華中科技大學

本科生毕业设计（论文）开题报告

题　　目：肠道噬菌体-细菌宿主关系的系统鉴定：方法评估与开发

院　　系	生命科学与技术学院
专业班级	生物信息学（国家基地班）201801 班
姓　　名	U201812416
学　　号	苏济雄
指导教师	陈卫华

2021 年　11 月

开题报告填写要求

一、 开题报告主要内容：

1.课题来源、目的、意义。

2.国内外研究现况及发展趋势。

3.预计达到的目标、关键理论和技术、主要研究内容、完成课题的方案及主要措施。

4.课题研究进度安排。

5.主要参考文献。

二、 报告内容用小四号宋体字编辑，采用 A4 号纸双面打印，封面与封底采用浅蓝色封面纸（卡纸）打印。要求内容明确，语句通顺。

三、 指导教师评语、教研室（系、所）或开题报告答辩小组审核意见用蓝、黑钢笔手写或小四号宋体字编辑，签名必须手写。

四、 理、工、医类要求字数在 3000 字左右，文、管类要求字数在 2000 字左右。

五、 开题报告应在第八学期第二周之前完成。

一、课题基本信息

1.1 课题来源

陈卫华教授课题组前期积累了大量肠道噬菌体和宏基因组数据。以此为基础，我们希望能对现有预测噬菌体-宿主作用关系的算法进行评估，通过整合现有算法，或开发新的算法，得到一套能准确预测噬菌体-细菌宿主作用关系的流程和工具。

1.2 课题目的

整合多种计算方法，或自行开发算法，得到一套可行的能准确预测噬菌体-细菌宿主作用关系的算法，并与其它方法比较具有优势。

1.3 课题背景及意义

肠道是人体最大的微生态系统，栖息着总数约 10^{14}、种类超过 1 000 余种、重量约为 1-2 公斤的微生物。这些肠道微生物编码基因的总数超过 330 万，约为人类编码基因总数的 100 倍，因此肠道微生物又被认为是人体的第二基因组，通过调节食物消化吸收、免疫、炎症、代谢等途径影响我们的健康[1]。

噬菌体是肠道病毒的主要成员，在调节肠道菌群的丰度和多样性方面发挥关键作用。噬菌体感染宿主细菌后，可杀死细菌或抑制细菌生长。并且由于其宿主范围窄，靶向性好，是理想的精准调控细菌的工具[2]。而其中的关键，是建立噬菌体与宿主的关系。

目前基于实验的方式可以研究噬菌体与宿主的作用关系，如噬菌斑测定、荧光标记噬菌体、液体检测、噬菌体 FISH、微流控 PCR 和单细胞测序等方法。但由于实验的方法只能使用可培养的微生物宿主和病毒，且成本高昂、耗费时间，存在一定的限制，难以大量推广应用。另一方面，目前测序技术的发展使得研究人员得以从环境样本中发现大量未培养的噬菌体及微生物，为全面研究病毒多样性和感染宿主范围提供了重要途径[3]。

本课题的研究意义在于通过生物信息学的手段，运用算法和模型，通过大规模计算，预测出噬菌体感染宿主范围。其研究成果后续可用于筛选出最有可能和最有价值的噬菌体-宿主关系对，经过实验验证用于临床，将大大加快将噬菌体成为理想的精准调控微生物工具的应用，以调节人体细菌的繁殖、生长，观察与

此微生物相关联的疾病的发生发展情况，将肠道微生物变化与疾病之间的关联转化为因果关系，达到通过调控微生物来干预疾病的目标，帮助医生进行精准的临床诊断和相关的治疗。

二、国内外研究现况及发展趋势

目前，国内外已开发出多种计算方法来预测噬菌体-宿主作用关系（以下简称为 PHI），可以大致分为三类：（1）基于序列同源性和序列相似性的序列比对方法；（2）基于序列组成和基因组特征的无比对方法；以及（3）基于机器学习的方法[4]。

