项目编号：_____

吉林大学“大学生创新创业训练计划”

创新训练项目

申请书

项目名称 基于胶囊神经网络的 RNA 二级结构预测新方法

项目负责人 雷浩洁

所在学院、年级、专业 计算机科学与技术学院 2017 级计算机科学与技术专业

联系电话 15518081176

电子邮箱 leihj2117@mails.jlu.edu.cn

指导教师姓名 张浩 职称 教授

填表日期 2019 年 4 月 20 日

吉林大学教务处制表

填表须知

一、**本表适用于创新训练项目**。本科生个人或团队，在校内导师指导下，自主完成创新性实验方法的设计、设备和材料的准备、实验的实施、数据处理与分析、总结报告撰写等工作。

二、申报书请按顺序逐项填写，实事求是，表达明确严谨。空缺项要填“无”。

三、申请参加大学生创新训练项目团队的人数为 3—5 人。

四、申请项目，必须聘请教师作为指导老师，并请指导教师在申请书中的指导教师意见栏中签署意见。

五、填写时可以改变字体大小等，但要确保表格的样式不变；不得随意涂改；A4 纸正反面打印，左侧装订。

六、本表由项目负责人报所在学院初审，学院签署初审意见后报送教务处实践教学科（一式 3 份原件）。

七、“项目编号”由教务处填写。

八、申报过程有不明事宜，请与教务处实践教学科联系，电话 85166413。

<table>
<tr><td colspan="2">项目名称</td><td colspan="8">基于胶囊神经网络的 RNA 二级结构预测新方法</td></tr>
<tr><td colspan="2">项目起止时间</td><td colspan="8">2019 年 4 月 至 2020 年 4 月</td></tr>
<tr><td rowspan="6">负责人
项目组成员</td><td>姓名</td><td>学院</td><td>专业</td><td>教学号</td><td>联系电话</td><td>E-mail</td><td>QQ</td><td colspan="2">各类实验班</td></tr>
<tr><td>雷浩洁</td><td>计算机科学与技术</td><td>计算机科学与技术</td><td>21172703</td><td>15518081176</td><td>leihj2117@mails.jlu.edu.cn</td><td>1301547369</td><td colspan="2">是□ 否☑</td></tr>
<tr><td>孙博文</td><td>计算机科学与技术</td><td>计算机科学与技术</td><td>21172501</td><td>13654319052</td><td>sunbw2117@mails.jlu.edu.cn</td><td>1289790957</td><td colspan="2">是□ 否☑</td></tr>
<tr><td>黄正坤</td><td>计算机科学与技术</td><td>计算机科学与技术</td><td>21172535</td><td>15526883781</td><td>huangzk2017@mails.jlu.edu.cn</td><td>1206198069</td><td colspan="2">是□ 否☑</td></tr>
<tr><td>张雨凡</td><td>计算机科学与技术</td><td>计算机科学与技术</td><td>21172537</td><td>18205614044</td><td>zhangyf1917@mails.jlu.edu.cn</td><td>527354728</td><td colspan="2">是□ 否☑</td></tr>
<tr><td>张瀚文</td><td>计算机科学与技术</td><td>计算机科学与技术</td><td>21172713</td><td>15154503645</td><td>Zhanghw2117@mails.jlu.edu.cn</td><td>962443318</td><td colspan="2">是□ 否☑</td></tr>
<tr><td rowspan="4">指导教师一</td><td>姓名</td><td colspan="4">张浩</td><td>职务/职称</td><td colspan="2">教授</td></tr>
<tr><td>所在单位</td><td colspan="7">吉林大学计算机科学与技术学院</td></tr>
<tr><td>联系电话</td><td colspan="4">13843148999</td><td>E-mail</td><td colspan="2">zhangh@jlu.edu.cn</td></tr>
<tr><td>对本课题相关领域研究情况</td><td colspan="7">从事过相关领域研究工作，对本项目很熟悉。</td></tr>
<tr><td colspan="2">项目性质</td><td colspan="8">1.小发明、小创作、小设计（ ）
2.开放实验室或实习基地中的创新性实验或新实验开发（ ）
3.基础性研究（ ）
4.应用性研究（√）
5.社会调研 （ ）</td></tr>
<tr><td colspan="2">项目选题来源</td><td colspan="8">1.自主立题（）
2.教师科研课题的子项目（√）</td></tr>
<tr><td colspan="2">项目学科类别</td><td colspan="8">计算机应用</td></tr>
</table>

