研究论文

DOI: 10.12211/2096-8280.2022-016

基于深度学习识别RiPPs前体肽及裂解位点

吕靖伟[1]，邓子新[1]，张琪[2]，丁伟[1]
([1]上海交通大学生命科学技术学院，微生物代谢国家重点实验室，上海 200030；[2]复旦大学化学系，上海 200243)

摘要：得益于基因测序技术的快速发展，基因组测序数据呈现爆炸式增长，核糖体合成和翻译后修饰肽（RiPPs）是近十年逐渐进入人们视野的一大类肽类天然产物。这类化合物在自然界中分布极其广泛，具有丰富的结构多样性和生物活性多样性，是天然药物的重要来源。RiPPs的发现主要依赖低通量生物实验，传统方法精确但成本高昂，随着新型计算机技术的更新迭代，包括antiSMASH、RiPP-PRISM等在内的生物信息学工具能够极大加速RiPPs挖掘进程，但依然无法突破基于同源性方法（例如搜索保守的生物合成酶）的限制——无法有效识别具有不同生物合成机制的新型RiPPs。在这里，本文首次基于自然语言处理预训练模型BERT，提出四种可以完全依赖序列数据识别RiPPs而非基于同源性及基因组上下文信息的深度学习模型，通过对各模型进行验证分析和对比，最终确定在RiPPs识别赛道上表现卓越的最佳模型BERiPPs（bidirectional language model for enhancing the performance of identification of RiPPs precursor peptides）。BERiPPs能够在不考虑基因组背景的情况下以无偏见的方式识别RiPPs前体肽，并可通过条件随机场生成对前导肽裂解位点的预测，为高通量挖掘全新RiPPs提供了思路，并在一定程度下揭示了前体肽和修饰酶间的生物学底层关系。

关键词：深度学习；RiPPs；前体肽；预训练模型；天然产物挖掘

中图分类号：Q819　　**文献标志码：**A

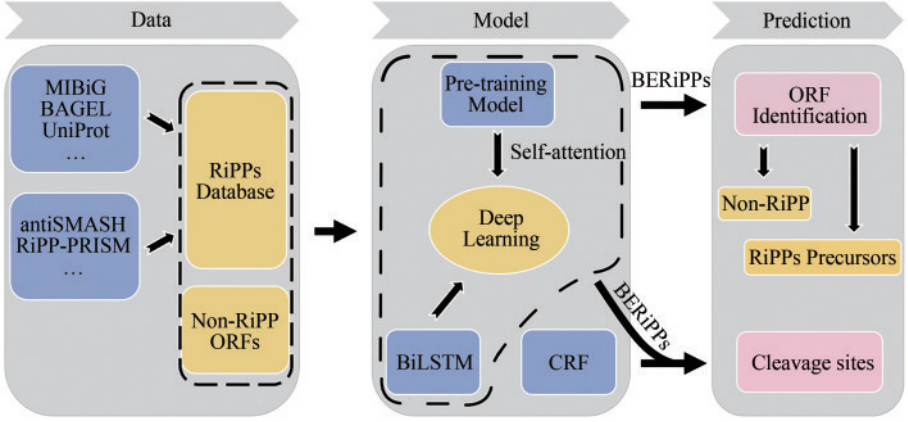
Identification of RiPPs precursor peptides and cleavage sites based on deep learning

LYU Jingwei[1]，DENG Zixin[1]，ZHANG Qi[2]，DING Wei[1]
([1]*State Key Laboratory of Microbial Metabolism*，*School of Life Sciences and Biotechnology*，*Shanghai Jiao Tong University*，*Shanghai*，*200240*，*China*；[2]*Department of Chemistry*，*Fudan University*，*shanghai 200243*，*China*)

Abstract: Genome sequencing data showed explosive growth attributed to the rapid development of DNA sequencing technology. Ribosomally synthesized and post-translationally modified peptides are a kind of natural peptide product that gradually came into people's view in the last decade. These compounds are widely distributed in nature, diverse in structure and bioactivity, and are important sources of natural drugs. The

收稿日期：2022-03-07 修回日期：2022-04-23
基金项目：国家重点研发计划（2018Y F A0900402）
引用本文：吕靖伟，邓子新，张琪，丁伟. 基于深度学习识别RiPPs前体肽及裂解位点[J]. 合成生物学，2022，3(6)：1262-1276
Citation：LYU Jingwei，DENG Zixin，ZHANG Qi，DING Wei. Identification of RiPPs precursor peptides and cleavage sites based on deep learning[J]. Synthetic Biology Journal，2022，3(6)：1262-1276

discovery of RiPPs mainly relies on low-throughput biological experiments, which are accurate but costly. With the development of new information technologies, bioinformatics tools such as antiSMASH and RIPP-Prism can greatly accelerate the process of RiPPs mining. However, methods based on gene homology, such as searching for conserved biosynthetic enzymes, are still unable to effectively identify novel RiPPs with different biosynthetic mechanisms. Here, for the first time, based on the natural language processing pre-training model BERT, four deep learning models that can fully rely on sequence data to identify RiPPs instead of homology and genomic context information are proposed and trained on the same RiPPs dataset. Through verification and comparison of these models, the best model BERiPPs performs well on the RiPPs identification track and is as accurate as the homology-based method. BERiPPs can identify RiPPs precursor peptides and RiPPs classes in an unbiased manner regardless of the genomic background, extending the range of novel RiPPs captured by approximately 60% compared to homology-based approaches. By combining BERiPPs with a conditional random field, the prediction of the cleavage site of the leader peptide can be indirectly generated with high accuracy by the recognition of each amino acid label in the sequence. The deep learning based on the pre-training model provides the possibility for high-throughput mining of novel RiPPs in a manner different from that of the gene context-dependent methods and reveals the underlying biological relationship between precursor peptides and modified enzymes.



Keywords: deep learning; RiPPs; precursor peptides; pre-training model; natural products mining

越来越多且不断进化的多重耐药菌对现代医学提出了持续的挑战，如何加快开发新型抗菌化合物来对抗危及人类健康的感染一直是困扰医学领域的难题[1]，也是亟待解决的重要议题。临床上广泛使用的抗生素大部分来源于微生物小分子代谢产物，如聚酮合酶家族和非核糖体肽合成酶家族，其中极具代表性的抗生素有四环素类、糖肽类等[2]。随着全世界范围内病菌不断变异及抗生素过度使用，病菌的耐药性上升，这些常见的抗生素对抗病菌的能力日益下滑，而生物合成机制限制了非核糖体肽合成酶和聚酮化合物合成酶大量生产实质上不同的类似物的能力[3]，因此具有与前者不同合成机制的核糖体合成和翻译后修饰肽（ribosomally synthesized and post-translationally modified peptides， RiPPs）引起了广泛关注，有希望成为抗生素队伍的强力补充。

通过核糖体合成的前体肽在大多数情况下由一个N末端区域（前导肽）和一个C末端区域（核心肽）组成，前导肽通常通过结合修饰酶的识别元件介导底物和酶的识别，修饰完成后会经蛋白水解去除，而核心肽以位点特异性方式进行翻译后修饰（图1），并最终转化为成熟的RiPPs[4]。目

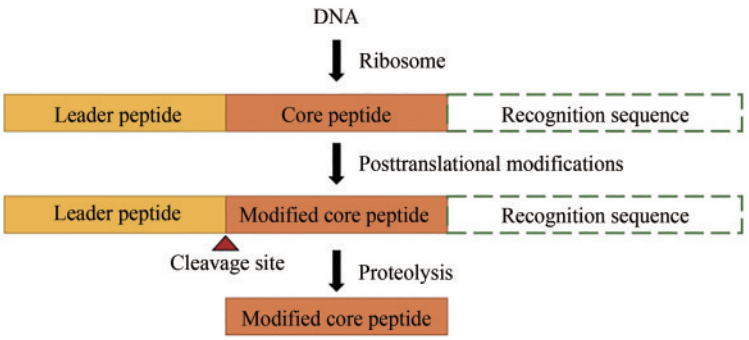


图1 RiPPs的生物合成途径

（图中各部分序列比例仅作图示用，不代表其实际长度）

Fig. 1 Biosynthetic pathways of RiPPs

(The proportion of each part of the sequence in the figure is for illustration only and does not represent its actual length)

