清 华 大 学

综 合 论 文 训 练

题目：基于深度神经网络的 RNA 二级结构预测

系　　别：生命科学学院

专　　业：生物科学

姓　　名：林祖迪

指导教师：鲁志 教授

2017 年 5 月 28 日

关于学位论文使用授权的说明

本人完全了解清华大学有关保留、使用学位论文的规定，即：学校有权保留学位论文的复印件，允许该论文被查阅和借阅；学校可以公布该论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存该论文。**(涉密的学位论文在解密后应遵守此规定)**

签　名：_____导师签名：_____日　期：_____

中文摘要

RNA是一种重要而特别的信号分子，其二级结构对RNA的各种生物学功能有着重要的意义。RNA二级结构测定方法成本很高，计算方法较之更加高效。常见的结构预测方法基于最小自由能模型和动态规划算法，其弱点在于需要依赖实验测定的热力学参数，以及不能预测含假结的RNA二级结构。为了解决这些问题，我们开发了基于深度神经网络的RNA结构预测工具，称为DeepFold。DeepFold仅从已知结构的序列中学习RNA的折叠规律。其不依赖能量模型和参数，并且能预测假结。在控制序列相似度的情况下，DeepFold在一维结构预测上超越了主流的RNA结构预测工具，而二维结构的预测表现也与之相接近。我们同时也用所有收集到的已知结构序列训练DeepFold模型，得到了更为实用的结构预测工具。

关键词：RNA二级结构；结构预测；深度神经网络

ABSTRACT

RNA is an important and unique information molecule whose secondary structures are crucial for its diverse biological functions. Deciding RNA secondary structures by experiments is time- and resource- consuming, while computational prediction is more efficient. Common RNA structures prediction methods are based on minimal free energy assumption and dynamic programming algorithm, which make them rely on experimentally decided thermodynamic parameters and unable to solve pseudoknot-containing structures. For surmounting those obstacles, we present a new approach, called DeepFold, to infer the secondary structure of RNA directly from sequences with deep neural networks. DeepFold can learn RNA folding mechanism directly from structure-known sequences without any energy assumption or parameters, and can also predict pseudoknots. While controlling the similarity between sequences, DeepFold out-perform common methods on one-dimensional structure prediction, and achieved similar-level performance on secondary structure predictions. We also used all collected structure-known sequences to train DeepFold, which finally gave us a practical prediction tool.

Key words: RNA secondary structure; Structure Prediction; Deep Neural Networks

目录

第 1 章　引言

1.1 RNA二级结构及其测定方法

RNA作为一种重要而特别的信号分子，广泛地参与到了各种各样细胞活动中，并在生物体信号传导的各条途径里处于一个中心的地位。在过去的很长一段时间里，RNA的功能被认为仅仅是实现遗传信息从DNA向蛋白质传递的载体。而近年来越来越多的研究表明，RNA分子除了直接参与翻译以外，还拥有翻译调控[1]、细胞内定位（localization）[3]、催化（catalysis）[4]和剪接（splicing）调控[5]等各种功能，并且这些功能在很大程度上是由RNA分子的二级结构所实现的。举例来说，mRNA的部分序列在与细胞内某些小分子结合的情况下可以通过二级结构的改变形成所谓的核糖开关（riboswitch），控制翻译的速率，甚至可以改变蛋白翻译的终止位点，以达到调节基因表达的目的[6]。另一个例子是第一类内含子（group-I intron）依赖于RNA分子精确的构象变化来催化自身的剪接[4]。

RNA分子的二维结构（或二级结构）是指RNA上碱基互补配所的模式。二维结构再经过折叠形成复杂的三维结构。RNA分子上有四种最常见的碱基，分别是腺嘌呤（Adenine，A），鸟嘌呤（Guanine，G），胞嘧啶（Cytosine，C）和尿嘧啶（Uracil，U）。在这四种碱基中，绝大多数碱基配对情况是A-U，C-G和G-U。所以这三种配对方式也被称为权威配对（canonical pairing）。不同于蛋白质的三维结构，目前人们对RNA的三维结构的了解还处于一个相对初级的阶段，这是因为RNA三维结构较蛋白质结构来说更为复杂，并且更加地不稳定。由于RNA的二级结构相对于三级结构来说稳定得多，并且RNA的二级结构已经能很好地解释它的许多生物学功能，再加上从RNA的二维结构中已经可以大致地构建其三维模型，所以目前RNA二维结构的高效测定和准确预测依然是RNA结构研究领域的重点。

RNA的二级结构可以由多种实验方法测定。X-ray晶体衍射实验可以精确地测定RNA的三维结构，并由三维结构得到其二级结构。但这种方法成本很高，操作难度也很大，以至于无法高通量地得到RNA的结构。SHAPE和DMS是低通量测定RNA二级结构的常用化学探测方法，其核心技术是通过分析DMS（dimethyl sulfate）等化学小分子修饰[9-11]或结构特异RNA酶降解[12]后得到的RNA产物的凝胶电泳结果来获取RNA的二级结构。随着高通量测序技术的完善和普及，这些低通量的

化学探测技术与RNA-seq技术相结合，产生了DMS-seq[9]，icSHAPE[13]和PARS[12]等高通量测定RNA二级结构的技术。但这些技术的主要缺点在于其只能用于测定RNA分子的一维结构（即单个碱基形成互补配对的概率），而无法测定RNA二维上碱基的相互配对情况。最近，PARIS技术[14]通过碱基的交联的方法，还有RPL[15]技术所使用的类似于3C技术的临近连接方法（proximity ligation），可以直接测定碱基互补配对的关系，得到RNA的二维结构，不过准确性与晶体衍射实验相比还有提升的空间。

与实验测定二级结构的方法相比，计算方法不仅成本低、通量高、稳定性强，更重要的是一个好的算法将适用于所有的RNA序列，包括很多因为低表达、蛋白干扰等因素而造成的传统实验方法无法测定的序列。因此，本课题的研究目标就是开发准确而高效的RNA二级结构预测算法。

1.2 预测RNA二级结构的算法

预测RNA二级结构的算法有很多种，其中最常用的是基于最小自由能（minimizing free energy）模型和动态规划（dynamic programming）算法的最优RNA结构预测，主要代表是Mfold[16], RNAfold[17]，RNAstructure[18]等。早期的预测模型使用动态规划算法来预测含最多氢键数目的RNA结构。之后随着生物物理学的发展，有了更多更精细的热力学参数，模型预测的准确度也相应地得到了提高。与其类似的方法是基于配分函数（partition function）的方法[19]。这种方法可以预测在多种可能结构中碱基互补配对的概率，并给出最大期望准确度（maximum expected accuracy）预测[20]。第三种方法是用自然语言处理中常用的随机上下文无关语法（stochastic context-free grammar，SCFG）来描述RNA二级结构[21]。这个方法可以用于模拟RNA结构的概率分布，并以此作为RNA结构预测的依据，同时它也可以用于RNA二级结构的保守性分析[24]。

以上提到的基于动态规划算法和随机上下文无关语法的预测工具从计算原理上来说有一个缺陷，那就是不能很好地预测含有假结（pseudoknot）的RNA二级结构。假结是一种特别的配对方式，其特点是一个RNA loop区域的碱基会和其自身stem-loop区以外的碱基配对，主要的功能是帮助RNA分子在空间中折叠成更稳定的三维结构，实现RNA的高级生物学功能。举例来说，RNase P的假结区域是其在进化上最为保守的区域，而telomerase RNA的假结对其正常的细胞活动也有重要的意义[26]。遗憾的是，由于假结是一种上下文敏感的结构，在计算复杂性理论

中已经证明了预测含有假结的最低自由能RNA二级结构是一个NP-hard的问题[27]，这也使得现有的基于动态规划算法和随机上下文无关语法的模型无能为力。

除了以上提到的几种算法外，还有一种不得了不提的二维结构预测方法就是序列比对分析法（comparative sequence analysis），其基于的假设是RNA的二级结构比其序列在进化上有更强的保守性。所以，如果RNA的两个可以配对的位点在进化上有很强的一致性（比如一条RNA上的一对A-U碱基在其同源序列上是C-G），那这两个位点就有很大的概率形成互补配对。基于序列比对的RNA结构预测算法是目前所有RNA二级结构预测算法中最准确的，其准确度能够达到90%以上。但是要成功使用这一分析方法预测结构需要有大量的同源RNA序列来提供序列比对信息；并且这一预测方法目前还没有自动化的软件问世，其主要还是依赖于有经验的研究人员的实现[31]。

针对前文提到的高通量RNA一维结构测定技术（PARS，icSHAPE等），近期也出现了相应的预测算法。其做法是用长度为奇数的滑动窗口在RNA上取样，并以训练集中窗口中间位点的配对概率作为输入的标签，然后用一个简单的三层神经网络进行拟合[32]。训练出的模型能对所有长度为13nt的RNA片段正中间的碱基给定一个配对的概率。虽然R.D. Ponti等人使用了非常小的滑动窗口，但是也取得了不错的准确度。其原因可能是：（1）没有去掉相似的序列；（2）不同的结构标定化学分子在结合碱基时有其内在的倾向性，而这一倾向性正好被模型所捕捉并用于了之后的结构预测。这一点可以从同一实验数据内预测（PARS预测PARS）准确度高而跨种类预测准确度不高（PARS预测icSHAPE）这一结果中看出。

RNA分子的三维结构比蛋白质更加地的复杂和多变，所以RNA三级结构的预测也更加困难。所以常用的RNA三级结构预测方法都会以RNA的二级结构为基础，再加上溶液中的RNA与溶剂相互作用的实验数据等信息作为约束条件，来提高RNA三级结构预测的准确度和效率[33][34]。所以，如果我们能提供高质量的RNA二级结构，RNA三级结构的准确性也能得到很大的提高。

1.3 深度神经网络

深度神经网络（Deep Neural Networks）是一种在近十年内快速发展起来的机器学习技术。到目前为止，深度人工神经网络（或称为深度学习技术）已经在图像识别[35]、自然语言处理[37]和游戏决策[38]中展现出了前所未有的能力。最早的神经网络的出现是人们为了通过模拟高等动物神经系统的物理连接与神经元计算

规律而让机器拥有“学习”的能力。而用数学的语言来说人工神经网络的作用是对复杂的映射进行参数化的拟合。相比于传统的机器学习方法，深度神经网络的优势在于其作为一种强大的表示学习（representation learning）方法能够适应真实原始的数据集，而不需要人工将输入通过精巧的方式转变成传统机器学习模型能够探测或者分类的模式。

