分　类　号：TP183　　单位代码：10183
研究生学号：2017534016　　密　　级：公 开

吉 林 大 学

硕士学位论文

(专业学位)

基于 GoogLeNet 模型的带假结的 RNA 二级结构预测方法

Method for predicting RNA secondary structure with pseudoknot based on GoogLeNet model

作 者 姓 名：李聪
类　　别：工程硕士
领域（方向）：计算机技术
指 导 教 师：张浩 教授
培 养 单 位：计算机科学与技术学院

2020 年 5 月

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

基于 GoogLeNet 模型的带假结的 RNA 二级结构预测方法

Method for predicting RNA secondary structure with pseudoknot based on GoogLeNet model

作 者 姓 名：李聪

领域（方向）：计算机技术

指 导 教 师：张浩　教授

类　　别：工程硕士

答 辩 日 期：2020 年 5 月 31 日

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

吉林大学硕士学位论文原创性声明

本人郑重声明：所呈交学位论文，是本人在指导教师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的作品成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

学位论文作者签名：李聪

日期：2020 年 5 月 31 日

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

关于学位论文使用授权的声明

本人完全了解吉林大学有关保留、使用学位论文的规定，同意吉林大学保留或向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅；本人授权吉林大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或其他复制手段保存论文和汇编本学位论文。

（保密论文在解密后应遵守此规定）

论文级别：■硕士 □博士

学科专业：计算机技术

论文题目：基于 GoogLeNet 模型的带假结的 RNA 二级结构预测方法

作者签名：李聪　　指导教师签名：张浩

2020 年 5 月 31 日

作者联系地址（邮编）：

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

摘 要

基于 GoogLeNet 模型的带假结的 RNA 二级结构预测方法

RNA 参与着生物体遗传信息的表达、蛋白质的翻译及基因调控等多个生物过程，在生物体内扮演着十分重要的角色。RNA 的结构与其功能紧密相关，只有确定 RNA 的结构才可深入研究 RNA 的功能。因此，研究 RNA 的二级结构具有极其重要的意义。传统的 RNA 结构获取主要有生物实验及计算机预测两种方法。传统的生物实验手段存在成本花费高，时间消耗多等问题。因此，计算机方法成为目前主要的研究手段。

现有的预测 RNA 二级结构的主要方法有：比较序列分析法、动态规划方法及启发式算法等。某种程度上来说，这些方法均取得较好的效果，但也存在着一定的不足。尤其是含假结的 RNA 结构复杂，使得预测难度加大，往往导致预测效果不理想。假结是一种特殊的 RNA 结构单元，也影响着 RNA 的功能。因此，假结的预测一直是 RNA 二级结构研究中的难点问题。

传统的深度学习方法在预测 RNA 二级结构时，虽然取得较好的效果，但随着网络层数的增加，会出现参数量增多、过拟合等问题。GoogLeNet 模型从网络的深度和宽度角度出发，在卷积神经网络模型的基础上进行改进，在提取出更多特征信息的同时，有效提高计算效率。因此，本文使用 GoogLeNet 模型并借助动态规划方法的思想来预测带假结的 RNA 二级结构。本文通过实验将现存的真实 RNA 数据进行处理，利用 GoogLeNet 网络模型从大量的 RNA 序列数据和结构数据中提取出有效的特征，然后对提取出的特征进行预测，得出各个碱基的配对概率。针对碱基的预测结果，利用 RNA 二级结构的定义及动态规划方法的思想，得出每一个碱基配对的概率之和最大的结构，此结构将作为最优的 RNA 二级结构。

本文首先将 GoogLeNet 模型基于 5sRNA、tRNA 数据进行评估，并与其他常见的预测算法进行对比，GoogLeNet 模型得出的预测精确度，其敏感性和特异性比其他算法中最好的预测结果高约 16%。其次该模型基于 tmRNA 数据进行评估，GoogLeNet 模型得出的预测结果比其他算法中最好的预测结果高约 9%。由于假结结构较复杂，因此后者得出的预测精度低，但该方法为后续研究 RNA 的二级

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

结构研究奠定了基础。此外，深度学习算法的性能与数据集大小有关，可推测出随着 RNA 数据量的增加，深度学习方法对 RNA 二级结构的预测精度也会有所提高。

关键词：

RNA 二级结构，假结，GoogLeNet 模型

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

Abstract

Method for predicting RNA secondary structure with pseudoknot based on GoogLeNet model

RNA plays an important role in the body by participating in many biological processes, such as the expression of genetic information, translation of proteins and gene regulation. The structure of RNA is closely related to its function. Only by determining the structure of RNA can we study the function of RNA in depth. Therefore, it is of great significance to study the secondary structure of RNA. Traditional methods of RNA structure acquisition include biological experiments and computer prediction. Traditional biological experiments have problems such as high cost and time consumption. Therefore, computer methods have become the main research methods at present.

The main methods for predicting the secondary structure of RNA include: comparative sequence analysis, dynamic programming, and heuristic algorithms. To some extent, these methods have achieved good results, but there are some shortcomings. In particular, the complex structure of RNA with pseudoknots makes the prediction more difficult, which often leads to poor prediction results. False junctions are a special RNA structural unit which also affects RNA function. Therefore, the prediction of false knots has always been a difficult problem in RNA secondary structure research.

Although traditional deep learning methods have achieved good results in predicting RNA secondary structure, but with the increase of the number of network layers, there will be problems such as increasing the number of parameters and overfitting. From the point of view of depth and width of the network, the GoogLeNet model is improved on the basis of the convolutional neural network model, which can extract more feature information and improve the calculation efficiency effectively. Therefore, this paper uses the GoogLeNet model and uses the idea of dynamic

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

programming method to predict the secondary structure of RNA with false knots. In this paper, the existing real RNA data is processed through experiments. The GoogLeNet network model is used to extract valid features from a large amount of RNA sequence data and structural data, and then the extracted features are predicted to obtain the matching probability of each base. According to the prediction result of bases, the definition of the secondary structure of RNA and the idea of dynamic programming method are used to obtain the structure of the maximum probability sum of each base pairing. This structure will be the optimal RNA secondary structure.

Firstly, the article evaluates the GoogLeNet model based on 5sRNA and tRNA data, and compares it with other common prediction algorithms. The prediction accuracy of the GoogLeNet model is about 16% higher than the best prediction results of other algorithms. Secondly, the model is evaluated based on tmRNA data. The prediction result from the GoogLeNet model is about 9% higher than the best prediction result of other algorithms. Due to the complex structure of the pseudoknot, the prediction accuracy of the latter is low, but this method lays the foundation for the subsequent study of RNA secondary structure. In addition, the performance of deep learning algorithms is related to the size of the data set, and it can be inferred that as the amount of RNA data increases, the accuracy of the deep learning method to predict the secondary structure of RNA will also be improved.

Key words :

RNA secondary structure, false knot, GoogLeNet model

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

目 录

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

第 1 章　绪论

1.1 研究背景与意义

RNA 是一种存在于细胞、病毒中的生物大分子。早期研究指出，RNA 可将 DNA 分子和蛋白质连接起来。其中，蛋白质是生物体发生反应的催化剂，是生物体进行一切生命活动的基础[1]。DNA 承载着生物体内的遗传信息，长期且稳定的存在于生物体内。RNA 则起到中间桥梁的作用，将 DNA 分子中的遗传信息转运到蛋白质中。在 Francis Crick 提出的中心法则中，便指出蛋白质、DNA 和 RNA 三者间的关系[2]。随着研究的深入，人们还发现 RNA 参与着催化、蛋白质的翻译、基因调控及基因表达等诸多过程[3, 4]。这些发现使得研究者对 RNA 在生物体具有的功能有了新的认识，从而引起各界研究人员的广泛关注，加速了人们对 RNA 的研究进程。

有研究表明，RNA 结构与功能有密切的关系，研究二者的关系不仅促进了基因研究的发展，而且揭示了 RNA 在生物体内扮演的重要角色。RNA 分子有自己独特的空间结构，并且 RNA 的结构若发生改变，往往会影响其生物功能的表达[5]；另外，RNA 的干扰功能受到 mRNA 的二级结构的影响[6]；当 RNA 折叠成为一个发卡结构时，会使生物体的转录过程发生终止[7]。因此，研究 RNA 分子的结构，对分析 RNA 的作用、研究 RNA 具有的功能等生物学意义起到关键性的作用，能极大推进人们对 RNA 分子的认知。

目前，研究 RNA 结构的传统方法主要有 X 射线、核磁共振，这些传统的实验方法在一定程度上可以取得较好的结果，但是从成本花费及实验难度[16]的角度而言，这些实验方法的带来的缺陷也是显而易见的。并且，RNA 分子在体外环境中存在不稳定，容易降解，难以结晶等问题[17]。因此，人们运用理论知识并借助计算机等高效、便捷的方法去获取和预测 RNA 的空间结构。然而，RNA 的空间结构存在一定的缺陷：缺少有效的表示方法，直接利用 RNA 的碱基序列去预测 RNA 的空间结构是难以实现的。因此，人们首先利用 RNA 的碱基序列去预测 RNA 的二级结构，进一步预测 RNA 的空间结构。另外，人们在已有预测

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

算法基础上进一步深入的研究，对其优化并提出新的预测模型，使人们可了解更多 RNA 的功能信息，对癌症治疗等疾病提供新的研究思路，对环境保护、医疗等多个领域产生深远的影响。

1.2 国内外研究现状

RNA 分子在生物体的基因调控、遗传信息的传递及细胞的生长等过程中发挥着重要的作用，在生命体的进化中扮演着关键的角色。然而，RNA 分子的结构变化往往会影响 RNA 的生物功能表达。因此，许多科研工作者致力于 RNA 结构的研究。

RNA 分子的结构较复杂，主要分为三层：一级结构是 RNA 的碱基序列组成；二级结构是一种茎环结构，由 RNA 的一级结构利用碱基配对原则形成；三级结构是一种三维的空间结构，由二级结构的基础上进一步折叠形成。大部分的研究人员认为难以直接利用 RNA 的线性碱基序列去预测 RNA 的三级结构，然而 RNA 的二级结构预测相对简单，因此先预测二级结构,然后再预测 RNA 的三级结构；另外，有热力学观点提出，在化学上 RNA 比 DNA 更不稳定，而且 RNA 的三维空间结构容易改变，因此研究 RNA 的结构主要是预测 RNA 的二级结构[8,9]。

RNA 结构中还存在一种特殊的结构，称之为假结。它是一个非常复杂且十分重要的结构。假结结构发现的较晚，其研究历史仅历经几十年的时间。在上世纪 70 年代，假结首次在植物病毒 tRNA 的研究中被发现。这一发现引起了人们的广泛关注，在随后的研究中，研究人员再次在细胞和病毒中发现了假结的存在[10]。假结由许多的拓扑链交叉折叠而成，具有几种不同的拓扑结构，其中最典型的是典型假结或 H-type[11]。假结在 RNA 功能上发挥着重要的作用，其存在于编码区和非编码区[12]。有研究人员发现，假结结构可在生物体内形成端粒酶并作为核糖酶催化的核心[13-15]。假结不仅影响着基因的表达、蛋白质的合成等多个过程，而且也影响着 RNA 分子的稳定性。由于假结在序列较短的 RNA 分子中个数极少，在序列较长的 RNA 分子中存在的个数也较少。因此在早期研究中，由于受到当时技术水平的限制及假结自身结构的特殊性，在预测 RNA 的结构时，科研

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

人员在很多情况下未将假结列入考虑进范围之内。但随着后续研究的加深，研究人员逐渐认识到假结在 RNA 分子中的重要性。因此，即使在预测假结的过程中遇到各种各样的难点问题，科研人员仍将假结预测作为预测 RNA 二级结构的一个重要研究课题。

关于预测 RNA 二级结构的研究，目前已有许多的生物计算方法被应用。例如：动态规划方法、比较序列分析法、启发式算法等诸多方法。下面简单介绍一下：比较序列分析法不是针对一条碱基序列，它是针对多条同源的 RNA 序列进行比较，用于寻找到其中共同拥有的二级结构。这些序列一般来自多器官且同源的序列，加上 RNA 结构具有保守性，因此该方法比其他方法有更好的预测结果。但该方法需要在具有样本量较大且功能大致相同的序列数据中进行研究，因此，针对样本量少或序列间的功能相差较大的情况，比较序列分析法无法取得很好的效果。

动态规划算法（Dynamic programming，简称 DP）与比较序列分析法的不同之处在于：动态规划算法主要用于研究单条的 RNA 序列[18]。通俗来讲，动态规划算法是指将某个问题划分为多个子问题，并利用子问题之间的联系，逐个求解，最终得出结果。即将某条 RNA 序列划分成若干个子序列，类似于若干个单元，对其中的每个单元寻找最优的解，并利用单元之间的关系，最终计算出 RNA 序列的最优结构。针对 RNA 二级结构的预测方法，Nussinov 提出的最大碱基配对算法是最早基于动态规划方法的。该方法认为 RNA 二级结构的最优结果是由单条 RNA 序列相互折叠成碱基对最多的结构构成[19]。但该算法未考虑到不同子结构间的能量差异，最终得出的预测准确度较低。Zuker 认为 RNA 分子的最佳二级结构是在其能量最低的状态下折叠而成，因为该状态下 RNA 的分子结构稳定、不易破坏。因此，Zuker 等人提出了使用最小自由能（Minimum free-energy，简称 MFE）的方法进行研究。RNA 二级结构分为内环、发夹环等多个结构，并且不同的子结构有其固定的能量计算法则，最后使用动态规划方法将子结构进行拼接，得到能量最低的二级结构[20]。但由于实验水平有限，未得出准确的自由能参数，另外能量最低的结构不一定是 RNA 分子的真实结构[21]。

