通过对比学习和全基因组序列表示零样本识别噬菌体-宿主关系

Yaozhong Zhang, Yunjie Liu, Zeheng Bai, Kosuke Fujimoto,
Satoshi Uematsu and Seiya Imoto

摘要

从基因组序列中准确识别噬菌体-宿主的关系仍然具有挑战性，特别是对于那些同源序列较少的噬菌体和宿主。在这项工作中，重点关注在物种水平上识别噬菌体-宿主关系，提出了一种基于对比学习的方法来学习噬菌体-宿主相互作用的全基因组序列嵌入。对比学习用于使具有相互作用关系的噬菌体和宿主在新的表示空间中彼此靠近。具体而言，用频率混沌博弈表示重新表述全基因组序列，并通过对比学习来学习噬菌体和宿主关系的潜在嵌入。基于学习到的嵌入，作者提出了名为 CL4PHI 的模型，它可以预测训练中已知的宿主和未知的宿主。将 CL4PHI 与最近提出的两种最先进的基于学习的方法在其基准数据集上进行比较。实验结果表明，所提出的使用对比学习的方法提高了预测宿主的精度，并且明显优于最先进的宿主预测方法。

关键词：噬菌体-宿主预测；对比学习；全基因组序列表示

1 引言

在过去的二三十年，抗生素在医学和农业中被滥用和过度使用，这导致越来越多的耐药病原体出现，给临床治疗带来了极大的挑战 [1, 2]。对于难以治疗的耐药性细菌感染，抗生素治疗方案具有一定的局限性，并可能导致患者发病率和死亡率增加 [3]。因此，迫切需要新的替代疗法来应对抗生素耐药性的问题，其中一种就是噬菌体疗法 [4]。

噬菌体是地球上最多样化和最丰富的生物体，能够通过将其基因组注入宿主细胞并感染它们来调节和影响生态系统 [5]。噬菌体疗法试图寻找有效的噬菌体来消除细菌，因此可以用来解决抗生素耐药性问题。

然而，由于宿主表面的特异性受体会直接影响噬菌体的宿主范围，导致大多数噬菌体只能感染细菌种内有限的一个或几个菌株 [6]。因此，常用一种“鸡尾酒疗法”来治疗耐药菌感染。噬菌体鸡尾酒是一种混合制剂，通过将多种具有不同宿主范围的噬菌体组合在一起，以覆盖更广泛的靶向菌 [7]。为了广泛应用于临床实践，可以采用合理的鸡尾酒配方。但在此之前，需要确定噬菌体和宿主之间是否存在相互作用关系。

噬菌体-宿主相互作用（Phage-Host Interaction，PHI）是指噬菌体与细菌（宿主）之间的生物学过程，包括噬菌体的吸附、感染、复制和释放等步骤。传统上检测噬菌体-宿主相互作用需要在严格的实验条件下进行，这不仅耗时而且昂贵，部分实验甚至可能具有相当的挑战

性 [8]。这在一定程度上阻碍了噬菌体疗法的发展。近年来，随着下一代测序 (Next Generation Sequencing, NGS) 技术的发展，以相对低成本的方式对大量 DNA 进行测序成为可能。此外，科学家能够利用宏基因组学研究微生物群落的遗传内容，从而发现许多以前未知的噬菌体，这进一步增加了可研究的噬菌体数量 [9,10]。因此，越来越多的研究开始探索使用基于计算的方法预测噬菌体-细菌相互作用关系。

基于计算方法的宿主预测主要面临两个挑战 [11]：第一个是缺乏已知的噬菌体与宿主的相互作用。例如，截至 2020 年，已知相互作用的数量仅占当时国家生物技术信息中心 (National Center for Biotechnology Information, NCBI) 参考序列中噬菌体的 40% (1940) 左右。与此同时，在 NCBI 参考序列的 60105 个细菌基因组中，只有 223 个注释了与 1940 个噬菌体的相互作用。有限的已知相互作用需要精心设计模型或算法来进行宿主预测。第二个是，尽管噬菌体和宿主之间的序列相似性对于宿主预测来说是一个有用的特征，但并非所有噬菌体都与其宿主基因组共享相同区域。例如，在参考序列数据库中，约 24% 的噬菌体与其宿主没有显著的序列相似性。因此，基于序列比对的方法无法识别某些噬菌体的宿主。

在这项工作中，作者提出了一种将已知 PHIs 纳入噬菌体和宿主基因组序列表示的新方法，并利用学习到的表示来识别 PHIs。为了有效处理全基因组序列，作者用频率混沌博弈表示法 (Frequency Chaos Game Representation, FCGR) [12] 重新表述噬菌体和宿主的 k-mer 信息。根据已知的 PHIs 和 FCGR 表示，训练了一个卷积神经网络（Convolutional Neural Network, CNN）来学习噬菌体和宿主的嵌入表示。通过对比学习将已知 PHIs 纳入新的嵌入表示中，使得具有相互作用的噬菌体与宿主在学习到的嵌入空间中尽可能接近。对比学习可以有效处理正样本明显多于负样本的不平衡数据集，其中正样本表示噬菌体和宿主存在相互作用，负样本表示噬菌体和宿主不存在相互作用。噬菌体和宿主的潜在嵌入表示用目前已知的 PHIs 训练，通过计算噬菌体和宿主嵌入的距离来识别 PHIs。这种方法不仅能够实现与其他先进的预测方法类似的高预测准确率，而且还能很容易地扩展到预测未知宿主或多个宿主的情况。

