DATA 620 Final Project Proposal

Text Mining and Gene Network Analysis with Non-small Cell Lung Cancer

Youqing Xiang

11/29/2016

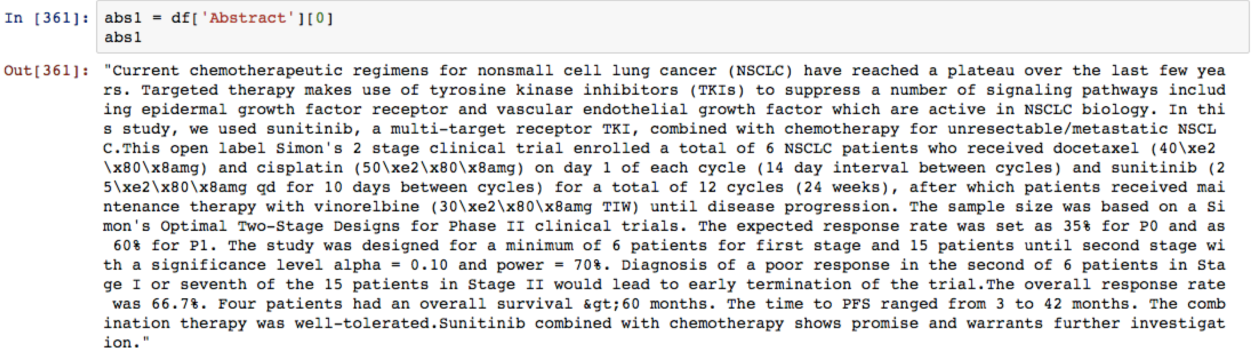
Introduction

Non-small cell lung cancer is the most common type of lung cancer. About 85% of lung cancers are non-small cell lung cancers (NSCLC) [1]. Different treatment options are available for people with NSCLC, including surgery, radiofrequency ablation, chemotherapy, targeted therapies, etc. Among them, targeted therapy is one of the most active medical research area. Targeted drugs work differently from standard chemotherapy drugs and they specifically target patient cancer cell changes. One of good example is the drugs that target cancer cells with EGFR changes, such as Erlotinib, Afatinib,etc. Biomedical researchers and drug companies are constantly searching for new genetic biomarkers which could potentially be the target for disease. New genetic biomarkers would bring new hope to the patients.

In this project, I will integrate text mining and network analysis for informative discoveries in NSCLC area. I will mine the available scientific papers' abstracts at Pubmed, which are relevant to the topic of NSCLC. And then I will extract co-occurrences of genes and build the network of gene-gene interaction. By analyzing the gene network, I will look for those important molecular markers relate to NSCLC. Further more, I will also search new and potential interesting genes or come up new bioinformatics research project ideas related to NSCLC.

Data

I got around 5000 of abstracts from Pubmed with `non small cell lung cancer` as searching keyword. And I show one of them as the following.



```
In [361]: abs1 = df['Abstract'][0]
          abs1
```

```
Out[361]: "Current chemotherapeutic regimens for nonsmall cell lung cancer (NSCLC) have reached a plateau over the last few yea
          rs. Targeted therapy makes use of tyrosine kinase inhibitors (TKIs) to suppress a number of signaling pathways includ
          ing epidermal growth factor receptor and vascular endothelial growth factor which are active in NSCLC biology. In thi
          s study, we used sunitinib, a multi-target receptor TKI, combined with chemotherapy for unresectable/metastatic NSCL
          C.This open label Simon's 2 stage clinical trial enrolled a total of 6 NSCLC patients who received docetaxel (40\xe2
          \x80\x8amg) and cisplatin (50\xe2\x80\x8amg) on day 1 of each cycle (14 day interval between cycles) and sunitinib (2
          5\xe2\x80\x8amg qd for 10 days between cycles) for a total of 12 cycles (24 weeks), after which patients received mai
          ntenance therapy with vinorelbine (30\xe2\x80\x8amg TIW) until disease progression. The sample size was based on a Si
          mon's Optimal Two-Stage Designs for Phase II clinical trials. The expected response rate was set as 35% for P0 and as
           60% for P1. The study was designed for a minimum of 6 patients for first stage and 15 patients until second stage wi
          th a significance level alpha = 0.10 and power = 70%. Diagnosis of a poor response in the second of 6 patients in Sta
          ge I or seventh of the 15 patients in Stage II would lead to early termination of the trial.The overall response rate
           was 66.7%. Four patients had an overall survival &gt;60 months. The time to PFS ranged from 3 to 42 months. The comb
          ination therapy was well-tolerated.Sunitinib combined with chemotherapy shows promise and warrants further investigat
          ion."
```

Tools and Techniques

1. Text Mining and Gene Extracting

Extracting abstracts from Pubmed was already finished in R by taking advantage of R package `RISmed`. And I also did basic data cleaning work in R. And then I move to Python and do the extensive text data processing work and extract genes from the abstracts.

2. Gene Newtwork Analysis

I will take advantage of Python package `networkx` to build and then analyze the gene network.

3. Data visualization

I will use `matplotlib` package for visualizing the gene network.

Goals

With this project, I try to accomplish three things:

1. I will gain extensive experience with PubMed database[2] and be able to webscrape all kinds of imformation from PubMed, such as abstracts, reviews, authors, methods, etc.
2. I also want to gain some experience with how to extract genes from literatures and then build the gene network.
3. At last, I am aiming to search for interesting genes relate to Non-small Cell Lung Cancer by combining text mining and gene network analysis techniques.

Overall, practising text mining and network analysis techniques is the main goal for this project. Besides that, I also would like to search for the potential bioinformatics project, which could be my capstone project next semester. As this paper [3] mentioned, by integrating text mining and network analysis we could potentiallly predict gene-gene relations and gene functions.

Reference Links

[1] http://www.cancer.org/cancer/lungcancer-non-smallcell/

[2] https://www.ncbi.nlm.nih.gov/pubmed/

[3] https://www.ncbi.nlm.nih.gov/pubmed/27112211