RNA 模式分析进展

赵英杰,王正志

(国防科技大学 机电工程与自动化学院,中国湖南 长沙 410073)

摘 要:研究表明,RNA 模式在基因表达调控方面起着重要作用. 由于 RNA 模式不仅与初级序列有关,更多的表现为高级结构 (一般为二级结构) 的保守性,所以 RNA 模式的识别比 DNA 模式的识别要复杂的多. 近十几年里,对 RNA 模式分析作了大量的计算方面的研究,包括:RNA 结构的预测、识别和已知的类型相似的 RNA 模式、在一组功能或进化相关的基因中找出共同的 RNA 模式. 这里对上述 3 个方面的计算方法的发展和研究进行了综述.

关键词:RNA 模式;二级结构;计算方法
中图分类号:Q527 文献标识码:A 文章编号:1007-7847(2006)S0-0032-05

The Development of RNA Motif Analysis

ZHAO Ying-jie, WANG Zheng-zhi

(*College of Mechatronics Engineering and Automation, National University of Defense Technology, Changsha* 410073, *Hunan, China*)

Abstract: RNA sequence motif have been found to play important roles in regulating gene expression. Because RNA motif are characterized by both primary sequence and higher order structural constraints, identification of RNA motif is more complicated than identification of DNA elements. During the last decade, many computational efforts have been made for RNA sequence analysis, including prediction of RNA structure, identification of RNA elements that are similar to known RNA elements and discovery of common RNA motif based on a group of functionally or evolutionally related genes. It's reviewed that the major methods developed for RNA sequence motif analysis in the above three areas.

Key words: RNA motif; secondary structure; computing method

(*Life Science Research*, 2006, 10(2): 032 ~ 036)

分子遗传学中基本的生物高聚物是蛋白质和核酸序列,也就是 DNA 和 RNA. 虽然真核和原核细胞的遗传信息储藏在 DNA 中,但 RNA 在基因表达中同样起着重要作用. 对于蛋白质合成来说,转录后调控是一个重要的控制点,它影响着许多生理和病理过程,包括正常的生长发育[1]、适应机能[2]及癌变反应[3]. 已经发现大量的 mRNA 模式在基因表达的转录后调控阶段起着重要作用,主要包括:转录的终止、mRNA 的定位和稳定、mRNA 的可变剪接及翻译效率[4~6]. 因此,对已知的及有待发现的 RNA 调控模式的预测是非常重要的,这将有助于我们理解调控机制,也将为药物

收稿日期:2006-04-04 修回日期:2006-05-08
基金项目:国家自然科学基金资助项目(60471003)
作者简介:赵英杰(1977-),男,山东曹县人,博士研究生,从事生物信息学研究,Tel:0731-4574991,E-mail:zhaoyingjie@nudt.edu.cn;王正志(1945-),男,上海人,国防科技大学教授,博士生导师,主要从事生物信息学、模式识别与信息处理等方面的研究.

设计提供一个潜在的靶点，以防止不正常的 mRNA 调控的发生.

研究显示，多数已知 mRNA 调控模式位于非翻译区，而非翻译区在人类和其他高等真核生物基因组中占有大多数. 多数的调控模式都是以茎环结构出现,它们在翻译过程中起的是消极作用,主要表现为阻止绑定或抑制核糖体子单元 40S 的移动,也可以通过绑定特殊的调控蛋白(通常是抑制翻译的,如铁离子调控蛋白)来实现. 一些模式是 microRNA 绑定位点，能够绑定附加的 microRNAs[7]. UTRsite——一个专业的数据库,搜集了真核生物 mRNA5′和 3′间非翻译区的功能模式[8]. 这些模式中的一部分,如 IRE(铁离子反应模式),被用于测试在 RNA 中发现一致结构的方法[9, 10]. 从一个相关的 RNA 序列集中找出保守的 RNA 模式的算法，能帮助我们发现潜在的未发现的 RNA 调控模式.

因为 RNA 模式在高级结构上比在初级序列上更保守[11],这使得采用计算方法来探测它们成为一个具有挑战性的问题，而用来识别转录因子绑定位点的模式发现方法用在这里并不合适.

本文将讨论识别已知的 RNA 模式的方法,及从单个序列或相关序列集中预测 RNA 二级结构的方法.

1　已知 RNA 调控模式的探测

从序列数据库中识别原始序列模式的算法已经有很多,最著名的程序是 GCG 软件包中 FindPattern 模块[12]. 对于 RNA 分子的模式发现算法,需要同时考虑空间结构和原始序列. 已经有了几个这样的算法,像 Palingol、PatSearch 和 RNAmotif.

