BMI203: HW2

Laura Shub

February 2020

Q1. *Explain your similarity metric, and why it makes sense biologically.*

My low-dimensional representation is an array containing the count of each of the twenty amino acids. I then calculate the center of each residue and create a distance matrix between all the residues. I append the mean value of this matrix to the array. Each value is then normalized to between 0 and 1. Distance is then scaled by a factor of 3 times the average sum of amino acid elements such that it is considered first before amino acid composition. Similarity is the Euclidean distance between these 21-element vectors. The encoding and similarity metric that I used puts clusters with a similar distance between all amino acids together, followed by amino acid composition. Theoretically, active sites of different sizes will have different catalytic activity because they will be interacting with differently-sized ligands, even if they have a similar amino acid composition. However, amino acid composition does still have an effect on catalytic activity and function within this group, so including these in the distance calculation is important biologically.

Q2. *Explain your choice of partitioning algorithm.*

I used k-means with k between 3 and max(9, number of points). K-means++ cluster initialization was used to try and avoid poor initial clustering. Clustering is also performed 10 times for each value of k tested to account for possible poor initialization. The best silhouette score is used to determine which of the resulting clusters to return. K-means was chosen due to its fast runtime, which allows for the multiple trials for values of k, and ease of understanding/implementation.

Q3. *Explain your choice of hierarchical algorithm.*

For hierarchical clustering, I used an agglomerative clustering algorithm that starts with each active site in its own cluster. At each step, the two closest clusters by complete-linkage are joined. This repeats until all active sites are together in a single cluster. I used complete-linkage instead of single-linkage because it avoids "chaining" together elements based on the proximity of a single member of each cluster.

Q4. *Explain your choice of quality metric. How did your clusterings measure up?*

The metric I used for the quality of the clusters was the average silhouette score across all points. This compares the distance between each element to other elements in its cluster with the distance to the elements of the nearest other cluster. The resulting clusters from both clustering methods had a silhouette score of 0.42-43. Because silhouette score ranges from -1 to 1, with 1 being ideal, this suggests that the clusters are ok, but not perfect. This could be due to the fact that the data does not have any underlying shape in my low-dimensional representation.

Q5. *Explain your function to compare clusterings. How similar were your two clusterings using this function?*

To compare the clusterings, I used the Jaccard index, which checks if two sites are in the same cluster across the two different clusterings. Comparing the clusters, I get an average Jaccard index of approximately 0.3 out of a possible range from 0 to 1, indicating that the two algorithms give different clusters. This can also be determined by visual inspection of the UMAP space for the two clusterings. Hierarchical clustering gave 3 clusters, with the main difference between the three being the average distance between the center of each residue. K-means further broke down these clusters based on amino acid composition.

Q6. *Did your clusterings have any biological meaning?*

Hierarchical clustering resulted in 3 clusters. The first of these only contains two active sites, 34088_A and 18773_A, which are unique in that they are very small (3 residues each) and contain proline, which is uncommon and likely resulted in these being distant from the other active sites. The next largest cluster consists of active sites with a large number of amino acids, some of which are very spread out, which lead to a larger average distance between them. These active sites probably bind larger ligands. The last cluster has shorter, condensed active sites.

In k-means, a similar trend was observed. The two clusters that in hierarchical were by themselves are still together, but other clusters with a similar inter-residue average distance are added (20856_A, 71389_A, 38081_A). The large cluster that contained the more condensed active sites is divided into a number of smaller clusters based on their amino acid composition, ex. 46495_A, 82212_A, 15813_X, 85232_A, 3458_A all consist mostly of aspartate, lysine, and glutamate, and are in a separate cluster from 4629_X, 91911_A, and 32088_A, which each contain a larger proportion of serines, arginines, and tyrosines.

While both resulting clusters likely have some biological relevance, I prefer the k-means clusters as they seem to take into account both the distance between residues and the amino acid composition, as opposed to grouping the site by distance alone.

Github: `https://github.com/laurashub/BMI203HW2`.