**BIO 434 Bioinformatics Exercise using Myostatin**

**Similarity in Developmental Biology**

We often use model organisms in developmental biology to learn about human development in health and disease. However, what basis is there for using these seemingly different organisms to study human development? Perhaps another way to think about this would be to find examples of proteins in model organisms and look for similar proteins in humans. One could determine the sequence of the protein produced from our gene of interest in humans, and compare this to the sequence in a chosen model organism. It makes sense that proteins with similar amino acid sequences likely have similar structures and hence functions. However, the actual function of the protein cannot be determined from its sequence alone. Rather, one would research the literature and perform laboratory experiments to test the function of a protein.

Let’s review an example from class and use bioinformatics to help guide us to an answer. While studying gene expression earlier in this course, myostatin was used as an example of a mutated gene that resulted in an obvious phenotype. It seems that differential RNA processing results in an unusual splice site, thus altering the mRNA in such a way, that produces a non-functional protein.

Here, let’s retrieve the human myostatin protein sequence and compare it to the mouse protein sequence using bioinformatics. Essentially the question we are asking is, **how similar are the myostatin proteins from human and mouse?**

**Bioinformatics** combines the power of computational science with the ability to retrieve and analyze biological data, such as DNA, RNA and protein sequences. Due to the vast quantity of data available, computer programs are required to find your “needle in a haystack”. In addition to finding comparative sequences from large datasets, bioinformatics can also remove unwanted information and decipher patterns in the remaining information. Meaningful outcomes are further resolved using statistical analyses.

**BLAST** (Basic Local Alignment Search Tool) is a bioinformatics tool that can be used to perform various functions, including comparison of amino acid sequences from different organisms. How does BLAST work? I’m glad you asked! Your protein sequence is sampled in short segments (termed **query words**), and used to search for matching sequences in a reference data base. This **Local Alignment** strategy is used to find similarity between sequences of different lengths, typically using only a smaller portion of the query sequence for the similarity comparison. Searching for shorter fragments of your sequence accomplishes 2 tasks: 1. It decreases the computational time and 2. Increases the chances of finding a sequence of biological relevance. After comparison to the database of sequences, the degree of similarity is calculated; the end result is a quantitative assessment (score) of similarity between your query sequence and each of the database sample sequences.

Although bioinformatics can quantitatively test for similarities between sequences, as aforementioned, bioinformatics cannot with certainty link sequence similarity to function. Function is difficult to quantify, as the same protein in different tissues may perform different functions based on, for example, substrate availability (Pearson, 2013) and/or partner proteins. Knowing that a similarity score does not equate to functional similarity, sequence similarity is still a good starting point when inferring structural and functional relationships. Should sequence similarity be detected, further literature searches and/or laboratory experiments could be performed to support (or refute) a functional link.

During sequence alignment against a database, statistical analyses are performed to indicate the number of times that the alignment of your sequence segments (query words) could happen by random chance, rather than through actual sequence similarity. The **E value** or expectation value, reports the number of times that alignment between your query sequence and sequences in the database, could occur by random chance. Hence, smaller E values would reflect a lower probability of the alignment occurring by chance, whereas larger E values, indicate that the alignment is likely due to chance rather than due to an actual sequence alignment. Larger databases and shorter query sequences will increase the likelihood of alignment due to random chance.

Our **main goal for this exercise is to determine the degree of similarity between the human and mouse myostatin proteins**. We will use sequence alignment to sequences acquired from a database to accomplish our goal. But first, we will do an exercise that illustrates how BLAST works.

**Exercise 1. Modeling a simplified version of BLAST**

We will use a single amino acid sequence as our query and calculate a similarity score using a BLOSUM-62 matrix (Henikoff and Henikoff, 1992). The BLOSUM-62 matrix shows 20 amino acids, using their single letter designations, and forms a matrix showing each possible amino acid combination, for your query compared to your subject (database) sequence. Each amino acid combination is given a numerical value, either positive or negative, which reflects the degree of similarity or difference of the substituted amino acid and the frequency of the amino acid in nature.

