摘　要

探究蛋白质结构与功能之间的关系是现代生物信息学领域中最为关键的课题之一，对工农业生产和生物医药发展等方面都有着十分重大的价值。由于新基因组时期的出现，蛋白质的信息量迅速增加，如贝叶斯、K 近邻和支撑向量机（Support Vector Machines, SVM）等技术都已不那么适合，基于神经网络的机器学习模型在大数据分析的新形势下有了良好的发展。

本文针对蛋白质 8 类二级结构预测提出了一种 Deep-BGRU 模型，通过多层双向门控循环单元系统（Bidirectional Gated Recurrent unit, BGRU）来深入获取氨基酸序列的全局信息，进而利用 Softmax 分类器实现蛋白质 8 类二级结构预测。较之当前应用广泛的长短期记忆神经网络（Long Short-term Memory, LSTM）模型和基于 LSTM 构成的融合模型，Deep-BGRU 模型除了改善传统的循环神经网络（Recurrent Neural Network, RNN）梯度消失无法处理极长依赖性的问题，还在蛋白质二级结构预测速度和预测结果精度上有着明显的提升。实验结果表明，Deep-BGRU 模型在基准数据集 CB513 上的 Q8 准确度达到了 70.6%。与其他模型方法相比，本文提出的模型能够很好地提高蛋白质 8 类二级结构的预测精度，具有很好的可扩展性和较低的训练成本。

由于氨基酸序列编码方式对搭建的蛋白质二级结构预测模型准确度有较大影响，本文设计了相关对比实验，研究了该领域常用的三种氨基酸编码方式对预测模型准确度的影响。实验结果表明，将进行了独热编码和 Profile 轮廓编码的氨基酸序列组合输入时，预测模型的准确度较好。

随着蛋白质数据库的完善和高通量技术的发展，基于深度学习的蛋白质二级结构预测算法得到了广泛研究，许多蛋白质结构预测平台应运而生。现有的大多数预测平台只能进行 3 类二级结构预测，不能进行预测算法的选择，基于此，本文设计和实现了基于 Vue 和 SpringBoot 的蛋白质二级结构预测算法平台 PSP（Protein Structure Prediction, PSP）。该平台内置了若干性能优良的基于深度学习的 8 类/3 类蛋白质二级结构预测模型供用户选择使用，包括 DeepACLSTM、DeepCNF、PSRSM、融合卷积神经网络和贝叶斯优化等模型以及本文提出的 Deep-BGRU 预测模型。PSP 平台采用 ECharts 可视化技术对预测结果进行可视化展示。用户还可参考平台内置模型和公共数据集来构建自己的预测模型。此外，PSP 平台中集成的算法种类可以更新，具有良好的可扩展性。

关键词：蛋白质二级结构；Q8；深度学习；预测平台；GRU

Abstract

Investigating the relationship between protein structure and function is one of the most critical topics in modern bioinformatics, and is of great value for industrial and agricultural production and biomedical development. Due to the emergence of the new genomic period, the amount of information about proteins has increased rapidly, and techniques such as Bayesian, K-nearest neighbor, and support vector machines (SVM) are no longer so suitable, and neural network-based machine learning models have developed well in the new situation of big data analysis.

In this dissertation, we propose a Deep-BGRU model for protein class 8 secondary structure prediction, which uses a multilayer bidirectional gated recurrent unit (BGRU) system to obtain in-depth global information of amino acid sequences, and then uses a Softmax classifier to achieve protein class 8 secondary structure prediction. Compared with the long short-term memory (LSTM) model and LSTM-based fusion models, the Deep-BGRU model not only improves the traditional recurrent neural network (RNN) gradient disappearance problem but also improves the speed of protein secondary structure prediction. The Deep-BGRU model not only improves the problem that the conventional recurrent neural network (RNN) gradient disappearance cannot handle very long dependencies but also significantly improves the speed and accuracy of protein secondary structure prediction. The experimental results show that the Deep-BGRU model achieves a Q8 accuracy of 70.6% on the benchmark dataset CB513. Compared with other modeling approaches, the model proposed in this dissertation can well improve the prediction accuracy of protein class 8 secondary structure, with good scalability and low training cost.

Since the amino acid sequence coding methods significantly impact the accuracy of the constructed protein secondary structure prediction models, this dissertation designed a relevant comparison experiment to investigate the impact of three amino acid coding methods commonly used in this field on the accuracy of the prediction models. The experimental results show that the accuracy of the model is better when a combination of amino acid sequences that have undergone unique thermal coding and Profile profile coding are fed into the prediction model.

With the improvement of protein databases and the development of high-throughput technologies, protein secondary structure prediction algorithms based on deep learning have been widely studied, and many protein structure prediction platforms have emerged. Most of the existing prediction platforms can only perform three types

of secondary structure prediction and cannot perform the selection of prediction algorithms. Based on this, this dissertation designs and implements PSP (Protein Structure Prediction, PSP), a protein secondary structure prediction algorithm platform based on Vue and SpringBoot. The PSP platform is built with several high-performance deep learning-based 8 classes/3 classes of protein secondary structure prediction models for users to choose from, including DeepACLSTM, DeepCNF, PSRSM, fused convolutional neural network, and Bayesian optimization models, as well as the Deep-BGRU prediction model proposed in this dissertation. The prediction results are displayed visually. Users can also build their prediction models by referring to the platform's built-in models and public datasets. In addition, the variety of algorithms integrated into the PSP platform can be updated and is highly scalable.

Key words: Protein secondary structure; Q8; deep learning; prediction platform; GRU

目　　录

1 绪论

1.1 研究背景及意义

随着多种生物基因测序技术的发展，全球生物数据库已经拥有了大量的基因和蛋白质序列信息[1]。为了深入探索这些信息背后的机制，科学家们开发出了多种数据分析方法，以便更好地理解生物信息学的内涵[2]。研究蛋白质结构和功能是生物信息学的核心课题，蛋白质的功能取决于它的组成，因此，深入探索蛋白质结构和功能，将有助于更好地理解生物体的运作机制。

蛋白质是生命系统中用途最广泛的大分子，在生物发展过程中发挥着重要作用。研究生物大分子蛋白质对于探索生物功能[3]至关重要，它们不仅可以充当催化剂、提供机械支持和免疫保护，还可以控制生长和分化等过程。高级结构是蛋白质功能多样性的基础，它也被称为蛋白质的空间构象，是蛋白质结构的重要组成部分。蛋白质是生物体中最重要的组成部分，它们的功能主要取决于它们的结构、运动和相互作用[4]。

深入了解相关蛋白质、复合物和组装体的高级结构以及它们之间的功能关系，是理解生命科学问题的关键所在。只有在分子和原子水平上对这些结构和关系有深入的了解，才能够解释生命体的增殖、分化和凋亡等现象。因此，研究这些结构和关系对于生命科学的发展至关重要。

蛋白质的二级结构预测是预测三级结构的基础，由三维结构的圈面所产生的构像，将决定其生物结构[5]。现实中，通过蛋白质的一级结构很难直接判断蛋白质的三级结构，而利用二级结构来判断三级结构进而判断蛋白质的作用则相对简单。所以，明确蛋白质的二级结构对于人体内蛋白质复合物的分析及其医学病变的防治和诊断具有重要意义。

此外，实验也表明无序蛋白质缺乏明确的三维特征，只是局部的二次构造，但仍具备生物特性[6]，所以通过蛋白质的二级结构在某种程度上就可确定其生物特性，这也突显出了蛋白质二级结构预测在生物信息方面有重要意义。

早期，蛋白质二级结构的预测研究主要集中在螺旋（H）、折叠（E）和卷曲（C）三种类型。但随着蛋白质领域研究的进步，现在已经将二级结构划分为 8 类，分别是3_{10}螺旋（G），α-螺旋（H），π-螺旋（H），β-桥（B），β-折叠（E），转角（T），弯曲（S）以及环状（L）[7]。

