DOI:10.16262/j.cnki.1000-8217.2000.02.014

·成果简介·

人类基因组计划中目前的几个重要研究领域

江虎军* 王钦南* 邓汉湘†

(*国家自然科学基金委员会生命科学部,北京 100083;
†湖南医科大学医学遗传国家重点实验室,长沙 410078)

[关键词] 人类基因组,测序,生物芯片,基因

自1990年10月1日美国在世界上第一个实施人类基因组计划以来,已整整过去了9年。这项被称之与"曼哈顿计划"、"登月计划"相媲美的巨大生命工程提出之初,便受到全世界的科学家、新闻媒体等各界的关注,并就此展开了激烈的讨论,沸沸扬扬中研究计划启动了。时至今日,当初的吵闹声不见了,惊喜代替了担忧,乐观代替了悲观,积极的参与代替了冷漠的旁观。由于新的技术和方法的不断出现,人类基因组计划极有可能要比原计划提前完成,或许多数人想急于了解该计划目前进展如何?现在主要研究些什么?

1 人类基因组测序

人类基因组计划的最终目的是希望用15年时间,完成人全部23对染色体的遗传图谱、物理图谱、序列图谱和转录图谱,提供一张精度达99.99%的人类基因组序列图是整个研究计划的重点和难点,因此,测序手段的改进和测序方法的创新一直是该计划的资助重点。

目前,世界上已有美国、英国、法国、德国和日本5个国家正在进行人基因组的大规模测序,进展速度异常之快。1999年12月2日英国的The Sanger Center在《NATURE》上正式公布完成了人第22号染色体的全序列测定(其中尚有11个序列间隙有待填补)[1]。这是待测序的24条人染色体中第一条完成测序的人染色体。该染色体不大,仅33.4 Mb,至少包括有545个基因和134个假基因。通过这次测序新发现148个基因。令人感兴趣的是在该染色体上分布有与人的一些遗传病有关的基因,如Cat Eye syndrome (CES), Velocardiofacial/DiGeorge Syndrome (VCFS), Schizophrenia susceptibility 和 Spinocerebellar Ataxia(SCA10)等。显而易见,这项工作的完成使得科学家们今后能较容易地从DNA序列中鉴定出而不是分离出所感兴趣的基因。这也说明开展人基因组大规模测序的重要性。

我国科学家继上述5国之后也加入了人类基因组大规模测序这一国际大协作。1999年7月在国际上注册承担了全部人基因组1%的测序任务。工作区域位于3号染色体的3 pter至D3S3610之间,约30 Mb,预计2000年3月底之前完成工作草图(working draft)。面临的任务是艰巨的,但更大的困难是研究经费的严重不足。

人类基因组计划实施9年以来,已完成了整个基因组约16.7%(截止于1999年12月16日)的测序任务。从现在的测序手段和方法的进展看,估计到2001将完成全部测序任务的1/3,到2003年将基本完成全部测序。比原计划预定的2005年提前2年完成,以此纪念DNA双螺旋结构发现50周年。值得一提的是,由于新的测序手段和方法在大规模测序中的使用,有专家认为至2001年即可完成人全基因组测序。

为了配合研究人基因的结构和功能,人类基因组计划中除完成人基因组测序外,还包括对大肠杆菌(*E. coli*)、酵母(*S. cerevisiae*)、线虫(*C. elegans*)、果蝇(*D. melanogaster*)和老鼠(*M. musculus*)等5种模式生物的测序。到目前为止,已完成了大肠杆菌、酵母和线虫的全部基因组测序。估计到2004年左右将完成果蝇的全基因组测序,2005年完成老鼠的全基因组测序(表1)。可以肯定,随着测序手段的完善和测序速度的加快,还将对其他与人类活动密

本文于1999年12月15日收到.

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

切相关的生物的基因组测序,从根本上揭示生命的本质。比如美国能源部(Department of Energy, DOE)自 1994 年开始了微生物基因组计划(Microbial Genome Project, MGP),估计到目前为止,约有 24 种以上经过精心选择的重要的微生物基因组完成了全部测序。可以肯定地说,今后将会完成越来越多生物的全基因组测序。

表 1　人类基因组计划模式生物基因组测序进展*

模式生物名称	基因组大小(*Mb*)	估计基因数	测序完成时间
大肠杆菌(*Escherichia coli*)	4.6	4 288	1997 年完成
酵母(*Saccharomyces cerevisiae*)	12.1	6 034	1997 年完成
线虫(*Caenorhabditis elegans*)	100	11 800—13 800	1998 年完成
果蝇(*Drosophila melanogaster*)	165	15 000—25 000	预计 2004 年完成
老鼠(*Mus musculus*)	3 000	5 000—10 0000	预计 2005 年完成