前RiPPs的发现还是以低通量的生物实验为主导，快速且廉价的基因测序技术使得基因组数据成倍增长，而绝大多数基因编码的天然产物仍然未知，因此大量生物信息学工具被开发用以根据序列数据来预测天然产物的种类和结构，揭示了数以千计未知的假定RiPPs[5]。

最经典且最流行的方式是通过保守的RiPPs合成酶编码基因簇挖掘已知类别中的新成员。研究表明，不同的RiPPs家族在结构和功能上均存在很大的差异，不同的修饰酶在其中发挥着至关重要的作用。值得注意的是，RiPPs通常有以低遗传成本获得高化学多样性的潜在进化优势[6]，某些RiPPs合成酶（例如Ⅰ型羊毛硫肽合成酶中独立的脱水酶LanB和环化酶LanC[7]）极其保守。基于此特性，对于那些同已知RiPPs共享保守的合成酶但在前体肽上与后者存在明显差异或同时存在其他合成酶的生物合成基因簇（biosynthetic gene clusters，BGCs），可以认为是存在已知RiPPs类型中的新型成员[8]，这为绝大部分用于挖掘RiPPs的生物信息学工具提供了思路。AntiSMASH是被广泛运用的BGCs预测和分析平台[9]，基于隐马尔可夫模型以及前体肽识别的“规则”（如前体肽中通常富含相关生物合成酶作用的残基[10]），在与已知RiPPs序列的数据库比较后可检测RiPPs，例如具有独特且保守修饰酶的羊毛硫肽和套索肽。

寻找新的前体肽是挖掘RiPPs的另外一种常用方法。前体肽是RiPPs生物合成的前体，即便与已知类型共享同类修饰酶，在前体肽差异较大的前提下，亦存在得到结构与功能完全不同的全新RiPPs的极大可能。然而编码RiPPs前体肽的开放阅读框（open reading frame，ORF）通常较小，经常被经典的基因预测算法如Prodigal[11]和Glimmer[12]所忽视，因此这种算法往往不适用于RiPPs的挖掘。Kuipers课题组[13]提出的BAGEL通过搜索核心的翻译后修饰酶附近小段ORFs来预测RiPPs前体肽，很好地解决了这一问题。Mitchell课题组[14]设计的RODEO算法则在前者的基础上更加完善，结合隐马尔可夫模型和Pfam数据库检索RiPPs的生物合成基因簇，并通过机器学习和启发式算法对RiPPs前体肽生成预测。

然而随着研究的不断深入，RiPPs的多样性和复杂性对挖掘算法计算策略提出了较高的挑战。以上提及的生物信息学方法在底层逻辑上均未离开基于基因同源性及上下文信息的范畴，即只有潜在的生物合成基因与已知的RiPPs相似时才会被识别，所以不可避免地面临两大问题——无法识别新的RiPPs家族，无法判断是否存在新的酶修饰机制[15]。可以肯定地说，利用这种策略挖掘RiPPs也只会得到已知RiPPs的类似产物。如果能够揭开RiPPs前体肽的内在规律和特性，不难预见挖掘RiPPs甚至其他天然产物的研究将跨入新纪元。随着计算机技术的飞速发展以及算力水平大幅提升，科学家们尝试将人工智能运用在生物化学领域，取得了不错的进展。Agrawal等尝试通过机器学习来对RiPPs进行预测，提出的RiPPMiner工具实际上是基于机器学习中支持向量机（support vector machine，SVM）算法的衍生，也是分类算法的一种。Agrawal等将经实验表征的RiPP组成的数据库

用于训练RiPPMiner，抽取RiPPs前体肽特征，从大量的组合可能性识别前体肽中正确的交联模式，并以此将RiPPs前体肽与其他肽类区分开来，分为RiPPs的12个子类，并预测前导肽可能的裂解位点。结果表明，RiPPMiner仅在羊毛硫肽类数据集上取得了良好的效果，展示了其在RiPPs大家族鉴定上的高灵敏度和特异性，其主要原因是机器学习的效果往往受到训练数据集的大小和质量的影响较大，训练RiPPMiner所采用的数据集过小，加之支持向量机算法的固有缺陷，RiPPMiner在对其他RiPPs家族识别及裂解位点预测上没有取得很好的效果，后者精度仅为0.69[16]。随后新的人工智能方法深度学习也被用于进一步提升预测的准确性，NeuRiPP将卷积神经网络（convolutional neural network，CNN）和长短期记忆网络（long short-term memory，LSTM）结合来判断短肽是否为RiPPs前体肽[17]，但CNN和LSTM在一定程度上依然存在局限性，如前者在训练中可能出现忽略局部和整体间关联性的情况[18-19]。DeepRiPP[15]在采用深度神经网络的基础上将基因组和代谢组学信息结合用于对新型RiPPs的鉴定。为了在现有研究基础上进一步探索适合处理蛋白质序列的深度学习方法，本文首次将基于自注意力机制的NLP预训练模型BERT[20]（bidirectional encoder representations from transformers）运用于天然产物的挖掘，并提出可以完全依赖序列数据而非基因组背景识别RiPPs的深度学习模型BERiPPs，其能够在全新的测试数据集上达到90%以上的预测准确率。

1 方法

氨基酸序列在某种程度上可以视作为大自然的语言，每一个氨基酸则是唯一、独特的词汇，可以通过针对自然语言开发的神经网络模型对它们进行建模。为了得到在天然产物挖掘领域中具有良好性能的深度学习模型，本文基于RiPPs前体肽数据库，利用预训练模型BERT提取特征，将BERT输出作为embedding输入到与RiPPs识别任务相关的其他主流深度学习模型中，并在该任务数据上进行微调，最后对所有模型的训练效果进行对比评估，其中预测效果最好、最稳定的模型本研究将其命名为BERiPPs。

1.1 用于模型训练的RiPPs数据集

在监督学习中，数据集的“好坏”直接影响到模型最终性能。数据集过小，深度神经网络容易过拟合，泛化能力欠缺。设计合适数据集是实现任务的关键一步。尽管微生物基因组的生物信息学分析显示存在大量的RiPPs生物合成基因簇，但只有其中一小部分的亚类、前导肽切割位点已被实验验证。本文从MIBiG、BAGEL、UniProt等数据库以及大量的RiPPs相关文献收集了13类共计527个RiPPs，其中主要集中在羊毛硫肽、蓝藻素、硫肽和套索肽等。在此基础上，本文又通过基于隐马尔可夫模型，根据已知RiPPs修饰酶谱来识别RiPPs的antiSMASH和RiPP-PRISM等工具进一步扩大了数据集，共3007个RiPPs前体肽正例样本，并按照1∶1的比例随机添加了非RiPPs前体肽的负例样本。除预测RiPPs前体肽裂解位点任务数据集仅有正例样本外，其他任务训练模型均使用同一个数据集，并按照8∶1∶1的比例分为训练集、验证集和测试集。

值得一提的是，本研究在处理RiPPs数据集的过程中发现绝大多数的RiPPs前体肽长度处于20～100个氨基酸之间，而随机采样的编码非RiPPs前体肽的ORF平均长度要远大于前者。本文基于简单经典的二分类模型支持向量机，以一个由2000个编码RiPPs前体肽和1000个随机采样的非RiPPs前体肽的ORF组成的数据集对ORF长度进行预测建模，该模型仅根据ORF长度就可以对识别RiPPs前体肽达到82.67%的精度（图2）。因此，为了防止模型将ORF长度作为判定是否为RiPPs前体肽的重要因素，仅根据ORF长度便可得到较高的RiPPs前体肽预测精度，本文将所有负例样本中的ORF长度控制在100以内。

1.2 数据集的标注

样本序列中的每个氨基酸（共20种天然氨基酸）可以看作是独一无二的词元，在此基础上，添加特殊类别词元［CLS］、分割词元［SEP］、填充词元［PAD］、未知词元［UNK］以及掩码词元

