COMP150CSB: Computational Systems Biology
Tufts University, Department of Computer Science
Prof. Soha Hassoun
rev1

Homework 2: Analyzing Molecular Similarities

Assignment Overview

Although traditionally assumed specific, many enzymes, if not all, have promiscuous activities by acting on substrates other than those for which they were evolved to transform [1, 2]. As a result, a given enzyme could catalyze the formation of more than one metabolic product. For example, a recent study found that about one-third of enzymes in a genome-scale model of *Escherichia coli* metabolism are responsible for two-thirds of the known nonspontaneous metabolic reactions [3].

One example potentially promiscuous enzyme is MCR in *E. coli*, which has Malonyl CoA (C00083) as a natural substrate as documented per R00740, but we hypothesize that it can also act on Succinyl CoA (C00091), which has a similar structure (see lecture slides for details).

In this assignment, you will compute the similarity of the natural substrate associated with a particular enzymatic reaction against the similarity of some alternate substrates (candidates).

In a prior lecture, several molecular representations were provided that primarily fall into two groups: graph-based similarity and fingerprint similarity. The lecture also provided several metrics for computing similarities including the Jaccard and Dice coefficients and several others. In this assignment, you will select a molecular representations and a similarity metric and analyze the similarity between four natural products and sets of candidate substrates. These substrates were chosen using the PROXIMAL [4] algorithm described in class, which makes the simplifying assumption that high local similarity between a natural substrate and a candidate molecule will translate to an affinity of the natural substrate's enzyme to the candidate molecule.

Data

The excel sheet provides four different datasets (Prom Set 1, 2, 3 and 4) for which to evaluate molecular similarities. The transformation pattern for each set is associated with a reaction and its natural substrate. A candidate substrate set is provided for each such natural substrate. The datasets are small and shown below.

DataSet Num	Reaction	Reference Molecule	Candidate Substrates
1	R02946	C00810	C00022,C00026,C00111,C00118,C00149,C00188,C00197,C00199,C00231,C00257,C00258,C00318,C00332,C00345,C04411
2	R01976	C00332	C00022,C00026,C00091,C00111,C00199,C00231,C05223
3	R01175	C00877	C00083,C00091,C00100,C00332
4	R01172	C00136	C00083,C00100,C00332

Assignment

- Select a fingerprint and a similiarty metric that you would like to use for the assignment. In a file called HW2.txt (or.pdf, or .doc), write a few lines or a paragraph to explain which fingerprints and metrics you considered, and to explain your choice of fingerprints. Label this section Q1.

- Each data set has a natural substrate, referred to as the referene molecule, and a set of candidate substrates.

For each data set, the assignment calls for the following steps:

1. Retrieve/generate smiles/mol file for list of the given molecules
2. Generate fingerprint from mol/smiles. specifying the desired fingerprint
3. Call function to compute pairwise similarity between each reference molecule and each molecules in the candidate set, specifying the metric.
4. Generate a summary of the results that will be returned in a data structure, and outputted to a file as well. The summary should contain a listing of the similarity between the reference molcule and each molecule in the candidate set. In addition, report the average similiarty score and the standard deviation.

Write and test four functions each implementing each of the steps above. Select appropriate names for the functions, and appropriate parameters and return types. Set the default verbose mode to false.

- Write one function computeSimiliartyOfRefMoleculeToCandidateSet that calls each of the functions you wrote already to implement steps 1-4 as specified above. This function should at minimum take the following arguments:
  referenceMolecule
  candidateSet
  summaryFile (name of file to write the summary of results)
  verboseMode (default is false, otherwise print results to the screen)

- Write a function getData (filename, setNumber) that opens the excel spreadsheet, and reads in the dataset specified by the setNumber.

- Write a function analyzeOneDataSet (setNumber, filename) that will read the correct data (using getData function), and call computeSimiliartyOfRefMoleculeToCandidateSet

An example output is as follows, where the first time specifies the dataset number, and each row thereafter, except the last one, prints the pairwise similarity between the reference molecules and those in the candidate set. The last row summaries the average and standard deviation for the dataset. (numbers below are not accurate).

Summary for Dataset 1
C00810, C00022 0.57
C00810, C00026 0.99
C00801, C00111 0.66
……..
Average Similarity Score: .67, Standard Deviation = .12

- Place the results in HW2.txt, and label this section Q2.

- Write on function runHW2 that will call analyzePromSets that calls analyzeOneDataSet on each of the datasets.

- In HW2.txt (or.pdf, or .doc), write an analysis of the summaries. For each reference molecule, are there any trends regarding the similarity scores? Were there any molecules that were more similar to the refernce more so than any of the others? How did the size of the molecules contribute to the simialrity scores? Label this section Q3.

- *PROXIMAL* conjectures that the specified reaction in each data set will operate on the candidate sets. Based on your analysis and any additional data that you might find, support or argue against this conjecture. Write your answer in HW2.txt

  Each dataset lists the KEGG ID of the reaction that is attributed to the transformation pattern suggested by PROXIMAL. For each such reaction, identify the enzyme responsible for the transformation. Further, lookup any K_M or K_{cat} (Turnover) or K_M / K_{cat} data for that enzyme(s) that catalyze this reaction in BRENDA (you can write a script, or do this manually). What are the functional parameters for the natural substrates? List their values and the organisms that were used to determine these parameters. Do any K_{cat} or K_M / K_{cat} values approach the theoretical limit? How do these values compare to the average values computed in the reading [5] ?

  Examine the same functional parameters for the candidate substrates, if available. Are any of the functional parameters similar in value to those for the natural substrate?

  Write a paragraph or two that summarizes your findings, and patterns or insights from the data. Label this section Q4.

Submission
All homework submissions should be through the web-based provide interface:
https://www.cs.tufts.edu/comp/150CSB/provide.cgi

All your code should be well documented, and you should have a "README" file that explains how one could run your code to genrate the results. You should also provide a specific listing of your environment. If you are using anaconda, you can do that as follows:

```
conda list --explicit > name_environment.txt
```

All discussion/questions should be provided in a .txt or .pdf form.

All files needed to run your code should be submitted. You can submit each file individually, or you can zip and/or tar your files into one file.

References

1. D'Ari, R. and J. Casadesus, Underground metabolism. Bioessays, 1998. 20(2): p. 181-6.
2. Tawfik, O.K. and S. Dan, Enzyme Promiscuity: A Mechanistic and Evolutionary Perspective. Annual Review of Biochemistry, 2010.
3. Nam, H., et al., Network context and selection in the evolution to enzyme specificity. Science, 2012. 337(6098): p. 1101-1104.
4. Yousofshahi, M., et al., PROXIMAL: a method for Prediction of Xenobiotic Metabolism. BMC Syst Biol, 2015. 9: p. 94.
5. Bar-Even, A., et al., The moderately efficient enzyme: Evolutionary and physicochemical trends shaping enzyme parameters. Biochemistry, 2011.