Processing of Biological Data

Prof. Dr. Volkhard Helms
Tutor: Markus Hollander

Saarland University
Chair for Computational Biology

Winter Semester 2021

Exercise Sheet 2

Due: 23.11.2021 10:15am

Submission

- Submit your solution by email to markus.hollander@bioinformatik.uni-saarland.de with the following two attachments:
  (a) A single PDF containing your answers, plots **and properly formatted source code**.
  (b) A ZIP archive containing all your source code files as well as potential other files.
- Do not forget to mention your names **and** matriculation numbers **in the PDF**.
- **Late submissions will not be considered.**

General Remarks

- Answer in complete sentences.
- In your implementation, use meaningful variable names and **add comments that describe what you are trying to do with different sections of your code**.
- You are free to use any programming language to solve the tasks. You may use plotting packages and libraries to create plots. However, the use of functions, packages or libraries that allow you to circumvent implementing the algorithms yourself will not grant you points. For example, when asked to implement clustering, do not use an already existing clustering function, package or library that solves the task for you.
- If you have questions, you can contact me via the email address mentioned above or in MS Teams.

Exercise 2.1: DNA Methylation in Haematopoiesis and Clustering (50 points)

In the first part of the assignment, you will implement and apply a classical clustering algorithm to preprocessed DNA methylation data by Bock et al. (2012).

In their study, bisulfite sequencing was applied to different mouse cell types in blood and skin cell differentiation lineages. The resulting DNA methylation data was mapped to genomic regions with a size of 1,000 basepairs (1kb) each. Regions with insufficient sequencing coverage were discarded. For each region, the average fraction of methylated cytosines was determined in each cell type.

(a) Before you start with the data analysis, familiarise yourself with DNA methylation and briefly describe which parts of the genome tend to be methylated and the general effect on gene expression. (2 points)

(b) Generally, DNA methylation increases with specification during cell differentiation. Briefly explain why this is the case. (2 points)

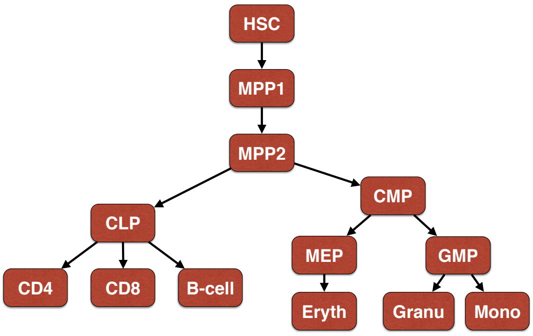
(c) Write a parser for the data in **methylation.csv**. Apart from the header row, each row corresponds to a genomic region. The first seven columns present meta information about the region, while all subsequent columns contain the methylation level of different cell types.

In practice, such files are never perfect. You may, for example, need to slightly rewrite the methylation values to enable a floating-point number conversion in your programming language of choice. Also, you may encounter missing data points. In the latter case, you should set missing methylation values in the data to 0.

Store the data in a way that makes sense to you for further tasks. You may use table reading functions or libraries to parse the file. (5 points)

(d) Determine and report the overall average methylation level for each cell type. (4 points)

(e) Have a look at the differentiation of different blood cells from haematopoietic stem cells to mature cells below. Does the overall methylation-level increase from haematopoietic stem cell to the mature cells? Are there exceptions? (2 points)



(f) Briefly explain why the overall methylation level could decrease during cell differentiation. (2 points)

(g) Hierarchical clustering consists of a **distance function** that describes how (dis–)similar two cell types are, and a **linkage criterion** that uses the distance function to determine the distance between *clusters* of cell types.

