RNA 二级结构预测方法综述

邹 权, 郭茂祖, 张涛涛
(哈尔滨工业大学计算机科学与技术学院, 黑龙江哈尔滨 150001)

摘 要: RNA 二级结构预测是计算分子生物学中的一个重要领域. 本文介绍了 RNA 二级结构的预测方法, 包括该问题的数学模型、主要算法思想以及每种算法对应的软件. 在 tRNA 和 RNase P RNA 数据库中随机选取了几组样例对目前主要的7种软件进行测试, 同时对每种软件的优缺点进行了详细比较. 实验证明, 当存在同源序列时, Pfold 的效果优于其它软件. 最后, 在总结分析现有算法的基础上探讨了该领域进一步的研究方向.

关键词: RNA 二级结构预测; 最小自由能; 比较序列分析; 假结
中图分类号: Q811 文献标识码: A 文章编号: 0372-2112 (2008) 02-0331-07

A Review of RNA Secondary Structure Prediction Algorithms

ZOU Quan, GUO Mao-zu, ZHANG Tao-tao
(*School of Computer Science and Technology, Harbin Institute of Technology, Harbin, Heilongjiang* 150001, *China*)

Abstract: RNA secondary structure prediction is one of the most important fields in computational molecular biology. The method to predict RNA secondary structure is introduced, including the mathematic models, main algorithms and softwares. Then seven main softwares are tested and compared detailedly with some groups of tRNA and RNase P. It showed that Pfold performed better than others when several homologous RNA sequences were predicted. At last some vital aspects that may be conducted in the future investigations are discussed.

Key words: RNA secondary structure prediction; minimum free energy; comparative sequence analysis method; pseudoknot

1 引言

RNA(脱氧核糖核酸)是生物系统内最为重要的分子之一, 它在生物体内行使多种功能, 尤其是在 HIV 等病毒体中, 遗传信息由 RNA 而不是 DNA 携带[1]. 2006 年诺贝尔生理或医学奖授予美国斯坦福大学医学院的安德鲁·菲尔(Andrew Fire)和麻省理工大学医学院的克雷格·梅洛(Craig Mello), 以表彰他们发现了双链 RNA 引发的基因沉默, 这也使国际上对 RNA 在生物信息上的相关研究升温. 生物分子的功能通常和结构有紧密的联系, 因此越来越多的研究人员开始关注 RNA 的二级结构和三级结构.

目前已经获得大量的 RNA 的一级结构信息, 但是用实验的方法(X 射线晶体衍射和核磁共振)来确定生物分子的具体的三维空间结构花费高、难度大, 而且并不是对所有分子都有效[2], 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. 因此就有必要利用生物信息学的手段, 加之对已有生物分子结构和功能特性的认识, 通过计算机模拟和计算来"预测"出这些信息. 这样就可以用较低的成本和较快的时间获得具有一定可信度的结果. 生物信息研究者认为: RNA 和蛋白质的三级结构很难通过一级序列直接得到, 预测二级结构是获取三级结构的必经之路[1].

预测 RNA 二级结构的本质就是找出一级序列的各个位点之间形成的配对关系. 对于一个给定的 RNA 序列, 如果按照 Watson-Crick 规则进行配对, 一个序列中可能出现很多茎区, 其中只有部分茎区是真实的. 由于 RNA 字母表大小只有 4, 会巧合出现很多"冗余茎区", 这些冗余茎区一般会和真实茎区不相容(incompatible). 在所有可能出现的茎区的集合中排除冗余茎区, 找出真实茎区组成的子集就是 RNA 二级结构预测的主要内容.

RNA 二级结构中茎区有时会形成一种特殊的位置关系, 称之为假结(pseudoknot). 对于假结可以如下定义: 在一段 RNA 序列中, 如果存在四个位点的碱基 i, i', j, j' ($i<i'<j<j'$)满足 i 和 j 配对且 i' 和 j' 配对,

收稿日期: 2007-05-22; 修回日期: 2007-08-30
基金项目: 国家自然科学基金(No. 60741001, No. 60761001, No. 60671011); 国家863高技术研究发展计划(No. 2007AA01Z171); 黑龙江省杰出青年科学基金(No. JC200611); 黑龙江省自然科学重点项目(No. ZJG0705); 哈尔滨工业大学校基金(No. HIT. 2003. 53)

那么称碱基对(i, j)和(i', j')形成了假结. 在序列长度小于 100 的简单的 *RNA*(如 *tRNA*)中, 一般不存在假结; 在较长较复杂的 *RNA*(如 *RNase P*)中, 假结的个数也比较少, 一般只有一两个. 因此在早期的 *RNA* 二级结构预测研究中, 一般不考虑假结. 假结数量虽然少, 但通常在功能上起重要作用[3], 因此越来越多的研究者[4~6]开始研究包含假结 *RNA* 的二级结构预测. 目前假结预测问题是 *RNA* 二级结构预测的难点和重点.

