PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2018, 45(1): 59~67
www.pibb.ac.cn

原核生物基因组肽编码 sORFs 分布及功能特征 *

陈宜亭[1,2)**] 张 凤[1,2)**] 赵 佳[1,3)**] 于家峰[1,2)***] 沙玉杰[1)] 王吉华[1,2)]

([1)] 山东省生物物理省级重点实验室，德州学院生物物理研究所，德州 253023；[2)] 德州学院物理与电子信息学院，德州 253023；[3)] 山东师范大学生命科学院，济南 250000)

摘要 近期，从非编码 RNA 中发现具有肽编码能力的小开放阅读框(sORFs)，激发了人们对这种长期被忽略的基因组元件的研究兴趣，sORFs 迅速成为当前重点研究领域．由于表达水平及丰度低、序列短等因素，对肽编码 sORFs 的有效研究方法及数据资源还很缺乏，现有研究仅集中在少数真核模式生物，对自然界中广泛存在的原核生物研究非常少，肽编码 sORFs 的发现为目前精准背景下的基因组注释提出严峻挑战．在此背景下，本文首先系统研究了 80 余种不同类型原核生物中长度小于 100 个氨基酸的肽编码 sORFs 分布及功能特征，并对不同长度区间 sORFs 的序列组成、分布及进化特征进行了对比分析．结果表明，肽编码 sORFs 在原核生物基因组普遍存在，随着序列长度的降低，其序列复杂度降低，行使的生物功能也相对集中．在此基础上，进一步结合当前肽编码 sORFs 研究现状，深入总结了肽编码 sORFs 研究存在的问题及挑战，为今后肽编码 sORFs 研究奠定了坚实理论基础．

关键词 小开放阅读框，原核生物基因组注释，蛋白质编码基因
学科分类号 Q61 **DOI**: 10.16476/j.pibb.2017.0109

2016 年，德克萨斯大学 Eric 教授团队在《科学》(*Science*)杂志报道了一种由位于 lncRNA 上的小开放阅读框(small open reading frame，sORF)编码的小肽 DWORF[1]，长度仅为 34 个氨基酸，在心肌收缩中发挥重要作用．DWORF 的发现引起了人们对 sORFs 及其编码多肽(sORFs encoded peptides，SEPs)这种长期被忽略的基因组元件的空前关注，激发了人们对非编码 RNA 的激烈争论，成为当前研究的重点领域[2]．一方面，能否编码蛋白质一直是区分非编码 RNA 与 mRNA 的金标准，而 DWORF 的发现引起人们对非编码 RNA 是否真的不编码蛋白质的讨论[3–5]；另一方面，SEPs 的发现打破了人们长期以来认为生物活性肽多是通过蛋白质前体修饰剪切而来的推论[6]．几十年来，sORFs 被认为不可能具备蛋白质编码能力，往往将它们归于错误预测为蛋白质编码基因的随机序列[7]，在之前数据库及蛋白质编码基因预测算法中通常将长度低于 100 个氨基酸的 sORFs 排除在外，进而降低基因预测结果的假阳性，但这样反而使得许多真正编码蛋白的 sORFs 被遗漏，进一步增加了预测结果的假阴性[6]．因此，肽编码 sORFs 的发现具有重要的学术意义，尤其在以“精准”为标签的组学测序快速发展的现在，肽编码 sORFs 的发现似乎表明人们在某种程度上对基因组的认识转了一圈又回到了原点．目前，已有大量研究证实了肽编码 sORFs 在生命活动中发挥重要生物学功能[8–9]，但由于其表达水平及丰度低、序列短、实验技术缺乏等诸多因素，对 sORFs 的研究还处于初级阶段[6, 10–11]，能够有效识别 sORFs 的生物信息算法及相关数据库资源也很缺乏[12–13]，其相关序列结构等生物特征认识亟待深入[11, 14]，多数研究集中在人、鼠、拟南芥等几种真核模式生物[6, 15]，对原核生物研究还较少[16–17].因此，本文借助已有数据资源，首次针对原核生物基因组中肽编码 sORF 序列组成及功能分布

* 国家自然科学基金(61771093, 61671107)和山东省自然科学基金重点项目(ZR2016JL027)资助.
** 并列第一作者.
*** 通讯联系人.
Tel: 0534-8982557, E-mail: jfyu1979@126.com
收稿日期：2017-10-13，接受日期：2017-11-10

特征开展了系统研究，并针对当前 sORFs 研究存在的问题开展了深入探讨，为今后原核生物 sORFs 研究提供了重要理论依据.

1 材料与方法

1.1 数据来源

目前仅有少数几个针对模式真核生物或特定区域的 sORFs 数据库发表[12-13, 18]，专门针对原核生物基因组的 sORFs 相关数据库资源鲜有报道. 尽管如此，基于转录组、蛋白质组等组学测序与生物信息技术，已有一些原核生物中的肽编码 sORFs 被发现并注释出来[16-17]，为本工作提供了良好的参考资源. 为了保障数据可靠性，本文所需肽编码 sORFs 信息(即 CDS 序列)从 RefSeq 数据库[19]获取.

