《信息可视化概论》

项目建议书

（2021-秋）

题目： **基于 H3N2 流感序列的**

可视化分析系统

小组名称：助力每一个梦想

成员姓名：宋文凯

苏超

尹航

秦阳

指导教师：朱敏

提交时间：：二〇二一年十月十九日

声 明

本团队声明所呈交的项目建议书是本团队成员进行的项目讨论的相关结果。

据我所知，除了文中特别加以标注引用和致谢的地方外，建议书中不包含其他人已经发表或撰写过的研究成果。

团队成员对本项目建议书所做的任何贡献均已在作了明确的说明。

本项目建议书是本团队在四川大学读书期间在任课教师指导下取得的，论文成果归四川大学所有，特此声明。

项目建议书团队成员签名：

签字日期：　　年　月　日

目 录

第 1 章　项目背景

1.1 选题背景及意义

流感病毒是一种能引起人类或动物发生急性呼吸道传染病的病原体，持续威胁着人类和动物的健康。二十世纪三次世界性流感大流行导致全球数千万人死亡，季节性流感每年导致全球数百万人感染和数十万人死亡（以老年人和小孩为主）[1]。流感病毒的核酸（遗传物质）是单股负链 RNA。流感病毒的结构一般为球形，从内到外分别是核心、衣壳和包膜，包膜表面镶嵌的两种糖蛋白或者是表面抗原分别是血凝素 HA 和神经氨酸酶 NA。病毒是通过表面的血凝素与宿主细胞膜上的唾液酸受体结合，进入宿主细胞进行繁殖扩增，从而引起宿主感染。病毒也是通过表面的血凝素和抗血凝素抗体结合，导致病毒失去感染能力[2]。流感病毒的基因组是分节段的，在复制过程中容易发生变异，比如基因突变或者基因重配，特别是甲型流感病毒，导致病毒的抗原属性发生漂移或转变[3]。因此探究流感病毒进化的情况，特别是地区分布的情况以及共进化特性，将会显著的影响疫苗的开发效果和保护区域效力的情况。

本项目基于流感病毒序列，探究流感病毒的可视化工作，主要包含对 H3N2 流感的序列的进化分析和病毒的地区分布的情况探究。其中对序列的可视化将包含病毒的聚类可视化、序列差异可视化、序列分子进化树的构建。从氨基酸序列角度对病毒的突变进行可视化分析。能够很好的帮助研究人员把控序列进化角度的情况。对于病毒的时间和空间上的分布，则能很好的探索不同病毒在进化过程中的时空特征，辅助研究人员探究疫苗开发中的抗原关系，辅助建立疫苗分配的机制[3]。

1.2 相关工作

本项目将基于 H3N2 的病毒数据进行序列进化分析和时空朔源工作。

1.2.1 序列进化分析

序列的进化伴随着位点的突变，往往序列的突变不会引起抗原的剧烈变化，但是探究不同病毒序列的相似性质以及序列进化的程度和方向将能很好的研究病毒的进化路径，特别是不同地区病毒间序列进化的层面的，对研究抗原预测和进化预测有重要的辅助作用，基于此出发点此项目将包含一下功能点[4]：

（1）基于序列的进化树：

（2）病毒序列聚类分析；

（3）序列差异分析；

（4）序列位点分析；

（5）抗原关系可视化；

（6）回溯性分析可视化；

1.2.2 病毒时空分布分析

病毒的进化有着时间和空间特性，病毒的进化是成簇的[5]。不同区域的病毒往往有着较强的相关性，但同时不同区域也可能出现病毒的共同进化特性。因此探究病毒进化过程中在时间和空间上的分布，将包含以下功能：

