Week 1 Worksheet:

Personal and Professional Goals:

What parts of scientific research interest you the most?

Why are you interested in computational/mathematical biology?

What skills/knowledge do you want to gain out of your summer research experience?

What are your current goals for your career?

How can we best help you achieve these goals?

Project Overview:

miRNAs are short non-coding RNAs (i.e., they are molecules in the cell that are not involved in creation of proteins) that function within the cell to post-transcriptionally regulate gene expression in the cell by degrading mRNAs (see figure 1). Note: miRNAs have other functions too, but mRNA degradation is their most common function. By degrading mRNAs within the cell, the miRNA prevents certain proteins from being translated, which changes how the cell functions.

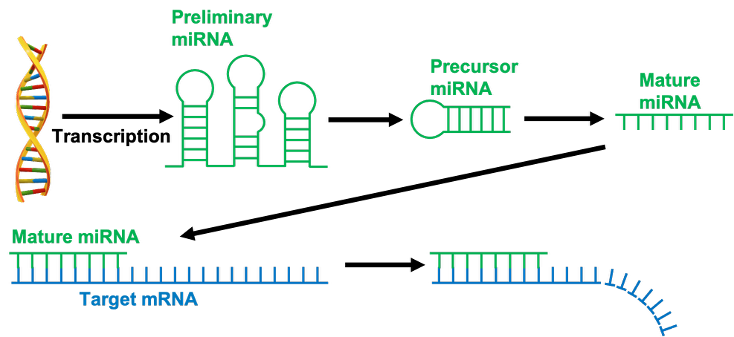


Figure 1: The (simplified) biological process of a miRNA. The miRNA is transcribed from the DNA of the organism into a preliminary miRNA, which is then modified into a precursor miRNA and finally into a mature miRNA. This mature miRNA binds to an mRNA that has a complementary sequence (Note: in animals the miRNA does not need to be fully complementary). After the miRNA binds to the mRNA, it causes the mRNA to degrade.

Because miRNAs are much shorter than mRNAs (about 20 nucleotides in miRNAs vs hundreds to thousands in mRNAs), an individual miRNA can target many different mRNAs. Similarly, each mRNA can be targeted by many different miRNAs. Combining all of this information can quickly result in a complex network of miRNAs and mRNAs interacting with each other (see figure 2).

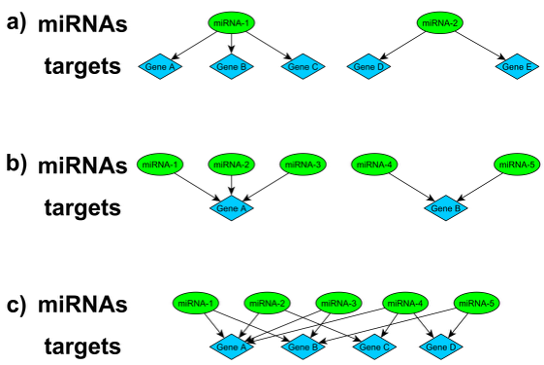


Figure 2: Interactions between miRNAs and mRNAs (i.e., genes). A) Each miRNA can target multiple genes. B) Each gene can be targeted by multiple miRNAs. C) If we combine these effects, we get a bipartite network (i.e., graph) of miRNAs and genes interacting with each other.

**The overarching goal of this research is to better understand how the miRNA functions within the cell through studying their interactions with mRNAs in various ways.**

Summer Project Goal: How can we best summarize a set of significant functions to understand the functions of miRNAs within a network community using a machine learning approach?

Background: To assess the functions of miRNAs within a system, we construct a bipartite miRNA-mRNA network consisting of miRNAs and the mRNAs they target. One of the biggest hurdles in understanding miRNA biology is that a miRNA has context-dependent function. That is, the targets of a miRNA change depending on many different factors that describe the system (e.g., environmental condition, location of tissue in body). To understand how the context affects the function of miRNAs, we have constructed separate bipartite networks using data from different cancer types. These networks typically have the same miRNAs, but the miRNAs have very different targets. We want to study the similarities and differences between the miRNAs and their functions in these networks.