We will construct alignments beginning with an exact query word match. Once the match is found, we will extend our alignment to the right, until a negative substitution is found. We will repeat our alignment to the left, again, until a negative substitution is found.

To illustrate how BLAST works, we will use the following sequence as our query, with the underlined triplet as the query word.

STWGERGLMPYRGLACEGHI

Let’s say that a search of the database revealed four protein sequences with possible similarity. Using the instructor provided **BLAST Alignment Template (class handout)**, align the query word with each of the protein sequences and extend the match in both directions. Then calculate a similarity score using the BLOSUM-62 matrix. To calculate the similarity score, add the similarity matrix values for the query word and each continuously aligned amino acid in both directions, until a negative value is encountered. A negative value is what will terminate the local alignment on both sides.

Q1. Which of the proteins from the database are the i) most and ii) least similar to our query?

**Exercise 2** **Protein BLAST**

Now that you have been introduced to the process of aligning sequences and scoring their similarity, let’s use BLAST to locate and compare two protein sequences. In this example we will be comparing the sequence of the human and mouse myostatin proteins.

Myostatin is a secreted protein growth factor belonging to the transforming growth factor-β superfamily. It is also known as GDF8 (growth and differentiation factor 8). Myostatin functions as a negative regulator of skeletal muscle growth, affecting both the number and size of muscle fibres, it functions to ensure that muscles do not grow too large. Interestingly, myostatin is also thought to maintain adult muscle stem cell quiescence; a state where damaged muscle cells would not be replaced. Hence, therapeutic inhibition of myostatin signaling in muscle cells could provide a welcomed remedy for muscle wasting. The sequence of myostatin is highly conserved, where mutations of the gene have resulted in increased muscle production in cattle, sheep, dogs and other organisms (Se-Jin Lee, 2012)

Like other TGF- β superfamily members, myostatin is thought to signal through activin receptors and intracellular Smad proteins (see signaling diagram). Specifically, myostatin binds to activin type II receptors, hence recruiting and phosphorylating activin type I receptors. This activates the activin type 1 receptor and initiates an intracellular signal transduction pathway using Smad proteins. Smad2, Smad3 and Smad4 form a protein complex, that translocates to the nucleus to activate the transcription of target genes. Through a negative feedback mechanism, myostatin induces the Smad7 (termed the inhibitory Smad), which blocks further canonical myostatin signaling pathway (Carnac, et al., 2007).

Q2. How similar do you expect the human and mouse myostatin proteins to be? Rationalize your answer. **(3)**

**Computational Procedure:**

1. Go to the NCBI home page at ncbi.nlm.nih.gov. This web site provides access to DNA and protein databases as well as BLAST.
2. At the top of the page, enter MSTN (the abbreviation for myostatin) *Homo sapiens* to search all databases for this protein. Click “Search.”
3. Under the “Proteins” heading, click on “Protein” – hits from your search within a protein sequence database.
4. One of the first few results will be for the MSTN protein of *Homo sapiens*. Click on the first MSTN [*Homo sapiens*] link to take you to a page that will provide details about the 375-amino acid protein, including its sequence.
5. At the top of this page, click on the “FASTA” link. This will display the amino acid sequence in a simple format. FASTA is a standard format, used for both genes and proteins, in which the first line begins with a > symbol, followed by any text. The text is sometimes longer than one line. This is OK so long as there is no “return” in the text. After this one line of text, there is a return, followed by the nucleotide sequence of the gene or the amino acid sequence of the protein.
6. In the upper right hand corner of the page, click on “Run BLAST” under the “Analyze this sequence” heading. This will take you to the BLAST home page.
7. On this page under “Database” select “RefSeq protein”. Then under “organism” type “*Mus musculus*” (the scientific name for the mouse).
8. Click the “BLAST” button at the bottom of the page. The BLAST algorithm is currently looking for similarity by comparing the human myostatin sequence to all annotated mouse protein sequences.
9. When the search process is completed, a Graphic Summary box will appear towards the top of the page. This will show you a graphical comparison of the query and subject sequences and as you scroll down you will see the amino acid alignments.
10. The first subject has a score of >200 and an identity of 96.2% which means that the human and mouse myostatin protein have relatively similar amino acid sequences.

