本周工作报告\_1

——2014.10.08

## 一、论文思路总结

因为是新学期第一篇工作报告，所以先总结一下之前工作的内容：

A Two-layer Transductive Learning for Retinal Disease Gene Prioritization 工作介绍

这篇论文的大概思路就是利用TTL方法对视网膜疾病基因进行相关性排序，大概方法如下：

1. 首先，对1,2,3三组数据都要转换成svm\_light所适应的格式。分别对1,2,3三个view的数据单独采用svm\_light进行学习，得到三组confidence
2. 将1中得到的三组confidence与proteomics and Crx相结合，再用svm\_light进行学习，得到最终的confidence，这个confidence就是用来评价基因与视网膜疾病的相关程度
3. 按照confidence降序排列，便得到一个有序列表。越靠前的基因与视网膜疾病相关程度就越高。

## 二、我的工作（这两周的工作）

1. 然后我就把大师兄的结果与一些排序网站进行一系列的比较，经过比较发现，toppGene这个网站的结果比大师兄的结果要好，所以我就想能不能利用ToppGene这个网站的排序结果，然后将其整合到自己的结果中，最终整合文件是“ToppGeneData\_results\_1\_17000.csv”，然后我删去了一些不必要的属性，转换成SVM-light支持的格式（这一步还没有做）。

2. 为了进行SVM训练，我的数据中还需要加入positive-gene的数据，本来打算把所有的属性都置为1，后来觉得这样做是鲁莽的，可能排序效果并不好，所以要把positive-gene的数据的属性通过网站来得出来。

3.为了验证这个方法的必要性，我设计了一个实验：

目标：判断其它167个positive基因在topgene网站的属性值

1.判断是否对于不同的positive gene输入，它们的属性值一样，所有的Training Gene Set：包含了 positive\_name\_human.txt 文件里的166个基因，所有的 Test gene set: 为全基因组Human\_genome\_random.txt

a. 输入数据分为三组：

第一组：Training Gene Set：前100个基因；Test gene set:后66个基因； 结果存入positiveGene SVM attribute/test1.csv

网站识别了全部基因

通过这组实验，我发现，每个positive基因的属性与输入的Training Gene Set的基因无关，而与Test gene set的大小有关，但差距并不是特别大。

第二组：Training Gene Set：前100个基因；Test gene set:后66个基因加Human\_genome\_random.txt中前1000个基因； 结果存入positiveGene SVM attribute/test2.txt，对比test1.txt,test2.txt中后66个基因的属性

网站识别了1066个基因中的1016个基因，有50个基因没有被识别

将一二组进行对比，选择 5+6 组基因输入 对比.xlsx 文件，进行人工对比，发现除了第一个属性GO: Molecular Function Score 值不同，其它值都是相同的

第三组：Training Gene Set：前160个基因；Test gene set:后6个基因； 结果存入positiveGene SVM attribute/test3.txt，并与一二组进行对比

第四组：Training Gene Set：前165个基因；Test gene set:最后1个基因； 结果存入positiveGene SVM attribute/test4.txt，并与一二三组进行对比，相同的实验进行6次，将最后六个基因都进行测试并对比

4.所以之后我就在写爬虫代码对网站进行自动填表与下载，然后发现网站把我的IP好像封了，然后我用其它网址登陆ToppGene这个网站，发现网站已经关闭数据下载功能了。

## 三、接下来的工作计划

因为topGene网站的数据下载被封掉了，所以只能使用之前下载的部分数据进行模拟来得到SVM的输入数据，所以接下来的工作就是：

1. 对positive-gene的各个属性进行模拟，具体模拟方式还需考察；
2. 将positive-gene 和之前整合的全基因组的数据进行整合；
3. 将所有数据转换为SVM-light支持的格式；
4. 利用SVM-light对这些数据进行训练；
5. 分析训练之后的结果，比较这些结果与简单的求平均之间的区别

这是我接下来的工作计划，具体细节还需要我自己思考，希望老师能多多指导