这两种蛋白质二级结构的关系如图 1-1 所示。

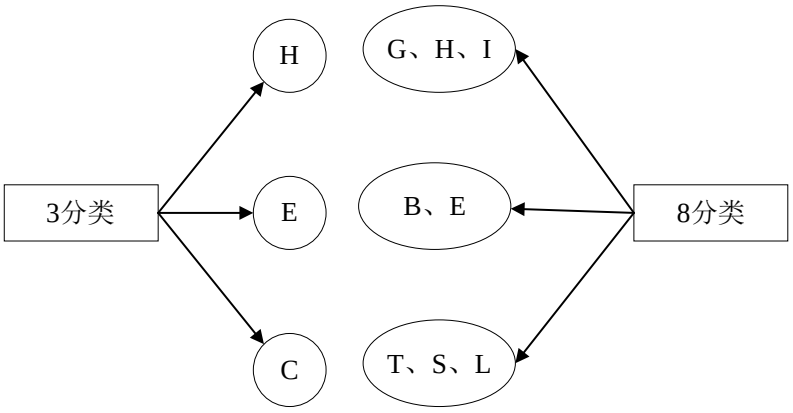


图 1-1　蛋白质二级结构类型

蛋白质二级结构预测包括 3 类和 8 类二级结构预测，其中 8 类是对 3 类的细化，能够提供更多且更加详细的信息，更为人们所需要，但是这使得预测更具有困难性。本文主要研究 8 类二级结构预测，如图 1-2 所示。

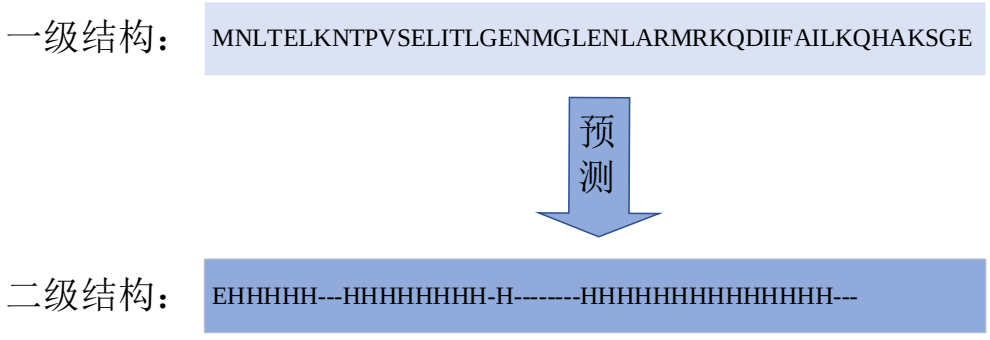


图 1-2　蛋白质二级结构预测示例

1.2 国内外研究现状

1.2.1 预测算法

对蛋白质二级结构的预测研究始于 1951 年，甚至在第一个蛋白质二级结构确定之前就有了关于其多肽主链的螺旋和折叠构象的预测[8]。蛋白质二级结构预测迄今为止历经三代的发展，第一代二级结构预测依赖分析单个氨基酸序列信息对二级结构的影响从而进行规律的统计分析，并根据手工分析的规则预测二级结构[9]，但这些方法在新的蛋白质结构上准确率通常不超过 60%[10]；二代二级结构预测提取氨基酸序列的多个信息，使用前馈神经网络（Feed Forward Neural Networks, FFNN）[11]、最小二乘法等回归分析方法对提取的信息进行分析来预测二级结构，相比第一代方法，该代方法准确率提升至 63-64%；第三代二级结构预测的特点是采用进化信息[12]，将多个能代表进化信息同源序列的比对形式输入基

于机器学习或深度学习的算法模型，该代早期算法中的 PHD 系统所使用的算法[12]关于二级结构预测的准确率首次突破了 70%。随着计算机性能的提高以及数据规模的激增，深度学习技术取得了巨大的进步，目前的蛋白质二级结构预测算法均在第三代的工作基础上改进而来，其中 3 类二级结构预测准确率达到 85%左右[13]，8 类二级结构预测准确率达到 70%左右[14]。

在第三代蛋白质二级预测方法中，蛋白质 8 类二级结构预测模型研究工作基本上是使用深度学习模型开展的。Gianluca Pollastri 等[15]采用双向朴素循环神经网络（B-RNN）提出了 SSpro8 蛋白质二级结构预测模型。Lin 等[16]利用蛋白质氨基酸序列的位置特异性打分矩阵（Position-specific Scoring Matrix, PSSM）编码和正交编码，结合卷积神经网络（Convolutional Neural Networks, CNN）和 LSTM 模型，建立 8 类蛋白质二级结构的预测模型。Wang 等人[17]基于氨基酸 PSSM 编码，利用条件神经域（Conditional Neural Fields, CNF）建立了 8 类蛋白质二级结构的预测模型。2016 年，Wang 等人进一步将深度学习技术应用于条件神经域，提出了 DeepCNF[18]。Zhou 等人[19]采用特殊位置重复 BLAST 工具（Position-Specific Iterated Basic Local Alignment Search Tool, PSI-BLAST）和 PSSM 对氨基酸序列完成编码，并使用基于监督学习的卷积生成随机网络（Deep Con-volutional Generative Stochastic Networks, DCGSN）进行蛋白质 8 类二级结构预测。Zhang Lei 等[20]采用深度双向 LSTM 构建模型改善了传统循环神经网络（Recurrent Neural Network, RNN）的梯度消失问题。Guo Yanbu 等[21]提出卷积长短时记忆神经网络，考虑到氨基酸序列内部残基之间的局部相关特征和远程相互作用，同时于 2019 年，Guo Yanbu 等人又提出了基于非对称卷积神经网络（ACNN）和双向 LSTM 的 DeepACLSTM[14]。

对于 8 类蛋白质二级结构的预测，氨基酸序列的长度有时会达到 700 多条，RNN 本身的梯度消失而导致其无法处理长期依赖问题的缺点就会很明显，因此 LSTM 比 RNN 更适合作为蛋白质二级结构基础模型方法。现有大家认可和使用的蛋白质数据集不是很大，相比 LSTM，GRU 作为一种变种模型，具有参数较少、训练速度较快、需要的泛化数据也较少等优点，而且在部分任务中 GRU 的表现比 LSTM 更好，因此 GRU 可能更适合于蛋白质 8 类二级结构的预测。

1.2.2 预测平台

随着神经网络的应用与发展，基于传统机器学习或深度学习的预测平台应运而生。国外广泛使用的蛋白质二级结构预测平台是 PSIPRED、JPred、和 Predict Protein。其中，PSIPRED（http://bioinf.cs.ucl.ac.uk/psipred）V.4 版本内置基于位置特异性评分矩阵的蛋白质二级结构预测方法使得其 Q3 得分可达 84.2%[22]；JP

red4（http://www.compbio.dundee.ac.uk/jpred/index.html）采用多序列比对图谱技术使得其 Q3（蛋白质 3 类二级结构预测准确度）从 JNet v.2.0 的 81.5% 提高至 82.0%[23]；PredictProtein（https://predictprotein.org/）采用自监督深度学习和语言计算的 Q3 得分为 82%[24]。国内的蛋白质二级结构预测平台数量极少，且大多为企业开发的附带产品。其中，德泰生物的蛋白质 3 类二级结构在线预测工具（http://www.detaibio.com/tools/chou-fasman-forecast.html）采用基于单个氨基酸残基统计 Chou-Fasman 方法[25]；纽普生物的预测工具（https://www.novopro.cn/tools/secondary-structure-prediction.html）的预测方法与 PSIPRED 采用的方法[23]相同；齐鲁工业大学推出的预测平台 QiluBio（http://210.44.144.20:82/protein_PSRSM/default.aspx）采用基于数据分区和半随机子空间方法的深度学习算法，其 Q3 得分高达 85%[13]。

上述平台基本上都没有实现蛋白质 8 类二级结构预测的功能，这使得最近提出的先进的蛋白质 8 类二级结构预测算法无法在平台中得到应用。此外，近年基于深度学习的预测算法实现了性能突破，而部分平台采用的仍为准确率较低的传统算法。针对这一问题，本文设计 PSP 平台实现蛋白质 8 类二级结构的预测功能，并集成五个性能优良的基于深度学习的二级结构预测算法。