Q3. If we were to compare the *nucleotide* sequences for the gene encoding this protein between humans and the mouse, do you think they would be relatively similar? Explain. **(4)**

**Exercise 3 Nucleotide BLAST**

A comparison of nucleotide sequences can also be done using BLAST. In this exercise, we will compare the human and mouse myostatin gene sequences.

**Computational Procedure:**

1. Return to the NCBI home page (ncbi.nlm.nih.gov).
2. At the top of the page, enter MSTN *Homo sapiens* (the abbreviation for human myostatin) to search all databases. Click “Search.”
3. Under the “Genomes” heading, click on “Nucleotide” – hits from your search within a nucleotide sequence database.
4. Near the top of the page, click on “Homo sapiens myostatin (MSTN), mRNA”. A new screen will appear that gives you a lot of information about this 2,823 bp transcript sequence. At this point we are only interested in obtaining the nucleotide sequence in a format that can be compared to that of the mouse.
5. At the top of this page, click on the “FASTA” link. This will display the nucleotide sequence in a simple format.
6. In the upper right-hand corner of the page, click on “Run BLAST” under the “Analyze this sequence” heading. This will take you to the BLAST home page.
7. Under “Organism” near the middle of the page, type “*Mus musculus*” into the text box. This will allow you to compare the human myostatin gene with the mouse genome, and not with the entire NCBI database. This saves time, because myostatins have been sequenced for several organisms. Other BLAST searches could allow you to compare your sequence with the entire NCBI database or with different subsets of it.
8. Scroll to the bottom of the page and click on “BLAST.”
9. On the resulting page, scroll down and look at the Graphic Summary. At the bottom of the box, you will see several bright red lines of two classes indicating that there are two basic matches (one of the matches has many transcript isoforms transcribed from the same genetic locus). The fact that the lines are bright red indicates that these are very close matches. Mouse over the lines, and it will tell you, in technical terms, which mouse gene your query (human myostatin) matches.
10. Scroll down to “Alignments.” The “Query” is human myostatin. The “Subject” is mouse myostatin. Notice that whenever the genes are identical, there is a vertical line between the identical base pairs.
11. Now take a **screen shot** of the top 400-500 nucleotides of the first variant of subject sequence that aligned to your query. Use the Microsoft Snipping Tool found by clicking on the Microsoft icon on the desktop; this can be found by searching through the Microsoft start menu. If using a Mac, use Grab from the Launchpad.
12. Paste your alignment screen shot in a new Microsoft Word file. This will give you a comparison of the two myostatin gene sequences. Print the alignment screenshots and **submit them as part of your laboratory worksheet**.
13. Highlight the differences in the query and subject sequences on the printed alignment.

Q4. Were your query and subject sequences identical? Provide evidence for your answer.

Q5. What percent identity is there between these sequences? Is this at all surprising to you? Explain. **(6)**

**References**

Carnac, G., Vernus, B., and Bonnieu, A. (2007). Myostatin in the pathophysiology of skeletal muscle. Curr Genomics *8*, 415-422.

Henikoff, S., and Henikoff, J.G. (1992). Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A *89*, 10915-10919.

Pearson, W.R. (2013). An introduction to sequence similarity ("homology") searching. Curr Protoc Bioinformatics *Chapter 3*, Unit3 1.

Se-Jin Lee (2012) Myostatin: regulation, function, and therapeutic applications. In Joseph A. Hill and Eric N. Olson (Eds.), Muscle: Fundamental Biology and Mechanisms of Disease (pp. 1077-1084). Academic Press.