（1）全球病毒空间分布展示；

（2）全球病毒时间分布展示；

（3）全球病毒流行趋势描述；

1.3 研究目的与可视化任务

本项目探究病毒进化的时空关系和序列进化路径，给出病毒在进化过程中，序列进化情况，特别是某些位点发生变化是发生的显著抗原漂移时，查看此类情况的序列进化情况，将会帮助研究人员确定某些重抗原位点，帮助构建抗原预测模型。而病毒进化的时空关系则有助于研究人员研究不同病毒进化在区域上的差别[6]，辅助开发针对不同区域的疫苗。同时本项目预计结合不同的抗原预测方法来，可视化不同序列之间抗原的相似性，并基于结果进行可视化展示与分析。

（1）序列的差异，包含序列位点差异，同时采用不同颜色进行表明差异位点；

（2）序列的分子进化树，利用树形，表示进化的路径，利用颜色表示序列的差异情况；

（3）序列聚类情况，依据聚类算法将不同的序列进行聚类分析，同时增添时空分布查看功能，将不同的聚类以不同颜色表明，映射在全球分布的地图中；

（4）抗原关系预测，上传序列文件，进行抗原距离预测，同时支持训练结果的时空回溯分析，帮助开发抗原预测工具。

（5）流感进化的时空分布，将全球地区和时间轴相结合，通过时间轴展示不同病毒分布情况。

1.4 创新点

生物信息学中，众多的流感可视化网站往往功能分散，例如序列差异对比和抗原预测分析。同时许多可视化的工具，都是以工具库的形式封装在具体语言中，对于刚入门的学者不利，特别是从事流感的非计算机人员。本项目基于生物信息学，利用大量的开源工具和可视化技术，集成针对流感病毒序列（以 H3N2 序列为例子）进行集成化的分

析，帮助人员快速了解其感兴趣的功能。

同时值得注意的是，抗原预测工具开发过程中，往往不会对具体的结果进行可视化分析，只有一个平均结果，而本项目将会展示每轮数据中具体的训练结果，帮助分析潜在的数据问题和模型改进潜在方向[7]。

第 2 章　数据说明

2.1 数据来源

本项目需要使用两类 H3N2 数据：HI 实验数据和病毒抗原 HA1 序列数据。HI 实验数据来源于学术报告、科学文献、世界卫生组织流行病学周报、世界卫生组织流感参考和研究合作中心和各国疾病预防控制中心网站。病毒的抗原 HA1 序列数据来源于 Influenza Research Database、NCBI Influenza Virus Database 和 Global Initiative on Sharing All Influenza Data[8]。

在文献中数据收集方面，主要在四篇论文的附录以及公开网站获取，如表 2-1 数据来源汇总所示。

表 2-1 数据来源汇总

数据来源	HI 滴定对数目	HA 序列数目	备注
CAO	3867	679	
LIAO	277	62	
PENG	791	621	不同簇为抗原变异，相同簇为抗原相似
WU	11 个抗原簇	195	
NCBI、PDB	4181	124	
数据合并	47633	791	
数据清洗	22401	791	序列差异<20

2.2 数据描述

将收集的 HI 和 HA 数据进行初步的数据清洗和汇总，整理出如表 2-2 格式所示的原始汇总数据。

表 2-2 原始 H3N2 数据

字段名称	示例	字段描述
Pair1/2	A/BRAZIL/1742/2005	病毒名称（型别/分离地区/毒株序号/分离年份）
Seq1/2	QKLPGN...EKQTR	329 长度的氨基酸序列
Diff-seq	7	序列对差异
Class	0	抗原变异为 1，相似为 0
Dij	0.495	抗原滴定数据

依据时空数据分析的需要对表 2-2 的数据进行进一步的解构，形成如表 2-3 所示的序列时空数据。

表 2-2 原始 H3N2 数据

字段名称	示例	字段描述
Name	CHINA/FUJIAN	分离地区(国家+地区)
Id	1742	病毒序列唯一标志 ID
Sequence	QKLPGN...EKQTR	病毒氨基酸序列
Cluster	203	所属簇分类
Year	2005	分离时间
Foreign Keys	101	有本毒株有抗原距离的其他毒株序列

