**Print your Name: James Jedediah Smith**

**MIDTERM EXAM**

**Due end of class**BIFX-550: Functional Genomics: Sequence Analysis and Structural Bioinformatics   
Hood College, Frederick, MD  
Dr. S. Ravichandran, Ph.D.

**INSTRUCTIONS:**  
The exam is worth a total of 100 points.   
Use GRCh38 assembly and latest transcript(s) for all the questions unless you are specifically asked to use other assembly versions. Please feel free to use NCBI, Ensembl and/or UniProt databases to answer the questions

**When using NCBI/UniProt, please clear all filters before you begin a new search.**In Ensembl, please use “**Configure this page**” menu to make sure the options that you want are the only ones that are turned on. Here are some useful links:   
NCBI: <https://www.ncbi.nlm.nih.gov/>   
Ensembl: <http://useast.ensembl.org/index.html>  
UCSC: <https://genome.ucsc.edu/>   
UniProt: <http://www.uniprot.org/>   
EBI: <http://www.ebi.ac.uk/Tools/msa/>   
TCOFFEE: <http://www.ebi.ac.uk/Tools/msa/tcoffee/>   
BioMart (available from Ensembl, <http://useast.ensembl.org/index.html> )   
Galaxy (Please create the free account at <https://usegalaxy.org/>   
SMS2 (basic bioinformatics manipulation): <http://www.bioinformatics.org/sms2/>   
E-Direct (Please make sure you can use E-direct on your computer before you begin the exam

Tips:

* Type your name at the top of this document and upload in BB before leaving the classroom
* For the question that requires you to show the work (multiple sequence alignments, e-direct commands etc.) you can copy and paste the output into this document.
* Please start working on the BLAST problem(s). For example, initiate the BLAST search and go on to work on other questions and come back when BLAST is done. If some component within BLAST is not working, use related software to get that information

**Section I: Total 20 points**

**If the question is not a Yes/No question, please provide brief explanations**

1. In general, if a gene has 10 exons, how many introns it will be expected to have? b) If a gene has 13 introns, how many exons would you expect to observe? (2 points)

A gene with 10 exons will probably have 9 introns. A gene with 13 introns will probably have 14 exons.

1. In a single-stranded hexanucleotide DNA molecule, where each nucleotide can occur only once, but the order of letters is not important (random). For this case, provide the total number of hexanucleotides one can expect? On the other hand, if the nucleotides can repeat, then how many combinations can you expect for this case? Please provide the answers along with the formulas you used to compute these quantities. (3 points)

If each nucleotides can only occur once, then there aren’t any hexanucleotides for this case because there are only four nucleotides, and hexanucleotide need to have six nucleotides in two codons. Along those lines, there are 64 possible codon combinations. Hexanucleotides are a sequence of 2 codons back to back. So to get the possible number of hexanucleotides, we would need a 64 x 64 matrix, which yields a whopping 4096 possible combinations.

1. If a double-stranded hypothetical DNA has 15% of Cytosine, what % of Adenine would you expect to be present in that DNA? Assume that the double-stranded DNA is evenly paired. (2 points)

We would expect 15% Adenine.

1. We had talked about Regular Expressions (or patterns/motifs) during our sequence alignment class and during one of the hands-on sessions. These patterns had been identified using multiple sequence protein alignments and are often useful in identifying new sequence homologs for the protein family. Look at the sequence (shown in blue font) below to see whether it contains the motif represented by the following regular expression (shown in red) and choose one of the choices. Also show where the pattern lies in the sequence. Note that the motif is shown in capital letters for clarity, please assume them to be small letters. (2.5 points)  
     
   **C-x-H-x-[YFMVIL]-C-x(2)-C-[AYMVIL]   
   mdlsalrvee vqnvinamqk ilecpiclel ikepvstk*cd* *hifckfcm*lk llnqkkgpsq**
   1. No
   2. **Yes (cdhifckfcm)**
   3. Not possible to tell
2. In a BLAST search, if you do two searches, one using a larger database and other using a smaller database, how will it impact the calculation of E values? Assume that you are using the same query along with all the other parameters for the two BLAST searches except for the database size. Please explain your response and don’t give Yes/No as an answer.   
   (2.5 points)

The E-Values will be larger for the larger database query. This is because the length of the database is multiplied in the formula. Larger database length means larger E-Value.

