Lesson 15: Unsupervised Classification / Clustering

15 PROTOFILAMENT MICROTUBULE

PEET uses a version of Principal Component Analysis (PCA) adapted to minimize missing wedge artifacts followed by k-means clustering to detect and manage heterogeneity in averaged subvolumes. First let's examine the initial 15 protofilament microtubule alignment described in an earlier exercise. Recall that in this example we used axial randomization to suppress missing wedge artifacts; some missing wedge artifacts will still be present, however.

1)
```
cd $WORKSHOP_HOME/PEET_Labs/MT/PEET/firstSearch
```

2)
```
pca *.prm 2 243 series4_8um_AvgVol_2P243.mrc
```
Principal component analysis is cpu and memory intensive. In this case, however, we're using binned data and only 243 particles, so the analysis will complete fairly quickly. Based on the resulting plots, it looks like the first 4 or 6 principal components will be useful as features for clustering. We'll go ahead and try using the first 6. Typically, you would try several combinations. In this case, the results are insensitive to the specific features chosen. Close the plot windows when finished examining them. The pca program automatically saves pdf copies of these plots for you.

3)
```
clusterPca *.prm pca243_series4_8um.mat 2 1:6
```
Generally, I recommend starting with a small number of features and clusters and working your way up as needed. Some users routinely use the first 20 features, although I consider this risky

and prone to overestimating heterogeneity. Notice from the output in the terminal window that both AIC and BIC indicate significant improvement in scores as a result of clustering. Roughly speaking, and under assumptions which are seldom fully satisfied, the likelihood of getting a score improvement *S* by random chance is approximately *exp(-S/2).*

4) Use the rotate tool to view the cluster plot from different perspectives. While we have clustered in a 6-dimensional space, things have been projected down to the 3-dimensional space of the first 3 features (principal components) for visualization. You'll see that the data look somewhat like the outline of a saddle or a potato chip. While the AIC and BIC suggest that the clustering is highly non-random, this figure strongly suggest that we are looking at continuous variation of a single parameter rather than discrete clusters. Notice that we could arbitrarily rotate the boundary between the 2 classes around this saddle / potato chip and still get highly significant results. As discussed in the lecture, axial position around this shape likely corresponds to axial orientation in the original tomogram. This is a case where, despite our efforts to suppress them, we've wound up clustering according to residual missing wedge artifacts. It's advisable to always be aware of this possibility. Missing wedge artifacts are almost always present in the tomograms and are frequently dominant. There is no indication that the subvolumes themselves are actually heterogeneous in this instance. Close the plot window when finished.

MUTANT CHLAMYDOMONAS RS2

Recall from the lecture that the density of Radial Spoke 2 (RS2) was greatly attenuated in averages from the 6E6 mutant strain compared to wild type. Let's analyze this case further to see whether this strong but incomplete reduction in density is due to conformational flexibility or because the majority of structures completely lack RS2.

5) `cd ../../../Axoneme/`

6)

```
3dmod PEET/pWT/pWT_AvgVol_3P159.mrc \
PEET/6E6/6E6_AvgVol_3P162.mrc
```

where “\” should be followed immediately by **Enter**. Compare the two averages and notice that RS2 is almost completely missing in the case of 6E6, as previously described. Exit 3dmod when finished. Because this is a large, somewhat variable structure, we’ll restrict pca’s attention to RS2 by using a binary mask. An appropriate mask has already been created by Tom Heuser, and is in *RS2MaskLong.mrc*. Let’s apply it to the average so we can visualize the results and verify that it selects RS2.

7)

```
applyBinaryMask RS2MaskLong.mrc \
PEET/pWT/pWT_AvgVol_3P159.mrc masked_pWT.mrc
```

8) `3dmod masked_pWT.mrc`

9) In the 3dmod info window, adjust the **Black** and **White** sliders to **200** and **255**, respectively, and page up and down to verify that the masked region includes most of RS2 and little else. Exit 3dmod when finished.

10) `cd PEET/6E6`

The alignment has already been run and intermediate files removed.

11) `gedit 6E6.prm`, scroll to the end of the file, and uncomment (**remove the “#”**) from the line #**pcaFnParticleMask = '../../RS2MaskLong.mrc'**. Normally, you would have to add this line manually. This tells the pca program to use this mask. See the pca man page for additional details. See the imodmop man page for one way to create such a binary mask. Save the modified prm file and exit gedit.

12) `pca *.prm 3 162 6E6_AvgVol_3P162.mrc`

Examine the figures after the program completes. Notice that the first principal component is significantly larger than any other, and that the first 6 principal components explain about 20% of the variance. We will use components 1-6 as features for clustering. Close all the figure windows.

13) `clusterPca *.prm pca162_6E6.mat 2 1:6`

Here, we've asked clusterPca to split the data into 2 classes using features 1 through 6. The AIC and BIC scores indicate significant results. Notice that the majority of the points are in class 1, which is reasonably compact. Class 2 is spread out and probably contains several additional subclasses. Because it contains so few particles, however, we won't try to further subdivide it. Close the figure window.

14) After backing up the original, we copy the motive list created by clustering into place.

```
cp 6E6_MOTL_Tom1_Iter4.csv 6E6_MOTL_Tom1_Iter4.csv.orig

cp pca_6E6_MOTL_Tom1_Iter4.csv 6E6_MOTL_Tom1_Iter4.csv
```

This version will have the class numbers assigned during clustering in column 20 of the motive list. (Verify this if you like).

15) `etomo *.epe`

16) On the **Run** tab, set **Average only members of classes** to **1** and press **Remake averages**. Wait for averaging to finish.

17) `mv 6E6_AvgVol_3P128.mrc class1_AvgVol_3P128.mrc`

Notice that the number of particles in the class averages matches the number reported by clusterPca. It's good practice to rename class averages to avoid confusing with the original averages. In fact, you should also take care that a class average does not accidentally over-write a prior average… *i.e.* if they happen to contain the same number of particles. When saving results from multiple attempts at clustering, I recommend moving the class averages and other results you wish to preserve… *e.g.* figures, motive lists with class labels, etc… to a subdirectory. In this case, for example, I might create a subdirectory named "*2ClassesFeatures1-6*".

18) On the **Run** tab, set **Average only members of classes** to **2** and press **Remake averages**. Once again, wait for averaging to finish.

19)
```
mv 6E6_AvgVol_3P034.mrc class2_AvgVol_3P034.mrc
```

20) Clear the entry for **Average only members of classes** and exit Etomo. The prm file will be automatically saved on exit. The Average only members of classes setting, which corresponds to selectClassID in the prm file, is an example of a parameter which can lead to very confusing results if accidentally left in an unintended state!

21)
```
3dmod class*.mrc
```

Notice that class 1, contains most of the particles and appears to be comprised of particles in which RS2 is completely absent. The minority class 2, on the other hand, consist of particles which all seem to contain RS2 in its normal configuration as it appears at full or near-full intensity. As mentioned above, class 2 may consist of particles with varying conformations. Because there are so few particles in this class, we will not attempt to analyze this heterogeneity further here. Exit 3dmod when finished.