1.3 论文主要研究内容

现有的蛋白质 8 类二级结构预测模型在一定程度上提高了蛋白质 8 类二级结构预测的准确率，但均基于模型叠加的方法实现，这增加了模型的复杂度以及模型的训练成本。同时现存的大多数蛋白质二级结构预测平台只能进行 3 类二级结构预测且预测时不能进行预测算法的选择。

针对上述问题，本文开展了基于深度学习的蛋白质二级结构预测模型及其平台研究，主要工作如下：

（1）提出了一种新型的基于深度双向门控循环单元网络（Bidirectional Gated Recurrent Unit, BGRU）的 Deep-GRU 蛋白质 8 类二级结构结构预测模型。该模型较之传统模型在预测结果和预测速度上都有很好提升，由于使用了 LSTM 变体 GRU 来构建模型，所以预测模型结构简单且易于扩展。

（2）研究氨基酸序列的编码方式对预测结果的影响。氨基酸序列编码方式有独热编码（one-hot 编码）、基团编码、PSSM 编码等。研究单种编码方式对模型准确度影响的同时还研究了多种编码方式组合情况下的编码方式对实验模型准确度的影响。

（3）设计并实现了基于 Vue 前端框架和 SpringBoot 后端框架技术的 PSP 平

台，以便能将 Deep-GRU 模型以及蛋白质二级结构预测领域具有代表性的模型进行应用。

1.4 论文组织结构

本文采用如下的章节结构进行阐述：

第一章绪论。简要探讨了蛋白质二级结构预测研究的背景和国内外研究现状，概述了本文的主要研究内容。

第二章相关理论和方法。介绍了深度学习中的 CNN、RNN、LSTM 等一些热门模型方法和开发预测平台所用的技术框架。

第三章研究氨基酸序列的编码方式对预测结果的影响，以及多种编码方式组合的情况对实验结果的影响，并设计了对比实验。

第四章深入研究了 8 类蛋白质二级结构预测模型，实现了一种高效的预测模型 Deep-BGRU，并与领域内其它算法做了对比。

第五章蛋白质二级结构预测算法平台研究。借助 Deep-BGRU 预测模型与现有模型，运用 Keras 人工神经网络库、SpringBoot 后端技术、Vue 前端技术和微信小程序技术实现了可扩充算法模型的蛋白质二级结构预测算法平台。

第六章对本文的研究工作进行了全面的概括，对下一步的工作进行了展望。

2 相关理论和方法

2.1 蛋白质简介

2.1.1 蛋白质基础介绍

蛋白质是构成细胞与生命体的基本有机物，参与了生命的所有过程，如遗传信息的表达、代谢反应的调节、机体的防御和生化反应的催化，神经信号的传递等[26]。蛋白质是一类复杂的有机物，由碳、氢、氧和氮组成。它们也含有其他元素，如硫、磷、铁和碘等其他成分。蛋白质是人体最重要的氮来源之一，与碳水化合物和脂肪相比，它们的最大特点是含氮量更高[27]。

蛋白质在人体中具有重要的地位，是构成人体机体组织的主要成分之一。人体内的各种组织器官都含有蛋白质，而在细胞内，除了水分外，蛋白质也占据了细胞内物质的约 80%[28]。蛋白质在人体内发挥着多种重要的作用，例如：构成机体细胞并与其他物质共同作用；维持机体正常的新陈代谢和不同物质在体内的运输，如血红蛋白可以运输氧气，载脂蛋白可以输送脂肪；构成许多必需酶，参与生化反应；具有激素调节作用，可以调节不同器官的生理活性，如胰岛素可以降低血糖；提供生命活动所需的能量；作为抗体参与体内免疫反应；从出生到死亡，蛋白质都在人类整个生命的代谢过程中发挥重要作用。因此蛋白质是生命的物质基础，没有蛋白质就没有生命。同样也可以通过对蛋白质的研究来揭露生命的密码，给医学、生物学和生物信息学等领域提供帮助。

2.1.2 蛋白质结构分类

美国化学家鲍林（Linus Pauling）是一位卓越的科学家，他的成就包括两次诺贝尔奖的获得。在 1954 年，他的研究团队发现了蛋白质的 α 螺旋结构，这一发现对于蛋白质的结构研究产生了深远的影响，确定蛋白质中存在空间折叠结构[29]。诺贝尔化学奖获得者佩鲁茨（M. F. Perutz）和肯德鲁（J. C. Kendrew）在 1959 年运用 X 射线衍射科学技术对血红蛋白和肌血蛋白开展了深入的研究，他们的研究成果[30]为蛋白质的三维空间结构提供了重要的理论支持，并证实了蛋白质的功能与结构之间存在着密切的联系。现今已经证实蛋白质是由 20 种不同的氨基酸组成的多肽链经过紧密折叠所构成[31]，且除了少部分无序蛋白，大多数蛋白质的功能是由其三维结构所决定的。

蛋白质的结构可以划分为一级、二级、三级和四级，其中一级结构由氨基酸组成，而不同的氨基酸序列具有不同的功能和特性。α-螺旋和 β-折叠片为蛋白质中不同的局部空间结构，这也是蛋白质结构中的二级结构。三级结构为多个二级

结构组成的整体空间结构。不同的是四级结构为多个蛋白质分子相互作用而形成的超大分子结构。对于蛋白质的特有性质，一般来说是根据它特有的空间结构决定的，其中空间结构由蛋白质分子中每个原子在空间中的位置而构成。需要注意的是，有些蛋白质不具备四级结构，比如单条肽链形成的蛋白质。

（一）蛋白质的第一级结构的决定因素是由氨基酸的排列顺序。理解蛋白质的结构、作用机制和生理功能的关键基础是了解蛋白质的第一级结构。氨基酸的结构由一个氨基、一个羧基、一个氢和一个 R 基连接在同一个 C 原子上，肽键是氨基酸序列中主要的化学键[32]。表格 2-1 中列举了 20 种常见氨基酸的名称和主要的物理化学性质。

表 2-1　20 种氨基酸及其理化性质

名称	英文名	缩写	支链	极性	带电性	酸碱性	亲水性	分子量
甘氨酸	Gly	G	aliphatic	nonpolar	中性	中性	-0.4	75.07
丙氨酸	Ala	A	aliphatic	nonpolar	中性	中性	1.8	89.09
缬氨酸	Val	V	aliphatic	nonpolar	中性	中性	4.2	117.15
亮氨酸	Leu	L	aliphatic	nonpolar	中性	中性	3.8	131.17
异亮氨酸	Iie	I	aliphatic	nonpolar	中性	中性	4.5	131.17
苯丙氨酸	Phe	F	aromatic	nonpolar	中性	中性	2.8	165.19
色氨酸	Trp	W	aromatic	nonpolar	中性	中性	-0.9	204.23
酪氨酸	Tyr	Y	aromatic	polar	中性	中性	-1.3	181.19
天冬氨酸	Asp	D	acid	acidic polar	负电	酸性	-3.5	133.1
天冬酰胺	Asn	N	amide	polar	中性	中性	-3.5	132.12
谷氨酸	Glu	E	acid	acidic polar	负电	酸性	-3.5	147.13
赖氨酸	Lys	K	basic	basic polar	正电	碱性	-3.9	146.19
谷氨酰胺	Gln	Q	amide	polar	中性	中性	-3.5	146.15
甲硫氨酸	Met	M	sulfur-containing	nonpolar	正电	中性	1.9	149.21
丝氨酸	Ser	S	hydroxyl-containing	polar	中性	中性	-0.8	105.09
苏氨酸	Thr	T	hydroxyl-containing	polar	中性	中性	-0.7	119.12
半胱氨酸	Cys	C	sulfur-containing	nonpolar	中性	中性	2.5	121.16
脯氨酸	Pro	P	cyclic	nonpolar	中性	中性	-1.6	115.13
组氨酸	His	H	basic aromatic	basic polar	正电	碱性	-3.2	155.16
精氨酸	Arg	R	basic	basic polar	正电	碱性	-4.5	174.2

