基于级联卷积和递归神经网络的蛋白质二级结构预测

电子与信息工程学院　　计算机科学与技术
118532014013　　袁 超　　指导教师：游文杰

【摘　要】 蛋白质二级结构预测（**PSSP**）是生物信息学中的一个重要问题。受深度学习在自然语言处理领域成功启发，本文提出了一种端到端的深度神经网络模型，用于预测八类蛋白质二级结构。模型主要包含四个层次：首先，基于氨基酸残基的类别信息、生物的进化信息以及蛋白质序列的组成信息进行编码组合，利用特征嵌入来消除 **0-1** 矩阵的稀疏性；其次，使用多尺度卷积提取氨基酸残基之间的局部相邻特征；然后，通过门控单元（**GRU**）双向递归神经网络提取蛋白质序列的远程上下文关系；最后，融合局部相邻特征以及远程上下文关系来进行八类蛋白质二级结构预测。实验结果表明，本文提出的深度神经网络模型易于收敛，具有较好的可扩展性。预测效果良好，在 **CB513** 数据集上达到了 **68.2%** 的 Q_8 预测精度。

【关键词】 蛋白质二级结构预测；深度神经网络

目录

1. 引言

1.1 背景及现状

1.1.1 背景知识

蛋白质是生命的物质基础，也是细胞中不可缺少的分子成份。各种不同形式的生命活动都与蛋白质发生着紧密的联系，如：催化化学反应的酶、控制体内信号的受体以及人体的肌肉和器官组织[21] 等。随着蛋白质研究的不断进行，先进的测序技术也随之发展，蛋白质组学应用而生。人们可以利用成熟的技术手段[12] 测量蛋白质的序列，或是从已有的开放数据库中查询具有相似组成的蛋白质。迄今为止，已有大量测序蛋白质不断被填入到数据库中，呈现出了指数级上升的趋势[13]。

与测序不同，测量蛋白质的结构困难重重。对于已知序列的蛋白质，可以被确定具体结构的微乎其微[18]。传统实验方法，例如 X 射线晶体学和核磁共振光谱[11]，精度高但耗时长，而且造价昂贵，不能够应对蛋白质序列快速增长的挑战。通过采用先进的计算机技术，高效地处理大量蛋白质序列数据，降低了经济成本和时间负担。因此，采用计算方法进行蛋白质结构预测成为了交叉学科生物信息学领域重要而有趣的课题。

蛋白质结构可划分为四个等级：一级为氨基酸序列，决定了蛋白质的折叠方式；二级描述了多肽的局部现象；三级为全局折叠模式；四级反应了蛋白质高度紧凑和相互作用的结果。蛋白质三级结构和四级结构是高度立体的三维状态，决定了蛋白质的功能特征。在一级结构到三级结构之间，二级结构起到了承接和桥梁的作用。因此，蛋白质二级结构预测（Protein Secondary Structure Prediction，PSSP）是蛋白质结构预测中必不可少的步骤。

1.1.2 发展现状

PSSP 最早追溯到 20 世纪 80 年代，利用统计模型估算不同二级结构中特定氨基酸出现的概率[17] 。考虑到该方法特征不充分，识别率较低，所以结合多种特征信息（亲/疏水性等理化性质），以及复杂统计方法（多肽等局部相关作用）进行了改进，但总体改善效果甚微[28] 。20 世纪 90 年代，位置特异得分矩阵（Position-specific Scoring Matrices，PSSM）通过提取蛋白质家族的进化信息[14] ，预测效果实现了显著提升。21 世纪，优秀的机器学习算法层出不穷，在共同考虑进化信息的前提之下结合不同机器学习算法，如：人工神经网络、支持向量机以及概率图模型等[27] [6] [1] 。近些年，集成多种模型的组合方法，可以克服单个模型固有的缺陷所带来的性能问题。不同种模型间互补，这是一种新的趋势[2]。然而，现有模型多为非序列模型，这些模型不是蛋白质二级结构的理想模型，这是因为蛋白质序列不能以固定维数作为载体。此外，几乎所有方法都是基于 3 类 PSSP，当涉及更具挑战的八类 PSSP 问题时，进展停滞不前。

近年来，得益于更强大的计算机、更庞大的数据集和训练更深网络的技术，一种通用的表征学习的方法：深度学习（Deep Learning）流行起来。传统技术的表示方法完全由手工设计，而深度学习中更复杂的表示是由简单表示组合而成，并使用通用的学习过程从数据中学习而来。事实证明，深度学习非常擅长发现高维数据中错综复杂的结构关系[16] 。

1.2 本文主要工作

本文从蛋白质问题的概述展开，介绍了相关的课题背景以及发展现状。对八类蛋白质二级结构预测的问题做了更精确的描述，并从探讨预测问题出发，介绍了机器学习相关背景知识，包括机器学习最核心的训练模式：监督学习，以及采用经验风险最小化策略和梯度下降算法来进行模型优化。然后，具体介绍了深度神经网络中的一些最基础、最重要的计算模型，并利用这些模型组建深度神经网络来做预测。接着，进行实验，测试该模型的具体效果。最后，做分析和总结。

1.3 论文组织结构

本课题共分为六章，章节内容安排如下：

第一章（引言） 介绍课题的相关背景和发展现状，主要工作，以及论文的结构和安排。

第二章（蛋白质预测概述） 介绍了蛋白质基础知识，八类二级结构预测的具体问题，监督学习以及优化策略等，还有常用的预测评估标准。

第三章（深度学习简介） 介绍了感知机模型，卷积神经网络模型以及 GRU 递归神经网络模型。

第四章（模型架构） 介绍了本文提出深度神经网络模型，包括：编码、组合嵌入层、多尺度卷积层、双向 GRU 层以及预测层。

第五章（实验分析） 介绍了数据集以及具体实验过程。

第六章（总结） 总结本文的工作成果。

2. 蛋白质预测概述

2.1 符号表

表 1 中列举了在本文中使用的符号及其意义

2.2 二级结构预测

2.2.1 组成及结构

蛋白质作为重要的生物大分子，其化学化学组成及比例：碳（50%）、氢（7%）、氧（23%）、氮（16%）、硫（0.3%）其他元素微量。所有蛋白质都来自 20 种不同的 L 型 α 氨基酸（A、C、D、E、F、G、H、I、K、L、M、N、P、Q、R、S、T、V、W 和 Y），氨基酸有氨基（$-NH_2$）和羧基（$-COOH$），两两间可以脱去一分子水（H_2O）缩合形成肽键（$-CO-NH-$）进而组成多肽，此时的氨基酸也被称作残基。通常认为，残基数目达到一定规模，便可称其为蛋白质的一级结构。

