



长非编码 RNA 生物特征研究与分析

郭杏莉^①, 高琳^{①*}, 刘永轩^①, 党合莹^①, 王羽^①, 杨晓飞^①, 罗海涛^②

① 西安电子科技大学计算机学院, 西安 710071;

② 中国科学院计算技术研究所, 北京 100190

* 联系人, E-mail: lgao@mail.xidian.edu.cn

2012-11-05 收稿, 2013-01-30 接受

国家自然科学基金(60933009, 91130006, 61070137)和中央高校基本科研业务费(K5051223005, K5051223003)资助

摘要 长非编码 RNA 因其独特的生物特征和复杂的生物功能引起国内外广泛关注. 本文对长非编码 RNA 的生物特征进行研究和分析, 涉及长非编码 RNA 的鉴定、数据特征和功能特异性. 首先从生物技术及计算预测两个层面对长非编码 RNA 的鉴定方法进行详细阐述, 并列出现主要的长非编码 RNA 数据库. 其次, 分析长非编码 RNA 的生物特征: 核酸序列开放阅读框的长度、密码子偏好性、密码子替换频率、序列保守性等, 以及长非编码 RNA 二级结构中茎发夹结构的特异性出现, 并论述了长非编码 RNA 的功能特异性, 包括表达的时空特异性, 复杂的代谢调控机制, 与复杂疾病的关系. 最后指出并探讨长非编码 RNA 未来需进一步研究的问题和主要研究方向.

关键词

长非编码 RNA
开放阅读框
密码子偏好性
密码子替换频率
二级结构
功能预测

对基因组、转录组及表观遗传组等测序数据的研究发现哺乳动物基因组中编码蛋白质的区域只占基因组区域的 1%~2%, 大量的区域不编码产生蛋白质, 但会转录产生 RNA 分子, 其中一类长度大于 200 nt 的 RNA 分子称为长非编码 RNA(long non-coding RNA, lncRNA)^[1,2]. 长非编码 RNA 的基因结构类似 mRNA, 具有 5'端帽子结构、3'端聚腺苷酸尾巴以及剪接现象等类似于蛋白编码基因的特性^[3], 因此有时也将其称为 mRNA-like 的非编码 RNA^[4], 然而, 长非编码 RNA 在序列、结构和功能等各个层面都体现了其独特的生物特征.

研究表明长非编码 RNA 分子在生命过程的转录、调控等阶段发挥了重要的生物功能. 如 Huarte 等人^[5]观察到 lincRNA-p21 在 p53 抑制基因表达的调控通路中与 hnRNP-K 相互作用发挥重要的功能; Cesana 等人^[6]发现鼠的长非编码 RNA 分子 linc-MD1 作为竞争性的内源 RNA 通过影响 miRNA 与靶基因的结合, 改变肌肉的分化过程, 在后转录水平行使其

调控功能. 长非编码 RNA 分子被发现可以通过顺式或反式调控影响基因的表达, 其在与表观遗传有关的生命过程中发挥了重要作用^[7], 如著名的长非编码 RNA 分子 H19^[8], Xist^[9]和 Jpx^[10]被发现与 X 染色体失活有关, Kcnq1ot1 被发现在细胞核中表达, 并且通过组蛋白修饰等表观遗传机制调控靶基因的表达^[11]. 长非编码 RNA 与胚胎干细胞的多能性及分化有关^[12], 可以通过与 PRC2(polycomb repressive complex 2)等蛋白质复合物相互作用调控基因表达^[13]. 针对长非编码 RNA 复杂多样的生物功能, Guttman 和 Rinn^[14]分析探讨了其潜在的分子机制, 提出长非编码 RNA 与其他生物分子相互作用形成某种模块结构发挥其生物功能. 长非编码 RNA 与人类疾病关系密切, 研究证实其参与癌症等复杂疾病的发生与发展过程^[15,16]. 总之, 长非编码 RNA 作为独立转录的 RNA 分子, 具有广泛的生物功能, 这一观点正在被逐渐证实.

当前有关长非编码 RNA 的研究已经引起国内外的广泛关注, 对其认识和研究正逐步深入. 本文对长

引用格式: 郭杏莉, 高琳, 刘永轩, 等. 长非编码 RNA 生物特征研究与分析. 科学通报, 2013, 58: 2779–2786

Guo X L, Gao L, Liu Y X, et al. Long non-coding RNAs: Insights into the biological property (in Chinese). Chin Sci Bull (Chin Ver), 2013, 58: 2779–2786, doi: 10.1360/972012-1023

非编码 RNA 的生物特征进行分析综述, 重点探讨长非编码 RNA 的鉴定方法和数据库、长非编码 RNA 数据的生物特征以及功能特异性, 在此基础上展望长非编码 RNA 研究的发展趋势, 指出研究中的几个关键问题。

1 长非编码 RNA 鉴定及数据库

真核生物的物种复杂性与其非编码区域的大小成正相关, 非编码基因在生物体内发挥重要的生物功能^[1,2], 长非编码 RNA 分子因其独特的生物学特征和重要的生物功能引起了广泛关注。长非编码 RNA 的鉴定是当前开展研究的首要任务, 基于不同的生物技术手段和生物信息学算法提出了多种长非编码 RNA 鉴定方法, 其主要区别在于: (1) 基于什么样的生物技术和数据来发掘潜在的转录本; (2) 对于给定的转录本序列如何通过生物信息学分析预测其编码能力, 从而可以判断它为非编码基因。在此对主要的鉴定方法进行分析比较, 最后给出长非编码 RNA 数据库。

1.1 长非编码 RNA 发掘

针对长非编码 RNA 的发现和转录本构建, 不同研究者基于各种实验技术和生物数据提出了多种方法。例如, 全基因组的从头构建方法发现新的长非编码 RNA^[17,18], 基于 cDNA 文库和 EST 数据借助计算方法构建^[19], 基于染色质修饰特征^[3,13], 基于 Tiling array 技术^[20], 基于长非编码 RNA 芯片技术^[21], 基于 RNA-seq 测序技术^[22], 基于 RIP-seq 技术^[23]等, 都发现了潜在的长非编码 RNA。下面分析主要的几种方法。