2 相关工作

Edwards 等人 [13] 总结了目前一些用来识别宿主的实验方法，包括斑点检测 [14]、液体测定 [15]、病毒标记 [16]、微流聚合酶链式反应 [17]、噬菌体荧光原位杂交 [18]、单细胞测序 [19] 以及高通量染色体构象捕获测序技术 [20]。然而，上述实验方法是对宿主和噬菌体的筛选标记，但是由于当前噬菌体的种类比较多，如果对每组噬菌体和宿主是否相互作用都进行实验验证，将会耗费大量的人力、物力以及财力。基于此，已有许多研究开始探索根据基因序列或者蛋白质序列的计算方法来预测噬菌体-宿主之间的相互作用关系，这些计算方法可大致分为两类，一类是基于序列相似性的方法，另一类是基于学习的方法。接下来分别对这两种方法进行简要介绍。

2.1 基于序列相似性的噬菌体-宿主相互作用预测方法

大多数基于序列相似性的方法主要利用噬菌体之间或噬菌体与宿主之间的序列相似性预测 PHIs。例如，HostPhinder [21] 计算查询噬菌体与数据库中参考噬菌体之间的相似性来预测宿主。VirHostMatcher [22] 使用病毒和宿主基因组中共有的相似性寡核苷酸频率来预测宿主，

但是在处理短序列时，该方法的性能会下降。vHULK [23] 考虑了噬菌体蛋白质序列与噬菌体蛋白质家族数据库之间的比对得分来预测宿主。鉴于病毒蛋白家族（Viral Protein Families, VPFs）信息对于病毒发现的重要性，VPFClass [24] 利用 VPFs 对病毒基因组进行分类，并根据 VPFs 进行宿主预测。

2.2 基于学习的噬菌体-宿主相互作用预测方法

基于学习的方法主要利用机器学习和深度学习技术发现数据中的模式。这种方法可以处理多样化和复杂的数据，并且当有足够的数据和可用的计算资源时，可以实现准确的预测和分类。例如，WIsH [25] 为每个潜在宿主基因组训练一个 8 阶同质马尔可夫模型，并计算其与查询噬菌体的相互作用概率，以准确预测宿主。Leite 等人 [26] 首先从噬菌体和细菌蛋白质序列中提取特征，然后采用四种类型的机器学习模型（K 最近邻 [27]、随机森林 [28]、支持向量机 [29] 和人工神经网络 [30]）进行宿主预测。基于这项工作，Leite 等人 [31] 做了进一步的探索，提出了基于传统多类学习和单类学习的方法来预测菌株水平上的 PHIs。PredPHI [32] 首先基于 K-Means 聚类选择高质量的负样本来构建平衡的训练集，然后从蛋白质序列中提取三种蛋白质特征的六个数学统计信息，最后构建 CNN 来预测 PHIs。PHIAF [33] 扩展了这项工作，首先使用生成对抗网络在噬菌体和宿主之间生成伪相互作用，然后，利用 CNN 和注意力机制将 DNA 序列和蛋白质序列特征结合起来，进一步提高模型的性能和可解释性。此外，图卷积网络（Graph Convolutional Network, GCN）也被用于预测 PHIs 任务。例如，HostG [34] 利用噬菌体-噬菌体蛋白相似性和噬菌体-宿主 DNA 序列相似性构建知识图谱，并训练基于 GCN 的半监督模型来预测宿主。CHERRY [11] 构建了一个以 k-mer 为节点特征的多模态图来增强学习能力，并采用编码器-解码器结构学习输入序列的最佳嵌入来预测 PHIs。

3 本文方法

3.1 本文方法概述

在本文的工作中，首先，使用频率混沌博弈表示（Frequency Chaos Game Representation, FCGR）作为噬菌体或宿主的基因组序列表示。其次，使用带有批归一化的两层 CNN 作为编码器来学习从 FCGR 到潜在嵌入的映射。此外，使用基于已知 PHIs 的对比学习训练编码器。最后，通过衡量噬菌体-宿主嵌入对之间的距离来预测给定噬菌体的宿主或预测可侵染宿主的噬菌体。本文的工作主要关注预测目标噬菌体的宿主，一方面，和许多预测方法一样，可以选择候选宿主列表中与噬菌体嵌入表示距离最小的宿主作为预测结果。另一方面，也可以选择候选宿主列表中与噬菌体嵌入表示距离小于 m 的所有可能宿主作为预测结果，从而适应噬菌体可能感染多个宿主的情况。本文所提出的方法称为 CL4PHI [35]，其架构如图 1 所示。

