Lecturer: Jan Baumbach Term: Winter 2015/2016

Class: Introduction to Bioinformatics

Exercise sheet – Sequence logos and operon prediction

1. Explain/define: "Operon" and "Transcription Unit".

2. Write a Java program SELO that reads a file with the following eight sequences. SELO shall compute and output a table with the characteristic numbers for a "standard" sequence logo (no HMM logo).

```
GAHODEFSEXRCCKSCSG
TAAODENSEHROCKSIFL
FAKOLENSEHROKKSAAK
UAFLDENSEFROCKSABE
DHAODEESEKROCKSACE
GSHODYNSEKROCKSYEE
OHHODENSELROCKSTEJ
KIAODENSTMTOCKSHEJ
```

a) What is the most likely word?

b) SELO shall also paint the sequence logo and output it to a PNG file.

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.

3. What is the main idea behind the unsupervised operon prediction method introduced in the class, i.e. why does it "work" without prior knowledge about concrete previously identified operons (sample/training data)?

4. Why are the features for adjacent genes computed separately for closely and distantly related organisms?