[5] Glass L. Physics Today, 1996(8): 40
[6] Comacho P, Lechleiter J D. Science, 1993, 260: 226
[7] Belmonte A, Ouyang Q, Flesselles J-M. J. Phys. II France, 1997, 7: 1425
[8] Murray J D. Mathematical Biology. Berlin: Spring-Verlag. 1989
[9] Ammelt E, Astrov Y A, Purwins H-G. Phys. Rev. E, 1997, 55: 6731
[10] Klausmeier C A. Science, 1999, 284: 1826
[11] Mendez V. Phys. Rev. E, 1998, 57: 3622
[12] Mandez V, Llebot J E. Phys. Rev. E., 1997, 56: 6557
[13] Fort J, Mendez V. Phys. Rev. E, 1999, 60: 5894
[14] Skinner G S, Swinney H L. Physica D, 1991, 48: 1
[15] Ouyang Q, Fellesselles J-M. Nature, 1996, 379: 143
[16] Li G, Ouyang Q, Petrov V *et al*. Phys. Rev. Lett., 1996, 77: 2105
[17] Winfree A T. Science, 1973, 181: 937
[18] Barkley D. Phys. Rev. Lett., 1992, 68: 2090

解读生命密码*

——人类基因组计划

戴 闻　　曾宗浩

(中国科学院理化技术研究所　北京　100080)　　(中国科学院生物物理研究所　北京　100101)

摘　要　DNA测序技术的自动化使得人类基因组测序工作在启动10年后就已接近完成.从物理学的角度来看,生物体是工作在单分子水平上的多层次综合的信息、能量和物质加工转换系统.关于生物生长发育和遗传的信息记录在线型分子——核酸的碱基序列中.每个基因是编码一个蛋白质的核酸片段.这些蛋白质分子是分子机器的主要零部件.首次人类基因组测序的完成,只是生命密码破译的开始,而不是结束.年轻的物理学家应积极地参与揭示生命本质的活动.

关键词　遗传,基因组,核酸,蛋白质,分子机器

TOWARDS UNDERSTANDING LIFE'S SECRET——THE HUMAN GENOME PROJECT

DAI Wen

(*Institute of Physical and Chemical Technology, Chinese Academy of Sciences, Beijing* 100080, *China*)

ZENG Zong-Hao

(*Institute of Biophysics, Chinese Academy of Sciences, Beijing* 100101, *China*)

Abstract　Since the automation of DNA sequencing technology, after about a decade the Human Genome Project is going to be fulfilled within the next year. From a physical point of view, an organism is an assembly of multiple levels of information, energy and material transformation systems working on the single molecule level. Information about the organism's development and heredity is encoded in the base sequence of linear molecules: nucleic acids. Each gene is a piece of nucleic acids encoding a protein, which is the main type of component of molecular machines. The complete sequencing of human genome signals the beginning, not the end of the road towards decoding life's secret. Yong physicists should actively participate in exploring the nature of life.

Key words　inheritance, genome, nucleic acid, protein, molecular machine

经过多国科学家的共同努力,启动于1990年的人类基因组计划(human genome project, HGP)取得了战略性进展——获得了人类基因组的"工作草图".在此之前,已有22和21号染色体的测序工作宣布完成.人类基因组的测序工作已进入尾声,预计2001年即可全部结束.

自古以来,生物就因其具有"活"的特征,而让我们的先辈们大为困惑,也使生物学成了现代科学的热点.所谓"活",就是能够生长、繁殖、进化.为什么生物的各种特征能够世代相传呢?通过对可观察的

* 2000-07-13收到初稿,2000-09-07修回

性状的遗传规律的研究,在近百年以前产生了“基因”的概念,专用来指负责生物特定性状遗传的物质.生物是否从其父母那里继承一种特定的性状,比如眼睛、皮肤、花朵的颜色,翅膀的形状,人的血型,甚至身体的某种缺陷等,都是由负责该种性状遗传的基因所决定的.大约 50 多年前,著名物理学家薛定谔就推测基因是一种可以发生突变的巨大分子.差不多同时,生物化学家们发现了染色体中的 DNA 是负责遗传的物质.

