HW2 - Solution

Scoring Matrices

1) In this question you will calculate part of a scoring matrix based on a set of proteins. The following formula defines the value in each cell in the substitution matrix:

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

\( P(i \rightarrow j) \) – the probability of AA (amino acid) i being substituted by AA j (out of all the possible substitutions in the multiple sequence alignment).
\( P(i) \) – the frequency of AA (amino acid) i (out of all the AAs in all the sequences).

Why is it important to incorporate the frequency of the AAs when computing M (the denominator in the equation above)? Give an example of how this value can make a difference in the final value.

These value are important because one would like to know not only how many times it was changed to j, but also what is the probability to find each aa independently. The denominator is normalization to the frequencies of the AAs and makes the matrix values comparable with one another. An example of how this value can change the final result is: for the same value of \( P(i \rightarrow j) \), say 0.3, if the frequency of i is high (0.9) the substitution score will be higher, compared to the score if i can almost not be found in the sequences.

2) What is the difference between the calculation of M above and the calculation that was presented in lecture 4 slides 19 (ignore the multiplication by 1,000 that is missing from the slide, assume it exists there as well)?

What is the assumption behind each calculation? When would you use each approach?

The calculation that was shown in class is for symmetric matrices:

\[ M_{i \rightarrow j} = \log \left( \frac{\text{Average}(P(i \rightarrow j), P(j \rightarrow i))}{P(i) \cdot P(j)} \right) \cdot 10 \]

In that calculation the numerator is the average of substituting A with B and B with A, whereas here we calculate the frequency of substitution only in one direction. In the lecture calculation the denominator includes frequencies of both AAs, whereas here we take into account only the origin AA frequency.

The current calculation can be used, when we know the reference sequence / the mutation direction. If the reference sequence / the mutation direction is unknown or the evolutionary distance between the sequences is big, it is better to use the symmetric formula.
3) Below you are given a multiple sequence alignment of 7 peptides. Use this alignment and the equation above to calculate the following substitution scores: 
M(P→A), M(P→N), M(N→P)

1: MPNPAPGKPS
2: MNNARNPGKPS
3: AANNARNPGKPS
4: APNARNPGKPS
5: MPNNRAGKAS
6: MNNPNAPGKPS
7: ANNNRPGKPS

Notes:
a) Each sequence is 10AA long. In total there are 10*7=70 AA. There are 10*6=60 different substitutions. We will use these values when computing the frequencies in the substitution matrix.
b) If the calculation in the log turns to be 0, convert it to epsilon=0.001 to allow a result.
c) Sequence number 1 is the reference sequence.

\[ P(P \rightarrow A) = \frac{\# \text{ of time } A \text{ substitute } P}{\# \text{ of total substitutions}} = \frac{5}{60} = 0.083 \]

\[ P(P \rightarrow N) = \frac{7}{60} = 0.1166 \]

\[ P(N \rightarrow P) = \frac{0}{60} = 0 \]

\[ P(P) = \frac{\# \text{ of time } P \text{ appears}}{\text{total } \# \text{ of AAs}} = \frac{16}{70} = 0.228 \]

\[ P(A) = \frac{8}{70} = 0.114 \]

\[ P(N) = \frac{14}{70} = 0.2 \]

\[ M(P \rightarrow A) = \log\left( \frac{P(P \rightarrow A) \cdot 1000}{P(P)} \right) = \log\left( \frac{83}{0.228} \right) \cdot 10 = 25.6 \]

\[ M(P \rightarrow N) = \log\left( \frac{166.6}{0.228} \right) \cdot 10 = 27.08 \]

\[ M(N \rightarrow P) = \log\left( \frac{0}{0.2} \right) \cdot 10 = \log(0.001) \cdot 10 = -30 \]
Another calculation option also gained full credit:

\[ M(N \rightarrow P) = \log \left( \frac{0.001 \cdot 10}{0.2} \right) \cdot 10 = 6.99 \]

A researcher is interested in studying the PUF gene family of RNA binding proteins that act as transcriptional repressors by binding the 3’-UTR of mRNA targets. The collaborating laboratory sent him the sequence of a functional region of the protein, but forgot to mention the organism it was taken from. Please help the researcher to answer a few important questions.

In this question, use the query sequence below (the one available to the researcher) and run PSI-BLAST. Change the following parameters:

- ‘Max target sequences’ to 100.
- ‘Matrix’ to BLOSUM62.
- ‘Database’ to nr.
- PSI-BLAST threshold to 0.005.