1.2 数据集

基于 RefSeq 数据库，随机选取了 80 余种具有不同基因组 G+C 含量的原核生物作为研究对象. 为了进一步研究肽编码 sORFs 的生物学特征，我们根据 RefSeq 中的注释信息，将 sORFs 分为 3 个长度区间，即 50～100 个氨基酸、30～49 个氨基酸、29 个氨基酸及以下.

1.3 序列复杂度分析

为了展现 sORFs 编码多肽序列中氨基酸使用偏好特征，我们利用序列复杂度[20]概念来定量描述：

$$K=-\sum_{i=1}^{20} f_i \log_2 f_i$$

式中，i=1，2，3，…，20，表示氨基酸种类，f_i 表示序列中第 i 种氨基酸的频率. 根据统计原理，当 20 种氨基酸平均使用时，K 值最大，为 4.32；而当序列中只有 1 种氨基酸时，K 值最小，等于 0，此时说明序列中氨基酸使用偏好最强. 因此，K 值能够很好地反映序列中氨基酸使用信息. 本文中，将利用复杂度 K 值来进一步分析各原核基因组中肽编码 sORFs 序列组成特征.

1.4 氨基酸使用偏好分析

氨基酸组成是蛋白质及多肽的生物学功能基础，直接计算各种氨基酸百分含量 f_i 是研究蛋白质序列的重要方法. 为了有效描述蛋白质序列中各种氨基酸的使用特征，可直接通过计算 $C_i=f_i-0.05$ 来完成. 由于 20 种氨基酸的随机使用概率为 1/20=0.05，因此本文将各种氨基酸的使用偏好用上式来简单描述，若 C_i 大于零，则表示相应氨基酸偏好使用，反之不偏好.

1.5 氨基酸分布特征研究

为了描述氨基酸在蛋白质序列中的分布特征，本文利用我们提出的一种蛋白质序列分析圆柱体模型[21]，通过计算 $d_i=n_i/N$ 来描述，这里 n_i 表示任意蛋白质序列中第 i 种氨基酸的位置，N 表示序列长度，因此 $d_i \in [0, 1]$，直接计算

$$D_i=\sum d_i/N_i$$

N_i 表示序列中第 i 种氨基酸个数，因此 D_i=[D_1，D_2，D_3，…，D_{20}]定量描述了 20 种氨基酸在蛋白质序列中的分布中心.

2 结果与讨论

2.1 肽编码 sORFs 基因组分布特征

基于 RefSeq 中的注释信息，附件表 S1 给出了 77 种不同 G+C 含量的原核生物基因组中肽编码 sORFs 分布信息，其中百分比表示各基因组中肽编码 sORFs 占该基因组所有蛋白质编码基因的比例. 可以看到，各个基因组中均有小于 100 个氨基酸的 sORFs 被不同程度注释出来. 从长度分布来看，已经有注释信息的 sORFs 还多集中在 50 个氨基酸以上，其次是 30～49 个氨基酸之间的 sORFs，而且有些模式基因组中甚至已注释出了长度在 29 个氨基酸以下的 sORFs. 例如，在大肠杆菌 *E. coli* str. K-12(NC_000913)中，有 35 个长度小于 29 个氨基酸的 sORFs 被注释以来，其中最短的 sORFs 仅有 14 个氨基酸. 可见，随着基因组信息研究的深入，sORFs 不再是之前认为的不具备蛋白编码能力的随机序列，sORFs 也是普遍存在的基因组元件. 由于目前对于原核生物基因组中肽编码 sORFs 的相关研究很少，为了揭示原核生物中 sORFs 的相关特征规律，我们进一步以 6 种研究相对较为广泛、深入的大肠杆菌基因组为例开展了序列分析. 从表 1 给出的 6 种大肠杆菌菌株 sORFs 的注释情况来看，每个大肠杆菌菌株中都有长度小于 29 个氨基酸的 sORFs. 根据表 1，计算了各种长度分布 sORFs 对应 G+C 含量与其编码的多肽序列复杂度之间的散点图，结果见图 1. 可以发现，不同长度肽编码 sORFs 相应的 G+C 含量分布差别不大，但与基因组 G+C 含量(～50%)相比，分布比较离散. 而对序列复杂度而言，不同长度区间 sORFs 分布区域具有明显差异，50～100 个氨基酸 sORFs 序列复杂度最大，其次是 30～49 个氨基酸 sORFs，0～29 个氨基酸 sORFs 序列复杂度整体最低，3 个长

度区间序列复杂度平均值及标准偏差见表 2，各样本间方差分析 P 值均 <<0.05，表明相应序列中氨基酸使用具有不同程度偏好特征.

Table 1　Peptide coding sORFs length distribution in genome of E. coli

Species	Ref No.	Genomic GC/%	Genome size	Peptide coding sORFs distribution and percent			
				50～100 AA	30～49 AA	0～29 AA	Percent/%
E. coli O157:H7str. Sakai	NC_002695	50.5	5.5	671	37	4	13.69
E. coli UMN026	NC_011751	50.7	5.20	495	41	8	11.27
E. coli IAI39	NC_011750	50.6	5.13	434	34	12	10.16
E. coli str. K-12 substr. MG1655	NC_000913	50.8	4.64	346	41	35	10.19
E. coli O83:H1 str. NRG 857C	NC_017634	50.7	4.75	435	72	2	11.50
E. coli O104:H4 str. 2011C-3493	NC_018658	50.7	5.27	571	90	18	13.68

