·热点聚焦·

PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2007, 34(7): 669~672
www.pibb.ac.cn

1%人类基因组 DNA 元件解译 基因组结构的传统认识受到挑战

变人类基因组“天书”为“百科全书”的重要一步

吴 涛 石宝晨 陈润生*
(中国科学院生物物理研究所，北京 100101)

继“人类基因组计划”后生命科学领域最大的国际合作计划之一——“DNA 元件百科全书”计划（Encyclopedia of DNA Elements，ENCODE）日前发表了一系列重要文章报告了其示范期完成的 1%人类基因组的解码情况. 这些成果由来自 11 个国家的 35 个小组合作完成，挑战了人们对于人类基因组的传统认识，即人类基因组不是由孤立的基因和大量“无用 DNA 片段”组成的，基因组本身就是一个复杂的网络系统，有着多层次的精细调控规则. ENCODE 计划示范期（1%人类基因组解码）的成果为进一步解译人类基因组这部“天书”开辟了道路，刷新了我们关于基因的概念，成为人类生物学研究史中又一座里程碑. 同时，大量的“非编码 RNA 基因”仍旧是生命科学天空中的“一朵乌云”，其未知的功能尚待探索.

ENCODE计划的最新进展——1%人类基因组解译的重大意义

继“人类基因组计划”之后，生命科学最大的国际合作计划——“DNA 元件百科全书”计划(Encyclopedia of DNA Elements，ENCODE) 日前发表了一系列重要研究成果，挑战了关于人类基因组的传统理论，即我们的基因组不是由孤立的基因和大量“无用 DNA 片段”组成的，而是一个复杂的网络系统. 单个基因、各种调控元件以及非编码的其他类型的功能 DNA 序列之间有着复杂的相互作用，共同控制着人类的生理活动. ENCODE 计划促使我们重新考虑长期以来关于“基因”的概念和对于基因组功能元件及组织机制的认识，这将对与人类疾病相关研究产生革命性的影响，为进一步认识整个人类基因组的功能蓝图开辟道路.

ENCODE 团队是由美国国家人类基因组研究所 (National Human Genome Research Institute，NHGRI) 发起组织成立的，包括全世界 11 个国家 80 家科研机构 35 个小组的研究人员. 该团队在 2007 年 6 月 14 日出版的《自然》杂志(Nature)上发表了一篇重要的论文，并在 2007 年 6 月出版的 Genome Research 上发表了由 28 篇相关论文组成的专集，报道了他们 4 年来的研究成果，即通过建立一个“功能元件目录”(ENCyclopedia of DNA Elements)，在人类基因组的 1%区域内详尽地描述了实现全部生理功能的分子基础. NHGRI 主任 Francis S. Collins (原人类基因组计划负责人) 表示，“这是人类生物学史上的一个里程碑.”不过，这部分工作仅仅是整个 ENCODE 工程的第一阶段的试验项目，目的是考察建立整个人类基因组生物功能详细“目录”的可行性.

2003 年人类基因组计划的完成仅仅是人类向着利用基因信息进行疾病诊断、治疗和预防的目标迈出的第一步，尽管这是非常重要的一步. 当时，很多人过低地估计了基因组结构、组织的复杂性. 近年来虽对基因组的研究已经取得了巨大进展，但截至目前，研究主要还集中在编码蛋白的基因上，而它们只占整个人类基因组的 2%. 而 ENCODE 计划首次系统地研究了人类基因组上面所有类型的功能元件的位点和组织方式. ENCODE 计划的研究对

* 通讯联系人. Tel: 010-64888549, E-mail: crs@ibp.ac.cn
收稿日期：2007-06-30，接受日期：2007-07-01

象包括：蛋白编码基因、非蛋白编码基因、调控区域、染色体结构维持区域和调节染色体复制动力的DNA元件. 到目前为止，ENCODE 主要集中研究了44 个靶标区域，共约 3 000 万个 DNA 碱基对. 负责该计划数据整合和分析工作的欧洲分子生物学实验室 (European Molecular Biology Laboratory, EMBL)主任 Ewan Birney 说，“我们的结论揭示了有关 DNA 功能元件组成的重要原理，为从 DNA 转录到哺乳动物进化的一切生物过程提供了新的认识和线索.”

ENCODE 最新一批研究成果是得到约 200 个实验数据集，主要着眼于基因组注释，RNA 表达分析和比较基因组学. 结果表明，人类基因组中有93%的 DNA 都会转录成 RNA，众多转录本为非编码 RNA, 有一些看来很像是**“融合转录本”**，这些转录本会发生相互作用. 而且还发现了许多新的转录起始位点，其中很多位点带有组蛋白修饰. 数据还显示，远端调控元件所带有的修饰与近端启动子是不同的. 总之，这一批新成果表明，组蛋白修饰、DNase Ⅰ超敏感度与转录及复制有着广泛的联系，这些联系强有力地支持着一种猜想，就是基因组有着更高层次的功能组织域. 因此，人类基因组本身就是一个极其复杂的网络，所谓的垃圾基因 (junk DNA) 实际上非常少. 蛋白编码基因只不过是众多具有特定功能的 DNA 序列元件的一种.

