《Pubmed 数据挖掘在 Covid-19 研究课题的应用》

姚岭君 [1], 徐琪 [1]

[1] 华中农业大学信息学院

摘要

新型冠状病毒肺炎（以下简称“新冠肺炎”）是由 2019 新型冠状病毒（以下简称“Covid-19”）感染导致的肺炎。到目前为止，由于 Covid-19 在印度和美国等地不断变异，全球感染者已接近 5 亿例，因新冠肺炎而死亡的人数已超过 600 万。幸运的是，各国科学家已经对新冠病毒进行了长达 3 年的研究，相关研究成果在科研论文数量上已经具备了一定的规模，相关基因的相互作用机制已逐渐明晰。不过，文献数量庞杂且稂莠不齐，人工阅读手段很难提取出有用的信息。NLP 在近些年快速发展，其拥有的快速检索数据信息和统计方便高效等优点，为我们分析文献提供了很大帮助。因此在本文中，我们利用 PubTator 提取 Pubmed 中的相关文献的实体信息，再对该信息使用 Python、R 语言等对数据进行词频统计以及基因关联分析等操作，得出了 xxxxxxxxxxx 的结论。这对于指导我们研究并应对 Covid-19 及其变异株的威胁非常有帮助。

课程数据及相关代码说明均在 github 中，链接如下`https://github.com/minghong/bionlp-Covid-19`

关键词: Covid-19, 文献检索, 词频统计, 关联分析

1 概况

1.1 选题说明（本章节作者：姚岭君）

新冠肺炎自 2019 年爆发以来，科学界已发表的相关研究文献有 24 万篇，我们无法对如此大量的文献进行人工分析，这个时候我们发现使用 BioNLP 的方法处理相关文件就是一个较好的的选项。因此本项目旨在使用生物医学自然语言处理的 Covid-19 相关文献，利用 PubTator 中识别文献实体进行提取，高效挖掘文献摘要中的实体信息并进行知识发现，为 Covid-19 的研究提供一定的参考。本项目计划挖掘出与 Covid-19 相关的高频基因、化合物、人类基因组和 Covid-19 基因组的突变情况；了解 Covid-19 相关高频基因的功能以及 Covid-19 相关突变对 Covid-19 侵染造成的影响。

2018 级学长学姐已经对这项课题进行了较为广泛的研究，得出了许多重要的结论。但仅此短短一年之间，有关 Covid-19 的文献数量翻倍，证明 Covid-19 的威胁仍然无处不在。特别是德尔塔和奥密克戎等传染性强的病毒的快速爆发，使我们认为重新对 Covid-19 文献分析刻不容缓，尤其是近一年间的数据与 2020 的数据进行比较十分必要。因此，我们决定继续这项课题，在前人已有的数据以及代码之上，优化代码检索文献速度，对比近年来病毒变异的方向与趋势，为抗击新冠疫情作出贡献。

1.2 算法基本原理（本章节作者：姚岭君）

我们在 Linux 上通过 edirect 工具获取了 PubMed 文献数据库中所有与“Covid-19”标题有关的文献的 uid，并通过调用 PubTator 的 API 接口，根据文献的 uid，提取了这些文献的实体识别信息。

本项目的实体识别过程使用的是 PubTator 工具。PubTator 是一款基于网络的医学文本挖掘工具，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，其提供了一站式服务来注释 PubMed 引用。与其他方法相比，PubTator 提供了已经完成识别的实体数据，

R 语言是一个功能强大的数据处理软件，并且

它还具有较好的画图功能，可以将结论可视化的展现在面前，大大加强了结论的可信度以及可读性。同时，我们利用市面上常见的统计词频网站对相同数据进行操作，并与 R 语言的结果进行比较，以期获得正确的结果。

本项目的数据处理部分使用到了 Linux Shell 系统，R 语言、Python 语言以及相关的拓展包。Linux Shell 可以方便地利用正则表达式快速进行数据的初步筛选，R 语言以及 Python 语言具有强大的拓展包生态、绘图能力、数据处理能力以及简洁的语法，使得数据处理过程方便且迅速。

1.3 数据来源（本章节作者：姚岭君）

本项目的文献数据来自于 pubmed 中的 242769 篇与 Covid-19 有关的文献，文献摘要中的实体信息来自于 PubTator。

获取的文献摘要实体数据由三个部分组成，第一个部分是文献的标题，第二个部分是文献的摘要，第三个部分是文献中识别出来的实体。实体内容包括基因、化合物、变异、细胞系、物种、疾病六个种类。本项目的研究将重点围绕实体数据中的摘要部分以及实体部分进行展开。

