Modeling Cell Fate

Prof. Dr. Volkhard Helms
Maryam Nazarieh, Thorsten Will
Summer Semester 2015

Saarland University
Chair for Computational Biology

Exercise Sheet 4

Due: June 30, 2015 10:15 am

Submission

- Submit your solutions on paper or send an email with a single PDF attachment to thorsten.will@bioinformatik.uni-saarland.de. Late submissions will not be considered.
- Start early. Both tutors are absent during the second week of the assignment.
- Do not forget to mention your names/matriculation numbers.
- Discussion of this exercise will be on Juli 7th at 12:45 in the lecture room (E2.1 007).

Exercise 4.1: Experimental methods to determine DNA methylation (30 points)

Methylation at the 5 position of cytosines within CpG pairs is an important epigenetic mark and a popular subject of research in computational biology. While there are many experimental assays to detect methylation signals, we will only consider three classes that still have practical relevance. **Briefly** explain the principle of experimental methods based on

(a) microarrays/chips,

(b) next-generation sequencing,

(c) and immunoprecipitation.

Which steps in their protocols do they have in common? What is the resolution of the methods? What are their advantages/disadvantages? Is any of those methods, slightly modified, maybe also able to detect histone modifications?

Exercise 4.2: DNA methylation in hematopoiesis (40 points)

In this assignment you are working with 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, or encounter missing datapoints. Treat missing methylation values in the data with 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.
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 former methylation?

(c) At last, you will implement an agglomerative hierarchical clustering approach that helps to group the data. Wikipedia contains a convenient introduction to the topic: http://en.wikipedia.org/wiki/Hierarchical_clustering.

Basically, 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 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 actual 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 based on their methylation patterns? Can you see the developmental succession of the blood cells?

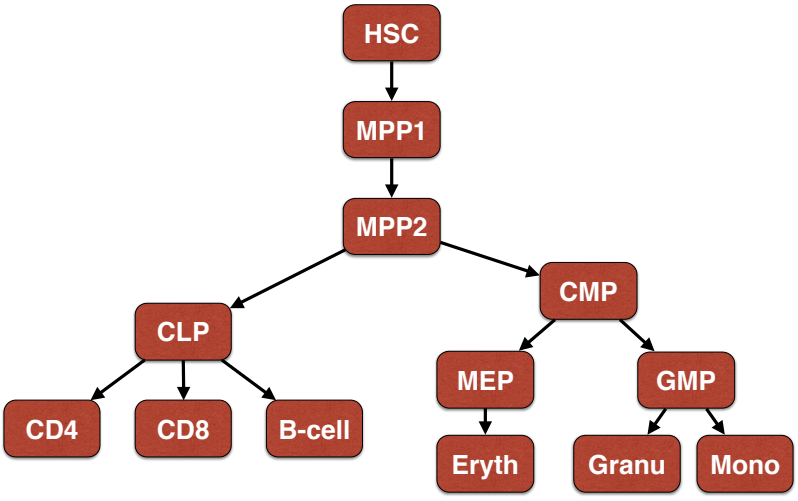

Figure 1: Development of different blood cells from hematopoietic stem cell to mature cells.

Exercise 4.3: Correlation of DNA methylation/expression (30 points)

Bock et al. (2012) also released matching gene expression data for the cell types of the previous exercise. The data is stored in **expression.csv**. Here, expression value are given for each gene (Ensembl gene identifiers in mouse: ENSMUSGx).

(a) Read the expression file and store the data in a convenient way.

(b) Modify your parser from Exercise 4.2 to obtain a methylation level per gene for each sample. Compute the methylation level of a gene as the average methylation of all regions associated with it.

(c) Calculate and report the correlation between gene expression and methylation of the genes for all samples. Only take genes into account that are annotated in both samples. Why could the averaging of the methylation data in part (b) have a huge impact an those results?