3.1.3 Conformational clustering

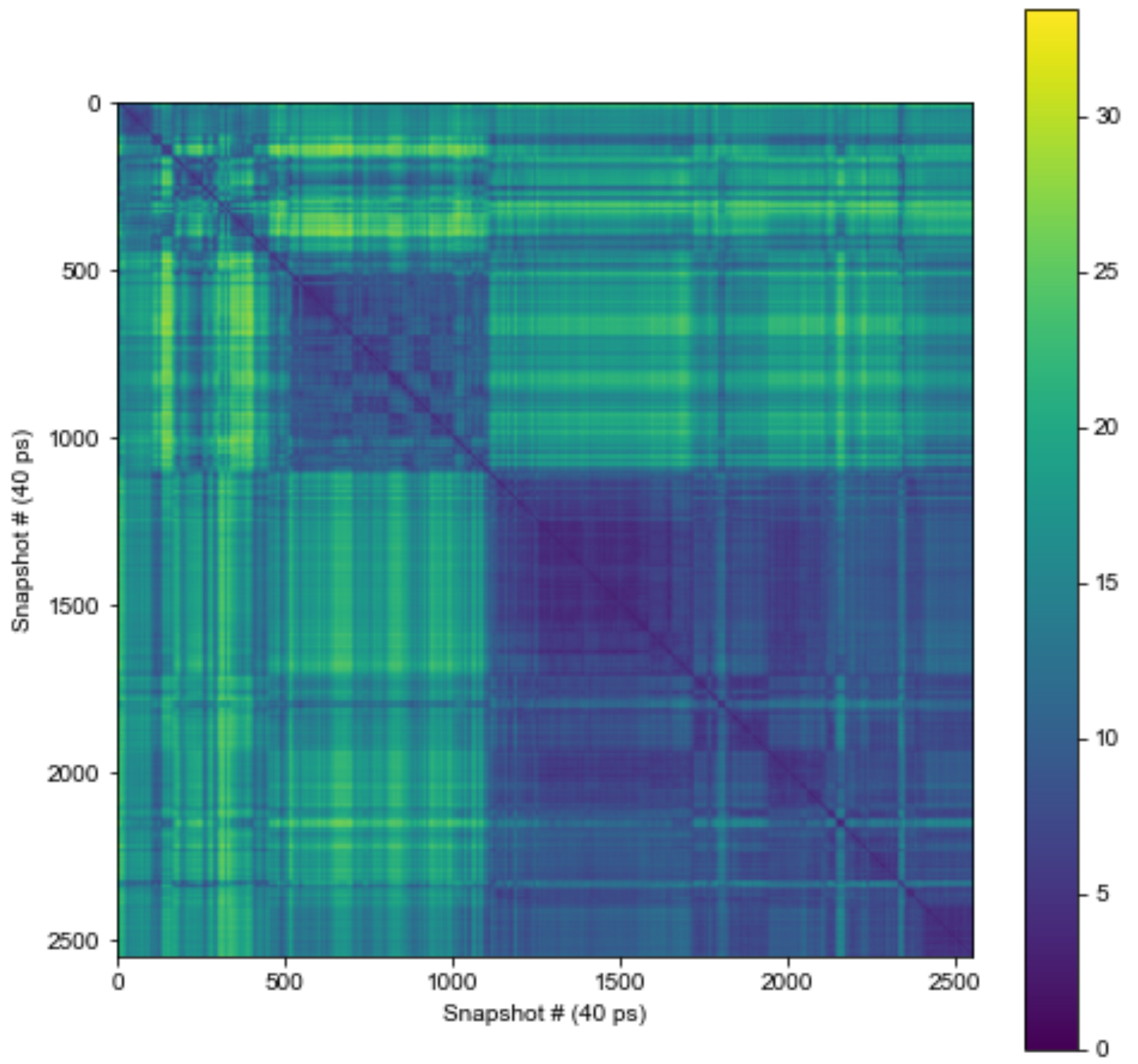
- This module will consist of a mini-lecture and exercise on conformational clustering
- At the end of this module, you should be able to answer the following questions:
  - What is clustering and why is it useful?
  - What distance matrices are there? How should they be selected?
  - How does agglomerative hierarchical clustering work? What is a linkage criterion?