2 RNA 二级结构预测的研究现状

RNA 二级结构预测问题, 根据不同的情况衍生了两种不同的预测思路. 在有些情况下, 生物研究者会提交一个 RNA 序列, 他们想知道该序列能否折叠成稳定的二级结构, 如果能的话在什么位置上会出现茎区、凸环或内环等模体(motif). 上世纪 70 年代, Tinoco 等[7]提出了最小自由能模型, Zuker 等针对该模型使用动态规划的方法来寻找最优结构[8,9]. 在另外一些情况下, 生物研究人员得到了一系列同源 RNA 分子序列, 想通过比较并分析多个同源序列, 然后确定该组同源序列的一致结构, 进而确定每一个 RNA 的二级结构. 基于这种思想的方法被称为比较序列分析法(comparative sequence analysis method). 研究和大量实验表明, 比较序列分析法预测的结果准确率更高[10].

实验预测的准确率通常由两个参数确定, 即敏感性(sensitivity)和特异性(specificity), 在许多文献中特异性也被称为选择性(selectivity)[11]. 目前衡量一种方法预测效果时通常使用这两个参数, 不过敏感性和特异性通常会出现此消彼涨的现象, 因此又引入了另一个参数 *MCC*(马休兹相互作用系数)来均衡它们. 下面依次详细介绍这些问题.

2.1 基于最小自由能的方法

当没有任何先验知识, 只给定了 RNA 的一级序列时, 预测 RNA 的二级结构一般采用最小自由能模型. 该模型假定真实的 RNA 会折叠成一个具有最小自由能的二级结构. 而二级结构中的每段模体(motif)都有相应的自由能计算方法, 一般茎区的自由能为负值, 环区自由能为正, 茎区越长其自由能越小. 因此可以近似的认为, 配对的碱基使自由能降低, 没有形成配对的碱基使自由能升高. 基于这种思想, Nussinov 使用了一种简单的动态规划方法[12], 用来寻找一个 RNA 序列在无假结的情况下形成最多配对的折叠方法. 该算法只考虑了配对碱基对自由能的影响, 忽略了碱基对之间的自由能影响和环区的自由能, 因此它的时间复杂性较低, 为 $O(n^3)$, 其中 n 为 RNA 一级序列的长度. 由于 Nussinov 算法假定 RNA 二级结构中所有碱基对的能量是相互独立的, 因此它输出的碱基对通常是不连续的, 不能够形成茎区.

自由能的大小不是和配对碱基的个数呈线性关系, 它与配对关系(A·U, G·C 还是 G·U)有关, 特别是它还受相邻碱基对的影响. 尤其是对于凸环、内环等模体, 其自由能往往与变异前相应的茎区相差很大. 基于这些考虑, Zuker, Wateman 等将模体简单分类为茎区、凸环、内环和发夹环, 针对不同的模体采用不同的自由能计算方法, 然后利用动态规划方法将模体组合, 得到一个最小自由能的二级结构[13]. Zuker 的动态规划算法假定分子的自由能等于各个模体自由能之和, 而且各个模体之间的自由能相互独立; 而 Nussinov 算法认为分子的自由能是各个碱基对的自由能之和, 而且各个碱基对的自由能是相互独立的.

Zuker 的动态规划算法考虑了凸环、内环等模体, 其中由于内环的自由能受到两个参数的影响, 因此其时间复杂性略高于 Nussinov 算法. 原始的 Zuker 算法的时间复杂性是 $O(n^4)$, 空间复杂性为 $O(n^2)$. 由于 Zuker 的自由能计算模型认为内环的大小相同时, 其自由能应该相同, 那么对于相同大小的内环可以只计算一次然后存储起来, 于是多消耗一些存储空间可以把算法的时间复杂性降低到 $O(n^3)$[14]. 进一步研究表明, 如果限定内环的大小至多为 k, 其时间复杂性可以降低到 $O(kn^2)$[15].

