组 别  **本科生**

题 目  **A**

队 号

****

2020年数学建模竞赛

新型冠状病毒的基因分析与变异进化研究

|  |  |
| --- | --- |
| 姓名及学号 | 汪劲翔17374168 |
| 姓名及学号 | 李昂17374158 |
| 姓名及学号 | 罗星月17374234 |
| 队伍联系电话 | 18810010572 |
| 队伍联系邮箱 | 2205779233@qq.com |

目 录（示例）

[一、问题重述 1](#_Toc45651868)

[二、问题分析 1](#_Toc45651869)

[2.1 问题1的分析 2](#_Toc45651870)

[2.2 问题2的分析 2](#_Toc45651871)

[三、模型假设 2](#_Toc45651872)

[四、模型建立与求解 2](#_Toc45651873)

[4.1 问题一的建模与求解分析 2](#_Toc45651874)

[4.1.1 预备知识： 2](#_Toc45651875)

[4.1.2 预处理及基因提取： 3](#_Toc45651876)

[4.1.3 突变模型 4](#_Toc45651877)

[4.1.4 数据代入模型 5](#_Toc45651878)

[4.2 模型二的建立与求解分析 5](#_Toc45651879)

[4.2.1进化树模型算法介绍及选择 5](#_Toc45651880)

[4.2.2 进化树模型建立 7](#_Toc45651881)

[4.2.3进化树分析： 7](#_Toc45651882)

[五、模型评价 8](#_Toc45651883)

[主要参考文献 9](#_Toc45651884)

致公共卫生部的一封信

敬爱的卫生部门工作人员：

你们好！

我们小组通过建模分析从NIH（National Institutes of Health）的GenBank下载2020年3月到7月（截止14日之前）的SARS-CoV-2的DNA序列，我们得出了基因变异的特征，同时绘制进化树，做出了关于进化树的结果分析。

关于基因变异特征方面的研究，选取S段蛋白质基因片段（具体位置在正文中有述），在结果数据中，发现2043例发生于23403位碱基，由A腺嘌呤替换为G鸟嘌呤，突变率高达为67.2%的基因突变。该点位的氨基酸由天冬氨酸变异为甘氨酸。超过半数的变异都为该变异，可以考虑将其作为一稳定平衡位点，进而推断出变异规律。

对于进化树而言，我们选取M段基因绘制进化树，在进化树中，MT325570这一样本序列比例最高，同时观察分支，分支代表变异方向不同。序列文字越大，其所占比例越大，所以应着重考虑所占变异比例大的几个变异方向。

希望我们的建议能够派上用场。祝工作、研究顺利！

敬祝

身体健康

2020年7月14日星期二

# 一、问题重述

2019年12月以来，湖北省武汉市持续开展流感及相关疾病监测，发现多起[病毒性肺炎](https://baike.baidu.com/item/%E7%97%85%E6%AF%92%E6%80%A7%E8%82%BA%E7%82%8E/2251212" \t "_blank)病例，均诊断为病毒性肺炎/肺部感染，称为新型冠状病毒。而后在全球各地相继爆发新冠，如今截止2020年7月11日。全球新型冠状病毒肺炎累计确诊病例已经超过1260万人。

本题主要侧重于基因分析及进化树模型的建立，利用匹配算法和动态规划建立模型，解决以下问题：

数据取自何处：

1. 问题一：利用匹配算法建立模型，分析新型冠状病毒的基因变异特征，进一步求出稳定变异点特征。

2．问题二：采用动态规划算法编辑最小距离，使用iTOL工具构建进化树模型，对结果进行相关分析，给出建议。

# 二、问题分析

## 2.1 问题1的分析

建立基因匹配分析模型。首先数据 从NIH（National Institutes of Health）的GenBank下载2020年3月到7月（截止14日之前）的SARS-CoV-2的DNA序列，以及下载参考序列。数据平均采集与三、四、五、六、七月份，同时从数据库获取了某一月采集的病毒DNA序列作为参考序列。研究其他数据相对于参考序列的基因突变特征以及稳定突变点的研究。采用匹配算法进行基因变异的研究，通过对基因各个点位碱基对的突变频率与相应氨基酸的变异情况研究基因变异特征和稳定变异点位。