Clustering

- MD simulations yield configurations in continuous space
- Clustering methods group together similar configurations (or, in a more general data science context, observations)
- Clustering is useful
    - interpreting simulation results
    - calculating thermodynamic and kinetic properties
        - predicted populations of conformations
        - predicted rates of transitions (e.g. Markov state models [1, 2])
    - selecting representative configurations for molecular docking [3]

Distance matrices in clustering

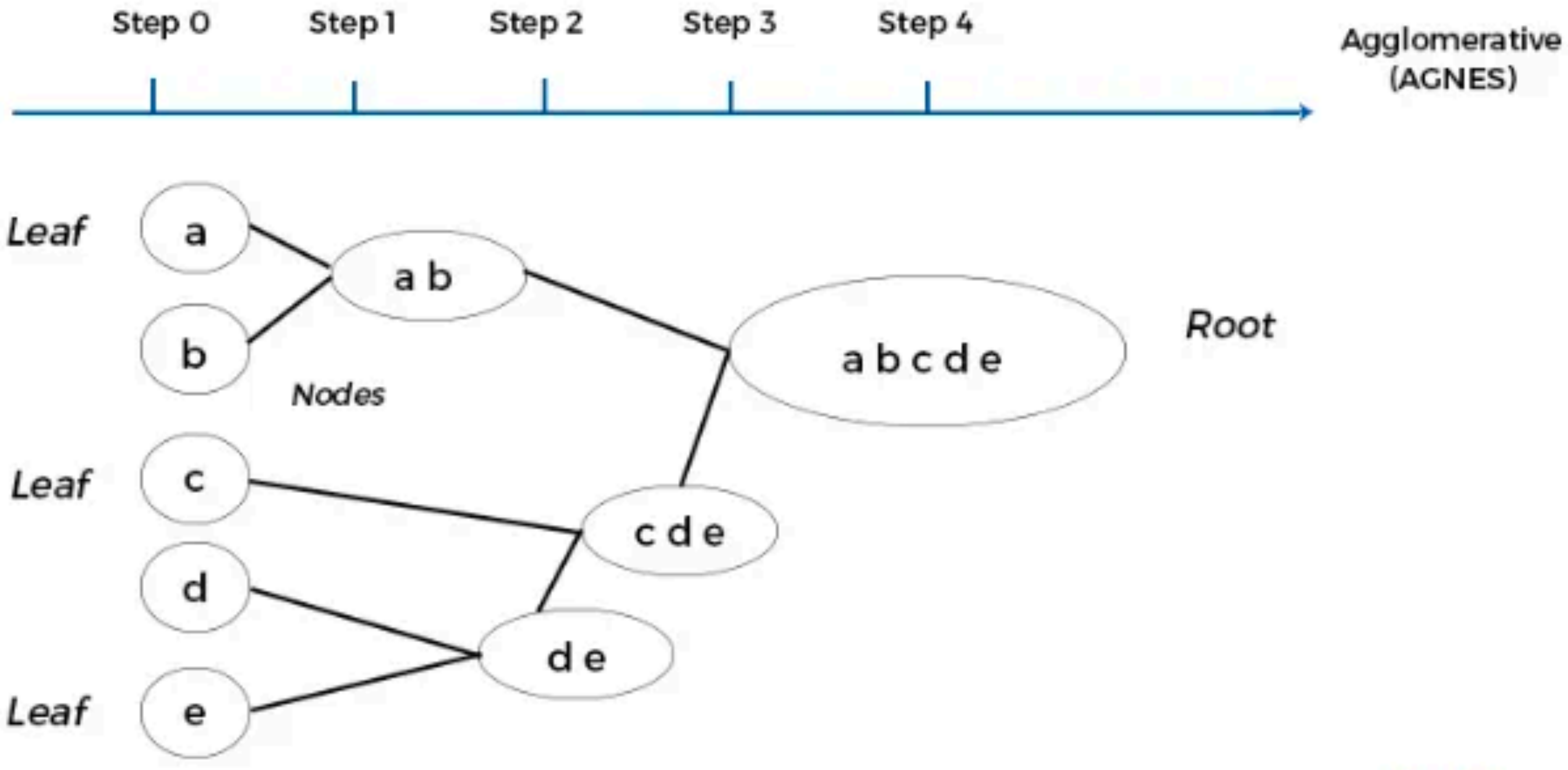
- Almost all clustering algorithms employ a distance matrix
- In a matrix **D**, D_{kl} denotes the distance between observation k and l
- Distance matrices include [3]
  - the RMSD
    - between alpha carbons/all heavy atoms
    - in a entire protein/in a region of the protein
  - Euclidean distance between principal components (like the RMSD, PCA can be based on different subsets of coordinates)
  - based on occupancy fingerprints
    - a 3D grid with zero or one depending whether a point is close to an atom
    - If M_{ab} is the number of points where one grid has *a* and the second *b*,
      - the overlap is $M_{10} + M_{01}$
      - the Tanimoto similarity is $-\log_2[M_{11}/(M_{11} + M_{10} + M_{01})]$
      - the Jaccard distance is $[(M_{11} + M_{01})/(M_{11} + M_{10} + M_{01})]$