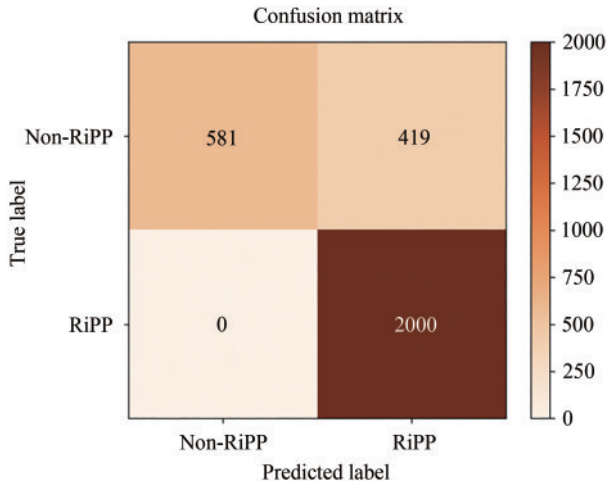


图2 根据ORF长度识别RiPPs前体肽

Fig. 2 Identification of RiPPs precursor peptides based on ORF length

[MASK]，共同组成模型训练所需的词汇表。在RiPPs前体肽识别任务中，编码候选RiPPs前体肽的ORFs将作为模型的输入，得到其分类预测的输出（RiPPs前体肽或非RiPPs前体肽），即为RiPP标签或Non-RiPP标签。在RiPPs前体肽类别的预测任务中，模型的输出是13个RiPPs家族类别以及非RiPPs，共计14个标签。RiPPs前体肽裂解位点的预测任务在数据集处理上则与前者不同，本文基于BIO（Begin、Internal、Other）标注规则将RiPPs前体肽裂解位点处的核心肽起始氨基酸标记为B，核心肽中除起始位点外的其他氨基酸标记为I，其余所有氨基酸均用O表示（图3），模型将会对序列中每个位置的氨基酸标签进行预测，以判断RiPPs前体肽裂解位点。

1.3 预训练模型BERT

回溯自然语言处理领域的发展史，2018年自然语言处理在深度学习飞速发展的浪潮下取得了巨大的突破，在全世界NLP开源社区的贡献下，各种运用深度学习技术处理文本的能力通过预训练模型极大地发挥了出来，其中Google提出的预训练模型BERT堪称划时代的作品。BERT首先在大规模无监督语料上进行预训练，通过预训练好的参数使模型具有挖掘数据中的因果关系及特征信息的能力，之后在下游任务上进行微调。这种预训练加微调的方法无疑是目前自然语言处理解决方案中的最佳范式[21]。

过去生物学领域有关深度学习应用中，循环神经网络（recurrent neural network， RNN）几乎被用在所有序列处理上。然而RNN有着比较明显的短板，例如基于RNN的seq2seq模型编码器将所有信息都编码到了一个固定长度的context向量中[22]，一方面单个向量很难包含所有序列的信息，另一方面RNN递归地编码序列使得模型在处理长序列时面临非常大的挑战，比如RNN处理到第500个氨基酸的时候，很难再包含前1～499个氨基酸中的所有信息。虽然RiPPs前体肽的ORFs相对比较短，受到RNN处理长序列短板的影响较小，但本研究希望模型足够泛化，可以覆盖到其他未知领域。Bahdanau等首次提出一种叫做注意力attention[23]的技术，让模型可以有区分度、有重点地关注输入序列。Vaswani等将注意力机制进一步完善，提出的Transformer模型引入“自注意力（self-attention）”“多头注意力（multi-head attention）”概念拓展模型关注不同位置的能力[24]。BERT采用了Transformer编码器，继承了其提取有效语义信息的能力。

自注意力机制如式（1）所示：

$$\mathrm{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \mathrm{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \qquad (1)$$

式中，$\boldsymbol{Q}$，$\boldsymbol{K}$，$\boldsymbol{V}$表示同一输入与不同参数计算后得到的3个矩阵，$\sqrt{d_k}$是k维度的调节平滑因子，防止相乘结果过大。

以往的NLP预训练通常是基于语言模型进行

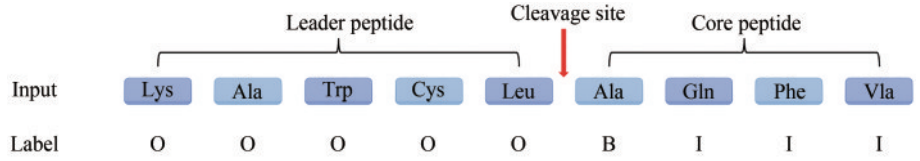
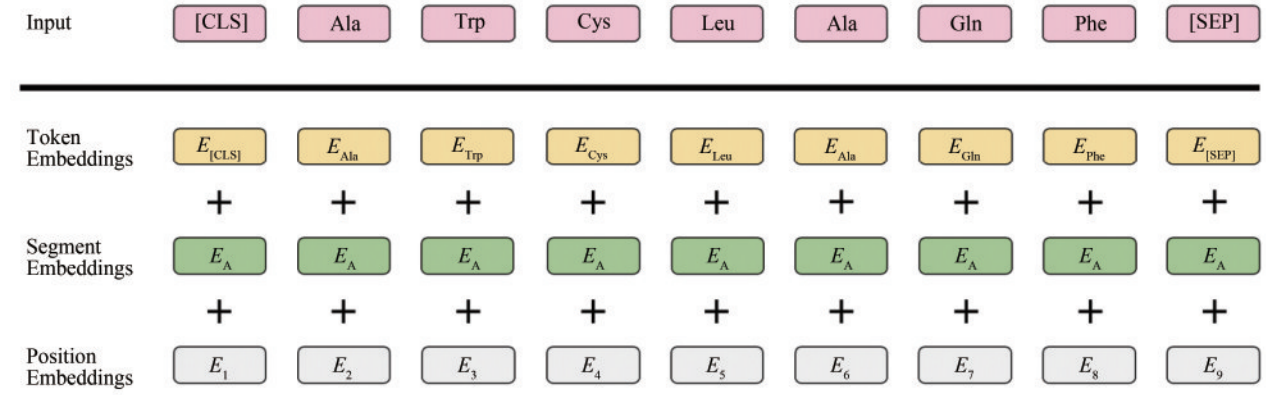


图3 基于BIO规则的序列标注示例

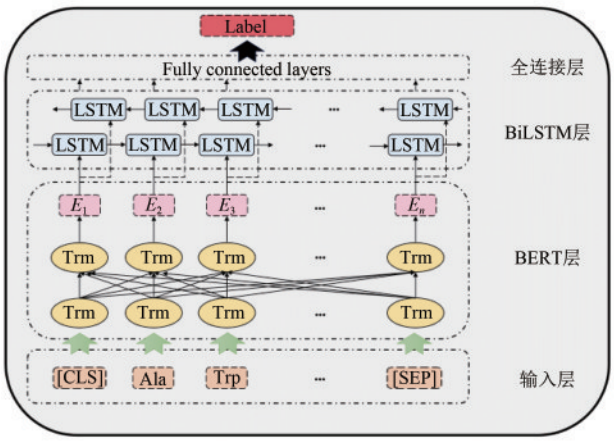
Fig. 3 Sample sequence annotations based on BIO rules

的，比如给定语言模型的前5个词，让模型预测第6个词，这也就意味着模型只能从单个方向上考虑语义间的依赖关系[25]，极大地限制了模型的表达能力。但BERT是基于掩码语言模型（masked language model，MLM）进行预训练，即将输入的序列中15%的词元随机遮掩，让BERT来预测这些被遮掩的词元[20]。这种类似于“完形填空”的任务让模型能够有效地学习到双向编码的能力，编码的结果可以同时包含前后双向的语境信息。本文将这种能力用于解析氨基酸序列，能够很好地捕捉每一个氨基酸残基与全局的关系，而非仅某一段序列。

