file containing the model to be refined in the “Target PDB(s)” field (e.g. “homolog_prepared.pdb”). Note: this is the same PDB file that you have been providing as the input to the Refmac5 interface.
- Provide the PDB file corresponding to the original homolog in the “Reference PDB(s)” field (e.g. “homolog.pdb”).
- Now run the job.

We now need to provide these restraints during refinement, in place of the jelly-body restraints:

- Clone the previous Refmac5 refinement job in CCP-EM.
- In “Refinement options”, reduce the number of “Refmac cycles” to “10” (this job will converge in fewer cycles).
- Ensure that “Auto weight” is selected.
- Ensure that the “Sharpen / blur” field is still set to “50.0”.
- Change “Jelly body” to “False”. Note that jelly-body restraints and external restraints work against each other – at present it is best to use one or the other, but not both. Ensure that “Add hydrogens” is set to “False”.
- Click on “External restraints”, and select the “Use restraints” tickbox (this is important!).
- In the “Restraints file” field, select the external restraints file. You will need to navigate to the correct directory, which corresponds to the CCP-EM ProSMART job – this directory will be called “~/ccpem_project/ProSMART_12” or similar. Within this directory will be another directory called “ProSMART_Output”, and within that directory there will be a “.txt” file that contains the restraints, e.g. “homolog_prepared.txt”.
- Directly below where the Restraints file is provided to the interface, there will be a “Weight” field that should be set to “10.0”, and a “GMWT” field that should be set to “0.02”.
- Run the job.
- Once it has finished, compare the refinement statistics from this job and the previous jobs.
- Click on the “Coot” button in the upper-left portion of the window. Look at the Ramachandran plot from this refinement run, and compare with the other jobs.

This latest job with ProSMART restraints hasn’t produced as good a model as was previously attained without these external restraints… why do you think this is? Hint: in Coot, load the “homolog.pdb” file that was used for external restraint generation and look at the Ramachandran plot corresponding to this homologous structure.

(d) Refinement with modified parameters – attempt three – using external ProSMART restraints from a re-refined model from PDB-REDO

When using ProSMART restraints, it is important to use the best quality reference model that you can. PDB-REDO provides re-refined models corresponding to PDB entries that were determined using X-ray diffraction – this is a useful resource when looking for reference models.
The 1.75 Å model with PDB code 3t09 is identical in sequence to the beta-galactosidase model we have been refining.

- Open a web browser and navigate to PDB-REDO (https://pdb-redo.eu).
- Enter the PDB code 3t09, to get to the relevant model summary.
- Download the PDB file corresponding to the “Re-refined and rebuilt structure”.

Now repeat the procedure followed in the previous step (c), to generate restraints using ProSMART with this new reference model, and subsequently re-refine the model using these additional restraints.
When refining the model, select the following options in the “Refinement options” tab:

- “Refmac cycles” = 20 (this job requires more cycles than in the previous step).
- “Auto weight” = ticked.
- “Sharpen / blur” = 50.0.
- “Jelly body” = False.
- “Add hydrogens” = False.
- Make sure to tick “Use restraints” in the “External restraints” tab, and provide the newly generated set of ProSMART restraints in the “Restraints file” field.

While this job is running, clone the job and start a new job that has the weight set to 0.02 (which should be a bit lower/tighter than the automatic weighting).
When the jobs have finished, inspect the refinement statistics and the Ramachandran plot in Coot. Which of the refinement protocol strategies produced the best model?

Part 5) Manual building and Validation

(A) Local interactive refinement
Unfortunately not all issues can be solved automatically. It is often necessary to inspect and correct the model in Coot in between rounds of refinement.
Since local resolution varies throughout the map, it is very useful to visualise multiple maps simultaneously when critiquing the model.

- Select the “MRC to MTZ” task. Provide the map, the PDB file corresponding to the output of the latest refinement run (this will be called “~/ccpem_project/Refmac5_13/refined.pdb” or similar), and specify the resolution (2.9 Å). Now run the job.
- Once the job has finished, click the “Coot” button in the upper-left corner of the window.
- Inspect the model and map, in order to identify any discrepancies. There are various tools in Coot to help with this, notably (but not exhaustively):
    - Validate -> Ramachandran Plot
    - Validate -> Unmodelled blobs
    - Validate -> Density fit analysis
    - Display Manager -> Colour by B-factors
- Go to residue 732. From looking at the overlaid blurred/sharpened maps, is it possible to use Coot’s real space refinement tool to correct the model in this region? *The real-space refinement tool is near the top of the right-hand toolbar. You will need to select a suitable map for Coot to use for the refinement, and possibly also adjust the refinement weight – see the “R/RC” button in the right-hand toolbar.*

When you have corrected the 732 loop region and any other problems you identified save the modified structure:

- File -> Save Coordinates

You can then re-run Refmac using the previously optimised parameters. Does the overall structure quality improve? Again consider statistics representing fit to the data (FSC average), as well as geometric quality (Rms bond/angle/chiral).

Look at the structures in Coot - have the B-factors in the 732 region changed?

(B) Building small sections
This is an example of model (re)building in Coot and these tools are useful for (re)building small sections of a protein. To demonstrate this, choose a section of the protein to remodel e.g. residues 285-289 (YADRV). In the right-hand toolbar use “Delete Item” (the bin icon), select “Residue/Monomer” and click on the residues to remove (tip - check “Keep Delete Active”). Then select the “Add Residue” button to replace the deleted residues by clicking on the nearest existing residue. Coot will add an alanine by default. Mutate it to the correct residue using the “Mutate and AutoFit” button (upper radiation sign). You may need to adjust the fit using “Real Space Refine Zone”.

