PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2013, 40(7): 634~640
www.pibb.ac.cn

新型长链非编码 RNA(lncRNA)的生物信息学研究进展

原佳沛　张浩文　鲁　志*
(生物信息学教育部重点实验室，清华大学生命科学学院，北京 100084)

摘要　非编码 RNA (noncoding RNA，ncRNA)是指不被翻译成蛋白质的一类 RNA，近几年来关于它们的功能研究越来越引起人们的重视．现在已经发现了一些中小型 ncRNA，比如 microRNA、snoRNA、tRNA 等，但是关于长 ncRNA(lncRNA)的研究还不够完善．本篇综述回顾了 ncRNA 特别是 lncRNA 的生物信息学研究进展，包括它们的研究历程、基本特点、与疾病的关系，以及对已有的预测非编码 RNA 的计算机方法进行了分析和比较，并且介绍了利用机器学习模型整合新一代高通量测序数据的方法．

关键词　非编码 RNA，长链非编码 RNA，非编码 RNA 预测，机器学习
学科分类号　Q6
DOI: 10.3724/SP.J.1206.2013.00266

1　非编码 RNA 及长非编码 RNA 简介

非编码 RNA(noncoding RNA，ncRNA)的发现历史可以追溯到 20 世纪 50 年代，在 1948 年，研究者普遍认为在给定的动物物种中，所有个体的每个细胞的 DNA 含量是恒定的[1]，直到 1951 年，由 Mirsky 和 Ris[2]证明了在同一物种内不同类型细胞的 DNA 含量不同．另外有其他证据证明物种间 DNA 含量差异非常大，并且 DNA 含量与组织的复杂度及基因数目没有关系，比如一些单细胞变形虫的基因组比人类基因组大 100 倍[3]．直到 1971 年，由 Thomas [4]提出 *C* 值悖论的概念，*C* 值悖论这个现象表明基因组除了基因和它们的调控序列外，可能还包含了相当大部分的其他类型的 DNA．*C* 值悖论与另一个令人费解的观察结果有关，称为“突变负荷”：考虑到人类基因的突变率，人类基因组似乎太大，如果整个人类基因组是功能性的(指受到选择压力)，我们每一代会产生太多的有害突变．在 20 世纪 70 年代早期，随着非编码 RNA 的发现，解决了 *C* 值悖论的问题，说明在真核生物中，基因组大小并不反映基因的数量，因为大部分的 DNA 序列是非编码的，根据 2008 年的最新文献报道，证明了人类的基因只占整个基因组的 2%[5]．在 1970 年以后，关于非编码转录的研究开始出现，50%的异核型 RNA (hnRNA)被证明仅限于细胞核并且不编码蛋白质[6]．在 20 世纪 90 年代早期，具有基因特异性调控作用的长链非编码 RNA (lncRNA)被发现，比如 H19[7]和 Xist[8]．随着 20 世纪 90 年代末期全基因组技术的发展，在 2002 年“普遍转录”的概念被提出，尤其是伴随着芯片和新一代测序技术的发展，研究者们认为在发育过程中的不同时间点，人类基因组约有 70%～90%进行了转录[9]．另外一些染色质特征方面的证据，如 DNase Ⅰ过敏性、组蛋白修饰、转录因子结合等证明了在基因间隔区存在非编码转录本[10]．发展到 2005 年，“转录噪声”的观点引起了大家的普遍认同[11]，这个观点的零假设是非编码 RNA 不执行功能．近几年关于长链非编码 RNA 的功能研究非常多，也有更多的非编码 RNA 被证实具有各种不同的生物学功 能[12]．这些长链非编码 RNA 中包括：a．相对独立的不与编码基因重叠的 RNA，如

* 通讯联系人.
Tel: 010-62789217, E-mail: zhilu@tsinghua.edu.cn
收稿日期：2013-06-14，接受日期：2013-06-26

Xist、H19、HOTAIR[13]和 MALAT1[14]；b．天然反义转录本，如 Xist 和 Tsix 共同控制 X 染色体的失活[15]；c．假基因；d．长的内含子区非编码 RNA，如 COLDAIR[16]; e．与启动子联系的转录本或增强子 RNA，如 pasRNAs[17]和 eRNAs[18-19].