* 资料截止日期:1999 年 10 月

在测序手段方面,Perkin-Elmer 推出了 ABI PRISM 3700 全自动 DNA 测序仪。它是在 ABI PRISM 377 的基础上改进而来。加样孔由原来的 36 孔增加到 384 孔,全部采用毛细管自动加样,无须人工制作测序胶,与 ABI PRISM 377 相比省时省力达 80% 以上。每台仪器每天可完成 1 000 个样品测序,测序长度达 0.5 Mb/d。

在测序方法学上,鸟枪法测序(Shotgun Sequencing)是目前的一种全新的测序方法[4]。传统的测序方法采用的策略是从克隆到克隆(clone-by-clone),即通过对人基因组 BAC 文库中各克隆的测序而获得全基因组序列图。测序过程中,须对每一个 BAC 克隆进行亚克隆,然后对亚克隆进行直接测序,再将各亚克隆的序列通过 DNA 片段两端的重叠性将所有亚克隆的序列连起来;采用同样的方法再将各 BAC 克隆的序列连起来,进而拼接成一连续的序列图。显然,采用传统方法测序的中间环节多,测序速度受到限制,但准确性高。

鸟枪法测序的基本原理是在人基因组随机文库的基础上,直接对随机文库中的各克隆进行测序,因为克隆中插入的人 DNA 片段只有 0.4—1.2 kb 或 5—20 kb 2 种类型,同时从 3′ 和 5′ 两端对同一克隆测序。对于插入片段短的克隆(0.4—1.2 kb),只一次即能完成该克隆的测序;对于插入片段长的克隆(5—20 kb),通过两端同时测序后可获得大量的两端序列已知而中间部分未知的克隆群,然后通过专门的计算机软件将测得的序列拼接成连续的序列图。理论上,通过测 10 倍于人基因组的序列,即 3×10^{10} bp,所测序列便可覆盖人的全部基因组。由于所构建随机文库的克隆可直接用于测序,且每个克隆只测序一次,省去了许多中间环节,简化了原来的测序方法,因而大大加快了测序速度。鸟枪法测序刚被提出来时,人们担心的是如何开发出专门用于拼接全部克隆序列的计算机软件,并使拼接错误尽可能低。现在看来问题已得到很好的解决。

基于鸟枪法测序原理,美国能源部已资助美国基因组研究所(TIGR)和华盛顿大学人基因组测序中心(University of Washington)用鸟枪法开展大规模测序,共配备 230 台 ABI PRISM 3700 DNA 自动测序仪,估计每天将完成 100Mb 原始序列。由于目前已积累有大量现成的人基因组 BAC 克隆的 3′ 端和 5′ 端的序列,所以计划只对 30 万个 BAC 克隆进行鸟枪法测序。此项工作的结束将使得人类基因组中每 5 kb 即能提供一个标记,无疑将加快测序速度,提高所测序列的精确性。

鸟枪测序法的另一个明显的优点是可以检测到人基因组中大量的 DNA 多态性(DNA Polymorphism),又称单核苷酸多态性(Single Nucleotide Polymorphism, SNPs)。对 SNP 的研究已被列入美国人类基因组计划最后的一个五年计划(1999—2003)的资助重点,并提出要建立一张包含至少 10 万个 SNPs 的遗传图谱,使得在人基因组中每 3 万个核苷酸中即有一个 SNP 标记[2,3]。这种高密度的 SNP 遗传图将有利于开展人全基因组研究,特别是对寻找复杂性状疾病如肿瘤、糖尿病、精神病、高血压等的相关基因,并据此开发出针对不同人的特效药物。

鸟枪测序法的测序成本也较传统的从克隆到克隆测序法便宜,估计每个核苷酸的测序费用可从0.3 美元降为 0.01 美元,完成全部人基因组的费用可从 9 亿美元降为 5 亿美元,测序时间由原来的 7 年缩短为 3 年。尽管鸟枪测序法提出之初受到多方面的质疑,但从现在的实际效果看,其优点已得到进一步的肯定。

2　生物芯片技术

随着大量基因的被克隆,人类基因组下一步的

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

工作重点和难点便是分析基因的功能，了解基因在时间和空间上的表达特征。生物芯片技术正是在这种需求下应运产生，它集生物技术，微电子加工技术，分析化学技术，激光技术，甚至计算机软件技术等于一体。生物芯片技术的发展被评为1998年美国十大科技进步之一。

