分 类 号 学号 511310200678216

学校代码10487 密级

****

**博士学位论文**

**核酸分子三级结构预测研究**

|  |  |  |
| --- | --- | --- |
| 学位申请人 | ： | 王剑 |
| 学科专业 | ： | 生物物理 |
| 指导教师 | ： | 肖奕 教授 |
| 答辩日期 | ： | 2017年5月30日 |

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Engineering

**Research on Information Management and Communication Technology of Wide Area Protection Coping with  
Power Grid Catastrophe**

|  |  |  |
| --- | --- | --- |
| **Ph. D. Candidate** | **：** | **Wang Jian** |
| **Major** | **：** | **Power System and its Automation** |
| **Supervisor** | **：** | **Prof. Yin Xianggen** |
|  |  | **Prof. You Dahai** |

Huazhong University of Science & Technology

Wuhan 430074，P.R.China

May, 2017

独创性声明

本人声明所呈交的学位论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除文中已经标明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的研究成果。对本文的研究做出贡献的个人和集体，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：

日期： 年 月 日

学位论文版权使用授权书

本学位论文作者完全了解学校有关保留、使用学位论文的规定，即：学校有权保留并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。本人授权华中科技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、缩印或扫描等复制手段保存和汇编本学位论文。

保 密□，在\_\_\_\_\_\_\_\_\_\_\_\_年解密后适用本授权书。

本论文属于

不保密□。

（请在以上方框内打“√”）

学位论文作者签名： 指导教师签名：

日期： 年 月 日 日期： 年 月 日

# 摘 要

摘要。

本文如何如何。

**关键词：**RNA；RNA三级结构预测；RNA三级结构打分；RNA二级结构预测；直接关联分析；DCA

# ABSTRACT

Abstract.

This paper.

**Keywords:** RNA; RNA 3D structure prediction

# 目 录

[摘 要 I](#_Toc475197470)

[ABSTRACT II](#_Toc475197471)

[目 录 III](#_Toc475197472)

[1 绪 论 1](#_Toc475197473)

[1.1 引言 1](#_Toc475197474)

[1.2 本文的研究内容及章节安排 2](#_Toc475197475)

[1.3 本章小结 2](#_Toc475197476)

[2 RNA三级结构预测 3](#_Toc475197477)

[2.1 引言 3](#_Toc475197478)

[2.2 原理与方法 3](#_Toc475197479)

[2.3 结果与分析 9](#_Toc475197480)

[2.4 本章小结 10](#_Toc475197481)

[3 RNA三级结构打分 11](#_Toc475197482)

[3.1 引言 11](#_Toc475197483)

[3.2 保护区域的划分原则及实现方法 11](#_Toc475197484)

[4 DNA三级结构预测 12](#_Toc475197485)

[5 RNA二级结构预测 12](#_Toc475197486)

[6 全文总结与工作展望 13](#_Toc475197487)

[6.1 本文工作总结 13](#_Toc475197488)

[6.2 下一步工作展望 13](#_Toc475197489)

[致 谢 14](#_Toc475197490)

[参考文献G 16](#_Toc475197491)

[附录1 攻读博士学位期间发表的主要论文 17](#_Toc475197492)

[附录2 博士生期间参与的课题研究情况 18](#_Toc475197493)

# 绪 论

## 引言

核酸是细胞最基本物质之一，是由核苷酸聚合而成的生物大分子。核酸可分为两类：脱氧核糖核酸（DNA）和核糖核酸（RNA）。传统上认为，核酸在细胞中的功能主要是存储和传递遗传信息以及参与蛋白质的合成，但是近十几年的研究表明，细胞内存在大量的非编码RNA（non-coding RNA），也就是不编码蛋白质的RNA，它们和蛋白质一样参与细胞内的各种生物学过程[1-5],特别是在各种生物调控过程中扮演着极其重要的角色［6］。例如,核糖开关(riboswitch)可以通过结合代谢产物在转录和翻译水平调控基因的表达[7]。除了我们熟知的转运RNA和核糖体RNA外，典型非编码RNA还有各种小RNA，RNA酶（ribozyme），核糖开关和很多功能未知的长非编码RNA等等。这些非编码RNA的发现使我们对生物进化、生命过程和疾病成因有了更深入和全面的认识[8]。

虽然现在已经确定了越来越多 RNA 分子的功能,但是仍然还有大量未知功能的RNA分子[9,10]。 RNA分子的功能不仅依赖于它的序列还依赖于它的三维空间结构，要深入理解 RNA 的功能需要确定它们的三维空间结构。目前，实验上确定 RNA 分子的序列相对容易,但由于RNA易于降解,实验上测定RNA空间结构比测定蛋白质空间结构更加困难，已测定的单体RNA空间结构目前还不到1000个,而蛋白质的已经有九万个。因此,从理论上预测和构建非编码 RNA 分子的空间结构非常必要。

相比蛋白质三级结构预测，RNA三级结构预测的研究还处于起步阶段，真正意义上的RNA分子三级结构预测方法近几年才出现。目前已有的RNA三级结构预测方法可以分为两类，一类是基于分子力场的从头预测方法，例如基于离散分子动力学的方法iFoldRNA［11，12］和基于统计势的分子动力学模拟方法NAST［13］ ，一类是基于知识（已知结构信息）的片段或模块组装的方法，其中包括通过连续片段组装的方法FARNA［14］和FARFAR［15］, 基于模块组装的方法RNA2D3D［16］，Vfold Model［17］， MC-Sym［18］等。从头预测方法一般都是采用粗粒化模型，准确性随链长下降比较快，因为全原子模型计算量太大，难于实现。基于知识的方法的准确性依赖于目前已知的实验结构的数目或者是模块结构库的可靠性。如果已知二级结构的信息，预测的准确性会大大提高。

系统评估显示［19］，现有的大部分RNA三级结构预测方法对于较小或者拓扑结构较为简单的RNA分子（<50nt）可以给出较高的精度，与天然结构的均方偏差（RMSD）约为4 Å左右；然而对于分子较大或者拓扑结构较为复杂的RNA分子（50-100nt）则很难给出较为精确的近天然态结构，RMSD约为20-25 Å左右；对于链更长、拓扑结构更复杂的（比如假节）RNA分子则很难预测或不能预测。另外多数方法构建成的RNA模型需要手工进一步调整，不能够自动完成预测。最近，这方面有了进一步的发展，比如FARFAR 对长度为20nt左右的RNA的三级结构的预测精度达到2 Å左右［15］。最近我们提出了基于二级结构基本单元拼装的自动化三级结构预测方法3dRNA［20］。3dRNA对小于50nt的RNA的空间结构预测精度为3 Å左右,对50-100nt RNA 分子结构的预测精度在 6 Å 左右。该方法的速度和精度均优于目前主流的结构预测方法，而且是全自动的。但总的来说RNA三级结构预测方法预测精度还有待进一步提高，对更大和拓扑结构更复杂分子难于预测的问题还亟需解决。

基于模块预测RNA三级结构的方法首先将提供的目标RNA的二级结构拆成各种子结构，对于每一个子结构从模板库中找到一个与其匹配的全原子三级结构模板，然后将这些模板依次组装起来就得到了最终的三级结构。目前拟解决的关键问题有：

（1） 如何自动地把任意长度和任意复杂拓扑结构分解成子结构并把对应模板重新组装成整体。

（2） 如何从理论上构建没有实验模板结构的子结构的三维结构模板，否则能够预测的RNA数量会收到极大的限制；

（3） 如何构建合适的RNA三级结构评估函数，是能否从构建的RNA三级结构候选模型中挑选出正确的结构和提高预测精度的关键。

5. 总体方案

(1)非编码RNA折叠机制研究

目前准确预测RNA空间结构的困难之一方面是对RNA分子的折叠机制，水和离子等对RNA分子结构的影响还不很清楚。RNA如何折叠成三级结构是一个亟待解决的重要难题，到目前为止还没有成功模拟RNA分子甚至RNA发卡折叠的例子。原因可能是原来核酸分子力场还不够准确，特别是电荷极化效应没有得到充分考虑，另外也没有考虑胞内的折叠方式和生理环境（包括溶剂水分子和各种金属离子）。本项目希望利用最新改进的分子力场，并考虑胞内的折叠方式，利用分子动力学模拟方法研究非编码RNA分子的折叠过程，对RNA空间结构形成的机制有所了解，为RNA空间结构预测以及精度的提高提供新的思路。

(2)RNA基本模块的自动分解和三维结构构建

基于模块拼接的RNA三级结构预测方法第一步是根据二级结构信息把RNA分解成二级结构基本模块，包括螺旋,发卡环、内环、突环、多分支环和假节环等。由于对长RNA(比如含上千个核苷酸的RNA)基本模块的自动分解实现很困难，而一般手动分解不可能，所以目前的预测方法一般只能处理100个核苷酸的RNA分子。本项目拟发展一套方法来处理任意长度和任意复杂拓扑二级结构RNA分子基本模块自动分解问题。

另一方面，由于目前实验测定的RNA三级结构有限，在预测中经常会遇到没有模板的序列,特别是各种环区，这样整个结构就无法构建，大大限制了3dRNA可能预测的对象。本项目拟对各种基本模块的实验结构特征进行分析，在此基础上发展基于物理的理论方法来自动产生无实验模板的基本模块的三级结构，克服实验结构不足的局限，同时也完善现有的基本模块的结构模板库。

(3)非编码RNA三级结构评价函数构建

在搭建一个具体RNA分子的三级结构时,对每一个基本模块,3dRNA只选取一个最优的 (序列相似性最好、出现概率最大)的三维空间结构作为模板,因此最后只给出一个预测结构。这种搭建方式采用局部优化，速度很快,但缺少整体结构的优化。为了进一步提高预测的精度,需要选取不同的模板来组装整体结构,从而能够构建出最优的整体结构。

对于同一基本模块，选取不同的空间结构模板，将提供大量的RNA空间结构模型。如何从这大量的空间结构模型中挑选出正确的结构仍然是一个难题。相比蛋白质分子结构而言, RNA分子的结构具有较大的柔性,已知的RNA三级结构的数目还很少,使得构建RNA分子空间结构评价函数更加困难。现有的基于经验的RNA评估函数挑选出正确的空间结构准确性还比较低,尤其是包含非标准碱基配对等特殊的RNA 分子。本项目将根据全原子模型,把距离势与扭角势结合，提出新的RNA三级结构评价函数，提高识别正确RNA空间结构的能力。

## 本文的研究内容及章节安排

研究内容。

章节安排。

## 本章小结

本章小结。

# RNA三级结构预测

## 引言

在最近几十年，非编码RNA（ncRNA）被发现在许多生物学过程中都起到非常重要的作用，例如在许多催化和调控过程中都发挥了不可替代的作用。为了行使各种各样的生物学功能，RNA分子需要形成一些特定的三级结构。与蛋白质一样，RNA分子的三级结构对于了解它们的功能非常重要。由于已经解析出来的RNA分子的三级结构非常稀少，越来越多的RNA三级结构预测方法开始出现。

在2008年之前，大多数的预测方法都不是自动的，并且需要人工的干预，例如MANIP [[1](#_ENREF_1)]，BARNACLE [[2](#_ENREF_2)]以及ASSEMBLE [[3](#_ENREF_3)]。在经过几年的发展之后，越来越多的自动的和精确的方法被提出来了。我们也设计了一个自动并且快速和精确的预测RNA三级结构的方法，3dRNA[[4](#_ENREF_4)]。

这些RNA三级结构预测方法可以大致的被分为2类：第一类是静态的，或者基于片段组装的方法进行预测，或者使用同源模建进行预测，代表性的方法是MC-Sym [[5](#_ENREF_5)]，它通过核苷酸周期性模块（NCM）来组装全原子的结构。其它的方法包括基于三向量虚键（three-vector virtual bond）的RNA折叠模型的Vfold [[6](#_ENREF_6), [7](#_ENREF_7)]；基于机器翻译系统的RNAComposer [[8](#_ENREF_8)]；使用粗粒化的螺旋中心的模型来预测RNA三级结构的Enrwin [[9](#_ENREF_9)]；基于核苷酸离散态模型的RNAnbds/SMC method [[10](#_ENREF_10)]；可以包含后翻译信息的基于同源模建的ModeRNA[[11](#_ENREF_11)]；允许对已知的RNA结构中的一段序列进行插入和替换的RLoom [[12](#_ENREF_12)]。

第二类是动态的，或者采样蒙特卡罗模拟，或者采用分子动力学模拟，代表性的方法是FARNA[[13](#_ENREF_13)]/FARFAR[[14](#_ENREF_14)]，它采用基于片段组装的蒙特卡罗模拟来进行预测，在与蛋白质三级结构预测竞赛CASP类似的RNA三级结构预测竞赛RNA Puzzles [[15](#_ENREF_15), [16](#_ENREF_16)]中取得了非常好的成绩。另外一个典型的方法是NAST，它构造了一个粗粒化的基于知识的统计势，然后使用分子动力学模拟来对RNA的三级结构进行预测。其它的方法包括基于DMD[[17](#_ENREF_17)]的iFoldRNA；RNA2D3D[[18](#_ENREF_18)] ；SimRNA[[19](#_ENREF_19)] ；oxRNA[[20](#_ENREF_20)]；可以考虑盐的效应的Shi的方法[[21](#_ENREF_21)]。

Liang and Schlick [[22](#_ENREF_22)]对这些方法的预测性能进行了比较并且发现对于小RNA，大多数的预测方法预测的结构与实验结构的RMSD值都能达到到6Å以内。对于大一点的RNA（50nt到130nt），平均RMSD值大约为20Å。并且，大多数方法都不是全自动的，需要人工的干预。

Sim等人[[23](#_ENREF_23)]对进行核算系统的模建的预测方法进行了总结。他们主要关注与核算预测的四个关键性的方面：分子表示方式，能量函数，自由度以及采样方法。他们发现通过在这四个方面进行合适的选择可以有效地结合它们来促进核酸结构的预测。

3dRNA可以基于RNA的二级结构自动并且快速的预测RNA的三级结构。由于RNA的三级形态很大程度上取决于二级结构和三级接触所组成的约束[[24](#_ENREF_24)]，因此我们可以通过最小二级结构单元（SSE）[[4](#_ENREF_4)]来分解这些约束，然后将这些SSE组成一个整体的结构。在3dRNA中，这些SSE被定义为发卡环（hairpin loop），内环（internal loop），凸环（bulge loop），以及多分支环（multi-branch loop）。

3dRNA有几个比较明显的缺点需要解决。首先，3dRNA每一次只能预测一个结构，这并不总能满足生物学研究的需求。并且，当我们解决了这个问题使得3dRNA可以一次预测大量的结构之后，我们需要一个打分函数来对预测的结构进行评估。第二个问题是3dRNA还无法处理太复杂的结构，例如长链RNA或者包含有太多的分支的多分支环的RNA。因此，我们从这几个方面对3dRNA进行了全方位的改进。新版的3dRNA可以通过（<http://biophy.hust.edu.cn/3dRNA>）进行访问。

## 原理与方法

### 二级结构树的构建与分解

通常，RNA的二级结构可以很方便的用点括号的形式来表示，圆点表示不与任何其它碱基配对的碱基，一对圆括号则表示一对配对的碱基。如果存在假结，则需要用方括号来表示假结中配对的碱基。【图】

在RNA的二级结构比较简单的情况下，这样表示比较方便而且也不会有什么问题。不过在长非编码RNA（lncRNA）中，假结的存在比较常见，一个长非编码RNA中可能会存在好几十个假结。例如在【图】中就存在【n】个假结。只用圆括号以及方括号无法用点括号的形式来表示这个二级结构。所以如果只用方括号来表示假结可能会引起歧义，因此我们引入了更多的符号来表示假结：‘[’， ‘]’， ‘{’， ‘}’， ‘<’， ‘>’， ‘A’， ‘a’， ‘B’， ‘b’, ‘C’, ‘c’, …

通过引入这些符号，我们可以将这个二级结构用点括号的形式来表示出来，如图【图】。

我们将RNA的二级结构用树来表示，树的每一个节点表示一个二级结构元素（SSE，secondary structure element）。SSE包括螺旋，发卡环，内环，凸环以及多分枝环等。基于二级结构把RNA分解成SSE的难点是任意长度和任意复杂拓扑结构RNA的自动化分解。利用计算机中的入栈出栈的思想，步骤如下：

首先，需要提供RNA二级结构的点括号形式，例如

..(((((...(((((...(((((...(((..((.....)))))...))))).....(((((...(((((.....)))))...))))).....)))))...)))))..