非编码 RNA(noncoding RNA，ncRNA)是所有不被翻译成蛋白质的功能性 RNA 的统称，短的只有 20 几个核苷酸，长的可以达到几千个核苷酸．近 10 年来，人们发现了很多种类的新型非编码 RNA，并揭示了一些非编码 RNA 在基因调控网络中发挥的重要作用．人类基因组中研究最清楚的是编码蛋白质的基因．然而，这些编码基因的外显子只占整个基因组的 1.5%，即使算上非编码区(untranslated region，UTR)比例也不过 2%．近年来，越来越多的研究表明，基因组中非编码蛋白质区段同样发挥着重要的作用[20]：不仅作用于正常的生长发育和生理过程，而且会参与和疾病有关的机制．比如，microRNA 通过降解 mRNA 或阻止转录造成基因沉默，在人类疾病特别是癌症中都有作用，它产生的表观遗传修饰和遗传缺陷和它们的加工过程构成一些疾病的普遍特点．然而，microRNA 只是冰山一角，还有很多不同的 ncRNA 可能也与人类疾病相关，比如核糖体小 RNA (snoRNA)、T-UCRs、piRNAs、长的基因间 ncRNA (lincRNAs)、多种类型拼接起来的 lncRNA.

长非编码 RNA(long non-protein coding RNA，lncRNA)是在真核生物中新发现的一类长度大于 200 个核苷酸、没有长阅读框架、但往往具有 mRNA 结构特征(帽式结构和 poly A 尾巴)的 RNA[21]. lncRNA 在基因组中存在普遍的转录现象，但较之编码蛋白质的基因，往往表达水平比较低[22]，lncRNA 自身的表达水平也受到转录及转录后调控机制的严密调节．它们都不编码蛋白质，被发现参与调节多种重要的细胞活动，如基因表达、招募染色质修饰物、调节 X 染色体失活(XIST)、基因组印记(H19)、蛋白质折叠和蛋白质活性等[21]．1996 年发现的人 Xist 基因就是这类 lncRNA 的代表．在女性中，它控制失活两条 X 染色体中的一条．使 X 染色体编码的蛋白质在两性生物中的表达量趋向一致．根据 lncRNA 与蛋白质编码基因的位置关系，可以分为 5 类[22]，分别是：a．正义；b．反义；c．双向；d．内含子；e．基因间隔区.

早期的研究阐明了 lncRNA 对邻近基因位点有顺式调控的作用，比如 AIR、XIST 和 Kcnqlot，它们招募染色质修饰物使邻近的位点沉默．随着 lncRNA HOTAIR 的发现，lncRNA 的反式调控作用也得到了验证[21]．除此之外，lncRNA 还可以作为分子伴侣调控蛋白质的构象和作为结构分子锚定蛋白质在细胞内的位置．lncRNA 调控基因表达的分子机制非常多样，不仅可以通过结合转录因子来激活或抑制靶基因的表达，还能参与组蛋白修饰、mRNA 拼接等．虽然目前已确定的 lncRNAs 很多，但对绝大部分 lncRNA 在生命活动过程中的具体调控机制及功能模式仍不清楚，有待进一步研究．它们的进化受到进化选择的限制，尽管这种表现相对较弱[23]．蛋白质编码基因的进化速度和表达水平具有负相关的关系：一般来说，高表达的基因比低表达的基因更为保守．这种负相关性在人类和小鼠的 lincRNA(如前所述，位于 intergenic 区的 lncRNA 又称 lincRNA)中也得到了验证，这种相关程度与具有相似序列保守性和大小的蛋白质编码基因是相似的.

2　ncRNA 及 lncRNA 与人类疾病的关系

在人类重大疾病发生机制和控制方面，RNA 的重要性也日趋明显．目前已发现很多复杂的疾病与非编码 RNA 有关：如致死性新生儿线粒体肌病、亚急性坏死性脑脊髓综合征等与 tRNA 有关[24-25]，自体免疫性疾病如红斑狼疮、斯耶格伦综合征等与 snRNA 有关[26]，一些自体免疫性疾病如硬皮病与 snoRNA 有关等[27-28]．近年来还发现，一些非编码 RNA 与肿瘤抑制和生成的细胞凋亡直接相关，其变异将导致细胞发育异常或产生癌变[29]. 例如：bantam RNA 和 mir-14 能直接或间接抑制细胞凋亡[30]，而 bantam 更可促进细胞增殖而起着癌基因的作用；内含子编码的 U76 snoRNA 水平的降低与胚胎细胞的大量凋亡相关[31]．在人慢性淋巴细胞白血病患者中，miRNA mir-15 和 mir-16 所定位染色体区域的缺失非常普遍，提示它们在人体中可能起着肿瘤抑制基因的作用[32].

