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. It should be noted that a large portion of validation is just knowing your system and what to expect, such as sequence, protein fold etc. The outlined routines here are helpful to ensure that the optimization process has both improved the model and that no modeling issues exist.

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).
```

phenix.map_model_cc

```
Usage: phenix.map_model_cc model.pdb map.mrc resolution=X

Computes the correlation between map and model for individual
subunits, and at the main-chain and side-chain level.
```

phenix.emringer

```
Usage: phenix.emringer input.pdb map.mrc

Calculates EMringer score, which assess model fitting at at the
side-chain level.
```

e2pdb2mrc.py

```
Usage: e2pdb2mrc.py input.pdb output.mrc --apix=X --res=X

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 original (fit, not refined) model compared to the experimental map:

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

Which should produce:

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

 Ramachandran outliers =   0.37 %
               favored =  97.18 %
  Rotamer outliers      =   2.72 %
  C-beta deviations     =    10
  Clashscore            =   4.16
  RMS(bonds)            =   0.0127
  RMS(angles)           =   1.43
  MolProbity score      =   1.68
```

2. We can then run our optimized model:

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

Which should produce something like the following:

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

Ramachandran outliers =   0.24 %
               favored =  96.73 %
  Rotamer outliers      =   0.51 %
  C-beta deviations     =     0
  Clashscore            =   3.10
  RMS(bonds)            =   0.0093
  RMS(angles)           =   1.30
  MolProbity score      =   1.31
  Refinement program    = PHENIX
```

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

```
Validate hryc$ coot molprobity_coot.py
1yar_fit_real_space_refine.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 alters the results. There are a few established methods to quickly assess map to model fit. The first being cross-correlation, which can be done at various levels, such as the side-chains, secondary structure elements, individual subunits, and whole complexes. When generating a correlation between map and model at the complex level, one typically computes the Fourier shell correlation or FSC between map and model. This is then reported at the publication stage. EMringer is also a quick and easy way to assess the model at the side-chain level and reveals a single score which would indicate if an improvement has occurred during the model optimization process (discussed below).

Correlation is an effective way of comparing map vs. model. One way to monitor cross-correlation during a Phenix real-space refinement is to check CC around atoms and CC within the unit cell. Both are displayed throughout the refinement process and can be computed by running the following:

```
Optimize hryc$ phenix.map_model_cc
1yar_fit_real_space_refine.pdb emd_6287.map resolution=2.8
```

If this is performed before and after model optimization, one should see an improvement in correlation at both the main, and side-chain level.

Furthermore, another 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 then computes the cross correlation between the model generated map and the experimental map.

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 1yar_fit_rsr_complex.pdb
rsr_2p8A_simiulated_map.mrc --apix=0.982 --res=2.8
```

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, experimental 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_6287.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_6287.map.

Then save the resampled map as rsr_2p8A_simiulated_map_rs.mrc.

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

```
Optimize hryc$ e2proc3d.py emd_6287_masked.mrc
rsr_simulated_map-vs-emd_6287.fsc
--calcfsc=rsr_2p8A_simulated_map_rs_masked.mrc --apix=0.982
--res=2.8
```

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.

Finally, another routine that will quickly reveal the improvement of map to model fit would be EMringer (Barad, 2015). This method assess the model fit at the side chain level. The overall EMringer score should improve when running in phenix. The following reveals the improvement before and after running our refinement. Note that the EMringer score improved from 2.00 to 4.56.

Before refinement -

```
Validate hryc$ phenix.emringer 1yar_fit.pdb emd_6287.map

     =====Final Statistics for Model/Map Pair=====
     Optimal Threshold: 0.037
     Rotamer-Ratio: 0.665
     Max Zscore: 8.295
     Model Length: 1715
     EMRinger Score: 2.002979
```

After refinement -

```
Validate hryc$ phenix.emringer 1yar_fit_real_space_refine.pdb
emd_6287.map
     =====Final Statistics for Model/Map Pair=====
     Optimal Threshold: 0.028
     Rotamer-Ratio: 0.799
     Max Zscore: 18.907
     Model Length: 1715
     EMRinger Score: 4.565486
```

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 **`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 (this method is a bit slower, but allows for the model to move more) feature:

```
Optimize hryc$ phenix.real_space_refine
1yar_fit_real_space_refine.pdb EVEN.map resolution=4.2
run=minimization_global+adp+simulated_annealing
```

Once done, a simulated map is generated from the model (in Chimera, like above) 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 produced in Phenix.real_space_refine (**run=adp**).

Helpful References:

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

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

Barad B.A., Echols N., Wang R.Y.-R., Cheng Y.C., DiMaio F., Adams P.D., Fraser J.S. EMRinger: side-chain-directed model and map validation for 3D electron cryomicroscopy. Nature Methods 12 943-946 (2015); doi:10.1038/nmeth.3541.

Campbell MG, Veesler D, Cheng A, Potter CS, Carragher B. 2.8 Angstrom resolution reconstruction of the Thermoplasma acidophilum 20S proteasome using cryo-electron microscopy. eLife (2015) 4, pp. E06380-e06380.

Goddard, T. D.; Huang, C. C.; Meng, E. C.; Pettersen, E. F.; Couch, G. S.; Morris, J. H.; Ferrin, T. E. (2017). "UCSF chimerax: meeting modern challenges in visualization and analysis". Protein Science. 27.

Hryc CF*, Chen D-H*, Afonine PV, Jakana J, et. al. Accurate Model Annotation of a Near-atomic Resolution Cryo-EM Map. Proceedings of the National Academy of Sciences of the United States of America doi:10.1073/pnas.1621152114 (2017).

PHENIX: a comprehensive Python-based system for macromolecular structure solution. P. D. Adams, P. V. Afonine, G. Bunkóczi, V. B. Chen, I. W. Davis, N. Echols, J. J. Headd, L.-W. Hung, G. J. Kapral, R. W. Grosse-Kunstleve, A. J. McCoy, N. W. Moriarty, R. Oeffner, R. J. Read, D. C. Richardson, J. S. Richardson, T. C. Terwilliger and P. H. Zwart. Acta Cryst. D66, 213-221 (2010).

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.