Rivas 和 Eddy 提出了可以预测假结的动态规划算法[16]. 除了凸环、内环等模体的影响, 递归关系中还加入了假结的自由能, 因此在时间和空间上都有消耗. 由于一个假结需要用四个参数来描述, 所以动态规划的递归关系中假结一项需要四层循环来寻找最优, 而计算动态规划表还需要两层循环, 故该算法的时间花费是 $O(n^6)$, 空间花费是 $O(n^4)$.

上述动态规划得到的结果是自由能最小时对应的二级结构. 而实验证明, 真实结构往往不是自由能最小的二级结构. 而且, 自由能迄今为止还没有完全精确的计算规则[9]. 为了解决真实结构与最小自由能结构不一致的尴尬, Zuker 等又提出了次优结构的概念[17]. 他们认为: 真实的二级结构的自由能也许不是最小的, 但也应该具有一个较小的值使其分子相对稳定. 因此可以人为设定一个阈值, 与自由能最小结构相差该阈值以内的所有二级结构都有可能是真实结构, 将它们全部输出, 送给生物研究人员再鉴定. 显然, 阈值设定的越大, 包容的二级结构越多, 因而覆盖到真实结构的概率越大, 而生物研究人员再鉴定时的花费也就越大; 而阈值设定的过小, 虽然节省了再鉴定的花费, 但却增大了漏掉真实结构的概率. 因此阈值的选择要适中. 目前由 Mathews 等开发的基于最小自由能模型的软件 RNAStructure(针对 windows 操作系统)和 Dynalign[18](针

对 Unix 和 Linux 系统)都可以由使用者自己限定阈值.

基于最小自由能模型预测单个序列二级结构的准确率还不能令人满意.当容忍"碱基对滑移"(base pair slippage)时,Mathews 等认为预测不同 RNA 的准确率可以达到 73%[9].而 Dowell 和 Eddy 在测试中指出:对于 RNase P,SRP 和 tmRNA 的预测,该模型仅能有 56%的敏感性和 46%的特异性[19].在生物实验中,通常是要处理一个数据集上的 RNA 结构,而数据集大多是一组或几组同源 RNA 序列.同源 RNA 通常具有相似的结构,针对这个特点,利用比较序列分析法可以提高预测的准确率.

2.2　基于比较序列分析的方法

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,这一点体现得尤为明显,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[20,21].基于这一点,提出了比较序列分析法预测分子结构.

比较序列分析法可以按照序列比对与结构预测的先后顺序分为 3 种.先比对后预测方法是假定了结构的保守性大于序列的保守性,这种思路的预测结果强烈依赖于多序列比对的效果,但针对保守区的多序列比对也是一个棘手的问题,目前大多使用 clustalw 等针对 SP 记分的多序列比对工具.先预测后比对的方法要求得到大量的次优结构(往往是多个局部最优),而多个次优结构中是否包含真实结构是不能确保的,并且对结构进行比对一直都是一个难题[22].结构预测与序列比对同时进行的主要是 Sankoff 算法[23],它结合序列比对和 Nussinov(最大碱基对)折叠进行循环,该算法极度消耗计算资源(时间复杂度为 $O(n^{3m})$,空间复杂度为 $O(n^{2m})$,其中 n 是序列的长度,m 是序列的数量).基于该算法的软件 Foldalign[24,25] 和 Dynalign[26] 都限制了子结构(substructure)的大小和形状.

先比对后预测 RNA 的主要软件有 Pfold[27,28] 和 Alifold[29].Pfold 使用结合了进化信息的随机上下文无关文法来预测 RNA 的二级结构.它首先需要人为写出随机上下文无关文法的规则,然后在已知结构的 RNA 数据库中进行训练,得到每条文法规则的概率.Pfold 的输入是一组比对好的同源 RNA 序列,计算每一列的概率(probabilities of columns),将列的概率进行乘积得到比对的概率(probability of an alignment),通过训练找到一颗进化树,使得比对的概率最大.然后利用进化信息和随机上下文无关文法,找到使每条序列概率最大的生成法则,即该序列对应的二级结构.由于文法规则的概率已经被学习到,Pfold 运行时间短,预测精度高.当序列较长,同源序列个数较多的时候,很适合用 Pfold 进行预测,不过基于随机上下文无关文法的 Pfold 最大的弱点是不能够预测假结.

Alifold 的目标是从 RNA 比对中计算出一致结构,它是 Zuker 算法的一个扩展.该方法首先计算一个平均能量矩阵和一个共变积分矩阵(covariation score matrix),用来增大不一致序列的罚分.然后使用一个标准的回溯过程来发现一致结构,该结构使得平均能量和共变积分之和是最优的.此算法的时间复杂性是 $O(N*n^2+n^3)$,空间复杂性为 $O(n^2)$,其中 N 为序列的个数,n 为比对后每个序列的长度.