Heat map of Euclidean distances between
top 20 principal components
in a simulation of ubiquitin

Agglomerative hierarchical clustering

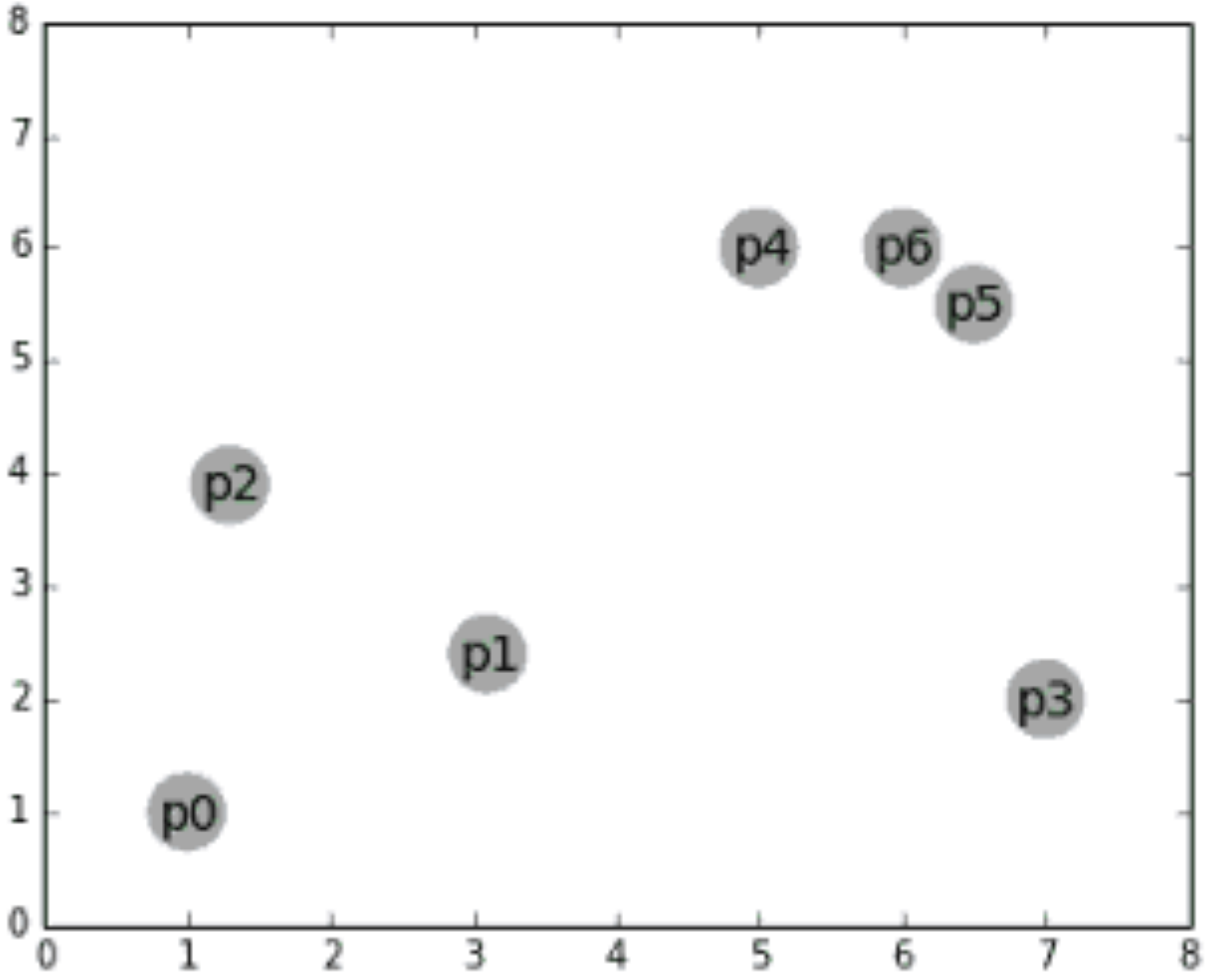
- Closest pair of observations (or clusters) are grouped together until all observations are in groups

