This tutorial will provide a general outline on how to validate a molecular model / map. This tutorial follows the previous model optimization tutorial.

The goal of this tutorial will be to validate our previously optimized model.

Before we start, here is some helpful information:

phenix.molprobity

```
Usage: phenix.molprobity model.pdb [data.mtz] [options ...]

Run comprehensive MolProbity validation plus R-factor calculation
(if data supplied).
```

e2pdb2mrc.py

```
Usage: prog [options] input.pdb output.mrc

Converts a pdb file into an electron density map. 0,0,0 in PDB
space will map to the center of the volume. Use e2procpdb.py to
adjust coordinates, apply symmetry, etc. Resolution is equivalent
to standard cryoEM definition, using 1/2 width of Gaussian in
Fourier space.
```

e2proc3d.py

```
usage: e2proc3d.py [options] <inputfile> <outputfile>

Generic 3-D image processing and file format conversion program.
All EMAN2 recognized file formats accepted (see Wiki for list). We
will be using --calcfsc.
```

Stereochemistry check with Molprobity

1. Navigate to the directory with models and run the following:

   ```
   Validate hryc$ phenix.molprobity 1DP0_fit.pdb
   ```

   Which should produce:

   ```
   =============== Summary ==============

     Ramachandran outliers =   0.17 %
                   favored =  96.51 %
     Rotamer outliers      =   3.53 %
     C-beta deviations     =   358
   ```

```
  Clashscore             =   5.55
  RMS(bonds)             =   0.0235
  RMS(angles)            =   2.84
  MolProbity score       =   1.95
  Resolution             =   1.70
  Refinement program     = TNT
```

2. We can then run our model:

   ```
   Validate hryc$ phenix.molprobity Complex_rsr_result.pdb
   ```

   Which should produce something like the following:

   ```
   ================ Summary ==============

     Ramachandran outliers =   0.00 %
                  favored =  97.42 %
     Rotamer outliers      =   0.00 %
     C-beta deviations     =     0
     Clashscore            =  11.82
     RMS(bonds)            =   0.0098
     RMS(angles)           =   1.13
     MolProbity score      =   1.70
     Refinement program    = PHENIX
   ```

3. Ideally, we would then do manual corrections in COOT based on our Molprobity results:

   ```
   Validate hryc$ coot molprobity_coot.py Complex_rsr_result.pdb
   ```

   This allows us to work through individual clashes and improve the ramachandran plot. This would be iterated with the real-space refinement process. To obtain percentiles, which would allow one to compare this structure and other structures, enter resolution in the header of the PDB file and run the structure at the Molprobity website (http://molprobity.biochem.duke.edu/).

Comparing Map vs. Model

R-values are poor approximation of fit-to-density since segmentation and masking can greatly alter the results. Correlation is an effective way of comparing map vs. model.

A quick an easy way to monitor correlation during a Phenix real-space refinement is to check CC around atoms and CC within the unit cell. Both are displayed throughout and most importantly before and after the refinement.

Another option for a quick and easy way to assess correlation is to use Chimera's Fit in Map too with advanced options. This map quickly generates a model at an assigned resolution and

A common method to assess correlation after refinement is to compute an FSC between map and model. To do this, one can use e2pdb2mrc.py (in the terminal) to create a simulated map, from the model, that **somewhat** resembles the actual density map.

```
Optimize hryc$ e2pdb2mrc.py Compelx_rsr_result.pdb
rsr_32A_simiulated_map.mrc --apix=0.637 --res=3.2
```

There will however be variation in the data which is attributed to the lack of B-factors per-atom (This has been added but is not fully functional). Once a map is generated from the model, a soft mask (10-15Å soft mask) should be **applied to the original density map** before an FSC is computed (using **e2proc3d.py**). Moreover, the simulated map needs to have the same origin as the raw data, as long as the same Å/pix. I resampled the data using Chimera similar to that as we did in the optimization tutorial:

> Resample your map onto the emd_5995.map grid. To do this open the Chimera command line and type "**vop resample #0 ongrid #1**", where #0 is your map and #1 is the emd_5995.map.

Then save the resampled map as rsr_32A_simiulated_map_rs.mrc.

One can then compute the FSC in the terminal with EMAN2 using the following command:

```
Optimize hryc$ e2proc3d.py emd_5995_masked.mrc
RSR_simulated_map-vs-EMD_5995.fsc
--calcfsc=rsr_32A_simulated_map_rs_masked.mrc
```

After computing the FSC, the resolution value to which the maps are correlated to should be read at 0.5, as opposed to the 0.143 for the gold standard resolution, since the model is directly computed from the original density map.

Comparing Maps and Models from Independent Data Sets

To ensure that our model optimization process does not over-fit the data we look to our half-data sets, with a slightly worse resolution than the combined data set. Using our final, optimized molecular model (optimized complex), we run Phenix.real_space_refine with a map using half the data set, for instance Data Set 1 or the Even Map (which we will call **EMAN2_threed4_EVEN.map,** data not in tutorial). At this step, we like to give the model the most amount of movement to fit the half data set. Thus, we typically use the simulated annealing feature:

```
Optimize hryc$ phenix.real_space_refine Complex_rsr_result.pdb
EMAN2_threed4_EVEN.map resolution=4.2
run=minimization_global+adp+simulated_annealing
```

Once done, a simulated map is generated from the model and an FSC is computed between the simulated map and the Even map. Following this FSC, an FSC is computed between the simulated map and the Odd map. Ideally, the optimized model, using the Even map, should result in a better FSC curve with the Even data than compared to the Odd data.

If we optimize another model with the Odd data set, and compare the variation that exist between the Even model and the Odd model, we obtain a rough estimate to the amount of variation that exist within the data. Moreover, this can be compared to B-factors that are now produced the Phenix.real_space_refine (**run=adp**).

Helpful References:

http://molprobity.biochem.duke.edu/

https://www.phenix-online.org/documentation/tutorials/molprobity.html

Bartesaghi, Alberto, Doreen Matthies, Soojay Banerjee, Alan Merk, and Sriram Subramaniam. "Structure of B-galactosidase at 3.2-Å Resolution Obtained by Cryo-electron Microscopy." *Proceedings of the National Academy of Sciences of the United States of America* 111, no. 32 (2014): doi:10.1073/pnas.1402809111.

Wang, Zhao, Corey F Hryc, Benjamin Bammes, Pavel V Afonine, Joanita Jakana, Dong-Hua Chen, Xiangan Liu, *and others.* "An Atomic Model of Brome Mosaic Virus Using Direct Electron Detection and Real-space Optimization." *Nature communications* 5 (2014): doi:10.1038/ncomms5808.