蛋白质结构可以分为四个层次，层与层之间可以转化，每一层的结构都与上一层的结构关系密切。图 1 为蛋白质 PDB 105M 的四层结构示意图。观察该蛋白质的二级结构和三级结构，可以发现在三级结构中存在有许多螺旋，这些螺旋都可以从二级结构中找到一一对应的位置，螺旋的数目和长度也完全被定义。并且，在二级结构中，这些螺旋都沿着蛋白质一级结构的序列方向，循环折叠而成。每个残基都有一个二级结构类别标签，连续相邻的残基具有相同的标签，则在空间中体现为一段蛋白质序列的空间折叠。这说明二级结构可以初步呈现出蛋白质的局部空间状态，想要从大量已知的蛋白质序列预测复杂多变的空间结构，可以将二级结构预测作为首要环节。

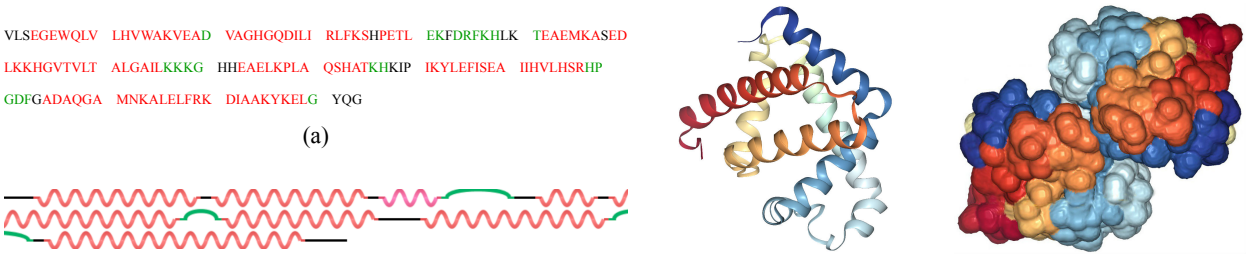

图 1: 数据来自 RCSB(https://www.rcsb.org/) 的蛋白质 PDB 105M：a）为该蛋白质氨基酸序列，字母代表不同种类的氨基酸残基；b）为该蛋白质的 3 类二级结构图，颜色代表的类别与 (a) 中的色彩相对应；c）代表其三级结构的彩虹带图；d）为两条 105M 蛋白质相互结合形成的四级结构表面图。

在早期的研究中，蛋白质的二级结构被划分为三类：螺旋、折叠和转角。之后分类方式不断丰富。根据二级结构定义词典[3] 可将残基分为八个类别：H（$\alpha-$ 螺旋）、G（$310-$ 螺旋）、I（$\pi-$ 螺旋）、E（$\beta-$ 折片）、B（$\beta-$ 桥）、T（转角）、S（卷曲）以及 C（其他）。基于八类的 PSSP 方法可以提供更详细的局部结构信息，也更具挑战性。本课题主要研究八类二级结构预测问题，所以之后提到的 PSSP 均指八类二级结构预测。

2.2.2 八类预测

给定一组蛋白质输入序列 $X = \{x_1, x_2, \cdots, x_N\}$，存在 $f : X \to Y$ 即，对于每个 $x_i \in X$，$y_i \in Y$，都有

$$y_i = f(x_i) \tag{2-1}$$

则称 f 为蛋白质二级结构的预测模型，Y 为输入集 X 的预测结果，N 为样本容量。特别的，令 $x = (x^{(1)}, x^{(2)}, \cdots, x^{(n)})^T$，$x \in X$，$x^{(i)} \in \{A, C, D, \cdots\}_{20}$ 为第 i 个残基，n 是该序列的长度。若存在模型 f，使得

$$y = f(x) = (y^{(1)}, y^{(2)}, \cdots, y^{(n)})^T \tag{2-2}$$

如果 $y^{(i)} \in \{H, G, I, E, B, T, S, C\}_8$，则称 y 为序列 x 的八类二级结构预测结果，$y^{(i)}$ 为第 i 个残基所对应的二级结构, 共八种类别。

我们已经形式化的定义了什么是 PSSP，即我们要学习什么样的模型，解决什么问题。下一小节将介绍一种最常见的学习形式——监督学习，以及如何用梯度下降实现该学习。

2.3 学习、策略以及评估

2.3.1 监督学习

首先，给定一组蛋白质二级结构训练样本

$$T = \{(x_1, y_1), (x_1, y_1), \cdots, (x_N, y_N)\} \tag{2-3}$$

表 1: 主要符号表

常用符号	意义
x	蛋白质序列
y	二级结构
X	蛋白质输入样本
Y	二级结构预测结果
f	二级结构预测模型
f_θ	带训练参数 θ 的预测模型
$\hat{f}$	估计模型
$x^{(i)}$	蛋白质序列第 i 个残基
$y^{(i)}$	二级结构中第 i 个位置所对应类别
x_i	样本中第 i 条蛋白质
y_i	第 i 条蛋白质的二级结构
θ	参数向量
η	学习率
σ	激活函数
$*$	卷积运算
I	矩阵
$h^{(t)}$	动态系统在 t 时刻的状态
$e^{(t)}$	表示序列在 t 位置的 fofe 编码
A	编码层的输出
B	B-GRUs 层的输出
C	多尺度卷积层的输出
J	模型复杂度
$\nabla_\theta J$	模型复杂度对参数 θ 的梯度值
$concatenate\{\}$	对应残基间进行拼接组合
$\{A, C, D, \cdots\}_{20}$	20 种氨基酸
$\{H, G, I, \cdots\}_8$	八类二级结构
$\mathbb{X} = \{A, C, D, \cdots\}_{20}{}^n$	输入空间，表示所有可能的蛋白质输入
$\mathbb{Y} = \{H, G, I, \cdots\}_8{}^n$	预测空间，表示所有可能的二级结构预测
$w = (w_1, w_2, \cdots, w_n)^T$	权值向量
$W_z, W_r, W_h, U_z, U_r, U_h$	GRU 训练参数