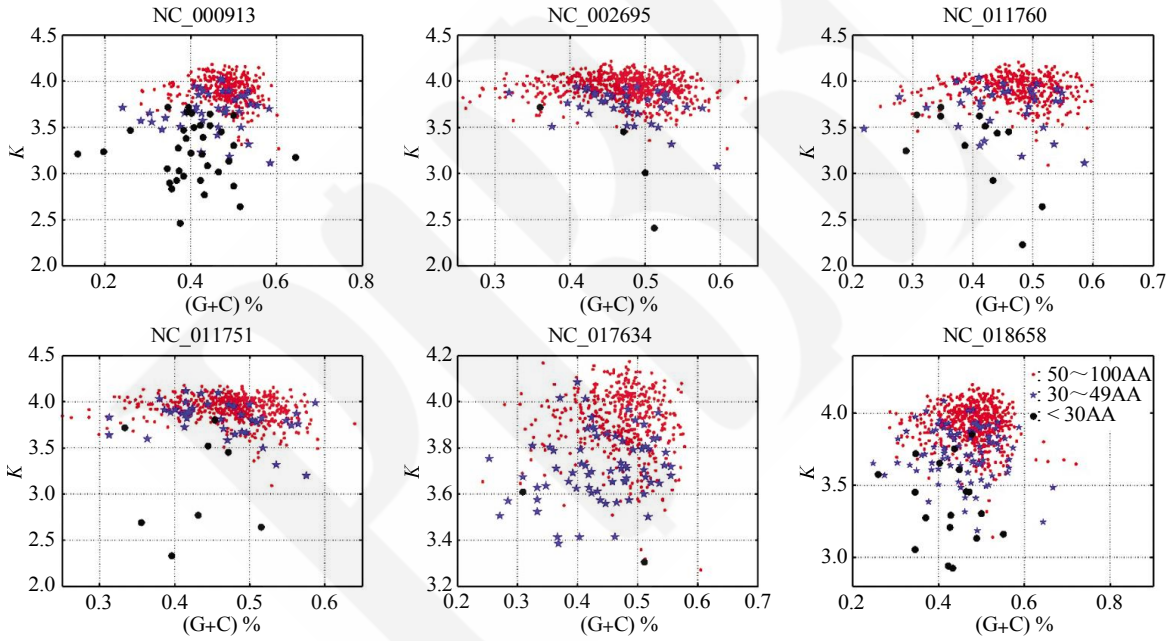

Fig. 1　Sequence (G+C) % and complexity of peptide coding sORFs

Table 2　Average and standard deviation of sequences complexity

Ref No.	50～100 AA	30～49 AA	0～29 AA
NC002695	3.93±0.13	3.73±0.18	3.15±0.57
NC011751	3.92±0.14	3.81±0.19	3.12±0.57
NC011750	3.92±0.15	3.71±0.25	3.28±0.46
NC000913	3.90±0.15	3.67±0.21	3.23±0.33
NC017634	3.91±0.14	3.72±0.15	3.46±0.21
NC018658	3.92±0.14	3.73±0.18	3.38±0.28

从信息熵的数理统计原理很容易理解图 1 中不同长度分布肽编码 sORFs 展现出的序列复杂度差异，而从生物学角度，则需要进一步揭示其背后的序列特征．为此，接下来本文又深入分析了各长度分布肽编码 sORFs 中氨基酸组成及分布特征，结果见图 2．为了便于观察和对比分析，图 2 横坐标中对各氨基酸按照其理化特征依次排列，即非极性氨基酸(A、V、L、I、F、W、M、P)、极中性氨基酸(G、S、T、C、Y、N、Q)、极正电氨基酸(H、K、R)和极负电氨基酸(D、E)．由图 2a 可以看出，不同长度 sORFs 在氨基酸使用偏好上具有一定程度的差别，就本文选择的 6 种大肠杆菌而言，3 种长度分布 sORFs 在非极性氨基酸使用上偏好性较为明显，且长度小于 29 个氨基酸的肽编码 sORFs

