微生物学报
Acta Microbiologica Sinica
2024, 64(2): 344–363
http://journals.im.ac.cn/actamicrocn
DOI: 10.13343/j.cnki.wsxb.20230418

Review 综 述

基于智能计算噬菌体的细菌宿主范围预测

王博宇[1,2]，杨孜孜[1]，SUN Fengzhu[3]，王颖[1,2,4*]

1 厦门大学自动化系，福建 厦门 361000
2 厦门大学国家健康医疗大数据厦门研究院，福建 厦门 361000
3 美国南加州大学量化与计算生物系，加利福尼亚州 洛杉矶 CA90089
4 厦门市大数据智能分析与决策重点实验室，福建 厦门 361000

王博宇, 杨孜孜, SUN Fengzhu, 王颖. 基于智能计算噬菌体的细菌宿主范围预测[J]. 微生物学报, 2024, 64(2): 344-363.
WANG Boyu, YANG Zizi, SUN Fengzhu, WANG Ying. Progress in predicting bacteriophage host ranges by intelligent computing[J]. Acta Microbiologica Sinica, 2024, 64(2): 344-363.

摘　要：针对噬菌体的细菌宿主范围预测对于深入理解噬菌体及将其作为抗生素替代用于生物疗法具有重要意义。传统生物实验方法确定噬菌体的细菌宿主范围受到极有限的噬菌体可培养性和严苛的培养条件限制，而高通量测序技术所提供的海量基因组或宏基因组序列提供了噬菌体及细菌重要的序列信息，因此智能计算为预测噬菌体的细菌宿主范围提供了可行方法。本文从智能计算的角度对噬菌体的细菌宿主范围预测研究进行系统梳理，从噬菌体感染细菌的过程入手，描述配对预测模型所依赖的特征及其生物合理性，归纳宿主范围预测的智能模型、建模原理及预测策略，总结建模训练和评估所依赖的参考数据集与真实数据及评价指标。本文特别注意挖掘和分析各信息手段、模型、方法其背后的生物合理性及其依赖的生物机理。本综述期望推动基于智能算法的噬菌体的细菌宿主范围预测研究发展，并探索将生物先验结合人工智能实现噬菌体侵袭细菌宿主的本质机理推断，同时也为基于噬菌体的临床应用提供参考与借鉴。

关键词：微生物组；噬菌体-宿主相互作用；预测模型；智能算法；机器学习；神经网络

资助项目：国家自然科学基金(62173282)；国家重点研发计划(2018YFD0901401)
This work was supported by the National Natural Science Foundation of China (62173282) and the National Key Research and Development Program of China (2018YFD0901401).
*Corresponding author. Tel: +86-592-2182338, E-mail: wangying@xmu.edu.cn
Received: 2023-06-15; Accepted: 2023-08-17; Published online: 2023-08-29

Progress in predicting bacteriophage host ranges by intelligent computing

WANG Boyu[1,2], YANG Zizi[1], SUN Fengzhu[3], WANG Ying[1,2,4*]

1 Department of Automation, Xiamen University, Xiamen 361000, Fujian, China
2 National Institute for Data Science in Health and Medicine, Xiamen University, Xiamen 361000, Fujian, China
3 Department of Quantitative and Computational Biology, University of Southern California, Los Angeles CA90089, California, USA
4 Xiamen Key Laboratory of Big Data Intelligent Analysis and Decision, Xiamen 361000, Fujian, China

Abstract: The prediction of bacteriophage host ranges is of great significance for the basic research and clinical application of bacteriophages. The conventional biological experimental methods are limited by the poor culturability of bacteriophages and strict cultivation conditions. The availability of massive genome or metagenome sequencing data provides the sequence signature of bacteriophages and bacteria. Therefore, intelligent computing serves as a feasible way to predict bacteriophage host ranges. This paper systematically reviews the studies about intelligent computing-based prediction of bacteriophage host ranges. Starting from the process of bacteriophage infecting bacteria, we describe the feature source and biological rationality of prediction models, analyze the typical intelligent models and their prediction principles, and list all the reference datasets, real-world datasets, and evaluation indicators. The review aims to improve the understanding on the mechanism of bacteriophages in invading bacterial hosts and promote the usage of bacteriophages as antibiotic substitutes in biological therapy.

Keywords: microbiome; bacteriophage-host interactions; prediction model; intelligent computing; machine learning; neural network

噬菌体是地球上最丰富、最多样化的生物实体[1-2]。噬菌体在核苷酸序列水平上具有显著的多样性，这很大程度是由于在噬菌体和细菌共同进化的过程中，噬菌体会进行基因交换以应对选择宿主的压力。因此，噬菌体具有严格的宿主特异性，某一种噬菌体并不会感染所有细菌，往往只能入侵特定的细菌菌株，且具有较高的感染率，随着感染周期的迭代，被感染的细菌数量会呈指数型增长。根据噬菌体的特性、感染的细菌以及互作的环境，感染暴发的大小会发生显著变化[3]。由于噬菌体疗法具有特异性杀菌、易生长、适合基因操作改造等特点，因此研究噬菌体的细菌宿主范围预测对于噬菌体的基础研究和临床应用都有重要的意义。

噬菌体由遗传物质(DNA 或 RNA)以及保护其遗传物质的形状各异的蛋白质外壳组成，没有细胞结构，只有寄生于宿主才能复制。虽然地球上噬菌体的数目估计为 10^{31} 个[1]，但科学家能够获取的噬菌体与细菌数据却很少，截至 2023 年 7 月 13 日，NCBI RefSeq 数据库仅包含 28 676 个病毒基因组和 45 712 个细菌基因组，其中噬菌体基因组的数目为 10 400 个，占全部病毒基因组的 36%。实验室培养是探索病毒与其细菌宿主关系的重要手段[4]，然而自然环境中只有 1% 的微生物细胞是可培养的[5]，可用于培养的细菌宿主也非常有限，且一些噬菌体的培养条件极为

苛刻，一些溶原性噬菌体即使培养成功也难以观察和检测。因此，通过实验室培养获得大量噬菌体的细菌宿主范围非常困难。

高通量测序(high-throughput sequencing)技术的出现使得微生物基因组及群落宏基因组(metagenomics)数据大量涌现，使得原本无法通过生物培养的噬菌体及宿主基因组在测序数据中体现。然而，测序数据中只能拼装获得噬菌体或细菌基因组序列片段，无法获得预测噬菌体对应的宿主。尽管噬菌体基因组相对较小，但基因组的频繁重组导致的镶嵌现象(mosaicism)使其表现出显著的基因组多样性和复杂的进化关系[6]。

因此，基于计算方法的噬菌体宿主范围预测具有重要意义。这些计算模型往往基于噬菌体及细菌宿主的基因组序列提取特征并进行关联分析。噬菌体与宿主在共同进化过程中可能出现基因交换，或在感染与被感染的对抗过程中产生分子信号，例如病毒与宿主基因的相同片段、病毒与宿主的规律间隔成簇短回文重复序列(clustered regularly interspaced short palindromic repeats, CRISPR)匹配，以及序列组成分布的相似性等[7]。以上这些信息都能作为智能计算的特征用于预测噬菌体的细菌宿主范围。本文对基于智能计算的噬菌体的细菌宿主范围预测研究方法进行系统梳理，从噬菌体感染细菌的过程入手，归纳出用于宿主预测的特征、智能模型以及配对策略。此外，本文注重模型所体现的生物机理和生物合理性，整理现有研究所用到的参考数据集、真实数据集及评价指标，并讨论目前研究中遇到的瓶颈，展望可能的发展前景。

1 噬菌体感染细菌的生物过程

了解噬菌体感染细菌的生物过程，可以为特征的选择和抽取提供生物先验知识，在模型构建中融入生物机理，提升模型的可解释性和生物合理性，有助于提升噬菌体与细菌配对关系预测的准确性和泛化性。

大多数噬菌体由核酸与蛋白质外壳构成，绝大多数噬菌体的遗传物质是 DNA[8]。蛋白质外壳的主要作用是保护噬菌体的遗传物质，此外，大部分噬菌体的蛋白质外壳还有尾部结构，用来将遗传物质注入宿主体内。对所有原核细胞的功能至关重要的结构有细胞膜、细胞质、核糖体以及核质体(nucleoid)，大多数还具有细胞壁和多种形式的表面涂层或糖萼(glycocalyx)，在一些细菌中发现的特殊结构还有鞭毛、纤毛、菌毛等。

噬菌体感染细菌大致可以分为吸附、注射、遗传物质复制与蛋白质合成、子代噬菌体的装配和子代噬菌体的释放这 5 个阶段，如图 1 所示。图 1 右侧描述的是溶源周期，噬菌体在吸附和侵入宿主细胞后，将自身遗传物质整合在宿主的遗传物质中(或以质粒形式存在细胞内)，随着宿主 DNA 的复制而同步复制自己的遗传物质，并随宿主细胞分裂而将遗传物质传递到 2 个子细胞中，宿主细胞则可正常繁殖。图 1 左侧描述的是溶菌周期，宿主细胞内进行噬菌体的遗传物质复制及蛋白质的合成，并组装成噬菌体颗粒，最终将宿主细菌裂解，释放子代噬菌体。

1.1 吸附

噬菌体通过受体识别，其蛋白质与细菌细胞表面受体之间发生相互作用，从而吸附到细菌的细胞壁表面，进而感染细菌。因此，噬菌体吸附细菌的特异性决定了噬菌体是否能够入侵特定的细菌[9]。细菌和病毒基因组序列包含蛋白质及受体信息，可以作为判断噬菌体宿主配对关系的信息来源。

1.2 注射

噬菌体与细菌受体结合后，通过肽聚糖降解酶来降解细菌宿主的细胞壁，通过细长尾壳进行类似注射器的运动将遗传物质注射进入宿主细胞内，有些没有细长尾壳的病毒则是在插入遗传物质之前利用细小的齿状纤维将部分细胞膜进行酶降解。