其中 (x_i, y_i)，$i = 1, 2, \cdots, N$，称作训练样本点，

$$x_i = (x_i^{(1)}, x_i^{(2)}, \cdots, x_i^{(n)})^T \tag{2-4}$$

$$y_i = (y_i^{(1)}, y_i^{(2)}, \cdots, y_i^{(n)})^T \tag{2-5}$$

$x_i \in X \subseteq \mathbb{X}$，$x_i$ 为输入的观测值，$\mathbb{X} = \{A, C, D, \cdots\}_{20}{}^n$ 为蛋白质序列的输入空间，$y_i \in Y \subseteq \mathbb{Y}$ 为输出观测值，$\mathbb{Y} = \{H, G, I, \cdots\}_8{}^n$，为预测空间。假设训练数据与测试数据依联合概率分布 $P(X, Y)$ 独立同分布产生，在未给出测试数据的情况下，监督学习模型可以从训练数据得到

$$\hat{f} = arg\ max\{P|\hat{P}(Y|X)\} \tag{2-6}$$

$\hat{P}(Y|X)$ 表示由训练样本观测到的条件概率分布，是从数据中学到的经验，也可称模型 f[32]。因为训练集有多组，所以我们学到的经验不唯一。我们从这组经验中选择“最优”（$arg\ max$）的分布 P，既为我们通过训练样本所估计的模型 $\hat{f}$。更普遍的，机器学习模型通常包含一个参数向量 θ

$$\hat{f}_\theta = arg\ max\{P|\hat{P}_\theta(Y|X), \theta \in \mathbb{R}^n\} \tag{2-7}$$

参数向量 θ 的维度由具体模型决定，并会在学习过程中不断被修正，直至模型收敛到最优。

2.3.2 梯度下降

PSSP 问题为典型的多类别分类问题，所以设估计模型 $\hat{f}_\theta$ 的损失函数为

$$L(y_j, \hat{f}_\theta(x_j)) = -\sum_{i=1}^{n} y_j^{(i)} \log \hat{f}_\theta(x_j)^{(i)} \tag{2-8}$$

则其代价函数：

$$J_\theta = \frac{1}{N} \sum_{j=1}^{N} L(y_j, \hat{f}_\theta(x_j)) \tag{2-9}$$

J_θ 反应了模型在观测样本中的拟合效果，当观测样本 N 趋近于无穷，我们可以学习到最理想的模型，代价函数也可以达到最低值 0。所以，我们说优化模型，就是训练参数 θ 使得 J_θ 趋于 0。我们也把这种策略称为经验风险最小化（Empirical Risk Minimization，ERM）策略[33] 。

若 $\hat{f}_\theta$ 为线性模型，最小化代价函数可以转化成凸优化问题，即求 ∂J_θ 的零界点。但 PSSP 问题是高维非线性复杂性问题，也称 NP（Non-deterministic, Polynomial）问题[15] ，无法在多项式时间内被确定地求解，所以我们只关注一些较小值或是局部极小值，并利用迭代、梯度下降来优化该解，具体的：

$$\theta \leftarrow \theta - \eta \frac{\partial J_\theta}{\partial \theta} \tag{2-10}$$

其中 η 为学习率。通常，在深度学习中模型非常复杂，难以求出 $\partial J_\theta / \partial \theta$ 的具体解析式，我们常用梯度 $\nabla_\theta J$ 来替代，并利用反向传播（Backpropagation，BP）的方法来近似地计算。将 η 设置为一个较小的值，每次进行小步的迭代。迭代过程中 θ 总会沿着梯度下降的方向进行更新，即 J_θ 越来越接近于最小值，从而到达局部极小值点。

2.3.3 预测评估

为了客观评价预测的性能，需要选取一些评估指标。PSSP 问题最常用的评估指标是 Q_8 准确率，设第 i 条蛋白质序列 x_i 有 n_i 个残基，其中正确预测有 m_i 个残基，则

$$Q_8 = \frac{\sum_{i=1}^{N} m_i}{\sum_{i=1}^{N} n_i} \tag{2-11}$$

第 i 条蛋白质的 Q_8 准确率

$$q_i = \frac{m_i}{n_i} \tag{2-12}$$

也可以参考准确率

$$r = \frac{1}{N} \sum_{i=1}^{N} \frac{m_i}{n_i} \tag{2-13}$$

2.4 本章小结

本章介绍了蛋白质二级结构预测的基础知识。下一章主要介绍一些计算模型方法。

3. 深度学习简介

首先回顾一些基本概念。

3.1 神经元与深度

神经元（Neuron）是神经网络中最基础的成分，各种复杂的网络结构都是由一个个小的神经元连接构成。其中，最简单的设计是感知机（Perceptron），一种包含有多个输入的感知机模型如图 3 所示。

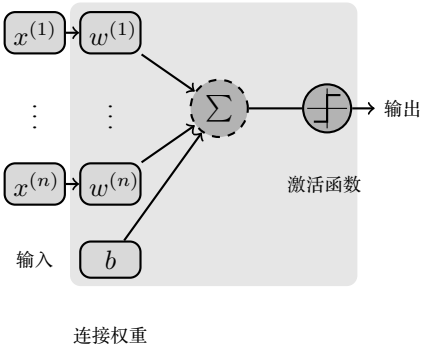

图 2: 多输入感知机模型

$$f_\theta = \sigma(\sum_{i=1}^{n} w^{(i)}x^{(i)} + b) \tag{3-1}$$

特别的，$\theta = (w^{(1)}, w^{(2)}, \cdots, w^{(n)}, b)^T$。$\sigma$ 为激活函数，常见的激活函数有:S 函数 $Sigmoid$、双曲正切函数 $tanh$，线性整流函数 $ReLU$ 以及归一化指数函数 $Softmax$[7] 。

$$sigmoid(x) = \frac{1}{1+e^{-x}} \tag{3-2}$$

$$tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{3-3}$$

$$ReLU(x) = \begin{cases} 0.01x, & x0 \\ x, & x \geq 0 \end{cases} \tag{3-4}$$

$$Softmax(x)^{(j)} = \frac{e^{x^{(j)}}}{\sum_n^{i=1} e^{x^{(i)}}} \tag{3-5}$$

感知机只在输出时进行一次激活，所以学习能力非常有限。事实上，感知机可以灵活组合，使用多个感知机学习可以得到多个输出，多分类神经网络模型中经常可以见到类似的结构。并且也可以将感知机的输出作为新的输入，这就是我们常说的多层神经网络，对应的总层数即是所谓的模型的深度（Depth）。

这种由一层的输出作为下一层的输入的结构是非常灵活的，我们可以用其他计算结构替换对应的子结构。下一小节将会介绍两种常用的计算结构：卷积神经网络（Convolutional Neural Network, CNN）和递归神经网络（Recursive Neural Networks），这是深度学习中最重要的，也是最基础技术[22] [10] 。

