CSci 3003: Introduction to Computing in Biology

Lab Assignment #3

15 points
Assigned: 09/19/19
Due: 09/26/19, before 11:55pm

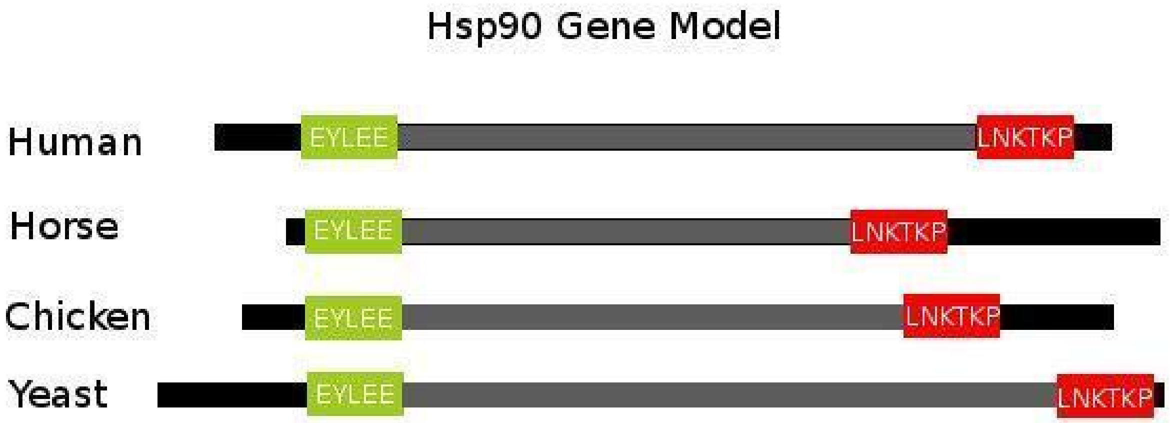
Goals of this lab:

(1) Practice Python syntax rules.
(2) Use numeric and string operations to manipulate variables.
(3) Practice file input/output operations.
(4) Combine all of the above to write your first Python script to accomplish a task.

Writing your first Python program

A protein's biological function results from its physical structure, which is determined by its sequence of amino acids encoded in the genome. Amino acid sequences naturally fold to form 3D structures, but this process is not perfect – misfolded proteins are sometimes created that can pose a risk to the cell. Many diseases including Huntington's, ALS, and Alzheimer's are associated with either genetic or environmental factors that result in an accumulation of misfolded proteins. To battle this, some genomes encode special proteins called chaperones that help proteins fold into their correct conformations and stabilize them during environmental stresses like heat shock. One of these chaperone proteins, Hsp90 (Heat shock protein 90), is extremely well conserved across the tree of life from bacteria to higher eukaryotes.

Chaperone proteins such as Hsp90 contain stretches of amino acids that are nearly perfectly conserved across large evolutionary times. Whether you look in humans, gorillas, dogs, chicken, fish, worms, or even yeast, you will find stretches of the Hsp90 protein with almost exactly the same amino acid sequence. These highly conserved regions typically include *domains*, which are short stretches of amino acids that fold into well-characterized, stable structures with specific biochemical functions. However, in addition to these highly conserved sequences associated with stable physical structure, some proteins also contain regions that take on many different conformations and do not fold into a fixed shape. These intriguing regions are called 'disordered', and they have been an intense area of research over the past decade as researchers try to understand the sequence properties that result in disorder and the biochemical properties that are enabled by protein disorder. The Hsp90 protein is unique in that it contains two extremely well conserved domains linked by a region that is extremely disordered. In addition to a lack of fixed structure and sequence conservation, these disordered regions have been found to have high occurrences of amino acids with special chemical properties.



The above protein model demonstrates the domain and ‘disordered’ linker regions of Hsp90. Domains are shown in green and red, and they are ‘linked’ together by the disordered region shown in gray. Notice that while the certain parts of the gene may change between organisms, there are specific regions that are highly conserved.