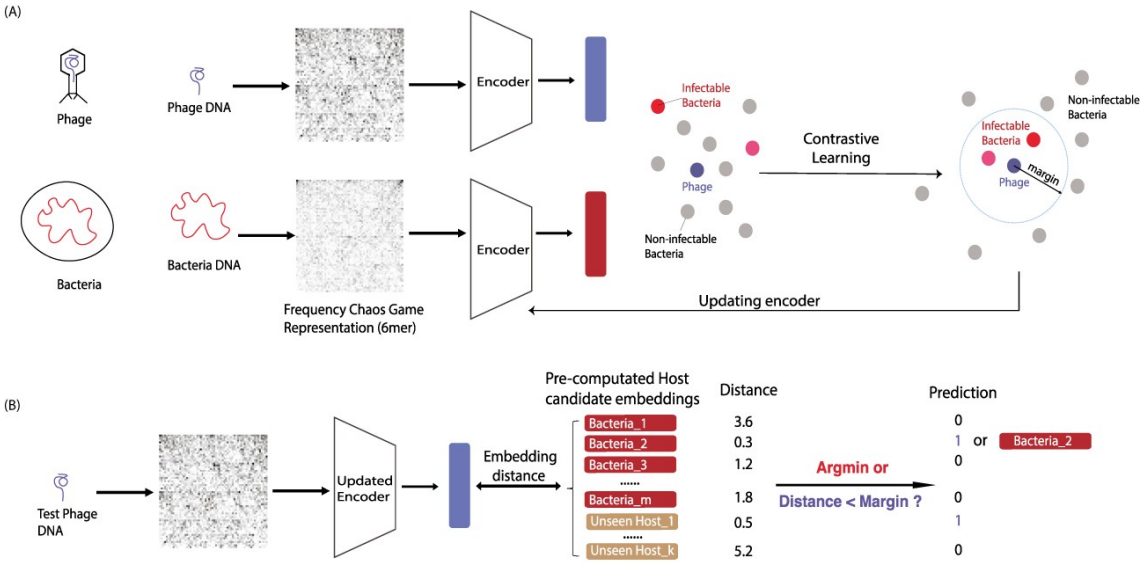

图 1. CL4PHI 的架构。CL4PHI 主要由两个部分组成（A）结合已知 PHIs 的噬菌体和宿主基因组序列表示，以及（B）基于学习到的嵌入表示识别 PHIs。

3.2 基于频率混沌博弈表示重述基因组序列

混沌博弈表示（Chaos Game Representation, CGR）是图形生物信息学的一个里程碑，已经成为机器学习中免对齐序列比较和特征编码的强大工具 [36]。该算法将序列映射到二维空间中，并具有如下独特的性质：序列被表示为唯一的模式、序列被映射到唯一的坐标以及单个坐标编码完整的输入 k-mer 等。由于其特性，CGR 已被用于无配对序列比较和机器学习编码等方面，并且在生物信息学的未来应用中也具有巨大潜力。作为 CGR 的扩展，即 FCGR，它将 k-mer 频率信息压缩为二维矩阵，其元素坐标根据 CGR [37] 排列。本文所采用的是 DNA 序列的 6-mer 频率信息。

给定 DNA 序列 S，获得其 FCGR 表示涉及三个关键的步骤。第一步，获得 DNA 序列 S 的 k-mer 及其频率。第二步，通过式 (1) 迭代计算每个 k-mer 在 FCGR 矩阵中的位置。第三步，将 FCGR 矩阵中的元素按照位置设置为相应 k-mer 的频率。

$$P_i = \frac{(P_{i-1} + PN_i)}{2}, \quad 1 \le i \le k \tag{1}$$

其中，$P_0 = (\frac{\sqrt{4^k}}{2}, \frac{\sqrt{4^k}}{2})$ 表示起始位置，P_k 表示 k-mer 在 FCGR 矩阵中的位置，PN_i 表示 k-mer 中第 i 个碱基对应 FCGR 矩阵中固定顶点的坐标，具体而言，碱基 C 的坐标为 $(0, 0)$、碱基 A 的位置为 $(0, \sqrt{4^k})$、碱基 T 的位置为 $(\sqrt{4^k}, \sqrt{4^k})$ 以及碱基 G 的位置为 $(\sqrt{4^k}, 0)$。

假设给定一个 k-mer 为 “ATGC”，获得其 FCGR 矩阵的过程如图 2 所示。首先计算出子串 “A” 的位置，然后计算出子串 “AT” 的位置，随后计算子串 “ATG” 的位置，最后计算子串 “ATGC” 的位置，即求得当前 k-mer 在 FCGR 矩阵中的位置。