3.2 卷积神经网络

卷积神经网络是一种专门处理矩阵的神经网络，卷积神经网络中每层都由若干卷积单元组成，其中的参数称为核函数，记为 K。

$$K = \{w_1, w_2, \cdots, w_k\} \tag{3-6}$$

设卷积核 $w \in K$，w 与矩阵 I 进行一次卷积运算，则矩阵 I 第 i 行第 j 列的运算结果为:

$$I'_{i,j} = (I * w)_{i,j} = \sum_m \sum_n I_{i+m,j+n} w_{m,n} \tag{3-7}$$

$(m \times n)$ 为卷积核 w 的尺寸，$*$ 为卷积运算，I' 为单个卷积的输出，上标 $'$ 表示输出 I' 的尺寸与输入不同，发生了改变。若 K 中卷机核数目为 k，则单层卷机的输出为 A:

$$A = ReLU((I * K) + b) \tag{3-8}$$

$$A = \{I'_1, I'_2, \cdots, I'_k\} \tag{3-9}$$

3.3 递归神经网络

与卷积神经网络的核函数不同，循环神经网络的子结构是一个动态系统

$$h^{(t)} = f_\theta(h^{(t-1)}, x^{(t)}) \tag{3-10}$$

$h^{(t)}$ 称为系统在 t 时刻的状态也可称为 t 时刻的输出，$x^{(t)}$ 是 t 时刻给定的输入。系统在 t 时刻展开形式

$$h^{(t)} = g_\theta^{(t)}(x^{(t)}, x^{(t-1)}, \cdots, x^{(1)}, h^{(1)}, h^{(2)}, \cdots, h^{(t-1)}) \tag{3-11}$$

若保存所有迭代的输出，循环神经网络模型可以构建出如下的映射关系：

$$f_\theta : \{x^{(1)}, x^{(2)}, \cdots, x^{(t)}\} \to \{h^{(1)}, h^{(2)}, \cdots, h^{(t)}\} \tag{3-12}$$

注意，这里映射两边都是集合的形式，表示该模型也和神经网络模型一样非常灵活。我们可以只保留第 t 时刻的输出状态 $h^{(t)}$，模型为 $n \to 1$；也可以保留第 i 到第 t 时刻输出 $\{h^{(i)}, \cdots, h^{(t)}\}$，模型变成了 $n \to m$；也可以令 $x^{(t+1)} \leftarrow h^{(t)}$，模型又变成了 $m \to n$ 或 $1 \to n$，其中 n 和 m 满足 $n > m > 1$。

因为循环神经网络会考虑 t 时刻的状态，所以结构具备学习“因果”的能力[23]。但蛋白质二级结构和序列的关系是整体依赖的，我们不能说蛋白质第 i 位置的二级结构只和从第 1 到第 $i-1$ 的序列相关而与第 $i+1$ 后面的序列无关。双向循环神经网络[24]则是为满足该需求而设计的，具体思路是沿着时间 t 的反方向再输入一次，训练一层新的网络，这样输出集比原来扩大了一倍，包含了更丰富的输出结果。

普通的循环神经网络无法处理随模型不断递归导致的权重爆炸或梯度消失的问题（Vanishing gradient problem）。近几年提出的具备“遗忘功能”的模块：长短时间记忆（Long-Short Term Memory，LSTM）和门控循环单元（Gated Recurrent Unit，GRU）被广泛用于训练循环网络[4][5]，具有类似模块的可以递归训练的神经网络又被称为递归神经网络（Recursive Neural Networks，RNN）。LSTM 和 GRU 在实验中表现效果相当[9]，但 GRU 的参数更少，约为 LSTM 的 $3/4$，所以训练速度更快。GRU 的结构如图 3。

$$z^{(t)} = \sigma(W_z x^{(t)} \oplus U_z h^{(t-1)} + b_z) \tag{3-13}$$

$$r^{(t)} = \sigma(W_r x^{(t)} \oplus U_r h^{(t-1)} + b_r) \tag{3-14}$$

$$h'^{(t)} = tanh(W_h x^{(t)} \oplus U_h(r^{(t)} \otimes h^{(t-1)}) + b_h) \tag{3-15}$$

$$h^{(t)} = z^{(t)} \otimes h^{(t-1)} + (1 - z^{(t)}) \otimes h'^{(t)} \tag{3-16}$$

$z^{(t)}$ 是控制更新的门控，$r^{(t)}$ 是控制重置的门控，$h'^{(t)}$ 包含了当前输入 $x^{(t)}$ 以及历史信息 $h^{(t-1)}$。当 $z^{(t)} = 0$ 时，当前状态 $h^{(t)}$ 和历史状态 $h^{(t-1)}$ 之间为非线性函数。若同时有 $z = 0$，$r = 1$ 时，GRU 单元退化为简单的循环网络；若同时有 $z = 0$，$r = 0$，当前状态 $h^{(t)}$ 只和当前输入 $x^{(t)}$ 相关，和历史状态 $h^{(t-1)}$ 无关。当 $z = 1$ 时，当前状态等于上一时刻的状态 $h^{(t-1)}$，和当前输入 $x^{(t)}$ 无关。

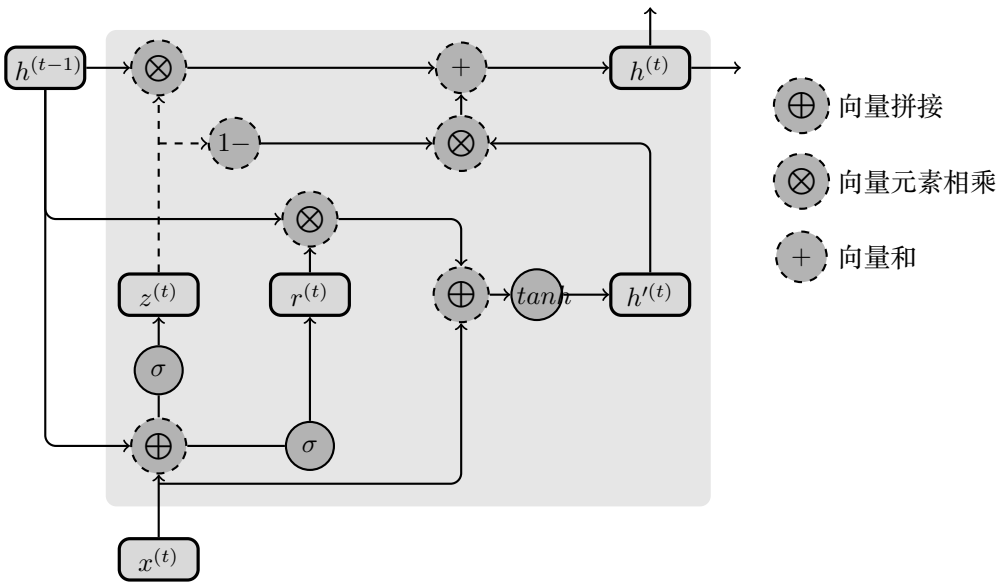

图 3: GRU 的结构图

3.4 本章小结

本章介绍了深度学习模型的基础知识，下一章将利用这些知识构建预测蛋白质二级结构的深度神经网络模型。

4. 模型架构

为解决 PSSP 问题，本文设计了一个深度神经网络模型架构，如图 4 所示。该模型主要包含：组合嵌入层、多尺度卷积层、B-GRUs 层以及预测层四个层次。下面将依次详细介绍各层。

