Algoritmos Bioinformática /Bioinformática 2021/2022

Assignment 6

Multi-omics Graph Analysis

Patient Networks from Genomic Data

The goal of this project is to create patient networks derived from different types of molecular data and compare its similarities and differences. We will use data from the TCGA project, in particular from patients with Lung Squamous Cell Carcinoma: https://portal.gdc.cancer.gov/

Squamous cell carcinoma (SCC) of the lung is a form of lung cancer. It is more common in men than in women. It's usually caused by smoking tobacco. Lung cancer is one of the most common types of cancer.

We will analyze 106 samples from three datasets (data types) containing:
gene expression, methylation and miRNA data from these 106 patients.

When compared with previous assignments this is a more exploratory project where you will be able to investigate different approaches to the problem and propose your own solutions. In this case I will be table to provide little feedback on the work!

Framework

The terms graph and network are used in an equivalent way throughout the text.

You can make use of the code provided in the class or in alternative you can explore and use the Python package NetworkX:

- https://networkx.org/
- https://networkx.org/documentation/stable/_downloads/networkx_reference.pdf

You can also use other standard packages (please indicate them in the report) for plotting, table management and machine learning including:
pandas, sklearn, matplotlib, seaborn, numpy.

Data

The data was obtained from TCGA and used in a previous study [1] (see Lung.zip).

It corresponds to Lung cancer data and contain three files:

- Gene expression data (12042 x 106) - LUNG_Gene_Expression.txt
- Methylation data (23074 X 106 - LUNG_Methy_Expression.txt
- MicroRNA data (352 x 106) - LUNG_Mirna_Expression.txt

What is DNA methylation? See [3]
What are miRNAs? See [4]

The columns in the matrices contain the sample identifiers (patients) and the rows are the features (genes, miRNAs, methylation probes).

The identifier of the sample or patient is given by the first three elements of the sample name:

In LUNG_Methy_Expression.txt the id:
TCGA.18.3406.01A.01T.0981.13

In LUNG_Gene_Expression.txt
TCGA.18.3406.01A.01D.0979.05

In LUNG_Gene_Expression.txt
TCGA.18.3406.01A.01R.1031.01

They all refer to the patient **TCGA.18.3406**. In order to associate data from the three datasets you will need to parse this data. and eventually change the labels in each dataset.

Task 1 - Patient-by-patient networks (12 values)

This task consists in creating 3 independent networks that express patient similarity.
Each patient (column in the matrix) is represented by vector of molecular of values (12042 values in gene expression, 23074 in methylation and 352 in miRNA). Using correlation measures (Pearson and Spearman) build a matrix of patients (pairwise correlation between every patient), where each cell $C_{i,j}$ corresponds to the correlation between patients *(i,j)*.

1) Compute Patient-by-patient similarity matrix for each of the three datatypes. Hint: Pandas offer a *corr()* function to compute the correlation matrix.

2) Build the Patient similarity networks for each datatype. Use a weighted network where each edge is weighted by the correlation between its two nodes. In the correlation matrix, we compute all pairs of patients. We can define a threshold for which a connection can be considered. Two patients *(i,j)* are connected in the graph: if |corr(i,j)| >= X. Thus, in the graph only edges corresponding to a correlation greater than X are considered.
Hint: use the NetworkX package.

3) Plot the matrix before and after the threshold selection. Suppose you select X = 0.3 as the threshold. How does the matrix looks after and before the cut-off.

4) The optimal value of X can be studied. For instance, plot several graph statistics with relation to the value of X. Create a table with different statistics (e.g. number of nodes and edges, average degree, average clustering coefficient, et...) in the columns, row as values of X for each of the networks.

5) Considering a threshold value of X = 0.2 in the $|corr(i,j)| > X$ test, indicate which type of network (random, scale-free or hierarchical) is most similar to the network obtained from each data type.

Task 2 - Merging data (4 values)

The Similarity Network Fusion (SNF) is a computational method for data integration [1,2]. It allows to create a fused network from several networks obtained from different datatypes of data (methylation, expression, miRNA, etc). In this task, you will be able to explore this method. You can use either the Python [6] or the R implementation [5].

The idea is to create a fused network from the existing networks. Once this network you can apply other data analysis. Clustering is an obvious choice. You can try to use some methods to infer the best number of clusters, and plot them. The goal is then to try to find a particular clinical characteristic on the patients within each cluster. You can inspect the files lusc.clinical.txt and Lung_survival.txt to see if there are significant differences for the available features in these two files in the different clusters. In particular, in the features, *age_at_diagnosis, pathologic_stage and Survival*.

Note that this step is more exploratory. You may need to look out in the documentation and literature further details to complement this analysis.

Task 3 - Report (4 values)

This task evaluates the organization and the quality of the report.

You need to deliver a report based on a presentation. You will have a slide for each of the above points (e.g. 5 slides for Task 1), plus:

- presentation slide - elements of the group and description of the assignment.
- methods slide - additional slide where you can describe some of the methods used and particular choices for the methods and implementation.
- conclusion slide - key points on what you learned (or not) from this work and things to do as possible work extension.

Per slide you can use figures, composition of figures and text. Try not each slide with too much information.

Deliverables

1) zip file with the code for the above implementation.
2) pdf file with the presentation exported to pdf format.

References

[1] - http://compbio.cs.toronto.edu/SNF/SNF/Software.html
[2] - http://mghassem.mit.edu/wp-content/uploads/2015/06/nmeth.2810.pdf Similarity network fusion for aggregating data types
on a genomic scale, Nat Biotech 2014.
[3] - https://www.youtube.com/watch?v=W-S84J4zK9E
[4] - https://www.youtube.com/watch?v=h4t-fhvAorA
[5] - https://cran.r-project.org/web/packages/SNFtool/SNFtool.pdf
[6] - Python implementation of SNF: snfpy by Markello
https://github.com/rmarkello/snfpy#:~:text=The%20similarity%20network%20generation%20and,retained%20via%20the%20fusion%20process.