1.4 本文的方法部分与前人的相同点与区别（本章节作者：徐琪）

在项目开始之前，我们详细研读了学长学姐在此课题上所做的文章分析以及代码逻辑，从中提炼了精华部分，并对一些不合理的部分进行改进。其中，我们的文献数据更加充分，与学长学姐的 12 万篇文献相比，我们使用了 242769 篇文献数据（截止 2022.4.6），是前者的 2 倍，这使我们的结论更具有说服力。

在代码方面，我们经过测试与估计发现，如果直接采用学长学姐或者老师教授的代码，获取文献实体的时间将会达到 60 多个小时，这显然是不能被接受的。因此，我们重新设计了程序，并采用的是 python 多线程并行计算，分布处理多条数据，使得运行时间大大缩短。经过运行，设计并优化后的程序可以将时间压缩到 6 个小时，这无疑大大提高了我们的工作效率。

拥有了六类实体信息后，我们需要对这些信息进行整合处理。与其他人类似，我们也使用了 R 语言去统计处理这些繁杂的信息。最后通过词频统计以及关联分析，得到了相应的结论。

最后，我们会把结论与 18 级学长学姐的结论做对比，这也是我们独有的优势。通过数据对比，我们拥有明确的目标，也能发现一些全新的观点，从而掌握病毒的变异趋势，为协助研究 Covid-19 而努力。

2 Pubmed 文献挖掘及其算法实现研究》

2.1 实验流程大纲（本章节作者：徐琪）

1、首先使用 Linux 系统爬取所有有关 Covid-19 的文献 uid 号

2、使用 Python 利用 Putator 的接口提取文献标题与摘要。

3、使用 Python 利用正则表达式分别提取六类实体数据并保存。

4、利用 R 语言对实体数据进行词频统计以及关联分析

5、得出相应结论

2.2 文章实体获取（本章节作者：徐琪）

第一步，是获取文献的 uid 号。在 Linux 使用 edirect 工具中的 esearch 命令获取与 Covid-19 相关的 uid，并将其存入 txt 格式的文件中。不过，有些错误会出现，需要对报错信息进行处理，并下载安装必须的文件，此时才能获取相关的 uid 号。

第二步，是编写 Python 脚本通过 PubTator 接口获取 uid 对应文献摘要及识别得到的实体信息。因为以前的脚本运行速度过慢，因此需要重新设计新的 Python 程序提取标题与摘要。在提取实体过程中，由于 uid 数量过于庞大，串行处理时间过长，不满足我们的预期，因此我们使用扩展包 concurrent 中的多线程处理数据。遗憾的是，实体信息过于庞大，在一次性处理过程中我们发现有乱码的情况产生，并且由于未知原因，在处理过程中会遗漏大量 uid。为解决这一问题，我们将 24 万条数据不均匀分成 12 份，每份文件约 2 万条 uid，这样可以保证每一步的数据都是正常的而没有乱码产生。针对遗漏 uid 的问题，我们使用 Excel 的

countif 函数来找到遗漏的 uid，然后将这些 uid 重新放到脚本上运行，直到所有 uid 都被运行。

2.3 实体数据分析（本章节作者：徐琪）

第一步，使用 python 利用正则表达式过滤所得文件中的标题、摘要等部分以及六类实体信息存入以其实体名称命名的文本中。为防止出现数据重复的情况，我们对所有数据进行去重操作，这样可以保证得到的结果是正确的。

第二步，是对获得实体进行知识挖掘。通过 R 语言编写脚本读取实体数据，按照不同的实体种类把实体数据拆分开，统计实体词汇的词频并排序，然后根据词频分别绘制不同实体的词云。统计基因实体出现频率并使用 ggplot() 命令作出柱状图，找出其在染色体上的分布情况，息相关。统计基因，突变，疾病实体出现频率，找出出现次数较多的化合物，通过文献摘要及 NCBI 相关信息进一步了解这些化合物的功能，并且与上届学生所做的数据进行横向比对，找到在这新的一年中 Covid-19 发生的种种变化进行分析，有助于我们对这个病毒进行更好的规律性掌握，从而提出较好的防治建议。

3 实验结果

3.1 相关基因频词云结果（本章节作者：姚岭君）

在分析中我们发现这些基因，ACE2 与 S 在出现频率上有着极大的相似度，且分别属于人类和新冠病毒，我们有理由怀疑这两个基因转录的蛋白是否存在互作的作用。因此我们查阅了相关文献，得到的结果验证了我们的猜测。周强研究团队证明了新冠病毒表面 S 蛋白受体结合结构域与细胞表面受体 ACE2 全长蛋白的复合物的存在，但为后续科学家的靶向药物研究提供了更多信息。该项研究成果的另一个意义在于，计算生物学的研究人员可以在此基础上去构建不同的模型，进而展开具有针对性的研究，判断什么样的突变可能会进一步提高 S 蛋白与 ACE2 的相互作用，从而设计针对 S 蛋白或者 ACE2 蛋白的药物和抗体；又或者设计小分子破坏它们之间的相互作用。这些都为药物设计和检测手段开发提供坚实的基础。