Sankoff 算法将序列比对和结构预测一起进行,同时可以得到一个比对和一个一致结构.它使用动态规划方法得到一个碱基列表和碱基权重的最大和.从根本上讲,这是将序列比对和 Nussinov 动态规划折叠方法的一个合并[30].运用 Sankoff 思想的软件主要有 CARNAC 和 RNAStructure.

CARNAC 不像 Sankoff 算法那么费时.它在序列集合中进行双序列比对时,使用了一个过滤集合(set of filters).然后扫描序列,找出那些高度相似的区域(high similarity regions),消除锚点(anchor points).最后根据共变信息(covariation information)和锚点来选择保守茎区[31,32].而 RNAStructure[26](Linux 下称为 Dynalign)是 Sankoff 算法在双序列比对上的应用,它用"完全能量模型"(full energy model)在局部寻找低能量结构(包括多分枝环),并且比对两个 RNA 结构,通过限制两个序列中的比对位置来减少 Sankoff 算法的时间复杂性.实际上相当于在局部上使用了 Sankoff 算法,因此它不能处理较长的 RNA 序列(目前在线预测限制 RNA 序列单个长度不能超过 80nt).

MARNA 是一种比较特殊的结构构预测软件[33],它采用先结构预测后比对结构的思想.当观察不到序列保守区时,比较适合使用该方法.MARNA 在第二步比对时,采取的不是多序列比对,也不是严格意义上的结构比对,而是一种结合了结构信息的多序列比对.然而,如何完美地融合入结构信息,并且尽量得到效果理想的多序列比对至今还没有被解决.关于 MARNA 以及其他几种软件的预测效果将会在第 3 节中详细比较.

2.3　检验标准

对预测准确率的度量目前绝大多数文献使用的都是敏感性(sensitivity)、特异性(specificity)[34]和马休兹相互作用系数(Matthews correlation coefficient)[35]这三个参数.

假阴性(false negative)和假阳性(false positive)的概念通常在衡量实验结果的时候使用.在 RNA 二级结构预测中,用 TP(true positive)表示正确预测碱基对的个数;*FN*(false negative)表示真实结构中存在但没有被正确预测出的碱基对个数;*FP*(false positive)表示真实结

构中不存在却被错误预测到的碱基对个数; *TN*(true negative)表示正确预测的不配对的碱基的个数, 由于 *TN* 一般远远大于 *TP*, *FN* 和 *FP*, 所以在实际衡量中很少用到.

敏感性(*X*)指真实结构中所有的碱基对中被正确预测到的百分比; 特异性(*Y*) 指在所有预测到的碱基对中正确预测的百分比. 一般的预测方法很难两者兼顾, 总是偏向于一边, 因此用马休兹相互作用系数(*MCC*)折中衡量. 具体计算公式如下:

$$X=\frac{TP}{TP+FN} \quad Y=\frac{TP}{TP+FP}$$

$$MCC=\frac{TP*TN-FP*FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

MCC 的取值范围从-1(*TP*= *TN*=0, 完全不正确)到 1(*FP*=*FN*=0, 完全正确). 在比较 *RNA* 结构时只有在极端情况下才有 *TN*=0, 因此 *MCC* 一般取值在 0 到 1 之间[34] . 一般情况下由于 *TN* 远远大于 *TP*, *FN* 及 *FP*, 因此 *MCC* 化简后可以用敏感性*X* 和特异性*Y* 的几何平均近似代替.

但也有研究者指出: *FP* 虽然都是不希望出现的结果, 但是对 *FP* 还有轻重之分. 如果预测到碱基 *i* 和碱基*j* 形成碱基对, 而真实结构中不存在, 显然它属于 *FP*; 如果碱基 *i* 或*j* 在真实结构中与其他碱基形成碱基对或 *i* 和*j* 形成的碱基对与真实结构中的某个碱基对形成假结, 则该预测结果影响到了另外的一对碱基的预测, 这直接影响到了预测的准确率, 由此而引发的链式反应大大的降低了预测的准确程度, 因此这种情况的 *FP* 对预测的准确性更有害; 而如果在真实结构中 *i* 和*j* 都是自由碱基, 并且即使 *i* 和*j* 结合成碱基对也不会和其他的真实碱基形成假结, 那么这种预测不会导致一连串的错误, 因此这种情况在预测中对准确率的影响程度远小于上一种. 如果设 *FP* 中这种碱基对的个数是 δ, 很多研究人员更倾向于利用 *FP*-δ代替 *FP*[34] .