对于抗原预测结果数据采用如表 2-3 的形式进行结果说明。

表 2-3 抗原预测结果数据

字段名称	示例	字段描述
Year	2003	训练数据集时间段
Accuracy	90.2%	训练集准确度
F1-score	56	F1 分数指标
Model-Name	CNN	模型名称
Epoch	23	训练轮数

第 3 章 项目主要内容

为了帮助科研人员更好的掌握病毒动态的进化过程，探究不同毒株之间序列水平的差异以及毒株系统计划的情况，辅助研究人员对不同区域的毒株开展特定的疫苗研发工作。本项目《基于 H3N2 流感序列的可视化分析系统》将作用于流感病毒，依托公共数据集，收集 1968 年以来自有明确记载的毒株数据[8]，探究 H3N2 病毒序列的可视化和病毒时空分布的相关内容。主要内容点如下：

3.1 分子进化分析可视化

本项目中 H3N2 的序列选自 HA 蛋白中的 HA1 多态序列，HA1 是领域中主要研究的区域，将其作为 H3N2 的主要抗原。可视化的内容包含常见的序列差异对比的相关内容，同时结合 H3N2 自身的特点，我们对工具进行定制化操作，结合开源工具例如 MAFFT、MEGA 等作为参考，构建一个针对 H3N2 病毒序列的可视化序列分子水平可视化平台，包含一下内容

（1）基于序列的进化树：

不同的病毒序列间有着进化的关系，每次病毒突变某一个位点，并且其毒性不会立马改变。通过探究分子水平的进化树，根据位点的突变规则，将序列之间的关系构建树状模型，完成对分子水平朔源的实现。通过进化树，可以实现分子水平进化的可视化功能。依托不同的层次和颜色表明不同的“家系”，展示出序列进化的不同层次。

（2）病毒序列聚类分析；

病毒序列进化有着层次性，对于第一步的连续性的分子进化，抗原水平表现出明显的成簇的现象。对于不同簇的划分，有利于探究抗原漂移的关系。使用聚类算法，对毒株进行聚类分析，并进行可视化操作。可以通过不同的颜色表明不同簇，同时簇间的距离通过显示的距离表明。

（3）序列差异分析；

两个毒株之间的差异对比，也是分子进化的重要功能，将两株 329 长度的序列，进行逐位的对比，不同位点给予重点的标注。同时也会对序列段进行分析，将两序列中相似的片段展示出。

（4）序列位点分析；

对某一个特定的序列，统计每个氨基酸的频率，同时进行可视化展示。

（5）抗原关系可视化；

点开某一抗原后，展示其在数据库中的所有信息，并且能通过标签访问其他功能。

（6）回溯性分析可视化；

抗原分析工具需要进行测试和验证。按照不同 epoch 训练出的结果结合数据的时间进行的回溯性分析结果可视化，利用三维试图和热力图的形式呈现出 F1-score 与 epoch 和数据集之间的关系[9]。

3.2 病毒时空分布分析

结合实际地图、时间分类对病毒的时间和空间的分布进行可视化操作，同时对于同一时间段的多个病毒，利用不同的大小表明此区域毒株的数目，利用不同的颜色区分出不同差异较大的毒株。同时每个毒株标记点，可以通过交互点开，获得此毒株的详细信息，也将通过虚线标注出在地图中属于同一簇的毒株。