Proceed as follows:

(i) Implement a function $d(a,b)$ that computes the Euclidean distance between the methylation patterns of two cell types a and b:

$$d(a,b) = \sqrt{\sum_{\text{region } r} (a_r - b_r)^2}$$

Report the the pairwise distances between HSC (hematopoetic stem cells), CD8 (T cells), and EDif (differentiated epidermis cell). (6 points)

(ii) Implement a function $L(A,B)$ that computes the average linkage criterion between two clusters of cell types A and B as:

$$L(A,B) = \frac{1}{|A||B|} \sum_{a \in A} \sum_{b \in B} d(a,b)$$

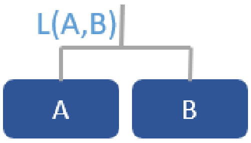
Report the pairwise linkages between the clusters (HSC, MPP1, MPP2), (CLP, CD4, CD8, B–cell) and (MEP, Eryth). (6 points)

(iii) Finally, implement the agglomerative hierarchical clustering. At the beginning, each cell type forms its own cluster. Iterate over the current list of clusters:

1. Compute the linkage criterion for all unique combinations of clusters.
2. Determine the cluster pair A, B with the smallest linkage.
3. Combine A and B into a new cluster and remove A and B from the list.

Stop when only one cluster is left. (10 points)

(h) Apply your clustering implementation to the data. Draw a dendrogram from the results by hand. Annotate each connection with the corresponding linkage value. (9 points)



(i) Can you separate the two lineages (blood and skin cells) based on their methylation patterns? Can you see the developmental succession of the blood cells? (2 points)

Exercise 2.2: Peak Detection in Cell–Cycle Gene Expression Data (50 points)

In the second part of the assignment, you will implement a simple but powerful peak detection algorithm and apply it to yeast cell–cycle gene expression data by Spellman et al. (1998).

Simply put, chemical treatments were used to synchronise the cell cycle of yeast cell cultures. The gene expression was then determined at various time points with DNA microarrays. The data is supplied in **yeast_cell_cycle.txt**.

The file is tab–separated and contains a header row that you can ignore. All other rows contain the systemic locus name of a gene in the first column, while all subsequent columns contain the expression values in chronological order.

(a) We are only interested in the genes **SWI4**, **MCM1** and **ACE2**. Determine and report their systemic locus name that is used in the data. Also state which source you used. (2 points)

(b) Parse the data file and set missing expression values to 0. Store the data in a way that makes sense to you for further tasks. You may use table reading functions or libraries to parse the file. (4 points)

(c) Plot the expression curves of **SWI4**, **MCM1** and **ACE2** in one plot, with the expression on the y–axis and time on the x–axis. Do not forget to label your axes. (4 points)

(d) Implement the watershed algorithm for peak detection. It takes the expression timeline of a single gene and an adjacency threshold parameter n. Consider the data points in descending order of their **absolute** value. In each iteration:

- If the current point has at least one neighbour with a label, it receives the label of the neighbour with the highest **absolute** value. Neighbours are the n data points right before and the n data points right after the current data point (in the input data, not in the sorted data).
- If the current point has no labelled neighbours, it is a peak and receives a new label.

The algorithm stops when all data points are labelled and returns the peaks. (20 points)

(e) Apply the peak detection method that you implemented to the expression data of **SWI4**. Use all adjacency thresholds from $n = 1$ (only considering directly adjecent points) to $n = 4$. For each adjacency threshold, create a plot with the expression curve of **SWI4** and mark the identified peaks with circles. (6 points)

(f) Discuss how the adjacency threshold affects peak detection. Which threshold seems to be best suited for this data? (4 points)

(g) Plot the expression curves of **SWI4**, **MCM1** and **ACE2** together with their detected peaks into one plot. Use an adjacency threshold that seems reasonable to you. (2 points)

(h) Look at the previous plot and the identified peaks. Briefly compare and discuss the expression patterns of the three genes. Are they expressed at the same time? Are they expressed in a cell–cycle dependent manner? Does it seem like their expression depends on each other? Given that they are transcription factors, what can you deduce about their general role in the cell–cycle context? (8 points)