7.36/7.91/20.390/20.490/6.802/6.874
PROBLEM SET 5. Network Statistics, Chromatin Structure, Heritability, Association Testing (24 Points)

Due: Thursday, May 1st at noon.

Python Scripts
All Python scripts must work on athena using /usr/athena/bin/python. You **may not assume availability of any third party modules** unless you are explicitly instructed so. You are advised to test your code on Athena before submitting. Please only modify the code between the indicated bounds, with the exception of adding your name at the top, and remove any print statements that you added before submission.

Electronic submissions are subject to the same late homework policy as outlined in the syllabus and submission times are assessed according to the server clock. **Any Python programs you add code to must be submitted electronically, as .py files** on the course website using appropriate filename for the scripts as indicated in the problem set or in the skeleton scripts provided on course website.

P1 – Network Statistics (10 points)

Assessing bias in high-throughput protein-protein interaction networks

Protein-protein interactions are often stored in databases that cover thousands of proteins and interactions between them. However, there are biases present in these data.

- Poorly studied proteins may be under-represented in these databases because few people have taken the time to identify interacting partners, thus the databases are over-represented among highly studied proteins.
- Evidence of protein-protein interaction is highly variable, as there exist diverse biochemical assays to identify protein-protein interaction and these assays in themselves are biased towards specific types of proteins.

In this problem, we will study these biases and the relationship between them.

You will be completing the script citationNetwork.py, using the networkX module which allows us to manipulate and study networks in python. NetworkX has been installed on Athena so we will follow a similar procedure as other problems.

Log on to Athena's Dialup Service:

```
ssh <your Kerberos username>@athena.dialup.mit.edu
```

Before running any python scripts, use the following command to add the networkX module we installed to your `PYTHONPATH`:

```
export
PYTHONPATH=/afs/athena/course/20/20.320/pythonlib/lib/python2.7/site-
packages/
```

otherwise, you will get an `ImportError`.

You will also need to get the .zip containing the files for this problem in the course folder.

```
cp /afs/athena/course/7/7.91/sp_2014/citationNetwork.zip ~
cd ~
unzip citationNetwork.zip
cd citationNetwork
```

Please submit the citationNetwork.py script online.

It should take a uniprot file and a network and print some scaffold text. The networks that the script expects as input are given in the .pkl files that will be used in part (b). Each of the .pkl files corresponds to a different network. Of course, feel free to modify the script in any way that is helpful to you to answer the questions, but have it conform to the above standard when you submit it.

(a) (2 points) Bias in protein studies: In the zip file, we have provided protein citation data from UniProt in uniprot.txt.

In citationNetwork.py, for each unique entry name in this file, collect the unique number of mapped PubMed ID (correlating to the number of times the protein has been cited).

a. Plot a histogram of the number of citations per protein. Attach the PDF to this write-up.
b. What are the median citation rate and maximum citation rate?
Median:
Maximum:

c. Which protein has been cited the most?

(b) (4 points) Bias in source of interaction evidence: STRING is a database of protein-protein interactions with confidence scores ascribed to each interaction based on distinct sources of evidence (http://www.string-db.org). There are seven sources of evidence, each with its own scoring contribution:

Evidence	Description
Neighborhood score	Computed from the inter-gene nucleotide count
Fusion score	Derived from fused proteins in other species
Co-occurrence score	Score of the phyletic profile (derived from similar absence/presence of genes)
Co-expression score	Derived from similar pattern of mRNA expression measured by DNA arrays and similar technologies
Experimental score	Derived from experimental data, such as, affinity chromatography
Database score	Derived from curated data of various databases
Text-mining score	Derived from the co-occurrence of gene/protein names in abstracts

For each source of evidence, we've provided you with a .pkl file containing a distinct NetworkX graph with the proteins (nodes) and interactions (edges) between them. The weight of the edge is the normalized score for that particular source of evidence (if it was greater than 0.25). You will find a function to load these, and directions for interacting with them, in the script.

a. How many edges (interactions) and nodes (proteins) are in each protein interaction network?

Network	Edges	Nodes
Neighborhood score		
Fusion score		
Co-occurrence score		
Co-expression score		
Experimental score		
Database score		
Text-mining score		

b. How many edges have a normalized score above 0.4? 0.8? What is the number of nodes in interaction networks with these score cutoffs?