氨基酸使用频率浮动范围较广，展现的更明显的氨基酸使用偏好特征．而由图 2b，3 种长度分布 sORFs 中氨基酸分布差别相对较大．

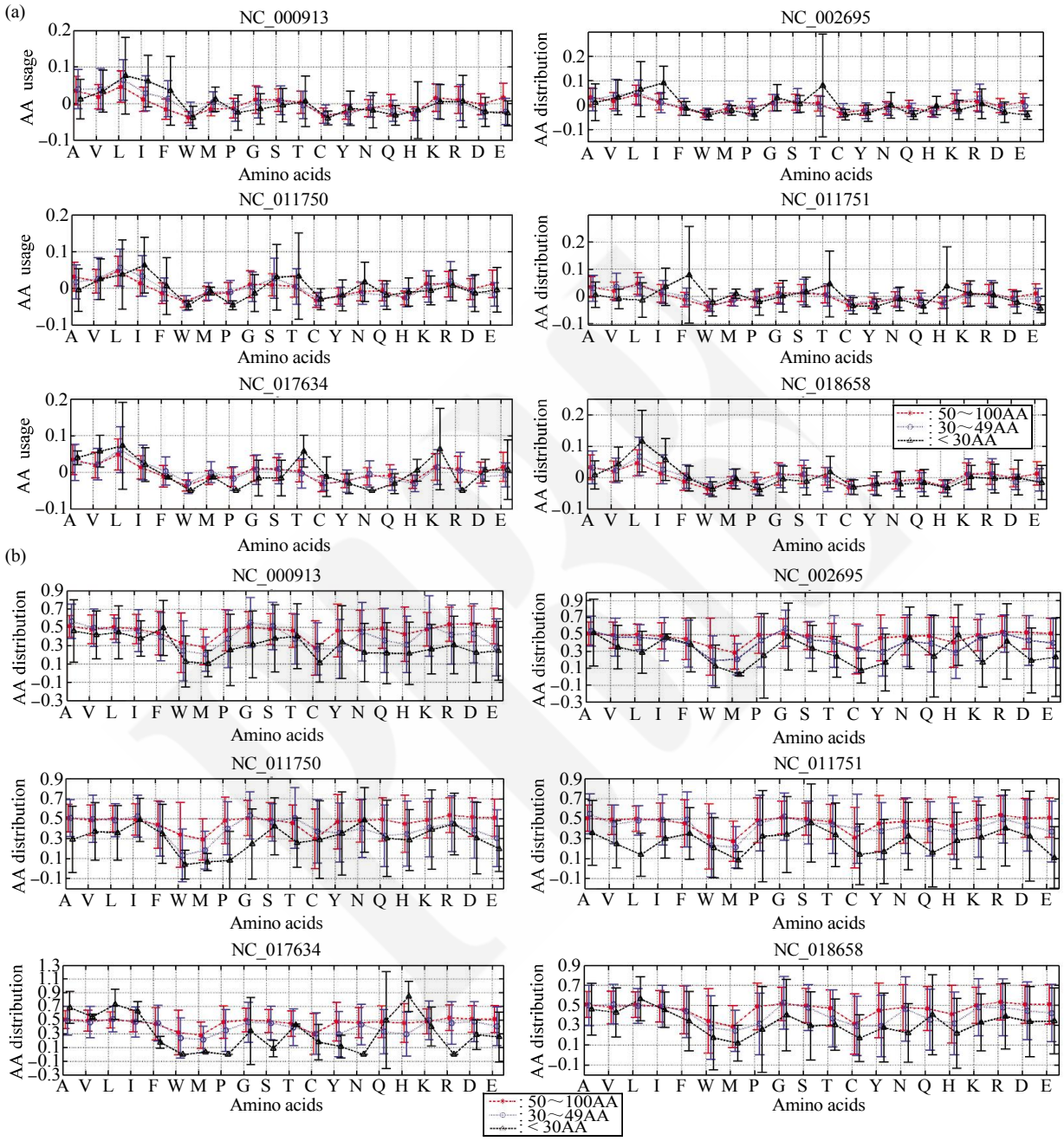

Fig. 2　Amino acids composition (2a) and distribution (2b) of peptide coding sORFs with different sequence lengths

2.2　肽编码 sORFs 功能分析

通常情况下，序列长度的降低增加了分子生物学实验中 PCR 引物设计的难度，实验研究 sORFs 难度也相应增加，因此对其功能的研究是目前生命科学的一项重要课题．在附件表 S2 中，我们整理了不同长度分布肽编码 sORFs 的具体功能信息，从目前注释结果来看，3 种长度区间 sORFs 分别有 35%(6337/18118)、25%(380/1498)和 53%(58/110)具有明确功能注释．从分布上来看，长度越大，涉及到的功能类型也越多．这里，我们依然以大肠杆菌中长度在 29 个氨基酸以下的 sORFs 为例，来进一步研究其功能特征．表 3 中给出了来自表 1 中所列 6 种大肠杆菌基因组的 79 个肽编码 sORFs(< 30 氨基酸)功能分布信息，除掉没有功能注释的 30 个 sORFs，其余 49 个 sORFs 的生物功能可以大致分为 6 大类，其中近 50%为前导肽(leader peptide)．

Table 3 Function analysis of the peptide coding sORFs less than 30 amino acids

sORFs number		Function
25	Leader peptide	his operon leader peptide (2)
		leu operon leader peptide (5)
		pheA gene leader peptide (2)
		phenylalanyl-tRNAsynthetase operon leader peptide (2)
		regulatory leader peptide for mgtA (1)
		ryhB-regulated fur leader peptide (2)
		thr operon leader peptide (4)
		trp operon leader peptide (2)
		tryptophanase leader peptide (5)
4	Membrane-associated protein	acid-inducible small membrane-associated protein(1)
		inner membrane-associated protein (2)
		putative membrane-bound BasSregulator (1)
3	Potassium-transporting	potassium ion accessory transporter subunit (2)
		potassium-transporting ATPase subunit F (1)
2	Stress response	stress response membrane (1)
		stress-induced small enterobacterial protein (1)
11	Toxic peptide	toxic peptide TisB (2)
		toxic membrane protein (7)
		toxic membrane persister formation peptide, LexA-regulated (1)
		UV-inducible membrane toxin, DinQ-AgrB type I toxin-antitoxin system (1)
4	Other	stationary phase-induced protein (1)
		protamine-like protein (1)
		3-hydroxypropionic acid resistance peptide (1)
		acetyltransferase (1)
30	Function unknown	Hypothetical protein