（二）蛋白质分子中的一段肽链的局部空间结构为二级结构。常见的 α-螺旋（α-helix）、β-折叠（β-sheet）、β 转角（β-turn）等都是蛋白质的二级机构。蛋白质的二级结构的主要化学键为氢键。由于蛋白质的二级结构为一段肽链的局部空间结构，所以在蛋白质分子中可以有多个二级结构，并且在相近的两个或以上的二级结构可以合作完成任务，这称为模体。在最初的时候，蛋白质二级结构只分为螺旋（H），折叠（E）和卷曲（C）三种，随着对蛋白质结构研究的深入，蛋白质的二级结构也从最开始的 3 类变成了 8 类，8 类二级结构分别是3_{10}螺旋（G），α-螺旋（H），π-螺旋（I），β-桥（B），β-折叠（E），转角（T），弯曲（S），环状（L）。

（三）肽链中所有氨基酸残基在空间中的相对位置关系为蛋白质的三级结构，也可以理解为肽链的折叠关系。在三级结构中，主要依赖于一些次级键，像盐桥、氢键和范德华力等，这些都是蛋白质三级结构稳定的因素。通常情况下，如果一个蛋白质的分子量比较大，那么它可以折叠多个肽链形成结构稳定的区域，这些区域各自具有不同的功能，称为结构域，结构域是蛋白质功能的基本单位，一个蛋白质可能包含一个或多个结构域。

（四）蛋白质分子中的多个亚基的空间排布和相互作用为蛋白质的四级结构，它是蛋白质功能和稳定性的关键因素。四级结构是通过亚基与亚基之间通过共价键相连接并呈特定的空间排布形成的，亚基是多肽链的完整的三级结构。

蛋白质结构可以分为基础结构和高级结构两部分，其中一级结构被认为是结构基础，而二级、三级和四级结构则被称为高级结构。一级结构是其他结构的基础，而高级结构则是蛋白质功能多样性的结构基础。这四种结构关系如图 2-1 所示。

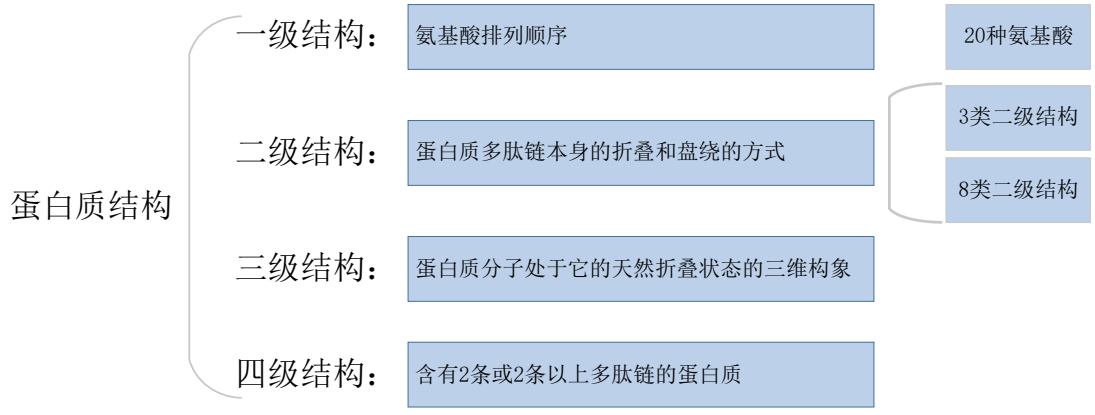


图 2-1　蛋白质四种结构关系图

2.2 深度学习预测方法

深度学习（Deep Learning），是一个多层神经网络的机器学习方法，它表示着神经网络从浅层到深层的发展。在深度学习出现之前，最早的神经网络基础是感知机，只有输入层和输出层两层，只能做正负判别，且不能做很复杂的判断，具有局限性且准确性不高。随后 1986 年，Rumelhart 和 McCelland 等人提出了 BP 神经网络，通过信号的向前传播和误差的向后传播从而更新权重矩阵。BP 神经网络一般为三层，其中包括输入层、中间隐藏层以及最后的输出层。BP 神经网络出现后也引起了不小轰动，它可以在大规模的数据样本中获得其规律，从而对未知事件做出预测，作为浅层神经网络的经典其在数据分析和预测领域都取得了不俗的成绩[33]。在 BP 神经网络中，如果想要获取更加贴合的拟合，那么就需要增加隐藏层的数量，但是这就产生了梯度扩散和局部最优解的问题，这也就导致了 BP 神经网络无法处理时间序列类似的问题。由于存在上面提到的问题，在深度神经网络被提出之前，没法对具有四层或更多层的深度神经网络进行充分的训练，并且训练出来的其性能也不佳。

2006 年，Hinton[34]和 Salakhutdinov 提出了通过人为干预的方式在添加新的隐含层和优化网络参数之间交替优化来逐渐增加神经网络深度，缓解了局部最优解问题，使得神经网络可以逐层学习[35]，深度学习因此才具有了真正的“深度”，这也被视为深度学习的开创性工作，为学习更深层的网络铺平了道路。庞大的数据集使人们无法直接优化整体性能，取而代之的是通过随机优化来训练深度神经网络。另外，为了改善优化性能，有许多常见的优化方法（例如 Adam，Adagrad，RMSprop）作为随机梯度下降算法的变体，通常与步长的自适应调节配合使用。

深度学习的目的是创建神经网络进行分析学习，这种网络可以像人类大脑一样处理各种数据，例如图像、声音和文本等，同时也能够用人类大脑的方式来解释这些数据。相对于传统机器学习算法而言，深度学习的优势在于其能够发现高维数据中的复杂结构，并且其精度远远超过传统的机器学习方法。

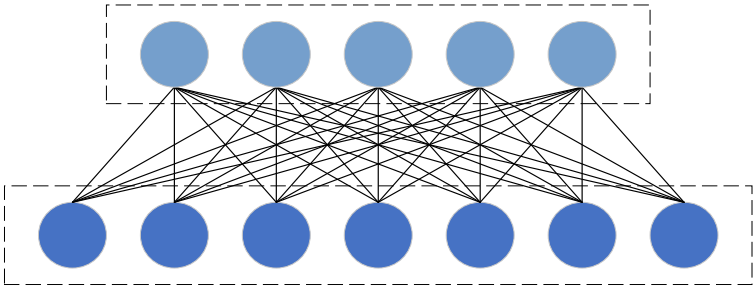
另外，经典的机器学习算法通常需要复杂的特征工程，而深度网络只需直接对原始数据进行处理，通常就可以实现良好的性能。

深度学习还具有适应性强，易于转换的优势，该技术可以更容易地适应不同的领域和应用。其中的迁移学习方法能够使事先训练的深度网络能够在各种各样的场合和应用中发挥最大的作用，从而提高系统的效率和准确性。例如，以语音识别领域的深度学习理论为基础，学习如何将深度神经网络应用于自然语言处理的难度就会有所降低，因为这些领域的基础知识非常相似。

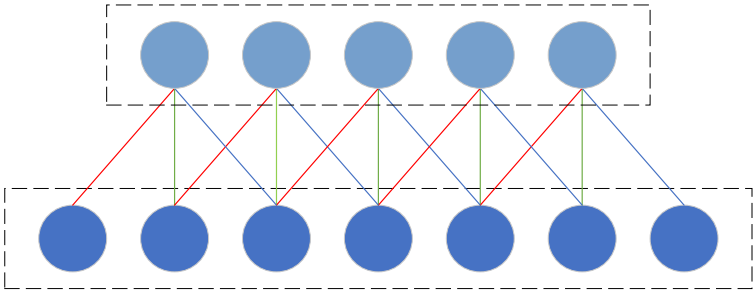
2.2.1 卷积神经网络 CNN

上世纪 60 年代，Hubel 和 Wiesel[36]通过对猫视觉皮层细胞的研究，提出了感受野（Receptive Field）这个概念，它指的是卷积神经网络每一层输出的特征图上的像素点在原始输入图像上映射的区域大小。到 80 年代，Fukushima 在感受野概念的基础之上提出了神经认知机的概念[37]，这可以看作是卷积神经网络（Convolutional Neural Networks, CNN）的雏形，神经认知机将一个视觉模式分解成若干个子模式，并将它们加入分层递阶式连接的特征水平面进行计算，以此来模型化视觉网络，这使得模型不但能够识别静止的物体，甚至还能识别运动中的物体。80 时代末，LeCun[38]结合反向传播算法与权值共享给出了首个经典卷积神经网络模型 LeNet-5，并再次提高手写字符识别的正确率，成功将其应用到美国邮局的手写字符识别系统中。

