Local vs global alignments

The goal of this assignment is to familiarize yourself with some approaches for identifying and comparing protein sequences using sequence alignment tools.

Part 1. Starting with a mystery sequence, you will use blast to search SwissProt and genomes of target animal species to identify potential homologs of the mystery sequence. Blast uses a strategy of creating **local alignments** to sub-regions of sequences in order to identify regions of conservation between a query sequence and the population of sequences stored in the database.

Part 2. You will download some of the sequences identified from the blast search, and you will create a multiple fasta file containing these potential homologs.

Part 3. Finally, you will use ClustalW to perform a **global alignment** of all 6 sequences, so that you can visually assess the level of conservation of each gene across the metazoa (animal kingdom).

Part 4. You will perform this procedure for two different mystery sequences, which have different degrees of conservation across the species that you will be examining, and you will examine the different levels of conservation of these two genes across the metazoa.

```
>MysterySequence1
MSARGPAIGIDLGTTYSCVGVFQHGKVEIIANDQGNRTTPSYVAFTDTERLIGDAAKNQVAMNPTNTIFD
AKRLIGRKFEDATVQSDMKHWPFRVVSEGGKPKVQVEYKGETKTFFPEEISSMVLTKMKEIAEAYLGGKV
HSAVITVPAYFNDSQRQATKDAGTITGLNVLRIINEPTAAAIAYGLDKKGCAGGEKNVLIFDLGGGTFDV
SILTIEDGIFEVKSTAGDTHLGGEDFDNRMVSHLAEEFKRKHKKDIGPNKRAVRRLRTACERAKRTLSSS
TQASIEIDSLYEGVDFYTSITRARFEELNADLFRGTLEPVEKALRDAKLDKGQIQEIVLVGGSTRIPKIQ
KLLQDFFNGKELNKSINPDEAVAYGAAVQAAILIGDKSENVQDLLLLDVTPLSLGIETAGGVMTPLIKRN
TTIPTKQTQTFTTYSDNQSSVLVQVYEGERAMTKDNNLLGKFDLTGIPPAPRGVPQIEVTFDIDANGILN
VTAADKSTGKENKITITNDKGRLSKDDIDRMVQEAERYKSEDEANRDRVAAKNALESYTYNIKQTVEDEK
LRGKISEQDKNKILDKCQEVINWLDRNQMAEKDEYEHKQKELERVCNPIISKLYQGGPGGGSGGGGSGAS
GGPTIEEVD*


>MysterySequence2
MKCILLASVVAVLIVSSYSLPLLELKPAVKLPKPSQNLGNFVEVANSLGLTDLISAAKAAGIAQFLVMGK
SLTLFGPTNEAFDTIPEAYKPINSTFLKEVLLFHVIKSVVYANAIKNELLVPSILEMPKKDIRFNVYGGG
KIVTAQCSPIIKVNQNASNGVIHVVSRVMIPPFGTVTDVVAMEKQYFSTLLTAVLAAKLQGVLAGPGPFT
VFAPTNEAFAKIPAEKLKEILKNIPLLTKILKYHVVSGTFCSAGLTNGATVPTLEGSDVTVHISGGSVTV
NNAVVVFVDIPVTNGVVHVIDTVLIPKDVEV*
```

Identify potential homologs from other species — Method: Blast search

A. NCBI BLAST to identify three potential homologs: use **blastp** to identify similar protein sequences:

- Go to the NCBI Website and select the link to the BLAST page
- Select blastp (protein blast) and, using MysterySequence#, search the **Swiss-Prot** database for potential homologs
- what are the top three blastp hits (what species)?
    - How similar is your query sequence with the top two hits? (Look at the alignments for the percent identity)
        - Is there a single alignment for each hit, or are there multiple alignments? Why do you think this might be?
    - Click on the hyperlink (the defline) to take you to the accession record for each of the top three hits
- Copy the fasta-formatted protein sequence of the top three hits and place them into a text file, below the mystery sequence, to create a multiple fasta file.

- Make sure that your deflines contain names that will make sense to you later! Include the species name as part of the defline, and keep the defline short (around 20 characters)

B. Species Genome Websites: Blast to identify homologs from selected species of interest

Now look for homologs from selected animal species of interest. NCBI can often be a terrible way to do this, because there is so much to wade through to find the specific genes of interest. For organisms that have had their entire genome sequenced, we can go directly to the Genome Website to find genes and proteins. Your goal is to determine whether the following animals possess homologs: *Nematostella vectensis* and *Drosophila melanogaster*

Identify homologs from *Nematostella (Sea anemone)*:

1. Go to the *Nematostella vectensis* Genome Browser website at the Joint Genome Institute
2. Perform a BLASTP search against the Nematostella genome (search against the protein database)
1. Examine the **top hit** from the blast search:
    - From the alignment, select the "more info" link and examine the information on the page.
    - Collect the protein sequence from the top hit (try selecting the green bar) and copy it into your multiple fasta file

Identify homologs from *D. melanogaster (Fruit fly)*

Drosophila melanogaster, an insect, is a model organism for studying animal development and genetics.

1. Start at the Genome website "Flybase".
2. Perform a blast search to find and copy the protein sequence into your MysterySequences file.

Store your multiple fasta file on bioinf.

Global alignments to assess homology — Method: Clustal Global Alignment

Now that you have collected 5 different potential homologs, you want to examine their similarity along the entire lengths of the protein sequences to assess the level of conservation. We will use a very common global alignment program called Clustal.

1. Perform a Google search for ClustalW and select the top hit
2. Input the sequences in multiple fasta format into the window (or upload your .txt file)
3. Click on the "run" button to start the clustal algorithm
4. Examine the alignment (select the .aln file)
    a. Are the 4 sequences all roughly the same length?
    b. What do the asterisks and colons represent?
    c. For the regions that do align, do you think this region is highly conserved across metazoan (animal lineage) or poorly conserved? Is there a pattern to the conserved regions (are the conserved amino acids randomly located, or clustered into strings of conserved regions)

*** What does the level of "conservation" suggest about homology? ***

Copy the .aln file into a text file and store on bioinf.

** Alignments such as these are used to identify protein domains. An alignment of homologs is constructed, and then assessed for the probabilities of particular amino acids occurring in each of the positions of the sequence, for example, using Hidden Markov Models.