在机器学习中最常见的一类学习问题是监督学习（supervised learning），其特点是先用有标签的数据对模型进行训练，而训练好的模型能对无标签的数据进行分类或者计算其对应的映射值（拟合）。对于监督学习的问题来说，以映射$\mathcal{F}$表示我们的模型，x为输入，l为其标签，当我们设定了一个损失函数（loss function）E以后（E需要满足数学上“度量”的定义），网络的“学习”过程用数学的语言来表述就是优化模型的参数使得E($\mathcal{F}$(x),l)的值尽量地小。常用的损失函数有均方误差函数和交叉熵函数等。本课题中所涉及的学习问题都是监督学习。

最简单的深度神经网络被称为多层感知机。如图1.1第一张图所示，多层感知机通过线形变换将当前输入的信息传到下一层，而下一层的人工神经元则会对输入做一个非线形地激活，并将激活值传输给再后一层的神经元。由于非线形激活函数的引入，神经网络在监督学习中所展现出的拟合能力大大地超过了传统的线形拟合模型。除了图中所示的多层感知机以外，常见的神经网络架构还有卷积神经网络（CNN，图1.1第二张图）和循环神经网络（RNN，图1.1第三张图）等。卷积神经网络的特点是权值共享和局部感知，便于提取位于输入的不同位置的相似特征；而循环神经网络的特点是加入了同一层神经元中的连接，可以用于整合当前输入以及其上下文的特征。

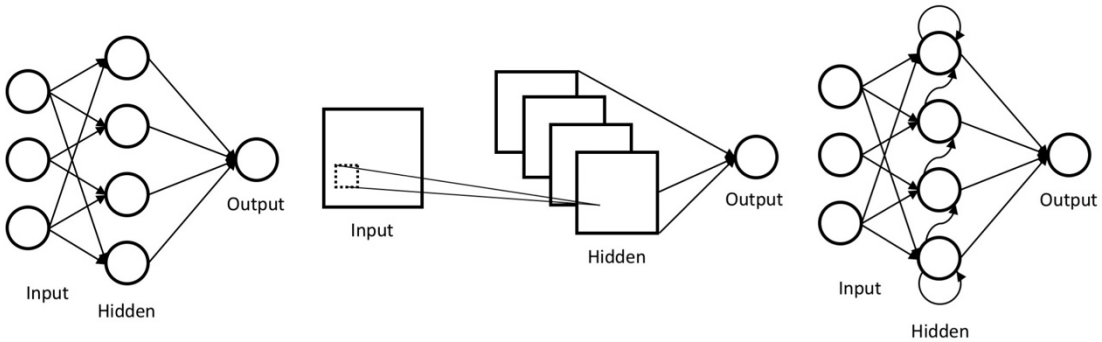

图1.1 常见的神经网络架构：MLP、CNN和RNN

神经网络的训练算法是基于反向传播（back-propagation）的随机梯度下降算法。早期的深度神经网络难以训练，主要是在训练中容易出现梯度消失和梯度爆炸的情况，导致模型不收敛。为了解决这些问题，计算机科学家们逐渐提出了修正线形单元（Rectified Linear Units, ReLU）[36]，Batch Normalization[30]，基于mini-batch的梯度下降算法（SGD），动量学习法（Nesterov momentum），残差学习（residual learning）[29]框架等一系列新的优化方法，使得神经网络在结构不断复杂化的同时的也能更加容易和高效地进行训练。近几年来得益于GPU并行计算的快速发展，神经网络的训练速度以及可处理的数据规模也得到了大幅的提高。

深度神经网络技术如今也慢慢地在基因组分析和生物影像的处理中得到了应用。在基于序列分析的生物学研究中，深度神经网络已经成功地实现了预测RNA可变剪接[28]，蛋白质与DNA和RNA的特异性结合[25]，以及预测非编码区基因突变的影响[22]，并且相比于其他方法达到了更高的准确度。

1.4 研究计划概述

本课题的研究目标是设计和训练基于深度人工神经网络的RNA二级结构预测方法，我们将其命名为DeepFold。DeepFold将能够做到纯序列预测，即在仅仅提供RNA一维序列的情况下就能由模型给定其二级结构，而不需要任何其他信息的输入，也不需要针对不同的输入序列设置相应的参数。我们也希望我们所设计的RNA结构预测流程能够将RNA二级结构预测的问题从一个传统的生物物理学问题转变成一个定义清晰的（well-defined）机器学习问题，这样将吸引更多机器学习领域的专家来从事这方面的研究。

传统的RNA二级结构算法依赖于人工设计的RNA结构形成的最近邻模型（nearest neighbor model）和通过热力学实验测定的能量参数。之后有ContraFold方法[23]用机器学习的方法从以已知的RNA二级结构作为训练数据集，学习到了模型所需的能量参数，但是RNA结构生成模型SCFG仍然需要人工设计。深度学习在图像识别、自然语言处理和游戏决策等复杂问题上的成功应用，使得我们相信人们很有可能完全通过机器学习方法，就从已知结构的RNA数据中学习到RNA二级结构形成的物理规律，并用参数化的模型来表示。就目前来看，基础的深度学习理论研究已经趋于成熟，而且有Keras、TensorFlow、Theano和scikit-learn等整合的深度学习软件包被开发出来。利用这些软件包，我们可以在无需完全熟悉RNA

折叠的物理和化学理论细节的情况下，针对我们要解决的问题快速搭建和训练模型，得到新型的结构预测工具。

我们所设计的RNA二级结构预测算法开发的研究方案分为以下几个主要的步骤。第一步是收集与处理RNA二级结构数据。我们需要的数据是有精确测定的二维结构的RNA序列，所以高通量实验给出的只含有一维结构信息的序列和二维碱基配对以概率的方式给出的序列都不能使用。第二步是设计序列的编码方式。虽然说深度神经网络有很强的利用原始数据的能力，但RNA序列毕竟不像图像那样可以直接让模型读取。并且在训练和预测时需要提供一条RNA多少的信息也是需要考虑的地方。第三步是，在设计好合适的序列编码方式后，我们将运用机器学习模型，特别是多种深度神经网络，对RNA的一维和二维结构进行预测，并比较不同方法预测准确度，计算效率等方面的优劣。在挑选出最准确最高效的模型架构之后，我们也会将其预测表现与常用的基于能量模型的RNA二级结构预测算法进行比较。最后，我们课题的目标不光是尝试和比较预测算法的好坏，我们还会完善训练好的模型，并将其做成一个开源的RNA结构预测工具。

1.5 课题的意义和价值

我们设计基于深度神经网络的RNA二维结构预测算法的最主要的目的是替代传统的能量模型。传统的基于最小自由能假设的模型从物理学原理上来说是没有任何问题的，但正如前文所提到的那样，基于最小自由能模型的RNA二级结构预测算法有两个最薄弱的环节：一是其预测依赖于实验测定的能量参数，而是其算法的核心为动态规划（dynamic programming）。热力学参数的测定首先需要消耗大量的时间和实验资源，同时所测定的热力学参数的微小误差在预测较长的RNA时会被放大很多，造成预测的准确度大幅降低。另外一方面，有些结构较为复杂的RNA如RNA病毒、rRNA还有RNase P中含有很多的假结，而这些假结对其三级结构的正常细胞功能的维持有关键作用。而传统模型并不能很好地预测假结。通过从已知结构的RNA数据中建立模型，预测RNA二维结构，我们不仅可以优化假结的预测，同时还有可能对未来RNA-RNA以及RNA-DNA相互作用的预测提供新的思路。

基于纯序列的RNA结构预测算法还可以帮助我们探索RNA序列在多大程度上决定RNA的二级结构。之前DeepBind的预测结果表明，仅仅从RNA序列出发就可以精确地预测DNA、RNA与结合蛋白的相互作用[48]。同样地，如果我们可以从

纯序列出发很准确地预测RNA的二级结构，就说明RNA的一维序列已经对RNA二级结构的决定提供了足够的信息。

目前卷积神经网络、循环神经网络等前沿的深度学习模型还没有在RNA二级结构预测领域得到应用。本项目将是第一次在该领域对深度学习模型的探索。如果深度学习可以达到与基于自由能模型的RNA二级结构预测工具类似或者更高的准确度，那么也可以证明从已知结构的RNA数据本身可以重建复杂的热力学模型，从而为深度学习在生物学领域更广泛、更复杂的问题上的应用提供了支持。

第 2 章　算法与结果

2.1 RNA结构数据的收集与处理

2.1.1 训练和测试数据的选择

我们共收集了3324条RNA，包括transfer RNA（tRNA）、ribosome RNA（rRNA）、telomerase RNA（tmRNA）和RNase P等种类。这些RNA的长度分布在28-780nt之间（其中16s和23s等长度超过1000的rRNA在收集的数据中已经被切分成了较小的结构域），平均长度为200.2nt。在统计了RNA二级结构中碱基的配对距离后我们发现绝对配对距离的均值为79.4nt，而中位数为33nt，这说明多数配对还是发生在一维序列上距离较近的碱基之间。不过在统计相对配对距离之后（即绝对配对距离除以其所在的RNA序列的长度，为一个0-1之间的数值），我们发现相对配对距离的分布在0.75-1.0之间的密度有一个显著的升高，这说明还是有不可忽视的一部分碱基会与其所在RNA上距离较远的碱基互补配对。同时，从配对位置的热图（heatmap）中可以看出，一条RNA的5’头端碱基和3’尾端碱基有很明显的配对倾向。这些现象说明常见的仅用要预测的碱基位点和其上下游临近的一段序列来预测配对情况的方法会损失掉许多信息。这也启发了我们去使用一种包含全序列信息的数据编码方式作为模型训练和预测的输入。