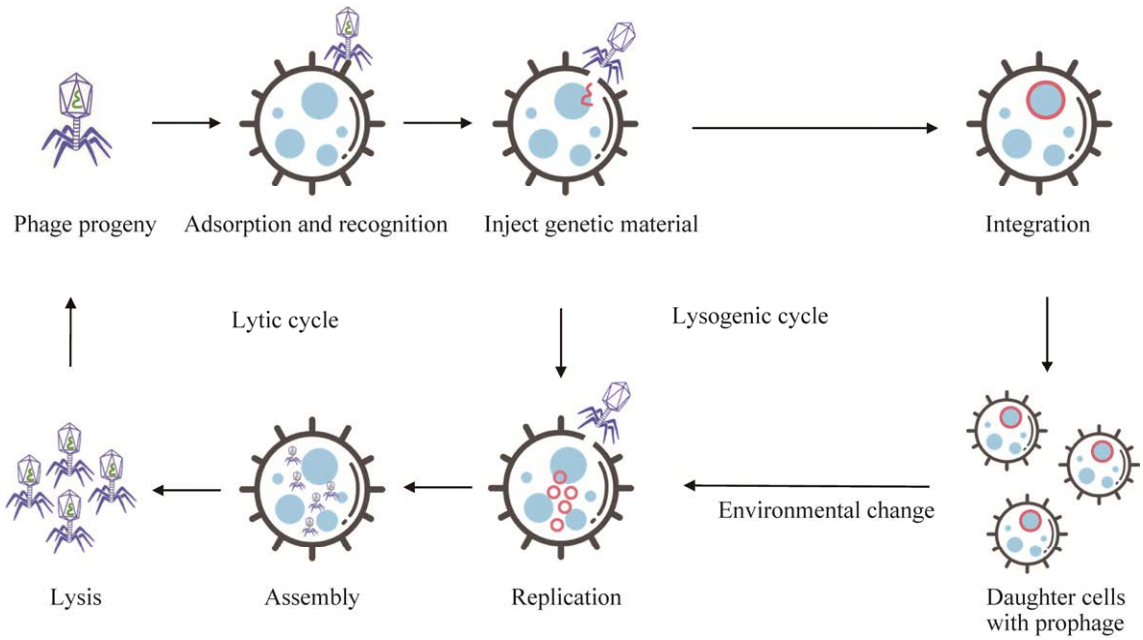

图 1　噬菌体感染细菌过程示意图

Figure 1　Diagram of bacteriophage infection process. Cycle left is the lytic process, and cycle right is the lysogenic process. Lytic cycle destroys bacteria and releases their progeny bacteriophages. Some lysogenic bacteriophages integrate their own genetic material into the nucleoid of the host bacteria, or exist in the form of plasmids within the cells. They replicate with the host cell and pass on to daughter cells as the host cell divides, known as the "lysogenic cycle".

1.3　遗传物质复制与蛋白质合成

遗传物质进入宿主细胞后，一方面，噬菌体遗传物质首先转录为 mRNA，细菌核糖体将噬菌体 mRNA 翻译为噬菌体所需的蛋白质，包括新噬菌体的蛋白质外壳、有助于新噬菌体组装的辅助酶或促使宿主细胞裂解的催化酶；另一方面，注射进宿主细胞的亲代噬菌体核酸被作为模板，在核酸聚合酶的作用下，大量复制子代噬菌体的遗传物质。

1.4　子代噬菌体的装配

新的遗传物质与蛋白质外壳被组装成新的子代噬菌体，蛋白质外壳的底板、头部、尾部与遗传物质在辅助蛋白质的作用下组装成新的子代，遗传物质被保护在蛋白质外壳的头部。

1.5　子代噬菌体的释放

溶菌周期是多数噬菌体的释放手段，这一过程通过摧毁细胞释放新合成的病毒粒子。当复制得到的子代噬菌体达到一定数量时，噬菌体会使宿主细菌开始表达内溶素，内溶素会攻击并破坏细胞壁肽聚糖，细菌细胞裂解释放出的噬菌体去感染其他目标细菌；溶源周期中一些溶原性噬菌体将自身基因组整合到宿主细菌的染色体上，或以质粒形式存在细胞内，随宿主 DNA 同步复制，随宿主细胞分裂而传递至 2 个子细胞中，这一行为造成部分噬菌体与其宿主具有极高相似性的 DNA 片段，通过捕捉这些相似性极高的 DNA 片段可以有效判断噬菌体的宿主范围。

综上所述，细菌被噬菌体感染时会被动地帮助病毒复制子代，并且多数情况下自身会裂解。但是细菌并不会完全任由噬菌体感染，在长期的进化过程中，细菌进化出了多种与原核生物进行斗争的免疫武器，包括 CRISPR[10]、限制-修饰(restriction-modification, RM)[11]、化学防御(chemical defense)[12]、流产感染(abortive infection, Abi)[13]等。

CRISPR 是原核生物基因组中的序列，来源于曾经感染过它们的噬菌体的 DNA 片段。此后，细菌被类似噬菌体入侵时，可以通过这些 CRISPR 序列检测并破坏噬菌体的 DNA。研究表明，在大约 45%的细菌基因组中发现了 CRISPR 序列[10]，大多数来自噬菌体基因组[14]，其中大约 7%的可检测 CRISPR 序列可以与已知的噬菌体序列匹配。因此，在已知噬菌体与宿主细菌的基因组序列信息时，可以通过与 CRISPR 序列比对情况识别噬菌体与宿主的亲缘关系。此外，RM 系统通过识别 DNA 的特殊位点来区分细菌自身 DNA 和噬菌体 DNA，并最终通过限制内切酶切割噬菌体所注入的 DNA，阻止噬菌体 DNA 复制[11]。化学防御是指细菌通过分泌小分子来抑制噬菌体的复制，且不会影响细菌的正常生长繁殖，如链霉菌分泌的阿霉素和柔红霉素[12]。流产感染是指在细菌感知到被噬菌体感染后，在噬菌体完成复制周期前自杀，以确保没有成熟的子代噬菌体颗粒被释放，从而避免新的细菌被感染；如果噬菌体在感染早期便被细菌的其他防御系统清除，那么流产感染将不会被启动[13]。

2 噬菌体的细菌宿主范围预测智能计算模型框架

通过智能算法预测噬菌体的细菌宿主范围构建思路如图 2 所示，首先获取噬菌体和细菌的

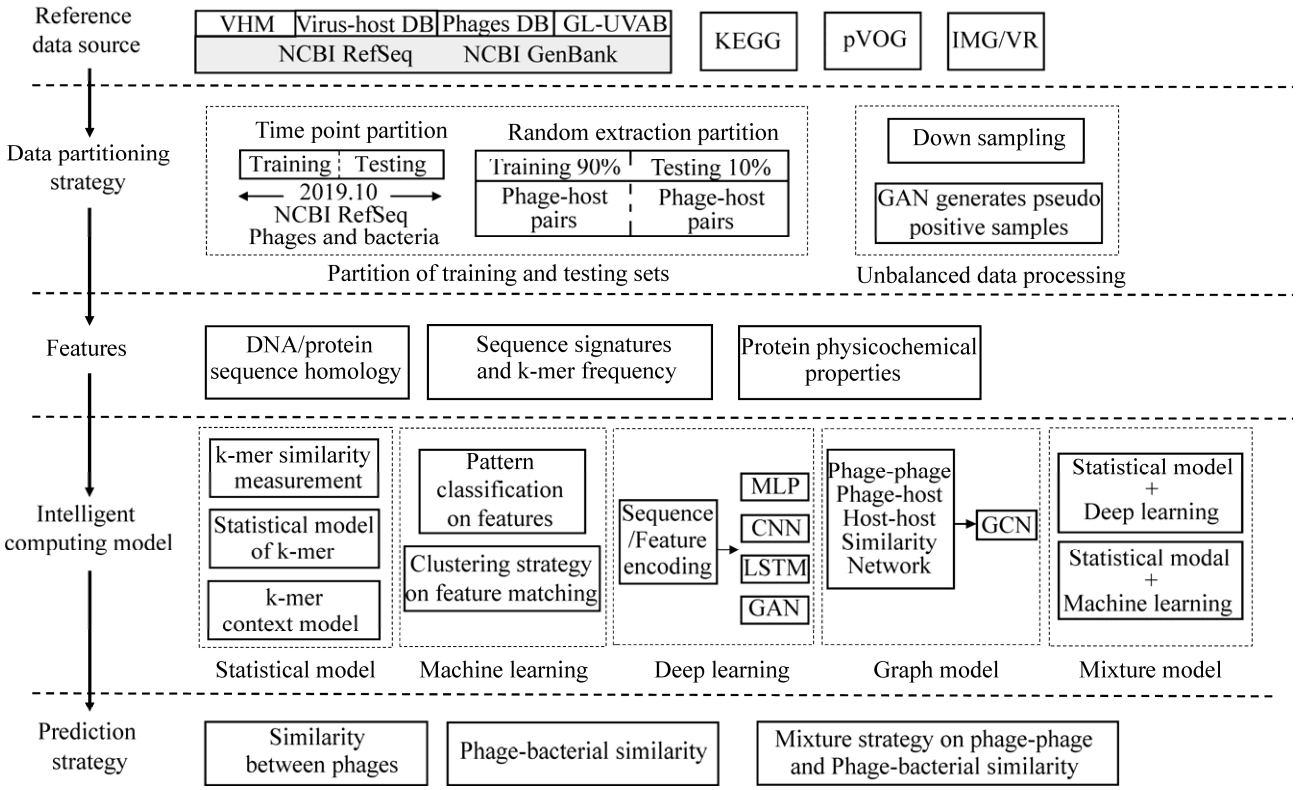

图 2　噬菌体宿主范围预测的智能计算模型框架

Figure 2　An intelligent computing model framework for bacteriophage host ranges prediction. The process of bacteriophage host ranges prediction is summarized from top to bottom. Based on various reference data sources, training and testing dataset are divided by different strategies. And then, different types of features are extracted, and different intelligent algorithms are used to construct computational models. Finally, different prediction strategies are used to predict the host range of new phages.

