中国科学: 生命科学 2010 年 第 40 卷 第 4 期: 294 ~ 310
SCIENTIA SINICA Vitae www.scichina.com life.scichina.com
《中国科学》杂志社
SCIENCE CHINA PRESS

评 述

计算 RNA 组学: 非编码 RNA 结构识别与功能预测

郑凌伶, 屈良鹄*

中山大学生物工程研究中心, 广州 510275
* 联系人, E-mail: lssqlh@mail.sysu.edu.cn

收稿日期: 2010-01-17; 接受日期: 2010-02-03
国家自然科学基金(批准号: 30830066, 30771151)、国家重点基础研究发展计划(批准号: 2005CB724600)资助项目

摘要 真核生物基因组中包含大量非编码 RNA 基因, 计算 RNA 组学采用信息科学等多学科方法解析 ncRNA 的结构与功能. 本文就 ncRNA 数据存储与管理、ncRNA 基因识别与鉴定、ncRNA 靶标识别与功能预测等问题, 对目前计算 RNA 组学的主要研究方法和内容进行了评述.

关键词
计算 RNA 组学
非编码 RNA
转录组
结构
功能

人类基因组计划揭示, 在高等哺乳动物基因组中蛋白质编码序列不足 2%, 而 98%以上的序列为非蛋白质编码 DNA. 这些巨大的非蛋白质编码区并非“垃圾 DNA”, 而是包含大量的转录调控元件和非编码 RNA(non-coding RNA, ncRNA)基因. 由美国国家人类基因研究所(The National Human Genome Research Institute, NHGRI)提出的“DNA 百科全书”(ENCyclopedia of DNA Elements, ENCODE)计划, 经过 4 年的研究完成了 1%人类基因组的详细解码, 证实了人类基因组中的绝大部分区域均发生转录, 且含有大量 ncRNA 基因[1]. 在此基础上, 新的 ENCODE 计划开始对整个基因组进行高通量的注释[2]. “哺乳动物基因组功能注释国际协作计划”也启动了第 4 期计划(FANTOM4), 旨在进一步研究哺乳动物中大量转录元件、功能 RNA 的相互作用及转录调控网络[3]. 由此可见, 在现代高等生物细胞中, 存在一个隐蔽的“RNA 世界”, RNA 组学已成为后基因组时代一个重要的科学前沿[4].

RNA 组学是在基因组水平系统地研究生命细胞中 ncRNA 结构与功能的一门科学, 计算 RNA 组学(computational RNomics)是其中重要的组成部分. 计算 RNA 组学采用信息科学等多学科方法解析 ncRNA 的结构与功能, 主要研究以下 3 方面内容.

(1) ncRNA 数据的存储与管理. 随着 ncRNA 研究的不断深入, 大量 ncRNA 及其相互作用的分子通过实验方法或计算机方法被鉴定和预测, 尤其是利用大规模测序技术对转录本进行分析, 更产生了海量的序列数据, 将这些数据进行有效存储与管理, 是进行 RNA 组学研究的首要步骤. 计算 RNA 组学采用数据库对这些海量数据进行存储与分类管理, 并建立它们之间的关系, 方便研究人员查询及使用.

(2) ncRNA 基因的识别与鉴定. 目前对任何一种生物中 ncRNA 的确切数目和种类都不清楚, 而现在所知道的数据仅为冰山一角. 在各种生物中发现和鉴定 ncRNA 基因, 并将其分类注释, 是 RNA 组学的重要目标之一. 因而, 计算 RNA 组学开发出大量的算法在基因组范围内进行 ncRNA 基因的预测与分类鉴定.

英文版见: Zheng L L, Qu L H. Computational Rnomics: structure identification and functional prediction of non-coding RNAs *in silico*. Sci China Life Sci, 2010, doi: 10.1007/s11427-010-0085-5

(3) ncRNA 的靶标识别与功能预测. 寻找 ncRNA 基因的最终目的是揭示其在细胞中的功能并阐明其生物学意义. 而 ncRNA 通过与其他分子进行直接或间接地相互作用而起到调控作用, 因此, 计算 RNA 组学开发出一系列算法用以寻找与 ncRNA 相互作用的靶标分子, 进而对其功能及调控网络进行有效地预测.

本文对现有的研究方法和软件进行了归纳和分类, 对目前计算 RNA 组学的主要内容与问题进行了综述.

1 ncRNA 数据资源

目前的 ncRNA 数据库主要分为 3 大类: 通用的 ncRNA 数据库、专门的 ncRNA 数据库与利用大规模测序技术获得的转录本中得到的 ncRNA 数据库. 表 1 列出了部分常用的 ncRNA 数据库及其网址, 用户可根据自己的需要选择合适的数据库进行使用.

1.1 通用的 ncRNA 数据库

此类数据库全面搜集各种类型的 ncRNA, 包括长 ncRNA 和短 ncRNA, 并根据一定的标准进行分类存储, 数据主要来源于文献挖掘, 实验结果及计算机预测结果.

(1) RNAdb. 该数据库全面收集哺乳动物的 ncRNA 序列及注释信息, 其信息来源有: 文献中得以证实的 ncRNA; FANTOM3 项目的 ncRNA 数据集; 人类 7 号染色体注释计划以及来自 NCBI, Ensembl 和 UCSC 数据库中的序列[5]. 用户可通过多种方式搜索自己感兴趣的信息: 包括在浏览器中进行简单浏览、采用关键字进行查询、根据数据库给出的提示在特定范围内进行搜索, 还可将已有序列输入数据库进行 BLAST 同源性搜索. 该数据库中存储的 ncRNA 序列及用户定义的搜索结果可以 FASTA 或 XML 格式下载查看. 而且可以将整个数据库或搜索得到的序列结果转化成 BED 格式在 UCSC 的基因组浏览器中直接查看.

(2) NONCODE. 该数据库是由中国科学院计算技术研究所和中国科学院生物物理研究所共同开发并维护的 ncRNA 基因综合数据平台. 该数据库收集了大量的 ncRNA, 并提出了一套以 ncRNA 基因所参与的细胞生化过程和在此过程中发挥的功能为标准的统一分类体系. 目前的 2.0 版本包括来自 861 个物种的 212527 条序列, 几乎涵盖了从病毒到真核生物的所有物种. 其数据来源主要有 3 方面: 从文献中手工挖掘、从 GenBank 数据库中自动获取并加工整理、经该实验室验证的序列[6]. 该数据库可以按照物种或 ncRNA 类别进行浏览, 也可以根据 GenBank 的登录号或其他关键词进行查询, 同时支持 BLAST 搜索及 UCSC 基因组浏览器格式转换. 整个数据库可以 FASTA 格式下载.

(3) Rfam. 该数据库利用多序列比对和共进化模型建立 RNA 家族的注释信息[7]. 用户可按照 RNA 家族进行浏览, 也可以利用 BLAST 对感兴趣的序列进行同源搜索. 此外, 用户可以匿名登录到网站(ftp://selab.janelia.org/pub/Rfam/)下载文件. 比对数据为 Stockholm 格式, 可采用 HMMER 软件包对其进

表 1 常用 ncRNA 数据库

数据库	网址	特点	参考文献
RNAdb	http://jsm-research.imb.uq.edu.au/rnadb/	哺乳动物 ncRNA 数据库	[5]
NONCODE	http://www.noncode.org/	包含真核生物、细菌、古细菌、病毒 ncRNA	[6]
Rfam	http://rfam.sanger.ac.uk/	搜集结构 RNA 家族	[7]
fRNAdb	http://www.ncrna.org/frnadb/	综合性 ncRNA 数据库	[8]
miRBase	http://www.mirbase.org/	microRNA 数据库	[9,10]
snoRNABase	http://www-snorna.biotoul.fr/index.php/	人类 snoRNA 数据库	[11]
microrna.org	http://www.microrna.org/microrna/home.do/	microRNA 表达和靶标数据库	[12]
Tarbase	http://diana.cslab.ece.ntua.gr/tarbase/	实验证实的 microRNA 互作靶标数据库	[13]
deepBase	http://deepbase.sysu.edu.cn/	大规模测序所得 ncRNA 数据库	[14]

行解析.

1.2 专门的 ncRNA 数据库

该类型数据库专门存储某一类别的 ncRNA, 如 microRNA, snoRNA 等, 另外一些数据库存储 microRNA 及其靶标数据, 还有一些数据库整合了 microRNA 表达数据, 方便用户查看不同组织中的 ncRNA 表达情况.

(1) miRbase. 该数据库搜集了大量 microRNA 前体、成熟 microRNA 的序列及预测的二级结构信息, 该数据库对 microRNA 进行统一命名, 并且还可进行跨物种靶标预测. 用户可以通过 microRNA 的登记号、名字、关键词等信息进行检索, 并且可以简便地下载 microRNA 及前体的序列信息及注释信息.

(2) microRNA.org. 该数据库存有 microRNA 序列及其表达谱的数据, 用户可在特殊的组织中搜索和查看感兴趣的 microRNA 的表达谱情况, 并且可以下载以热点图(heat map)和条形图显示的结果. 另外还给出了采用 Miranda 软件预测到的 microRNA 靶标分子的信息.