序列比对的方法依赖于病毒和宿主相似或同源序列来计算预测病毒的宿主范围，比较经典的方式是使用局部比对工具 BLAST[5]。Edwards 等人[3]用一个由 820 个噬菌体和 153 个已知宿主组成的基准数据集以比较五种 PHI 预测方法，其中包括宏基因组中的噬菌体与细菌的丰度变化特征、遗传同源、精确匹配、CRISPR spacer 和寡核苷酸图谱等。这五种方法各有优劣，序列同源性方法在预测 PHI 方面最为有效，且发现基于核酸的序列同源寻找比基于蛋白质的序列同源寻找准确率要高，但该方法基于事先建立的参考数据库来确定哪些细菌与给定的噬菌体最相似，受制于数据库的全面性和完整性。同时指出基于 CRISPR spacer 预测 PHI 的方法，由于只有大约 40% 的细菌和 70% 的古细菌编码 CRISPR 系统，并且 CRISPR 阵列中的 spacer 可在环境中迅速变化，其假阳性率低，假阴性率高。

有些情况下，病毒序列和宿主序列可能缺乏序列同源性，因此不太适合基于比对的方法。这时，无比对方法通过研究序列组成模式的相似性（例如密码子使用模式或寡核苷酸频谱的相似性），为推断 PHI 提供了一种替代方案[3]。其方法原理是认为噬菌体基因组通过匹配其宿主的核苷酸组成、密码子使用模式，可以避免宿主限制-修饰系统（RM）的识别、更好合成病毒蛋白。Crane 等人[6]使用 COUSIN[7] 确定了 129 个感染分枝杆菌噬菌体的密码子使用偏好。VirHostMatcher(VHM) [8] 通过计算噬菌体和宿主的寡核苷酸频率模式之间的距离来预测 PHI；WIsH[9]在 VHM 的基础上，使用马尔可夫模型来减少运行时间；HostPhinder[10]则使用病毒-病毒相似性策略，假设病毒之间相似的寡核苷酸频谱则表示有共同宿主或相似的宿主。

近些年来预测 PHI 的计算方法倾向于应用机器学习的方法。为了推断 PHI，机器学习的方法利用提取“特征”来训练模型，例如病毒基因组的核苷酸和氨基酸含量，氨基酸特性和蛋白质结构域等。有研究人员使用各种传统机器学习分类器[11–13]，如随机森林、支持向量机、逻辑回归、k 近邻、多层感知器和朴素贝叶斯，基于核苷酸特征或蛋白质结构特征预测 PHI；PHP[14]计算病毒和宿主基因组序列之间的 k-mer 频率差异来训练高斯模型；PredPHI[15]使用氨基酸频率、化学组成、氨基酸相对分子量作为特征表示来预测 PHI。VirHostMatcher-Net[16]集成了多种特征，包括病毒-病毒相似性、病毒-宿主无比对的相似性、病毒-宿主比对的相似性以及病毒-宿主 CRISPR 的相似性，以预测 PHI。BacteriophageHostPrediction[17]使用 200 多个特征来表征噬菌体受体结合蛋白训练机器学习模型以预测噬菌宿主，特征包括基因组序列特征（如核苷酸和密码子频率和 GC 含量）、蛋白质序列特征（如氨基酸频率）、蛋白质二级结构（如 α 螺旋和 β 折叠）和物理化学特性（如分子量和等电点）。PHIAF[18]则融合了源自 DNA 和蛋白质序列的特征，应用了基于 GAN （生成对抗网络）的数据增强模块，该模块生成伪 PHI 以缓解数据稀缺性，利用注意力机制来考虑 DNA/蛋白质序列衍生特征的不同贡献，增强预测模型的可解释性。

虽然目前预测 PHI 的方法已有许多，但目前并没有方法涉及到只从基因的角度来试图联系噬菌体和宿主。本次研究可以尝试从水平基因转移的层面切入，通过寻找可能在 PHI 作用过程中多个重要的基因来预测 PHI，或许能进一步探索 PHI 背后的生物学机制。

目前通过噬菌体分类和培养得到的实验证据依然是确定 PHI 的黄金标准，因此有关噬菌体-宿主关系的信息仍然很少。NCBI Refseq 和 Genbank 数据库的部分噬菌体信息中有提供宿主名称，但这些记录大多只在属或种水平上识别宿主，没有确定菌株[19]。

