

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 1<sup>st</sup> 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. 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. This file was generated using the following procedure (you do not have to generate the file):

Download protein citation data from <http://www.uniprot.org>. Use the 'Advanced Search' to select only those proteins that are human and their status is 'reviewed'. Click on 'customize' to make sure the table has ONLY the 'Mapped Pubmed ID' field, and then download the tab-delimited file.

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).

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

**Median: 3**

**Maximum: 5256**

- Which protein has been cited the most? **P04637 (p53)**

**(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.

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

Network	Edges	Nodes
Neighborhood score	19976	1207
Fusion score	296	240
Co-occurrence score	3944	1148
Co-expression score	149333	3967

Experimental score	112787	10487
Database score	233910	6422
Text-mining score	1251381	14765

- 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	8354	866
Fusion score	144	118
Co-occurrence score	1110	550
Co-expression score	49874	2390
Experimental score	95245	9772
Database score	233910	6422
Text-mining score	620858	14481

Cutoff = 0.8:

Network	Edges	Nodes
Neighborhood score	412	154
Fusion score	18	16
Co-occurrence score	0	0
Co-expression score	5172	642
Experimental score	16523	4035
Database score	233910	6422
Text-mining score	89560	9416

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

Many of the data types have a low proportion of high confidence interactions, indicating that much of the data may be of low quality. As long as a comment on data quality or reliability was made, points were awarded. In many cases, students commented on specific types of evidence, which was also accepted.

- (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	0.058	0.075	
Fusion score	0.104	0.182	0.370
Co-occurrence score	0.017	-0.002	

<b>Co-expression score</b>	0.008	0.012	0.051
<b>Experimental score</b>	0.416	0.429	0.247
<b>Database score</b>	0.104	0.104	0.034
<b>Text-mining score</b>	0.706	0.685	0.564

- 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?

For some forms of evidence, there is a correlation between number of citations and edge degree. This may be due to bias in how much certain proteins have been studied. The more a protein has been experimented on, the more true interactions (and false positive) are likely to be discovered. Also, if a protein has lots of connections, it is likely to be mentioned in the context of its partners. Many people said there was a positive correlation, despite the fact that in many cases it was very low. It was quite open to interpretation, depending on what people considered a strong correlation.

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

Yes, fusion, experimental, and text-mining score have higher correlations than the other data types. This is likely because the data for these experiments comes from a large pool of smaller studies, which have bias in choice of protein to study. Genome-wide assays that look at all proteins at once do not have this problem. Again, this was open to interpretation.

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?

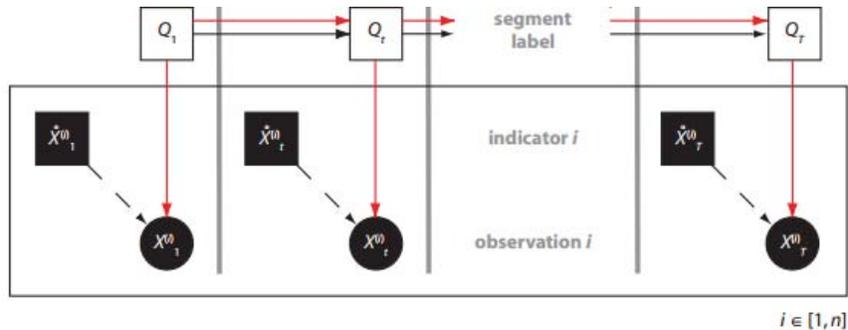
As we are attempting to model the same data using fewer labels, we might expect that labels with similar chromatin mark profiles would be merged.

(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?

A traditional HMM doesn’t allow tuning for state duration. In the case where we train with a large number of labels, we might expect results to be incomprehensible with frequent switching between states. The countdown variables allow us to enforce our beliefs on how long a segment should be – such as allowing a long minimum segment length.

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 that either forces the label to change or prevents it from changing 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.

Previously,  $Q_t$  (the segment label variable) was dependent on  $J_t$  and  $Q_{t-1}$ . As a result, the conditional probability table could be viewed as consisting of  $2 \cdot 50 - 2 = 98$  parameters (normalize along each ‘row’ of the conditional probability table).