(i) 基于 cDNA 文库和 EST 数据。长非编码 RNA 分子是独立转录的一类 RNA 分子, 具有独特的基因结构, 通过发掘相关的转录数据, 可以从中筛选长非编码 RNA, 首先构建全长 cDNA 文库, 结合表达序列标签数据(EST)发掘潜在转录单元, 基于基因组注释信息通过生物信息分析筛选潜在的、有功能的长非编码 RNA^[19,24]。

(ii) 基于染色质修饰特征。研究发现 RNA 聚合酶 II 起始转录的基因其启动子区域会观察到 H3K4me3 的修饰信号, 在整个转录区域会观察到 H3K36me3 的修饰信号^[25]。基于这一观察事实, Guttman 等人^[3]借助 ChIP-seq 测序数据在全基因组范围内, 通过计算分析从基因间区域筛选成对的“K4-

K36”修饰信号, 发现了约 1600 条小鼠的长非编码 RNA, 将其命名为 lincRNA。Khalil 等人^[13]用同样的方法从人的 6 个细胞系中发现了约 3300 条长非编码 RNA。

(iii) 基于 RNA-seq 数据分析。随着新一代测序技术的发展, 高通量测序数据大量涌现, 为全基因组范围研究基因的转录水平和生命活动提供了丰富信息, 对不同样本的测序数据进行生物信息学分析, 可以发现各种组织特异或者表型特异的长非编码 RNA, 如 Cabili 等人^[26]基于人的 24 个组织和细胞系的 RNA-seq 测序数据鉴定了约 8000 条长非编码 RNA。

1.2 长非编码 RNA 编码能力预测

转录本的构建是长非编码 RNA 鉴定的第一步, 进一步对转录本编码能力进行预测以发现潜在的长非编码 RNA。给定转录本的核苷酸序列, 预测编码能力, 从而确定其为 mRNA 或者长非编码 RNA, 其中涉及转录本的序列特征分析, 包括 ORF(open reading frame)预测、密码子替换频率和进化特征分析等, 所采用的计算模型包括 SVM 模型、概率模型等。以下针对主要的几种编码能力预测软件进行探讨。

(i) CPC 预测软件。CPC(coding potential calculator)^[27]由北京大学生物信息学中心开发, 采用支撑向量机(supporting vector machines, SVM)模型, 将编码能力预测建模成二分类问题, 通过已知编码基因与非编码基因的序列特征来训练 SVM 分类器。选取正负样本的 ORF 相关特征及与已知蛋白库序列比对结果的相关特征训练分类器, 通过计算得到一个参考分值来衡量其编码能力。

(ii) CSF 预测软件。CSF 由 Lin 等人^[28]提出, 通过判断核酸序列的密码子替换频率来鉴定编码序列或非编码序列。蛋白质为了在物种进化过程中保持其功能而受到自然选择压力, 表现出特有的密码子替换频率, 基于此, CSF 通过大量数据训练得到编码区和非编码区密码子替换频率矩阵, 基于两种密码子替换频率矩阵的概率模型来计算给定序列是属于编码区域或非编码区域。

(iii) PhyloCSF 预测软件。随后 Lin 等人^[29]通过引入比对序列的进化模型设计实现了 PhyloCSF 预测软件, 借助比对序列中包含的进化信息, 建立比对序列的密码子模型, 计算给定序列分别由两种密码子模型生成的概率值, 以此衡量其编码能力。

1.3 长非编码 RNA 数据库

随着实验技术和计算分析方法的发展,大量的长非编码 RNA 数据不断涌现,主要收录在各种综合数据库、非编码数据库和文献中。例如,RefSeq 数据库、Ensemble、UCSC 等综合数据库中都包含长非编码 RNA 数据。非编码 RNA 数据库的整理见表 S1。值得一提的是 NONCODE 数据库(第三版),是目前数据收集最全的非编码 RNA 数据库,包括 73372 条长非编码 RNA,涵盖了几乎所有鼠和人的长非编码 RNA 数据^[4]。文献中也陆续报道了基于高通量测序数据和生物信息分析发现的长非编码 RNA 数据^[3,13,26],Prensner 等人^[30]通过带 PolyA 的 RNA-seq 测序技术从 102 个前列腺组织和细胞系中新发现了 121 条与前列腺癌有关的长非编码 RNA。

2 长非编码 RNA 生物特征

对于长度小于 200 nt 的小非编码 RNA 生物特征已经开展了广泛的研究。小非编码 RNA 在序列和结构上存在很多标志性的特征,例如,miRNA 前体可以形成发卡结构,在酶的作用下,在茎结构处产生成熟的 miRNA。因此不同类型小非编码 RNA 预测、靶位点查找和功能分析等都取得了快速的进展。与之不同,长非编码 RNA 没有明显的、独特的生物特征,一般不能通过单一的方法鉴定和识别,其长度大于 200 nt,产生方式类似 mRNA,具有许多与其类似的特征,但从序列、结构到功能又体现了特异性。因此,本节通过与 mRNA 对照分析来论述长非编码 RNA 的生物特征,包括 ORF 长度、密码子偏好性、密码子替换频率和序列保守性等,并给出我们近期在长非编码 RNA 二级结构方面的一些研究结果。

2.1 ORF 长度

研究者通过对 FANTOM3 中的长非编码 RNA 进行分析,发现大多数都不存在较长的 ORF^[31],众所周知,编码基因转录本中的 ORF 会翻译产生氨基酸序列,进而合成蛋白质。因此这被作为一个区分编码基因与非编码基因的重要指标。在编码基因的鉴定中,Dinger 等人^[32]通过限定 ORF 长度来提高 mRNA 的预测准确性;在 FANTOM 项目计划中,包含不小于 300 nt 的 ORF 这一标准被用来鉴定新的 mRNA^[19]。与之相反,在长非编码 RNA 的鉴定中可以通过对