3 目前主要的 RNA 二级结构预测软件的比较

RNA 二级结构预测发展速度很快. NCBI 提供了大量的 RNA 序列数据, 然而提供结构的数据库并不多. 目前用来为预测算法提供测试数据的结构数据库主要有三个, 分别是 tRNA 数据库[20, 21] 、RNase P 数据库[36] 和 Gutell 实验室比较 RNA 站点, 如表 1 所示. 由于目前大多数预测算法还不能够很好地处理长度在 1knt 左右的长 RNA 分子, 因此大部分研究者都采用 tRNA 和 RNase P 来测试他们的算法或软件.

表 1 可以免费获得 RNA 二级结构的数据库列表

数据库名称	网址	说明
tRNA 数据库	http://lowelab.ucsc.edu/GtRNAdb/	长度大多在 70nt 到 80nt
RNase P RNA 数据库	http://jwbrown.mbio.ncsu.edu/RNAseP/home.html http://www.mbio.ncsu.edu/RNaseP/home.html	长度大多在 300nt 到 400nt
Gutell 实验室比较 RNA 站点	http://www.rna.icmb.utexas.edu/	有更长的 RNA, 需免费注册后才可使用

目前 RNA 二级结构预测的软件和提供在线预测的网址很多, 本文列举了 7 个主要的软件. 它们分别是: RNAfold[29] , mfold[37] , Srna[38] , CARNAC[31, 32] , MARNA[33] , Pfold[27, 28] 和 RNAStructure[18] . 这 7 种软件被公认为效果比较好, 很多 RNA 结构预测相关的论文都与它们其中的几个进行对比, 或利用某个软件预测结构用于其它方面[39, 40] . 本文使用 tRNA 和 RNase P 对这 7 种软件进行测试, 总结了各自的优点和限制(如表 2 所示), 并且对比了预测的准确率, 实验结果列表及详细数据数据下载见 http://nclab.hit.edu.cn/zq/table.htm. 其中 CARNAC 由于使用的是 Sankoff 算法, 无法处理长为 80 以上的 RNA 序列, 因此较短的 tRNA 的测试实验使用了 7 种软件, 而长度在 300nt 左右的 RNase P RNA 的测试实验只对比了 6 种.

在预测单个序列时, RNAfold 和 mfold 使用的都是动态规划计算最小自由能的方法, 因此预测效果近似. 特别是在预测 Lake Griffy B ♯41、Volunteer ESH212C 和 Pond Scum ♯26 三种 RNase P 时, 得到的结果完全相同. 在预测一组同源序列时, 基于比较分析法的 Pfold 效果明显优于基于最小自由能的 RNAfold, mfold, Srna 和 RNAStructure. 在预测 tRNA 时 Pfold 效果最好. 这是因为同源的 tRNA 结构高度相似, Pfold 准确地找到了一致结构, 使得敏感性和特异性都达到 90%以上.

而在预测 RNase P 时, 虽然平均效果 Pfold 最好(三组平均 *MCC* 值 63.57%, 在 6 种软件中最高), 但与 mfold 等基于最小自由能模型的软件相差不大. 这是由于 RNase P 的结构保守性较差, 而且提交给 Pfold 的同源序列较少(只有 4 个). 由此可见, 当存在同源 RNA 序列时, Pfold 适合进行二级结构预测; 当只有一个序列需要预测结构时, 可以考虑使用 mfold, 因为其不但准确率相对较高, 而且附加功能强大, 可以把二级结构图形化输出.

