華中科技大學

本科毕业设计（论文）参考文献译文本

译文出处： LI M, ZHANG W. PHIAF: Prediction of Phage-Host Interactions with GAN-Based Data Augmentation and Sequence-Based Feature Fusion[J]. Briefings in Bioinformatics, 2021, 23(1): bbab348.

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

译文要求

一、　译文内容须与课题（或专业内容）联系，并需在封面注明详细出处。

二、　出处格式为

图书：作者.书名.版本（第×版）.译者.出版地：出版者，出版年.起页～止页

期刊：作者.文章名称.期刊名称，年号，卷号（期号）：起页～止页

三、　译文不少于 5000 汉字（或 2 万印刷符）。

四、　翻译内容用五号宋体字编辑，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。

五、　译文及其相应参考文献一起装订，顺序依次为封面、译文、文献。

六、　翻译应在第七学期完成。

译文评阅

导师评语

应根据学校“译文要求”，对学生译文翻译的准确性、翻译数量以及译文的文字表述情况等做具体的评价后，再评分。

评分：___________________(百分制)　　　　指导教师(签名)：___________________

年　　月　　日

PHIAF：使用基于 GAN 的数据增强和基于序列的特征融合预测噬菌体-宿主相互作用

摘要

噬菌体疗法已成为抗生素治疗细菌性疾病最有希望的替代方法之一，鉴定噬菌体-宿主相互作用（PHI）有助于了解噬菌体感染细菌的可能机制，从而指导噬菌体疗法的发展。与湿实验相比，使用计算方法来识别 PHI 可以降低成本，节省时间，更加有效和经济。在本文中，我们提出了一种 PHI 预测方法，该方法具有基于生成对抗网络 (GAN) 的数据增强和基于序列的特征融合 (PHIAF)。首先，PHIAF 应用了基于 GAN 的数据增强模块，该模块生成伪 PHI 以缓解数据稀缺性。其次，PHIAF 融合了源自 DNA 和蛋白质序列的特征以获得更好的性能。第三，PHIAF 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，这也提供了预测模型的可解释性。在计算实验中，通过 5 折交叉验证评估，PHIAF 优于其他最先进的 PHI 预测方法（AUC 和 AUPR 分别为 0.88 和 0.86）。消融研究表明，数据增强、特征融合和注意力机制都有助于提高 PHIAF 的预测性能。此外，最近的文献证实了案例研究中 PHIAF 得分最高的四个新 PHI。综上，PHIAF 可成为帮助加快噬菌体疗法探索的工具之一。

1. 前言

现有报告表明细菌感染可能参与各种疾病的生长和发展，包括霍乱[1]、炎症性肠病[2]、结肠癌[3-5]、破伤风[6] 和不同类型的癌症[7]。抗生素于在 1928 年被发现，随后在临床实践中使用它们来治疗严重的细菌性疾病，挽救了无数生命[8]。然而，由于抗生素的过度使用，细菌已经形成了一些耐药机制[9]。2019 年，美国疾病控制与预防中心报告称，美国每年发生约 280 万例抗生素耐药性感染病例，导致超过 35,000 人死亡[10]； 在欧洲，大约每年有 33,000 人死于抗生素耐药性感染[11]。我们迫切需要开发新的抗生素或替代疗法，以避免抗生素耐药性感染的进一步恶化。但由于抗生素生产成本高、预期效益不理想、研发时间长，许多制药公司不再开发新的抗生素[12,13]。因此，研究人员寄希望于寻找替代疗法来减少抗生素耐药性感染和治疗细菌性疾病。

噬菌体和细菌的基因组在分子和生态共同进化过程发生了改变，并在其基因组序列中留下信号，使得研究人员得以据此预测 PHI[15]，目前已经开发了各种基于噬菌体和宿主基因组序列的 PHI 计算方法[16-18]。例如，Ahlgren 等人[19] 提出了基于 DNA 序列的 VirHostMatcher (VHM)，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。

然而，VHM 的运行时间阻碍了它在大型数据集上的发展，因此 Galiez 等人[20]提出 WIsH 通过构建马尔科夫模型来预测噬菌体的原核宿主来减少运行时间。与 VHM 相比，WIsH 的运行时间减少了数百倍。除了通过计算噬菌体和宿主之间的相似性来预测 PHI 的 VHM 和 WIsH 之外，研究人员还使用了各种机器学习分类器，包括逻辑回归 (LR)、支持向量机 (SVM)、随机森林 (RF) 和朴素贝叶斯（NB）以预测 PHI[21]。此外，有研究人员开发了 PHP[22] 和 VIDHOP [23] 来增强 PHI 预测性能。PHP 通过计算病毒和宿主基因组序列之间 k-mer 频率的差异训练了一个高斯模型，而 VIDHOP 使用深度神经网络来预测三种不同病毒（甲型流感病毒、狂犬病狂犬病毒和 A 型轮状病毒）的宿主。

一些研究表明，蛋白质在噬菌体和宿主的生物学过程中发挥着重要作用[24, 25]； 因此，研究人员提出了基于蛋白质序列的 PHI 预测方法[26, 27]。例如，Leite 等人[28, 29] 利用来自噬菌体和宿主蛋白的一级结构序列和经典分类器，包括随机森林（RF）、支持向量机（SVM）、逻辑回归（LR）、k 近邻（KNN）、多层感知器（MLP） 和朴素贝叶斯（NB）等，来预测 PHI。在上述方法的基础上，Li 等人[30] 使用卷积神经网络（CNN） 来提高 PHI 预测的性能。

尽管现有方法在 PHI 预测中取得了良好的性能，但仍然存在一些挑战。首先，数据库中有数千个经过实验验证的 PHI [31-34]，但只有几百个非冗余 PHI 可用并可用于构建预测模型。这种限制阻碍了高性能预测模型的发展。其次，大多数现有方法使用噬菌体和宿主的 DNA 序列或蛋白质序列来构建预测模型，但很少结合两种类型的序列特征。第三，尽管已经使用了多种特征和机器学习技术来构建预测模型，但这些模型往往缺乏足够的可解释性，这阻碍了对 PHI 机制的阐述。

近年来，深度学习技术在生物信息学领域受到广泛关注，研究人员已将此类技术应用于处理不同的任务[35, 36]。生成对抗网络 (GAN) 作为深度学习技术的一个分支，最初用于图像处理[37, 38]，后来在数据增强方面表现出优异的性能。例如，Wan 等人[39] 成功地使用 GAN 根据蛋白质序列生成生物物理特征。同时，研究人员为深度学习开发了一种注意力机制[40]，以增加预测模型的可解释性和以提高预测性能。这些深度学习技术的发展促使我们进一步增强和改进 PHI 预测。在当前的研究中，我们提出了一种新的 PHI 预测方法，简称 PHIAF，基于 GAN 数据增强和基于序列的特征融合来解决 PHI 预测的各种挑战。首先，PHIAF 使用 GAN 构建数据增强模块，生成高质量的伪样本，以克服 PHI 数据稀缺的瓶颈。其次，PHIAF 融合了噬菌体和宿主的 DNA 和蛋白质序列编码的不同特征，以提高预测性能。第三，PHIAF 利用 CNN 构建 PHI 预测模块，并将注意力机制整合到 CNN 中，以提供预测

模型的可解释性。实验结果表明，PHIAF 优于最先进的 PHI 预测方法。消融研究和讨论表明，数据增强模块生成的伪样本、DNA 和蛋白质序列衍生特征的融合以及 CNN 中的注意力机制有效地提高了 PHIAF 的性能。

2. 材料和方法

2.1 数据集

我们于 2021 年 3 月从四个广泛使用的数据库（包括 MVP[31]、PhagesDB[32]、VHDB[33] 和 NCBI[34]）下载数据（包括噬菌体、宿主及其相互作用），并合并这些数据以构建具有更多 PHI 的数据集以供我们研究。这四个数据库中的数据经过以下处理：首先，我们删除未在文献中发表或未包含在 NCBI 记录中的 PHI，以确保 PHI 可靠。其次，我们根据噬菌体的定义删除错误标记为噬菌体/宿主的数据（噬菌体定义为是在细菌和古细菌中感染和复制的病毒）。删除的数据包括不属于病毒的噬菌体和不属于细菌或古细菌的宿主。第三，我们从 NCBI 数据库中提取过滤后的噬菌体和宿主的全基因组序列和编码蛋白序列。

经过上述过程，我们将四个数据库中剩余的噬菌体和宿主结合起来，去除重复，得到 5331 个噬菌体和 235 个宿主之间的共 5399 个相互作用。由于噬菌体的数量远大于宿主的数量，一个宿主可能与多个噬菌体相互作用。我们使用算法 1 去除每个宿主具有高相似性的冗余噬菌体（不同相似性测量之间的比较以及噬菌体 对预测性能的影响分别在补充材料的第 1 节和第 2 节中提供）。我们将 0.90 设置为高相似度阈值，这与 CD-HIT 工具[41] 的默认阈值相同。减少冗余后，我们获得了一个基准数据集，其中包含 304 个噬菌体和 235 个宿主之间的 312 个相互作用，可用于更好地评估预测模型的性能。在这个数据集中，我们将 312 个已知 PHI 设置为正样本，并从所有未知 PHI 中选择负样本，同时确保正样本和负样本的数量相等。

算法 1：数据处理以去除多余的噬菌体。

Require: 宿主集合，H ={h_1, h_2, ..., h_n}, n 代表宿主数量; 对应不同宿主的噬菌体集合, P ={P_1, P_2, ..., P_n}, Pi = [p_1, p_2, ..., p_m]代表感染 host Hi 的噬菌体集合, i ∈ [1, n]; 不同噬菌体的相似性矩阵, S = {$s_{p1,p2}$, $s_{p1,p3}$, ..., $s_{pm-1,pm}$ }, m 代表噬菌体的数量;

Ensure: 非冗余噬菌体和宿主之间的相互作用，I；

```
function main(P, H, S)
    I ← []
    for k ← 1 to n do
        Ik ← []
        Ik ← del_redundant(Pk)
        I = I + Ik
    end for
return I
end function
function del_redundant(Pk)
    Pdel ← [], Pnew ← []
    Ik ← Ik + (Pk[0], hk)
    Pdel ← Pdel + Pk[0]
    for j ← 1 to m do
        if sPk[0],Pk[j] > 0.90 then
        Pdel ← Pdel + Pk[j]
        end if
    end for
    Pnew ← Pk − Pdel
    if length(Pnew) > 1 then
        del_redundant(Pnew)
    end if
    if length(Pnew) = 1 then
        Ik ← Ik + (Pnew[0], hk)
    end if
    return Ik
end function
```

2.2 PHIAF

PHIAF 由三个主要模块组成：特征提取、数据增强和 PHI 预测。PHIAF 的示意图如图 1 所示。首先，将噬菌体和宿主的 DNA 和蛋白质序列编码为特征（图 1A）。其次，基于 GAN 的数据增强模块用于生成伪 PHI（图 1B）。最后，在使用注意力机制的 CNN 框架下构建了一个 PHI 预测模块，并利用来自 DNA 和蛋白质序列的特征在重塑为适当的形式后预测 PHI（图 1C）。

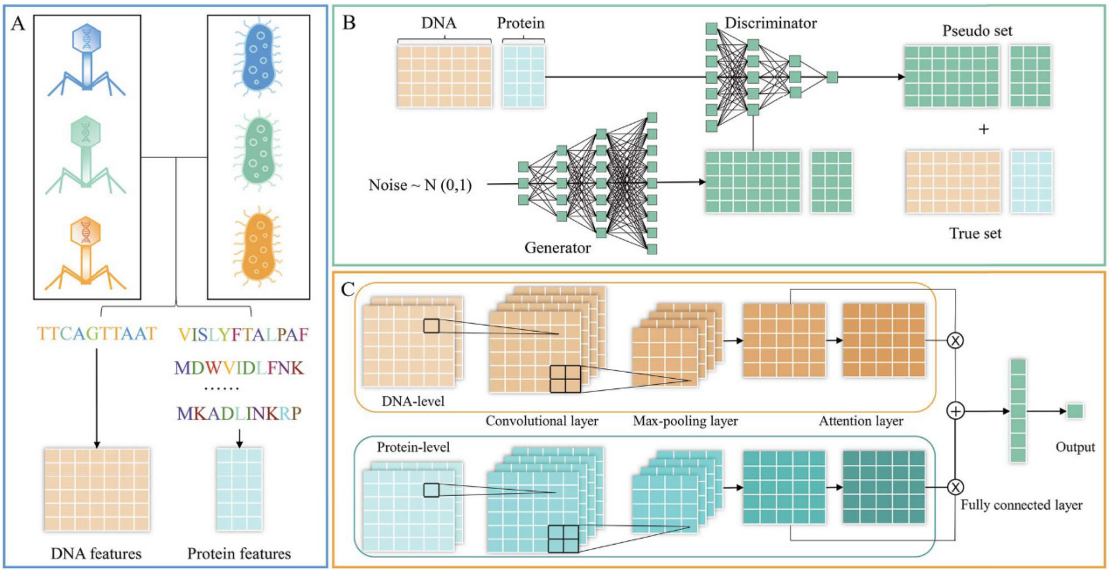

图 1 PHIAF 概览。(A) 特征提取模块。(B) 数据增强模块。(C) PHI 预测模块。

特征提取模块

一些研究表明，DNA 和蛋白质序列在噬菌体和宿主的生物进化中发挥着基础性作用。为了获得更全面和有效的信息，我们从噬菌体和宿主的 DNA 和蛋白质序列中提取特征，如表 1 所示； 这些功能的详细信息在补充材料的第 3 节中提供。由于噬菌体和宿主的 DNA 序列具有不同的长度（对噬菌体和宿主的 DNA 和蛋白质序列长度分布的分析在补充材料的第 4 节中提供），我们考虑了几个与序列长度无关的 DNA 序列衍生特征，包括 Kmer、反向互补 Kmer (RCKmer)、核酸组成 (NAC)、二核苷酸组成 (DNC)、三核苷酸组成 (TNC)、k 间隔核酸对组成 (CKSNAP) 和电子离子相互作用赝势三核苷酸（PseEIIP）。我们使用软件 iLearn[42] 计算这些特征，并获得每个噬菌体/宿主 DNA 序列的 340 维特征向量。我们按照之前的研究[28-30] 提取广泛使用的蛋白质序列衍生特征，包括氨基酸组成 (AAC)、组成蛋白质的化学元素的丰度 (AC) 和蛋白质的分子量 (MW)。由于每个噬菌体/宿主有多个蛋白质序列，我们考虑六个算子（平均值、最大值、最小值、标准差、方差和中值）来整合蛋