(3) snoRNABase. 该数据库搜索来自人类的 box C/D 和 H/ACA snoRNA. 用户可通过 snoRNA 的名字、GenBank 登录号等信息进行搜索. 另外, 还可以根据 rRNA 和 snRNA 寻找靶向它们的向导 snoRNA. 该数据库与 UCSC 相互连接, 在 UCSC 中搜索到 snoRNA 后也可以链接到该数据库进行查看.

1.3 整合大规模测序结果的 ncRNA 数据库

最近, 大规模测序技术的发展, 提供了对转录本数据进行高通量测序的技术, 同时也积累了海量的序列信息. 因此, 合理有效地存储和整合这些数据, 并挖掘其中的信息也成为当前研究中的一项重要任务. 本实验室 Yang 等人[14]整合了现有的转录本测序数据, 开发了具备高通量和深度注释的小 RNA 数据库——deepBase 数据库.

deepBase 数据库的基本框架如图 1 所示. Yang 等人[14]搜集了来自于人类(*Homo sapiens*)、小鼠(*Mus musculus*)、鸡(*Gallus gallus*)、海鞘(*Ciona intestinalis*)、黑腹果蝇(*Drosophila melanogaster*)、线虫(*Caenhorhabditis elegans*)及拟南芥(*Arabidopsis thaliana*)7 个物种的不同组织和细胞的 230 个小 RNA 数据库. 其中包含 1400 多万的特定短序列片段(unique reads)唯一匹配到基因组的 2 亿 8 千余个座位上, 并进一步分析和鉴定出大批已知和未知的 ncRNA, 发现这些短片段可以聚成 120 万个簇. 用户可以将结果以 FASTA 和 BED 格式进行下载. 该数据库不仅为用户提供了良好的访问界面、许多相关的数据挖掘和搜索工具, 而且还给出了大量统计信息, 包括各文库中 reads 长度分布, 各文库中 5′端序列碱基分部情况, ncRNA 在不同文库中的表达情况等. 同时, 该数据库可以方便地与其他网站进行信息交互. 用户若想进一步了解感兴趣的序列, 还可以点击进入 UCSC 等综合性网站进行查看.

2 用于 ncRNA 基因发现的算法

各种模式生物基因组计划的实施, 为全基因组范围内识别和鉴定 ncRNA 提供了前所未有的机遇. 需要采用计算机方法, 开发出合适的算法, 进行大规模高通量的 ncRNA 识别. 在基因组中进行 ncRNA 基因识别的主要难点在于, ncRNA 序列不像编码序列那样具有明确的特征(可读框、起始密码子、终止密码子、特异性剪切位点等). 因此, 传统的基因预测软件都无法适用, 需根据 ncRNA 基因的特征开发出新的预测算法. 目前, 计算 RNA 组学已经发展了许多算法进行 ncRNA 基因的预测. 本文根据它们的特征将其分为 3 大类: 基于结构的预测方法、基于序列信号的预测方法及基于大规模测序结果的预测方法.

2.1 基于结构的预测算法

早期的预测算法基于这样一个假设: ncRNA 序列更易折叠成稳定的二级结构而存在, 因而采用结构预测软件在基因组范围内预测出具有稳定二级结构的序列即被认为是 ncRNA 基因. 进行 RNA 二级结构预测的软件有很多, 如 Mfold[15,16], RNAfold[17], Sfold[18~20]等, 它们也常用在 ncRNA 的预测中. 然而, 后来的研究发现, 这一特征并不足以作为 ncRNA 的判定标准, 因为许多随机序列同样具有较低的自由能[21]. 因此, 在自由能参数的基础上又加入了一个新的参数, 即二级结构保守性, 这是由于许多 ncRNA 存在进化上的协同变异现象, 即茎上的配对碱基存在共同变异, 从而保持结构的不变. 采用多基因组比对可以发现一致的二级结构. 现有的基于结构特征的 ncRNA 基因识别算法主要是将以上 2 个参数相结

合进行预测.

2001 年, Rivas 和 Eddy[22]基于序列比对分析方法开发了 QRNA 软件进行 ncRNA 基因预测. 基于 ncRNA 协同进化思想, 他们建立了 3 个统计配对方法, 根据结构 RNA 进化比对建立的一个配对随机上下文无关文法模型, 根据编码序列进化比对建立的一个隐马尔科夫模型, 根据位置无关性进化信息所构建的一个作为零假设的配对隐马尔科夫模型. 当输入一个双序列比对结果时, 算法根据模型的贝叶斯后验概率将序列归类为编码序列、RNA 序列或其他. 采用该算法在大肠杆菌基因组中进行测试, 将大肠杆菌(*Escherichia coli*)基因组和沙门氏菌(*Salmonella typhi*)基因组的比对数据作为输入, 得到的结果与已知的编码基因、ncRNA 和基因间区域进行比较. 结果显示, 程序的敏感性很大程度上取决于合适的比对数据. 他们认为如果采用多个基因组的比对结果会优于双基因组的结果.

QRNA 在细菌和酵母的 ncRNA 识别中表现出了良好的效果, 但对于大基因组则表现欠佳. 同时其输入必须是双序列比对结果, 特别是当输入为两个进化关系较远的基因组时, 可信度较差[23]. 随后开发的几款软件同样存在各种不足, 使得对 ncRNA 的预测存在限制. 面对这些问题, Washietl 等人[23]在 2004 年基于保守结构的热力学稳定性开发了 RNAz 软件, 该算法包含两个部分: 根据一致性二级结构得出 RNA 二级结构保守性, 根据 *z* 分值计算热力学稳定性. *z* 分值的计算根据序列长度和碱基组成的标准化值通过 SVM 的回归模型得出. RNAz 软件可分析双序列及多序列比对结果. 他们采用该软件对来自人类、大鼠(*Rattus norvegicus*)、小鼠、河豚(*Takifugu rubripes*)及

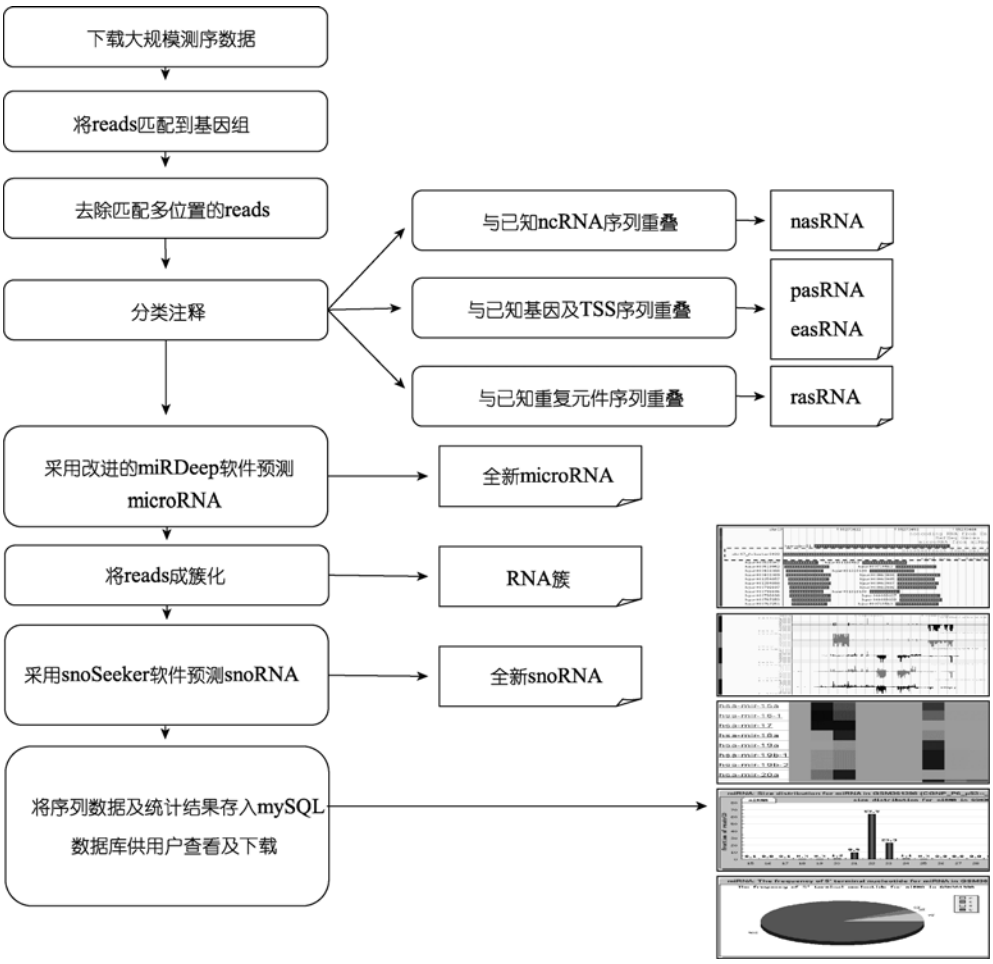

图 1　deepBase 数据库的基本框架

斑马鱼(*Danio rerio*)的直系同源基因的上游区域非编码元件进行预测, 找到了当时所有已知的 ncRNA 和顺式调控元件, 同时还发现了一批全新的保守 RNA 二级结构.