高娜等人[20]构建了 MVP 数据库，提供了一个噬菌体-宿主相互作用的全面目录，并能帮助用户选择能够选择噬菌体来靶向感兴趣的特定微生物。其 PHI 主要来源于四个方面：1. 从 NCBI 的参考病毒数据库页面提取 host 字段；2. 从参考原核基因组中鉴定出原噬菌体建立噬菌体-宿主关系；3. 从宏基因组鉴定出原噬菌体 contig，再从 contig 的两边侧翼提取序列与原核基因组比对，确定宿主；

4. 收集来自其他公开数据集和数据库的 PHI 信息。最终基于 30,321 个证据建立了 18,608 个病毒簇和 9,245 个原核生物之间的 26,572 个相互作用。

Viral Host Range database[21]旨在收集相关实验数据，截至 2021 年 11 月，数据库已有 771 种病毒和 1955 种宿主共 17067 种相互作用。Virus-Host Database[22]数据库则整合 Refseq、Genbank、Uniprot、Viralzone 以及文献信息中含有的病毒与宿主信息，截至 2021 年 11 月，数据库已收集了 17297 条病毒宿主信息。本次研究中将利用这些公开数据库收集已有并得到验证的 PHI 数据。

三、关键理论和技术

（1）噬菌体基本知识

烈性噬菌体和温和噬菌体：根据噬菌体和宿主菌的关系，可将噬菌体分为两类：一类噬菌体在宿主菌细胞内迅速增殖，产生许多子代噬菌体，并最终使宿主菌细胞破裂，这类噬菌体被称为烈性噬菌体（virulent phage）；另一类噬菌体感染宿主菌后不立即增殖，而是将其核酸整合到宿主菌染色体中，随宿主核酸的复制而复制，并随细胞的分裂而传代，这类噬菌体被称作 温和噬菌体 （temperate phage)或溶原性噬菌（lysogenic phage）[23]。

Prophage：原噬菌体（prophage）指的是某些温和噬菌体侵染细菌后，其核酸整合到宿主细菌染色体中。噬菌体所整合的核酸称为 prophage。通过鉴定细菌基因组中包含的 prophage 能够较为准确地构建噬菌体-宿主作用关系。使用 Phage_Finder[24]、Virsorter[25]等软件可以鉴定出 prophage。

（2）研究 PHI 的基本方法和工具

BLAST：BLAST（Basic Local Alignment Search Tool）[5]是一套在 DNA 数据库或数据库中进行相似性比较的分析工具。利用 BLAST 工具可以针对细菌数据库寻找噬菌体相似的核苷酸、蛋白质序列。常用的 BLAST 方法有 BLASTN、BLASTX、BLASTP。BLASTN 是核酸序列到核酸库中的一种查询。库中存在的每条已知序列都将同所查序列作一对一地核酸序列比对。BLASTX 是核酸序列到蛋白库中的一种查询。先将核酸序列翻译成蛋白序列（一条核酸序列会被翻译成可能的六条蛋白），再对每一条作一对一的蛋白序列比对。BLASTP 是蛋白序列到蛋白库中的一种查询。库中存在的每条已知序列将逐一地同每条所查序列作

一对一的序列比对。

CRISPR：在某些古菌和细菌中，存在 CRISPR 系统（规律成簇的间隔短回文重复序列），可认为是原核生物中可获得性免疫功能器官。外源的遗传物质（病毒 DNA 或质粒）的片段会经过剪切放入 CRISPR 中间隔的 repeats 中成为 spacer，当外源遗传物质再次进入细菌之时，CRISPR 可以通过 CRISPR 的特异性识别直接攻击外源遗传物质，使其失去正常功能，从而达到免疫的作用。利用细菌和古菌含有的 CRISPR spacer 序列与噬菌体进行序列比对，可以预测 PHI，可使用的软件有 SpacePHARER[25]。