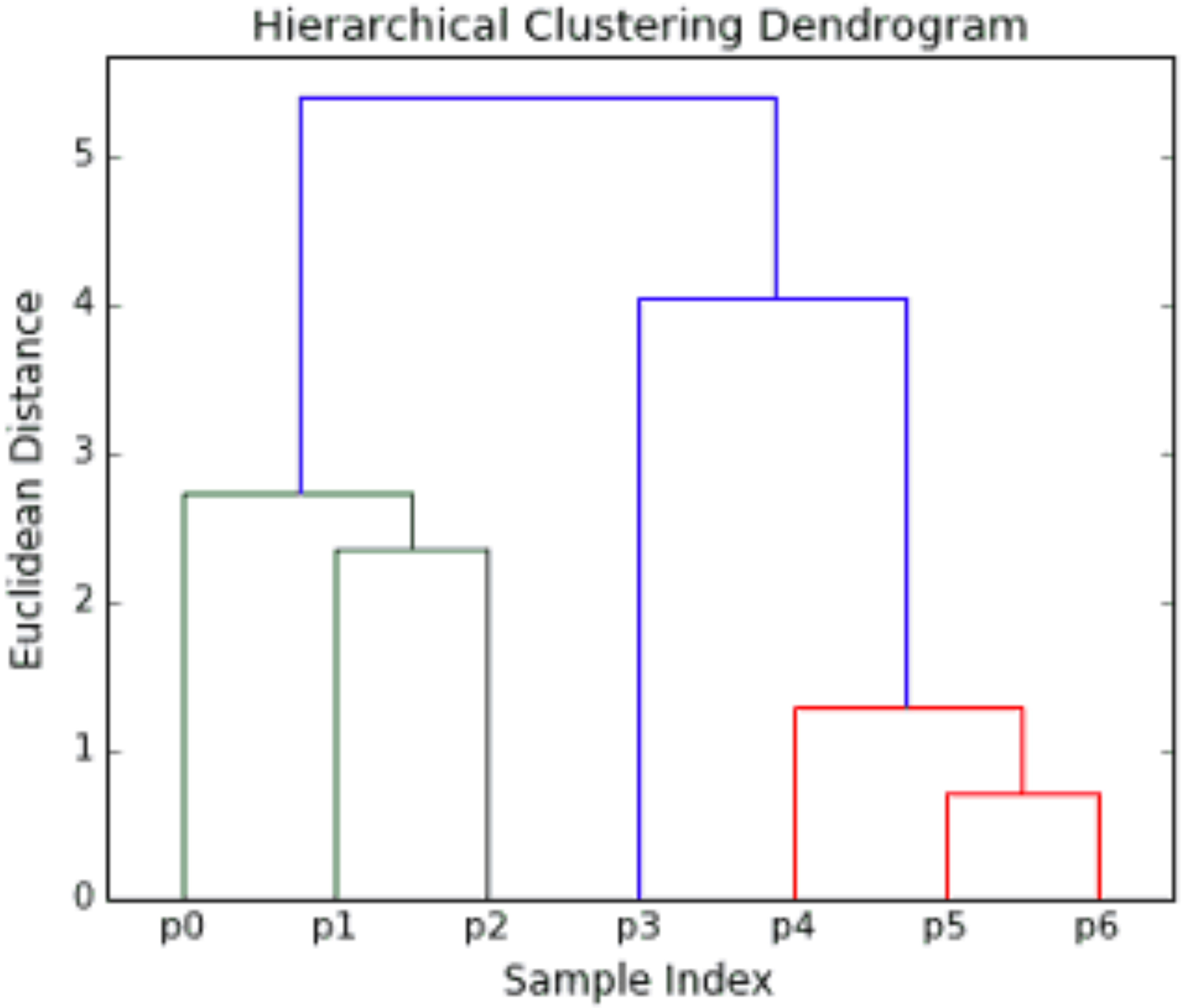


http://primo.ai/index.php?title=Hierarchical_Clustering;_Agglomerative_(HAC)_%26_Divisive_(HDC)

