Exercise 1, solutions

1. Sequence databases — paralogs


Solution:

2. Sequence databases — orthologs

Find insulin orthologs from human and mouse in NCBI sequence database. Compare sequences using some aligner.

Solution:

3. Alignment scores

Give a scoring scheme that yields score 6 for the alignment below.

CAGCA-CGTACAACAGCTACCA
CATCACCG--C--CA--TAG-A

Solution:
Example scoring:
- Match = 2
- Mismatch = -1
- Indel = -2

The alignment score is then calculated as: 12 \times 2 + 2 \times -1 + 8 \times -2 = 6.

4. Justification for score matrices.

Prove that Kullback-Leibler divergence (see lecture slides) is always non-negative.

Solution:
Let \( P = \{p_1, \ldots, p_n\} \) and \( Q = \{q_1, \ldots, q_n\} \) be probability distributions. The Kullback-Leibler divergence is defined as

\[
D_{KL}(P||Q) = \sum_{i=1}^{n} p_i \log_2 \frac{p_i}{q_i}.
\]
Lemma. For any $P$ and $Q$ it holds $D_{KL}(P||Q) \geq 0$.

Proof. Recall that $\log_2 x = \ln x / \ln 2$ and $-\ln x = \ln 1 / x$. Since for any $x > 0$ it holds $\ln x \leq x - 1$, we have

$$-D_{KL}(P||Q) = \frac{1}{\ln 2} \sum_{i=1}^{n} p_i \ln \frac{q_i}{p_i} \leq \frac{1}{\ln 2} \sum_{i=1}^{n} p_i \left( \frac{q_i}{p_i} - 1 \right) = \frac{1}{\ln 2} \left( \sum_{i=1}^{n} q_i - \sum_{i=1}^{n} p_i \right) = 0$$

\[\square\]

5. PWMs and PSSMs

Some binding sites for hematopoietic transcription factor GATA-1 from *H. sapiens* are listed below:

AGATAA
TGATAA
AGATAG
TGATAG
TGATCA
TTATCA

Compute the consensus sequence, positional weight matrix (PWM), and position-specific scoring matrix (PSSM) for the sites as described at the lecture (using pseudocounts for the latter). Compute also the sequence logo heights for the letters at each position.

Solution:

Consensus sequence: TGATAA.

<table>
<thead>
<tr>
<th>counts</th>
<th>counts incl. pseudocounts</th>
</tr>
</thead>
<tbody>
<tr>
<td>[2 0 6 0 4 4]</td>
<td>[3 1 7 1 5 5]</td>
</tr>
<tr>
<td>[0 0 0 0 2 0]</td>
<td>[1 1 1 3 1]</td>
</tr>
<tr>
<td>[0 5 0 0 0 2]</td>
<td>[1 6 1 1 1 3]</td>
</tr>
<tr>
<td>[4 1 0 6 0 0]</td>
<td>[5 2 1 7 1 1]</td>
</tr>
</tbody>
</table>

\[
\text{PWM} = \begin{bmatrix}
\frac{1}{3} & 0 & 1 & 0 & \frac{2}{3} & \frac{2}{3} \\
0 & 0 & 0 & 0 & \frac{1}{3} & 0 \\
0 & \frac{5}{6} & 0 & 0 & 0 & \frac{1}{3} \\
\frac{2}{3} & \frac{1}{6} & 0 & 1 & 0 & 0
\end{bmatrix}
\]

\[
\text{PSSM} = \begin{bmatrix}
0.26 & -1.32 & 1.48 & -1.32 & 1 & 1 \\
-1.32 & -1.32 & -1.32 & -1.32 & 0.26 & -1.32 \\
-1.32 & 1.26 & -1.32 & -1.32 & -1.32 & 0.26 \\
1 & -0.32 & -1.32 & 1.48 & -1.32 & -1.32
\end{bmatrix}
\]

Note: PSSM is given under the assumption that $q_a = 1/4$ for all $a \in \{A,C,G,T\}$.

Relative entropies for all columns (include pseudocounts):

\[
\begin{bmatrix}
0.314 & 0.429 & 0.643 & 0.643 & 0.314 & 0.314
\end{bmatrix}
\]

2
Heights in the sequence-logo of all symbols for every column (include pseudocounts):

\[
\begin{bmatrix}
0.094 & 0.042 & 0.450 & 0.064 & 0.157 & 0.157 \\
0.031 & 0.042 & 0.064 & 0.064 & 0.094 & 0.031 \\
0.031 & 0.257 & 0.064 & 0.064 & 0.031 & 0.094 \\
0.157 & 0.085 & 0.064 & 0.450 & 0.031 & 0.031 \\
\end{bmatrix}
\]

6. **Searching with palindrome PSSM**

   Modify the example given in the first lecture (and below) to work with palindrome PSSMs like AGAACAnnnTGTTCT.

   **Solution:**


7. **Motif discovery and statistical significance** *

   Given a set of \( N \) promoter sequences each of length \( L \), an exact motif finding problem can be formulated as the task of finding \( k \)-mers that occur in \( n \) out of \( N \) promoter sequences (at least once in each) and have small probability of occurring that many times in a random set of sequences following the same distribution as the promoter sequences.

   Let \( C_w \) denote the number of promoter sequences containing \( k \)-mer \( w = w_1w_2\cdots w_k \).

   a) Derive an estimate for the expected value of \( C_w \) assuming the background follows the i.i.d. model.

   b) Why \( C_w \) divided by its expected value does not give a good ranking for reporting the statistically most significant \( k \)-mer motifs?

   c) Find out what kind of different rankings (statistical tests) are used in this kind of contexts. What do you need to know about the distribution of values \( C_w \) to use them?

   **Solution:**


8. **Generating DNA sequences with higher-order Markov chains**

   Write a program (e.g. in python) to read the \( k \)-th order distribution of a given DNA sequence (for given \( k \)), and to generate a new sequence of the same length simulating the same distribution.

   **Solution:**