In this lab, we will write a program to read and analyze amino acid sequences from 268 Hsp90 orthologs across a diverse set of organisms. You will perform sequence analysis on these proteins to show that while the linker sequence between the two structural domains in Hsp90 seems to be disordered, some biochemical properties are conserved. For more background on the motivation for today’s lab, please check out the references listed at the end of the assignment.

Your task:
We would like to measure the proportion of amino acids that are charged in the linker region of each Hsp90 ortholog. The position of this region varies in each ortholog sequence, but can be identified based on the well-conserved sequences, EYLEE and LNKTKP, near the start and the end of the disordered region, respectively. Write a script that reads in each of the 268 Hsp90 protein sequences, extracts the subsequence associated with the disordered regions, and computes the fraction of negatively and positively charged amino acids in the disordered region. Then, compute the fraction of both positively and negatively charged amino acids across the whole sequence for comparison. Finally, print all of this information to an output file and perform any necessary clean up.

We have provided a skeleton script for you with comments outlining the specific tasks for which you need to write code.

Your script should do all of the following:

(1) Open the sequence file and an output file called *aaoutput.txt*.

(2) Read in the provided sequence file, *Hsp90_conserved.txt,* line by line, being careful to remove any newline characters.

(3) Use built in string methods to split the line by tabs into two variables: name and seq.

(4) Calculate the indices for the conserved domains: EYLEE and LNKTKP and extract the sequence between them (i.e. disordered linker sequence).

(5) One characteristic of disordered regions is that they have a high occurrence of amino acids with 'charged' side chains. The charge on a side chain can either be positive, negative, or neutral. Look up the Side Chain Charges for positive and negative amino acids here: http://en.wikipedia.org/wiki/Amino_acids

(6) Calculate the percentage of amino acids in the disordered region that are either positively or negatively charged.

(7) For comparison, calculate the percentage of amino acids in the entire sequence that are either positively or negatively charged.

(8) Print out the following information about each sequence into the output file:

a. The name of the amino acid sequence taken from file.

b. Percentage of positively charged amino acids in the disordered region.

c. Percentage of negatively charged amino acids in the disordered region.

d. Length of disordered region.

e. Percentage of positively charged amino acids in the whole sequence.

f. Percentage of negatively charged amino acids in the whole sequence.

Please output one line of information per sequence, separating the properties of the sequence with tabs (a tab-delimited table). In addition, the first line of the file should contain the column titles.

(9) Perform any necessary I/O cleanup.

Follow-up questions: (please include answers in your script or in a separate document)

(1) Are there differences in the fraction of charged amino acids in the linker region relative to the rest of the protein? What are they?

(2) How consistent is this property across the 268 Hsp90 orthologs?

Optional extension: (1 extra credit point)
When reading data in from an outside source, it isn't always guaranteed to be “clean”. For example, in the case above, one thing that might have gone wrong would be a sequence that did not contain the conserved domains. Add a statement in your script above that will catch this case and skip over sequences not containing the conserved domains.

Submit to Canvas:
When you're finished with the lab, make a report of any questions you answered plus any requested output, and gather the scripts that you modified.

Submit your homework files using the online Canvas submit page. You should include the following items:

- The completed Python Script.
- A text file containing all the information regarding the sequences.
- A text (or doc) file containing the answers (or you can include it in the other text file or even inside the script).

Grading rubric:

- Writing the script and generating correct output: 13 points
- Follow-up questions: 2 points

References:

Perspective paper on protein disorder: Babu et al. Versatility from Protein Disorder. *Science* 2012: Vol. 337 no. 6101 pp. 1460-1461.
http://www.sciencemag.org/content/337/6101/1460.full

Review of Hsp90: Taipale et al. HSP90 at the hub of protein homeostasis: emerging mechanistic insights. Nature Reviews Molecular Cell Biology 11, 515-528 (July 2010).
http://www.nature.com/nrm/journal/v11/n7/full/nrm2918.html

Study on Hsp90's charged linker region: Tsutsumi et al. Charged linker sequence modulates eukaryotic heat shock protein 90 (Hsp90) chaperone activity. Proc Natl Acad Sci U S A. 2012 Feb 21;109(8):2937-42. http://www.pnas.org/content/109/8/2937.long