However, due to the nature of  $J_t$ , we can be more specific. If  $J_t = 0$  (we prevent a label change), we know that the entry for  $Q_{t-1}$  will be 1 and all the remaining entries will be 0. Likewise, if  $J_t = 1$  (we force a label change), we know that the entry for  $Q_{t-1}$  will be 0 and the remaining entries must sum to 1 giving 48 parameters for this row.

Now, with the dependence on  $J_t$  removed, the table only consists of  $50 - 1 = 49$  parameters – that is, it only depends on the previous state. Many students said the number of parameters was halved, which was also accepted.

Furthermore, some students also considered the fact that there must be one such table for each of the 50 possible values  $Q$  can take on, whereas the above analysis basically assumes  $Q$  is a binary variable.

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

We still retain the observed variable which indicates whether data at a time point is defined or undefined. As a result, we still retain the ability to handle missing data.

**P3 – Heritability (5 points)**

**(A) (3 points)**

For 3A, many students simply copied from the lecture slides, which was accepted, but students should understand how to calculate these quantities.

- (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.

$$V_G = 0.5(1 - 0.5)^2 + 0.5(0 - 0.5)^2 = 0.25$$

- (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  $V_G$ .

$$\mu_G = \frac{1}{8} * 0 + \frac{3}{8} * \frac{1}{3} + \frac{3}{8} * \frac{2}{3} + \frac{1}{8} * 1 = \frac{1}{2}$$

$$V_G = \left(\frac{1}{8}\right)\left(0 - \frac{1}{2}\right)^2 + \left(\frac{3}{8}\right)\left(\frac{1}{3} - \frac{1}{2}\right)^2 + \left(\frac{3}{8}\right)\left(\frac{2}{3} - \frac{1}{2}\right)^2 + \left(\frac{1}{8}\right)\left(1 - \frac{1}{2}\right)^2 = \frac{1}{12}$$

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

The variance for a single allele, which has 50% chance of contributing value 0 and 50% chance of contributing value 1, is (considering it as an independent Bernoulli trial):

$$\frac{1}{2}\left(0 - \frac{1}{2N}\right)^2 + \frac{1}{2}\left(\frac{1}{N} - \frac{1}{2N}\right)^2 = \frac{1}{(2N)^2}$$

Since the alleles are independent, we can add the variances for the  $N$  alleles to get:

$$V_G = \frac{1}{4N}$$

How many possible values are there for the phenotype?

$N+1$  – the phenotype can take on the values 0 and all multiples of  $1/N$  up to/including 1

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

$R^2$  is equal to  $h^2$  (narrow sense heritability).  $R^2$  in a linear regression is defined as the regression sum of squares divided by the total sum of squares (sample variance).

In the case where the predictors are genotypic components, this regression is the same as the additive model covered in class. The numerator is therefore  $\sigma_a^2$  and the denominator is  $\sigma_p^2$ , giving the same equation as narrow sense heritability.

- (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.

$1-R^2$

We were looking for a quantity relating to  $R^2$ , which wasn't made clear in the instructions. Many students used the  $\sigma_p^2 = \sigma_e^2 + \sigma_a^2$  formula and used the  $1/4N$  term for the genotypic variance, which was accepted. We were looking for a ratio of the environmental contribution.

$$\frac{\sigma_e^2}{\sigma_p^2} = \frac{\sigma_p^2 - \sigma_a^2}{\sigma_p^2} = 1 - \frac{\sigma_a^2}{\sigma_p^2} = 1 - h^2 = 1 - R^2$$

**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	56	144
Control	84	216

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

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i} = \frac{(90 - 56)^2}{56} + \frac{(50 - 84)^2}{84} + \frac{(110 - 144)^2}{144} + \frac{(250 - 216)^2}{216} = 47.8$$

The result is significant (there is an association between the SNP and the disease).

**(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?

Segway outputs a segmentation of the genome into its constitutive elements. We can look where SNPs cluster in particular cell types to identify particular genes and regulatory elements, as annotated by Segway. For example, a significant SNP may lie in an enhancer element or affect the motif of a regulator involved in the disease.

**(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.

When genotypes are known, we can just do a chi-square test. Uncertainty in the variant calls adds another layer of error in determining association, especially with sequencing at low depth.

MIT OpenCourseWare  
<http://ocw.mit.edu>

7.91J / 20.490J / 20.390J / 7.36J / 6.802J / 6.874J / HST.506J Foundations of Computational  
and Systems Biology  
Spring 2014

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.