白质序列的特征，并为每个噬菌体/宿主获得一个 162 维的特征向量。

表 1 描述从 DNA 和蛋白质序列提取特征

序列	特征	描述
DNA	Kmer	k 个相邻核酸的出现频率（k = 3）
	RCKmer	Kmer 的一种变体，它去除了反向补码 Kmers
	NAC	核苷酸序列中每种核酸类型（A、C、G、T）的频率
	DNC	核苷酸序列中两种核酸类型的频率
	TNC	核苷酸序列中三种核酸类型的频率
	CKSNAP	被任何 p 个核酸分开的核酸对的频率 (p = 5)
	PseEIIP	每个序列中三核苷酸的平均 EIIP 值 (A: 0.1260, C: 0.1340, G: 0.0806, T: 0.1335)
Protein	ACC	蛋白质序列中每个氨基酸的频率
	AC	组成蛋白质的选定化学元素的丰度
	MW	蛋白质序列的分子量

数据增强模块

GAN[37] 是一种新型的生成模型，旨在通过精确学习真实样本的底层分布来生成高质量的伪样本。该模型受到了大量关注，并在许多领域取得了出色的表现[43, 44]。在这项研究中，我们使用 GAN[39]来解决模型训练数据集中 PHI 的数据稀缺性。

我们首先将真正的正样本设置为$\mathrm{I} = \{(\mathrm{p}_1, \mathrm{h}_1), (\mathrm{p}_2, \mathrm{h}_2), \ldots, (\mathrm{p}_\mathrm{m}, \mathrm{h}_\mathrm{n})\}$，$V = \{V_{p1}, h_1, V_{p2}, h_2, \ldots, V_{pm}, h_n\}$ 来表示这些样本的特征向量，由上面编码的噬菌体和宿主的 DNA 和蛋白质特征组成。m 和 n 分别代表噬菌体和宿主的数量。这些正样本（V）的特征向量被输入到 GAN 中以生成高质量的样本伪特征向量，其中 GAN 由两个神经网络（生成器和判别器）组成，它们相互“对抗”以学习分布 的真实样本。一个网络（生成器）尝试通过五个全连接层（公式 1）生成伪样本。另一个由四个全连接层（公式 2）组成的网络（鉴别器）试图区分给定样本是否真实。每个网络的任务越来越好，直到达到平衡，此时生成器无法制作更好的样本，鉴别器无法区分真实样本和伪样本。

$$\mathrm{O_{ge}} = \mathrm{FC_t}\left(\mathrm{FC_r}\left(\mathrm{FC_r}\left(\mathrm{FC_r}(\mathrm{FC_r}(\mathrm{V}))\right)\right)\right) \tag{1}$$

$$\mathrm{O}_{\mathrm{di}} = FC_l\left(FC_l\left(FC_l\left(FC_l\left(V^{'}\ ,V\right)\right)\right)\right) \tag{2}$$

其中 V′是伪样本的特征向量，O_{ge}是生成器输出，O_{di}是判别器输出，FC_{l}（或 FC_{r}）表示具有 LeakyReLU（或 ReLU）激活函数的 MLP，FCt 表示具有 Tanh 激活函数的 MLP。

我们使用 KNN(k = 1) 和留一法交叉验证 (LOOCV) 进行分类器双样本测试 (C2ST) 方法[45] 来区分真实样本和伪样本。C2ST 方法涉及接受或拒绝 P 等于 Q 的零假设（其中 P 和 Q 是两个相等大小的样本集的分布）。如果接受原假设，则预测被保留样本的标签（0 或 1）的分类准确度将接近随机水平（即 0.50）。因此，在 LOOCV 下 KNN 分类准确率最接近 0.50 时，真实样本和伪样本无法区分。最后，我们选择无法区分的伪样本来放大我们的数据集，表示为$I' = (p_1', h_1'), (p_2', h_2'), \ldots, (p_m', h_n')$。

上述处理用于放大我们数据集中的正样本。由于所有未知的 PHI 都是候选负样本，并且远远超过正样本，因此我们从这个候选集中随机选择负样本，以确保正负样本的数量相等。最后，我们结合真实的正样本、选择的负样本和伪正样本来构建一个增强数据集，用于训练预测模型。

PHI 预测模块

基于增强数据集，我们使用噬菌体和宿主的 DNA 和蛋白质序列衍生特征来构建预测模型

如上一节所述，我们为每个噬菌体/宿主提供两种类型的特征向量，分别从 DNA 和蛋白质序列中提取。由于序列衍生特征通常具有复杂的短程/长程依赖性，我们将 DNA/蛋白质特征向量重塑为“图像”，以捕捉它们维度之间的复杂关系[46]。我们首先采用 Min-Max 归一化将特征向量中的值归一化到 0 到 1 的范围内。令 N 表示 DNA/蛋白质的特征向量的维数；然后，我们通过按行放置值，将特征向量重塑为 n×n 的“图像”，其中 n 满足条件：(n − 1) × (n − 1) < N 且 N ≤ n × n. 当 N < n × n 时，我们通过向剩余的 n × n - N 个条目添加零来填充空值。最后，我们获得每个噬菌体（或宿主）的 DNA 和蛋白质序列衍生特征矩阵，分别表示为 M_d^P 和 M_p^P（ M_d^H 和 M_p^H ）。

然后，我们构建了一个双层架构（DNA 和蛋白质水平）以从 DNA 和蛋白质特征矩阵中提取更深层次的特征。对于 DNA 水平特征，我们将噬菌体和宿主的 DNA 序列特征矩阵跨通道堆叠形成组合矩阵，然后将该组合矩阵输入到两层 CNN 中，以生成具有更有意义信息的特征图O_{d}。CNN 包括一个卷积层和一个最大池化层。

$$O_d = MaxPool\left(Conv2D([M_d^P, M_d^H])\right) \tag{3}$$

其中 Conv2D 和 MaxPool 分别表示卷积层和最大池化层，[·,·] 表示跨通道堆叠。类似地，噬菌体和宿主的蛋白质衍生特征矩阵跨通道组合并馈送到同一个 CNN 以在蛋白质水平上产生特征图$\mathrm{O_p}$ 。

$$O_p = MaxPool\left(Conv2D([M_p^P, M_p^H])\right) \tag{4}$$

注意机制旨在模仿人脑的动作，在机器学习任务中选择性地专注于几个重要部分，而忽略其他部分[47]。我们拥有的 DNA 和蛋白质水平的特征可能对 PHI 预测做出不同的贡献。因此，我们在模型中引入了注意力机制，添加了注意力层来捕获重要特征，然后整合这些特征。

在注意力层中，将 DNA 和蛋白质级别的特征图（$\mathrm{O_d}$ 和 $\mathrm{O_p}$）输入全连接层，分别计算权重向量（α_d和α_p）；然后，特征图乘以相应的权重向量。最后，注意力层的输出计算如下：

$$\mathrm{O_{att}} = \mathrm{O_d} \otimes \alpha_d + \mathrm{O_p} \otimes \alpha_p \tag{5}$$

其中⊗是逐元素乘法。最后，注意力层的输出被输入到一个两层的 MLP 中，以产生样本是 PHI 的概率。

$$\mathrm{Pred} = FC_s\left(FC_r(O_{att})\right) \tag{6}$$

其中FCs表示具有 Sigmoid 激活函数的 MLP。

PHIAF 优化

PHIAF 的数据增强和 PHI 预测模块中有几个重要的超参数。在数据增强模块中，我们将生成器的不同全连接层中的神经元数量设置为 128、256、512、1024 和 1004，将鉴别器中的神经元数量分别设置为 512、256、128 和 1。在 PHI 预测模块中，我们在卷积层中设置了 32 个过滤器和一个过滤器的大小为 3×3，并将 DNA 和蛋白质水平的最大池化层的大小分别设置为 3×3 和 2×2。这两个最大池化层的大小差异确保了$\mathrm{O_d}$ 和 $\mathrm{O_p}$的维度相等。我们还考虑了学习率（可选值为 0.1、0.01、0.001 和 0.0001）、一次训练所抓取的数据样本数量（可选值为 8、16、32、64 和 128）、dropout 层的丢失率（可选值包括 0.25、0.5 和 0.75）和全连接层的神经元个数（可选值包括 16、32、64、128、256 和 512）来确定最优参数。给定不同参数的模型性能在补充材料的第 5 节中提供。

此外，我们在 PHI 预测模块中利用 dropout 和批量归一化层来防止过度拟合并提高泛化性[48, 49]。数据增强模块和 PHI 预测模块是独立训练的。此外，我们采用 Wasserstein 损

失添加梯度惩罚作为数据增强模块的损失函数，并为 PHI 预测模块使用二元交叉熵损失函数。所有上述损失函数都由 Adam 优化器[50] 进行了优化。

3. 结果

3.1 性能评估

在这项研究中，我们采用 5 折交叉验证 (5-CV) 来评估 PHIAF 的性能。 我们数据集中的所有样本都被随机分成五个大小相等的子集。 交叉验证过程重复五次，每个子集依次作为测试集，其余四个子集作为训练集。通过平均五个测试集结果生成最终的 5-CV 结果。我们采用几种常用的评估方法[51,52] 来评估 PHIAF 的性能和最先进的方法，包括特异性 (Spe)、灵敏度 (Sen)、准确度 (Acc)、F1 分数 (F1) ，接收者操作特征曲线下面积（AUC）和精确召回曲线下面积（AUPR）。 措施计算如下：

$$\text{Spe} = \frac{TN}{TN+FP} \tag{7}$$

$$\text{Sen} = \frac{TP}{TP+FN} \tag{8}$$

$$\text{Acc} = \frac{TP+TN}{TP+TN+FP+FN} \tag{9}$$

$$F1 = \frac{2*TP}{2*TP+FP+FN} \tag{10}$$

其中 TP（TN）表示在预测中正确分类的正（负）样本的数量，FP（FN）表示错误识别的负（正）样本的数量。

3.2 与最先进的方法比较

为了证明 PHIAF 的有效性，我们将其与以下最先进的方法进行了比较，包括基于 DNA 序列的算法和基于蛋白质序列的方法

- VHM [19] 是一种基于 DNA 序列的方法，它计算噬菌体和宿主的寡核苷酸频率模式之间的距离，并根据该距离获得 PHI 的可能性。

- WIsH [20] 是一种基于 DNA 序列的方法，它为每个宿主训练马尔可夫模型并计算所有噬菌体的相互作用概率。

- PHP [22] 是一种基于 DNA 序列的方法，它构建了一个高斯模型，使用病毒和宿主基因组序列之间的 k-mer 频率来预测 PHI。

- RF、SVM、KNN 和 MLP 是广泛使用的机器学习分类器； Leite 等人[28]利用蛋白质序列衍生的特征和这些分类器来预测 PHI。

- PredPHI [30] 是一种基于蛋白质序列的方法，可在 CNN 框架下预测 PHI。

我们首先使用 5-CV 将 PHIAF 与这些方法进行比较。如图 2 所示，PHIAF 在 AUC 和 AUPR 方面优于所有比较方法。一般来说，基于 DNA 序列的方法（VHM、WIsH 和 PHP）比基于蛋白质序列的方法（RF、SVM、KNN、MLP 和 PredPHI）产生更好的结果，表明来自 DNA 序列的信息在 PHI 中可能发挥更重要的作用。我们的 PHIAF 模型融合了源自 DNA 和蛋白质序列的信息，优于只基于 DNA 序列或蛋白质序列的模型，在 AUC 和 AUPR 方面平均提高了 13.63% 和 14.75%。

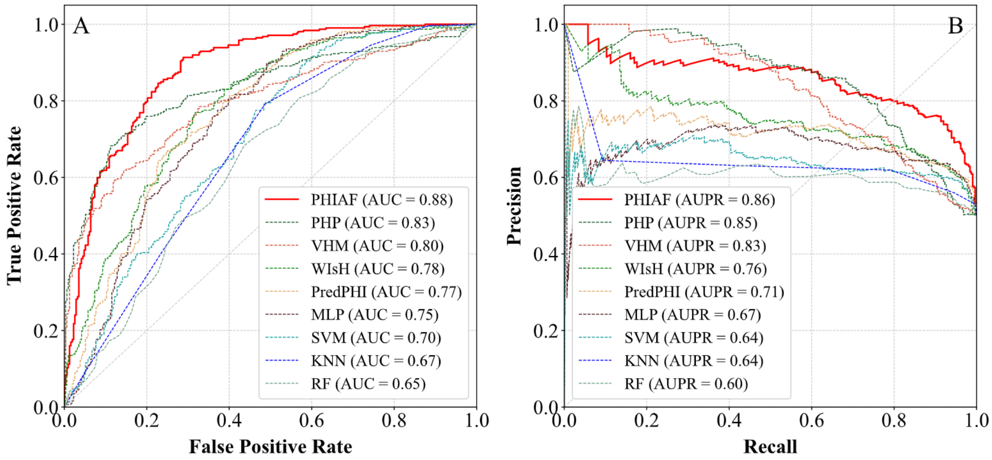

图 2 使用 5 折交叉验证比较 PHIAF 和其他最新方法的性能

模型对未知数据的预测能力也很重要。因此，我们随机选择数据集中三分之一的噬菌体和宿主，并将它们及其相互作用作为测试集。剩余的噬菌体和宿主及其相互作用用于训练预测模型。在此实验设置下，测试集中的噬菌体或宿主不包括在训练集，它们被视为未知数据。我们基于训练集训练预测模型，然后将训练好的模型应用于测试集，以评估预测模型在未知数据上的性能。为了避免评估偏差，我们重复训练和测试 10 次，并采用平均/中值来评估性能。如图 3 所示，PHIAF 优于其他方法，产生更高的平均 AUC (0.86) 和 AUPR 得分 (0.85) 以及合理的标准差，AUC 和 AUPR 得分的中位数分别为 0.88 和 0.87。与 5-CV 结果类似，基于 DNA 序列的方法比基于蛋白质序列的方法产生更好的结果。此外，5-CV 的结果与测试未知数据的结果比较表明，我们的方法在两种情况下都实现了非常相似的性能（AUC 和 AUPR 相差 2% 和 1%），这些结果表明我们提出的方法是稳健的并且可以在未知的数据上表现良好，这表明 PHIAF 是一种很有前景的工具，可以从序列数据中识别 PHI。