Cutoff = 0.4:

Network	Edges	Nodes
Neighborhood score		
Fusion score		
Co-occurrence score		
Co-expression score		
Experimental score		
Database score		
Text-mining score		

Cutoff = 0.8:

Network	Edges	Nodes
Neighborhood score		
Fusion score		
Co-occurrence score		
Co-expression score		
Experimental score		
Database score		
Text-mining score		

c. Comment on what these results say about the different types of evidence.

(c) (4 points) Relationship between number of citations and node degree? The size and distribution of an interaction network measured by distinct types of evidence varies greatly. For each interaction network collect the node degree of each protein. Then, after removing proteins without interactions and interacting nodes without data in UniProt, calculate the Spearman rank correlation to determine if the node degree is correlated with the number of citations of that protein collected in the first section.

a. What is the node citation/interaction correlation for each of the sources of evidence?
b. What are the correlation values when you restrict the interactions to those with at least a score of 0.4? 0.8?

Network	No cutoff	0.4	0.8
Neighborhood score			
Fusion score			
Co-occurrence score			
Co-expression score			
Experimental score			
Database score			
Text-mining score			

c. Is there a relationship between the number of citations and degree? If so, what is the relationship and why do you think this is?

d. Does this relationship vary between sources of evidence? If so, why?

To copy your script and histogram from Athena onto your own computer, use SCP:

```
<in a new Terminal on your computer, cd into your local computer's
directory where you want to download the PDF>

scp –r <your Kerberos
username>@athena.dialup.mit.edu:~/citationNetwork/citationNetwork.py .

scp –r <your Kerberos
username>@athena.dialup.mit.edu:~/citationNetwork/histogram.pdf .
```

P2 – Analysis of Chromatin Structure (5 points)

(A) Suppose we reduced the number of Segway states to be fewer than the true number of distinct patterns of chromatin marks. How might the resulting labels under this model be different?

(B) The *C, M*, *t,* and *J* variables in the Segway model implement a 'countdown' function, one of the core features of Segway. How might these countdown variables improve on a simple HMM model in modeling the underlying genomic states?

Suppose we remove the *C*, *M*, *t*, and *J* countdown variables from the Segway model for the remainder of this problem.

(C) Draw the resulting graphical model.

(D) Assuming that *J* is a binary variable and we allow for 50 segment labels, describe how the conditional probability table for the segment label variables has changed between the old model and this new model in terms of the number of parameters.

(E) Which other core feature of the Segway model does this new model retain that is not present in a simple HMM model?

P3 – Heritability (5 points)

(A) (3 points)

(i) Suppose there is a single locus in a haploid organism controlling a trait with a positive allele for which the phenotype is 1 and a neutral allele for which the phenotype is 0. Calculate V_G for this trait in an infinite population of F_1 children from these two parents.

(ii) Now, suppose there are three unlinked loci each with a positive allele contributing $\frac{1}{3}$ to the phenotype and neutral allele contributing 0 to the phenotype. Calculate V_G.

(iii) Generalize the previous results to calculate V_G for *N* unlinked loci contributing 0 or $\frac{1}{N}$ to the phenotype.

How many possible values are there for the phenotype?

(B) (1 point) You perform linear regression to predict a phenotypic trait (y) on a set of binary genotypic variables (x_1, x_2, … , x_N) for a model system. Show how the R^2 that results relates to the narrow sense heritability of the trait.

(C) (1 point) Assume that all of the genetic components from part (a) are additive. Give the environmental contribution to the observed phenotype variance assuming that the covariance between the genetic and environmental components is zero.

P4 – Association Studies (5 points)

(A) (3 points) Consider the following data case-control data. We will perform a chi-square test for association with a SNP.

	A	T
Case	90	110
Control	50	250

(i) Fill in the following table with the counts you would expect if you assumed independence. Show your work.

	A	T
Case		
Control		

(ii) Now, compute the Chi-Square statistic and state the conclusion for the p-value cutoff of 0.05.

(B) (1 point) You perform a large scale analysis and generate a list of significant SNPs and would now like to prioritize SNPs for further study. How might you use what you have learned from using the Segway model to do so?

(C) (1 point) Describe why it is better to do association tests using a likelihood test based on reads instead of first calling variants and then using a statistical test on the binary variant calls.

7.91J / 20.490J / 20.390J / 7.36J / 6.802J / 6.874J / HST.506J Foundations of Computational and Systems Biology

Spring 2014