伴随着基因组计划的逐步完成, 越来越多的物种全基因组序列得以被确定, 这也为通过比较基因组学建立进化关系来寻找 ncRNA 提供了更多的可用数据. 2006 年, Pedersen 等人[24]开发的 Evofold 软件根据比较基因组建立了进化上下文无关文法模型用来寻找功能 RNA. 采用该模型根据来自于人类、黑猩猩(*Pan troglodytes*)、大鼠、小鼠、鸡、斑马鱼及河豚的全基因组比对数据, 寻找人类中保守的功能 RNA. 最终他们找到了大量已知的功能 RNA 和各类反转座元件(genetic recoding elements)及许多全新的 ncRNA.

表 2 中列出了目前较为普遍的基于结构方法进行 ncRNA 识别的软件, 它们主要根据结构稳定性及跨物种结构保守性两个参数进行识别, 对于长、短 ncRNA 都达到了较好的识别效果, 因而是一种通用的 ncRNA 识别方法. 然而该方法的不足之处在于: 首先, 预测的结果可能是那些在 mRNA 和 ncRNA 中都保守的元件, 如在 5′和 3′UTR 区域的调控蛋白结合位点的结构即十分保守[25]. 另外, 无法找出不具备保守结构的 ncRNA; 其次, 该方法不能给出预测结果具体属于哪一类 ncRNA. 所以采用结构方法进行预测会导致显著的假阳性和假阴性结果的出现[26], 同时不能进行专门类别的 ncRNA 预测.

2.2 基于序列特征的预测算法

采用通用的基于结构特征的 ncRNA 预测方法虽然可以达到普适性的效果, 但其并非放之四海而皆准的策略. 而且 ncRNA 的种类之繁杂, 数目之庞大, 找寻其中的共性本身就是一项艰巨的任务. 而人们对 ncRNA 的认识尚处于刚刚起步的阶段. 因此, 许多研究者针对各类型的 ncRNA 开发出了专用的识别和注释软件, 根据特定类别的 ncRNA 的已知序列, 分析其特征并建立模型, 用以全基因组范围内搜索新的 ncRNA.

随着人们对 ncRNA 认识的逐步深入, 已有数千小 ncRNA 得以被鉴定, 包括 microRNA, snoRNA, siRNA, piRNA 等. 另外, 通过全长 cDNA 克隆和 tilling arrays 技术也鉴定了成百上千的长 ncRNA. 由于长、短 ncRNA 的序列特征差异很大, 所以对其特征提取方式也不尽相同, 下面分别介绍这两种 ncRNA 的预测方法.

(1) 长 ncRNA 的预测算法. 根据实验可以获得大量的转录本数据(cDNA 或 EST), 这些序列中有的是编码蛋白序列, 也有非编码序列. 由于蛋白编码序列具有明确的可读框(ORF), 且长度有一定的范围, 并且不同物种之间的 ORF 较保守, 可以根据它们的特征进行编码序列与非编码序列的区分. 如一些程序采用 ORF 长度作为判别标准, FANTOM 项目采用 300 nt(100 个密码子)长度作为判别编码与非编码的标准, 在这个长度以下的转录本即可能为非编码序列[29]. 当然, 这一条件会造成误判. 如有一些较长的 ncRNA 恰巧具有 ORF 特征, 如 *H19*, *Xist*, *Mirg*, *Gtl2* 和 *KcnqOT1* 都含有超过 100 个密码子长度[30]. 另一方面, 一些短的编码序列会被误判成 ncRNA, 如 *tal*(*tarsal-less*)基因, 最初也被误判为 ncRNA[31]. 因此, 需要增加限制条件, 如寻找序列的保守功能域、特定的受体/供体位点、氨基酸组成偏好、多腺苷酸信号等特征. 判别编码与非编码序列的软件有很多, 如 ESTScan[32], CSTminer[33], CRITICA[34]. Frith 等人[35]选用 10 个此类软件对 FANTOM 项目中的小鼠 cDNA 序列进行编码与非编码的判别, 经过比较显示, CRITICA 软件的一致性程度最高, 并且识别错误最少, 该软件将采用比较分析与统计特征分析相结合

表 2 基于结构的 ncRNA 识别软件

软件	网址	网络版	本地版	参考文献
Mfold	http://mfold.bioinfo.rpi.edu/	√	√	[15,16]
RNAfold	http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi/	√	√	[17]
QRNA	ftp://selab.janelia.org/pub/software/qrna/		√	[22]
RNAz	http://www.tbi.univie.ac.at/~wash/RNAz/	√	√	[23]
Randfold	http://bioinformatics.psb.ugent.be/software/details/Randfold/		√	[27]
RNAstrand	http://www.bioinf.uni-leipzig.de/Software/RNAstrand/		√	[28]

的方法进行编码与非编码序列的判别.

采用机器学习方法也可以对 mRNA 与 ncRNA 进行判别, 将已证实的 ncRNA 和 mRNA 序列作为输入, 提取它们各自的特征, 后采用机器学习或统计方法进行训练及预测. 近年来许多方法都青睐于采用支持向量机(SVM)来完成. CONC[36]判别软件提取了序列的特征作为特征向量, 如长度、与已知蛋白间的同源性、氨基酸组成、二级结构、表面溶解性和热力学信息等, 之后采用真核生物的编码蛋白和已知 ncRNA 作为正负数据集对 SVM 模型进行训练, 并针对 102801 个小鼠 cDNA 数据进行预测, 找到了 14000 个可信的 ncRNA, 并预测总 ncRNA 数目为 28000, 该算法最终可达 97%的敏感度和 98%的特异性. CONC 软件提取了 180 个特征向量进行计算, 随后北京大学同样采用 SVM 开发了 CPC[37]软件, 该软件提取的特征向量仅有 6 个, 但判别的敏感度和特异性与 CONC 不相上下, 而耗费时间则远远小于 CONC.

(2) 短的 ncRNA 基因预测方法. 短的 ncRNA 包括 microRNA, snoRNA, siRNA, piRNA 等, 且每一类都有独特的特征. 对于这些 ncRNA 的预测算法, 所基于的参数主要有: 特殊的长度、位置、结构、自由能、序列模式、GC 含量、同源信息等.

在基因组范围内进行 ncRNA 的搜索可以得到大量的结果, 因为很多序列都能够折叠成茎环结构, 但未必是真实的 ncRNA. 由于真实的 ncRNA 折叠成的二级结构比随机产生的茎环结构具有更低的自由能, 而且 microRNA 二级结构的物种间保守性优于一级序列, 因而需考虑序列的自由能特性及结构保守性. 来自清华大学的 Wang 等人[38]基于序列和结构的保守性开发了 MiRAlign, 并在 *Anopheles gambiae* 中找到了 59 个新的 microRNA. MiPred[39]提取了序列的二级结构最小自由能特征并采用机器学习方法区分真实的 microRNA 前体和随机的茎环结构, 并得到了 95.09%的敏感度和 98.21%的特异性. MiRFinder[40]通过比较相近的两个基因组数据并采用 SVM 进行全基因组范围内的 microRNA 前体预测, 得到的准确率超过 99%.

snoRNA 的预测比 microRNA 更为复杂, 由于它包含两类对靶 RNA 有着不同修饰作用的分子, 分别是 C/D box snoRNA 和 H/ACA box snoRNA. 最早的 snoRNA 搜索程序是 Snoscan[41]和 snoGPS[42], 它们主要寻找向导 snoRNA, 如专门靶向 rRNA 或 snRNA. 随后, 本实验室 Yang 等人[43]开发了 snoSeeker, 该软件可以进行全基因组范围内的各类 snoRNA 的寻找, 包括没有靶标的‘orphan’ snoRNAs, 并且可以进行靶位点的预测. 该软件包含两个程序: CDseeker 和 ACAseeker, 该算法首先剔除基因组中的重复序列, 接着基于比较基因组学寻找同源序列作为候选 snoRNA, 然后利用一系列特征建立统计模型. 采用该方法在人类基因组中进行搜索, 最终找到 266 个已知 snoRNA 及 54 个全新的 snoRNA, 其中包括 26 个‘orphan’ snoRNA. 而另一个程序 snoReport[44]结合 RNA 二级结构预测信息与机器学习方法同时预测两类 snoRNA, 并且不依靠比对信息, 而是在整个基因组范围内全新搜索结果或对其他方法找到的 snoRNA 进行验证.

