Lecture Title and Date

Unsupervised Data Mining - Clustering 2/26/25

Objectives of the Lecture

- By the end of this lecture, students should be able to
    1. Understand the conceptual calculations that can be used from matrices
    2. Understand the different clustering methods

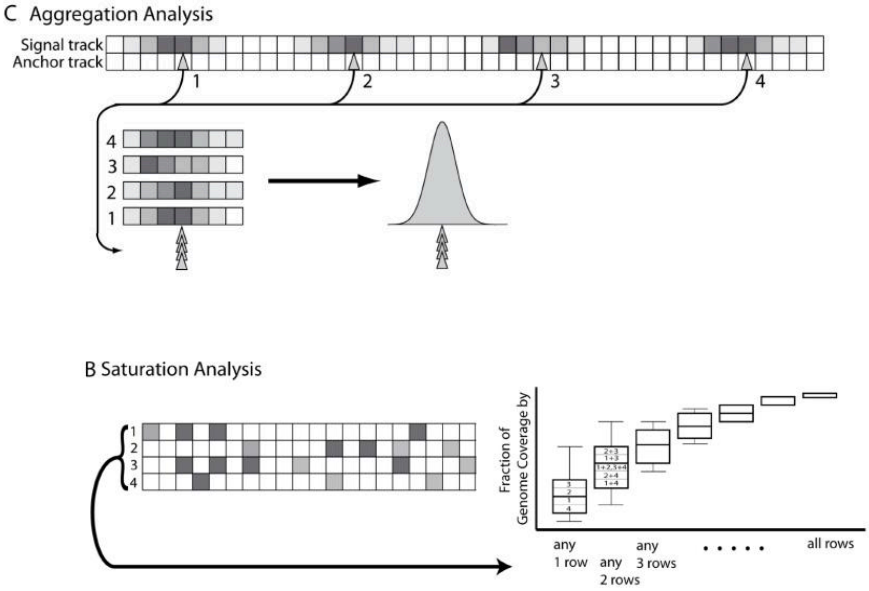
Key Concepts and Definitions

- Unsupervised data- unlabeled data
- Aggregation Analysis- anchor sites in genome, and recenter the matric around the site and compute distribution around area to see level of gene expression, chromatin levels etc.
- Saturation Analysis- different sites of genomes are compared to different conditions in gene expression
- Expression Clustering- marks activity of the area in genome from the matrix by correlating each row to each other or columns to each other
- Signal profile- takes a gene from matrix where signals (expression levels) are graphed to show relationship between different genes
- Clustering Methods
    - Centroid- based method- clusters based on randomly chosen centers and find data points that fit with those centers
        - K-means clustering- finding the data points that are closest to the centers, and recompute the centers until center stops
            1. Pick k random points as putative cluster centers
            2. Group the points to be clustered by the center that are the closest
            3. Take the mean of each group
            4. Repeat until center stays the same
    - Distribution-based method- clusters data by Gaussians (based on distribution/std/ etc)
        - Ex. LDA and tSNE
    - Connectivity-based method- clusters based on hierarchy, finding two closest points, merging them together etc.
        - Ex. Hierarchical clustering, Agglomerative clustering
        - Agglomerative clustering- clusters data points based on whether there is a threshold for connection
    - Density-based method- looking at point density and look for sparse regions
        - Ex. DBSCAN, MSB

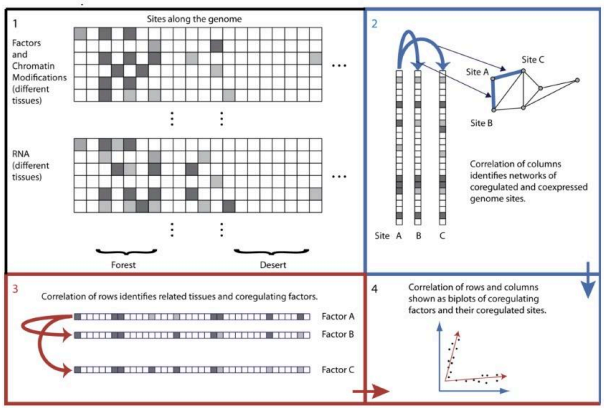
Main Content/Topics

Unsupervised data refers to data without labels, requiring clustering and pattern recognition techniques to extract meaningful insights. In genomics, data is often structured as matrices, where populated regions are termed "**forests**" and unpopulated areas as "**deserts**." Various computational analyses help interpret genomic matrices.