CNN 是一个独特的神经网络，它由卷积层和池化层组成，可以从大量输入数据信息中抽取出有用的特征信息，这与普通神经网络有着显著的不同。在普通神经网络中，各个神经元都与下层级的全部神经元相连接，而在 CNN 中，各个神经元只与其邻层的部分神经元相连接，如图 2-2 所示。这样可以有效地降低网络的复杂度，同时也可以从输入中抽取出有用的特征信息，且不会让模型变得过于庞大，能够降低模型的复杂度和训练成本，这使得 CNN 能够活跃在图像处理领域。



（a）全连接层



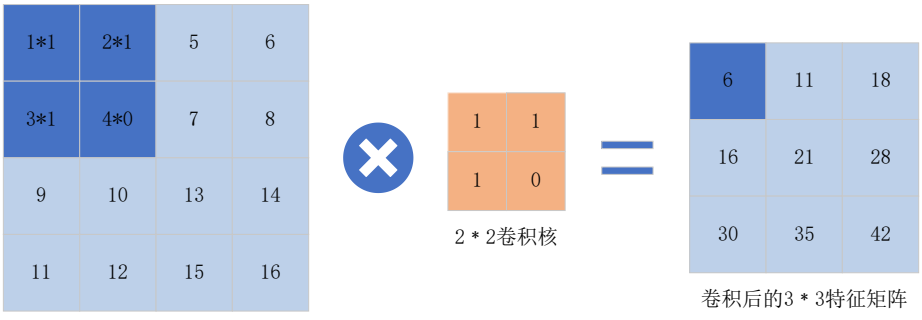
（b）卷积层

图 2-2　全连接层和卷积层的神经元连接图

CNN 卷积神经网络通常由输入层、卷积层、池化层、全连接层和输出层组

成。在卷积层中，通常包含多个特征图（Feature Map），各个特征图由若干神经元构成，同一特征图的神经元共用一组权值，即为卷积核。卷积核以随机小数矩阵的形式初始化，在网络的训练过程中，将学习到正确的权值。通过使用卷积核，可以有效地减少网络各层之间的联系，同时也大大降低了过拟合的可能性。根据卷积过程可知，第j单元的输出值a_j可通过公式（1）计算出来，公式中M_j代表选定的输入特征图的集合，k_{ij}代表可学习的卷积核，f函数是激励函数，a_i是输入，而b_j则是激励函数中的参数。图 2-3 展示了卷积层的具体计算过程。

$$a_j = f\left(\sum_{i \in M_j} a_i \times k_{ij} + b_j\right) \text{公式（1）}$$



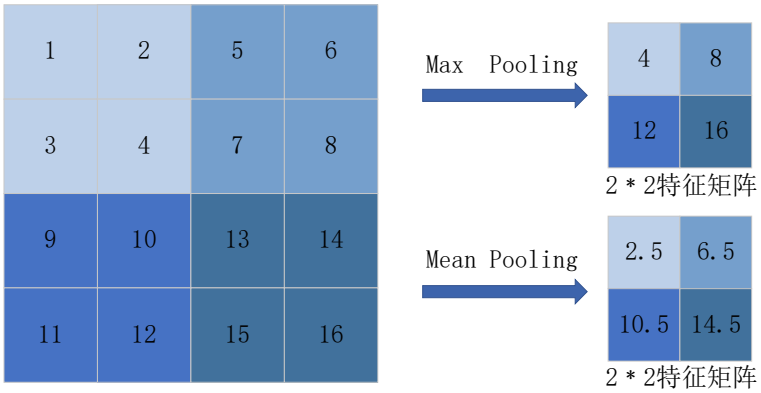
输入的4 * 4特征矩阵　　2 * 2卷积核　　卷积后的3 * 3特征矩阵

$$a_j = Rrlu\left(a_i \times k_{ij} + b_j\right)$$

$$6 = Rrlu(1 \times 1 + 2 \times 1 + 3 \times 1 + 4 \times 0)$$

图 2-3　卷积层卷积操作示意图

卷积层之后的特征信息会经过池化层（子采样层），池化层有均值池化（Mean Pooling）或最大值池化（Max Pooling），两种形式的池化层如图 2-4 所示。



4 * 4特征矩阵　　2 * 2特征矩阵

2 * 2特征矩阵

图 2-4　两种池化方式计算过程

通常输入层的数据经过卷积层之后得到的特征维度会很大，通过取其最大值或平均值，池化层可以将输入的特征降维，得到新的特征。池化操作是一种常见的深度学习操作，它可以被视为一种特殊的卷积过程。和卷积操作一样，池化操

作可以减少模型的复杂度和参数数量。除此之外，池化操作还有其他优势，例如可以防止过拟合、保留主要特征、间接增大感受野等，从而最终改善模型的性能。

最后将通过将池化层中得到的局部特征信息与全连接层结合，得到完整的特征信息，以统计每个类型的分数。CNN 的全链接层结构与多层感知机（Multilayer Perceptron, MLP）相似，因此 CNN 的训练算法也多使用 BP 算法[39]，为了进一步提高 CNN 网络的稳定性，全连接层的激励函数通常使用$Relu$函数，以实现更高效的特征信息获取。

2.2.2 循环神经网络 RNN

神经网络是一种能够拟合任意函数的模型，主要由输入层、多个隐藏层和一个输出层构成，其网络结构如图 2-5 所示。经过训练后，神经网络模型可以在给定输入层的 x 后，通过网络计算得到对应的输出层 y。因此，神经网络可以被看作是一个黑盒子，通过学习数据的模式，实现输入和输出之间的映射关系。

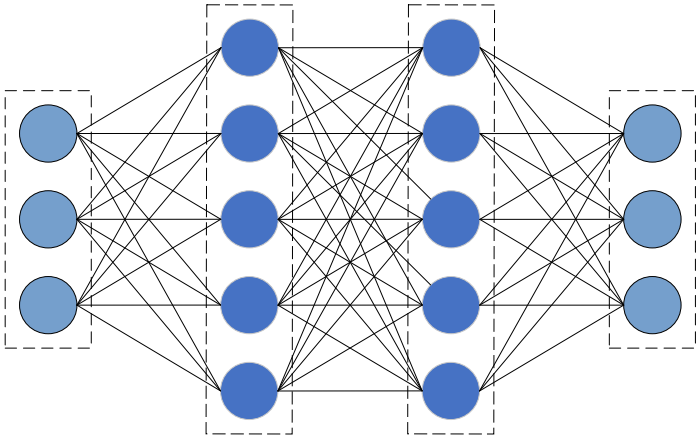


图 2-5　简单神经网络结构图

传统的神经网络通常只能处理一个输入，而且前后输入之间没有直接的联系或关系[40]。但是现实生活中许多问题的输出结果不但依赖当前的输入，还依赖于历史时期的输入，即前面的输入和后面的输入是有关系的。如：自然语言处中的文本翻译、情感分析、股票走势预测和温度变化趋势预测等。于是 20 世纪 80 年代，循环神经网络（Recurrent Neural Network, RNN）被提出，并在 21 世纪初逐渐发展成为深度学习的一个重要神经网络结构。RNN 是一种基于序列数据建模的人工神经网络，所谓的序列数据就是指这一些列的数据中前后的输入数据是有关联的，序列前面的数据特征有可能影响到序列后面的数据。如文字数据“我喜欢吃苹果”，在“我喜欢”的后面最有可能加的是名词或者动词，这就是所谓的序列数据中序列数据之间的影响作用。RNN 的特别之处在于，它的隐藏层神经元相互连接，这使得它可以传递与时间相关的输入信息，并考虑时间维度上距离较远的事件之间的时间相关性。RNN 能处理序列的输入，发现前后输入序列的关