4.1 组合嵌入层

4.1.1 PSSM

位置特异性得分矩阵（Position Specific Scoring Matrix，PSSM）通过一组相似蛋白质序列进行位置特异性匹配，计算得分获得。相似蛋白质是通过多序列比对从相似度较高的排名中选取，再根据序列长度分组，计算各个位置对应每一种氨基酸的可替换性得分，这个得分称为位置特异性得分。对于一条蛋白质序列的一个位置，也就是一个残基，替换 20 种不同类型的氨基酸就对应着 20 个得分。所以一条长度为 n 的蛋白质序列对应的 PSSM 是一个 $n \times 20$ 的矩阵。

PSSM 通过 NCBI(`https://www.ncbi.nlm.nih.gov/`)PSI-BLAST 软件获取，计算工作量非常大，需要先对蛋白质数据库中数量庞大的蛋白质进行搜索和匹配，每一条蛋白质都需要搜索一遍数据库。通常认为，PSSM 可以挖掘到远缘亲属关系和生物的进化信息[20] 。

4.1.2 FOFE 编码

传统的蛋白质编码 one-hot[30] ，将蛋白质序列映射成一个稀疏矩阵，矩阵中每一行对应一个残基，且每一行都只有一个 1 值，其他全为 0。one-hot 可以编码对应残基的值的信息，但无法体现上下文关系的信息，这是序列模型所具有的天然属性。为了利用这一属性，我们采用 FOFE 编码（Fixed-size Ordinally-Forgetting Encoding，FOFE)[29] 。

给定一条蛋白质序列 x 及其 one-hot 编码 x'，则 x 的 FOFE 编码矩阵 $A_{fofe} = (e^{(1)}, e^{(2)}, \cdots, e^{(t)})^T$

$$e^{(t)} = \alpha e^{(t-1)} + x'^{(t)} \tag{4-1}$$

α 为遗忘因子，$e^{(t)}$ 表示序列 x 中残基 $x^{(t)}$ 的 FOFE 编码，且 $1 \geq e^{(t)} \geq 0$。我们可以进行正反两次编码也称双向 FOFE，这样编码既保留了上下文信息，也保留了前后依赖关系。

4.1.3 组合嵌入

分别计算出蛋白质序列的三种编码矩阵：ont-hot 编码矩阵 A_{0-1}，位置特异性得分矩阵 A_{pssm} 以及双向 FOFE 编码矩阵 A_{bifofe}。为了保持数值大小的一致性，将 A_{pssm} 做一次 S 函数激活：$sigmoid(A_{pssm})$，然后根据残基的位置，将三种不同的编码拼接形成新的组合编码矩阵：

$$A = concatenate\{A_{0-1}, sigmoid(A_{pssm}), A_{bifofe}\} \tag{4-2}$$

因为新的组合编码矩阵 A 中有 one-hot，稀疏的形式不利于我们进行训练，所以在输出到下一层前还需要做一次嵌入操作来消除编码矩阵的稀疏性。嵌入的方法有很多种，例如：look-up、word2vec 等[8] ，本文主要采用训练的方式实现嵌入。在组合编码后添加一层具有 128 个神经元的全连接层，使用 ReLU 进行激活，该层的参数会随整个深度模型一同进行训练。输出结果即为残基的组合嵌入编码，长度为 128，整个嵌入过程执行结束，会输出对应于每个位置所有残基的组合嵌入编码。

4.2 多尺度卷积层

很显然，相邻残基间的二级结构具有相似的特征，称为局部依赖关系。为了模拟相邻残基间的局部依赖关系，我们使用具有滑动窗口特性的 CNN 和 ReLU 激活函数来学习局部特征，特别的采用扩张技术使卷积后的数据尺寸保持不变，即蛋白质链的长度不会发生损失。我们选择 3 种不同的尺寸卷积核函数 $K = \{K^{(1)}, K^{(2)}, K^{(3)}\}$，卷积核 $w^{(1)}, w^{(2)}, w^{(3)}$ 分别属于这三种核函数，每一种核函数都包含 64 个卷积核。所以每一种核函数 $K^{(i)}$ 输出的结果

$$C^{(i)} = ReLU((A * K^{(i)}) + b^{(i)}) \tag{4-3}$$

$i = 1, 2, 3$，其中卷积核 $w^{(1)}$ 大小为 (3×128)，$w^{(2)}$ 为 (7×128)，$w^{(3)}$ 为 (11×128)。和之前的操作类似，将三种卷积的输出并排连接在一起：