最新发现越来越多的证据表明 lncRNA 的突变和异常调节与很多人类疾病相关[33]．lncRNA 一级结构、二级结构、表达水平的改变，以及同源的 RNA 结合蛋白是引发从神经退行症到癌症相关疾病的基础．最近发现的一种 lncRNA(HOTAIR)介导的染色质改变和癌症转移之间的关系进一步说明了 lncRNA 与人类疾病之间的关系[21]，HOTAIR 与 PRC2(polycomb repressive complex 2)的互作导致染

色质的三甲基化状态改变而加快癌细胞转移[34]. 这些 lncRNA 占了哺乳动物中 ncRNA 的大多数，并被揭示在非常多的重要生物过程中起着极其关键的调控作用[35].

3 预测 ncRNA 的方法比较

现阶段 ncRNA 的预测仍依赖于计算机方法[36]. 对于特定类型的 ncRNA，可以使用现成的特定数据库和预测方法，比如针对 tRNA 的有 tRNA-SE 和 GtRNAdb, 针对 snoRNA 的预测方法有 SnoScan、snoSeeker、snoGPS、snoReport 和 snoRNABase，预测 miRNA 的有 miRBase[37]. 预测全基因的 ncRNA，常用的方法有两类：一类是利用 RNA 二级结构的保守性，这类方法有 QRNA、DDBRNA 和 MSARI；第二类是把二级结构稳定性和保守性联系起来，成为更加综合的方法，例如 Infernal、RNAz、Evofold 和 REAPR(表 1).

以上方法虽然可以对特定类型 ncRNA 进行有效的预测，但还是不能满足预测 lncRNA 的要求，存在各种缺陷，比如 tRNA-SE 虽然可以依据 DNA 序列预测出 99%～100%的 tRNA，并且保证假阳性率小于 1.5×10^{-8}，但是受限于速度并且不能预测其他种类的 RNA[38]；GtRNAdb 提供了由 740 个物种预测得到的 74 000 个 tRNA 基因，提供的信息包括亚型、基因位点、一级序列、二级结构图、多序列比对[39]，但是这个数据库能够提供的基因组范围内的数据搜索能力还是有限的；SnoScan 可以成功地在很多不同的真核和古细菌中检测到甲基化诱导的 snoRNAs、snoGPS 可以在酵母和哺乳动物中鉴定出假尿苷化诱导的 snoRNAs[40]，但它们的应用环境只能局限在 UNIX 系统；QRNA 利用比较序列的算法预测有结构的 RNA 基因，它的主要思路是在两个同源序列中，测试观察到的替代模式[41]. 保守的编码区域往往显示同义替换的模式，而保守的有结构的 RNA 表现出一定的补偿性突变，这与一些配对碱基的二级结构是一致的. 但它只对有结构的 RNA 会表现出相对较高的可信度； ddbRNA 同样是基于比对预测保守的 ncRNA，分为序列和结构的比对，其中根据结构比对得到的预测结果敏感性更高[42]；MSARI 通过搜索同源的多序列比对集的反向互补区域，找到保守的 RNA 二级结构，进而预测 ncRNA，这是一个精度比较高的方法，然而对于单独的 ncRNA 并不适用[43]. Infernal 可以根据协同变化模型(covariance model)在全基因组的范围内扫描，但其结果被限制在已知的结构中，不能找到新类型的非编码 RNA[44]. RNAz 和 Evofold 的算法是基于结构保守性，能够成功预测成千上万个有结构的非编码 RNA. 然而，这些方法都是基于全基因组的序列比对上，但很大部分的基因组是比对不上的[45-47]. 虽然 REAPR 对已知的序列比对结果通过结构进行了重新比对，由此可以找到更多的有结构的非编码 RNA，但这个方法的前提仍然是全基因组序列比对，并且这个方法并不支持全基因组范围的扫描(表 1)[48].

Table 1 The comparison of prediction methods of ncRNA

表 1 ncRNA 的预测方法比较

预测方法	适用范围	优点	缺点
tRNA-SE	tRNA	敏感度高，假阳性率很小	速度慢，仅限于 tRNA
GtRNAdb	tRNA	提供的信息全面，包括亚型、基因位点、一级序列、二级结构图、多序列比对	仅限于 tRNA
SnoScan	snoRNA	可在很多不同的真核和古细菌中预测到甲基化诱导的 snoRNAs	信息单一，仅限于 snoRNA
snoGPS	snoRNA	可在酵母和哺乳动物中鉴定出假尿苷化诱导的 snoRNAs	信息单一，仅限于 snoRNA
miRBase	miRNA	提供综合全面的 miRNA 序列信息、注释信息和预测的基因靶标位点信息	仅限于 miRNA
QRNA，DDBRNA	全基因组的 ncRNA	预测保守的有结构的 ncRNA 时可信度较高	局限于已知结构
Infernal	全基因组的 ncRNA	内置共变模型，根据结构对 RNA 划分家族	依赖多物种序列比对信息，局限于已知结构
RNAz	全基因组的 ncRNA	成功地预测成千上万个有结构的非编码 RNA	依赖多物种序列比对信息
Evofold，REAPR	全基因组的 ncRNA	对已有的序列比对结果进行重新比对，可以找到更多的有结构的非编码 RNA	依赖多物种序列比对信息，不支持全基因组范围的扫描
incRNA	全基因的 ncRNA	结合了基因序列、结构、表达、修饰方面的数据，精确度高，覆盖面广	依赖多物种序列比对信息