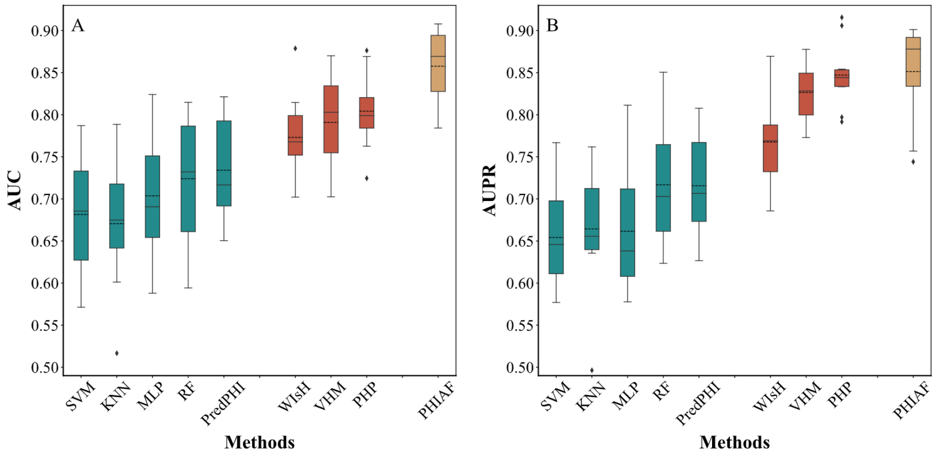

图 3 针对未见数据比较 PHIAF 和其他最新方法的性能。箱式图的实线表示中位数，虚线表示平均值。

3.3 消融研究

上述比较说明了 PHIAF 的有效性，而 PHIAF 的成功是其设计的结果：基于 GAN 的数据增强模块，可生成高质量的伪样本，以及 PHI 预测模块，将 DNA 和蛋白质序列衍生特征与一种注意力机制。在这里，我们进行了消融研究，以详细说明这些组件的贡献。我们考虑以下 PHIAF 变体：

- PHIAF-D 是一种不使用 DNA 级特征的变体。
- PHIAF-P 是一种不使用蛋白质水平特征的变体。
- PHIAF-A 是一种不使用注意层的变体。
- PHIAF-G 是一种不使用伪样本的变体。

表 2 显示了 PHIAF 及其四种变体在 5-CV 下的结果。当移除任何组件时，PHIAF 的性能会降低，这意味着所有组件对于 PHIAF 都是关键的。

表 2 PHIAF 及其四种变体在 5-CV 下的性能比较

	AUPR	AUC	F1	Acc	Sen	Spe
PHIAF-D	0.80	0.85.	0.77	0.77	0.79	0.75
PHIAF-P	0.84	0.86	0.79	0.79	0.80	0.79
PHIAF-A	0.82	0.86	0.64	0.73	0.75	**0.88**
PHIAF-G	0.79	0.82	0.73	0.73	0.75	0.71
PHIAF	**0.86**	**0.88**	**0.81**	**0.81**	**0.83**	0.78

PHIAF-G 的性能下降幅度最大，AUC 和 AUPR 分别下降了 6% 和 7%，其次是 PHIAF-D（AUC 和 AUPR 分别下降了 3% 和 6%）和 PHIAF-A（AUC 和 AUPR 分别下降了 2% 和 4%）。PHIAF 和 PHIAF-G 的结果比较表明，伪样本的使用有效地增强了 PHI 预测。此外，PHIAF-D 和 PHIAF-P 之间的比较表明，DNA 序列衍生的特征对于 PHI 预测比源自蛋白质序列的特征更有效。去除注意力层也会导致性能变差，这表明必须考虑特征的差异。

4. 讨论

消融研究表明，PHIAF 的主要成分对 PHI 预测做出了重要贡献。此外，我们从三个方面分析 PHIAF。为了证明真实样本和伪样本无法区分并且伪样本可以用作训练正样本，我们使用 t 分布随机邻域嵌入 (t-SNE) [53] 来可视化在 GAN 的不同训练迭代中真实和伪正样本的 2D 分布（图 4）。在第 1 次迭代中（图 4A），真实样本的分布远离伪样本，导致 LOOCV 精度为 1.00，这表明生成器尚未学习真实样本的分布。如图 4B 所示，生成器开始捕捉真实样本的特征，而判别器在 1000 次迭代后无法完全区分真实样本和伪样本（LOOCV 精度为 0.93）。然后，分布逐渐收敛(图 4C，LOOCV 精度达到 0.69)。经过 20000 次迭代后，生成器和判别器达到平衡，真实和伪样本遵循相似的分布（图 4D，LOOCV 精度为 0.50）。因此，可知数据增强生成了高质量的伪正样本。

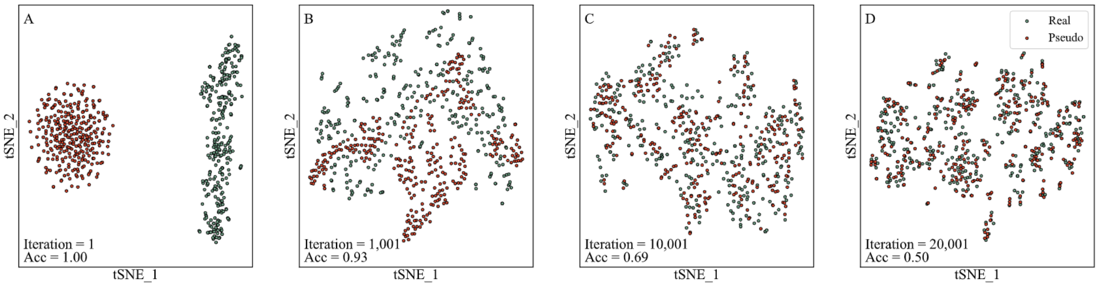

图 4 GAN 不同训练迭代期间真实和伪样本的 t-SNE 变换二维可视化（A，1 次迭代；B，1001 次迭代；C，10 001 次迭代；D，20 001 次迭代）

此外，我们分析了分配给不同特征的注意力层中的权重，以研究注意力机制学习到的特征的重要性。图 5A 和图 5 B 显示正负样本的注意力权重分布相似，这表明某些特征在正负样本中起着同样重要的作用。此外，我们比较了 DNA 和蛋白质水平的注意力权重，其中蛋白质水平的特征通常被分配比 DNA 水平的特征更低的权重。这些结果证实，对于 PHI 预测，DNA 水平的特征比蛋白质水平的特征更重要，并且可知 CNN 中的注意力层不仅提

高了预测性能，而且还能够根据不同特征的重要性有效地分配了权重。

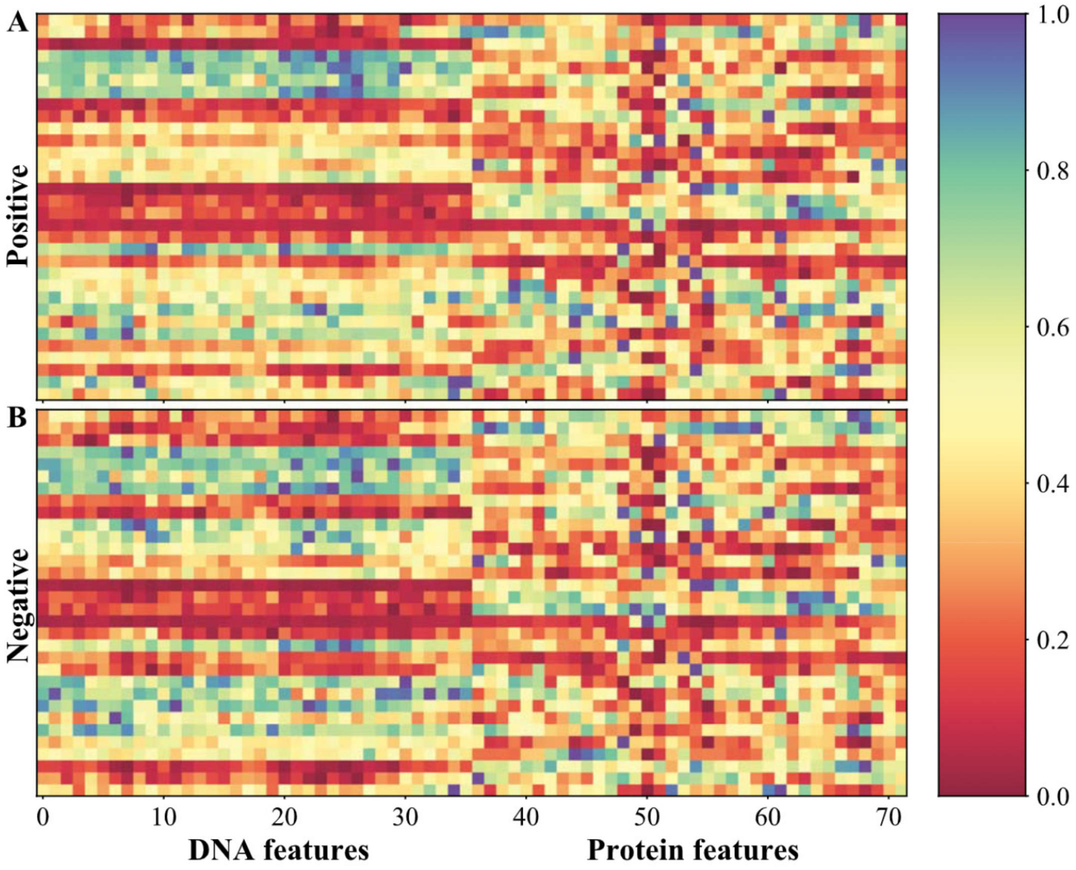

图 5 注意力层数据集中正负样本的平均权重（A 为正样本，B 为负样本）。

最后，我们基于具有不同正负样本比例（1:1、1:2、1:3、1:4、1:5、2:1、3:1、4:1 和 5:1) 分析负样本或伪正样本的数量如何影响 PHIAF 的性能。如图 6A 和 6C 所示，PHIAF 在具有不同的正负样本比例的数据集上产生相似的 AUC 和 AUPR 分数（变化在 2% 以内）。随着阴性样本数量的增加（图 6B），PHIAF 的灵敏度降低（从 0.83 降低到 0.30）和特异性更高（从 0.78 提高到 0.98）。相比之下，当假阳性样本数量增加时，PHIAF 产生更高的灵敏度（从 0.79 增加到 0.95）和更低的特异性（从 0.78 降低到 0.62）（图 6D）。这些结果表明，不平衡的数据集未能显著提高 PHI 预测的性能，并且会增加样本错误判断的概率。

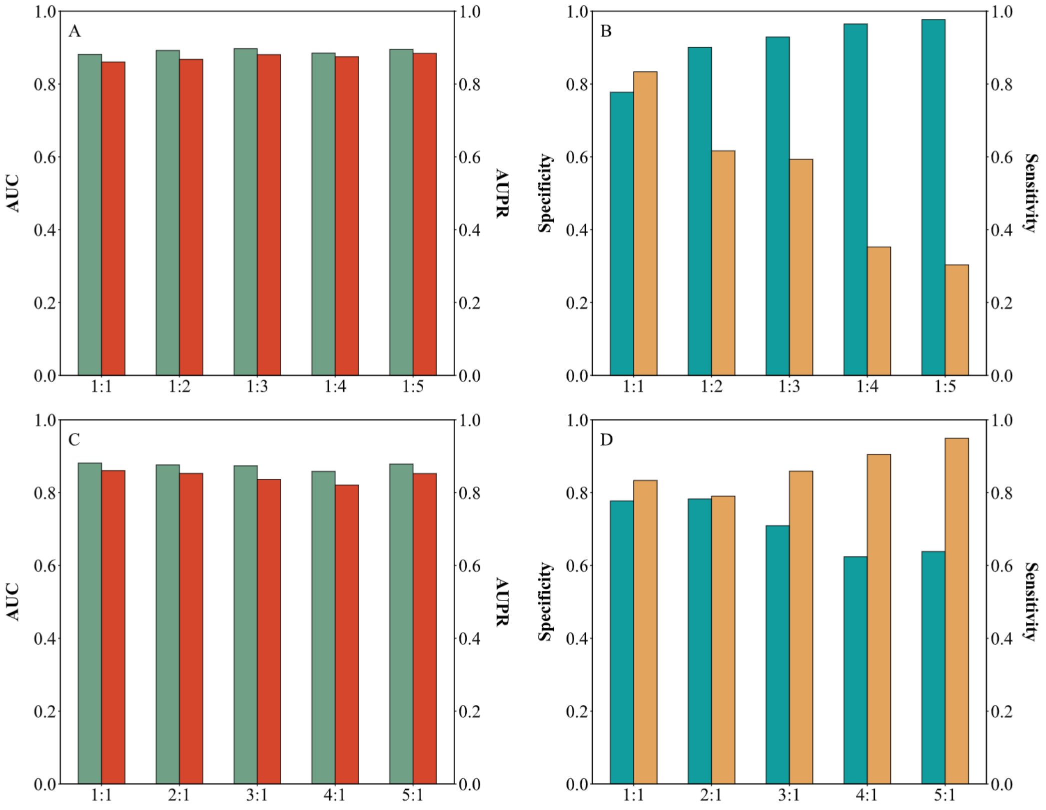

图 6 模型在 5-CV 下基于数据集不同比例的正负样本的性能（x 轴代表不同的比例）

案例分析

在本节中，我们进行了一个案例研究，以估计 PHIAF 预测未知新 PHI 的能力。我们首先使用 2021 年 1 月 1 日之前出现在 NCBI 中的所有噬菌体和宿主的已知相互作用来训练 PHIAF，以识别剩余噬菌体和宿主之间的所有相互作用（指 2021 年 1 月 1 日之后出现在 NCBI 中的噬菌体和宿主相互作用）。然后，我们根据预测分数对 PHI 进行排序，并搜索新发表的文献，以验证预测的 PHI 是否已被生物实验证实。我们在表 3 中列出了这些预测的 PHI（按预测分数排序）；其中四对最近发表的文献所验证。例如，Kokobel1（NCBI 登录号：NC_052979）中的分析结果[54]描述了可以侵染 Providencia rettgeri（NCBI 登录号：NZ_CP029736）和 Providencia stuartii（NCBI 登录号：NC_017731）的某些菌株。Zhan 等人[55]报道了五种感染 Ruegeria pomeroyi DSS-3 的噬菌体（NCBI 登录号：NC_003911），其中一种是 vB_RpoS-V16（NCBI 登录号：NC_052969）。本案例研究的结果表明，PHIAF 可以帮助识别新的 PHI 并缩小进一步生物实验的候选范围。