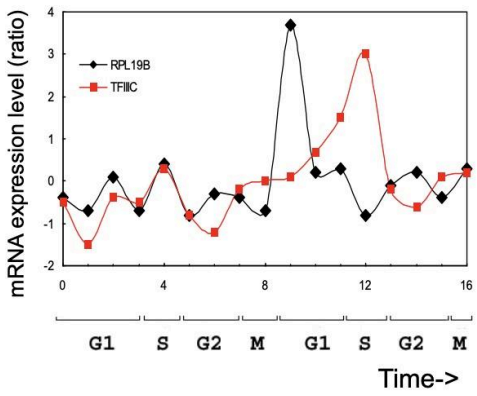
Aggregation analysis focuses on anchor sites in the genome, recentering the matrix around these sites to study gene expression or chromatin levels in surrounding regions. **Saturation analysis** compares different genomic sites under varying conditions to determine whether complete gene coverage has been achieved. A challenge in this analysis is that the order of genomic sites is arbitrary.



Expression clustering identifies activity patterns in genomic regions by correlating rows or columns within a matrix, forming an affinity matrix. This is useful for studying gene expression changes over time, such as RNA expression at different stages.

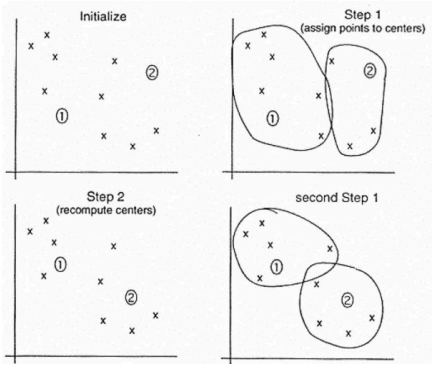


Similarly, a **signal profile** extracts specific genes from a matrix and graphs their expression levels, helping to visualize gene interactions and construct regulatory networks. The following image shows an example of how signal profile can be used to show the relationship between different genes. This can help make a network to see how different genes are linked to each other.



Unsupervised learning methods, particularly clustering techniques, are widely used in genomic analysis to identify patterns within large datasets. These methods help group similar genomic regions based on various characteristics.

- **Centroid-based clustering** assigns data points to randomly chosen cluster centers and iteratively adjusts them until convergence. A common example is **k-means clustering**, where k initial cluster centers are selected randomly. Data points are assigned to the nearest center, and new centers are computed based on the mean of assigned points. This process repeats until the centers stabilize, making it effective for partitioning data into well-defined clusters.



- **Distribution-based clustering** groups data according to statistical distributions. Methods such as **Latent Dirichlet Allocation (LDA)** and **t-Distributed Stochastic Neighbor Embedding (t-SNE)** assume data points follow specific probability distributions, typically Gaussian. These methods are useful for reducing dimensionality

and visualizing high-dimensional genomic data while preserving meaningful relationships between points.

- **Connectivity-based clustering** builds hierarchical structures by iteratively merging or splitting data points based on distance metrics. **Hierarchical clustering** and **agglomerative clustering** start by treating each point as its own cluster and successively merging the closest clusters until a stopping criterion is met. This approach is particularly useful for capturing nested relationships in genomic features.
- **Density-based clustering** identifies clusters by searching for dense regions of data points while filtering out sparse areas. **DBSCAN (Density-Based Spatial Clustering of Applications with Noise)** groups data based on the number of neighboring points within a defined radius, making it effective for detecting irregularly shaped clusters. Other methods, such as **Mean Shift Clustering (MSB)**, iteratively shift cluster centers towards regions of higher density.

While unsupervised methods are effective, they can be combined with supervised learning to form **semi-supervised methods**, leveraging both labeled and unlabeled data for improved accuracy. These hybrid approaches offer flexibility in analyzing complex genomic datasets where full annotation is unavailable.

Discussion/Comments

Overall, this lecture covers an overarching view of the different clustering methods available in unsupervised learning. Each clustering method has its strengths and effectiveness and should be utilized depending on the tasks required from a given data.

References ESL/ISL

- ISL refers to An Introduction to Statistical Learning (https://www.statlearning.com/)
    - Chapter 12: Unsupervised Learning, specifically 12.1 and 12.4
- ESL refers to The Elements of Statistical Learning (https://hastie.su.domains/ElemStatLearn/)
    - Chapter 14: Unsupervised Learning, specifically 14.3 (Cluster Analysis)

Additional suggested Reference

- 2.3 Clustering - scikit learn
    - This link provides a more in depth view on the types of clustering and how it can be implemented, Although this goes into a lot more depth in terms of potential code, it provides a fundamental understanding of the range of available clustering techniques.
    - https://scikit-learn.org/stable/modules/clustering.html