>PUF_member
AHGGASDTNSGNAGILSPRDSCTAIVVEYVSPTTMDSSLSG
LEPHLRNLKDDDKSEDKEKNSPDTNKLDDQVTNSNGVNGIDDDKGPNRTPGS
RQPSPAESQPRPNNLLPFPFTNHLMDHQQGMGGLGGVSNGVGGSGSGGAGGA
YAAHQQMAQMQSOLQPPMMNVVGCGMPMAAQSPFLNQAAGPNNHESPGNLLQQQNDVQ
QLFRSQNPQGAAVATNAAAAAAAAAAAAAATSAASAAAAAVGAPPVNGSLQSOQQQQQQQQQQQQQ
QQQQQMMHAAAASOFQFAAQQAQNAAYAAQQATSYVIVNPQEAAPYMGMLAAQMPYGV
APWGMYPGNNLIPQQGQRTQPRPLTIPQQGQAENQPYQVIPAFDLHTGLMSGRPSCTIPRML
VSPAPVLPQGATAGGPPQCPQQLYQPQPQTAQQQNLSSQQNGSSVGGGLALTSSLTGR
DSFTRSTSAFSPSTMGYYSSVGAANNAV

4) Explore the results that were found after the first iteration and answer the following questions based on the results you obtained: What organism was the original sequence taken from? What is the length of the original sequence? Explain

The organism is Drosophila Melanogaster. We cannot claim the exact length of the original protein, since we don’t know what isoform was sent to the researcher. There are two possible isoforms with 100% identity to the query sequence: isoform B – 1185 aa long, and isoform A – 1533 aa long.

5) Continue running psi-blast for two more iterations (total 3 iterations)

Based on the Psi-blast result can you conclude if the PUF protein has homologs in primates or/and carnivores? did your answer changes after 1, 2 or 3 iterations? explain

Attach a screenshot of the top 10 matches results at the first and the third iteration.
No results were found for the carnivores/primates taxa after first or third iteration. The reason for it can be either that there is no known homologues for this protein in these taxa, or we should use another distance matrix, that allow us to search in more evolutionary distant species.

First iteration:

6) Change the ‘Max target sequences’ parameter in PSI-BLAST to 500 and answer question 2 when using the new parameter. Compare between your answer in 2 and 3 and explain what could be the reason for these differences between your answer in question 2 and 3.

Carnivores were found after the first iteration. Primates and carnivores were found after second iteration, and as a result of it also in the third iteration.

For every iteration of PSI-BLAST we use PSSM built from the results of the previous iteration. Since we included much more sequences when building PSSM after the first iteration and carnivores (evolutionary closer to primates, then fruit flies) were among those sequences – it allowed search among sequences that were more evolutionary distant then in previous search.
**First iteration:**

<table>
<thead>
<tr>
<th>Description</th>
<th>Max Score</th>
<th>Total Score</th>
<th>Query Cover</th>
<th>E Value</th>
<th>Identity</th>
<th>Accession</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Second iteration:**

<table>
<thead>
<tr>
<th>Description</th>
<th>Max Score</th>
<th>Total Score</th>
<th>Query Cover</th>
<th>E Value</th>
<th>Identity</th>
<th>Accession</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Third iteration:**

<table>
<thead>
<tr>
<th>Description</th>
<th>Max Score</th>
<th>Total Score</th>
<th>Query Cover</th>
<th>E Value</th>
<th>Identity</th>
<th>Accession</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
7) A researcher generated a chimeric protein in the lab (for more information on chimeric proteins see https://en.wikipedia.org/wiki/Fusion_protein), consisting of two human proteins. Given the chimera sequence (below) and the knowledge you acquired, answer what proteins did the researcher use to generate the chimer? What are the possible functions of this chimeric protein? Describe the procedure you used to obtain the answer including the method and the parameters you used and provide the screenshot.

```
>chimera
EATMVILEPGQPVAATTALELPQPSVTGVPFLGLPSATRALESQGVPATGALFLPGPLMAAGAL
EFSQGSAAGALELLELQPIATVLELPQPGPAELPGQPVTAVALEISVQSVTTLTMQVSQ
LEVPSATTALESNTAQELPTTLVGETSVTGVVDIPMAPESHILASNMTMTIHSLNTMDSQMLAS
NTMDQLASNTMDSQMLASSTMDQMLATSMSDMQMLATSMSDMQMLATSMSMDQMLATSMSDMQMLATSMSDMQMLASSGT
MDMSQMLASGTMDAQMLASGTMDAQMLASSTFRIYGESADAVKARGFLEFVEDFIQVPRLVVKV
GKNGKVIEIVDKSAGVVRIEGDNENKLPREDMVVFVQVGEKQGVQQLYHIAAYLKEVEQ
LRMERLQIDEQLRQIGSRSYSRGRGRGPRGNYTSGYNTSGLSELSPSETESERKDELSWLAGEDD
RDSRHRQDSRRRPQGRGSRVSQGGRGFRGSKISSVILKDPDNSFYSLNDTESDQATTDASE
SHHSTNRRRRSRRRTDEAVLDGMENTEDTASVNGTVAVYISRAESQSRQNLRETLAKN
KREMAKDVIEEHGPEKAIINGPTSAEEDDISKLQRTPGEKINTLKEENTQEAIVLNGVS
```

The chimera consists of two proteins: SON and FXR1.