表 3. PHIAF 预测的噬菌体-宿主相互作用

噬菌体（登录号）	宿主（登录号）	证据
NC_052979	NZ_CP029736	[54]
NC_053009	NZ_CP029736	[54]
NC_052969	NC_003911	[55]
NC_052979	NC_017731	[54]
NC_053009	NC_017731	NA
NC_052979	NC_003911	NA
NC_053009	NC_003911	NA
NC_052969	NZ_CP029736	NA
NC_052969	NC_017731	NA

NA 代表该相互作用没有文献报道的证据。

5. 总结

抗生素的过度使用给细菌性疾病的治疗带来了一些严峻的挑战。作为治疗细菌性疾病最有希望的抗生素替代品之一，噬菌体疗法受到了广泛关注。确定 PHI 对于了解噬菌体是否可用于治疗细菌性疾病极为重要。在本研究中，我们提出了一种用于 PHI 预测的 PHIAF 方法，该方法利用 GAN 生成高质量的伪样本，融合来自 DNA 和蛋白质序列的特征以获得更好的性能，并使用注意力机制来提供预测模型的可解释性。通过 5-CV 与最先进的方法进行比较表明，PHIAF 实现了最好的 PHI 预测（就 AUC 和 AUPR 而言，性能平均提高了大约 13.64% 和 14.75%）。此外，消融研究说明了 PHIAF 的每个组件的贡献，数据增强模块对预测模型的贡献最大。此外，还进行了案例研究以证明我们方法的实用性。实验结果表明，PHIAF 是一种很有前途的 PHI 识别工具

尽管我们的模型具有良好的预测性能，但仍有一些限制需要解决。例如，我们设置的多个超参数的初始值和范围源自之前的研究，仅在有限的实验中粗略确定。未来，我们可以通过训练更多的预测模型来获得最优的超参数。此外，作为一种旨在预测 PHI 的网络架构，PHIAF 可用于处理生物信息学中其他基于序列的分类任务。

参考文献

[1] Chin CS, Sorenson J, Harris JB, et al. The origin of the Haitian cholera outbreak strain. N Engl J Med 2010; 364:33–42.

[2] Khan S, Imran A, Malik A, et al. Bacterial imbalance and gut pathologies: association and contribution of E. coli in inflammatory bowel disease. Crit Rev Clin Lab Sci 2019; 56:117.

[3] Khan S. Potential role of Escherichia coli DNA mismatch repair proteins in colon cancer. Crit Rev Oncol Hematol 2015; 96:475–82.

[4] Khan S, Zaidi S, Alouffi AS, et al. Computational proteomewide study for the prediction of Escherichia coli protein targeting in host cell organelles and their implication in development of colon cancer. ACS Omega 2020; 5(13): 7254–61.

[5] Li J, Zakariah M, Malik A, et al. Analysis of Salmonella typhimurium protein-targeting in the nucleus of host cells and the implications in colon cancer: an in-silico approach. Infect Drug Resist 2020; 13:2433–42

[6] Hassel B. Tetanus: pathophysiology, treatment, and the possibility of using botulinum toxin against tetanus-induced rigidity and spasms. Toxins (Basel) 2013; 5:73–83.

[7] Khan S, Zakariah M, Rolfo C, et al. Prediction of mycoplasma hominis proteins targeting in mitochondria and cytoplasm of host cells and their implication in prostate cancer etiology. Oncotarget 2017; 8:30830–43.

[8] Davies J, Davies D. Origins and evolution of antibiotic resistance. Microbiol Mol Biol Rev 2010; 74:417–33.

[9] Gorski A, Miedzybrodzki R, Wegrzyn G, et al. Phage therapy: current status and perspectives. Med Res Rev 2020; 40: 459–63.

[10] Kadri SS. Key takeaways from the U.S. CDC’s 2019 antibiotic resistance threats report for frontline providers. Crit Care Med 2020; 48:939–45.

[11] Cassini A, Högberg LD, Plachouras D, et al. Attributable deaths and disability-adjusted life-years caused by infections with antibiotic-resistant bacteria in the EU and the European Economic Area in 2015: a population-level modelling analysis. Lancet Infect Dis 2019; 19:56–66.

[12] Towse A, Hoyle CK, Goodall J, et al. Time for a change in how new antibiotics are reimbursed: development of an insurance framework for funding new antibiotics based on a policy of risk mitigation. Health Policy 2017; 121: 1025–30.

[13] Stokes JM, Yang K, Swanson K, et al. A deep learning approach to antibiotic discovery. Cell 2020; 180:688–702.e13.

[14] Pires DP, Costa AR, Pinto G, et al. Current challenges and future opportunities of phage therapy. FEMS Microbiol Rev 2020; 44:684–700.

[15] Edwards RA, McNair K, Faust K, et al. Computational approaches to predict bacteriophage-host relationships. FEMS Microbiol Rev 2016; 40:258–72.

[16] Villarroel J, Kleinheinz KA, Jurtz VI, et al. HostPhinder: a phage host prediction tool. Viruses 2016; 8:116.

[17] Liu D, Ma Y, Jiang X, et al. Predicting virus-host association by kernelized logistic matrix factorization and similarity network fusion. BMC Bioinformatics 2019; 20:594.

[18] Wang W, Ren J, Tang K, et al. A network-based integrated framework for predicting virus-prokaryote interactions. NAR Genom Bioinform 2020; 2: lqaa044.

[19] Ahlgren NA, Ren J, Lu YY, et al. Alignment-free d＊2 oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences. Nucleic Acids Res 2017; 45:39–53.

[20] Galiez C, Siebert M, Enault F, et al. WIsH: who is the host? predicting prokaryotic hosts from metagenomic phage contigs. Bioinformatics 2017; 33:3113–4.

[21] Zhang M, Yang L, Ren J, et al. Prediction of virus-host infectious association by supervised learning methods. BMC Bioinformatics 2017; 18:60.

[22] Lu C, Zhang Z, Cai Z, et al. Prokaryotic virus host predictor: a Gaussian model for host prediction of prokaryotic viruses in metagenomics. BMC Biol 2021; 19:5.

[23] Mock F, Viehweger A, Barth E, et al. VIDHOP, viral host prediction with deep learning. Bioinformatics 2021; 37:318–25.

[24] Hauser R, Blasche S, Dokland T, et al. Bacteriophage proteinprotein interactions. Adv Virus Res 2012; 83:219–98.

[25] Alguwaizani S, Park B, Zhou X, et al. Predicting interactions between virus and host proteins using repeat patterns and composition of amino acids. J Healthc Eng 2018; 2018:1391265.

[26] Young F, Rogers S, Robertson DL. Predicting host taxonomic information from viral genomes: a comparison of feature representations. PLoS Comput Biol 2020; 16:e1007894.

[27] Boeckaerts D, Stock M, Criel B, et al. Predicting bacteriophage hosts based on sequences of annotated receptor-binding proteins. Sci Rep 2021; 11:1467.

[28] Leite DMC, Brochet X, Resch G, et al. Computational prediction of inter-species relationships through omics data analysis and machine learning. BMC Bioinformatics 2018; 19: 420.

[29] Leite DMC, Lopez JF, Brochet X, et al. Exploration of multiclass and one-class learning methods for prediction of phagebacteria interaction at strain level. In: International Conference on Bioinformatics and Biomedicine (BIBM). Madrid, Spain: IEEE, 2018, 1818–25.

[30] Li M, Wang Y, Li F, et al. A deep learning-based method for identification of bacteriophage-host interaction. IEEE/ACM Trans Comput Biol Bioinform. 10.1109/TCBB.2020.3017386.

[31] Gao NL, Zhang C, Zhang Z, et al. MVP: a microbe-phage interaction database. Nucleic Acids Res 2018; 46:D700–7.

[32] Russell DA, Hatfull GF. PhagesDB: the actinobacteriophage database. Bioinformatics 2017; 33:784–6.

[33] Mihara T, Nishimura Y, Shimizu Y, et al. Linking virus genomes with host taxonomy. Viruses 2016; 8:66.

[34] Pruitt KD, Tatusova T, Brown GR, et al. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. Nucleic Acids Res 2012; 40: D130–5.

[35] Deng Y, Xu X, Qiu Y, et al. A multimodal deep learning framework for predicting drug-drug interaction events. Bioinformatics 2020; 36:4316–22.

[36] Li J, Pu Y, Tang J, et al. DeepATT: a hybrid category attention neural network for identifying functional effects of DNA sequences. Brief Bioinform 2021; 22: bbaa159.

[37] Goodfellow IJ, Pouget-Abadie J, Mirza M, et al. Generative Adversarial Nets. In: Conference on Neural Information Processing Systems (NeurIPS). Montreal, Quebec,

Canada: Curran Associates, Inc., 2014, 2672–2680.

[38] Zhu JY, Park T, Isola P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: IEEE International Conference on Computer Vision (ICCV).Venice, Italy: IEEE, 2017, 2223–32.

[39] Wan C, Jones DT. Protein function prediction is improved by creating synthetic feature samples with generative adversarial networks. Nat Mach Intell 2020; 2:540–50.

[40] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: Conference on Neural Information Processing Systems (NeurIPS). Long Beach, CA, USA: Curran Associates, Inc., 2017, 5998–6008.

[41] Huang Y, Niu B, Gao Y, et al. CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics 2010; 26:680–2.

[42] Chen Z, Zhao P, Li F, et al. iLearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data. Brief Bioinform 2020; 21:1047–57.

[43] Yi X, Walia E, Babyn P. Generative adversarial network in medical imaging: a review. Med Image Anal 2019; 58:101552.

[44] Gao X, Deng F, Yue X. Data augmentation in fault diagnosis based on the Wasserstein generative adversarial network with gradient penalty. Neurocomputing 2020; 396:487–94.

[45] Lopez-Paz D, Oquab M. Revisiting classifier two-sample tests. In: The International Conference on Learning Representations (ICLR). Toulon, France: OpenReview.net, 2017.

[46] Xu Y, Zhang Z, You L, et al. scIGANs: single-cell RNA-seq imputation using generative adversarial networks. Nucleic Acids Res 2020; 48:e85.

[47] Wei L, Ye X, Xue Y, et al. ATSE: a peptide toxicity predictor by exploiting structural and evolutionary information based on graph neural network and attention mechanism. Brief Bioinform 2021;bbab041. 10.1093/bib/bbab041.

[48] Xu W, Zhu L, Huang DS. DCDE: an efficient deep convolutional divergence encoding method for human promoter recognition. IEEE Trans Nanobioscience 2019; 18:136–45.

[49] Zhang Q, Zhu L, Huang DS. High-order convolutional neural network architecture for predicting DNA-protein binding sites. IEEE/ACM Trans Comput Biol Bioinform 2019; 16:1184–92.

[50] ChuaiG,MaH,YanJ,et al. DeepCRISPR: optimized CRISPR guide RNA design by deep learning. Genome Biol 2018; 19:80.

[51] Zhang S, Zhao L, Zheng CH, et al. A feature-based approach to predict hot spots in protein-DNA binding interfaces. Brief Bioinform 2020; 21:1038–46.

[52] Tang X, Zhang T, Cheng N, et al. usDSM: a novel method for deleterious synonymous mutation prediction using undersampling scheme. Brief Bioinform 2021;bbab123. 10.1093/bib/bbab123.

[53] LVD M, Geoffrey H. Visualizing data using t-SNE. JMachLearn Res 2008; 9:2579–605.

[54] Rakov C, Ben Porat S, Alkalay-Oren S, et al. Targeting biofilm of MDR Providencia stuartii by phages using a catheter model. Antibiotics 2021; 10:375.

[55] Zhan Y, Huang S, Chen F. Genome sequences of five bacteriophages infecting the marine Roseobacter bacterium Ruegeria pomeroyi DSS-3. Microbiol Resour Announc 2018; 7:e00959–18.