人体细胞多达数万亿,除血液里的红细胞外,每个细胞都有 23 对染色体,每对两条,分别来自父和母.这种染色体成对出现的状态称为双倍体.例如性染色体是第 23 对.男子为 XY,女子为 XX.生殖细胞经减数分裂成为单倍体的精子或卵子.精子里的性染色体可能是 X 也可能是 Y,卵子里的性染色体只可能是 X.带 Y 染色体的精子与卵子结合,就会发育为男孩,带 X 染色体的精子与卵子结合则发育为女孩.遗传物质 DNA 就存在于染色体中.DNA 与 RNA 都是由核糖核酸共价连成的线形分子.每种核糖核酸包含碱基、糖环和磷酸三部分.DNA 的糖环少一个氧原子,所以称为脱氧核糖核酸.DNA 可以携带四种不同的碱基,即 A,G,C,T;RNA 也可以携带四种不同的碱基,即 A,G,C,U.人体细胞单套染色体的 DNA 中约有 30 亿个碱基,总长约为 1 m.

1953 年,英国物理学家 Crick 和美国生化学家 Watson 发现 DNA 可形成双螺旋结构.碱基 A 与 T,G 与 C 可以分别靠氢键配对(称为碱基对,简称 bp),形成螺旋“楼梯”上的台阶.糖-磷酸骨架则形成“楼梯”上的“扶手”.双螺旋链绕在蛋白质颗粒上,再装配成染色体.只有这样,长度为米量级的 DNA 才有可能放入直径为微米量级的细胞核中.每条染色体包含一条双螺旋 DNA 链.构成双螺旋的两条 DNA 链是互补的:如果一条为 CGGAGACG,另一条一定为 GCCTCTGC.这种互补结构向人们暗示了遗传物质复制的秘密:由一条 DNA 链靠互补规律精确地复制出另一条链来.

生物体内的主要工作都是由另一种线形分子——蛋白质——组成的分子机器来完成的.遗传物质的主要任务就是编码各种蛋白质.蛋白质由 20 种不同的氨基酸通过肽键连接而成.各种氨基酸在线形链中的顺序,决定了这种分子的三维结构,也决定了它的功能.一种蛋白质是消化食物的淀粉酶,还是合成 DNA 的聚合酶,是完全由它的氨基酸序列决定的.现在的问题是,由 4 种碱基组成的 DNA 链如何去编码由 20 种氨基酸组成的肽链(蛋白质)?这也是一个曾经让人大为困惑的问题.天文学家 G.Gamow 在破解这一秘密的过程中做出了贡献.他于 1954 年首次提出了三联体密码的假说:三个相连的碱基编码一个氨基酸.由四种碱基组合成的三联体共有 $4^3 = 64$ 种,比氨基酸的种类多.所以,某些氨基酸可以对应几种不同的编码.这种三联体密码的假说和具体的编码方式,经过生物化学家们的努力,终于在 20 世纪 60 年代弄清楚了.他们发现,破解这种密码的器具是一种叫做 tRNA(翻译器 RNA,也译为转移 RNA)的分子.它的一端携带一种氨基酸,另一端则带有编码这种氨基酸的三联体密码.

在合成蛋白质之前,先要根据 DNA 上的碱基序列按互补规律合成另一种线形分子 mRNA(信使 RNA)(RNA 中有 U 无 T.碱基 U 与 T 相比只少一个甲基,并不影响它与 A 配对),这一过程叫做“转录”.然后,一种叫做核糖体的由蛋白质和 rRNA 组成的分子机器,以 mRNA 为模板合成蛋白质.携带不同氨基酸的 tRNA 不断地把各种氨基酸带向核糖体.如果某种 tRNA 上的三联体与 mRNA 上待读出的三联体可以互补,则这种 tRNA 就可以与核糖体结合,它所携带的氨基酸就被连接到正在合成的肽链上.如此,完成了一个密码子的“翻译”.例如,RNA 三联体 CAU、CAC 都被翻译为组氨酸,GUU、GUC、GUA、GUG 都被翻译为缬氨酸,等等.这样我们就知道了遗传信息的两种流动方向:由 DNA 到 DNA 的“复制”过程,使遗传信息可以由父代传给子代;由 DNA 到 RNA 再到蛋白质的“转录”和“翻译”过程使遗传信息转化为生物功能.

所有这些过程都是在分子水平上发生的.细胞内的主要过程都是在分子水平上发生的,这决定了现代生物学的研究方法:如果你发现了一种新的分子,那么就去了解它有什么作用和如何起作用;如果你关心某种现象,那么就去寻找与这种现象有关的分子,并研究它是如何导致这种现象的.现在我们对“基因”的认识也是在分子水平上的:基因是编码一种蛋白质(可能还有 RNA)的 DNA 片段.在这个片段中可能插入一些非编码区,也可能与另外的基因共享部分区段.所谓基因组就是单套染色体中的全部 DNA.在 1—22 号染色体中,每种基因都存在于成对染色体的每一条上,因而是双份的.第 23 号染色体有 X 和 Y 两种.因此,HGP 的工作实际上要测定 24 条染色体中的碱基对顺序.生物化学家们用酶把 DNA 分子打断成长短不一的片段,用克隆技术

复制,以使每种片段都有足够的量,然后用自动化的测序仪测定每种片段的碱基顺序,最后用超级计算机把所得到的片段序列照原样连接起来.目前已经得到"工作草图",对片段的测序工作已基本完成,剩下的主要是如何将它们正确地连接起来.这项工作是靠国际合作完成的,作为发展中国家,我们中国也承担了1%的任务.