ORF 长度进行筛选得到潜在长非编码 RNA(ORF 长度不大于 300 nt)。也有例外,著名的长非编码 RNA *H19*, *Xist*, *Mirg*, *Gtl2* 和 *KcnqOT1* 都含有大于 100 个密码子的 ORF。另外翻译较小肽段的转录本也会被误以为非编码 RNA,如果蝇的 *tarsal-less* 基因转录产生长度为 1.5 kb 的转录本,其 ORF 的长度很短。随着 RNA 分子功能双向性的发现与研究^[33],这一传统论点正在接受事实的挑战。

2.2 密码子偏好性(codon usage bias)

在遗传密码中共有 64 个密码子,对应着 20 种氨基酸。除少数氨基酸仅由一个密码子编码之外,大多数可由 2~6 个密码子编码,编码同一种氨基酸的密码子被称为同义密码子。在蛋白质合成过程中,同义密码子的使用频率不同,这种现象叫密码子偏好性。对其解释大致可分为两类,其一基于自然选择理论,认为密码子偏好性与蛋白质表达的效率和准确性有关;其二基于变异理论,假设密码子偏好性源于变异模式中的非随机性,一些密码子更容易变异,因此其平均使用频率偏低^[34]。目前广为接受的是 selection-mutation-drift 模型,该模型同时考虑自然选择压力和变异作用^[35]。近期对人类长非编码 RNA 的密码子偏好性进行研究,发现其密码子偏好性与内含子区域相似,与 mRNA 的密码子偏好性不同^[36]。

2.3 密码子替换频率(codon substitution rate)

通过定义非同义/同义替换比率 ω 来量化核酸序列在进化过程中密码子所受到的自然选择压力。 $\omega=1$ 表明序列发生自然变异,即没受到自然选择压力; $\omega<1$ 表明序列受到纯化选择压力; $\omega>1$ 意味着序列受到了正向选择压力^[37]。如果一段核酸序列在很长的自然选择过程中被保留,这段序列上的纯化选择结果应该非常明显,研究结果显示绝大部分的编码序列都具有 $\omega>1$ 的特征,如 Endo 等人^[38]研究发现,3539 个蛋白质编码序列中只有 17(0.47%)个序列的 $\omega>1$ 。蛋白编码基因在进化过程中,对于可以保留蛋白质功能的变异(同义密码子替换)更容易保留下来,而改变蛋白质功能的变异(非同义密码子替换)则更倾向于被淘汰。编码能力预测软件 CSF 判断一个区域是否在进化过程中受到与蛋白质编码区域相同的选择压力,并以此为依据来确定该区域是否属于编码区域^[28]。Ponjavic 等人^[39]使用与此类似的度量

dRNA/dAR, 即一段核算序列的突变速率与其邻近的无自然选择压力的序列片断的突变速率之比, 对 3122 条来自 FANTOM 的小鼠长非编码 RNA 进行测试, 分别在(小鼠-大鼠)和(小鼠-人)的对照数据中得到突变速率之比为 0.899 和 0.948, 也就是相比无自然选择压力的中性进化, 其差异分别为 10%和 5%, 并且他们分析认为长非编码 RNA 中存在的非中性突变速率是因为作用于其上的纯化选择压力。

2.4 序列保守性

序列的保守性往往暗示着其可能具有生物功能。Guttman 等人^[3]对基因间的长非编码 RNA(lincRNA)进行序列分析, 发现 lincRNA 的外显子与随机选取的基因间区域相比具有明显的序列保守性, 但低于蛋白质编码基因的外显子区域, 高于其内含子区域; lincRNA 的启动子区域和剪接位置具有极高的保守性, 类似蛋白质编码基因。Pang 等人^[40]指出, 保守性的缺乏并不意味着功能的缺失, 如 Xist 和 Air 等已经被证实具有重要功能, 其保守性非常低, 相反, 这可能意味着它们在快速进化下承受着正向选择(positive selection)压力。

2.5 二级结构特征

核酸序列的二级结构中包含茎、发夹环、内环、突环和分支环等各种丰富的结构元件, 如图 S1 所示。我们比较分析 mRNA 和长非编码 RNA 二级结构中的发夹环, 研究分析长非编码 RNA 二级结构可

能具有的特异性。为此采用二级结构预测软件 RNAfold(http://rna.tbi.univie.ac.at/Vienna_RNA_Package, 该方法通过计算核酸序列折叠所产生的最小自由能来预测其最优二级结构), 预测了小鼠的 36070 条长非编码 RNA 和 35125 条 mRNA 的二级结构。分别选取了长度分布完全一致的 32321 条长非编码 RNA 和 mRNA 的二级结构预测结果进行统计分析, 研究发现: (1) 平均每条序列中各种茎长发夹结构的分布趋势在 mRNA 和长非编码 RNA 中是一致的, 平均每条长非编码 RNA 中茎长偏长的发夹结构数量比 mRNA 中多, 但是茎长较短的发夹结构数量较之 mRNA 中要少, 如图 S2 所示; (2) 各种茎长发夹结构在 32321 条 mRNA 和长非编码 RNA 中的分布趋势相同, 但是包含较长茎长发夹结构的长非编码 RNA 的数量较之 mRNA 的数量多, 包含较短茎长发夹结构的长非编码 RNA 的数量较之 mRNA 偏少, 其中包含有长茎(茎长大于 22)发夹结构的长非编码 RNA 有 143 条, 而 mRNA 有 56 条, 如图 1 所示。Maenner 等人^[41]在对 Xist 的研究中暗示长茎发夹结构可能与某些特异功能有关, 因此对长非编码 RNA 二级结构进行预测, 其二级结构中丰富的长茎发夹结构或许暗示了其功能特异性。