然后建立一个符号栈，从左向右开始读取符号并放进栈中，每当遇到一个右括号’)’的右方不是右括号的情况，则暂停读取，将栈中最右方的形如(((((.....)))))的结构进行出栈，从而分离出一个发卡环(…..)和一个螺旋((((()))))，然后从暂停的位置继续读取符号并入栈，如此循环下去即可分离出所有的环和螺旋。

我们对于圆括号，方括号，大括号，尖括号以及db和qp分别构造二级结构树。这样，所有的二级结构树在一起构成了二级结构森林。

为了方便组装三级结构模块，我们可以将分离出的二级结构的基本模块放在一颗树型数据结构中。树的根节点是RNA分子序列的最开头几个核苷酸与最后几个核苷酸形成的环，如果形成的是螺旋，那么根节点就是该螺旋。然后找到与根节点接触的螺旋或者环作为根节点的子节点，依此递归下去得到整棵树。

### 模板库的构建

首先从PDB/NDB数据库中获取所有的RNA，然后对它们进行序列比对以删除掉同源的RNA，并剔除实验精度大于3.5Å的结构。对于剩下的RNA根据其二级结构将三级结构进行拆分，从而得到对应于二级结构基本模块的三级结构模板，将这些模板及其相关信息存进数据库中得到最终的模板库。为了有利于下一步的搜索三级结构模板，这些相关信息应该包括二级结构的序列、点括号形式、长度以及家族信息等。

我们从蛋白质数据库（PDB）中下载了所有的单体RNA，从中剔除了所有的包含转录后修饰核苷酸的RNA，剩余957个单体RNA。由于在某些RNA中，存在一些核苷酸的某些原子的缺失，为了保证程序的运行不会出现错误，我们剔除掉了这些缺失一些原子的核苷酸，从而使所有的核苷酸都是标准的A、U、G、C核苷酸。这样剔除掉核苷酸会造成RNA的断链，但是这不会对预测造成影响，因为我们在后续构建模板库的过程中，不会将存在断链的模板放进模板库中。

为了将这957个RNA根据它们各自的二级结构分解为对应的二级结构模板，我们还需要得到它们的二级结构。在已知三级结构的情况下，RNA的二级结构可以通过分析它们的三级结构中的碱基配对的情况来获得。这里，我们将存在标准的Watson–Crick碱基配对的区域作为螺旋区域，而对于存在非标准的碱基配对的区域和不存在配对的区域，我们都作为环区。通过这样的手段，在组装三级结构的过程中，我们不需要考虑螺旋的柔性，因为我们这里的螺旋都是最为标准的A螺旋，甚至可以不需要使用螺旋的模板，而是直接根据标准螺旋的参数来直接搭建螺旋。这样RNA分子的柔性就被限制在了环区，我们在最后的大规模采样过程中，只需要对环区进行采样，就能有效的搜索RNA分子的构象空间。

我们使用X3DNA程序(Lu & Olson, 2003)来分析RNA分子的碱基配对的情况，从而获得RNA分子的二级结构。

由于同家族的RNA往往具有相似的空间结构以及二级结构，例如tRNA的空间结构都是很相似的类似字母T的形状，核糖开关的空间结构都是很相似的Y形结构，因此为了在接下来组装的过程中选取最优的模板，我们将RNA所属的家族信息也加进了模板库中。

我们使用HD-RNAS数据库(Ray, Halder, Kaypee, & Bhattacharyya, 2012)来找到所有RNA的所属家族信息，HD-RNAS数据库将所有的RNA分为8大类：rRNA、tRNA、mRNA、Riboswitch、Ribozyme、Ribonuclease、SRP以及other类，其中other类是所有其它类别的RNA的集合。我们通过HD-RNAS数据库找到所有RNA单体所属的家族，其中有些RNA单体在该数据库中找不到，我们也将其归为other类。

通过X3DNA程序获得所有957个RNA分子的二级结构以后，通过分解二级结构来生成最终的模板库

### 组装与模板采样

在寻找基本模块的三维结构模板的过程中，对于模板库中的每一个模板都需要提取信息与之进行初步比较，以确定是否可以作为它的模板。初步比较主要看模板的序列长度是否与基本模块的长度一致，二级结构类型是否与基本模块一致（即5nt长度的发卡环必须对应于5nt长度的发卡环；5-6nt的内环，即一条支链的长度为5nt而另一条的长度是6nt，必须对应于5-6nt的内环）。初步挑选完毕后会出现两种情况，第一种是有多种与该二级结构的子结构基本模块匹配的模板，第二种是找不到与之对应的模板。

对于存在多种匹配模板的情况，需要进一步对这些模板进行评估。本项目评估初步设想包含三个部分，第一部分是与模板的序列的相似度，第二部分是与模板的接口的碱基配对是否相同，第三部分是与模板所属的家族是否相同。

对于找不到模板的情况，本项目拟采用距离几何的算法来创造模板。距离几何法（distance geometry）是一种用来进行分子模建和构象搜索的方法，最初是由Crippen和Havel提出的（Distance geometry and molecular conformation. Vol. 74. 1988: Research Studies Press Taunton, UK）。为了应用距离几何法，需要知道每两个原子之间的距离的上界和下界。对于相邻的两个原子之间的距离分布、相隔一个原子的两个原子之间的距离分布可以直接从天然态的RNA中经过统计分析得出；对于相隔两个原子的两个原子之间的距离分布，可以直接计算得出；对于相隔n(n>2)个原子的两个原子之间的距离分布需采用距离几何算法中的三角优化方式计算。根据原子之间距离的上界和下界可以得到上下界矩阵，然后根据距离几何算法得到原子的坐标矩阵，最后利用共轭梯度法优化可得最终的结构模板。

找到所有与二级结构的子结构基本模块对应的模块三维模板以后，需要将这些模块依次组装起来。首先，需要一种合适的算法将两个模块最优的拼接在一起，其次，需要找到一种合理的顺序来组装这些模块。

首先，为了将两个模块最优的拼接在一起，需要事先在每个模块的两端多留出1-2nt的长度，这样两个待拼接的模块多出来的这部分就可以采用Kabsch算法[[25](#_ENREF_25)]最优拼接重叠在一起了，从而完成两个模块的拼接。

其次，根据二级结构分解产生的基本模块树，从根节点开始进行三级结构的组装。组装的过程是一个遍历基本模块树的过程，具体过程如下：首先根据根节点的基本模块找到对应的模板，然后寻找第一个子节点的模板，找到后与根节点进行拼接，然后寻找第一个子节点的第一个子节点的模板，与第一个子节点进行拼接，如此向下递归下去，一直到某个叶节点停止，然后寻找该叶节点的父节点的下一个子节点的模板与其父节点拼接，再寻找该子节点的第一个子节点的模板与它拼接，如此递归下去，直到整棵树被遍历完成，这时整体的三级结构也就被组装好了。

在组装三级结构的过程中，如果对于某一个基本模块，存在多个模板与之匹配，那么对这些模板进行评估排序，从分数最高的往下排，选择最前面的几个模板，所选模板的数量由采样数决定，例如假如采样数是1000，RNA分子中存在4个环，那么在对于每个环，都挑选10000.25 约等于6个模板，然后将这些模板组装，从而产生64个结构，选择1000个结构作为最终的采样结果

### 三级结构结构评估

由于可以进行大规模采样，需要对构造出的每一个结构模型进行评估，以比较所有的结构模型，从而对它们进行聚类并挑选出符合要求的结构。RNA三级结构打分数，目前主要有两种方式，一种是基于物理的评估函数，另外一种是基于知识的评估函数（又称为统计势），具体表述形式是

T为温度（298K）， 为波尔兹曼常数， 观测态中原子i和原子j的频率， 为参考态。

基于物理的评估函数具有明确的物理意义，但不能有效计算构象熵，评估效果不如基于知识的统计势。而基于知识的统计势是从已有的天然态RNA中进行统计分析得到的，包含部分构象熵，评估效果较好，并且计算速度也很快。本项目将发展一种基于知识的统计势，它将同时考虑传统的距离依赖的统计势和新的两面角依赖的统计势，能够比较好地考虑RNA的柔性。这种统计势不仅具有基于知识的统计势的优势，同时又能够在一定程度上与基于物理的统计势相对照，也就是说具有一些基于物理的统计势的特征。

### 训练集的构建

为了评估预测精度，本项目采用RMSD和INF（interaction network fidelity）(Sci Rep. 2:734, 2012)两个参数相结合的方式来比较预测结构与实验结构之间的偏差。计算RMSD值是为了比较RNA之间整体几何构象的偏差，计算INF是为了比较整体的拓扑结构（即氢键网络）之间的偏差

## 结果

### 通过采样来提高预测精度

For simple hairpins and duplexes, 3dRNA-1.0 can give high prediction accuracy. We’ve took two representative RNAs as examples: 28SP and 1I9X. The hairpin 28SP (28 nt) contains a large internal loop. 3dRNA gave the predicted model with the interaction network fidelity (INF) of 0.89 and RMSD of 2.99 Å, while the promoted new version 3dRNA-2.0 increased the INF to 0.93 and decreased the RMSD to 2.05 Å; the duplex 1I9X (26nt) contains two bulge loops and the structure predicted by 3dRNA has an INF of 1.00 and RMSD of 2.73 Å. 3dRNA-2.0 could predict it with INF of 1 and RMSD of 1.89 Å.

In last work, we’ve compared the performance of 3dRNA-1.0 with several methods including iFoldRNA, RNA2D3D and MC-sym in a dataset consisting of 185 RNAs. These RNAs includes duplexes (12–56 nt), hairpins (10–63 nt) and molecules (26–101 nt) with complex topology containing pseudo-knots or junctions. It turns out that 3dRNA-1.0 performs better than the other methods in this dataset. In this work, we’ve tested the prediction performance of 3dRNA-2.0 with respect to 3dRNA-1.0 in the same dataset. Supplementary Table 2 shows that the overall average RMSD of the 185 RNA structures predicted by 3dRNA-2.0 is 3.81 Å which is smaller than 3.99 Å of 3dRNA-1.0. Figure n shows the RMSD distribution of predictions of 185 RNA molecules. All these suggest that there is a slightly promotion in the prediction accuracy from 3dRNA-1.0 to 3dRNA-2.0.

For RNA, RMSD is not enough to reflect the difference between two molecule 3D structures. RMSD is calculated by superposing two molecule structures. It can perfectly designates how far the geometrical coordinates of two molecules differ from each other. But RMSD is not able to reflect the difference of the base interaction network between two RNA molecules. In this sense, INF is preferable. So we further test if 3dRNA-2.0 can recapitulate the whole base interaction network in native state. As we’ve done in last work, we use the 32 representatives of 185 RNA structures to further test the performance of 3dRNA-2.0. To get the 32 representative structures, we clustered the 185 RNA structures into 32 classes with the sequence identity between any two of them being less than 50%. One representative for each class is then selected for comparison. These representatives include simple hairpins such as 1ZIH and complex junctions such as typical L-shaped tRNA with four-way junction.

Supplementary Table 1 shows that the overall average RMSD of predictions of the 32 RNA molecules with 3dRNA-2.0 is 3.91 Å which is also slightly better than 3.95 Å of 3dRNA-1.0, while the INF is 0.87 which is slightly worse than 0.88 of 3dRNA-1.0. Furthermore, the average RMSD decreased to 3.60 Å and the INF increased to 0.89 if we sampled 1000 models. The RMSD and INF is calculated for the center structure of the cluster with lowest energy in all the clusters from the 1000 models. Once we sampled 10000 models, the RMSD decreased further to 3.13 Å and INF increased to 0.90. These result verified the effectiveness of the sampling procedure.

### 减少预测时间

A prominent feature of 3dRNA-2.0 is that it can predict even a large set of long non-coding RNAs in a very short time. We picked out 14 RNA molecules with the length from 10 nt to 168 nt. Figure 4 (a) shows all the time elapsed for the prediction of the 14 RNA molecules. It suggests a weak relationship between the time and the length of RNA. The length longer, the time elapsed more. We also delved into the relationship between the time and the complexity of the 2D topology of RNA. We use the number of loops to represent the complexity of the 2D topology. For example, there are 3 loops in the 2D structure in Figure 2 (a). Figure 5 (b) shows a stronger relationship between the time and the number of loops than that between the time and the length. For all of these RNA molecules, the running time are less than 3 seconds.

To dig into the time consumed for the sampling procedure, we’ve predict from 1, 2, 4, … , to 214 number of models for RNA 1KXK with the length of 70 nt. Figure 5 (c) suggests a perfect linear relationship between the time and the sampling number. It is worth noting that even for 214 (16384) number of predictions, the running time is only 346.6 seconds.

### 捕捉非标准碱基配对

FARFAR can predict accurate RNA 3D structures with low RMSD, while recapitulating non-Watson–Crick base pairs seen in the native structures [[14](#_ENREF_14)]. Here we also tested if 3dRNA-2.0 can recapitulate non-Watson-Crick base pairs. For the convenience of comparison, we select the FARFAR models [[14](#_ENREF_14)] which is composed of 32 RNA motifs as the test set. Supplementary table 3 shows that the average RMSD of the cluster center of the models predicted by 3dRNA-2.0 is 3.81 Å, and the average RMSD of the cluster center of the models predicted by FARFAR is 3.84 Å. The two are almost equal. Nevertheless, the average RMSD of the lowest energy cluster member of the models predicted by 3dRNA-2.0 is 3.83 Å, which is smaller than 4.01 Å by FARFAR. This suggests that 3dRNAscore which is plugged in 3dRNA-2.0 is able to pick out the structure more close to the native. Within the total 92 non-Watson–Crick base pairs, 3dRNA-2.0 can recapitulate 41 from the cluster centers and 45 from the lowest energy cluster members. FARFAR can recapitulate 37 from the cluster centers and 42 from the lowest energy cluster members. This suggests that 3dRNA-2.0 performs better in the ability of recapitulating non-Watson-Crick base pairs.

### 对更复杂的RNA的支持

3dRNA-1.0 is not limited to predict RNA 3D structures with simple topology and small size. It can also give reliable accuracy (5.7 Å on average) for RNA molecules with complex topology and larger size such as pseudo-knot and multi-branch junction. With 3dRNA-1.0, the topology of pseudo-knot couldn’t be too complicated and the number of pseudo-knot or multi-branch junction in one structure should only be one. We’ve solved this problem in 3dRNA-2.0 with the help of the new tree representation of RNA 2D structure and distance geometry. For example, 3ZEX is an rRNA of 4199 nt. Figure 6 (a) shows the extremely complicated 2D structure plotted by VARNA [[26](#_ENREF_26)]. Figure 6 (b) is the native 3D structure of 3ZEX. Figure 6 (c) shows the predicted 3D structure of 3ZEX with 3dRNA-2.0. The total running time is only 8.4 seconds, while the all-atom RMSD is 81.55 Å. This suggests that it need a promotion to predict large RNA with more accuracy.

## 分析

### 假设

In the last work, we illustrated that in different molecules the 3D conformations of the loops of a given type and length (e.g., 4 nt hairpin loop) are close to each other if they have the same sequence. Even their sequences are different the backbone conformations are still similar. In this work, we found that for long loops having same sequence, the conformations of them are close in most cases. But there still existed circumstances that the conformations are far away from each other. Longer the loops, the more the case that conformations of loops having same sequence are far away from each other. Therefore, just like FARNA [[13](#_ENREF_13)], in 3dRNA-2.0 we also took the assumption that the distribution of conformations observed in known RNA structures with same sequence provides a reasonable approximation to the conformations of a specific SSE with the same sequence.

### 采样程序

In most other methods, the sampling procedure is done on the level of atoms or residues. For instance, iFoldRNA utilized a simplified three-bead per nucleotide model of RNA to do replica-exchange DMD simulation. NAST does Monte Carlo simulation with the model that each residue is replaced by one atom. The sampling is done on residue level. For FARFAR, it also does a Monte Carlo simulation on the three residue fragments. All of these methods need a large time to generate a large set of models. For 3dRNA-2.0, the sampling procedure is almost a moment for small RNA. The short time benefits from the sampling on the level of SSEs. The number of residues in one SSE varies from several to dozens. Hence the sampling is very fast. On the other hand, the substitution templates for SSE are all extracted from experimental structures according to the assumption above. So the final structure can recapitulate a lot of non-Waston-Crick base pairs.

### 长RNA的问题

Also 3dRNA-2.0 can generate models for arbitrary long and complex sequence and its corresponding 2D structure, the accuracy is very bad for RNA the size of which reaches thousands of nucleotides long. The bad accuracy comes from two aspects. First, the long sequence accumulates the inaccuracy for short sequence. For locally short sequence fragments, even the precision of the predicted model is very high such as less than 1 Å, the total precision may become very low after a large number of assembly of these high accuracy fragments. That is to say that the inaccuracy accumulates. Second, the given 2D structure is not able to provide so much information about the tertiary interaction in RNA. The tertiary interactions have a great influence on the space arrangement of the SSE templates. Nevertheless, the 2D structure can provide with only the base-pair information. So much tertiary interactions especially in large RNA are omitted. Therefore, to solve the problem of large RNA 3D structure prediction problem, we may move along these two directions.