系，对比起其它神经网络如 CNN 只能接受独立的输入，这是其与 CNN 等网络的本质区别。RNN 具有记忆特性，而 CNN 不具有，RNN 的这一特性是根据其网络结构来实现的。

RNN 的构成包括输入层、隐藏层以及输出层，如图 2-6 所示。

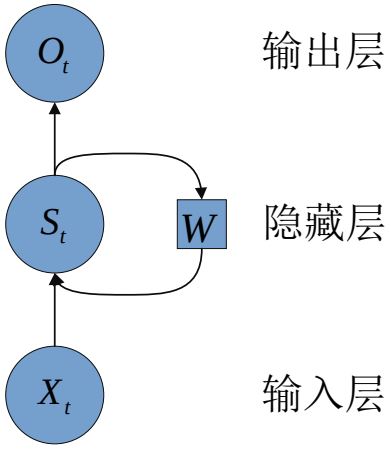


图 2-6　循环神经网络 RNN 结构图

从图 2-6 中可知 RNN 的结构与传统神经网络结构大致相似，但是 RNN 的隐藏层里的神经元还有一个权值 W 连接，就是这个连接使得 RNN 能够获取上下文信息，这也是 RNN 拥有记忆特性的根本原因。

将 RNN 按输入时间进行展开后得到 RNN 网络的展开结构，如图 2-7 所示。

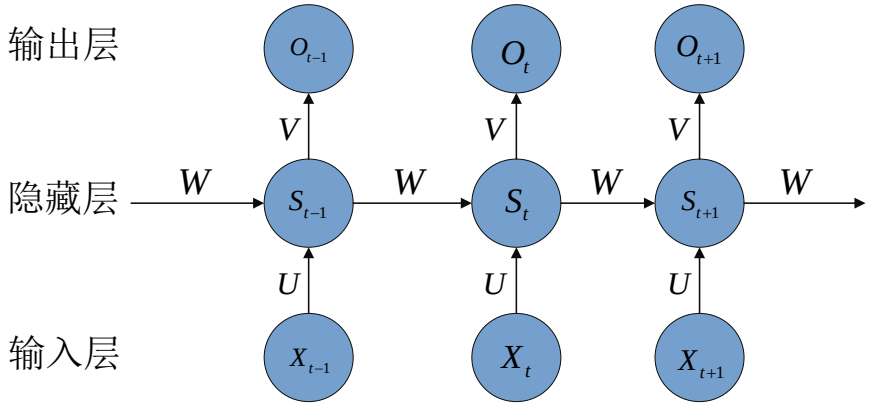


图 2-7　循环神经网络 RNN 展开结构图

根据上面的 RNN 展开结构图图，可以得到得 RNN 的t时刻的输出O_t的计算为公式（2）。其中S_t表示t时刻隐藏层的输入，W表示隐藏层的权值，U 和 V 也表示权重，X_t表示t时刻的输入，b为参数，以下公式总结了 RNN 输出层以及隐藏层神经元之间的输入输出关系：

$$O_t = g\left(V \times S_t + b_q\right) \text{公式（2）}$$

$$\text{其中} S_t = f(U \times X_t + W \times S_{t-1} + b_s)$$

在 RNN 的隐藏层之间，当（X_{t-1}, X_t, X_{t+1}）序列输入时，每个隐藏层的前一神经元的输出也作为当前神经元的输入，这样每一隐藏层神经元就有S和X两个输入，这是实现 RNN 记忆特性的基本原理。例如，在输入一个句子“我爱吃苹果”的时候，这种网络结构就可以发现“我”、“爱吃”、“苹果”三个词组之间的关系，这也使得 RNN 网络适用于 NLP 领域问题的研究。

RNN 可以获取输入序列之间的相关性，从而实现短期记忆。但是对于长序列来说，RNN 就存在梯度消失和梯度爆炸问题[41]。为优化上面提出的问题，在 20 世纪 90 年代，Hochreiter 和 Schmidhuber[42]提出了长短期记忆神经网络（Long Short Term Memory, LSTM）。LSTM 是一种被广泛使用的时间序列算法，它是一种特殊的 RNN，能够学习长期的依赖关系。LSTM 专门用于解决长序列训练过程中梯度消失和梯度爆炸问题，是一种表现更加优异的循环神经网络，尤其在处理更长的序列时表现更加出色。它的结构如图 2-8 和图 2-9 所示。

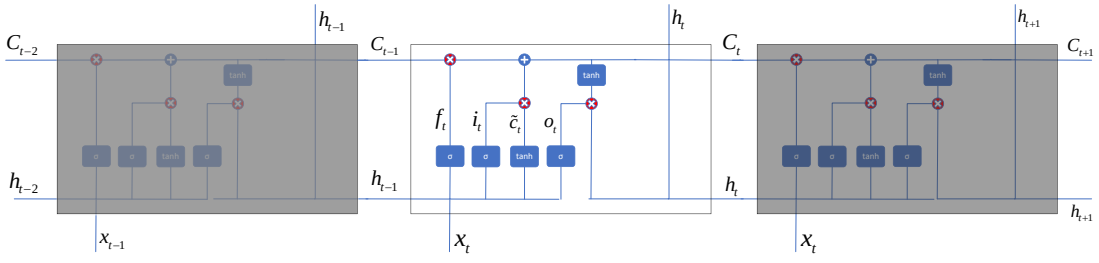


图 2-8　LSTM 神经网络

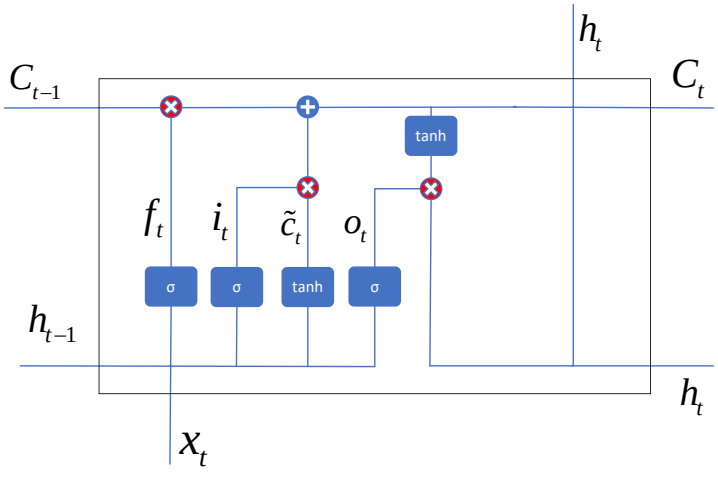


图 2-9　LSTM 神经元结构图

LSTM 参考人类的记忆模式，记住重要的信息、遗忘相对不重要的信息，为了实现这个功能，LSTM 在 RNN 的基础上增加了细胞状态（Cell State）和“门”的结构，它们的作用是调节关于信息的“记忆”，上一层的细胞状态经过遗忘和新的记忆的选择存储后，继续流向下一个细胞。根据图 2-9 可知，细胞状态C_t在最上面传播，隐藏层状态h_t在下面传播，他们的初始状态为全 0，隐藏

层状态h_{t-1}与新的输入x_t对细胞状态进行修改。在 LSTM 神经元细胞中，有三个门：遗忘门、输入门和输出门，它们从左到右依次排列。

（1）遗忘门

$$f_t = \sigma\left(W_f \times [h_{t-1}, x_t] + b_f\right)$$

遗忘门将上一层的隐藏状态h_{t-1}和当前的输入x_t进行拼接，传入$sigmod$函数中，映射到[0,1]中，越接近 0 则意味着越应该被丢弃，越接近 1 则意味着越应该被保留。与上一层的细胞状态C_{t-1}相乘，就是完成了对C_{t-1}中信息的选择，对不重要信息的遗忘。