**SON:** RNA-binding protein that acts as a mRNA splicing cofactor by promoting efficient splicing of transcripts that possess weak splice sites. Specifically promotes splicing of many cell-cycle and DNA-repair transcripts that possess weak splice sites. Also binds to DNA. May indirectly repress hepatitis B virus (HBV) core promoter activity and transcription of HBV genes and production of HBV virions.

**FXR1:** RNA-binding protein required for embryonic and postnatal development of muscle tissue. Interacts with the functionally-similar proteins FMR1 and FXR2. These proteins shuttle between the nucleus and cytoplasm and associate with polyribosomes, predominantly with the 60S ribosomal subunit.

The possible function of the chimera may be a protein with multiple functions, for example DNA binding and also RNA binding. However, it is very important to notice, that the chimera doesn’t necessarily consists of functional domains of those two proteins,
and even if it does, the functionality of one of them or even both may be disturbed due to specific folding of the chimeric protein.

The possible parameters for the search are:

<table>
<thead>
<tr>
<th>Search Parameters</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Program</td>
<td>blastp</td>
</tr>
<tr>
<td>Word size</td>
<td>6</td>
</tr>
<tr>
<td>Expect value</td>
<td>1</td>
</tr>
<tr>
<td>Hitlist size</td>
<td>500</td>
</tr>
<tr>
<td>Geneprobs</td>
<td>11.1</td>
</tr>
<tr>
<td>Matrix</td>
<td>BLOSUM62</td>
</tr>
<tr>
<td>Filter string</td>
<td>F</td>
</tr>
<tr>
<td>Genetic Code</td>
<td>1</td>
</tr>
<tr>
<td>Window Size</td>
<td>40</td>
</tr>
<tr>
<td>Threshold</td>
<td>21</td>
</tr>
<tr>
<td>Composition-based stats</td>
<td>2</td>
</tr>
</tbody>
</table>

The result:
BLAST – Theoretical questions

8) A student ran BLASP with a specific protein as a query against 'nr' database and found the homolog protein with an e-value of $10^{-7}$. If you would run the same query against a different database that has much less sequences than "nr", but includes the same protein that the student found in 'nr', do you expect to get a higher or lower E-value for the same homolog protein? Explain your answer.

E-value is proportional to the length of the database. If the database contains less sequences, we expect a smaller e-value.

9) A researcher managed to isolate a 456 AA long protein, from a new species of purple ants that was found in a rain forest in South America. He would like to find the possible function of the protein and to check if it’s related to the unique color of the ant. What parameters should he use (BLAST program, matrix) when searching for homologs? Explain.

Because this is a new specie, we would like to search as far as possible to recover the possible function of the protein. Using PSI-BLAST with BLOSSOM45 will allow us to find more distant homologs.

10) A researcher isolated a 15 AA long polypeptide from human kidney cell line. He knows that the polypeptide is a part of longer protein and would like to find the function of the original protein using blast. What parameters would you suggest the researcher to use (BLAST program, matrix)?

The researcher should use PAM30 matrix with BLASTP. Since we use very short sequence, the probability to find this sequence in large number of proteins is getting higher. We must be very stringent in our search and allow as less as possible mismatches to discover the source protein. As an alternative, he may also use BLOSSOM90.
11) Paralogs are homolog proteins that are found in the same species (for example myoglobin and hemoglobin in human) and orthologs are homolog proteins that are found in different species (for example hemoglobin in Human and hemoglobin in Monkey). If you would run human hemoglobin in blast against nr, what do you expect will give a lower E-value:
   a. Human myoglobin.
   b. Monkey hemoglobin

Explain your answer.

We would expect that monkey hemoglobin will give a lower e-value. We expect orthologs to be more similar than paralogs, since orthologs preserve the same function through evolution. Paralogs are created by gene duplication and are more likely to evolve to be functionally divergent. As a result of it we assume that the sequence similarity between orthologs will be higher than between paralogs.