## 2.2 问题2的分析

建立最小距离模型构建进化树，首先根据增减与替换的概率确定增减与替换变异的权值来确定最小距离的具体计算。其次，选取单个蛋白质基因的进化树进行分析。采用动态规划中编辑距离问题算法对突变代价进行参数解算，得到有序突变过程。构建邻接矩阵之后，采用Prim算法进行最小生成树的提取，最后使用iTOL工具完成进化树的绘制。

# 三、模型假设

1. 假设一月的病毒序列为参考序列。
2. Y，N中对于不确定的碱基序列中没有发生突变的碱基删去。
3. 按照基因变异中的发生增删和替换的概率，将增删和替换的权值之比设为10：1。

# 四、模型建立与求解

## 4.1 问题一的建模与求解分析

### 4.1.1 预备知识：

(1)密码子

密码子（codon）是指信使[RNA](https://baike.baidu.com/item/RNA" \t "_blank)分子中每相邻的三个[核苷酸](https://baike.baidu.com/item/%E6%A0%B8%E8%8B%B7%E9%85%B8/2115624" \t "_blank)编成一组，在蛋白质合成时，代表某一种[氨基酸](https://baike.baidu.com/item/%E6%B0%A8%E5%9F%BA%E9%85%B8/303574" \t "_blank)的规律。

(2)转录

转录（Transcription）是遗传信息从DNA流向[RNA](https://baike.baidu.com/item/RNA" \t "_blank)的过程。即以双链DNA中的确定的一条链（模板链用于转录，编码链不用于转录）为模板，以A,U,C,G四种核糖核苷酸为原料，在[RNA聚合酶](https://baike.baidu.com/item/RNA%E8%81%9A%E5%90%88%E9%85%B6" \t "_blank)催化下合成RNA的过程。

### 4.1.2 预处理及基因提取：

#### 预处理

从NIH（National Institutes of Health）的GenBank下载2020年3月到7月（截止14日之前）的SARS-CoV-2的DNA序列，以及下载参考序列。因数据量过大按月份进行分组。

按如下规则过滤数据：

1. DNA序列全场在28500到30300之间（参考序列长度为29903）
2. N等待确定的“碱基”在序列中出现次数不多于25

#### 基因提取

使用匹配算法，根据起始密码子和终止密码子对应的DNA序列来切割获得的病毒DNA序列。这里使用该算法将基因分为：M，E，N等等，如图所示：

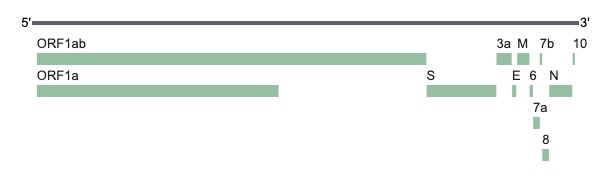


图 1 基因切割样式示例

匹配算法如下：

1. 设置起始位点S初始值为-1
2. 由S+1正序查找起始子并更新位点为S
3. 从S位点开始，间隔三个碱基进行终止子匹配，找到第一个终止子，标记位点为E
4. 对S：S+25和E-25：E的片段进行匹配，若重合率大于80%，即认为匹配到目标基因。否则回到步骤2。
5. 匹配到目标基因后，回到步骤2匹配下一个目标基因，直到所有基因匹配完成或者待匹配序列匹配完成。