3 长非编码 RNA 功能特异性

随着研究不断深入, 大量的长非编码 RNA 序列数据和表达数据陆续产生, 其作为有功能的 RNA 分子这一观点正在被不断证实, 然而绝大多数长非编

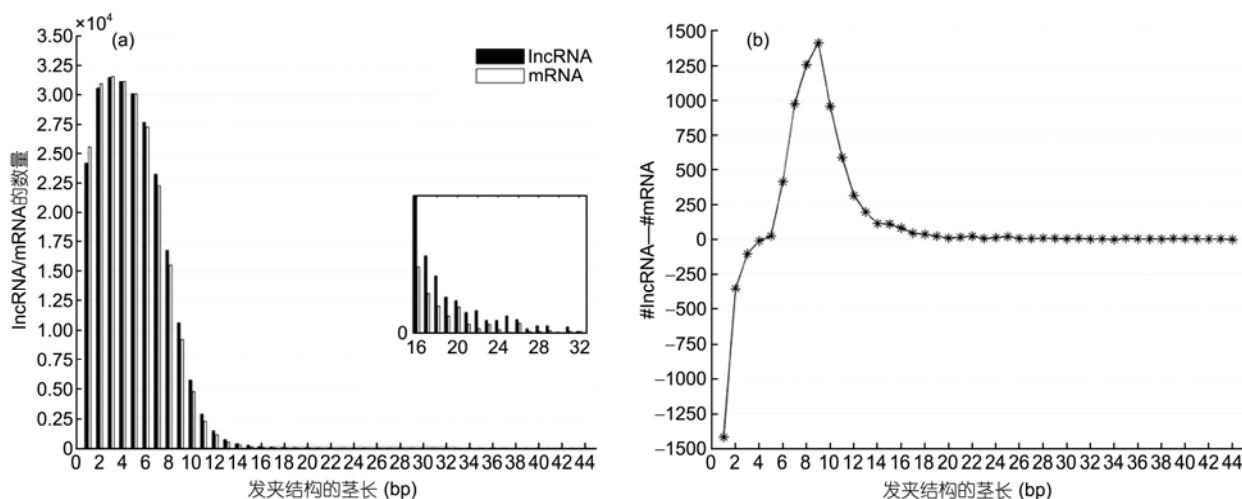


图1 不同茎长发夹结构在 lincRNA/mRNA 中的分布

码 RNA 的功能未知. 低通量的分子实验可以对其进行功能研究, 此外通过计算方法预测长非编码 RNA 的功能是当前一个重要的研究内容, 目前计算预测长非编码 RNA 的功能主要基于其与编码基因之间的共表达关系. 如 Guttman 等人^[3]基于 lincRNA 与编码基因在多个细胞系和多个组织中的表达数据构建 lincRNA 与编码基因之间的相关性矩阵, 通过双向聚类将 lincRNA 与某些基因集合关联起来, 对基因集合进行功能富集分析, 进而预测与之关联的 lincRNA 的功能, 结果显示长非编码 RNA 参与了多样的生物功能和调控过程. Liao 等人^[42]通过基因芯片重注释, 建立长非编码 RNA 与编码基因的共表达网络, 基于网络中的 hub 节点和模块来预测长非编码 RNA 的功能, 其预测结果表明长非编码 RNA 具有非常多样的生物功能.

研究表明长非编码 RNA 生物功能的多样性, 同时其生物功能也体现了特异性: (1) 长非编码 RNA 与特定的生物过程相关, 其表达具有时空特异性, 如长非编码 RNA 被发现于脑部^[43]和胚胎干细胞^[12]中特异表达, 在特定发育条件下表达^[44]; (2) 长非编码 RNA 具有复杂的调控功能, 其在染色质改变、转录调控及后转录调控中发挥重要作用^[2]; 作为重要的调控因子参与复杂的分子调控机制^[45]; (3) 长非编码 RNA 可能具有复杂的代谢机制, 如 Clark 等人^[46]基于芯片技术在全基因组范围对照分析了长非编码 RNA 与 mRNA 的稳定性, 发现大多数长非编码 RNA 是稳定的, 其半衰期的变化范围较大, 半衰期的平均值要小于 mRNA, 这可能暗示长非编码 RNA 具有更复杂的代谢机制和更广泛的功能; (4) 长非编码 RNA 与疾病存在密切关系, Howard 等人^[16]综述了长非编码 RNA 与疾病之间的关系, 探讨了其与癌症发展的关系, 并指出了长非编码 RNA 对于癌症诊疗和作为潜在药物靶标的意义^[15].

4 长非编码 RNA 研究的关键问题

新一代测序技术可以提供全基因组范围非编码基因的表达注释图谱, 这些数据是深入开展长非编码 RNA 研究的基础, 从中鉴定各种表型相关或者组织特异的长非编码 RNA, 对其进行定量分析, 发现隐藏的生物学知识和规律. 在此指出有关长非编码 RNA 的几个关键研究问题, 并对其研究发展趋势进行展望.

(i) 长非编码 RNA 的鉴定. 随着各种高通量测序数据的不断出现, 通过对转录组数据的定量和定性分析来发现新的长非编码 RNA 是当前一个重要的研究问题. 但是通过计算方法进行长非编码 RNA 的鉴定有以下几个难点和重点, 首先其基因结构类似 mRNA, 存在可变剪接现象, 一个长非编码 RNA 基因可以转录多个 RNA 分子, 发现多种可能的剪接形式对于准确鉴定长非编码 RNA 至关重要; 其次, 长非编码 RNA 的表达量相比 mRNA 较低, 当进行转录本构建的时候, 借助统计方法降低转录噪音以发现表达水平较低的转录本是转录本构建中亟待解决的重要难题; 再次, 长非编码 RNA 与 mRNA 在基因位置上存在重叠, 怎么正确的区分他们, 不混淆、不遗漏地发现长非编码 RNA 也是要解决的重要问题.