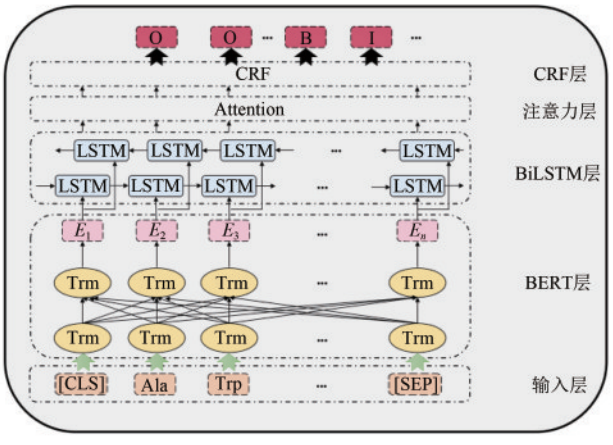
在BERT中，作为输入的Embedding层由Token Embeddings、Segment Embeddings、Position Embeddings三部分构成：Token Embeddings表示词元向量；Segment Embeddings对序列进行编码，用于表征序列的全局语义信息，区分不同序列；Position Embeddings对位置信息进行编码，作为对不同位置的词元的附加信息[20, 21]，这三部分相加组成了每个词元最终的向量表示形式［图4(a)］。模型通过查找词汇表将输入的氨基酸序列转换为向量形式，其中，特殊标记［CLS］表示序列的开始，特殊标记［SEP］表示序列的间隔或结束。值得注意的是，在自然语言中不同的词可能含有同样的含义，而同样的词在不同的语境中也可能存在完全不同的理解。这套规则同样适用于生物领域，不同的氨基酸可能具有相似的结构和功能角色，相同的氨基酸也可能出现生物学意义上的差



(a) BERT模型的输入表示
(a) Input representation of BERT



(b) BERiPPs模型结构
(b) Model structure of BERiPPs



(c) BERiPPs-CRF模型结构
(c) Model structure of BERiPPs-CRF

图4 模型输入表示及结构

Fig. 4 Model input representation and structure

别。BERT核心的自注意力机制会通过文本中的其他词来增强目标词的语义表示，因此，即便对于相同的氨基酸残基，在不同的上下文背景下其最终对应的输出也是不同的，从而能够解决“一词多义”的问题。

1.4 识别RiPPs前体肽的最佳模型BERiPPs

如何将氨基酸序列通过数值完整地表示出来是利用计算机在蛋白质领域进行高通量计算的关键所在。尽管包括BERT在内的绝大多数基于序列的语言模型设计初衷都是为了处理以英文为主的自然语言，但蛋白质序列在结构上与自然语言存在一定的相似之处，序列整体信息由单个氨基酸本身含义及各氨基酸间的关系共同决定。基于预训练模型BERT提取底层特征的优良性能，本文将整个BERT作为Embedding层接入到其他主流模型，从而组成多个新深度学习模型，并在同一个RiPPs前体肽数据集上进行训练和验证，最后将各个模型进行比较、评估，得到基于BERT及双向长短时记忆网络（bi-directional long short-term memory，BiLSTM）的最佳模型BERiPPs［图4(b)］。BERiPPs能够不考虑基因组背景，仅通过候选前体肽序列就能较准确地识别RiPPs，这种能力既是发现全新RiPPs家族所必需的，又是依靠保守的修饰酶簇或其他同源性方法所不具备的。

为了进一步探究在较大的领域偏移下基于大规模通用域文本数据集的预训练参数对分类性能的影响，且考虑到在小规模样本上靠近输出层的BERT高层参数会带来负面作用的可能性，本文分别将BERT顶部1层、2层以及整个BERT的预训练参数按照BERT默认的标准差为0.02的截断正态分布进行随机初始化（图5）。参考post-norm[26]的设计，偏小的标准差在一定程度上有利于缓解梯度消失的问题。同时，为了使模型更“贴近”下游任务，本文也尝试基于未标注的RiPPs数据集通过MLM任务对在通用域文本语料上预训练的BERT再次训练。

1.5 利用BERiPPs-CRF预测RiPPs前体肽裂解位点

鉴定某个蛋白家族通常是利用隐马尔可夫模型基于对蛋白质序列多重比对结果中的保守区域进行搜索，考虑不同保守度的氨基酸在相应位置的权重，从而可以捕获进化距离较远的蛋白相关性[27]，但其极为严格的独立性假设条件限制了描述序列全局信息的能力。本文将构成RiPPs前体肽的前导肽和核心肽视作两组含有不同语义信息的序列，即ORFs中的每一个氨基酸只有两种状态，核心肽构成成员和非核心肽构成成员，这可以被认为是在氨基酸水平上的序列标注。由BERiPPs输出的序列标签是由词本身以及上下文特征所决定，但基于BIO标注规则，如实体起始标签为B、标签Ⅰ一定在B之后等，不同的标签中存在硬性约束[28]。所以本文在提取序列特征上表现优秀的BERiPPs模型基础上加入条件随机场（conditional random field，CRF）来引入一些约束关系。CRF是一种经典的判别式概率无向图模型，在隐马尔可夫模型的基础上，可以容纳任意的上下文信息，得到全局的概率分布，该模型主要应用于自然语言处理中的序列标注任务[29]。

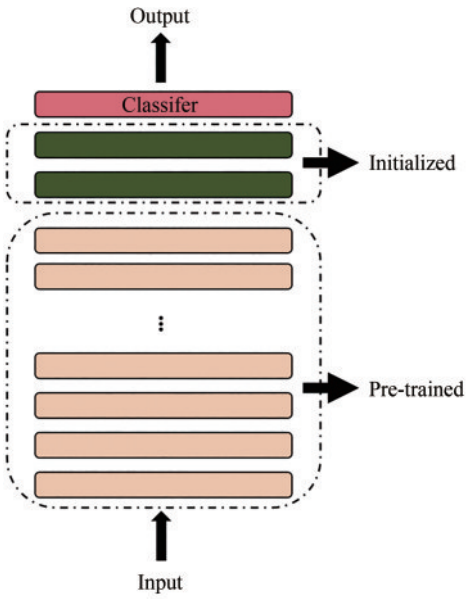


图5 BERT预训练层的初始化

Fig. 5 Initialization of pre-trained layers of BERT

通过引入转移得分矩阵$\boldsymbol{A}$来表示各标签之间的关系，矩阵元素$\boldsymbol{A}_{i,j}$表示从标签转移到标签j的概率。给定输入序列$X=(x_1, x_2, \cdots, x_n)$，输出标签序列$Y=(y_1, y_2, \cdots, y_n)$，输入序列$X$得到输出序列$Y$的得分公式如式（2）：

$$\mathrm{Score}(X,Y)=\sum_{i=1}^{n}P_{i,y_i}+\sum_{i=1}^{n+1}A_{y_{i-1},y_i} \quad (2)$$

式中，P_{i,y_i}表示第i个词预测为标签y_i的得分。

同时，为了进一步更好地筛选BERiPPs输出的特征信息，突出对核心肽实体识别起关键作用的特征，注意力机制被作为BERiPPs和CRF层中间的结构来对输出的特征向量进行权重分配[30]，模型结构如图4(c)所示。

BERiPPs-CRF模型会对输入序列中的每个氨基酸生成相应的标签，从而得到RiPPs前体肽裂解位点的预测。实际上，酶切反应也是一种极具约束性的化学反应，切割的位点受到酶、底物以及反应环境等多因素的影响会呈现一定的特征。比如本研究发现，套索肽前体常在甘氨酸、半胱氨酸或丝氨酸处断开，而BERiPPs-CRF在预测套索肽切割位点上表现得极为出色。

1.6 模型评估指标

为了评估各个模型的性能，本实验使用“提前停止”（early stopping）技术（一旦训练效果停止改善，立即自动停止训练过程）[31]，这可以更好地避免过拟合问题。主要采用3种常用的评估指标：精确率（Precision）、召回率（Recall）以及精确率和召回率的调和均值（F1 score）以上指标计算方法如式（3）～式（5）：

$$\mathrm{Precision}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \quad (3)$$

$$\mathrm{Recall}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \quad (4)$$

$$\mathrm{F1}=\frac{2\times\mathrm{Precision}\times\mathrm{Recall}}{\mathrm{Precision}+\mathrm{Recall}} \quad (5)$$

式中，TP表示数据集中被判定为正类，实际也为正类的样本数；FP表示数据集中被判定为正类，实际却为负类的样本数；FN表示数据集中被判定为负类，实际却为正类的样本数。