动态规划方法将时间复杂度降到多项式时间，虽然在一定程度上降低了时间复杂度，但是该方法的时间复杂度仍然较高。因此，科研人员提出运用启发式算

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

法，来降低最小自由能模型的复杂度。启发式算法认为 RNA 受折叠机制及环境等原因的干扰，自由能最小的结构不一定是真实的 RNA 二级结构，用于解决假结和长序列等问题。例如，遗传算法通过效仿自然界的进化秩序，遵循适者生存的原则，在解空间中逐步求得最优解。遗传算法在上述问题时，将 RNA 分子中全部可能的茎区集设置对应的编码格式，并借助能量参数的指引预测 RNA 的二级结构[14]。启发式算法除遗传算法外，仍有多种算法被应用在预测 RNA 二级结构的研究中，如 Annealing algorithm [15]、Particle swarm optimization [16]等算法取得的效果也比较好。然而，启发式算法具有随机性，计算得出的结果是否会收敛到全局最优解具有不确定性。此外，机器学习方法中的 SCFG[17]、Hopfield neural network [18]、Artificial neural network [19]和 SVM [20]等方法也逐渐应用致 RNA 二级结构的研究领域，并取得了较好的结果。然而，上述机器学习方法仅针对单类且数据量少的样本进行研究，实际的应用价值不高。

由此可见，关于预测 RNA 二级结构的研究仍需要进一步的探索。随着研究的深入，为解决传统方法的弊端，深度学习方法开始逐渐应用到 RNA 二级结构的研究中，有效提高 RNA 二级结构的预测准确度，为后续研究 RNA 二级结构的预测提供了新的思路。

1.3 论文组织架构

第一章：绪论。本章介绍了 RNA 二级结构预测的研究背景与意义，阐明 RNA 结构与功能之间的关系。讨论了预测假结的国内外研究现状，简要概述了 RNA 二级结构常见的预测方法。

第二章：RNA 及其二级结构相关内容概述。首先，本章介绍了 RNA 相关的基本知识、RNA 的二级结构及组成 RNA 二级结构的几种基本结构，并着重介绍了常见的表示 RNA 结构的方法。最后，介绍了 RNA 的假结，常见的假结类型及假结的功能，为后续的研究奠定了理论基础。

第三章：常用的 RNA 二级结构预测方法。本章介绍了几种传统的 RNA 二级结构预测方法：比较分析法、动态规划法及启发式算法。另外，对各种预测方法的分类进行简单的介绍。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

第四章：基于深度学习的 RNA 二级结构预测方法。本章首先简单介绍了深度学习方法的相关知识，提出使用 GoogLeNet 模型预测 RNA 的二级结构。同时，介绍了一种表示假结结构的方法，最后对 GoogLNet 模型进行介绍。

第五章：基于 GoogleNet 模型的带假结的 RNA 二级结构预测方法。本章主要介绍实验的核心算法流程及步骤：数据处理、设计 RNA 序列的表示方法、及概率和最大的修正算法。

第六章：总结与展望。本章对整个论文的内容进行了总结，并指出本论文存在的不足之处，同时本文对使用深度学习方法预测 RNA 二级结构的研究进行了简要的阐述和展望。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

第 2 章 RNA 及其二级结构相关内容概述

RNA 的结构与功能有着亲密的联系，结构的不同往往表现出不同的功能。因此，研究 RNA 的结构可帮助我们了解 RNA 在生物体内发挥的作用，为 RNA 在各个领域的应用提供理论支撑。

2.1 RNA 的基本知识

2.1.1 RNA 的构成

RNA 是由四种核糖核苷酸构成，经过磷酸二酯键相互连接而成。其中，每一种核糖核苷酸由三部分组成：磷酸、核糖及碱基。嘌呤衍生物和嘧啶衍生物则组成含氮碱基，嘌呤衍生物由 A（adenine、腺嘌呤）和 G（guanine、鸟嘌呤）组成，嘧啶衍生物由 T（thymine、胸腺嘧啶），U（uracil、尿嘧啶）和 C（cytosine、胞嘧啶）组成。其中 DNA 和 RNA 均含有的碱基为：A、G、C，RNA 特有的碱基为 U，DNA 特有的碱基为 T。并且，DNA 与 RNA 在结构上有差异，DNA 的结构是双链螺旋结构，而 RNA 通常为单链分子，但也有极少数的环状或双链 RNA，如在哺乳动物、古细菌中存在大量的环状 RNA[27]。通常情况下，RNA 分子通过碱基互补配对形成 RNA 的二级结构。而互补的碱基配对主要由胞嘧啶 C 和鸟嘌呤 G，尿嘧啶 U 和腺嘌呤 A 之间通过氢键连接而成。其中，A 和 U 形成两个氢键，可表示为 A=U，G 和 C 形成三个氢键，可表示为 G≡C，氢键越多，其表现越稳定，因此 G 与 C 配对和 A 与 U 配对相比前者稳定性更高。除以上两种碱基配对之外，RNA 分子中也存在鸟嘌呤 G 与尿嘧啶 U 配对的情况，但 G 与 U 之间的氢键不稳定，容易发生断裂，故将二者的配对称为摇摆配对（wobble base pairs）。因此若按稳定性排序:GC 最稳定，AU 次之，GU 最不稳定。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

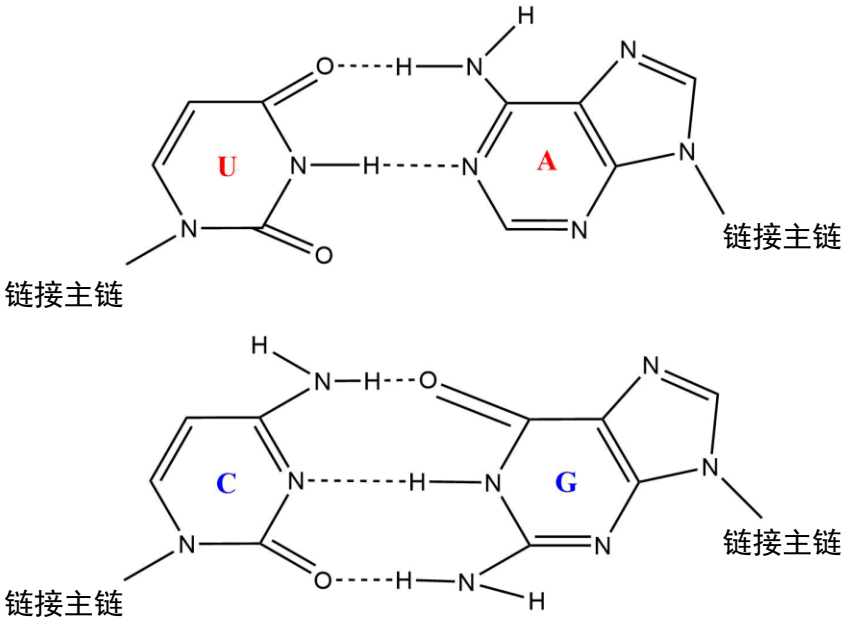


图 2.1 RNA 配对碱基中氢键的构成

通常情况下，单链的 RNA 分子根据碱基互补配对原则，其部分区域经过折叠形成平面的结构，称为 RNA 的二级结构。除 RNA 的二级结构外，RNA 分子还存在更复杂的空间结构，如 RNA 的三级结构、RNA 的四级结构等。另外，RNA 分子可产生多个相互作用力，如碱基互补配对、主链与碱基间、主链与主链间以及孤立氢键间等均可产生相互作用力[28]。以上相互作用力会帮助 RNA 的二级结构形成 RNA 的空间结构，即 RNA 的三级结构。但 RNA 的三级结构易受温度等环境的影响，故 RNA 的三级结构不稳定，研究较困难[29]。在 RNA 三级结构的基础上，RNA 的四级结构是 DNA 和 RNA，蛋白质和 RNA 及 RNA 和 RNA 之间经过相互作用形成。其中，核糖体便是最有代表性的一种 RNA 四级结构。

虽然 RNA 分子存在一级结构、二级结构、三级结构甚至四级结构，但 RNA 的三级及三级以上的结构都十分复杂，目前尚未找到描述其结构的有效表示方法，然而了解 RNA 的二级结构足以了解大部分 RNA 分子的功能。因此，RNA 分子的二级结构仍是目前研究的重中之重。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

2.1.2 RNA 的种类及功能

RNA 是一种存在于细胞质、细胞核中的遗传信息载体，在细胞的蛋白质合成、基因表达等过程中起到关键的作用[30]。若按是否为编码 RNA，可将 RNA 分为两类：一类是编码 RNA，用于蛋白质的翻译，如 mRNA；另一类是非编码 RNA，无法进行蛋白质的翻译，如：tRNA、rRNA 和 lncRNA 等。若按参与蛋白质合成的过程，可将 RNA 根据功能的不同分为三类： mRNA、tRNA 和 rRNA。三者在细胞中约占 RNA 总量的 2%、12%、85%[31]。这些分子在生物体内发挥着极其重要的生物学功能。

在中心法则中，mRNA（信使 RNA）是由 DNA 分子的一条单链经过转录而成的单链 RNA 分子，可将遗传信息从 DNA 上转录下来，作为蛋白质合成时的模板，控制着蛋白质中的氨基酸的排列顺序。此外，蛋白质在翻译的过程中，tRNA（转运 RNA）充当着适配器的功能，与 mRNA 的密码子相对应，形成相应的反密码子，把不同种类的氨基酸运到核糖体并加工成蛋白质，如图 2.2。rRNA 是细胞中数量最多的 RNA，可与多种蛋白质相互结合从而形成核糖体，mRNA 可在其上面形成肽链，充当装配机的作用。

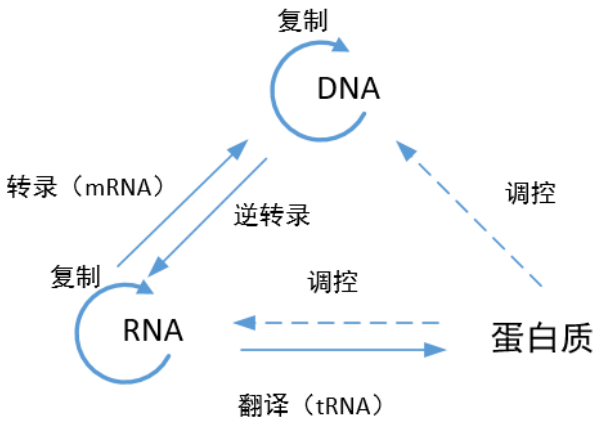


图 2.2 中心法则中 RNA 的作用

由此可见，RNA 分子具有携带遗传信息、传递遗传信息、参与蛋白质的合成等多个过程的生物学功能。除上述三种 RNA 分子之外，仍有许多其他类型的 RNA 在生物体内发挥着重要的功能。如 microRNAs（简称为 miRNAs）是一种非编码 RNA，主要存在于真核生物中。miRNAs 具有调控基因表达的功能，成熟的 miRNAs 可识别靶向目标基因 mRNA[32]。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

2.2 RNA 二级结构及其表示方法

RNA 分子的线性序列构成 RNA 的一级结构，是由 4 种碱基按照不同的组合方式排列而成，一般情况下开始的位置为 5’磷酸末端，结束的位置为 3’羟基末端。RNA 一级结构中碱基的数量不固定，数量范围大概在几个或几十个到成千上万个碱基之间。RNA 的线性序列经过碱基互补配对的原则会形成 RNA 的二级结构，包含茎区、环区等结构。RNA 的三级结构则是由 RNA 的二级结构经过进一步的相互折叠形成的空间结构。由于 RNA 的一级结构无法充分表达 RNA 的空间结构信息，并且 RNA 的三级结构也存在稳定性差等问题。因此，本章主要从 RNA 二级结构入手，介绍 RNA 二级结构及 RNA 的基础知识。

2.2.1 RNA 二级结构

单链的 RNA 分子经过自身的回折，使得 A 与 U，G 与 C 之间的碱基互补配对（G 与 U 也可能发生配对）形成的双股螺旋区，称为 RNA 的二级结构。RNA 的二级结构由以下部分组成：自由单链区域、茎区、环区三种。

自由单链也可称为单链结构，是指在 RNA 分子折叠后，其序列末端未发生配对的区域，称为单链结构。

茎区是指 RNA 结构中两个等长且不相交的区域中全部的碱基进行互补配对及相互之间的连接，最终经过折叠形成的双螺旋区域。由于单个碱基对在 RNA 二级结构中存在不稳定性，因此，通常情况下互补碱基对是成串出现的。

环区是茎区内部的一种区域，由未发生碱基互补配对的单链碱基组成。按照种类的不同，可将环区分为如下类型：凸环、内环、发夹环等多种结构。凸环是指在茎区的一条单链上存在未配对的碱基，另一条单链上均为配对的碱基，并且该单链上存在一个或一个以上的自由碱基。凸环会造成 RNA 二级结构中的茎区弯折，从而影响到 RNA 的三级结构。与凸环不同的是，内环是 RNA 茎区的两条单链上均存在未发生配对的碱基。且内环会受到碱基间互相作用力等因素的影响，形成向外侧突出的圆环结构。若每条单链上存在相等数目的未配对的碱基，内环可称为对称的。发夹环是指 RNA 序列在形成茎区后，中间有一段空出来的未

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

发生配对的碱基串构成的结构，该结构近似发卡形状，因此将该段未配对的结构称作发夹环。另外，发夹环需要四个或四个以上未配对的碱基组成。多分支环又称多重环，是由 RNA 分子结构中两个或两个以上的螺旋茎区的环连接而成，在螺旋之间存在未配对的碱基。

RNA 分子的结构除了上述的茎区、发夹环、凸环、内环等几种结构外，还存在着一种特殊的结构：假结结构。如图 2.3 所示，本图不仅包含以上六种基本结构，而且列举出两种常见的假结结构：H-type 假结、kissing hairpins 假结。此两种假结将在 2.2.2 节详细介绍。

