CCP4 Tutorial: Molecular Replacement

We will solve the hypF structure by molecular replacement, using several programs and approaches. Other MR examples can be found at the end of this tutorial.

When this tutorial is obtained as part of the CCP4 distribution, $MR_TUTORIAL corresponds to $CCP4/examples/mr_tutorial_2006 (Linux and Mac). It is a good idea to set up a new project in CCP4i that corresponds to this directory.

Note for Windows users: The windows distribution does not contain the mr_tutorial_2006 folder. A zip file containing the folder can be downloaded from:

http://www.ccp4.ac.uk/MrBUMP/mr-tutorial-2006.zip

Once downloaded, move the zip file to the $CCP4/examples folder and unzip it. The folder with all necessary files will now be in the correct place.

Checking the data

Target is the acylphosphatase-like domain of hydrogenase maturation factor HypF from E.coli, see Rosano et al, *JMB*, **321**, 785 (2002). HypF-ACP sulphate and phosphate complexes deposited as 1gxt and 1gxu respectively.

This protein has a Hg derivative. You may have processed this data in a preceding tutorial. We have prepared a reflection file for you including the data from 1gxu, 1gxt, the Hg derivative, and some experimental phases based on the Hg sites.

There is native data in H32 to 1.3 A resolution. The target has 91 residues and a Matthews calculation strongly suggests only one molecule in the asymmetric unit.

We first use **Sfcheck** to check a few things about the data:

1. Select the Program List module and open the sfcheck task window.
2. Enter a title.
3. Un-check Run Rampage to analyse structure geometry and Run Procheck to analyse structure geometry (we do not yet have any coordinates)
4. Select Run Sfcheck to analyse experimental data only
5. Enter MTZ in $MR_TUTORIAL/data/hypF/hypF-1gxu-1gxt-HG_scaleit1.mtz and select the labels F FP1gxu, SIGF SIGFP1gxu and Free FREE
6. Check that a suitable filename has been generated for Sfcheck Output PS
7. Keep all defaults, and click Run -> Run Now.

Sfcheck produces a postscript file with some useful things (see under View Files from Job):

- Anisotropy of data (it is not very anisotropic)
- Overall B from Wilson plot of 21.9 A**2
- Pseudo-translation not detected (from analysis of the native Patteron map)

Also check the log file View Files from Job then View Log File:

- This includes the results of the a twinning test:
- `Perfect twinning test <I^2> / <I>^2 :  2.0573`

  A value of 2.0 indicates untwinned data, whereas perfectly twinned data would have a second moment of 1.5

Choice of search models

The target is an acylphosphatase-like domain. A search of the PDB reveals two acylphosphatases with a sequence identity to the target of about 31% - 1v3z and 1w2i. Each has two chains in the asymmetric unit, either of which could be used as the basis of a search model.

Normally you would use something like Chainsaw at this point to prepare a search model from the template. As an exercise, we are going to try MR straightaway. We will return to Chainsaw later before running Phaser.

Notes on Sequence Alignment

There are many ways of approaching this, and the different tools will give slightly different assessments. The sequence identity depends on the definitions used (i.e. treatment of gaps and alignment length), the specific alignment technique, and whether bits have been chopped out of the model.

Molrep

Run 1

We will use chain B of 1v3z as the search model (file $MR_TUTORIAL/data/hypF/1v3z_B.pdb).

1. Select the Molecular Replacement module and open the Run Molrep - auto MR task window.
2. Enter a title.
3. Do molecular replacement performing rotation and translation function should be already selected.
4. Enter the MTZ file $MR_TUTORIAL/data/hypF/hypF-1gxu-1gxt-HG_scaleit1.mtz and select the labels FP FP1gxu and SIGFP SIGFP1gxu
5. Enter the PDB file $MR_TUTORIAL/data/hypF/1v3z_B.pdb
6. (Optional) You can use an upper resolution cut off of 3A to speed up the calculation, see folder Experimental Data.
7. Keep all defaults, and click Run -> Run Now.

When the job has finished, look at the log file (View Files from Job -> View Log File). Note the following:

- Molrep automatically estimates:

- `INFO: expected number of monomers :  1  Vmol:  61.4%`

  which is correct. The estimate may be unreliable when there are many monomers in the asymmetric unit, in which case it can be set explicitly with the keyword NMON (see folder Search Parameters in the Molrep GUI).

- Molrep checks whether or not an anisotropy correction is necessary:
- `INFO: Anisotropicy will not be used`
- The first table is a list of peaks of the Cross Rotation Function (CRF), sorted according to their heights.
- The second table shows the best Translation Function (TF) for each of the CRF peaks (scored according to the correlation coefficient * PKmax). Other TF solutions can be viewed in the file View Files from Job -> Output Files ... <proj_dir>_<job_no>_molrep.doc
- The final table gives a list of solutions, sorted according to the score.
- Molrep reports:
- `INFO: contrast is good enough. Stop this run`

  based on a contrast of 3.29 (the precise value you get will depend on Molrep version, resolution limits used, etc.)