### RNA Puzzles

We’ve evaluated 3dRNA-1.0 by ourselves in the problem 1-3 of the RNA-Puzzles[[15](#_ENREF_15), [16](#_ENREF_16)]. RNA-Puzzles is a CASP-like contest for blind RNA three-dimensional structure prediction. This time, we’ve took participate in the problem 5 of the RNA-Puzzles. The target is the lariat-capping ribozyme of 188-nt long. It catalyzes a distinct reaction involving formation of a 3-nt 2’, 5’ lariat. It’s a disappointment that the two models we’ve submitted gained a bad results. The RMSD of the two are 32.53 Å and 36.54 Å, respectively. We’ve analyzed the results and found that the large RMSD is due to the large difference between the 2D structure we used and that of the experimental structure. Once we used the experimental structure to predict the structure, the RMSD decreased rapidly to 12.37 Å. Another result also supports the conclusion that the INF taking into account only non-Watson–Crick interactions of our two models rank the third and the fourth, respectively. This suggests that the local structure of our models are verisimilar to the native structure. And it’s because that the decomposition of the total structure into individual SSEs can result in locally accuracy of each SSEs. The main deviation comes from the space arrangement of these SSEs, which largely depends on the 2D structure. Another worth noting result is that the clash score of our two models rank the first and the fourth. And it’s due to the delicate assembly of the SSE templates which are extracted directly from the experimental structures.

## 本章小结

# RNA三级结构优化

## 引言

## 原理与方法

### 优化

在蒙特卡罗模拟的每一步，需要对分子的结构进行采样。我们决定在采样的过程中不破坏分子的二级结构，同时也不对5个长度以下的发卡环进行采样。所以我们在每一步当中随机对处于环上的某一个残基进行位移和旋转或者随机对某一个螺旋上的所有残基进行整体的位移和旋转。这样可以保证不破坏螺旋的结构，从而不破坏整体的二级结构。

我们在模拟的过程中不破坏分子的二级结构，这样做的好处是，由于RNA的体系非常大，这样做可以提高模拟速度， 因为省去了减少形成二级结构的配对所需要的时间，而只关心比较大的分枝在空间中的排布。

我们构建了一个简单的势能函数用以在蒙特卡罗算法的每一步判断接受或是拒绝。这个势能函数由三个部分组成：



其中，



vb表示虚键，这是因为优化是基于残基级别的，所以最小的采样单元是残基，虚键表示的是残基与残基之间的虚拟的键。是与虚键长相关的能量， 是与虚键角相关的能量，是与主链虚扭角相关的能量，是与约束是否满足相关的能量。*kl, ka, kt, kc*分别是相关的参数，*l0, a0, c0*分别是标准键长，标准键角，给定的约束。

与其它文章所用的势能函数相比，这个势能函数非常简单，主要区别是省略了pairing项和stacking项，我们可以这样做的信心来源于不破坏二级结构的采样方式，而其它的势能函数则不能这么做。

### 恢复成全原子结构

我们首先从核糖体RNA的三级结构从第一个碱基开始，每4个碱基组成一个碱基片段，放入库中，从而构成了一个片段库。

由于前面的步骤都是在残基级别进行的，所以我们需要将这个粗粒化的结构转换为全原子的结构。我们对于每一个残基用主链上的C4’原子来表示。

我们以4个残基作为一个窗口，然后从N端开始将窗口滑到C端，每滑动一步，就把窗口中的4个残基的坐标提取出来，然后从库中搜索与其结构差异最小（RMSD最小）的4残基模板进行替换。这样当窗口滑到C端，所有的残基就都被替换成为了全原子。

为了进一步减少空间冲突，我们最后用Amber进行最陡下降和共轭梯度的优化。

## 结果与方法

# RNA三级结构打分

## 引言

RNA分子在生物体内具有非常重要的作用，除了作为DNA与蛋白质之间的信使（mRNA），转运氨基酸到核糖体来合成蛋白质（tRNA），作为核糖体的组成成分（rRNA）等功能，它还能起到调控作用。如同蛋白质，RNA的三级结构的信息对于更好的理解RNA的功能非常重要。由于目前解析出来的RNA的实验结构的数目非常少，现在已经有一些利用计算机对RNA的三级结构进行预测的方法。这些预测方法通常会产生非常多的候选结构，通常需要对这些结构进行评估来挑选出合适的结构。

基于知识的统计势已经被证实是一个评估蛋白质三级结构的非常有力的方法。目前，已经出现了一些基于知识的方法来对RNA的三级结构进行评估与打分。例如Capriotti等人提出的RASP（Ribonucleic Acids Statistical Potential）以及Bernauer等人提出的KB。RASP的全原子版本（RASP-ALL）使用了经过聚类后的23类原子，其参数是通过由MODELLER构造的非冗余的训练集训练出来的。RNA KB也是一个基于距离的统计势。RNA KB的粗粒化版本使用了5个原子（P，C4’在骨架上，C2，C4，C6在碱基上）来表示一个核苷酸。在RNA KB的全原子版本中，所有的85个位于A U G C四个核苷酸中的不同位置上的原子都被用来进行打分，并且不像RASP-ALL中把这85个原子聚类成23类再使用，而是把每个不同的核苷酸上的不同原子都作为一个独立的个体。并且，RNA KB统计势使用了迪拉克过程混合模型（Dirichlet process mixture model）来对距离分布进行拟合，而不是直接用装箱法(23)（bin counting）进行统计。因此RNA KB势是可导的，这使得它可以被用来进行分子动力学模拟。相关的测试结果(17,18)表明RNA KB势可以有效地识别出正确的天然态结构。然而，为了提高对近天然态结构进行排序的能力以及从包含非标准碱基配对的结构中识别出最近天然态的结构的能力，我们需要进一步的努力。另外，在一些三级结构预测的程序中也包含了三级结构打分的功能，它们的打分能力在文献(17)中被测试过了。例如，ROSETTA套件中的针对RNA的全原子的统计势FARFAR（fragment assembly of RNA with full-atom refinement）已经成功地应用在了RNA三级结构的从头预测和设计。FARFAR中包含了氢键相关以及溶解能相关的能量项，以及碱基和主链氧原子之间可能存在的氢键的相互作用。

在这个工作中，我们开发了一个新的全原子的基于知识的统计势，3dRNAscore。与前面提到的2个与距离有关的基于知识的统计势（RASP，KB）不同，3dRNAscore引进了一个新的基于主链二面角的能量项。基于主链二面角的能量项可以更有效地来描述RNA分子的柔性(24,25)。另外，与其他的打分势不同，我们在3dRNAscore的基于距离的能量项中还考虑了相邻2个碱基之间的碱基堆积的作用。我们在接下面的测试中发现3dRNAscore在识别出正确的天然态以及对近天然态结构进行排序两方面的能力要强于RASP和RNA KB。

## 原理与方法

构造3dRNAscore的过程如下。首先，我们根据玻尔兹曼分布来设计3dRNAscore的函数形式。3dRNAscore包含两个能量项：与距离有关的能量和与主链二面角有关的能量。第二部，为了训练打分函数的参数，我们构造了一个包含非冗余RNA三级结构的训练集，其中含有高相似性并且包含有相似模板的结构都被去掉了。通过这个训练集，我们可以得到打分函数的参数。第三部，我们使用这个测试集来测试3dRNAscore的性能。这里，我们选择了3个已知的训练集。我们用不同的指标来比较与其它的方法相比3dRNAscore的性能。详细的构造3dRNAscore的过程如下。

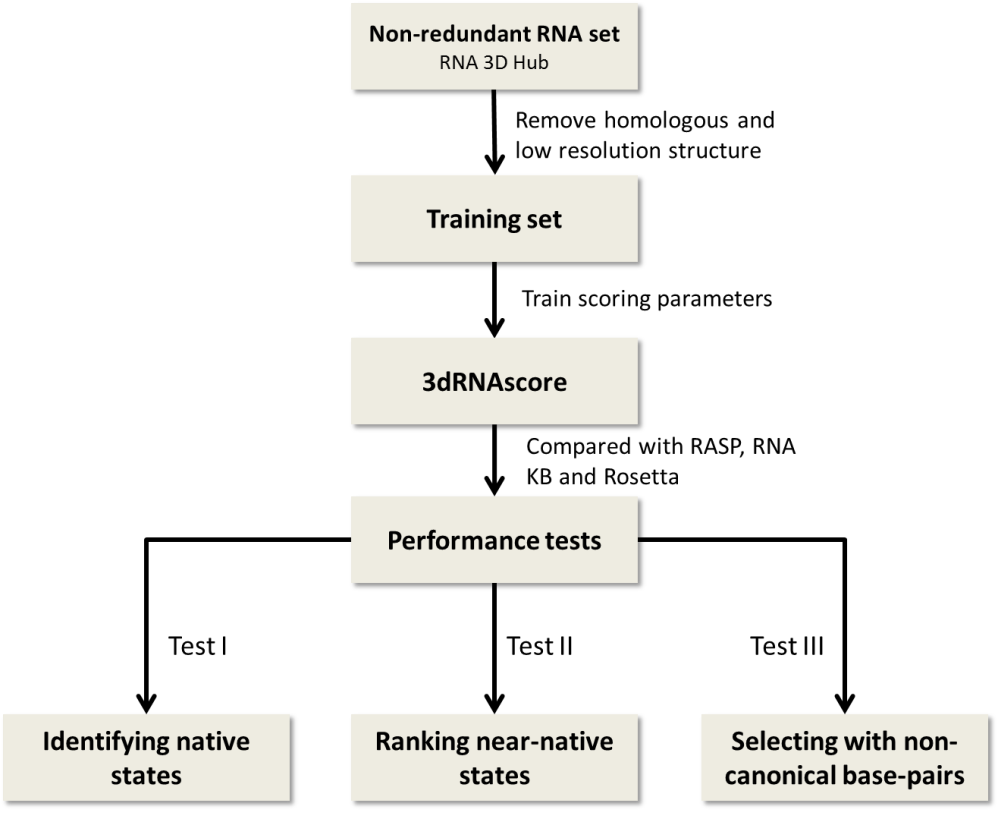


Figure 1 Flow charts of the building steps of 3dRNAscore.

### 与距离有关的能量项

我们的基于知识的能量函数3dRNAscore由两部分组成：第一部分是基于分子当中两个不同残基的任意两个非键相互作用的重原子之间的距离的；第二部分是基于主链的扭转角（二面角）的。其中，基于距离的项是从波尔兹曼分布推导而出的。

Samudrala指出了基于平均力势的打分函数的三个假设。第一个假设是一个分子的总的自由能，，可以被表示为一系列单独的分布的自由能的和。其中，R表示反应坐标的值。反应坐标可以是分子的任意可测量的值。当用距离原子*i*（类型为*a*）核*j*（类型为*b*）之间的距离*d*作为反应坐标，和可以被表示为：



第二个假设是相对自由能可以从波尔兹曼分布分布反向推导而得：



将公式（2）带入公式（1），可得：





其中，T是温度（设置为298K），是玻尔兹曼常数，是天然态RNA中观测到的类型*a*和*b*的两个原子之间的距离的概率，是参考态结构中两个类型*a*和*b*的原子之间的距离的概率。

第三个假设是热动力学假设【文献】：天然态的构象具有最低的自由能。因此公式（3）可以被用来对近天然态的结构进行排序。

概率可以通过统计实验结构中的相关信息得到：



概率不能只由实验结构得到。它依赖于我们选择哪个参考态。RASP【文献】和KB【文献】分别使用平均参考态（RAPDF）和准化学势参考态（KBP）。当计算原子对之间的距离的时候，这些参考态考虑了原子之间的共价键约束，并且通常需要更大的实验结构数据集来得到准确的参考势。在平均参考态中，不考虑原子的类型：



将公式（4）和（5）带入公式（3）可得：



其中，是类型*a*和类型*b*的两个原子之间的距离*d*出现的数目。是不考虑类型时两个原子之间的距离为*d*的数目。是所有的类型为*a*和*b*的原子对的数目。

通常，分别位于序列上相邻两个核苷酸中的两个原子将不被考虑。这是因为相邻核苷酸中的两个原子之间可能会存在键的相互作用。基于距离的统计势只考虑非键的相互作用。然而，相邻碱基中的原子不会存在键相互作用，只有碱基堆积作用。因此，不像其它的统计势，在3dRNAscore考虑了相邻碱基中的原子对。另外，在统计过程中3dRNAscore所取的最大的距离d（通常称为截断）取为20Å。

我们的基于重原子之间的距离的统计势用到了四种核苷酸（腺嘌呤，胞嘧啶，鸟嘌呤，尿嘧啶）中的所有的85种原子类型【表】。我们将这85种原子类型从1编号到85.对于从1-1到85-85的每个原子对，我们在最小间隔为0.15Å的离散空间中来统计距离分布。我们用一个矩阵来表示距离分布的信息，其中每一行表示一个原子对，每一列表示对应的小间隔中的距离的数目。因此矩阵的第i行和第j列的数值就表示天然态中原子类型为第i个原子对的两个原子之间的距离为j的数目。最后，我们将矩阵保存到一个参数文件当中，它将在后续的打分过程中用到。

Table 1 Atom types in 3dRNAscore potential

|  |  |
| --- | --- |
| A | P OP1 OP2 O5’ C5’ C4’ O4’ C3’ O3’ C2’ O2’ C1’  N9 C8 N7 C5 C6 N6 N1 C2 N3 C4 |
| U | P OP1 OP2 O5’ C5’ C4’ O4’ C3’ O3’ C2’ O2’ C1’  N1 C2 O2 N3 C4 O4 C5 C6 |
| C | P OP1 OP2 O5’ C5’ C4’ O4’ C3’ O3’ C2’ O2’ C1’  N1 C2 O2 N3 C4 N4 C5 C6 |
| G | P OP1 OP2 O5’ C5’ C4’ O4’ C3’ O3’ C2’ O2’ C1’  N9 C8 N7 C5 C6 O6 N1 C2 N2 N3 C4 |

### Bin距离的选择

概率和保存为以为间隔的离散的距离区间内。间隔的大小对概率分布有极大的影响。如果间隔太大了，保存在离散空间内的概率和就会太粗糙。如果间隔太小了，那么某些间隔区间内的点数可能会过少，这会造成概率分布的误差过大。间隔的大小应该与样本的总数相匹配。随着样本的数目的增长，间隔的大小可以越来越小，同时要保持每个间隔内的样本点足够多，并且每个间隔内的样本点必须只是总的样本的一部分。

Sippl【文献】使用大小为1Å的距离间隔，Samudrala【文献】也使用距离为1Å的距离间隔，并且进行了spline拟合，Capriotti（RASP）也使用了大小为1Å的距离间隔，Bernauer（KB）使用了一个Dirichlet过程混合模型来得到一个可进行微分分析的势能以取代拥有固定间隔的离散势能。3dRNAscore是使用有固定间隔的离散势。由于训练集具有大量的样本，间隔大小可以取一个非常小的值。为了选取一个合适的间隔大小，我们对间隔大小的值取0.1Å, 0.15Å, 0.2Å, 0.25Å, 0.3Å, 0.35Å, …, 0.95Å, 1.0Å分别进行了测试。测试的结果将在结果部分进一步进行讨论。

### 稀疏数据

在数学统计中，一个大的样本数可以得到一个更精确的分布。小样本数往往导致分布的不确定度太大。因此某个原子类型对的距离的分布的精度将极大的依赖于这个原子对类型出现的总数目。

1990年，Sippl设计了一个算法来解决小数据集的问题【文献】。它通过计算总密度和统计频率来估算真实频率：



其中，*m*是，表示自定义常数。当，收敛到。当，收敛到。公式（7）可以保证当遇到稀疏数据的情况时，真实概率分布可以类似于总的分布。

为了避免稀疏数据所带来的错误（某个原子对类型的观测数太少了），我们检查了所有的原子对类型的观测数目的大小。结果显示即使是最小的观测数目也达到了39030，这足够用来产生一个非常精确的概率分布。这表明3dRNAscore没有必要使用Sippl的方法来处理稀疏数据。

### 基于二面角的能量项

我们不仅使用基于距离的能量项，也使用基于二面角的能量项，包括RNA的7个二面角（主链上的α, β, γ, δ, ε, ζ 以及碱基有关的 χ ）。首先，我们再训练集上计算它们的统计分布【图】。

