Databases Assignment 4
Challenge on
Predicting Cancer-Related Proteins
in Protein-Protein Interaction Networks

Hossein Rahmani, Erwin M. Bakker

LIACS

Due: Tuesday 31-1 2012
Grading: This assignment will be graded from 0 to 10

– **Notes**:
- Groups of 1-2 students are allowed.
- Write down your report for this assignment in a .pdf file with the following name " < your student number><your name>-4.pdf", e.g., "012345janjansen-4.pdf", or "012345janjansen-678910-ansjansen-4.pdf" if you are working in a team of 2.
- Put your report together with your *PredictionResultsTest2.txt* file in one 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-4 to erwin@liacs.nl.
- Do not use more than 8 A4 (font size 10 pt) for your report.
- Grading will be based on
  * the quality of your data mining strategy and results, and
  * the argumentation, validity, and clarity of your report.
- Do not distribute the dataset!
  * The dataset may only be used by members of your team, and only if you intend to participate in this challenge.

1 Introduction

We model the Protein-Protein Interaction (PPI) network as a graph where each node is a protein and each edge is a physical interaction between two proteins. There are two types of annotation information for each protein p in the PPI network. A 'functional annotation' shows the biological functions of p in the PPI network and a 'cancer-relatedness annotation' shows if p is involved in cancer or not. The task of predicting cancer-related proteins in PPI networks is trying to predict the new proteins that are involved in cancer.

In this assignment, you are asked to design, develop, and evaluate a strategy for predicting cancer-related proteins in PPI networks. A PPI network as described above is given in a dataset. You will be asked to evaluate your method with one test-set and give your predictions for a second test-set.

2 Input Files

The following files should be used for training your proposed algorithms.

1. *HumanPPI.txt*: In this file each row $< Prot_i, Prot_j >$ represents one edge (modeling a physical interaction) between two proteins $Prot_i$ and $Prot_j$ in the PPI network. Please note, the edges are undirected, this means every edge in the PPI network will be represented in the file twice, by $< Prot_i, Prot_j >$ and $< Prot_j, Prot_i >$.
2. *Functions.txt*: Here each row $< Prot_i, Func_j >$ adds the function $Func_j$ to the function set of protein $Prot_i$. Please be careful that, there exist some proteins in the dataset which are not functionally annotated yet.
3. *Cancers.txt*: Each row $< Prot_i >$ in this file indicates that protein $Prot_i$ is involved in cancer.

3 Evaluation

You should evaluate your algorithms based on the following files:

1. *Test1.txt*: This file contains proteins that are known to be involved in cancer (as given in the file *Cancers.txt*) or not. The file should be used to evaluate the performance of your proposed prediction algorithm. You should predict the cancer-relatedness of each of the proteins in the file *Test1.txt* using your methods/algorithms, and report the evaluation results according to the Precision, Recall and F-measures:

$$Precision = \frac{tp}{tp + fp} \tag{1}$$

$$Recall = \frac{tp}{tp + fn} \tag{2}$$

$$F - measure = \frac{2 * Precision * Recall}{Precision + Recall} \tag{3}$$

   where proteins involved in cancer are considered as the positive class, and tp, fp and fn denote the number of true positives, false positives, and false negatives, respectively.
2. *Test2.txt*: Determine for each protein p in Test2.txt file, if p is involved in cancer or not and report your predictions in a file *PredicionResultsTest2.txt*. For example if there are three proteins p_1, p_2 and p_3 in Test2.txt and your algorithm predicts just p_1 as a cancer-related protein, then *PredictionResultsTest2.txt* should be as follows:

   p_1, Cancer
   p_2, nonCancer
   p_3, nonCancer

4 Reports

- Wu et al., [1] could be good starting point for exploring the current methods.
- It is allowed to use and/or modify any node-classification algorithm or machine learning tool (e.g., WEKA, MatLab).
- It is only allowed to use the provided anonymized dataset, while the use of external information sources is prohibited.
- The final report (one single pdf) should be structered as a scientific paper, containing, amongst others, the sections "Developed Algorithms", " Evaluation Results of Test1.txt". Furthermore, all the code/information used in the proposed algorithms should be described. Code that you developed for the assignment should be added to the zipfile. The predictions for the proteins listed in the file *Test2.txt* should be reported in another separate file *'PredictionResultsTest2.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.