To identify the functions of miRNAs within each network, we separate the network into sub-network communities, and then find functions by identifying which functions have high amounts of overlap between the genes that perform the function and the genes in the community (see figure 3). Once all of the functions with high overlap are identified for a community, we build a function graph (see figure 4) that connects two functions if they share a lot of mRNAs. We can separate this graph into different communities (the different colors) to find functions that are more strongly related to each other (i.e., we find which functions have very similar genes). **We then manually summarize all functions of a specific color into a single function (this is the part of the process we want to automate).** For the community analyzed in figure 4, we identify 5 different communities, and thus 5 different functions for this community (Glycolysis, Phosphatase, Cell Cycle, Metabolism, and RNA Splicing). We also rank the miRNAs within this community to understand which miRNAs are the most strongly linked to these identified functions.

Note: For more information on this process, see the miRNA-mRNA manuscript draft that I sent you.

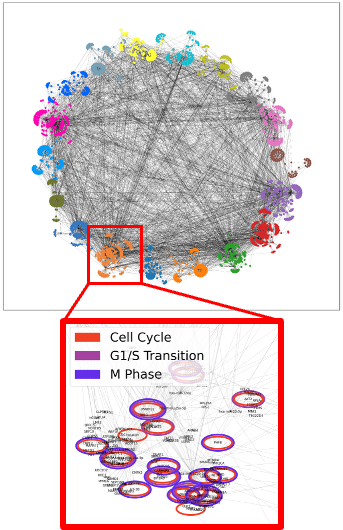
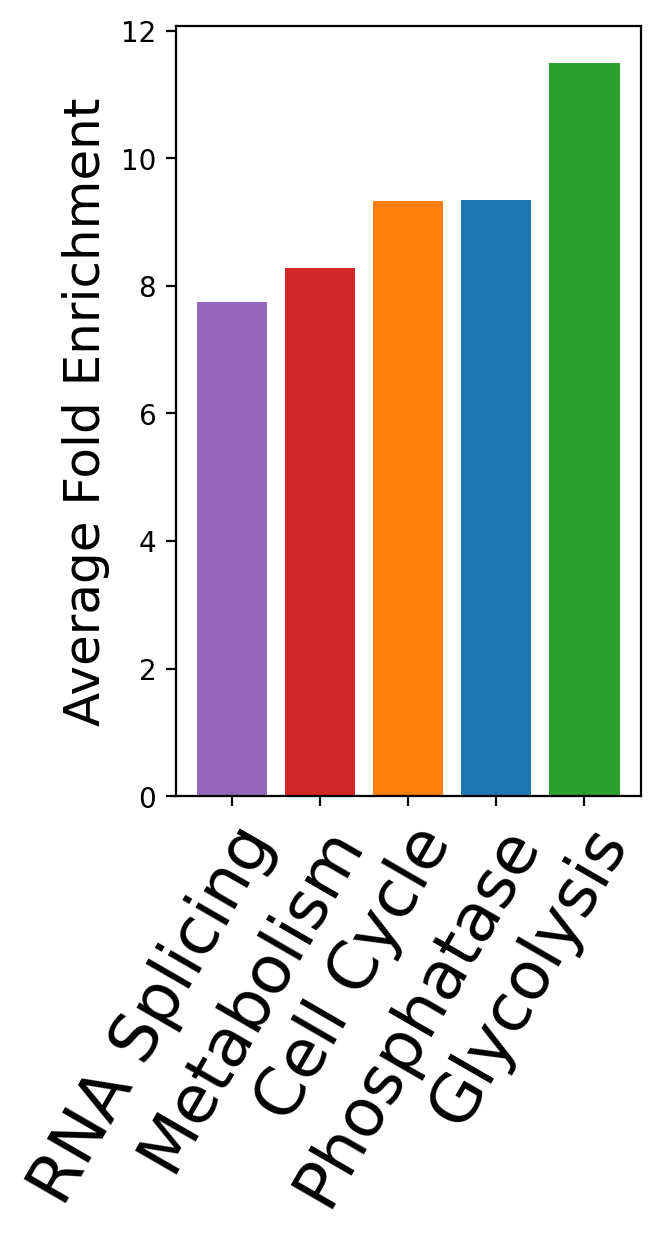
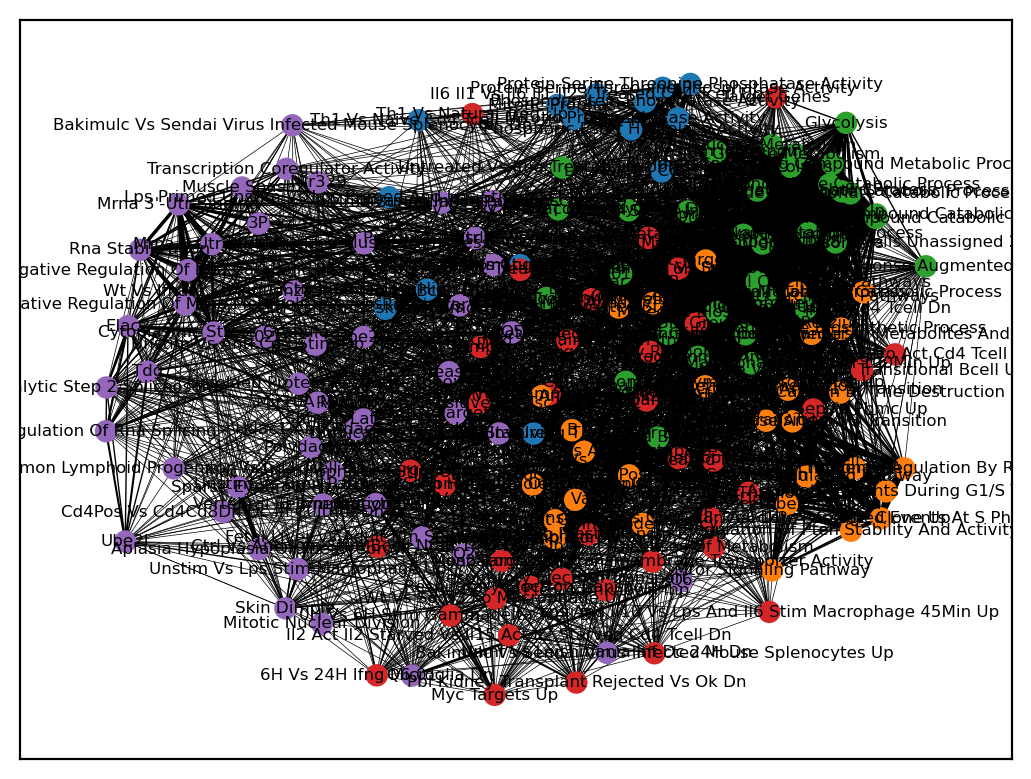


Figure 3: Breast cancer network separated into individual communities. By breaking the network into smaller communities, we are able to better understand the functions of individual miRNAs. To identify functions, we find which functions have high amounts of gene overlap with the genes in a community. For example, we show that the genes that drive the cell cycle, G1/S Transition, and M Phase have high amounts of overlap with the zoomed in community, indicating that the miRNAs in this community likely drive the cell cycle.