Run 2

In fact, we can make use of our knowledge of the target, and this will often improve the solution. The search model has a moderately low sequence identity with the target and therefore the majority of the side chains are incorrect. Molrep can make use of the target sequence to improve the search model.

1. Select the previous job, and click ReRun Job
2. Most of the parameters should be set correctly, but you should change the title, and the name of the Coords out file, so that it is different from the first job.
3. This time, select Use sequence in the protocol section. A folder will open below where you can specify the name of the target sequence file $MR_TUTORIAL/data/hypF/hypF_Ndom.seq
4. Click Run -> Run Now

Look at the log file of this job.

- After a section about the input MTZ file, there are details of the sequence alignment between the target sequence you have supplied and the sequence of the search model (i.e. the PDB file).
- Molrep reports a sequence identity of about 30%. This is lower than other estimates because Molrep is more conservative in introducing gaps into the alignment.
- Molrep outputs tables for the CRF and TF as before.
- The contrast is now 3.53 and Molrep reports:
- `INFO: contrast is good enough. Stop this run`

Checking the solution

The top MR solution is applied to the input coordinates, and the positioned PDB file is written out as 1v3z_B_molrep2.pdb. The contrast indicates that this is probably a correct solution, but this should be checked!

The positioned model can be submitted for a few cycles of automated refinement, then checked manually against 2mFo-DFc and mFo-DFc maps, using a graphics program such as Coot. Since we have a good resolution dataset, the model can also be passed to ARP/wARP for rebuilding. Refinement, validation and model re-building are covered in other tutorials.

In fact, the Molrep solution is related to the deposited structure 1gxu by the symmetry operation -Y+2/3, X-Y+1/3, Z+1/3 Comparison of the structures in CCP4mg shows that the beta sheet and one of the two helices are well matched, but there are significant differences elsewhere.

In general, we can compare an MR solution to the deposited structure (or compare different MR solutions) as follows:

1. check for change of origin with cphasematch utility (Phase Comparison task in module Reflection Data Utilities).
2. apply any origin shift with Pdbset (Edit PDB File task in module Coordinate Utilities).
3. apply symmetry shift with csymmatch utility (not yet available in GUI).

Chainsaw

Search models can also be prepared using Chainsaw. Chainsaw takes an external sequence alignment, which can be generated by many bioinformatics tools and/or manually adjusted. In this job, we will create a model based on chain B of 1v3z, using a previously prepared alignment to the target.

1. Select the Molecular Replacement module and open the Create Search Model task window in the Model Generation folder.
2. Enter a title.
3. Leave Create search model using Chainsaw unchanged.
4. Leave Prune non-conserved residues to gamma atom unchanged.
5. Enter PDB in $MR_TUTORIAL/data/hypF/1v3z_B.pdb
6. Use the sequence alignment format PIR and enter the file Alignment in $MR_TUTORIAL/data/hypF/1v3z_B_to_target.pir
7. Click Run -> Run Now

Chainsaw produces a coordinate file 1v3z_B_chainsaw1.pdb which is an edited version of the input PDB file. 6 residues that do not align to the target sequence have been deleted. Of the rest, 34 have been left unchanged and 50 have had their side chains cut back to the gamma atom. The output PDB file uses the naming and numbering of the target sequence.

Have a look at the log file:

- At the top, the alignment used is confirmed.

- Then there is a listing of all the model residues, with the action applied (deleted, conserved, mutated).
- Finally, there is a summary of the changes made. This includes the estimated sequence identity. Note that this is not unique, but depends on the particular sequence alignment used.

Now repeat this exercise using the other search model, based on chain A of 1w2i. We can overlap the two models and use the ensemble as input to Phaser (in place of individual search models).

- Enter PDB in $MR_TUTORIAL/data/hypF/1w2i_A.pdb
- Use the sequence alignment format PIR and enter the file Alignment in $MR_TUTORIAL/data/hypF/1w2i_A_to_target.pir

Aligning the models

These models can be aligned and the overlapped structures used as input to Phaser.

1. Select the Coordinate Utilities module and open the Superpose Molecules task window.
2. Enter a title.
3. Change mode to Superpose using Secondary Structure Matching.
4. Enter Moving 1w2i_A_chainsaw1.pdb
5. Enter Fixed 1v3z_B_chainsaw1.pdb
6. Enter PDB out 1w2i_A_to_1v3z_B_chainsaw1.pdb
7. Click Run -> Run Now

The 1w2i_A_chainsaw1.pdb has been moved to overlap 1v3z_B_chainsaw1.pdb. The log file shows the transformation used, and gives an RMSD = 0.305 A between 84 C-alpha atoms of the superposed structures.