1. Please use the NCBI gene ID 2 to answer the following question. Take the protein coding transcripts from NCBI and Ensembl and compare/contrast them. Please report your findings in a table (gene name, what chromosome, which transcripts reported from Ensembl and NCBI and why). If you are looking to use a transcript in your research which one (report the ID) will you use and why? (4 points)

Gene Name: A2M alpha-2-macroglobulin; Chromosome #: Chromosome 12; Positive/Negative stranded gene: Negative

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Genomic DB | ID | Protein coding transcripts | AA length |  |  |  |  |  |
| NCBI | 2 | 5 | 1474 |  |  |  |  |  |
| Ensembl | ENSG00000175899 | 2 | 1474 |  |  |  |  |  |

Explain which transcript you will use and why?

I will use the one from Ensembl because they seemed to have narrowed it down better and clearly labeled which one is the highest quality copy.

1. Take the following beginning coding segment of a DNA sequence, transcribe, and translate (codon table is shown below) into a protein (peptide) sequence. Please provide the mRNA sequence and the corresponding protein sequence. Hint: The beginning coding segment of the gene is on the negative strand. (4 points)

**5’ GACTACATGCCAAGACAGCTCCATTTTCTAGGAA 3’  
 ||||||||||||||||||||||||||||||||||  
 3’ CTGATGTACGGTTCTGTCGAGGTAAAAGATCCTT 5’**

mRNA: GACUGCGUGCCAAGACAGCUCCAUUUUCUAGGAA  
Protein: DCVPRQLHFLG

**Section II: Each question is worth 20 points. (Total 80 points).**

1. **(25 points)** Use human HBB protein sequence to find homologs. Please show that you can discover globin homologs (alpha, gamma (-1, -2), theta (-1), delta, epsilon, mu, myoglobins) from human and other organisms (esp. from primates, rodents and bacteria). For this exercise, use any BLAST flavor within NCBI and choose any parameters. Please share your successful BLAST flavor and parameter options in a table (what BLAST flavor, matrix, etc.). Please note that I should be able to redo your BLAST search with the parameters you shared to rediscover your results.

Settings for BLASTx to find homologs in primates, rodents, and bacteria.

Graphical user interface, text, application, email

Description automatically generated

Results for BLASTx to find HBB homologs in primates, rodents, and bacteria.

Graphical user interface

Description automatically generated

Settings for DELTA-BLAST to find globin subunits in primates, rodents, and bacteria.

Graphical user interface, text, application, email

Description automatically generated

Results for DELTA-BLAST to find globin subunits in primates, rodents, and bacteria.

Graphical user interface, application

Description automatically generated

Settings for DELTA-BLAST to find globin homologs in primates, rodents, and bacteria.

(Same as previous DELTA-BLAST, but with the following additional changes.)

Graphical user interface, application

Description automatically generated

1. **(30 points)** You will be using e-direct and **Galaxy** for this question. For this problem, you will use **Gene ID, 2**, as your query. Use e-direct to extract the **homologous sequences** for gene id 2 from NCBI **homologene** database. Use e-direct to also extract the alignment scores with human sequence and the domain information for each of the homologous sequences. Take the extracted homologous sequences to carry out multiple sequence alignments using Galaxy sequence alignment tools, **MAFFT** and **ClustalW**. Please use the alignment options (**MAFFT** (**flavor:** L-INS-i). After completing the alignments, please answer the following questions:
2. What is the query gene symbol and what organism does this belong to? Please **copy and paste below** the edirect command you had used to extract the protein homologs. In this step, you will extract the fasta sequences of homologous sequences (I should be able to cut-and-paste your e-direct commands and get the same output). **(10 points)**

esearch -query "2[Gene ID]" -db homologene | efetch -format fasta

The gene symbol is A2M and the organism is *Homo sapiens*.

1. What two organism sequences produce the top alignment score? Provide the score and sequence IDs.   
   **(5 points)**

The format seemed to be a little broken for one of the ClustalW methods. The other processes displayed okay but didn’t have any values or scores in them, so I went with what I have here, even though it seems shockingly low.  
XP\_697477.6 : 0.27685 (27%)

XP\_002938550.2 : 0.24050 (24%)

1. What edirect command can we use to extract the conserved domain information for this gene (from Homologene DB)? How many conserved domains were found in these gene’s protein products? Please cut-and-paste the edirect commands below (I should be able to cut-and-paste your e-direct commands and reproduce your output) **(5 points)**

The conserved domains database accessible through the gene page for A2M hat 16 different conserved domains. I wasn’t sure how to grab this information in Linux, so I just looked it up on the NCBI website.

