 Due by 2:30pm, 24^{th} Feb 2020

CMPSC 301
Data Analytics
Spring 2020

Lab 3: Exploring Gene Expression Data With R

Objectives

To enhance the understanding code for data gene expression analysis in Bioinformatics. In this lab, you will learn to read basic plots and to understand the code that produced the plots.

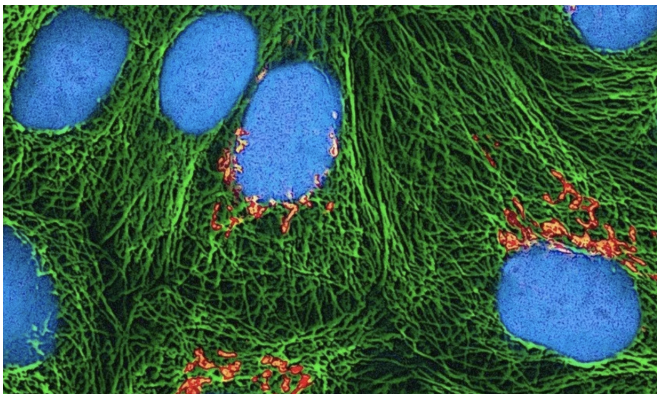

Figure 1: The NCI60 data set was was developed by the National Cancer Institute in the 1980's to allow for the study of potential anti-cancer drugs. Concerns probes, which are used to detect molecular compounds, monitor cellular activity, and to deliver therapy in cells, the data allows researchers to investigate potential anti-cancer drugs in a wide variety of cell lines which have been organized (i.e., arrogated) by types of cancer

Reading Assignment

Please read Chapters assigned for this week's lessons which you will find in the class slides, in addition to reviewing your notes. Please take some time to gain experience with using Markdown to write your work. See *Mastering Markdown* `https://guides.github.com/features/mastering-markdown/` for more details about Markdown. In addition, you may consult your notes on the talk given by Dr. Thu of the Department of Biology who discussed motivations and methods for the study of gene expression data.

Part of the motivation for this lab may be found in the YouTube video that discusses the gene expressions of Planarian worms at the following link: `https://www.youtube.com/watch?v=roZeOBZAa2Q`. Much of the knowledge that we have about the curious genetics of the stem cell function in these organisms was the result of gene expression analysis. Interested readers are invited to learn more about the biology of stem cell growth and regeneration in Planarians from Sánchez *et al* [1].

Handed out: 10^{th} February 2020

GitHub Starter Link

https://classroom.github.com/a/v0PNt4yQ

To use this link, please follow the steps below.

- Click on the link and accept the assignment.
- Once the importing task has completed, click on the created assignment link which will take you to your newly created GitHub repository for this lab.
- Clone this repository (bearing your name) and work on the lab locally.
- As you are working on your lab, you are to commit and push regularly. You can use the following commands to add a single file, you must be in the directory where the file is located (or add the path to the file in the command):
  - `git commit <nameOfFile> -m ‘‘Your notes about commit here’’`
  - `git push`

  Alternatively, you can use the following commands to add multiple files from your repository:
  - `git add -A`
  - `git commit -m ‘‘Your notes about commit here’’`
  - `git push`

Instructions

When undergoing a project in data science, there will be computer code to write to manage data, calculate statistics and produce plots. Although this computer code provides the evidence of underlying trends in the data, the causality of any observed patterns is largely unknown. In fact, spotting trends is the task of the analyst who must then decide how to understand them, and decide whether a result is actually a relevant pattern to study. Further decisions must be made, based on this learning, to determine the next courses of action: to explain how the trends became apart of the data.

Understanding the R Code and Making Sense of the Output

Imagine that you are discussing the code and its output with someone who is new to the project. It is your task to explain to them what each part of the code is doing and what knowledge it brings. In deliverable you are to explain in clear and meaningful language *what the code is doing and how to make some sense of its output.*

You have been given an R source code file `src/geneExpAnalysis.r` which you are to run using rStudio. The questions to which you are to respond, are also listed below in this assignment in blue.

Questions in Blue

1. What does gsub() do in this code block?
2. What does the plot on the next line show?
3. Plot 2 is a scaled view of the same information from Plot 1. Why is scaling helpful when viewing these results of "Plot 2 ?
4. What is the function "mutate()" doing in this code block?
5. What is the function "summarise()" doing in this code block?
6. Interpret the results of Plot 3.
7. Interpret the results of Plot 4.
8. In Plot 5, both of the previous plots have been combined together on the same canvas. How does this, either help or confuse the output of these lines of code?
9. What do the whiskers (i.e., the big "I" shaped structures above each histogram bar) convey in this plot? Why might we need to know this information?
10. The word, "function" is below in the code block. What does this part of the code do?
11. Some time is necessary to generate the variable, "ProbeR2". What is taking this code block so long to complete? What is the "lm()" function in R?
12. With the above ideas about R-squared values in mind, which Probe in Plot 8 appears to be able to best explain the model?
13. In Plot 9 we see two curves. What can you infer about the relationships between the Actual cases of relationships and those cases for which no relationship exists? Remember, the blue curve corresponds to the variable "r2" and the red curve corresponds to the Null values.
14. Plot 10, is very important to determine a result concerning the Actual data sets and the Null data sets. How do we use this plot to conclude that the 30 probes in Plot 8 are relevant results (i.e., these cases demonstrate relationships)?

Ethics

The response to the following questions are to be included at the end of your `writing/report.md` submission.

Medical Records

An electronic health record (EHR) (or a medical record) is a report in which doctors store extremely private information concerning a patient's medical details. Some of these details are the patient's medical history, a physical examination information, previous investigations and past treatments.

Often this information is encrypted to prevent non-authorized peopls from viewing it, however, technology may fail and the information may be leaked.

Discussed in Ozair *et al.* [2], (link: `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4394583/`) are some of the dilemmas in ethics that arise from the use of medical records. In their article, the authors wrote that,

> *When patient's health data are shared or linked without the patients' knowledge, autonomy is jeopardized.*

The below questions concern this premise.

1. Discuss what is meant by the term *autonomy*.
2. What can go wrong in technology to create a such a leak that they mention.
3. The authors mention that paper records could only be viewed by one person at a time. In this case, each medical record would have to be check-out at the medical records room in a hospital. In your opinion, do you think that this would be a better form of security for the records, even though the process of getting to the records would be extremely slow for the doctors who require this information?
4. The authors also mention that, *Clinical personnel often have little knowledge of the clinic's workflow and the roles others play in care delivery.* How could this lack of coordination be an ethical problem?

Required Deliverables

1. File `writing/report.md`: Your **labeled** answers to the questions in blue (included in this assignment sheet and also in the source code itself `src/geneExpAnalysis.r`
2. Your response to the ethical question is to be included at the end of your `writing/report.md`:

When you have finished, please ensure that the GitHub web site has your pushed work by visiting your repository at the site. Please see the instructor if you have any questions about assignment submission.

References

[1] P. W. Reddien and A. S. Alvarado, "Fundamentals of planarian regeneration," *Annu. Rev. Cell Dev. Biol.*, vol. 20, pp. 725–757, 2004.

[2] F. F. Ozair, N. Jamshed, A. Sharma, and P. Aggarwal, "Ethical issues in electronic health records: A general overview," *Perspectives in clinical research*, vol. 6, no. 2, p. 73, 2015.