在得到了统计分布之后，类似于基于距离的势，可以假设所有的这些角遵从玻尔兹曼统计分布，然后仍然可以使用平均力势【文献】：



其中，T是温度（298K），是玻尔兹曼常数，是实验结构中角度*i*（α, β, γ, δ, ε, ζ 和χ）的大小为*θ*并且类型为*a*的观测概率，是参考态中角度*i*的大小为*θ*并且类型为*a*的期望概率。

然后，我们可以用与基于距离的势的同样的方式来得到一个总的分数：





与基于距离的能量一样，我们仍然用一个矩阵来储存所有的角度分布的信息。每一行表示一个二面角对的类型，每一列表示大小在对应的间隔区间内（间隔大小为1度）的二面角的数目。因此这个矩阵含有7\*7行和360列。

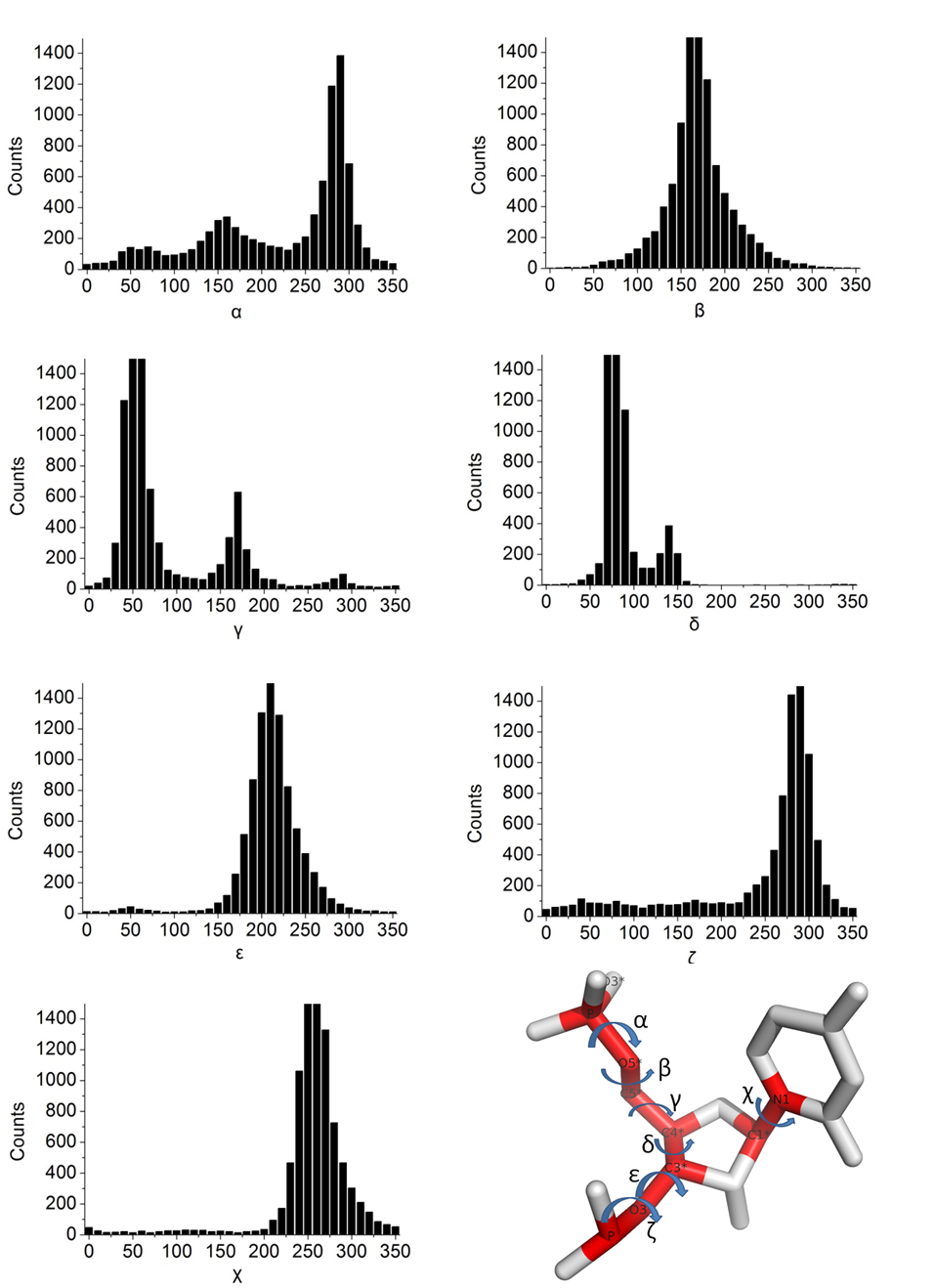


Figure 2 The frequency distributions of seven RNA torsion angles (α, β, γ, δ, ε, ζ for the nucleotide backbone, and χ for the base).

### 结合两个能量项

以一定的权重将3dRNAscore的这2个能量项进行求和可以得到一个总的新的能量：



其中，是总的能量，是基于距离的能量，是基于二面角的能量，是权重。

基于二面角的能量考虑了主链上相邻四个连续的原子之间的相互作用。基于距离的能量考虑了分子内部原子与原子之间的相互作用，不过它不考虑主链上相邻两个残基中的2个原子之间的相互作用。因此，基于二面角的能量与基于距离的能量并没有相互重叠的部分。

### 训练集

我们通过RNA 3D Hub的非冗余数据集来训练3dRNAscore。RNA 3D Hub从PDB中挑选出来那些含有RNA的三级结构，然后根据【文献】的第13章的方法进行去冗余。Leontis’组总结了两类PDB/NDB数据库中存在的结构冗余性。这些冗余性包括单个PDB文件中的冗余性以及PDB/NDB数据库中的冗余性。他们设计了一些方法来去除这些冗余性，然后对这些结构依据结构的冗余性进行聚类为749个类（Release 1.32, 2013-10-12）。对于他们如何去除冗余性，可以参考【文献】的第13章。对于每个类，他们选择了一个结构来代表这个类并且赋予了一个单独并且稳定的id给它。这个数据集可以从<http://rna.bgsu.edu/nrlist/oldsite.html>下载。

为了构造训练集，我们首先从RNA 3D Hub非冗余RNA集获取了749个类的代表性结构。RNA 3D Hub确保了任何两个序列直接的序列相似度小于95%。另外，RASP也用了95%的序列相似度和80%的覆盖率来减少训练集的冗余度。RNA KB势用了一个更低的序列相似度80%类去掉训练集中的同源性。我们然后使用blastn【文献】来去掉所有序列相似度在80%以上并且覆盖率在80%以上的序列。然后，我们使用三级结构比对的方法ARTS程序【文献】来去掉所有比对上的三级结构的覆盖率在80%以上的RNA。然而，在这些处理之后仍然会存在重复的模块，例如sarcin模块以及GNRA模块。这些模块往往会具有相同的三级结构，但是具有完全不同的序列并且通过序列比对很难发现。这些重复的模块也会影响打分函数的训练并且应该去掉来减少冗余。根据RNA 3D Motif Atlas【文献】（一个RNA 3D 模块的集合），我们可以找到训练集中的每个结构所包含的模块。然后我们将训练集中的那些包含了其他的结构也包含的模块的RNA去掉。这一步可以确保训练集和测试集没有序列和结构上的重叠。最后，将那些低分辨率（精度大于3.5Å）的结构也都去掉。这样训练集最后将剩下317个结构，这些结构不包含同源的序列或者模块，并且实验结构精度也很高（小于3.5Å）。

### 测试集

我们再三个不同的测试集上测试了我们的基于知识的打分函数。测试集I是randstr测试集【文献】，它是对85个天然的结构通过MODELLER【文献】软件对二面角和距离进行高斯约束然后进行采样产生的。这是我们用的最大的一个数据集，可以在<http://melolab.org/supmat.html>下载。

测试集II是由Bernauer小组【文献】构造的一个测试集以及FARNA【文献】测试集构造而成。前者是通过位置约束动力学，REMD模拟以及正则模式微扰方法【文献】来产生的。在REMD模拟当中，对于每个RNA结构都进行1纳秒的REMD模拟。温度分布在285K至592K中的50个不同的温度上。在正则模式微扰方法中，这些结构包含了空间上没有冲突的正确的键长以及键角，但是并没有正确的接触。这些方法可以产生RMSD从0到10Å的测试集，可以从http://csb.stanford.edu/rna下载。FARNA测试集包含了用FARNA预测的许多近天然态的三级结构。

测试集III是FARFAR的数据集【文献】，它是由Rosetta-3.1通过RNA预测程序预测出来的。FARFAR数据集对32个包含非标准配对的模块预测三级结构然后进行聚类得到了5个最低能量的类。FARNA和FARFAR测试集可以通过<http://www.stanford.edu/~rhiju/data.html>下载。

### 评估RNA结构的指标

为了量化地比较两个RNA结构，我们应该用一些指标来评估RNA的三级结构。最常用的指标是均方根偏差（RMSD），RMSD描述了两个RNA三级结构之间的全局的几何差异性，但是却不利于描述RNA结构内的氢键网络的差异性。因此一些用来考虑RNA分子内的氢键网络的差异性的指标被提出来用以评估RNA的结构。一个经常被使用的这样的指标是变形指数（DI），它是由Parisien【文献】提出来的。DI的定义是：



其中，RMSD反映了几何的差异性，INF反映了拓扑的差异性。INF(A, B)是两个结构之间的相互作用网络保真度。碱基配对和碱基堆积是RNA分子内的两个主要的相互作用。这两个相互作用组成了RNA分子内部的氢键相互作用网络。假设是参考态中的相互作用的集合，是预测的结构中的相互作用的集合，那么真阳性被定义为，假阳性被定义为，假阴性被定义为，INF被定义为马修斯相关系数（MCC），也就是说：



马修斯相关系数（MCC）在文献【文献】中可以被估计为：



其中，并且。

在这个工作中，我们同时使用了RMSD和DI来估计一个RNA的结构与其实验结果的相似程度。

## 结果

### 识别出天然态的能力

打分方法的一个重要的功能就是从大量的结构中挑选出最近天然态的结构。为了公平地比较3dRNAscore与其它已知的RNA的基于知识的打分方法（RASP 【文献】, KB 【文献】 and Rosetta (FARFAR)）在识别出天然态结构方面的能力，我们在相同的测试集中进行测试，并且每一个打分方法都使用了全原子表示。

我们再测试集I和测试集II上分别进行了测试【图】。当使用3dRNAscore时，测试集I中的85个天然态结构中的84个都被挑选出来了，并且测试集II中的39个天然态结构中的36个都被挑选出来了。对于RASP，测试集I中的85个天然态结构中的79个被挑选出来了，测试集II中的39个天然态结构中的34个被挑选出来了。KB可以从测试集I的85个天然态结构中挑选出80个出来，从测试集II的39个天然态结构中挑选出33个出来。Rosetta可以从测试集I的85个天然态结构中挑选出53个出来，从测试集II的39个天然态结构中挑选出26个出来。这些结果表明了3dRNAscore与其它方法相比在识别出天然态结构方面的能力有着较为明显的优势。

值得注意的是Capriotti的小组验证了RASP在打分方面有着比NAST和AMBER的分子力场更为优异的表现，所以我们并未测试NAST和AMBER立场。

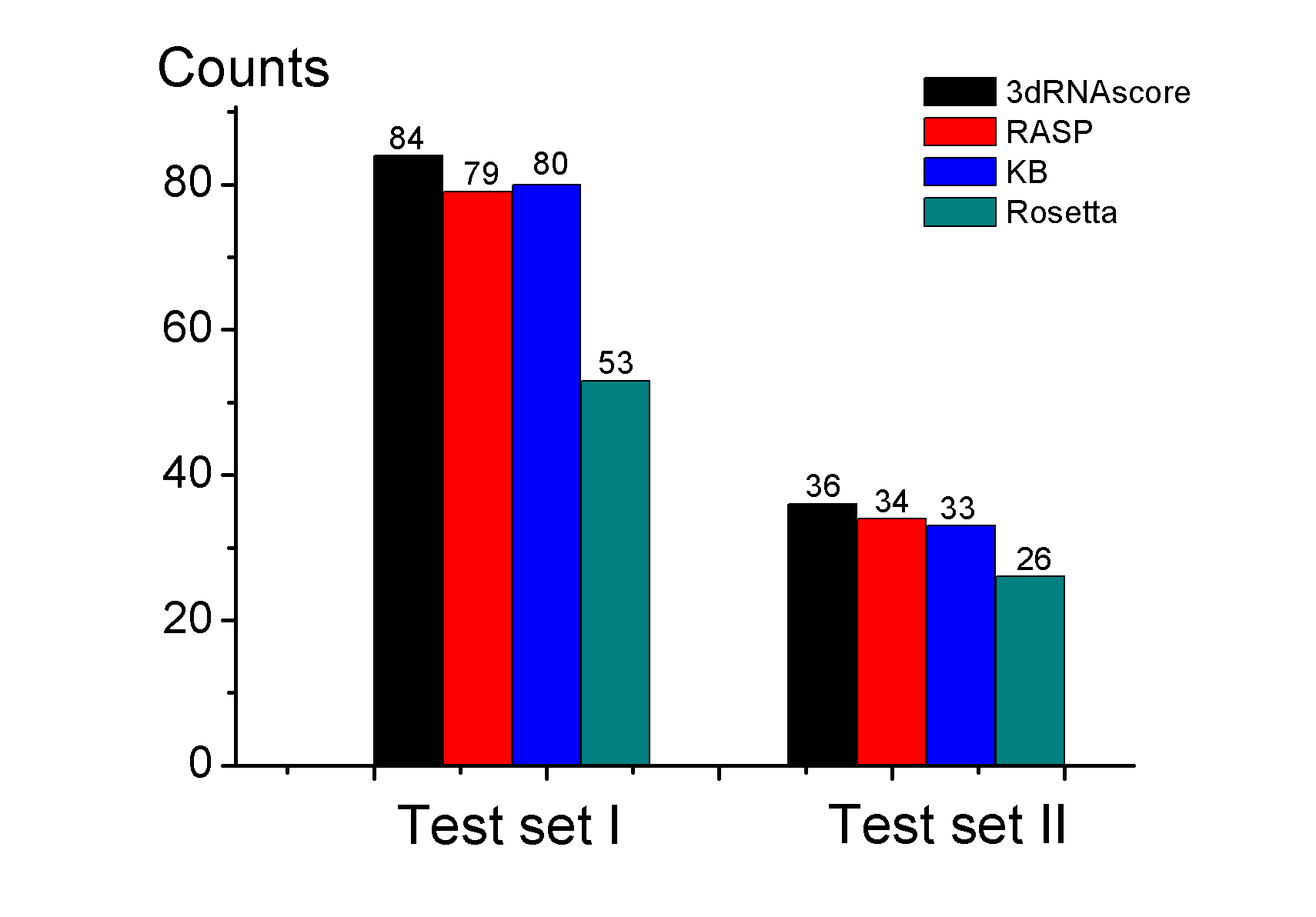


Figure 3 Counts of native states identified correctly from Test set I and Test set II by 3dRNAscore, RASP, KB and Rosetta, respectively.