这些依靠结构来寻找非编码 RNA 的整体思路是比较基因组学，并同时考虑 RNA 二级结构折叠自由能与二级结构的保守和稳定性．虽然部分发挥功能的 RNA 往往具有稳定的结构，但仅通过最低折叠自由能的计算不足以在全基因组范围内将 ncRNA 鉴别出来，这是因为特定序列的折叠自由能与在随机测试中相应的序列自由能差异不足以将具有稳定结构的非编码 RNA 识别出来[49]．因此不能局限于通过最低自由能寻找有结构的非编码 RNA．由于很多非编码 RNA 并没有常规非编码 RNA 那么稳定的二级结构或序列特征，所以基于二级结构稳定性和保守性的非编码 RNA 预测在策略上需要有所调整．

自 2008 年以来，新一代高通量芯片和测序技术快速发展，产生了大量的 RNA 表达数据[50]．新一代测序技术主要特点是测序通量高、测序时间和成本显著下降[51]．该技术的应用体现在不同层面上：a．全基因组范围内的重测序或者更有针对性的检测点突变和核苷酸多态性；b．染色体结构重组的映射(mapping)分析，其中包括拷贝数变异、染色体平衡易位断点和倒置；c．RNA-seq，类似于表达序列标签(EST)或连续分析的基因表达(SAGE)，对 mRNA 或小 RNA 进行深度测序，测到的读段数(reads)可以用来量化每个物种在广泛的动态范围内的基因表达，测到的序列本身也可以用来对基因组进行注释；d．深度测序经过亚硫酸氢盐(bisulfite)处理后的 DNA，对 DNA 甲基化进行大规模分析；e．染色质免疫沉淀测序(ChIP-Seq)，或全基因组 DNA- 蛋白质相互作用分析，方法是对染色质免疫共沉淀实验得到的 DNA 片段进行深度测序．目前，在已经推出的几种新一代测序平台中，Illumina/Solexa 测序平台上 RNA-seq 应用最广．伴随着新类型表达数据的产生，预测 ncRNA 的方法也得到了发展，比如 *incRNA*[36]的方法，它可以综合序列、结构和表达数据，其中表达数据来源于高通量芯片或测序技术．除此之外，该模型还包括其他特性，它可以把 ncRNA 从其他基因元件中分离出来(图 1)．这个方法与之前的方法相比，在确定 ncRNA 中具有更高的准确性，多种类型数据经整合后，相互补充产生了综合方法特有的优势，这是只利用单一类型数据所不能实现的．之后，利用这些数据，得到训练集，通过机器学习的方法(SVM 和 Random Forest)进行新的 ncRNA 的预测(图 2)．

