1. Give at least two examples of concrete problems where clustering could be useful.

Problem one: scientists have RNA-seq data from a large-scale time-series experiment analyzing circadian cycles in plasma-derived lymphocytes. They would like to find and distinguish which transcripts vary over a period of 24h from those that vary randomly, or with a different frequency, as well as the ones that maintain their baseline expression level. While it is possible to do manually with very simple statistical methods, we would really want to speed up the process by not testing every possible hypothesis manually, and to get a mathematically unbiased representation of those fluctuations.

Problem two: Ash would like to quickly determine which pokemon to choose for his battles. Most pokemon have strengths and weaknesses, and have attacks of different elements, which can have varying effect of the enemy (based on their strengths and weaknesses). Ash would like to be able to sort his collection by those properties, in order to consider his options more quickly and correctly, and finally win a championship after so many years of trying.

2. Name two types of clustering methods and describe the concept behind them?

Hierarchical clustering – (Phylogenetic trees, protein families, day-to-day terminology)
Agglomerative: datapoints will be compared based on similarity of their features, and the most similar will be grouped together. Smaller clusters are then in turn compared to others, and larger hierarchical order is formed, based on the measure of their relatedness, and so on.
Divisive: the bottom-up clustering. Datapoints will be considered all together at first, but the two most distinctive groups will split from each other. Moving on, the data will be split until all groups are distinguished(or only one member left in group).

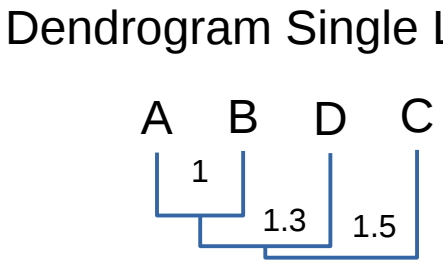
Density-based clustering - (DBSCAN)
This method is used to distinguish continuous groups of datapoints in close proximity to each other. Datapoints will be evaluated to be core point, border point, or noise, unless using a modified approach. Core points are usually determined by being in (predetermined) range from more core points. Borderpoints can then be reached form at least one core point, but not enough to be assigned as core. Borderpoints belong to the cluster, but themselves cannot reach out further. Points that cannot be reached by core points are determined to be noise.

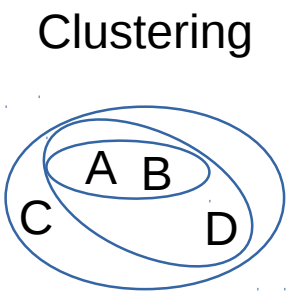
3. Give three examples of distance metrics commonly employed in clustering.

Hamming distance: measure of dissimilarity between sequences at the same position.
Levenshtein distance: similar, but not to be confused with Hamming distance. Levenshtein distance is the amount of changes(insertions, deletions, mutations) required to transform one string into another.
Euclidean distance: distance between two points in euclidean space. The euclidean distance is easy to imagine as a straight line connecting two points on a plane or a rope between two trees in 3D space. Despite the “Euclidean” definition, this distance can be imagined/measured/calculated in multidimensional space for clustering purposes.

4. Assume we have four data points, A, B, C, D, with the distances between them given in the matrix below. Build the dendogram obtained when applying hierarchical single linkage clustering.

4.





6. Open the file clustering_scikit.py. It loads the data from a text file and performs the k-means algorithm. Do you understand what it does? How many clusters does it create by default?

> K-means attempts to sort data into groups, by assigning points to groups where they collectively will have lowest inertia (generalized similarity in all dimenstions).
> By default, initially 3 clusters(centroids) are created.

7. Run the program. Modify it as needed so you get the names of the sequences belonging to each cluster.

YBL024W | J 0
YBR126C | G 0
YBL052C | B 0
YBL087C | J 0
YBL088C | TBLD 0
YBL092W | J 0
YBL097W | BD 0
YBR048W | J 0
YBR081C | BK 0
YBR103W | B 0
YBR199W | G 0
YBR241C | G 0
YCL040W | G 0
YCR036W | G 0
YDL021W | G 0
YDL115C | S 0
YBR009C | B 1
YBR010W | B 1
YBL002W | B 1
YBL003C | B 1
YAR007C | L 2
YNL102W | L 2
YBR088C | L 2
YBR073W | L 2

8. What does k in k-means mean?

It refers to cluster number/amount, or centroids K which to which points are assigned. Earlier articles refer to the method as K-means with large K, and it is unlikely it stems from unit vector k or matrix index k, however I cannot tell what is the origin of K.

9. If you change k=2, how many sequences do you get in the smallest cluster?

4.

10. Change to k=3. How many sequences do you now get in the smallest cluster(s)? You can also try other values for k.

4.

11. For k=3, what is the function of the genes in the two smaller clusters? Look at the functional classification and also use your database search skills. NB: the numbering of the clusters may change between runs.