ENCODE 计划的示范期，通过完成对 1%人类基因组的解码，获得了很多具有突出生物学意义的发现[1]:

1. 人类基因组绝大部分区域均发生转录，许多区间至少对应于一种初级转录本，并且许多距离较远的转录本之间相互联系而组合成为蛋白编码序列.

2. 发现了众多原来未知的非蛋白编码转录本，有一些位于蛋白编码区，而许多位于原先人们认为是转录沉默的位点.

3. 探测到了许多原先没有被发现的转录起始位点，其染色质结构性质以及蛋白绑定特性与已知的启动子(promoter)很相似.

4. 靠近编码区的调控位点相对于转录起始位点是对称分布的，而并不是原先人们所认为的它们有向起始位点上游方向偏移的倾向性.

5. 染色质可接近性和组蛋白修饰方式与转录起始位点的位置及活性状态极为相关，通过前面两种特征可以很好地预测出转录起始位点.

6. 相对于编码区很远的 DNA 酶Ⅰ超敏感位点的组蛋白修饰方式与近端启动子有着很不同的特征. 一些远端位点表现出了间隔子 (insulator) 的特征.

7. DNA 复制时间段与染色质结构有关.

8. 在哺乳动物基因组当中约有 5%的碱基处于“进化限制(evolutionary constraint)”当中，到目前为止，在这些碱基中约 60%有着明确的被实验所证实了的生物学功能.

9. 虽然，基因组当中有功能的区域，有一部分处于进化限制当中，但也发现有一部分例外，这部分功能区并没有受到所谓的进化限制.

10. 不同的功能元件，在不同的群体当中变化很大，而且一部分元件落在了基因组结构可变区.

11. 更为不寻常的是，在哺乳动物进化过程当中，很多功能元件并不保守. 这一点提示我们，似乎存在着一个巨大的进化中性的功能元件库，它们有着某些生化活性，但对生物体却并没有表现出特别的进化选择上的特殊偏好性(benefit).

特别值得提出的是，几年前在人类基因组计划完成的时候，人们关注的是基因组中编码蛋白质的区域(被称为基因)，因为它们有明确的生物学功能. 随着 ENCODE 计划的逐步展开，基因组复杂的散在调控序列，大量的非蛋白编码 RNA 基因以及非蛋白编码区域的保守性元件渐渐出现，对于传统的基因观念产生了重大的挑战. 最近，耶鲁大学 Gerstein 等对基因提出了一个新的概念，他们认为“基因”就是一个基因组上面编码潜在相关联的一系列功能元件的基因组序列的“联合体”(A gene is a union of genomic sequences encoding a coherent set of potentially overlapping functional products.)[2]. 尽管，从孟德尔和摩尔根开始到今天，基因的概念一直在变化[2]，但我们相信随着 ENCODE 计划的深入，传统“基因”的概念，甚至于分子生物学的“中心法则”都将会再次被改变.

ENCODE计划的历史，任务和特点

ENCODE计划是 2003 年 9 月由美国国家人类基因组研究所 (NHGRI) 提出的. 其目的是希望找出人类基因组序列中所有的结构和功能元件，形成一个完整的人类基因组的“元件目录”，包括：蛋白编码基因；非蛋白编码基因；转录调控元件；其他调节染色体结构和动态活动的功能序列，如 DNA 复制起始序列. 这一计划分为 3 个部分，即示范期

(pilot phase)、技术发展期(technology development phase)和产出期(production phase)[3]. 与人类基因组计划相比，“ENCODE 计划”有三个明显的特点：一是采用综合性研究策略，二是重视新技术的研发，三是将计划向学术界和公司开放.

1. 综合性研究策略

人类基因组有不同种类的结构和功能元件，这些元件涉及到 DNA 代谢和染色体构成等生命活动的各个方面，如 DNA 的复制起始与终止位点、基因、启动子、RNA 剪切位点、DNA 甲基化位点和 DNA 酶Ⅰ超敏感位点等. 显然，要想实现“ENCODE 计划”拟定的科学目标，不可能像人类基因组计划那样只依赖于 DNA 测序仪一种手段，而要尽可能地采用和整合现有的各种实验以及理论手段.

