rif\_correct\_alpha\_v0.py

Purpose

* This program uses CDS, (optional) transcriptome information and Rif-Seq Data points (from 3 to 20 time points) to calculate mRNA decay

Python Version

* “rif\_correct\_alpha\_v0.py” can run in python2 or python3

Packages

* The below packages must be installed on your computer with python before you run the script
  + SciPy python package
  + NumPy python package

Input

* CDS file
  + Input in the form of a tab delimited CSV file for CDS
    - See “ \20010311 Alpha\Data Files for rif\_correct\cds\_ordered.csv” for file structure and order
    - Or use “cds\_orderer\_module\_alpha\_v0.py” to order CDS file correctly
      * Readme for “cds\_orderer\_module\_alpha\_v0.py” has the necessary data structure
* Data file
  + Input in the form of a tab delimited CSV file for the data set
    - See “ \20010311 Alpha\Data Files for rif\_correct\ dataset\_\*\*\*.csv” for file structure
    - Only 1 data set can be used at a time
      * This dataset can have between 3 and 20 datapoints
* (optional) transcriptome data
  + Input in the form of a tab delimited CSV file for the transcriptome data
    - See “ \20010311 Alpha\Data Files for rif\_correct\ Transcriptome\_optional\_input.csv” for the file structure
      * Transcriptome file is entirely OPTIONAL, this program can be run with two or three arguments (the third being the transcriptome file)
      * Without the transcriptome file there will be no information on operons

Linux command line to run:

$python2 rif\_correct\_alpha\_v0.py cds\_ordered.csv dataset\_b.csv Transcriptome\_optional\_input.csv

Or

$python2 rif\_correct\_alpha\_v0.py cds\_ordered.csv dataset\_20point.csv Transcriptome\_optional\_input.csv

Requested inputs:

* “In Quotation Marks: Enter your filename. Filename must end in .csv (output is tab delimited CSV file):”
  + Enter a file name in quotes:
    - “half life output.csv”
* “In ascending order enter your time points (in minutes) seperated by commas (no quotes): “
  + Enter time points in minutes without quotes separated by commas starting at 0 (need 3 to 20 time points):
    - 0,0.5,1,1.5,2,2.5,3,3.5,4,4.5,5,5.5,6,6.5,7,7.5,8,8.5,9,9.5
* "Enter Polymerase mRNA elongation rate in nucleotides per Second (E. coli has been measured at 41, C. Crescentus has been measured at 19.36) entering a value of 0 nt/sec will stop any lag correction: "
  + Enter mRNA polymerase transcription rate in nucleotides per second
  + This provides the lag correction from rifampicin
  + If you enter 0 nt/sec, then no lag correction will occur, and mRNA decay will start at 0 minutes
    - 19.23
* "Enter at what r value would you like the transformed linear regression to be flagged above (r = -1 is perfect fit, r = 0 is a flat line, r = 1 is perfect INVERSE correlation) r = -0.65 is recommended: "
  + This sets what value above will an r\_value\_flag show up in your output, this merely sets a flag
    - -0.65
* "Enter at what percent of the 0 minute RPKM for each gene would you like to set as a cutoff (100% to 0%). For each individual gene RPKM datapoints below the cutoff will be excluded (2.71828 is recommended): "
  + This sets at what percent of the first time point (0 minutes) would you like to no longer consider data for regression
  + This functionally sets any data below this percentage to a null, so that time points and data point are not used for the regression for that gene
  + If you choose 0%, it is instead entered as (1\*10^-100)% as the log of 0 cannot be taken
    - 2.71828

Output

* Output in the form of a tab delimited CSV file
  + “half life output.csv ”

Columns of the tab delimited output file

1. Gene name
2. Slope of the linear regression from the natural log transformed data
3. Y intercept of the linear regression from the natural log transformed data
4. R value of the linear regression from the natural log transformed data
5. Half-Life Calc (minutes)
6. Number of data pts. After user inputted cut off was applied
7. Number of leading data points affected by lag calc, if no poly elongation rate then “0”, if any rate then at least “-1”
8. Natural Log adjusted value of last data point used in the regression
9. In operon (y or n): if you did not use the optional transcriptome file, will always be “n”
10. Operon number
11. Complex or simple: if you did not use the optional transcriptome file, will always be “simple”
12. Total number of data points inputted for that particular gene
13. R value Flag: if the r value of the linear regression is above user inputted r value for a gene, this flag is raised
14. Half-Life Flag: if the half-life is more than twice as long as the last time point used in the regression, or the half-life is negative (indicating the level of mRNA is not decreasing, but rather increasing) this flag is raised

Columns 0. Through AH. Are headered by the user inputted “time points” and for each gene will report the natural log transformed value of your data, including data below cut off