Grand Valley State University
CIS 635 – Knowledge Discovery and Data Mining
Project 1
Instructor: Guenter Tusch

Overview of this Project

1. Article
2. Project description
3. Deliverables and timeline
4. Additional information

1. Article:

The project is based on the following article:
Xu G, Zhang M, Zhu H, Xu J. A 15-gene signature for prediction of colon cancer recurrence and prognosis based on SVM. Gene. 2017 Mar 10;604:33-40.

2. Project description

The authors analyze two data sets to find a gene signature to classify patients with high risks from those with low-risks for colon cancer recurrence. They use a support vector machine (SVM) model for the prediction. You are expected to perform only the microarray analysis and SVM. You might try the PPI using, e.g., Genemania (5 pts. extra credit), not the Cox model. You can use GEO2R (see lab 2) for the first part of the project. The article mentions five microarray datasets of colon cancer samples from the Gene Expression Omnibus database and one that was obtained from The Cancer Genome Atlas (TCGA), but you'll use only GSE17537 and GSE17538 from NCBI GEO. The GSE17537 analysis is using the Linear Models for Microarray data (LIMMA) to identify the differentially expressed genes (DEGs). The DEGs are used for PPI network-based neighborhood scoring and support vector machine (SVM) analyses, which are validated by using GSE17538. Cox regression analysis is not part of this class and not required.
The authors identify a total of 1207 genes as DEGs between recurrence and no-recurrence samples, including 726 downregulated and 481 upregulated genes. Using SVM analysis and data validation, 15 genes (HES5, ZNF417, GLRA2, OR8D2, HOXA7, FABP6, MUSK, HTR6, GRIP2, KLRK1, VEGFA, AKAP12, RHEB, NCRNA00152 and PMEPA1) are identified as a predictor of recurrence risk for colon cancer patients.

Your task is to repeat the analysis and confirm the 15 genes. Use the same R and Bioconductor models.

Hints: You'll probably notice that there is a discrepancy of the reported number of DE genes in the paper (>1,000) and the actual number you'll find, if you apply the authors' methods. Since they couldn't use the FDR (no genes would be selected), the authors chose the method that is employed for volcano plots: p<.05 AND (FC > 2 OR FC < ½)
Note: they use abs(log(FC)), therefore the threshold is log(2).

To apply the method, you need to be sure to get all genes with p < .05 (toptable with 250 in the R code doesn't get all, you need to increase the number significantly, closer to 50,000).

When I ran the code, the number of the DE genes that I found was closer to 109. Then there is no need to reduce that number again as they did in the paper (to 100). Just stay with that number and create the SVM. Use the SVM we used in lab 3 (unless you definitely want do variable reduction or recursive feature elimination).

If you want use Genemania as in lab 2 (since we are not reducing variables here, it is not really necessary), use the same mechanism as in lab 2 to copy and paste the genes into Genemania. You can then download the Interactions Data (see fig 1.)

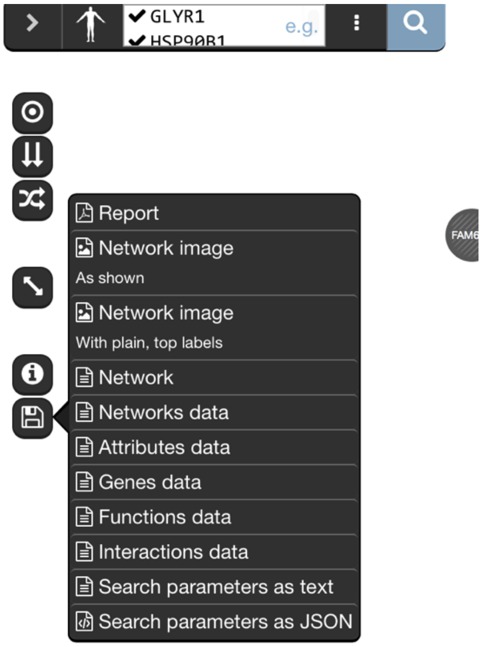

Fig. 1

You could use the maximum weight in the co-expression group to rank all genes similarly as the authors did with their score and reduce the number of genes that way.

3. Deliverables and timeline:

1st deadline: Tuesday, January 30

- Project Success Chart
- Work Breakdown Structure
- Task Assignment Matrix *(who is doing what, each student has to prepare* ***at least one third*** *of the paper and write about half of the code by himself)*.
- Task duration table
- Gantt (bar) chart schedule (see BB).

2nd deadline: Friday. February 23

- Project report (750-1,500 words)
- R script with complete code
- Team member evaluation

4. Additional Information:

Here is a possible **outline** of the paper:

1. Abstract
2. Introduction - A short statement that presents the purpose of the paper.
3. Background - Motivate and place the work in context. Subsections might address the significance of the problem being addressed, other approaches to the problem, and prior work on the approach being reported.
4. Formulation process - the biological experiment design.
5. Model description – describe the statistical or machine learning model
6. Analysis – describe the authors statistical/ML analysis, and what you did different and why.
7. Results – Report your results and potential differences to the authors' results.
8. Discussion - Subsections might address significance, limitations, and problems with the analysis.

For grading criteria see BB.

Make sure that you understand the clinical (or biological) problem in the selected publication and the methodology completely. You might need additional resources besides the textbooks. It is assumed that you are familiar with the use of the library and the online resources. Personal in the library can assist you with identifying resources. Some additional resources are listed on Blackboard. Your paper can only be as good as your understanding of the material. Your research paper needs to explain the clinical problem and its importance as well as the technology involved. Focus on the important parts that are needed for the understanding of the problem. Don't state trivial aspects. You can assume that the reader of your paper took the class and is familiar with its basic concepts. Be specific and concise. Conduct on the graduate level is assumed including that you take full responsibility for the entire process. Start immediately with the project, there might be unexpected problems in the process. An extension of the deadline cannot be granted. See the instructor early in the project; by end of the project office hours are very crowded.

References need to be a trusted source. Although the web is a common information resource for student research papers, and students frequently believe that the web contains only accurate information, finding reliable digital references is not as easy as selecting a book from a library shelf. Though highly popular, Wikipedia is not considered a trusted source for research references. The majority of your references should be books and articles from peer-reviewed scientific publications like those you find in PubMed (http://www.ncbi.nlm.nih.gov/pubmed/) or the GVSU library website.

Your research paper should be in correct English and free of spelling and grammatical errors, orphan and widow lines and headers, etc. The writing style should be consistent with that of a scientific paper. You can either use passive voice or “we”. If you refer to the publication, use “the authors state” or similar.

Code requirement:
You should create a well-documented R script file that I can execute and run with the supplied data set after I set the working directory according to the following criteria:

- You must write your source code from scratch. You may not copy code from another file or use another file as a starting point except code generated by GEO2R. You must type all the source code yourself. The documentation needs to include the author’s name.
- All work is to be done independently. You may not copy any code from any other student or Internet resource unless you give them credit. That credit will be deducted from your grade.
- R code will be graded on compilation, execution, structure, and documentation. Points will only be awarded for correct work. The R code must run without any errors. The program must perform as described in the assignment. The R code must be properly documented and written in a well-structured manner. See also General Grading Criteria for Programming Projects on BB.

Have fun!