**非编码RNA二级结构预测与挖掘方法**

钟伟东 1140340116

**摘要**：非编码RNA除了已知的rRNA，tRNA，snRNA，snoRNA和microRNA等，已经发现与生命密不可分。还有大量未知的RNA序列对生命的作用，不能一一求解。利用飞速进步的计算能力，挖掘预测可能的RNA功能，对生物研究起重要的引导作用。

**关键词**：ncRNA，二级结构，机器学习，分类不平衡

**一，概论**

所谓编码RNA可以简单的理解为mRNA，就是用来翻译形成蛋白质作用来进行基因表达的，在真核细胞中这类能编码蛋白的基因仅占10%左右。非编码RNA（ncRNA，non-codingRNA）则是指例如熟知的tRNA，rRNA等一系列已知功能的或者未知功能的RNA，这些不会进行翻译编码形成蛋白质的。

ncRNA有已知功能的许多例如rRNA，tRNA，snRNA，snoRNA和microRNA。这些已经发现与生命有着密不可分的关系。其余还有大量的ncRNA，有待进一步挖掘功能。

学化学时经常听到一句话，结构决定功能。到了大学生物化学不分家，更是有了深刻的体会，故研究其结构对研究其功能有这重要的意义。RNA的一级结构是核糖核酸链，用含氮碱基序列来标注，例如AUCG等。一级结构决定更高级的结构。RNA的二级结构的信息大多隐藏在一级结构中，这给算法进行预测提供了理论基础。而真正发挥功能的是在RNA的三级结构，预测三级结构的意义不言而喻。而为三级结构预测提供信息的就是二级结构，故做好二级结构的预测是非常关键的。一级结构由测序得出，不进行讨论。

目前已经获得大量的RNA序列信息，但是用实验的方法来确定生物分子的三维空间结构花费高、难度大，而且并不是都行之有效。导致了已知的序列知识与结构知识之间形成了巨大的差距。因此就有必要向计算机寻求帮助，通过给出一级结构及部分空间结构信息，来预测其结构。低成本且具有一定可信度，可作为进一步实验参考依据。

**二，RNA二级结构信息**

RNA的二级结构主要是指RNA链中碱基的互补配对的情况。部分逆向配对的子串形成发卡环，其中配对的部分叫做茎区，没有配对的地方叫做环。有些短序列可能进行逆向配对的称为假结。

对这些特征进行描述表示的方法有很多，如半圆表示、圆形表示、点阵表示等。但不同的结构信息可能存储在茎区的碱基中，故不能丢失配对的茎区碱基信息。可以用“质心”来衡量特征。

**三，部分算法模型与理解**

1，最小自由能模型

当没有任何先验知识，只给定了RNA的一级序列，预测RNA的二级结构一般采用小自由能模型。RNA结构一般满足热力学规则，结构所具能量较低，才稳定。就可以利用一级结构信息，用算法与自由能的计算搜索能量最低的二级结构。

但穷举所有可能的结构对于现在的计算机计算能力是不可能的。为了在合理的时间内获得结果，搜索最小自由能结构时会用到最近邻假设，并结合动态规划算法降低搜索复杂度。在最近邻假设中，若两个碱基互相配对，它们之间的序列会自行折叠成自由能最低的结构，不再受外部序列的影响。

早期算法只考虑，碱基配对使自由能减少，时间复杂度较低。后来加逐步加入对碱基对之间的自由能，环区的自由能的考虑，还有对假结的预测。时间空间复杂度逐渐升高。

但是实验证明，RNA实际结构往往不是自由能小的二级结构。迄今为止，还没有找到自由能与结构的绝对的计算规律。

2，同源对比模型

在生物同源分子中，结构保守性一般大于序列的保守性。在RNA分子中，这一点体现得犹为明显，比如绝大多数tRNA分子的二级结构都是三叶草型结构，三级结构呈倒L型状，而它们的一级序列却存在部分差异。

也就是说，RNA的一级结构可能截然不同，但是结构却相近。因此，通过从同源序列中寻找共突变的碱基对，我们可以进行结构预测。相比于热力学模型，同源比对方法的一个优势是直接反映出RNA在细胞中的结构状态，但需要额外提供同源序列作为输入。