### 对近天然态结构进行排序的能力

打分函数的另外一个重要的功能是对近天然态结构进行排序。在RNA结构预测中，一个很自然的问题使如何知道一个结构比另外一个结构更接近天然态。对于一个RNA来说，一个结构可以从几何和拓扑两个方面比另外一个结构更加接近天然态。因此，在这个工作当中我们同时使用了RMSD和DI来评估结构之间的差异性。正如前面所提到的，RMSD用来评估结构之间的几何的差异性，DI用来评估结构之间的拓扑的差异性。如果我们能够知道一个结构比另外一个结构更加接近天然态，那么我们就能验证打分函数是否给出了正确的分数。

我们使用富集分数（ES，Enrichment Score）来评价打分函数在对近天然态结构进行排序的方面的能力：



其中是能量最小的10%的结构的集合，是RMSD（或者DI）最小的10%的结构的集合，是和的交集。如果打分与RMSD（或者DI）之间呈完全的线性关系，那么ES的值为10.如果它们之间的关系是随机的，那么ES为1，因此



我们同时使用RMSD和DI两种指标在测试集II中对3dRNAscore，RASP，KB以及Rosetta在对近天然态结构进行排序的方面的能力进行了测试。表格2表明，如果使用DI来计算ES，那么3dRNAscore打分的结果的平均ES要比其余三种方法的结果好(RASP(ES=3.8)，KB(ES=3.7)，Rosetta(ES=2.7))。在测试集II的三个子集REMD，正则模式以及FARNA测试集中，3dRNAscore的结果 也要比其余三种方法要好。如果将测试集II分为NMR和X-Ray两部分，3dRNAscore的结果也要比其余三种方法好。如果使用RMSD来计算ES，最终的结果也同样如此。这些结果表明3dRNAscore在对近天然态结构进行打分排序的方面的能力要比其余方法好。

Table 2 The near-native structures ranking results of 3dRNAscore, KB potential, and Rosetta methods.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  | **Enrichment Score (RMSD)** | | | | **Enrichment Score (DI)** | | | |
| **Decoy** | **RNA** | **Length** | **Method** | **3dRNAscore** | **KB** | **RASP** | **Rosetta** | **3dRNAscore** | **KB** | **RASP** | **Rosetta** |
| Position restrained dynamics and REMD (A) | 1duq | 26 | X-ray | 8.5 | 7.6 | 7.6 | 7.1 | 8.4 | 7.5 | 7.6 | 7.0 |
| 1f27 | 30 | X-ray | 8.3 | 7.9 | 6.6 | 6.2 | 8.2 | 7.8 | 6.6 | 6.2 |
| 1msy | 27 | X-ray | 7.5 | 5.7 | 5.7 | 3.6 | 7.4 | 6.0 | 5.6 | 3.5 |
| 1nuj | 24 | X-ray | 7.7 | 7.3 | 5.2 | 6.9 | 7.7 | 7.2 | 5.2 | 6.7 |
| 434d | 14 | X-ray | 8.0 | 7.7 | 7.0 | 6.8 | 7.9 | 7.7 | 6.9 | 6.8 |
| Normal modes (B) | 1duq | 26 | X-ray | 7.5 | 7.0 | 5.7 | 3.8 | 6.3 | 7.0 | 5.7 | 3.5 |
| 1esy | 19 | NMR | 4.9 | 5.4 | 4.5 | 5.6 | 6.3 | 5.5 | 4.7 | 5.7 |
| 1f27 | 30 | X-ray | 5.9 | 5.8 | 3.7 | 2.6 | 6.1 | 5.8 | 3.7 | 2.5 |
| 1i9v | 76 | X-ray | 6.1 | 2.6 | 5.3 | 3 | 5.9 | 2.7 | 5.1 | 3.0 |
| 1kka | 17 | NMR | 5.7 | 4.6 | 4.1 | 4.6 | 6.5 | 4.3 | 4.1 | 4.8 |
| 1msy | 27 | X-ray | 5.9 | 5.6 | 2.2 | 4.6 | 6.3 | 5.4 | 2.7 | 4.5 |
| 1nuj | 24 | X-ray | 7.1 | 7.4 | 5.9 | 2.4 | 4.5 | 7.0 | 5.9 | 2.2 |
| 1qwa | 21 | NMR | 3.5 | 3.2 | 2.0 | 3.8 | 3.5 | 3.3 | 2.2 | 3.9 |
| 1x9k | 62 | X-ray | 6.7 | 1.6 | 5.2 | 3 | 2.5 | 1.9 | 5.2 | 2.8 |
| 1xjr | 46 | X-ray | 7.7 | 5.4 | 7.9 | 2.2 | 6.0 | 5.0 | 7.9 | 2.5 |
| 1ykq | 19 | X-ray | 4.6 | 3.4 | 3.5 | 2.8 | 4.8 | 3.3 | 3.8 | 2.5 |
| 1zih | 12 | NMR | 7.7 | 5.4 | 5.7 | 6.6 | 7.5 | 5.3 | 5.7 | 6.5 |
| 28sp | 28 | NMR | 5.7 | 4 | 6.5 | 1.8 | 5.7 | 4.5 | 6.7 | 2.9 |
| 2f88 | 34 | NMR | 6.8 | 5.4 | 4.9 | 4.4 | 4.3 | 5.4 | 4.7 | 4.2 |
| 434d | 14 | X-ray | 7.7 | 7.4 | 7.4 | 5.2 | 6.9 | 7.4 | 7.6 | 5.2 |
| FARNA (C) | 1a4d | 41 | NMR | 2.4 | 3.8 | 2.0 | 0.8 | 2.4 | 4.0 | 2 | 0.7 |
| 1csl | 28 | X-ray | 2.0 | 1.5 | 1.6 | 1.3 | 2.6 | 1.4 | 1.6 | 1.4 |
| 1dqf | 19 | X-ray | 4.2 | 1.8 | 2.8 | 1 | 4.4 | 2.0 | 3.2 | 1.0 |
| 1esy | 19 | NMR | 3.2 | 3.7 | 4.8 | 1.2 | 3.6 | 3.5 | 4.2 | 1.1 |
| 1i9x | 26 | X-ray | 3.4 | 1.3 | 2.2 | 1.5 | 4.4 | 1.5 | 3.6 | 1.6 |
| 1j6s | 24 | X-ray | 0.2 | 1.4 | 0.2 | 0.6 | 0.2 | 1.7 | 0.8 | 0.5 |
| 1kd5 | 22 | X-ray | 3.2 | 0.3 | 0.8 | 0.2 | 3.0 | 0.5 | 1.6 | 0.3 |
| 1kka | 17 | NMR | 1.4 | 1.2 | 0.6 | 0.6 | 1.8 | 0.8 | 0.8 | 0.6 |
| 1l2x | 27 | X-ray | 0.6 | 3.2 | 1.0 | 1.8 | 0.8 | 3.2 | 1.2 | 1.8 |
| 1mhk | 32 | X-ray | 1.6 | 1.2 | 1.6 | 1 | 1.8 | 1.2 | 1.4 | 1.2 |
| 1q9a | 27 | X-ray | 2.6 | 0.5 | 0.8 | 0.8 | 3.6 | 0.5 | 1.2 | 0.8 |
| 1qwa | 21 | NMR | 2.2 | 1.2 | 0.4 | 1 | 1.8 | 1.3 | 0.6 | 1.0 |
| 1xjr | 46 | X-ray | 2.4 | 2 | 2.4 | 1.2 | 2.8 | 1.9 | 3.4 | 1.0 |
| 1zih | 12 | NMR | 4.8 | 5 | 4.8 | 2 | 6.8 | 5.5 | 7.2 | 1.9 |
| 255d | 24 | X-ray | 2.4 | 0.7 | 0.6 | 1.3 | 2.6 | 0.7 | 0.6 | 1.1 |
| 283d | 24 | X-ray | 1.4 | 0.8 | 0.8 | 0.7 | 3.8 | 0.8 | 1 | 0.7 |
| 28sp | 28 | NMR | 2.8 | 1.5 | 3.0 | 1.7 | 3.8 | 1.2 | 4.2 | 1.8 |
| 2a43 | 26 | X-ray | 2.2 | 2 | 1.4 | 0.6 | 2.4 | 1.8 | 2 | 0.6 |
| 2f88 | 34 | NMR | 3.2 | 1.3 | 2.2 | 1.3 | 3.6 | 1.3 | 1.6 | 1.2 |
| AVERAGE VALUES | (A) | | | 8.0 | 7.2 | 6.4 | 6.1 | 7.9 | 7.2 | 6.4 | 6.0 |
| (B) | | | 6.2 | 4.9 | 5.0 | 3.8 | 5.5 | 4.9 | 5.0 | 3.8 |
| (C) | | | 2.3 | 1.8 | 1.8 | 1.1 | 2.8 | 1.8 | 2.2 | 1.1 |
| X-Ray | | | 4.8 | 3.8 | 3.7 | 2.8 | 4.7 | 3.8 | 3.9 | 2.7 |
| NMR | | | 4.2 | 3.5 | 3.5 | 2.7 | 4.4 | 3.5 | 3.7 | 2.8 |
| All | | | 4.5 | 3.7 | 3.6 | 2.7 | 4.5 | 3.7 | 3.8 | 2.7 |

### 挑选出正确的包含非标准碱基配对的结构.

非标准的碱基配对经常出现在RNA的三级结构的功能性模块中【文献】。挑选出包含非标准碱基配对的近天然的构象在目前的RNA三级结构预测当中还是一个不可避免和不易解决的问题【文献】。测试集III（FARFAR测试集）就是用来测试3dRNAscore在挑选包含非标准碱基配对的近天然的构象方面的能力。测试集III包含了从包含非标准碱基配对的32个小模块，对每个小模块来预测三级结构并把预测的结构聚成5个类。我们分别对RASP， KB，ROSETTAmin和3dRNAscore进行了测试。然后挑选出最低打分值的结构并与具有最低DI值的结构作比较。表3是测试的结果。3dRNAscore，RASP，KB和Rosetta分别能够从测试集III的32个模块中挑选出10,9,8,和2个具有最低DI值的结构。3dRNAscore挑选出的结构分别有8，12，21个结构比RASP，KB和Rosetta挑选出来的结构的DI值要低，并且分别有22，16，1个结构与RASP，KB和Rosetta挑选出来的结构的DI值相等，只有2，4,和10个结构比RASP，KB和Rosetta挑选出来的结构的DI值高。

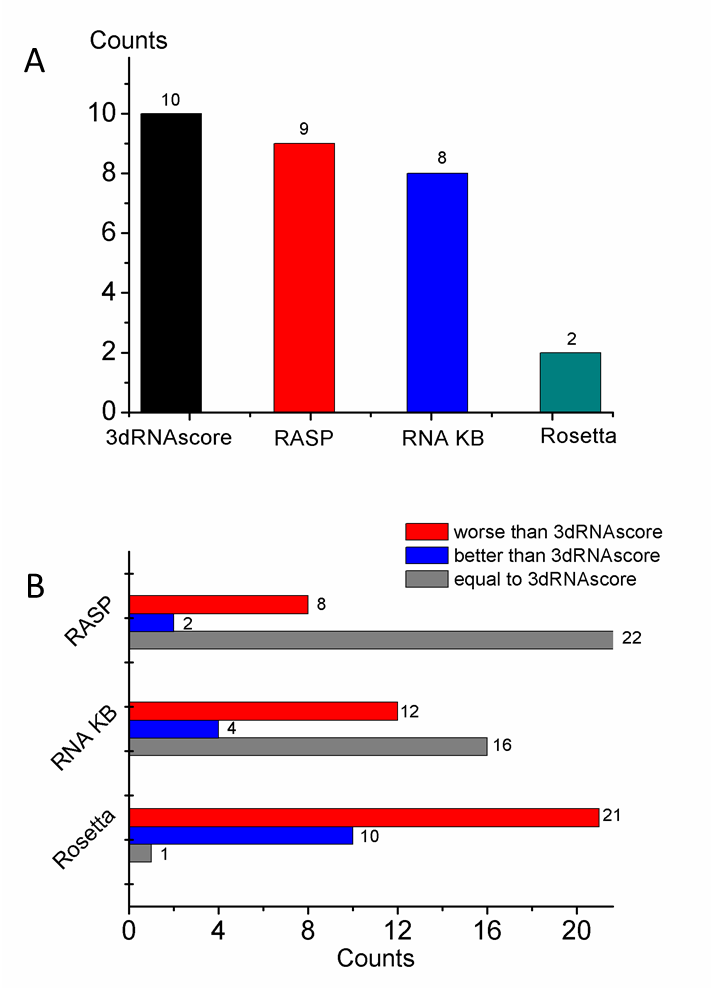


Figure 4 Results on test III. (A) The counts of models having the lowest energy selected by 3dRNAscore, RASP, RNA KB and Rosetta, respectively. (B) The red part represents the counts of models that are worse than those selected by 3dRNAscore. Blue represents the counts of models are that better than those selected by 3dRNAscore. Gray represents the counts of models that are the same as those selected by 3dRNAscore.

Table 3 The ranking results of Rosetta, RASP-ALL, KB, and 3dRNAscore methods on the FARFAR decoy set.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Motif name** | **DI of the Lowest Energy Model** | | | | **Minimum DI**  **Model** |
| **ROSETTAmin** | **RASP-ALL** | **RNA KB** | **3dRNAscore** |
| G-A base pair | 4.026 | 1.408 | 2.455 | 1.408 | 1.408 |
| Fragment with G/G and G/A pairs, SRP helix VI | 13.910 | 6.073 | 6.073 | 6.073 | 5.267 |
| Helix with A/C base pairs | 2.414 | 3.262 | 5.410 | 5.866 | 2.414 |
| Four-way junction, HCV IRES | 21.236 | 21.236 | 21.526 | 20.067 | 15.040 |
| Loop 8, A-type Ribonuclease P | 8.421 | 4.290 | 4.290 | 4.290 | 1.615 |
| Helix with U/C base pairs | 6.012 | 3.136 | 3.136 | 3.136 | 3.136 |
| Curved helix with G/A and A/A base pairs | 1.062 | 1.929 | 3.565 | 1.929 | 0.998 |
| Pre-catalytic conformation, hammerhead ribozyme | 30.097 | 12.830 | 12.830 | 12.830 | 12.830 |
| Loop E motif, 5S RNA | 2.404 | 2.561 | 2.561 | 2.561 | 1.986 |
| UUCG tetraloop | 1.284 | 1.278 | 1.403 | 1.278 | 1.278 |
| Rev response element high affinity site | 7.802 | 5.826 | 4.227 | 5.826 | 4.227 |
| Fragment with A/C pairs, SRP helix VI | 3.930 | 2.424 | 2.424 | 2.424 | 2.424 |
| Signal recognition particle Domain IV | 5.516 | 3.224 | 3.224 | 3.224 | 1.338 |
| Bulged G motif, sarcin/ricin loop | 1.902 | 7.066 | 7.838 | 7.066 | 1.588 |
| Tertiary interaction, hammerhead ribozyme | 25.215 | 24.647 | 24.647 | 24.647 | 19.453 |
| GAGA tetraloop from sarcin/ricin loop | 1.075 | 0.956 | 0.956 | 0.956 | 0.956 |
| Pentaloop from conserved region of SARS genome | 4.147 | 4.145 | 4.147 | 4.145 | 1.068 |
| L2/L3 tertiary interaction, purine riboswitch | 16.371 | 19.160 | 19.160 | 19.160 | 16.371 |
| L3, thiamine pyrophosphate riboswitch | 6.076 | 2.262 | 2.262 | 2.262 | 2.262 |
| Kink-turn motif from SAM-I riboswitch | 1.592 | 13.086 | 12.107 | 12.107 | 1.454 |
| Active site, hammerhead ribozyme | 20.394 | 23.413 | 22.927 | 22.927 | 17.947 |
| P1/L3, SAM-II riboswitch | 14.885 | 18.103 | 10.478 | 18.174 | 10.478 |
| J4/5 from P4-P6 domain, Tetrahymena ribozyme | 2.504 | 2.919 | 2.919 | 2.919 | 2.199 |
| Stem C internal loop, L1 ligase | 3.479 | 3.479 | 3.226 | 3.479 | 3.226 |
| J5/5a hinge, P4-P6 domain, Tetr. ribozyme | 25.063 | 23.920 | 28.689 | 23.920 | 20.910 |
| Three-way junction, purine riboswitch | 13.693 | 13.734 | 10.616 | 10.616 | 10.387 |
| J4a/4b region, metal-sensing riboswitch | 6.237 | 5.012 | 6.549 | 5.012 | 5.012 |
| Kink-turn motif | 15.0735 | 23.997 | 28.729 | 16.805 | 11.695 |
| Tetraloop/helix interaction, L1 ligase crystal | 1.398 | 0.948 | 1.398 | 0.948 | 0.948 |
| Hook-turn motif | 7.598 | 16.240 | 15.354 | 1.985 | 1.985 |
| Tetraloop/receptor, P4-P6 domain, Tetr. ribozyme | 16.803 | 11.112 | 4.477 | 4.477 | 3.818 |
| Pseudoknot, domain III, CPV IRES | 10.097 | 7.063 | 7.771 | 4.377 | 4.274 |

### 不同的bin宽的比较

在前面提到过，我们测试了不同的bin宽对3dRNAscore的打分效果的影响。受益于训练集中的大量的样本数目，我们可以选取一个比较小的bin宽来获得一个非常精确的统计分布。图4展示了不同bin宽时3dRNAscore在测试集II中的打分结果的平均ES值。它表明当bin宽取0.15Å时，3dRNAscore的表现最好。图4还展示了随着bin宽度的增大，3dRNAscore对近天然态的排序能力也随之下降。