1. Using the multiple sequence alignments, please comment about the following questions: **(10 points)**
   * 1. Do you see any differences between MAFFT and ClustalW methods? Answer the question and please explain.

Yes they have a very different ways of sorting the data. ClustalW shows you each sequence compared to each other line-by-line, while MAFFT seems to show the conserved regions of each sequence one-by-one.

* + 1. Can you comment about the conservation of key residue(s) that are essential either for the query gene function or its variants (ex Met1Ala) that are suspected to cause a disease between the following alignment procedures, **MAFFT** and **ClustalW**? What are the key residues and where (what DB?) did you find them?

MAMMEISVWKW seems to be a conserved residue. I went into the ClustalW alignment and several of the NP (as opposed to XP) entries all started with this same sequence.

* + 1. What disease(s) (if any) are they associated with?

They are associated with Alzheimer's disease.

1. **(25 points)** Tomato’s 1-aminocyclopropane-1-carboxylate synthase 2 (ACS2) gene is involved in ripening. If you have time, please read about the latest developments in the horticulture research area in the following review article, “Understanding development and ripening of fruit crops in an ‘omics’ era”, Gapper et al, Horticulture Research (2014); **free article**; <http://www.ncbi.nlm.nih.gov/pubmed/?term=26504543>*.* Please note that reading the mentioned manuscript is not necessary to answer this question.

Use the tomato ACS2 mRNA transcript to identify whether there is a homologous transcript in the Potato genome. What BLAST will you use for this step? Make sure the two transcripts (one from Tomato and Potato) also produce similar proteins. If you find several transcripts within the potato ACS2 gene, report the one with the largest score. Also report Total score, query sequence coverage, Identity\_score and E-value. Show the alignment (Font: Courier New; size 9) and explain in a few sentences how you carried out the alignments.

Hint: The hit should cover almost the whole region of the query.

BLASTn with ACS2 mRNA from Tomato to find homologs in Potato species. (next page)

Graphical user interface, text, application, email

Description automatically generated

Results from BLASTn

Graphical user interface, text, application, email

Description automatically generated

Solanum tuberosum 1-aminocyclopropane-1-carboxylate synthase mRNA, 3' end of cds has the largest score. Total score was 2308. Query sequence coverage was 85%. Identity score was 94%. E-Value was 0.0.

Query: Solanum lycopersicum 1-aminocyclopropane-1-carboxylate synthase 2 (ACS2), mRNA Query ID: NM\_001247249.3 Length: 1857

They are both some kind of aminocyclopropane-1-carboxylate synthase and should therefore produce similar proteins. I found this by simply looking up the ACS2 for tomatoes, grabbing the mRNA ID, then pasting that into BLASTn, and restricting my results to potatoes.

>Solanum tuberosum 1-aminocyclopropane-1-carboxylate synthase mRNA, 3' end of cds

