Exercise 3: Consensus sequences and recognition matrices

Deadline exercise 3: Tuesday Dec. 4th, 23.59

The sequence motif to which a certain protein binds can be identified experimentally through foot printing of real binding sites from the genome. With a set of already identified binding sequences one can extract a common sequence motif, the consensus sequence, to be used when searching for binding sequences in the genome.

A set of known binding sequences can be lined up under each other (see table below). At each position, *k*, in the sites, the number of occurrences, n_{bk}, of base *b* (b = A,C,T,G) can be calculated. The most occurring base is called the ***consensus base***, and the sequence of the consensus bases at each position in the binding sites is called the ***consensus sequence***. All n_{bk} ($0 < k \leq$ size of binding sequence) constitute a ***recognition matrix*** (see table below).

Gene	**Sequence**	**D**
recA	TACTGTATGAGCATACAGTA	6.4781
uvrA	TACTGTATATTCATTCAGGT	5.2859
uvrB	AACTGTTTTTTTATCCAGTA	6.2238
sulA	TACTGTACATCCATACAGTA	4.1920
uvrD	ATCTGTATATATACCCAGCT	5.3257
mucAB	TACTGTATAAATAAACAGTT	2.3917
clo13	TACTGTGTATATATACAGTA	1.7579
lexA-1	TGCTGTATATACTCACAGCA	5.9664
lexA-2	AACTGTATATACACCCAGGG	4.2489
cle1-1	TGCTGTATATAAAACCAGTG	3.5579
cle1-2	CAGTGGTTATATGTACAGTA	10.8461
Col1b	TACTGTATATGTATCCATAT	6.2857
ColA-1	TACTGTATATAAACACATGT	4.1082
ColA-2	ACATGTGAATATATACAGTT	9.1825
ColE2	ATCTGTACATAAAACCAGTG	5.8670
UMUDC	TACTGTATATAAAAACAGTA	0.6478
recN-1	TACTGTATATAAAACCAGTT	1.1094
recN-2	TACTGTACACAATAACAGTA	6.0218
recQ	GCCTGTTTTTATTT-CAGGC	-

Recognition matrix:

b\k	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
A	5	**13**	1	0	0	0	**14**	1	**16**	2	**14**	6	**15**	6	**10**	0	**19**	0	1	**8**
C	1	2	**17**	0	0	0	0	3	0	1	1	5	0	4	7	**19**	0	0	2	1
G	1	2	1	0	**19**	1	2	0	1	0	2	0	1	0	0	0	0	**17**	4	3
T	**12**	2	0	**19**	0	**18**	3	**15**	2	**16**	2	**8**	3	**9**	1	0	0	2	**12**	7

Consensus sequence:
TACTGTATATATATACAGTA

Let *c* denote the consensus base at position *k* and *b* the actual base at position *k* in the particular sequence considered. Then one can construct a ***weight matrix*** where each element is given by:

$$d_{bk} = \ln\left(\frac{n_{ck} + 0.5}{n_{bk} + 0.5}\right)$$

that provides a measure of the ***dissimilarity*** with consensus at position *k*. The extra 0.5 terms are statistical corrections that make d_{bk} finite also for a base that does not occur at that position in the sample sites. Note that when $b = c$, $d_{bk} = 0$, i.e. no dissimilarity. The sum over all positions, $k = 1,2,3,\ldots,s$, in the sequence (where *s* is the sequence size) is called the ***dissimilarity index***:

$$D = \sum_{K=1}^{s} \ln\left(\frac{n_{ck} + 0.5}{n_{bk} + 0.5}\right)$$

which is a measure of the differences from the consensus sequence. *D* is defined as a positive number that becomes larger the more different a sequence is from the consensus sequence. The larger *D*, the weaker is the expected recognition (binding strength) of the sequence. It is common to use a ***dissimilarity threshold,*** just above the largest dissimilarity index of the known binding sequences, so that identified binding sequences with dissimilarity index below the threshold will be considered as potential binding sites.

Task

In the file *consensus_lab.scm*, some procedures are implemented to help you in this lab. The procedures that are implemented are: **make-sliding-window** and **parse-sites**. **Make-sliding-window** implements a procedure object for a sliding window, that slides over the characters (bases) in a genome-file. **Parse-sites** is a procedure that parses a file with known binding sequences and returns a list of sites, where each site is a list of characters. Two files: *lexA.txt*, with known binding sequences in E. Coli for the *lexA* protein, and *NC_000913.fna*, with the complete genome for E. Coli are to be used in this lab.

1. Implement procedures to build the recognition matrix from the sites returned by parse-sites.
   Tip: a recognition matrix is conveniently represented as a list of ACGT tuples, where each ACGT tuple can be implemented as a procedure object (message-driven) with four local variables: A, C, G, and T (Use **set!** to update them). Each local variable can represent the number of occurrence for each base at position *k*.

2. Implement a procedure that extracts the consensus sequence from the recognition matrix.

3. Implement the dissimilarity calculation procedures.

4. Finally, implement a procedure that slides a window over the whole genome of E. Coli and prints the identified sequences with sufficiently low dissimilarity threshold together with their start position in the genome.