Nate Evans

Michael Mooney, PhD

10/15/18

BMI665

HW#3: XML [Uniprot], RE [Prosite]

To run this script, it is necessary to have your desired xml file a relative path location:

/data/xml/

The xml file in question should contain all the data that you wish to analyze, multiple xml files will result in only the first being analyzed. To combine your data into a single xml:

navigate to uniprot.org

search for relevant/desired proteins

click 'add to cart'

once all desired proteins are in the cart

click 'cart'

download all the files as an uncompressed xml to the relative (to this script) directory:

/data/xml/

This script can then be run from command line by navigating to script folder location and using the command:

$python evans\_hw3.py

No additional arguments are required.

Output csv (tab delimited) files will be located in /data/

“common\_ids-commonto2proteins.csv” and “common\_ids-commontoall.csv”

Protein pattern matching will be written to the console.

This can be saved to a text file by adding command:

$python evans\_hw3.py >> your\_file.txt

Results:

There were a significant number of GO ids that are shared between at least two proteins and a subset of that membership that shared ids among all four proteins. All GO ids are prefaced with “GO:“

There are also three of proteins whose sequence had matches to the given prosite patterns. Listed below:

-----------------------------------

PROSITE CONVERTED SEQUENCE MATCHING - pattern id: PS00109

-----------------------------------

Epidermal growth factor receptorsequence pattern matches:

--------------------------------------

pattern match string: LVHRDLAARNVLV

pattern match indices (start, stop): (832, 845)

No further matches found

-----------------------------------

PROSITE CONVERTED SEQUENCE MATCHING - pattern id: PS00109

-----------------------------------

Receptor tyrosine-protein kinase erbB-2sequence pattern matches:

--------------------------------------

pattern match string: LVHRDLAARNVLV

pattern match indices (start, stop): (840, 853)

No further matches found

-----------------------------------

PROSITE CONVERTED SEQUENCE MATCHING - pattern id: PS00109

-----------------------------------

Receptor tyrosine-protein kinase erbB-4sequence pattern matches:

--------------------------------------

pattern match string: LVHRDLAARNVLV

pattern match indices (start, stop): (838, 851)