Sequence ID: L20634.1 Length: 1571

Range 1: 1 to 1570

Score:2308 bits(5312), Expect:0.0,

Identities:1487/1586(94%), Gaps:23/1586(1%), Strand: Plus/Plus

Query 266 AAAGCATACGATAGTGATCCTTTCCACCCTCTAAAAAACCCCAACGGAGTTATCCAAATG 325

|||||||||||||| |||||||||||||||||||| ||||| || |||||||||||||||

Sbjct 1 AAAGCATACGATAGCGATCCTTTCCACCCTCTAAAGAACCCAAATGGAGTTATCCAAATG 60

Query 326 GGTCTTGCTGAAAATCAGCTTTGTTTAGACTTGATAGAAGATTGGATTAAGAGAAACCCA 385

|| ||||||||||||||||||||||||||||||||||| |||||||||||||||||||||

Sbjct 61 GGACTTGCTGAAAATCAGCTTTGTTTAGACTTGATAGAGGATTGGATTAAGAGAAACCCA 120

Query 386 AAAGGTTCAATTTGTTC---TGAAGGAATCAAATCATTCAAGGCCATTGCCAACTTTCAA 442

|||| |||||||||||| |||||||||||||||||||| |||||||||||||||||||

Sbjct 121 AAAGCTTCAATTTGTTCCAATGAAGGAATCAAATCATTCAGGGCCATTGCCAACTTTCAA 180

Query 443 GATTATCATGGCTTGCCTGAATTCAGAAGAGCGATTGCGAAATTTATGGAGAAAACAAGA 502

||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||

Sbjct 181 GATTATCATGGCTTGCCTGAATTCAGAAGAGCGATTGCGAAATTTATGGAGAAAACAAGA 240

Query 503 GGAGGAAGAGTTAGATTTGATCCAGAAAGAGTTGTTATGGCTGGTGGTGCCACTGGAGCT 562

||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||

Sbjct 241 GGAGGAAGAGTTAGATTTGATCCAGAAAGAGTTGTTATGGCTGGTGGTGCCACTGGAGCT 300

Query 563 AATGAGACAATTATATTTTGTTTGGCTGATCCTGGCGATGCATTTTTAGTACCTTCACCA 622

|||||||||||||||||||||||||||||||| |||||||||||||||||||||||||||

Sbjct 301 AATGAGACAATTATATTTTGTTTGGCTGATCCAGGCGATGCATTTTTAGTACCTTCACCA 360

Query 623 TACTACCCAGCATTTAACAGAGATTTAAGATGGAGAACTGGAGTACAACTTATTCCAATT 682

|| ||||||||||||||||||||| |||||||||||||||||||||||||| ||||||||

Sbjct 361 TATTACCCAGCATTTAACAGAGATCTAAGATGGAGAACTGGAGTACAACTTCTTCCAATT 420

Query 683 CACTGTGAGAGCTCCAATAATTTCAAAATTACTTCAAAAGCAGTAAAAGAAGCATATGAA 742

||||||||||||||||| ||||||||||||||||||||||||||||||||||||||||||

Sbjct 421 CACTGTGAGAGCTCCAACAATTTCAAAATTACTTCAAAAGCAGTAAAAGAAGCATATGAA 480

Query 743 AATGCACAAAAATCAAACATCAAAGTAAAAGGTTTGATTTTGACCAATCCATCAAATCCA 802

|||||||||||||||||||||||||||| |||||||||||||||||||||||||||||||

Sbjct 481 AATGCACAAAAATCAAACATCAAAGTAAGAGGTTTGATTTTGACCAATCCATCAAATCCA 540

Query 803 TTGGGCACCACTTTGGACAAAGACACACTGAAAAGTGTCTTGAGTTTCACCAACCAACAC 862

||||| ||||||||||||||| |||||||||||||| |||||||||||||||||||||||

Sbjct 541 TTGGGTACCACTTTGGACAAATACACACTGAAAAGTCTCTTGAGTTTCACCAACCAACAC 600

Query 863 AACATCCACCTTGTTTGTGACGAAATCTACGCAGCCACTGTCTTTGACACGCCTCAATTC 922

||||||||||||||||| |||||||||||||||||||| ||||| |||||||||||||||

Sbjct 601 AACATCCACCTTGTTTGCGACGAAATCTACGCAGCCACGGTCTTCGACACGCCTCAATTC 660

Query 923 GTCAGTATAGCTGAAATCCTCGATGAACAGGAAATGACTTACTGCAACAAAGATTTAGTT 982

||||| ||||||||| ||||||||||| ||||||||||||| ||||||||||||||||||

Sbjct 661 GTCAGCATAGCTGAAGTCCTCGATGAAAAGGAAATGACTTATTGCAACAAAGATTTAGTT 720

Query 983 CACATCGTCTACAGTCTTTCAAAAGACATGGGGTTACCAGGATTTAGAGTCGGAATCATA 1042

||||||||||| |||||||||||||||||||||||||||||||||||| |||||||| ||

Sbjct 721 CACATCGTCTATAGTCTTTCAAAAGACATGGGGTTACCAGGATTTAGAATCGGAATCGTA 780

