

人基因组顺序测定蓝图

——一项直观探测人体本质的巨大科学工程

THE BLUEPRINT OF HUMAN GENOME SEQUENCES

——A GIGANTIC SCIENCE PROJECT OF DIRECT
DETECTION OF THE NATURE OF HUMAN BODY

李以莞 顾方舟

编者按：近年来，国际生物学界对人全基因组图谱绘制和顺序测定的必要性及可行性进行了广泛的讨论。鉴于这一讨论及其展现的动向和前景对生物科学和生物工程有深远的影响，故特约此稿以飨读者。这项讨论的内容，主要涉及：进行人基因组顺序测定的可能性如何？要用多长时间？花多少钱？有什么用？生物学家的聪明才智用在这个大课题上是否值得？本文作者将讨论中涉及的主要问题及技术开发、实施阶段及其应用等方面的讨论意见归纳整理如下。

生命现象最主要的特征之一是遗传性。对遗传规律的研究有助于对生命本质的认识。遗传信息是通过染色体传递的，染色体存在于人的每个细胞的核内，它的基本组成成分是双链脱氧核糖核酸（DNA）分子。每条DNA链是由四种脱氧核糖核苷酸的线性多聚体组成。这四种核苷酸中的碱基是腺嘌呤（A）、胞嘧啶（C）、鸟嘌呤（G）及胸腺嘧啶（T）。两条DNA链上的A和T，C和G是互补配对的。在每个细胞中由这两两配对的互补链组成的染色体就是人的基因组，它约含有30亿个碱基对，包括了人的全部遗传信息。这30亿个碱基对大约可划分为100000个遗传信息片段，每个片段称为基因。一般说来每个基因编码一种蛋白质，蛋白质是由20种不同的氨基酸组成的线性多聚体，每三个碱基编码20种氨基酸中之一。一般说来，蛋白质由300个左右氨基酸以特定的顺序连接起来组成多肽链，多肽链又盘旋折叠形成蛋白质的三维空间结构。蛋白质活动是生命活动的基础，如已知的数千种高效专一的生物催化剂——酶就是蛋白质，人的皮肤、肌肉，动物的蹄、角也是由蛋白质组成，维持人

生命所必需的氧也是靠红血球中的血红蛋白所输送。因此，蛋白质赋予我们机体的外形和功能，并进行着生命的化学反应。由此，我们可以说研究人染色体和基因的结构与功能是朝向了解人类本身的存在，阐明人类生命现象迈出的首要的一步。这项拟议中的课题，由两个密切相关的部分组成。第一部分是绘制人基因组的物理图谱，即将人基因组切割为小片段，进行克隆，将克隆按人基因组的顺序排列，编号成一条直线。这个按顺序排列的克隆就可以作为研究某些特定基因的原始材料。第二部分是确定每个克隆的DNA碱基顺序，经过分析、排列，最后确定人基因组约30亿个碱基的排列顺序。

关于第一部分，根据科学家的估计，大约要用5年左右的时间才可以完成。国际上绝大多数生物学家认为这是一项有意义的工作，值得为之而努力。但对第二部分，即测定约30亿个碱基顺序问题，争论较大，一些生物学家认为是时间和金钱的浪费，并且是无意义的。他们认为人细胞中大约10万个基因，只由5%的DNA编码，其余95%的DNA，他们认为是“没用的东西”。

李以莞（Li Yiwan）中国医学科学院肿瘤研究所副教授
顾方舟（Gu Fangzhou）中国医学科学院院长

美国加州理工学院生物学部的史密斯(Smith)和胡德(Hood)对此有不同的看法。他们最近发表文章指出:使用现有的技术方法进行人基因组顺序测定的确代价昂贵,要消耗大量时间。他们的战略方针是首先建立测序自动化技术,以加快速度降低代价,并且他们及其他一些科学家已在自动化测序技术方面有了进展。他们认为现在还没有什么方法可以用来区别哪些DNA是“有用的”;哪些DNA是“没用的”。那些被称为“没用的”DNA中,可能具有尚未被了解的功能。染色体可自我复制并转录信息,它依照自然律进行有丝分裂和减数分裂,并参与重组;染色体的结构在区域突变率中起作用,并由此影响进化过程,那些所谓“没用的”DNA顺序可能和上述染色体的这些功能有关。他们举例说人基因组顺序在生物学中的作用就象粒子加速器在物理学中的作用,一旦这个顺序测定出来,就是一个强有力的工具,人们可以利用它进行新的前沿知识的探索。