Palingol[13]是一个公开的编程语言,它能描述 RNA 二级结构、扫描序列数据库. 它对结构的描述是基于螺旋集合(一个螺旋意味着一个发卡)和一系列限制条件.

PatSearch[14]是另一个模式发现程序,它允许错配或是错配数低于用户设定的一个固定阈值,它能通过 Markov 链的模拟来得到这些错配出现的统计特性. PatSearch 中用到的模式描述和正则表达规则相似.

RNAmotif[15]不仅能找出和所描述模型匹配的模式，而且能用热力学特性和序列复杂性将这些候选模式归类(序列区域的复杂性越低,它所包含的碱基类型越少，它们可能和模式以一个很高的分值匹配，不过这不是因为碱基顺序而是因为这一段的碱基组成).

和这些 RNA 模式搜索工具相关的,还有一些搜索特定类型结构 RNA 的算法,如 tRNAscan-SE[16]和 FAStRNA[17]是用来搜索 tRNA 的;CITRON[18]是用来识别 I型接触反应内含子的核酶,MIRscan[19]是用于搜索 microRNA 绑定位点.

2　单个 RNA 序列二级结构预测

单个 RNA 二级结构预测是基于最小能量模型. 1994 年,Zucker 用动态规划法来预测 RNA 全局最优二级结构，他的方法是基于最小自由能模型. Nussinov[20]改进了这种方法,他采用了最邻近能量模型，这种方法的准确性通常在 70% ~ 80%[21].

PKNOTS[22]是另一种预测优化 RNA 二级结构的动态规划算法，它包括了基于最小自由能模型的假结结构，在最糟情况下的时间复杂性是 O(N^6),空间复杂性是 O(N^4).

第 3 种方法是 NUPACK[23]，它可以预测包括(或不包括)假结的二级结构. 这个算法采用了最小能量结构、划分函数和碱基配对概率.

第 4 种方法是 Dyalign[24],它通过结合最小自由能和比较序列分析，来预测两个序列的共同低自由能结构,它用了动态规划算法,计算复杂性在时间上为 O(M^3N^3),这里 M 是两个序列中,对齐的碱基的最大距离,N 是较短序列的长度.

基于自由能模型的 RNA 二级结构预测只能处理短的序列，而 RNA 折叠通常包括中间状态,这意味着一些 RNA 分子的真实结构的自由能并不总是最小的[25].

3　共同 RNA 序列模式预测

3.1　比较研究

通常认为一致 RNA 二级结构预测和多序列比对的最可靠方法是比较研究,它能通过人工改进比对的方法来识别出多序列比对中强相关位置[26]. 但这个方法要求这些序列要有明显的序列相似性,而且也不是自动的. 它通常要求一个专家来研究比对以获得最终的结果.

3.2　随机上下文无关语法

随机上下文无关语法(stochastic context-free grammar-SCFG)是用于同时进行 RNA 一致结构预

测和全局结构比对的自动方法. 1994 年，Eddy 工作组和 Sakakibara 工作组都采用 SCFG 提出了识别一致 RNA 二级结构的方法.

COVE[27] 是一个基于 SCFG 的程序，它也叫 RNA 协方差模型 (covariance model, CM)，采用了一个有序树，树中的成对节点被指定为 RNA 结构中的碱基对，而单个节点被指定为未配对碱基. 因此，可以通过遍历 (从根到叶、从左到右)所有节点来得到原始序列. 协方差模型是隐马尔科夫模型 (hidden Markov model, HMM)的一般化. CM 模型中，产生概率被指定为 16 种可能的配对碱基或是 4 个单体碱基，转移概率是从一个当前状态转到几个可能的新状态 (匹配、插入、删除)之一的分值. 特殊的非产生状态描述了树结构本身，例如强迫转移到两个新状态的分支状态和哑的开始和结束状态.

在模型训练时，通过给训练序列集合指派高的比对分值 (极大似然)以得到优化模型的参数. 多序列比对是将训练集合中的单个序列和模型用一个三维动态规划法进行比对得到. 一致的 RNA 二级结构可以通过一个利用比对中所有配对列的互信息值的算法来得到. 这个算法类似于 Nussinov/Zuker 的 RNA 动态规划折叠算法，但不是优化基于热力学堆积能的分值，而是优化互信息函数值.

Sakakibara 改进了这个算法，通过利用可用的碱基配对信息，并从 RNA 二级结构的先验知识中构造了一个 RNA 随机上下文无关语法 (SCFG)[28].