最近刚完成的21号染色体的测序是由日、德、美、英、法和瑞士的科学家合作完成的.共测定了3300多万个bp,占人类基因组大小的1%.研究者们精巧地设计了酶切和克隆实验,使复制出的DNA小片段间保持有部分的重叠关系,构成"连续克隆系"(contig).严谨的工作保证了最长的连续DNA测序覆盖了2500多万bp.迄今为止这是有报告的最长连续测序.在整条序列中只有三处间断,漏测的长度仅为10万bp.分析表明,在21号染色体中共有225个基因(包括127个已知基因和98个预测基因),以及59个赝基因.127个已知基因的平均长度为57000bp,其中有22个的长度大于10万bp,最长的(符号为DSCAM)含有833627个bp.98个预测基因的平均长度为27000bp.这项工作以详细的列表给出了所有这些基因的符号、登记号、基因描述、类别、取向、起始和终止碱基对位置、基因长度和相应的基因克隆名称.

遗传物质DNA在生命现象中占有中心地位,基因的缺陷往往导致疾病,甚至是致命的.比如,第21号染色体异常(如果多一条,有三条)就会导致"塘氏综合症".这种病症的发生率大约是七百分之一,多由于母亲年纪过大(>40岁)而发生,主要表现为智力发展障碍或心脏畸形.某些人群可能带有对某些疾病(如癌症、心血管病、老年痴呆等)敏感的或具有抵抗性的基因,某些基因突变也可能导致另外一些疾病.识别出这些基因并研究他们的致病机理是关系人类健康的大事.所以除政府大力资助外,私立机构和公司也都积极地投入和关注这项工作.比如成立于1998年的美国公司"Celera"就要单独完成人类基因组测序工作.英国"Wellcome Trust"基金会也支持英国医药委员会的"Sanger Centre"完成人类基因组三分之一的测序工作.

人类基因组测序工作接近尾声,不光是在科技界,在社会上也带来了对基因的热情.是否像有些人认为的那样,测序的完成就意味着遗传密码的"破译",人类的一切疾病就都可以治愈,人类的寿命也可以延长到200岁,或者更多?现实的答案远没有如此乐观.事实上,在这里任何的盲目乐观都有炒作之嫌.人类基因组的测序工作只是把记录着生命之谜的天书逐字地读了一遍(也许是数遍,因为为了可靠,每个碱基对平均都测了7次).也像是把这本天书打开,摆在了世人的面前.里面的文字清晰可见,但谁能读懂它呢?正像洛克菲勒基因组研究所的科学家所说,"谁也不完全知道该怎么做."为了理解它,给它做出全面的注译,科学家们需要找到里面包含的所有基因和调控这些基因的序列.而目前根据碱基序列预测基因的各种方法都有自己的缺陷,可靠性尚有待提高.测序公司正期望从对DNA序列的正确解释中得到它们巨额投资的回报.还有许多被称为"垃圾"DNA的区段,它们真的是"垃圾"吗?关于基因的总数,目前就是一个很有争议的问题.多数人的估计是在3.5万到15万之间,许多科学家甚至还为此而打赌.还有科学家建议,通过国际互联网,全世界的科学家们都来注译这部用DNA的碱基写成的记录着全部生命奥秘的天书.此外,更多的问题是关于这些基因的表达是如何受哪些因素调控的,他们的表达产物——蛋白质是如何形成一定的结构并互相识别或装配进而发挥特定的生物功能的.但是,无论如何,HGP把我们带入了基因的世纪,我们正在向着理解生命之谜前进.

参 考 文 献

[1] Macilwain C. Nature,2000,405:983

[2] Butler D,Smaglik P. Nature,2000,405:984

[3] 陈竺.自然科学进展,2000,10:208[CHEN Zhu. Advances of Natural Sciences,2000,10:208(in Chinese)]

[4] 陈竺等.自然杂志,2000,22:125[CHEN Zhu *et al*. Nature Journal,2000,22:125(in Chinese)]

[5] Reeves R H. Nature,2000,405:283

[6] 21号染色体定位和测序组. Nature,2000,405:311