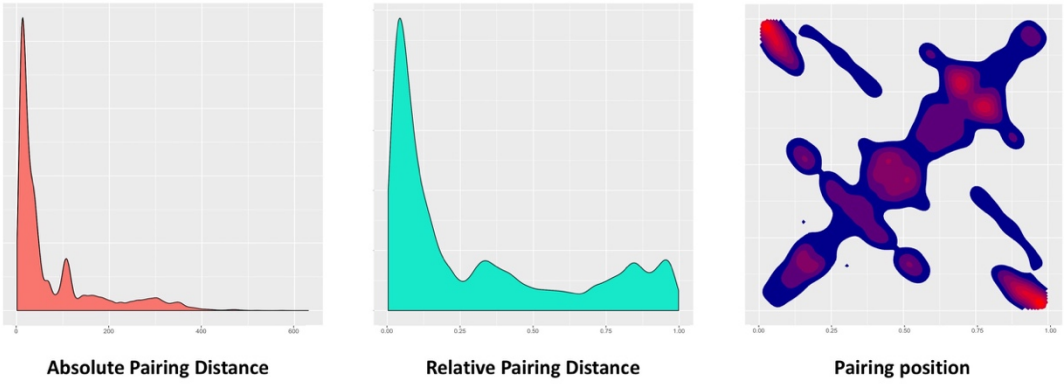

图2.1 RNA的绝对配对距离，相对配对距离和配对位置

基于高通量测序的RNA结构测定技术（PARS，icSHAPE等）可以较准确地测定细胞内（in vivo）真实的RNA结构状态，但是在应用上也有一些局限性。首先，

这些高通量结构测定实验给出的结果仅仅是RNA的一维结构，所以使用这些数据训练的模型只能用于一维结构的预测。其次，实验测定过程中的各部分操作容易影响RNA结构测定的结果，尤其是在表达量低的RNA结构测定上较不准确很难测到，因此不同批次的实验结果会有一定的差异。常见的高通量二级结构测定方法测定的都是大量同序列RNA的平均结构。由于处在不同状态下的RNA会在温度和结合蛋白等因素的影响下形成不同的二级结构，所以平均以后的结构并不能精确地反映RNA真实的二级结构状态，不适合于我们算法的构建目标。另外，一些化学小分子（如DMS）只能修饰特定的碱基，或者对不同碱基的修饰具有选择性，这些会进一步增加结构测定数据的误差和可用数据的数量。同样，高通量二级结构测定技术（如PARIS，RPL）所产生的数据也有类似的问题。这些实验技术的缺陷使得我们在使用这些数据进行网络训练时会让网络产生碱基倾向性上的偏差，影响网络的预测效果。所以在本课题中，我们没有使用高通量结构探测技术的数据，而是使用X-ray衍射和低通量化学探测法所得到的学术界公认的精确二级结构来进行算法的测试。

2.1.2 RNA序列相似度的控制

在我们收集的三千余条RNA中，许多同一种类的RNA之间的相似度非常高，这将会导致算法的预测结果没有说服力。因为就算给出了精确的预测结果，也不能说明训练出来的模型是真正学到了RNA形成二级结构的重要特征，还是仅仅是由序列高相似度以及过拟合（over-fitting）而造成的虚高准确度。为此我们在设计算法之前首先对原始数据集基于相似度进行了筛选。具体的做法是，首先将所有RNA随机排列成一个有序的序列，然后按顺序将序号较大的序列与序号较小的序列用基于Needleman-Wunsch算法的EMBOSS Needle[2]全局序列比对工具进行比对，计算出这两条序列的相似度。如果相似度大于60%，则将序号较大的那条序列给删除掉，反之则保留。我们不断重复这个过程，直到最后剩下来的序列满足：任意两条序列都完成了相似度的比对，且相似度都小于60%。经过这样的筛选之后我们最后保留下了仅540条已知结构的RNA序列。这个数据集又被随机分成了训练集（400条）、验证集（40条）和测试集（100条）用于之后的算法研究。对于深度学习来说，成功训练深度神经网络的一大基础是要有足够量的训练数据。由于已知结构RNA序列数据的稀缺性，我们曾经考虑过能否设计一个算法，使之在去掉相似度高的序列的同时保留最多条的RNA。但是在之后的分析中我们发现，如果将每条RNA视为无向图（undirected graph）中的一个点，而将相似度大于60%

的RNA用一条边（edge）连接起来的话，那保留最多的两两相似度小于60%的RNA序列就正好对应了一个最大独立集（Independent Set）问题。最大独立集问题在计算复杂性里属于*NP*-complete问题，在现有计算机架构下无法高效求解，其计算时间随着数据规模的扩大会成指数上升。考虑到算法的运行效率，再加上多出来的序列对最终的预测结果不会有太大的影响，我们最后选择采用前一方法筛选出的数据集进行研究。

2.2 RNA序列数据的编码

2.2.1 一维结构训练和预测所使用的编码方式

要想顺利地使用深度神经网络，很重要的一点在于数据的编码。对于本课题来说，我们只有400条已知二级结构的RNA用于模型的训练。如果将每一条RNA当作一个训练样本，而用神经网络直接去拟合RNA一维序列到RNA二级结构（以邻接矩阵表示）这一复杂的映射的话，这个数据量是远远不够的。同时，在前面的分析中我们也提到过，为了达到一个全局最优的结构，RNA上的部分碱基会倾向于和一维上距离较远的碱基形成互补配对，所以仅用长度为十几个碱基的滑动窗口取样的方式会遗失掉许多有用的信息。基于以上两点，此我们设计了一个新的数据编码方式，我们将其称为循环one-hot编码。这个编码方式的原理如图所示。One-hot编码方式将A、U、C和G四种碱基所形成的核苷酸分别对应到四维空间四个标准正交向量中的一个（$[1,0,0,0]^T$、$[0,1,0,0]^T$、$[0,0,1,0]^T$、$[0,0,0,1]^T$），这是在计算生物学领域编码DNA和RNA序列的一种常用的方法。但于基于滑动窗口的编码方式不同，我们为了让每个输入样本包含全序列信息而使用一种循环编码方式，其做法为：首先设置一个长度为w的序列容器（为了保证对称性，w为奇数），即一个全零的矩阵，然后将要进行训练／预测的核苷酸放在这个容器的正中央，其所在的RNA序列在容器中自然延伸。如果3’端的RNA序列超过了容器右面的边界，那么就将多出来的这部分序列移到容器的左端，形成一个循环的结构。对于5’端也是一样地处理。接下来就将每一个碱基变成其对应的四维向量，而没有序列信息的部分还是0。具体的编码结果如图2.2所示。这个编码方式有至少三个好处。第一个好处是将每一个RNA上的核苷酸编码成了一个样本，并且可以认为每个样本在统计上是属于独立同分布的。这是因为以每一个核苷酸为训练／预测目标的样本都包含了所在RNA的全序列信息，而我们的假设是不同RNA所遵循的折叠规律是一致的。第二个好处是将直接预测RNA二级结构的问题变成了一个易于处理

的二分类问题。在前文中我们也谈到过，直接用神经网络拟合一维序列到二级结构的映射在原理上是可行的，但是由于此映射的复杂度太高，能收集到的有二级结构信息的RNA序列的数量远远不足以训练如此复杂的网络。比如，之前有工作用深度神经网络从蛋白质一维序列出发直接预测蛋白质的二维接触图（contact map），其做法是将长度为*L*的蛋白质一维序列（同样以矩阵表示）通过神经网络的卷积、非线形激活和采样等操作直接映射成一个*L*x*L*的二维矩阵，并用已知的蛋白质的二维接触图作为拟合的目标[7]。这是一个非常聪明的办法，但是这种在蛋白质预测上取得成功的办法不能运用到RNA结构的预测中是因为已知二级结构的RNA数目远远小于已知二维接触图的蛋白质数目，因此我们能使用的RNA数据量远远达不到训练类似蛋白质二维接触图的预测的深度神经网络所需要的数据规模。除此之外，Wang et al.在预测蛋白质接触图的过程中加入共进化（co-evolution）、成对接触（pairwise contact）和距离势能（distance potential）等信息，并且这些信息对提升预测表现有很大的帮助。由于我们这个算法的设计目标是完全基于一维序列进行预测，所以我们也没有采纳这一方法。不过我们在以后的二维结构预测模型中也使用了Wang et al.所使用的残差学习框架。

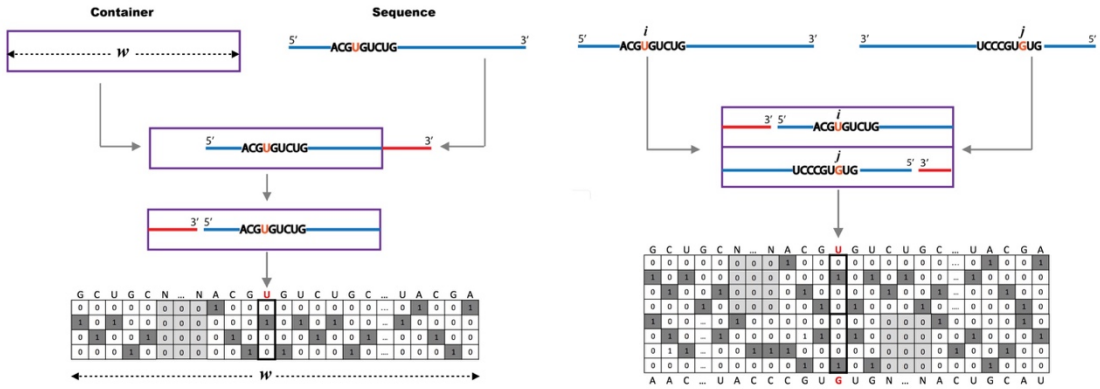

图2.2 RNA序列的循环编码方式

2.2.3 二级结构训练和预测所使用的编码方式

对于二级结构预测情况，我们使用的编码方式和一维结构预测有类似之处。如图2.2的第二幅图所示，对于二维模型的预测，如果模型要判断一条RNA上第*i*号位的碱基和第*j*号位的碱基是否配对的话，我们只需要将第*i*个碱基的一维编码和第*j*个碱基的一维编码上下拼接起来，就得到了一个预测二维配对情况的输入。在训练的时候，如果位于输入矩阵中间的碱基对形成了互补配对，那这个输入的

标签就是1，反之就是0。在实际操作中，我们将下方的序列反向了，因为RNA的二级结构中多数的简单环（loop）结构会使序列在局部的碱基配对时满足一个反向互补的关系。经过这样的编码操作以后，二维结构的预测问题也被抽象成了一个二分类问题。这样的二维结构模型训练的方式并不依赖于一维结构预测的结果，只有在预测的时候才结合一维预测信息，便于处理和优化。并且由于每一个输入的样本都包含了其所在RNA的全序列信息，所以样本相互之间依然是独立的。

2.3 评判预测结果的标准

RNA一维结构中配对和不配对的碱基数目相差不大，所以直接以总体准确度来衡量预测表现就可以。而RNA二级结构预测算法的预测表现通常用两个度量方式来评判，分别是灵敏度（sensitivity）和阳性预测值（positive predictive value，简写为PPV）。它们的表达式为：

$$\text{Sensitivity} = \frac{TP}{TP+FN} \quad (2\text{-}1)$$

$$PPV = \frac{TP}{TP+FP} \quad (2\text{-}2)$$

在公式中，TP（真阳性结果）表示模型所预测出来的真实碱基配对的个数，即能在RNA的已知结构中找到的碱基配对。FN（假阴性结果）表示RNA真实结构中有的但模型没有预测出来的配对碱基对。FP（假阳性结果）表示的是模型预测出来的但是在真实结构中并不存在的配对情况。因此，灵敏度反映的是在RNA所有已知的碱基配对中有多大比例被模型正确的预测了出来，而阳性预测值反映的是正确预测的配对碱基对占所有预测出的碱基对的比例。同时考虑这两个统计量才能全面地刻画一个预测模型的表现。在有些情况下研究人员也会使用F1分数（F1-score），即灵敏度和阳性预测值的调和平均数来衡量模型的预测能力。