Eddy 和 Sakakibara 的方法在进行 tRAN 和其他的小 RNA 的预测二级结构和多序列比对时，都有一个好的结果. 但是，这些方法的计算非常耗时、耗空间，它们不能分析超过 150 ~ 200 bp 碱基的序列. 这些方法的另一个限制是，它们需要大量的序列以构造一个满意的有判别力的模型.

Knudsen 用 SCFG 模型和进化历史提出了一种预测 RNA 二级结构的方法[29]，是对 SCFG 模型的改进. 这种方法要求序列有相同的二级结构，也就是说要有一个好的结构比对. 这种方法的计算复杂性是 $O(N^3)$，其中 N 是比对的长度. Holmes 提出了 Stemloc 算法[30]，它是一个 RNA 序列的双序列比对程序. 它首先预测每条序列的二级结构，然后预测基于初级序列的比对. 候选二级结构和比对被用作双序列比对 SCFG 的约束条件，这有效的减少了计算复杂性.

上面提到的 4 种方法都不能预测假节，为解决这个问题，Cai 提出了基于平行通讯语法系统 (parallel communicating grammar systems，PCGS)的语法模型方法[31]，这种方法通过添加更多的符号扩展了 SCFG. 和多数随机上下文自由语法相似，这种方法非常耗时、耗内存，所以这种方法只能用在小的数据集.

3.3 Sankoff 算法和贪婪 (greedy)算法

在 1985 年，Sankoff 提出了一种同时进行优化折叠和比对 RNA 序列的算法[32]. 这个算法的时间复杂性是 $O(N^6)$，空间复杂性是 $O(N^4)$，这使得它没有多少实际意义.

基于SCFG 的方法通常用于全局比对的序列，但许多 RNA 中的调控元件只是序列的一部分. 于是专门提出了 FOLDALIGN 以识别 RNA 序列中的局部元件，这些元件由序列和结构限制组成[33]. FPLDALPGN 用Sankoff 和贪婪算法在序列集中同时折叠和对齐局部茎环结构. 它采用动态规划算法以发现两序列间最高分值的局部比对，或是一条序列和其他对齐序列间的最高分值局部比对. 如果这些两两比对中的一个是正确的，那么当加入另一个序列进入比对后，这个效果将会加强，以此来识别序列集中真正的局部模式. 但是，FOLDALIGN 在计算上比 COVE 更敏感.

SLASH，Stem – Loop Align Search[9]程序，结合了 FOLDALIGN 和 COVE 以确定相关 RNA 序列集中的一致结构. 它首先用 FOLDALIGN 获得结构比对的预测，然后用这个比对来训练 COVE 的 SCFG 模型. SCFG 模型可以用来对齐除 FOLDALIGN 比对的其他序列，也能用于 FOLDALIGN 比对的更新. 这种方法能在随机 UTR 序列中找到像 IRE 这样的茎环模式.

对于 L 条 N 个碱基的序列集，FOLDALIGN 和 SLASH 算法的复杂性大概都是 $O(L^4N^4)$，显然也不能用于大的数据集.

3.4 图论方法

Cary[34] 和 Tabaska[35] 提出了将图论方法中的最大权重匹配用于预测含有假节的一致 RNA 结构的方法，这个方法的复杂性，时间上是 $O(N^3)$，空间上是 $O(N^2)$. 为了得到满意的结果，这种方法需要一个可靠的结构比对.

3.5 茎比较

最近提出来许多基于茎比较的方法. 这些方法首先从每个输入序列中抽取出候选区集合，这

些候选区基于碱基配对或是热力学规则的预测最优二级结构包含有茎环结构．然后，对这些选出的区域进行相互比较，以发现最相似的一组．一种方法是将结构中的茎模式表示成二元的，然后将 RNA 一致结构预测问题变成搜索一致的二元结构模式问题．在处理短元件时这种方法很快，但它没有考虑序列的相似性[36].

另一种方法是用遗传算法来递归进化和合成来自不同序列的茎，以形成最好的一致结构，例如，Chen[37] 和 Hu (GPRM)[10] 提出的算法．这种方法在计算大数据集是非常耗费资源．

第 3 种方法是用动态规划法来得到最好的茎比对[38, 39]，然后预测两序列上的一致 RNA 结构，CARNAC (Computer Alignment of RNA by Cofolding)就是采用这种方法．

第 4 种是 comRAN 所用的图论方法，它能在功能或进化相关的序列集中预测出一致的 RNA 二级结构模式[40]．这种方法首先找出所有可能的、稳定的茎，然后两两比较来自不同序列中的茎，以发现穿过任意两条序列的保守茎．然后用最大团发现算法来找出至少出现在 k 条序列中的所有有意义的茎．最后，将所有可能的、保守的茎 (至少在 k 条序列中出现)组装，最好的组装集作为一致结构模式的候选．这种方法能预测假节，也能发现序列子集中的一致 RNA 元件．正如作者指出的，这个方法的限制是在最坏的情况下，此算法的时间复杂性不是多项式的，这使得它不能用于大数量 (< 20)的长序列 (< 300).

