

MBI 3100: Introduction to medical bioinformatics

Lecture # 5

Diversity, distance and clusters

AMJAD KHAN

*Department of Pathology and Laboratory Medicine
Schulich School of Medicine & Dentistry
Western University*

October 2023

Lecture #5

Aligned sequences

- ▶ Now that we can align sequences, we can make biologically meaningful comparisons.
 - Which parts of the gene/genome are more variable? more conserved?
 - Which sequences are more closely related than others?
- ▶ It is far easier to measure similarity when the sequences are aligned.

Lecture #5

Measuring sequence diversity

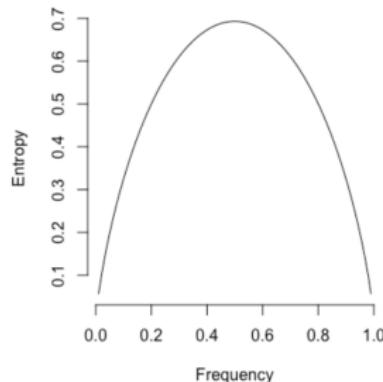
- ▶ Which regions of the genome are the most conserved (least diverse)?
- ▶ Variable regions can reveal targets of diversifying selection, e.g., major histocompatibility complex (MHC).
 - MHC is a large locus on vertebrate DNA containing a set of closely linked polymorphic genes that code for cell surface proteins essential for the adaptive immune system.
- ▶ Conserved regions can make good targets for sequencing primers, antibodies.

- ▶ There are several ways to measure sequence diversity.
 - Fraction of polymorphic sites - what counts as a polymorphism?
 - Minor allele frequency (MAF): the frequency of the second-most common residue
 - Sequence entropy
- ▶ Convention is to label a site as polymorphic if MAF is greater than 1% and less than 5%.

- ▶ The concept of entropy comes from information theory.
- ▶ For each site, we calculate:

$$S = - \sum_i p_i \log p_i$$

where p_i is the frequency of the i -th residue at that site.



- Entropy is highest when residues appear at equal frequency.

- ▶ Calculate the entropy for:

Seq 1: A G G C

Seq 2: A G C G

Seq 3: A C G G

Seq 4: C A G C

for each column

$$S = - \{ p_A \times \log p_A + p_C \times \log p_C + p_G \log p_G + P_T \log p_T \}$$

- ▶ Frequency of polymorphic sites.
- ▶ Mean nucleotide or amino acid entropy - calculate entropy at each site, and then take the average:

$$\bar{S} = \sum_{j=1}^L \frac{S_j}{L}$$

- ▶ Nucleotide diversity (π): the average number of differences between two randomly sampled sequences

$$\pi = \frac{2 \sum_{i=1}^{n-1} \sum_{j=i+1}^n \pi_{ij}}{n(n-1)}$$

Here π_{ij} is the number of nucleotide differences between sequences i and j .

Lecture #5

Sliding windows

- ▶ Site-wise diversity measures can be too noisy to be useful.
- ▶ Averaging diversity by gene requires knowledge of gene coordinates, may be too coarse.
- ▶ A "sliding window" takes the average of a statistic for a given window size and step size.

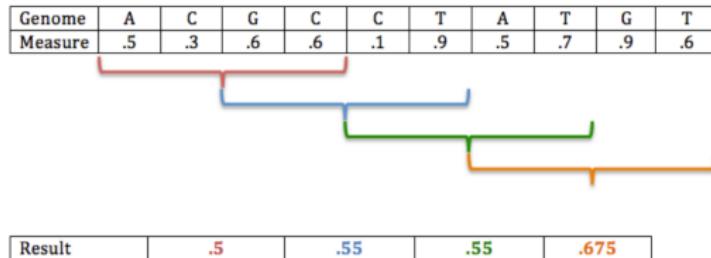


Image source: <http://coleoguy.blogspot.com/2014/04/sliding-window-analysis.html>.

Lecture #5

Genetic distances

- ▶ Another approach to quantify diversity is to use a distance measure (comparing pairs of sequences).
- ▶ A genetic distance is a function $d(x, y)$ that maps sequences x and y to some non-negative value.
- ▶ A distance function $d(x, y)$ should have the following properties:
 - $d(x, y) \geq 0$ for all $x, y \in \Omega$
 - $d(x, y) = 0$ if $x = y$.
 - $d(x, y) = d(y, x)$ (symmetry)

Lecture #5

Genetic distances

- ▶ The simplest distance is to count the number of different residues, i.e., the Hamming distance (HD):

Seq 1:	G G G T T G C G C T C G T T G
Seq 2:	G G G A T G C A C T C G C T G

- ▶ Hamming distance (HD) is 3.
- ▶ HD increases with sequence length.
- ▶ We can divide the HD by sequence length. This gives us the p-distance (p is for proportional).
 - What is the p-distance for the above example?