表 3 中的 sORFs 呈现出相对集中的生物学功能特征，在对真核生物学研究中发现 sORFs 编码肽具有一定程度的序列保守性[9]，接下来我们分别以具有 leu operon leader peptide、thr operon leader peptide、tryptophanase leader peptide 及 toxic membrane protein 功能的 4 组 sORFs 编码肽为例进一步分析其序列保守特征．借助多序列比对程序 Clustal，可以得到图 3 中的序列分析结果．可以看到，尽管具有前导肽功能的序列来自不同大肠杆菌菌株，但展现出高度的序列保守性，其中 leu operon leader peptide 和 tryptophanase leader peptide 对应的 sORFs 在不同菌株中完全一致．同样，在具有 toxic membrane protein 功能的 sORFs 序列中也展现出一定程度的序列保守特征．在图 3e 中，我们对上述 4 种不同功能的 sORFs 编码肽进行了序列比对，其中每条序列前标记的数字 1～4 分别表示图 3a～d 的功能类别，可以看出在这些序列中缺少保守序列．因此，图 3 的结果表明尽管 sORFs 序列较短，在依然具备了同蛋白质编码基因类似的保守性序列特征．

2.3 讨论

截止到 2017 年 2 月，GenBank 收录已完成基因组计划 70 036 个[22]，其中超过 90%为细菌(62 720 个)和古细菌(713 个)．作为与日常生活及工农业生产最为密切的原核生物，一直都是科学研究及生物技术应用的重点领域．本文从全基因组角度证明了原核生物基因组同样普遍存在具有重要功能的肽编码 sORFs，而且数量分布也非常广泛，这可为今后基于多肽的抑制剂、药物设计等提供丰富的资源．同时，肽编码 sORFs 的普遍存在也进一步说明之前基因预测研究中将长度小于 100 个氨基酸的 sORFs 排除掉缺少理论依据，反而加剧了原核生物

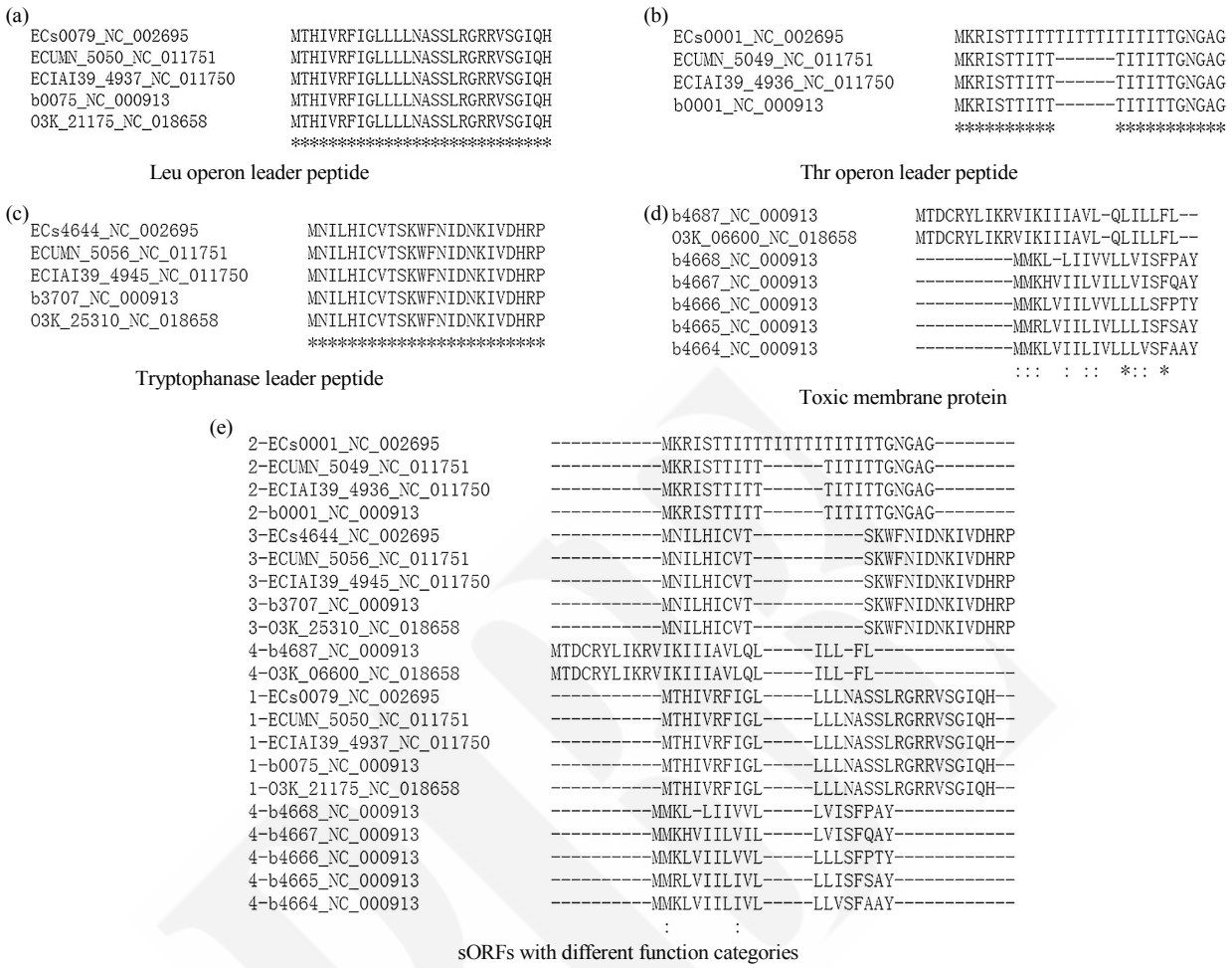

Fig. 3　Sequence conservation analysis of the peptide coding sORFs with different types of functions