第 4 章 预期效果

4.1 分子进化树与时空分布

分子进化树与 MAFFT 功能相近，预期通过页面的跳转，从进化树跳转到具体序列差异的界面，如图 4-1 进化树与分子差异界面。

图 4-1 分子进化树与序列差异对比

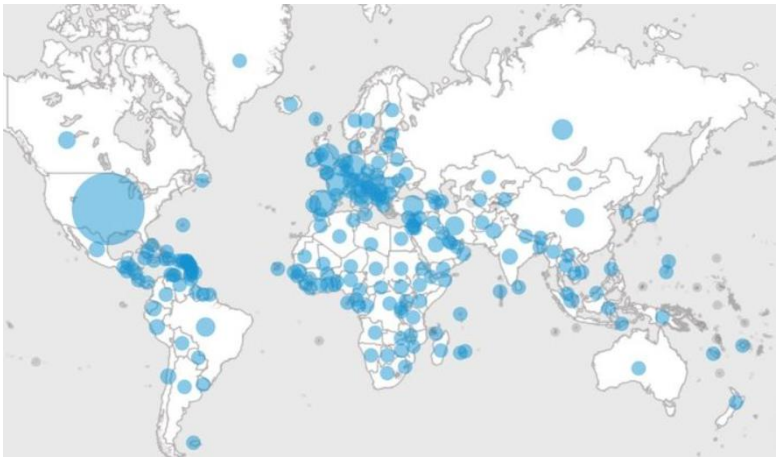

图 4-2 病毒时空分布

4.2 训练效果 3D 展示

序列预测训练集随时间与 F1-score 分数的 3D 可视化如图 4-3 所示

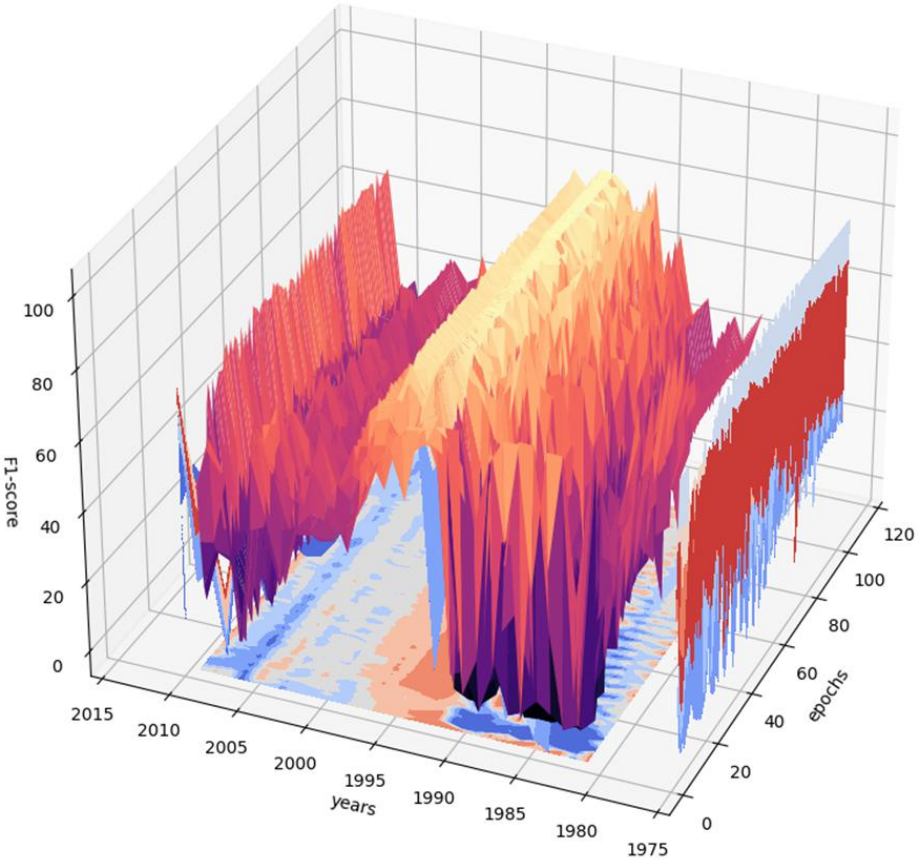

图 4-3 训练效果 3D 展示

第 5 章　预期效果与进度计划

5.1 成员分工

本小组成员分工表如表 5-1 所示

表 5-1 成员分工表

任务名称		小组成员
整体设计	发现问题转化需求	宋文凯
	绘制视图草图	宋文凯
项目建议书撰写	项目建议书撰写	宋文凯、秦阳、苏超、尹航
	项目建议书审校	宋文凯、秦阳、苏超、尹航
数据获取与处理	数据搜集与清洗	宋文凯、尹航
	数据格式整理与规范化	秦阳、苏超
	序列进化分析	宋文凯、尹航
	时空溯源	秦阳、苏超
系统基础架构	成员协作、版本控制与代码托管	苏超
	后端框架搭建	秦阳
	前端静态元素设计与实现	尹航
	在线环境部署部署	秦阳、苏超、尹航
可视化系统实现	基于序列的进化树	宋文凯
	病毒序列聚类分析	尹航、苏超
	序列差异分析	秦阳
	序列位点分析	秦阳
	抗原关系可视化	尹航
	回溯性分析可视化	苏超
系统整合与测试	项目整合	宋文凯、秦阳、苏超、尹航
	对系统进行测试与评估	宋文凯、秦阳、苏超、尹航
撰写项目总结报告	生成总结文档	宋文凯、秦阳、苏超、尹航

5.2 项目进度计划

项目进度计划表如表 5-2 所示