和 comRNA 相似的另一个算法是 RNAProfile，它为每条序列预测茎，然后用贪婪启发算法来构造多序列比对[32]．这个程序也能发现小的序列子集中的保守元件，但它不能预测假节．这个算法是非常有效的，当输入序列是 8 条尺寸为 880 ~ 2 000 bp 时，所耗时不超过 3 min.

4　结论

RNA 序列元件包含了初级序列和二级结构的信息，已经提出了许多方法来识别潜在的 RNA 序列元件，以及对相关的 RNA 序列集进行多序列比对. 这些方法中的大多数都要求集合中的序列有较高的相似性，或是要求预先进行一个好的多序列比对. ComRNA 和 RNAProfile 是两个能在未对齐序列子集中发现局部保守 RNA 元件的程序. 但所有这些方法仍然不能用于真核基因组范围的分析．在 RNA 序列元件预测方面的期望将来能有突破．

参考文献(References)：

[1] HAKE L E, RICHTER J. D. Translational regulation of maternal mRNA [J]. Biochim Biophys Acta, 1997, 1332(1): 31-38.

[2] PAIN V M. Initiation of protein synthesis in eukaryotic cells [J]. Eur J Biochem, 1996, 236(3): 747-771.

[3] SONENBERG N. Translation factors as effectors of cell growth and tumorigenesis [J]. Curr Opin Cell Biol, 1993, 5(6): 955-960.

[4] PESOLE G, MIGNONE F, GISSI C, *et al*. Structural and functional features of eukaryotic mRNA untranslated regions [J]. Gene, 2001, 276(1-2): 73-81.

[5] STORMO G D, JI Y. Do mRNAs act as direct sensors of small molecules to control their expression? [J]. Proc Natl Acad Sci U S A, 2001, 98(17): 9465-9467.

[6] MANDAL M, BOESE B, BARRICK J E, *et al*. Riboswitches control fundamental biochemical pathways in Bacillus subtilis and other bacteria [J]. Cell, 2003, 113(5): 577-586.

[7] WILKIE G S, DICKSON K S, GRAY N K. Regulation of mRNA translation by 5′- and 3′-UTR-binding factors[J]. Trends Biochem Sci, 2003, 28(4): 182-188.

[8] PESOLE G, LIUNI S, GRILLO G, *et al*. UTRdb and UTRsite: specialized databases of sequences and functional elements of 5′ and 3′ untranslated regions of eukaryotic mRNAs [J]. Nucleic Acids Res, 2002, 30(1): 335-340.

[9] GORODKIN J, STRICKLIN S L, STORMO G D. Discovering common stem-loop motifs in unaligned RNA sequences[J]. Nucleic Acids Res, 2001, 29(10): 2135-2144.

[10] HU Y J. Prediction of consensus structural motifs in a family of coregulated RNA sequences[J]. Nucleic Acids Res, 2002, 30(17): 3886-3893.

[11] EDDY S R. Non-coding RNA genes and the modern RNA world [J]. Nat Rev Genet, 2001, 2(12): 919-929.

[12] DEVEREUX J, HAEBERLI P, SMITHIES O. A comprehensive set of sequence analysis programs for the VAX[J]. Nucleic Acids Res, 1984, 12(1): 387-395.

[13] BILLOUD B, KONTIC M, VIARI A. Palingol: a declarative programming language to describe nucleic acids' secondary structures and to scan sequence database[J]. Nucleic Acids Res, 1996, 24(8): 1395-1403.

[14] PESOLE G, LIUNI S, D' SOUZA M. PatSearch: a pattern matcher software that finds functional elements in nucleotide and protein sequences and assesses their statistical significance [J]. Bioinformatics, 2000, 16(5): 439-450.

[15] MACKE T J, ECKER D J, GUTELL R R, *et al*. RNAMotif, an RNA secondary structure definition and search algorithm [J]. Nucleic Acids Res, 2001, 29(22): 4724-4735.

[16] LOWE T M, EDDY S R. tRNAscan-SE: a program for im-

proved detection of transfer RNA genes in genomic sequence [J]. Nucleic Acids Res, 1997, 25(5): 955-964.

