**\*\*\*\*\* Assignment 2 \*\*\*\*\***

**(Full marks = 100)**

**Sample Answer**

Answer-1: [50 points. Result = 20 points, E-value = 10 points, Alignment = 20 points]

Blast 1:

Closest: Macaca mulatta (Rhesus macaque)

Furthest: Papio anubis (Olive baboon)

Blast 2:

For this query, BLAST was only able to find Homo sapiens as a close organism under the taxonomy report.

Blast 3:

Closest: Pan troglodytes (Common chimpanzee)

Farthest: Zea mays (corn)

Blast 4:

Closest: Papio anubis (Olive baboon)

Farthest: Halogeometricum pallidum

Blast 5:

Closest: Papio anubis

Farthest: Opitutus terrae

In addition to this students need to show the alignment with the E-value.

Q. 2 [15 points]

Answer-2: So why the slight difference among the different BLAST queries...? Well, the first two have the same parameters, but the only difference between them is the word size. Increasing the word size will lead to increase the sensitivity of the hits because the program is expecting to match more of the bases together for the bigger word size. The difference between BLAST 1 and 3 is the program used. For blast1, we used mega blast, which allows one to retrieve highly similar sequences. On the other hand, blastn is not as selective for high similarity, therefore, the hits that one obtains are not as sensitive as those in the mega blast. Blast4 and 5 use different substitution matrices, for blast4, we used BLOSUM45 and for the last blast, we used BLOSUM80. Blosum45 allows one to find distantly related proteins, so the sensitivity of this matrix is much lower than blosum80. Blosum80, on the other hand, is able to find more related proteins, so the sensitivity for this matrix is far more stringent than blosum45. All these difference in parameters allow one to change the sensitivity for better hits when performing BLAST alignments.

Alternative answer:

After looking over the 5 outputs it will be observed, there were different outputs in the number of hits for each BLAST. This could be because BLAST is a program that has been designed for speed to search for sequences that have alignments with the query but has minimal sensitivity to distant sequences. Differences in hits for these BLAST’s could also be due to the following explanations. From the results BLAST1 contained 44 hits, in contrast to BLAST 2, which only had 1 hit. Therefore, since the only difference between these two BLAST’s was the word size; one can deduce that increasing word size results with fewer hits because fewer sequences will align a larger word size. Thus, a larger word size resulted in fewer hits that were very close homologs. Differently, from BLAST 1 and BLAST 2 that used mega blast in BLAST 3 the program blastn was used, which resulted in 104 hits; with blastn, the word size is a lot smaller which is why it is slower and sensitive because it search’s for distant homologs resulting with many hits. Megablast is a program that is fast and insensitive with a larger word size, which is why it gives hits that are very close homologs. Therefore this is why we see that BLAST1 and BLAST 2 have fewer hits compared to BLAST 3 that used blastn. Lastly, when comparing BLAST4 that used the algorithm BLOSUM45 to BLAST5 that used algorithm BLOSUM80, it was noted how for BLAST4 there were 100 hits on 293 subject sequences, in contrast to BLAST5 that had 100 hits on 211 subject sequences. Substitution matrix BLOSUM45 has a low degree of identity that gives hits that are distantly related, which is why for BLAST4 there are more subject sequences than for BLAST5. However, substitution matrix BLOSUM80 is a high degree identity that gives hits that are very closely related with fewer substitutions, which explains why for BLAST5 there were fewer subject sequences because they are more closely related to the query.

3. Neuroglobin [Gallus gallus] Sequence ID: NP\_001026722.1 [35 points]

Hypothesizing human B-globin and the neuroglobin are biological homologs, is a bad hypothesis because the sequence for the neuroglobin of this organism has an E-value of 0.44. We know that any E-value that falls in between 0.1 and 1 is the iffy range, which is in this case we are not 100 percent sure that the human B-globin and the chicken neuroglobin are homologs or not. Also since the E-value of 0.44 falls within that iffy range, we cannot actually determine if this hit was due to chance. Therefore in order to determine if the human B-globin and the chicken neuroglobin are true homologs, there would have to be further research and studies done. The length of the BLAST hit neuroglobin [Gallus gallus] Sequence ID: NP\_001026722.1 is 160 amino acids which only has a 23% identity to the query sequence. The percent identity is very low, thus one can deduce that the sequence of this hit does not have a lot of amino acids that align at the same position as the query sequence.

On PubMed, the article “Neuroglobin, cytoglobin, and a novel, eye-specific globin from chicken” states that neuroglobin is commonly expressed in the chicken’s brain and eye. It is also mentioned that neuroglobin of domestic chickens [Gallus gallus] is different to mammalian proteins by about 30%. When comparing the chicken and mammalian gene order it was revealed that; neuroglobin is located on conserved syntenic chromosomal segments (Kugelstadt et al).

References:

Kugelstadt, D., M. Haberkamp, T. Hankeln, and T. Burmester. "Neuroglobin, Cytoglobin, and a Novel, Eye-specific Globin from Chicken." Biochemical and Biophysical Research Communications. U.S. National Library of Medicine, 17 Dec. 2004. PubMed. Web. 04 Mar. 2017.