最早研制出来的生物芯片是cDNA芯片(cDNA Chip)。在一个约为20 mm×45 mm大小的芯片上，可以将数千个甚至上万个cDNA密集排列并固定在由特殊材料制成的微小芯片表面，通过一次原位杂交反应(Fluenscence in situ Hybridization，FISH)，即可同时获得所有cDNA的杂交信号，根据信号的有无强弱，去判断不同组织中基因的表达情况，描绘出该组织中基因的表达特征，进而鉴定与该组织功能相关的基因。

组织芯片(Tissure Chip)是芯片技术的进一步发展。1998年美国人类基因组研究所的科学家报道[5]，先将病变组织在冷酒精中固定再用石蜡包埋，以保证组织中DNA和RNA保持在原位，然后用微小取样器(直径0.6 mm，高3—4 mm)在包埋好的组织的不同部位打孔取样，每次取样的厚度仅为4—8 μm，在组织的同一位置可连续取样达200次，样品排列在20 mm×45 mm大小的石蜡块上，然后进行原位杂交，即能获得病变组织中某基因的空间表达特征。用乳腺癌、前列腺癌为研究对象，不仅验证了这种方法的科学性和准确性，还获得了新的发现。

1998年底，美国加州一生物技术公司宣布已成功开发出世界上第一块蛋白质芯片(Protein Chip)[6]。具体做法是将已知的抗体、受体和合成的化学分子排列在微小芯片上，利用生物分子间的相互作用，原位杂交之后即可从一组织中捕捉到与这些分子相互结合的蛋白质分子，再用Laser Zaps技术除去芯片上未结合的蛋白质分子，再用质谱仪(Mass Spectrometer)读出结合在芯片上各蛋白质的分子量，得到被测样品中蛋白质指纹图(Protein Fingerprint)。比较正常人与病人的蛋白质指纹图，便能找到与疾病可能有关的蛋白质分子。专家们分析了一种细菌(Haemophilus influenzae)分裂状态和静止状态细胞，将该细菌全部1 740种蛋白质中的600种做成蛋白质芯片，几个小时后便得到结果，发现细胞在这2种状态下约有30—60种蛋白质的差异，提示这些蛋白质可能与细胞的分裂有关。蛋白质芯片技术还可用于蛋白质纯化和氨基酸序列的测定。

生物芯片技术的特点在于快速和大规模的"搜索"性研究，灵敏度高省时又省力。估计美国约有10家左右的生物技术公司从事生物芯片技术的研究和开发，商业化的组织芯片和蛋白质芯片不久将大量投放市场。可以预见神经细胞的生物芯片将是今后生物芯片研究领域新的热点。

3 人基因的定位和克隆

人究竟有多少基因，目前尚未有确切的定论。有人认为约5万个基因，也有人认为在6万—8万个之间，甚至有人认为应在10万个以上，也许准确的数字即便到2003年测序全部结束也不一定完全清楚。在世界各地的科学期刊上，每天都有关于人新基因的报道。目前，已定位人基因和基因片段3万个以上，已克隆基因1.8万个左右，其中定位人疾病相关基因约5 000个以上，已克隆约1 300个左右，估计人约有1万个以上的疾病相关基因。

在基因的克隆过程中，有2种类型的基因格外引人注意：一是具有重要社会行为功能的基因；二是与疾病相关的基因。因为根据这些基因可以开发出新的药物，所以这些基因一旦被克隆便会被专利保护其知识产权。

所谓重要社会行为功能基因是指能影响人的日常生活等但并不一定表现为一种具体疾病的基因，如与人的寿命、肥胖、睡眠、记忆、行为、智商等相关基因。目前只有人生物钟基因、肥胖基因和苗条基因等少数几个这方面的基因被克隆，主要原因是由于缺少研究这些基因的家系，实际上也很难搜集这样的家系，无法通过正常的途径去定位这些基因，更谈不上克隆该基因。目前主要依赖对模式生物相应的基因保守序列的同源分析去获取这些基因。1998年《NATURE》上报道[7]，科学家们已从果蝇中鉴定出一种与其寿命相关的基因，携带这种基因的果蝇较普通果蝇的寿命长35%，为将来鉴定人寿命相关基因打下了基础，这也说明了为什么在人基因组计划中需要对老鼠、果蝇等模式生物同时开展研究。

人的疾病相关基因的研究一直是遗传学领域的一个研究热点。目前已克隆到的疾病相关基因几乎全是以孟德尔方式遗传的质量性状基因。然而，不少严重威胁人生命的疾病，如肿瘤、高血压、糖尿病等，均表现为多基因性状，发病的早晚轻重与环境、个人生活方式等有关，研究这些疾病相关基因已成为今后的一个重点和难点。与神经系统功能有关的基因也是目前研究的焦点，如与Alzheimer Disease (AD)，Huntington Disease (HD)，Parkinson Disease