B = chromatin structure dynamics
L= replication, recombination, repair

BLAST | Align | Download | Add to basket | Columns — 1 to 8 of 8 — Show 25

All 8 results selected.(Clear selection)

Entry	Protein names	Gene ontology (molecular function)	Gene ontology (biological process)
P02309	**Histone H4**	DNA binding; histone binding; protein heterodimerization activity	chromatin assembly or disassembly; histone H3-K79 methylation; nucleosome assembly; sexual sporulation resulting in formation of a cellular spore; transfer RNA gene-mediated silencing
P61830	**Histone H3**	DNA binding; nucleosomal DNA binding; protein heterodimerization activity	chromatin assembly or disassembly; global genome nucleotide-excision repair; nucleosome assembly; rRNA transcription; sexual sporulation resulting in formation of a cellular spore
P02294	**Histone H2B.2**	DNA binding; protein heterodimerization activity	chromatin assembly or disassembly
P04912	**Histone H2A.2**	DNA binding; protein heterodimerization activity	chromatin assembly or disassembly; chromatin silencing; DNA repair
P22336	**Replication factor A protein 1**	double-stranded DNA binding; metal ion binding; sequence-specific DNA binding; single-stranded DNA binding	DNA repair; DNA replication; DNA topological change; DNA unwinding involved in DNA replication; double-strand break repair via homologous recombination; establishment of protein localization; heteroduplex formation; mitotic recombination; nucleotide-excision repair; protein ubiquitination; reciprocal meiotic recombination; sporulation; telomere maintenance via recombination; telomere maintenance via telomerase; telomere maintenance via telomere lengthening
P13382	**DNA polymerase alpha catalytic subu...**	3'-5' exonuclease activity; 4 iron, 4 sulfur cluster binding; DNA binding; DNA-directed DNA polymerase activity; metal ion binding; nucleoside binding; nucleotide binding	DNA replication; DNA replication initiation; DNA synthesis involved in DNA repair; double-strand break repair; double-strand break repair via nonhomologous end joining; lagging strand elongation; leading strand elongation; premeiotic DNA replication; RNA-dependent DNA biosynthetic process
P15873	**Proliferating cell nuclear antigen**	DNA binding; DNA polymerase processivity factor activity; identical protein binding	chromatin silencing at silent mating-type cassette; chromatin silencing at telomere; error-free translesion synthesis; establishment of mitotic sister chromatid cohesion; lagging strand elongation; leading strand elongation; maintenance of DNA trinucleotide repeats; meiotic mismatch repair; mismatch repair; mitotic cell cycle; mitotic sister chromatid cohesion; nucleotide-excision repair; positive regulation of exodeoxyribonuclease activity; positive regulation of phosphodiesterase activity, acting on 3'-phosphoglycolate-terminated DNA strands; postreplication repair; regulation of DNA replication
P38086	**DNA repair and recombination protei...**	ATP binding; DNA-dependent ATPase activity; DNA topoisomerase activity; DNA translocase activity; double-stranded DNA binding; helicase activity	DNA duplex unwinding; DNA geometric change; DNA recombination; DNA repair; heteroduplex formation; meiotic sister chromatid segregation; reciprocal meiotic recombination

12. There is a sequence (YDL115C) of unknown function in the dataset. Can you make a prediction of the functional class of this protein (according to the classification in functioncodes.txt)?

K – transcription or U – intracellular trafficking

13. For comparison, take a look at the result of a neighbour joining clustering (nj_clustering).This tree has been created from the same data set. (However before calculating the tree the correlation between the profiles of each gene pair has been calculated.) What do you think are the advantages/disadvantages of the two clustering methods? When is k-means clustering useful and when is NJ preferable?

NJ in this case does illustrate some useful information, since it groups the genes that are co-expressed together. However the tree is measuring relative relatedmess between the datapoints, and is mainly used for representing evolutionary relationships. This is a great model for representing flow genetic lineages, under assumption that the differences are caused by random mutations and no assumption about selection and evolution. Such model could represent recent migrations of birds/humans or other populations where most changes are due to drift and not active selection.
When it comes to gene expression, I feel like it would fail to distinguish between correlated expression, negative correlation, and no correlation efficiently, since the tree displays a relative relationship between members. A negative/anticorrelation would be “farthest away” relative to the correlated gene, while genes of no correlation would be fitted somewhere inbetween. This type of polarisation is not ideal since the relatedness of independently expressed genes is inflated, and anticorrelated genes do share pattern similarity, yet it is not very evident form plot.

K-means is not run on non-numeric data, such as evolutionary lineages, but would be able to detect and distinguish anticorrelated values from uncorrelated in this case. K-means has poor performance on elongated clusters of datapoints, and k-value is determined arbitrarily, letting the algorithm calculate the optimal clustering strategy – which might be inaccurate if k-value is chosen poorly. One can apply k-mean method to study heterogeneity in tumors based on their gene expression or metabolism

14. Modify the program to use a different clustering algorithm. Do you get significantly different results? Which one would you trust most?

DBSCAN – Groups into 2 clusters, but the result at threshold distance 4 for core is almost the same as with k-means, distinguishing members of B, L processes from others.

Birch – Also consistent with k-mean at lower threshold, but slightly more accurate at identifying other groups (J).

AffinityPropagation – at thresholds closer to 1 will distinguish B, L members, otherwise the output does not coincide with groupings.

SpectralClustering – after manipulating parameters, never had a result consistent with k-means or annotations. However the settings and tuning is more complex than for others, might still be very good at clustering this type of data (judging by description of what it does)

Out of the ones I mentioned, Birch was the most accurate when trying to distinguish more ambiguous groupings than B and L.

Overall, I can see that most of the algorithms have strengths and weaknesses, and one needs to select one that fits distribution pattern and research question the best.