(3) 基于序列信号的 ncRNA 预测方法总结. 表 3 列出了采用序列信号进行长、短 ncRNA 基因预测的软件. 采用序列特征的预测方法进行单独类别的 ncRNA 基因预测较为可行, 但也存在一些缺陷. 首

表 3 基于序列信号的 ncRNA 识别软件

软件	网址	网络版	本地版	参考文献
ESTScan	http://www.ch.embnet.org/software/ESTScan2.html/	√	√	[32]
CSTminer	http://t.caspur.it/CSTminer/	√		[33]
CRITICA	http://rdpwww.life.uiuc.edu/	√	√	[34]
CPC	http://cpc.cbi.pku.edu.cn/	√	√	[37]
MiRAlign	http://web.archive.org/web/20080803072157/bioinfo.au.tsinghua.edu.cn/miralign/	√		[38]
Snoscan	http://lowelab.ucsc.edu/snoscan/	√	√	[41]
snoGPS	http://lowelab.ucsc.edu/snoGPS/	√	√	[42]
snoSeeker	http://genelab.zsu.edu.cn/snoseeker/	√	√	[43]
snoReport	http://www.bioinf.uni-leipzig.de/Software/snoReport/		√	[44]

先, 这一方法的假设前提就值得商榷, 即假设 RNA 分为编码和非编码, 然而有些 RNA 本身就是双功能的, 它们既可以作为调控 RNA 又可以被翻译成蛋白[45]; 另外, 利用转录本数据的一个决定性因素是输入数据的完整性, 即要求全长的序列. 然而实验数据往往存在各种各样的错误, 而使这一要求不能满足, 如反转录不完全、pre-mRNA 的内部启动(internal priming of pre-mRNAs)和基因组 DNA 的污染, 都有可能产生错误或截短的转录本, 而这些序列则可能被误判为 ncRNA[46]; 其次, 采用同源序列分析的方法, 仅能找到已知的 ncRNA, 而无法发现新的或不保守的 ncRNA 基因. 同时, 由于实验方法的限制, 对于一些低表达量的或组织特异性的 ncRNA 的预测和验证都存在一定困难.

2.3 基于大规模测序的预测算法

随着测序技术的不断发展, 许多软件利用转录本数据反匹配到基因组上寻找小 ncRNA. 近来采用该方法在 microRNA 的研究方面取得了初步成效. 图 2 列出了利用大规模测序结果寻找 microRNA 的基本流程[26,47], 大规模测序往往得到大量的序列片段(reads), 经过初步处理后, 将这些片段匹配到基因组上, 如果匹配到多个位置上则不选择, 接下来剔除掉已知的 mRNA、其他的 ncRNA(如 tRNA, snoRNA,

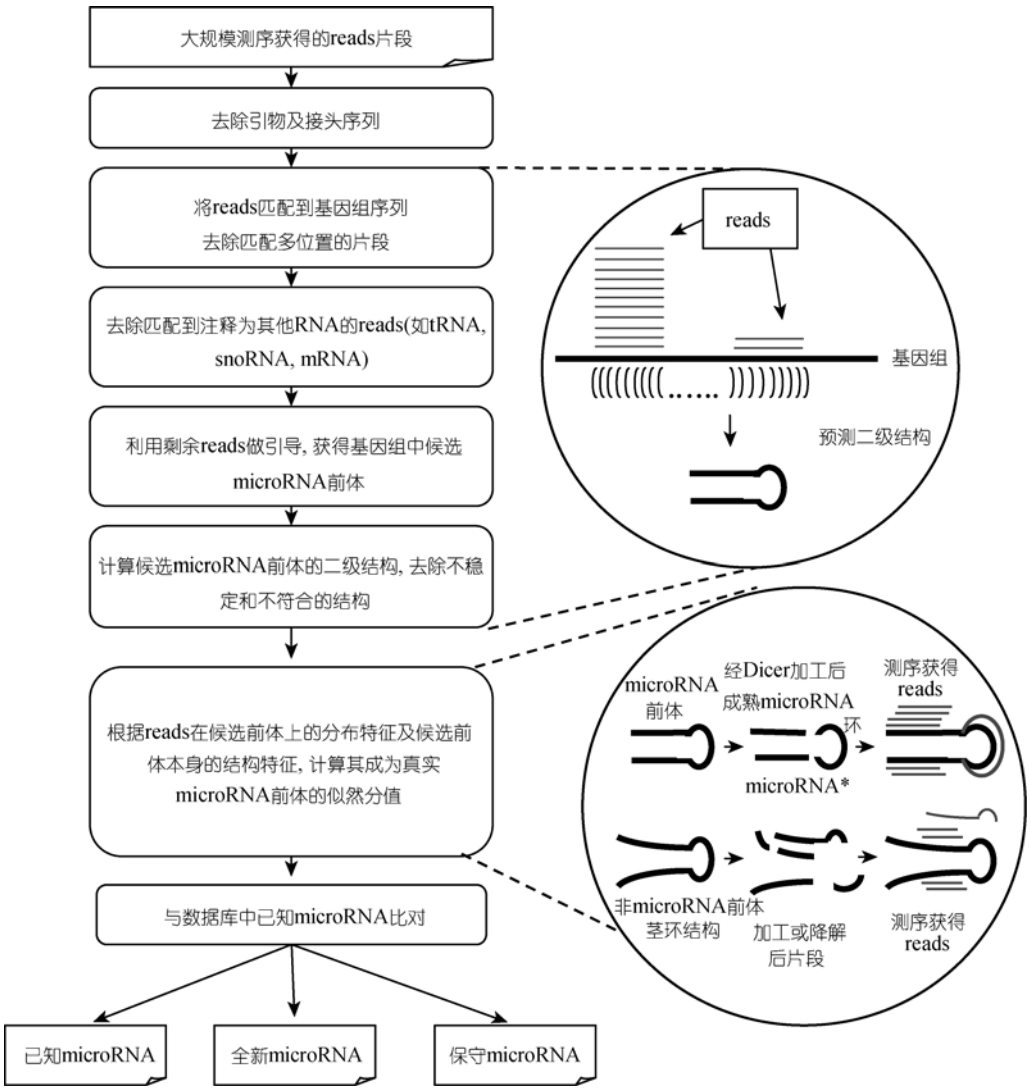

图 2 利用大规模测序结果寻找 microRNA 的基本流程图

piRNA等), 再根据二级结构的自由能等信息, 提取其中可能的microRNA前体. 接下来将这些序列与已知的microRNA进行聚类, 最后与已知的microRNA进行比对, 找出已知的、保守的和全新的microRNA. miRDeep[47]根据Illumina/Solexa的测序结果, 依据reads在具有microRNA前体二级结构序列的匹配位置和丰度建立概率模型, 预测可能的microRNA, 采用线虫、狗(*Canis familiaris*)和人类的测序数据进行实验, 最终找到230个全新的microRNA, 并将其中的一小部分进行Northern blot验证.

最近, FANTOM4 项目组的研究人员利用大规模测序技术取得了几项令人瞩目的研究成果. Taft 等人[48]分析了人类、鸡和果蝇的大规模测序数据, 发现位于基因转录起始位点(transcription start sites, TSS)−60~+120 区域存在着一段 18 nt 的保守序列模式. 他们将该类小分子命名为 tiRNAs(transcription initiation RNAs), 认为该类小分子 RNA 普遍存在于多细胞生物甚至整个真核生物的细胞中. 另一个研究成果来源于同一项目的 Carninci 实验室[49], 他们利用 FANTOM3 时期获得的 CAGE(cap analysis gene expression)数据, 即对全长 cDNA 的 5′端 20~21 nt 标签切割后进行大规模测序得到的数据进行分析, 发现人类和小鼠中有 6%~30%的 RNA 转录起始位点含有重复元件. 他们在全基因组范围内鉴定出 23000 个候选反转录转座子调控序列, 其中超过 2000 个具有双向转录, 认为这些调节序列对哺乳动物的转录结果具有重要影响.

表 4 列出了利用大规模测序技术进行 ncRNA 预测的几款软件, 大规模测序技术为 ncRNA 预测问题提供了良好的数据, 采用这些数据结合 ncRNA 的生物学特征和加工机制开发出的算法可以达到高效准确的预测效果, 将是未来发展的重要方向.

2.4 *ncRNA* 基因发现算法小结

ncRNA 基因的识别和分类算法, 将会发现新的 ncRNA 类别及其在基因表达调控中的作用. 然而, 目前还没有一个通用的工具可以实现准确地鉴别并注释所有的 ncRNA, 多数计算机方法都是针对特异的情况而开发的. 对于研究已知的 ncRNA, 在使用时根据 RNA 所属的类别而采用特定的方法, 或采用几种方法对比结合来预测结果, 可以得到较高的敏感性. 然而, 由于大量的 ncRNA 类别还不为人知, 开发出通用的方法进行普适性的预测仍然势在必行. 另外, 随着实验技术的不断发展, 会有许多新的类别加入到 ncRNA 家族中来, 找寻这些 ncRNA 的特征, 建立合适的算法进行全基因组范围内的识别和鉴定, 将是未来的发展方向之一. 目前, 小 RNA 的作用机制和功能在被一步步解开, 加之大规模测序技术提供了良好的转录本分析资源, 使得小 RNA 的识别和注释研究都取得了长足的进步. 相比之下, 长的 ncRNA 研究方面则略显缓慢, 未来还需要进一步研究和探索. 同时, 需要认识到 ncRNA 的识别之路仍然任重而道远. 现在所认识的主要是保守的和普遍表达的 ncRNA, 而还有许多组织或物种特异低表达的 ncRNA 有待发掘. 因此, 开发出更精确和快速的识别算法仍然是未来发展的重要方向.

3 ncRNA 靶标预测算法

对于 *ncRNA* 基因的功能注释中非常重要的一步是寻找与其相互作用的分子, 并最终阐明其在调控过程中的整个作用通路. 许多 ncRNA 都是通过与其他分子进行直接作用而发挥功能的. 并且 ncRNA 与靶标分子的作用结果取决于它们之间作用时的序列互补程度. 当完全或几乎完全互补时, 如 siRNAs 或植物中的 microRNA, 靶标分子将会被剪切; 而大多数动物体内的 microRNA 与靶分子之间存在部分互补, 则 microRNA 主要对 mRNA 翻译进行抑制. 且 microRNA 与其靶标之间的互作还取决于靶位点的位置和局部二级结构. 因此, 寻找并鉴定 ncRNA 的靶标分子是揭示其功能的基础和关键.