参考文献原文

Briefings in Bioinformatics, 00(0), 2021, 1–10

https://doi.org/10.1093/bib/bbab348
Problem Solving Protocol

PHIAF: prediction of phage-host interactions with GAN-based data augmentation and sequence-based feature fusion

Menglu Li and Wen Zhang

Corresponding author: Wen Zhang, College of Informatics, Huazhong Agricultural University, Wuhan 430070, China. E-mail: zhangwen@mail.hzau.edu.cn

Abstract

Phage therapy has become one of the most promising alternatives to antibiotics in the treatment of bacterial diseases, and identifying phage-host interactions (PHIs) helps to understand the possible mechanism through which a phage infects bacteria to guide the development of phage therapy. Compared with wet experiments, computational methods of identifying PHIs can reduce costs and save time and are more effective and economic. In this paper, we propose a PHI prediction method with a generative adversarial network (GAN)-based data augmentation and sequence-based feature fusion (PHIAF). First, PHIAF applies a GAN-based data augmentation module, which generates pseudo PHIs to alleviate the data scarcity. Second, PHIAF fuses the features originated from DNA and protein sequences for better performance. Third, PHIAF utilizes an attention mechanism to consider different contributions of DNA/protein sequence-derived features, which also provides interpretability of the prediction model. In computational experiments, PHIAF outperforms other state-of-the-art PHI prediction methods when evaluated via 5-fold cross-validation (AUC and AUPR are 0.88 and 0.86, respectively). An ablation study shows that data augmentation, feature fusion and an attention mechanism are all beneficial to improve the prediction performance of PHIAF. Additionally, four new PHIs with the highest PHIAF score in the case study were verified by recent literature. In conclusion, PHIAF is a promising tool to accelerate the exploration of phage therapy.

Key words: phage-host interactions; generative adversarial network; feature fusion; attention mechanism.

Introduction

Available reports have demonstrated the possible involvement of bacterial infection in the growth and development of various types of diseases, including cholera [1], inflammatory bowel disease [2], colon cancer [3–5], tetanus [6] and different types of cancer [7]. Researchers discovered antibiotics in 1928 and have since used them in clinical practice to treat serious bacterial diseases and save countless lives [8]. Unfortunately, due to the overuse of antibiotics, bacteria have developed a few resistance mechanisms [9]. In 2019, the US Centers for Disease Control and Prevention reported that approximately 2.8 million cases of antibiotic-resistant infections occur each year in the United States, resulting in more than 35 000 deaths [10]; in Europe, about 33 000 people die from antibiotic-resistant infections each year [11]. Thus, it is urgent to develop new antibiotics or alternative therapies to avoid further deterioration of antibiotic-resistant infections. However, many pharmaceutical companies no longer develop new antibiotics because of their high production costs, unsatisfactory expected benefits and long research and development time [12, 13]. Therefore, researchers want to look for alternative therapies to reduce antibiotic-resistant infections and treat bacterial diseases.

Bacteriophages can not only destroy specific bacteria hosts but also replicate exponentially, and these characteristics make bacteriophages one of the most promising therapies in the treatment of bacterial diseases and address antibiotic-resistant infections [14]. Determining phage-host interactions (PHIs) helps to

Menglu Li is a PhD candidate in the College of Informatics at Huazhong Agricultural University.
Wen Zhang is a professor in the College of Informatics at Huazhong Agricultural University.
Submitted: 2 January 2021; **Received (in revised form):** 5 July 2021

understand whether phages can be used to treat bacterial diseases. However, experimental verification of PHIs requires considerable time, manpower and money. Therefore, researchers have attempted to develop computational PHI prediction methods to screen out target phages for treating bacterial diseases and to guide the *in vivo* validation, thereby greatly reducing the required time and costs [15].

Molecular and ecological coevolutionary processes shape phage and bacterial genomes and leave signals in their genomic sequences that allow researchers to predict PHIs [15], so various PHI computational methods based on phage and host genomic sequences have been developed [16–18]. For example, Ahlgren *et al.* [19] proposed VirHostMatcher (VHM) based on DNA sequences to predict PHIs by calculating the distance between the oligonucleotide frequency patterns of phages and hosts. However, the running time of VHM hindered its development on large datasets, so Galiez *et al.* [20] proposed WIsH to reduce the running time by constructing a Markov model to predict the prokaryotic host of bacteriophages. Compared with that of VHM, the running time of WIsH was reduced by a factor of several hundred. In addition to VHM and WIsH, which predict PHIs by calculating the similarity between phages and hosts, researchers have used various machine learning classifiers, including logistic regression (LR), support vector machine (SVM), random forest (RF) and naive Bayesian (NB), to predict PHIs [21]. Further, PHP [22] and VIDHOP [23] were developed to enhance the PHI prediction performance. PHP trained a Gaussian model by calculating the differences in k-mer frequencies between viral and host genomic sequences, and VIDHOP used deep neural networks to predict phages related to three different viruses (influenza A virus, rabies lyssavirus and rotavirus A).

Some studies have shown that proteins play a fundamental role in the biological processes of phages and hosts [24, 25]; thus, researchers have proposed PHI prediction methods based on protein sequences [26, 27]. For example, Leite *et al.* [28, 29] utilized the primary structure sequences from phage and host proteins and classic classifiers, including RF, SVM, LR, k-nearest neighbor (KNN), multi-layer perceptron (MLP) and NB, to predict PHIs. On the basis of the above method, Li *et al.* [30] used a convolutional neural network (CNN) to improve the performance of PHI prediction.

Although existing methods achieve good performance in PHI prediction, some challenges remain. First, there are thousands of experimentally verified PHIs in databases [31–34], but only a few hundred non-redundant PHIs are available and can be used to build predictive models. This limitation hinders the development of predictive models with high performance. Second, most existing methods use either the DNA sequences or protein sequences of phages and hosts to construct predictive models but rarely combine two types of sequences. Third, although a variety of features and machine learning techniques have been used to build prediction models, these models often lack sufficient interpretability, which obstructs elaborating the mechanism of PHIs.

In recent years, deep learning technology has received extensive attention in the field of bioinformatics, and researchers have applied such techniques to handle different tasks [35, 36]. The generative adversarial network (GAN), as a branch of deep learning technology, was originally used for image processing [37, 38] and later showed excellent performance in data augmentation. For instance, Wan *et al.* [39] successfully used a GAN to generate biophysical features based on protein sequences. Meanwhile, researchers developed an attention mechanism [40] for deep learning to increase the interpretability of predictive models and to improve prediction performance. The development of these deep learning technologies motivates us to further enhance and improve PHI prediction.

In the current study, we propose a novel PHI prediction method, abbreviated as PHIAF, based on GAN data augmentation and sequence-based feature fusion to solve the various challenges of PHI prediction. First, PHIAF uses GAN to construct a data augmentation module, which generates high-quality pseudo samples to overcome the bottleneck of the PHI data scarcity. Second, PHIAF fuses different features encoded by the DNA and protein sequences of phages and hosts to enhance the prediction performance. Third, PHIAF utilizes CNN to build a PHI prediction module and incorporates an attention mechanism into CNN to provide interpretability of the prediction model. Experimental results show that PHIAF is superior to the state-of-the-art methods of PHI prediction. The ablation study and discussion indicate that the pseudo samples generated by the data augmentation module, the fusion of DNA and protein sequence-derived features and the attention mechanism in CNN effectively improve the performance of PHIAF.

Algorithm 1 : Data processing to remove redundant phages.

Require: The set of hosts, $H = \{h_1, h_2, ..., h_n\}$, n is the number of hosts; the set of phages corresponding to different hosts, $P = \{P_1, P_2, ..., P_n\}$, where $P_i = [p_1, p_2, ..., p_m]$ is the set of phages of host h_i, $i \in [1, n]$; similarity matrix of different phages, $S = \{s_{p_1,p_2}, s_{p_1,p_3}, ..., s_{p_{m-1},p_m}\}$, m is the number of phages;
Ensure: The interactions between non-redundant phages and hosts, I;

```
function main($P, H, S$)
  $I \leftarrow []$
  for $k \leftarrow 1$ to $n$ do
    $I_k \leftarrow []$
    $I_k \leftarrow del\_redundant(P_k)$
    $I = I + I_k$
  end for
  return $I$
end function
function del_redundant($P_k$)
  $P_{del} \leftarrow [], P_{new} \leftarrow []$
  $I_k \leftarrow I_k + (P_k[0], h_k)$
  $P_{del} \leftarrow P_{del} + P_k[0]$
  for $j \leftarrow 1$ to $m$ do
    if $s_{P_k[0],P_k[j]} > 0.90$ then
      $P_{del} \leftarrow P_{del} + P_k[j]$
    end if
  end for
  $P_{new} \leftarrow P_k - P_{del}$
  if $length(P_{new}) > 1$ then
    $del\_redundant(P_{new})$
  end if
  if $length(P_{new}) = 1$ then
    $I_k \leftarrow I_k + (P_{new}[0], h_k)$
  end if
  return $I_k$
end function
```

Materials and methods

Dataset

We download data (including phages, hosts and their interactions) from four widely used databases on March 2021, including

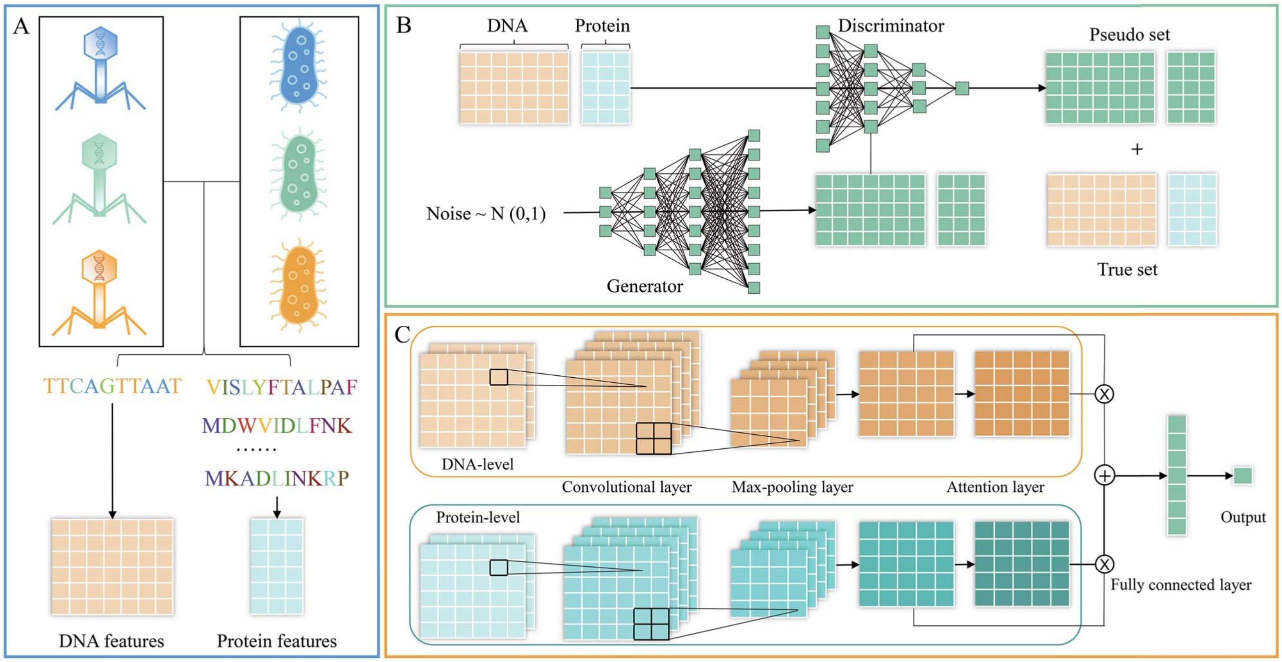

Figure 1. An overview of PHIAF. (**A**) Feature extraction module. (**B**) Data augmentation module. (**C**) PHI prediction module.

MVP [31], PhagesDB [32], VHDB [33] and NCBI [34], and merge these data to construct a dataset with more PHIs for our study. The data in these four databases are subjected to the following processing. First, we delete the PHIs that are not published in the literature or not included in NCBI records to ensure that the PHIs are reliable. Second, we remove data incorrectly marked as phages/hosts based on the definition of phages (phages are viruses that infect and replicate within bacteria and archaea). The removed data include phages that do not belong to viruses and hosts that do not belong to bacteria or archaea. Third, we extract whole-genome sequences and coding protein sequences from the NCBI database for remaining phages and hosts.

After the above processes, we combine the remaining phages and hosts of the four databases and remove duplicates, resulting in a total of 5399 interactions between 5331 phages and 235 hosts. The number of phages is much larger than the number of hosts, and one host may interact with multiple phages. We use Algorithm 1 to remove redundant phages with high similarity for each host (the comparison between different similarity measures and the impact of phage $P_k[0]$ on prediction performance are provided in Sections 1 and 2 of Supplementary Materials, respectively). We set 0.90 as a high similarity threshold, which is the same as the default threshold of the CD-HIT tool [41]. After redundancy reduction, we obtain a benchmark dataset with 312 interactions between 304 phages and 235 hosts, which can be used to better evaluate the performance of the prediction models. In this dataset, we set 312 known PHIs as positive samples and select negative samples from all unknown PHIs while ensuring that the numbers of positive and negative samples are equal.

PHIAF

PHIAF consists of three main modules: feature extraction, data augmentation and PHI prediction. A schematic diagram of PHIAF is shown in Figure 1. First, DNA and protein sequences of phages and hosts are encoded into features (Figure 1A). Second, a GAN-based data augmentation module is used to generate pseudo PHIs (Figure 1B). Finally, a PHI prediction module is built under a CNN framework with attention to utilize the features derived from DNA and protein sequences after reshaping into appropriate forms to predict PHIs (Figure 1C).

Feature extraction module

Some studies have shown that DNA and protein sequences play a fundamental role in the biological evolution of bacteriophages and hosts. To obtain more comprehensive and effective information, we extract features from DNA and protein sequences of phages and hosts, as summarized in Table 1; detailed information of these features is provided in Section 3 of Supplementary Materials.