参考序列及配对信息，主要来自 NCBI RefSeq 数据库[15]；接着通过训练集和测试集的划分、不平衡的数据的处理，为后续的模型训练提供平衡、无信息泄露的数据集合；进一步基于生物先验知识、序列特性等提取具有生物合理性的特征，并转化为数值向量，输入各类智能模型进行训练，不同模型使用的特征不同、优势不同、所适应的场景侧重点也不同；基于训练好的模型，可以通过不同的预测策略进行未知噬菌体的宿主范围预测。

接下来，将从特征抽取、模型构建及配对策略 3 个步骤来描述用于噬菌体的宿主范围预测的智能模型技术，并在实验设计与评估环节讨论参考数据、测试数据、评估方案等环节中的方法与策略。

3 基于智能计算的噬菌体宿主范围预测的特征提取

提出反映噬菌体与宿主关系的特征对于噬菌体的宿主范围预测十分重要，特征主要来源于噬菌体及细菌的基因组 DNA 序列、蛋白质序列及理化信息，如表 1 所示。

表 1 噬菌体宿主范围预测使用的特征来源、形式、机理及抽取方式

Table 1 Sources, forms, mechanisms and extraction methods of features used for bacteriophage host ranges prediction

Source of features	Features representation	Mechanism of features	Extracting methods	References
DNA sequence homology	CRISPR	Prokaryotic immune system that confers resistance to phages	Short sequence alignment	PHISDetector[16], SpacePHARER[17], HostG[18], CHERRY[19]
	AMG	Duplication of phage components	Short sequence alignment	RaFAH[20]
DNA sequence	k-mer frequency	Phages share similar patterns in coden usage with their hosts	Counting k-mer of genomes	HostPhinder[21], VHM[22], PHP[23], PHIAF[24], WIsH[25]
	DNA encoding	Phages share similar patterns in coden usage with their hosts	Encoding genome	VIDHOP[26], DeepHost[27], ContigNet[28]
Protein sequence	RBP	Specific receptors on the bacterial surface	Protein-related features	Boeckaerts[29]
	HMM profile	Phages share similar proteins are more likely to infect similar hosts	Search against protein databases	RaFAH[20], vHULK[30], VPF-Class[31]
	Amino acid frequency	Interactions between phage and host due to interactions between their encoded proteins	Counting amino acid frequency of genomes from phage and host	Leite[32], PredPhi[33]
Protein physicochemical properties	Protein abundance of the chemical elements, protein molecular weight	Interactions between phage and host due to interactions between their encoded proteins	iFeature[34]	Leite[32], PredPhi[33], PHIAF[24]

3.1 基于 DNA 序列同源性的特征

细菌、古细菌和噬菌体的基因组序列可用于衡量噬菌体与候选宿主之间 DNA 序列的局部相似性，包括宿主编码的 CRISPR 区域、辅助代谢基因(auxiliary metabolic genes, AMG)、噬菌体的受体结合蛋白(receptor binding protein, RBP)等，它们分别反映噬菌体感染宿主细菌不同的生物步骤。

细菌序列中的 CRISPR 序列源于曾经感染过细菌的噬菌体规则间隔短回文重复 DNA 序列[35]，在遇到同样噬菌体入侵时，可通过 CRISPR 序列识别噬菌体。因此，DNA 到 DNA 或者 DNA 到氨基酸的序列比对可以作为有效的特征，判断细菌序列中是否包含 CRISPR 区域[17,36]，从而确定噬菌体与细菌的匹配关系。

AMG 是一类起源于细菌并存在于许多噬菌体中的基因，绝大多数 AMG 被认为是从宿主细菌中获得[37]。这类基因在感染过程中调节宿主细胞的代谢，从而使噬菌体能够更有效地复制。因此，可通过序列比对判断噬菌体是否包含相关 AMG 作为宿主预测的特征。

噬菌体吸附到宿主细菌表面的结构称为受体结合蛋白(RBP)。同一噬菌体可具有多个 RBP，用于识别细菌特异性受体并与之结合，RBP 识别宿主细菌表面的特异性受体，如多糖、蛋白质、鞭毛等，帮助噬菌体将遗传物质注入宿主[29]。因此，基于噬菌体蛋白家族数据库如 pVOGs (prokaryotic virus orthologous groups)[38]及其对应的细菌宿主信息作为参考信息，可将噬菌体 DNA 匹配到受体蛋白质数据库中，并将匹配结果作为特征，通过这些特征对具有类似 RBP 的噬菌体进行宿主范围预测。然而已知的 RBP 数据较少，在 Boeckaerts 等构建的数据库中共有 887 条，且仅与 7 种细菌宿主关联[29]，该方法依赖于参考信息，只能对有 RBP 数据的细菌宿主进行判断，因此现有方法中利用 RBP 进行广泛的宿主范围预测的模型较少，能够预测的宿主范围有限。

3.2 基于碱基或氨基酸序列的特征

大多数噬菌体不包含已标记的 CRISPR 等信息，无法通过序列同源性匹配进行宿主范围预测。由于大多数噬菌体自身不包含 tRNA，导致其基因翻译严重依赖于宿主提供的 tRNA 库，因此噬菌体基因组会体现其宿主的密码子偏好[39]。基于该生物前提，可通过直接提取噬菌体与细菌的碱基或编码氨基酸序列的 k-mer 频来比较噬菌体与细菌的序列分布相似性，进而预测噬菌体的宿主范围。所谓 k-mer 是指序列中长度为 k 的子序列，其频度反映序列密码子模式的特征。不同长度 k-mer 可得到不同尺度、不同侧重点的信息特征，例如 k 值为 2−10 时，k-mer 主要度量微生物在宏基因组、基因组层面的统计分布差异[40]，而 k≥15 的 k-mer 主要识别基因组序列在局部差异[41]，因此可以通过合适的度量模型学习任意已知、未知的噬菌体和细菌的配对关系度量[42]。

除了 k-mer 频度以外，碱基和蛋白质的序列也可通过直接序列编码进行表征。VIDHOP[26]采用 one-hot 编码基因组，DeepHost[27]考虑碱基插入、删除、突变的影响采用不同间隔的 k-mer，ContigNet[28]利用 one-hot 编码 DNA 序列和氨基酸序列。

3.3 基于蛋白质及其理化性质的特征

蛋白质的性质和功能特异性很大程度来源于其结构特异性。这种特异性不仅体现在氨基酸序列层面，也体现在物理化学性质上，包括蛋白质长度、分子量、等电点、芳香性以及描述蛋白质二级结构的存在于 α-螺旋、β-折叠或 β-转角中的氨基酸组分等，可作为辅助特征用于宿主预测。PredPhi[33]使用氨基酸的比例(20 种氨基酸加

未知氨基酸，21 维)、化学元素的丰度(碳、氢、氧、氮和硫，5 维)、蛋白质的分子量(1 维)组合成特征向量，并计算平均值、标准差、最大值、最小值、中位数和方差等统计量，作为深度网络的输入向量。iFeature[34]是用于提取蛋白质/肽序列的各种序列特征、结构特征和物理化学特征的 Python 工具包。

4 噬菌体宿主范围预测的智能模型

基于所抽取的噬菌体与细菌特征，可以通过不同的模型设计算法和预测逻辑建立模型。模型的类型大体可分为 4 类：统计模型、传统机器学习、深度学习以及图网络模型。图 3 给出不同类型的噬菌体宿主范围预测模型的提出与发展时间线。

4.1 基于统计模型的噬菌体宿主范围预测

统计模型通过计算噬菌体与细菌基因组序列的匹配度或序列相异度进行宿主范围预测。一方面主要基于噬菌体和细菌的基因组序列 k-mer 频度向量，用不同的相异度度量计算基因组序列的相似性；另一方面基于序列建立隐马尔可夫等统计模型衡量噬菌体和细菌序列的匹配度。基于统计模型的噬菌体宿主范围预测最大的优点是计算量小，无需训练复杂的模型，可以针对未见过的噬菌体和细菌直接计算其相异度或匹配度。

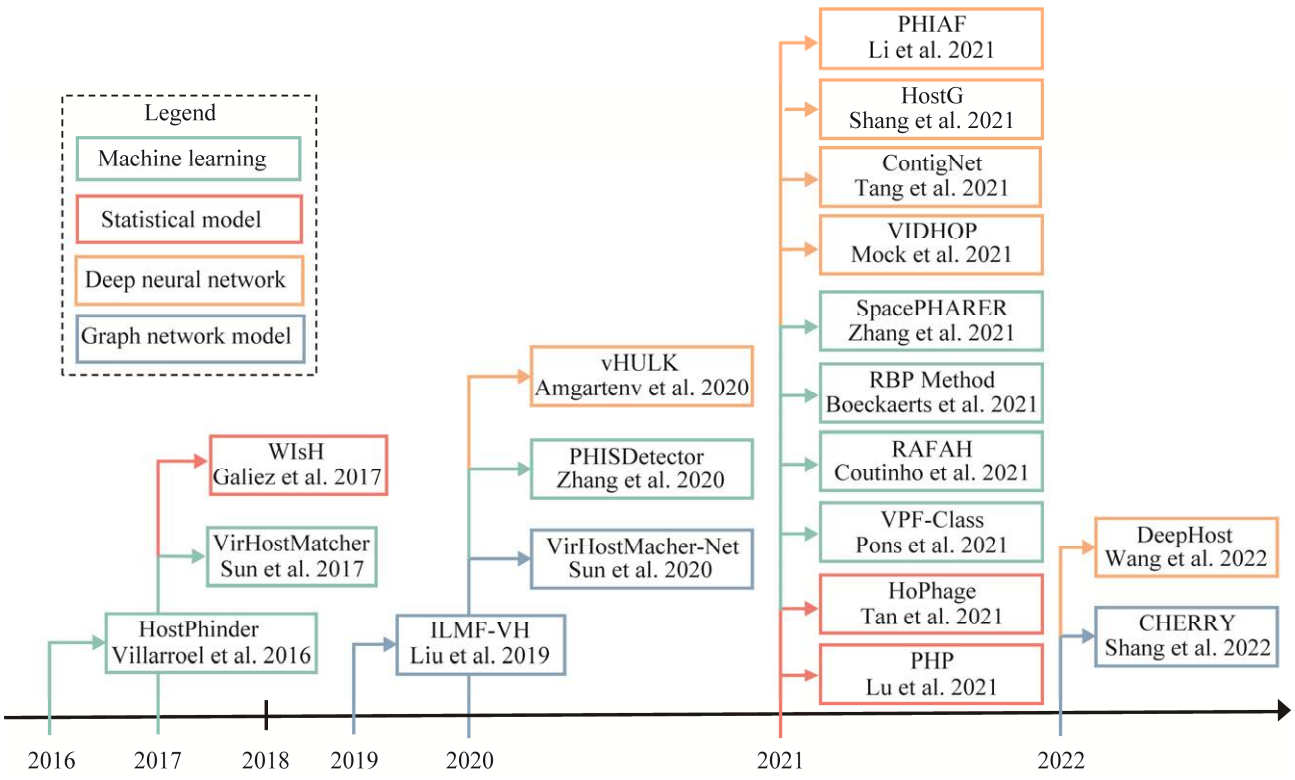

图 3 噬菌体宿主范围预测模型的发展线索

Figure 3 The development of bacteriophage host range prediction models. The horizontal axis is the time of model construction, and colors represent model categories.