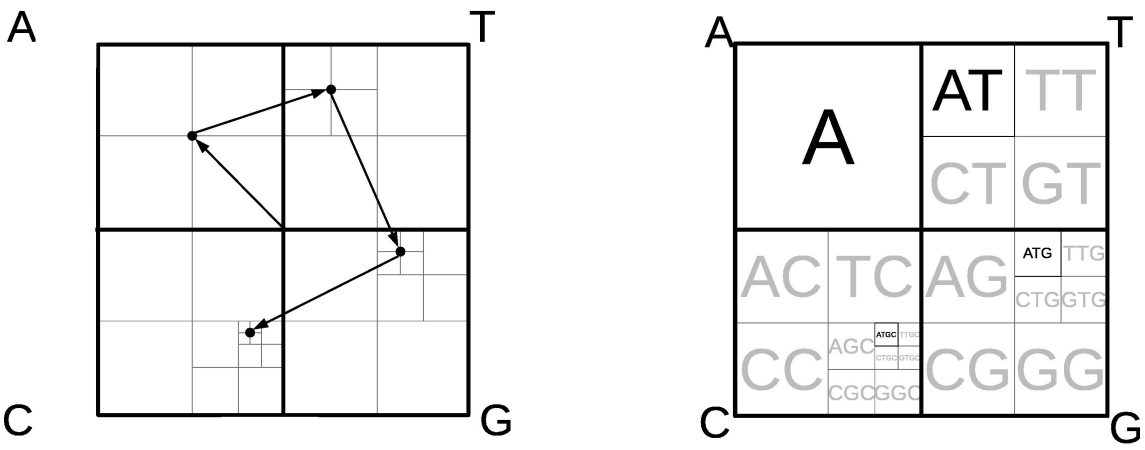

图 2. DNA 序列的频率混沌博弈表示。左图表示 FCGR 矩阵的四个顶点（C、A、T 和 G）。右图表示计算 k-mer “ATGC” 位置的迭代过程。

最终 DNA 序列的每个 k-mer 都对应 FCGR 矩阵中的一个位置，FCGR 矩阵中的数值表示当前 k-mer 的频率。给定 DNA 序列 S 的 FCGR 矩阵灰度可视化效果如图 3 所示，其中颜色越深表示该位置对应的 k-mer 频率越高。

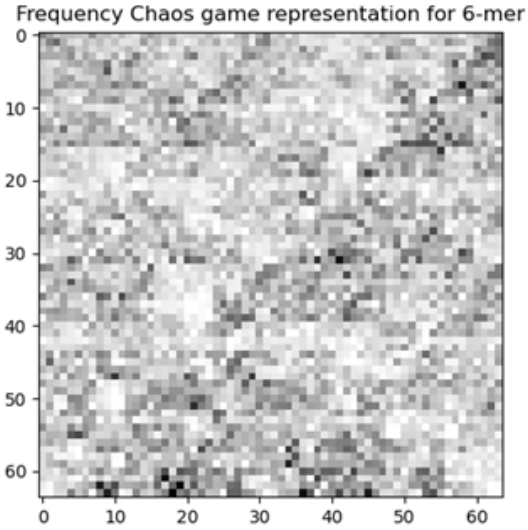

图 3. DNA 序列的 FCGR 矩阵灰度图

3.3 编码器

CL4PHI 的编码器使用带有批量归一化的两层卷积层作为骨架模型来学习从 FCGR 到嵌入潜在嵌入向量的映射。编码器将 DNA 序列的 FCGR 表示映射为 512 维的嵌入向量表示。编码器的架构如图 4 所示。

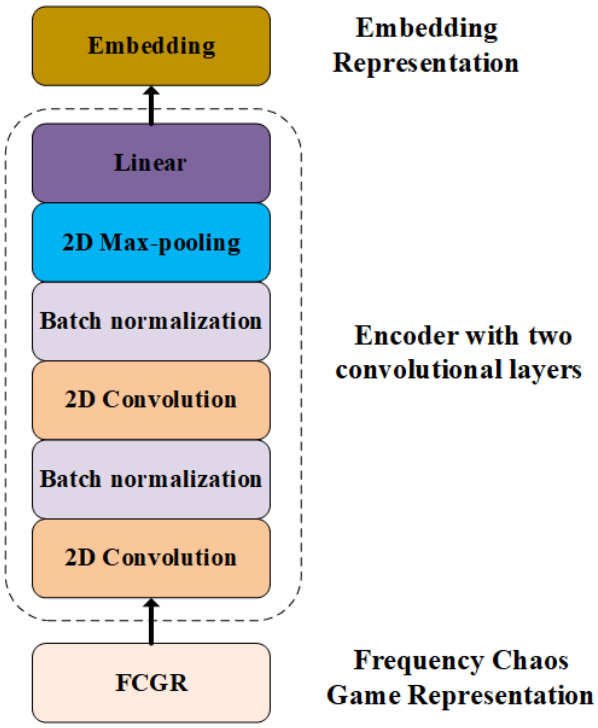

图 4. CL4PHI 的编码器

3.4 对比学习

FCGR 提供了噬菌体基因组序列和候选宿主基因组序列的一般表示。为了进一步将噬菌体-宿主关系考虑到嵌入表示中，可以使用基于已知 PHIs 的对比学习来训练编码器。根据给定的噬菌体-宿主训练数据集准备对比训练样本 $data_{train} = \{X = (p_i, h_j), Y = \{y_{ij}|y_{ij} = \{0,1\}\}\}$。具体而言，将每个噬菌体 $P = \{p_i|i = 1, \ldots, m\}$ 与所有已知候选宿主 $H = \{h_j|j = 1, \ldots, n\}$，其中 m 和 n 分别是噬菌体和宿主的数量。存在相互作用的噬菌体-宿主对被标记为 1（$y_{ij} = 1$），不存在相互作用的噬菌体-宿主对被标记为 0（$y_{ij} = 0$）。FCGR 的潜在嵌入表示是通过最小化对比损失来学习的，如下所示：

$$
\begin{aligned}
L(data_{train}) &= \sum_{i=1}^{m}\sum_{j=1}^{n} L(p_i, h_j, y_{ij}), \\
L(p_i, h_j, y_{ij}) &= y_{ij}\frac{1}{2}Dist(embed(p_i), embed(h_j))^2 \\
&\quad + (1 - y_{ij})\frac{1}{2}max\{0, m - Dist(embed(p_i), embed(h_j))\}^2,
\end{aligned}
\tag{2}
$$