针对RNA二级结构预测算法的测试，评价预测结果的严格程度被放开了一些。特别之处是，在衡量RNA二级结构预测模型的表现时，弹性配对（flexible pairing）是被允许的。即在*i'*和*i*与*j'*和*j*的误差都不超过一个碱基的情况下，那么预测出的碱基对（i'，j'）就会被认为是对真实结构中的配对碱基对（i，j）的正确预测[18]。

2.4 DeepFold的预测流程

DeepFold的预测流程如图2.3所示。其预测分为两部分，分别是一维结构预测和二维结构预测。对于一条输入模型的长度为L的RNA序列，DeepFold会对L个碱基参与配对的概率分别进行预测，得到RNA分子的一维结构。由于二维的碱基互补配对只会出现在一维上参与配对的碱基之间，所以在二维预测时我们只需要将一维结构中拥有高配对概率的碱基单独拿出来进行预测就行。当完成二维结构预测以后，我们会得到一个矩阵，称为受限制的碱基配对概率矩阵（restricted BPPM），即配对概率矩阵只限制在一维上有高配对概率的碱基中。之后我们会对矩阵中的概率进行一个排序，并将配对概率大的互补配对情况先加入到结构中。这里我们基于的假设是深度神经网络模型给出的高配对概率的碱基对更有可能是正确的。如果一个碱基已经加入到了结构之中，那之后还和其配对的情况都会被跳过。这种从配对概率矩阵得到结构的方法不再使用传统结构预测中所使用的动态规划算法，再加上所有的配对可能性都是独立考虑的，所以可以预测含有假结的RNA结构，以及RNA序列上的远距离配对。同时这种预测流程相当于先在预测一维结构时对二维预测做了简化，这样可以大大地提高二维预测的效率。

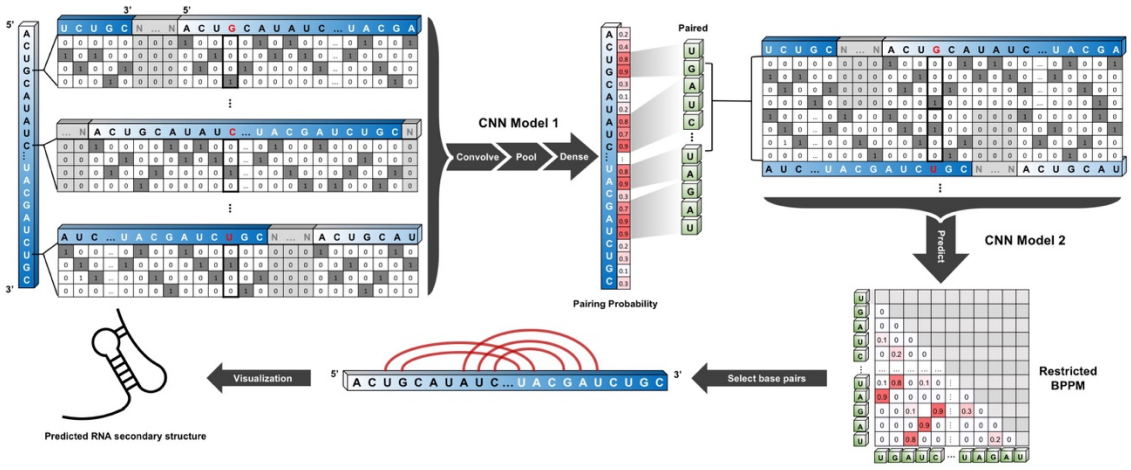

图2.3 DeepFold结构预测流程图

2.5 RNA一维结构的预测

2.5.1 基于Logistic回归的结构预测

对于RNA一维结构的预测，即预测每个碱基是否参与配对，我们首先尝试了最常用的机器学习方法——Logistic回归。Logistic回归作为一种广义的线性回归模型，会对输入向量做一个线性组合，再将结果经Logistic函数的处理，得到一个

（0，1）之间的值。对于Logistic回归常处理的二分类问题来说，如果标签1表示A类样本，0表示B类样本，那这个输出值就可以理解为输入的向量属于A类的概率。在RNA结构预测问题中，标签1表示参与配对，标签0表示不参与配对，而模型的预测值即为碱基参与配对的概率。

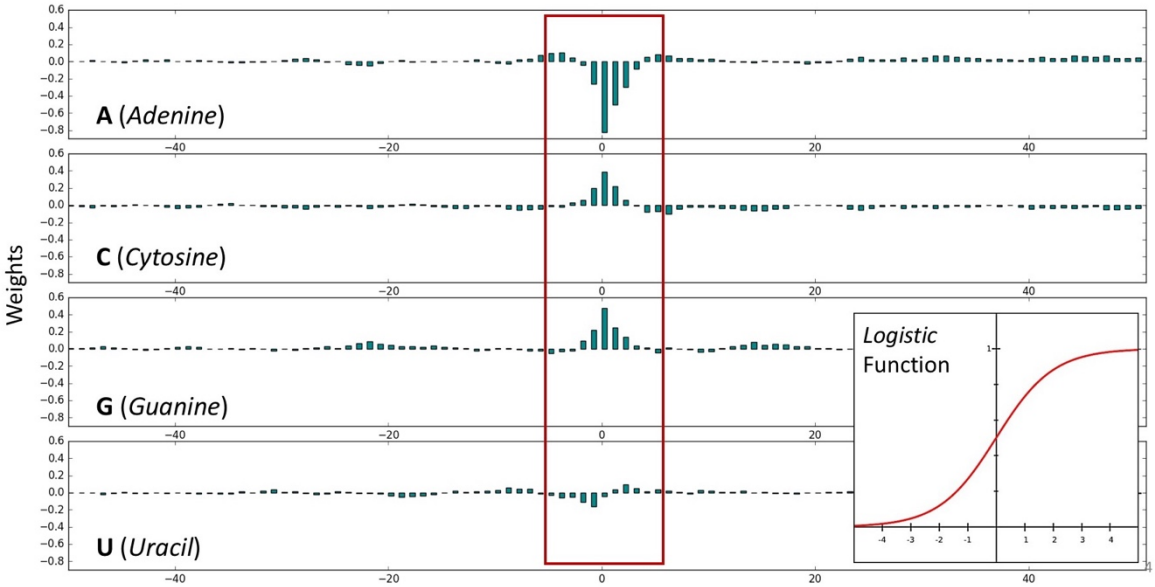

图2.4 Logistic函数和Logistic回归模型的训练结果

对于一维结构的预测，我们将原来的4x*L*的矩阵重构成了长度为4*L*的向量，作为Logistic回归模型的输入。当模型在训练集上的准确度趋于饱和的时候，我们得到的在测试集RNA上的预测准确度为0.645。在对训练得到的模型权值的分析中我们可以发现，绝对值较大的权值，及那些对模型做出分类判定更重要的权值，主要都分布在对应中央的五个碱基的位点。这说明基于Logistic回归的模型认为一个碱基本身及其下上游各两个位点就能决定其配对状态，这与Ponti et al.使用多层感知机只需13nt的序列窗口就能预测PARS和icSHAPE实验数据中RNA的单链概率相类似[32]。在对Logistic模型权值进行分析的时候我们首先用权值的每一列减去了其所在列的均值（长度为4*L*的向量被重构成了4x*L*的矩阵，与编码好的一维RNA序列相对应）。这样的处理去掉了那些由于随机训练产生的对分类的结果影响很小的权值。从权值的分布方式中我们可以看出，腺苷酸（A）倾向于降低配对的概率，而鸟嘌呤（G）和胞嘧啶（C）倾向于增加配对的概率。尿嘧啶（U）在决定配对中起的作用比较小。这一结果符合生物学常识，因为G和C配对将形成三个氢键，多于A-U和G-U配对所形成的两个氢键，而且从物理规律上来说更多的氢键能

更好地减小RNA的自由能而使二维结构达到稳定的状态。但是，这一合理的结果没能达到较好的预测效果的原因在于：（1）从权值的分布情况来说，模型在处理全序列的输入时仅仅考虑了中间5个左右的碱基，而没有考虑远程的信息，没有从全局上考虑碱基配对的各种可能性；（2）Logistic模型的架构决定了最终结果仅仅取决于各个碱基位点信息的线形组合，而没有考虑RNA二级结构决定中的非线形关系，以及不同层次信息，如RNA模体（motif）和局部二级结构等。总地来说，预测准确度较低的原因是Logistic回归模型将RNA结构预测的问题考虑得太“简单”了，这也是我们之后引入深度神经网络来处理这一问题的动机。我们使用Python编程语言的机器学习软件包scikit-learn搭建了这一个Logistic回归的模型。

2.5.2 基于深度神经网络的结构预测

为了使模型能够考虑更复杂的非线形特征，我们引入了深度神经网络。我们最先设计的是一个多层感知机模型，在隐藏层的神经元中使用*tanh*函数来做非线性激活，而预测结果与Logistic回归相比没有明显的提高。由于卷积神经网络已经在生物序列分析中取得了一些不错的成果，我们也将研究的重点放在了卷积神经网络模型上。正如在引言部分所描述的那样，卷积神经网络可以通过卷积、非线形激活、重采样等操作来获取非线形信息。这些高维抽象的信息最终由一个预测器（predictor）整合，得到样本的分类结果。我们设计的网络结构如图2.5所示。每一个卷积神经网络的模块（图中的CNN Block）由卷积、修正线形单元（ReLU）、Batch-normalization和最大值池化（Maxpooling）组成。卷积操作会对当前的输入做一个线形变换，其特点是权值共享，即同一个卷积核会在输入矩阵上以长度为1的步长滑动并对数值进行线性组合，直到处理完所有位点得到新的输出矩阵。之后会使用一个非线形的激活函数——修正线形单元——对线形变换产生的值做一个非线形映射。其中，修正线形单元的数学表达式为：

$$\mathrm{ReLU}(x) = \begin{cases} x & if\ x \geq 0 \\ 0 & if\ x < 0 \end{cases} \quad (2\text{-}3)$$