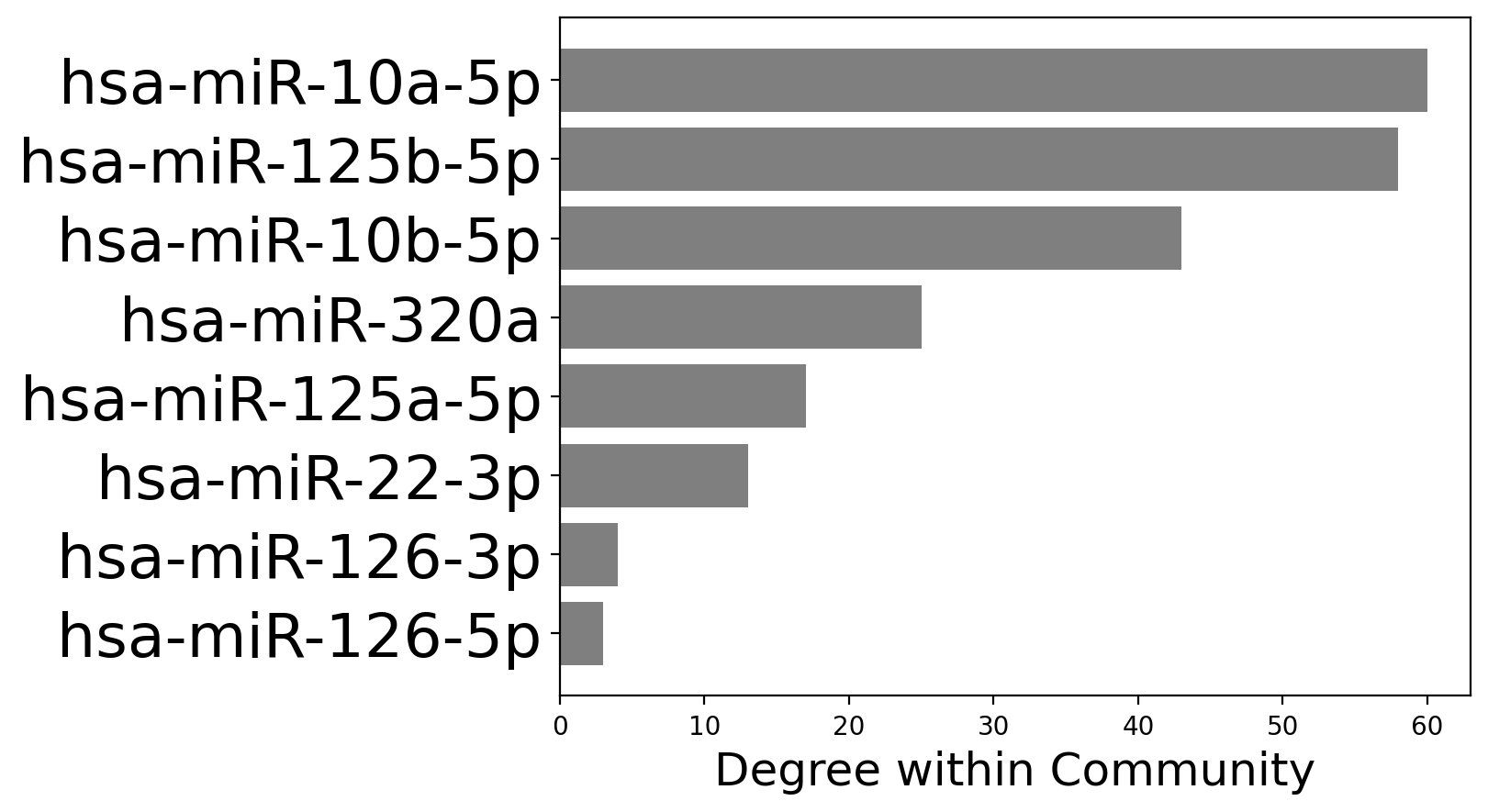


Figure 4: Once functions are identified for a single community, we then need to summarize these functions to determine the overall functions of the community. We do this by constructing a function graph (top left), where the identified functions are connected if they share genes, and the strength of the connection depends on the number of shared genes. We cluster this graph to identify communities of more strongly connected functions (each color), and then summarize the functions of each color into a single overall function (in this case, the green functions correspond to glycolysis, the blue functions correspond to the phosphatase, and so on), which we show on the right. On the bottom, we show the miRNAs within this community ranked by their influence over the community, so we can determine which miRNAs have the largest effect on these functions.