其中，m 是控制嵌入空间中噬菌体与可感染宿主紧密程度的边距，$Dist$ 表示两个嵌入之间的欧几里得距离，$embed(p_i)$ 和 $embed(h_j)$ 分别为通过编码器得到的噬菌体和宿主的 FCGR 潜在嵌入表示。在对比损失中，边距用于限制噬菌体的可感染宿主，同时分离所有不可感染宿主。

3.5 预测噬菌体-宿主相互作用

在考虑已知 PHIs 的基因组序列嵌入后，可以通过如下步骤预测噬菌体-宿主的相互作用。首先，输入噬菌体的 DNA 序列获得 FCGR，其次，通过对比学习得到的编码器获得 FCGR

的嵌入向量，随后，通过计算特征空间中噬菌体嵌入向量与候选宿主嵌入向量的距离，最后，可以通过取距离的最小值或者判断距离是否小于临界值，来预测噬菌体和宿主是否存在相互作用。

4 复现细节

4.1 与已有开源代码对比

本次复现工作代码主要参考原文提供的开源代码 https://github.com/yaozhong/CL4PHI，复现细节如下：

1、复现原文提出的 CLPHI 方法。

2、在编码器模型架构方面，增加了 1 层的卷积层，探索卷积层数对于最后预测的影响。

3、原文的研究主要集中在物种及物种以上的生物学分类水平进行噬菌体-宿主相互作用预测，但这在指导噬菌体疗法的临床应用方面仍具有一定的局限性。因此，我整理了华大基因提供的肺炎克雷伯菌噬菌体和肺炎克雷伯菌相互作用的菌株水平数据（104 株噬菌体和 125 株宿主），将原文的预测噬菌体-宿主相互作用的生物学分类拓展到了菌株水平。

4、优化代码结构，去除了一些冗余代码，使得代码具有更高的可读性。

4.2 实验环境搭建

本实验所使用的主要环境列表如下：

python == 3.8.10

pyfaidx == 0.7.2.2

scikit-learn == 1.3.2

torch == 1.11.0

torchvision == 0.12.0

numpy == 1.24.4

pandas == 2.0.3

4.3 创新点

为了使编码器能够更好地表示 FCGR，在原文模型（记为 CNN2）的基础上添加了一层卷积层，改进的模型记为 CNN3。其中原文模型（CNN2）是基于我的复现，其实验结果与原文结果基本吻合。实验结果表明，CNN3 的预测效果略优于 CNN2。CNN3 的模型如图 5 所示。

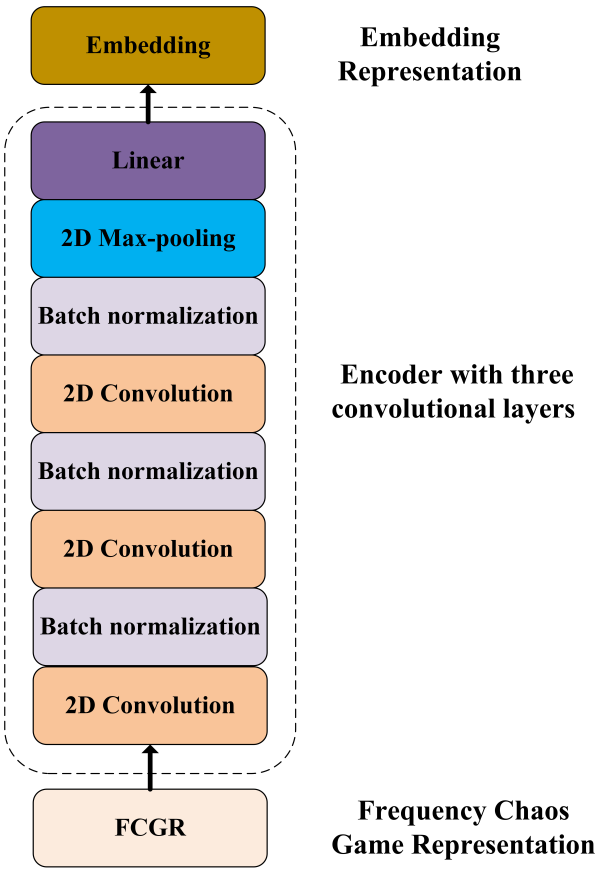

图 5. 改进的编码器 CNN3

5 实验结果分析

5.1 数据集

本文共采用了三个数据集，其中 DeepHost [38] 和 CHERRY [11] 来自于原文，Strain 来自华大基因（剔除了与任何噬菌体都没有相互作用的 5 个宿主）。DeepHost 数据集共有 6734 个噬菌体和 207 个宿主，CHERRY 数据集共有 1306 个噬菌体和 187 个宿主，Strain 共有 104 个噬菌体和 120 个宿主，将噬菌体-宿主之间的相互作用数据分割成训练集、验证集和测试集，数据集的详细信息如表 1 所示。