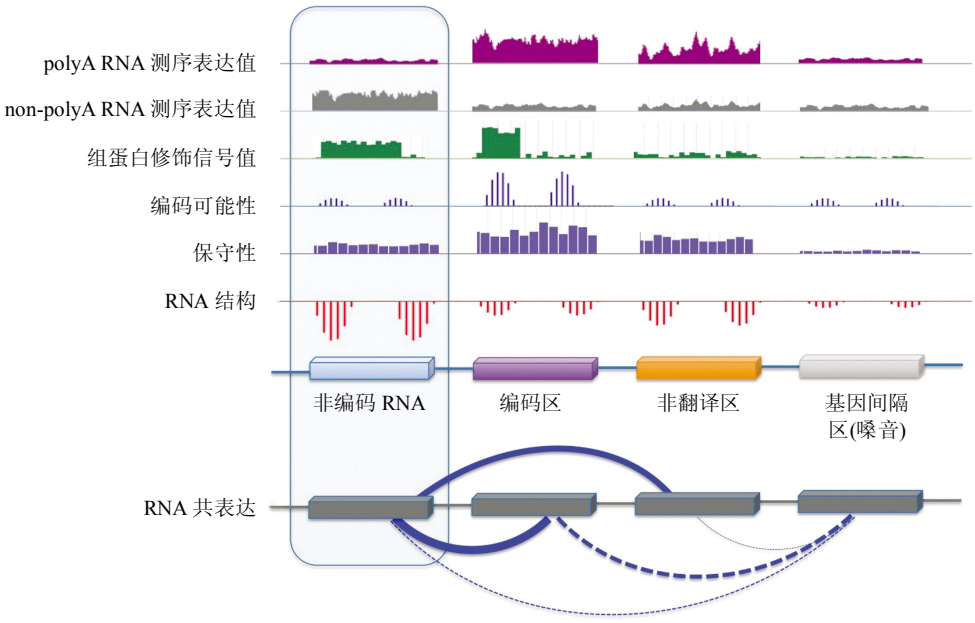

Fig. 1　Data integration for lncRNA characterization

图 1　利用数据整合对长链非编码 RNA 进行鉴定

该图展示了对长链非编码 RNA 进行鉴定时，采取的策略是收集不同类型的数据，包括 polyA RNA sequencing、nonpolyA RNA sequencing、表观遗传信号值、编码可能性、保守性和 RNA 结构等，并对其进行分析，例如 CDS 的 RNA-seq polyA 的表达值比较高，而 ncRNA 的 RNA-seq non-polyA 表达值比较高．通过对不同类型数据的整合，还可以进一步得到不同类型基因元素的网络调控关系．

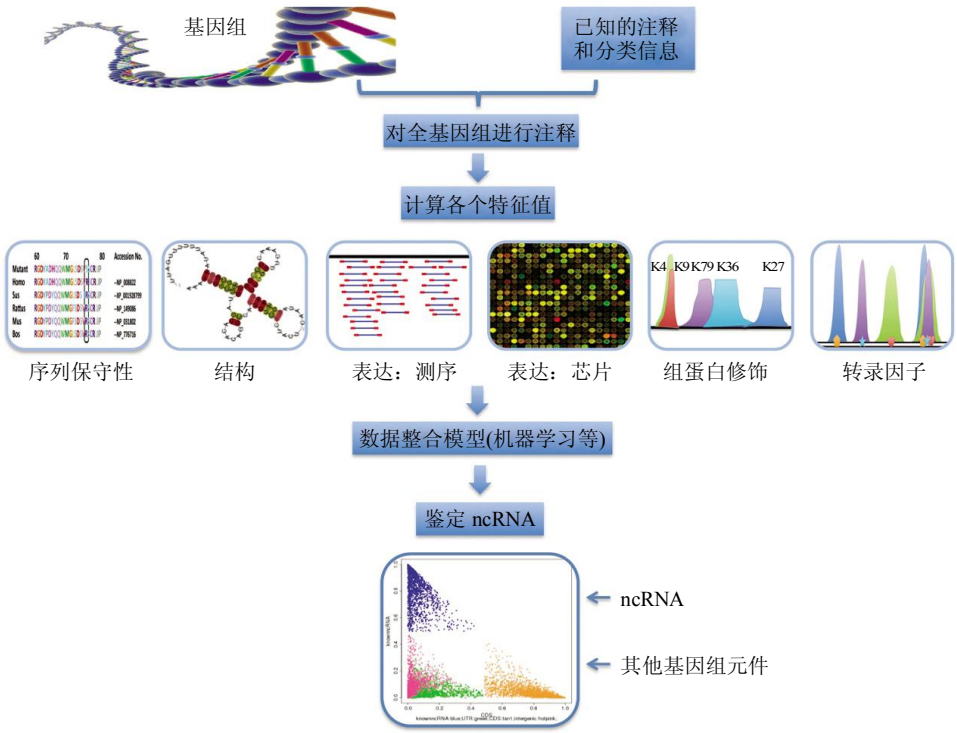

Fig. 2　The example of integrated analysis of lncRNA

图 2　lncRNA 综合分析方法流程示例

该图展示了对 lncRNA 进行综合分析的流程，首先将基因组划分成小的单位(bin)，根据 Gencode 的注释信息对每个 bin 进行注释后，分别计算每个 bin 的特征值，这些特征值包括序列保守性、结构稳定性、RNA 表达值、组蛋白修饰、转录因子结合等，进而利用机器学习的模型，将 lncRNA 与其他基因类别区分开，并且对新的 lncRNA 进行预测.

4　机器学习与非编码 RNA 的预测

有的时候我们的专业知识不足以完成分析和预测．尤其在面对高通量数据时，从中挖掘有用的信息尤为关键．这时可以用到机器学习(machine learning)的方法，令机器自动分析数据，比如特征提取或是分类．机器学习应用在生物信息学主要有两大分支，即监督学习(supervised learning)和非监督学习(unsupervised learning).