1.7 实验环境设置

本实验研发环境基于Google Colab平台，深度学习框架为pytorch1.9.0版本，所采用的BERT为Github的公开开源版本。

2 结果与分析

2.1 RiPPs前体肽的识别

在同一个数据集上，分别采用BERT、BERT-CNN、BERT-DPCNN、BERT-RCNN、BERT-BiLSTM模型进行训练，在对模型超参数多次优化后得到各模型在RiPPs前体肽识别任务上的预测性能测试结果，如表1所示。

表1 不同算法在RiPPs前体肽识别任务上的效果对比

Tab. 1 Comparison of different algorithms in identification of RiPPs precursor peptides

Model	Precision	Recall	F1
BERT (top 1 layers initialized)	0.9056	0.8962	0.9009
BERT (top 2 layers initialized)	0.8938	0.9031	0.8985
BERT (fully initialized)	0.8710	0.8408	0.8556
BERT (pre-trained)	0.9031	0.9031	0.9031
BERT-CNN	0.9123	0.8997	0.9059
BERT-DPCNN	0.9126	0.9031	0.9078
BERT-RCNN	0.9127	0.8685	0.8901
BERiPPs (BERT-BiLSTM)	0.9331	0.9170	0.9250

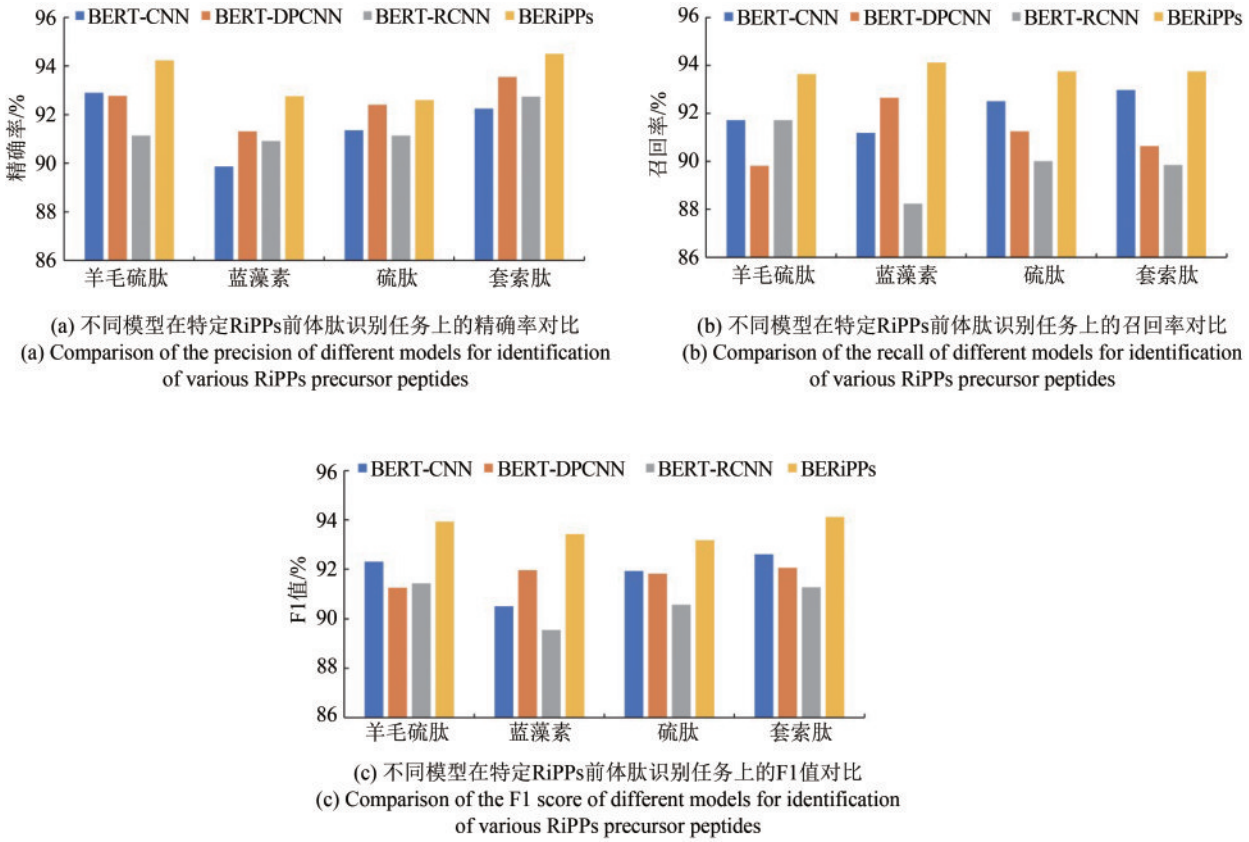
可以看到，直接将预训练模型BERT用于RiPPs前体肽识别也能取得较好的预测效果，这不仅体现了BERT在自然语言处理上的强大，也有理由相信BERT在基于特定领域调整后会得到理想的效果。但令人意外的是，不管是在将BERT顶部1层还是2层参数重新初始化后，模型的预测性能都没有较大的变化，只是在训练中收敛速度略微加快，而基于完全随机初始化后的BERT重新训练，最终结果比预训练后的BERT精确率低了3.2%，基于对比结果不妨可以合理地做出推测，BERT在特殊领域下的较好性能表现不仅源于通过预训练得到的初始化参数相较于随机初始化对特征抽取性能的提升，其强大的学习能力也是关键因素之一。在将BERT作为embedding层，与其他主流模型组合后发现识别效果较BERT原模型有了一定提升，且与RNN模型组合后的预测性能整体上要优于CNN，这也从侧面反映了RNN在处理文本序列

上较CNN具有更好的性能表现。其中，由BERT与BiLSTM组合的BERiPPs模型明显优于其他组合模型，其精确率、召回率和F1值分别为93.31%、91.70%、92.50%。

为了探究BERiPPs在特定RiPPs类别上预测性能，本文将数据集中占较大比重的四类经典RiPPs，即羊毛硫肽、蓝藻素、硫肽和套索肽，分别用于各模型的训练。同样地，通过精确率、召回率和F1值这3个指标，对比各模型在特定类别上的预测能力（图6）。可以直观地看到，即使在特定类别的RiPPs数据集上，BERiPPs依然保持着最佳的预测成绩，而且相比于对整个RiPPs数据集进行训练，预测性能略有提升，这与同一RiPPs家族在氨基酸序列上有更多的保守特征是吻合的，展现了BERiPPs在提取RiPPs前体肽序列全局信息上的出色性能。

2.2 RiPPs类别的预测

像antiSMASH这样的基因簇分析工具可以根据BGCs中修饰酶来预测RiPPs的类别，例如合成羊毛硫肽的脱水环化酶以及套索肽的天冬酰胺合成酶，这样以修饰酶为中心来搜索可能的相邻前体短肽的策略可以使antiSMASH快速分析BGCs中是否存在RiPPs以及RiPPs的类别，但这种策略只能局限于已知的生物合成途径范围内，而BERiPPs仅通过RiPPs前体的氨基酸序列训练的深度学习模型来生成对RiPPs类别的预测，在对RiPPs精准识别的基础上进一步拓展。BERiPPs训练的结果符合预期，在例如各型羊毛硫肽、套索肽及硫肽等大型RiPPs类的识别上准确率较高，但受限于小类RiPPs数据集十分有限，对林那肽类、赛克肽类等的预测精度较低。对此，本文采用k折交叉验证以及留出法（hold-out method）[32]结合的方式，将原始数据集按照9∶1的比例分为A、B数据集。之后将数据集A等分为9部分，每次采用不同的部分作为验证集，其余数据均作为训练集以此训练并验证模型，重复9次后再将作为hold-out set的数据集B分别用于测试模型BERiPPs。这样可保证数据集



(a) 不同模型在特定RiPPs前体肽识别任务上的精确率对比
(a) Comparison of the precision of different models for identification of various RiPPs precursor peptides

(b) 不同模型在特定RiPPs前体肽识别任务上的召回率对比
(b) Comparison of the recall of different models for identification of various RiPPs precursor peptides

(c) 不同模型在特定RiPPs前体肽识别任务上的F1值对比
(c) Comparison of the F1 score of different models for identification of various RiPPs precursor peptides