表 1. **数据集详细信息**

Dataset	Phage	Host	Train	Validation	Test
DeepHost	6734	207	709020	60660	68159
CHERRY	1306	187	213850	6812	60230
Strain	104	120	9984	1248	1248

5.2 实验设置

本文采用的三个数据集的超参数设置如表 2 所示。

表 2. **模型在不同数据集的超参数设置**

Dataset	Learning Rate	Epoch	BatchSize
DeepHost	1e-3	150	3744
CHERRY	1e-3	150	5984
Strain	1e-3	200	256

5.3 实验结果及分析

在本次的复现工作中，将 CNN2 和 CNN3 分别用三个不同的训练集进行训练，在训练中每轮用验证集评估，最终训练好的模型在测试集上进行测试。CNN2 和 CNN3 模型在不同数据集上训练的损失曲线如图 6 所示。

从图 6 可以发现，CNN3 模型在训练过程中比 CNN2 模型更平滑，收敛更好，这表明通过堆叠卷积层能够更好地表示 FCGR。

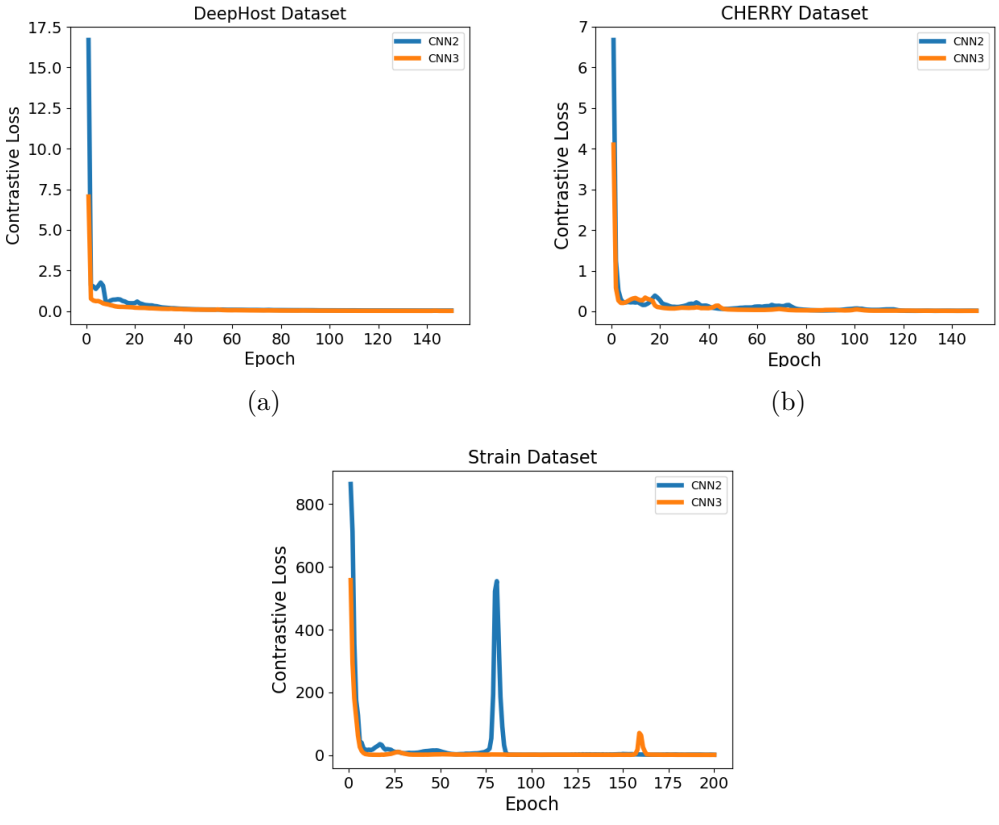

图 6. CNN2 和 CNN3 在 DeepHost（a）、CHERRY（b）和 Strain（c）数据集上的损失曲线

CNN2 和 CNN3 在不同的验证集和测试集上的准确率如表 3 所示。

表 3 的结果表明，CNN3 在大部分的结果中略优于 CNN2。在噬菌体和宿主种类相对较多的 DeepHost 数据集中，CNN3 表现好的原因可能是堆叠多层的 CNN 能够更好嵌入表示噬菌体和宿主。在 CHERRY 数据集中，由于正负样本相对较少，可能的原因是两层堆叠的 CNN2 已经能够很好地表示噬菌体和宿主，因此在测试集中 CNN3 表现不如 CNN2。在 Strain 数据

集中，两个模型均达到了最优值，这体现了原文所提出的编码器能够很好地嵌入表是噬菌体和宿主的 FCGR。

表 3. **模型在不同数据集的超参数设置**

Dataset	Hightest accuracy on the validation set		Average accuracy on the validation set		Accuracy on the test set	
	CNN2	CNN3	CNN2	CNN3	CNN2	CNN3
DeepHost	**0.927**	**0.927**	0.883	**0.901**	0.917	**0.919**
CHERRY	**0.779**	0.733	0.539	**0.599**	**0.584**	0.438
Strain	**0.968**	**0.968**	0.756	**0.825**	**0.966**	**0.966**