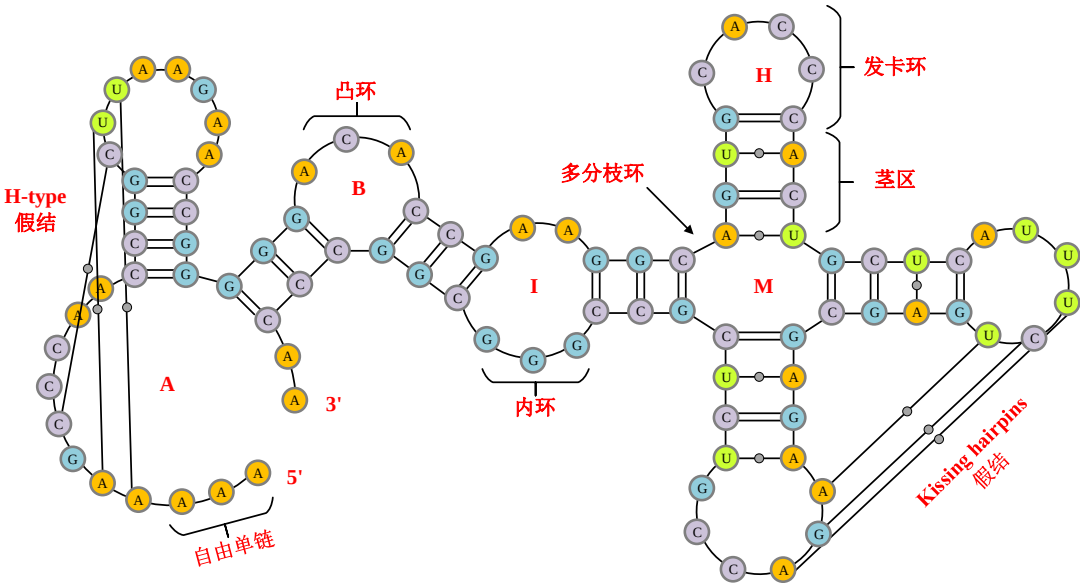


图 2.3 典型的 RNA 二级结构

2.2.2 RNA 二级结构表示方法

由于 RNA 的二级结构比较复杂，为方便研究人员对 RNA 的二级结构有更直观和深入的了解，目前已有多种方法用于表示 RNA 的二级结构。常见的表示方法是点括号表示法、平面图表示法及 CT 文件等表示方法[33]。这几种方法各有利弊，下面逐一介绍：

点括号表示法是使用一对“（”和“）”表示 RNA 序列中发生碱基互补配对的两个碱基，使用“.”表示未发生碱基互补配对的自由碱基。其中“（”表示相对接近 5’端的碱基，“）”表示相对接近 3’端的碱基。针对未产生交叉关系的情况使用“（”和“）”表示，其相应的表现形式主要为两种：相互嵌套

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

方式(())及相互间隔方式()()。若发生交叉情况，则使用中括号或者大括号来表示，如字符串([)]出现了小括号和中括号相互交叉的情况，即小括号中间的部分（“[”）与括号外的部分（“]”）相匹配，即表明该区域存在假结。

```
GCCCCAUGGAGGUGGCUGGGGCCAGCCUCAUGGAGGUGGCUGGGG
...[[[[[[............((((((]]]]]]....))))))..
```

图 2.4　点括号表示法

平面图表示法是 RNA 二级结构表示方法中最常见的一种方法，主要使用图形来描述 RNA 序列（A、U、G、C）的碱基配对情况。其中的圆点图（弧图）表示法是将 RNA 的碱基序列使用水平排列的方式表示，发生碱基互补配对的两个碱基使用半圆弧表示，未发生配对的碱基不标注，弧线接连出现的区域表示茎区，发生交叉情况的表示假结结构。如图 2.5 所示，圆点图十分直观的表示出 RNA 二级结构中碱基的配对情况。

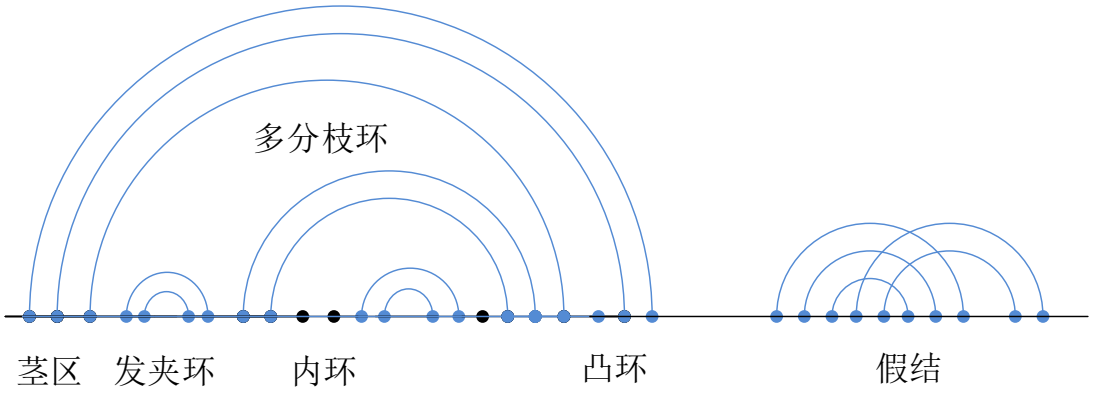


图 2.5　圆点图表示法

另外，CT 文件表示法是由 Zuker 提出的一种 RNA 二级结构表示方法，其主要包括 RNA 的序列信息和 RNA 的结构信息[34]，CT 文件可用于预测 RNA 的类别。关于 CT 文件的处理将在后续章节详细讲述。

2.3 RNA 假结结构

假结（pseudoknots）是 RNA 中一种特殊的结构，在 RNA 二级结构的研究中扮演重要的角色，因此针对假结的探究一直是预测 RNA 二级结构领域的研究重点。假结的定义如下：在某条 RNA 序列中，若存在 a ，b，c ，d(a＜b＜c＜d)四个位置的碱基，其中 a 与 c 配对，b 与 d 配对，则称(a ,c) 和(b ,d)碱基对形成

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

的结构为假结结构。

假结可以有多种不同的折叠排序，因此研究人员猜想假结的结构种类不止一种。针对这种猜想，1990 年，Pleij 在理论上提出假结的结构种类为 14 种，其中有 10 种是各种环与环之间通过碱基互补配对形成的，另外 4 种是各种环与自由单链通过碱基间相互配对形成的。然而，以上假结类型是基于理论提出的，在实际的研究中，经常受到热力学或结构化学等环境的影响，许多假结无法出现。并且实验数据有限，无法为假结提供可靠的参数；另外，无法通过实验证明这 14 种假结结构中的某些结构是真实存在的。因此，下面将对两种比较常见的假结结构进行详细的介绍。

H 型假结是所有的假结结构中比较简单及常见的一种结构，它是由发夹环未配对的碱基与茎环之外的自由单链的碱基发生碱基互补配对形成的。这种 H 型假结一般是由 2 个茎区和 3 个环区组成。其中茎区使用 S1 和 S2 表示，环区使用 L1、L2 和 L3 表示。其中 S1 和 L1 分别为接近 5’端的茎区与环，S2 和 L2 分别为接近 3’端的茎区与环，L3 是将不同茎区连接起来的环。

除 H 型假结之外，假结结构仍存在一种结构一直受到人们的广发关注：kissing hairpins 假结，kissing hairpins 假结的结构比 H 型假结的结构更加复杂。kissing hairpins 假结表示发夹环与发夹环形成的假结，它是由两个发夹环的碱基发生碱基互补配对形成的。kissing hairpins 假结是由 3 个茎区和 5 个环区组成，其中茎区使用 S1、S2 及 S3 表示，环区使用 L1、L2、L3、L4、L5 表示。L2 和 L4 环区的长度可为 0，其他环区至少包含一个碱基。由于 kissing hairpins 假结结构较复杂，在计算中研究人员通常将一个 kissing hairpins 假结划分成两个 H 型假结来计算[34]。如图 2.6 表示，其中图片的左边为 H 型假结，图片的右边为 kissing hairpins 假结。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

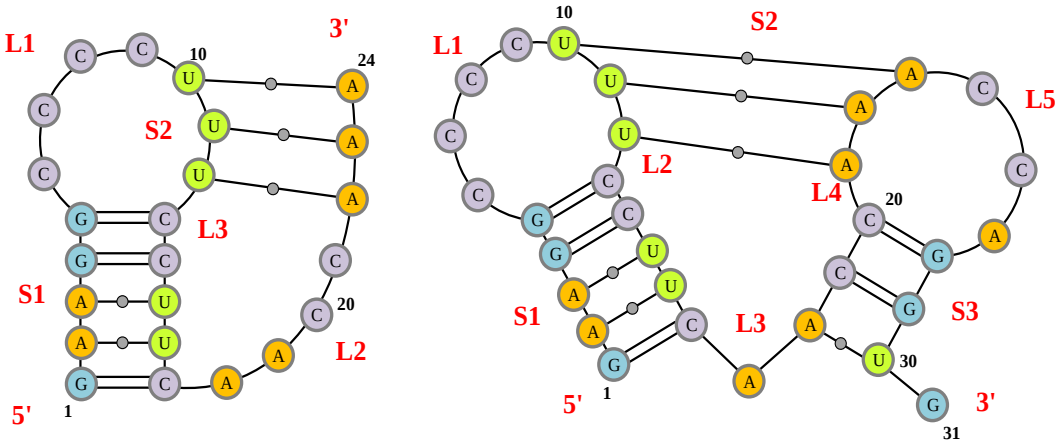


图 2.6 H-type 假结（左）及 kissing hairpins 假结（右）

假结由于其特殊的配对位置，在 RNA 表达的生物学功能中起到关键作用。因此，假结的研究逐渐成为研究人员的一个热门课题。假结在不同的 RNA 生物大分子中扮演着不同的角色，参与着一系列的生物过程，如假结可催化、可形成端粒酶以及可进行自我剪接的内含子，并且假结在改变细胞、病毒的基因表达中也发挥着关键的作用。早期关于假结的预测精确度不太理想，为提高关于 RNA 二级结构预测的精确度，研究人员对现有的预测算法进行改进和优化，并且不断提出关于结构预测的新模型。本文提出一种新型的基于 GoogleNet 模型的 RNA 结构预测方法，为 RNA 二级结构的研究提供一种新的思路。

2.4 本章小结

本章主要对 RNA 及其二级结构的相关内容进行了介绍，首先概述了 RNA 的构成和 RNA 的种类及功能，其次主要介绍了 RNA 的二级结构和几种 RNA 二级结构的表示方式，最后对 RNA 的假结进行论述，主要介绍了两种常见的 RNA 假结结构，为本文的后续研究做了理论支撑。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

第 3 章 常用的 RNA 二级结构预测方法

3.1 比较序列分析法

RNA 二级结构预测的方法有多种，如比较序列分析法、动态规划方法等。比较序列分析法是其中准确度较高的一种，可用于预测假结及某些三级结构。比较序列分析法依据结构保守性高于序列保守性的原则，通过已知的同源序列推测出新序列的结构及生物学作用。即使某些同源 RNA 的一级序列排序不同，但若其结构相似，则其表达的生物学功能也相似。例如，不同序列的 tRNA 分子构成的二级结构均为三叶草形态。

利用比较序列分析法进行 RNA 二级结构预测时，其主要由序列比对和结构预测两个过程构成。根据这两个过程不同的执行顺序，可将比较序列分析法分为三类：1.先进行序列比对后进行结构预测 2.先进行结构预测后进行序列比对 3.二者同时进行[35]。此三类过程如图 3.1 所示：

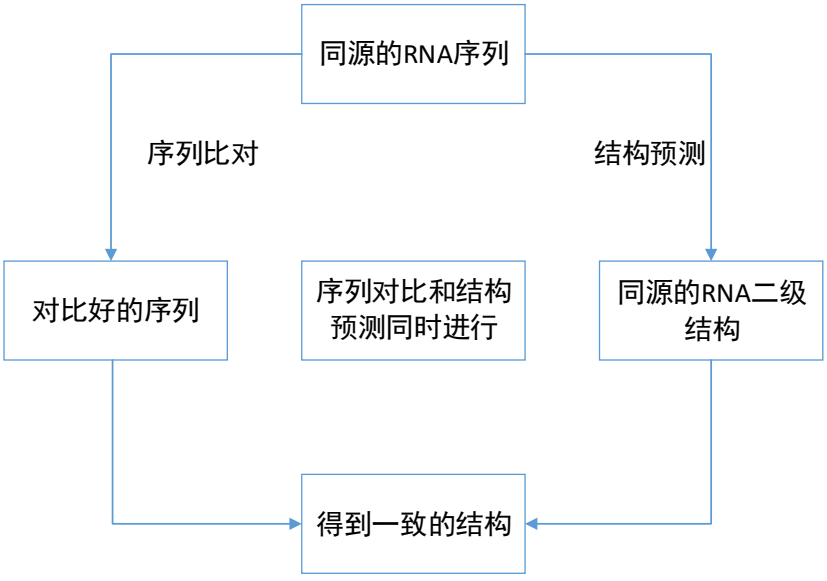


图 3.1 比较序列分析的三种方法

1. 先进行序列比对后进行结构预测方法利用序列比对工具，如 ILM[36]、RNAalifold[37] 和 Pfold[38]等得出相应的序列比对结果，然后进行结构预测得出多种不同 RNA 序列间共有的结构。该方法以结构保守性比序列保守性高为前提，即预测结果依赖于序列比对的精确性。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

2. 先进行结构预测后进行序列比对将数据集中的每条 RNA 序列的结构进行预测，并针对得出的预测结果进行分析对照，以此得出 RNA 序列的保守结构。该方法也需要相应的结构预测软件（如 RNAfold，mfold），但这些结构中可能不存在真实的 RNA 结构。该方法的准确性依赖于结构比对的准确性，另外结构预测比序列比对难度更大，无法保证可以得出很好的准确性[39]。

3. 结构预测和序列比对同时进行，该方法中 Sankoff 方法是最有代表性的，将序列比对方法及最大碱基对方法整合完成序列比对及结构预测操作。该方法需将二者同时进行，因此该方法的时间复杂度和空间复杂度与之前两种方法相比明显增加，花费成本较高。因此该方法无法预测结构复杂的 RNA，可尝试预测结构简单序列长度较短的 RNA[40]。