(ii) 序列特征发掘. 随着各种长非编码 RNA 数据的逐步积累, 对其核酸序列进行特征发现和模式识别, 以发现长非编码 RNA 所独有的、更本质的序列特征. 这有助于研究长非编码 RNA 的鉴定、分类及其所特有的转录和剪接机制等.

(iii) 二级结构预测. 通过对长非编码 RNA 二级结构的初步预测分析发现, 其二级结构相比 mRNA 体现了一定的特异性, 这启发着我们对其二级结构进行更深入的研究. Khaitan 等人^[47]的研究发现长非编码 RNA 分子 SPRY4-IT1 在黑色素瘤中表达上调, 并且与细胞凋亡和侵入机制有关, 其二级结构中包含一个特异的长发夹结构, 启示着特异性二级结构可能具有的特殊调控功能. 二级结构作为长非编码 RNA 的一个重要生物特征值得深入研究, 为进一步研究其功能机制提供指引.

(iv) 功能预测. 长非编码 RNA 的功能预测是一个重要研究内容, 是分子序列、结构与表型相关联的重要纽带, 目前大多数长非编码 RNA 的功能未知. 基于高通量测序数据开展功能预测研究, 构建长非编码 RNA 的功能图谱, 为通过分子实验研究长非编码 RNA 的功能提供有价值的参考.

(v) 调控关系预测. 长非编码 RNA 受到细胞内复杂调控机制的作用, 体现出特异表达^[2]; 另外一方面长非编码 RNA 与转录因子相互作用影响其他基因的表达^[12]. 通过计算方法预测长非编码 RNA 与其他分子之间的调控关系, 构建长非编码 RNA 的调控网络, 进一步认识长非编码 RNA 的分子机制; 通过 CHIP-seq, RIP-seq 等手段挖掘转录因子与长非编码

RNA之间的调控关系,构建长非编码RNA的调控网络.

(vi) 长非编码 RNA 生物网络构建. 哺乳动物的细胞中长非编码 RNA 与蛋白质复合物之间相互作用是一个普遍现象,它们通过相互作用共同行使特定的生物功能^[1],以长非编码 RNA 与其他生物分子的相互作用数据为基础,类似蛋白编码基因的蛋白质相互作用网络和调控网络,预测和构建长非编码 RNA 的相互作用网络,并构建其网络数据库是当前亟待解决的重要问题.以生物网络为基础,从系统生物学的观点深入开展长非编码 RNA 的研究.当前可行的方法:(1) 挖掘已有的表达数据及 RNA-seq 测序数据,构建长非编码 RNA 与其他分子的共表达网络,收集整理非编码 RNA 与其他分子的相互作用数据,构建各种通用的长非编码 RNA 参与的生物网络;(2) 有待高通量实验技术测定长非编码 RNA 与其他生物分子的相互作用数据;(3) 构建表型特异的长非编码 RNA 的生物网络,针对特定应用开展研究.

(vii) 长非编码 RNA 与疾病关系的研究. 在基因与疾病关系的研究中,非编码区域之前一直被忽略,近年来研究发现长非编码 RNA 与疾病之间关系密切^[1,15,16].长非编码 RNA BACE1-AS 被发现与阿尔茨海默病发生发展有关^[48],HOTAIR 被发现可以通过重编程染色质状态引发癌症的转移^[49],Gibb 等人^[50]通过 SAGE 技术对癌症组织和正常组织对照分析发

现大量可能与癌症相关的长非编码 RNA.通过对各种生物实验数据分析判断长非编码 RNA 中发生的 SNP 是否可以影响其功能,并进一步导致表型的变异和疾病的发生;或者从长非编码 RNA 的表达分析来判断其异常表达是否可以导致生物功能的改变,将是当前长非编码 RNA 与疾病关系研究的核心.因此,借助全基因组关联分析方法,基于高通量测序数据以及各种疾病相关的 SNP 数据等,长非编码 RNA 异常表达及基因区域内异常变异的发现以及长非编码 RNA 与疾病关系研究的开展,进一步促进了复杂疾病的分析诊疗,发现了潜在的药物靶标.

5 总结

长非编码 RNA 作为一类特殊的 RNA 分子,在生命体内发挥着重要的生物功能,引起了国内外的广泛关注,对其研究正逐步深入.为此本文分析论述了长非编码 RNA 的生物特征并指出几个关键研究问题.文章首先分析了主要的长非编码 RNA 鉴定方法和数据库;其次从核酸序列、二级结构两个层面对长非编码 RNA 的生物特征进行分析探讨,并对其生物功能特异性进行阐述;最后指出了长非编码 RNA 的几个关键研究问题,对其研究发展趋势进行展望.希望本文的分析可以促进长非编码 RNA 研究的深入开展.

参考文献

- 1 Khalil A M, Rinn J L. RNA-protein interactions in human health and disease. *Semin Cell Dev Biol*, 2011, 22: 359–365
- 2 Mercer T R, Dinger M E, Mattick J S. Long non-coding RNAs: Insights into functions. *Nat Rev Genet*, 2009, 10: 155–159
- 3 Guttman M, Amit I, Garber M, et al. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature*, 2009, 458: 223–227
- 4 Bu D, Yu K, Sun S, et al. Noncode v3.0: Integrative annotation of long noncoding RNAs. *Nucleic Acids Res*, 2012, 40: D210–D215
- 5 Huarte M, Guttman M, Feldser D, et al. A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. *Cell*, 2010, 142: 409–419
- 6 Cesana M, Cacchiarelli D, Legnini I, et al. A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA. *Cell*, 2011, 147: 358–369
- 7 Ponting C P, Oliver P L, Reik W. Evolution and functions of long noncoding RNAs. *Cell*, 2009, 136: 629–641
- 8 Bartolomei M S, Zemel S, Tilghman S M. Parental imprinting of the mouse *H19* gene. *Nature*, 1991, 351: 153–155
- 9 Brockdorff N, Ashworth A, Kay G F, et al. Conservation of position and exclusive expression of mouse *Xist* from the inactive X chromosome. *Nature*, 1991, 351: 329–331
- 10 Tian D, Sun S, Lee J T. The long noncoding RNA, *Jpx*, is a molecular switch for X chromosome inactivation. *Cell*, 2010, 143: 390–403
- 11 Redrup L, Branco M R, Perdeaux E R, et al. The long noncoding RNA *Kcnq1ot1* organises a lineage-specific nuclear domain for epigenetic gene silencing. *Development*, 2009, 136: 525–530
- 12 Guttman M, Donaghey J, Carey B W, et al. lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature*, 2011, 477: 295–300

