作业四. PubTator Covid-19相关基因突变实体识别和Shell编程

生信1802-余思克-2018317220208

实验目的

PubMed是从MEDLINE,生命科学期刊和在线书籍中收录了超过3000万的生物医学文献引文的生物医学文献数据库。其数据库中包含了许多带挖掘的生物医学知识。PubTator是一个基于Web,可通过使用高级文本挖掘技术来加快人工文献的管理(例如,注释生物实体及其关系)的工具。作为一个多合一的系统,PubTator提供了一站式服务来注释PubMed引用。本次实验将基于PubTator工具对PubMed中Covid-19相关文献的摘要进行实体抽取,针对Covid-19基因突变以及与Covid-19有关的人类基因组SNP信息进行分析与知识发现。

实验步骤

1. 数据获取

在Linux环境中使用shell脚本通过edirect工具中的 esearch 命令从PubMed数据库中获取与"Covid-19" 关键词有关的文献的uid,然后使用 curl 命令根据文献的uid调用PubTator工具网站中的API接口,将文献摘要对应的生物医学实体信息下载下来(近12000篇文献摘要)。

```
esearch -db pubmed -query "covid-19" | efetch -format uid > ./covid_19_uid.txt # 获取uid

echo 'I am curating the result.\n'
echo "\n" > covid_19_entity_result.txt
for pmid in `cat ~/NLP/task3/covid_19_uid.txt`
do
        curl "https://www.ncbi.nlm.nih.gov/research/pubtator-
api/publications/export/pubtator?pmids=${pmid}" >>
        ~/NLP/task3/covid_19_entity_result.txt
        sleep 5.8s
done # 根据uid下载生物医学实体信息
```

2. 数据清洗

通过上述步骤获取的数据包含三个部分的内容,分别为:标题、摘要与实体信息。本次实验关注的是第三个部分:实体信息中的基因实体。固需要通过Linux正则表达式提取出每篇文献的实体信息,并整合为一个数据框格式的文件以便后续处理。

```
grep -E "^[0-9]{8}\s" covid_19_entity_result.txt > covid_19_entity_result_entity.txt # 提取实体信息
```

3. 基因突变实体分析

PubTator提供的生物医学实体中包含的生物学概念有很多种,包括基因、蛋白质、疾病、物种以及变异等。本次实验关注的是变异相关实体,包括DNA变异、蛋白质变异以及SNP三个生物学概念,前两者为Covid-19病毒基因组出现的变异,后者为人类基因组中与Covid-19有关联的SNP。此步骤使用R语言抽取上述三种突变相关实体信息,并统计实体中各名称的词频,整合为一个由生物学概念、名称、词频三列元素组成的数据框。

```
data = read.csv(file.choose(), sep = "\t", header = F, stringsAsFactors = F,
quote = "")
entity = data.frame(entity = data$v4, bioconcep = data$v5)
entity$entity = tolower(entity$entity)
entity_list = c(entity$entity)
freq = table(entity_list)
freq = data.frame(freq)
names(freq)[1] = "entity"
freq$entity = as.character(freq$entity)
freq$Freq = as.integer(freq$Freq)
freq = merge(freq, entity, by = "entity")
freq = freq[!duplicated(freq$entity), ]
order_temp = order(freq$Freq, decreasing = T)
freq = freq[order_temp,]
rownames(freq) = seq(1,nrow(freq)) # 实体数据的读入及词频统计
DNA_mutation = freq[freq$bioconcep == "DNAMutation", ]
order_temp = order(DNA_mutation$Freq, decreasing = T)
DNA_mutation = DNA_mutation[order_temp,]
rownames(DNA_mutation) = seq(1,nrow(DNA_mutation)) # DNA变异实体抽取并按词频排序
protein_mutation = freq[freq$bioconcep == "ProteinMutation", ]
order_temp = order(protein_mutation$Freq, decreasing = T)
protein_mutation = protein_mutation[order_temp,]
rownames(protein_mutation) = seq(1,nrow(protein_mutation)) # 蛋白质变异实体抽取并按词
频排序
snp = freq[freq$bioconcep == "SNP", ]
order_temp = order(snp$Freq, decreasing = T)
snp = snp[order_temp,] # SNP实体抽取并按词频排序
```