史密斯和胡德提出的人基因组图谱绘制及顺序测定的方案是将每个染色体按顺序切割为一个一个小片段,使之能进行顺序测定。例如:第一步,先提纯单个染色体,将染色体片段依次克隆进粘性质粒载体中,这种载体是迄今已有的重组DNA载体中可容纳最大插入片段(40 000碱基)的载体。一系列这样按顺序排列的粘性质粒就包装了全部染色体,组成染色体图谱。每个粘性质粒可以进一步切割成为较小的片段,大约1 000碱基左右,使其大小适合于顺序测定。他们强调这个方案能否有效实施取决于全过程中各个阶段所使用的先进技术。这些技术包括以下几个方面:

一、染色体分离

目前已经研制了自动化的激光激活染色体分选仪(automated laser-activated chromosome sorting machines)。这种仪器可以较好的分离出人的小染色体,但对较大的染色体有点困难。对较大的染色体可以从人一啮齿动物杂交细胞株中选择出只含一个人染色体和全部啮齿动物染色体的细胞,经培养后可以用来分离较大的人染色体。

二、物理图谱的绘制

绘制人DNA物理图谱目前通用的方法是用限制性内切酶在染色体DNA特异碱基位点上将染色体切割为片段。将切割后的片段,插入载体,制成重组DNA文库,期望此文库可包括人基因组绝大部分DNA。通过限制性内切酶切点的选择,可以产生一系列其DNA序列首尾部分重叠的克隆。利用重叠的部分,可以将染色体的克隆排列起来。这样,人基因组的物理图谱就由一组首尾相互部分重叠的,容纳了最大DNA片段的一

个个载体组成。目前,以粘性质粒作为载体,最大的克隆片段为40 000碱基左右。最近对一种叫作*Caenorhabditis elegans*的蠕虫的基因组进行了研究,它的基因组长度约为8 000万碱基对,大约相当于一个中等大小的人染色体的长度,这个虫子的80%的基因组已克隆进粘性质粒,制成了物理图谱,用了约三年的时间。人的基因组除了比它大30多倍外,还有大量的重复DNA顺序,这对粘性质粒克隆和绘制图谱都将增加一些困难,但从目前的情况来看,不管怎样,经过努力,人的基因组的物理图谱有可能在五年左右完成。然后就可以利用遗传图谱技术在染色体上进行基因定位了。

上面提到了,目前最大的克隆片段为40 000碱基左右,假若在40 000碱基中有10 000碱基的长度是重叠的,那么共需制出约12万个克隆才能将人基因组全部克隆进载体。而且,开始时要建立比所需要的数目多得多的克隆,才能通过限制性内切酶切点分析、比较后,确定一组数目最小的、包括了绝大部分人基因组的克隆。这就需要建立快速的、自动化的制图及分析比较技术,建立可容纳更长的,如上百万碱基的DNA插入片段的克隆载体,研制可以切割DNA链上任何靶点的DNA切割反应剂。这些技术,有的已有很大的进展,一旦成功,将大大简化和加速人基因组物理图谱的建成。

三、DNA测序

目前广泛使用的DNA顺序测定的两种方法是酶法和化学降解法。两种方法都是制备出一系列具有相同起始点,但长短不一的放射性标记的DNA片段。两种方法不同之处在于用以产生DNA片段的方法不同。在酶法中,在以被测序列为模板合成DNA新链的反应液中,加入经修饰的碱基(即双脱氧核糖核苷三磷酸)。这样,在DNA链合成过程中,一旦双脱氧核糖核苷三磷酸整合到新合成链上,DNA链的合成就终止。化学降解法是直接用化学方法处理获得与上述相同的一组DNA片段,在凝胶电泳中分离同位素标记的、长度不等的DNA片段。这种凝胶可以分辨出一个碱基长度的区别,用自显影法获得DNA片段在凝胶中的图像,读出4个测序反应所获得的图像,就产生了DNA片段的顺序。