表 4 利用大规模测序技术进行 ncRNA 识别的软件

软件	网址	网络版	本地版	参考文献
snoSeeker	http://deepbase.sysu.edu.cn/	√		[14]
miRDeep	http://www.mdc-berlin.de/en/research/research_teams/systems_biology_of_gene_regulatory_elements/projects/miRDeep/		√	[47]
MiRanalyzer	http://web.bioinformatics.cicbiogune.es/microRNA/	√		[50]

microRNA 由于其重要的生物功能和特殊的靶标结合机制, 成为靶标预测算法的主要研究对象. microRNA 的生物学功能包括控制细胞增殖、细胞死亡、脂肪代谢及在植物中控制叶子和花的发育. 预测人类中有超过 30%的基因受到 microRNA 的调控. 而 microRNA 的功能主要是通过与靶 mRNA 的 3′UTR 区相结合, 进而参与细胞增殖、凋亡、分化、代谢、发育、肿瘤转移等多种生物过程. 因此找寻其靶基因是揭示 microRNA 功能的重要步骤. 然而, 由于互作方式的多样性及靶标分子本身的二级结构特性都增加了识别的难度. 且动植物中 microRNA 与靶标的结合机制有较大的不同, 植物的 microRNA 预测较为容易, 因为在植物中 microRNA 与 mRNA 的配对几乎是完全互补的. 因而大多数研究把目光聚焦在动物 microRNA 靶基因的识别问题上. 如今, 已诞生了数十种同类的预测软件, 并总结出很多互作规律. 本文根据算法所运用的靶标识别原则的不同, 将现有算法分成 3 类: 基于常规原则的识别方法、突破常规原则的识别方法与整合其他原则的识别方法.

3.1 基于常规原则的靶标识别方法

大多数动物 microRNA 靶标识别软件都主要遵循 3 个基本原则: 即 microRNA 5′端种子区与 mRNA 3′UTR 序列的互补性, microRNA-mRNA 二聚体的结构及能量特性, 靶位点的物种间保守特性. 图 3 列出了采用常规方法进行 microRNA 靶基因识别的流程.

2003 年, Enright 等人[51]开发了 microRNA 靶标预测算法 miRanda, 该算法主要依据 3 个标准: 序列间的互补性、形成 RNA 二聚体分子的热稳定性、靶位点的保守性. 他们利用黑腹果蝇(*Drosophila melanogaster*)中 73 个已知的 microRNA, 在 9805 个基因的 3′UTR 区进行搜索, 最终找到了 535 个潜在的靶基因. 由于黑腹果蝇和冈比亚按蚊(*Anopheles gambia*)的 microRNA 基因具有极高的保守性[52], 采用同样的过程针对 *A. gambiae* 进行搜索, 找到 150 个潜在靶基因, 其中 40%与黑腹果蝇该位点的一致性大于 60%. 接下来, 他们用已证实的数据进行验证, 对于线虫和黑腹果蝇的 10 个已知靶标, 正确识别出 9 个. 另外, 还采用随机构造的 microRNA 与真实 microRNA 数据进行靶基因预测结果比较, 然后利用一种简单的方法评估假阳性率(R^-/R^+), 其中 R^-表示随机构造的 microRNA 预测到的靶基因数目, R^+表示真实 microRNA 预测到

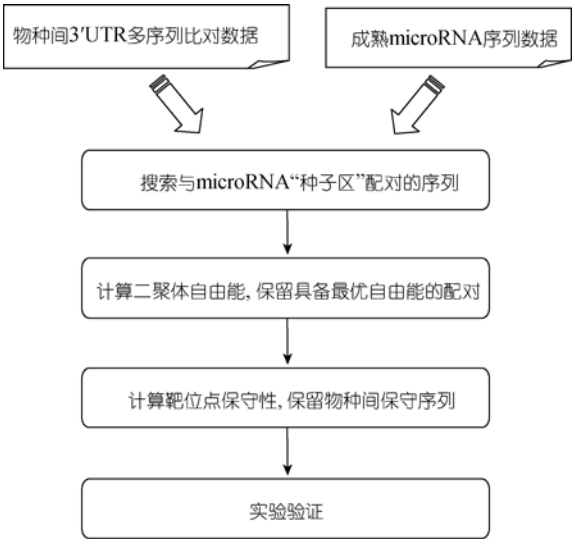

图 3 利用常规原则进行 microRNA 靶基因识别流程图

的靶基因数目, 最终的结果为 35%. 接着采用 Gene Ontology[53]数据库提供的生物过程和分子功能信息, 对经过预测后得到的靶基因进行进一步的分析试图揭示 microRNA 的功能. 最终得出结论认为, microRNA 参与众多的控制基因活性的过程, 尤其是转录、翻译及蛋白降解.

同年, Lewis 等人[54]开发了 TargetScan 软件用于哺乳动物 microRNA 靶标的预测. 他们不仅对该算法的假阳性率进行了精确的计算, 并且对其预测结果进行了实验验证. 该软件同样根据由互补配对构成的 RNA 二聚体的自由能特性和靶位点在不同物种间的保守性设计而成. 在计算互补配对的过程中, 他们提出 5′端的 2~8 个碱基, 对于靶标识别起至关重要的作用, 该片段称为“种子区”. 因此, 首先寻找 mRNA 的 3′UTR 中与“种子区”的 7 个核苷酸能够完全匹配的片段, 接着向片段两端不断延伸, 直到出现错配为止. 对于 3′UTR 中余下的部分采用 RNAFold 程序预测二级结构寻找最优的匹配关系以保证得到的靶位点最长. 他们从 Ensembl 数据库获得人类、大鼠、小鼠的直系同源基因的 3′UTR 区域, 最终预测到 451 个 microRNA: 靶基因互作关系, 平均每个 microRNA 有 5.7 个靶基因, 将该数值作为预测的“信号”. 另外采用随机方法构造 microRNA 进行靶标预测, 结果作为对照组, 平均每个 microRNA 有 1.8 个靶基因, 该数目作为“噪声”, 由此得到“信噪比”为 3.2∶1(5.7/1.8),

即假阳性率约为 31%(1.8/5.7). 另外, 对 15 个预测到的靶标进行了实验验证, 得到了 11 个真实的靶标. 接下来, Lewis 等人采用与 Enright 等人类似的方法, 对 TargetScan 预测得到的 mRNA 进行了分子功能分析, 同样发现 microRNA 靶基因在某些分子功能类中显著分布, 其中以转录和转录调控功能类别最为明显.

2005 年, Lewis 等人[55]又对 TargetScan 做了进一步优化, 开发了 TargetScanS. 该算法不要求输入全基因组比对数据, 而是更注重 3′UTR 中与种子区配对序列的保守性. 种子区的定义也有所改变, 先前定义 microRNA 5′端的 7 个碱基作为种子区, 而改进后的算法注重 5′端 8 个核苷酸的情况, 要求第 2~7 位的 6 个碱基与靶序列完全互补, 而第 1 位为腺嘌呤或第 8 位可与靶标序列互补. 他们采用该算法对 4 个脊椎动物基因组进行了预测, 得到了约 13000 个互作关系, 并得出人类中超过 30%的基因受 microRNA 调控的结论.

Krek 等人[56]引用概率思想开发了用于脊椎动物 microRNA 靶基因预测的 PicTar 软件, 计算一段序列成为 microRNA 靶位点的最大似然率. 该软件分成两个部分, 第 1 部分识别单个 microRNA 的靶位点, 对于一个给定的 microRNA, 计算一个 mRNA 成为其靶位点的概率; 第 2 部分, 对于一个给定的 RNA 序列(多为 mRNA 的 3′UTR), 计算多个 microRNA 靶向该序列的最大似然概率. PicTar 遵循 Doench 与 Sharp[57]的结论, 认为 microRNA 以竞争方式结合到靶基因上. 他们采用了 8 个脊椎动物全基因组的比对信息进行测试, 正确找到了所有已知的 microRNA 靶位点, 并估计假阳性率约为 30%. 另外他们还提出脊椎动物平均每个 microRNA 的靶基因有 200 个, 同时还存在协同作用.

随后的几年中, 人们不断采用 PicTar 软件在多个物种中进行预测. Grün 等人[58]于 2005 年采用该算法在 7 个果蝇基因组中进行预测, 最终结果显示, 对于一个 microRNA 平均有 54 个基因受其调控. 2006 年, Lall 等人[59]将该算法应用于线虫基因组的 microRNA 靶基因预测问题中, 根据 3 个线虫物种的比对数据, 得出至少 10%的线虫基因受到 microRNA 调控, 还预测受 microRNA 调控的基因间存在功能相关性. 接着, Chen 与 Rajewsky[60]采用该算法进行了人类 microRNA 结合位点的分析. 结果显示, 多数保守 microRNA 结合位点比其他保守性 3′UTRs 序列经历更多的负选择(negative selection), 对于非保守 microRNA 结合位点而言, 估计当 mRNA 和 microRNA 内源性共表达时, 其中 30%~50%的 microRNA 结合位点是有效的. 这说明预测到的 microRNA 结合位点的多态性是趋向于有害的, 因此在检测人类疾病方面可能是候选靶标.