表 5-1 成员任务分配计划表

整体设计	发现问题转化需求	2021/10/09 - 2021/10/11
	绘制视图草图	2021/10/12 - 2021/10/14
项目建议书撰写	项目建议书撰写	2021/10/15 - 2021/10/18
	项目建议书审校	2021/10/18 - 2021/10/19
数据获取与处理	数据搜集与清洗	2021/10/20 - 2021/10/22
	数据格式整理与规范化	2021/10/23 - 2021/10/25
	序列进化分析	2021/10/26 - 2021/10/28
	时空溯源	2021/10/29 - 2021/10/30
系统基础架构	成员协作、版本控制与代码托管	2021/11/1 - 2021/11/2
	后端框架搭建	2021/11/3 - 2021/11/5
	前端静态元素设计与实现	2021/11/6 - 2021/11/9
	在线环境部署部署	2021/11/9 - 2021/11/10
可视化系统实现	基于序列的进化树	2021/11/11 - 2021/11/30
	病毒序列聚类分析	2021/11/11 - 2021/11/30
	序列差异分析	2021/11/11 - 2021/11/30
	序列位点分析	2021/11/11 - 2021/11/30
	抗原关系可视化	2021/11/11 - 2021/11/30
	回溯性分析可视化	2021/11/11 - 2021/11/30
系统整合与测试	项目整合	2021/12/1 - 2021/12/10
	对系统进行测试与评估	2021/12/11 - 2021/12/15
撰写项目总结报告	生成总结文档	2021/12/16 - 2021/12/23

