Lecturer: Jan Baumbach Term: Winter 2015/2016

Class: Introduction to Bioinformatics

Exercise sheet – *De novo* sequence motif discovery

1. Explain/define: "Sequence motif" and "gene/transcription factor knock-out".

2. Imagine we performed at transcription factor (TF) knock-out study and identified 25 differentially expressed genes. Now, we aim to identify a binding motif for the knocked-out TF in the upstream sequences of these 25 genes *de novo*. We further assume the TF to dock a 19-bp sequence within each of the upstream sequences.

2a. Download the upstream sequences in FASTA format from URL1 (see below). Write a JAVA program SEQMOTIF that implements an Expectation Maximization algorithm on DNA sequences. What are the 25 most likely 19-bp binding sequences of the TF? For simplification: You may use the nucleotide content of all 25 upstream sequences as background distribution, i.e. you don't need to update the background model; assume it's static.

2b. Afterwards, use the sequence logo painter from our exercise sheet or the publicly available WebLogo painter (URL2, see below) to paint the sequence logo of the binding motif.

2c. Give the consensus sequence.

Hint: The consensus sequence of the most likely motif should start with TTAGG and end with CCTAA.

URL1:
http://www.imada.sdu.dk/~jbaumbac/download/teaching/ws15-16/DM847/exercises/bioinformatics_intro_class_de_novo_sequence_logo_discovery_upstreams.fas

URL2: http://weblogo.berkeley.edu

Please send the JAVA program as well as the source code and the input file via email to your TAs. Also email the names of all

group members and a short tutorial on how to execute the program with the input file.