ReLU激活函数计算简洁，并且不像传统的sigmoid激活函数那样会在参数学习的过程中出现梯度消失的问题。Batch Normalization功能是对于一个批次（batch）的输入，分不同的特征对输入进行标准化，保持网络的每一层输出在输入下一层时保持分布的不变性，使网络的训练更加地高效[30]。最大值池化层功能是计算所给

定的窗口中元素的最大值，并用这一个值来代替整个窗口中的序列信息，降低数据的维度。这样的操作不仅能大大减少卷积神经网络最后一层分类器的参数数量，更重要的是使得网络在之后的卷积和激活操作时能够在一个更高维的层级上考虑序列中的特征信息。

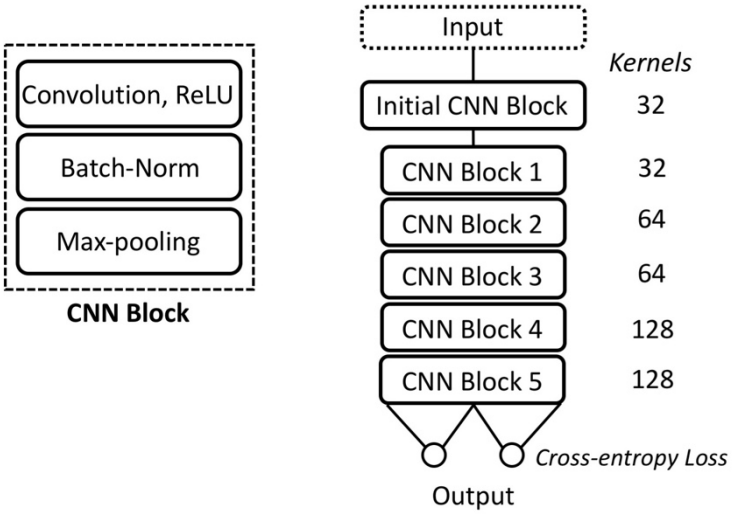

图2.5 用于一维结构预测的CNN模型架构

通常来说，对于第一层的卷积和激活操作可以理解为：模型使用一个RNA序列模体（motif）扫描元件对输入序列进行扫描，然后只将激活值大于0的位置通过修正线形单元保留下来，即决定结构的重要motif所处的位置。激活值小于0地部分会变成0，并在后一层网络执行卷积操作时不提供额外的信息。根据我们的模型设计方式，除了在网络的第一层使用了二维单特征通道（feature channel）卷积外，后面的网络进行的都是一维多特征通道的卷积。深度神经网络的多个隐藏层会对RNA模体的分布图做进一步地卷积和非线形变换，以便对原始序列进行逐层的抽象和特征提取。运用卷积神经网络还有一个好处是其在处理序列时能够保证内部元素相对位置关系的不变性。值得注意的是，在神经网络的最后一层通常会使用一个全连接网络来作为最终的分类器，这能让模型在最终的预测时整合之前多层网络中通过卷积操作并行学习到的各种信息。各个不同特征通道的信息的线性组合将用一个Softmax函数将其值映射到（0，1）区间，输出碱基配对的概率。我们在进行训练时所使用的损失函数是交叉熵函数。网络训练的流程如图2.6所示。

除了搭建多层感知机和卷积神经网络以外，我们还搭建了一个长短期记忆网络（Long-Short Term Memory，LSTM）模型。长短期记忆网络是一种循环神经网络。循环神经网络在处理输入时每次只处理序列的一个元素，但同时会在隐藏层的神经元中保持一个状态向量（state vector），用于记录之前处理过的元素的信息，能够使模型考虑远距离信息，非常适合序列信息的处理。但传统的循环神经网络在处理长序列时的表现并不好，因为较早的输入元素的信息会慢慢消失掉。

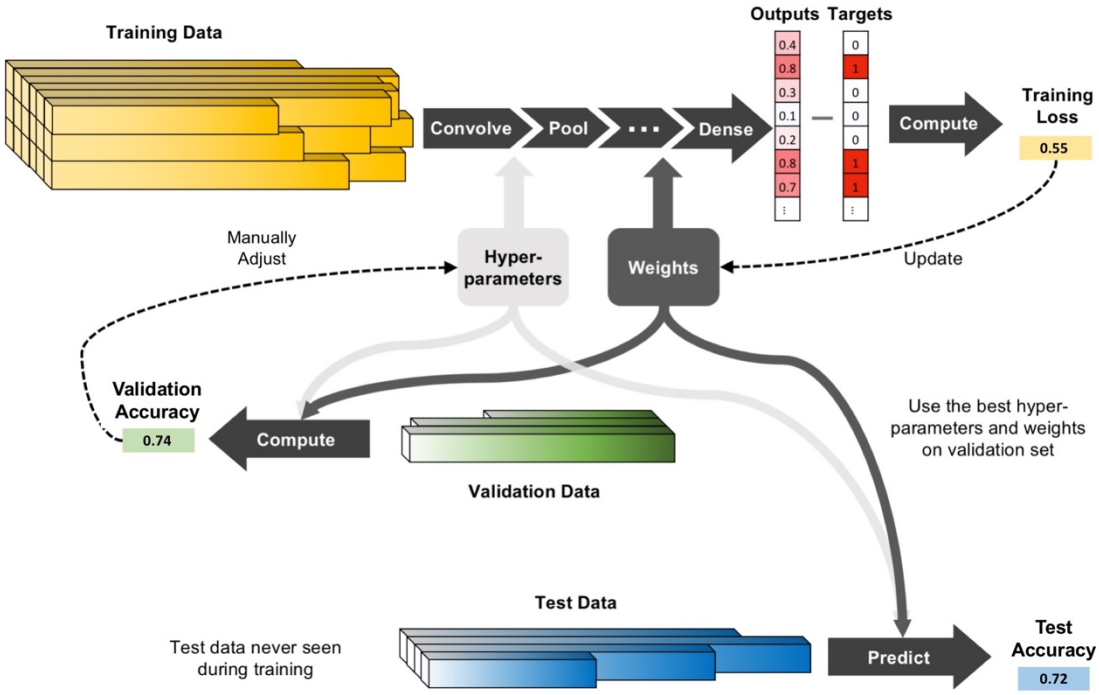

图2.6 DeepFold的训练流程示意图

为了解决这一问题，计算机科学家们发明了长短期记忆网络，用一条网络的信号流专门来保存前面处理过的元素的状态信息[8]。这一网络架构已经在机器翻译、文本下文预测等领域展现出了强大的能力。由于RNA结构预测问题与自然语言处理有很多相似的地方，我们也训练了LSTM网络。由于RNA序列的信息不像自然语言那样有单向性，所以我们除了传统的LSTM还使用了含双向LSTM层的神经网络。双向LSTM层先从RNA的5’向3’端做一次处理，再从3’端向5’做一次处理，得到整合的输出值。除此之外，我们还训练了一个简单的CNN-LSTM网络，即用一个CNN网络提取RNA的序列特征并降低数据维度后再用一个LSTM网络来考虑特征的上下文信息，并最终输出分类概率。在这里我们利用了CNN处理后的序列依

然保有原始的上下文关系的特点。表2.1展示了多个神经网络模型的训练结果（在验证集上的准确性）和训练时间（每个epoch所用的时间）。最初设计的模型在测试集上的预测准确性相比与Logistic回归模型来说都没有太大的提高，并且LSTM等模型的训练效率极低。综合考虑模型的预测准确度和效率，我们决定以当前预测效果最好和训练时间较短的卷积神经网络作为之后的研究重点。我们利用了Python编程语言的深度学习资源包Keras搭建了这两个深度卷积神经网络(基于TensorFlow backend)。

表2.1 不同神经网络在一维验证集上的表现

	MLP	LSTM	Bi-LSTM	CNN	CNN-LSTM
Validation Accuracy	0.641	0.622	0.634	**0.682**	0.626
Epoch Time	**15s**	~2000s	~3600s	47s	~1000s

2.5.3 提升模型的泛化能力的方法

如图2.7（1）所示，在最初的网络训练过程中我们遇到了很严重的过拟合问题，即网络在训练集上的预测准确度很高，而在验证集上的准确度很低。这个现象的产生很大程度上是由于数据量太少，导致网络学到了一些训练集上特有的特征，而缺乏泛化的能力。为了提升网络的泛化能力，我们使用了四种方法。第一种方法是在输入数据中加入更多的有用信息。为此我们在原有的循环one-hot序列编码中假入了两行新的特征，分别是预测的目标碱基，和目标碱基可能的配对对象。对于目标碱基，我们赋给了1.0的权值，而考虑到真实RNA结构中碱基的配对多数不是单一事件，而是相邻的几个碱基都参与配对，所以我们也将目标碱基上下游各两个碱基也赋予了一个较小的权值。对于可能的配对对象这一特征，我们将目标碱基在RNA序列上所有在权威RNA配对规则（AU，CG，GU）中有可能配对的位置用1.0进行了标记。所添加的两个新特征都是从原始序列中得出来的，所以依然是基于纯序列信息的预测。第二个用于缓解过拟合的方法是给输入的样本加上一个高斯（Gaussian）噪音，其方法是在输入样本的每一个矩阵单元上加上一个随机实数α，这个随机数服从均值为0，标准差为δ的高斯分布。在图像处理中，这一方法会将输入的图片变得模糊，以此来让模型学到泛化能力更强的结构或纹理特征，提高模型的鲁棒性（robustness）。我们给每个输入的矩阵加入了标准差

为0.3的高斯噪音。除了这两种方法外，我们还使用了L2正则化（L2 regularization）方法。在前面的实验结果中我们看到，Logistic回归得到的参数在RNA中间的几个位点上的值相对于平均的参数值来说非常地大，这就导致模型只关注少数的序列特征，导致过拟合。为此我们希望能避免网络中出现权重特别大的参数。L2正则化方法在原来的损失函数里加上了一个正则项，其数学表达式为：

$$C = C_0 + \frac{\lambda}{2n}\sum w^2 \quad (2\text{-}4)$$

其中n表示的是样本数量，λ是给定的超参，w是网络的参数，而C_0是原来的损失函数。从直观理解上可以看出，如果网络中出现了权重很大的参数，会使得总体损失函数C的值变得很大，有悖于网络的训练目标。而从式中可以看出，L2正则项使得网络在通过随机梯度下降算法时倾向于减小网络参数，达到权重衰减（weight decay）的效果。我们为减小过拟合问题采用的最后一种方法被称为Dropout，其作用是在训练的时候随机屏蔽掉一些神经元的连接，使网络变得稀疏，让网络从更少的输入中学到分类的特征，避免网络权值的更新依赖于某些固定的并且会造成网络输出偏差的节点组合关系，同时也能在一定程度上提升训练速度。从图2.7