表 2　各种主要的 RNA 二级结构预测软件比较

软件名称	优点	限制	主要原理	特殊说明	软件使用情况	网址
Alifold/ RNAfold	(1)提供选择是否支持GU配对的选项 (2)提供选择茎区边缘是否支持GU配对的选项 (3)容纳错误字符 (4)可以预测单一序列;也可以预测多个序列	(1)预测单个结构时序列长不能超过300 (2)预测多个序列时,只能给出一致结构,不能预测每一个序列的二级结构 (3)预测一致结构时比对后的序列单个长度不能超过2K;总长度不能超过10K	预测单一序列依靠最小自由能模型;预测多个序列依靠比较序列分析模型	输入的是clustalw比对完的aln文件,输出的是一致结构	提供在线预测;提供Unix/Linux版本软件下载;提供界面简单的windows版本	http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi http://rna.tbi.univie.ac.at/cgi-bin/alifold.cgi
mfold	(1)可以人为设定先验知识(比如序列的哪两段子串需要形成茎区或不能形成茎区,哪段必须形成单链等) (2)支持环形RNA的预测 (3)可以设置内环/凸环的最大值 (4)可以设置内环的最大不对称值 (5)可以设置碱基对之间的最大距离 (6)每次提供多个可选择结构 (7)提供图形化界面输出	只能预测单个序列	最小自由能模型的Zuker动态规划算法	预测的是单一序列	提供在线预测和Linux版本软件	http://www.bioinfo.rpi.edu/applications/mfold/rna/form1.cgi
Srna	(1)提供反转功能(预测其逆补序列) (2)可以设置碱基对之间的最大距离 (3)可以人为设定先验知识 (4)可以图形化输出二级结构,界面友好 (5)可以提供多个可选择结构	(1)只能预测单个序列 (2)序列的长度不能超过5K (3)预测速度慢	结合统计方法的最小自由能模型	预测的是单一序列	只提供在线预测,不提供单机软件	http://www.bioinfo.rpi.edu/applications/sfold/srna.pl
CARNAC	(1)可以选择是否允许长为1的茎区存在; (2)可以图形化输出二级结构	每个序列最长为80	比较序列分析法	预测的是一组同源序列	只提供在线预测,不提供单机软件	http://bioinfo.lifl.fr/carnac/carnac.php
MARNA	该软件自带多序列比对功能,而且序列比对和结构预测中的绝大多数参数都可以由用户自行设定	(1)总长度不超过10000nt(如果20个序列,则每个序列长度不超过500;如果100个序列,则每个序列长度不超过100) (2)不能包含非法字母,比如N	先利用最小自由能原理折叠序列,然后在同源序列中进行结构比对	输入的是比对前的初始序列,要求fasta格式	提供在线预测和Linux版本软件	http://biwww2.informatik.uni-freiburg.de/Software/MARNA/index.html
Pfold	输出结果的同时提供进化树	不能预测假结	结合进化树的上下文无关文法	输入要求是比对好的多个序列	只提供在线预测	http://www.daimi.au.dk/~compbio/pfold/
RNAStructure/ Dynalign	(1)操作界面友好,功能强大 (2)可以给出良好的图形输出	输入字母表只有AGCU,其他字母或小写字母都不会被预测	Sankoff算法和动态规划算法	可以预测单一序列,也可以比较两个序列的结构	提供windows和Linux/Unix版本;不提供在线预测	http://rna.urmc.rochester.edu/

4　RNA 二级结构预测算法研究的发展方向

分子结构的预测是分子功能发现的基础,而分子功能的发现又是生物信息学的最终目标,即利用生物信息学产生的结果来指导医疗、制药、卫生等方面.因此,生物信息研究人员需要提供更加准确、更加快速的结构预测软件.而二级结构预测作为三级结构预测的基础,吸引着众多生物信息研究者去完善它.目前,RNA二级结构预测领域的主要发展方向包括以下几个方面.

(1)假结的预测.在真实结构中有些 RNA 没有假结,有些 RNA 拥有大量的假结,更多的 RNA 是只存在少量假结.目前研究人员还无法从序列层上区分到底会存在多少假结,而且假结一旦预测错误,将会连锁地引起正常茎区预测错误,因此很多研究者都回避假结.正确地预测假结是 RNA 二级结构预测中最富有挑战性的问题.

(2)最小自由能模型与比较序列分析法的结合.越来越多的研究者在运用比较序列分析时结合入最小自由能模型,来加强预测精度.目前,寻找最小自由能与比较序列分析大多是分开进行的,一方为另一方提供预处理或后处理.而如何将这两者和谐地融合在一起,还有待进一步地研究.

(3)与多序列比对、构建进化树等问题的结合.以往的计算分子生物学研究者大多认为多序列比对是手段,构建进化树和预测二级结构是目标.但自从 Pfold 将进化树作为手段来指导 RNA 二级结构预测获得了突破性成功后,更多的研究者试图寻找这三者之间的联系.每一个问题都不再单纯是手段或者是目标,将三者结合寻找生物学意义、解决生物问题将会是未来研究的重点.