比较序列分析法是在相应的数学模型基础上建立的，主要包括两种模型：随机上下文无关语法模型（SCFG）和共变模型（CM）。其中 SCFG 无法预测假结结构，共变模型可用于预测假结结构。

3.2 动态规划方法

动态规划是一种从某一个问题的多个可行解中寻求最优解的方法。动态规划方法的核心思想是将某一个阶段的问题分为多个子阶段问题，利用每个子阶段的关系对问题逐一求解。动态规划方法解决了计算过程中存在的冗余问题，将需消耗大量时间的复杂问题简化为在多项式时间内解决，解决了子问题反复运算的问题，提高了计算效率。目前，动态规划方法在生信、计算机等领域有着普遍的应用。

动态规划方法借助分治法的概念，将某个原问题解析为多个具有一定关系的子问题，分别对各个子问题求出相应的解，然后子问题的解组合求出原问题的最优解。与动态规划方法不同的是，分治法用于解决多个子问题不重叠的情况。分治法通过将问题分解为许多不重叠的子问题并求出相应的解，然后将这些子问题组合作为最终的解。然而，动态归法方法主要用于解决子问题重叠的情况。当不同的子问题间存在子子问题相同的情况时，动态规划方法会将每个子子问题进行求解，并将得到的结果保存，无需再重复运算，从而提高计算的效率。从根本来

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

讲，动态规划方法是将空间替换时间，但其在计算过程中需存储不同的状态，故就空间复杂度而言，动态规划方法高于其他算法。

在对原问题求解时，动态规划方法要在对原始问题有充分了解的前提下对其进行分析，通过分解多个子问题最终找到最优解。然后，动态规划方法需满足无后效性及最优化原理这两个条件。无后效性是指某一阶段的状态确定后，其后续演化便不再受之前状态的限制。最优化原理是无论过去的状态如何，仅针对当前的状态进入下一步决策。在预测 RNA 的二级结构研究中，动态规划方法也逐渐被运用致该问题的研究。其中 Nussinov 和 Zuker 等研究人员分别对该方法进行了扩展和改进，下面将逐一介绍这些方法。

（1）最大碱基配对算法

最大碱基配对算法是由 Nussinov 提出的，此算法针对 RNA 二级结构预测方法的发展起到十分重要的影响。该算法的基本假设是碱基间的氢键会使得发生碱基互补配对的碱基结合更紧密，因此若某条 RNA 序列中发生碱基互补配对的碱基对越多，则氢键越多，结构也将越稳定。因此，最大碱基配对算法将某条 RNA 序列中形成碱基对个数最多的结构状态作为最终真实的 RNA 二级结构。

最大碱基配对算法在预测 RNA 二级结构的研究中使用动态规划的基本思想，将整个结构预测问题分解为多个子问题。假设某一条 RNA 序列为 $x=(x_1,\ldots,x_n)$，其中 $x_1,\ldots x_n$ 表示 RNA 的四种碱基（A、U、G、C），n 表示 RNA 序列的长度。最大碱基配对算法使用动态规划方法的思想将整个 RNA 序列的结构预测问题转为计算其子序列问题，然后将得到的所有预测结果组合在一起作为最终的结果。另外，$\gamma(i,j)$ 用于表示第 i 个和第 j 个碱基是否发生配对的情况，若二者发生了碱基互补配对，则 $\gamma(i,j)=1$，否则 $\gamma(i,j)=0$。$\partial(i,j)$ 表示从第 i 个位置到第 j 个位置的子序列中形成的最大碱基配对个数。在计算 RNA 序列的最大碱基对时，通常我们要考虑四种情况，并将其中的最大值作为最终的计算结果：（1）若第 i 个位置到第 j 个位置的子序列中碱基 i 为自由碱基，则该子序列的最大碱基对数目 $\partial(i,j)$ 等于 $\partial(i+1,j)$。（2）若第 i 个位置到第 j 个位置的子序列中碱基 j 为自由碱基，则该子序列的最大碱基对数目 $\partial(i,j)$ 等于 $\partial(i,j-1)$。（3）若第 i 个位置到第 j

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

个位置的子序列中碱基 i 和碱基 j 均为自由碱基，则该子序列的最大碱基对数目 $\partial(i,j)$ 等于 $\partial(i+1,j-1)+\gamma(i,j)$。（4）若第 i 个位置到第 j 个位置的子序列中碱基 i 和碱基 j 分别与二者间的碱基发生配对，则该子序列的最大碱基对数目 $\partial(i,j)$ 等于 $\partial(i,m)+\partial(m+1,j)$。由上述情况可得出如下的计算公式：

$$\partial(i,j)=\max\begin{cases}\partial(i+1,j),\\ \partial(i,j-1),\\ \partial(i+1,j-1)+\gamma(i,j),\\ \max_{i<m<j}\{\partial(i,m)+\partial(m+1,j)\}\end{cases} \quad \cdots\cdots（3.1）$$

根据以上公式从 1 到 n 进行迭代计算，即 $i=1, j=n$ 时，$\partial(i,j)$ 的值便为整个 RNA 序列形成的最大碱基对数。最后，通过回溯过程得到最终的预测结果。然而，由于最大碱基配对算法在预测 RNA 的二级结构时，考虑的情况比较简单，因此得出的预测精确率并不高。

（2）最小自由能算法

Zuker 提出的最小自由能算法(简称 MFE)适用于预测单条 RNA 二级结构的情况。与最大碱基配对算法不同之处在于自由能算法使用能量数据进行相关的计算。最小自由能算法是依靠自由能的大小来判定。由于 RNA 中发生碱基互补配对的碱基依靠氢键进行连接，破坏 RNA 分子中的氢键需要一定的能量，而这些需要消耗的能量称为自由能。由于 RNA 分子中能量越低其对应的结构稳定性越高，并且某条 RNA 序列的结构中的茎区结构会产生值为负的自由能，而未形成碱基对的结构如凸环、凹环等结构会产生值为正的自由能。若某个 RNA 分子的茎区结构中碱基对个数越多，则其对应的自由能越小。因此，在预测 RNA 的二级结构时，通过 MFE 算法得到最小自由能，其对应的结构作为 RNA 分子真实的二级结构。

自由能的计算不仅与发生碱基互补配的碱基对有关，而且与碱基对的类型、RNA 分子中的茎区、环区等不同结构相关。在计算 RNA 二级结构的最小自由能时，Zuker 假定 RNA 分子的茎区、环区等不同的结构是完全不受彼此影响的，将实验得到的各个结构的自由能通过动态规划方法组合起来，得出的自由能之和最小的结构即为 RNA 分子真实的二级结构。由综上可知，最小自由能算法通过计算各个结构的自由能来获得整体自由能最小的值。计算公式如下所示：

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

$$E_{i,j}=\min\begin{cases}E_{i+1,j-1}+\alpha_{i,j}\ ,\\ \min(E_{i+k,j}+\beta_k)\ ,\\ \min(E_{i,j-k}+\beta_k)\ ,\\ \min(E_{i+k,j-1}+\gamma_{k+1})\ ,\\ \min(E_{i+k,j}+E_{i,j-1}+\varepsilon_{k+1,i-j})\ ,\\ \delta_{j-i}\end{cases} \quad \cdots\cdots(3.2)$$

其中从第 i 个碱基至第 j 个碱基子序列最小自由能的值使用 $E_{i,j}$ 表示，第 i 个碱基和第 j 个碱基发生碱基互补配对时产生的自由能使用 $\alpha(i,j)$ 表示。另外，β、γ、ε、δ 分别表示凸环、内环、多分支环及发夹环等多个基本结构产生的最小自由能[38]。最后，经过过迭代运算得到的 $E_{1,n}$ 值作为整条 RNA 序列的最小自由能，并使用回溯法得到带有最小自由能的 RNA 二级结构。

另外，有相关研究人员在最小自由能方法研究的基础上，提出了可预测 RNA 分子中带假结结构的方法如的 PKNOTS、NUPACK、LP 等算法。这些算法不仅计算了基本结构的自由能，同时计算了假结产生的自由能。虽然这些算法都各有各的优点，但这些算法也提高了计算整体上时间和空间的复杂度。

3.3 启发式算法

之前提到了两种关于预测 RNA 二级结构的方法：比较序列分析法和动态规划方法，这两种方法各有各的优点，也各有各的不足。然而，与之前两种方法不同之处的是启发式算法不仅计算灵活而且效率更高，并且可用于预测 RNA 序列的假结结构等相关问题。目前，随着计算机技术的日益成熟，启发式算法中的遗传算法、贪婪算法等多种算法被应用到预测 RNA 二级结构的研究中。本章将简要描述启发式算法的贪婪算法、遗传算法。

（1）贪心算法

贪心算法又称为贪婪算法，该算法在预测 RNA 二级结构的问题研究中寻求价值最高的结果，贪心算法在求解过程中将全局最优解中的局部最优解作为最终的结果[33]。在预测 RNA 二级结构的研究时，贪心算法将茎区作为基本单元，在所有的茎区中选取可在最大程度上降低初始结构的茎区，然后添加进初始结构中。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

在选取能量低的茎区时不仅要考虑茎区的能量而且要考虑和茎区相邻环的能量，二者的能量之和越低则对降低初始结构能量产生的影响越大。将以上过程进行重复操作，直到初始结构能量稳定为止，此时得到的结果便作为最终的预测结果。

贪心算法无法保证对所有问题求取全局最优解，但该算法有其特定的作用。贪心算法在求解问题时需要限定搜索的范围，一旦不对问题的搜索范围加以限定，该算法将求解出全部的解，会造成资源的浪费和时间的消耗。贪心算法不是一步完成的，它需要一步步的过程，若在计算的过程中本不应该形成的茎区被选择的话，将对后续的选择造成很大的影响，从而使得最终的预测结果精确率不高。

（2）遗传算法

美国 Holland 研究学者在 1975 年提出了遗传算法(Genetic Algorithm，GA)。遗传算法将达尔文进化论及遗传变异的学说作为根本，其目的在于求解全局最优解。遗传算法主要模仿自然界中存在的“适者生存，优胜劣汰”的生物进化过程，并进行物种遗传学上的选择、交叉、基因变异等过程，与此同时在全局搜寻出最优个体，得出最终的结果。关于遗传算法的流程如图 3.2 所示：

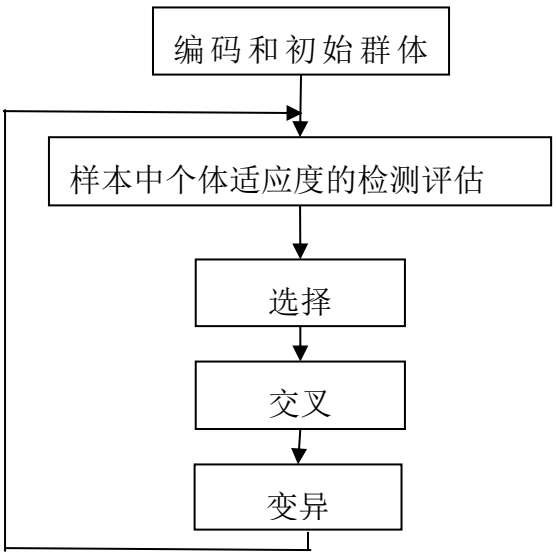


图 3.2 遗传算法基本流程

遗传算法在预测 RNA 的二级结构时，需要经过如下步骤：第一步构建初始群体:将所有 RNA 序列中的茎区构建一个茎区容器，采用随机的方式，从中选择一个茎区作为我们的初始茎区，并使用螺旋区堆积法将相容茎区添加至起始茎区

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

中作为初始的 RNA 二级结构，重复以上操作，形成一个初始群体。选取的茎区中若出现交叉的茎区，可用于预测假结结构。以自由能为适应度函数，使初始群体一步步进化，直到能量达到最低且结构稳定为止[41]。第二步交换：根据适应度从初始群体中选取两个父本并将父本的两个结构从一个、两个或多位点中进行交叉操作。然后按照轮盘赌的方式从父代组成的茎区容器选取合适的茎区，以便子代形成较好的结构。第三步为变异：变异是以剔除尽可能能量大的茎区为准则，将茎区中自由能值为正的某些茎区剔除，而负的茎区依然遵循轮盘赌方法，然后从茎区池中添加符合条件的相容茎区。第四步选择：选择是指从按照适者生存法则从初始群体、交换和变异个体中选取 n 个个体成为下一代继续新的循环。

3.4 算法优缺点分析

比较序列分析法是预测精确度很高的一种算法，并且可以预测含有假结的 RNA 二级结构。然而，即使比较序列分析法的精确率较大，但该方法在很多情况下仍不适用。例如，比较序列分析方法需要大量的同源 RNA 的序列信息，若在同源样本量太少的情况下，该方法将无法得出非常准确的预测结果。另外该方法需占用大量的时间和空间，因此比较序列分析方法的应用受到一定的限制。

动态规划方法在预测 RNA 二级结构时可获得能量最优的 RNA 二级结构。然而，动态规划方法的准确性依赖于其中的能量参数。若能量参数较少或不准确，其预测结果将受到很大的影响。另外，针对序列较长的数据，动态规划方法也需消耗大量的计算时间。

启发式算法在预测 RNA 二级结构时将茎区作为基本单位，并不针对全局进行搜索，仅对其部分茎区进行搜索来寻求最优解。因此，该方法极大的降低了搜索量，可留出更多的计算时间进行假结预测。然而，启发式算法随机性较大，无法保证得出的局部最优解与全局最优解相似。

3.5 本章小结

本章我们主要对关于 RNA 二级结构预测的方法进行了介绍和分析，包括：

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

比较序列分析法、几种常见的动态规划方法及几种常见的启发式算法。比较序列分析法虽然预测精确度较高，但需要大量的同源序列数据，且计算量较大。详细介绍了动态规划算法中的几种方法，其中本文的修正方法将以最大碱基配对算法作为理论基础。另外，启发式算法的运算比较灵活，可用于预测 RNA 的假结。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