$$C = concatenate\{C^{(1)}, C^{(2)}, C^{(3)}\} \tag{4-4}$$

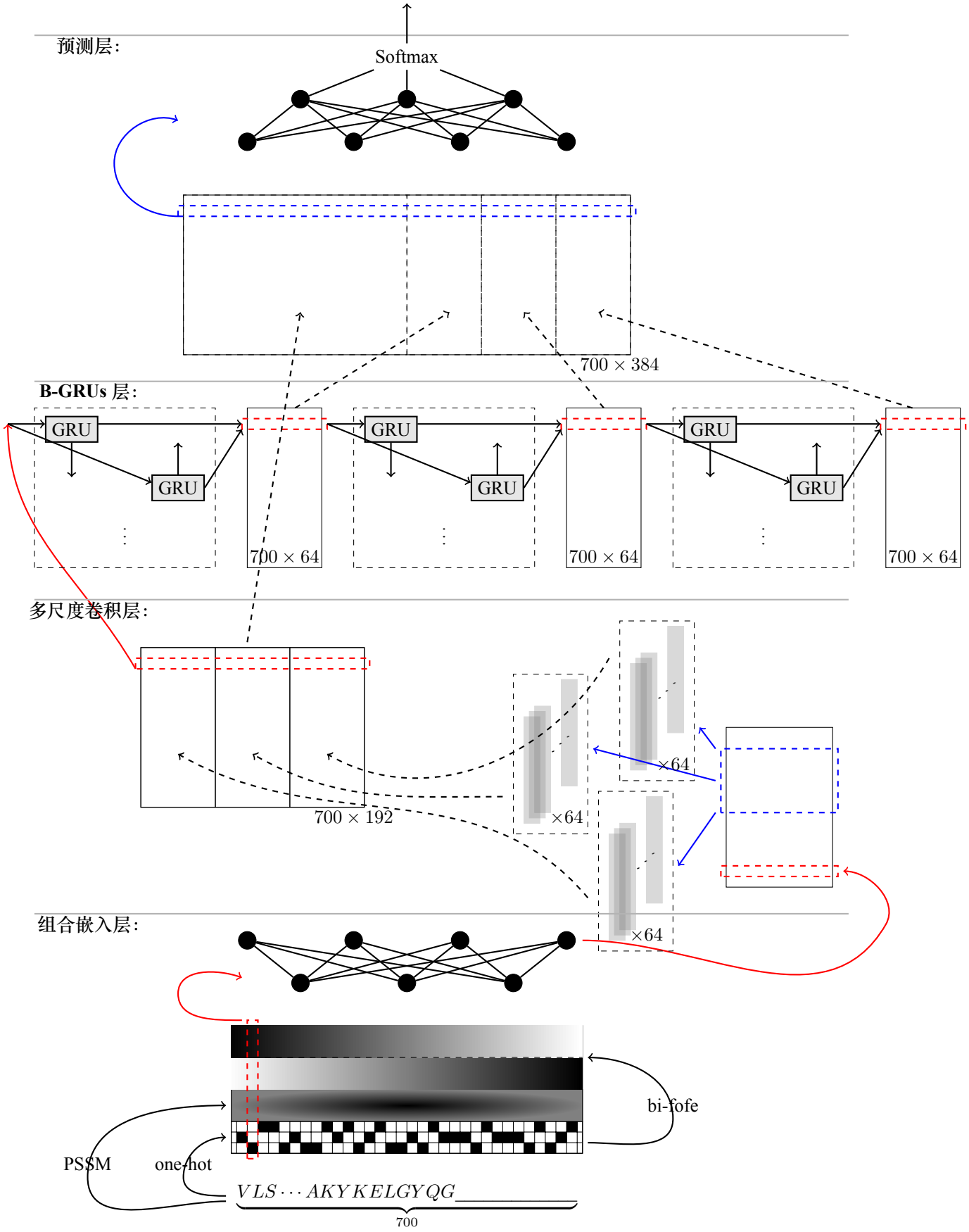
预测层:
Softmax
700 × 384
B-GRUs 层:
GRU
GRU
GRU
GRU
GRU
GRU
700 × 64
700 × 64
700 × 64
多尺度卷积层:
700 × 192
×64
×64
×64
组合嵌入层:
bi-fofe
PSSM
one-hot
VLS · · · AKYKELGYQG
700

图 4: 模型架构

4.3 B-GRUs 层

另一方面，二级结构不只和相邻残基有关，两个距离较远的残基也可能因为蛋白质空间折叠而相互接触，产生作用关系，这种情况也称作长程依赖关系。由于卷积神经网络的卷积核具有固定的尺寸，只能学习到窗口内有限的局部特征。而递归神经网络会将上一层递归层的输出 $h^{(t)}$ 作为参数输入到新的递归层当中进行计算，所以 $h^{(t)}$ 携带了所有上层的信息，即前半段序列的信息。

我们使用三层的 BGRU，每一层的输出用作下一层的输入，每层输出长度都为 64，即正反向各输出 32。设三层 BGRU 的输出分别为 $B^{(1)}, B^{(2)}, B^{(3)}$，其中 $B^{(2)} = bgru(B^{(1)})$，$B^{(3)} = bgru(B^{(2)})$。将这三层的输出 $B^{(1)}, B^{(2)}, B^{(3)}$ 取出和上一层卷积层的输出 C 拼接在一起，组合成 B-GRUs 层的输出 B：

$$B = concatenate\{B^{(1)}, B^{(2)}, B^{(3)}, C)\} \tag{4-5}$$

4.4 预测层

至此，已经得到了表示局部以及长程依赖关系的特征，其中局部特征和全局特征长度各为 192。使用两层全连接层混合这两种特征，用 $ReLU$ 来激活。通过 dropout[19] 随机失活部分神经元，来防止模型过拟合陷入到局部极值。最后使用一个具有八输出的 softmax 做分类。

4.5 本章小结

本章介绍了深度神经网络模型的细节，下一章将利用该模型进行实验和分析。

5. 实验分析

5.1 数据集

本文采用公开数据集 CB6133 和 CB513。CB6133 是利用 PSICES CullPDB（http://dunbrack.fccc.edu/PISCES.php）生成非同源蛋白质结构和序列的数据集，CB513 是公共基准测试数据集。因为 CB513 和 CB6133 之间存在冗余数据，所以通过去除相似性大于 25% 的序列来生成 CB6133 的过滤版本，过滤后的 CB6133 包含 5534 条蛋白质，每条序列的长度都是不固定的。过滤 CB6133 和 CB513 中的序列长度统计如图 5。由于大多数序列都少于 700 个氨基酸，我们选择 700 作为模型训练的信道长度。即，所有长度不满足 700 的蛋白质用 0 来填充空余位置，超出 700 的部分进行截断。

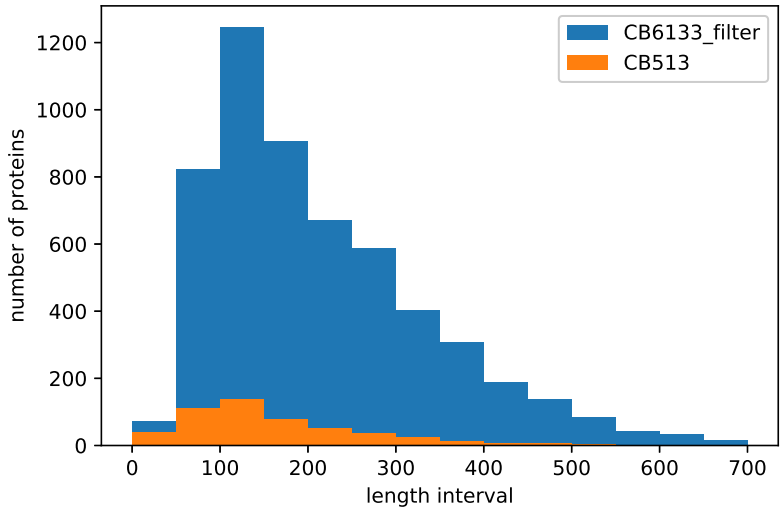

图 5: 过滤 CB6133 和 CB513 中序列长度分布图

对于上述的数据集，每个残基有 57 长度的数据通道。‘[0,22)’通道是氨基酸残基；‘[22,31)’通道是八类二级结构标签；‘[31,33)’通道是 N-和 C 末端；‘[33,35)’通道是相对和绝对溶剂可及性；‘[35,57)’通道是序列谱，通过逻辑函数重新定标 PSSM（PSSM 通过 PSI-BLAST 从 UniRef90 数据库导出，E 值阈值为 0.001 和迭代次数为 3 次）。