Phaser

Using the superposed search models generated by Chainsaw, we will now use Phaser to solve hypF. Phaser is designed to use ensembles of models to improve the signal.

1. Select the Molecular Replacement module and open the Run Phaser task window.
2. Enter a title.
3. Leave Mode for molecular replacement automated search unchanged.
4. Enter $MR_TUTORIAL/data/hypF/hypF-1gxu-1gxt-HG_scaleit1.mtz and select the labels FP FP1gxu and SIGFP SIGFP1gxu
5. In the folder Define ensembles ..., enter the PDB #1 1v3z_B_chainsaw1.pdb. Set the similarity to be sequence identity 0.38
6. To add another model click Add superimposed PDB file to the ensemble, enter the PDB #2 1w2i_A_to_1v3z_B_chainsaw1.pdb. Set the similarity to be sequence identity 0.38
7. In the folder Search details, select Perform search using ensemble1

8. In the folder Composition of the asymmetric unit, enter the SEQ file $MR_TUTORIAL/data/hypF/hypF_Ndom.seq and leave Number in asymmetric unit 1 unchanged.
9. Click Run -> Run Now

Have a look at the log file:

- After details about the input parameters, there is information on the anisotropy correction used (compare to the output of Sfcheck above). This is followed by a Matthews coefficient calculation.
- Phaser then calculates a Fast Rotation Function (FRF). It finds 5 peaks greater than 75% of the top peak (this threshold can be changed with the option Final selection criterion for ROTATION search peaks in the folder Search details).
- These peaks are passed to the Fast Translation Function (FTF). Detailed results for each rotation peak are given, followed by a summary table: Beware - these numbers may differ slightly for different versions of Phaser.
-
-
-
-
-
-
-
-
-

```
     Fast Translation Function Table: Space Group R 3 2 :H
     -----------------------------------------------------
     #SET #TRIAL      Top    (Z)    Second    (Z)     Third    (Z)
Ensemble
        1      1    44.70 ( 6.93)        -      -         -      -
ensemble1
        1      2    25.81 ( 4.74)    24.29 ( 4.56)    22.62 ( 4.36)
ensemble1
        1      3    22.91 ( 4.75)    22.02 ( 4.64)    21.27 ( 4.55)
ensemble1
        1      4    21.72 ( 4.36)    19.58 ( 4.12)    19.42 ( 4.10)
ensemble1
        1      5    20.93 ( 4.55)    20.92 ( 4.55)    20.31 ( 4.48)
ensemble1
     ---- ------
```

The first trial (based on the 1st peak of the FRF) gives a clear solution, with a good Z-score, and a single significant peak of the FTF.

- Next is a check on packing for this good solution. Phaser finds 2 clashes between a C-alpha and a C-alpha of a symmetry-related molecule. Because the threshold is set to 10 clashes, this solution is accepted.
- Finally, Phaser refines the MR solution, and displays the improvement in the log-likelihood gain (LLG).
- Phaser outputs a .sol file containing the MR solution, a .pdb file containing the correctly positioned model, and .mtz file containing the original data plus a calculated structure factor from the model and columns of map coefficients.

Checking the solution:

- In fact, the Phaser solution is related to the deposited structure 1gxu by the symmetry operation -X+Y+2/3, -X+1/3, Z+1/3 We are lucky - the spacegroup H32 has two possible origins (see $CHTML/alternate_origins.html), and the solution could equally

well have been on the origin (0.0,0.0,0.5). Comparison of the structures in CCP4mg shows that the beta sheet and one of the two helices are well matched, but there are significant differences elsewhere.

- The .pdb and .mtz files from Phaser can be inspected directly in Coot. This shows good agreement in most places, but also highlights problem areas.
- Do 20 cycles of restrained refinement in REFMAC.
- Run ACORN to remove all phase biase.
- Re-build in arp/warp using the ACORN phases as restraints.

MrBUMP

You have now prepared three search models based on 1v3z, and used Molrep and Phaser to do the molecular replacement. These steps, and the initial discovery of 1v3z and other related proteins, are automated in the program MrBUMP.

Depending on what you want to do, MrBUMP can make use of web-based services. The following tutorial deliberately does *not* make use of the web, so that it can be run anywhere. At the end of the tutorial, there are suggestions for web-based options. The use of a few local PDB template files also means that the tutorial is fairly quick. Beware that a full run of MrBUMP might take longer than is reasonable for a tutorial.