4.1.1 基于 k-mer 频度的相异度度量指标

多数统计预测模型采用 k-mer 频度作为特征，这些基于统计模型和 k-mer 的方法前提假设是噬菌体及其宿主的共同进化导致它们共享某种相似的 k-mer 频度分布。许多研究基于噬菌体和细菌的 k-mer 频度向量，通过给定的相似性度量方法计算噬菌体与宿主、噬菌体与噬菌体或宿主与宿主之间的序列相似性，进而实现宿主范围预测。

由 Sun 团队所提出的 VirHostMatcher (VHM)[22]最早探索基于 k-mer 频度向量距离预测噬菌体的宿主范围。VHM 针对不同长度 k-mer，采用 11 种距离度量方式计算其相异度，主要分为两大类：一类是直接基于观测到的 k-mer 频度计算欧氏距离、曼哈顿距离、切比雪夫距离、d_2[43]和 JS 距离 (Jensen-Shannon divergence)[44]等，实验表明这类距离测量方式能够将噬菌体及其各自的宿主聚在一起[45]；另一类是考虑噬菌体和宿主的基因组背景 k-mer 频率分布的统计模型，包括 d_2^*、d_2^S [46-47]、Hao[48-49]、Teeling[50]、EuF[45]、MetaGO[51]和 Willner[52]等，能够去除序列中由背景模型带来的偏差。

4.1.2 基于 k-mer 频度差异的统计分布

对于 2 组向量，可通过计算这 2 组向量来自同一个统计分布的似然比来判断其匹配性。PHP[23]直接使用噬菌体与其宿主基因组 DNA 序列的 4-mer 频率差异作为特征，通过 k-means 聚类，假设相同类别下频度差异服从同一个高斯分布，计算未知噬菌体与 31 918 个备选细菌的 4-mer 频度差异在高斯分布中的似然比得分进行宿主范围预测。

4.1.3 基于 k-mer 频度的马尔可夫模型

马尔可夫模型假定系统第 k+1 个状态只与前 k 个状态有关。在噬菌体宿主范围预测中，针对细菌基因组序列建立 k-mer 频度的马尔可夫模型[53]，而后计算噬菌体出现在相应宿主模型下的概率，进而判断两者的相互作用关系。

WIsH[25]针对每个候选细菌基因组 H 建立 8 阶齐次马尔可夫模型，计算给定 k-mer 序列 $x_1x_2\cdots x_k$ 下观察到下一个碱基为 x_{k+1} 的概率 $p_H(x_{k+1}|x_1x_2\cdots x_k)$，而后对于噬菌体 Φ 的序列 $y_1y_2\cdots y_N$，计算 H 下的对数似然概率 $L(\Phi|H)=\frac{1}{N-k}\sum_{i=1}^{N-k}\log_{10} p_H(y_{i+k}\,|\,y_i\cdots y_{i+k-1})$，最终得到每一对噬菌体细菌匹配的概率。

统计模型通过构建噬菌体与宿主的相似性模型预测噬菌体的宿主范围，计算相似性度量、高斯模型、马尔可夫模型等量化噬菌体宿主的匹配度，无需将噬菌体宿主对进行训练，可以对任意噬菌体进行宿主范围预测。

4.2 基于机器学习的噬菌体宿主范围预测

机器学习对噬菌体宿主范围的预测主要分为有监督分类和无监督聚类 2 种策略，有监督分类将噬菌体宿主对看作训练样本，将细菌宿主作为类别，进行有监督分类；无监督聚类通过聚类信息量化噬菌体与聚类中心的距离，通过现有聚类所对应的已知宿主信息进行新噬菌体的宿主范围预测。

4.2.1 基于不同特征的传统机器学习模型

用于噬菌体宿主范围预测的传统机器学习模型主要包括逻辑回归(logistic regression)、支持向量机(support vector machines, SVM)、随机森林(random forest)和朴素贝叶斯(naive Bayes)模型等，主要将宿主视为有监督问题中的类别，使用已知的噬菌体宿主配对关系作为训练样本，基于不同的特征和模型，训练分类器预测噬菌体的宿主类别。

在 Sun 团队的另一项研究中[54]，使用噬菌体基因组数据对细菌宿主的 9 个属分支建立分类模型，并针对不同 k-mer 长度(k=4, 6, 8)的特

征设计逻辑回归、支持向量机、随机森林、高斯朴素贝叶斯和伯努利朴素贝叶斯 5 种机器学习模型，并发现采用观测 k-mer 频度减去 k-mer 的期望频度，可以去除基因组背景分布，提高分类性能。Young 等[55]基于核苷酸、氨基酸、氨基酸特性和蛋白质结构域所提取的 20 类特征，包括 k 值为 1−9 的 DNA k-mer 频度、k 值为 1−4 的编码区氨基酸 k-mer 频度、k 值为 1−6 的氨基酸序列对应的理学性质(每种氨基酸残基根据其物理化学性质分装到 7 个 bin 中)以及蛋白质结构域数目，训练支持向量机进行宿主范围预测。

PHISDetector[16]基于序列组成相似性、CRISPR 匹配、原噬菌体、遗传同源性(BLASTn, BLASTp)和蛋白质相互作用关系(protein-protein interaction)或蛋白域相互作用关系(domain-domain interaction)分析计算出 5 类 18 个特征，训练决策树、逻辑回归、支持向量机、高斯朴素贝叶斯和伯努利朴素贝叶斯等机器学习模型，并通过集成学习进行噬菌体的宿主范围预测。

Boeckaerts[29]团队基于受体结合蛋白的 887 条噬菌体序列数据来预测噬菌体宿主范围，通过提取 133 个 DNA 和氨基酸序列特征(核苷酸频率、GC 含量、密码子频率和密码子使用偏差)、20 个氨基酸相对丰度特征、15 个序列物理化学性质特征(蛋白质长度、分子量、等电点和芳香性等)和 3 个蛋白质二级结构特征，以及 Chen 等[34]提出的 47 个描述蛋白质序列的特征(组成、转换和 Z-scale 特征及对应的蛋白质序列等)共 218 维的特征向量，设计线性判别分析(linear discriminant analysis, LDA)、逻辑回归、随机森林和梯度提升(gradient boosting)模型，但该模型只能适用于受体结合蛋白已知的情况。

Boeckaerts 团队在另一项研究中[56]讨论了反映噬菌体感染细菌生物过程的多层机器学习概念模型，第一层建模噬菌体受体结合蛋白和细菌宿主受体之间相互作用；第二层建模噬菌体和细菌之间相互拮抗过程的进化动力学特性，包括 CRISPR-Cas、DISARM、BREX 和流产感染(abortive infection)等；第三层关注噬菌体对宿主代谢的操纵(hijacking)和转化，通过层次化模型集成噬菌体宿主范围的最终预测，并可以评估不同层次的相对贡献。

4.2.2 基于特征匹配的聚类策略

无监督的聚类方法首先针对已知宿主的噬菌体进行聚类，而后计算待查询噬菌体与每个聚类之间的距离。

HostPhinder[21]基于基因序列相似性越高的噬菌体所对应的细菌宿主越接近的前提，以 16-mer 频度向量为特征对噬菌体进行聚类；每个聚类产生一个代表基因组列表，称为种子(seeds)；通过每次加入的新序列与列表中种子序列的相似性(重叠的 16-mer)判断其所属的聚类或成立新的聚类；待查询噬菌体通过与已知聚类代表种子的相似性判断其对应的细菌宿主；但该模型无法处理宿主未出现在参考噬菌体-宿主配对数据中的情况。

SpacePHARER[17]在蛋白质水平上比较 CRISPR 区域和噬菌体的匹配关系，用集合 Q 表示一个细菌基因组待查询的 CRISPR 区域集合，元素 q 为细菌基因组 CRISPR 间隔区翻译所得的开放阅读框(open reading frame, ORF)序列，集合 T 为噬菌体蛋白质组目标集，由若干噬菌体蛋白序列 t 组成。该方法统计 q 与 t 序列蛋白质水平比对配准一致性和配上(hit)次数评估噬菌体和细菌的匹配程度。

针对宏基因组测序数据，VPF-Class[31]首先通过 HMM 对病毒蛋白家族(viral protein families, VPF)进行聚类，能侵袭相同宿主的蛋白质属于同一聚类，从而建立蛋白质簇与细菌宿主的映射关系，进而将宏基因组拼装获得的片段进行基因

预测，通过对应的蛋白质簇预测可能的宿主。该方法对与参考数据有关的噬菌体预测准确性很高，但该方法适用与已知宿主的参考噬菌体共享至少 1 个 VPF 的噬菌体，可扩展性较差。

基于机器学习的噬菌体宿主范围预测模型将宿主作为类别，基于相互作用对提取特征，这类模型基本采用常规的机器学习模型，重点常放在特征的提取和整合。

4.3 基于深度学习神经网络的噬菌体宿主范围预测模型

近年来，深度网络在噬菌体宿主范围预测研究得到广泛关注。如何将噬菌体和其宿主序列转换为神经网络的数值输入，以及如何设计网络结构以更好适配生物机理，都是重点研究的问题。

vHULK[30]将噬菌体的蛋白质序列与 pVOG 数据库[38]的标记基因进行匹配，将 9 504 个标记基因蛋白质命中与否的逻辑值输入神经网络，通过 2 层全连接网络，实现 52 个种水平和 61 个属水平的分类，并集成分别使用 ReLu 激活函数、Softmax 激活函数的种水平预测网络、属水平预测网络，通过启发式决策树获得最终的预测结果。

VIDHOP[26]将基因组切为等长的 100 bp 或 400 bp 子序列，通过 one-hot 编码将噬菌体基因组处理成相同长度的输入向量，并通过双向 LSTM、CNN+双向 LSTM 分别抽取不同尺度的特征，并以甲型流感病毒、狂犬病病毒和轮状病毒及其已知的 49 种、19 种、6 种细菌宿主为训练数据，训练不同的神经网络分别预测这 3 类病毒的潜在宿主。