5.2 实验测试

具体的，深度学习模型用 Keras(https://keras.io/) 库实现，该库基于深度学习开源框架 TensorFlow(https://www.tensorflow.org/) 的高级的 API。使用 Keras 默认设置来初始化网络权重，并用自适应学习率的优化器 Adam 训练所有层，批量大小为 32。整个神经网络都在一台带有 32GB 内存，NVIDIA GTX 1080 Ti GPU，Intel Core i7-8700k CPU 的服务器上训练。同时，使用提前停止来防止模型发生过拟合，训练大约需要半天时间。

图 6 反应了训练损失和验证损失随迭代次数的变化关系，数据来自过滤 CB6133 数据集，0.4% 的数据用于验证，其余全做训练。

图 7 是用过滤 CB6133 迭代 20 次进行训练，将每次迭代的模型用 CB513 测试其 Q8 准确率。可以观察到，在训练集迭代到第 10 次，测试集的 Q8 准确率局部最优为 68.24%。

表 2 给出了算法在 CB513 测试集上的混淆矩阵。

表 3 给出了本文算法与三个算法 CNF[26] ，SC-GSN[31] 和 LSTM large[25] 在 CB513 数据集上的 Q8 预测精度对比。

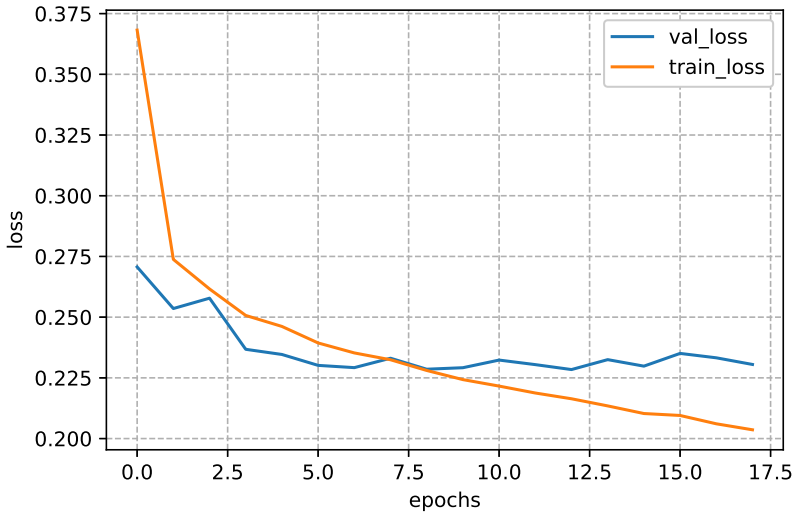

图 6: 训练损失和验证损失随迭代次数的变化。使用过滤 CB6133 的 99.6% 做训练数据，剩余的 0.4% 做验证数据。

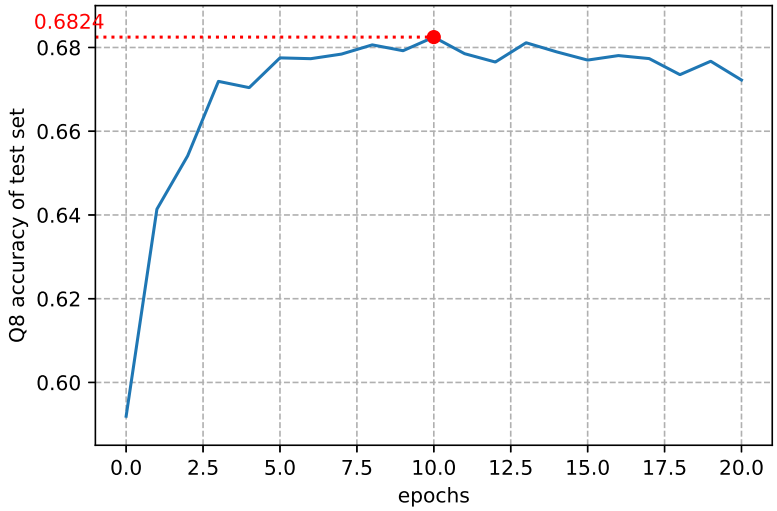

图 7: CB513 Q8 准确率随训练迭代次数的变化。使用过滤 CB6133 做训练数据，每轮迭代测试一次 CB513 Q8 准确率。

表 2: 算法在 CB513 测试集上的混淆矩阵

	H_{pred}	B_{pred}	E_{pred}	G_{pred}	I_{pred}	T_{pred}	S_{pred}	L_{pred}
H	**92.62**%	0.0%	0.97%	1.1%	0.0%	2.46%	0.3%	2.55%
B	9.14%	**1.27**%	27.18%	0.93%	0.0%	10.58%	5.59%	45.3%
E	2.2%	0.05%	**82.08**%	0.46%	0.0%	2.38%	1.53%	11.31%
G	27.14%	0.0%	6.1%	**23.63**%	0.0%	23.21%	2.68%	17.24%
I	66.67%	0.0%	3.33%	0.0%	**0.0**%	13.33%	3.33%	13.33%
T	18.37%	0.0%	5.38%	3.49%	0.0%	**53.07**%	4.49%	15.2%
S	8.35%	0.02%	11.1%	1.74%	0.0%	21.37%	**21.38**%	36.04%
L	6.17%	0.06%	17.7%	1.03%	0.0%	8.97%	5.48%	**60.59**%

表 3: 算法在 CB513 测试集上的分类性能比较

算法	Q_8(%)
CNF	64.9%
SC-GSN	66.4%
LSTM large	67.4%
本文算法	**68.2**%

6. 总结

本文将深度神经网络应用于蛋白质二级结构预测，融合多种信息：氨基酸残基的类别信息、生物的进化信息以及蛋白质序列的组成信息，提出了级联多尺度卷积和双向 GRU 的端到端模型，用于八类蛋白质二级结构预测。实验结果表明，我们的模型预测效果良好。这种结合局部相邻特征和远程上下文关系的模型架构，可以推广应用到生物信息学领域的其他结构预测问题中。

参考文献

[1] K. Asai, S. Hayamizu, and K. Handa. Prediction of protein secondary structure by the hidden markov model. *Bioinformatics*, 9(2):141–146, 1993.