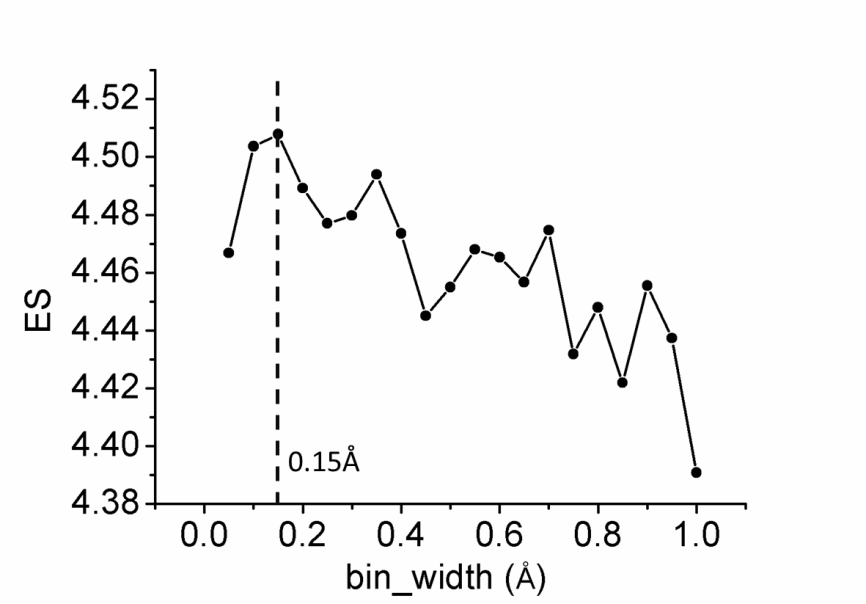


Figure 5 ES of 3dRNAscore versus bin width over test set II

### 与二面角有关的势的贡献

大多数RNA三级结构的统计势的参数是根据原子对或者残基对之间的距离分布得到的。由分子力场得到的灵感，3dRNAscore根据主链二面角的分布增加了一项能量项。要考察这个新加的与二面角相关的能量项的贡献可以通过分别测试只有依赖于距离的能量项，只有依赖于二面角的能量项以及两者都有这三种情况下的打分性能。

在测试集II上使用根据RMSD计算出来的ES作为评判标准，则只使用依赖于距离的能量项的打分结果的平均ES值是4.48，只使用依赖于二面角的能量项的打分结果的平均ES值是2.1，如果综合这两项，则平均ES值是4.57。从这个结果可以看出来增加了依赖于二面角的能量项可以提高整体的打分性能。

### 3dRNAscore与RNA内的物理相互作用之间的关系

Capriotti发现RASP的全原子版本RASP-ALL可以捕捉到RNA内的碱基配对和碱基堆积相互作用。我们用3dRNAscore也观察到了同一现象，从结果看来碱基配对和碱基堆积的相互作用隐含在全原子的与距离有关的统计势之中。例如，图5描述了腺嘌呤的N9原子与尿嘧啶的N1原子之间的距离的统计分布，其中有3个非常明显的峰。第一个峰出现于4.65Å处，这个峰来源于核苷酸链上相邻两个核苷酸之间的碱基堆积作用。第二个峰出现于7.05Å处，它来源于第i个核苷酸和第i+2个核苷酸之间的非直接的相互作用。第三个峰出现于8.7Å处，它来源于腺嘌呤和尿嘧啶之间的碱基配对相互作用。

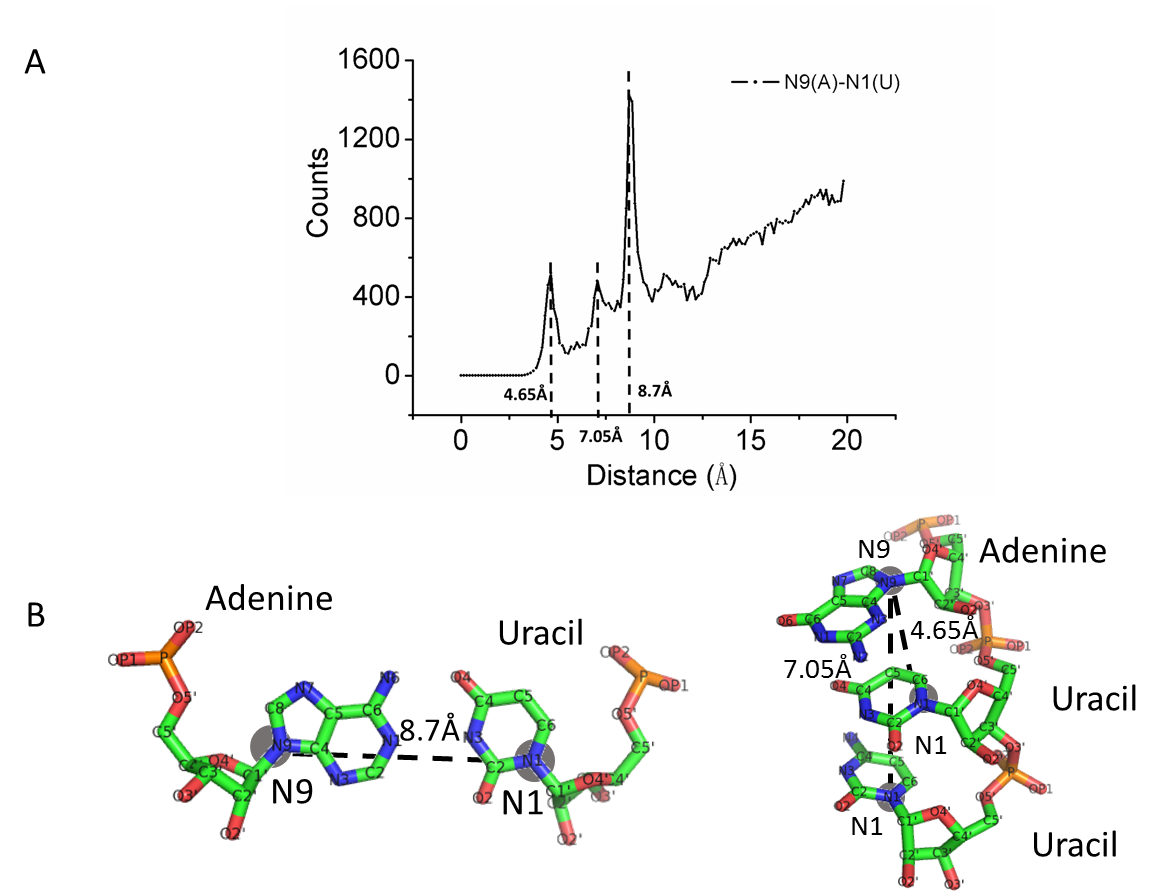


Figure 6 (A) Counts distribution of the distance between N9 of adenine and N1 of uracil. (B) Diagram of the three representative distance between N9 of adenine and N1 of uracil.

### 用更健壮的方式选择合适的结构

打分函数主要是被用来选择合适的结构模型。这里我们给出了一个关于如何在RNA三级结构预测中用一个健壮的方式来选择一个合适的结构的例子。我们首先使用RNA三级结构预测程序3dRNA来预测了RNA 1FQZ的500个结构。然后根据INF将这500个结构聚类成15组。然后用3dRNAscore对这15个组的中心结构进行打分。结果发现组II的中心结构的打分值最低，那么组II的中心结构应该是最接近于天然态结构的一个。然后将组II中的所有的结构进行打分，选出打分值最低的那一个作为最终的挑选出来的结构。表4中展示出了5个最大的组，它表明具有最低的打分值的结构的DI正好是最低的DI。

Table 4 Five largest groups divided from the 500 models predicted by 3dRNA for 1FQZ.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Group I | Group II | Group III | Group IV | Group V |
| Size | 190 | 168 | 56 | 23 | 19 |
| Center DI | 4.636 | 3.153 | 5.621 | 5.055 | 10.183 |
| Minimum DI | 4.116 | 1.429 | 5.355 | 4.373 | 9.758 |
| Minimum score DI | 5.129 | 1.429 | 5.930 | 5.048 | 10.307 |

## 分析

根据Anfinsen提出的热动力学假说，天然态结构往往含有最低的自由能。因此理论上来说，我们需要使用自由能来评估一个结构。对于经典的分子力场来说，用它来计算焓非常容易，但是用它来计算熵却非常耗时。与立场相反，从实验结构当中获取信息的基于知识的统计势本身就包含有焓和熵的信息。尽管基于知识的统计势不是自由能，但是它们之间有着强烈的关系。更进一步地，基于知识的统计势非常容易计算并且非常快速。因此基于知识的统计势比分子力场更多地用来对蛋白质和RNA的结构进行打分。

3dRNAscore与其它的打分函数的一个主要的区别是它包含了一个依赖于距离的能量项和一个依赖于二面角的能量项。上面的结果已经表明了3dRNAscore的打分性能比其余的方法要好。我们知道分子力场是进行分子动力学的基础。许多分子力场都包含有4项：键拉伸项，键角弯曲项，键的旋转项以及非键相互作用。依赖于距离的能量项对应于分子力场中的非键相互作用项，而依赖于二面角的能量项对应于分子力场中的键的旋转项。后者可以有效地描述RNA分子的柔性，而这正是RNA的结构很大得区别于蛋白质的地方。最初我们打算设计4个不同的能量项来对应于分子力场的4个组成项，但是我们发现键长和键角项并没有起到很大的作用，因为测试集中并不包含通过对键长和键角进行改变而得到的结构。

RNA主链二面角具有非常稳定的分布，这可以帮助我们识别出天然态的结构。Murray利用品质过滤技术例如精度，晶体B因子和全原子空间冲突等对一个包含有8636个核苷酸的RNA的数据库的主链二面角的分布进行了处理。当去除掉大部分噪声时，可以看到非常清晰的信号。这表明天然态的结构的主链二面角有着非常明显的倾向性的分布。3dRNAscore的依赖于二面角的能量项可以利用这一点来从二面角不同于天然态的分布的测试集中选出最接近于天然态的结构。

Rosetta的针对于RNA的打分函数（FARFAR）比大多数其余的基于知识的RNA的打分函数要更加复杂和精确。FARFAR里面所用的RNA的能量函数包含了与回转半径有关的项，惩罚空间冲突的项以及一些其它的专门为RNA分子内部的相互作用而设计的项。由于FARFAR里面包含了如此多的项，它的精度与测试集的关系很大。在上面的测试中FARFAR的打分结果不好可能与测试集有关。

测试集I是把分子内的距离和角度根据高斯分布进行采样。测试集II是通过进行一段短时间的分子动力学模拟来得到的。因此这两个测试集中的结构的复杂性较低，很有可能分布在一个自由能的局部最低点。测试集I和II中的测试表明，3dRNAscore非常擅于找出局部自由能最低的结构以及对位于局部自由能最低点附件的结构进行排序。而测试集III是一个更接近实际的测试集，它是用FARFAR进行RNA预测得到的，因此这个测试集包含的RNA的复杂性非常高，这归功于FARFAR所采用的蒙特卡罗算法。测试集III中的结果表明3dRNAscore对于广泛的分布在自由能面各个地方的结构的排序的性能不是特别好，但是仍然比其它的方法要好。

基于知识的统计势目前仍然有一些其它的限制。目前已经解析出来的RNA的结构的数目太少是基于知识的统计势遇到的主要的困难，因为过少的数目会导致过拟合的问题。我们尝试使用一个非冗余的RNA三级结构数据集来作为训练集并且与去掉测试集中的RNA相同或含有相同的模块的RNA来尽量减少过拟合的影响。进一步的，由于RNA的磷酸基团带有强烈的负电荷，导致RNA对于电的相互作用非常敏感。在缺少相应的阳离子的时候，RNA很可能形成不了三级结构。而目前的RNA的基于知识的统计势知识隐式的考虑了电的影响，我们将在未来对这一问题进行进一步的研究。

## 本章小节

在这篇文章中，我们开发了一个新的基于知识的统计势3dRNAscore来识别天然态RNA以及对近天然态结构进行排序。它包含了一个基于距离的能量项和一个基于二面角的能量项。我们构造了一个非冗余的训练集来训练3dRNAscore的参数。我们然后使用了三个测试集来全方位的对3dRNAscore的打分性能进行测试。测试结果表明我们的方法的打分性能比目前已知的其它的打分方法要好。同时我们对3dRNAscore的一些缺点进行了分析。

# DNA三级结构预测

## 引言

引言。

## 原理与方法

### ssDNA的三种类型的3D结构

First, we downloaded all the DNA 3D structures from PDB database [[27](#_ENREF_27)]. 我们对所有的这些结构进行了分析，发现这些结构可以分为3类，第一类是类似于非编码RNA的存在双链螺旋的结构，在这篇文章中为了表述的方便，我们将这类DNA称为D-DNA (double helix based DNA)。这种双链螺旋是指由A-U，G-C，G-U这三种标准碱基配对堆积形成的连续的碱基配对。一个ssDNA分子当中可能存在有多个螺旋区域，而没有与其它碱基形成配对的那些碱基形成环，如图1所示，这些环包括发卡环，内环，凸环与多分枝环。我们经常使用点括号的形式来表示核酸分子的二级结构，例如 “(((((….)))))”就代表一个二级结构，它包含了一个螺旋“((((()))))”与一个发卡环“(….)”。

第二类ssDNA是存在有三链螺旋的结构，这种三链螺旋是由包含有3个碱基的连续的三碱基对堆积而成的。为了表述方便，在这篇文章中我们将这类RNA称为T-DNA (triple helix based DNA)。如图1，这种三碱基对的三个碱基之间形成了两种碱基配对，一个是标准配对，另外一个是非标准配对。在这里我们将同时与另外两个碱基形成了配对的那个碱基称为1类碱基，将与1类碱基形成了标准配对的那个碱基称为2类碱基，将与1类碱基形成了非标准配对的那个碱基称为3类碱基，将不与其它任何碱基形成配对的碱基称为0类碱基。因此，我们通过将碱基的类型转为为这种分类来表示三链螺旋的二级结构，例如 “003333330001111110000222222000”就表示一个三链螺旋的二级结构。我们在T-DNA的天然态结构中发现了3种类型的三碱基对，按照“1类碱基-2类碱基-3类碱基”这种方式表述，这3种3碱基对分别是G-C-G，G-C-C，A-T-T。其中G-C-G占(?)，G-G-C占(?)，A-T-T占(?)。同时按从5’到3’的顺序，3碱基对中3个碱基的顺序有4种：1-2-3,2-1-3,3-1-2,3-2-1，也就是说3类碱基不可能出现在中间的位置。

第三类ssDNA是存在有四链螺旋的结构，这种四链螺旋是由4个G形成的G-quadruplex堆积形成的。为了表述方便，在这篇文章中我们将这类DNA称为Q-DNA。与D-DNA和T-DNA不同的是，大部分的Q-DNA通常由2条链组成，只有极少数与D-DNA和T-DNA一样，是一条完整的链。这是由于在G-quadruplex中的4个G所在的链的方向是相同的，故而没有办法连成一条完整的链。在Q-DNA当中，我们将形成了配对了的碱基称为1类碱基，将没有形成配对的碱基称为0类碱基，那么如图1所示，我们可以用“1111000011110000111100001111”这种形式来表示一个Q-DNA的二级结构。

We have created a little database online (<http://biophy.hust.edu.cn/CADNAS>) to deposit the classified information. It includes all the 761 DNAs which we’ve downloaded from PDB database. There are 373 ‘non-helix’ D-DNAs, 184 ‘helix’ D-DNAs, 15 T-DNAs, 84 Q-DNAs, 101 ‘unstructured’ DNAs and 4 other types of DNAs, respectively.

### ssDNA的二级结构预测

与RNA一样，DNA的二级结构有助于预测三级结构。对于D-DNA，由于它与非编码RNA类似的结构，因此可以用RNA结构预测的方法来预测其二级结构。由于a lot of state-of-the-art methods for the prediction of RNA/DNA 2D structure (mainly for base-pair type) prediction exist, we didn’t focus on this field. 我们将采用ContextFold方法[[28](#_ENREF_28)]来预测ssDNA的二级结构。

对于T-DNA，我们改进了用来预测RNA和D-DNA的二级结构的基于自由能最小的动态规划算法，将其用于预测三碱基对，而不是碱基配对。The total 2D prediction process is divided into two steps. The first step is to find the minimum free energy (MFE) and corresponding 2D structure, and the second step is to find the sub-optimal free energies (SFE) and corresponding 2D structures.

Before interpreting the algorithm, we should first give some notation for the convenience. We denote *Si,j* as the sequence from the *ith* to the *jth* nucleotide along the ssDNA sequence, *fi,j* as the minimum free energy of *Si,j*, *Si,j|k,l* as a compound sequence by concatenating the *Si,j* and *Sk,j*, *fi,j|k,l* as the free energy of *Si,j|k,l*, *f(i, j, k)* as the free energy of the base triple *(i, j, k)*.

The first step is to find the MFE of the integral sequence (*f1,N*), where N is the number of the last nucleotide along the sequence. We devised a recursive algorithm to calculate *f1,N*. The core idea can be represented by these two formulas:

where *1’* means the first nucleotide along the new compound sequence *Sa,b|c,d*, *N’* means the last nucleotide.