Lecture #5

Multiple hits

- ▶ A big problem with the Hamming and p-distances is that they tend to underestimate the amount of evolution.
 - Suppose we are tracking the evolution of a sequence A A A A
 - A single mutation occurs resulting in A G A A ($p = \frac{1}{4}$)
 - As we continue to accumulate mutations, the chance that we mutate a site that has already undergone a mutation increases.
- ▶ Multiple hits mask evidence of previous evolution
 $(A \rightarrow G \rightarrow A)$.

Lecture #5

Modeling multiple hits

- ▶ Let's make a few lousy assumptions:
 - Each residue in a sequence evolves independently of the others.
 - A residue mutates to another at a rate that is constant over time.
 - A residue is equally likely to mutate to any of the other residues.
 - The frequency of every residue is the same.
- ▶ These define the Jukes-Cantor model.

Lecture #5

Markov property

- ▶ The Jukes-Cantor model describes a Markov process.
- ▶ A process has the Markov property if the probability of state at time t depends only on the state at a previous time and no further i.e., the system has no memory.
- ▶ For example, Snakes and Ladders is a Markov process.

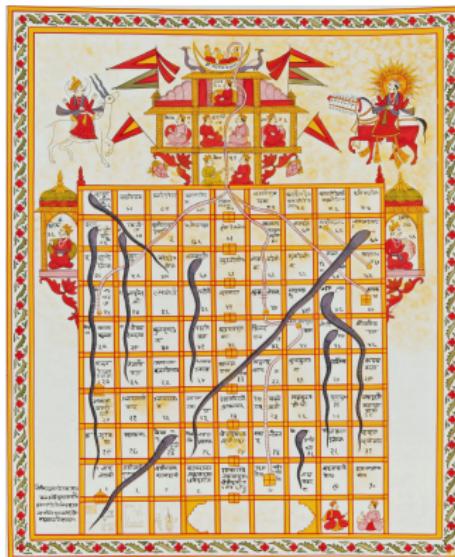
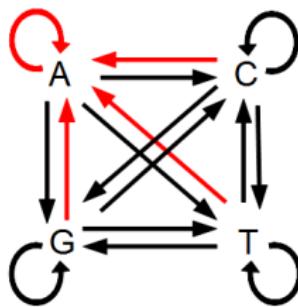


Image source: Jain version Game of Snakes & Ladders called jnana bazi or Gyan bazi, India, 19th century, Gouache on cloth.

Lecture #5

Markov models

- ▶ Jukes-Cantor is an example of a continuous time Markov model.
- ▶ A system is in one of two or more discrete states. After some random amount of time, it switches between states.



Lecture #5

Jukes-Cantor formula

- ▶ Because of multiple hits, the actual number of mutations tends to be greater than the number of visible differences.
- ▶ Given a p-distance (p) between two sequences, the JC estimated number of mutations (d) is:

$$d = -\frac{3}{4} \ln\left(1 - \frac{4}{3}p\right)$$

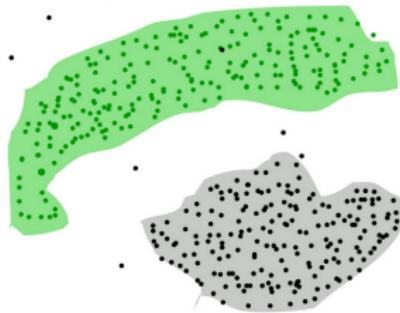
- ▶ This is the mean (expected) estimate — evolution is stochastic, so there will be variation around the mean!
- At what p-distance does the Jukes-Cantor formula fail?

- ▶ The Jukes-Cantor model enables us to estimate the divergence time of two populations (species or infections) more accurately.
- ▶ Two distantly related species might otherwise look about the same as more closely related species.
 - The expected p-distance asymptotes to a maximum value.
 - A small change in p-distance can imply an enormous change in evolutionary time.

Lecture #5

Clustering

- ▶ A cluster is a subset (group) of objects that are more similar to each other than objects outside the cluster.
 - Similarity is just the opposite of distance!
- ▶ Clustering is subjective. Our brains are wired to see patterns where none exist.