表 5 列出了基于常规原则进行 microRNA 靶标识别方法的常用软件及其使用的范围, 基于常规原则进行预测可以得到较好的结果, 然而许多研究者提出这些原则不足以作为 microRNA 靶标识别的判别标准. Grimson 等人[61]结合计算机和实验方法提出了靶位点的其他 5 个特征对 microRNA 识别较为有效: (1) 靶位点附近的 AU 富集区; (2) 靶位点与共表达的 microRNA 邻近; (3) 与 microRNA 有约 13~16 个碱基的配对; (4) 位于 3′UTR 与终止密码子距离至少 15 nt 区域; (5) 处于长 UTR 中远离中心的位置. 因此, 许多软件在常规的识别原则上进行了改进和突破.

表 5 利用常规原则进行动物 microRNA 靶标预测的算法

软件	适用范围	网址	网络版	本地版	参考文献
miRanda	脊椎动物	http://www.microRNA.org/	√	√	[51]
TargetScan, TargetScanS	哺乳动物	http://www.targetscan.org/	√		[54,55]
PicTar	脊椎动物	http://pictar.mdc-berlin.de/	√		[56]
DIANA-microT	哺乳动物	http://diana.cslab.ece.ntua.gr/microT/	√		[62]
RNAhybrid	哺乳动物	http://bibiserv.techfak.uni-bielefeld.de/rnahybrid/	√	√	[63]
mirTarget2	人类、大鼠、小鼠、鸡、狗	http://mirdb.org/miRDB/		√	[64]

3.2 突破常规原则的靶标识别方法

虽然大多数 microRNA 的靶标结合机制都遵循常规原则, 然而, 许多其他特征也同样对 microRNA 靶基因的识别起作用, 包括靶位点数目、靶位点附近的 AU 含量、microRNA 与 mRNA 的表达量等都被作为一定的判别标准进行研究. 甚至有许多研究发现, 常规的识别原则在某些情况中不能适用, 并且各个物种中 microRNA 靶标结合机制不尽相同[65], 从而提出了许多新的判别标准, 这些标准可以作为常用规则的补充规则, 甚至有的对常规原则做出了挑战.

(1) 3′UTR 规则. 大多数预测算法都遵循一个普遍规则, 即认为动物的 microRNA 靶向 mRNA 的 3′UTR 区域, 因此主要集中在这一区域进行靶位点的搜索, 并找到了大量真实的靶标. 然而, 已有许多文献报道, microRNA 可以靶向 mRNA 的启动子区[66]、基因编码区[67]及 5′UTR 区[68].

Forman 等人[69]开发了一个算法用于在编码区寻找保守序列模式, 分析了 17 个脊椎动物基因组的多序列比对数据, 首先在编码区域中, 搜索完全保守的 8 bp 长度的序列, 给每一个保守模式分配一个序列水平保守得分(sequence level conservation score, SLCS)代表该模式在核苷酸水平上超过氨基酸水平的保守程度. 分析这些高分保守模式, 发现其中有很多是已知的 microRNA 靶标序列. 之后通过实验证实了 let-7 microRNA 直接靶向 microRNA 加工过程必须的 Dicer 酶的基因编码区, 由此构成了一个负反馈环. 接下来, 他们采用 ViennaRNA 软件包中的 RNAcofold 程序计算 let-7 的种子区与编码区, 3′UTR, 5′UTR 配对的二级结构. 通过比较发现, 与 3′UTR 区配对的 microRNA 在 10~12 nt 位置多具有环结构, 13~16 nt 位置具备互补配对结构, 然而与编码区配对的 microRNA 常在 13~15 nt 位置及 17~19 nt 位置的配对结构, 而无环结构偏好. 由此, 他们认为 microRNA 在编码区和 3′UTR 区具有不同的结合机制.

(2) “种子区”配对规则. microRNA 与靶位点的结合规则中, 很重要的一条是“种子区”配对规则. 在早期的研究中发现, microRNA 的 5′端约 6 nt 序列片段的完全互补对靶标分子的识别非常重要[70]. 并且通过对人类、果蝇及拟南芥(*Arabidopsis thaliana*)细胞的诱变分析, 认为种子区可以允许错配出现[57,71,72]. 接下来的研究将该规则进一步完善, 认为 microRNA 5′端第 2~7 位序列片段为“种子区”, 目标序列与其匹配的程度决定了靶标分子的识别能力[55]. 现有的算法基本都遵循“种子区”规则进行评分, 许多预测软件都识别到真实的靶标分子. 并且高通量转录本和蛋白质组学的研究也表明在 microRNA 过表达或失活的人类细胞中, 与种子序列配对的区域富集了大量的基因[73~75]. 虽然众多研究都表明“种子区”的结合对 microRNA 靶标识别确实具有重要作用, 但不能说明所有 microRNA 的靶标分子都必须遵循“种子区”规则, 如在早期的研究中即发现 let-7 microRNA 拥有许多 lin-41 和 lin-14 mRNA 上的非“种子区”配对结合位点[18,76], 因此, 将其作为寻找靶标的强制条件会漏掉许多真实靶标.

清华大学深圳研究生院张雅鸥实验室[77]采用血管内皮生长因子(VEGF)为研究对象, 分析了 microRNA 与靶标分子构成二聚体的“中心环”区序列对靶标识别的影响. 他们开发的预测程序——FindTar 在 2.0 版本中, 引入了“中心环”规则. 即考虑中心环的 3 个指标: 环位置、环大小及环优先级. 结果显示, microRNA 与靶序列构成的二聚体中心区域的环状结构对靶标识别有重要影响. 他们认为, 在现有的识别规则中加入中心环的位置及大小特征, 将显著降低识别算法的假阳性率, 并提高 microRNA 靶基因预测的特异性.

(3) 靶位点保守性原则. 常规原则都需要考虑 microRNA 靶基因物种间保守性的特征. 然而, 已有研究证实, 大量的靶基因不存在跨物种保守性[78], 而采用这一原则进行搜索就会漏掉许多真实的非保守靶基因. 因此, 人们开发了许多不依赖靶位点保守性的软件同样达到了较好的预测效果.

2006 年, Miranda 等人[79]一改前人的思路, 开发了一种全新的预测 microRNA 靶基因的软件——RNA22. 该软件基于模序方法进行 microRNA 靶基因预测, 同时不依赖物种间的保守性, 因此可以进行任何单一基因组的预测. 该算法的另一独特之处在于, 并没有从已知的 microRNA 出发去寻找能够与其结合的靶标, 而是从感兴趣的序列入手, 判断其是否为 microRNA 的靶标, 进而确定其被哪条 microRNA 所调控[80]. 他们在几个物种进行的实验显示, 许多先前

得出的关于 microRNA 靶基因预测的结论受到了挑战: 预测人类基因组中的一个 microRNA, 其仅靶向 3′UTR 区域的靶标即可有数千个之多. 同时, 他们认为, microRNA 还可以靶向包括基因编码区和 5′UTR 在内的其他部位, 并且这一观点现已得到了证明[69,81]. 该软件还可以进行 microRNA 前体的预测, 分析显示, microRNA 前体的数量远高于目前估计的数目. 表 6 列出了突破常规原则所进行 microRNA 靶基因预测方法的主要软件.

3.3 整合其他原则的识别方法

最近的一些算法将更多的特征整合进靶标预测方法中, 其中包括靶标分子的二级结构特征、microRNA 及靶基因的表达水平及靶基因的相关信息(如基因的功能、作用通路和相关疾病)等信息.

常规的规则中已经引入了对二级结构的分析, 主要做法是通过二级结构预测软件分析 microRNA 靶序列构成的二聚体结构, 计算双链的自由能, 认为具有稳定结构的位点更有可能成为真实的靶标. 然而, 真实的情况往往比想象中复杂. 已有研究显示, 虽然 microRNA 与目标序列构成良好的序列匹配及稳定的二级结构, 但是由于目标序列自身可以构成茎环结构, 导致 RISC 无法结合上去, 从而可以逃脱 microRNA 的调控[83]. 因此, 结合位点上、下游区域构成的二级结构也应作为靶标预测算法的考虑因素之一.

Long 等人[84]分析了靶标分子的二级结构对 microRNA 抑制效果的影响. 他们采用 Sfold 软件进行 mRNA 二级结构的预测, 建立了 microRNA 与结构位点之间的两步杂交互作模型. 采用该模型进行线虫中 let-7 microRNA 对其互作靶标 lin-41 的 3′UTR 区的多个突变体的抑制能力, 及其他线虫和果蝇中经过实验验证的 microRNA 靶标互作关系的分析. 结果显示, 靶标分子的二级结构对于 microRNA 识别靶基因具有重要影响, 建立基于结构模型的方法可以提高动物全基因组 microRNA 靶标分子识别算法的精确度.

Kertesz 等人[83]开发的 PITA 软件采用了自由参数模型计算了 microRNA-mRNA 二聚体的自由能和靶向非匹配 mRNA 所消耗能量之间的差别, 并考虑了协同靶向. 为每一个 microRNA 计算了靶标分值用以代表其与所有给定 UTR 上预测的可能靶标位点间的结合能力. 结果显示, 靶位点的突变会降低靶标分子的可趋近性(accessibility), 而靶位点的可趋近性对于 microRNA 分子的功能起重要作用.