### 4.1.3 突变模型

#### 突变代价

首先确定基因中发生替换和删减的权值。根据前文有关基因突变中增删与替换的比例的假设，记一次增删或一次替换为一次基本突变，设置突变的代价如下。

当为一次增删时：

当为一次增删时：

设基因片段到的突变可由一系列基本突变有序构成，即

有

#### 编辑距离

编辑距离是指由串到串所需要的最小“编辑”次数。在本模型中将其中的“编辑次数”扩展为“突变代价”。

编辑距离问题是典型的动态规划问题：

我们用表示的前个碱基和的前个碱基之间的突变代价。自底向上由子问题向父问题的推导规则：

1. 若串和串的最后一个碱基不相同：
2. 若串和串的最后一个碱基相同：

式中为增删对应的代价，为替换对应的代价

即为所求的最小突变代价。

根据此方法求得的结果一定是最小突变代价，具体参考动态规划的证明，本文不再赘述。

#### 突变过程

保存上一步中获取的动态规划数组，由依次追溯，和至即可恢复出突变的流程，再将其倒序排列即可得到有序突变过程。

通过计算参考序列对应基因到目标基因的距离，对每一条待统计基因，可得到符合最小突变代价要求的基因突变列表。

### 4.1.4 数据代入模型

根据切割出的长度，同时对每一段基因实际操作后，以变异次数及有效变异（次数最多的变异是否会使对应氨基酸改变）为考虑因素，最终选取S基因作为研究序列。

将所有序列代入模型解算，得到表格，此处只给出变异次数大于10次的变异类型：

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| count | position | source | source\_pair | target | target\_pair |  |  |
| 2043 | 23403 | A | GAT | G | GGT |  |  |
| 49 | 21575 | C | CTT | T | TTT |  |  |
| 18 | 23010 | T | GTT | C | GCT |  |  |
| 22 | 24694 | A | GGA | T | GGT |  |  |
| 57 | 24034 | C | AAC | T | AAT |  |  |
| 11 | 22444 | C | GAC | T | GAT |  |  |
| 16 | 23929 | C | TAC | T | TAT |  |  |
| 24 | 24904 | C | ATC | T | ATT |  |  |
| 13 | 23604 | C | CCT | T | CTT |  |  |
| 11 | 24382 | C | TCC | T | TCT |  |  |

表 1 数据集中S基因突变类型汇总

基因变异特征及稳定变异位点解答部分：

本研究共对3038例产生变异的新冠病毒进行了DNA序列分析，截取s段基因（给出基因切割图），发现2043例相对于参考序列产生变异的病毒序列，此变异发生于23403位碱基，由A腺嘌呤替换为G鸟嘌呤，突变率为67.2%，此时，该点位的氨基酸由天冬氨酸变异为甘氨酸。此变异比例高，且产生氨基酸变异，故可以认定该点位为稳定变异点位。此外，在24034位碱基处，发生了57例由C胞嘧啶替换为T胸腺嘧啶的突变，虽然变异的频率较高，但氨基酸依然是天冬氨酸，不足以作为稳定变异位点。另外一个比较常见的发生突变的基因位于21575位碱基，发生了49次C胞嘧啶到T胸腺嘧啶的转换，此处氨基酸由亮氨酸替换为苯丙氨酸。变异散乱分布于S基因序列，并无某一特定的聚集变异区域。