项目受其他渠道资助情况（填“无”或具体资助来源和经费，包括获奖情况）	无

一、立项背景和依据（包括研究目的、国内外研究现状分析与评价、研究意义，应附主要参考文献及出处）

1、研究目的

近年来，探知 RNA 二级结构一直是 RNA 研究领域的重点和难点。目前虽然部分 RNA 的二级结构可以通过实验手段获取，但在大多数情况下仍然需要采用计算方法来预测 RNA 二级结构。当前 RNA 二级结构的预测方法主要是基于最小自由能的动态规划类算法，这种算法通过迭代的方式找到满足能量最小或其他限制条件的 RNA 体内折叠最佳状态来预测 RNA 结构。但是由于生物体内环境的复杂性使得 RNA 的真实结构并不能满足能量最小的最佳折叠状态，而是一种基于生物势能的平衡状态。对于序列较短的 RNA，折叠生物势能平衡状态接近最小自由能状态，所以最小自由能算法仍可以获得较高的准确率。然而对于较长的 RNA 序列，其结构复杂，在生物体内环境中存在反复折叠的现象导致其生物势能平衡状态远远偏离最小自由能状态，使得传统能量方法预测精度严重下降，无法准确预测 RNA 二级结构。

深度学习方法是一种常见的表示学习方法，能够自动的从数据中挖掘出有效分类的隐藏特征。本文基于深度学习和现有真实 RNA 二级结构数据，提出了一种应用胶囊神经网络的 RNA 二级结构预测新方法。该方法基于现有实验已经得出的真实 RNA 结构数据，构建胶囊网络模型，从大规模 RNA 序列数据及其结构数据中挖掘有效分类的隐含特征，得到满足 RNA 二级结构定义且各个碱基的匹配概率之和最大的 RNA 二级结构作为预测的最终结果。由于深度学习方法的性能与数据量的大小直接相关，可以推测出随着经过生物实验验证的真实 RNA 结构数据的不断增加，应用本文所提出的方法对各类 RNA 家族的预测精度也会不断的提高。本项目的顺利实施将为 RNA 结构预测及其他相关领域的相关研究提供新的思路。

2. 国内外研究现状分析与评价

RNA,即核糖核酸（Ribonucleic Acid），它是由核糖核苷酸经磷酸二酯键缩合而形成的长链大分子，是存在于生物细胞以及部分病毒、类病毒中的遗传信息载体。

RNA 的功能并不仅限于从 DNA 到蛋白质的遗传信息传递，还可能具有参与蛋白质生物合成、作为生物催化剂、参与基因表达调控、影响生物体的进化等相关生物学功能[1]。RNA 功能多样性的物质基础是 RNA 丰富的结构多样性。

生物体内的 RNA 分子结构分为一级、二级、三级、四级。RNA 的一级结构是指 RNA 分子中的核苷酸排列顺序；RNA 的二级结构是指核糖核酸单链自身折叠而形成的，由单链区、茎环结构、双链结构等原件组成的平面结构；RNA 的三级结构是指在 RNA 二级结构的基础上，核糖核酸链再折叠形成的三维空间的高级结构；RNA 的四级结构是指不同生物大分子之间的相互作用，已远远超出了本项目的研究范围[2]。

RNA 的三级结构本身就较为复杂，且极易受到温度、环境等因素的影响，同时还缺乏有效的表示方法来描述其结构，这使得根据 RNA 一级结构直接研究 RNA 三级结构变得极为困难。而作为沟通 RNA 一级结构（RNA 序列）与 RNA 三级结构（空间结构）的桥梁，RNA 二级结构的研究就成为研究 RNA 三级结构的先导工作。RNA 有许多二级结构原件：单链、双链（茎区）、错配、发卡环、内环、突环、两茎连接、四茎连接等[2]。RNA 的二级结构主要由 RNA 链的不同区段碱基间的氢键维系。通常的碱基配对方式有 AU 和 GC，也存在少数 GU 摇摆碱基对[2]。

目前已有的研究 RNA 二级结构的传统方法可以大致分为三种：生物实验方法、计算机数学方法、二者相结合的算法等，具体优缺点可见表 1。

表 1 RNA 二级结构传统研究方法比较

种类	技术	优点	缺点
生物实验	X 射线、核磁共振等	准确、真实	RNA 在体外环境极易降解，难以结晶[3]；实验方法花费高，难度大，不是对所有 RNA 分子都有效[4]
计算机数学预测	比较序列分析法[5]-[7]	若 RNA 功能相似，其结构一般也相似，因此结果较为	需要一定数量二级结构一致或相似的同源 RNA 序列作为