表 7 列出了几款整合其他特征的 microRNA 靶标预测软件, 靶标预测的最终目的是揭示 microRNA 的功能. 因此, 新的方法在遵循靶标识别原则的基础上更多地整合了靶标分子的相关信息, 包括表达量、GO 功能类及参与的作用通路等.

3.4 ncRNA 靶标预测算法小结

寻找 ncRNA 基因的靶标分子, 对揭示其功能和具体作用机制至关重要. 就目前研究来看, 虽然在各

表 6 突破常规原则的 microRNA 靶基因预测软件

软件	适用范围	网址	网络版	本地版	参考文献
FindTar	人类、大鼠、小鼠、黑猩猩、牛	http://bio.sz.tsinghua.edu.cn/findtar/	√		[77]
RNA22	哺乳动物	http://cbcsrv.watson.ibm.com/rna22.html/	√	√	[79]
microTar	线虫、果蝇、小鼠	http://tiger.dbs.nus.edu.sg/microtar/		√	[82]

表 7 整合其他原则的 microRNA 靶标预测软件

软件	适用范围	网址	网络版	本地版	参考文献
PITA	人类、小鼠、果蝇、线虫	http://genie.weizmann.ac.il/pubs/mir07/mir07_prediction.html/	√	√	[83]
GenMir++	人类	www.psi.toronto.edu/genmir/		√	[85]

个物种中发现的 ncRNA 数量不少, 但是能够给出确实的靶标分子及其功能的很少, 这说明在基因识别和功能预测方面存在脱节现象. 目前, microRNA 的功能预测是研究的重点. microRNA 靶标预测的软件主要依靠 microRNA 与靶基因的互补性、配对稳定性及靶位点的保守性进行预测. 虽然这些特征对靶标结合至关重要, 但这些特征间的关系及如何影响 microRNA 与靶标分子的结合都不清楚[26]. 而且, 其他的特性也有可能构成影响, 同时靶位点的位置也需要放宽, 不仅局限在 mRNA 的 3′UTR 中, 而且有可能出现在其他区域. 目前的 microRNA 靶基因预测软件对于已知的 microRNA 靶基因有着很高的预测特异性和敏感度, 而对于未知的靶基因预测情况, 各软件之间的交集很小, 假阳性率也较高[86]. 这需要对 microRNA 与靶基因的作用机制进行更深入的研究, 使得预测时有更多的可信参数, 以此提高预测准确率. 另一方面, 随着实验技术的不断发展, 得到验证的 microRNA 与靶基因互作数据将不断累积, 这就为识别算法提供了良好的训练集数据, 使得算法得到不断的改进和优化.

4 机遇与挑战

21 世纪的 ncRNA 研究日新月异, 随着 Roche/454, Illumina/Solexa, ABI/SOLiD 等新一代测序技术的迅猛发展, 可以一次性获得海量的测序数据. 这为后基因组时代的 ncRNA 研究提供了强大的动力, 为计算 RNA 组学带来了新的机遇与挑战.

一方面, 新的实验技术的发展极大地扩展了计算 RNA 组学的研究思路及领域. 除了模式生物外, 人们将进行大量非模式生物的遗传与发育研究, 也可以根据需要设计各种正常细胞或组织与病理样品的转录组分析, 揭示与之相关的遗传特性或病理机制. 比较基因组学和比较转录组学将在不同的层次和更大的规模展开, 从而揭示出 ncRNA 及其介导的遗传信息表达调控在生物进化中的作用. 另外, 将不同的测序结果进行整合, 并结合基因组学与蛋白质组学的数据, 可以构建出整体的调控网络, 最终将各种计算机方法整合起来用以解决复杂的生物学问题.

另一方面, 各种基因组和转录组序列数据以指数级的形式增长, 这就对计算机硬件方面(如计算和储存能力不断扩展、更新和维护)提出新的要求. 如何有效增强数据库间的交互能力, 并降低维护成本、提高传输效率和保障信息安全, 是目前计算 RNA 组学面临的一大挑战.

ncRNA 研究是后基因组时代的一个重要科学前沿, 采用计算 RNA 组学与“实验 RNA 组学”相结合的方法, 将使 ncRNA 的探索更加广泛和深入, 最终揭示出一个完整的决定生命遗传信息表达和进化的“RNA 世界”及其调控网络.

致谢 感谢中山大学生物工程研究中心杨建华、邵鹏、廖建友、官道刚、陈华成与李俊豪提供相关资料并给予建设性意见.

参考文献

1 Birney E, Stamatoyannopoulos J A, Dutta A, et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature, 2007, 447: 799—816

2 Rosenbloom K R, Dreszer T R, Pheasant M, et al. ENCODE whole-genome data in the UCSC Genome Browser. Nucl Acids Res, 2010, 38: 620—625

3 Maeda N, Kasukawa T, Oyama R, et al. Transcript annotation in FANTOM3: mouse gene catalog based on physical cDNAs. PLoS Genet, 2006, 2: 62

4 屈良鹄. RNA 组学: 后基因组时代的科学前沿. 中国科学 C 辑: 生命科学, 2009, 39: 1—2

5 Pang K C, Stephen S, Engstrom P G, et al. RNAdb-a comprehensive mammalian noncoding RNA database. Nucl Acids Res, 2005, 33: 125—

130

6 Liu C, Bai B, Skogerbo G, et al. NONCODE: an integrated knowledge database of non-coding RNAs. Nucl Acids Res, 2005, 33: 112—115

7 Gardner P P, Daub J, Tate J G, et al. Rfam: updates to the RNA families database. Nucl Acids Res, 2009, 37: 136—140

8 Kin T, Yamada K, Terai G, et al. fRNAdb: a platform for mining/annotating functional RNA candidates from non-coding RNA sequences. Nucl Acids Res, 2007, 35: 145—148

9 Griffiths-Jones S, Saini H K, van Dongen S, et al. miRBase: tools for microRNA genomics. Nucl Acids Res, 2008, 36: 154—158

10 Griffiths-Jones S, Grocock R J, van Dongen S, et al. miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Res, 2006, 34: 140—144

11 Lestrade L, Weber M J. snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. Nucleic Acids Res, 2006, 34: 158—162

12 Betel D, Wilson M, Gabow A, et al. The microRNA.org resource: targets and expression. Nucleic Acids Res, 2008, 36: 149—153

13 Papadopoulos G L, Reczko M, Simossis V A, et al. The database of experimentally supported targets: a functional update of TarBase. Nucleic Acids Res, 2009, 37: 155—158

14 Yang J, Shao P, Zhou H, et al. deepBase: a database for deeply annotating and mining deep sequencing data. Nucleic Acids Res, 2010, 38: 123—130

15 Zuker M. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res, 2003, 31: 3406—3415

16 Zuker M, Stiegler P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucl Acids Res, 1981, 9: 133—148

17 Hofacker I L. Vienna RNA secondary structure server. Nucleic Acids Res, 2003, 31: 3429—3431

18 Chan C Y, Lawrence C E, Ding Y. Structure clustering features on the Sfold Web server. Bioinformatics, 2005, 21: 3926—3928

19 Ding Y, Chan C Y, Lawrence C E. Sfold web server for statistical folding and rational design of nucleic acids. Nucl Acids Res, 2004, 32: 135—141

20 Ding Y, Lawrence C E. A statistical sampling algorithm for RNA secondary structure prediction. Nucl Acids Res, 2003, 31: 7280—7301

21 Rivas E, Eddy S R. Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. Bioinformatics, 2000, 16: 583—605

22 Rivas E, Eddy S R. Noncoding RNA gene detection using comparative sequence analysis. BMC Bioinformatics, 2001, 2: 8

23 Washietl S, Hofacker I L, Stadler P F. Fast and reliable prediction of noncoding RNAs. Proc Natl Acad Sci USA, 2005, 102: 2454—2459

24 Pedersen J S, Bejerano G, Siepel A, et al. Identification and classification of conserved RNA secondary structures in the human genome. PLoS Comput Biol, 2006, 2: 33

25 Meyer I M, Miklos I. Statistical evidence for conserved, local secondary structure in the coding regions of eukaryotic mRNAs and pre-mRNAs. Nucl Acids Res, 2005, 33: 6338—6348

26 Solda G, Makunin I V, Sezerman O U, et al. An Ariadne's thread to the identification and annotation of noncoding RNAs in eukaryotes. Brief Bioinform, 2009, 10: 475—489

27 Bonnet E, Wuyts J, Rouze P, et al. Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. Bioinformatics, 2004, 20: 2911—2917

28 Reiche K, Stadler P. RNAstrand: reading direction of structured RNAs in multiple sequence alignments. Algorithm Mol Biol, 2007, 2: 6

29 Frith M C, Forrest A R, Nourbakhsh E, et al. The abundance of short proteins in the mammalian proteome. PLoS Genet, 2006, 2: 52

30 Dinger M E, Pang K C, Mercer T R, et al. Differentiating protein-coding and noncoding RNA: challenges and ambiguities. PLoS Comput Biol, 2008, 4: 1000176