[2] N. P. Bidargaddi, M. Chetty, and J. Kamruzzaman. Combining segmental semi-markov models with neural networks for protein secondary structure prediction. *Neurocomputing*, 72(16-18):3943–3950, 2009.

[3] E. Bulut and I. Korpeoglu. Dssp: a dynamic sleep scheduling protocol for prolonging the lifetime of wireless sensor networks. In *Advanced Information Networking and Applications Workshops, 2007, AINAW'07. 21st International Conference on*, volume 2, pages 725–730. IEEE, 2007.

[4] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[5] F. A. Gers, J. Schmidhuber, and F. Cummins. Learning to forget: Continual prediction with lstm. 1999.

[6] J. E. Gewehr and R. Zimmer. Ssep-domain: protein domain prediction by alignment of secondary structure elements and profiles. *bioinformatics*, 22(2):181–187, 2005.

[7] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256, 2010.

[8] Y. Goldberg and O. Levy. word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*, 2014.

[9] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

[10] Y. Guo, B. Wang, W. Li, and B. Yang. Protein secondary structure prediction improved by recurrent neural networks integrated with two-dimensional convolutional neural networks. *Journal of bioinformatics and computational biology*, 16(5):1850021–1850021, 2018.

[11] H. K. Ho, L. Zhang, K. Ramamohanarao, and S. Martin. A survey of machine learning methods for secondary and supersecondary protein structure prediction. In *Protein Supersecondary Structures*, pages 87–106. Springer, 2012.

[12] D. F. Hunt, J. R. Yates, J. Shabanowitz, S. Winston, and C. R. Hauer. Protein sequencing by tandem mass spectrometry. *Proceedings of the National Academy of Sciences*, 83(17):6233–6237, 1986.

[13] Q. Jiang, X. Jin, S.-J. Lee, and S. Yao. Protein secondary structure prediction: A survey of the state of the art. *Journal of Molecular Graphics and Modelling*, 76:379–402, 2017.

[14] D. T. Jones. Protein secondary structure prediction based on position-specific scoring matrices1. *Journal of molecular biology*, 292(2):195–202, 1999.

[15] R. H. Lathrop. The protein threading problem with sequence amino acid interaction preferences is np-complete. *Protein Engineering, Design and Selection*, 7(9):1059–1068, 1994.

[16] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *nature*, 521(7553):436, 2015.

[17] M. Levitt and C. Chothia. Structural patterns in globular proteins. *Nature*, 261(5561):552, 1976.

[18] D. Li, T. Li, P. Cong, W. Xiong, and J. Sun. A novel structural position-specific scoring matrix for the prediction of protein secondary structures. *Bioinformatics*, 28(1):32–39, 2011.

[19] Z. Li and Y. Yu. Protein secondary structure prediction using cascaded convolutional and recurrent neural networks. *arXiv preprint arXiv:1604.07176*, 2016.

[20] Z. Lin, J. Lanchantin, and Y. Qi. Must-cnn: a multilayer shift-and-stitch deep convolutional architecture for sequence-based protein structure prediction. *arXiv preprint arXiv:1605.03004*, 2016.

[21] L. Liu, L. Tang, W. Dong, S. Yao, and W. Zhou. An overview of topic modeling and its current applications in bioinformatics. *SpringerPlus*, 5(1):1608, 2016.

[22] Y. Liu, J. Cheng, Y. Ma, and Y. Chen. Protein secondary structure prediction based on two dimensional deep convolutional neural networks. In *Computer and Communications (ICCC), 2017 3rd IEEE International Conference on*, pages 1995–1999. IEEE, 2017.

[23] T. Mikolov, M. Karafiát, L. Burget, J. Černockỳ, and S. Khudanpur. Recurrent neural network based language model. In *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[24] S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, and M. Pantic. End-to-end audiovisual speech recognition. *arXiv preprint arXiv:1802.06424*, 2018.

[25] S. K. Sønderby and O. Winther. Protein secondary structure prediction with long short term memory networks. *arXiv preprint arXiv:1412.7828*, 2014.

[26] Z. Wang, F. Zhao, J. Peng, and J. Xu. Protein 8-class secondary structure prediction using conditional neural fields. In *Bioinformatics and Biomedicine (BIBM), 2010 IEEE International Conference on*, pages 109–114. IEEE, 2010.

[27] C. Wu, M. Berry, S. Shivakumar, and J. McLarty. Neural networks for full-scale protein sequence classification: Sequence encoding with singular value decomposition. *Machine Learning*, 21(1-2):177–193, 1995.

[28] G.-Z. Zhang, D.-S. Huang, Y. Zhu, and Y.-X. Li. Improving protein secondary structure prediction by using the residue conformational classes. *Pattern Recognition Letters*, 26(15):2346–2352, 2005.

[29] S. Zhang, H. Jiang, M. Xu, J. Hou, and L. Dai. A fixed-size encoding method for variable-length sequences with its application to neural network language models. *arXiv preprint arXiv:1505.01504*, 2015.

[30] L. Zheng, H.-l. Li, N. Wu, and L. Ao. Protein secondary structure prediction based on deep learning. *DEStech Transactions on Engineering and Technology Research*, (ismii), 2017.

[31] J. Zhou and O. G. Troyanskaya. Deep supervised and convolutional generative stochastic network for protein secondary structure prediction. *arXiv preprint arXiv:1403.1347*, 2014.

[32] 周志华. 机器学习. Qing hua da xue chu ban she, 2016.

[33] 李航 et al. 统计学习方法, 2012.

Protein Secondary Structure Prediction Base on Cascaded Convolutional and Recurrent Neural Networks

College of electronic and information engineering
Major:Computer Science and Technology
118532014013 Yuan Chao Supervisor: You Wenjie

【Abstract】 **Protein secondary structure prediction (PSSP) is an important issue in bioinformatics. Inspired by the successes of deep learning in the field of natural language processing (NLP), in this paper, we propose an end-to-end deep neural network model for prediction of eight types protein secondary structure. The model includes four layers: The first layer, the category information of amino acid residues, the evolutionary information of organism and the composition information of protein sequences are encoded and combined, and use the feature embedding to eliminate the sparsity of the 0-1 matrix; The second layer, use the multi-scale convolution to extract the local adjacent features between the amino acid residues; The third layer, use the bidirectional recurrent neural network with GRU to extracts the remote context of protein sequences; The last layer, combine local adjacent features and remote context relationships to predict the secondary structure of eight types. The experimental results show that the deep neural network model proposed in this paper is easy to converge and has good scalability. The model worked well, and achieved 68.2% Q_8 accuracy on the public benchmark CB513.**

【Keywords】 **Protein Secondary Structure Prediction; Deep Neural Network**