Example projects

The goal of the project is for you to get some hands-on experience in applying some of the network analysis approaches we have seen in class on some real data. Below I list a few project ideas. You are welcome to define your own research project, which should be relevant to the contents of the class. It is completely OK to apply algorithms covered in class on non-biological data. The main thing is that your project must have "network" component to it. For any of the projects below, you are interested to do the project but not sure about the dataset, please contact Prof. Roy.

1 Comparative analysis of different integrative network learning algorithms

The goal of this project would be to compare how well different strategies for integrating prior knowledge into network inference works. For example, one can compare the Dependency network with priors to the Bayesian network approach where one imposes priors on the graph structure. This project can go in many directions. An example might be to implement the MCMC approach for multiple priors and see if the priors can be learned reliably.

2 Multi-task learning of multiple context-specific networks

We will soon be covering a paper which talks about learning multiple Graphical Gaussian Models to learn tissue-specific networks. See `https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004220` for details. Implement this algorithm and apply it to additional gene expression datasets, for example different cancer datasets available from TCGA.

3 Identification of signature subgraphs

A key problem in analyzing gene expression measurements across multiple conditions is to identify differentially expressed genes between two conditions. However, since genes don't function independently, an important analysis task is to identify if there is a subnetwork that is differentially active in one condition versus another. One approach to address this problem is to integrate gene expression data with an underlying molecular network by coloring the nodes based on the expression levels and then clustering the graph based on the network and the expression values. Possible extensions could be to apply this approach to $k > 2$ conditions. See `https://academic.oup.com/bioinformatics/article/18/suppl_1/S233/232152` and `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2063581/` for some ideas.

4 Graph diffusion for smoothing

An important type of dataset that is emerging is called Hi-C, which measures the genomic contact count of two regions. You can think of the Hi-C matrix entry telling us how likely is one region to interact with another region. The output from a Hi-C experiment is often very noisy and sparse. Graph diffusion based methods could be used to "smooth" this matrix. See for example `https://academic.oup.com/bioinformatics/article-abstract/34/16/2701/4938489?redirectedFrom=fulltext`. Apply graph diffusion to smooth a given Hi-C matrix and compare how it works to simple Gaussian smoothing or fixed window averaging. The smoothed matrix can be further analyzed to define clusters of interacting regions. Compare the clusters to those obtained pre- smoothing. This project could be increased in scope by considering other types of data such as scRNA-seq matrices or microbiome data.

5 Inference of cell-cell networks

This is a network inference problem however, instead of nodes as genes, these networks connect cells. For each cell we have a vector of gene expression values and we are going to infer the cell-cell network using the measured values of genes as joint assignments. Apply at least three network inference algorithms to infer cell-cell networks and compare the results. This can be extended to further cluster the network using graph clustering approaches.

6 Inference of region-region networks

This is a network inference problem, however, instead of nodes as genes, these networks connect arbitrary genomic regions using their activities measured across many cell types and many different regulatory signals. Such regulatory signals could be transcription factor binding or histone marks. The goal of this project is to predict which genomic region "interacts" with another genomic region based on the predictive/statistical dependence between these two regions. All algorithms that don't use priors are valid choices to examine this problem.

7 Graph diffusion and clustering to integrate different types of genomic data

Often times for the same biological system two complementary types of measurements are made. For example, for a given tissue one can measure protein and mRNA levels, or for a cancerous sample, one can measure genotype data as well as gene expression. Further more one can collect these data for multiple samples or subjects. This project would entail using diffusion and clustering approaches to integrated different complementary measurements.

8 Comparison of graph clustering algorithms

This project will entail performing a systematic comparison of different graph clustering algorithms. An interesting paper to check out for this is `https://www.biorxiv.org/content/early/2018/02/15/265553`.