Since DNA sequences of phages and hosts have different lengths (analysis of the distribution of the DNA and protein sequence lengths of phages and hosts is provided in Section 4 of Supplementary Materials), we consider several DNA sequence-derived features that are unrelated to sequence length, including Kmer, reverse compliment Kmer (RCKmer), nucleic acid composition (NAC), di-nucleotide composition (DNC), tri-nucleotide composition (TNC), the composition of k-spaced nucleic acid pairs (CKSNAP) and electron-ion interaction pseudopotentials of trinucleotide (PseEIIP). We calculate these features with the software iLearn [42] and obtain a 340-dimensional feature vector for each phage/host DNA sequence.

We follow previous studies [28–30] to extract widely used protein sequence-derived features, including amino acid composition (AAC), the abundance of chemical elements composing a protein (AC) and the molecular weight of a protein (MW). Since each phage/host has multiple protein sequences, we consider six operators (mean, maximum, minimum, standard deviation, variance and median) to integrate features from protein sequences and obtain a 162-dimensional feature vector for each phage/host.

Table 1. Description of the features that originated from DNA and protein sequences

Levels	Features	Descriptions
DNA	Kmer	the occurrence frequencies of k neighboring nucleic acids ($k = 3$)
	RCKmer	a variant of Kmer, which removes the reverse complement Kmers
	NAC	the frequency of each nucleic acid type (A, C, G, T) in a nucleotide sequence
	DNC	the frequency of two nucleic acid types in a nucleotide sequence
	TNC	the frequency of three nucleic acid types in a nucleotide sequence
	CKSNAP	the frequency of nucleic acid pairs separated by any p nucleic acid ($p = 5$)
	PseEIIP	mean EIIP values (A: 0.1260, C: 0.1340, G: 0.0806, T: 0.1335) of trinucleotides in each sequence
Protein	AAC	the frequency of each amino acid in a protein sequence
	AC	abundance of selected chemical elements composing a protein
	MW	molecular weight of a protein sequence

Data augmentation module

The GAN [37] is a new type of generative model that aims to generate high-quality pseudo samples by precisely learning the underlying distribution of real samples. This model has received considerable attention and achieved outstanding performance in many fields [43, 44]. In this study, we use a GAN [39] to address the data scarcity of PHIs in our dataset for model training.

We first set the real positive samples as $I = \{(p_1, h_1), (p_2, h_2), \ldots, (p_m, h_n)\}$, $V = \{V_{p_1,h_1}, V_{p_2,h_2}, \ldots, V_{p_m,h_n}\}$ to represent the feature vectors of these samples, which are composed of the DNA and protein features of phages and hosts encoded above. m and n represent the numbers of phages and hosts, respectively. These feature vectors of positive samples (V) are input into the GAN to generate high-quality pseudo feature vectors of samples, where the GAN is composed of two neural networks (generator and discriminator) that 'fight' against each other to learn the distribution of real samples. One network (generator) tries to generate pseudo samples via five fully connected layers (formula 1). Another network (discriminator), which is composed of four fully connected layers (formula 2), tries to distinguish whether a given sample is real. Each network's task gets better and better until equilibrium is reached, where the generator cannot make better samples, and the discriminator cannot separate real and pseudo samples.

$$O_{ge} = FC_t\left(FC_r\left(FC_r\left(FC_r\left(FC_r(V)\right)\right)\right)\right) \tag{1}$$

$$O_{di} = FC_l\left(FC_l\left(FC_l\left(FC_l(V', V)\right)\right)\right) \tag{2}$$

where V' is the feature vectors of pseudo samples, O_{ge} is the generator output, O_{di} is the discriminator output, FC_l (or FC_r) represents MLP with the LeakyReLU (or ReLU) activation function and FC_t means MLP with the Tanh activation function.

We conduct the classifier two-sample tests (C2ST) method [45] using the KNN ($k = 1$) and leave-one-out cross-validation (LOOCV) to distinguish the real and pseudo samples. The C2ST method involves accepting or rejecting a null hypothesis of P being equal to Q (where P and Q are the distributions of two equal-sized sets of samples). If the null hypothesis is accepted, the classification accuracy for predicting the binary labels of held-out samples will be near the level of chance (that is 0.50). Therefore, the real and pseudo samples are indistinguishable when the KNN classification accuracy is the closest to 0.50 under LOOCV. Finally, we choose the indistinguishable pseudo samples to amplify our dataset, represented by $I' = \{(p'_1, h'_1), (p'_2, h'_2), \ldots, (p'_m, h'_n)\}$.

The above processing is used to amplify positive samples in our dataset. Since all unknown PHIs are candidate negative samples and are much more than positive samples, we randomly select negative samples from this candidate set to ensure that the numbers of positive and negative samples are equal. Finally, we combine the real positive samples, selected negative samples and pseudo positive samples to construct an augmented dataset, which is used to train the prediction model.

PHI prediction module

Based on the augmented dataset, we use the DNA and protein sequence-derived features of phages and hosts to build the prediction model.

As described in the previous section, we have two types of feature vectors for each phage/host, which are extracted from DNA and protein sequences, respectively. Since sequence-derived features usually have complicated short-/long-range dependency, we reshape the DNA/protein feature vectors into 'images' to capture the complicated relationship between their dimensions [46]. We first adopt the Min-Max normalization to normalize values in feature vectors to the range from 0 to 1. Let N denote the dimension of the feature vector of a DNA/protein; then, we reshape the feature vector into an $n \times n$ 'image' by placing values by row, where n satisfies the condition: $(n-1) \times (n-1) < N$ and $N \le n \times n$. When $N < n \times n$, we implement padding by adding zeros to the remaining $n \times n - N$ entries. Finally, we obtain the DNA and protein sequence-derived feature matrices for each phage (or host), denoted as $\mathbf{M}_d^P$ and $\mathbf{M}_p^P$ ($\mathbf{M}_d^H$ and $\mathbf{M}_p^H$), respectively.

Then, we construct a bi-level architecture (DNA- and protein-level) to extract deeper features from the DNA and protein feature matrices. For the DNA-level, we stack the DNA-derived feature matrix of phages and hosts across channels to form a combined matrix and then input this combined matrix into a two-layer CNN to produce a feature map O_d with more meaningful information. The CNN includes a convolutional layer and a max-pooling layer.

$$O_d = MaxPool\left(Conv2D([\mathbf{M}_d^P, \mathbf{M}_d^H])\right) \tag{3}$$

where *Conv2D* and *MaxPool* represent the convolutional layer and max-pooling layer, respectively, and $[\cdot, \cdot]$ represents stacking across channels. Similarly, the protein-derived feature matrices of phages and hosts are combined across channels and fed to the same CNN to produce a feature map O_p at the protein-level.

$$O_p = MaxPool\left(Conv2D([\mathbf{M}_p^P, \mathbf{M}_p^H])\right) \tag{4}$$

The attention mechanism aims to imitate the action of the human brain to selectively concentrate on a few important parts while ignoring others in machine learning tasks [47]. The DNA- and protein-level features we have may make different contributions to the PHI prediction. Thus, we introduce an attention mechanism into our model, add an attention layer to capture important features and then integrate these features.

In the attention layer, the DNA- and protein-level feature maps (O_d and O_p) are input into a fully connected layer to calculate weight vectors (α_d and α_p), respectively; then, the feature maps are multiplied by the corresponding weight vectors. Finally, the output of the attention layer is calculated as follows:

$$O_{att} = O_d \otimes \alpha_d + O_p \otimes \alpha_p \tag{5}$$

where ⊗ is the element-wise multiplication. At last, the output of attention layer is feed into a two-layer MLP to yield the probability of samples being a PHI.

$$Pred = FC_s\left(FC_r\left(O_{att}\right)\right) \tag{6}$$

where FC_s means the MLP with Sigmoid activation function.

PHIAF optimization

There are several important hyper-parameters in the data augmentation and PHI prediction modules of PHIAF. In the data augmentation module, we set the numbers of neurons in the different fully connected layers of the generator as 128, 256, 512, 1024 and 1004 and those in the discriminator as 512, 256, 128 and 1, respectively. In the PHI prediction module, we set 32 filters and the size of a filter as 3 × 3 in the convolutional layer and set the size of the max-pooling layers of the DNA- and protein-level as 3 × 3 and 2 × 2, respectively. The difference in the size of these two max-pooling layers ensures that the dimensions of O_d and O_p are equal. We also consider the learning rate (optional values are 0.1, 0.01, 0.001 and 0.0001), batch size (optional values are 8, 16, 32, 64 and 128), loss rate of the dropout layer (optional values include 0.25, 0.5 and 0.75) and the number of neurons in the fully connected layer (optional values include 16, 32, 64, 128, 256 and 512) to determine the optimal parameters. The model performance given different parameters is provided in Section 5 of Supplementary Materials.

In addition, we utilize dropout and batch normalization layers in the PHI prediction module to prevent overfitting and improve generalizability [48, 49]. The data augmentation module and PHI prediction module are trained independently. Additionally, we adopt the Wasserstein loss add gradient penalty as the loss function for the data augmentation module and use the binary cross-entropy loss function for the PHI prediction module. All of the above loss functions are optimized by the Adam optimizer [50].

Results

Performance assessment

In this study, we adopt 5-fold cross-validation (5-CV) to evaluate the performance of PHIAF. All samples in our dataset are randomly divided into five equal-sized subsets. The cross-validation process is repeated five times, and every subset is used as the test set in turn while the remaining four subsets are used as the training set. The final 5-CV results are generated by averaging the five test set results. We employ several commonly used evaluation measures [51, 52] to assess the performance of PHIAF and state-of-the-art methods, including specificity (Spe), sensitivity (Sen), accuracy (Acc), F1-score (F1), area under the receiver-operating characteristic curve (AUC) and area under the precision-recall curve (AUPR). The measures are calculated as follows:

$$Spe = \frac{TN}{TN + FP} \tag{7}$$

$$Sen = \frac{TP}{TP + FN} \tag{8}$$

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

$$F1 = \frac{2 * TP}{2 * TP + FP + FN} \tag{10}$$

where *TP* (*TN*) denotes the number of positive (negative) samples correctly classified in the prediction and *FP* (*FN*) represents the number of incorrectly identified negative (positive) samples.

Comparison with state-of-the-art methods

To demonstrate the effectiveness of PHIAF, we compare it with the following state-of-the-art methods, including DNA sequence-based algorithms and protein sequence-based methods.

- VHM [19] is a DNA sequence-based method that computes the distance between the oligonucleotide frequency patterns of phages and hosts and obtains the possibility of PHIs based on this distance.
- WIsH [20] is a method based on DNA sequences that trains a Markov model for each host and calculates the probability of interactions for all phages.
- PHP [22] is a DNA sequence-based method that constructs a Gaussian model to predict PHIs using k-mer frequencies between virus and host genomic sequences.
- RF, SVM, KNN and MLP are widely used machine learning classifiers; Leite *et al.* [28] utilizes protein sequence-derived features and these classifiers to predict PHIs.
- PredPHI [30] is a protein sequence-based method that predicts PHIs under a CNN framework.

We first compare the PHIAF with these methods using 5-CV. As shown in Figure 2, PHIAF outperforms all comparison methods in terms of AUC and AUPR. In general, the DNA sequence-based methods (VHM, WIsH and PHP) produce better results than the protein sequence-based methods (RF, SVM, KNN, MLP and PredPHI), indicating that the information from the DNA sequences may play a more important role in PHIs. Our PHIAF model, which fuses information originating from DNA and protein sequences, is superior to models based on either DNA sequences or protein sequences, achieving 13.63% and 14.75% improvement, on average, in terms of the AUC and AUPR.

The predictive capability of models for unseen data is also important. Thus, we randomly select one-third of the phages and hosts in our dataset and use them and their interactions as a test set. The remaining phages and hosts and their interactions are used to train the prediction models. Under this experimental setting, phages or hosts in the test set are not included in

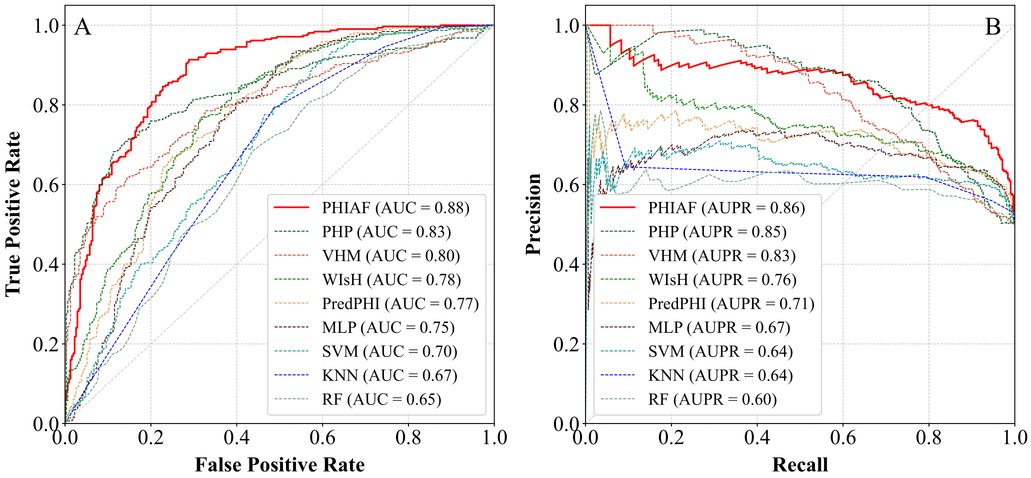

Figure 2. The performance of PHIAF and state-of-the-art methods using 5-fold cross-validation.