第 4 章 基于深度学习的 RNA 二级结构预测方法

4.1 现有深度学习简述及存在问题

人工智能（ArtificialIntelligence，缩写为 AI）的概念是在 1956 年达特茅斯的会议上提出的，其目的是针对人类的智能，创造出有理解能力的机器，这将为人类的生产、生活和科技进步带来巨大的推进作用。目前，人工智能的主要实现方法包括机器学习、深度学习方法等。其中，机器学习方法的目的是将获得的原始数据进行数据处理和分析，最后发现数据潜藏的规律。当数据经验有了一定的累积后，机器学习方法会通过不断改进自身的方式，使其预测能力不断提升。机器学习方法将已收集的数据进行操作、预处理，对数据进行人工的特征提取和特征选择，并使用 SVM、LR 等机器学习模型训练处理好的数据，以此解决预测或识别问题。机器学习方法的主要思路：数据预处理、特征提取、特征选择及预测或识别问题。其中，将前三项称为特征工程，特征工程比算法选择更能影响模型的准确率。然而，许多的特征工程需要有专业知识的科研人员完成，并且选取的特征十分敏感难以应用到其他任务中。因此，特征工程是十分耗时耗力的。

针对上述特征工程的困难，科研人员尝试使用计算机手段仿效特征工程。由此，产生了深度学习方法。常见的深度学习模型主要有循环神经网络、卷积神经网络等。深度学习中模型的输入是现有的原始数据，模型将原始数据从底层特征一层又一层地抽象到高层特征，最终得到所需要的特征表示，再将得到的特征映射到目标上。以上整个过程均由计算机参与，无需任何人工操作。

目前为止，深度学习不仅在农业、医疗、自然语言处理等多个领域得到应用，而且科研人员也将深度学习方法应用到了 RNA 二级结构的预测研究中。此前，本人所在实验室的相关科研人员提出一种使用卷积神经网络模型来预测 RNA 二级结构的方法，该方法将 RNA 数据中的隐含特征提取出来，并将得到的预测结果作为中间结果，使用基于动态规划思想的修正方法对结果进行修正，最终得到 RNA 的最优二级结构。经实验表明，该方法得到的预测效果较好。虽然该方法在预测 RNA 二级结构的过程中取得了较好的效果，但是该实验仍然存在一些问

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

题。例如，该实验采用仅含三个标签的点括号表示法表示 RNA 的结构信息，但该点括号表示法无法有效的表示 RNA 结构中的假结结构。为解决以上问题，本文采用含七个标签的点括号表示法，用于表示含有假结的 RNA 二级结构，并选取了 GoogLeNet 模型进行相关实验研究。

4.2 GoogLeNet 建模

针对上述提到的机器学习和深度学习算法，虽然这些方法在许多应用领域取得较好的效果，但是针对一些更复杂的情况，仍存在特征提取不全面而影响网络的性能等问题。因此，面对更复杂的实际问题时，科研人员想从网络规模角度来提高网络的性能，即扩大网络的深度和宽度，从而增加网络的规模。但随着网络规模的增大，会引起网络的参数增多的问题，使得计算的难度加大。若实验中用到的数据集有限，会更容易导致过拟合的情况。此外，若网络的层数越深，更易造成梯度消失的现象，难以优化模型。因此，本文提出了将 GoogLeNet 模型运用致预测 RNA 的二级结构研究中。

GoogLeNet 方法是 2014 年在 ImageNet 挑战赛上提出的，与其一同产生的还有 VGG 方法。GoogLeNet 方法的创新性在于从网络的结构入手，增加网络的深度和宽度的同时，提高网络的计算效率。有研究表明，将稀疏矩阵聚类为相对密集的子矩阵可提高计算的性能，如同人的的大脑可看做由神经元重复堆积而成。GoogLeNet 方法正是借助这种思想，创新的搭建了一个稀疏且计算性能很高的“基础神经元”结构，称之为 Inception 网络结构。Inception 网络结构经过了 Inception v1、Inception v2、Inception v3 及 Inception v4 等多个版本，这些版本各有各的优点，但也存在相应的缺点。本文使用 GoogLeNet 的 Inception v1 版本，下面将对其进行介绍。如图 4.1 所示，该结构为 Inception v1 的原始版本。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

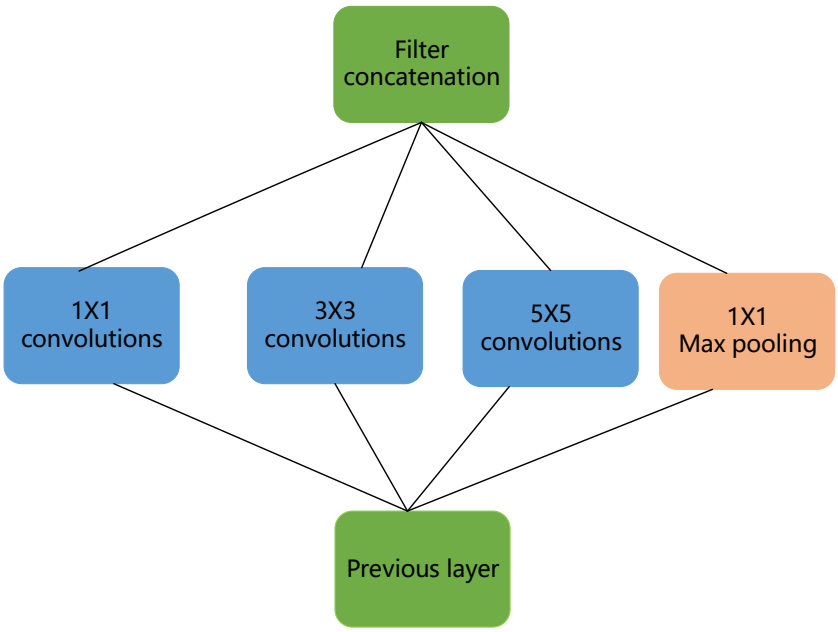


图 4. 1　Inception v1 初级结构

通过将卷积神经网中的多个卷积核（1×1，3×3，5×5）和池化层（3x3）以并行的方式放在同一层，卷积和池化之后的尺寸相同，将各个通道相加，这样不仅增加了网络的宽度，同时也增加了网络对尺度的适应性。不同的卷积核可挖掘出输入中存在的每一个细节特征。另外，池化操作主要用于减少空间大小。因此，Inception 结构无需人为确定是否添加卷积层、池化层，网络可自行决定是否需要什么样的参数。然而，这个 native Inception 仍然存在很大的缺陷：将以上三个卷积层与池化层拼接后输出的 feature map 数量较大，若网络层数不断增加，模型将变得十分复杂，难以训练和优化。并且，5×5 的卷积核会带来计算量过大导致特征图厚度过大的情况。

为解决上述问题，Inception 在此基础上进行优化，提出一种新的 Inception 网络结构，即在 3×3 卷积核和 5×5 卷积核的前面以及 max pooling 的后面分别加上 1×1 的卷积核，这样不仅减少了维度，而且会大大减少参数的个数。例如，某一层的输出为 100×100×128，一种方式通过具有 256 个通道的 5x5 卷积层，另一种方式使其先经过有 32 个通道 1×1 卷积层，再经过具有 256 个输出的 5x5 卷积层。经计算，两种方式得到的输出数据虽然一致，但后者参数数量比前者的参数数量减少了大约 4 倍，大大提高了计算效率。Inception v1 的网络结构如图 4.2 所

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

示。整个 GoogLeNet 模型的层数为 22 层，其中使用 9 个 Inception 结构，由于网络的层数加深，可能出现梯度消失的情况。为避免网络层数过多引起梯度消失的问题，GoogLeNet 网络在训练的过程中，在其中间层添加了两个辅助的 softmax(辅助分类器),使网络的梯度信号可以反向传播，对整个 GoogLeNet 模型的训练大有帮助[42]。

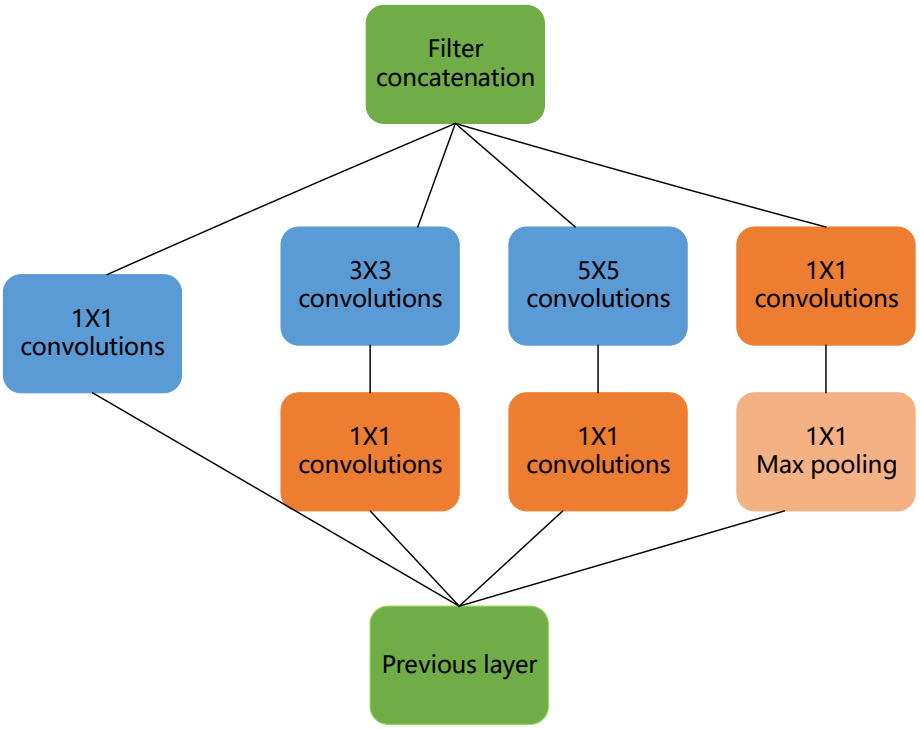


图 4. 2　Inception v1 结构

4. 3 基于深度学习的假结结构表示方法

若存在某条 RNA 序列，RNA 的部分区域如图 4.3 所示，该图上方表示 RNA 的部分区域不含假结，该图下方表示 RNA 的部分区域含有假结，若使用三个标签的点括号方法表示，结果均为“(( ))”。因此传统的点括号方法无法准确表示某条带假结 RNA 的真实结构，这将导致无法准确有效表示实验的分类问题，会对实验的精确度及后续的实验研究造成很大的影响。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

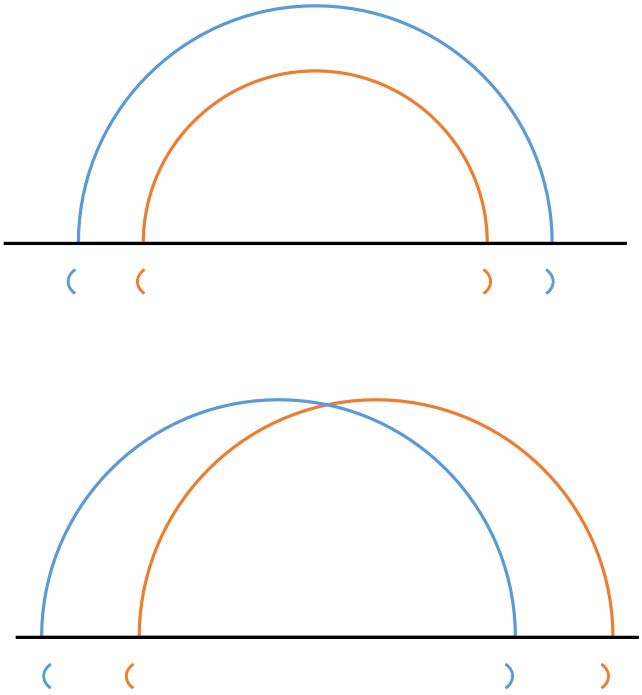


图 4.3 点括号表示法

针对以上问题，本论文采用较复杂的点括号表示法--即用“.”、“(”、“)”、“[”、“]”、“{”、“}”的组合表示 RNA 的结构信息[43]。其中，“(”、“)”表示不含假结的结构，“[”、“]”、“{”、“}”表示含有假结的结构。七个标签点括号表示法的大致思想是将总问题化成各个子问题进行解决，最后将子问题获得的解整合作为总问题的解。即将 RNA 中的假结结构逐步拆分为不含假结的结构，最后再将其整合起来表示完整的 RNA 结构。步骤如下：在某条 RNA 序列中，未发生碱基互补配对的标记为“.”，按序列顺序将第一对发生碱基互补配对的碱基对标记为“(”、“)”，找到所有与其不相交的碱基对标记为“(”、“)”，去除这些结构。之后找到按序列顺序将第一对发生碱基互补配对的碱基对及与其不相交的碱基对标记为“[”、“]”，以此类推，最后得到 RNA 结构的标签表示，如图 4.4 所示。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

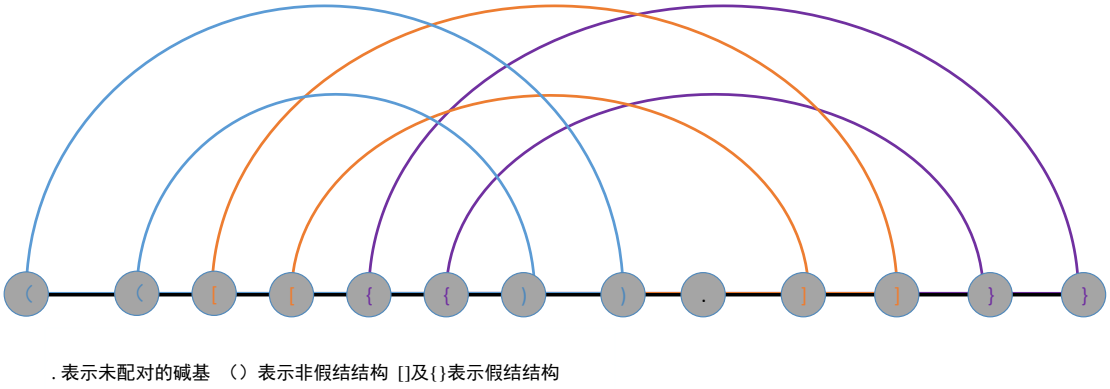


图 4.4 RNA 的标签表示方法

4.4 本章小结

本章首先对机器学习与深度学习的相关知识进行了简单的描述，阐述了机器学习方法存在的不足，以及深度学习方法存在的优势。另外，本章介绍一种七个标签的点括号表示法，有效表示带假结的 RNA 结构信息。最后，本章对 GoogLeNet 模型的 Inception v1 版本进行了相关的概述，在面对网络规模扩大的情况下，阐明其特有的优势。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