在与上届的前辈的结果进行比较的时候我们发现我们其他方面大致相同，但是在第十个高频基因出现了一些区别，其命名为“RNA 结合调节肽”（RBRP）在经过相关文献查阅，我们发现这个蛋白该肽具有结合 IGF2BP1 的能力，并增强其识别 RNA 上 m6A 的能力，随后在肿瘤发生中起到致癌作用发现了之前注释过的 lncRNALINC00266-1 实际上编码了一种未标记的肽 RBRP ，该肽在肿瘤发生过程中具有致癌作用。这一发现改变了人们对于非编码 RNA 无法编码蛋白的传统认知，为今后对非编码 RNA 的研究打开了新的思路。

RBRP 可以与 m6A 编码器 IGF2BP1 结合，结合后与 YTH 结构域家族蛋白促进 mRNA 降解不同，而是使 mRNA 更加稳定。其原理是通过 RBRP 与 IGF2BP1 的结合，增强了 m6AIGF2BP1 对 RNA 的识别作用【1】，如 C-Myc 基因，通过增加 C-Myc 的稳定性和表达，从而促进肿瘤发生。很巧合的是，我们在对相关疾病进行统计词频的时候也发现“cancer”居然跃居前十，这不得不令人展开联想了。短文准备时间过短数据量过大，此处长文在进行深度挖掘。

表 1: 基因词频 top9 来源姚岭君自制 []

基因	功能
ACE2	血管紧张素转换酶 2
S	刺突糖蛋白
IL-6	白细胞介素 6
CRP	C 反应蛋白
TMPRSS2	跨膜丝氨酸蛋白酶 2
Mpro	2 ORF1a 多蛋白
CD4	4 CD4 抗原
CD8	CD8K 抗原
RBRP	RNA 结合调节肽

3.2 相关疾病分析（本章节作者：姚岭君）

此外在对 top20 的疾病进行统计中，我们发现了了一个很值得关注的地方：“anxiety”，“depression”等心里意义上的负面词语居然取代了许多生理上的疾病，要知道在去年的前辈们分析时还是疾病实体富集于“侵染”，“呼吸道”，“致命”，“死亡”，“糖尿病”，“感冒”，“疼痛”，“气喘”，“急性肾损伤”，“中风”，“神经”等词汇上

而今年的词汇却开始出现了“压力”，“不安”，“沮丧”等心理词汇，这无疑说明了在新冠疫

情时期，人们的心理健康状况在逐步的恶化，这也是我们应该深思的一个问题。要知道，身体的疾病还可以治愈，心灵的创伤我们又该如何抚慰呢？如果这个疫情下的心里问题无法得到我们重视的话，那么整个社会将遭受前所未有的灾难！

但是由于短文时间过于短，我们暂时不从得知究竟是新冠通过使机体产生病变而产生心理疾病还是由于新冠影响正常生活影响的人类心理健康，这个问题我们也将在长文进行深度探索。但是无论如何我们必须做好心理健康和身体健康的统筹兼顾。

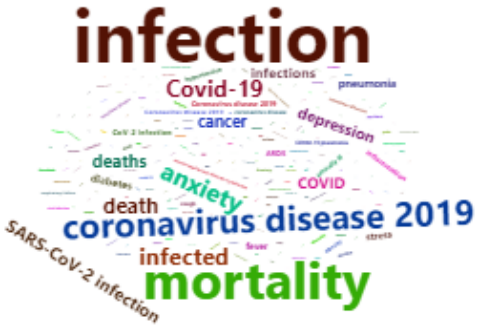

图 1: 相关疾病词云。图片来源 [姚岭君]

3.3 高频突变位（本章节作者：姚岭君）

对高频的基因突变位点进行了分析，在预实验中，我们用最新的 2 万篇文章的数据和一共 24 万篇文章进行对比，初步发现在最近的 2 万篇文献数据中 D614G 仍然是变异病毒的主要突变位点，但是 N501Y 的突变数量几乎已经与 D614G 持平。

