Databases and Data Mining 2018
Data Mining Assignment 2

Topology based Protein-Protein Interaction Network Analysis for Prioritization of Candidate Genes associated with Menière's Disease.

Due: Monday 17-1 2018

Grading: This assignment will be graded from 0 to 10

Notes:

- Groups of 1-6 students are allowed.
- Write down your report for this assignment in a .pdf file with the following name “<your student number><your name>-dm02.pdf”, e.g., “012345janjansen-dm02.pdf”, or “012345janjansen-678910-jansjansen-dm02.pdf” if you are working in a team of 2, etc.
- Put your report and *C.txt* file together with any code you developed for this assignment in a zip-file (using the same naming convention as for the pdf) and send the zip-file as an attachment of an e-mail with subject *DBDM-DM02* to erwin@liacs.nl.
- Do not use more than 8 pages (A4, font size 11 pt) for your report.
- Grading will be based on
  - the quality of your PPI Network Analysis strategy and results, and
  - the argumentation, validity, and clarity of your report.

1 Introduction

Computational protein function prediction and identifying novel genes associated with diseases are a very important and actively studied research area. Recent methods use techniques such as homology modeling (for related proteins), sequence analysis and analysis of weighted and annotated Protein-Protein Interaction (PPI) networks [2-5].

In this assignment a PPI network *N* of Homo Sapiens is modeled as a graph where each node is a protein and each edge is a physical interaction between two proteins. The links are weighted with a score indicating the strength of the interaction. Furthermore a gene set *S* is given, corresponding to 106 proteins that are known to be associated with Menière's Disease (Li et al. [4]). In [4] a set *G* of 43 novel genes were identified that probably are associated with the disease also, 15 of which were biologically validated (set *V*).

The task of this assignment is to identify and prioritize novel candidate genes associated with Menière's Disease using only the weighted topology of the PPI network *N* and the set *S* of known proteins associated with the disease as input information. Using these input data an ordered set *C* of novel candidate genes in *N* have to be identified that may be associated with Menière's Disease. The ordering should reflect the likeliness (from high to low) of a novel gene being associated with the disease. Furthermore, *C* should have size equal to 100, and contain as many of the novel genes (gene set *G*) found in [4] as possible.

2 Input Data

The following input files should be used when executing your proposed algorithms.

1. The STRING database (string-db.org) [6] is a well-known public database for high quality weighted and annotated PPI Networks of thousands of organisms. In this assignment we use the PPI Network for Homo Sapiens which is represented in the file *9606.protein.links.v10.5.txt.gz* (65.9 MB). The file contains the protein network data as scored links between proteins. It can be downloaded by using the following link and subsequently selecting the organism 'Homo Sapiens': https://string-db.org/cgi/download.pl?sessionId=2FWULbtZb2d4&species_text=Homo+sapiens
2. Each PPI contains two proteins, represented by Ensembl IDs, and a score that indicates the strength of the interaction. A larger score means that the interaction is more likely to occur.
3. At https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0182592 [4] in the supporting information section table S1 containing the Ensemble IDs of the 106 genes associated with Menière's disease can be found. You can cut and paste the list from the given document.

3 Evaluation Data

Your proposed algorithms should generate an ordered list *C* (of size 100) of prioritized candidate genes associated with Menière's Disease. The quality of your algorithms should be evaluated using the following data:

1. At https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0182592 [4] in the supporting information section table S4 with the Ensemble IDs of the 43 inferred genes that are probably associated with Menière's disease can be found.
2. In Table 1 of the report by Li et al. [4] the 15 biologically validated novel genes associated with Menière's disease are listed.

You should report for each of these genes if they are listed in *C*, and if so, which rank (position) it takes.

4 Report

- Li et al. [4] and Wu et al. [1] could be good starting points for exploring reasonable methods.
- It is allowed to use and/or modify any random walker, node-classification algorithm or machine learning tool (e.g., PageRank, RWR, MatLab, Python-libs, R-scripts, etc.). Be sure to list the appropriate references!
- The final report (one single pdf) should be structured as a scientific paper, containing, amongst others, the sections "Abstract", "Introduction", "Methods", "Evaluation", "Conclusions" and "References".
- All the code/information used in the proposed algorithms should be described. Use a formal writing style. Code that you developed for the assignment should be added to the zip-file.
- The identified and prioritized 100 novel gene candidates that may be associated with Menière's disease should be given as an ordered list of Ensembl IDs in a file *C.txt*.

References

1. Xuebing Wu and Shao Li, *Cancer Gene Prediction Using a Network Approach. Chapter 11 Mathematical and Computational Biology*. Cancer Systems Biology (Ed. Edwin Wang). Series: Chapman and Hall/CRC, 2010.
2. Renzhi Cao, and Jianlin Cheng, *Integrated protein function prediction by mining function associations, sequences, and protein–protein and gene–gene interaction networks*, Methods, Vol. 93, January 2016, pp. 84 – 91. (Available here: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4894840/)
3. Predrag Radivojac et al., *A large-scale evaluation of computational protein function prediction*, Nat Methods, Vol. 10(3), March 2013, pp 221 - 227.
4. Lin Li, YanShu Wang, Lifeng An, XiangYin Kong, Tao Huang, *A network-based method using a random walk with restart algorithm and screening tests to identify novel genes associated with Menière's disease.* PLOSL ONE, August 7, 2017
5. Lei Chen, Jing Yang, Zhihao Xing, Fei Yuan, Yang Shu, YunHua Zhang, XiangYin Kong, Tao Huang, HaiPeng Li, Yu-Dong Cai, *An integrated method for the identification of novel genes related to oral cancer.* PLOSL ONE, April 6, 2017
6. Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, Santos A, Doncheva NT, Roth A, Bork P, Jensen LJ, von Mering C., *The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible.* Nucleic Acids Res. 2017 Jan; 45:D362-68