第 5 章 基于 GoogLeNet 模型的带假结的 RNA 二级结构预测方法

5.1 数据收集及处理

本论文将 Mathews lab 中的数据集作为本次实验的数据，目前已有许多科研人员利用 Mathews lab 数据库进行 RNA 二级结构的相关研究。本文选取的 RNA 结构数据集共包含 RNA 序列 3957 条，总共 10 个家族。下表分别为 10 个家族包含 RNA 的数量：

表 5.1 各家族 RNA 数量

RNA 家族	Number
5sRNA家族	1283
srpRNA家族	928
tRNA家族	557
tmRNA家族	462
RNasePRNA家族	454
16sRNA家族	110
grp1RNA家族	98
telomeraseR家族	37
25sRNA家族	35
grp2RNA家族	11

本文从 10 个家族中选取 5sRNA、tmRNA、tRNA 三个家族进行研究，其中 5sRNA、tRNA 不含假结结构，而 tmRNA 含有假结。以上三个家族数量较多且分布均衡，因此将其作为实验对象。通过对 5sRNA、tmRNA、tRNA 数据进行分析，三个家族均存在相似或相同的数据，其中相似数据是指序列相同但名称不同的数据。这些数据会影响以后实验的精确性，因此需要去除数据中的相似序列，即对数据进行去冗余操作。经过去冗余操作后 5sRNA、tmRNA、tRNA 数据分别

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

为 1059 个、486 个及 378 个。

原始数据集是使用 CT 文件格式表示 RNA 的二级结构，CT 文件不仅包含着数据集中 RNA 的序列信息及结构信息，而且包含与本实验无关的信息。其中，使用碱基“A”、“U”、“G”、“C”组合形成的序列表示 RNA 的序列信息，使用“.”、“(”、“)”、“[”、“]”、“{”、“}”点括号表示法表示 RNA 的结构信息。因此，本文需要将可用的 RNA 序列信息和结构信息抽取出来。CT 文件表示的 RNA 二级结构如图 5.1 所示，在 CT 文件的第一行包含 RNA 序列的长度及名称等描述信息，数字 M 表示某条 RNA 序列的长度，数字 M 后面的字符串表示该 RNA 的名称。除去首行，CT 文件对的每一行都包括 6 列数据：第 1 列和第 6 列表示该条 RNA 序列碱基的位置；第 2 列表示该条 RNA 序列从起始至结束位置各个碱基的排列顺序；第 3 列表示该 RNA 序列中与某一位置碱基相邻的前一个碱基所在位置；第 4 列表示该 RNA 序列中与某一位置碱基相邻的后一个碱基所在位置；第 5 列表示该 RNA 序列中与该位置碱基是否发生碱基互补配对的碱基，其中数字非“0”表示该位置碱基与第 1 列或第 6 列相应位置的碱基发生了碱基互补配对，数字“0”表示该位置碱基与第 1 列或第 6 列相应位置的碱基未形成碱基对。

```
115 5s_Achromobacter-denitrificans-1
1   U   0   2   0   1
2   G   1   3   115 2
3   C   2   4   114 3
4   C   3   5   113 4
5   U   4   6   112 5
6   G   5   7   111 6
7   A   6   8   110 7
8   C   7   9   109 8
9   G   8   10  108 9
10  A   9   11  107 10
11  C   10  12  0   11
12  C   11  13  0   12
13  A   12  14  0   13
14  U   13  15  0   14
15  A   14  16  0   15
16  G   15  17  68  16
17  C   16  18  67  17
18  G   17  19  65  18
19  A   18  20  64  19
20  G   19  21  63  20
21  U   20  22  62  21
22  U   21  23  61  22
23  G   22  24  60  23
24  G   23  25  0   24
25  U   24  26  0   25
26  C   25  27  0   26
27  C   26  28  0   27
28  C   27  29  0   28
29  A   28  30  55  29
30  C   29  31  54  30
```

图 5.1 CT 文件

RNA 的序列信息在 CT 文件的第 2 列，可直接抽取出来。本文使用七个标

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

签的点括号表示法，因此在提取 RNA 的结构信息时，将 CT 文件的第 1 列和第 5 列的数字相互比较，若第 5 列的数据为“0”，表示未发生碱基互补配对，将其标记为“.”，若第 5 列数字非“0”，如果第 5 列数字大于第 1 列的数字，表示第 5 列数字对于位置上的碱基在第 1 列相应位置的碱基之后，则使用“（”表示，相反使用“）”表示。同样地，中括号和大括号也使用相应规则进行表示。

经过以上操作，将 RNA 数据的序列信息和结构信息抽取出来放入相应的 csv 文件中。另外，本实验将处理的数据集拆分成三部分：训练集、验证集以及测试集。RNA 在三者的数量比例为 7：2：1。其中训练集用于将训练实验需要的实验模型。验证集的作用是选择合适的模型，确定最终优化的模型。测试集是在训练集、验证集之后使用的，可衡量模型针对实际问题的泛化能力。

5.2 算法流程概述

5.2.1 核心算法流程

目前，已有许多算法应用到 RNA 二级结构的预测研究中，如动态规划算法、比较序列分析法以及深度学习方法等。其中动态规划方法针对 RNA 的碱基序列信息去预测 RNA 的结构，然而比较序列分析法则是针对多条 RNA 序列进行预测的。二者各有各的好处，但也存在着相应的缺陷。动态规划算法在预测 RNA 的二级结构时，若 RNA 的序列过长则会使最终的预测精度过低。另外，序列比对分析方法需要多条同源序列的数据进行预测，若数据集中存在的同源序列数据数量太少，也会导致预测精确度偏低。因此，本文提出使用 GoogLeNet 模型，预测带假结的 RNA 二级结构的方法，为以后的研究提供一种新的思路。

本文提出一种使用 GoogLeNet 模型与将动态规划方法相结合的方法预测带假结的 RNA 二级结构。由于 RNA 的二级结构是由碱基对和非碱基对组成，即由连续的碱基对形成的茎区以及未形成碱基对的环状结构组成。因此，若要预测 RNA 的二级结构只需确定该条 RNA 序列含有的所有碱基对即可。针对存在的大量 RNA 数据，GoogLeNet 模型善于从中挖掘出其中隐含的有效特征信息，预测该条 RNA 上每个碱基相应的匹配概率。针对得到的匹配概率结果，基于 RNA 二

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

级结构的定义并借助动态规划方法的概念，将碱基概率相加，将和最大的结果作为最终的 RNA 二级结构。在此过程中，设计到诸多的核心算法。如：RNA 序列表示方法的设计、模型搭建及修正算法等。以上核心算法将逐一进行介绍：

前一节提到将 RNA 的结构信息和序列信息提取出来，使用点括号方法表示 RNA 的结构信息，那么 RNA 的序列信息也需要一种方式进行表示。由于 “A” “U”“G”“C” 四种碱基会组成 RNA 的序列信息，在模型训练的过程中，GoogLeNet 模型无法将 “A”“U”“G”“C” 等字符作为数据输入。最常见的 one-hot 编码方式虽然简单易操作，但无法准确表示出 RNA 碱基间的配对关系，因此本文利用矩阵设计一种有效表示 RNA 序列的方法。主要思路如下：对每个 RNA 序列建立一个相应的二维矩阵，矩阵的每一行具有特殊的含义。例如，矩阵的第 i 行表示第 i 个位置上的碱基与其他位置的碱基发生碱基互补配对的可能性。此外，配对碱基间的氢键个数决定权值的大小，其中 A 与 U 设为 2，G 与 C 设为 3，U 与 G 为摇摆配对，设为 x。公式如下：

$$P(R_i,R_j)=\begin{cases}2, & (\text{若i与j位置的碱基为(A,U)或(U,A)})\\ 3, & (\text{若i与j位置的碱基为(G,C)或(C,G)})\\ x,x\in(0,2) & (\text{若i与j位置的碱基为(G,U)或(U,G)})\\ 0, & \text{其他情况}\end{cases} \quad \cdots\cdots (5.1)$$

由于 RNA 二级结构又称为茎环结构，因此对于 i，j 位置上的碱基，不仅要考虑二者是否发生配对，而且要考虑二者是否为茎区上的碱基。此外，由于越靠近茎区中间位置的碱基对比其两侧的碱基对要稳定，稳定性会对碱基间的配对情况产生不同的影响。因此，针对 RNA 序列不仅考虑 i，j 位置上的碱基，而且需将 i 和 j 各自左右两侧碱基的配对情况考虑进去。针对以上情况，本文借助局部加权线性回归中的概念，在 i，j 两侧的碱基添加高斯函数，越接近 i，j 位置的碱基对，其权值越高，越远离 i，j 位置的碱基对，其权值越低。针对矩阵 $W_{i,j}$ 的每一个位置上的权值计算，其中 (i, j) 碱基对内外两侧位置的碱基对情况相同，如图 5.2 所示。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

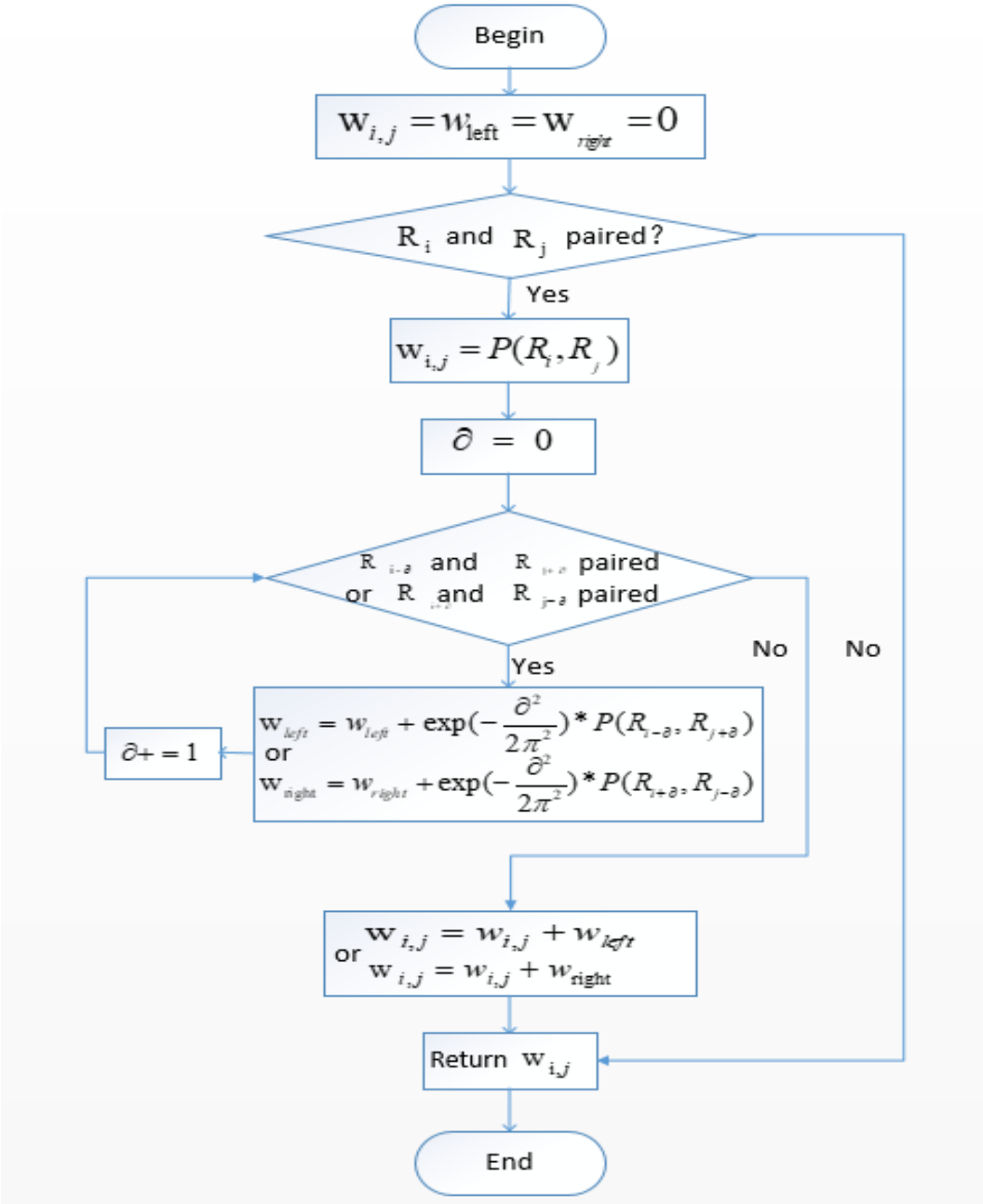


图 5.2 RNA 序列的表示方法设计

通过上述 RNA 序列表示方法将某条 RNA 序列转换为一个 RNA 序列的二维矩阵，由于 GoogLeNet 模型需要预测每一个碱基的配对状况，因此需要拆分 RNA 序列的二维矩阵。若将某条 RNA 序列的长度定义为 m，经过 RNA 序列表示方法将其转化为一个 $m \times m$ 的二维矩阵。针对以上情况，借助滑动窗口的思想将矩阵拆分为 m 个小矩阵，使用 a 表示滑动窗口的大小，这样每个小矩阵的尺寸则为 $a \times m$。因此，一个大小为 $a \times m$ 的矩阵代表着该 RNA 序列的一个碱基。其中，滑动窗口的尺寸会影响实验模型的精确度。若设置的滑动窗口过大，会提取矩阵