图6 各模型对特定RiPPs家族识别结果对比

Fig. 6 Comparison of identification results of various RiPPs families by different models

除hold-out set以外的所有数据都能参与模型的训练，使模型对于数据的划分不那么敏感。在目前采用深度学习挖掘RiPPs的主流工具中，NeuRiPP将重点更多地聚焦于在大量非RiPPs负例样本中准确地识别RiPPs[17]，为了更直观地评估BERiPPs在RiPPs类别预测上的性能，本研究选取了DeepRiPP在同样的测试集上的预测结果作为对比，如图7所示训练方式的优化在一定程度上提高了BERiPPs的预测性能和泛化能力。同时，DeepRiPP也展现出了强大的RiPP识别能力，与BERiPPs在对不同类别的识别性能上各有优劣，但比较明显的是在对林那肽类的预测能力上明显低于优化后的BERiPPs。

RiPPs样本的不均衡带来的问题是显而易见的。在某种意义上，模型训练的本质可以理解为损失函数的最小化，大多数分类任务上的损失函数会采用交叉熵损失（cross entropy loss， CEL）[33]，当训练样本中的某类样本数量远多于其他类，损失函数的输出会极大地受该类样本所影响，最终导致模型分类结果向该样本类别倾斜。反之，数据量较少的类别对于总损失影响较低，模型容易忽视该类样本。基于以上分析，对损失函数的优化可能会是解决RiPPs类别不平衡问题的有效方式，故本文引入提出于图像处理领域的focal loss（FL）[34]，通过减少易分类样本的损失权重，让模型在训练中更专注于稀疏的难分类样本。加入了平衡参数α的focal loss表示公式［式(6)］：

$$\mathrm{FL}(p_t) = -\alpha_t(1-p_t)^{\gamma}\log(p_t) \qquad (6)$$

式中，α_t表示该类样本对应的权重参数，在深度学习多分类任务中对应形状为［1,categories］的张量（tensor），其中categories为类别数；p_t表示预测结果对应标签的概率，$p_t \in [0,1]$；$(1-p_t)^{\gamma}$为调制因子（modulating factor）；γ为聚焦参数（focusing parameter），$\gamma \geqslant 0$。

在实际训练过程中，参数γ和α的取值会直接影响focal loss的效果。本研究对两类参数的设置做了简单的评估，先将所有类对应的权重均设置为1，分别将$\gamma = 0$（即等同于交叉熵损失）和$\gamma = 2$代入训练后发现，基于focal loss的BERiPPs仅对自诱导肽及硫肽类样本识别精度有一定提高，但对林那肽类、LAPs（linear azole-containing peptides）等小类样本没有明显改善，且对包括羊毛硫肽、套索肽在内的大类样本预测性能有所下降，模型整体加权平均精确率下降了0.4%。因此，考虑到γ取值较大，故取每个RiPPs类别占总样本比例的值

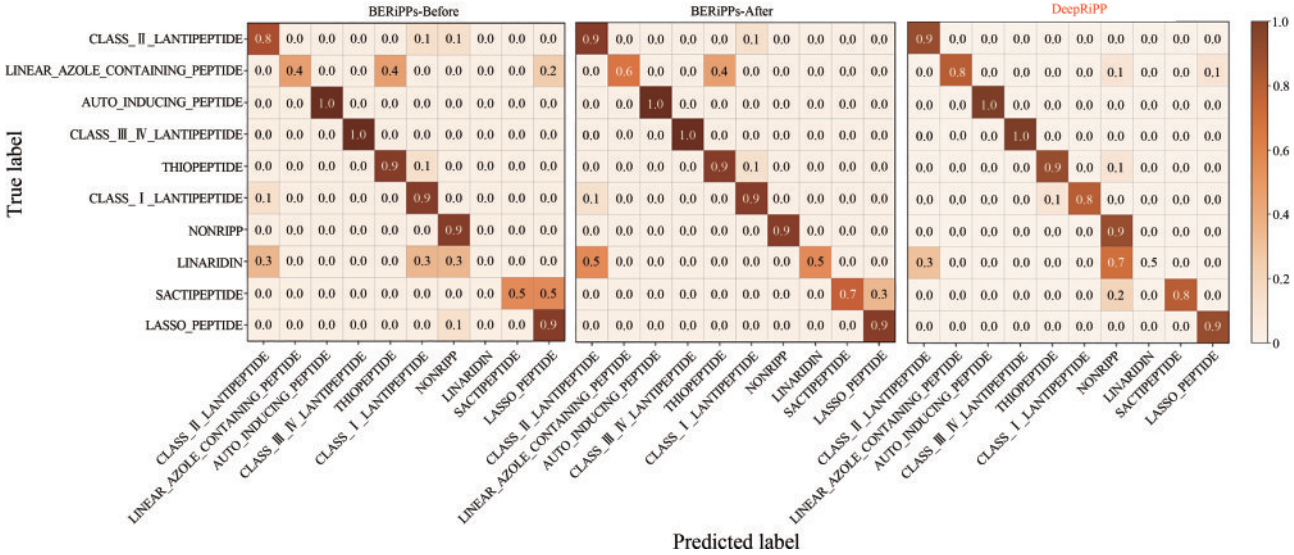


图7 不同训练方式下的BERiPPs和DeepRiPP在预测RiPPs类别上的结果对比

（因测试集中各类RiPPs样本数量不同，故将混淆矩阵中的数值进行归一化处理，再根据四舍五入原则精确到小数点后一位）

Fig. 7 Comparison of prediction results of RiPPs classes between BERiPPs under different training methods and DeepRiPP

(Due to the different number of various RiPPs samples in the test set, the values in the confusion matrix are normalized and then accurate to one decimal point according to the rounding principle.)

表2 不同损失函数下的部分RiPPs类别预测结果对比

Tab. 2 Comparison of prediction results of partial RiPPs classes under different loss functions

Class	Loss Function	Precision	Recall	F1
Autoinducing peptides	Focal Loss	1.0000	0.9583	0.9787
	Cross Entropy Loss	0.9231	1.0000	0.9600
Thiopeptides	Focal Loss	0.8947	0.8947	0.8947
	Cross Entropy Loss	0.8500	0.8947	0.8718
Lasso peptides	Focal Loss	0.8873	0.9130	0.9000
	Cross Entropy Loss	0.8889	0.9275	0.9078
Class Ⅲ_Ⅳ Lanthipeptides	Focal Loss	0.9483	0.9649	0.9565
	Cross Entropy Loss	0.9655	0.9825	0.9739

作为α中的对应元素值以反向平衡，降低对大类样本的负面影响，再次进行实验后，如表2所示的四类RiPPs的测试结果前后对比较明显。

总体来说，focal loss的采用在局部上对RiPPs类别预测产生了积极影响。本文没有对参数γ和α进行更为细化的对比实验，虽然这可能会进一步小幅提升模型在类别预测上的性能，但重点是通过对损失函数的优化来缓解RiPPs类别不平衡问题的想法得到了一定的验证。更值得一提的是通过以上尝试有所启发，本研究注意到损失函数的改变对不同RiPPs类别的影响并不一致。显然，对于RiPPs各类别预测的差异化结果，各类样本数量间的巨大差异总是会吸引更多的关注，从而容易忽视其他可能存在的影响因素，比如部分RiPPs可能相比于其他家族具有更丰富的多样性以至于模型很难捕获完整特征，或是在收集样本之前被错误地分类，又或是属于尚未被充分表征的未知类别，而这还有待进一步探究。但可以预见的是，随着对RiPPs研究的不断深入及RiPPs数据库的扩充，以人工智能驱动RiPPs挖掘的策略能够展现出更好的性能。

