Modeling Cell Fate

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

Saarland University
Chair for Computational Biology

Exercise Sheet 5

Due: Juli 14, 2015 10:00 pm

Submission

- Submit your solutions on paper at the beginning of the lecture in the lecture hall or in Room 3.02, both E2 1. Alternatively you may send an email with a single PDF attachment to thorsten.will@bioinformatik.uni-saarland.de. Late submissions will not be considered.
- Do not forget to mention your names/matriculation numbers.
- Discussion of this exercise will be on Juli 21th at 12:45 in the lecture room (E2.1 007).

Exercise 5.1: Experimental methods for gene regulatory networks (30 points)

(a) What are DNAse I hypersensitive sites and what information do they provide? Briefly describe an experimental protocol to obtain such data. (10 points)

(b) Transcription factor binding can be investigated experimentally with ChIP-seq. How? (10 points)

(c) Binding motifs allow to find transcription factor binding sites (TFBSs) computationally. How are they derived and what are the advantages/disadvantages of a motif search compared to experimental data? How can DNAse data help there? (10 points)

Exercise 5.2: DNAse and promoters (35 points)

In this exercise you will map DNAse I hypersensitive sites to gene promoters. ENCODE DNAse data for human embryonic stem cells can be found in **DNAse_H1hESC.narrowPeak** and for CD14 monocytes in **DNAse_CD14.narrowPeak**. More information regarding this data can be found in the corresponding documentation. Transcriptional start sites (TSSs) are taken as defined by the Eukaryotic Promoter Database in the file **human_epdnew_hg19.bed**.

(a) All data is given in the BED file format and reported relative to the hg19 assembly of the human genome. Describe what this format can be used for and why knowledge of the precise assembly is very important for genomic data. (5 points)

(b) Write a small program that uses the TSS definition file and determines all genes whose TSS is within 2kb of a hypersensitive site in a given DNAse file. Note that some genes may have several TSSs, denoted by GENE_#TSS in the annotation. Only report the gene names. You are allowed to use the popular toolset 'BEDTools' for this task or may code it on your own. (20 points)

(c) Apply your tool to the two DNAse files provided in the supplementary material. How many gene promoters are DNAse accessible in stem cells and how many in monocytes? (10 points)

Exercise 5.3: Gene regulatory networks and expression data (35 points)

Gene regulatory networks (GRN) are directed graphs that depict the regulatory interplay between gene products. Classically they associate transcription factors (TFs) with their regulated targets if the TF binds to their promoter and thus likely regulates the corresponding gene.
TF_target_map.txt contains in each line a TF and all proteins to whose gene's promoters the TF binds to. All binding events were derived computationally by motif search and all proteins are reported as UniProt accessions. Furthermore, **expr_H1hESC.gtf** contains h1ESC gene expression data from the ENCODE project in the GENCODE GTF format. Here, genes are given as Ensembl identifiers.

(a) Build an initial human GRN using the TF/target association given in TF_target_map.txt. Report the number of proteins in the network and the number of interactions. Is this network helpful for analyses regarding a specific cellular state? (10 points)

(b) Read the ENCODE H1hESC expression data. First, separate all protein-coding genes into the ones with FPKM value above 1 and those below 1. The former ones are then said to be expressed. Explain in few lines what FPKM means and how gene abundances are quantified from RNA-seq data.
Next, convert the gene identifiers to UniProt accessions using mapping data from the HGNC webservice found at this link. How many of the genes can be mapped to UniProt proteins? How many of them are expressed in stem cells? (10 points)

(c) Refine your initial GRN by integrating the specific expression data. For simplicity, assume that all promoters are accessible by the TFs and that a TF always affects the expression of the gene it binds to. Consequently, all TFs that are expressed are assumed to interact with all their target genes. Report the number of proteins in this network and the number of interactions.
Furthermore, discuss what additional data would be needed to overcome the strong assumptions above. (10 points)

(d) Crucial regulatory drivers are often referred to as 'master regulators'. While there is no fixed definition of the term, it sometimes means the TFs on the highest level of the regulatory hierarchy. Use the notion of topological sorting to determine putative master regulators in the refined network. Do the TFs in the highest level of the hierarchy have parents as well? How many equivalent TFs are on this level? (5 points)