K-mer：原始的 DNA 序列数据通常长短不一，常存在长序列。为了减少处理数据的成本，通常使用 K-mer 预处理序列，同时能够使 DNA 序列更接近普通文本的词句结构。其方法是从原始序列第一个碱基开始，以一个碱基为单位每次向后一位，每次取一个长度为 K 的短序列。经过 K-mer 之后，一条长度为 L 的长序列就被转换成了 L-K+1 个短序列。通过调节 K 的值，可以提高模型的准确性。转换后的短序列可以更好地进行数据读取、特征提取、向量化，以适用于大数据分析与机器学习。

K-mer 频谱：K-mer 频谱可以评估基因组大小、杂合度、重复序列比例等[26]，也可用于鉴定两个序列的相似度、宏基因组的分箱中[27]。

四、课题实施方案

（1）具体方案

1. 通过公共数据库搜索，得到已知可靠的噬菌体-细菌作用关系，作为阳性数据集。同时收集没有侵染关系的阴性数据集。下载相关的噬菌体和细菌的基因组、蛋白质组、基因组注释等序列数据；

2. 了解和掌握目前已有的噬菌体和细菌关系计算预测方法，在阳性数据集上运行，以评估其性能、优劣。在此过程中需要有批判性思维，总结鉴定方法的优势和不足之处，做好记录，为开发一个性能更好的算法做好准备；

3. 结合已有的噬菌体-细菌作用关系预测算法优缺点，整合多种计算方法，或开发自己的算法，以准确预测噬菌体-宿主关系，提高预测精准性；

4. 收集人类肠道噬菌体和细菌的序列数据，将数据进行相关预处理。将训练好的算法应用于数据，为肠道噬菌体鉴定相应的宿主。通过文献检索对输出的评分

较高的噬菌体-宿主作用关系进行评估，判定算法在真实情况的实用性；

（2）预计达到的目标

1. 得到可靠的、去冗余的噬菌体-细菌作用关系数据库；

2. 掌握已有的判定噬菌体-细菌作用关系的预测算法，开发出自己的算法，并和已有算法相比有良好的预测性能、独特优点，以及尝试探究噬菌体-细菌作用关系背后隐藏的生物学意义；

3. 研究成果形成研究论文；

五、课题研究进度安排

由于本课题是一个研究性的课题，难以给出一个详细的时间计划表。在研究该课题的过程中，预计会花大量的时间进行文献阅读和评估已有鉴定算法，大致的进度安排如表 1 所示。本人会尽力提高学习效率，尽可能按计划完成项目。如果遇到困难会及时进行相应调整，力争顺利完成毕业设计。

表 1 课题研究进度安排表

学期	周次	工作任务
2021-2022 第一学期	第 1 周	指导教师布置任务，确定任务书基本框架；指导教师与毕设学生共同填写任务书的电子文档；学生开始做查阅资料和翻译工作
	第 2 － 10 周	完成英文翻译，开题报告初稿；
	第 11 － 19 周	使用上手数据了解宏基因组组装和功能注释流程；在导师指导下进行毕业设计工作。
2021-2022 第二学期	第 1 - 2 周	项目回顾，整理中期检查材料
	第 3-11 周	解读研究结果，形成数据库，完成毕业论文的撰写、定稿
	第 12-13 周	完成论文查重工作，准备毕设答辩

六、主要参考文献

[1] Thursby E, Juge N. Introduction to the human gut microbiota[J]. Biochemical Journal, 2017, 474(11): 1823-1836.

[2] Hyman P, Abedon S T. Bacteriophage host range and bacterial resistance[J]. Advances in applied Microbiology, 2010, 70: 217-248.

[3] Edwards R A, McNair K, Faust K, et al. Computational approaches to predict bacteriophage–host relationships[J]. FEMS microbiology Reviews, 2016, 40(2):