31 Tupy J L, Bailey A M, Dailey G, et al. Identification of putative noncoding polyadenylated transcripts in *Drosophila melanogaster*. Proc

Natl Acad Sci USA, 2005, 102: 5495—5500
32 Lottaz C, Iseli C, Jongeneel C V, et al. Modeling sequencing errors by combining Hidden Markov models. Bioinformatics, 2003, 19: 103—112
33 Castrignanò T, Canali A, Grillo G, et al. CSTminer: a web tool for the identification of coding and noncoding conserved sequence tags through cross-species genome comparison. Nucleic Acids Res, 2004, 32: 624—627
34 Badger J H, Olsen G J. CRITICA: coding region identification tool invoking comparative analysis. Mol Biol Evol, 1999, 16: 512—524
35 Frith M C, Bailey T L, Kasukawa T, et al. Discrimination of non-protein-coding transcripts from protein-coding mRNA. RNA Biol, 2006, 3: 40—48
36 Liu J, Gough J, Rost B. Distinguishing protein-coding from non-coding RNAs through support vector machines. PLoS Genet, 2006, 2: 29
37 Kong L, Zhang Y, Ye Z, et al. CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. Nucleic Acids Res, 2007, 35: 345—349
38 Wang X, Zhang J, Li F, et al. MicroRNA identification based on sequence and structure alignment. Bioinformatics, 2005, 21: 3610—3614
39 Jiang P, Wu H, Wang W, et al. MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. Nucleic Acids Res, 2007, 35: 339—344
40 Huang T, Fan B, Rothschild M F, et al. MiRFinder: an improved approach and software implementation for genome-wide fast microRNA precursor scans. BMC Bioinformatics, 2007, 8: 341
41 Lowe T M, Eddy S R. A computational screen for methylation guide snoRNAs in yeast. Science, 1999, 283: 1168—1171
42 Schattner P, Barberan-Soler S, Lowe T M. A computational screen for mammalian pseudouridylation guide H/ACA RNAs. RNA, 2006, 12: 15—25
43 Yang J, Zhang X, Huang Z, et al. snoSeeker: an advanced computational package for screening of guide and orphan snoRNA genes in the human genome. Nucl Acids Res, 2006, 34: 5112—5123
44 Hertel J, Hofacker I L, Stadler P F. SnoReport: computational identification of snoRNAs with unknown targets. Bioinformatics, 2008, 24: 158—164
45 Lanz R B, McKenna N J, Onate S A, et al. A steroid receptor coactivator, SRA, functions as an RNA and is present in an SRC-1 complex. Cell, 1999, 97: 17—27
46 Ravasi T, Suzuki H, Pang K C, et al. Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome. Genome Res, 2006, 16: 11—19
47 Friedlander M R, Chen W, Adamidi C, et al. Discovering microRNAs from deep sequencing data using miRDeep. Nat Biotech, 2008, 26: 407—415
48 Taft R J, Glazov E A, Cloonan N, et al. Tiny RNAs associated with transcription start sites in animals. Nat Genet, 2009, 41: 572—578
49 Faulkner G J, Kimura Y, Daub C O, et al. The regulated retrotransposon transcriptome of mammalian cells. Nat Genet, 2009, 41: 563—571
50 Hackenberg M, Sturm M, Langenberger D, et al. miRanalyzer: a microRNA detection and analysis tool for next-generation sequencing experiments. Nucleic Acids Res, 2009, 37: 68—76
51 Enright A J, John B, Gaul U, et al. MicroRNA targets in *Drosophila*. Genome Biol, 2003, 5: 1
52 Aravin A A, Lagos-Quintana M, Yalcin A, et al. The small RNA profile during *Drosophila melanogaster* development. Dev Cell, 2003, 5: 337—350
53 Ashburner M, Ball C A, Blake J A, et al. Gene ontology: tool for the unification of biology. Nat Genet, 2000, 25: 25—29
54 Lewis B P, Shih I, Jones-Rhoades M W, et al. Prediction of mammalian microRNA targets. Cell, 2003, 115: 787—798
55 Lewis B P, Burge C B, Bartel D P. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell, 2005, 120: 15—20

56 Krek A, Grun D, Poy M N, et al. Combinatorial microRNA target predictions. Nat Genet, 2005, 37: 495—500

57 Doench J G, Sharp P A. Specificity of microRNA target selection in translational repression. Gene Dev, 2004, 18: 504—511

58 Grün D, Wang Y, Langenberger D, et al. MicroRNA target predictions across seven *Drosophila* species and comparison to mammalian targets. PLoS Comput Biol, 2005, 1: 13

59 Lall S, Grün D, Krek A, et al. A genome-wide map of conserved microRNA targets in *C. elegans*. Current Biology, 2006, 16: 460—471

60 Chen K, Rajewsky N. Natural selection on human microRNA binding sites inferred from SNP data. Nat Genet, 2006, 38: 1452—1456

61 Grimson A, Farh K K, Johnston W K, et al. MicroRNA targeting specificity in mammals: determinants beyond seed pairing. Mol Cell, 2007, 27: 91—105

62 Maragkakis M, Reczko M, Simossis V A, et al. DIANA-microT web server: elucidating microRNA functions through target prediction. Nucl Acids Res, 2009, 37: 273—276

63 Rehmsmeier M, Steffen P, Höchsmann M, et al. Fast and effective prediction of microRNA/target duplexes. RNA, 2004, 10: 1507—1517

64 Wang X, El Naqa I M. Prediction of both conserved and nonconserved microRNA targets in animals. Bioinformatics, 2008, 24: 325—332

65 Watanabe Y, Tomita M, Kanai A. Computational methods for microRNA target prediction. Meth Enzymol, 2007, 427: 65—86

66 Place R F, Li L, Pookot D, et al. MicroRNA-373 induces expression of genes with complementary promoter sequences. Proc Natl Acad Sci USA, 2008, 105: 1608—1613

67 Duursma A M, Kedde M, Schrier M, et al. miR-148 targets human DNMT3b protein coding region. RNA, 2008, 14: 872—877

68 Lee I, Ajay S S, Yook J I, et al. New class of microRNA targets containing simultaneous 5′-UTR and 3′-UTR interaction sites. Genome Res, 2009, 19: 1175—1183

69 Forman J J, Legesse-Miller A, Coller H A. A search for conserved sequences in coding regions reveals that the let-7 microRNA targets Dicer within its coding sequence. Proc Natl Acad Sci USA, 2008, 105: 14879—14884

70 Lai E C. Micro RNAs are complementary to 3' UTR sequence motifs that mediate negative post-transcriptional regulation. Nat Genet, 2002, 30: 363—364

71 Mallory A C, Reinhart B J, Jones-Rhoades M W, et al. MicroRNA control of PHABULOSA in leaf development: importance of pairing to the microRNA 5′ region. EMBO, 2004, 23: 3356—3364

72 Brennecke J, Stark A, Russell R B, et al. Principles of microRNA-target recognition. PLoS Biol, 2005, 3: 85

73 Lim L P, Lau N C, Garrett-Engele P, et al. Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. Nature, 2005, 433: 769—773

74 Selbach M, Schwanhäusser B, Thierfelder N, et al. Widespread changes in protein synthesis induced by microRNAs. Nature, 2008, 455: 58—63

75 Baek D, Villén J, Shin C, et al. The impact of microRNAs on protein output. Nature, 2008, 455: 64—71

76 Ha I, Wightman B, Ruvkun G. A bulged lin-4/lin-14 RNA duplex is sufficient for Caenorhabditis elegans lin-14 temporal gradient formation. Gene Dev, 1996, 10: 3041—3050

77 Ye W, Lv Q, Wong C A, et al. The Effect of Central Loops in microRNA: MRE Duplexes on the Efficiency of microRNA-Mediated Gene Regulation. PLoS ONE, 2008, 3: 1719

78 Farh K K, Grimson A, Jan C, et al. The widespread impact of mammalian microRNAs on mRNA repression and evolution. Science, 2005, 310: 1817—1821

79 Miranda K C, Huynh T, Tay Y, et al. A pattern-based method for the identification of microRNA binding sites and their corresponding heteroduplexes. Cell, 2006, 126: 1203—1217

80 夏伟, 曹国军, 邵宁生. MicroRNA 靶基因的寻找及鉴定方法研究进展. 中国科学 C 辑: 生命科学, 2009, 39: 121—128

81 Tay Y, Zhang J, Thomson A M, et al. MicroRNAs to Nanog, Oct4 and Sox2 coding regions modulate embryonic stem cell differentiation. Nature, 2008, 455: 1124—1128

82 Thadani R, Tammi M T. MicroTar: predicting microRNA targets from RNA duplexes. BMC Bioinformatics, 2006, 7: 20

83 Kertesz M, Iovino N, Unnerstall U, et al. The role of site accessibility in microRNA target recognition. Nat Genet, 2007, 39: 1278—1284

84 Long D, Lee R, Williams P, et al. Potent effect of target structure on microRNA function. Nat Struct Mol Biol, 2007, 14: 287—294

85 Huang J C, Babak T, Corson T W, et al. Using expression profiling data to identify human microRNA targets. Nat Meth, 2007, 4: 1045—1049

86 Xu X. Same computational analysis, different microRNA target predictions. Nat Methods, 2007, 4: 191