(C) Building large sections - “Baton building”
This section is based on Paul Emlsey & Paul Bond’s Coot tutorial (https://paulsbond.co.uk/coot-workshop/part2.html)
Sometimes you may be required to build large sections of structure by hand e.g. the resolution is too low for automated *de novo* building, or there is a lack of good homologous structures or predictions. In these cases you can use Coot’s Baton building method.
To test this go to the starting residue MET1A. This allows you to build the complete chain in the correct direction and you can directly compare it to the real structure afterward. Once you are at residue MET1A, use the Display Manager to turn off the display of any existing molecules. Don’t look at it again until you have finished building, validating and refining!

The backbone trace through the density can be more easily visualised by using a skeleton map.

- Calculate -> Map Skeleton and change it to On.

To start to “baton build”

- Calculate / Other Modelling Tools -> C-alpha Baton Mode

A new menu will be displayed, as well as a white baton and a set of pink points that show possible places for a C-alpha that are 3.8 Å away from the current position. Note that when you start, you are placing a CA at the baton tip for residue 1 (the N-terminus). After placing CA for residue 1, it will switch to building in the C-terminal direction and you will get a choice of positions for residue 2, which is currently at the centre of the view. This might seem that you are “doubling-back” on yourself, which can be confusing the first time, but it is useful for extending existing chains.

Press “Try Another” and “Previous Tip Position” to cycle through the points, Lengthen and Shorten to change the baton length, and Accept to place a CA and move onto the next residue. If none of the guide points are suitable you can use b to toggle baton swivel mode.

Build from the N-terminus to the C-terminus. Try to build the first ~25-50 residues, which will probably take ~10-15 minutes. If you make a mistake you can press Undo to move back one residue. Click Dismiss on the baton controls when done.

Now to turn these CA positions into mainchain:

- Calculate / Other Modelling Tools -> Ca Zone to Mainchain

Click on the Baton model. It may take several seconds while it builds (note that you need at least 6 residues for this to work). Two new molecules will be created that are traced in different directions. We know the forward direction is correct (see how much better the carbonyls fit) so the “mainchain-backwards” and Baton Atoms models can be deleted.

To optimise the new chain’s fit-to-density use:

- Refine -> Chain Refine

Accept the results if the fit is better.

The next step is to convert the poly-alanine chain into a fully sequenced chain by assigning the sequence. There are two options for this. Option 1 is useful if you know the identity for all the residues you have built. Option 2 can be used if you know the sequence for the whole chain but have only built part of it.

Assigning the sequence - Option 1:

- Calculate -> Mutate Residue Range

Assigning the sequence - Option 2:

- Calculate -> Assign Sequence -> Dock Sequence (py)

Remember to select the “mainchain-forwards” model to mutate.
The full sequence is:

```
>homolog_prepared.pdb|Chain=A
MITDSLAVVLQRRDWENPGVTQLNRLAAHPPFASWRNSEEARTDRPSQQLRSLNGEWRFAWFPAPEAVPESWLECDLPEA
DTVVVPSNWQMHGYDAPIYTNVTYPITVNPPFVPTENPTGCYSLTFNVDESWLQEGQTRIIFDGVNSAFHLWCNGRWVGY
GQDSRLPSEFDLSAFLRAGENRLAVMVLRWSDGSYLEDQDMWRMSGIFRDVSLLHKPTTQISDFHVATRFNDDFSRAVLE
AEVQMCGELRDYLRVTVSLWQGETQVASGTAPFGYADRVTLRLNVENPKLWSAEIPNLYRAVVELHTADGTLIEAEACDV
GFREVRIENGLLLLNGKPLLIRGVNRHEHHPLHGQVMDEQTMVQDILLMKQNNFNAVRCSHYPNHPLWYTLCDRYGLYVV
DEANIETHGMVPMNRLTDDPRWLPAMSERVTRMVQRDRNHPSVIIWSLGNESGHGANHSRPVQYEGGGADTTATDIICPM
YRPLILCEYAHAMGNSLGGFAKYWQAFRQYPRLQGGFVWDWVDQSLIKYDENGNPWSAYGGDFGDTPNDRQFCMNGLVFA
DRTPHPALTEAKHQQQFFQFRLSGQTIEVTSEYLFRHSDNELLHWMVALDGKPLASGEVPLDVAPQGKQLIELPELPQPE
SAGQLWLTVRVVQPNATAWSEAGHISAWQQWRLAENLSVTLPAASHAIPHLTTSEMDFCIELGNKRWQFNRQSGFLSQMW
IGDKKQLLTPLRDQFTRAPLDNDIGVSEATRIDPNAWVERWKAAGHYQAEAALLQCTADTLADAVLITTAHAWQHQGKTL
FISRKTYRIDGSGQMAITVDVEVASDTPHPARIGLNCQLAQVAERVNWLGLGPQENYPDRLTAACFDRWDLPLSDMYTPY
VFPSENGLRCGTRELNYGPHQWRGDFQFNISRYSQQQLMETSHRHLLHAEEGTWLNIDGFHMGIGGDDSWSPSVSAEFQL
SAGRYHYQLVWCQK
```

You can then use Coot’s other validation and rebuilding tools to optimise your new structure followed by automated refinement with Refmac. You can also unhide the pre-existing model and see how well your de novo structure agrees with it.

Building models correctly is a time consuming process but it is necessary to give you and any others who may use it in the future the best possible structure to work with.