Processing of Biological Data

Prof. Dr. Volkhard Helms
Tutor: Markus Hollander

Saarland University
Chair for Computational Biology

Summer Semester 2020

Exercise Sheet 2

Due: 04.06.2020 10:15

Submission

- You are advised to work in groups of two people.
- Submit your solution by email to markus.hollander@bioinformatik.uni-saarland.de **as a single PDF attachment with your answers and all source code.** Late submissions will not be considered.
- Do not forget to mention your names/matriculation numbers.
- You are free to use any programming language to solve the problems. The usage of libraries that allow you to circumvent implementing the algorithms asked for will not grant you points. For example, when asked to implement clustering, do not use a library that basically solves the task for you.

Exercise 2.1: DNA methylation in hematopoiesis and clustering (50 points)

In the first part of the assignment you will implement and apply a classical clustering algorithm to preprocessed methylation data for different cell types across blood and skin development in mouse by Bock et al. (2012).
The data can be found in the file **methylation.csv** and features the average methylation level of genomic regions of size 1kb size that were sufficiently covered across all samples. If the region overlaps with a gene it is annotated with an Ensemble gene identifier in the 6th column.

(a) First, write a parser for the file and store the data in a way that makes sense to you for further tasks. 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 datapoints. In the latter case, you should set missing methylation values in the data to 0.

(b) Determine the overall average methylation-level per cell type in the context of hematopoiesis. Compare your results to the developmental succession shown in Figure 1 (see next page). Generally, methylation increases with specification during development. Is this also the case here? Discuss your results and elaborate on the difference you would expect in different genomic regions. Are there regions that may lose the methylation they had in earlier developmental stages?

(c) Finally, you need to implement an agglomerative hierarchical clustering approach that helps to group the data. Wikipedia provides a convenient introduction to the topic: `http://en.wikipedia.org/wiki/Hierarchical_clustering`.

Such a method consists of two variable parts: a **distance function** that depicts how (dis)similar two samples are and a **linkage criterion** that uses this function to determine the distance between sets of samples.

Proceed as described:

(1) Implement the Euclidean distance between the methylation patterns of two cell types a and b as:

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

What are the pairwise distances between HSC (hematopoetic stem cells), CD4 (T cells) and TBSC (from skin lineage)?

(2) Implement the average linkage criterion between two sets of cell types A and B as:

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

(3) Implement the agglomerative clustering method and apply it to all cell types across blood and skin development that are part of the dataset.
At the beginning, every cell type forms its own cluster. Until only one cluster is left, iterate over all current pairs of clusters and merge the pair with minimal $L(A,B)$ in each step. In each iteration, print the clusters that are merged, their linkage value $L(A,B)$ and the current cluster assignment.
Draw a dendrogram from the result (by hand). 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?

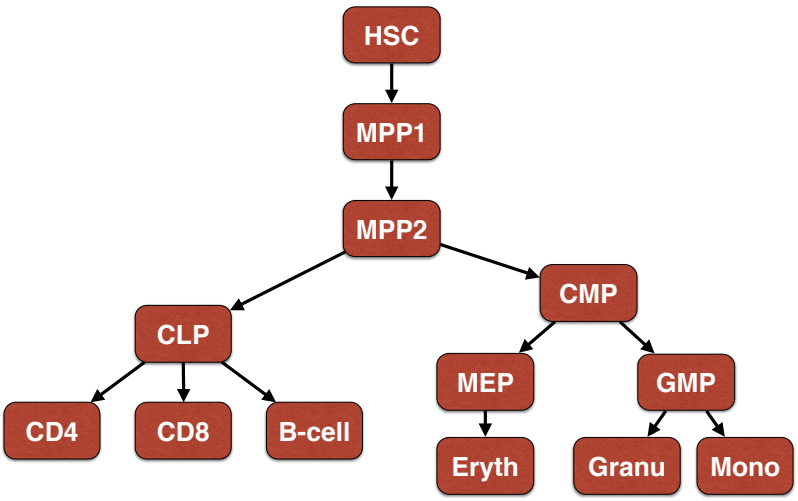

Figure 1: Development of different blood cells from hematopoietic stem cell to mature cells.

Exercise 2.2: Detecting peaks in cell-cycle 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 expression data by Spellman et al. (1998).
The tab-separated data as provided by the authors is supplied in the file **yeast_cell_cycle.txt**.

(a) Since we are only interested in the genes **SWI4**, **MCM1** and **ACE2**, first determine their systematic/locus name. Please report your source and findings in that regard. Then, parse the corresponding rows in the file and treat the columns as successive datapoints. Ignore the column annotations and fill missing expression values in the data with 0. Visualize the expression of the three genes in one plot.

(b) Next, implement a peak detection method that is based on the idea of the watershed algorithm which originates in image processing, see https://en.wikipedia.org/wiki/Watershed_(image_processing). Imagine the expression timepoints as a landscape that includes hills and valleys. The algorithm starts with a "water level" above the highest **absolute** expression value (we also want to find negative peaks) followed by a stepwise lowering of this level while uncovering more and more local maxima. In each step:

(1) Points that are sufficiently adjacent to an already labeled neighboring point are annotated with the same label. This property is controlled by the parameter *adjacency_threshold* that specifies which distance between points is considered near enough (here: how many datapoints earlier or later in the cell cycle).

(2) Points that remain unlabeled are identified as a new peak and thus receive a new label.

The algorithm stops if all data points are labeled. The highest datapoint among a set of positions with the same label defines the peak coordinate. All such peaks are then reported.

(c) Apply the peak detection method that you implemented to the expression data of SWI4. Hereby, use all *adjacency_threshold* parameters from 1 (only considering directly adjecent points) to 4. Plot the expression of SWI4 together with the four different peak detection results into separate plots. How do the detection results differ and which parameter choice seems to be best suited for this data?

(d) At last, plot the expression of SWI4, MCM1 and ACE2 together with their individually annotated peaks into one plot. Use a value for *adjacency_threshold* that seems reasonable to you. What can the progression of the peaks tell you about the role of those transcription factor genes in the cell-cycle context?