Lecture #5

Clustering methods

- ▶ Clustering is useful:
 - for finding real patterns, e.g., biological pathways
 - to reduce a large database to a representative subset
 - to define species, other taxonomic groupings
 - to detect anomalies (outbreaks)
- ▶ There are an enormous number of methods (algorithms) for clustering data.
 - It is easiest to talk about different categories of clustering methods.

- ▶ Terms associated with machine learning.
- ▶ Supervised clustering means that you have assigned some data to clusters yourself, and leave the rest to the machine.
- ▶ Unsupervised clustering means that the machine has to figure it all out itself.

Lecture #5

Unsupervised



- ▶ A non-parametric clustering method uses the observed distribution of one or more characteristics to cluster the data.
 - For example, if we look at cars on a one-lane road, we can build up clusters from any two cars closer than some cut-off distance of each other.
- ▶ A parametric clustering method fits a model to the data to define clusters.
 - If we have a model on the distance between cars, we can identify groups of cars that are consistent with a “close following” mode.

Lecture #5

k-means clustering

- ▶ An unsupervised nonparametric method.
- ▶ k refers to the number of clusters defined by "means".
- ▶ Assign each point to the closest mean, while locating the optimum locations of means.

Lecture #5

Gaussian mixture models

- ▶ An unsupervised parametric method
- ▶ Find the assignments of each data point to one of k Gaussian distributions.
- ▶ Also find the mean and variance of each Gaussian that maximizes likelihood.
- ▶ Method can determine for itself the optimal number of clusters.

Lecture #5

Hierarchical clustering

- ▶ Another class of unsupervised, nonparametric clustering methods.
 - Acts on a distance matrix d relating observations.
- ▶ Hierarchical clustering can be agglomerative or dissociative.
- ▶ An agglomerative method starts with every item in its own cluster, and progressively merges clusters that are the most similar.
 - ▶ Choosing which clusters to merge is determined by linkage criteria.

Lecture #5

Linkage criteria

- ▶ Merging clusters updates the distance matrix (remove two rows and columns, add a new row and column).
- ▶ Complete linkage clustering

$$d(AB, x) = \max(d(A, x), d(B, x))$$

- ▶ Single linkage clustering,

$$d(AB, x) = \min(d(A, x), d(B, x))$$

- ▶ Unweighted pair group method with arithmetic mean, also known as average linkage clustering.
- ▶ Every sequence starts out as a cluster of one ($n_X = 1$).
 - Join clusters X , Y , with shortest average distance:

$$d(X, Y) = \sum_{x \in X} \sum_{y \in Y} \frac{d(x, y)}{n_X n_Y}$$

- Replace X and Y with cluster XUY , where:

$$d(XUY, Z) = \frac{n_X d(X, Z) + n_Y d(Y, Z)}{n_X + n_Y}.$$

- Go to step 1 until only one cluster remains (the root).

Lecture #5

Dendrograms

- ▶ The end result of hierarchical clustering is a tree or “dendrogram”.
- ▶ Lengths of branches connecting x and y to their “ancestor” are calculated by splitting the distance $d(x, y)$ in half.

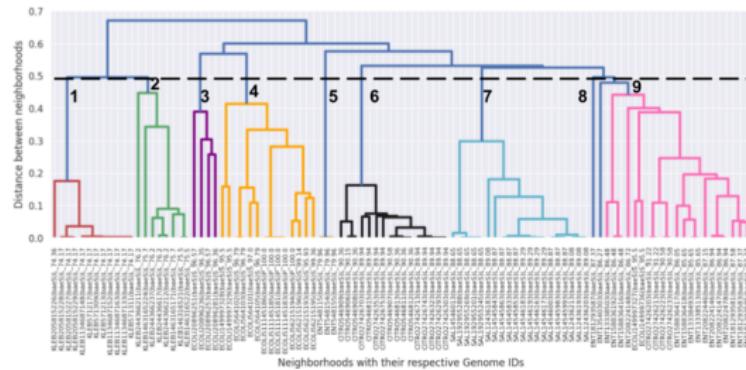


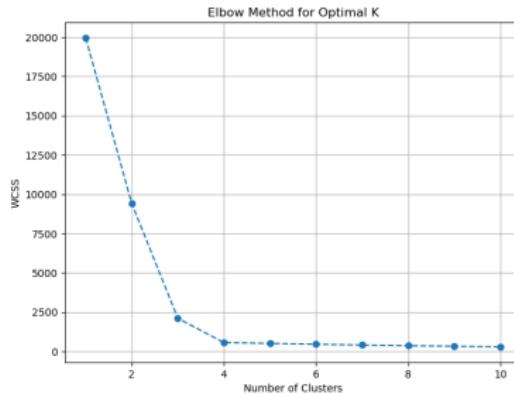
Image source: NEIGHBORHOOD CLUSTERING TO ANALYSE ANTIMICROBIAL RESISTANCE IN BACTERIAL GENOMES by Rudrappa.

