MIDTERM DEMO-1

REVIEW OF CONCEPTS

1. INTRODUCTION TO NETWORK MODELS
   * What is a social network?
   * What is a genome network?
   * What is the concept of genome architecture mapping?
   * What are the major steps of the GAM method?
   * What is the genome network that is captured by GAM?
2. CHROMATIN ORGANIZATION
   * What is chromatin and what is its function?
   * What is the functional reason for a DNA loop?
   * What is the name of the protein that often helps to form DNA loops?
   * What are the functions of DNA ‘marks’?
   * What is a nuclear sub-compartment?
   * What is the purpose of genome folding?
   * What is “the radial position of a nuclear profile (NP)”? How can it be estimated by using GAM data?
   * How can the degree of compaction of a genomic region (a locus) be estimated by using GAM data?
3. INTRODUCTION TO NUCLEAR ORGANIZATION AND FUNCTION
   * What is the most important function of the nucleus?
   * What is the nucleolus and what does it do?
   * What is the nuclear pore and what is its function?
   * Which chromosomes contain the histone genes in the mouse genome?
   * What is the Histone Locus Body (HLB)?
   * Where is the Hist1 region of the HLB located on chromosome 13?
4. INTRODUCTION TO CLUSTERING
   * What are some examples of how clustering can be used?
   * What are the major steps of the k-means clustering algorithm?
5. UCSC GENOME BROWSER
6. GENOMIC AND EPIGENOMIC FEATURES
   * Histone genes
   * Lamina associated domains (LADs)
   * Binary feature table

REVIEW OF ACTIVITIES

1. SEGMENTATION TABLE
   1. Obtain the data
   2. Write a program to read the file and compute the following:

* Number of genomic windows
* Number of NPs
* On average, how many windows are present in an NP?
* What is the smallest number of windows present in any NP? The largest?
* On average, what is the number of NPs in which a window is detected? The smallest? The largest?
  1. Evaluate data quality
     + Identify windows that have an unusually high detection frequency (the detection frequency of a genomic window is the number of NPs that detect the window). One way to do this is to sort the windows by detection frequency and then to look for outliers at the top of the sorted list. Alternately, this can be done by generating a scatter plot and visually determining what threshold constitutes an unusually high detection frequency.

1. ESTIMATE RADIAL POSITION OF EACH NP
   1. The GAM paper describes how to estimate radial position of an NP.
   2. Use these insights to estimate radial position for each NP on a scale of 1-5 (1 – strongly apical; 2 – somewhat apical; 3 – neither apical nor equatorial; 4 – somewhat equatorial; and 5 – strongly equatorial).
2. ESTIMATE COMPACTION OF EACH GENOMIC WINDOW
   1. The GAM paper describes how to estimate the degree of compaction of a genomic window (see below for the relevant excerpt from the paper).
   2. Use these insights to assign each window a compaction rating between 1-10 (10 is most condensed; 1 is least condensed).
3. SINGLE CELL ANALYSIS OF THE HISTONE LOCUS BODY USING GAM
4. Extract the Hist1 region from the segmentation table. HIST1 is located on mouse chromosome 13 at the following coordinates: [Start: 21.7 Mb, Stop: 24.1 Mb].
5. Extract relevant NPs for Hist1 (NPs which contain at least one window in the region of interest).
6. Calculate basic statistics for the Hist1 region

* Number of genomic windows
* Number of NPs
* On average, how many windows are present in an NP?
* What is the smallest number of windows present in any NP? The largest?
* On average, what is the number of NPs in which a window is detected? The smallest? The largest?
* What are the most common radial positions of the Hist1 region? (Based on the NPs that captured the region.)
* What are the typical compactions of the windows within the Hist1 region?

1. NP similarities

* compute the Jaccard *index* (aka the Jaccard *similarity* coefficient) for each pair of relevant NPs
* store the computed values in a matrix; the matrix should contain *similarities* for all pairs of relevant NPs

1. NP difference

* compute the Jaccard *distance* (use the method for asymmetric binary attributes, with normalization for GAM data) for each pair of relevant NPs
* store the computed values in a matrix; the matrix should contain *distances* for all pairs of relevant NPs
* store the computed values in a matrix; the matrix should contain *similarities* for all pairs of relevant NPs

1. Normalized Jaccard similarity and difference for GAM data

* Implement the technique for computing similarity of asymmetric binary attributes[[1]](#footnote-1).
* Avoid skewing result due to differences in #windows detected by 2 NPs
* Let the denominator in the Jaccard equation = min{|A|, |B|}
  + |A|: represents the total number of attributes where the attribute of *A* is 1
  + |B|: represents the total number of attributes where the attribute of *B* is 1

1. Visualize the similarity and difference matrices as heatmaps
2. Using k-means clustering, form 3 clusters from the NPs that captured Hist1.
3. Generate (or pick) k=3 random, distinct NPs. These are the centers of the initial clusters. Remove these NPs from the model after step 3.
4. Measure the distance between each point and the centers of the ‘k’ clusters.
5. Assign each point to the nearest cluster.
6. Find the center of each cluster.
   * + - For each cluster, find the NP whose average dissimilarity to all the objects in the cluster is minimal.
       - These are the centers of the new clusters.
       - Note: The NP selected by this method is called the *medoid* (see <https://en.wikipedia.org/wiki/Medoid>).
7. Repeat steps 2-4 until the clusters no longer change.
8. Assess the quality of the clustering by adding up the variation within each cluster.
9. Keep track of the clusters and their total variance, and perform steps 1-6 repeatedly, using different starting points in each iteration.
10. Pick the best set of clusters (the set with the smallest total variance).
11. ~~8-Pick the best value for ‘k’ by finding the “elbow” in the plot.~~

Do not perform step 9. Let k = 3.

1. Visualize EACH cluster as a heatmap
   * + Rows: NPs
     + Columns: genomic windows
     + Cells: values from the segmentation table
2. Study the Hist1 region of the mouse genome (mm9) in the UCSC genome browser
3. Characterize the k clusters according to genomic and epigenomic features
   * + Extend your program to read the feature table and determine if there is a pattern of features in each cluster. Do the features allow you to discriminate between the clusters?
     + Extend your program to characterize each cluster by the radial positions of the NPs in the clusters. Is there a pattern?
4. Infer the chromatin structure of *each cluster*
   * + look for common windows across NPs in a cluster
     + draw a line (manually) that interconnects blocks of prominent windows
5. Infer the chromatin structure and state of *sets of clusters*
   * + compare and contrast prominent blocks of windows in pairs of clusters
     + draw a line (manually) that interconnects complimentary clusters
     + make a genomic interpretation (e.g., a chromatin loop, cell states)

1. <https://en.wikipedia.org/wiki/Jaccard_index#Similarity_of_asymmetric_binary_attributes> [↑](#footnote-ref-1)