1. Select the Molecular Replacement module and open the Run MrBUMP task window.
2. Enter a title.
3. Leave Program Mode Model search and Molecular Replacement unchanged.
4. Enter SEQ in $MR_TUTORIAL/data/hypF/hypF_Ndom.seq
5. Enter MTZ in $MR_TUTORIAL/data/hypF/hypF-1gxu-1gxt-HG_scaleit1.mtz and select the labels F FP1gxu, SIGF SIGFP1gxu and Free FREE
6. Leave the rest of the files folder unchanged, and move to the Template Search Options folder.
7. Un-check Do a FASTA search for possible template models. Instead we are going to use some known local templates.
8. Un-check Update local copies of search databases
9. Un-check all Additional search methods, i.e. SCOP, PQS and SSM
10. The folder User specified search models will have opened. Because we have switched off all search options, we are required to use local files. Click on Add PDB file 3 times to add 3 local PDB files. The first file is $MR_TUTORIAL/data/hypF/1w2i_A.pdb and specify Chain identifierA. The second file is $MR_TUTORIAL/data/hypF/1v3z_B.pdb and specify Chain identifierB. The third file is $MR_TUTORIAL/data/hypF/2acy.pdb and specify Chain identifierA.
11. In the folder Search Model Preparation Options, keep the default which is to use Molrep and Chainsaw. This means there will be 6 search models in total. Turn one off to make the job quicker.
12. In the folder Molecular Replacement and Refinement Options, keep the default which is to use Molrep only. If you want, you can use Phaser instead or both.
13. Click Run -> Run Now

After a few minutes, have a look at the MrBUMP log file (do not wait for the job to finish).

- At the top, it echoes the options selected.

- Under Target Information, it estimates that there is 1 molecule in the target asymmetric unit.
- Under Template Model Search Results, it lists the three local files entered. They are named "loc0", "loc1", "loc2" for internal use.
- Under Search Model Preparation Results, details of the Molrep and Chainsaw methods are given.
- Finally, the section Molecular Replacement and Refinement gives details for every MR job tried.

By default, it will finish when it finds a solution. For example, it may finish with model loc1_B_MOLREP, which corresponds to template 1v3z_B.pdb with a search model created with the Molrep editing features. The Rfree drops from 0.549 to 0.436 (precise numbers may vary!) indicating that the MR solution is refinable, and likely to be correct. If you want to try all search models in MR (a good idea unless you are in a rush), select Finish when all of the search models have been tried in MR in the folder Molecular Replacement and Refinement Options.

If there are no problems accessing web-based services, then you can search for templates rather than use local PDB files. Run as above, with the following differences:

1. In the folder Template Search Options, check Do a FASTA search for possible template models.
2. Check Run the FASTA search locally. This refers just to the search step - the PDB files are still downloaded from the web.
3. Check all of the Additional search methods, i.e. SCOP, PQS and SSM
4. Do not enter anything into the folder User specified search models.

For comparison, here are some example results from MrBUMP (you may not get exactly the same):

PDB chain	sequence identity	source / release date	Rfree from MrBUMP
1w2i_B	0.310	OCA - released Apr 2005	chainsaw 0.447 molrep 0.442
1w2i_A	0.310	OCA	chainsaw 0.471 molrep 0.527
1v3z_B	0.310	OCA - released Mar 2005	chainsaw 0.430 molrep 0.453
1v3z_A	0.310	OCA	chainsaw 0.474 molrep 0.470
2bje_G	0.287	OCA - released Nov 2005	chainsaw 0.458 molrep 0.442
2bje_E	0.287	OCA	chainsaw 0.468 molrep 0.486
2bje_C	0.287	OCA	chainsaw 0.491 molrep 0.481
2bje_A	0.287	OCA	chainsaw 0.448 molrep 0.443

2bjd_B	0.287	OCA - released Nov 2005	chainsaw 0.468 molrep 0.529
2bjd_A	0.287	OCA	chainsaw 0.544 molrep 0.466
1y9o_A	0.275	OCA - released Jan 2006 (NMR)	(not tried)
1ulr_A	0.286	OCA - released Nov 2004	chainsaw 0.476 molrep 0.471
2acy_A	0.264	SSM - released Nov 1997 (authors tried)	chainsaw 0.539 molrep 0.564

Advanced tutorial (OPTIONAL)

Other search models for hypF

Another possible search model is chain A of 1w2i. This is a different structure of the same protein as 1v3z. Try repeating the above steps using $MR_TUTORIAL/data/hypF/1w2i_A.pdb as the search model.

You should find that this is more difficult! Modifying the search model using the target sequence is now necessary. Adjusting the resolution limits also helps.

Check your solutions against those produced from 1v3z_B.

Phased TF

As an exercise use Molrep or Amore with the Hg phases to solve the structure using the phased translation function. You will need to modify the input to Amore from the mtz file to read FP=FP1gxu PHI=PHIB_mlphare1 W=FOM_mlphare1

Other structure soltuions

See separate document for 3 more example MR problems.