## 4.2 模型二的建立与求解分析

### 4.2.1进化树模型算法介绍及选择

#### 邻接法（NJ法）

基于最小进化原理，进行类的合并时，不仅要求待合并的类是相近的，而且要求待合并的类远离其他的类。

适用于远缘序列，但进化距离不大，信息位点少的短序列。

优点：假设少，树的构建相对准确，计算速度快，只得一颗树，可以分析较多的序列，运行速度优于最大简约法

缺点：序列上的所有位点等同对待，且所分析的序列的进化距离不能太大

#### 最大简约法（MP法）

基于进化过程中碱基替代数目最少这一假说，不需要替代模型，对所有可能的拓扑结构进行计算，并计算出所需替代数最小的那个拓扑结构，作为最优树

优点：善于分析某些特殊的分子数据如插入、缺失等序列有用。

缺点：存在较多回复突变或平行突变时，结果较差。推测的树不是唯一的，变异大的序列会出现长枝吸引而导致建树错误。

#### 贝叶斯法（BI法）

是目前主流的建树方法

基因进化模型的统计推论法，通过后验概率直观反映出各分支的可靠性，而不需要自举法检验。

适合大而复杂的数据集

优点：具有坚实的数学和统计学基础，可以处理复杂和接近实际情况的进化模型

缺点：对进化模型敏感，BI法中指定的每个氨基酸的后验概率建立在许多假设条件下，现实中可能不成立。

#### 最大似然法（ML法）

依赖于某一个特定的替代模型来分析给定的一组序列数据，使得获得的每一个拓扑结构的似然率都为最大值，然后再挑出其中似然率最大的拓扑结构作为最优树。

适用于特定的替代的模型，有模型有模型的情况下ML是与进化事实最吻合的树。

优点：很好的统计学基础，大样本时似然法可以获得参数统计的最小方差，在进化模型确定的情况下，ML法是与进化事实吻合最好的建树算法.

缺点：所有可能的系统发育树都计算似然函数，计算量大，耗时时间长。依赖于合适的替代模型。

对于新冠病毒，因序列数量庞大以及变异复杂，我们依据各种建树方法的优缺点，选择最大简约法与距离法结合，建立进化树模型。

### 4.2.2 进化树模型建立

进化树的构建的主要依据是基因之间的突变关系网络。最小生成树在图中提取联通图子图并使的边权值之和最小。因此进化树的求解可以转化为，将所有基因构建成一个图，基因节点之间的边权值为两基因之间的突变代价。

考虑到数据的量和计算的复杂度，在其基础之上进行优化：

1. 根据前步突变过程的构建，提取出所有相对于参考序列有变异的序列，记为集合（大量序列是与参考序列一致的，利用进行过滤可减少大量冗余计算）
2. 优化后的图中每个节点不代表单个DNA序列，而是目标基因相同一簇DNA序列，因此每个节点还应该保存序列簇中序列的个数，序列的名称等基本的信息。
3. 设序列簇，。中含有序列。中含有序列。到的边权值为