同源比对方法可以进一步划分为人为指导和自动比对。人为指导获得的结构准确率极高，在核糖体ITS(internaltranscribedspacers)结构的预测中，所预测的碱基对有95%的置信。自动比对预测包括先比对后折叠、先折叠后比对和边折叠边比对。在RNAalifold中，序列比对直接获得的多重序列被看成一个碱基位有多种赋值的RNA序列，在碱基形成配对时，所有的赋值都参与自由能贡献。利用修改后的自由能规则，仍然使用最小自由能算法，可以有效地对同源序列进行结构预测，这种先比对后折叠的算法速度快，近似等同于折叠一条序列所需的时间。TurboFold则是边折叠边比对，通过迭代，利用同源序列不断修正碱基配对概率。M条同源序列中的每条序列在每次迭代时都被用配分函数计算了一次碱基配对概率，其他序列的配对概率会对当前序列配分函数的计算进行修正。这一方法的优势是并不要求每条同源序列

3，统计学习模型

现在，已知结构的RNA越来越多。RNASTRAND收集了可信的4666个RNA序列和二级结构，这些二级结构来自晶体衍射、磁共振或RNA同源序列比对[17]。统计学习模型从已知的结构进行学习，从而进行结构预测。统计学习模型在预测准确率上达到或超过最小自由能法，但统计学习模型不可避免会对训练数据集过学习，即对训练数据集和相似数据集的预测效果超过普遍预测效果，因此，普遍情况下统计学习模型的准确率仍待确定。利用训练数据集帮助预测主要分为两种思路。第一，假设RNA折叠仍旧服从热力学规律，通过数据集对热力学参数进行重新估计。优化后的两组热力学参数的预测性能在测试集上均超过了优化前的热力学参数。第二，不再假设RNA折叠服从热力学规律，而是认为真实结构来自概率模型的一个采样，通过最大化似然函数估计模型参数。在最近邻假设下，语言学中的随机上下文无关语法(SCFG)非常适合作为描述RNA结构的概率模型。在CONTRAFold这一工作中，使用了类似于SCFG的语言学模型，预测性能在其  
测试集上也超过了最小自由能方法

**四，正负样例不平衡的分类方法**

在许多机器学习问题上，都会存在正负样例不平衡的情况。在有些时候，可以通过多花一些代价获取更大样例。但对于部分问题，是不存在更多的样例信息。

在生物信息学研究中，存在大量这样类别不平衡分类问题。由于大多数分类问题中的正例来自于实验验证，而反例通常不需要实验验证，因此获取反例的成本低、正例的成本高，从而在训练集中通常出现反例远远多于正例的情况。通常使用单分类的方法，但是会失去部分信息，在部分情况下效果不好。

早期常采用随机采样法，对较少的正例进行随机过采样，对较多的负例进行随机降采样。通常降采样法效果较好，过采样容易造成过度拟合。后有人进行人工过采样，也就是对正例进行人工添加，可能会引入噪声。

除了采样策略外，还有一些其它的策略但都未能很好的解决问题。

论文中提出一种基于投票机制的集成学习方法，将反例等分N份，每份个数与正例相同，形成N组数据。对每组数据采取不同的机器学习分类方法，最后对分类的结构进行统计，以权值高的为准。但是各分类器没有进行参数调优，可能效果不是很好。所以要进行下一步，基于重复训练错分样本的优化策略，就是利用已经错误的信息，再来训练，使最后预测的结果愈好（还没能理解这个）。

**参考文献**

[1]. 邹权 基于二级结构的非编码RNA挖掘方法研究[学位论文]博士 2009

[2]. 李鹏. 王晓龙. 刘远超. 王宝勋. LI Peng. WANG Xiao-long. LIU Yuan-chao. WANG Bao-xun 一种基于混合策略的失衡数据集分类方法[期刊论文]-电子学报 2007(11)

[3]. 邹权. 郭茂祖. 刘扬. 王峻. Zou Quan. Guo Maozu. Liu Yang. Wang Jun 类别不平衡的分类方法及在生物信息学中的应用[期刊论文]-计算机研究与发展 2010(8)

[4]. 尤明聪 基于高通量测序的非编码RNA挖掘、分析和功能注释平台mirTools2.0的开发及应用[学位论文]硕士 2013

[5]. 郭杏莉 基于网络模型的基因相关预测问题算法研究[学位论文]博士 2013

[6]. 杨建华 基于新一代测序数据的非编码RNA基因的系统挖掘[学位论文]博士后 2010

[7]. 刘永轩 基于MRF的长非编码RNA功能预测方法[学位论文]硕士 2013

[8]. 张浩文. 杨禹丞 非编码RNA的生物信息学研究方法：RNA结构预测及其应用 [J].生命科学,2014(3)26.