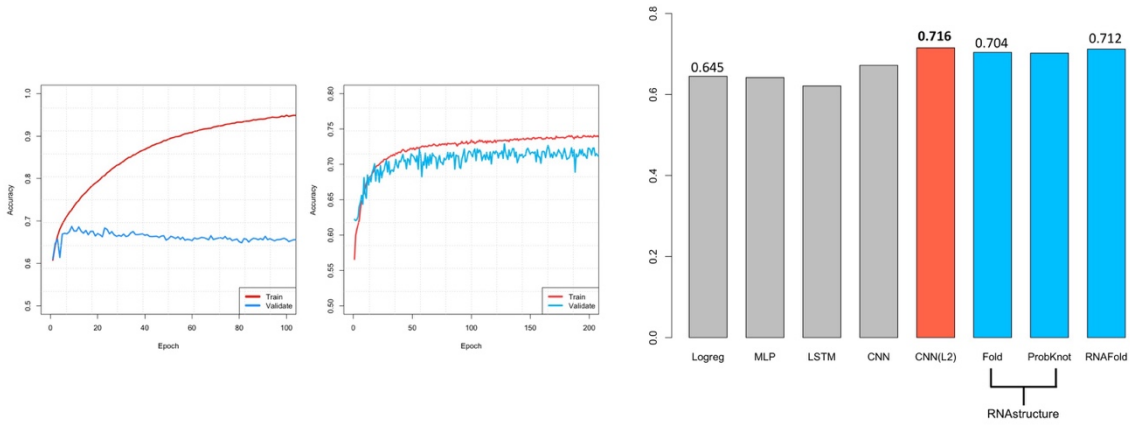

图2.7 一维预测模型的训练曲线和预测结果

的前两幅子图中可以看出，在加入这些限制网络过拟合的方法之后，虽然训练集上的准确度下降了不少，但是验证集上的预测准确度从0.682提升到了0.738。同时我们计算了测试集上的预测准确度，并将一维的预测结果与主流RNA二级结构预测软件进行了比较。可以看到，加入了各种正则化方法之后的DeepFold一维模型

的预测准确度已经略高于RNAstructure和RNAfold等基于能量模型的主流RNA结构预测软件。

2.6 RNA二维结构的预测

二维结构的预测较一维结构的预测来说更加地困难，其原因有二。第一是样本的不均衡分布。在没有对输入样本进行控制和筛选的情况下，负样本（即在真实结构中不配对的碱基对组合）是正样本的300余倍。第二是数据分布空间的复杂性。二维网络的输入维度比一维网络的输入维度更高，并且样本的数量也有数量级上的变化。

为了缓解正负样本不均衡的问题，我们首先对输入的样本进行了控制和筛选。根据对已知结构RNA的统计，平均每条RNA序列中都大约有一半的碱基会参与配对，所以在DeepFold的进行二维结构预测时，由于我们只考虑一维上配对的碱基，所以在这些碱基上产生的可能的配对情况就变成了一条RNA上两两配对可能的组合数的1/4。同时，由于已知结构的RNA序列中超过99.9%的配对是权威配对（AU，CG和GU），所以我们在训练二维结构模型和进行最后的结构预测时都只考虑权威配对。这样的操作不仅不会对二维结构预测的准确度带来太大的影响，还有能因为样本空间复杂度的降低而让模型更容易拟合，并且也节省了很多计算时间。对于我们使用的含540条RNA的数据集，当只考虑一维上形成双链的碱基和权威配对的情况下，我们将负样本对正样本的比值降到了78。虽然进行了输入的筛选，但是数据分布不均的问题依然严重，训练出的模型会有很强的偏差。可以想象，如果模型将所有样本都判定为负样本，那么模型的整体准确度会达到98%以上，但是这对最后结构的预测没有任何作用。为此我们使用了一个循环训练的方法。对于一次循环，我们先将负样本随机分成78份，然后从第一份开始依次和所有正样本输入网络中进行训练。当78份样本都完成训练之后，负样本会重新被随机分成78份，开始下一轮的训练。这个训练模式的好处是既在每次训练时使用了平衡的正负样本，又利用上了所有的样本，优于直接的升采样和降采样方法。

对于二级结构的预测问题，我们同样先搭建了多层感知机和卷积神经网络模型。模型在正、负样本上的预测准确度分别如表2.2中所示。考虑到二维结构的复杂性，我们需要使用更深的人工网络来提取特征。为此，在标准的CNN模型之上，我们引入了残差学习（residual learning）框架。Kaiming He等研究人员提出残差学

习框架是为了解决神经网络在层次加深的过程中所遇到的训练困难，并使预测精度能够受益于网络深度的增加。残差学习的核心是使用了残差学习单元。我们所

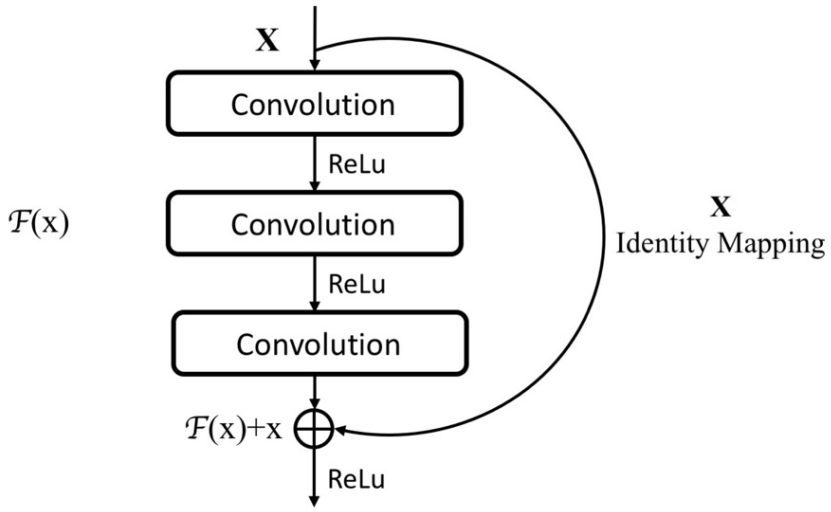

图2.8 残差学习模块

使用的残差学习单元的示意图如图所示。从图中可以看出，相比于传统的单一路线卷积神经网络，我们在残差学习单元最后一步非线形激活前将当前输出和残差学习单元的原始输入进行了一次加法运算，其目的在直观上可以理解为让网络将后期的抽象的特征和前期特征相结合，使模型得到更多的信息而提高预测精度。我们在网络中使用了14个残差学习单元，加上第一层的二维卷积，DeepFold的二级结构预测模型共使用了43个卷积层，包含了五百万个可训练权值。在残差网络的最后一个隐藏层，我们不再按之前网络中那样使用最大值池化，而使用了全局平均池化（Global Average Pooling）的方法，人为得减少最后一层预测器的神经元连接数目，防止过拟合。

表2.2 不同神经网络在二维验证集上的表现

	Positives Accuracy	Negatives Accuracy	Over-all Accuracy	Weighted Accuracy	$P(S^+\|T^+)$
MLP	0.75	0.81	0.81	0.78	0.06
CNN	**0.89**	0.70	0.70	0.80	0.04
Residual CNN	0.72	**0.98**	**0.98**	**0.85**	**0.31**

如何从二维结构的预测结果中选出合适的网络也是一个较困难的问题。我们讨论过，在样本不均衡的情况下考虑全局准确度没有任何意义。为此我们考虑了两种度量方法，一是直接计算正样本和负样本预测准确度的均值；二是计算一个正样本上预测准确度的贝叶斯后验条件概率，即在网络判定样本属于正样本的情况下样本确实为正样本的概率，用公式表示为：

$$P(S^+|\,T^+)=\frac{P(T^+|S^+)P(S^+)}{P(T^+|S^+)P(S^+)+P(T^+|S^-)P(S^-)}=\frac{P(T^+|S^+)\frac{1}{79}}{P(T^+|S^+)\frac{1}{79}+P(T^+|S^-)\frac{78}{79}} \qquad (2\text{-}5)$$

其中T^+表示网络将样本判定为正样本，而S^+表示样本为正样本。几种模型的计算结果如表2.2所示。由于从整体准确度、加权平均准确度（权值为样本比的倒数）和$P(S^+|\,T^+)$三种度量方式看来残差学习网络都有最好的表现，所以我们使用了这种网络架构，并以最大的$P(S^+|\,T^+)$作为最终模型参数选取的标准。将训练好的一维模型和二维模型整合起来，我们得到了最终的基于序列的RNA二级结构预测结果。DeepFold的灵敏度和PPV分别为48.26和42.14，而Fold软件的的灵敏度和PPV分别为55.91和49.92。从结果中可以看出，DeepFold预测的灵敏度和阳性预测值和主流结构预测软件RNAstructure-Fold的结果相比还是有一定的差距。不过考虑到一维上DeepFold的预测准确度已经超过了RNAstructure-Fold，再加上二维结构模型有很多超参可以优化，我们有信心在接下来的研究中能够让模型达到和主流结构预测软件同一水平的预测表现。

2.6 训练更为实用的预测模型

对于一个机器学习算法的测试，我们要严格地控制训练集和测试集的相似性，以验证模型的鲁棒性和泛化能力。但对于一个实用的算法工具来说，只要达到好的预测效果就算成功。为此，我们用所有收集到的RNA进行训练和测试，得到了一个新的模型。虽然没有控制相似度，但由于现实中多数新发现的RNA都和以往出现的RNA有较高的同源性，所以反而可以利用序列的相似度来提升预测效果。在我们收集到的3000余条RNA中，我们先随机挑出了750条作为测试集，300条作为验证集，并将剩下的两千余条RNA作为训练集训练DeepFold的一维和二维结构预测模型。最终的二级结构预测结果如图2.9所示。可以看出，全数据训练的DeepFold结构预测模型的灵敏度和阳性预测值已经达到了90%的水平，大大超越

了现有的基于最小自由能模型的结构预测算法，接近序列比对分析法的准确度。在RNA的分种类测试中也取得了不错的结果。不控制序列相似性的DeepFold与基于序列比对分析的预测方法有一定的相似性，即都利用了序列的相似性。而DeepFold的优势在于这个算法是完全自动的，不需要人力的投入，并且在没有大量同源序列作为预测基础的情况下也能进行一定程度上的结构预测。