- 13 Khalil A M, Guttman M, Huarte M, et al. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc Natl Acad Sci USA*, 2009, 106: 11667–11672
- 14 Guttman M, Rinn J L. Modular regulatory principles of large non-coding RNAs. *Nature*, 2012, 482: 339–346
- 15 Tsai M C, Spitale R C, Chang H Y. Long intergenic noncoding RNAs: New links in cancer progression. *Cancer Res*, 2011, 71: 3–7
- 16 Wapinski O, Chang H Y. Long noncoding RNAs and human disease. *Trends Cell Biol*, 2011, 21: 354–361
- 17 Hiller M, Findeiss S, Lein S, et al. Conserved introns reveal novel transcripts in *Drosophila melanogaster*. *Genome Res*, 2009, 19: 1289–1300
- 18 Rose D, Hiller M, Schutt K, et al. Computational discovery of human coding and non-coding transcripts with conserved splice sites. *Bioinformatics*, 2011, 27: 1894–1900
- 19 Okazaki Y, Furuno M, Kasukawa T, et al. Analysis of the mouse transcriptome based on functional annotation of 60770 full-length cDNAs. *Nature*, 2022, 420: 563–573
- 20 Li L, Wang X, Stolc V, et al. Genome-wide transcription analyses in rice using tiling microarrays. *Nat Genet*, 2005, 38: 124–129
- 21 Hung T, Wang Y, Lin M F, et al. Extensive and coordinated transcription of noncoding RNAs within cell-cycle promoters. *Nat Genet*, 2011, 43: 621–629
- 22 Zhang G, Guo G, Hu X, et al. Deep RNA sequencing at single base-pair resolution reveals high complexity of the rice transcriptome. *Nat Genet*, 2010, 20: 646–654
- 23 Zhao J, Ohsumi T K, Kung J T, et al. Genome-wide identification of polycomb-associated RNAs by RIP-seq. *Mol Cell*, 2010, 40: 939–953
- 24 Jia H, Osak M, Bogu G K, et al. Genome-wide computational identification and manual annotation of human long noncoding RNA genes. *RNA*, 2010, 16: 1478–1487
- 25 Mikkelsen T S, Ku M, Jaffe D B, et al. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, 2007, 448: 553–560
- 26 Cabili M N, Trapnell C, Goff L, et al. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev*, 2011, 25: 1915–1927
- 27 Kong L, Zhang Y, Ye Z Q, et al. CPC: Assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res*, 2007, 35: W345–W349
- 28 Lin M F, Carlson J W, Crosby M A, et al. Revisiting the protein-coding gene catalog of *Drosophila melanogaster* using 12 fly genomes. *Genome Res*, 2007, 17: 1823–1836
- 29 Lin M F, Jungreis I, Kellis M, et al. PhyloCSF: A comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics*, 2011, 27: i275–i282
- 30 Prensner J R, Iyer M K, Balbin O A, et al. Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. *Nat Biotechnol*, 2011, 29: 742–749
- 31 Nordström K J, Mirza M A, Almén M S, et al. Critical evaluation of the FANTOM3 non-coding RNA transcripts. *Genomics*, 2009, 94: 169–176
- 32 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: e1000176
- 33 Ulveing D, Francastel C, Hubé F. When one is better than two: RNA with dual function. *Biochimie*, 2011, 93: 633–644
- 34 Hershberg R, Petrov D A. Selection on codon bias. *Annu Rev Genet*, 2008, 42: 287–299
- 35 Bulmer M. The selection-mutation-drift theory of synonymous codon usage. *Genetics*, 1991, 129: 897–907
- 36 Niazi F, Valadkhan S. Computational analysis of functional long noncoding RNAs reveals lack of peptide-coding capacity and parallels with 3' UTRs. *RNA*, 2012, 18: 825–843
- 37 Yang Z H, Nielsen R, Goldman N, et al. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*, 2000, 155: 431–449
- 38 Endo T, Ikeo K, Gojobori T. Large-scale search for genes on which positive selection may operate. *Mol Biol Evol*, 1996, 13: 685–690
- 39 Ponjavic J, Ponting C P, Lunter G. Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome Res*, 2007, 17: 556–565
- 40 Pang K C, Frith M C, Mattick J S. Rapid evolution of noncoding RNAs: Lack of conservation does not mean lack of function. *Trends Genet*, 2006, 22: 1–5
- 41 Maenner S, Blaud M, Fouillen L, et al. 2-D structure of the A region of *Xist* RNA and its implication for PRC2 association. *PLoS Biol*, 2010, 8: e1000276
- 42 Liao Q, Liu C, Yuan X, et al. Large-scale prediction of long non-coding RNA functions in a coding–non-coding gene co-expression network. *Nucleic Acids Res*, 2011, 39: 3864–3878