由 Sun 团队提出的 ContigNet[28]使用 one-hot 编码噬菌体序列和宿主序列的碱基和密码子，形成四维特征矩阵，通过 CNN 网络的卷积层和池化层将不同长度的序列生成固定长度的输出，并将 4 类编码序列对应的网络输出首尾相连输入全连接网络，得到输入噬菌体与细菌序列的相互作用概率。

PHIAF[24]针对噬菌体宿主对正样本稀少的问题，提出基于生成对抗网络(generative adversarial network, GAN)的数据增强，从噬菌体和细菌的 DNA 和蛋白质序列中提取出基于不同长度 k 的 k-mer 频度及其衍生的 340 维特征向量；针对蛋白质序列的氨基酸频度及蛋白质的化学元素丰度(chemical element abundance, AC)和蛋白质的分子量(molecular weight, MW)等 162 维的特征向量。通过 6 个统计特征(均值、最大值、最小值、标准差、方差和中位数)整合的蛋白质序列特征。通过输入 GAN 进行数据增强，并随机选取负样本进行训练，输入 2 层 CNN 网络，最后引入注意力机制利用全连接层得到噬菌体宿主对的匹配度。

DeepHost[27]通过 N 个不同间隔距离的 k-mer 来适应序列的插入、删除和突变，得到 $2^k \times 2^k \times 2N$ 的三维特征输入矩阵，应用卷积神经网络预测宿主为各细菌的概率。

在噬菌体宿主范围预测深度网络模型中，最重要的是输入网络的特征，往往需要合理的网络结构提取不同尺度的特征信息。深度学习方法往往基于有标签的监督学习，因此很难预测新出现的宿主，可扩展性较差。

4.4 基于网络图的噬菌体宿主范围预测

噬菌体与其细菌宿主具有天生的连接关系，因此采用网络图的结构来表示这种关系具有天然的适应性。

Liu 等[57]基于噬菌体与宿主关联、噬菌体间相似度、宿主之间相似度构建异构网络：基于已知的噬菌体-宿主对构建噬菌体-宿主关系网络；基于 k-mer 频度的 d_2^* [22]构建噬菌体之间的相似度网络；基于相似网络融合(similar network fusion, SNF)将 k-mer 频度相似性指标 d_2^* 与基于

邻接矩阵的高斯互作谱核相似度(Gaussian interaction profile kernel similarity)结合构建宿主网络[58]。提出一种核化逻辑矩阵分解模型(logistic matrix factorization with integrating different information to predict potential virus-host associations, ILMF-VH)[57]预测异构网络上潜在的噬菌体宿主关联。

由Sun团队提出的VHM-Net[59]将噬菌体宿主对作为节点构建无向网络图，边为噬菌体宿主对(virus-host pairs, VHPs)之间的相似性(2个噬菌体之间的相似性加上2个细菌之间的相似性)。基于以下3个前提假设进行噬菌体宿主范围预测：噬菌体之间的基因组序列相似性可能表明共同的宿主或密切的宿主相关性；细菌宿主之间的基因组序列相似性表明有可能被同一种噬菌体感染；由于噬菌体依赖宿主的细胞机制进行复制，噬菌体在全基因组特征方面通常与其感染的宿主更相似。VHP的相互作用状态成立与否取决于每个VHP节点本身的特性以及每个VHP与邻居VHP之间的联系。VHP本身的特征包括该对噬菌体宿主的相似性度量(基于d_2^*和d_2^s提出的s_2^*)、BLASTn得分和CRISPR区域配准得分。VHP与其他VHP之间的联系根据噬菌体与感染同一宿主的其他噬菌体之间的基因组相似性来定义。

HostG[18]基于图卷积网络(graph convolutional networks, GCN)进行半监督模型的构建，图的节点为噬菌体和细菌，细菌节点带有分类标签，噬菌体节点包括已知宿主和未知宿主的噬菌体，GCN最终将为未知宿主的噬菌体节点分配标签，所有节点特征使用基于CNN预训练的序列嵌入表示。HostG包括2种类型的边，噬菌体-噬菌体和噬菌体-细菌。噬菌体间的边通过序列相似性和共享蛋白家族之间的相似性获得，噬菌体与细菌之间的边通过基因组之间的局部相似性获得，通过卷积神经网络训练获得节点间的相似性，预测未知宿主节点对应的宿主信息。

CHERRY[19]同样基于图卷积网络进行噬菌体宿主范围预测，其节点为噬菌体或细菌DNA序列的4-mer频度向量，噬菌体与噬菌体的连接使用蛋白质相似性来构建，噬菌体与细菌的连接基于CRISPR和BLASTn进行构建。该方法通过编码器-解码器(AutoEncoder)结构，构建网络图结构，得到集成网络图中考虑噬菌体间相似性和噬菌体与细菌相似性，得到的噬菌体节点和细菌节点的嵌入向量，而后进一步使用2层全连接神经网络分类器来解码预测噬菌体的宿主或预测感染细菌的噬菌体。

网络图可以包含噬菌体之间的相似性，也可以包含噬菌体与细菌之间的基因组相似度。网络图模型整合所有连接关系和相似关系，可以得出更为可靠的结果，但缺点是消耗的计算资源过大。

4.5 基于混合模型的噬菌体宿主范围预测

4.5.1 统计模型与机器学习混合的策略

RaFAH[20]基于CRISPR间隔、同源性匹配和共享的tRNAs三类特征，对噬菌体基因组预测的蛋白质序列进行聚类，得到43 644个蛋白质簇(protein clusters , PCs)，并建立隐马尔可夫模型(hidden Markov model, HMM)，将25 879个噬菌体序列(其中709个被判断为完整基因组，其余为序列片段)映射到HMM得到噬菌体基因组对蛋白质聚类的得分表(25 879×43 644)；进而使用随机森林分析噬菌体基因组中存在的蛋白质簇，并根据与已知宿主之间的相似性给出预测得分。该模型还通过重要性分析来量化蛋白质簇对于宿主预测的贡献，具有较好的可解释性。

4.5.2 统计模型与深度学习混合的策略

HoPhage[60]集成深度学习和马尔科夫链模型，其中HoPhage-G (genus)模块基于深度学习实现属水平上的配对预测；HoPhage-S (strain)模

块利用每个候选细菌基因组的编码序列(coding sequence, CDS)构建马尔可夫链模型，通过 2 个模块的加权平均进行宿主的预测。

5 噬菌体的细菌宿主范围预测策略

噬菌体宿主范围的预测策略分为基于噬菌体间相似性的配对策略和基于噬菌体与宿主相似性的配对策略，前者是通过噬菌体之间的相似性进行判断，认为相似度高的噬菌体更有可能感染同类型宿主；后者是认为噬菌体与其宿主应该具有较高的相似度，通过计算噬菌体与细菌的相似度来预测噬菌体宿主范围。

5.1 基于噬菌体间相似性度量的噬菌体宿主范围预测策略

该策略基于数据库中已有的噬菌体和宿主的配对关系，通过建模具有相同宿主及不同宿主的噬菌体之间的相似性度量模型，计算与已知宿主噬菌体的相似性来预测新噬菌体的宿主范围。这类策略无需使用宿主的基因组序列，但对于不出现在已有宿主列表中的宿主预测情况无法适用。同时该思路的前提是具有相同宿主的噬菌体基因组序列在某种空间和度量下具有很高的相似性，但该策略仅利用噬菌体的基因组参考序列，而没有充分利用宿主的基因组信息。其代表方法包括 HostPhinder[21]、ILMF-VH[57]、vHULK[30]、VPF-Class[31]和 RaFAH[20]等。

5.2 基于噬菌体与细菌间相似性度量的噬菌体宿主范围预测策略

该策略直接将噬菌体与宿主的特征进行相似性度量，通过 CRISPR、AMG 等短序列匹配信息，或是 k-mer 频度等相似性度量方法来评估噬菌体与宿主的相似性，其代表方法包括 VHM[22]、WIsH[25]、PHP[23]等。

5.3 混合噬菌体相似性与噬菌体-细菌相似性的噬菌体宿主范围预测策略

该策略在建模中既考虑噬菌体间的相似性，又考虑噬菌体与宿主的配对关系，以最大限度利用噬菌体与细菌的基因组信息，其代表性方法包括 VirHostMatcherNet[59]、PHISDetector[16]以及 CHERRY[19]，CHERRY 结合噬菌体之间蛋白质相似性、噬菌体和宿主的 CRISPR 区域和 BLASTn 配准及基于 k-mer 频度的相似性度量进行预测。

6 实验设计与评估

6.1 噬菌体和细菌配对数据库

6.1.1 NCBI RefSeq 数据库[15]

在噬菌体宿主范围预测中，使用的绝大多数噬菌体和细菌的基因组信息都来自 NCBI RefSeq 数据库，而不同研究中所构建的针对噬菌体与细菌宿主的专门数据库基本也是从 NCBI RefSeq 数据库中抽取和整理获得。

6.1.2 NCBI GenBank 数据集[61]

GenBank 包含 40 万个物种的核苷酸序列，这些序列主要通过单个实验室的提交和大规模测序项目的批量提交获得。GenBank 包括物种的分类、基因组、蛋白质序列和结构等信息，RaFAH[20]利用该数据库的噬菌体序列信息作为测试集。

6.1.3 VHM 数据集[22]

抽取自 NCBI RefSeq 数据库，在 2015 年 5 月 8 日 NCBI 数据库中筛选 1 427 个具有宿主信息的噬菌体基因组和对应的 31 986 个原核基因组，构成了一个噬菌体宿主配对预测的基准数据集。PHP[23]根据国际病毒分类委员会(International Committee on Taxonomy of Viruses, ICTV)[62]和 NCBI 分类数据库[63]更新了 VHM 数据集中病毒和原核生物基因组的分类信息，更新的 VHM 数据集包含 1 426 个噬菌体基因组和 31 918 个原核

基因组。使用该参考数据集的模型包括 VHM[22]、PHP[23]、HostG[18]和 CHERRY[19]。

6.1.4 Virus-Host DB[64]

挖掘和整合现有的数据库和文献，Virus-Host DB 提供了一个全面的、手动筛选的噬菌体及其细胞宿主之间分类联系的数据库，从 NCBI RefSeq 病毒基因组条目和 UniProtKB 中的蛋白质序列条目中提取了天然宿主和实验室宿主信息。此外，还提供了到外部参考资源的链接，如 ViralZone[65]、NCBI 分类数据库、KEGG[66]和 ICTV。使用该数据集的模型包括 PHIAF[24]、VPF-Class[31]、HoPhage[60]和 ContigNet[28]。