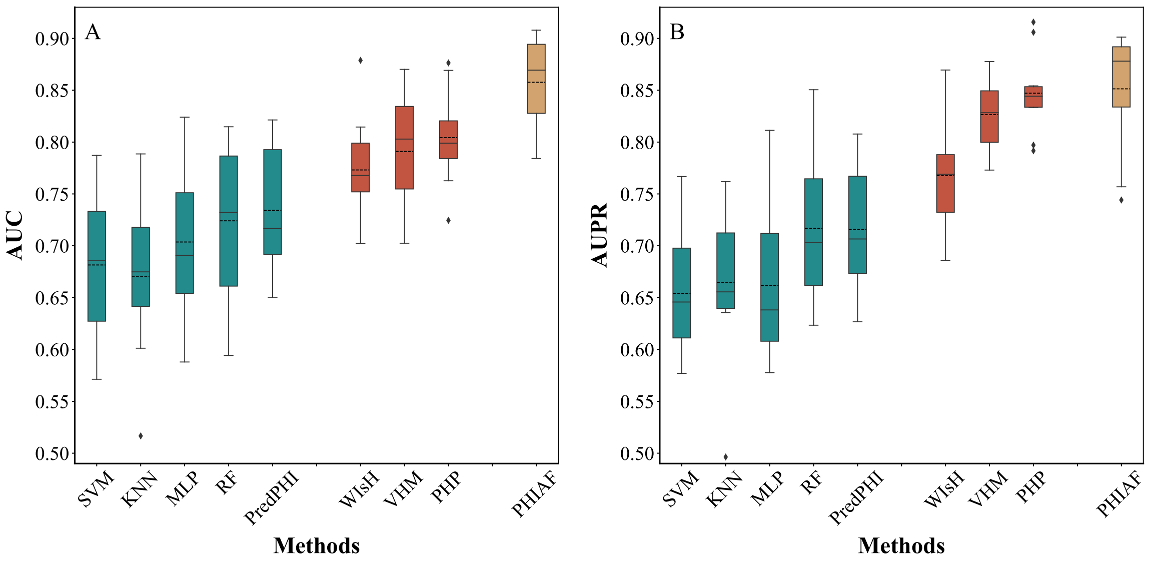

Figure 3. The performance of PHIAF and state-of-the-art methods for the unseen data. The solid line is the median of results, and the dashed line represents the mean.

the training set, and they are taken as unseen data. We train the prediction model based on the training set and then apply the trained model to the test set to evaluate the performance of the prediction model on unseen data. To avoid evaluation bias, we repeat the training and testing 10 times and adopt the average/median performance. As shown in Figure 3, PHIAF outperforms the other methods, generates higher average AUC (0.86) and AUPR scores (0.85) as well as reasonable standard deviation, and the medians of the AUC and AUPR scores are 0.88 and 0.87. Similar to the 5-CV results, DNA sequence-based methods produce better results than protein sequence-based methods. Moreover, a comparison between the results of the 5-CV and testing on unseen data indicates that our method achieves very similar performance in both cases (the AUC and AUPR differ by 2% and 1%), and these results demonstrate that our proposed method is robust and can perform well on unseen data, indicating that PHIAF is a promising tool for identifying PHIs from sequence data.

Ablation study

The above comparison illustrates the effectiveness of PHIAF, and the success of PHIAF is a result of its design: a GAN-based data augmentation module that generates high-quality pseudo samples and a PHI prediction module that fuses DNA and protein sequence-derived features with an attention mechanism. Here, we conduct an ablation study to elaborate the contribution of these components. We consider the following variants of PHIAF:

- PHIAF-D is a variant that does not use DNA-level features.
- PHIAF-P is a variant that does not use protein-level features.
- PHIAF-A is a variant that does not use the attention layer.
- PHIAF-G is a variant that does not use pseudo samples.

Table 2 shows the results of PHIAF and its four variants under 5-CV. The performance of PHIAF decreases when any component is removed, which means that all the components are critical

Table 2. The performance of PHIAF and different variants using 5-fold cross-validation

	AUPR	AUC	F1	Acc	Sen	Spe
PHIAF-D	0.80	0.85	0.77	0.77	0.79	0.75
PHIAF-P	0.84	0.86	0.79	0.79	0.80	0.79
PHIAF-A	0.82	0.86	0.64	0.73	0.60	**0.88**
PHIAF-G	0.79	0.82	0.73	0.73	0.75	0.71
PHIAF	**0.86**	**0.88**	**0.81**	**0.81**	**0.83**	0.78

Note: The highest value in each column is bold.

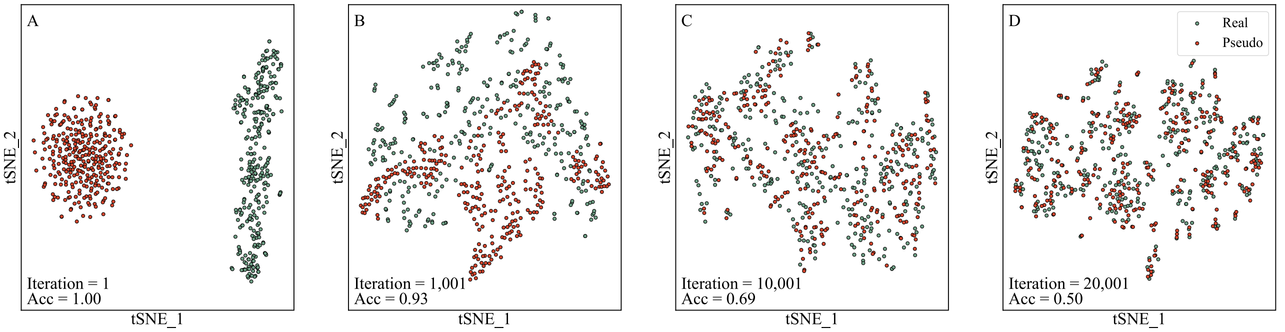

Figure 4. The t-SNE-transformed two-dimensional visualization of real and pseudo samples during different training iterations of GAN (A, 1 iteration; B, 1001 iterations; C, 10 001 iterations; D, 20 001 iterations).

for PHIAF. PHIAF-G suffers the greatest performance decrease, with the AUC and AUPR decreased by 6% and 7%, respectively, followed by PHIAF-D (the AUC and AUPR decreased by 3% and 6%, respectively), and PHIAF-A (the AUC and AUPR decreased by 2% and 4%, respectively). Comparison between the results of PHIAF and PHIAF-G shows that the use of pseudo samples effectively enhances PHI prediction. In addition, the comparison between PHIAF-D and PHIAF-P indicates that the DNA sequence-derived features are more effective for PHI prediction than are the features originating from protein sequences. Removing the attention layer also leads to poorer performance, indicating that the differences in features must be taken into account.

Discussion

The ablation study shows that the main components of PHIAF make important contributions to PHI prediction. Further, we analyze PHIAF from three aspects.

To demonstrate that the real and pseudo samples are indistinguishable and that the pseudo samples can be used as training positive samples, we use t-distributed stochastic neighbor embedding (t-SNE) [53] to visualize the 2D distribution of real and pseudo positive samples at different training iterations of GAN (Figure 4). In the 1st iteration (Figure 4A), the real samples are distributed far from the pseudo samples, leading to a LOOCV accuracy of 1.00, which suggests that the generator has not learned the distribution of real samples. As shown in Figure 4B, the generator begins to capture the characteristics of real samples, and the discriminator cannot fully distinguish between real and pseudo samples after 1000 iterations (LOOCV accuracy of 0.93). Then, the distributions continue to converge gradually (Figure 4C, LOOCV accuracy reaching 0.69). After 20 000 iterations, the generator and discriminator reach equilibrium, and the real and pseudo samples follow similar distributions (Figure 4D, LOOCV accuracy of 0.50). Thus, high-quality pseudo positive samples are generated for data augment.

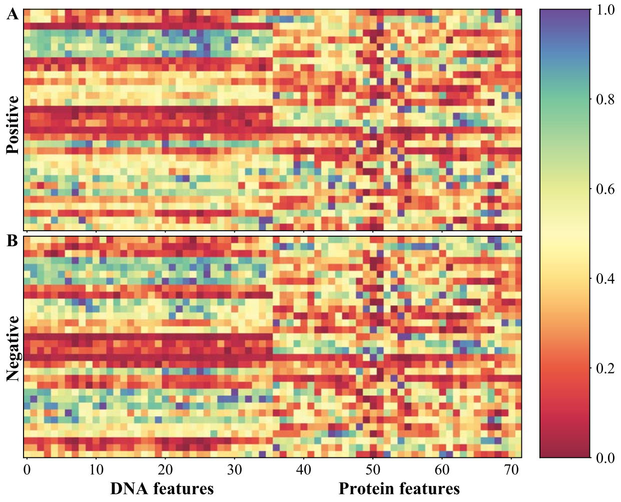

Figure 5. The average weights of positive and negative samples in our dataset in the attention layer (A is positive samples, B is negative samples).

Further, we analyze the weights in the attention layer assigned to different features to investigate the importance of features learned by the attention mechanism. Figure 5A and B shows that the attention weight distributions of positive and negative samples are similar, which indicates that some features play the same important role in positive and negative samples. Additionally, we compare the attention weights in the DNA- and protein-level, where protein-level features are generally assigned lower weights than are DNA-level features. These results confirm that the DNA-level features are more important than protein-level features for PHI prediction and that the attention layer in the CNN not only enhances the prediction performance but also effectively assigns weights based on the importance of different features.

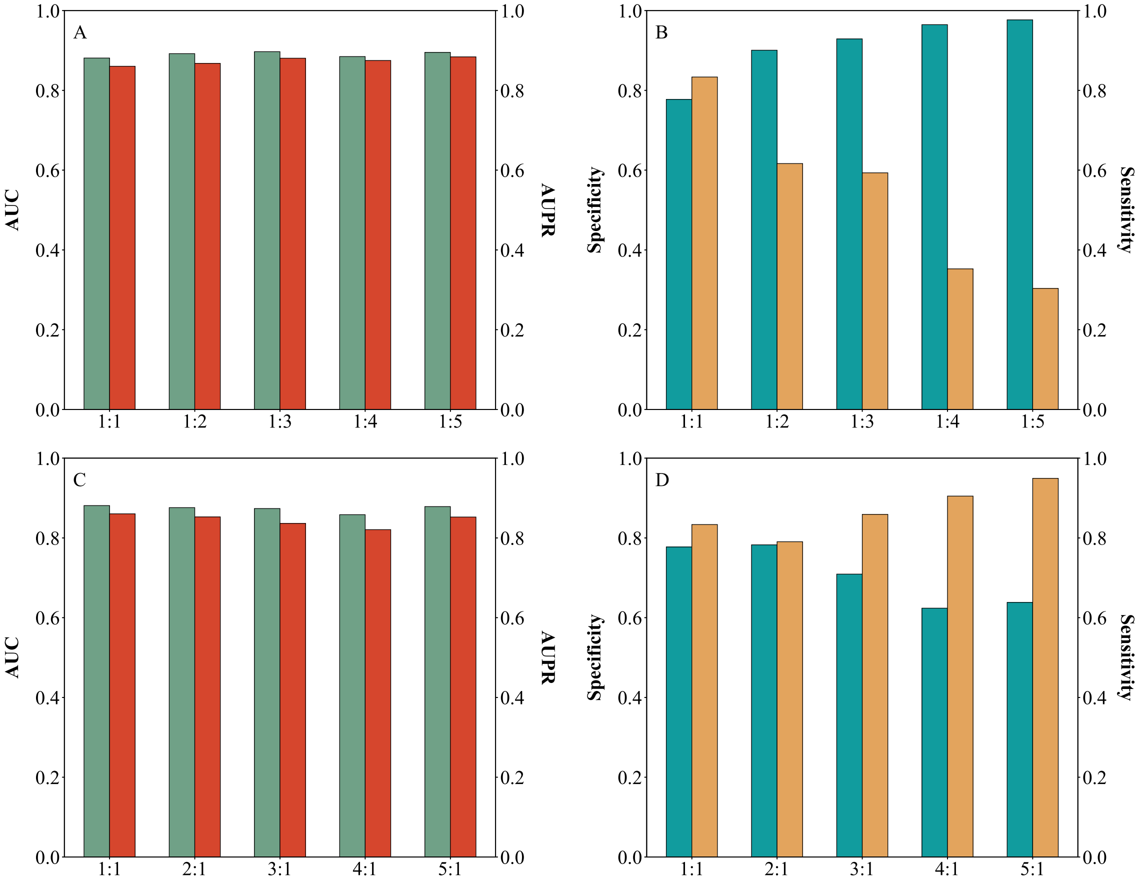

Figure 6. The performance of models based on datasets with different ratios of positive and negative samples under 5-CV (x-axis represents different ratios).

Finally, we build PHIAF models based on datasets with different ratios of positive and negative samples (1:1, 1:2, 1:3, 1:4, 1:5, 2:1, 3:1, 4:1 and 5:1) to analyze how the number of negative or pseudo positive samples influences the performance of PHIAF. As shown in Figures 6A and 6C, PHIAF produces similar AUC and AUPR scores (vary within 2%) on datasets with different data imbalance ratios. As the number of negative samples increases (Figure 6B), PHIAF achieves lower sensitivity (decreased from 0.83 to 0.30) and higher specificity (increased from 0.78 to 0.98). By contrast, PHIAF produces higher sensitivity (increased from 0.79 to 0.95) and lower specificity (decreased from 0.78 to 0.62) when the number of pseudo positive samples increases (Figure 6D). These results suggest that unbalanced datasets failed to significantly enhance the performance of PHI prediction, and samples are likely to be predicted as negatives/positives.