		准确	先验知识；不能预测功能差异较大的 RNA 二级结构
	启发式算法（遗传算法[8]、退火算法[9]、粒子群优化算法[10]等）	模拟自然界规律，如优胜劣汰等，逐步形成预测解	结果随机性很大，无法保证收敛到全局最优解
	动态规划法	最大碱基配对算法计算碱基尽可能配对时的 RNA 序列结构[11]	忽略了碱基配对类型、其他二级结构如双链、环等的影响，准确性较差
		最小自由能算法计算 RNA 自由能最小时的 RNA 序列结构[12]	真实二级结构自由能可能并非最小[13]
	传统机器学习法（非神经网络算法，如随机上下分无关文法[14]，支持向量机[15]等）	预测结果较为准确	建立在单一的数据样本基础之上，很难有大幅度提升实际应用价值的机会
生物实验和计算机算法相结合	PARS 技术[16]等	用酶切割 RNA 序列，计算机分析各个序列片段后获得序列的二级结构	可能破坏 RNA 天然结构，导致预测结果错误

在大数据时代，各种信息的获取成本变得越来越低，可以轻易地获取大量的 RNA 分子的序列信息，这使得通过以数据为驱动力的深度学习技术来预测 RNA 二级结构成为可能。近年来，深度学习领域中的神经网络技术在语音识别，自动驾驶，计算机视觉等诸多领域均取得了很大的成就。

神经网络技术在生物信息学领域也有相当可喜的表现，卷积神经网络及卷积神经网络和

循环神经网络的共同使用在用于预测蛋白质二级结构时均取得了实验成果[18]。相比于一旦形成便很少有结构变化的蛋白质，RNA 在生物体内往往有多次结构变化[2]，这也使得使用深度学习技术预测 RNA 二级结构的难度加大。

目前，已经有循环神经网络应用到 RNA 二级结构预测的研究，研究人员使用双向 LSTM 神经网络对 RNA 二级结构进行打分[19]。而生成对抗网络和胶囊神经网络在 RNA 二级结构预测的研究还未见发表。

胶囊神经网络（Capsule Network）是一种近期刚刚提出的神经网络算法。不同于卷积神经网络（CNN），循环神经网络（RNN）等主流的神经网络架构，它对样本特征的空间位置关系更为敏感，且不需要极为庞大的数据集就可以达到很好的训练效果[20]，在预测 RNA 二级结构方面有着其他神经网络所不具备的独特的优势。

3. 研究意义

传统 RNA 二级结构预测方法有着不可避免的缺陷，而使用深度学习进行预测的新方法亟待发现。作为深度学习领域的最新研究成果，胶囊神经网络在各个领域的应用比较稀少。本项目尝试使用胶囊神经网络架构提出一种全新的 RNA 二级结构预测方法，不仅可以有效避免传统方法的缺陷，为 RNA 二级结构预测提供新的方法和思路，而且有助于发展胶囊神经网络在各个不同领域的泛用性，。

4. 参考文献

[1] 赵亚华. 分子生物学教程[M]. 科学出版社, 2006.

[2] 金由辛. 核糖核酸与核糖核酸组学[M]. 科学出版社, 2005

[3] Irina V. Novikova, Scott P. Hennelly and Karissa Y. Sanbonmatsu, Sizing up long non-coding RNAs: do lncRNAs have secondary and tertiary structure[J], 2012, BioArchitecture, 2:6, 1-11.

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

[5] Gorodkin J, Heyer L J, Stormo G D. Finding the most significant common sequence and structure motifs in a set of RNA sequences[J]. Nucleic Acids Research, 1997, 25(18):3724-32.

[6] Knudsen B, Hein J. Pfold: RNA secondary structure prediction using stochastic context-free grammars[J]. Nucleic Acids Research, 2003, 31(13):3423-3428.

[7] Hofacker I L, Bernhart S H, Stadler P F. Alignment of RNA base pairing probability

matrices[J]. Bioinformatics, 2004, 20(14):2222.

[8] WIESE, Kay C, GLEN, et al. 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):29-41.

[9] Ren J , Rastegari B , Hoos H H . HotKnots: heuristic prediction of RNA secondary structures including pseudoknots.[J]. Rna-a Publication of the Rna Society, 2005, 11(10):1494-504.

[10] 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.

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

[12] Zuker M, Stiegler P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information.[J]. Nucleic Acids Research, 1981, 9(1):133-148.

[13] Pace N R, Thomas B C, Woese C R. Probing RNA Structure, Function, and History by Comparative Analysis[J]. Gesteland R.f. & Atkins J.f.the Rna, 1999.

[14] Dowell R D, Eddy S R. Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction[J]. Bmc Bioinformatics, 2004, 5(1):71.

