**\*\*\*\*\* Assignment 1 \*\*\*\*\***Top of Form

**Solution**

**Part 1**

**(Full marks = 100)**

The query we use to investigate polymorphisms in the **CAPN10 gene relating to Type II diabetes in different human subpopulations** utilizing the AND/OR conjunctions – while also using certain filters to include only human species studies and publications from the last ten years. When using these filters, the following example returned 22 results:

Correct query should be like this: (**noninsulin-dependent diabetes OR Type II diabetes) AND (calpain-10 OR CAPN10) AND human AND polymorphism**

True Positive: 19

False Positive: 1

Sensitivity: 19/20 = 0.95

This will give you a 19 true and 1 false positive ratio. We define true positive as all the **returned** articles that mentioned variation in the CAPN10 gene and how they are correlated to human Type II diabetes. Most of them also make references to subpopulations – like Argentinians, Iranians, or obese children. The only false positive will inform about the genetics of polycystic ovary syndrome and briefly mentioned CAPN10 and diabetes in the introduction, but was not very relevant to our goal.

**Part 2**

**(Full marks = 100)**

**Sample Answer**

**Q1. [50 points. Result = 20 points, E-value = 10 points, Alignment = 20 points]**

Blast #1:

Closest: Macaca mulatta (Rhesus macaque), E-value=0

Furthest: Papio anubis (Olive baboon), E-value=9e-70

Blast #2:

Only finds the identical query sequence.

Blast #3:

Closest: Pan troglodytes (Common chimpanzee), E-value=0

Farthest: Xenopus laevis, E-value=0.008, 50% coverage

Blast #4:

Closest: Papio anubis (Olive baboon), E-value=8e-107

Farthest: Halogeometricum pallidum, E-value=9.6

Blast #5:

Closest: Papio Anubis, Evalue=3e-99

Farthest: Opitutus terrae, E-value=6.0

In addition, the students need to show the alignments with the E-values.

**Q2. [15 points] Order of sensitivity in detecting differences between sequences: #2 < #1 < #3 < #4 < #5. However, note that #4 > #5 in sensitivity is detecting the effects of evolutionary time “look-back”.**

1. BLAST #1: Megablast-28
   1. Blastn less sensitive than blastp
   2. Megablast has large default word size of 28bp so is less sensitive than blastn
   3. Large word size will find only very similar and closely related sequences.
2. BLAST #2: Megablast-256
   1. Blastn less sensitive than blastp
   2. Megablast with largest word size of 256bp is less sensitive than BLAST #1
   3. Very large word size will find only very similar and closely related sequences.
3. BLAST #3: Blastn-11
   1. Blastn less sensitive than blastp
   2. Default blastn word size of 11bp is more sensitive than megablasts BLAST #1 & #2
4. BLAST #4: Blastp-Blosum45
   1. Blastp more sensitive than blastn due to amino acid substitution matrix
   2. BLOSUM45 contains substitution rates for distantly related sequences with only 45% identity, so less sensitive than higher numbered BLOSUM matrices in detecting differences between sequences. However, BLOSUM45 is more sensitive in detecting the effects of evolutionary time “look-back”.
   3. More sensitive than BLASTS #1, #2, & #3.
5. BLAST #5 Blastp-Blosum80
   1. Blastp more sensitive than blastn due to amino acid substitution matrix
   2. BLOSUM80 contains substitution rates for closely related sequences with 80% identity, so more sensitive than lower numbered BLOSUM matrices in detecting differences between sequences.
   3. Most sensitive BLAST search among the 5 searches.

**Q3. [5 points] Find BLAST #4 hit neuroglobin [Gallus gallus] Sequence ID: NP\_001026722.1**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Score** | **Expect** | **Method** | **Identities** | **Positives** | **Gaps** |
| 38.4 bits(114) | 0.77 | Compositional matrix adjust. | 32/147(22%) | 72/147(48%) | 7/147(4%) |

Query 4 LTPEEKSAVTALWGKVNVDEV--GGEALGRLLVVYPWTQRFFE-SFGDLSTPDAVMGNPK 60

L+ +++ + W +V+ V G + RL+ + P +F+ + +++P + P+

Sbjct 6 LSRTQQALIRESWRRVSGSPVQHGVVLFSRLFDLDPDLLPLFQYNCKRFASPQECLAAPE 65

Query 61 VKAHGKKVLGAFSDGLAHLDN---LKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAH 117

H +KV+ ++HL++ L+ + L + H + V E+F +G L+ +L

Sbjct 66 FLDHIRKVMLVIDAAVSHLEDLPCLEEYLCNLGKKH-QAVGVKVESFSTVGESLLYMLEK 124

Query 118 HFGKEFTPPVQAAYQKVVAGVANALAH 144

+G F+P V+ A+ ++ + V A+

Sbjct 125 CLGAAFSPDVREAWIELYSAVVKAMQR 151

**Q4. [20 points] Neuroglobin [Gallus gallus] Sequence ID: NP\_001026722.1**

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 also 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.

**Q5. [10 points] BLAST search #4 is the best to find distant relatives of the human beta-globin.**

1. BLASTP better than the BLASTNs and megablasts because more sensitive
   1. Amino acid substation matrix more sensitive than DNA substitutions
2. BLOSUM45 is better than BLOSUM80 to find distant relatives
   1. Distant relatives will have lower percent identity than 80%

Part 2:

Identical sequences to your query sequence are not homologous to your query sequence. They ARE your query sequence.

Distant homologs have E-values in the range of 0.1 to 10.﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿

Need to post the actual alignments as well﻿﻿﻿﻿﻿﻿.

Posted the wrong Gallus gallus sequence.

Did not include the Gallus gallus alignment.

Also, many questions went on answered.

BLOSUM45 is better for DISTANTLY related homologs.

Some key concepts are missing from Part 2.  
Please come see me about this HW so that we can talk about it.