Agglomerative hierarchical clustering

- Closest pair of observations (or clusters) are grouped together until all observations are in groups





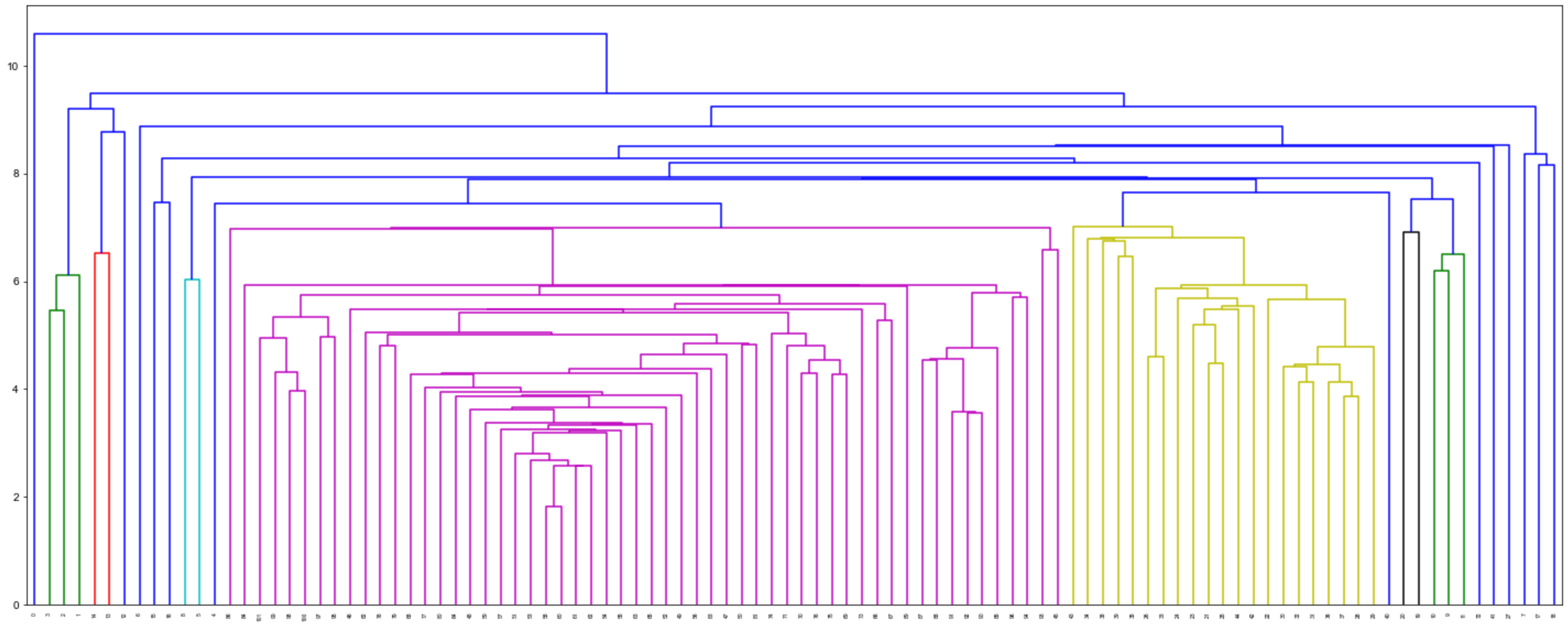
http://primo.ai/index.php?title=Hierarchical_Clustering;_Agglomerative_(HAC)_%26_Divisive_(HDC)

Linkage

- The distance matrix provides distances between observations
- What is the distance between clusters?
- Different linkage algorithms are available in scipy: https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.linkage.html. In different algorithms, the distance between clusters is the
  - single - minimum distance between observations
  - complete - maximum distance between observations
  - centroid - distance between centroids (the mean of the cluster)
- Which linkage algorithm will yield the smallest and largest apparent distance between clusters?

Agglomerative hierarchical clustering

- There are
  - Different definitions of distances between observations and clusters
  - Different ways to go from linkage matrix to clusters
- See Clustering.ipynb, which shows clustering analysis for a simulation of Mpro



Dendrogram of hierarchical clustering for every 1 ns for a simulation of ubiquitin

Review

- What is clustering and why is it useful?
- What distance matrices are there? How should they be selected?
- How does agglomerative hierarchical clustering work? What is a linkage criterion?