[15] Jing-Yuan H E. The Model Research of Support Vector Machines in the RNA Secondary Structure Prediction[J]. Computer Science, 2008, 35(4):181-183.

[16] Kertesz M, Wan Y, Mazor E, et al. Genome-wide measurement of RNA secondary structure in yeast[J]. Nature, 2010, 467(7311):103.

[17] WANG S, PENG J, MA J, et al. Protein secondary structure prediction using deep convolutional neural fields[J]. Scientific Reports, 2016, 6: 18962. DOI:10.1038/srep18962.

[18] LI Z, YU Y. Protein secondary structure prediction using cascaded convolutional and recurrent neural networks[A]. Proceedings of the twenty-fifth international joint conference on artificial intelligence[C]. AAAI Press, 2016: 2560-2567.

[19] 王帅, 蔡雷鑫, 顾倜, 吕强. 运用双向 LSTM 拟合 RNA 二级结构打分函数[J]. 计算机应用与软件, 2017, 34(9):232-239.

[20] S. Sabour, N. Frosst, and G. E. Hinton. Dynamic routing between capsules[C]. In Advances in Neural Information Processing Systems, pages 3856–3866, 2017.

二、项目研究内容（项目主要研究内容；拟解决的关键问题、重点和难点）

1. 项目主要研究内容

本项目提出一种全新的 RNA 二级结构预测方法，即应用改进的胶囊神经网络模型对

RNA 二级结构进行预测，并检验和评估预测效果。

首先，本项目收集不同家族的真实的 RNA 二级结构数据作为实验集。将 RNA 二级结构进行适当的表示方法转化，以满足胶囊神经网络的输入要求。

然后设计合适的胶囊神经网络架构，将数据集按照一定比例划分为训练集、评估集、测试集，输入构造完毕的胶囊神经网络，对其进行迭代训练。

接着对神经网络输出的概率分布做进行进一步处理，用动态规划的思想做最后一步修正，得出符合 RNA 真实二级结构，且概率之和最大的结果，以此得出最终的预测结果。

最后以预测碱基对的准确率为标准对神经网络的预测结果进行检验和评估，量化出该方法的预测效果，并不断改进算法。

2. 拟解决的关键问题、重点和难点

（1）**编码 RNA 序列与其二级结构。**RNA 序列是由 A、C、G、U 四种碱基类型排列而成的序列，并通过碱基互补配对规则形成二级结构。网络上对 RNA 序列及其二级结构数据的保存方式为“CT 文件表示法”，而这一表示方法无法作为神经网络的输入。为此需要找到一种合适的编码方案，将 RNA 二级结构的表示方法由“CT 文件表示法”转化为矩阵表示，以满足胶囊神经网络的输入要求。

（2） **胶囊神经网络架构的设计与实现。**胶囊神经网络的正式提出是在 2017 年，不同于其它较早的神经网络，如卷积神经网络（CNN），循环神经网络（RNN），生成式对抗网络（GAN）等，在 TensorFlow、pyTorch 等深度学习框架内已有较为成熟的代码可以直接调用，目前主流的对胶囊神经网络的认知大多还停留在思想层面，而为数不多的实现代码最多也只能作为大致的参考，难以复用。因而胶囊神经网络的架构设计和代码编写等细节均需小组实现。

（3） **预测结果统计分析。**胶囊神经网络模型输出各个碱基配对情况的概率分布，且得到的结果并不是一定符合 RNA 二级结构的定义，如果把这样的输出直接作为预测的结果，实际的应用价值不大。因此从模型的输出出发，如何得到一个满足 RNA 二级结构定义，且最为合理的预测结果，也是一个需要解决的问题。

（4） **胶囊神经网络架构的改进。**使用神经网络进行研究是一种高度经验化的工作，架构的设计细节，超参数的设置不可能在一开始就处理的非常完美。需要根据训练出的模型的测试结果，灵活地调整和改进实施方案。

三、项目特色及创新点

本项目特色在于使用深度学习神经网络来进行 RNA 二级结构预测，是一种较为前沿的 RNA 二级结构预测的思路，也是计算机神经网络在生物信息领域的应用，具有学科交叉融合的特点。本项目创新点在于使用深度学习神经网络中最新的研究成果——胶囊神经网络完成预测工作，其一有效地规避了传统预测方法的缺点，其二可以进一步佐证深度学习神经网络在生物大分子结构预测上的优势，其三将有效地填补胶囊神经网络在 RNA 二级结构预测领域应用的空白。

