Refinement against cryo-EM data

Rob Nicholls, Rangana Warshamanage, Tom Burnley, Colin Palmer, Garib Murshudov

EMBO Workshop Birkbeck, September 2019

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.

(1) Map and Model Inspection

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

- Fix nomenclature error warning -> Click Yes
- 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 low – try increasing the map contour level (e.g. using the scroll wheel).
- It is often useful to display the box that encapsulates the map:
  Draw -> Cell & Symmetry -> Show Unit Cells.
  Then zoom in/out until you see the whole box (using the "m" and "n" keys).
- Increase the map radius in order to display more of the map (e.g. try 70 Å in this case):
  Edit -> Map Parameters -> Map Radius. [if Coot becomes unresponsive drop this to 30 Å]
- When zoomed out, it is often useful to display the model as a CA trace (via the Display Manager).

(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 not, 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.

- 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).

- The model includes a large number of water molecules. Should these be included in the model? (Validate -> Check/Delete Waters)
- Are there any ligands in the model (click the ligand icon, which is next to the "Display Manager" and "Go To Atom" buttons)? Does the map support the modelled ligand? If you are viewing the output of Molrep, consider whether there is a ligand in both the input and output models.
- Are there any side chain rotamers that can be fixed easily? (Don't spend time fixing them!)

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) Following any 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

(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 map, and specify the resolution (2.9 Å). Now run the job.
- Once the job has finished, look at the output plots. Is there an argument for blurring/sharpening the map for refinement?
- 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 the model was provided as an input to the "MRC to MTZ" task).

(b) Now try to refine the model using Refmac5, blurring the map using a B-factor of 50 Å.

- In the CCP-EM interface, select the Refmac5 task.
- Provide the map, your fitted model, and specify the resolution (2.9 Å).
- Open the "Refinement options", and type "50.0" into the "Sharpen / blur" field (this will blur the map by 50 $Å^2$ prior to refinement. Make sure to hit Enter, or otherwise click away from the field, in order to ensure that the interface registers the parameter value change.
- Now run the job.
- Did the job fail? If so, investigate why – ensure the "Pipeline" tab is selected, and click the red "Refmac refine (global)" job. Why did the job fail? Continue on to the next step…

(c) Obtain a ligand description – this is required in order to create a ligand dictionary. The ligand in the model has the three-letter-code: PTQ.. There are a few ways to obtain a ligand description:

- Use the PDBeChem online resource:
    1. Search for "pdbechem" in an online search engine.
    2. Search for the code "PTQ" using PDBeChem.
    3. Select "Download Links" from the left of the screen, and download the mmCIF file.
- **Or** open a terminal and download the mmCIF file directly using the command:
  `curl -o PTQ.cif ftp://ftp.ebi.ac.uk/pub/databases/msd/pdbechem/files/mmcif/PTQ.cif`
- **Or** continue with the provided file: PTQ.cif

(d) Now create a dictionary for the ligand using AceDRG:

- Open the command line interface, and navigate to your working directory.
- To get AceDRG usage instructions, type: **`acedrg --help`**
  (If you get a "command not found" error then it means that CCP4 is not set up as required)
- We want to generate a ligand dictionary for PTQ, using our mmCIF file, so the appropriate command is:
  `acedrg --mmcif=PTQ.cif -o PTQ_acedrg`

Running this command will create a dictionary called "PTQ_acedrg.cif", along with coordinates for a low-energy conformer "PTQ_acedrg.pdb", in your working directory.

Note: If an mmCIF file were not available, we could instead have used a SMILES string corresponding to the ligand, which is also available on the PDBeChem website. 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.

(3) Refinement

(a) 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).
- Open the "Refinement options", and type "50.0" into the "Sharpen / blur" field.
- Also in "Refinement options", provide your PTQ_acedrg.cif file in the "Ligand dictionary" field.
- Now run the job (this should only take a couple of 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, as well as geometric quality.
- 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 (in the Validate menu). Are any differences due to overall trends or individual residues, and are they substantial or minor?

(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). As can be seen in the Results page corresponding to the previous refinement run, the weight is ~0.015. In order to loosen the geometry / improve fit-to-density, we need to increase this weight.

- 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 higher weight of 0.2. Ensure that the "Sharpen / blur" field is still set to "50.0". Make sure to hit Enter, or otherwise click away from the field, in order to ensure that the interface registers the parameter value change.
- 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 weight 0.2?

(b) Refinement with modified parameters – attempt two.
Increasing the weight to 0.2 meant that the model suffered from overfitting – the model sank into the density, but ended up with excessively distorted geometry. Now let's try a more conservative weight:

- Clone the previous Refmac5 refinement job in CCP-EM.
- Ensure that "Auto weight" is unticked in "Refinement options", and specify a weight of 0.05.
- Ensure that the "Sharpen / blur" field is still set to "50.0". Again, make sure to hit Enter, or otherwise click away from the field, in order to ensure that the interface registers the parameter value change.

- Run the job.
- Once it has finished, compare the refinement statistics from this job, the previous job with weight 0.2, 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 weight 0.2, or the one with weight 0.05?

(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 left.
- Ensure that the "Alignment mode" is set to "Reference model".
- 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" (since we won't be using jelly-body restraints, which slow refinement, we can achieve convergence in fewer cycles).
- Ensure that "Auto weight" is unticked in "Refinement options", and specify a weight of 0.05, to match the previous job.
- 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.
- 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_X" 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". Again, make sure to hit Enter, or otherwise click away from the field, in order to ensure that the interface registers the parameter value change.
- 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.
- Which of the models do you prefer – the one with auto-weight, the one with weight 0.2, the one with weight 0.05, or the one with external restraints & weight 0.05? What are the benefits/drawbacks of each?

(5) Manual Intervention and Validation

Unfortunately, not all issues can be solved automatically, in general. It is often necessary to inspect and manually 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_X/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?

Note: It is important to perform additional validation, in order to ensure that overfitting during refinement is limited. If half-maps are available then they should be used for this purpose. In our example, half-map validation is too computationally expensive to do quickly in a tutorial session. However, in practice this can be achieved in the Refmac5 interface by scrolling down to "Validate options", selecting "Validate", and providing the two half-maps.