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

(PD), Amyotrophic Lateral Sclerosis(ALS), Spinocerebellar Ataxia(SCA)等相关的基因。令人感兴趣的是有些神经系统疾病的发病机制比较特殊,正常人与病人间的遗传差异表现在相关基因中某一段核苷酸序列的重复单元次数的不同(如 CAGCAGCAGCAG···),且存在一定的阈值,当重复次数超过阈值时即表现出病状,且发病的早晚轻重随重复次数的增加而增加。这种核苷酸重复序列多存在外显子(Exon)中,但也存在内含子(Intron)中,甚至在基因的表达调控区也有发现。病人的神经细胞中往往形成包含体(Inclusion Body),发病机制仍不清楚。

4 基因组流行病学

在现实生活中,不同人群间由多基因控制疾病的发生概率明显不同,如我国南方广东人群的鼻咽癌发病率明显较其他地区人群高,而山西、河南人群的食管癌的发病率则显著高于其他地区人群。要真正认清这些疾病的机制,仅从个体出发进行研究是远不够的,需要以整个地区人群为研究对象,要研究疾病相关基因及其基因频率、生活环境、当地人群的饮食习惯及其相互间的作用等,从基因水平开展流行病学研究。专家们呼吁应尽快开展如下四方面的工作[8]:

(1)研究不同人群间疾病相关基因的变异频率;

(2)确定不同人群间基因变异与疾病发生间的密切程度;

(3)研究不同人群间基因与基因、基因与环境互作对疾病发生的影响;

(4)寻找针对不同人群有效的遗传学检测方法,并制定相应的措施以提高预防疾病的能力。

5 讨　论

人类基因组计划的实施正在改变着整个生命科学研究,使人们能从分子水平认识生命的本质。美国花巨资开展这一庞大的生命工程确是人类的一件喜事,蕴藏在其中的财富一时是无法衡量的。正因为如此,估计约有18个国家相继启动了本国的人类基因组计划。

1993年我国也加入了这一行列,并根据我国的具体情况选取了几个有限的领域开展研究。面对国际上人类基因组计划进展的加快,我国科学家对自己的战略也作出了相应的调整,并加大了投资力度,已在急性早幼粒白血病研究、神经性耳聋基因的克隆、不同民族的多样性研究等领域取得令世界注目的成就。这些成绩的取得与当初所制定的战略及后来及时的调整是分不开的。

人类基因组计划的重点——全基因组测序估计在2003年全部完成,甚至更多的其他重要生物的基因组的测序也将陆续完成,鉴于此,以后的研究重点将转向寻找基因,并研究其功能。如何从基因组DNA序列库中获取基因并开展功能研究,不仅需要有研究的材料,还须具有有效的研究手段。收集遗传病家系是寻找基因的基础,因为只有通过对家系的分析才能从序列库中找到基因,否则面对巨大的序列库会感到无从着手。开展基因的功能研究更需要有强有力的研究方法。在这两个方面,尽管我们已有一些工作基础,但仍很不够,还须作出较大的努力才行。

参 考 文 献

[1] Dunham A R, Hunt j, Collins E et al. The DNA Sequence of Human Chromosome 22. Nature, 1999, **402**:489—495.

[2] Collins S, Ari P, Jordan E et al. New goals for the U. S. human genome project: 1998—2003. Science, 1998, **282**:682—689.

[3] Deloukas P, Schuler G D, Gyapay G et al. A physical map of 3000 human genes. Science, 1998, **282**:744—746.

[4] Venter J, Craig D, Adams G et al. Shotgun sequencing of the human genome, Science, 1998, **280**:1 540—1 542.

[5] Juha K, Bubendorf L, Kallioniemi A et al. Tissue microarrays for high-througput molecular profiling of tumor specimens. Nature Medicine, 1998, **4**: 844—847.

[6] Strauss E. New Ways to Probe the Moleculars of Life. Science, 1998, **282**, 1 406—1 407.

[7] Lin YiJyun, Seroude L, Benzer S. Extended lige-span and stress resistance in the drosoghila mutant methuselah. Science, 1998, **282**: 943—945.

[8] Khoury M J. The human genome epidemiology network. American Journal of Epidemiology. 1998, **148** (1):1—3.

THE PROGRESS OF MAIN RESEARCH AREAS IN HUMAN GENOME PROJECT

Jiang Hujun* Wang Qinnan* Deng Hanxiang†

(* *Department of Life Sciences, NSFC, Beijing* 100083; †*National Laboratory of Medical Genetics, Hunan Medical University, Changcha* 410078)

Key words human genome, sequence, biological chip, gene

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net