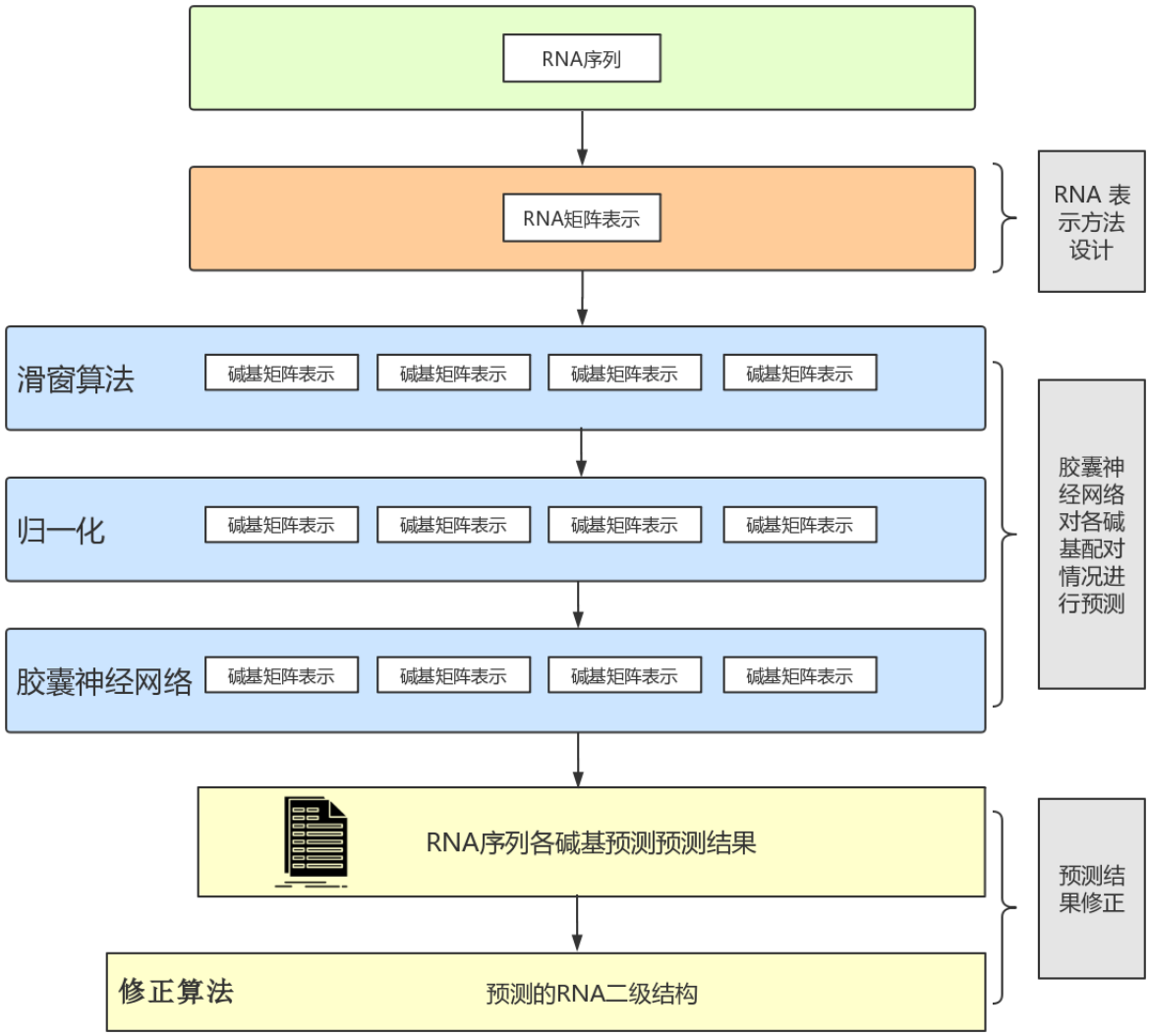
四、申请理由（1、团队条件——自身/团队具备的知识、素质、能力、特长、兴趣；2、前期准备基础等）

项目组成员 5 人均为 2017 级计算机科学与技术专业学生，且都是 25 班，27 班学生。成员之间早已熟识，之前已共同合作进行过科研项目和课程设计，团队合作能力较强，分工明确，配合默契。