- 43 Mercer T R, Dinger M E, Sunkin S M, et al. Specific expression of long noncoding RNAs in the mouse brain. *Proc Natl Acad Sci USA*, 2008, 105: 716
- 44 Amaral P P, Mattick J S. Mattick. Noncoding RNA in development. *Mamm Genome*, 2008, 19: 454–492
- 45 Wilusz J E, Sunwoo H, Spector D L. Long noncoding RNAs: Functional surprises from the RNA world. *Genes Dev*, 2009, 23: 1494–1504
- 46 Clark M B, Johnston R L, Inostroza-Ponta M, et al. Genome-wide analysis of long noncoding RNA stability. *Genome Res*, 2012, 22: 885–898
- 47 Khaitan D, Dinger M E, Mazar J, et al. The melanoma-upregulated long noncoding RNA spry4-it1 modulates apoptosis and invasion. *Cancer Res*, 2011, 71: 3852
- 48 Faghihi M A, Modarresi F, Khalil A M, et al. Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of β -secretase. *Nat Med*, 2008, 14: 723–730
- 49 Gupta R A, Shah N, Wang K C, et al. Long non-coding RNA hotair reprograms chromatin state to promote cancer metastasis. *Nature*, 2010, 464: 1071–1076
- 50 Gibb E A, Vucic E A, Enfield K S S, et al. Human cancer long non-coding RNA transcriptomes. *PLoS One*, 2011, 6: e25915

Long non-coding RNAs: Insights into the biological property

GUO XingLi¹, GAO Lin¹, LIU YongXuan¹, DANG HeXuan¹, WANG Yu¹,
YANG XiaoFei¹ & LUO HaiTao²

¹School of Computer Science and Technology, Xidian University, Xi'an 710071, China;

²Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China

Long non-coding RNAs (lncRNAs) draw a widespread attention on their study because of their specific properties and complex functions. Here we analyze the biological properties of lncRNAs, describing their identification, properties of their data and their roles in cells, and the key problems in the study of lncRNAs are brought forward. First, the methods applied into the identification of lncRNAs are analysed and compared carefully according to the two aspects, including the biological technology and data used to construct transcripts, and the computational methods to predict the coding potential of a transcript. The databases of lncRNAs are also listed. Then the biological properties of lncRNAs are analyzed and discussed, including the length of Open Reading Frame (ORF), codon usage bias and codon substitution rate in their nucleotide sequences, and our study on their secondary structure is also discussed. Next, specific functions of lncRNAs are also described and analyzed. Finally, the key problems in the study of lncRNAs are brought forward and the progress in their future study is also indicated.

long non-coding RNA(lncRNA), Open Reading Frame(ORF), codon usage bias, codon substitution rate, secondary structure, function prediction

doi: 10.1360/972012-1023

补充材料

图 S1 RNA 二级结构元件

图 S2 每条序列中平均发夹结构的分布

表 S1 长非编码 RNA 数据库

本文的以上补充材料见网络版 csb.scichina.com. 补充材料为作者提供的原始数据, 作者对其学术质量和内容负责.

附录:

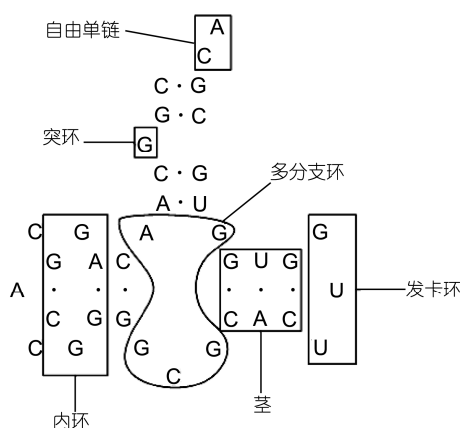


图 S1 RNA 二级结构元件

图 S1 中序列末端没有形成配对的单链叫做自由单链(Unstructured single strand), 连续的碱基配对相互堆积构成茎(Stem), 不形成互补配对的单链结构称为环, 依据形态的不同环又可分为发卡环(hairpin loop)、内环(internal loop)、突环(bulge loop)和多分支环(multibranched loop)等. 茎和环一起形成发夹样的结构被简称茎环(stem-loop)或者发卡环(Hairpin).

RNAfold 预测 RNA 的二级结构

(1) RNAfold 的工作原理: 该方法基于热力学(Thermodynamics)原理, 假定生物体内的环境为恒定, 以其自由能最低(MFE)的 RNA 二级结构最为稳定作为总原则, 通过在解空间中搜索自由能最低的二级结构来预测给定序列的二级结构.

(2) RNAfold 的使用: 在我们的研究中, 从网站 <http://www.tbi.univie.ac.at/~ivo/RNA/> 下载 RNAfold 的程序包, 在命令行直接运行程序来预测 RNA 的二级结构, 具体命令如下:

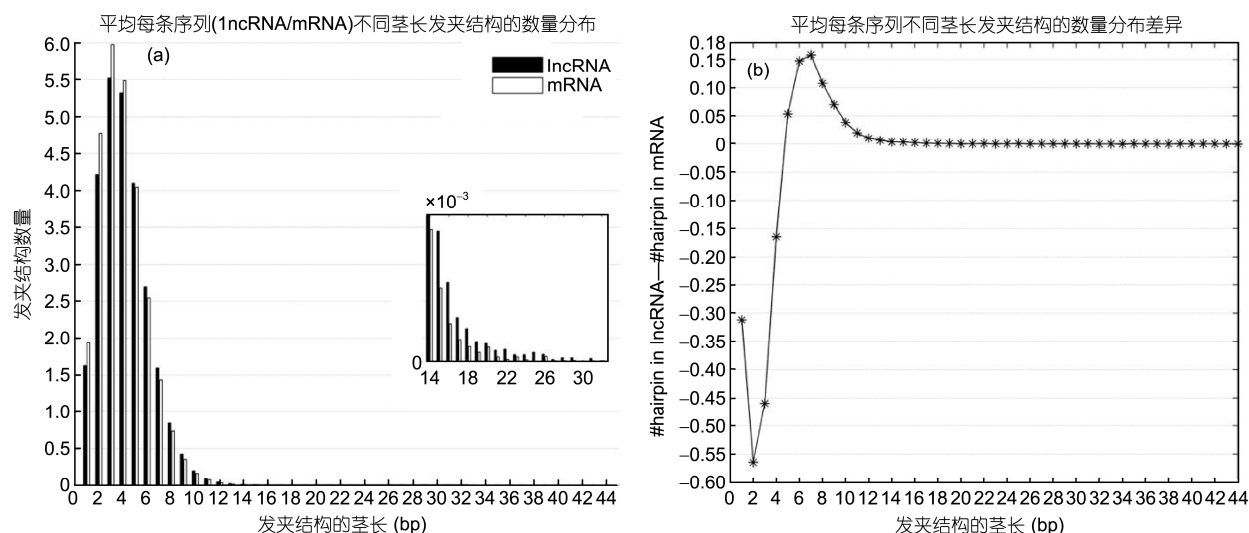


图 S2 每条序列中平均发夹结构的分布

RNAfold.exe <seq.fasta>result.txt

其中“<”代表序列以文件形式作为输入，允许序列的批处理，输入文件中存放多条待预测的 fasta 序列；“>”代表结果以文件形式输出，预测结果存放在 result.txt 中。

另外，RNAfold 也提供在线服务，网址为：<http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi>。

表 S1 长非编码 RNA 数据库

数据库名称	链接	参考文献
RNAdb	http://research.imb.uq.edu.au/rnadb/	[1]
FANTOM3	http://fantom3.gsc.riken.jp/	[2,3]
H-Invitational	http://h-invitational.jp/hinv/ahg-db/index.jsp	[4]
NONCODE	http://www.noncode.org/NONCODERV3/	[5]
lncRNAdb	http://lncrnadb.com/	[6]
fRNAdb	http://www.ncrna.org/frnadb/index.html	[7]
Rfam	http://rfam.sanger.ac.uk/	[8]
NRED	http://jsm-research.imb.uq.edu.au/nred/cgi-bin/ncrnadb.pl	[9]
ncRNAdb	http://biobases.ibch.poznan.pl/ncRNA/	[10]

参考文献

1 Ken C. Pang, Stuart Stephen, Pär G. Engström, Khairina Tajul-Arifin, Weisan Chen, Claes Wahlestedt, Boris Lenhard, Yoshihide Hayashizaki and John S. Mattick. RNAdb—a comprehensive mammalian noncoding RNA database. *Nucleic Acids Research*. 33(Database issue): D125–D130, 2005

2 Hayashizaki Y, Carninci P. Genome Network and FANTOM3: assessing the complexity of the transcriptome. *PLoS Genetics*. 2(4): e63, 2006

3 Maeda N, Kasukawa T, Oyama R, Gough J, Frith M, Engström PG, Lenhard B, Aturaliya RN, Batalov S, Beisel KW, Bult CJ, Fletcher CF, Forrest AR, Furuno M, Hill D, Itoh M, Kanamori-Katayama M, Katayama S, Katoh M, Kawashima T, Quackenbush J, Ravasi T, Ring BZ, Shibata K, Sugiura K, Takenaka Y, Teasdale RD, Wells CA, Zhu Y, Kai C, Kawai J, Hume DA, Carninci P, Hayashizaki Y. Transcript annotation in FANTOM3: mouse gene catalog based on physical cDNAs. *PLoS Genetics*. 2(4), e62, 2006

4 Yamasaki C, Murakami K, Fujii Y, Sato Y, Harada E, Takeda J, Taniya T, Sakate R, Kikugawa S. and Shimada M. The H-Invitational Database (H-InvDB), a comprehensive annotation resource for human genes and transcripts. *Nucleic Acids Research*. 36(Database issue): D793–D799, 2008

5 Dechao Bu, Kuntao Yu, Silong Sun, Chaoyong Xie, Geir Skogerbo, Ruoyu Miao, Hui Xiao, Qi Liao, Haitao Luo, Guoguang Zhao, Haitao Zhao, Zhiyong Liu, Changning Liu, Runsheng Chen and Yi Zhao. NONCODE v3. 0: integrative annotation of long noncoding RNAs. *Nucleic Acids Research*. 40(D1): D210–D215, 2012

6 Paulo P. Amaral, Michael B. Clark, Dennis K. Gascoigne, Marcel E. Dinger and John S. Mattick. lncRNAdb: a reference database for long noncoding RNAs. *Nucleic Acids Research*. 39(Database issue): D146–D151, 2011

7 Taishin Kin, Kouichirou Yamada, Goro Terai, Hiroaki Okida, Yasuhiko Yoshinari, Yukiteru Ono, Aya Kojima, Yuki Kimura, Takashi Komori and Kiyoshi Asai. fRNAdb: A platform for mining/annotating functional RNA candidates from non-coding RNA sequences. *Nucleic Acids Research*. 35(Database issue): D145–D148, 2007

8 Paul P. Gardner, Jennifer Daub, John G. Tate, Eric P. Nawrocki, Diana L. Kolbe, Stinus Lindgreen, Adam C. Wilkinson, Robert D. Finn, Sam Griffiths-Jones, Sean R. Eddy and Alex Bateman. Rfam: updates to the RNA families database. *Nucleic Acids Research*. 37(Database issue): D136–D140, 2009

9 Marcel E. Dinger, Ken C. Pang, Tim R. Mercer, Mark L. Crowe, Sean M. Grimmond and John S. Mattick. NRED: A database of long noncoding RNA expression. *Nucleic Acids Research*. 37(Database issue): D122–D126, 2009

10 Maciej Szymański, Volker A. Erdmann, and Jan Barciszewski. Noncoding regulatory RNAs database. *Nucleic Acids Research*. 31(1): 429–431, 2003