6.1.5 Phages DB[67]

Phages DB 收集并共享了与感染放线菌宿主的病毒相关数据，包括发现、表征和基因组学相关的信息。到文章发表时为止，已有 8 000 多个噬菌体被输入数据库，包括 1 600 多个已测序的基因组。使用该数据集的模型包括 PHIAF[24]、DeepHost[27]。

6.1.6 GL-UVAB[68]

古菌和细菌的未培养病毒的基因组谱系(genomic lineages of uncultured viruses of archaea and bacteria, GL-UVAB)进行了近 20 万个病毒核苷酸序列的系统基因组分析，已鉴定谱系的泛基因组内容揭示了它们的一些感染策略、调节宿主生理行为的潜力以及逃避宿主抵抗噬菌体侵略的机制。使用该数据集的模型包括 RaFAH[20]。

6.1.7 KEGG[66]

KEGG 数据库项目包含基因组、生化反应、生化物质、疾病与药物，以及最常用的 PATHWAY 通路信息。在噬菌体宿主预测研究中，KEGG 数据库主要提供完整基因组信息，在未来的研究中，如何将 KEGG 数据库中最常用的通路信息加入预测问题中值得进一步研究。使用该数据集的模型包括 WIsH[25]。

6.1.8 pVOGs[38]

原核病毒直系同源群(prokaryotic virus orthologous groups, pVOGs)数据库包括病毒蛋白的功能注释、未表征 DNA 样本中基因和病毒的鉴定、系统发育分析和大规模比较基因组学项目等同源基因家族信息。由于一些基因家族会同时存在于感染某种细菌或古菌的多个病毒基因组中，因此这些基因所翻译的蛋白可以作为特征。将噬菌体序列是否包含这些标记基因作为特征用于宿主预测。使用该数据集的代表模型是 vHULK[30]。

6.1.9 IMG/VR[69]

整合微生物基因组/病毒(integrated microbial genome/virus, IMG/VR)数据库提供病毒基因组和基因组片段、基因和基因簇、蛋白质功能以及相关宿主和栖息地数据的多层次集成。IMG/VR 与 pVOGs 类似，它提供了病毒蛋白家族信息，可以作为特征进行宿主预测。使用该数据集的代表模型是 VPF-Class[31]。

6.2 训练测试集的划分及样本不平衡的实验策略

训练集和测试集的划分主要采取 2 种策略。第一种策略是按照时间点进行划分，例如 HostG 和 CHERRY 以 2015 年以前的 1 426 个噬菌体宿主相互作用作为训练集，以 2015−2020 年的 671 个数据作为测试集；RaFAH 将 2019 年 10 月前的 25 879 个噬菌体序列(其中 709 个是完整基因组)作为训练集，2019 年 10 月后的 3 427 个数据作为测试集；第二种策略是随机抽取，例如 HoPhage 随机抽取 90%的噬菌体-宿主配对信息作为训练集，剩余 10%作为测试集。PHISDetector[16]的测试集将建模阶段未使用的噬菌体宿主配对作为正样本，噬菌体与其宿主以外的细菌的配对作为人工建立的负样本。

噬菌体在不同的生物分类学下的样本数目极不均衡，有些属仅包含 1 个噬菌体，样本分布的不均衡影响了模型的训练，因此数据划分过程

中需要随机选取负样本或对正样本进行数据增强，以得到平衡的训练数据。例如可以将感染每种宿主的噬菌体按训练集和测试集的比例分配到训练集和测试集。HostPhinder[21]首先将噬菌体基因组按照 16-mer 相似性聚类，最终得到了 1 414 个聚类，其中 1 121 个聚类只包含 1 个基因组，将这 1 414 个聚类按照宿主字母和基因组大小依次排序，并且交替分布到 5 个部分，这保证了每个部分有相对平均的宿主类型以及数据大小。

另一方面，噬菌体宿主预测的正样本数目有限，负样本数目远大于正样本数目，造成噬菌体宿主互作的正样本数目远小于负样本数目。一方面，可以通过负样本抽样获得平衡数据，许多方法采用了此类采样策略，如 VirHostMatcherNet、PHISDetector、ContigNet 等；另一方面，通过数据增强来产生更多可用的正样本，例如 PHIAF[24]使用 GAN 生成新的伪正样本。

6.3 测试噬菌体及细菌配对智能算法的真实数据集

宏基因组测序数据常用于测试各模型在真实应用场景下的性能。例如基于前人研究中对宏基因组数据通过序列比对、生物实验等方案鉴定出的噬菌体宿主对，分析模型所能预测的宿主的准确性；或直接针对宏基因组中装配获得的未知噬菌体进行宿主范围预测。CHERRY[19]在健康人类肠道的 MetaHiC[70]样本中预测出 6 545 个噬菌体的宿主。VPF-Class[31]针对基于海洋宏基因组装配构建的全球海洋病毒(global marine viruses, GOV)数据库[71]中的 1 380 523 个病毒序列进行宿主预测，其中 834 023 条序列能够得到属水平的宿主预测。RaFAH[20]对热带和亚热带温暖海洋样本[72]中的 4 751 个病毒基因组预测宿主；对来自永久冻土、海洋、人类肠道、淡水、土壤、高盐度湖泊和热液泉[69]宏基因组中的 61 647 个病毒序列进行宿主预测。

6.4 评估指标和评估方式

6.4.1 模型评估

最常使用的评估指标是噬菌体宿主范围预测的准确率，即判断正确的噬菌体宿主对占全部的百分比，此外，在机器学习中经常使用的查准率(precision)、查全率(recall)、F1 score、ROC 曲线、AUROC、AUPRC、Marco F1-score 和加权 F1-score 等评价指标也可以被用来评价噬菌体宿主范围预测方法的好坏。

6.4.2 特征重要性评估

许多研究通过消融实验对特征的重要性进行评估。例如 RaFAH[20]对使用的蛋白质聚类进行重要性分析，最重要的预测因子被确定为 Rz-like 噬菌体裂解蛋白，另外，多种赖氨酸、尾部和尾部纤维蛋白等在噬菌体对宿主识别的生物过程中发挥着重要作用。PHIAF[24]分析神经网络注意力机制中分配给不同特征的权重，表明正样本和负样本的注意力权重分布相似，说明某些特征在正样本和负样本中具有同样重要的作用；此外，蛋白质水平的特征通常权重低于 DNA 水平的特征，表明对于噬菌体宿主范围预测，DNA 水平的特征比蛋白质水平的特征更重要。

6.4.3 k-mer 参数的影响

基于 k-mer 频度的方法中，k 值的选择也会显著影响预测的结果。Young 等[55]发现针对 DNA k-mer、氨基酸 k-mer 以及蛋白质理化性质 k-mer 的预测结果都会随着 k-mer 长度的增加而提高，其实验最长为 9-mer。VHM[22]也发现长度在 4–8 bp 的 DNA 序列 k-mer 中，随着 k 值的提高各个距离度量方式在不同水平的分类结果均有所提高。同时，研究还表明 k-mer 频度计算时考虑互补链可以提高噬菌体宿主范围预测任务的准确性[73]。

6.5 代表模型在 VHM 数据集上的性能比较

VHM 数据集最早由 Sun 的团队在 2017 年构建[22]，在之后由 PHP[23]和 CHERRY[19]进行了更新，包含 2020 年之前 NCBI RefSeq 中的噬菌

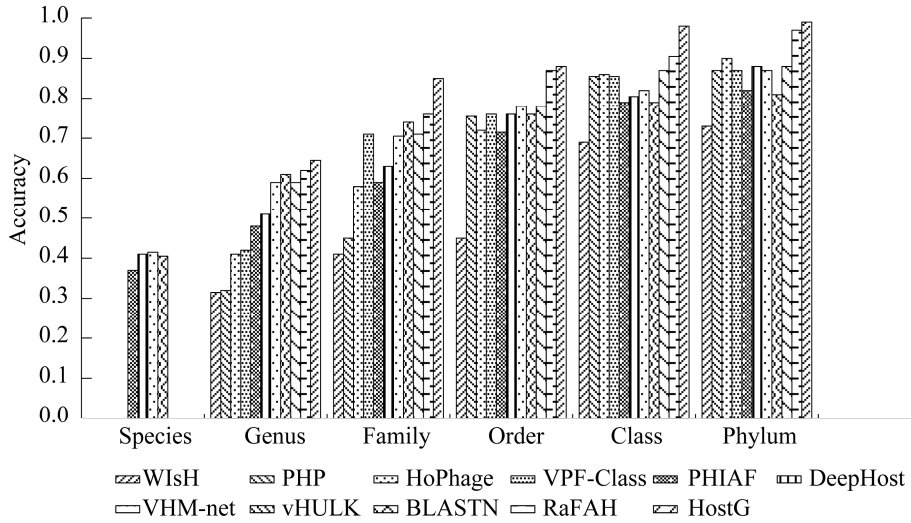

图 4　不同方法在 VHM baseline 数据集上的表现

Figure 4　The performance of different methods on VHM baseline datasets. The horizontal axis represents different classification levels, with phylum, class, order, family, genus, and species from right to left. The vertical axis represents the accuracy of prediction, and different patterns represent different methods. For species level prediction, only the methods that can be used for species level host prediction are listed.