[17] EL-MABROUK N, LISACEK F. Very fast identification of RNA motifs in genomic DNA. Application to tRNA search in the yeast genome [J]. J Mol Biol, 1996, 264(1): 46-55.

[18] LISACEK F, DIAZ Y, MICHEL F. Automatic identification of group I intron cores in genomic DNA sequences[J]. J Mol Biol, 1994, 235(4): 1206-1217.

[19] LEWIS B P, SHIH I H, JONES-RHOADES M W, *et al*. Prediction of mammalian microRNA targets [J]. Cell, 2003, 115(7): 787-798.

[20] NUSSINOV R, JACOBSON A B. Fast algorithm for predicting the secondary structure of single-stranded RNA [J]. Proc Natl Acad Sci U S A, 1980, 77(11): 6309-6313.

[21] ZUKER M. Prediction of RNA secondary structure by energy minimization [J]. Methods Mol Biol, 1994, 25: 267-294.

[22] RIVAS E, EDDY S R. A dynamic programming algorithm for RNA structure prediction including pseudoknots[J]. J Mol Biol, 1999, 285(5): 2053-2068.

[23] DIRKS R M, PIERCE N A. A partition function algorithm for nucleic acid secondary structure including pseudoknots[J]. J Comput Chem, 2003, 24(13): 1664-1677.

[24] MATHEWS D H, TURNER D H. Dynalign: an algorithm for finding the secondary structure common to two RNA sequences [J]. J Mol Biol, 2002, 317(2): 191-203.

[25] CHEN S J, DILL K A. RNA folding energy landscapes [J]. Proc Natl Acad Sci U S A, 2000, 97(2): 646-651.

[26] AKMAEV V R, KELLEY S T. STORMO G D. A phylogenetic approach to RNA structure prediction [J]. Proc Int Conf Intell Syst Mol Biol, 1999: 7-10.

[27] EDDY S, DURBIN R. RNA sequence analysis using covariance models [J]. Nucleic Acids Res, 1994, 22: 2079-2088.

[28] SAKAKIBARA Y, BROWN M, HUGHEY R, *et al*. Stochastic context-free grammars for tRNA modeling[J]. Nucleic Acids Res, 1994, 22(23): 5112-5120.

[29] KNUDSEN B, HEIN J. RNA secondary structure prediction using stochastic contextfree grammars and evolutionary history [J]. Bioinformatics, 1999, 15(6): 446-454.

[30] HOLMES I, RUBIN G M. Pairwise RNA structure comparison with stochastic context-free grammars [C]// World Scientific. Singapore Pac Symp Biocomput, 2002. 163-174.

[31] CAI L, MALMBERG R L, WU Y. Stochastic modeling of RNA pseudoknotted structures: a grammatical approach[J]. Bioinformatics, 2003, 19(Suppl 1): 166-173.

[32] PAVESI G, GIANCARLO Mauri, MARCO Stefani, *et al*. RNAProfile: an algorithm for finding conserved secondary structure motifs in unaligned RNA sequences[J]. Nucleic Acids Res, 2004, 32(10): 3258-3269.

[33] GORODKIN J, HEYER L J, STORMO G D. Finding the most significant common sequence and structure motifs in a set of RNA sequences[J]. Nucleic Acids Res, 1997, 25(18): 3724-3732.

[34] CARY R B, STORMO G D. Graph-theoretic approach to RNA modeling using comparative data [C]// Proc Int Conf Intell Syst Mol Biol. United States: AAAI Press. 1995. 3: 75-80.

[35] TABASKA J E, CARY R B, GABOW H N, *et al*. An RNA folding method capable of identifying pseudoknots and base triples [J]. Bioinformatics, 1998, 14(8): 691-699.

[36] BOUTHINON D, SOLDANO H. A new method to predict the consensus secondary structure of a set of unaligned RNA sequences [J]. Bioinformatics, 1999, 15(10): 785-798.

[37] CHEN J H, LE S Y, MAIZEL J V. Prediction of common secondary structures of RNAs: a genetic algorithm approach [J]. Nucleic Acids Res, 2000, 28(4): 991-999.

[38] PERRIQUET O, TOUZET H, DAUCHET M. Finding the common structure shared by two homologous RNAs[J]. Bioinformatics, 2003, 19(1): 108-116.

[39] HELENE Touzet, OLIVIER Perriquet. CARNAC: folding families of related RNAs [J]. Nucleic Acids Research, 2004, 32(2): 142-145.

[40] JI Y, XU X, STORMO G D. A graph theoretical approach for predicting common RNA secondary structure motifs including pseudoknots in unaligned sequences[J]. Bioinformatics, 2004, 20(10): 1591-1602.

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