According to equation (1) and (2), the calculation of MFE is a recursively process. In the meantime, we would calculate a large bunch of minimum free energies for various sequences. These free energies are all maintained after they were calculated, because they would be re-used in the next step of calculating the sub-optimal free energies.

While we were calculating free energies, we could extract the base triple information. We maintained a list L to store all the base triples. In the first, for all the possible *i* - *j* pairs, we could find one that leads to the MFE of *S1,N* based on equation (1), or we found that *f1,N-1* is equal to *f1,N*. If we find, we put the (*i*, *j*, *N*) triad into the list L. If not, we do nothing. We then searched for the triad that make the free energy of *S1,N-1* minimum. Recurred like this, and finally we could find all the triads.

This process is equivalent to the backtracking step in DP algorithm. In our algorithm, the process of calculating MFE and extracting base pair/triple/quadruple information is simultaneous. Whereas in the DP algorithm, the two procedures are separated, the backtracking is after the calculating of MFE.

In the second step, we need to calculate the sub-optimal free energies. We should first set a cutoff which means the maximum part of free energy that exceeds the minimum free energy. We used a percentage to represent the cutoff, e.g. a cutoff of 20% means that all the possible free energies that are not less than the 80% of the MFE are regarded as sub-optimal free energies. After that, we searched all the sub-optimal free energies based on below formulas:



where *F1,N* is a set of *f* which is either *f1,N-1* or *f1,i-1|j+1,N-1+fi+1,j-1+f(i,j,N)* and satisfying *1≤i≤N-1*and *f-f1,N<cutoff*.

We maintained a set V to deposit all the possible 2D structures, each of which is represented by a list L, the elements of which are triads involving all the nucleotides forming the same base triple. In the first, for all the possible *i* - *j* pairs, we could find one that leads to the MFE of *S1,N* based on equation (1), or we found that *f1,N-1* is equal to *f1,N*. If we find, we put the (*i*, *j*, *N*) triad into the list L. If not, we do nothing. We then searched for the triad that make the free energy of *S1,N-1* minimum. Recurred like this, and finally we could find all the triads.

对于Q-DNA，与T-DNA类似, the algorithm for G-quadruplex is:



First, we still calculated the MFE of the integral sequence according to equation (3) and (2). The next steps are almost the same with the aforementioned algorithm for base-triple type DNA except that this time we expected to find the list of tetrads that makes the free energy of *S1,N* minimum.

### 自动构造双链螺旋，三链螺旋和四链螺旋

首先，We’ve performed statistical analysis of the parameters of DNA double helices, triple helices and quadruple helices. 然后，We found a pure mathematics method to build a coarse grained standard DNA double helix.

对于双链螺旋，如图2所示，Let us designated the first chain of the double helix by A1A2A3…An, and the second chain by Bn…B3B2 B1. Then we performed statistical analysis of all the DNA double helices about the distances a1, a2, a3, …, an and b1, b2, b3, …, bn, where an means the distance between A1 and An, bn means the distance between A1 and Bn. Figure A shows the statistic of an and Figure B shows the statistic of bn.

对于三链螺旋，Similar to double helix, we designated the first chain of the triple helix by A1A2A3…An, the second chain by Bn…B3B2B1, and the third chain by C1C2C3…Cn. Then we did a statistic on all the DNA triple helices about the distances a1, a2, a3, …, an, b1, b2, b3, …, bn and c1, c2, c3, …, cn, where an means the distance between A1 and An, bn means the distance between A1 and Bn, and cn means the distance between A1 and Cn.

对于四链螺旋，We designated the first chain of the quadruple helix by A1A2A3…An, the second chain by B1B2B3…Bn, the third chain by C1C2C3…Cn, and the third chain by D1D2D3…Dn. Then we did a statistic on all the DNA triple helices about the distances a1, a2, a3, …, an, b1, b2, b3, …, bn, c1, c2, c3, …, cn and d1, d2, d3, …, dn, where an means the distance between A1 and An, bn means the distance between A1 and Bn, cn means the distance between A1 and Cn, and dn means the distance between A1 and Dn.

### 构造初始三级结构

不管是D-DNA，T-DNA还是Q-DNA，预测它们的三级结构时，我们首先都会构造一个初始的三级结构，然后通过后续的蒙特卡罗分子模拟的方法来对其进行优化以产生最终的结构。我们所构造的初始三级结构会是一个满足给定的二级结构约束的三级结构，只不过这个结构当中可能会存在断链的情况，并且也不会满足提供的一些三级约束。

1. D-DNA

与我们对非编码RNA进行三级结构预测的方法一样，对于D-DNA，我们首先采用模板组装的方式来搭一个结构。

我们首先从PDB数据库中提取出所有的D-DNA，然后用3DNA程序[[29](#_ENREF_29)]得到这些D-DNA的二级结构。根据二级结构，3dRNA可以构造这些二级结构的二级结构树。二级结构树中的每个节点代表一个二级结构元素，这些二级结构元素包括螺旋，发卡环，内环，凸环和多分枝环。这样，根据二级结构树，可以将一个完整的DNA分解成为很多二级结构元素。这样我们可以将二级结构元素对应的三级结构提取出来放入我们的模板库中。

在构造完了模板库之后，我们就可以来预测D-DNA的三级结构了。给定一条DNA的序列以及它的二级结构，我们首先构造它的二级结构树，然后遍历树中的所有二级结构元素，在遍历的过程中，我们会从模板库中查找每个二级结构元素所对应的三级结构模板。如果在模板库中找不到某个二级结构元素对应的三级结构模板，那么就需要来构造模板。如果这个二级结构元素是螺旋，那么就用上面介绍的构造标准螺旋的方法构造一个螺旋来作为三级结构模板；如果这个二级结构元素是环，那么就用距离几何方法构造一个三级结构模板。这样在遍历完成的时候，每个二级结构元素都会有一个与之匹配的三级结构模板。然后再一次遍历二级结构树，在遍历的过程中，每经过一个节点（二级结构元素），就将它的三级结构模板与它的父节点的三级结构模板拼接在一起。这样在遍历完整个二级结构树的时候，所有的三级结构模板就被组装在了一起，成为了一个完整的DNA的结构。

由于PDB数据库中所存有的非螺旋的D-DNA非常的少，所以模板库中的D-DNA的三级结构模板也非常的少，因此对于某个二级结构元素找不到对应的三级结构模板的情况非常常见，这时就需要用距离几何方法来构造一个三级结构模板。由于距离几何方法只考虑了几何的效应，而没有考虑物理的效应在里面，所以用这种方法构造的模板往往不够精确，因此用这种构造出来的模板所组装成的完整的DNA的结构只能够作为一个初始结构，需要通过后续的蒙特卡罗模拟的方法来对其进行优化。

1. T-DNA和Q-DNA

与D-DNA不同，对于三链螺旋和四链螺旋类的结构，我们并不是采用组装的方式，而是先用上面介绍的构造标准螺旋的方法构造一个很长的标准的三链或四链螺旋作为起始结构，然后去掉其中多余的残基，从而让起始结构的残基数与需要预测的序列的残基数一致。这样得到的起始结构会存在很多断链的现象，不过可以保证所有的三碱基配对的结构与天然态一致，然后对这个起始结构进行分子模拟，就可以修复所有的断链。

### 维持二级结构不变的蒙特卡罗分子模拟

我们进行分子模拟的原因来自于3个方面，

1. 缺少模板

由于已经解析出晶体结构的ssDNA的数量非常少，导致能够提取出来的模板就非常的稀少。对于D-DNA，我们采用的是模板组装的方法，由于模板的数量非常稀少，因而常常碰到找不到SSE的模板的情况。在这种情况下，3dRNA将会用距离几何的方法为SSE自动构造一个模板。由于DG方法只考虑了两两原子之间的距离，而没有考虑其它的诸如键角，主链扭角，碱基堆积以及碱基配对等效应，所以用DG直接造出来的模板往往与天然态结构存在着较大的差距。这个问题可以通过分子模拟得到解决，因为分子模拟所用的能量函数包含了键长，键角，主链扭角，碱基堆积，碱基配对以及空间冲突等项，可以将预测的结构慢慢拉到与天然态比较接近的程度。

1. 假结以及三级相互作用

对于含有假结或者三级相互作用的duplex类的ssDNA，在组装的过程中没有办法考虑假结以及三级相互作用的影响，因此组装完成后的结构极有可能不满足假结以及三级相互作用的约束。通过分子模拟，可以将预测的结构拉到满足符合这些约束的状态。

1. 初始结构中的断链

不管是D-DNA，T-DNA还是Q-DNA，尽管它们会符合给定的二级结构的约束，但是它们的初始的三级结构当中都很有可能会存在断链的现象，这时就需要动态的分子模拟，来修复所有的断链的情况。

由于全原子的分子模拟非常花费时间，并且需要辅以一个非常复杂的势能函数。而如果采用一个残基用一个虚原子来表示的粗粒化模型，就很难考虑碱基堆积和碱基配对的作用。因此我们采用了一个用三个虚原子来表示一个残基的模型，这三个虚原子分别对应主链上的P原子，糖环上的C4’原子以及碱基的中心。采用这种模型可以很方便的考虑碱基堆积和碱基配对的作用，并且进行模拟的速度也比全原子要快很多。

在蒙特卡罗模拟的每一步，我们随机选取一个移动元素，然后对其进行平移或旋转，再根据一个势能函数进行判断是否接受更新。

其它的方法在每一步选取的移动元素要么是单个的原子，要么是固定长度的一段残基片段。例如iFoldRNA每一步移动一个虚原子（？），FARFAR每一步选取3个残基长度的片段进行替换。而我们的方法3dRNA每一步选择的移动元素既可能是单个的残基，也可能是某个SSE包含的所有的残基。对于双链螺旋类的DNA，每一步选择的移动元素要么是某个双链螺旋上的所有残基，要么是单个的并且处于环上的残基（也就是说，这个残基不与其它的残基形成配对）。对于三链螺旋类的DNA，每一步选择的移动元素要么是某个三链螺旋上的所有残基，要么是单个的不与其它的残基形成配对的残基。很明显，在这样的采样过程中，我们将不会破坏双链或是三链螺旋的结构，这样可以保证不破坏整体的二级结构。与别的模拟算法比起来，这样做的好处是可以大大的节省采样的时间，不用花很多时间用在形成二级结构上，同时也会避免在模拟的过程中破坏二级结构（例如双链螺旋的两条链打开了）。

我们构建了一个势能函数用以在模拟的每一步判断接受这个移动或是拒绝这个移动。这个势能函数由六个部分组成：



其中，



vb表示虚键，这是因为优化是基于残基级别的，所以最小的采样单元是残基，虚键表示的是残基与残基之间的虚拟的键。是与虚键长相关的能量， 是与虚键角相关的能量，是与主链虚扭角相关的能量，是与约束相关的能量。*kl, ka, kt, kc*分别是相关的参数，*l0, a0, t0, c0*分别是标准键长，标准键角，标准二面角，指定的约束。

和是3dRNAscore [[30](#_ENREF_30)] 中的能量项。

### 打分函数

由于研究人员进行结构预测的目的往往是得到具有代表性的几个结构，而并不需要获得大量重复的结构。因此，在预测了大量的结构之后，3dRNA v2.0可以自动地对这些结构进行聚类。3dRNA v2.0采用的聚类算法是k-means算法。使用者可以自行设定所需要预测的结构数N与聚类数k，如果不设定，我们将默认只产生1个结构，也就是说N与k均为1。

最后，我们采用3dRNAscore[[30](#_ENREF_30)]对所有产生的结构模型进行打分，以供使用者参考。打分值越低，结构越接近天然态。3dRNAscore是一个基于在全原子层面并结合了两两原子之间的距离与主链二面角的统计势的打分函数，可以较为精确的区分RNA分子与天然态的接近程度，我们在文献[[30](#_ENREF_30)]中将3dRNAscore与其它几种打分函数进行了比较。

### 网站

In order to facilitate the using of our method, we provide an online web server at <http://biophy.hust.eud.cn/3dDNA>.

1. 2D prediction

The user just needs to select an ssDNA type (D-DNA or T-DNA or Q-DNA) and then input the sequence. After the submission, the results would show below after a while. If you wanted to get more suboptimal results, just increase the cutoff in the advanced options. The results of base-pair type ssDNA are represented with the dot-bracket form, whereas the results of base-tuple and G-quadruplex can’t be represented with the dot-bracket form. What is more, the existing RNA/DNA 2D structure representation programs are not able to give an intuitive graphical view of base-triple and G-quadruplex type ssDNA 2D structure.

1. 3D prediction

用户可以选择只输入序列，然后通过预测二级结构来自动获得一个二级结构，然后就可以提交来预测三级结构；也可以自己提供二级结构，然后直接预测三级结构。进行提交后，将会进入到预测的结果的页面，每5秒刷新一次，在等待一段时间之后，预测的结果将会显示在结果页面上。用户可以在高级设置中选择是否进行打分，是否用Amber进行全原子的优化等等。

## 结果与分析

### T-DNA以及Q-DNA的二级结构预测

由于T-DNA与Q-DNA的数目非常少，我们挑选了4个T-DNA与4个Q-DNA来分别预测它们的二级结构与三级结构。附表1给出了这8个DNA的二级结构的信息。

对于预测的二级结构，我们计算了STY，PPY与MCC，其中







TP是预测的二级结构中的三碱基对或四碱基对当中为真的数目，FP是预测的三碱基对或四碱基对当中为假的数目，FN是没有被预测出来的天然态二级结构中的三碱基对或四碱基对的数目。

MCC的精确定义[[31](#_ENREF_31)]是

不过由于三碱基对或四碱基对的TN很难计算，所以我们采用的是MCC的近似计算。

从表1可以看出4个T-DNA的二级结构的预测精度的平均STY，PPV和MCC都是0.933。其中，1AT4和134D的二级结构完全预测正确，1WAN和1GN7都只有1个三碱基对预测错误。这说明我们的方法可以比较精确的预测T-DNA和Q-DNA的二级结构。

### T-DNA以及Q-DNA的三级结构预测

我们预测了4个T-DNA与4个Q-DNA的三级结构，其中，对每个DNA我们预测了5个结构。如表2所示，所有这些结构的平均RMSD为3.22 Å。其中，1AT4的预测精度可以达到1.80 Å。1WAN的预测精度最大，不过也只有4.22 Å。这4个T-DNA的预测精度比较好，主要是因为它们的长度都比较短，最长的也只有32 nt，不过由于PDB数据库中T-DNA的结构非常少，这是我们挑选的最长的一个T-DNA。

### D-DNA的三级结构预测

PDB数据库中的DNA大部分都是D-DNA，并且大部分D-DNA都是非常标准的A螺旋，只有一小部分D-DNA具有与RNA类似的茎-环结构。我们从中挑选了15个具有不同类型的二级结构和长度的D-DNA，然后通过预测它们的三级结构来检测3dRNA的预测能力。在这15个DNA中，有7个发卡环，3个内环，2个多分枝环以及3个假结。长度最长的是1SLS，有50个核苷酸，最短的是1LA8，只有13个核苷酸。

对于每个DNA，我们都预测了5个结构。从表3可以看出，这15个DNA的预测结果的平均RMSD是7.62 Å。其中1JVE，1QE7和2VAI三个DNA的预测精度最好，它们的RMSD只有4 Å多。这3个DNA都是发卡环。1JUA，1SLS，3GBI和3Q5C这4个DNA的预测精度在所有的DNA当中最差，它们的RMSD为10-12 Å。这4个DNA中1JUA，1SLS和3Q5C都是假结，3GBI是一个长度为42的发卡环。对于其它的8个包括发卡环，内环和多分枝环的DNA，它们的RMSD均为5-9 Å。这些结果说明3dRNA对于非常简单的发卡环DNA，可以预测出较为精确的结构，而对于含有假结的DNA预测的精度就偏低。

与用3dRNA对RNA进行的三级结构预测[[4](#_ENREF_4)]相比，对D-DNA的三级结构预测的精度要明显比较差。这主要是由于PDB数据库中，非螺旋的D-DNA的数量要远远少于非螺旋的RNA 的数量，从而导致D-DNA的三级结构的模板要远远少于RNA的三级结构模板。因此预测程序主要是靠蒙特卡罗模拟来预测D-DNA的结构，而3dRNA的蒙特卡罗模拟预测的结构没有用模板组装来预测的结构精确，所以最终对D-DNA的预测的结果的精度要差于对RNA的三级结构预测的结果的精度。