甘地图如图 5-1 所示

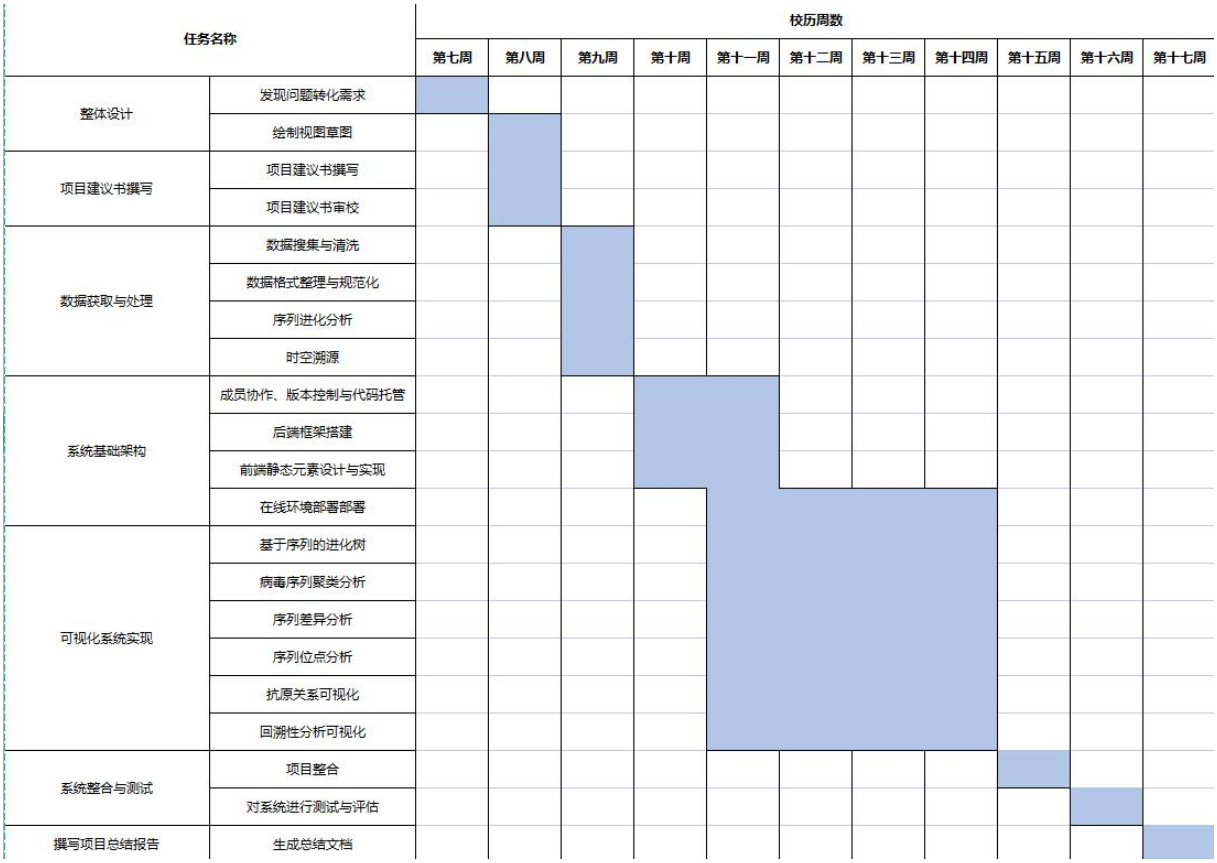

图 5-1 成员任务分配甘地图

参考文献

[1] Organization W H . Influenza (Seasonal): fact sheet. 2014.

[2] Trevor Bedford, Riley Steven, Barr Ian-G, et al. Global circulation patterns of seasonal influenza viruses vary with antigenic drift.[J]. Nature, 2015, 523(7559).

[3] M-Bouvier Nicole, Peter Palese. The biology of influenza viruses[J]. Vaccine, 2008, 26.

[4] J-L Virelizier. Host defenses against influenza virus: the role of anti-hemagglutinin antibody.[J]. Journal of immunology (Baltimore, Md. : 1950), 1975, 115(2).

[5] Kalus,Stöhr. Influenza—WHO cares[J]. Lancet Infectious Diseases, 2002.

[6] Claudia T , Daniele P , Stuart M , et al. Overview of Serological Techniques for Influenza Vaccine Evaluation: Past, Present and Future[J]. Vaccines, 2014, 2(4):707-734.

[7] Shu Y , Mccauley J . GISAID: Global initiative on sharing all influenza data – from vision to reality[J]. Eurosurveillance, 2017, 22(13).

[8] Seonwoo M , Byunghan L , Sungroh Y . Deep Learning in Bioinformatics[J]. Briefings in Bioinformatics, 2017(5):851.

[9] Nakaya H I , Wrammert J , Lee E K , et al. Systems biology of vaccination for seasonal influenza in humans.[J]. Nature Immunology.