（2）输入门

$$i_t = \sigma(W_i \times [h_{t-1}, x_t] + b_i)$$

$$\tilde{C}_t = tanh(W_c \times [h_{t-1}, x_t] + b_C)$$

输入门决定加入多少新输入的信息到细胞状态中来。可以看到输入门分为i_t和$\tilde{C}_t$两部分：前者同遗忘门类似，将其映射到[0,1]之间，0 表示不重要，1 表示重要；后者输入$tanh$函数。i_t与后者相乘，决定了$tanh$输出结果的保留与舍弃，作为新输入信息的保留。对于细胞状态的更新，经遗忘门与上层细胞状态点乘后得到的结果，与输入门得到的结果相加，就完成了对上层不重要信息的遗忘和新加入信息的选择保留，即结束了对细胞状态的更新。

（3）输出门

$$o_t = \sigma(W_o \times [h_{t-1}, x_t] + b_o)$$

$$h_t = o_t + tanh(C_t)$$

输出门是确定将细胞状态中的哪些部分能够输出为隐藏状态的值h_t。首先将已更新的细胞状态经过一个$tanh$函数的映射，与h_{t-1}和x_t经$sigmod$函数映射后的门相乘，确定隐藏状态应携带的信息，最后将更新后的隐藏状态h_t和细胞状态C_t输入下一时刻。

2.3 小结

本章主要介绍了本文所使用的理论和方法。首先对生物领域蛋白质的功能和结构进行了介绍，讲述了蛋白质及其结构的基本概念，然后介绍了蛋白质四类结构之间的关系，为蛋白质的结构预测提供了理论基础。最后重点阐述了实验中所用到的一些神经网络的基本模型，如 RNN、LSTM、CNN 等。

3 氨基酸序列编码对比实验

蛋白质的一级结构氨基酸序列是由 20 个字母来表示的，蛋白质二级结构也是用字母表示的。因此，蛋白质一级结构到二级结构的预测任务可以理解为序列到序列（Sequence to Sequence）的任务，序列的编码方式对模型的预测结果会有很大的影响。

3.1 数据集

本文在进行对比实验时，使用了 CB6135 和 CB513 两个数据集，两者均基于 PISCES Cull PDB 服务器的蛋白质结构数据集。CB6135 数据集中一共有 6128 条蛋白质，经过重复项的剔除，最后保留 5926 条不重复的蛋白质，在保留的 CB6133 数据集中选择 5430 条为训练集，255 条为验证集，236 条为测试集。CB513 数据集中一共有 513 条蛋白质，其全部作为模型的验证集。

3.2 实验设计

3.2.1 氨基酸编码方式

本章对蛋白质二级结构预测领域最为常见的三种氨基酸编码方式进行介绍，并进行模型实验对比。

蛋白质一级结构氨基酸序列由丙氨酸、半胱氨酸、谷氨酸、天冬氨酸、甘氨酸、苯丙氨酸、异亮氨酸、组氨酸、赖氨酸、甲硫氨酸、亮氨酸、天冬酰胺、谷氨酰胺、脯氨酸、丝氨酸、精氨酸、苏氨酸、色氨酸、缬氨酸和酪氨酸这 21 种氨基酸构成。 蛋白质 8 类二级结构分由3_{10}螺旋、α-螺旋、π-螺旋、β-桥、β-折叠、转角、弯曲和环状组成。

在蛋白质一级结构到二级结构预测领域内，对于一级结构氨基酸的编码方式最为常见的是独热编码，基团编码和 profile 编码，二级结构一般用独热编码的方式进行编码。

独热编码（One-Hot 编码）也称正交编码[43]，是一种用于将离散型变量表示为二进制向量的编码方式。它通常采用 N 位状态寄存器来对 N 个状态进行编码，每个状态都由独立的寄存器位来表示，并且在任意时刻只有一位有效。在进行独热编码时，首先需要将离散型变量的取值映射到整数值，然后每个整数值被表示为一个二进制向量，除了整数值对应的索引位置上的元素为 1，其余元素都为 0，这样就可以将离散型变量表示为一个由 0 和 1 组成的二进制向量，便于在机器学

习算法中使用。

对 20 种氨基酸采用独热编码方式进行编码，即利用 20 位二进制数唯一表示某一种氨基酸。每个氨基酸被编码为一个长度为 20 的二进制向量，其中只有对应该氨基酸的位置为 1，其余位置都为 0。其编码结果如表 3-1 所示。

表 3-1　二十种氨基酸独热编码表示

氨基酸	独热编码	氨基酸	独热编码
丙氨酸(A)	10000000000000000000	亮氨酸(L)	10000000000000000000
半胱氨酸(C)	00000000000000000000	天冬酰胺(N)	00000000000000000000
谷氨酸(E)	00000000000000000000	谷氨酰胺(Q)	00000000000000000000
天冬氨酸(D)	00000000000000000000	脯氨酸(P)	00000000000000000000
甘氨酸(G)	00000000000000000000	丝氨酸(S)	00000000000000000000
苯丙氨酸(F)	00000000000000000000	精氨酸(R)	00000000000000000000
异亮氨酸(I)	00000000000000000000	苏氨酸(T)	00000000000000000000
组氨酸(H)	00000000000000000000	色氨酸(W)	00000000000000000000
赖氨酸(K)	00000000000000000000	缬氨酸(V)	00000000000000000000
甲硫氨酸(M)	00000000000000000000	酪氨酸(Y)	00000000000000000000

同时采用独热编码方式对 8 类蛋白质二级结构进行编码就是用 8 位二进制数唯一表示某一种二级结构，其编码结果如下表 3-2 所示。

表 3-2　八种蛋白质二级结构独热编码表示

二级结构	独热编码
3_{10}螺旋（G）	10000000
α-螺旋（H）	01000000
π-螺旋（I）	00100000
β-桥（B）	00010000
β-折叠（E）	00001000
转角（T）	00000100
弯曲（S）	00000010
环状（L）	00000001

基团编码，是 Zhang 等人 2017 年提出的一种氨基酸编码方式[44]。在氨基酸序列中，氢原子与氢原子之间，或者非氢原子之间形成的结构稳定的官能团就是

基团。根据基团中所含有的原子之间相互连接的共价键的分类，一共分为 42 种基团，从而生成基团信息表[45]。每一个氨基酸可根据是否拥有这些基团来对其进行编码。基团编码方式采用的是 42 位二级制编码，对其赋值 1 或者 0 的赋值取决于是否拥有其对应的基团，表 3-3 是根据 20 是种氨基酸实际拥有的基团所得到的基团编码表。

表 3-3　二十种氨基酸基团编码表示

氨基酸	基团编码
丙氨酸（A）	0101011000000000000000000000000000000000000
半胱氨酸（C）	0101010001000000000000000000000000000000000
谷氨酸（E）	0101010100000000000010000010000000000000000
天冬氨酸（D）	0101010100000000011000000000000000000000000
甘氨酸（G）	1001010000000000000000000000000000000000000
苯丙氨酸（F）	0101010100000000000000000000000000000000100
异亮氨酸（I）	0101010000101010000000010000000000000000000
组氨酸（H）	0101010100000000000000000000000000000001000
赖氨酸（K）	0101010100000000000010000000000010100000000
甲硫氨酸（M）	0101010100000000000001001000010000000000000
亮氨酸（L）	0101010100000000000100110000000000000000000
天冬酰胺（N）	0101010100000000101000000000000000000000000
谷氨酰胺（Q）	0101010100000000000010000100000000000000000
脯氨酸（P）	0101010100000000000000000000000000000010000
丝氨酸（S）	0101010010000000000000000000000000000000000
精氨酸（R）	0101010100000000000010000001101010110000000
苏氨酸（T）	0101010000010101000000000000000000000000000
色氨酸（W）	0101010100000000000000000000000000000000010
缬氨酸（V）	0101010100000000000000000000000000000100001
酪氨酸（Y）	0101010000101100000000000000000000000000000

Profile 编码是采用 PSSM 矩阵进行编码得到的一个能够体现生物进化相关信息的氨基酸序列信息编码方式。它指在氨基酸序列的各个部位上 1 个氨基酸类型出现的相对概率[46]。例如，若氨基酸序列的为“AAAACCCEDD”，那么氨基酸 A 的 Profile 编码是 0.4，因为在这个序列中 A 出现的几率是 4/10=0.4，相应的氨基酸 C、E 和 D 的 Profile 编码是 0.3、0.1 和 0.2，其他没出现的氨基酸如 G 等