小组全员计算机相关领域的专业知识基础扎实，均有较强的代码编写能力和框架搭建能力，富有创新精神，责任心强，且对深度学习在交叉学科领域的应用有着浓厚的兴趣。

姓名	特长	负责内容	
雷浩洁	学习成绩优秀，大一总成绩加权绩点在 3.7 以上。担任 2017 级计算机科学与技术学院 27 班学习委员，曾获 2017-2018 级国家励志奖学金。院优秀学生干部，曾参加大数据下推理算法的崩溃实验项目，并取得优秀的成绩。对机器学习已有一定了解。多次担任课程设计组长，有较强的组织分工能力。	统筹实验项目进度安排，搭建胶囊神经网络，测试并改进模型。	
孙博文	学习成绩优秀，曾获 2017-2018 校一等奖学金，院优秀干部。有校内创新项目的参加经验。掌握基本的编程知识，擅长构建程序的逻辑结构框架，有良好的数学基础，有积极学习新知识的热情，有积极参与团队合作的意识。	搭建胶囊神经网络，测试并改进模型。	
张瀚文	学习成绩优秀，有 C++、MySQL 和一部分 Python 基础，有作为计算机专业学生所需要具备的一定	进行实验数据的预处理工作，测试并改进模型。	

	的数理逻辑分析能力，也有一定的对新兴领域的知识的好奇心、求知欲以及对新知识的快速学习掌握能力。在论文写作方面，也有一定的实践经历和比较扎实的基本功。			
张雨凡	学习成绩优秀，有 C/C++，数据结构和 Python 基础知识，对深度学习的算法实现和神经网络架构有一定了解，具有较强的逻辑思维和编程能力，有责任心，能与队内成员团结协作。	进行实验数据的预处理工作，测试并改进模型。		
黄正坤	学习成绩优秀，曾获 2018 数学建模国赛全国二等奖，院优秀学生，系统学习过 Python 和 C/C++，具备一定编程能力，较强学习能力和团队合作意识。和组内组员熟悉，能协调团队完成工作	搭建胶囊神经网络，测试并改进模型。		
五、项目实施方案（研究思路和方法，实施计划、技术路线、人员分工等）				



项目流程图

1. 研究思路和方法

第一步是对 RNA 的二级结构数据进行预处理。由于实验数据里存在含有假结的 RNA 数据，而假结属于 RNA 的三级结构范畴，不在本项目的研究范围之内，因此首先通过筛选，去掉所有含有假结的 RNA 数据。由于不同族的 RNA 功能不同，结构差异也十分明显，因此可先将 RNA 二级结构数据按照不同族划分，各自作为单独的数据集使用。

第二步是对 RNA 二级结构数据进行编码。由于原始数据是使用 CT 文件表示法来存储结构数据，因此首先需要编写程序读入 CT 文件，将 CT 文件中的 RNA 序列信息使用“点括号表示法”来表示，然后设计算法，将序列信息转化为矩阵的形式。

第三步为构建并训练胶囊神经网络。使用典型的胶囊神经网络架构，即三层编码器+三

层解码器的六层网络架构。分别为第一层卷积层，第二层主胶囊层，第三层数字胶囊层，第四五六层为三个全连接层，最后使用 Softmax 方法输出结果。适当地划分数据集为训练集和测试集，对架构好的神经网络进行训练，并通过测试集测试训练效果。

第四步为对神经网络得到的预测结果进行修正。通过已有的算法对神经网络输出的概率结果进行优化，使之能够符合真实的 RNA 二级结构情况，并将其作为本模型最终的预测结果。

2. 实施计划和技术路线

（1）数据的获取与预处理

选择 Mathews lab 作为实验数据集的来源，因为很多研究人员都选用该 RNA 结构数据库进行 RNA 二级结构预测的研究工作，将有利于与其他 RNA 二级结构预测方法进行比较。考虑到作为 RNA 三级结构的假结对预测效果的影响，选择不含假结并且数量尽可能多、分布尽可能集中的 5sRNA 家族作为初步研究对象。在分析其序列结构时可能会发现数据集中的 RNA 存在部分相似或相同的序列数据，因此预处理过程应额外加入去除数据集冗余数据的操作，以避免影响实验结果的准确性。之后选取一定量的数据作为训练集，来训练胶囊神经网络模型，确定最佳的参数。再选取另外少量的数据作为验证集，用来进一步优化和修正胶囊神经网络模型。其余数据作为最终的测试集进行结果的评估，与其他 RNA 二级结构预测方法进行比较。

（2）数据的编码

考虑到胶囊神经网络的输入为矩阵，因此应对原始数据集进行表示方式的转化。原始数据集使用 CT 文件表示法提供 RNA 序列结构数据，先将其转化为点括号表示法，再转化为矩阵。

```
28      dG = -11.50 [Initially -11.50] 18Dec30-09-03-58
1       C       0       2       26      1
2       C       1       3       25      2
3       G       2       4       24      3
4       U       3       5       23      4
5       C       4       6       0       5
6       A       5       7       0       6
7       G       6       8       0       7
8       G       7       9       0       8
9       U       8       10      18      9
10      C       9       11      17      10
11      C       10      12      16      11
12      G       11      13      0       12
13      G       12      14      0       13
14      A       13      15      0       14
15      A       14      16      0       15
16      G       15      17      11      16
17      G       16      18      10      17
18      A       17      19      9       18
19      A       18      20      0       19
20      G       19      21      0       20
21      C       20      22      0       21
22      A       21      23      0       22
23      G       22      24      4       23
24      C       23      25      3       24
25      G       24      26      2       25
26      G       25      27      1       26
27      U       26      28      0       27
28      A       27      0       0       28
```