蛋白质编码基因错误注释的进一步积累．因此，原核生物肽编码 sORFs 还有待于系统研究、开发，这也是当前测序技术快速发展背景下提升基因组注释质量急需解决的重要课题．

3　总结与展望

开放阅读框本是分子生物学中一个非常基础概念，其定义是从起始密码子到终止密码子结束的一段碱基序列．虽然定义很简单，但从已有参考书中却很难找到对 ORF 的深入解释，在一些专业论坛及学术资料中，许多研究人员甚至认为 ORF就是 CDS 序列．而实际上，只有很少数 ORF 才具备蛋白质编码能力，这也是利用 ORFfinder 等预测程序得到的 ORF 数量要远多于基因预测程序得到的蛋白质编码基因数量的主要原因．因此，在非编码 RNA 中存在 ORF 序列也已是人们习以为常的现象，但很少有人想到过这些 ORF 能够编码蛋白质，因而非编码序列中肽编码 sORFs 的发现引起人们很大兴趣[2, 23]．在首个公开发表的 sORFs 数据库 sORF.org 中[12]，收集了人、鼠和果蝇等几种真核生物中共计 348 844 条由核糖体谱识别得到的肽编码 sORFs．其中，仅人类基因组中收集的新 sORFs 就达到了 190 195 个，而一直以来，人们对人类基因组蛋白质编码基因数目估计仅为 20 000～25 000 个．尽管原核基因组基因预测工作已持续了 30 余年，近年来大量工作表明，目前原核生物基因组蛋白质编码基因普遍存在错误注释[24]，已有的基因预测程序还有很大发展空间，而对具有肽编码能力的 sORFs 的有效预测算法还有很多难点．因此，刚刚处于起步阶段的肽编码 sORFs 研究带来的挑战及面临的系列问题还需要进一步系统、深入研究．

3.1　肽编码 sORFs 对各种组学测序带来严峻挑战

继基因组测序之后，转录组、蛋白质组(质谱分析)等测序手段已成为当前生物医学领域研究的

标配技术[25-26]. 基因组测序为人们打开了基因组这本天书，而转录组、蛋白质组等测序技术为解读这本天书提供了有力工具. 长期以来，转录组与蛋白质组在 mRNA 与非编码 RNA 研究中发挥了极为重要作用[16]，因而人们对借助这些组学测序技术来研究 sORFs 一度寄予厚望[26]. 然而，最终测序结果表明距离预期存在很大差距[14]. 这其中的主要原因就在于 sORFs 整体表达水平低、丰度低，有些仅在某种条件下选择性表达[27]. 借助转录组测序，人们无法进一步判断转录产物是停留在 RNA 水平还是继续被翻译，而以质谱分析为代表的蛋白质组测序虽然能够检测到 mRNA 翻译产物，但对于低丰度、低表达、不表达的 sORFs 没有效果，因而实际应用有限[16]. 近几年发展的核糖体谱(ribosomeprofiling)技术被认为是继质谱分析之后检测 RNA 翻译与否的更有效测序技术[28-29]，并被广泛应用于肽编码 sORFs 识别[12, 30-31]，其原理是通过分析结合在核糖体的分子来判断 mRNA 的翻译情况，但近期发现许多 lncRNA 等非编码分子也会与核糖体结合[4, 27]，因而依然存在许多问题. 因此，肽编码 sORFs 的发现为当前各种组学测序技术带来了更大的挑战，将测序技术与生物信息技术有效结合成为今后 sORFs 研究的必然趋势[2, 10, 23, 32].

3.2 肽编码 sORFs 研究还缺少有效计算和实验方法

缺少有效的高通量研究方法是目前肽编码 sORFs 遇到的首要问题[33]. 通过转录组测序可以获得大量的 RNA 转录产物，但无法判断这些转录产物是否能够翻译. 尽管某些分子生物学和生物化学方法(如实时荧光定量 PCR 等)也可以不同程度用于 sORFs 研究，但也仅能停留在转录层次[27]，也无法实现高通量实验研究. 因此，计算方法研究 sORFs 同样成为当前的主要研究手段[26-27]. 目前对 sORFs 计算研究包括两个层次，一个是通过分析 sORFs 序列组成[23, 34]，包括碱基组成、密码子和二联密码子(六聚体碱基)[35]等；另一种是借助 BLAST 等相似性分析程序来获取 sORFs 特征[11, 14, 36]. 然而，由于 sORFs 序列相对较短，有些传统的序列分析方法可能无法直接应用，但一些应用于其他类型短肽或 RNA 的计算方法可以为 sORFs 研究提供有效参考[37]. 另一方面，通过进一步发展序列结构分析方法来深入揭示肽编码 sORFs 的固有特征[38]，与实验研究 sORFs 有效互补[10]，有望为肽编码 sORFs 研究提供新思路.