体宿主相互作用数据，其中训练集包括 2015 年前发现的 1 306 个噬菌体宿主对，其细菌宿主分布在 187 个种；测试集包括 2015−2020 年发现的 634 个噬菌体宿主对，其细菌宿主分布在 95 个种。整合论文[18-19,28]中罗列出基于 VHM 数据集进行训练和测试的相关的模型性能评估。如图 4 所示，图中列出了 11 个模型在不同分类(门纲目科属种)层面的测试集准确率，在属层面的宿主预测准确率在 30%−60%，其中仅有 PHIAF、DeepHost、VHM-net 和 vHULK 四个模型能在种层面进行宿主预测，其准确率在 40%左右。

7　结论与展望

噬菌体在自然界中具有重要地位，基于智能计算的噬菌体宿主范围预测将有助于更好地理解噬菌体的感染机制，抗生素的新型疗法提供重要的参考和工具。基于智能计算的噬菌体的细菌宿主范围预测研究显著扩大了噬菌体的宿主信息集合，也为发现新的配对关系提供了重要的工具。

本文对基于智能计算的噬菌体宿主范围预测研究中存在的尚待解决的难点进行分析，并尝试给出在智能技术与生物技术发展下可能的解决方案和发展前景。

1) 目前大多数方法的宿主预测分辨率在种层面，针对菌株水平的宿主预测并不多见，精度也有待提高，菌株之间的高度相似性可能影响预测精度。可能的改进方式包括：

(1) 融合噬菌体、宿主多特征、多模态数据，集成统计、深度等不同策略的多智能模型可能优化噬菌体-宿主相互作用预测的性能和可靠性。

(2) 发展不同尺度的预测模型。使其不仅具有更加精细的处理策略，同时也具有更加良好的泛化性能。

2) 目前已标注的噬菌体宿主配对信息较少，宿主标签的生物分类学分布极为不平衡，感染某些宿主的噬菌体可能只有一条或几条记录，使得对应类别无法获得充分的训练。

(1) 引入迁移学习、预训练大模型等，将 few-shot 及 zero-shot 等智能策略引入模型；

(2) 采用针对不平衡数据的模型、负样本采样构建策略以及正样本的生成模型。

3) 噬菌体的宿主特异性愈发被认为高度可变：即一些噬菌体具有很强的宿主特异性，只能感染单个宿主或非常窄的宿主范围；另一些噬菌体能够感染大量宿主菌株，这种特异性的高度变化特性也影响了模型的预测性能。

(1) 针对不同分类学层次、分支的分布建模；

(2) 引入蛋白质三维构型、三维基因组 Hi-C 等序列以外的特征。

4) 智能计算常基于单纯的数据驱动，结果缺乏可解释性和泛化性能。

(1) 将生物机理、先验知识融入模型构建，使得模型能够真正学习和推导出噬菌体感染宿主的本质模式。

(2) 将统计模型与深度学习有机结合，引入可解释性 AI 算法，建立新一代具有更强泛化性和可解释性的噬菌体宿主预测框架。

智能计算方法在预测噬菌体的细菌宿主范围的研究中发挥了重要的作用。由于计算模型、训练所依赖的基因组、蛋白质及数据库等参考信息的限制以及噬菌体在不同的分类分支下的多样性，导致计算模型无法在所有情况下达到较高的精度。因此，计算方法主要为噬菌体研究人员提供其宿主范围，需要通过生物实验等方式进一步验证和确定噬菌体对应的细菌宿主。

参考文献

[1] BREITBART M, ROHWER F. Here a virus, there a virus, everywhere the same virus?[J]. Trends in Microbiology, 2005, 13(6): 278-284.

[2] BREITBART M, SALAMON P, ANDRESEN B, MAHAFFY JM, SEGALL AM, MEAD D, AZAM F, ROHWER F. Genomic analysis of uncultured marine viral communities[J]. Proceedings of the National Academy of Sciences of the United States of America, 2002, 99(22): 14250-14255.

[3] PRINCIPI N, SILVESTRI E, ESPOSITO S. Advantages and limitations of bacteriophages for the treatment of bacterial infections[J]. Frontiers in Pharmacology, 2019, 10: 513.

[4] LUONG T, SALABARRIA AC, EDWARDS RA, ROACH DR. Standardized bacteriophage purification for personalized phage therapy[J]. Nature Protocols, 2020, 15(9): 2867-2890.

[5] STEEN AD, CRITS-CHRISTOPH A, CARINI P, DeANGELIS KM, FIERER N, LLOYD KG, CAMERON THRASH J. High proportions of bacteria and archaea across most biomes remain uncultured[J]. The ISME Journal, 2019, 13(12): 3126-3130.

[6] DION MB, OECHSLIN F, MOINEAU S. Phage diversity, genomics and phylogeny[J]. Nature Reviews Microbiology, 2020, 18(3): 125-138.

[7] COCLET C, ROUX S. Global overview and major challenges of host prediction methods for uncultivated phages[J]. Current Opinion in Virology, 2021, 49: 117-126.

[8] SALMOND GPC, FINERAN PC. A century of the phage: past, present and future[J]. Nature Reviews Microbiology, 2015, 13(12): 777-786.

[9] LETAROV AV, KULIKOV EE. Adsorption of bacteriophages on bacterial cells[J]. Biochemistry (Moscow), 2017, 82(13): 1632-1658.

[10] GRISSA I, VERGNAUD G, POURCEL C. CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats[J]. Nucleic Acids Research, 2007, 35(suppl_2): W52-W57.

[11] ROSTØL JT, MARRAFFINI L. (Ph)ighting phages: how bacteria resist their parasites[J]. Cell Host & Microbe, 2019, 25(2): 184-194.

[12] KRONHEIM S, DANIEL-IVAD M, DUAN Z, HWANG S, WONG AI, MANTEL I, NODWELL JR, MAXWELL KL. A chemical defence against phage infection[J]. Nature, 2018, 564(7735): 283-286.

[13] LOPATINA A, TAL N, SOREK R. Abortive infection: bacterial suicide as an antiviral immune strategy[J]. Annual Review of Virology, 2020, 7(1): 371-384.

[14] SHMAKOV SA, SITNIK V, MAKAROVA KS, WOLF YI, SEVERINOV KV, KOONIN EV. The CRISPR spacer space is dominated by sequences from species-specific mobilomes[J]. mBio, 2017, 8(5): e01397-17.

[15] PRUITT KD, TATUSOVA T, MAGLOTT DR. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins[J]. Nucleic Acids Research, 2007, 35(suppl_1): D61-D65.

[16] ZHOU FX, GAN R, ZHANG F, REN CY, YU L, SI Y, HUANG ZW. PHISDetector: a tool to detect diverse *in silico* phage-host interaction signals for virome studies[J]. Genomics, Proteomics & Bioinformatics, 2022, 20(3): 508-523.

[17] ZHANG RS, MIRDITA M, LEVY KARIN E, NORROY C, GALIEZ C, SÖDING J. SpacePHARER: sensitive identification of phages from CRISPR spacers in prokaryotic hosts[J]. Bioinformatics, 2021, 37(19): 3364-3366.

[18] SHANG JY, SUN YN. Predicting the hosts of prokaryotic viruses using GCN-based semi-supervised learning[J]. BMC Biology, 2021, 19(1): 250.

[19] SHANG JY, SUN YN. CHERRY: a computational metHod for accuratE pRediction of virus-pRokarYotic interactions using a graph encoder-decoder model[J]. Briefings in Bioinformatics, 2022, 23(5): bbac182.

[20] COUTINHO FH, ZARAGOZA-SOLAS A, LÓPEZ-PÉREZ M, BARYLSKI J, ZIELEZINSKI A, DUTILH BE, EDWARDS R, RODRIGUEZ-VALERA F. RaFAH: host prediction for viruses of bacteria and archaea based on protein content[J]. Patterns, 2021, 2(7): 100274.

[21] VILLARROEL J, KLEINHEINZ KA, JURTZ VI, ZSCHACH H, LUND O, NIELSEN M, LARSEN MV. HostPhinder: a phage host prediction tool[J]. Viruses, 2016, 8(5): 116.

[22] AHLGREN NA, REN J, LU YY, FUHRMAN JA, SUN FZ. Alignment-free d_2^* oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences[J]. Nucleic Acids Research, 2017, 45(1): 39-53.

[23] LU CY, ZHANG Z, CAI ZN, ZHU ZZ, QIU Y, WU AP, JIANG TJ, ZHENG HP, PENG YS. Prokaryotic virus host predictor: a Gaussian model for host prediction of prokaryotic viruses in metagenomics[J]. BMC Biology, 2021, 19(1): 5.

[24] LI ML, ZHANG W. PHIAF: prediction of phage-host interactions with GAN-based data augmentation and sequence-based feature fusion[J]. Briefings in Bioinformatics, 2022, 23(1): bbab348.

[25] GALIEZ C, SIEBERT M, ENAULT F, VINCENT J, SÖDING J. WIsH: Who is the host? Predicting prokaryotic hosts from metagenomic phage contigs[J]. Bioinformatics, 2017, 33(19): 3113-3114.

[26] MOCK F, VIEHWEGER A, BARTH E, MARZ M. VIDHOP, viral host prediction with deep learning[J]. Bioinformatics, 2021, 37(3): 318-325.

[27] WANG RH, ZHANG XLL, WANG JP, LI SC. DeepHost: phage host prediction with convolutional neural network[J]. Briefings in Bioinformatics, 2022, 23(1): bbab385.

[28] TANG TQ, HOU SW, FUHRMAN JA, SUN FZ. Phage-bacterial contig association prediction with a convolutional neural network[J]. Bioinformatics, 2022, 38(supplement_1): i45-i52.

[29] BOECKAERTS D, STOCK M, CRIEL B, GERSTMANS H, de BAETS B, BRIERS Y. Predicting bacteriophage hosts based on sequences of annotated receptor-binding proteins[J]. Scientific Reports, 2021, 11: 1467.

[30] AMGARTEN D, IHA BKV, PIROUPO CM, da SILVA AM, SETUBAL JC. vHULK, a new tool for bacteriophage host prediction based on annotated genomic features and neural networks[J]. PHAGE, 2022, 3(4): 204-212.

[31] PONS JC, PAEZ-ESPINO D, RIERA G, IVANOVA N, KYRPIDES NC, LLABRÉS M. VPF-class: taxonomic assignment and host prediction of uncultivated viruses based on viral protein families[J]. Bioinformatics, 2021, 37(13): 1805-1813.

[32] LEITE DMC, BROCHET X, RESCH G, QUE YA, NEVES A, PEÑA-REYES C. Computational prediction of inter-species relationships through omics data analysis and machine learning[J]. BMC Bioinformatics, 2018, 19(14): 151-159.

[33] LI ML, WANG YN, LI FY, ZHAO Y, LIU MY, ZHANG SJ, BIN YN, SMITH AI, WEBB GI, LI J, SONG JN, XIA JF. A deep learning-based method for identification of bacteriophage-host interaction[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2021, 18(5): 1801-1810.

[34] CHEN Z, ZHAO P, LI FY, LEIER A, MARQUEZ-LAGO TT, WANG YN, WEBB GI, SMITH AI, DALY RJ, CHOU KC, SONG JN. iFeature: a Python package and web server for features extraction and selection from protein and peptide sequences[J]. Bioinformatics, 2018, 34(14): 2499-2502.