在监督学习问题中，每个数据拥有一个对应标签，我们希望通过数据建立一个模型，根据数据预测标签．传统的监督学习方法包括线性判别分析(LDA)、决策树(decision tree)、最近邻法(nearest neighbor)和神经网络(neural network)．20 世纪 90 年代后，诞生了一批很有影响力的工作，包括[52]支持向量机(SVM)、Adaboosting 和随机森林(random forest)，相比于传统的方法，上述方法更好地处理了过拟合(overfitting)的问题，从而在实际应用中有很好的预测效果.

近年来，这些方法被应用于 ncRNA 的预测．以 *incRNA*[36]为例，*incRNA* 首先将线虫的基因组分成一个个小区域，作为预测的基本单元．所有小区域中的一些已经被标注，这些被标注过的区域被进一步分为训练集和测试集．*incRNA* 整合了不同发育阶段的芯片和测序数据，以及此区域的结构信息(如果被转录成 RNA)，通过训练 SVM 和 Random Forest，可以识别大部分测试集中 ncRNA 的区域(图 2)；类似的思想在 microRNA 前体(microRNA precusor)的预测中也有应用，MiPred[53]以 miRNA registry database 中记录的 microRNA 前体为正样本，以编码蛋白区域提取的类似 microRNA 前体的颈环(hairpins)为负样本，综合了局部序列结构序列信息和最小自由能，使用 random forest 建立了预测模型，在这组数据集上的表现超过了 Triple-SVM 和 miR-abela．上述工作的共同特点是收集更多对分类有帮助的特征，使用机器学习算法自动对特征进行整合，实现分类.

非监督学习不再假设数据 x 拥有标签，而是自

动化寻找数据之间内在的关系. 常用的方法有主成分分析(PCA)、聚类分析(clustering)、独立成分分析(ICA)、自组织映射网(SOM)和隐马尔可夫链模型(HMM)等. Ernst 和 Kellis[54]使用隐马尔可夫链模型(HMM), 从组蛋白测序数据中自动识别出象征启动子(promoter)、增强子(enhancer)等基因调控元件的组蛋白图案. 在非编码 RNA 的分析和预测中也有非监督学习的应用. 一个使用非监督学习进行分类的例子是, 使用上下文敏感的隐马尔可夫链模型(context-sensitive HMM)预测 microRNA 前体[55]. 模型的构建是机器通过阅读已知的 microRNA 前体的序列, 去寻找这些序列的共同特征. 当新给一个未知序列时, 模型给出这个序列是 microRNA 前体的概率.

这些机器学习领域的最新突破可能会改变我们未来的建模思路, 极大推动生物信息学的发展.

综上所述, lncRNA 研究是基因组时代重要的科学前沿, 因为它有可能揭示一个全新的由 RNA 介导的遗传信息表达调控网络, 从不同于蛋白质编码基因的角度来注释和阐明基因组的结构与功能, 并为人类的疾病研究和治疗提供新的思路和方法. 同时, 新一代测序技术的发展也为鉴定 lncRNA 的计算机方法提供了强大的支持. 关于新型 lncRNA 有一些比较好的数据库资源(如 NONCODE[56]等)可以供大家进一步参考.

参 考 文 献

[1] Vendrely R, Vendrely C. The content of the cell nucleus in deoxyribonucleic acid through the organs, individuals and species. Experientia, 1948, **4**(11): 434−436

[2] Mirsky A E, Ris H. The desoxyribonucleic acid content of animal cells and its evolutionary significance. J Gen Physiol, 1951, **34**(4): 451−462

[3] Eddy S R. The C-value paradox, junk DNA and ENCODE. Curr Biol, 2012, **22**(21): R898−899

[4] Jr. Thomas C A. The genetic organization of chromosomes. Annu Rev Genet, 1971, **5**: 237−256

[5] Elgar G, Vavouri T. Tuning in to the signals: noncoding sequence conservation in vertebrate genomes. Trends Genet, 2008, **24** (7): 344−352

[6] Holmes D S, Mayfield J E, Sander G, *et al*. Chromosomal RNA: its properties. Science, 1972, **177**(4043): 72−74

[7] Brannan C I, Dees E C, Lngram R S, *et al*. The product of the H19 gene may function as an RNA. Mol Cell Biol, 1990, **10**(1): 28−36

[8] Brockdorff N, Ashworth A, Kay G F, *et al*. The product of the mouse Xist gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus. Cell, 1992, **71** (3): 515−526