- ▶ The output of hierarchical clustering is a tree, not clusters.
 - It remains to “cut” the tree at some point to extract clusters.
 - e.g., cutting near the root tends to yield two large clusters.
- ▶ Location of the cut point is a subjective decision.
 - User-specified number of clusters
- ▶ Some automated methods for selecting number of clusters
 - “knee” / “elbow” method, plot merge distance with number of clusters.

Lecture #5

Knee/ Elbow Method

- ▶ The elbow method, also known as the knee method, is used for determining the optimal number of clusters in a dataset.
- ▶ It plots within-cluster sum of squares (WCSS) against cluster numbers and identifies the elbow point, indicating the optimal cluster count.



Lecture #5 Heatmaps

- ▶ A popular method for visualizing a matrix of intensities, e.g., gene expression.
 - ▶ Hierarchical clustering can be used to reorder rows/columns to bring together similar observations/variables.
 - ▶ Dendograms can be displayed along respective axes.

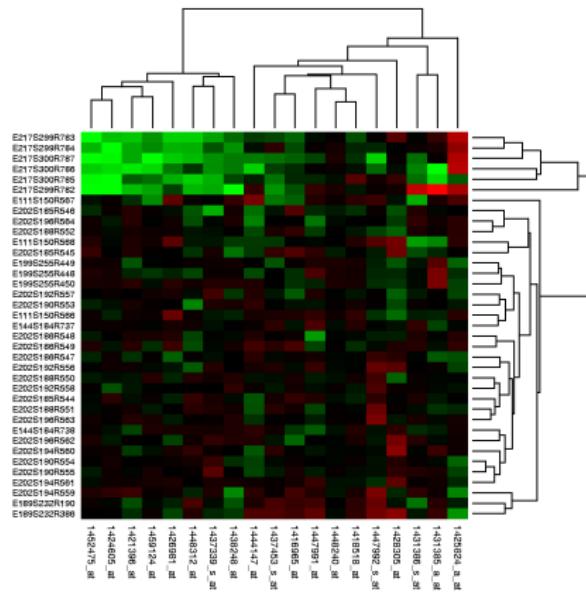


Image source: <https://commons.wikimedia.org/wiki/File:Heatmap.png>.

Lecture #5

From trees to networks

- ▶ Remember a genetic distance is used to quantify the difference between two sequences.
 - e.g., Jukes-Cantor (JC69)
 - Pick a distance threshold — any pair of sequences with a distance below that threshold are assigned to the same cluster.
- ▶ Clusters are often visualized as networks (graphs) where each node represents a sequence.
 - Similar sequences are connected by edges.

- ▶ The International Committee on the Taxonomy of Viruses allows the definition of a new virus species based on genetic clustering, although this remains controversial.
 - Unfortunately, in recent years, ICTV Study Groups [...] have created large number of species on the basis of a single criterion, namely a certain percentage of genome similarity between individual viruses.

Lecture #5

Epidemic structure from clustering

- ▶ Tuberculosis is one of top 10 causes of death worldwide.
- ▶ Caused by lung infection by *Mycobacterium tuberculosis*.
- ▶ Clustering of whole-genome sequence data can identify high-risk groups and detect undiagnosed cases.

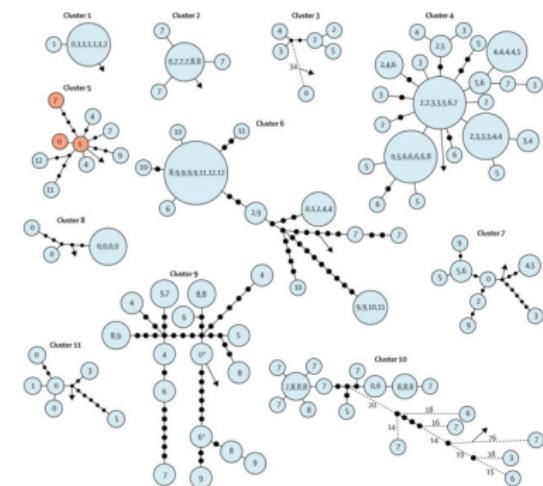


Image source: Cluster diagram by TM Walker et al. (2013) Lancet Inf Dis 13: 137.

Lecture #5

Measuring sequence diversity



Thank You