[35] KAMINSKI MM, ABUDAYYEH OO, GOOTENBERG JS, ZHANG F, COLLINS JJ.

CRISPR-based diagnostics[J]. Nature Biomedical Engineering, 2021, 5(7): 643-656.

[36] EDWARDS RA, McNAIR K, FAUST K, RAES J, DUTILH BE. Computational approaches to predict bacteriophage-host relationships[J]. FEMS Microbiology Reviews, 2016, 40(2): 258-272.

[37] REN J, AHLGREN NA, LU YY, FUHRMAN JA, SUN FZ. VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data[J]. Microbiome, 2017, 5(1): 69.

[38] LAURA GRAZZIOTIN A, KOONIN EV, KRISTENSEN DM. Prokaryotic virus orthologous groups (pVOGs): a resource for comparative genomics and protein family annotation[J]. Nucleic Acids Research, 2017, 45(D1): D491-D498.

[39] CARBONE A. Codon bias is a major factor explaining phage evolution in translationally biased hosts[J]. Journal of Molecular Evolution, 2008, 66(3): 210-223.

[40] WANG Y, LIU L, CHEN LN, CHEN T, SUN FZ. Comparison of metatranscriptomic samples based on k-tuple frequencies[J]. PLoS One, 2014, 9(1): e84348.

[41] ZHENG YL, SHI JL, CHEN Q, DENG C, YANG F, WANG Y. Identifying individual-specific microbial DNA fingerprints from skin microbiomes[J]. Frontiers in Microbiology, 2022, 13: 960043.

[42] MA Z, LU YY, WANG YW, LIN RH, YANG ZZ, ZHANG F, WANG Y. Metric learning for comparing genomic data with triplet network[J]. Briefings in Bioinformatics, 2022, 23(5): bbac345.

[43] BLAISDELL BE. A measure of the similarity of sets of sequences not requiring sequence alignment[J]. Proceedings of the National Academy of Sciences of the United States of America, 1986, 83(14): 5155-5159.

[44] NARLIKAR L, MEHTA N, GALANDE S, ARJUNWADKAR M. One size does not fit all: on how Markov model order dictates performance of genomic sequence analyses[J]. Nucleic Acids Research, 2013, 41(3): 1416-1424.

[45] PRIDE DT, WASSENAAR TM, GHOSE C, BLASER MJ. Evidence of host-virus co-evolution in tetranucleotide usage patterns of bacteriophages and eukaryotic viruses[J]. BMC Genomics, 2006, 7: 8.

[46] REINERT G, CHEW D, SUN FZ, WATERMAN MS. Alignment-free sequence comparison (I): statistics and power[J]. Journal of Computational Biology, 2009, 16(12): 1615-1634.

[47] LIAO WN, REN J, WANG K, WANG S, ZENG F, WANG Y, SUN FZ. Alignment-free transcriptomic and metatranscriptomic comparison using sequencing signatures with variable length Markov chains[J]. Scientific Reports, 2016, 6: 37243.

[48] QI J, LUO H, HAO BL. CVTree: a phylogenetic tree reconstruction tool based on whole genomes[J]. Nucleic Acids Research, 2004, 32(suppl_2): W45-W47.

[49] QI J, WANG B, HAO BI. Whole proteome prokaryote phylogeny without sequence alignment: a K-string composition approach[J]. Journal of Molecular Evolution, 2004, 58(1): 1-11.

[50] TEELING H, WALDMANN J, LOMBARDOT T, BAUER M, GLÖCKNER FO. TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences[J]. BMC Bioinformatics, 2004, 5: 163.

[51] WANG Y, FU L, REN J, YU ZX, CHEN T, SUN FZ. Identifying group-specific sequences for microbial communities using long k-mer sequence signatures[J]. Frontiers in Microbiology, 2018, 9: 872.

[52] KARLIN S, MRÁZEK J, CAMPBELL AM. Compositional biases of bacterial genomes and evolutionary implications[J]. Journal of Bacteriology, 1997, 179(12): 3899-3913.

[53] LU YY, BAI JX, WANG YW, WANG Y, SUN FZ. CRAFT: compact genome representation toward large-scale alignment-free daTabase[J]. Bioinformatics, 2021, 37(2): 155-161.

[54] ZHANG MG, YANG LP, REN J, AHLGREN NA, FUHRMAN JA, SUN FZ. Prediction of virus-host infectious association by supervised learning methods[J]. BMC Bioinformatics, 2017, 18(3): 143-154.

[55] YOUNG F, ROGERS S, ROBERTSON DL. Predicting host taxonomic information from viral genomes: a comparison of feature representations[J]. PLoS Computational Biology, 2020, 16(5): e1007894.

[56] LOOD C, BOECKAERTS D, STOCK M, de BAETS B, LAVIGNE R, van NOORT V, BRIERS Y. Digital phagograms: predicting phage infectivity through a multilayer machine learning approach[J]. Current Opinion in Virology, 2022, 52: 174-181.

[57] LIU D, MA YJ, JIANG XP, HE TT. Predicting virus-host association by Kernelized logistic matrix factorization and similarity network fusion[J]. BMC Bioinformatics, 2019, 20(16): 1-10.

[58] HOU YJ, ZHANG X, ZHOU QY, HONG WX, WANG Y. Hierarchical microbial functions prediction by graph aggregated embedding[J]. Frontiers in Genetics, 2021, 11: 608512.

[59] WANG WL, REN J, TANG KJ, DART E, IGNACIO-ESPINOZA JC, FUHRMAN JA, BRAUN J, SUN FZ, AHLGREN NA. A network-based integrated framework for predicting virus-prokaryote interactions[J]. NAR Genomics and Bioinformatics, 2020, 2(2): lqaa044.

[60] TAN J, FANG ZC, WU SF, GUO Q, JIANG XQ, ZHU HQ. HoPhage: an *ab initio* tool for identifying hosts of phage fragments from metaviromes[J]. Bioinformatics, 2022, 38(2): 543-545.

[61] BENSON DA, CAVANAUGH M, CLARK K, KARSCH-MIZRACHI I, OSTELL J, PRUITT KD, SAYERS EW. GenBank[J]. Nucleic Acids Research, 2018, 46(D1): D41-D47.

[62] WALKER PJ, SIDDELL SG, LEFKOWITZ EJ, MUSHEGIAN AR, ADRIAENSSENS EM, DEMPSEY DM, DUTILH BE, HARRACH B, HARRISON RL, CURTIS HENDRICKSON R, JUNGLEN S, KNOWLES NJ, KROPINSKI AM, KRUPOVIC M, KUHN JH, NIBERT M, ORTON RJ, RUBINO L, SABANADZOVIC S, SIMMONDS P, et al. Changes to virus taxonomy and the statutes ratified by the International Committee on Taxonomy of Viruses (2020)[J]. Archives of Virology, 2020, 165(11): 2737-2748.

[63] SCHOCH CL, CIUFO S, DOMRACHEV M, HOTTON CL, KANNAN S, KHOVANSKAYA R, LEIPE D, MCVEIGH R, O'NEILL K, ROBBERTSE B, SHARMA S, SOUSSOV V, SULLIVAN JP, SUN L, TURNER S, KARSCH-MIZRACHI I. NCBI taxonomy: a comprehensive update on curation, resources and tools[J]. Database: the Journal of Biological Databases and Curation, 2020, 2020: baaa062.

[64] MIHARA T, NISHIMURA Y, SHIMIZU Y, NISHIYAMA H, YOSHIKAWA G, UEHARA H, HINGAMP P, GOTO S, OGATA H. Linking virus genomes with host taxonomy[J]. Viruses, 2016, 8(3): 66.

[65] MASSON P, HULO C, de CASTRO E, BITTER H, GRUENBAUM L, ESSIOUX L, BOUGUELERET L, XENARIOS I, Le MERCIER P. ViralZone: recent updates to the virus knowledge resource[J]. Nucleic Acids Research, 2013, 41(D1): D579-D583.

[66] KANEHISA M, FURUMICHI M, TANABE M, SATO Y, MORISHIMA K. KEGG: new perspectives on genomes, pathways, diseases and drugs[J]. Nucleic Acids Research, 2017, 45(D1): D353-D361.

[67] RUSSELL DA, HATFULL GF. PhagesDB: the actinobacteriophage database[J]. Bioinformatics, 2017, 33(5): 784-786.

[68] COUTINHO FH, EDWARDS RA, RODRÍGUEZ-VALERA F. Charting the diversity of uncultured viruses of archaea and bacteria[J]. BMC Biology, 2019, 17(1): 1-16.

[69] PAEZ-ESPINO D, ROUX S, CHEN IM A, PALANIAPPAN K, RATNER A, CHU K, HUNTEMANN M, REDDY TBK, PONS JC, LLABRÉS M, ELOE-FADROSH EA, IVANOVA NN, KYRPIDES NC. IMG/VR v.2.0: an integrated data management and analysis system for cultivated and environmental viral genomes[J]. Nucleic Acids Research, 2019, 47(D1): D678-D686.

[70] MARBOUTY M, THIERRY A, MILLOT GA, KOSZUL R. MetaHiC phage-bacteria infection network reveals active cycling phages of the healthy human gut[J]. eLife, 2021, 10: e60608.

[71] ROUX S, BRUM JR, DUTILH BE, SUNAGAWA S, DUHAIME MB, LOY A, POULOS BT, SOLONENKO N, LARA E, POULAIN J, PESANT S, KANDELS-LEWIS S, DIMIER C, PICHERAL M, SEARSON S, CRUAUD C, ALBERTI A, DUARTE CM, GASOL JM, VAQUÉ D, et al. Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses[J]. Nature, 2016, 537(7622): 689-693.

[72] PACHIADAKI MG, BROWN JM, BROWN J, BEZUIDT O, BERUBE PM, BILLER SJ, POULTON NJ, BURKART MD, la CLAIR JJ, CHISHOLM SW, STEPANAUSKAS R. Charting the complexity of the marine microbiome through single-cell genomics[J]. Cell, 2019, 179(7): 1623-1635.e11.

[73] REN J, SONG K, DENG C, AHLGREN NA, FUHRMAN JA, LI Y, XIE XH, POPLIN R, SUN FZ. Identifying viruses from metagenomic data using deep learning[J]. Quantitative Biology, 2020, 8(1): 64-77.