[9] Okazaki Y, Furuno M , Kasykawa T, *et al*. Analysis of the mouse transcriptome based on functional annotation of 60 770 full-length cDNAs. Nature, 2002, **420**(6915): 563−573

[10] 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**(7235): 223−227

[11] Huttenhofer A, Schattner P, Polacek N. Non-coding RNAs: hope or hype?. Trends Genet, 2005, **21**(5): 289−297

[12] Kung J T, Colognori D, Lee J T. Long noncoding RNAs: past, present, and future. Genetics, 2013, **193**(3): 651−669

[13] Rinn J L, Kertesz M, Wang J K, *et al*. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell, 2007, **129**(7): 1311−1323

[14] Ji P, Diederichs S, Wang W, *et al*. MALAT-1, a novel noncoding RNA, and thymosin beta4 predict metastasis and survival in early-stage non-small cell lung cancer. Oncogene, 2003, **22** (39): 8031−8041

[15] Lee J T, Davidow L S, Warshawsky D. Tsix, a gene antisense to Xist at the X-inactivation centre. Nat Genet, 1999, **21**(4): 400−404

[16] Heo J B, Sung S. Vernalization-mediated epigenetic silencing by a long intronic noncoding RNA. Science, 2011, **331**(6013): 76−79

[17] Kanhere A, Viiri K, Araújo C C, *et al*. Short RNAs are transcribed from repressed polycomb target genes and interact with polycomb repressive complex-2. Mol Cell, 2010, **38**(5): 675−688

[18] Wang D, Garcia-Bassels I, Benner C, *et al*. Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. Nature, 2011, **474**(7351): 390−394

[19] Kim T K, Hemberg M, Gray J M, *et al*. Widespread transcription at neuronal activity-regulated enhancers. Nature, 2010, **465** (7295): 182−187

[20] Esteller M. Non-coding RNAs in human disease. Nat Rev Genet, 2011, **12**(12): 861−874

[21] Hung T, Chang H Y. Long noncoding RNA in genome regulation: prospects and mechanisms. RNA Biol, 2010, **7**(5): 582−585

[22] Ponting C P, Oliver P L, Reik W. Evolution and functions of long noncoding RNAs. Cell, 2009, **136**(4): 629−641

[23] Managadze D, Rogozin Z B, Chernikova D, *et al*. Negative correlation between expression level and evolutionary rate of long intergenic noncoding RNAs. Genome Biol Evol, 2011, **3**: 1390−1404

[24] Goto Y, Nonaka I, Horai S. A mutation in the tRNA(Leu)(UUR) gene associated with the MELAS subgroup of mitochondrial encephalomyopathies. Nature, 1990, **348**(6302): 651−653

[25] Chalmers R M, Lamont P J, Nelson I, *et al*. A mitochondrial DNA tRNA (Val) point mutation associated with adult-onset Leigh syndrome. Neurology, 1997, **49**(2): 589−592

[26] Marshak-Rothstein A. Toll-like receptors in systemic autoimmune disease. Nat Rev Immunol, 2006, **6**(11): 823−835

[27] Herrera-Esparza R, Kruse L, von Essen M, *et al*. U3 snoRNP associates with fibrillarin a component of the scleroderma clumpy nucleolar domain. Arch Dermatol Res, 2002, **294**(7): 310−317

[28] Yang J M, Hildebromdt B, Luderschmidt C, *et al*. Human scleroderma sera contain autoantibodies to protein components specific to the U3 small nucleolar RNP complex. Arthritis Rheum, 2003, **48**(1): 210−217

[29] Croce C M. Causes and consequences of microRNA dysregulation in cancer. Nat Rev Genet, 2009, **10**(10): 704−714

[30] Kumarswamy R, Chandna S. Inhibition of microRNA-14 contributes to actinomycin-D-induced apoptosis in the Sf9 insect cell line. Cell Biol Int, 2010, **34**(8): 851−857

[31] Newton K, Petfulski E, Tollervey D, *et al*. Fibrillarin is essential for

early development and required for accumulation of an intron-encoded small nucleolar RNA in the mouse. Mol Cell Biol, 2003, **23**(23): 8519-8527

[32] Garzon R, Calin G A, Croce C M. MicroRNAs in Cancer. Annu Rev Med, 2009, **60**: 167-179

[33] Wapinski O, Chang H Y. Long noncoding RNAs and human disease. Trends Cell Biol, 2011, **21**(6): 354-361

[34] Kogo R, Shimamura T, Mimori K, *et al*. Long noncoding RNA HOTAIR regulates polycomb-dependent chromatin modification and is associated with poor prognosis in colorectal cancers. Cancer research, 2011, **71**(20): 6320-6326