**Week #1 Goals:**

* Familiarize yourself with the current data and functional annotation methods
* Use DAVID to manually recreate previous result
  + In the DAVID Bioinformatics Functional Annotation Tool (<https://davidbioinformatics.nih.gov/>), upload the gene sets from the breast cancer network, and then select only the following annotation categories:
    - Under Functional\_Annotations:
      * UP\_KW\_BIOLOGICAL\_PROCESS
      * UP\_KW\_MOLECULAR\_FUNCTION
    - Under Gene\_Ontology:
      * GOTERM\_BP\_DIRECT
      * GOTERM\_MF\_DIRECT
    - Under Pathways:
      * BBID
      * BIOCARTA
      * KEGG\_PATHWAY
      * REACTOME\_PATHWAY
      * WIKIPATHWAYS
  + Familiarize yourself with the output from each of the buttons under the Combined View for Selected Annotation section.
    - These results are explained in the manual (<https://davidbioinformatics.nih.gov/content.jsp?file=functional_annotation.html#summary>)
    - Note: Our analysis does the analysis with both David and MSigDB (<https://www.gsea-msigdb.org/gsea/msigdb/>) pathways, so you will not get the full output from only looking at DAVID
  + Do this manually once or twice to better understand the analysis. The code I sent you does this automatically.
* Get previous code working on your machine (see data/code section below). To get everything to be running correctly, you will probably need to download some packages and do some editing to the paths for them to be applicable to your machine.
* Look at BRCA dataset (this is the output file from the [graphToFunctions.py](http://graphtofunctions.py) code)
  + These are the datasets that will be inputs for your analysis, which we need to summarize
  + For each module (the various sheets of the excel document), we have a set of lists of functions (in the functions column), and the average fold enrichment of the lists of functions (in the Average FE column). We want to convert each list of functions (i.e., each cell in the functions column) into a single function.
  + For the BRCA network, compare some of these lists to the manually determined functions in Table 1 of the manuscript.
* Begin brainstorming ways to summarize function lists into a single function.

Data/Code:

The code titled [graphToFunctions.py](http://graphtofunctions.py) is Python code that takes as input the name of one of the generated graphs, and outputs an excel file (titled “{cancerName}\_moduleFunctions.xlsx”) that lists the identified functions for each community of the network. The only input file required for the [graphToFunctions.py](http://graphtofunctions.py) script is the network in the form of an edgelist saved as a file “{cancerName}\_EdgeList2.txt” (I have provided you with these edgelists). The output excel sheet has two columns of data for each community: one has the list of functions in each of the clusters of the function graph, and the other has the average fold enrichment of these functions. Note: this code outputs some other files that can be used for intermediate analyses, but the “{cancerName}\_moduleFunctions.xlsx” is the final output file that contains all of the data that you will need. I have been running this code to generate all of the output files that you will need, but I am providing it to you, so you can reference it in case you have any questions/need any of these functions for your project.

To run the [graphToFunctions.py](http://graphtofunctions.py) you will also need to install R on your device because one of the functions is executed in R instead of Python. R can be installed here: <https://cran.r-project.org/>. It will also require the Python package rpy2 to be installed (<https://rpy2.github.io/doc.html>). This package allows us to run the analyzeModuleFunctions.R script within python. If you would like to look at/edit this script I would recommend also downloading RStudio (<https://posit.co/download/rstudio-desktop/>). The R script also reads in the functions that are compared to the communities, so I have provided you with these files too. All of these files are .gmt files, and they consist of lists of functions that are the function’s name followed by the (entrezID of) genes associated with that function.