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

中含有的冗余信息。若设置的滑动窗口过小，则会无法提取有效且全面的特征。经过一系列分析，我们发现滑动窗口的大小与 RNA 中茎区长度有很大的关系。因此，我们通过统计茎区信息设置了滑动窗口的大小。另外，由于 GoogLeNet 网络模型需要输入的数据大小必须保持一致，然而不同长度的 RNA 序列会形成不同大小的 RNA 序列二维矩阵。因此，需要对 RNA 数据进行归一化操作。GoogLeNet 方法要求输入的数据大小为 224×224，因此本实验将所有的实验数据设置同样的大小便可作为标准的数据输入到 GoogLeNet 模型中。

RNA 二级结构的预测问题属于分类问题的一种，但不是简单的二分类问题。针对分类问题，虽然 GoogLeNet 方法精确度较高，但无法避免其存在误差，因此无法确保通过该方法得出的预测结果是否满足定义 RNA 二级结构的要求。RNA 二级结构定义要求如下：针对小括号、中括号及大括号中的任一情况，预测出的左括号和右括号个数相等，且其对应的碱基可以配对。另外，RNA 二级结构预测是将碱基对的精确度作为评价标准，并不是以每个碱基的精确度为评价标准。针对以上情况，需要使用修正方法得出的概率结果进行操作，使得最终的结果满足 RNA 二级结构的定义。

由于 RNA 大分子中的假结结构在其功能表达上发挥十分重要的作用，若假结结构预测错误，将对正常茎区的预测结果产生影响。若同时对 RNA 的假结结构和不含假结的结构进行修正操作，可能会影响最终的预测结果。因此，本文将假结结构和正常茎区的预测结果分别进行修正，之后再将二者进行整合。如图 5.3 所示。

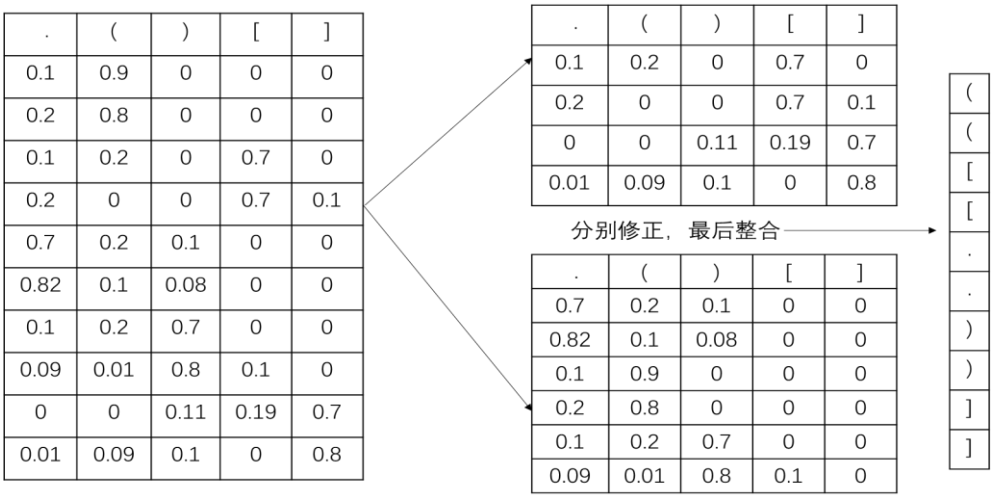


图 5.3 修正操作

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

由于含假结结构与不含假结结构的修正操作相同，因此本文仅针对不含假结的修正操作进行说明。GoogLeNet 模型在预测 RNA 二级结构的过程中，其输出层将返回前最大目的值和索引的一个函数，即 Tensorflow 中的 tf.nn.top_k()函数。本文将该函数删除，将得到每个碱基所属的标签“.”、“（”和“）”所对应的概率为 p_{point}、p_{left}、p_{right}。本文的目的是寻找满足如下条件的可表示 RNA 二级结构的点括号序列：首先，序列的括号相互匹配。其次，若序列存在已匹配的括号，并且其对应位置的碱基也互相配对。最后，满足上述两个条件后，在 GoogLeNet 网络中输出的每个标签所对应的碱基概率和最大。

针对上述情况，本文通过对最大碱基配对算法进行了相应的修改，提出了最大概率和修正的算法，即将最大碱基配对算法中不断累加配对碱基的个数修改为不断累加碱基对应的概率之和。该方法借助动态规划的思想，通过不断的递归得到满足上述要求的 RNA 二级结构。其公式如下：

$$N(i,j)=\max\begin{cases}N(i+1,j)+p_{point}(R_i)\ ,\\ N(i,j-1)+p_{point}(R_j)\ ,\\ N(i+1,j-1)+\delta(R_i,R_j)\ ,\\ \max\limits_{i<k<j}[N(i,k)+N(k+1,j)]\end{cases} \quad \text{…（5.2）}$$

$$\delta(R_i,R_j)=\begin{cases}p_{left}(R_i)+p_{right}(R_j)\ (i\text{与j位置碱基配对})\\ p_{point}(R_i)+p_{point}(R_j)\ (i\text{与j位置碱基未配对})\end{cases}$$

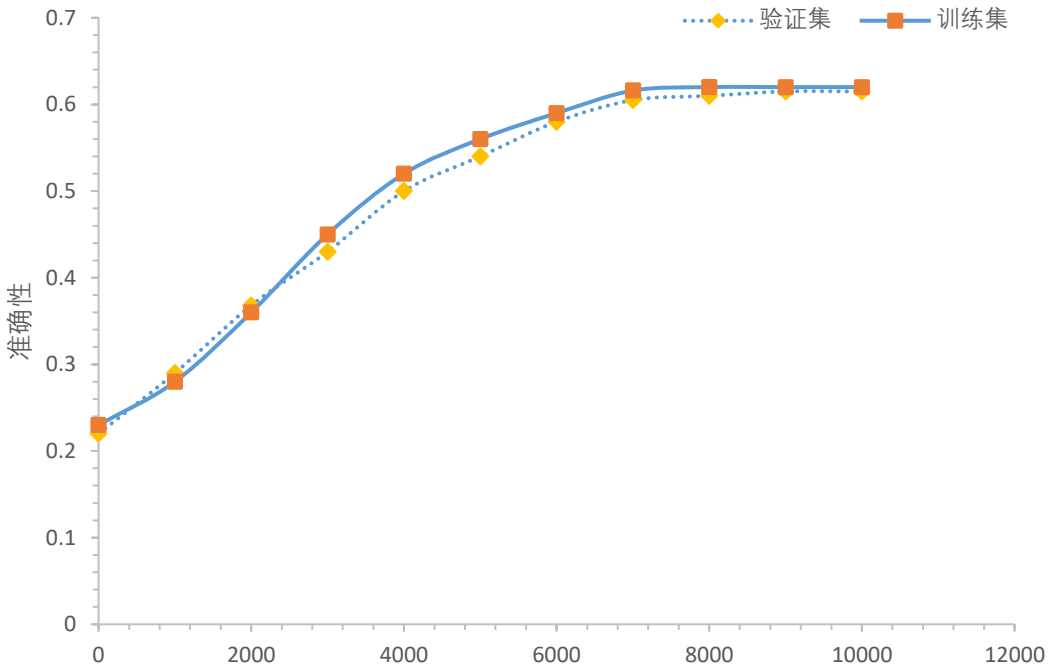
其中，$p_{point}(R_i)$、$p_{left}(R_i)$和$p_{right}(R_i)$ 分别表示第 i 个位置的碱基在使用 GoogLeNet 模型时输出的三个标签“.”、“（”和“）”所对应的概率，$N(i,j)$则表示该 RNA 序列中第 i 个碱基至第 j 个碱基对应的最大概率和。

5.2.2 实验结果

本文将 tmRNA 数据和 5sRNA、tRNA 数据分成两组实验数据，经过上述的数据处理及核心算法等流程得到标准的输入样本，将其分别输入到 GoogLeNet 模型中，通过多次迭代训练，最终的模型基本固定。训练结果如图 5.4 所示。含 tmRNA 数据的一组 GoogLeNet 模型在经过约 8000 次迭代后，实验准确度趋于稳定。相应的，含 5sRNA、tRNA 数据的模型也在一定的迭代次数后，实验的准

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

确度趋于平稳。上述情况的出现，在于 RNA 分子的假结结构十分复杂，因此通过 GoogLeNet 模型训练含假结结构数据得出的结果与不含假结结构数据相比，前者的准确率比后者准确率较低。从图中可以看出，带假结的 tmRNA 数据经过训练得到的结果约为 62%，不带假结的 5sRNA 及 tRNA 数据得到的结果约为 86%。虽然本文使用 GoogLeNet 模型得出含假结结构数据的准确性比使用 GoogLeNet 模型得出不含假结结构数据的准确性要低，但该模型为带假结的 RNA 二级结构的预测研究提供了一个新的突破口，为后续的深入研究奠定了坚实的基础。



(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

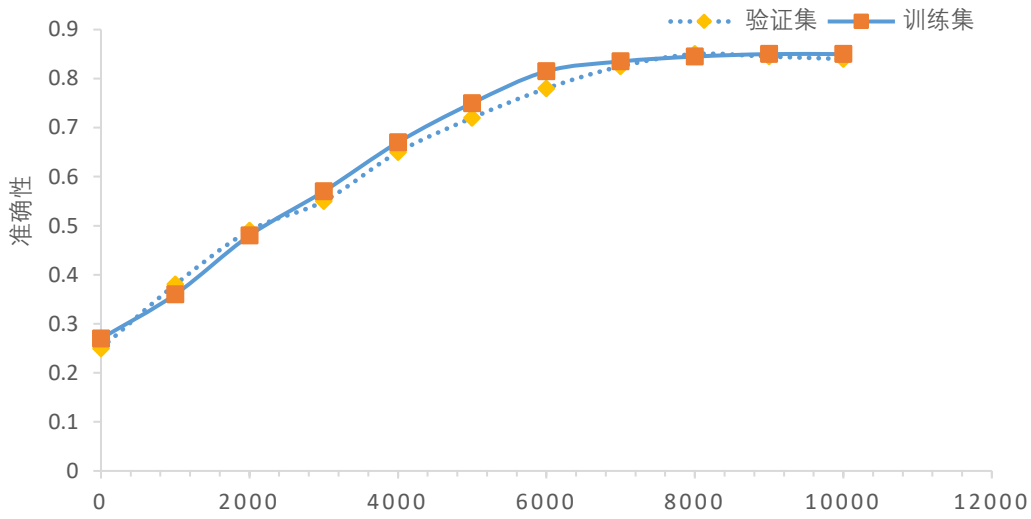


图 5.4 模型在训练集和验证集的准确度

5.3 预测结果及对比

5.3.1 评价标准

目前，主要将敏感性、特异性作为预测 RNA 二级结构的精准度的衡量指标。假设 a 是正确预测的碱基对数量，b 是存在于 RNA 真实结构中的碱基对数量，c 是全部预测出的碱基对数量。敏感性表示 a 在 b 中所占的比例，特异性表示 a 在 c 中所占的比例。具体公式如下：

$$R = \frac{TP}{TP + FN} \quad \cdots\cdots (5.3)$$

$$P = \frac{TP}{TP + FP} \quad \cdots\cdots (5.4)$$

其中 TP、FP、FN 分别代表真阳性、假阳性及假阴性。由上式可知，TP 阳是指真实存在且被正确预测到的碱基对数量，FP 是指全部预测出的碱基对中被错误的预测出的碱基对数量，FN 指真实存在但未被预测到的碱基对数量。

一般情况下，在预测 RNA 二级结构时无法使二者平衡，总是倾向于其中一个，可以使用 $F1-score$ 平衡二者，其公式如下：

$$F1\text{-}score = \frac{2PR}{P + R} \quad \cdots\cdots (5.5)$$

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

5.3.2 结果比较

本文根据上述衡量标准，使用相同的测试数据将 GoogLeNet 模型与 Mfold，RNAfold 得出的实验结果进行比较。

其中，表 5.2 表示 GoogLeNet 模型与其他算法在 tRNA 和 5sRNA 数据集上的预测精度。由此表可以看出，无论从敏感性还是特异性角度来说，GoogLeNet 模型的效果明显高于其他算法。因此，GoogLeNet 模型对 tRNA 和 5sRNA 数据集的预测是有效的。

表 5.2 tRNA 及 5sRNA 的预测精度对比

	tRNA		5sRNA	
算法	敏感性	特异性	敏感性	特异性
Mfold	0.687	0.685	0.683	0.679
RNAfold	0.698	0.693	0.699	0.699
GoogLeNet	0. 863	0.853	0.861	0.858

表 5.3 表示 GoogLeNet 模型与其他算法 tmRNA 数据集上的预测精度。由此表可以看出，在预测含有假结结构的 tmRNA 数据时，所有算法的预测精度均有所降低，但 GoogLeNet 模型得出的预测精确度与其他算法相比仍处于较大优势，预测精度提高了近 9%。由表 5.2 和表 5.3 可知，虽然含假结的数据预测精确度比不含假结的数据预测精确度要低，但是也验证了 GoogLeNet 模型对预测假结结构的数据是有效的。

表 5.3 tRNA 的预测精度对比

	tmRNA	
算法	敏感性	特异性
Mfold	0.539	0.516
RNAfold	0.535	0.512
GoogLeNet	0.629	0.612

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

5.4 本章小结

本章对基于 GoogLeNet 方法的带假结的 RNA 二级结构方法进行了详细描述，包括数据的收集和处理、设计 RNA 的表示方法及预测结果的修正等核心算法的流程。虽然含假结的数据比不含假结结构的数据的预测精确度低，但该方法为预测带假结的 RNA 二级结构的研究提供了一个新的突破口。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