总之，作为一种长期被忽略的基因组“新元件”，从其在基因组中的识别定位到功能研究，肽编码 sORFs 承载了极为重要的学术意义，也为多肽应用提供了丰富的生物资源[39]. 面对肽编码 sORFs 带来的机遇与挑战，充分融合生物信息技术、各种组学测序技术及生物实验技术将是今后切实可行的研究趋势. 因此，希望本文能够为今后原核生物基因组生物活性多肽资源挖掘及应用提供可靠的理论支持.

附件 表 S1 和表 S2 见本文网络版附录(http://www.pibb.ac.cn).

参 考 文 献

[1] Nelson B R, Makarewich C A, Anderson D M, *et al*. A peptide encoded by a transcript annotated as long noncoding RNA enhances SERCA activity in muscle. Science, 2016, **351**(6270): 271-275

[2] Oehler D, Haas J. Hide and Seek: protein-coding sequences inside “non-coding” RNAs. Genomics Proteomics Bioinformatics, 2016, **14**(4): 179-180

[3] Beermann J, Piccoli M, Viereck J. Non-coding RNAs in development and disease: background, mechanisms, and therapeutic approaches. Physiological Reviews, 2016, **96**(4): 1297-1325

[4] Guttman M, Russell P, Ingolia N T, *et al*. Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. Cell, 2013, **154**(1): 240-251

[5] Schmitz J F, Bornberg-Bauer E. Fact or fiction: updates on how protein-coding genes might emerge *de novo* from previously non-coding DNA. F1000 Research, 2017, **6**: 57

[6] CrappéJ, CriekingeW V, Menschaert G. Little things make big things happen A summary of micropeptide encoding genes. EuPA Open Proteomics, 2014, **3**: 128-137

[7] Lawrence J. When ELFs are ORFs, but don’t act like them. Trends in genetics, 2003, **19**(3): 131-132

[8] Cabrera-Quio L E, Herberg S, Pauli A. Decoding sORF translation - from small proteins to gene regulation. RNA Biology, 2016, **13**(11): 1051-1059

[9] 田 原, 杨金娥. 短肽编码基因的研究进展. 世界华人消化杂志, 2015, **23**(31): 4954-4960
Tian Y, Yang J E. World Chinese Journl of Digestology, 2015, **23**(31): 4954-4960

[10] Crappé J, Criekinge W V, Trooskens G, *et al*. Combining in silico prediction and ribosome profiling in a genome-wide search for novel putatively coding sORFs. BMC Genomics, 2013, **14**: 648

[11] Hellens R P, Brown C M, Chisnall M A W, *et al*. The emerging world of small ORFs. Trends in Plant Science, 2016, **21** (4): 317-328

[12] Olexiouk V, Crappe J, Verbruggen S, *et al*. sORFs.org: a repository of small ORFs identified by ribosome profiling. Nucleic Acids

Research, 2016, **44**(D1): D324-D329

[13] Hazarika R R, Coninck B D, Yamamoto L R, *et al*. ARA-PEPs: a repository of putative sORF-encoded peptides inArabidopsis thaliana. BMC Bioinformatics, 2017, **18**(1): 37

[14] Couso J P. Finding smORFs: getting closer. Genome Biology, 2015, **16**: 189

[15] Zanet J, Chanut-Delalande H, Plaza S, *et al*. Small peptides as new comer in the control of *Drosophila* development. Current Topics in Developmental Biology, 2016, **117**: 199-219

[16] Neuhaus K, Landstorfer R, Fellner L, *et al*. Translatomics combined with transcriptomics and proteomics reveals novel functional, recently evolved orphan genes in *Escherichia coli* O157:H7 (EHEC). BMC Genomics, 2016, **17**:133

[17] Wadler C S, Vanderpool C K. A dual function for a bacterial small RNA: SgrS performs base pairing-dependent regulation and encodes a functional polypeptide. Proc Natl Acad Sci USA, 2007, **104**(51): 20454-20459

[18] Hao Y, Zhang L, Niu Y, *et al*. SmProt: a database of small proteins encoded by annotated coding and non-coding RNA loci. Briefings in Bioinformatics, 2017, doi: 10.1093/bib/bbx005

[19] O'Leary N A, Wright M W, Brister J R, *et al*. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Research, 2016, **44**(D1): D733-745

[20] Wootton J C. Statistic of local complexity in amino acid sequences and sequence databases. Computers & Chemistry, 1993, **17** (2): 149-163

[21] Yu J F, Dou X H, Wang H B, *et al*. A novel cylindrical representation for characterizing intrinsic properties of protein sequences. Journal of Chemical Information and Modeling, 2015, **55**(6): 1261-1270

[22] Mukherjee S, Stamatis D, Bertsch J, *et al*. Genomes OnLine Database (GOLD) v.6: data updates and feature enhancements. Nucleic Acids Research, 2017, **45**(D1): D466-D456

[23] Pueyo J I, Magny E G, Couso J P. New peptides under the s(ORF) ace of the genome. Trends in Biochemical Sciences, 2016, **41**(8): 665-678