258-272.
[4] Versoza C J, Pfeifer S P. Computational Prediction of Bacteriophage Host Ranges[J]. Microorganisms, 2022, 10(1): 149.
[5] Johnson M, Zaretskaya I, Raytselis Y, et al. NCBI BLAST: a better web interface[J]. Nucleic acids Research, 2008, 36(suppl_2): W5-W9.
[6] Crane A, Versoza C J, Hua T, et al. Phylogenetic relationships and codon usage bias amongst cluster K mycobacteriophages[J]. G3, 2021, 11(11): jkab291.
[7] Bourret J, Alizon S, Bravo I G. COUSIN (COdon Usage Similarity INdex): a normalized measure of codon usage preferences[J]. Genome Biology and Evolution, 2019, 11(12): 3523-3528.
[8] Ahlgren N A, Ren J, Lu Y Y, et al. Alignment-free oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences[J]. Nucleic acids Research, 2017, 45(1): 39-53.
[9] Galiez C, Siebert M, Enault F, et al. WIsH: who is the host? Predicting prokaryotic hosts from metagenomic phage contigs[J]. Bioinformatics, 2017, 33(19): 3113-3114.
[10] Villarroel J, Kleinheinz K A, Jurtz V I, et al. HostPhinder: a phage host prediction tool[J]. Viruses, 2016, 8(5): 116.
[11] Leite D M C, Brochet X, Resch G, et al. Computational prediction of inter-species relationships through omics data analysis and machine learning[J]. BMC Bioinformatics, 2018, 19(14): 151-159.
[12] Zhang M, Yang L, Ren J, et al. Prediction of virus-host infectious association by supervised learning methods[J]. BMC Bioinformatics, 2017, 18(3): 143-154.
[13] Leite D M C, Lopez J F, Brochet X, et al. Exploration of multiclass and one-class learning methods for prediction of phage-bacteria interaction at strain level[C]//2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2018: 1818-1825.
[14] Lu C, Zhang Z, Cai Z, et al. Prokaryotic virus host predictor: a Gaussian model for host prediction of prokaryotic viruses in metagenomics[J]. BMC Biology, 2021, 19(1): 1-11.
[15] Li M, Wang Y, Li F, et al. A deep learning-based method for identification of bacteriophage-host interaction[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2020, 18(5): 1801-1810.
[16] Wang W, Ren J, Tang K, et al. A network-based integrated framework for predicting virus–prokaryote interactions[J]. NAR genomics and Bioinformatics, 2020, 2(2): lqaa044.
[17] Boeckaerts D, Stock M, Criel B, et al. Predicting bacteriophage hosts based on sequences of annotated receptor-binding proteins[J]. Scientific Reports, 2021, 11(1): 1-14.
[18] Li M, Zhang W. PHIAF: prediction of phage-host interactions with GAN-based data aug mentation and sequence-based feature fusion[J/OL]. Briefings in Bioinformatics, 2021, 23(1). https://doi.org/10.1093/bib/bbab348. DOI:10.1093/bib/bbab348.
[19] Sayers E W, Beck J, Bolton E E, et al. Database resources of the national center for biotechnology information[J]. Nucleic acids Research, 2021, 49(D1): D10.

[20] Gao N L, Zhang C, Zhang Z, et al. MVP: a microbe–phage interaction database[J]. Nucleic acids Research, 2018, 46(D1): D700-D707.
[21] Lamy-Besnier Q, Brancotte B, Brancotte H M, et al. Viral Host Range database, an online tool for recording, analyzing and disseminating virus–host interactions[J]. Bioinformatics, 2021, 37(17): 2798.
[22] Mihara T, Nishimura Y, Shimizu Y, et al. Linking virus genomes with host taxonomy[J]. Viruses, 2016, 8(3): 66.
[23] Echols H. Developmental pathways for the temperate phage: lysis vs lysogeny[J]. Annual review of Genetics, 1972, 6(1): 157-190.
[24] Fouts D E. Phage_Finder: automated identification and classification of prophage regions in complete bacterial genome sequences[J]. Nucleic acids Research, 2006, 34(20): 5839-5851.
[25] Roux S, Enault F, Hurwitz B L, et al. VirSorter: mining viral signal from microbial genomic data[J]. PeerJ, 2015, 3: e985.
[26] Hozza M, Vinař T, Brejová B. How big is that genome? Estimating genome size and coverage from k-mer abundance spectra[C]//International Symposium on String Processing and Information Retrieval. Springer, 2015: 199-209.
[27] Dubinkina V B, Ischenko D S, Ulyantsev V I, et al. Assessment of k-mer spectrum applicability for metagenomic dissimilarity analysis[J]. BMC Bioinformatics, 2016, 17(1): 1-11.