Query 1043 TATTCTTTTAACGACGATGTCGTTAATTGTGCTAGAAAAATGTCGAGTTTCGGTTTAGTA 1102

|||||||||||||| || ||||||||||| |||||||||||||||||||||||||||||

Sbjct 781 TATTCTTTTAACGATGACGTCGTTAATTGCGCTAGAAAAATGTCGAGTTTCGGTTTAGTG 840

Query 1103 TCTACACAAACGCAATATTTTTTAGCGGCAATGCTATCGGACGAAAAATTCGTCGATAAT 1162

|| || |||||||||||||||||||| || ||||||||||||||||||||||||||||||

Sbjct 841 TCAACTCAAACGCAATATTTTTTAGCCGCTATGCTATCGGACGAAAAATTCGTCGATAAT 900

Query 1163 TTTCTAAGAGAAAGCGCGATGAGGTTAGGTAAAAGGCACAAACATTTTACTAATGGACTT 1222

||||| | |||||| ||||| ||||||| |||||| |||||||||||||| ||||||||

Sbjct 901 TTTCTGACAGAAAGTGCGATAAGGTTAGCTAAAAGACACAAACATTTTACCAATGGACTC 960

Query 1223 GAAGAAGTGGGAATTAAATGCTTGAAAAATAATGCGGGGCTTTTTTGTTGGATGGATTTG 1282

||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||

Sbjct 961 GAAGAAGTGGGAATTAAATGCTTGAAAAATAATGCGGGGCTTTTTTGTTGGATGGATTTG 1020

Query 1283 CGTCCACTTTTAAGGGAATCGACTTTCGATAGCGAAATGTCGTTATGGAGAGTTATTATA 1342

||||| |||||||||||||||||||||||||| |||||||||||||||||||||||||||

Sbjct 1021 CGTCCGCTTTTAAGGGAATCGACTTTCGATAGTGAAATGTCGTTATGGAGAGTTATTATA 1080

Query 1343 AACGATGTTAAGCTTAACTTCTCGCCTGGATCTTCGTTTGAATGTCAAGAGCCAGGGTGG 1402

||||| || ||||| ||| ||||||||||||| |||||||||||||||||||||||||||

Sbjct 1081 AACGACGTAAAGCTCAACGTCTCGCCTGGATCATCGTTTGAATGTCAAGAGCCAGGGTGG 1140

Query 1403 TTCCGAGTTTGTTTTGCAAATATGGATGATGGAACGGTTGATATTGCGCTCGCGAGGATT 1462

||||||||||||||||| |||||||||||||||||||| ||||| ||||| ||| |||||

Sbjct 1141 TTCCGAGTTTGTTTTGCGAATATGGATGATGGAACGGTGGATATCGCGCTAGCGCGGATT 1200

Query 1463 CGGAGGTTCGTAGGTGTTGAGAAAAGTGGAGATAAATCGAGTTCGATGGAAAAGAAGCAA 1522

|||||||| ||| |||||||||||||||||||| ||||||| |||||||||||||||||

Sbjct 1201 CGGAGGTTTGTACGTGTTGAGAAAAGTGGAGATGAATCGAGCGCGATGGAAAAGAAGCAA 1260

Query 1523 CAATGGAAGAAGAATAATTTGAGACTTAGTTTTTCGAAAAGAATGTATGATGAAAGTGTT 1582

|||||||||||||||||||| |||||||||||||||||||||||||||||||||||||||

Sbjct 1261 CAATGGAAGAAGAATAATTTAAGACTTAGTTTTTCGAAAAGAATGTATGATGAAAGTGTT 1320

Query 1583 TTGTCACCACTTTCGTCACCTATTCCTCCCTCACCATTAGTTCGTTAAGACTTAATTAAA 1642

||||||||||||||||| |||||||| ||||||||| ||||||| || ||||||||||||

Sbjct 1321 TTGTCACCACTTTCGTCTCCTATTCCACCCTCACCACTAGTTCGATAGGACTTAATTAAA 1380

Query 1643 AGGGAAGAATTTAATTTATGtttttttATATTTTGaaaaaaaTTTGTAAGAATAAGATTA 1702

|||||||||||||||||||||||||||||| ||||||||| |||||||||||||||||||

Sbjct 1381 AGGGAAGAATTTAATTTATGTTTTTTTATA-TTTGAAAAATATTTGTAAGAATAAGATTA 1439

Query 1703 TAATAGGAAAAGAAAATAAGTATGTAGGATGAGGAGTATTTTCAGAAATAGTTGTTAGCG 1762

|| |||||| | ||| ||||||||||||||||||||||||||||||

Sbjct 1440 TAGAAGGAAA------------TCTAG---GAGGAGTATTTTCAGAAATAGTTGTTAGCG 1484

Query 1763 TATGTATTGACAACTGGTCTATGTACTTAGACATCATAATTTGTCT---TAGCTAATTAA 1819

|||||||||||||||| ||||||||||| ||||||||||||||||| || |||||||

Sbjct 1485 TATGTATTGACAACTGATCTATGTACTTTGACATCATAATTTGTCTATCTAATTAATTAA 1544

Query 1820 TG-AATGCAAAAGTGAAGTTATGTTA 1844

|| |||| |||||| |||||||||||

Sbjct 1545 TGAAATGTAAAAGTAAAGTTATGTTA 1570