4. 分析结果可视化

通过R语言中的 wordcloud2 包中的 wordcloud2() 函数对抽取出的实体中词频较高的名称进行可视化词云的绘制,从而值观反映出基因突变相关实体中的特征信息。由于实体中不同名称对应的词频大小差距过大,可视化效果不好,于是对词频进行了取根号处理以使词频差距变得柔和,便于展示与观察。

```
DNA_mutation$Freq = DNA_mutation$Freq \land 0.8 wordcloud2(DNA_mutation[seq(1,75),], size = 0.8) # DNA变异实体词云绘制 protein_mutation$Freq = protein_mutation$Freq \land 0.3 wordcloud2(protein_mutation[seq(1,75),], size = 0.6) # 蛋白质变异实体词云绘制 wordcloud2(snp[seq(1,75),], size = 0.7) # SNP实体词云绘制
```

实验结果

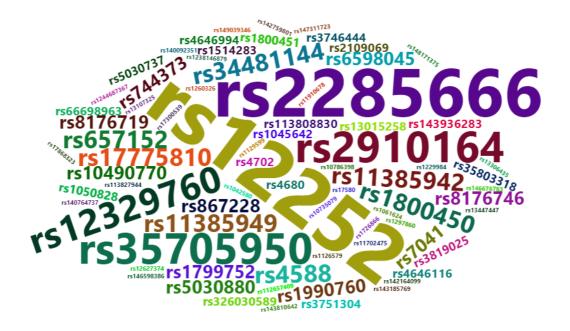
1. DNA变异实体可视化词云



2. 蛋白质变异实体可视化词云



3. SNP实体可视化词云



讨论

通过对Covid-19中的DNA、蛋白质突变实体词云的观察可以很明显地发现一些频率较高的突变信息,这些突变之所以发生频率如此之高,可能与Covid-19病毒的适应性进化有着关联,它们可能影响Covid-19病毒的存活能力、侵染能力、侵染宿主种类、侵染途径、传染性、致命性等,这些基因突变造成的病毒性状的改变使得病毒进化出了许多的亚型。通过PubTator提供的Covid-19基因突变实体信息的大致分析,可以筛选出一些病毒亚型的决定性突变特征,结合临床数据辅助进行病毒亚型的检测与鉴定,从而更好地辅助流行病学研究,为临床治疗与疫苗的研制提供参考。

通过对Covid-19相关的人类基因组SNP实体词云的观察可以明显发现一些出现频率较高的SNP。结合 NCBI SNP数据库的检索查询,可以发现这些SNP富集于诸如ACE2(编码血管紧张素转换酶2)、IFITM3(编码干扰素诱导的跨膜蛋白3)、MIR146A(编码microRNA等)、TMPRSS2(编码跨膜丝氨酸蛋白酶2)以及LZTFL1(亮氨酸拉链转录因1)等人类基因。通过这些高频SNP对应的人类基因信息可以大致了解到它们与Covid-19病毒有着密切的关联,进而可以辅助进行Covid-19病毒对于人类的入侵机理的研究,推测Covid-19病毒入侵人类时在分子生物学层面的攻击对象,人类的哪些基因对应蛋白质会与Covid-19病毒的侵染有关,以及SNP导致的这类蛋白质的表达量影响对Covid-19病毒侵染起到的促进或抑制作用,从而可以为临床治疗策略的选取提供一定的参考。

PubTator作为一款生物医学实体注释工具,有着强大的文本挖掘功能,能够辅助科研人员快速从海量数据中获取重要信息从而发现潜在的知识。虽然在本实验的进行过程中,我们发现此工具的注释效果不是特别好,存在许多实体信息被遗漏的情况,但是其作为一款简单易用的实体注释工具,依旧能够挖掘出文献摘要中的大量信息,是生物自然语言处理方向中辅助各类科学研究的利器。