2.3 RiPPs前体肽裂解位点的预测

BERiPPs-CRF模型通过对氨基酸序列的标注进行识别，从而间接生成对RiPPs前体肽裂解位点的预测。从实体识别的角度来说，BERiPPs-CRF依然表现了较高的水平，精确率、召回率和F1值分别为90.45%、91.33%和90.88%。但从对RiPPs前体肽裂解位点预测的角度来看，其重点在于能否准确判断标签B（即核心肽起始氨基酸）所在的位置。同样受限于数据集中各类RiPPs样本的数量以及不同RiPPs家族所展现的前体切割规则的差异，在对RiPPs前体肽裂解位点的预测上准确率出现了明显的两极差异。例如对套索肽裂解位点的准确率达到了70%以上，Ⅰ、Ⅱ型羊毛硫肽裂解位点的预测准确率也超过了60%，而小型的RiPPs家族的预测结果则不太理想。如果把与真实裂解位点相差±5个氨基酸的预测也纳入考虑范围之内，那么整体预测的准确率为80.67%。本文将基于机器学习的RiPPMiner用同样的测试数据集进行预测，与BERiPPs-CRF模型对比结果如图8(a)所示。对于没有准确识别真实裂解位点的样本而言，模型预测的位点与实际位点相隔越近，则越有RiPPs研究的借鉴意义，如果模型能够做到把预测与真实位点之间的间隔控制在±5甚至±3、±1个氨基酸以内，其模型的价值也将按倍数增加。因此，为了更直观地展现预测位点与实际位点的偏差程度，本文根据统计学规则引入一个新的评估指标偏位度V，计算公式如式（7）：

$$V = \frac{\sum_{i=0}^{n} x_i^{\,2}}{n} \tag{7}$$

式中，x_i为各测试样本的预测裂解位点与实际位点之间的间隔；n为样本总数。V越小，代表模型对RiPPs前体肽裂解位点预测的能力越强。

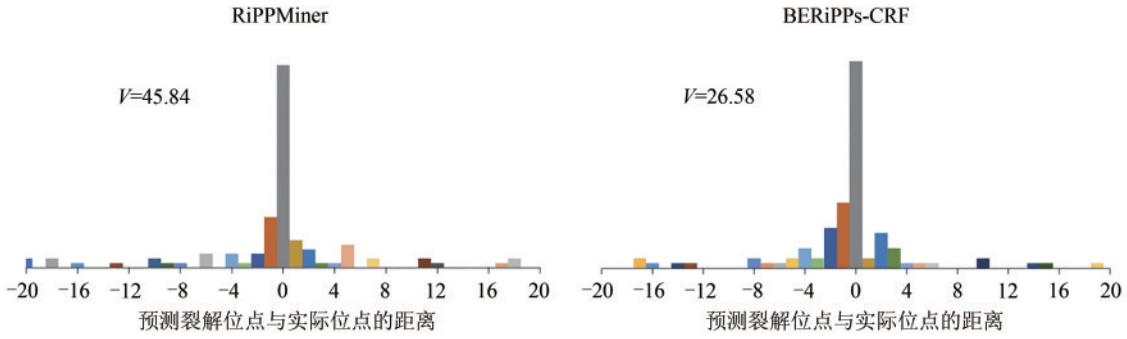
2.4 对BERiPPs模型挖掘RiPPs的性能评估

为了进一步验证基于预训练模型的深度学习方法在天然产物挖掘领域的可行性以及评估BERiPPs在处理未知RiPPs前体肽上的能力，本研

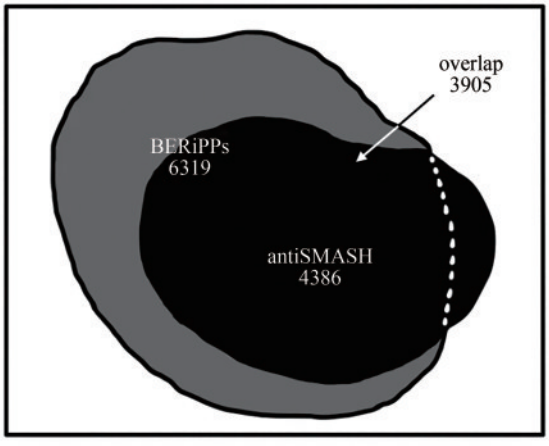
究于美国国家生物技术信息中心检索了1000个原核生物基因组，通过ORF分析工具找到ORF所在区域并翻译为相应的蛋白序列，将所有长度短于200个氨基酸的ORF作为BERiPPs的输入，共识别得到6319个RiPPs前体肽。同时，将上述1000个原核生物基因组通过antiSMASH进行全基因组分析，基于基因同源性找到了4386个RiPPs，其中3905个RiPPs所在区域与BERiPPs预测结果重合[图8(b)]。通过对比可以发现BERiPPs不仅识别到了89.03%的由antiSMASH进行全基因组分析得到的RiPPs，还将可能的未知RiPPs范围进一步扩大了60%左右，展现了在不考虑基因组背景的情况下仅基于候选前体肽序列挖掘全新RiPPs的潜力，并从一定程度上揭示了未知的新型RiPPs家族可能不完全适用于现有天然产物生物合成途径的规则。本文基于上述6319个由BERiPPs检测到的未经标注的RiPPs前体肽尝试对在通用域语料上预训练的BERT模型进行深入预训练，仅采用MLM任务且只对MLM计算损失，通过预测随机遮掩的氨基酸单元来进一步增加BERT在特定领域下抽取特征的能力，但这一点并没有在下游任务中得以体现。不难推测，实质性的提升依然需要大量的未标注样本。

3 讨论与展望

测序技术的快速发展极大促进了新型天然产物的挖掘进程，大量具有应用价值的化合物在各种生物信息学工具的帮助下得以发现。RiPPs以其独特的生物合成模式和生物活性而著称，研究人



(a) BERiPPs与RiPPMiner在预测裂解位点上的准确性对比
(a) Comparison of accuracy of BERiPPs and RiPPMiner in predicting cleavage sites



(b) BERiPPs与antiSMASH挖掘RiPPs结果对比
(b) Comparison of RiPPs mining results between BERiPPs and antiSMASH

图8 BERiPPs与RiPPMiner及antiSMASH对比

Fig. 8 Comparison of BERiPPs with RiPPMiner and antiSMASH

员通过不同的保守生物合成酶机制来针对性地挖掘不同类的RiPPs，然而这种基于同源性的经典基因组挖掘策略很大程度上依赖于与已知数据库的相似性，因此阻碍了它们靶向新家族的能力。随着对RiPPs的相关研究不断深入，研究人员发现在少数RiPPs生物合成途径中存在前体肽与修饰酶相距甚远的情况，例如van der Donk课题组在浮游海洋蓝藻中发现的prochlorosin，其部分前体肽与ProcM酶相距近1 Mbp[35]。显然对于与前者类似的RiPPs生物合成机制，依靠识别翻译后修饰酶附近前体肽的策略也是存在局限性的。因此，如何在不考虑基因组背景的情况下准确识别RiPPs前体肽是高通量挖掘新型RiPPs家族的最佳范式。本文通过在图像识别和自然语言处理领域取得巨大进展的深度学习来探究RiPPs生物合成的底层逻辑，提出基于BERT预训练模型的组合模型BERiPPs，仅根据候选前体ORFs即可较为准确地识别是否属于RiPPs，并且结合CRF后在预测RiPPs裂解位点任务上依旧取得了较好的效果，这意味着基于BERT的深度学习方法能够帮助发现未知生物合成机制的新型RiPPs。

更值得关注的是，BERiPPs也在某种程度上解释了RiPPs前体肽和修饰酶间的依赖关系。尽管在大多数认知中，修饰酶在某种程度上决定了最终产物的化学结构及生物活性，但仅基于RiPPs前体肽序列挖掘RiPPs的深度学习策略似乎证明了即使在没有修饰酶介入的情况下，RiPPs前体肽序列已经决定了其化学特性。例如，羊毛硫肽cypemycin的环状结构本质上是由前体肽序列所决定，而并非是有了脱水和成环的修饰酶才形成Avicys环，不过以上结论还需进一步实验验证。目前，因为深度学习模型对数据量的要求很高，基于通用域语料的预训练方式并不能完全克服这一难点，所以在RiPPs挖掘领域中基于深度学习的策略精度依旧略低于预定义规则的传统方法，这一点在样本数量极少的稀有RiPPs家族上表现得尤为明显。解决这一问题最直接的方法无疑是尽可能扩大RiPPs数据库以及在同领域或相关领域中的大规模样本上的训练模型，DeepMind团队发布的AlphaFold2[36]提供了一些参考，借鉴其思路可以在对RiPPs挖掘的研究中用已知的RiPPs样本先训练模型，再基于训练好的模型在未经过表征的未知氨基酸序列上进行预测并打分，保留其中预测得分高的样本生成新的大规模数据集，最终将两部分数据集混合后再次对模型进行训练，这种基于带噪声的自训练（self-training with noise student）方法[37]可能是未来人工智能运用于生化领域的主流，尽管这对计算机算力提出了一定要求。可以设想的是，未来生物学和计算机科学的跨学科交叉研究会极大推动基于RiPPs的药物发现，代谢组学、基因组学和深度学习的交互将为天物产物挖掘和生物合成研究提供新的思路。

参 考 文 献

[1] MARTENS E, DEMAIN A L. The antibiotic resistance crisis, with a focus on the United States[J]. The Journal of Antibiotics, 2017, 70(5): 520-526.