(4)不稳定因素的处理.由于结构的决定因素不止是序列,还有细胞组成、碱基修饰以及转录过程等,目前还没有任何计算方法考虑这些因素.甚至还存在"核糖转换"(ribo-swithes)等现象[41,42],使得对于一个给定的序列,存在两种以上可能的结构.这时结构预测模型本身就失去了意义,因为序列和二级结构之间不再是一一映射关系.在问题层次上重新抽象模型,将是结构预测问题得到完善后进一步面临的问题.

总之,结构和功能是息息相关的;在物理方法测定结构没有得到质的飞跃之前,结构预测问题仍旧是生物学和生物信息学研究者所共同关心的问题.随着模型和算法的改进、速度和精度的提高,RNA 二级结构预测方法将会越来越多地应用到 RNA 分子的结构和功能研究领域,为分子功能的研究和预测提供有力的参考.

参考文献:

[1] Tao Jiang, Ying Xu, Michael Q. Zhang. Current Topics in Computational Molecular Biology[M]. 北京: 清华大学出版社, 2002.

[2] Furtig B, Richter C, Wohnert J, Schwalbe H. NMR spectroscopy of RNA[J]. Chembiochem, 2003, 4(10): 936—962.

[3] E Ten Dam, K Pleij, D Draper. Structural and functional aspects of RNA pseudoknots[J]. Biochemistry, 1992, 31(47): 11665—11676.

[4] JE Tabaska, RB Cary, HN Gabow, GD Stormo. An RNA folding method capable of identifying pseudoknots and base triples [J]. Bioinformatics, 1998, 14(8): 691—699.

[5] Xiaolu Huang, Hesham Ali. High sensitivity RNA pseudoknot prediction[J]. Nucleic Acids Research, 2007, 35(2): 656—663.

[6] David P Giedroc, Carla A Theimer, Paul L Nixon. Structure, stability and function of RNA pseudoknots involved in stimulating ribosomal frameshifting[J]. Journal of Molecular Biology, 2000, 298(2): 167—185.

[7] Tinoco I, Uhlenbeck O C, Levine M D. Estimation of secondary structure in ribonucleic acids[J]. Nature, 1971, 230(5293): 362—367.

[8] J A Jaeger, D H Turner, M Zuker. Improved predictions of secondary structures for RNA[J]. Proceedings of the National Academy of Sciences, 1989, 86(20): 7706—7710.

[9] David H Mathews, Jeffret Sabina, Michael Zuker, Douglas H Turner. Expand sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure[J]. Journal of Molecular Biology, 1999, 288(5): 911—940.

[10] Woese C. Pace N: The RNA World, Chap. Probing RNA Structure, Function, and History by Comparative Analysis [M]. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY; 1993. 91—117.

[11] Baldi P, Brunak S, Chauvin Y, Andersen C, Nielsen H. Assessing the accuracy of prediction algorithms for classication: an overview[J]. Bioinformatics, 2000, 16(5): 412—424.

[12] Ruth Nussinov, George Pieczenik, Jerrold R Griggs, Daniel J Kleitman. Algorithms for loop matchings[J]. SIAM Journal on Applied Mathematics, 1978, 35(1): 68—82.

[13] Waterman M S, Smith T F. RNA secondary structure: A complete mathematical analysis[J]. Mathematical biosciences, 1978, 42(2): 257—266.

[14] Waterman M S, Smith T F. Rapid dynamic programming methods for RNA secondary structure[J]. Advances in Applied Mathematics, 1986, 7(1): 455—464.

[15] R B Lyngsø, M Zuker, C N S Pedersen. Fast evaluation of internal loops in RNA secondary structure prediction[J]. Bioinformatics, 1999, 15(6): 440—445.

[16] Elena Rivas, Sean R Eddy. A dynamic programming algorithm for RNA structure prediction including pseudoknots[J]. Journal of Molecular Biology, 1999, 285(5): 2053—2068.

[17] Zuker M. On finding all suboptimal foldings of an RNA molecular[J]. Science, 1989, 244(4900): 48—52.

[18] Mathews D H, Turner D H. Dynalign: An algorithm for finding the secondary structure common to two RNA sequences [J]. Journal of Molecular Biology, 2002, 317(2): 191—203.

[19] Dowell R, Eddy S. Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction [J]. BMC Bioinformatics, 2004, 5(1): 71. http://www.

biomedcentral.com/1471-2105/5/71.

[20] Mathias Sprinzl, Carsten Horn, Melissa Brown, Anatoli Ioudovitch, Sergey Steinberg. Compilation of tRNA sequences and sequences of tRNA genes[J]. Nucleic Acids Research, 1998, 26(1): 148—153.