[24] Yu J, Guo Z, Sun X, *et al*. A review of the computational methods for identifying the over-annotated genes and missing genes in microbial genomes. Current Bioinformatics, 2014, **9**(2): 147-154

[25] Housman G, Ulitsky I. Methods for distinguishing between protein-coding and long noncoding RNAs and the elusive biological purpose of translation of long noncoding RNAs. Biochimicaet Biophysica Acta, 2016, **1859**(1): 31-40

[26] Yang X, Tschaplinski T J, Hurst G B, *et al*. Discovery and annotation of small proteins using genomics, proteomics, and computational approaches. Genome Research, 2011, **21** (4): 634-641

[27] Andrews S J, Rothnagel J A. Emerging evidence for functional peptides encoded by short open reading frames. Nature Reviews Genetics, 2014, **15**(3): 193-204

[28] Baek J, Lee J, Yoon K, *et al*. Identification of unannotated small genes in salmonella. G3: Genes Genomes Genetics, 2017, **7** (3): 983-989

[29] 赵 晶, 张 弓. 翻译组学: 方法及应用. 生命的化学, 2017, **37**(1): 70-79
Zhao J, Zhang G. Chemistry of Life, 2017, **37**(1): 70-79

[30] Ingolia N T, Brar G A, Stern-Ginossar N, *et al*. Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. Cell Reports, 2014, **8**(5): 1365-1379

[31] Bazzini A A, Johnstone T G, Christiano R, *et al*. Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. EMBO Journal, 2014, **33**(9): 981-993

[32] Kageyama Y, Kondo T, Hashimoto Y. Coding *vs* non-coding: Translatability of short ORFs found in putative non-coding transcripts. Biochimie, 2011, **93**(11):1981-1986

[33] Vivek S, Mayank K, Santosh N, *et al*. ORFpred: a machine learning program to identify translatable small open reading frames in intergenic regions of the plasmodium falciparum genome. Current Bioinformatics, 2016, **11**(2): 259-268

[34] Ruiz-Orera J, Verdaguer-Grau P, Villanueva-Cañas J L, *et al*. Functional and non-functional classes of peptides produced by long non-coding RNAs. BioRxiv, 2016, 064951

[35] Hanada K, Akiyama K, Sakurai T, *et al*. sORF finder: a program package to identify small open reading frames with high coding potential. Bioinformatics, 2010, **26**(3): 399-400

[36] Lin M F, Jungreis I, Kellis M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. Bioinformatics, 2011, **27**(13): i275-i282

[37] Lin H, Deng E Z, Ding H, *et al*. iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition. Nucleic Acids Research, 2014, **42**(21): 12961-12972

[38] Cheng H, Chan WS, Li Z, *et al*. Small open reading frames: current prediction techniques and future prospect. Current Protein & Peptide Science, 2011, **12**(6): 503-507

[39] Saghatelian A, Couso J P. Discovery and characterization of smORF-encoded bioactive polypeptides. Nature Chemical Biology, 2015, **11**(12): 909-916

Sequence and Function Analysis of Peptide Coding Small Open Reading Frames in Prokaryotic Genomes*

CHEN Yi-Ting[1,2)**], ZHANG Feng[1,2)**], ZHAO Jia[1,3)**], YU Jia-Feng[1,2)***], SHA Yu-Jie[1)], WANG Ji-Hua[1,2)]

([1)] *Shandong Provincial Key Laboratory of Biophysics, Institute of Biophysics, Dezhou University, Dezhou* 253023, China;

[2)] *College of Physics and Electronic Information, Dezhou University, Dezhou* 253023, China;

[3)] *College of Life Science, Shandong Normal University, Jinan* 250000, China)

Abstract Whether encoding protein is the golden standard for distinguishing protein coding genes and non-coding RNA (ncRNA), while recent detected peptide coding small open reading frames (sORFs) from lncRNA challenged this standard. Now, more and more studies have shown that peptide coding sORFs exist in different regions of eukaryotic genomes universally, which play important roles in biological activities. Because of the low expression level as well as low abundance and the short sequence length, there are few computational and experimental methods or data resources exploited for peptide coding sORFs, then study of peptide coding sORFs is in its early phase. At present, most studies of peptide coding sORFs are concentrated on several model eukaryotes，people know little about its intrinsic features, therefore the peptide coding sORFs bring more challenges for genome annotation under the precision medicine era. In this work, comprehensive sequence and function analysis of the peptide coding sORFs were firstly performed based on more than 80 prokaryotic genomes. The results show that peptide coding sORFs also exist in prokaryotic genomes universally and many peptide coding sORFs sequences are conserved among different genomes. Further analysis indicates that the sequence complexity decreases and their functions are relatively centered with the decrease of sequence length of peptide coding sORFs. Finally, we summarized the problems and challenges psoposed by peptide coding sORFs, which will provide solid theoretical basis for future sORFs related studies.

Key words small open reading frame, prokaryotic genome annotation, protein-coding RNA

DOI: 10.16476/j.pibb.2017.0109

* This work was supported by grants from The National Natural Science Foundation of China (61771093, 61671107) and Shandong Natural Science Foundation (ZR2016JL027).

**These authors contributed equally to this work.

***Corresponding author.

Tel: 86-534-8982557, E-mail: jfyu1979@126.com

Received: October 13, 2017 Accepted: November 10, 2017