6 总结与展望

在这项工作中，提出了一种结合频率混沌博弈表示基因组序列和对比学习嵌入表示来预测噬菌体-宿主相互作用的方法（CL4PHI）。首先，利用频率混沌博弈表示噬菌体和宿主的 DNA 序列，然后，通过对比损失和已知的噬菌体-宿主相互作用训练编码器获得频率混沌博弈表示的嵌入向量，最后，可以通过判断噬菌体和宿主的嵌入向量在特征空间中的距离是否小于临界值，来预测噬菌体和宿主是否存在相互作用。

基于原文，我改进了其中的编码器模块并且将 CL4PHI 方法的生物学分类拓展到了菌株水平。在三个不同数据集上的实验结果表明改进后的模型 CNN3 略优于原模型 CNN2。未来将探索如蛋白质序列和裂解酶等其他信息的特征来预测噬菌体-宿主相互作用。

参考文献

[1] GJ Staats, SJ Mc Carlie, B Van der Walt, and RR Bragg. The linkage between antibiotic and disinfectant resistance. In *Antimicrobial Research and One Health in Africa*, pages 241–274. Springer, 2023.

[2] Md Mominur Rahman, Mst Afroza Alam Tumpa, Mehrukh Zehravi, Md Taslim Sarker, MD Yamin, Md Rezaul Islam, Md Harun-Or-Rashid, Muniruddin Ahmed, Sarker Ramproshad, Banani Mondal, et al. An overview of antimicrobial stewardship optimization: the use of antibiotics in humans and animals to prevent resistance. *Antibiotics*, 11(5):667, 2022.

[3] Akshita Thakur, Akanksha Sharma, Hema K Alajangi, Pradeep Kumar Jaiswal, Yongbeom Lim, Gurpal Singh, and Ravi Pratap Barnwal. In pursuit of next-generation therapeutics: Antimicrobial peptides against superbugs, their sources, mechanism of action, nanotechnology-based delivery, and clinical applications. *International Journal of Biological Macromolecules*, 2022.

[4] DR Harper, J Anderson, and MC Enright. Phage therapy: delivering on the promise. *Therapeutic delivery*, 2(7):935–947, 2011.

[5] Natalya Yutin, Kira S Makarova, Ayal B Gussow, Mart Krupovic, Anca Segall, Robert A Edwards, and Eugene V Koonin. Discovery of an expansive bacteriophage family that includes the most abundant viruses from the human gut. *Nature microbiology*, 3(1):38–46, 2018.

[6] Mikeljon P Nikolich and Andrey A Filippov. Bacteriophage therapy: Developments and directions. *Antibiotics*, 9(3):135, 2020.

[7] Benjamin K Chan, Stephen T Abedon, and Catherine Loc-Carrillo. Phage cocktails and the future of phage therapy. *Future microbiology*, 8(6):769–783, 2013.

[8] Ra' l R Raya and Elvira M H' bert. Isolation of phage via induction of lysogens. *Bacteriophages: Methods and Protocols, Volume 1: Isolation, Characterization, and Interactions*, pages 23–32, 2009.

[9] Anne Chevallereau, Benoît J Pons, Stineke van Houte, and Edze R Westra. Interactions between bacterial and phage communities in natural environments. *Nature Reviews Microbiology*, 20(1):49–62, 2022.

[10] Frederik Schulz, Simon Roux, David Paez-Espino, Sean Jungbluth, David A Walsh, Vincent J Denef, Katherine D McMahon, Konstantinos T Konstantinidis, Emiley A Eloe-Fadrosh, Nikos C Kyrpides, et al. Giant virus diversity and host interactions through global metagenomics. *Nature*, 578(7795):432–436, 2020.

[11] Jiayu Shang and Yanni Sun. Cherry: a computational method for accurate prediction of virus–prokaryotic interactions using a graph encoder–decoder model. *Briefings in Bioinformatics*, 23(5):bbac182, 2022.

[12] Patrick J Deschavanne, Alain Giron, Joseph Vilain, Guillaume Fagot, and Bernard Fertil. Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. *Molecular biology and evolution*, 16(10):1391–1399, 1999.

[13] Robert A Edwards, Katelyn McNair, Karoline Faust, Jeroen Raes, and Bas E Dutilh. Computational approaches to predict bacteriophage–host relationships. *FEMS microbiology reviews*, 40(2):258–272, 2016.

[14] Mathias Middelboe, Amy M Chan, and Sif K Bertelsen. Isolation and life cycle characterization of lytic viruses infecting heterotrophic bacteria and cyanobacteria. *Manual of aquatic viral ecology*, 13:118–133, 2010.

[15] Matthew Henry, Biswajit Biswas, Leah Vincent, Vishwesh Mokashi, Raymond Schuch, Kimberly A Bishop-Lilly, and Shanmuga Sozhamannan. Development of a high throughput

assay for indirectly measuring phage growth using the omnilogtm system. *Bacteriophage*, 2(3):159–167, 2012.

[16] Li Deng, J Cesar Ignacio-Espinoza, Ann C Gregory, Bonnie T Poulos, Joshua S Weitz, Philip Hugenholtz, and Matthew B Sullivan. Viral tagging reveals discrete populations in synechococcus viral genome sequence space. *Nature*, 513(7517):242–245, 2014.

