**Introduction to Bioinformatics**

Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

According to Wikipedia, amyloid precursor protein is the source of plaque in the brains of people with Alzheimer’s disease. Go to <http://www.rcsb.org> and search for the structure of “amyloid precursor protein”. Click above the search results on the tab labelled “3 News & PDB-101 Articles”, and then click on the entry, “Beta-secretase”. This is one of the enzymes responsible for forming the short protein chains that aggregate to form plaque. In the paragraph, “Tethered Trimmer”, click on the link to the structure of the catalytic domain of the protein, 1sgz. (The other pages can now be closed.) What are the two E.C. entries in the “Macromolecules” section under “Details”?

EC#: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_

Middle-click on icon after the first EC#, which will take you to the IUBMB Enzyme Nomenclature website for that enzyme. (Middle-clicking opens the page in a new window; Ctrl-click does, too.) In the “Reaction” section, what is the first sentence?

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Back at the PDB web site, find the entry, , and middle-click on the UniProt icon (not the number). That brings up the UniProt page about this enzyme.

“The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.”

In which two organs is this enzyme expressed at high levels? (Search on the page, Cntl-F, for “Expression”.)

organs \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_

Back at the PDB web site, still in the “Macromolecules” section, find the words (on the left) “SCOP domains”. On the bar to its right, click on the word “Acid proteases”, and, from the the pop-up menu, select “Show at SCOP website”.

SCOP stands for “Structural Classification of Proteins”. It categorizes proteins based on their structure.

Which family does this protein belong to, according to the SCOP site? \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Besides classifying proteins based on structure, they can also be classified based on “domain”, which is what the “Pfam database does.

“The Pfam database is a large collection of protein families, each represented by multiple sequence alignments…”

Back at the PDB web site, still in the “Macromolecules” section, click on “Full Protein Feature View for P56817” (the page opens in a new tab). Notice above “SCOP domains” is the word, “Pfam”, and to its right is “Asp – Eukaryotic aspartyl protease”. Click on that word, which opens a pop-up window. Click on “Go to Pfam site for [PF00026.22](http://pfam.xfam.org/family/PF00026.22)“, which will take you to the Pfam website. Click on “Domain organization”, on the left menu. How many sequences have the following architecture: A1\_Propeptide, Asp x 2? (It’s about the fifth one down.)

\_\_\_\_\_\_\_

If the mouse cursor is placed on the sequence, the parts are described. Click on “Show all sequences with this architecture”. (A scroll bar is now present on the right.) Let’s compare the sequences of amino acids in this protein from three different species of your choice. To do this, the sequence of amino acids is needed. Click on, say, the one ending in “9TREM”, then, on the left side of the page, click on “Sequence”, and, under the formatted sequence, click on “Show the unformatted sequence.” The bottom, unformatted sequence will be pasted in to <http://www.ebi.ac.uk/Tools/msa/clustalo/>, so open that site (control-click works in Microsoft Word). The data must be entered in FASTA format.

**FASTA –** Fast Alignment Search Tool-All (since it works on both nucleotide and amino acid sequences)**.** Associated with this software is a style of formatting a nucleic acid or protein sequence. It is important because many bioinformatics programs require that the sequence be in FASTA format. **The FASTA format has a title line for each sequence that begins with a “>” followed by any needed text to name the sequence. The end of the title line is signified by a paragraph mark (hit the return key).** Bioinformatics programs will know that the title line isn’t part of the sequence if you have it formatted correctly. The sequence itself does NOT have any returns, spaces, or formatting of any kind. The sequence is given in one-letter code. An example of two proteins in correct FASTA format is shown below:

>Homo sapiens like spinach

MTEYKLVVVGAGGVGKSALTIQLIQNHFVDEYDPTIEDSYRKQVVIDGETCLLDI

LDTAGQEEYSAMRDQYMRTGEGFLCVFAINNTKSFEDIHHYREQIKRVKDSEDVP

MVLVGNKCDLPSRTVDTKQAQDLARSYGIPFIETSAKTRQGVDDAFYTLVREIRK

HKEKMSKDGKKKKKKSKTKCVIM

>Mouse poop

MLPGLALLLLAAWTARALEVPTDGNAGLLAEPQIAMFCGRLNMHMNVQNG
KWDSDPSGTKTCIDTKEGILQYCQEVYPELQITNVVEANQPVTIQNWCKR
GRKQCKTHPHFVIPYRCLVGEFVSDALLVPDKCKFLHQERMDVCETHLHW
HTVAKETCSEKSTNLHDYGMLLPCGIDKFRGVEFVCCPLAEESDNVDSAD
AEEDDSDVWWGGADTDYADGSEDKVVEVAEEEEVAEVEEEEADDDEDDED

So, make up an appropriate title (use the name of the organism; only the first word is used by the software) and paste your data into the Clustal Omega – Multiple Sequence Alignment window. Do this for each of at least three species. (Don’t do the same three species that anyone else does.) Click the “Submit” button. When your results are displayed, click the “Show colors” button (that just colors amino acid residues that have similar properties). You can make some sense out of the line with asterisks, etc. in it, by looking at “Conservative mutation” on Wikipedia. If a warning appears, it may be because java needs to be updated on that computer; either give permission for the program to run, or update java (Control panel; programs; java; update tab).

Let’s look at a “phylogenetic tree” of the data. Click on “Phylogenetic Tree”, which is near the top of the screen. A “phylogram” is shown. The number (represented as distances in the tree) give an estimate of how much the sequences differ. Paste the phylogram below. (To do that, click in the browser to activate it, right-click in the tree and select “copy picture”, and paste it into Word. Email me this document (cking@troy.edu).