Exercises for the lecture
"Data Mining in Bioinformatics"

Organizer: Peter Meinicke
Submission Date: 14.06.2025, 12 p.m.

Sheet 3: Dimensionality reduction of Codon Usage Vectors

Task 1: RSCU data and PCA

Load the file `EcoliRSCU.mat.csv` from StudIP. The data matrix contains for each gene of *Escherichia coli* (E.coli) a row vector with so-called RSCU values. The abbreviation "RSCU" stands for "relative synonymous codon usage". Find out what the RSCU values mean and how they are calculated! are calculated! Why are there only 59 values (dimensions) per vector? The vector in `EcoliRSCU.label.csv` characterizes the genes according to the known annotation and the prediction with a Hidden Markov Model (SIGI-HMM[1]), which detects *genome islands* using the position of a gene on the the genome. In this process, for each gene there is an integer label (same order as the RSCU vectors), which has the following meaning:

0: "normal" gene

1: putative "alien" gene

2: highly expressed gene (ribosomal protein)

Check the literature/internet: what exactly are "alien" (foreign) genes and genomic islands? Why do highly expressed genes often show a codon usage that is different from the rest of the genome?

b) Perform a principal component analysis (PCA) of the data vectors. First analyze the eigenvalue spectrum of the covariance matrix (bar chart). What is the smallest possible number of eigenvalues that represent at least 70 or 90 percent of the total variance of the data? Then make a scatterplot of the first two principal components, representing the points in different colors according to the labels of the corresponding genes. Include a legend into the plot.
Repeat the whole analysis for the data of *Bacillus subtilis* (`BsubRSCU.mat.csv` and `BsubRSCU.label.csv`)!

Task 2: Multidimensional Scaling

a) Download the files `distance_mat1.npy` and `distance_mat2.npy` from the folder on StudIP. Note that they are stored in binary format and can be loaded using the `np.load()` function in numpy. These files contain distance matrices derived from two separate studies. Do you guess why we did not store them in

[1] `http://www.biomedcentral.com/1471-2105/7/142`

a human-readable CSV format? Determine the number of observations (genes) represented in each dataset.

b) Implement the metric multidimensional scaling (MDS) technique to compute a two-dimensional representation of the data (scatterplots), similar to the principal component analysis visualization in Task 1.

c) Generate 2D scatterplots for both distance matrices using the MDS approach. Compare the resulting scatterplots with the PCA-based plots from Task 1!

d) Use the "suitable" labels from Task 1 to color the data points in the scatterplots generated by the MDS approach.

Submission instructions

Please submit your solution by **14th June, 12 p.m.** using the corresponding homework folder in StudIP. Upload a **zip** compressed file containing all documents and data that you use in your solution. The file name should indicate the author, for example *Peter_Meinicke_Sheet3.zip*.
With your solution you should provide the Python source code together with the documentation within a **Jupyter Notebook** answering all the questions above. Use Markdown cells for your answers and explanations well separated from the Python code.
For the programming in Python/numpy do not use any code or functions from special bioinformatics libraries or toolboxes! Make sure that your code is readable and understandable for others, i.e. use descriptive names for the variable definitions according to the conventions on **Sheet 0**, do not use implicit variables (hard-coded values without a variable) and make use of comments!
For reasons explained in the lecture, we do not recommend the use of **A.I. tools** (such as ChatGPT) but, in principle, it is not prohibited. However, if parts of the solutions or code were generated with A.I. tools, it is **mandatory to specify** to what extent (percentage) and for what purpose it was used. Not indicating the actual use of A.I. tools can lead to disqualification from the exercises. If you are unsure about the specification(s) please ask us and do not make any assumptions!