一个 CT 文件表示法的 RNA 序列信息

CT 文件表示法已包含 RNA 碱基序列信息和结构信息：第一列数字为索引；第二列字母为 RNA 序列碱基构成；第五列数字表示与该碱基配对的碱基的索引，其中 0 表示没有配对的碱基。据此，对于每个碱基，若无配对碱基则为“.”，若有配对碱基，则只需比较索引和配对碱基的索引，前者小于后者的用“（”表示，因为该碱基相对其配对碱基在序列的前面，反之则用“）”表示。

矩阵的构造应突出 RNA 碱基配对关系，同时还应以不同的方式表达不同的碱基配对关系。因此不妨从配对的氢键个数入手，以碱基配对时氢键个数作为原始权值（不配对视为 0）。注意到不仅要考虑当前两碱基配对情况，还要考虑这两个位置是否能构成茎上的配对碱基，设 RNA 链上有两个位置 i, j，需要考虑 i 的左（右）侧与 j 的右（左）侧碱基的相互配对情况。此外，茎中间的配对碱基较为稳定，两侧的配对碱基则相对不稳定，据此可设定一个修正函数，降低远离 i 和 j 位置的碱基权值以减小影响。最终可以计算出编码后矩阵每个位置上的具体数值。考虑到最终目标是预测 RNA 序列上每个碱基的配对情况，还应选用适当的方法将编码后的矩阵进行拆分。

（3）构建和训练胶囊神经网络

使用 Python 语言构造胶囊神经网络，在数据集中选取适量的数据作为训练集，以 RNA

序列上各个碱基的矩阵表示形式输入网络。第一层卷积层接受这些矩阵，最终输出一个多维张量来表示输入的 RNA 二级结构 2D 图像的基本局部特征。考虑到胶囊更适合表达物体整体的特征，第一步采用了擅长抽取局部特征的 CNN 卷积操作。第二层主胶囊层接收该张量，使用该层包含的主胶囊生成 RNA 二级结构 2D 图像基本特征的组合，仍用多维张量表示，与卷积层很相似，仍主要储存低级别特征。第三层数字胶囊层接受多维张量，首先该层的每个胶囊通过权重矩阵将低维输入空间映射为高维输出空间，然后经过数次动态路由迭代和损失函数修正，最终得到一个二维矩阵，即一组多维向量。后三层在训练时每个全连接层依次从正确的数字胶囊中接受一个多维向量，学习将其解码重建为 RNA 二级结构 2D 图像，损失函数与重建图像和输入图像的欧氏距离有关，重建图像与输入图像越接近越好。考虑到中间结果应为点括号表示法的一段 RNA 序列表示，还应额外增加一个输出层，将胶囊神经网络输出的矩阵转化为一组概率向量。每个向量包括三个数据，即该位置是“（”、“.”或“）”的概率。

在胶囊神经网络的训练上，对于部分参数放弃传统神经网络使用的 BP 算法，使用名为“囊间动态路由”（Routing-By-Agreement）的算法迭代训练胶囊层。

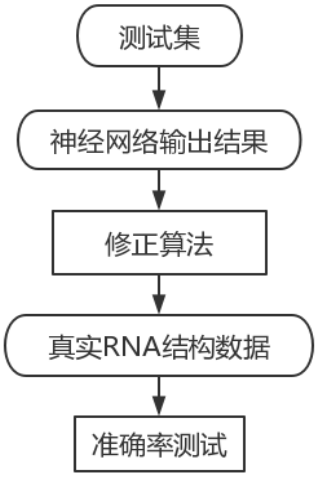
Procedure 1 Routing algorithm.