这有可能表示有大量病毒的突变是 N501Y 突变。N501 位氨基酸是新冠病毒 RBD 蛋白的关键接触残基之一，直接参与 RBD 和 ACE2 的结合作用，而 N501Y 突变增强了 RBD 和 ACE2 的结合亲和力。N501 属于亲水性残基，病毒结合 ACE2 时，N501 靠近 ACE2 的 Y41 疏水苯环和 K353 疏水烷烃链；但当 N501 突变为疏水残基 Y 时，Y501 则可以通过疏水作用与 ACE2 中的 Y41 和 K353 更好的配位，改善 RBD 和 ACE2 互作构象，将结合亲和力提高约 0.81kcal/mol，同时使原来不感染小鼠的新冠病毒毒株获得感染能力。N501Y 突变对于靶向 RBD 中和表位的中和抗体的结合效力影响较小。研究发现，N501 毒株与中和抗体结合 VH-Fc ab8 后（该分子融合表达了可变重链区域 VH 和人类 IgG1 的 Fc 片段），S 蛋白呈现两种构象即两个 RBD 分子处于“向上”位置和仅有一个 RBD 分子处于“向上”位置。我们担心这个基因突变的增多会导致新冠病毒在其他物种的传染力增强，从而导致更严重的后果。

此外我们还发现 E478K,L452R 的比例明显上升经分析相关文献我们找到了原因，Delta 变异株相较于其他变异株，则在 S 蛋白上新增了 3 个重要突变“L452R”、“T478K”和“P681R”。恰好 Delta 这个名词集中出现在去年 7 月之后出现在文献中，而其毒株也在去年下半年成为了新冠病毒的主力军。

图 2: 24 万篇文章高频突变与最新的 2 万篇文章对比。图片来源 [姚岭君]

4 附录

4.1 Checklist for ethics, societal impact and reproducibility issues (请检查并简要回答)

4.1.1 本文是否提到了工作的局限性？如果有，请指出相关章节。

答：有待学习。

4.1.2 本研究工作有无任何潜在风险，是否在文中被提到？

答：实体信息过于庞大，在一次性处理过程中我们发现有乱码的情况产生，并且由于未知原因，在处理过程中会遗漏大量 uid。为解决这一问题，我们将 24 万条数据不均匀分成 12 份，每份文件约 2 万条 uid，这样可以保证每一步的数据都是正常的而没有乱码产生。针对遗漏 uid 的问题，我们使用 Excel 的 countif 函数来找到遗漏的 uid，然

后将这些 uid 重新放到脚本上运行，直到所有 uid 都被运行

4.1.3 文章的摘要是否能较好概括本文的工作？

答：比较好的得到宏观数据，但是具体概念有待观察

4.1.4 本工作是否产生了自己的数据或方法？如果有相关介绍，请指出相关章节。

答：利用 r 制作词云，利用 excel 出处理相关表格等数据

4.1.5 本工作是否使用并介绍了已有的数据或方法？如果有，请指出相关章节。

答：存在

4.1.6 如果使用了已有的数据或方法，是否介绍了相关使用条款？

答：介绍了并行，esearch 说明书

4.1.7 本文是否介绍了所使用数据的基本统计结果，例如 train/test/dev？

答：table 等等有待丰富

4.1.8 本工作是否使用了代码计算，如果有，是否介绍了运行环境和运行时间？如果有，青指出相关章节。

答：使用 python 和 R 以及 excel 等等

```
代码来源：esearch说明书
esearch -db pubmed -query "covid-19" |
    efetch -format uid > pubmed.txt #提
    取所有文献的uid
```

```
代码来源：https://github.com/minghong/
    bionlp-Covid-19 徐琪自制
def get_data(url):#利用接口得到文献题目及
    摘要
        data=urllib.request.urlopen(url).
            read()
        data=data.decode("utf-8")
        file2.write(data)

if __name__=='__main__':#多线程并发处理
        executor = futures.
            ThreadPoolExecutor(
            max_workers=10)
        for url in urls:
                f=executor.submit(
                    get_data, url)
```

```
代码来源：https://github.com/minghong/
    bionlp-Covid-19 徐琪
#利用正则表达式提取六类实体并保存
Species = re.compile('Species'+'\t')
for line in file2.readlines():
        if(re.search(Species, line, flags
            =0)):
                species.write(line)
```

```
代码来源：姚岭君
entitylist <- lapply(filelist, function(x) read.
    csv(x,header = F,sep = "\t",col.names = c("
    pmid","start","end","name","type","ID")))
names(entitylist)<-c("cell_lines","chemical","
    disease","gene","mutation","species")
gene<-(entitylist[["gene"]]) #找出基因
gene <- gene[!duplicated(gene),]#去重
name<-as.vector(gene$name)
word<-table(name)
gene_order<-as.data.frame(word)
wordcloud2(word,size = 2, fontWeight = "bold")
```