Case study

In this section, we conduct a case study to estimate the ability of PHIAF to predict unknown new PHIs. We first train the PHIAF using the known interactions of all phages and hosts that appeared in NCBI before 1 January 2021, to identify all pairs between the remaining phages and hosts (the remainder represents phages and hosts that appeared in NCBI after 1 January 2021). Then, we rank the PHIs according to the prediction score and search the newly published literature to verify whether the predicted PHI has been confirmed by biological experiments. We list these predicted PHIs in Table 3 (sorted by prediction scores); four of these pairs have been verified by recently published literature. For example, the analysis results in [54] described that Kokobel1 (NCBI accession number: NC_052979) can kill some strains of *Providencia rettgeri* (NCBI accession number: NZ_CP029736) and *Providencia stuartii* (NCBI accession number: NC_017731). Zhan *et al.* [55] reported five bacteriophages infecting *Ruegeria pomeroyi* DSS-3 (NCBI accession number: NC_003911), one of which is vB_RpoS-V16 (NCBI accession number: NC_052969). The results of this case study demonstrate that PHIAF can help to identify novel PHIs and narrow the scope of candidates for further biological experiments.

Table 3. The phage-host interactions predicted by PHIAF (these phages and hosts appeared in NCBI after 1 January 2021)

Phages (accession number)	Hosts (accession number)	Evidence
NC_052979	NZ_CP029736	[54]
NC_053009	NZ_CP029736	[54]
NC_052969	NC_003911	[55]
NC_052979	NC_017731	[54]
NC_053009	NC_017731	NA
NC_052979	NC_003911	NA
NC_053009	NC_003911	NA
NC_052969	NZ_CP029736	NA
NC_052969	NC_017731	NA

NA represents this interaction without the evidence reported in literature.

Conclusion

The overuse of antibiotics has led to several severe challenges in the treatment of bacterial diseases. As one of the most promising alternatives to antibiotics for the treatment of bacterial diseases, phage therapy has received widespread attention. Determining PHIs is extremely important for understanding whether phages can be used to treat bacterial diseases. In the present study, we propose a PHIAF method for PHI prediction that utilizes a GAN to generate high-quality pseudo samples, fuses the features derived from DNA and protein sequences for better performance and uses an attention mechanism to provide interpretability of the prediction model. A comparison with state-of-the-art methods via 5-CV demonstrates that PHIAF achieves the best PHI prediction (performance approximately 13.64% and 14.75% improvement, on average, in terms of AUC and AUPR). Moreover, an ablation study illustrates the contributions of each component of PHIAF, the data augmentation module makes the greatest contribution to the prediction model. Further, a case study is performed to prove the practicable capability of our method. The experimental results indicate that the PHIAF is a promising tool for identifying PHIs.

Despite the good prediction performance of our model, some limitations remain to be addressed. For example, the initial values and ranges of multiple hyper-parameters we set are originated from previous studies and only roughly determined in limited experiments. In the future, we are able to obtain optimal hyper-parameters by training more prediction models. In addition, as a network architecture designed to predict PHIs, PHIAF can be applied to handle other sequence-based classification tasks in bioinformatics.

Key Points

- A generative adversarial network-based data augmentation module is developed to generate high-quality pseudo samples to alleviate the data scarcity problem of phage-host interactions.
- DNA and protein sequence-derived features are combined to effectively improve the phage-host interaction prediction performance.
- The different contributions of DNA and protein features are taken into account through an attention layer, and an attention mechanism provides interpretability of the prediction model.

Supplementary Data

Supplementary data are available online at *Briefings in Bioinformatics*.

Data Availability

The data set and source code can be freely downloaded from https://github.com/mengluli-web/PHIAF or https://github.com/BioMedicalBigDataMiningLab/PHIAF.

Funding

National Natural Science Foundation of China (62072206, 61772381); Huazhong Agricultural University Scientific & Technological Self-innovation Foundation; the Fundamental Research Funds for the Central Universities (2662021JC008).

Acknowledgments

We thank anonymous reviewers for their valuable suggestions.

References

1. Chin CS, Sorenson J, Harris JB, *et al*. The origin of the Haitian cholera outbreak strain. *N Engl J Med* 2010; **364**:33–42.
2. Khan S, Imran A, Malik A, *et al*. Bacterial imbalance and gut pathologies: association and contribution of E. coli in inflammatory bowel disease. *Crit Rev Clin Lab Sci* 2019; **56**:1–17.
3. Khan S. Potential role of Escherichia coli DNA mismatch repair proteins in colon cancer. *Crit Rev Oncol Hematol* 2015; **96**:475–82.
4. Khan S, Zaidi S, Alouffi AS, *et al*. Computational proteome-wide study for the prediction of Escherichia coli protein targeting in host cell organelles and their implication in development of colon cancer. *ACS Omega* 2020; **5**(13): 7254–61.
5. Li J, Zakariah M, Malik A, *et al*. Analysis of Salmonella typhimurium protein-targeting in the nucleus of host cells and the implications in colon cancer: an in-silico approach. *Infect Drug Resist* 2020; **13**:2433–42.
6. Hassel B. Tetanus: pathophysiology, treatment, and the possibility of using botulinum toxin against tetanus-induced rigidity and spasms. *Toxins (Basel)* 2013; **5**:73–83.
7. Khan S, Zakariah M, Rolfo C, *et al*. Prediction of mycoplasma hominis proteins targeting in mitochondria and cytoplasm of host cells and their implication in prostate cancer etiology. *Oncotarget* 2017; **8**:30830–43.
8. Davies J, Davies D. Origins and evolution of antibiotic resistance. *Microbiol Mol Biol Rev* 2010; **74**:417–33.
9. Gorski A, Miedzybrodzki R, Wegrzyn G, *et al*. Phage therapy: current status and perspectives. *Med Res Rev* 2020; **40**: 459–63.
10. Kadri SS. Key takeaways from the U.S. CDC's 2019 antibiotic resistance threats report for frontline providers. *Crit Care Med* 2020; **48**:939–45.
11. Cassini A, Högberg LD, Plachouras D, *et al*. Attributable deaths and disability-adjusted life-years caused by infections with antibiotic-resistant bacteria in the EU and the European Economic Area in 2015: a population-level modelling analysis. *Lancet Infect Dis* 2019; **19**:56–66.
12. Towse A, Hoyle CK, Goodall J, *et al*. Time for a change in how new antibiotics are reimbursed: development of an insurance framework for funding new antibiotics based on a policy of risk mitigation. *Health Policy* 2017; **121**: 1025–30.
13. Stokes JM, Yang K, Swanson K, *et al*. A deep learning approach to antibiotic discovery. *Cell* 2020; **180**:688–702.e13.
14. Pires DP, Costa AR, Pinto G, *et al*. Current challenges and future opportunities of phage therapy. *FEMS Microbiol Rev* 2020; **44**:684–700.
15. Edwards RA, McNair K, Faust K, *et al*. Computational approaches to predict bacteriophage-host relationships. *FEMS Microbiol Rev* 2016; **40**:258–72.
16. Villarroel J, Kleinheinz KA, Jurtz VI, *et al*. HostPhinder: a phage host prediction tool. *Viruses* 2016; **8**:116.

17. Liu D, Ma Y, Jiang X, *et al*. Predicting virus-host association by kernelized logistic matrix factorization and similarity network fusion. *BMC Bioinformatics* 2019; **20**:594.
18. Wang W, Ren J, Tang K, *et al*. A network-based integrated framework for predicting virus-prokaryote interactions. *NAR Genom Bioinform* 2020; **2**: lqaa044.
19. Ahlgren NA, Ren J, Lu YY, *et al*. Alignment-free d_2^* oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences. *Nucleic Acids Res* 2017; **45**:39–53.
20. Galiez C, Siebert M, Enault F, *et al*. WIsH: who is the host? predicting prokaryotic hosts from metagenomic phage contigs. *Bioinformatics* 2017; **33**:3113–4.
21. Zhang M, Yang L, Ren J, *et al*. Prediction of virus-host infectious association by supervised learning methods. *BMC Bioinformatics* 2017; **18**:60.
22. Lu C, Zhang Z, Cai Z, *et al*. Prokaryotic virus host predictor: a Gaussian model for host prediction of prokaryotic viruses in metagenomics. *BMC Biol* 2021; **19**:5.
23. Mock F, Viehweger A, Barth E, *et al*. VIDHOP, viral host prediction with deep learning. *Bioinformatics* 2021; **37**:318–25.
24. Hauser R, Blasche S, Dokland T, *et al*. Bacteriophage protein-protein interactions. *Adv Virus Res* 2012; **83**:219–98.
25. Alguwaizani S, Park B, Zhou X, *et al*. Predicting interactions between virus and host proteins using repeat patterns and composition of amino acids. *J Healthc Eng* 2018; **2018**:1391265.
26. Young F, Rogers S, Robertson DL. Predicting host taxonomic information from viral genomes: a comparison of feature representations. *PLoS Comput Biol* 2020; **16**:e1007894.
27. Boeckaerts D, Stock M, Criel B, *et al*. Predicting bacteriophage hosts based on sequences of annotated receptor-binding proteins. *Sci Rep* 2021; **11**:1467.
28. Leite DMC, Brochet X, Resch G, *et al*. Computational prediction of inter-species relationships through omics data analysis and machine learning. *BMC Bioinformatics* 2018; **19**: 420.
29. Leite DMC, Lopez JF, Brochet X, *et al*. Exploration of multiclass and one-class learning methods for prediction of phage-bacteria interaction at strain level. In: *International Conference on Bioinformatics and Biomedicine (BIBM)*. Madrid, Spain: IEEE, 2018, 1818–25.
30. Li M, Wang Y, Li F, *et al*. A deep learning-based method for identification of bacteriophage-host interaction. *IEEE/ACM Trans Comput Biol Bioinform*. 10.1109/TCBB.2020.3017386.
31. Gao NL, Zhang C, Zhang Z, *et al*. MVP: a microbe-phage interaction database. *Nucleic Acids Res* 2018; **46**:D700–7.
32. Russell DA, Hatfull GF. PhagesDB: the actinobacteriophage database. *Bioinformatics* 2017; **33**:784–6.
33. Mihara T, Nishimura Y, Shimizu Y, *et al*. Linking virus genomes with host taxonomy. *Viruses* 2016; **8**:66.
34. Pruitt KD, Tatusova T, Brown GR, *et al*. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res* 2012; **40**: D130–5.
35. Deng Y, Xu X, Qiu Y, *et al*. A multimodal deep learning framework for predicting drug-drug interaction events. *Bioinformatics* 2020; **36**:4316–22.
36. Li J, Pu Y, Tang J, *et al*. DeepATT: a hybrid category attention neural network for identifying functional effects of DNA sequences. *Brief Bioinform* 2021; **22**: bbaa159.
37. Goodfellow IJ, Pouget-Abadie J, Mirza M, *et al*. Generative Adversarial Nets. In: *Conference on Neural Information Processing Systems (NeurIPS)*. Montreal, Quebec, Canada: Curran Associates, Inc., 2014, 2672–2680.
38. Zhu JY, Park T, Isola P, *et al*. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: *IEEE International Conference on Computer Vision (ICCV)*. Venice, Italy: IEEE, 2017, 2223–32.
39. Wan C, Jones DT. Protein function prediction is improved by creating synthetic feature samples with generative adversarial networks. *Nat Mach Intell* 2020; **2**:540–50.
40. Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. In: *Conference on Neural Information Processing Systems (NeurIPS)*. Long Beach, CA, USA: Curran Associates, Inc., 2017, 5998–6008.
41. Huang Y, Niu B, Gao Y, *et al*. CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* 2010; **26**:680–2.
42. Chen Z, Zhao P, Li F, *et al*. iLearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data. *Brief Bioinform* 2020; **21**:1047–57.
43. Yi X, Walia E, Babyn P. Generative adversarial network in medical imaging: a review. *Med Image Anal* 2019; **58**:101552.
44. Gao X, Deng F, Yue X. Data augmentation in fault diagnosis based on the Wasserstein generative adversarial network with gradient penalty. *Neurocomputing* 2020; **396**:487–94.
45. Lopez-Paz D, Oquab M. Revisiting classifier two-sample tests. In: *The International Conference on Learning Representations (ICLR)*. Toulon, France: OpenReview.net, 2017.
46. Xu Y, Zhang Z, You L, *et al*. scIGANs: single-cell RNA-seq imputation using generative adversarial networks. *Nucleic Acids Res* 2020; **48**:e85.
47. Wei L, Ye X, Xue Y, *et al*. ATSE: a peptide toxicity predictor by exploiting structural and evolutionary information based on graph neural network and attention mechanism. *Brief Bioinform* 2021;bbab041. 10.1093/bib/bbab041.
48. Xu W, Zhu L, Huang DS. DCDE: an efficient deep convolutional divergence encoding method for human promoter recognition. *IEEE Trans Nanobioscience* 2019; **18**:136–45.
49. Zhang Q, Zhu L, Huang DS. High-order convolutional neural network architecture for predicting DNA-protein binding sites. *IEEE/ACM Trans Comput Biol Bioinform* 2019; **16**:1184–92.
50. Chuai G, Ma H, Yan J, *et al*. DeepCRISPR: optimized CRISPR guide RNA design by deep learning. *Genome Biol* 2018; **19**:80.
51. Zhang S, Zhao L, Zheng CH, *et al*. A feature-based approach to predict hot spots in protein-DNA binding interfaces. *Brief Bioinform* 2020; **21**:1038–46.
52. Tang X, Zhang T, Cheng N, *et al*. usDSM: a novel method for deleterious synonymous mutation prediction using undersampling scheme. *Brief Bioinform* 2021;bbab123. 10.1093/bib/bbab123.
53. LVD M, Geoffrey H. Visualizing data using t-SNE. *J Mach Learn Res* 2008; **9**:2579–605.
54. Rakov C, Ben Porat S, Alkalay-Oren S, *et al*. Targeting biofilm of MDR Providencia stuartii by phages using a catheter model. *Antibiotics* 2021; **10**:375.
55. Zhan Y, Huang S, Chen F. Genome sequences of five bacteriophages infecting the marine Roseobacter bacterium Ruegeria pomeroyi DSS-3. *Microbiol Resour Announc* 2018; **7**:e00959–18.