本周工作报告

——2014.10.15

## 之前的工作

对大师兄的实验进行完善，在对ToppGene这个网站进行分析的基础上准备用自动化代码将网站里的数据自动填表并且下载，但是用自动化下载网站遇到了一些挫折，导致网站暂时关闭了ToppGene: Candidate gene prioritization 的功能，严重影响了实验的进度。

当初想的方法是通过已知数据模拟出positive gene的数据，后来觉得这个想法太异想天开了，太不规范。

## 关于大师兄的论文

我大概问过了我大师兄的一些近况，这篇论文还可以继续发下去，为了老师和大师兄，我得加快进度了。加油！

大师兄毕业后在腾讯公司工作很忙很累，也没有时间关心这篇论文了，大师兄托我为他结尾，而且因为没有拿到毕业证，只领的是实习一半工资，如果在这学期期末还拿不到毕业证，上海落户都可能有问题。

## 我的工作（本周的工作）

1. 经过多次实验，后来发现ToppGene ：Candidate gene prioritization的这个功能已经可以使用了，但是不能用代码自动化下载。
2. 根据我的判断，Test set大小为6与大小为1的各个属性的差别并不大，后来发现大小为20与大小为1的各个属性差距也不大，考虑到本身利用网站自己来判断它们的属性就是本身不精确的，为了提高速度，我决定利用手动的方法，每次测出20个基因的数据，然后对数据进行整合，加入到全基因组的数据中。
3. 将下载好的数据存入positivegene文件夹中，依次命名为：ToppGeneData1.csv到ToppGeneData9.csv，（利用一个小的merge.bat代码将其合并）然后整合成ToppGeneData\_result\_positive.csv中，并转化成支持svmlight的格式：toppgene\_svmlight\_positive\_1.txt，（转化的代码存在toppgene\_svmlight\_raw\_to\_svm.R 文件中）。
4. 对已由的全基因组的基因转换为svm支持的格式，原始数据放在toppgene\_svmlight\_raw.xlsx，整理后的数据放在toppgene\_svmlight\_1.csv文件中，然后用R语言进行处理，转换成svmlight支持的格式：toppgene\_svmlight\_genome\_1.txt。
5. 准备这周五的组会，这周五的组会是我讲，我决定看“ToppGene Suite for gene list enrichment analysis”这篇文献，并将里面的核心思想搞懂，并讲给其他人听

## 接下来的工作计划

1. 认真准备周五的组会
2. 将svmlight格式的positive gene 的数据（toppgene\_svmlight\_positive\_1.txt）和之前整合的全基因组的数据（toppgene\_svmlight\_genome\_1.txt）进行整合，随机确定negativegene，并将它们调整成svmlight支持的格式；
3. 利用SVM-light对这些数据进行训练；
4. 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O(∩\_∩)O~