[2] HUTCHINGS M I, TRUMAN A W, WILKINSON B. Antibiotics: past, present and future[J]. Current Opinion in Microbiology, 2019, 51: 72-80.

[3] HUDSON G A, MITCHELL D A. RiPP antibiotics: Biosynthesis and engineering potential[J]. Current Opinion in Microbiology, 2018, 45: 61-69.

[4] WANG F T, WEI W Q, ZHAO J F, et al. Genome mining and biosynthesis study of a type B linaridin reveals a highly versatile α -*N*-methyltransferase[J]. CCS Chemistry, 2021, 3(3): 1049-1057.

[5] SKINNIDER M A, JOHNSTON C W, EDGAR R E, et al. Genomic charting of ribosomally synthesized natural product chemical space facilitates targeted mining[J]. Proceedings of the National Academy of Sciences of the United States of America, 2016, 113(42): E6343-E6351.

[6] ARNISON P G, BIBB M J, BIERBAUM G, et al. Ribosomally synthesized and post-translationally modified peptide natural products: overview and recommendations for a universal nomenclature[J]. Natural Product Reports, 2013, 30(1): 108-160.

[7] YU Y, ZHANG Q, VAN DER DONK W A. Insights into the evolution of lanthipeptide biosynthesis[J]. Protein Science, 2013, 22(11): 1478-1489.

[8] ZHONG Z, HE B B, LI J, et al. Challenges and advances in genome mining of ribosomally synthesized and post-translationally modified peptides (RiPPs)[J]. Synthetic and Systems Biotechnology, 2020, 5(3): 155-172.

[9] BLIN K, SHAW S, STEINKE K, et al. antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline[J]. Nucleic Acids Research, 2019, 47(W1): W81-W87.

[10] HETRICK K J, VAN DER DONK W A. Ribosomally synthesized and post-translationally modified peptide natural product discovery in the genomic era[J]. Current Opinion in Chemical Biology, 2017, 38: 36-44.

[11] HYATT D, CHEN G L, LOCASCIO P F, et al. Prodigal: prokaryotic gene recognition and translation initiation site identification[J]. BMC Bioinformatics, 2010, 11: 119.

[12] DELCHER A L, BRATKE K A, POWERS E C, et al. Identifying bacterial genes and endosymbiont DNA with Glimmer[J]. Bioinformatics, 2007, 23(6): 673-679.

[13] VAN HEEL A J, DE JONG A, MONTALBÁN-LÓPEZ M, et al. BAGEL3: automated identification of genes encoding bacteriocins and (non-) bactericidal posttranslationally modified peptides[J]. Nucleic Acids Research, 2013, 41(W1): W448-W453.

[14] TIETZ J I, SCHWALEN C J, PATEL P S, et al. A new genome-mining tool redefines the lasso peptide biosynthetic landscape[J]. Nature Chemical Biology, 2017, 13(5): 470-478.

[15] MERWIN N J, MOUSA W K, DEJONG C A, et al. DeepRiPP integrates multiomics data to automate discovery of novel ribosomally synthesized natural products[J]. Proceedings of the National Academy of Sciences of the United States of America, 2020, 117(1): 371-380.

[16] AGRAWAL P, KHATER S, GUPTA M, et al. RiPPMiner: a bioinformatics resource for deciphering chemical structures of RiPPs based on prediction of cleavage and cross-links[J]. Nucleic Acids Research, 2017, 45(W1): W80-W88.

[17] DE LOS SANTOS E L C. NeuRiPP: Neural network identification of RiPP precursor peptides[J]. Scientific Reports, 2019, 9: 13406.

[18] SHIN H C, ROTH H R, GAO M C, et al. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning[J]. IEEE Transactions on Medical Imaging, 2016, 35(5): 1285-1298.

[19] SUNDERMEYER M, SCHLÜTER R, NEY H. LSTM neural networks for language modeling[C]// 13th Annual conference of the International Speech Communication Association 2012 (INTERSPEECH 2012). Portland, OR, USA: International Speech Communications Association, 2012:194-197.

[20] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of Deep Bidirectional Transformers for Language Understanding[C]// Proceedings of NAACL-HLT. 2019: 4171-4186.

[21] TENNEY I, DAS D, PAVLICK E. BERT rediscovers the classical NLP pipeline[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy. Stroudsburg, PA, USA: Association for Computational Linguistics, 2019: 4593-4601.

[22] CHO K, VAN MERRIENBOER B, GULCEHRE C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar. Stroudsburg, PA, USA: Association for Computational Linguistics, 2014: 1724-1734.

[23] Bahdanau D, Cho K H, Bengio Y. Neural machine translation by jointly learning to align and translate[C]//3rd International Conference on Learning Representations, ICLR 2015. 2015.

[24] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in neural information processing systems. 2017: 5998-6008.

[25] SHERSTINSKY A. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network[J]. Physica D: Nonlinear Phenomena, 2020, 404: 132306.

[26] WANG Q, LI B, XIAO T, et al. Learning deep transformer models for machine translation[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy. Stroudsburg, PA, USA: Association for Computational Linguistics, 2019: 1810-1822.

[27] SÖDING J. Protein homology detection by HMM-HMM comparison[J]. Bioinformatics, 2005, 21(7): 951-960.

[28] LIU L Y, REN X, SHANG J B, et al. Efficient contextualized representation: Language model pruning for sequence labeling [C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium. Stroudsburg, PA, USA: Association for Computational Linguistics, 2018: 1215-1225.

[29] Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging[EB/OL]. arXiv preprint: 2015, arXiv: 1508.01991. https://doi.org/10.48550/arXiv.1508.01991

[30] ZHAO H S, JIA J Y, KOLTUN V. Exploring self-attention for image recognition[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2020, 10073-10082.

[31] Dodge J, Ilharco G, Schwartz R, et al. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping[EB/OL]. arXiv preprint: 2020, arXiv: 2002.06305. https://doi.org/10.48550/arXiv.2002.06305

[32] YADAV S, SHUKLA S. Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification[C]//2016 IEEE 6th International Conference on Advanced Computing. Bhimavaram, India: IEEE, 2016: 78-83.

[33] ZHANG Z L, SABUNCU M. Generalized cross entropy loss for training deep neural networks with noisy labels[J]. Montréal: NeurIPS, 2018, 31.

[34] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[C]//2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017: 2999-3007.

[35] LI B, SHER D, KELLY L, et al. Catalytic promiscuity in the biosynthesis of cyclic peptide secondary metabolites in planktonic marine cyanobacteria[J]. Proceedings of the National Academy of Sciences of the United States of America, 2010, 107(23): 10430-10435.

[36] JUMPER J, EVANS R, PRITZEL A, et al. Highly accurate protein structure prediction with AlphaFold[J]. Nature, 2021, 596 (7873): 583-589.

[37] XIE Q Z, LUONG M T, HOVY E, et al. Self-training with noisy student improves ImageNet classification[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE,2020 : 10684-10695.

通讯作者：丁伟（1981—），男，博士，副教授，博士生导师。研究方向为微生物代谢及合成生物学。

E-mail：weiding@sjtu.edu.cn

第一作者：吕靖伟（1996—），硕士研究生。研究方向为天然产物合成基因挖掘。

E-mail：jingwei_lv@sjtu.edu.cn