第 6 章 总结与展望

本文首先对 RNA 二级结构的研究背景及国内外研究现状进行了阐述，对 RNA 的构成、RNA 的二级结构等相关的基础知识进行了详细的介绍，着重强调 RNA 二级结构对 RNA 分子功能表达上起到重要的作用，并对假结结构的种类、目前存在的预测难点问题进行了分析。然后本文介绍了常用的 RNA 二级结构预测方法如：比较序列分析法、动态规划方法以及启发式算法等，这些算法各自有各自的优势所在，但也存在着各自的不足。因此，本文利用计算机技术结合生物信息学的知识，提出了一种基于 GoogLeNet 模型来预测带假结的 RNA 二级结构的新方法。本文实验的工作主要包括如下几点：

1.从 Mathews lab 中的数据集中收集到 RNA 真实的数据样本。为避免数据中存在的同源序列对实验准确性造成影响，本文对收集到的原始 RNA 数据进行去冗余操作。然后对 RNA 数据进行预处理，即对保存在 CT 文件中的数据提取出有效的序列信息和结构信息，分别保存在不同的 csv 文件中，其中 RNA 的二级结构使用七个标签的点括号表示法进行表示。

2. 设计一种有效的 RNA 序列表示方法。由于算法模型无法识别由“A”“U”“G”“C”字符组成的输入数据，而 one-hot 编码无法准确表示 RNA 序列的碱基配对情况。因此，本文使用矩阵设计一种 RNA 序列表示方法，以碱基间配对的个数作为权值，矩阵的每行代表着该位置碱基与其他位置碱基发生配对的可能性。此方法可有效表示 RNA 序列中碱基是否发生配对的情况。

3. 设计预测结果的修正方法。由于假结结构影响着 RNA 的功能，若在预测假结的过程中出现错误，则其正常茎区的预测精确度也将受到影响。因此，本文将假结结构与不含假结结构分开，二者分别进行修正操作。通过对最大碱基配对算法进行修改，借助基于动态规划的思想，将预测过程中得到的中间结果使用概率和最大的修正方法进行修正，最后将二者得到的结果整合得到最终的修正结果。

本文将 5sRNA、tmRNA 和 tRNA 三个家族作为实验数据，其长度分布较均匀，通过 GoogLeNet 模型的效果比其他算法好。然而，针对序列更长的 RNA 数据集的预测效果尚未体现出来。通过实验了解到，本文选取的 Mathews lab 提供的数据集中长序列的 RNA 数据集的数量较少，其中 RNA 序列长度超过 1000nt

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

的个数不到 150 个，约占整体数据集的 10%左右。由于深度学习方法预测准确性的高低受到样本数据量的影响，加上本实验中长序列 RNA 数据量匮乏，可用于分析的数据十分有限，因此本实验未能完成对长序列 RNA 数据的深入研究。若在今后的工作中，我们通过实验测出的长序列数据量不断增加，我们将获得更多的结构特征，深度学习方法将会发挥其巨大的优势，从而提高我们的预测精确度。

总的来说，虽然本文提出的基于 GoogLeNet 方法的带假结的 RNA 二级结构的预测效果比较理想，但是仍存在着许多的问题，需要进一步的完善。本文选取的算法模型版本相对较低，未对其他较高版本的算法模型效果进行比较，这是一个小小的遗憾。针对 RNA 二级结构的研究，尤其是带假结的 RNA 结构预测的研究仍存在诸多的疑点和难点，需要我们更进一步的改进。但本文提出的新的预测方法为后续 RNA 二级结构的相关研究开辟了一个新的思路，奠定了良好的基础，有助于研究人员深入后续 RNA 二级结构的研究。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

参考文献

[1] 李敏. 蛋白质和氨基酸[C]// 中国营养学会 DRIs 修订专家委员会第二次会议样板稿汇编. 2011.

[2] 杨桢. microRNA 相关的生物信息学与进化分析[D]. 复旦大学, 2011.

[3] 苑寅. 带假结 RNA 二级结构预测研究[D]. 成都:电子科技大学,2013.

[4] Suzanne S . Handbook of Computational Molecular Biology.Edited by Srinivas Aluru[J]. Briefings in Bioinformatics,2007(3):3.

[5] 赵亚华. 分子生物学教程[M]. 科学出版社, 2006.

[6] Vickers, T. A . Efficient Reduction of Target RNAs by Small Interfering RNA and RNase H-dependent Antisense Agents. A COMPARATIVE ANALYSIS[J]. Journal of Biological Chemistry, 2003, 278(9):7108-7118.

[7] Mooney R A , Artsimovitch I , Landick R . Information processing by RNA polymerase: recognition of regulatory signals during RNA chain elongation[J]. Journal of Bacteriology, 1998, 180(13):3265.

[8] Onoa,B;Tinoco,I.RNA folding and unfolding[J]. Current Opinion In Structural Biology .2004:374-379.

[9] Tinoco, I; Bustamante, C.How RNA folds [J]. Journal of The American Chemical Society. 1999:271-281.

[10] Rietveld K, Poelgeest R V, Pleij C W A, et al. The tRNA-Uke structure at the 3′ terminus of turnip yellow mosaic virus RNA. Differences and similarities with canonical tRNA [J]. Nucleic Acids Research, 1982, 10(6): 1929.

[11] Lim N C H, Jackson S E. Molecular knots in biology and chemistry [J]. Journal of Physics Condensed Matter, 2015, 27(35):

[12] Liu, H.; Xu, D.; Shao, J.; Wang, Y. An RNA folding algorithm including pseudoknots based on dynamic weighted matching.Comput Biol Chem 2006(30):72-76.

[13] Silverman S K, Zheng M,Wu M. Quantifying the energetic interplay of RNA tertiary and secondary structure interactions[J]. RNA, 1995,5:1665-1674.

[14] David W. Staple, Samuel E. Butcher. Pseudoknots: RNA Structures with Diverse

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

Functions[J]. Plos Biology, 2005,3(6):0956-0959.

[15] LeoYean ling. A study on RNA Pseudoknot predictions[D]. Malaysia :Universiti Tunku Abdul Rahman,2003,1-110.

[16] Fürtig B, Richter C, Wöhnert J, et al. NMR spectroscopy of RNA.[J]. Cheminform, 2003, 34(49):936-962.

[17] Novikova I V , Hennelly S P , Sanbonmatsu K Y . Sizing up long non-coding RNAs: Do lncRNAs have secondary and tertiary structure?[J]. BioArchitecture, 2012, 2(6):189-199.

[18] Seetin M G, Mathews D H. RNA structure prediction: an overview of methods [J]. Methods in Molecular Biology, 2012, 905(905): 99.

[19] Nussinov R, Pieczenik G, Griggs J R, et al. Algorithms for Loop Matchings[J]. Siam Journal on Applied Mathematics, 1978, 35(1):68-82.

[20] 胡名刚. 基于真实结构特征的 RNA 二级结构预测方法研究 [D]; 吉林大学, 2014.

[21] 邹权, 郭茂祖, 张涛涛. RNA 二级结构预测方法综述[J]. 电子学报, 2008, 36(2):331-337.

[22] 胥杰. 基于混沌模拟退火的 RNA 二级结构预测的研究[D]. 电子科技大学, 2010.

[23] 胡桂武, 彭宏. 基于免疫粒子群集成的 RNA 二级结构预测算法[J]. 计算机工程与应用, 2007, 43(3):26-29.

[24] Wiese K C , Glen E . A permutation-based genetic algorithm for the RNA folding problem: a critical look at selection strategies, crossover operators, and representation issues[J]. Biosystems, 2003, 72(1-2):29-41.

[25] 常征, 孟军, 施云生, et al. 多特征融合的 lncRNA 识别与其功能预测[J]. 智能系统学报, 2018, 13(6):68-74.

[26] 张秀苇, 邓志东, 宋丹丹. RNA 二级结构预测的神经网络方法[J]. 清华大学学报（自然科学版）, 2006, 46(10).

[27] Ruan J, Stormo GD, Zhang W. An iterated loop matching approach to the prediction of RNA secondary structures with pseudoknots[J]. Bioinformatics, 2004,20:58–66.

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

[28] 陈家文. RNA 三级结构折叠以及 RNA 干扰相关的动力学[D]. 武汉大学, 2013.

[29] Grilley D , Soto A M , Draper D E . Mg^{2+}-RNA Interaction Free Energies and Their Relationship to the Folding of RNA Tertiary Structures[J]. Proceedings of the National Academy of Sciences of the United States of America, 2006, 103(38):14003-14008.

[30] 王魁祎. 基于能量优化的 RNA 二级结构预测[D]. 吉林大学, 2013.

[31] 张玉静. 分子遗传学[M]. 北京: 科学出版社, 2000.

[32] 樊龙江. 生物信息学札记第三版(M). 杭州: 浙江大学, 2010.

[33] 吴建英, 王淑琴. RNA 二级结构点括号图与 CT 文件表示法的相互转换算法研究[J]. 天津师范大学学报(自然科学版), 2012, 32(4):32-36.

[34] Sperschneider J, Datta A, Wise M J. Heuristic RNA pseudoknot prediction including intramolecular kissing hairpins [J]. RNA, 2011, 17(1): 27.

[35] Eddy.S.R. How do RNA folding algorithms work [J].Nature Biotechnology. 2004,22(11): 1457- 1458.

[36] Ruan J, Stormo G D, Zhang W. An Iterative Loop Matching Approach to the Prediction of RNA Secondary Structures with Pseudoknots [J]. Bioinformatics, 2004, 20(1): 58-66.

[37] Bernhart S H, Hofacker I L, Will S, et al. RNAalifold: improved consensus structure prediction for RNA alignments [J]. BMC Bioinformatics, 2008, 9(1): 1-13.

[38] Knudsen B. Pfold: RNA secondary structure prediction using stochastic context-free grammars [J]. Nucleic Acids Research, 2003, 31(13): 3423.

[39] Zuker M . Calculating nucleic acid secondary structure.[J]. Curr Opin Struct Biol, 2000, 10(3):303-310.

[40] Hofacker I L , Bernhart S H F , Stadler P F . Alignment of RNA base pairing probability matrices[J]. Bioinformatics, 2004, 20(14):2222-2227.

[41] Batenburg F H D V , Gultyaev A P , Pleij C W A . An APL-programmed genetic algorithm for the prediction of RNA secondary structure[J]. journal of theoretical biology, 1995, 174(3):269-280.

[42] Szegedy C , Liu W , Jia Y , et al. Going Deeper with Convolutions[J]. 2014.

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

[43] Padideh D , Mason R , Michelle W , et al. bpRNA: large-scale automated annotation and analysis of RNA secondary structure[J]. Nucleic Acids Research(11):11.

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

作者简介及在学期间所取得的科研成果

作者简介：

李聪，女，汉族，1994 年 2 月 17 日出生于山东省菏泽市。2017 年 9 月入学，就读于计算机科学与技术学院，2017 级硕士研究生，学制三年，研究方向为生物信息学。

硕士期间学术成果：

[1] Hao Zhang, Chunhe Zhang, Zhi Li, **Cong Li**, Xu Wei, Borui Zhang,Yuanning Liu. A New Method of RNA Secondary Structure Prediction Based on Convolutional Neural Network and Dynamic Programming. Frontiers in Genetics, 2019（JCR 生物 2 区)

[2] 专利：基于卷积神经网络和规划动态算法的 RNA 二级结构生成器（201810851933.X）

[3] 软件著作权：基于卷积神经网络的 RNA 二级结构预测平台 V1.0（2018SR541770）

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

致 谢

时光飞逝，转眼间已经到了研三下学期。此刻我依稀记得自己刚走进吉大校园的那一刻，自己内心十分的激动。回忆起在吉林大学三年的学习和生活，不得不说收获了很多。在吉林大学我不仅学习到了专业知识，而且收获了良师益友，给我的人生道路指明了正确的前进方向，在我人生的画册上增添了浓墨重彩的一笔。借此机会，在本文的最后我要表达对他们最真诚的谢意。

首先，我要非常感谢我的研究生导师张浩教授。张浩教授主要致力于生物信息学、人工智能及生物信息领域数据分析等多个方向的研究，并做出了大量的贡献。本论文是在张浩教授十分细心的专业指导下完成的。起初，本论文在算法的设计和实现方面遇到了困难，张浩教授总是悉心指导，十分耐心的为我提供不同的新思路以及许多创新性的建议，为我指出正确的研究方向和思考方向，鼓励我勇敢的做出尝试，不厌其烦的对我的专业知识进行学术指导。在张浩教授细心的教育和引导下，我的学习能力和科研能力都有了很大的提升，而且在思考问题、解决问题等综合素质方面也进步很大。在实验室的学习过程中，张浩教授始终关注每一个人的成长，经常举行各种集体活动，大家一起探讨、共同进步。而且张浩教授经常邀请国内外著名研究学者来做演讲，以丰富我们的专业知识。另外，张浩教授也经常带领学生参加国内甚至国际的学术会议，拓展了我们的视野。张浩老师这种对工作认真负责的态度将激励我在以后的工作中不断地前行。因此，我向给予我巨大帮助的张浩老师，表达我最诚挚的感谢与敬意。

与此同时，也要感谢实验室给予我大力的指导和帮助的众位师兄和师姐，尤其是张春鹤师兄、王林宇师兄。王林宇师兄给我的论文提供了数据支持，另外在数据处理方面，林宇师兄也为提供了有效的解决方案。张春鹤师兄无论在论文的思路和多个算法思路的设计与尝试的过程中，都给予了我巨大的支持和鼓励，使我克服了许多的困难问题。借此机会，我要向实验室的师兄师姐表达我最真诚的谢意。此外，实验室的各位同学，也在生活和学习方面中给了很大支持和鼓励，丰富了我三年的研究生生涯。在与各位同学交流的过程中，他们表现出的睿智和机敏也激发了我的许多思考与灵感。

最后，我要感谢我的家人，是他们一直无怨无悔的供我读书，在学业上给予

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

我支持，在生活上给予我无微不至的照顾和关怀，正是家人的爱才让我成为了现在的自己。在这里，我要发自内心地向他们说一声谢谢！

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net