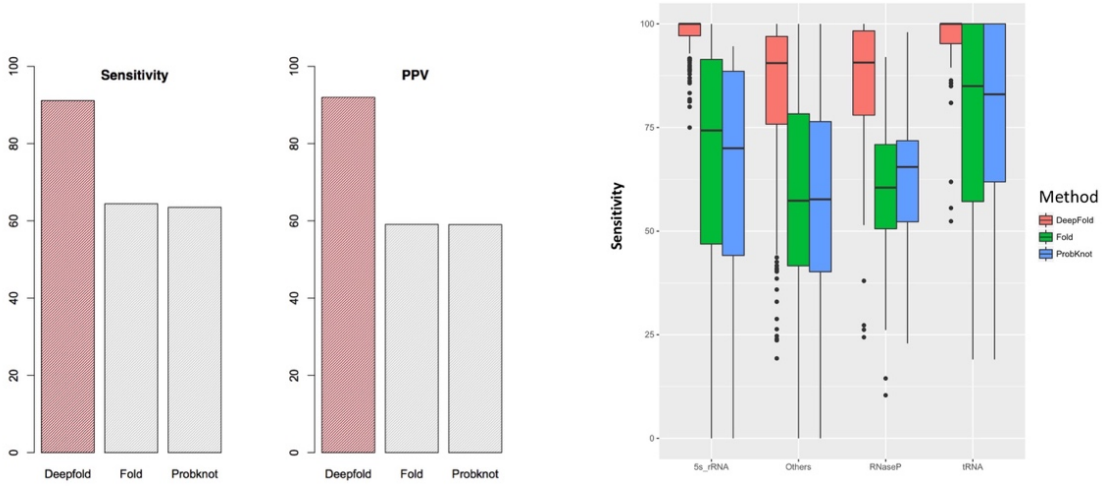

图2.9 使用全数据集模型的二级结构预测结果

第 3 章　结论

我们设计和训练了基于深度神经网络这一机器学习方法的RNA二级结构预测模型——DeepFold。本课题的创新点是设计了新的序列编码方式，以及将RNA结构预测问题转变成二分类问题。而开发基于深度学习算法的最主要目的是替代传统的基于能量模型和动态规划算法的预测工具。在控制序列相似性小于60%的条件下，DeepFold的一维结构的预测准确度超越了和RNAstructure、RNAfold等主流的基于能量模型的算法的预测能力。虽然二维结构的预测准确度与这些工具还有一定的差距，不过由于网络的超参还有很大的优化空间，所以在不久之后DeepFold应该能达到和主流以能量模型为基础的算法相近的预测准确度。在不控制序列相似性的条件下，我们的算法准确度达到了0.9，接近了序列比对分析法的准确度水平，并且我们工具的好处在于：（1）不需要大量的同源序列来提供碱基进化上是否保守的信息；（2）完全自动的预测，不需要人力的投入，甚至在预测时不需要模型的使用者进行参数的调整。

同样，我们的模型也有几个明显的缺点，其中包括：（1）由于我们使用的卷积神经网络模型的输入是固定长度的，所以我们目前还不能预测长度大于801nt的RNA序列。（2）DeepFold算法对于一条输入的RNA只能给定一个预测的结构，所以我们的模型不能像传统基于能量模型的算法那样预测不同温度状态下的RNA结构。当然，这两个问题都是有解决的方法的，现在不能解决是因为数据的缺失。如果有更多的长RNA的结构被解析出来，那我们就可以扩大序列编码中所使用的容器的大小，使更长的RNA也能被考虑进来。对于第二个问题，只要有足够多的同一条RNA在不同温度状态下的结构数据，我们就可以在输入中加上一个维度的特征来指定温度状态，并让网络从中学习不同温度下的折叠规律。所以说，随着已知二级结构RNA序列数目的不断增加，DeepFold模型将获得越来越多的训练样本，预测将变得越来越准确。这一种基于“学习”的思路在任何时候都不会过时。

除了已知精确结构的RNA外，对于PARS，DMS-seq和icSHAPE等高通量RNA结构探测技术，我们依然可以使用相同的DeepFold架构来预测实验的结果。只需要在训练时换上不同的数据集，就可以得到针对某种特定实验技术的结果预测模型。只要能收集到足够的标记好的训练样本，我们二维的序列编码方式和网络架构也同样可以用于RNA-RNA、RNA-DNA相互作用的预测。

在DeepFold网路的训练中同样有许多问题值得讨论。从网络的原理来说，LSTM应该比卷积神经网络更适合RNA结构预测这一类分析上下文关联的序列问题，但在我们的尝试中发现其效果并不好。其原因可能是：首先，对于我们所使用的长度为801的序列输入，LSTM的训练时间极长，并且网络的收敛性也不好，所以我们得到的可能是模型的不充分训练的结果。第二种可能是，在RNA二级结构决定的各种因素中，有局部性特征的序列motif很有可能起了很大的作用，而卷积神经网络的优势就在于探测这些motif，并能整合motif的位置分布信息得到更高维的结构信息用于之后预测器的分类。虽然我们尝试的CNN与LSTM结合的模型并没有取得比CNN本身更高的预测效率，但是CNN与LSTM结合的网络很有可能即能提取motif信息，又考虑远距离配对和序列上下文关系，提升模型的准确性。

我们在前文中谈到了课题中所遇到的最大的困难，那就是已知结构的RNA序列量太稀缺。由于通过实验精确测定RNA二级结构的成本很高，今后一段时间里能够用于训练的RNA序列依然会处于一个非常稀少的阶段。但是我们已知的生物体中的RNA序列却有很多。如何利用大量的没有已知结构的RNA序列进行半监督学习（semi-supervised learning）而不仅是监督学习也很有可能成为未来RNA结构预测算法准确度大幅提升的突破口。

插图索引

表格索引

参考文献

[1] Kozak, Marilyn. "Regulation of translation via mRNA structure in prokaryotes and eukaryotes." Gene 361 (2005): 13-37.

[2] EMBOSS: the European Molecular Biology Open Software Suite. (2000 June) Trends in genetics : TIG 16 (6) :276-7

[3] Martin, Kelsey C., and Anne Ephrussi. "mRNA localization: gene expression in the spatial dimension." Cell 136.4 (2009): 719-730.

[4] Fedor, Martha J., and James R. Williamson. "The catalytic diversity of RNAs." Nature Reviews Molecular Cell Biology 6.5 (2005): 399-412.

[5] Warf, M. Bryan, and J. Andrew Berglund. "Role of RNA structure in regulating pre-mRNA splicing." Trends in biochemical sciences 35.3 (2010): 169-178.

[6] Nudler, E, and A. S. Mironov. "The riboswitch control of bacterial metabolism." Trends in Biochemical Sciences 29.1(2004):11-17.

[7] Wang, S., et al. "Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. " Plos Computational Biology 13.1(2016):e1005324.

[8] Hochreiter, S. & Schmidhuber, J. Long short-term memory. Neural Comput. 9, 1735–1780 (1997).

[9] Rouskin, Silvi, et al. "Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo." Nature 505.7485(2014):701-5.

[10] Cordero, Pablo, et al. "Quantitative dimethyl sulfate mapping for automated RNA secondary structure inference. " Biochemistry 51.36(2012):7037.

[11] Ding, Y., et al. "In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. " Nature 505.7485(2014):696.

[12] Kertesz, Michael, et al. "Genome-wide Measurement of RNA Secondary Structure in Yeast." Nature 467.7311(2010):103.

[13] Spitale, Robert C., et al. "Structural imprints in vivo decode RNA regulatory mechanisms." Nature 527.7577(2015):486-90.

[14] Lu Z, et al. "RNA Duplex Map in Living Cells Reveals Higher-Order Transcriptome Structure." Cell 165.5(2016):1267.

[15] Ramani, V, R. Qiu, and J. Shendure. "High-throughput determination of RNA structure by proximity ligation." Nature Biotechnology 33.9(2015):980-4.

[16] Zuker, M., D. H. Mathews, and D. H. Turner. Algorithms and Thermodynamics for RNA Secondary Structure Prediction: A Practical Guide. RNA Biochemistry and Biotechnology. Springer Netherlands, 1999:11-43.

[17] Tafer, Hakim, et al. "ViennaRNA Package 2.0." Algorithms for Molecular Biology 6.1(2011):: 26.

[18] Reuter, J. S., and D. H. Mathews. "RNAstructure: software for RNA secondary structure prediction and analysis." BMC Bioinformatics 11.1(2010):129.

[19] Mathews, D. H. "Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. " RNA 10.8(2004):1178.

[20] Lu, Z. J., J. W. Gloor, and D. H. Mathews. "Improved RNA secondary structure prediction by maximizing expected pair accuracy. " Rna-a Publication of the Rna Society 15.10(2009):1805-13.

[21] Ding, Y., and C. E. Lawrence. "A statistical sampling algorithm for RNA secondary structure prediction." Nucleic Acids Research 31.24(2003):7280-301.

[22] Zhou, J., and O. G. Troyanskaya. "Predicting effects of noncoding variants with deep learning-based sequence model." Nature Methods 12.10(2015):931.

[23] Do, Chuong B, D. A. Woods, and S. Batzoglou. "CONTRAfold: RNA secondary structure prediction without physics-based models." Bioinformatics 22.14(2006):90-8.

[24] Yao, Zizhen, Z. Weinberg, and W. L. Ruzzo. "CMfinder—a covariance model based RNA motif finding algorithm." Bioinformatics 22.4(2006):445.

[25] Alipanahi, B, et al. "Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning." Nature Biotechnology 33.8(2015):831-8.

[26] Chen, J. L., and C. W. Greider. "Functional analysis of the pseudoknot structure in human telomerase RNA. " Proceedings of the National Academy of Sciences of the United States of America 102.23(2005):8077-9.

[27] Lyngsø, R. B., and C. N. Pedersen. "RNA pseudoknot prediction in energy-based models." Journal of Computational Biology A Journal of Computational Molecular Cell Biology 7.3-4(2000):409.

[28] Xiong, Hui Y., et al. "The human splicing code reveals new insights into the genetic determinants of disease." Science 347.6218(2015):1254806.

[29] He, Kaiming, et al. "Deep Residual Learning for Image Recognition." (2015):770-778.

[30] Ioffe, Sergey, and C. Szegedy. "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift." Computer Science (2015).

[31] Seetin, M. G., and D. H. Mathews. RNA Structure Prediction: An Overview of Methods. Bacterial Regulatory RNA. Humana Press, 2012:99-122.

[32] Ponti, Riccardo Delli, et al. "A high-throughput approach to profile RNA structure." Nucleic Acids Research (2016).