则为 0。再比如氨基酸序列的为“MMMMM”，则 M 的 Profile 编码为 1，因为“MMMMM”序列中只有 M。因为 Profile 编码不仅仅代表自身信息，还代表着其他序列比较信息，因此被认为带有较高的生物进化信息。表 3-4 是某一氨基酸序列的 Profile 编码例子，其中 Seq 是目标序列，AM 表示其他序列。

表 3-4　Profile 编码示例表

Seq	Q	S	E	P	E	D	L	L	K
AM	QQQQQ	SSQAA	KKKKE	PPPPP	EE-GEE	DEDDD	LLAAA	LLIVI	KKEKH
V	0	0	0	0	0	0	0	0.2	0
L	0	0	0	0	0	0	0.4	0.4	0
I	0	0	0	0	0	0	0	0.4	0
M	0	0	0	0	0	0	0	0	0
F	0	0	0	0	0	0	0	0	0
W	0	0	0	0	0	0	0	0	0
Y	0	0	0	0	0	0	0	0	0
G	0	0	0	0	0	0	0	0	0
A	0	0.4	0	0	0.2	0	0.6	0	0
P	0	0	0	1	0	0	0	0	0
S	0	0.4	0	0	0	0	0	0	0
T	0	0	0	0	0	0	0	0	0
C	0	0	0	0	0	0	0	0	0
H	0	0	0	0	0	0	0	0	0.2
R	0	0	0	0	0	0	0	0	0
K	0	0	0.8	0	0	0	0	0	0.6
Q	1	0.2	0	0	0	0	0	0	0
E	0	0	0.2	0	0.8	0.2	0	0	0.2
N	0	0	0	0	0	0	0	0	0
D	0	0	0	0	0	0.8	0	0	0

3.2.2 单编码方式实验

单种氨基酸编码方式对蛋白质二级结构预测精度影响研究实验中，以由 RNN 构成的深度学些模型作为实验的控制变量，然后将以各种氨基酸编码方式对 CB6135 的训练集进行编码从而得到不同氨基酸序列信息的表示，作为基础模型的输入，训练得到预测模型。最后将 CB6135 的测试集，CB513 数据集输入并得到模型的 Q8 准确率，从而分析出三种编码方式对模型结果的影响，以及氨基酸序列信息的表达效果。

在单编码方式实验中所使用的模型是，由一个输入层，一个单向 RNN，一个全连接层，以及一个输出层构成，整个模型结构如图 3-1 所示。

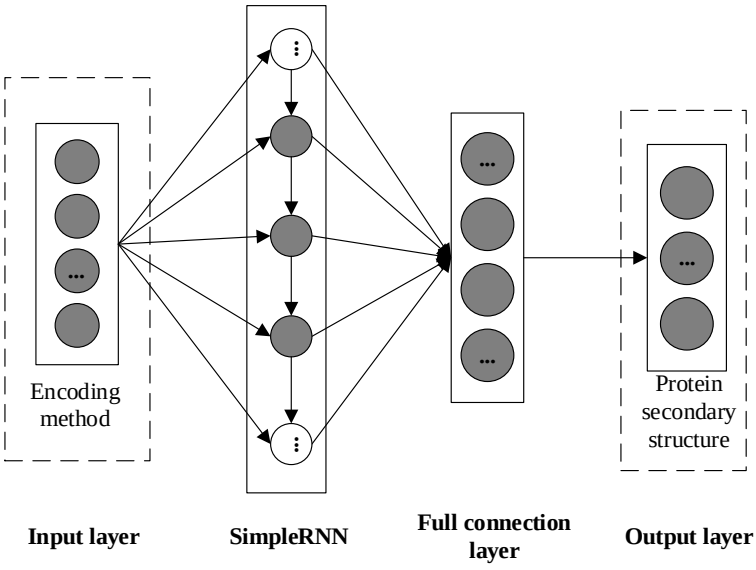


图 3-1　基于 RNN 的单编码蛋白质二级结构预测模型

在模型的输入中，由于 CB6135 和 CB513 数据集中氨基酸序列除了有上述除 20 中氨基酸，还存在着实验中无法准确测出的氨基酸种类，数据集收集的过程中用字母 X 对这种氨基酸进行表示。因此在实际实验中，对氨基酸的编码会比上述介绍的编码方式多一位，用于对 X 类型的氨基酸进行表示。同时由于实验中使用的都是 700 个单位长度以内的氨基酸构成的蛋白质，因此对于不足 700 长度的，还需要进行补足至 700 长度，在具体实验中是采用了全为 0 的情况作为补足的表示。

3.2.3 双编码方式实验

在双编码实验中，将各种编码方式编码的氨基酸序列两两组合输入 2 层 RNN 深度学习模型，得到在两种不同编码方式构造的输入对模型准确率的影响。在实验中，使用的数据集与单编码方式实验中的一致，唯一变化的是在输入方便增加了一种编码方式，使得输入信息更加丰富。这样就可以分析出两种不同编码方式构造的组合输入对模型准确率的影响。

在双编码方式实验中所使用的模型，由一个输入层，一个串联（连结）层，一个单向 RNN，一个全连接层，以及一个输出层构成，整个模型结构如下图 3-2 所示，整体结构与单编码方式实验使用模型一致，只是输入层多了氨基酸序列的另一种表示方法的输入，使得模型可以获取更多的信息。

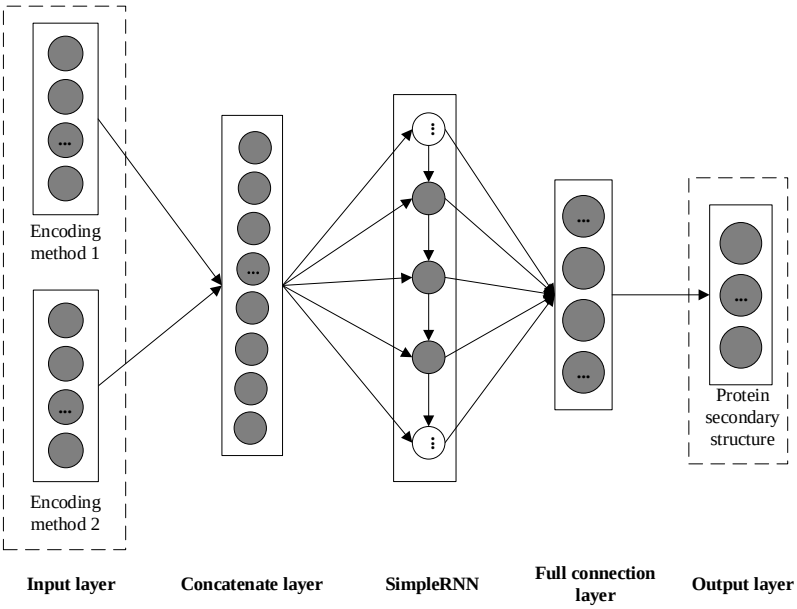


图 3-2　基于 RNN 的双编码蛋白质二级结构预测模型

在实际实验中，分别将独热编码和 Profile 编码组合，基团编码和独热编码组合，基团编码和 Profile 编码组合，三种组合情况输入模型中进行训练，得到预测模型，最后使用 CB6135 验证集和 CB513 数据集对模型的 Q8 准确率进行验证。

3.3 实验结果

单编码方式实验中，分别得到独热编码、基团编码和 Profile 编码的模型。最后将 CB6133 数据集和 CB513 数据集的验证集输入到三个模型中得到如图 3-3 所示结果。独热编码、基团编码和 Profile 编码这三种编码方式编码氨基酸序列训练出的模型在 CB6133 验证集上的准确率分别为 46.84%、23.88%和 59.11%，CB513 数据集的准确率分别为 44.72%、42.31%和 55.39%。由此可知，单编码方式对模型准确率的提升效果由高到低分别是 Profile 编码、独热编码、基团编码。