[35] Mercer T R, Dinger M E, Mattick J S. Long non-coding RNAs: insights into functions. Nature Reviews Genetics, 2009, **10**(3): 155-159

[36] Lu Z J, Yip K Y, Wang G, *et al*. Prediction and characterization of noncoding RNAs in *C. elegans* by integrating conservation, secondary structure, and high-throughput sequencing and array data. Genome Res, 2011, **21**(2): 276-285

[37] Griffiths-Jones S, Grocock R J, van Dongen S, *et al*. miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Res, 2006, **34**(Database issue): D140-144

[38] Lowe T M, Eddy S R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res, 1997, **25**(5): 955-964

[39] Chan P, Lowe T M. GtRNAdb: a database of transfer RNA genes detected in genomic sequence. Nucleic Acids Res, 2009, **37**(Database issue): D93-97

[40] Schattner A, Brooks N, Lowe T M. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. Nucleic Acids Res, 2005, **33**(Web Server issue): W686-689

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

[42] di Bernardo D, Down T, Hubbard T. ddbRNA: detection of conserved secondary structures in multiple alignments. Bioinformatics, 2003, **19**(13): 1606-1611

[43] Coventry A, Kleitman D J, Berger B. MSARI: multiple sequence alignments for statistical detection of RNA secondary structure. Proc Natl Acad Sci USA, 2004, **101**(33): 12102-12107

[44] Nawrocki E P, Kolbe D L, Eddy S R. Infernal 1.0: inference of RNA alignments. Bioinformatics, 2009, **25**(10): 1335-1337

[45] 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**(4): e33

[46] Washietl S, Hofacker I L, Stadler P F. Fast and reliable prediction of noncoding RNAs. Proc Natl Acad Sci USA, 2005, **102**(7): 2454-2459

[47] Washietl S, Pedersen J S, Korbel J O, *et al*. Structured RNAs in the ENCODE selected regions of the human genome. Genome Research, 2007, **17**(6): 852-864

[48] Will S, Yu M, Berger B. Structure-based whole genome realignment reveals many novel non-coding RNAs. Genome Research, 2013, **23**(6): 1018-1027

[49] Rivas E, Eddy S R. Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. Bioinformatics, 2000, **16**(7): 583-605

[50] Schuster S C. Next-generation sequencing transforms today's biology. Nat Methods, 2008, **5**(1): 16-18

[51] Shendure J, Ji H. Next-generation DNA sequencing. Nat Biotechnol, 2008, **26**(10): 1135-1145

[52] Hastie T, Tibshirani R, Friedman J. The Elements of Statistic Learning. 2nd. New York: Springer, 2009: 9-22

[53] 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**(Web Server issue): W339-344

[54] Ernst J, Kellis M. Discovery and characterization of chromatin states for systematic annotation of the human genome. Nat Biotechnol, 2010, **28**(8): 817-825

[55] Agarwal S, Vaz C, Bhattacharga A, *et al*. Prediction of novel precursor miRNAs using a context-sensitive hidden Markov model (CSHMM). BMC Bioinformatics, 2010, **11**(Suppl 1): S29

[56] Bu D, Yu K, Xie C, *et al*. NONCODE v3.0: integrative annotation of long noncoding RNAs. Nucleic Acids Res, 2012, **40**(Database issue): D210-215

Progress on Bioinformatic Research of lncRNA

YUAN Jia-Pei, ZHANG Hao-Wen, LU Zhi*

(*MOE Key Laboratory of Bioinformatics, School of Life Sciences, Tsinghua University, Beijing* 100084, China)

Abstract Noncoding RNA (ncRNA) refers to any RNA that is functional without being translated into proteins. Recently, people have paid more and more attention on various types of novel ncRNAs. Some small ncRNAs have already been identified and well studied, such as microRNAs, snoRNAs, tRNA, *etc*. However, the fields of long noncoding RNA (lncRNA) have not been well explored. This paper reviewed the research progress of ncRNA, especially the lncRNA, including the historical overview, the basic characteristics and the relationship with diseases. We also compared the prediction methods of ncRNA, and introduced an integrative strategy to combine the high throughput sequencing data in the prediction of ncRNA.

Key words noncoding RNA (ncRNA), long noncoding RNA (lncRNA), prediction of ncRNA, machine learning

DOI: 10.3724/SP.J.1206.2013.00266

*Corresponding author.

Tel: 86-10-62789217, E-mail: zhilu@tsinghua.edu.cn

Received: June 14, 2013 Accepted: June 26, 2013