Introduction to Bioinformatics (236523)

HW 2 – Winter 2019

General Instructions:

- Submission according to published pairs only.
- The submission is electronic only in the course website.

Scoring Matrices

1) For the following set of sequences:

<table>
<thead>
<tr>
<th></th>
<th>ref: KVWFQNRRAKWK</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>QIWFQNKRAAK</td>
</tr>
<tr>
<td>2</td>
<td>RVWFQNRRTQLK</td>
</tr>
<tr>
<td>3</td>
<td>RVWFQNRQKEK</td>
</tr>
<tr>
<td>4</td>
<td>RVWFQNRQKEK</td>
</tr>
<tr>
<td>5</td>
<td>RVWFQNRQKEK</td>
</tr>
<tr>
<td>6</td>
<td>RVWFQNRQKEK</td>
</tr>
<tr>
<td>7</td>
<td>ANFFMNKRAKER</td>
</tr>
<tr>
<td>8</td>
<td>VVWFQNRQKQR</td>
</tr>
<tr>
<td>9</td>
<td>VVWFQNRQKAR</td>
</tr>
</tbody>
</table>

a. Compute the scoring function for the substitutions of an amino acid K to R (and vice versa)

b. Compute the PSSM score for the substitutions K→R (position 1 and 12) and R→K (position 7 and 8)

Show the calculations for each result and use the following score function:

\[ M_{i \rightarrow j} = \log_{10} \left( \frac{P(i \rightarrow j) \cdot 1000}{P(i) \cdot P(j)} \right) \]

Note – In this exercise we omit the factor of multiplying the score by 10 to adjust the score to the small set of sequences.

c. Answer the following questions:

i. Looking at the provided alignment, can you explain the original differences you observe for replacing a K to R versus replacing an R to K?

What could the differences in the results indicate regarding the functional role of amino acids R and K in general and in this specific region of the protein?
ii. The score for substituting R and K in the BLOSUM tables are:
BLOSUM45(K,R) = 3, BLOSUM62(K,R) = 2, BLOSUM80(K,R) = 2
How is the result you got from the provided alignment compare to the
different scores in the tables? Based on what we learned regarding the
different BLOSUM matrices, what can you learn from these comparisons
about the evolutionary distances between the sequences in the
alignment you are analyzing? Explain your answer

iii. Compare the PSSM score you have received for individual positions to
the general scores. What additional information can you learn from the
comparison about the function importance of the specific positions in
this protein region? Explain.

**BLAST and PSI-BLAST**

2) The sequences from Q1 are all part of a specific protein domain.
The following sequence contains the protein domain in a human gene:
RIYPWMRSGTDRKRGRQTYTRQLELEKEPHYNRYLTRRRRIEIAHTLCLTERQIKIWFQNRMRKK

a) Which human gene is that? Where is it located in the genome? What is the
function of this gene?
b) What is the protein domain? What is biological function of this domain?
Describe what specify the genes that contain this domain briefly.
c) In the course website, you can see two files that are the results of a BLAST run of
the sequence from section (a) against the protein DB containing only human
proteins: BLAST_results.csv and BLAST_results.txt.
i. Describe the two file briefly, what does each file contain?
ii. What is the average percent identity?
iii. How many DB sequences have an alignment that covers more than 80%
of the query sequence?
iv. Select a DB sequence with 70-80% identity to the query and show its
alignment against the query.
d) Find a homologous gene to that gene in the fruit fly – [Drosophila
melanogaster].
i. Suggest two approaches for finding this gene.
ii. Where is the gene located?
iii. What is the gene function?
iv. Describe a known phenotype related to a mutation / alteration in this
gene

e) In each of the following sections, you are asked to compare 2 BLAST runs and state (no
need to run the experiment yourself) whether the expected E-value would be: (a) the
same for both runs, (b) higher in one of the runs, or (c) there is not enough information
to determine.
Explain your answers.
i. Two nucleotide sequences are searched against a specific DB using BLASTn, with the same parameters. Both sequences are 20 bases long. For both sequences BLAST finds an exact match.

ii. Two protein sequences are searched against a specific DB using BLASTp, with the same parameters. Both sequences are 20aa long. For both sequences BLAST the best match contains a single mismatch from the query.

iii. A nucleotide sequence is searched using BLASTn against DB 1, the best match has 3 gaps and the rest of the alignment is matches. The same sequence is searched against DB 2, which is bigger than DB 1, and the best match has 3 miss-matches and the rest of the alignment are matches. The parameters are the same in both runs and the scoring system is match=+2, miss-match=-1, indel =-2.

iv. The same scenario as in (iii) with the following scoring system: match=+2, miss-match/indel=-1