procedure ROUTING($\hat{\boldsymbol{u}}_{j|i}, r, l$)
  for all capsule i in layer l and capsule j in layer $(l+1)$: $b_{ij} \leftarrow 0$.
  for r iterations **do**
    for all capsule i in layer l: $\mathbf{c}_i \leftarrow \texttt{softmax}(\mathbf{b}_i)$ ▷ `softmax` computes Eq.
    for all capsule j in layer $(l+1)$: $\mathbf{s}_j \leftarrow \sum_i c_{ij}\hat{\mathbf{u}}_{j|i}$
    for all capsule j in layer $(l+1)$: $\mathbf{v}_j \leftarrow \texttt{squash}(\mathbf{s}_j)$ ▷ `squash` computes Eq.
    for all capsule i in layer l and capsule j in layer $(l+1)$: $b_{ij} \leftarrow b_{ij} + \hat{\mathbf{u}}_{j|i}.\mathbf{v}_j$
  return $\mathbf{v}_j$

囊间动态路由算法的一种实现

（4）结果的评估与修正

预测结果可能存在如下两个问题：1.预测的结构自身已违反 RNA 碱基互补配对定律，即左括号与右括号个数不相等；2.预测的结构能做到相互匹配，但是括号对应的碱基不能配对。需要设计概率和最大修正算法来解决这两个问题，即找到一条表示 RNA 二级结构的点括号序列，使得：1.序列括号匹配；2.序列中互相匹配的括号对应的位置上的碱基也相互配对；3.在满足 1 和 2 的条件下，序列中各个符号对应的碱基在胶囊神经网络下输出的概率之和达到最大。得到满足 RNA 二级结构定义的最优二级结构后，将其与对应的真实结构对比，衡量预测的准确性。



针对 RNA 二级结构预测的准确性的衡量，目前的方法是根据敏感性、特异性这两个指标进行衡量。通常的机器学习算法的衡量标准有真阳性（True Positive）、真阴性（True Negative）、假阳性（False Positive）和假阴性（False Negative），分别简写为 TP、TN、FP 和 FN。在 RNA 二级结构预测中，TP 表示正确预测的碱基对个数；FN 表示真实结构中存在但没有预测出来的碱基对个数；FP 表示真实结构不存在却被错误预测的碱基对个数。TN 表示正确预测的不配对碱基的个数，由于 TN 在 RNA 二级结构预测没有实际意义，在衡量准确率上很少使用。敏感性是指真实结构中所有碱基对被预测出来的百分比，对应着机器学习中的查全率（Precision）；而特异性指所有预测到的碱基对中正确的百分比，对应着机器学习中的查准率（Recall）。RNA 二级结构预测算法在一般情况下很难做到两者兼顾，总是偏向一边，可以使用 F1 值来衡量查准率与查全率。具体计算公式如下：

$$R=\frac{TP}{TP+FN}\quad P=\frac{TP}{TP+FP}\quad F1=\frac{2PR}{P+R}$$

3. 人员分工

数据预处理，筛选和编码：张雨凡，张瀚文。

神经网络模型的架构与训练：孙博文，雷浩洁，黄正坤。

后期测试，评估与改进：由 5 人共同完成。

六、项目进度安排（文献查阅、社会调查、方案设计、开题报告、实验研究、数据处理与分析、研制开发、填写结题表、撰写论文和研究报告、结题答辩和成果推广等时间安排）

- 2019 年 1 月-2 月：文献查阅、方案设计阶段。
- 2019 年 3 月-5 月：开题报告阶段。
- 2019 年 6 月-12 月：实验研究、数据处理与分析阶段。
- 2020 年 1 月-3 月：填写结题表、撰写论文和研究报告阶段。
- 2020 年 4 月：结题答辩和成果推广阶段。

七、项目研究所需资源（实验室、仪器设备、实验材料、资料等）

仪器设备：计算机。

资料：分子生物学，机器学习，深度学习领域的论文，书籍资料。

八、项目经费预算与用途（购置实验消耗材料、低值品、资料、加工测试、打字复印、调研、市内公交、论文发表、专利申请等经费开支）

1. 资料，打字复印等：2000 元

2. 论文发表：800/页，共 8000 元

总计 10000 元

九、项目完成预期成果（成果形式：研究论文、专利、设计、产品、软件、研究或调研报告等）

1. 申请软件著作权一份。

2. 提交研究报告一份。

3. 发表高水平论文一篇。

十、项目诚信承诺

本项目负责人和全体成员郑重承诺：该项目研究不抄袭他人成果，不弄虚作假，按项目研究进度保质保量完成各项研究任务。

项目负责人签名：　　年　月　日

项目组成员签名：

年　月　日

十一、指导教师意见（从项目科学性、前沿性、可行性、研究性、可操作性和成效性进行评价，是否同意立项）

同意立项　　签 名：

年　月　日

十二、学院评审意见（学术价值、预期效果、研究方案可行性、是否同意立项）

工作组组长签名（公章）： 

2019 年　4 月 24 日

十三、学校意见

年　月　日