构建邻接矩阵之后，采用Prim算法进行最小生成树的提取，才用prim算法的优点可以方便的将数据量较大的序列簇优先作为父（根）节点，与实际情况更符合。

最后利用iTOL工具进行进化树的绘制。

### 4.2.3进化树分析：

这里选取病毒DNA序列中的M段基因来进行进化树构建

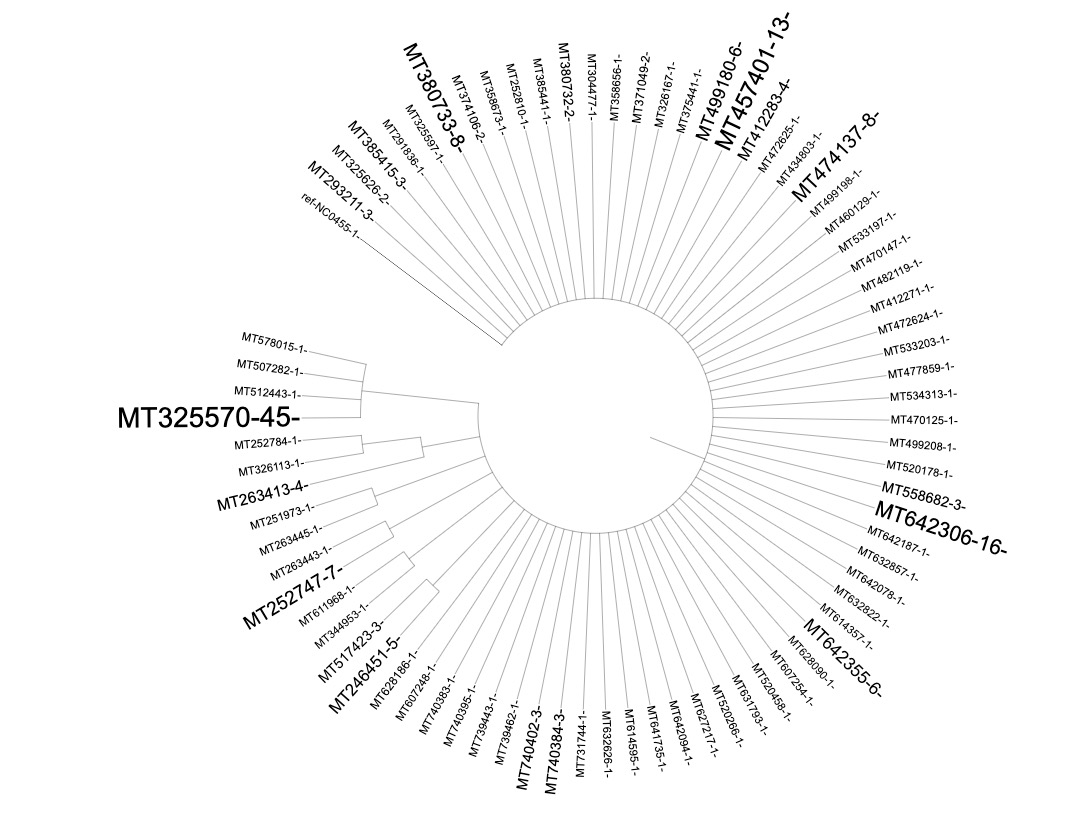


图 2 关于M基因建立的进化树

总共含有从数据平台下载的近3500条新型冠状病毒全基因组序列。其中选取M段基因采用上述模型构建进化树，得到图中所示进化树。其中序列序号的大小表示出现次数占所有采集次数的比例。这里，变异的远近关系通过层级积累。枝节点层数越大表示样本差异的拓扑距离越大。

# 五、模型评价

1. 此模型从最大简约法的理论原则“找出所需突变代价最小的拓扑结构作为最优树出发”出发，能够快速得出一个合理的进化树，对于新冠病毒这样变异迅速且基因序列样本庞大的情况很适用。
2. 缺点是缺少对照，不能及时发现不合理的进化分支，需要人工或者利用其它模型的结果来找出一场结果。
3. 可以通果加入新的模型，用多种模型的结果对照，保证结果的正确性。

主要参考文献

1. [广西钦州地区结核分枝杆菌相关耐药基因变异特征分析](https://kns-cnki-net-443.e2.buaa.edu.cn/kcms/detail/detail.aspx?filename=GWSQ201808012&dbcode=CJFQ&dbname=CJFD2018&v=)[J]. 李崇建,鲍金圭,陈焯彬,李仕娟,吴海松,梁玲.  国际检验医学杂志. 2018(08)
2. [新型冠状病毒实验室核酸检测方法及实践](https://kns-cnki-net-443.e2.buaa.edu.cn/kcms/detail/detail.aspx?filename=ZHYY202006009&dbcode=CJFQ&dbname=DKFX2020&v=)[J]. 匡慧慧,于梅,于帅,周汉王,王依函,朱乔,肖征,李慧灵.  中华医院感染学杂志. 2020(06)
3. [2019新型冠状病毒信息库](https://kns-cnki-net-443.e2.buaa.edu.cn/kcms/detail/detail.aspx?filename=YCZZ202002009&dbcode=CJFQ&dbname=CJFD2020&v=)[J]. 赵文明,宋述慧,陈梅丽,邹东,马利娜,马英克,李茹姣,郝丽丽,李翠萍,田东梅,唐碧霞,王彦青,朱军伟,陈焕新,章张,薛勇彪,鲍一明.  遗传. 2020(02)
4. [福建省新型冠状病毒的分离和全基因组特征分析](https://kns-cnki-net-443.e2.buaa.edu.cn/kcms/detail/detail.aspx?filename=ZRSZ202005002&dbcode=CJFQ&dbname=CJFD2020&v=)[J]. 张炎华,陈炜,何文祥,吴冰珊,朱颖,王金章,游丽斌,黄枝妙,陈宏彬,俞婷婷,鄢育青,吴晶晶,修文琼,袁平,张小鸿,骆婧,翁育伟,郑奎城.  中国人兽共患病学报. 2020(05)

.