# RNA二级结构预测

## 引言

## 原理与方法

### 将DI值转换为二级结构

DCA分析计算出来的DI值总共有L\*(L-1)/2个，我们首先将DI值从大到小进行排序，然后取前n个进行二级结构预测。为了研究n对二级结构预测精度的影响，我们将n分别取，，，L，然后进行预测。

我们首先将L个碱基的位置上都标为’.’。然后从最大的DI值开始，每次读取一个DI值，将这个DI值对应的两个碱基位置上分别标记为‘(’和‘)’。如果在某一次标记的过程中，某一个碱基已经被标记过，那么就放弃这次标记。像这样处理完所有的DI值后，就得到了二级结构的点括号形式。举例来说，在第N步，会读取第N大的DI，如果它所对应的两个碱基是i和j(i<j)，那么就将碱基i标记为’(’，将碱基j标记为’)’。在第N+1步，第N+1大的DI所对应的两个碱基是j和k，由于j已经被标记为’)’，所以就放弃这次标记。这么做的理由是DI值更大的碱基配对比DI值小的碱基配对更有可能是真阳。

如果在某一步标记的过程中，将两个碱基分别标记为’(’和’)’会引起歧义，那么就将它们标记为’[’和’]’。如果将两个碱基分别标记为’[’和’]’会引起歧义，那么就将它们标记为’{’和’}’，以此类推。我们使用的符号是’(’，’)’，’[’，’]’，’{’，’}’，’<’，’>’，’A’，’a’，’B’，’b’，’C’，’c’...等等。会引起歧义是指类似于下面的情形：如果序列Si+1,j-1的二级结构已经被标记为

’((((’，

那么i和j就不应该标记为’(’和’)’，因为如果这样标记的话，就成了j-1和j配对，

’((((()’，

这时应该将i和j分别标记为’[’和’]’

’[((((]’，

### 碱基对的生长与湮灭

在将DI转换为点括号形式的二级结构之后，如图Figure【】，我们会发现有一些碱基配对是假阳，因为这两个碱基不可能形成配对。不可能配对是指这两个碱基不属于A-U,G-C,G-U等。对于这种情况，我们将这两个碱基重新标记为’.’和’.’，我们将这种情况称之为碱基对的湮灭。对于另外一些碱基配对，例如i和j，如果i+1和j-1可以形成配对，那么就将i+1和j-1标记为配对，这种情形称之为碱基对的生长；如果i-1和j+1可以形成配对，那么就将i-1和j+1标记为配对，这也是碱基对的生长；如果这两个都不能形成配对，那么仍然取消i和j的配对，这种情况也属于碱基对的湮灭。

这种碱基对的生长与湮灭的方法的原理在于，在真实的RNA的二级结构的形成过程中，由于分子的热运动会导致碱基可能会与不同的碱基形成配对，如果与这两个碱基邻近的两个碱基也能形成配对，就会在这两个碱基对之间形成堆积效应，从而将这两个碱基配对稳固下来，这就相当于是碱基对的生长，因为如果周围的碱基能够形成配对与之形成堆积效应，碱基对就会一直生长下去；相反，如果这个碱基对两边的碱基不能形成配对，由于热运动就会导致这个碱基对被破坏掉，这就是碱基对的湮灭，这两个碱基会分别去寻找其它的碱基以形成更稳固的配对。

从Table【】可以看出经过这一步，会大大的提升STY和PPV以及MCC，这证明了这种方法的有效性。

### 删除短发卡

如图【】，有些比较短的发卡环是假阳的，这种发卡环的特点是螺旋部分不超过两个碱基配对，环的长度小于6。这一步将这种发卡环都去掉，从Table【】可以看出经过这一步，MCC和STY没有明显的变化，但是PPV大大的提升了。

### 预测最终的二级结构

从Table【】可以看出，经过了前3步之后，仍然存在很多假阴，从点括号的形式来看就是存在很多比较长的环区，这在真实的二级结构中是不可能存在的，因为很长的环区的柔性特别大。所以我们在这一步用动态规划算法来预测这些环区的二级结构。我们采用的是与mfold一样的自由能数据。

与之前的算法不同，我们并不是对整条序列用动态规划算法进行二级结构的预测，因为对于lncRNA来说，对整条序列进行动态规划算法太花费时间。在这一步，我们开发了一种新的迭代动态规划算法。由于我们已经预测出了相当一部分的螺旋区，这种迭代算法可以利用这些预测的信息来减少时间复杂度。

我们首先对序列中的每个发卡环的环区用动态规划算法进行预测，待预测完了之后，就将这整个发卡环（包括螺旋区）用XXXX来代替。然后再次对新的序列中的每个发卡环的环区用动态规划算法进行预测，然后再将预测好的发卡环用XXXX来代替，如此往复，直到最后的序列变成XXXX。

这样做的原理是：在真实的RNA的二级结构的形成的过程中，在形成了一个发卡环后，这个发卡环的5’端和3’端又靠在一起了，并且它们形成了配对，所以别的碱基不可能再与它们形成配对，这对应在算法中就是将它们标记为XXXX，然后剩下的序列再来形成二级结构。

从Table【】可以看出经过这一步，STY得到了极大的提升。

真实的二级结构往往会形成比较短距离的配对。

## 结果与分析

# 3dRNA与NSP网站搭建

## 引言

## 原理与方法

## 结果与讨论

# 全文总结与工作展望

## 本文工作总结

本文工作总结。

## 下一步工作展望

下一步展望。

# 致 谢

时光荏苒，白驹过隙，攻读博士学位的四年转瞬即逝。在即将结束求学生涯，迈入人生新的征程之际，回首走过的这四年，思绪万千。四年来，从老师与同学那里收获了太多的知识、太多的爱、太多的友谊。这些是我一辈子最宝贵的人生财富，必将激励和支撑着我奏响更加绚丽的人生乐章。对于所有帮助和关心过我的人，只能向你们道一声“谢谢”，但这份感激之情却永远铭记于心。

首先要衷心感谢我的导师肖奕教授。本论文是在肖老师的精心指导下完成的，在论文的选题、开题、研究、撰写和修改的每一个环节，无不倾注着两位老师大量的心血。肖老师渊博的学识、坦荡的胸襟、敏锐的洞察力、严谨的治学态度、平易近人的为师之道都深深地影响着我，并将使我终生受益。四年来，两位老师不仅传授了我做学问的方法和态度，还教授了我做人的准则和道理。孔子说“三十而立”，就是一个人在三十岁时要确立自己做人的原则。我想我在两位老师的直接影响下，已经形成了自己的做人原则，那就是真诚，勤奋，谦虚，谨慎。未来的人生之旅，漫长而又坎坷，但有两位老师教授给我的知识与方法，我充满信心与憧憬，必将追求更高，成就一番事业，不辜负老师的殷切之望。

感谢物理学院刘士勇教授、陈长军教授、段献忠教授对我的启发与指导；感谢电气学院研究生科与研究生工作组的所有老师，感谢你们多年来的帮助与关怀。

非常幸运的是，四年来我一直生活在一个团结友爱、朝气蓬勃的和谐大家庭中。无论在生活上还是在学习上大家都给予了我最无私的帮助。他们是赵蕴杰、龚洲、黄阳玉、李昊田、王军、王二成、张小利、何小玲、童晓雪、郑近芳、相建金、张仪、陶鹏、毛康昆、洪旭、谢娟、王艳杰、满建芬、刘鹏宇，我会永远记住你们，记住那片属于我们的天地—西九楼519室。还要感谢在西一区29栋401一同朝夕相处四年的室友们，他们是撖奥洋、马杰、陈卫民、何耀耀。与你们在一起的生活是美好的、快乐的，值得我用一辈子去珍藏！

感谢国家自然科学基金委员会与湖南省电力公司对我所做工作的资助，没有你们的支持，就无法完成本文。

感谢含辛茹苦养育了我的父母，一直以来你们都在背后默默地支持和鼓励着我，你们给予了我最无私的爱。当你们渐渐老去时，我发誓要努力工作，让你们安度一个幸福的晚年。感谢我的兄长王东光，无论是物质还是精力，你都为我付出了太多，你向我诠释了“兄弟”最好的定义。感谢我的姐姐、姐夫，还有那可爱的小外甥女，我们一起努力展现了一个幸福美满的和谐之家。

感谢我的女友张栋艳女士，谢谢你一直以来默默的付出，你给予了我最大的理解与包容。在我心中，你永远是最美的！

还有许多需要感谢的人，族繁不及备载。尽管在此不能一一列出你们的大名，但多年来你们对我的关心与照顾，我将永远铭记于心。

即将毕业，踏上一段新的旅程，心中有喜悦、有感慨，更有不舍。仔细思量父母家人的恩情、老师的教诲、同学的友谊，心存感激，唯有从谦卑做起，从感恩做起，在今后的学习工作中，以实际行动回报你们对我的爱，回馈学校、回馈祖国，回馈社会。

王剑

2017年4月30日于喻园

# 参考文献G

[1] C. Massire, E. Westhof, MANIP: an interactive tool for modelling RNA, J Mol Graph Model, 16 (1998) 197-205, 255-197.

[2] J. Frellsen, I. Moltke, M. Thiim, K.V. Mardia, J. Ferkinghoff-Borg, T. Hamelryck, A probabilistic model of RNA conformational space, PLoS Comput Biol, 5 (2009) e1000406-e1000406.

[3] F. Jossinet, T.E. Ludwig, E. Westhof, Assemble: an interactive graphical tool to analyze and build RNA architectures at the 2D and 3D levels, Bioinformatics, 26 (2010) 2057-2059.

[4] Y. Zhao, Y. Huang, Z. Gong, Y. Wang, J. Man, Y. Xiao, Automated and fast building of three-dimensional RNA structures, Scientific Reports, 2 (2012) 734.

[5] M. Parisien, F. Major, The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data, Nature, 452 (2008) 51-55.

[6] S. Cao, S.-J. Chen, Physics-Based De Novo Prediction of RNA 3D Structures, The Journal of Physical Chemistry B, 115 (2011) 4216-4226.

[7] X.J. Xu, P.N. Zhao, S.J. Chen, Vfold: A Web Server for RNA Structure and Folding Thermodynamics Prediction, Plos One, 9 (2014).

[8] M. Popenda, M. Szachniuk, M. Antczak, K.J. Purzycka, P. Lukasiak, N. Bartol, J. Blazewicz, R.W. Adamiak, Automated 3D structure composition for large RNAs, Nucleic Acids Res, 40 (2012) e112.

[9] P. Kerpedjiev, C.H. Zu Siederdissen, I.L. Hofacker, Predicting RNA 3D structure using a coarse-grain helix-centered model, RNA, 21 (2015) 1110-1121.

[10] J. Zhang, Y. Bian, H. Lin, W. Wang, RNA fragment modeling with a nucleobase discrete-state model, Phys Rev E Stat Nonlin Soft Matter Phys, 85 (2012) 021909.

[11] M. Rother, K. Rother, T. Puton, J.M. Bujnicki, ModeRNA: a tool for comparative modeling of RNA 3D structure, Nucleic Acids Res, 39 (2011) 4007-4022.

[12] C. Schudoma, P. May, D. Walther, Modeling RNA loops using sequence homology and geometric constraints, Bioinformatics, 26 (2010) 1671-1672.

[13] R. Das, D. Baker, Automated de novo prediction of native-like RNA tertiary structures, Proc Natl Acad Sci U S A, 104 (2007) 14664-14669.

[14] R. Das, J. Karanicolas, D. Baker, Atomic accuracy in predicting and designing noncanonical RNA structure, Nature Methods, 7 (2010) 291-294.

[15] J.A. Cruz, M.F. Blanchet, M. Boniecki, J.M. Bujnicki, S.J. Chen, S. Cao, R. Das, F. Ding, N.V. Dokholyan, S.C. Flores, L. Huang, C.A. Lavender, V. Lisi, F. Major, K. Mikolajczak, D.J. Patel, A. Philips, T. Puton, J. Santalucia, F. Sijenyi, T. Hermann, K. Rother, M. Rother, A. Serganov, M. Skorupski, T. Soltysinski, P. Sripakdeevong, I. Tuszynska, K.M. Weeks, C. Waldsich, M. Wildauer, N.B. Leontis, E. Westhof, RNA-Puzzles: a CASP-like evaluation of RNA three-dimensional structure prediction, RNA, 18 (2012) 610-625.

[16] Z. Miao, R.W. Adamiak, M.-F. Blanchet, M. Boniecki, J.M. Bujnicki, S.-J. Chen, C. Cheng, G. Chojnowski, F.-C. Chou, P. Cordero, J.A. Cruz, A.R. Ferré-D'amaré, R. Das, F. Ding, N.V. Dokholyan, S. Dunin-Horkawicz, W. Kladwang, A. Krokhotin, G. Lach, M. Magnus, F. Major, T.H. Mann, B. Masquida, D. Matelska, M. Meyer, A. Peselis, M. Popenda, K.J. Purzycka, A. Serganov, J. Stasiewicz, M. Szachniuk, A. Tandon, S. Tian, J. Wang, Y. Xiao, X. Xu, J. Zhang, P. Zhao, T. Zok, E. Westhof, RNA-Puzzles Round II: assessment of RNA structure prediction programs applied to three large RNA structures, RNA, (2015).

[17] F. Ding, S. Sharma, P. Chalasani, V.V. Demidov, N.E. Broude, N.V. Dokholyan, Ab initio RNA folding by discrete molecular dynamics: from structure prediction to folding mechanisms, RNA, 14 (2008) 1164-1173.

[18] H.M. Martinez, J.V. Jr. Maizel, B.A. Shapiro, RNA2D3D: a program for generating, viewing, and comparing 3-dimensional models of RNA, J Biomol Struct Dyn, 25 (2008) 669-683.

[19] K. Rother, M. Rother, M.L. Boniecki, T. Puton, K. Tomala, P.L. L Ukasz, J.M. Bujnicki, Template-based and template-free modeling of RNA 3D structure: Inspirations from protein structure modeling, (2012) 67--90.

[20] P. Ulc, F. Romano, T.E. Ouldridge, J.P.K. Doye, A.A. Louis, A nucleotide-level coarse-grained model of RNA, The Journal of chemical physics, 140 (2014) 235102.

[21] Y.Z. Shi, F.H. Wang, Y.Y. Wu, Z.J. Tan, A coarse-grained model with implicit salt for RNAs: predicting 3D structure, stability and salt effect, J Chem Phys, 141 (2014) 105102.

[22] C. Laing, T. Schlick, Computational approaches to RNA structure prediction, analysis, and design, Curr Opin Struct Biol, 21 (2011) 306-318.

[23] A.Y. Sim, P. Minary, M. Levitt, Modeling nucleic acids, Current opinion in structural biology, 22 (2012) 273--278.

[24] M.H. Bailor, X. Sun, H.M. Al-Hashimi, Topology links RNA secondary structure with global conformation, dynamics, and adaptation, Science, 327 (2010) 202-206.

[25] W. Kabsch, A discussion of the solution for the best rotation to relate two sets of vectors, Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography, 34 (1978) 827--828.

[26] K. Darty, A. Denise, Y. Ponty, VARNA: Interactive drawing and editing of the RNA secondary structure, Bioinformatics, 25 (2009) 1974.

[27] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The Protein Data Bank, Nucleic Acids Res, 28 (2000) 235-242.

[28] S. Zakov, Y. Goldberg, M. Elhadad, M. Ziv-Ukelson, Rich Parameterization Improves RNA Structure Prediction, Journal of Computational Biology, 18 (2011) 1525-1542.

[29] X.J. Lu, W.K. Olson, 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures, Nucleic Acids Res, 31 (2003) 5108-5121.

[30] J. Wang, Y. Zhao, C. Zhu, Y. Xiao, 3dRNAscore: a distance and torsion angle dependent evaluation function of 3D RNA structures, Nucleic Acids Res, 43 (2015) e63.

[31] B.W. Matthews, Comparison of the predicted and observed secondary structure of T4 phage lysozyme, Biochimica et Biophysica Acta (BBA)-Protein Structure, 405 (1975) 442--451.

# 附录1 攻读博士学位期间发表的主要论文

1. 王阳光, 游大海, 徐天奇, 等. 微机保护装置中一种基于DSP的嵌入式以太网接口设计[J]. 电气应用, 2007, 26(4): 81-85.
2. Wang yangguang, Yin xianggen, You dahai, et,al. Development of wide area current differential protection IED based on IEC 61850[C]. Proceeding of 2008 IEEE/PES Transmission and Distribution Conference & Exposition, Chicago, IL, USA,2008: 833-841.

# 附录2 博士生期间参与的课题研究情况

* 1. 中国科学技术部国家级火炬计划项目：光纤电力互感器及发（变）电站自动化设备（No.2003EB041150）。