过去几十年中，研究者一直通过经典的实验生物学方法对基因组内的各种结构和调控元件进行单独的研究. ENCODE 提出的每一类元件都曾经被发现过，所不同的是现在要在全基因组的范围内进行系统地整合研究. 因此，目前在实验生物学中常用的研究基因组的结构和调控元件的方法，如染色质免疫共沉淀(chromatin immunoprecipitation，ChIP)，基因组 DNA 对 DNase Ⅰ酶的超敏感位点探测(DNase Ⅰ hypersensitive sites detection)等，也成为了该工程的重要研究手段. 然而，为了满足同时进行基因组内成千上万个元件的大规模、高通量的分析需求，这些用于研究基因组内一两个元件研究的实验生物学技术，必须与高通量研究方法进行整合. 例如，在该计划中染色质免疫共沉淀技术采用的是与芯片相结合的研究策略，称为“ChIP-chip”技术，DNase Ⅰ超敏感位点探测可以与高通量测序技术“MPSS”或“454 GS_20”相结合. 可以说，传统经典实验手段与现代高通量研究方法的结合，是该工程的重要特征. 大规模、高通量的分析必然会产生海量的数据. 但值得注意的是，与人类基因组计划产出单一的测序数据相比，在 ENCODE 计划实施过程中获得的数据不仅量大，而且种类非常繁杂. 如何将这些数据进行分类、整合和展现，是该工程面临的巨大挑战. 为此，示范期大部分研究项目中都有计算生物学的参与. 而且，该工程将比较基因组研究列为一个主要的内容，“要发展更强大的计算工具用来进行序列比较，从而推导出生物学功能”[3].

2. 发展新方法和新技术

“工欲善其事，必先利其器”，过去在基因组领域的实验工作是典型的低通量小规模研究，针对的是个别的结构或功能元件，在这些工作中所使用的研究手段显然难以满足 ENCODE 计划的需求. 因此，ENCODE 计划的制定者专门提出了一个与示范期平行的技术发展期，用来发展能够提供给未来产出期进行工作的新式仪器与设备，或是将多种技术结合在一起. 例如发展一种结合了 ChIP、SAGE 和 FAIRE 三种技术的综合性新技术——基因组富集的序列标签分析(sequence tag analysis of genomic enrichment，STAGE)，用来确定在染色质上的转录调控元件[3].

在新方法和新技术的开发中，发展计算生物学方面的新手段也同样受到关注. 高等真核生物基因组的大部分基因都是不连续的，并且常常有不同的剪切方式. 有研究发现，有些具有不同剪切方式的基因拥有不同的启动子，称为备选启动子(alternative promoter). 另外，将支持向量机(SVM)，隐马氏模型(HMM)，以及小波分析等成熟的数理方法适当地应用到生物学问题当中也显得极为重要.

3. 向社会开放

ENCODE 计划还具有一个与以往科学计划不同的特点，即向社会开放. 最早的 ENCODE 联合体(ENCODE Consortium)，是由获得美国国家人类基因组研究所 ENCODE 计划资助的科学家所组成. 但是，ENCODE 联合体随后宣布，“ENCODE 联合体向所有对该项目有兴趣的学术机构、政府部门和私人公司的研究者开放”[3]. 也就是说，任何研究者都可以申请成为 ENCODE 联合体的成员. 当然，要想成为其成员必须同意遵守该联合体的有关规定.

对于科学研究计划来说，在研究过程中获得的数据是否被及时公布并被无偿使用，一直是科学界最为关心的问题. ENCODE 计划被美国国立人类基因组研究所界定为“公共资源项目”(community resource project)，所有的研究数据一经核实，必须放入公共数据库并供所有研究者无条件使用[3]. 因此，凡是 ENCODE 联合体的成员都要遵循这一原则，并且按照制定的数据管理规则操作.

(注：ENCODE 详细情况，请登录 http://www.genome.gov/ENCODE)

ENCODE 计划的展望

ENCODE计划不单单只是另一个大规模的生命科学数据产出，收集与分析计划，而是对传统经典的生物学研究方式方法和一些经典的分子生物学及进化生物学核心观念的革命性挑战. “中心法则”以及“有功能即保守”等等一批人们形成的观念，在基因组大批非蛋白编码转录本以及间区调控元件大量涌现的非传统现象面前，显得过于简单了. 该计划示范期关于1%人类基因组解码任务的完成证实了该计划的可行性和重大的生物学意义. 同时，示范期1%工作的结束也就意味着ENCODE计划的全面开始，预计在不久之后便会有针对人类全基因组的新数据和新结论出现. 随着技术手段的飞速更新和ENCODE计划的不断深入，我们相信会有越来越多的非传统实验现象和数据出现，而这些线索也正是我们打开生命核心规律之门的新钥匙. ENCODE会为我们提供越来越多的新线索.

参 考 文 献

1 The ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature, 2007, **447** (7146): 799～816

2 Gerstein M B, Bruce C, Rozowsky J S, *et al*. What is a gene, post-ENCODE? History and updated definition. Genome Research, 2007, **17** (6): 669～681

3 The ENCODE Project Consortium. The ENCODE (ENCyclopedia of DNA Elements) Project. Science, 2004, **306**: 636～640