[17] Arbel D Tadmor, Elizabeth A Ottesen, Jared R Leadbetter, and Rob Phillips. Probing individual environmental bacteria for viruses by using microfluidic digital pcr. *Science*, 333(6038):58–62, 2011.

[18] Elke Allers, Cristina Moraru, Melissa B Duhaime, Erica Beneze, Natalie Solonenko, Jimena Barrero-Canosa, Rudolf Amann, and Matthew B Sullivan. Single-cell and population level viral infection dynamics revealed by phage fish, a method to visualize intracellular and free viruses. *Environmental microbiology*, 15(8):2306–2318, 2013.

[19] Roger S Lasken and Jeffrey S McLean. Recent advances in genomic dna sequencing of microbial species from single cells. *Nature Reviews Genetics*, 15(9):577–584, 2014.

[20] Joshua N Burton, Ivan Liachko, Maitreya J Dunham, and Jay Shendure. Species-level deconvolution of metagenome assemblies with hi-c–based contact probability maps. *G3: Genes, Genomes, Genetics*, 4(7):1339–1346, 2014.

[21] Julia Villarroel, Kortine Annina Kleinheinz, Vanessa Isabell Jurtz, Henrike Zschach, Ole Lund, Morten Nielsen, and Mette Voldby Larsen. Hostphinder: a phage host prediction tool. *Viruses*, 8(5):116, 2016.

[22] Nathan A Ahlgren, Jie Ren, Yang Young Lu, Jed A Fuhrman, and Fengzhu Sun. Alignment-free oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences. *Nucleic acids research*, 45(1):39–53, 2017.

[23] Deyvid Amgarten, Bruno Koshin Vázquez Iha, Carlos Morais Piroupo, Aline Maria Da Silva, and João Carlos Setubal. vhulk, a new tool for bacteriophage host prediction based on annotated genomic features and deep neural networks. *bioRxiv*, pages 2020–12, 2020.

[24] Joan Carles Pons, David Paez-Espino, Gabriel Riera, Natalia Ivanova, Nikos C Kyrpides, and Mercè Llabrés. Vpf-class: taxonomic assignment and host prediction of uncultivated viruses based on viral protein families. *Bioinformatics*, 37(13):1805–1813, 2021.

[25] Clovis Galiez, Matthias Siebert, François Enault, Jonathan Vincent, and Johannes Söding. Wish: who is the host? predicting prokaryotic hosts from metagenomic phage contigs. *Bioinformatics*, 33(19):3113–3114, 2017.

[26] Diogo Manuel Carvalho Leite, Xavier Brochet, Grégory Resch, Yok-Ai Que, Aitana Neves, and Carlos Peña-Reyes. Computational prediction of inter-species relationships through omics data analysis and machine learning. *BMC bioinformatics*, 19:151–159, 2018.

[27] Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27, 1967.

[28] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[29] Bernhard Scholkopf. Support vector machines: A practical consequence of learning theory. *IEEE Intelligent systems*, 13, 1998.

[30] Ian H Witten and Eibe Frank. Data mining: practical machine learning tools and techniques with java implementations. *Acm Sigmod Record*, 31(1):76–77, 2002.

[31] Diogo Manuel Carvalho Leite, Juan Fernando Lopez, Xavier Brochet, Miguel Barreto-Sanz, Yok-Ai Que, Grégory Resch, and Carlos Pena-Reyes. Exploration of multiclass and one-class learning methods for prediction of phage-bacteria interaction at strain level. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1818–1825. IEEE, 2018.

[32] Menglu Li, Yanan Wang, Fuyi Li, Yun Zhao, Mengya Liu, Sijia Zhang, Yannan Bin, A Ian Smith, Geoffrey I Webb, Jian Li, et al. A deep learning-based method for identification of bacteriophage-host interaction. *IEEE/ACM transactions on computational biology and bioinformatics*, 18(5):1801–1810, 2020.

[33] Menglu Li and Wen Zhang. Phiaf: prediction of phage-host interactions with gan-based data augmentation and sequence-based feature fusion. *Briefings in Bioinformatics*, 23(1):bbab348, 2022.

[34] Jiayu Shang and Yanni Sun. Predicting the hosts of prokaryotic viruses using gcn-based semi-supervised learning. *BMC biology*, 19:1–15, 2021.

[35] Yao-zhong Zhang, Yunjie Liu, Zeheng Bai, Kosuke Fujimoto, Satoshi Uematsu, and Seiya Imoto. Zero-shot-capable identification of phage–host relationships with whole-genome sequence representation by contrastive learning. *Briefings in Bioinformatics*, 24(5):bbad239, 2023.

[36] Hannah Franziska Löchel and Dominik Heider. Chaos game representation and its applications in bioinformatics. *Computational and structural biotechnology journal*, 19:6263–6271, 2021.

[37] H Joel Jeffrey. Chaos game representation of gene structure. *Nucleic acids research*, 18(8):2163–2170, 1990.

[38] Wang Ruohan, Zhang Xianglilan, Wang Jianping, and LI Shuai Cheng. Deephost: phage host prediction with convolutional neural network. *Briefings in Bioinformatics*, 23(1):bbab385, 2022.