[33] Ding, F., et al. "Three-Dimensional RNA Structure Refinement by Hydroxyl Radical Probing." Nature Methods 9.6(2012):603.

[34] Tullius, T. D., and J. A. Greenbaum. "Mapping nucleic acid structure by hydroxyl radical cleavage. " Current Opinion in Chemical Biology 9.2(2005):127-134.

[35] Krizhevsky, Alex, I. Sutskever, and G. E. Hinton. "ImageNet classification with deep convolutional neural networks." International Conference on Neural Information Processing Systems Curran Associates Inc. 2012:1097-1105.

[36] Glorot, X., Bordes, A. & Bengio. Y. Deep sparse rectifier neural networks. In Proc. 14th International Conference on Artificial Intelligence and Statistics 315–323 (2011).

[37] Collobert, Ronan, et al. "Natural Language Processing (Almost) from Scratch." Journal of Machine Learning Research 12.1(2011):2493-2537.

[38] Mnih, Volodymyr, et al. "Human-level control through deep reinforcement learning." Nature 518.7540 (2015): 529-533.

致谢

感谢鲁志教授在过去的两年多时间内对我的悉心培养和指导，以及在课题上给予我的支持。感谢鲁志教授实验室的史斌斌，胡博钦和李洋对我的课题和帮助。感谢田原、叶丞中和曲日浩同学对我毕业设计课题给出的许多的启发性的建议。感谢清华学堂生命科学实验班的指导老师和同学们的引导和鼓励。最后，感谢清华大学生物科学项目四年来对我的培养。

声明

本人郑重声明：所呈交的学位论文，是本人在导师指导下，独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人享有著作权的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。

签 名：_____ 日 期：_____

附录A 外文资料的调研阅读报告

An active role of RNA in nearly every aspects of cellular processes and gene regulation proves the centrality of this unique information molecule[1]. And its secondary structure, which refers to the canonical base pairing (AU, GC, and GU) pattern of an RNA sequence, is fundamental to its diverse functions in splicing, translation, localization and catalysis[2-5]. For example, the combination of non-coding RNA and RNA binding proteins (RBP) contribute a lot to the post-transcription regulation of mRNA, while the specificity of such RNA-protein combination was shown to be determined by not only the primary sequences, but also the secondary structures[6]. However, comparing to the three-dimensional structures of proteins, the complexity and flexibility of high-order RNA structures make it difficult to solve the accurate spacial configuration of RNA molecules. Besides, the secondary structures of RNAs are already sufficient for scientist to understand many cellular process involving RNA-protein and RNA-RNA interactions. So the investigation of RNA secondary structures is currently the central topic of RNA structure biology.

For decades, the secondary structures of RNA molecules were determined by experimental methods including X-ray diffraction and chemical probing. X-ray crystallography can be used to decide to structure of RNAs with extremely high resolution, but such experiment is expensive and also difficult to accomplish[7]. Chemical probing methods utilizing DMS molecule[8-10] and RNase[11] are also common structure determination techniques. With the development of sequencing technology, those low-throughput chemical probing methods are now combined with RNA-seq to create high-throughput secondary structure determination methods including PARS[11], DMS-seq[9], icSHAPE[12] and PARIS[13]. However, those experiments still suffer from considerable sequencing expense and low signal-to-noise ratio. In the meanwhile, some low-abundance RNAs cannot even be detected by those techniques. Thus, an accurate and efficient RNA secondary structure prediction tool is obvious of enormous importance for scientist to know their interested structures easily and help them interrogate the underlying mechanism of RNA-involved cellular activities.

Currently dominant structure prediction approaches are mainly based upon the concept of minimizing free energy and dynamic programming algorithms, while the most widely used tools are RNAfold[14, 15] and RNAstructure[16]. Others are based upon partition functions[16] that predict the base pairing probability in different configurations and calculate the maximum expected accuracy[17] . Besides, stochastic context-free grammar (SCFG) algorithms are also implemented by scientist to conduct RNA structure comparison and analysis the conservation of RNA molecules from a evolutionary perspective[18-20].

However, those widely used methods still fail to show a satisfying predictive power on either long-distance pairs or pseudoknot-containing structures (a pseudoknot is a featured structure where the nucleotides in a loop pair with other bases outside its own stem-loop region). On the one hand, energy-based algorithms largely rely on experimentally determined thermodynamic parameters, where a tiny error would result in a totally different structure when the sequence length increases. On the other hand, due to the context-sensitivity nature, the problem of predicting RNA structures with pseudoknots (e.g. RNase P and telomerase) has been proved to be an extremely hard problem (NP-hard in the language of computational complexity) which would enervate approaches based on dynamic programming[21].

So in this project, the two main scientific problems we want to answer here are that (i) can we predict RNA secondary structures directly from RNA sequences without using any energy assumptions or experimentally determined thermodynamic parameters, and (ii) can we predict pseudoknots within reasonable time and computational resource consumption? The solving of these two scientific problems would make a large progress in the field of RNA structure prediction. Here we believe that a machine learning technology called deep neural networks would help in surmounting those obstacles and finally show us a road to the answers.

Machine learning technologies have already powered many aspects of human life, from website recommendation to spam email filtering. However, because conventional machine learning techniques is limited in handling data in their raw form, building a conventional pattern-recognition or machine-learning model requires considerable effort to design a feature extractor which can transform the natural data into a suitable representation from which the system can process[22]. Fortunately, the emergence of a

new representation-learning method called deep neural networks (deep learning) now allows machine to discover the proper representation for detection and classification from raw data automatically. Apart from image recognition[23], speech recognition[24], predicting the activity of potential drug molecule[25] and reconstruction brain circuits[26], deep learning-based methods have achieved unprecedented performance on biological sequence-based prediction tasks including predicting RNA-protein interaction[27] and the effects of non-coding variants on RNA[28]. In those two project, the researchers use pure sequence input and convolutional neural network architectures, where the convolutional filters can be interpreted into sequence motif detectors[27]. We will introduce such methodology into our new computational model and designed our predictor to learn base pairing features exclusively from RNA sequences and corresponding known structures, without pre-setting any energetics assumption.

调研阅读报告的参考文献

[1] Sharp, Phillip A. "The centrality of RNA." Cell 136.4 (2009): 577-580.

[2] Warf, M. Bryan, and J. Andrew Berglund. "Role of RNA structure in regulating pre-mRNA splicing." Trends in biochemical sciences 35.3 (2010): 169-178.

[3] Kozak, Marilyn. "Regulation of translation via mRNA structure in prokaryotes and eukaryotes." Gene 361 (2005): 13-37.

[4] Martin, Kelsey C., and Anne Ephrussi. "mRNA localization: gene expression in the spatial dimension." Cell 136.4 (2009): 719-730.

[5] Fedor, Martha J., and James R. Williamson. "The catalytic diversity of RNAs." Nature Reviews Molecular Cell Biology 6.5 (2005): 399-412.

[6] Lunde BM, Moore C, Varani G. RNA-binding proteins: modular design for efficient function. Nature Reviews Molecular Cell Biology. 2007;8:479-90.

[7] Ke, Ailong, and Jennifer A. Doudna. "Crystallization of RNA and RNA–protein complexes." Methods 34.3 (2004): 408-414.

[8] Rouskin S, Zubradt M, Washietl S, Kellis M, Weissman JS. Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. Nature. 2014;505:701-5.

[9] Ding Y, Tang Y, Kwok CK, Zhang Y, Bevilacqua PC, Assmann SM. In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. Nature. 2014;505:696-700.

[10] Cordero P, Kladwang W, VanLang CC, Das R. Quantitative Dimethyl Sulfate Mapping for

Automated RNA Secondary Structure Inference. Biochemistry. 2012;51:7037-9.

[11] Kertesz M, Wan Y, Mazor E, Rinn JL, Nutter RC, Chang HY, et al. Genome-wide measurement of RNA secondary structure in yeast. Nature. 2010;467:103-7.

[12] Spitale RC, Flynn RA, Zhang QC, Crisalli P, Lee B, Jung J-W, et al. Structural imprints in vivo decode RNA regulatory mechanisms. Nature. 2015;519:486-90.

[13] Lu Z, Zhang Qiangfeng C, Lee B, Flynn Ryan A, Smith Martin A, Robinson James T, et al. RNA Duplex Map in Living Cells Reveals Higher-Order Transcriptome Structure. Cell. 2016;165:1267-79.

[14] Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M, Schuster P. Fast folding and comparison of RNA secondary structures. Monatshefte für Chemie / Chemical Monthly. 1994;125:167-88.

[15] Lorenz R, Bernhart SH, Höner zu Siederdissen C, Tafer H, Flamm C, Stadler PF, et al. ViennaRNA Package 2.0. Algorithms for Molecular Biology. 2011;6(1):26.

[16] Reuter JS, Mathews DH. RNAstructure: software for RNA secondary structure prediction and analysis. BMC Bioinformatics. 2010;11:129.

[17] Mathews DH. Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. RNA. 2004;10:1178-90.

[18] Yao Z, Weinberg Z, Ruzzo WL. CMfinder—a covariance model based RNA motif finding algorithm. Bioinformatics. 2006;22:445-52.

[19] Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics. 2013;29:2933-5.

[20] Nawrocki EP, Kolbe DL, Eddy SR. Infernal 1.0: inference of RNA alignments. Bioinformatics. 2009;25:1335-7.

[21] Lyngsø, Rune B. "Complexity of pseudoknot prediction in simple models." Automata, Languages and Programming. Springer Berlin Heidelberg, 2004. 919-931.

[22] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.

[23] Krizhevsky, A., Sutskever, I. & Hinton, G. ImageNet classification with deep convolutional neural networks. In Proc. Advances in Neural Information Processing Systems 25 1090–1098 (2012).

[24] Mikolov, T., Deoras, A., Povey, D., Burget, L. & Cernocky, J. Strategies for training large scale neural network language models. In Proc. Automatic Speech Recognition and Understanding 196–201 (2011).

[25] Ma, J., Sheridan, R. P., Liaw, A., Dahl, G. E. & Svetnik, V. Deep neural nets as a method for quantitative structure-activity relationships. J. Chem. Inf. Model. 55, 263–274 (2015).

[26] Helmstaedter, M. et al. Connectomic reconstruction of the inner plexiform layer in the mouse retina. Nature 500, 168–174 (2013).

[27] Alipanahi, Babak, et al. "Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning." Nature biotechnology (2015).

[28] Zhou, Jian, and Olga G. Troyanskaya. "Predicting effects of noncoding variants with deep learning-based sequence model." Nature methods 12.10 (2015): 931-934.