[21] Lowe TM, Eddy SR. tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence[J]. Nucleic Acids Research, 1997, 25(5): 955—964.

[22] Julien Allali, Marie-France Sagot. A new distance for high level RNA secondary structure comparison[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2005, 2(1): 3—14.

[23] Sankoff D. Simultaneous solution of the RNA folding, alignment and protosequence problems[J]. SIAM Journal on Applied Mathematics, 1985, 45(5): 810—825.

[24] Gorodkin J, Heyer L, Stormo G. Finding the most significant common sequence and structure motifs in a set of RNA sequences[J]. Nucleic Acids Research, 1997, 25(18): 3724—3732.

[25] Gorodkin J, Stricklin SL, Stormo G. Discovering common stemloop motifs in unaligned RNA sequences[J]. Nucleic Acids Research, 2001, 29(10): 2135—2144.

[26] Mathews D, Turner D. Dynalign: An algorithm for finding the secondary structure common to two RNA sequences[J]. Journal of Molecular Biology, 2002, 317(2): 191—203.

[27] Knudsen B, Hein J. Using stochastic context free grammars and molecular evolution to predict RNA secondary structure [J]. Bioinformatics, 1999, 15(6): 446—454.

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

[29] Ivo L Hofacker. Vienna RNA secondary structure sever[J]. Nucleic Acids Research, 2003, 31(13): 3429—3431.

[30] Hofacker IL, Bernhart S, Stadler P. Alignment of RNA base pairing probability matrices[J]. Bioinformatics, 2004, 20(14): 2222—2227.

[31] Touzet H, Perriquet O. CARNAC: folding families of non coding RNAs[J]. Nucleic Acids Research, 2004, 32(suppl—2): W142—145.

[32] Perriquet O, Touzet H, Dauchet M. Finding the common structure shared by two homologous RNAs[J]. Bioinformatics, 2003, 19(1): 108—116.

[33] Siebert S, Backofen R. MARNA: multiple alignment and consensus structure prediction of RNAs based on sequence structure comparisons[J]. Bioinformatics, 2005, 21(16): 3352—3359.

[34] Gardner PP, Giegerich R. A comprehensive comparison of comparative RNA structure prediction approaches[J]. BMC Bioinformatics, 2004, 5(1): 1—32.

[35] Baldi P, Brunak S, Chauvin Y, Andersen C, Nielsen H. Assessing the accuracy of prediction algorithms for classication: an overview[J]. Bioinformatics, 2000, 16(5): 412—424.

[36] Brown J W. The ribonuclease P database[J]. Nucleic Acids Research, 1999, 27(1): 314.

[37] M Zuker. Mfold web server for nucleic acid folding and hybridization prediction[J]. Nucleic Acids Research, 2003, 31(13), 3406—3415.

[38] Ding Y, Chan C Y, Lawrence C E. Sfold web server for statistical folding and rational design of nucleic acids[J]. Nucleic Acids Research, 2004, 32(suppl—1): 135—141.

[39] Christina Witwer, Ivo L. Hofacker, Peter F. Stadler. Prediction of Consensus RNA Secondary Structures Including Pseudoknots[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2004, 1(2): 66—77.

[40] Jihong Ren, Baharak Rastegari, Anne Condon, Holger H. Hoos. HotKnots: Heuristic prediction of RNA secondary structures including pseudoknots[J]. RNA, 2005, 11(10): 1495—1504.

[41] Mandal M, Breaker R. Gene regulation by riboswitches[J]. Nature Reviews Molecular Cell Biology, 2004, 5(6): 451—463.

[42] Soukup J, Soukup G. Riboswitches exert genetic control through metabolite-induced conformational change[J]. Curr Opin Struct Biol, 2004, 14(3): 344—349.

作者简介:

邹　权　男,1982 年生于黑龙江佳木斯,博士研究生.主要研究领域为 RNA 结构预测方法、MiRNA 的识别与分类算法、序列比对.
E-mail: guoer713108@163.com

郭茂祖　男,1966 年生于山东夏津,博士后,教授,博士生导师,中国人工智能学会机器学习专业委员会常委.主要研究方向为机器学习与数据挖掘、计算生物学与生物信息学、新型计算模型.
E-mail: maozuguo@hit.edu.cn

张涛涛　女,1983 年生于山东临沂,硕士.主要研究领域为遗传算法、RNA 二级结构预测和参数比对问题.
E-mail: ztt317@126.com