用上述两种方法已测了将近1 200万个碱基,这是美国目前DNA顺序资料库GENEBANK的储存量,其中约100万个碱基是人的基因组。这个储存量正以每年30%的量在增加。对一个熟手来讲,用化学法在单链DNA上测序时误差约为0.5%,用酶法则误差较大。现今,在最好的

条件下,一个熟练的化学家,每年约可测定50 000个碱基。到目前为止,人的基因组已经测定的顺序仅占全基因组的0.03%。显然,需要更快速的测定方法。一些科学家和工程师正在设计、制造非同位素的自动化的DNA测序方法。如美国加州的史密斯等研制了用于DNA顺序测定的自动化仪器。他们的设计原理是将用酶法产生的一组DNA片段与荧光染料相结合,在测序的四个反应中,每个反应使用不同颜色的染料,混合在一起,在聚丙烯酰胺凝胶中电泳。凝胶底部装有高度敏感的荧光探针,在一定的间隔时间内测量荧光点。从按时间顺序出现的、不同颜色染料所特有的波型可以判定DNA的顺序。美国应用生物系统(Applied Biosystem)公司根据这个原理所制造的仪器已经上市。使用这种仪器,通过电脑可同时从16个聚丙烯酰胺凝胶泳道中读出DNA顺序。目前,每个泳道每14小时可测出大约300个碱基。从理论上讲,这个仪器有可能每天测10 000个碱基,准确率达98~99%。美国特拉华州杜邦公司一组工程技术人员于去年10月发表了他们研制的荧光标记DNA自动测序仪,原理上与上述基本相同。他们用此仪器测了M₁₃mp18第3~第234碱基,未发现错误,也没有漏测或多测的碱基。在日本,以东京大学秋吉和田为首的科学和工业联营集团采取的是装配线生产方式。建立测序工厂,全过程的每一个步骤都是自动化的,整个装配线的运转不需要非熟练的技术人员。

自动化DNA测序的程序比起常规所使用的方法,大大减少了劳力,并且不使用同位素,减少放射性公害。但是,从目前的进展来看,还是相当不成熟的技术,有待进一步完善和发展。

四、资料贮存和分析

目前,贮存DNA顺序资料的设备远远不能满足新资料的要求。这里有两方面的问题,一个是贮存的容量,一个是贮存的资料的注释和分析。由于对入库资料未作充分注释或注释方法各实验室不一致,因而直接影响这些资料的利用。因此需要设计更好的程序以便将作了充分注释的DNA顺序输入资料库。建立有相当贮存能力的资料库,还必须有更好的方式能对不断增加的资料进行快速分析。特制的集成电路芯片(chips)和超电脑在这方面将发挥极大作用。研制新的软件,必要时应用人工智能(artificial intelligence)方法都在考虑和酝酿中。

人基因组顺序的测定包括三个阶段:第一阶段,约需5~10年,是新的自动化技术发展阶段,旨在能快速、大量、准确地进行DNA的克隆、

制图、测序和分析。与此同时,进行人基因组遗传和物理图谱的制备。同时可对一些在医学上有重要意义的DNA区段,如多型现象,或对人类进化有意义的DNA区段进行顺序测定。第二阶段,完成人基因组顺序测定。这一阶段一定要在DNA顺序测定的自动化程度、误差、速度均达到较理想的水平时开始。开头的5~10%的顺序测定,可能由一些研究室进行,那里的研究人员对分析某些基因感兴趣。但由于研究人员没有那么多的经费、人力和兴趣完成全基因组的顺序测定,余下的顺序测定可由几个大的中心站一类性质的机构来完成。在此阶段中,对其他动物基因组顺序的测定也将广泛开展。第三阶段的任务是继续对基因组进行分析,包括人的和非人的,并对获得的资料进行开发和应用。

人基因组顺序一旦测出,将在以下几个方面的工作中产生深远的影响:

1. 遗传病

人的遗传病是由一个或一个以上的缺陷基因引起的。对全基因组的研究可使人们确定哪些基因与疾病有关。举Huntington氏病为例,这种病发病较晚(在30~40岁间),表现为进行性的精神失常及肌肉运动失控,常是致死性的。与此病有关的基因已经定位在人4号染色体上大约100万个碱基的区域内。目前的挑战是,在这100万碱基范围内,这个基因的确切位置在哪里。由于基因具有与非编码DNA区域不同的特点,人们可使用已建立的电脑程序,浏览DNA顺序资料,确定编码基因的可能位点。由此人们便可以在这100万个碱基中确定Huntington氏病的基因。一旦这个可能的基因被确定后,就可以使用重组DNA技术和生物化学的方法在相应的组织中找出这个基因编码的mRNA和蛋白质,继而对相应的基因进行分析,看是否发生了点突变。目前已知的遗传病约有3 000多种,但是已确定基因位点的没有几种。随着医学的发展,人类绝大多数传染病被控制或被消灭后,在肿瘤、心血管疾病的研究受到高度重视的同时,遗传病的研究将提到日程上来,并受到更多的关注。

2. 多型现象(Polymorphisms)

个人与个人之间约有0.1%的核苷酸不同,这就是所谓的多型现象,或翻译为“多型性”。因此除孪生子外,每个人都是不同的,有如每个人的指纹都不同一样。多型现象与医学有深奥的关连,例如组织相容性复合物基因,这是与人免疫有关的基因,这个基因的多型现象常常与后天自家免疫疾病有关:如糖尿病、多发性硬化(Sclerosis)。对人基因组顺序的研究,将使人们能对某些

感兴趣的基因的多型现象进行分析,并了解它们与疾病的关系。另外,人基因组的多型现象还可应用于法医学,对可疑的凶犯进行生物学的鉴别。

3. 生理学

人基因组顺序对阐明人的生理学是一个有力的工具。例如:T细胞,这是免疫系统中淋巴细胞中之一,它在各种自家免疫疾病中起着重要作用。T细胞表面的受体蛋白,使细胞具有识别能引起疾病的微生物的能力,也可能诱导自家免疫疾病。人基因组顺序测定后,可以确定编码受体蛋白的基因区,合成探针,进行基因定位;还可合成抗体探针,以鉴定各种T细胞受体蛋白,用于诊断和治疗与T细胞异常有关的疾病。

4. 癌

癌细胞与正常细胞的主要区别是癌细胞的分裂处于失控状态。癌的研究的主要内容之一是研究癌细胞与正常细胞的区别的分子机理。可利用凝胶电泳分析技术,将细胞的近5 000种蛋白,按其电荷及分子大小加以区分。利用电脑程序,将蛋白这两项“指纹”进行比较,确定癌细胞所特有的蛋白。分离癌蛋白,以高度敏感的微化学仪器——蛋白质测序仪确定其氨基酸顺序,译出其DNA顺序。利用电脑比较分析此疑为癌基因的DNA片段的顺序与人基因组顺序中是否有相似之处。若能发现相似之处,则可进一步研究此癌蛋白对细胞的作用及致癌机理。这样就可节省大量时间用于未知基因的克隆。

5. 进化

将人的基因组顺序与其他动物的基因组顺序进行比较分析,其DNA顺序相同的那部分是高度保守的序列。这部分包括编码区、调控区和其他可能与真核细胞染色体特殊功能相关的区域。高度保守的编码区基因是经受了进化压力选择后保留下来的基因。这些资料对研究进化机理是非常有用的。

6. 蛋白质折叠

蛋白质结构与功能的研究可从两方面着手:实验性的及理论性的。实验性的研究中包括利用DNA重组技术改变某个基因的核苷酸顺序,研究其编码的蛋白与正常基因编码的蛋白二者之间在结构与功能方面的异同。理论性的研究包括研究蛋白质折叠三维空间结构。绝大部分人的基因由2~3个被非编码DNA区段间隔开的外显子组成。外显子可能编码一个特定的折叠方式,它具有特定的功能。可能有一定数量的外显子,它们为蛋白质的折叠方式编码。人基因组顺序测定后,可深入了解蛋白质折叠三维空间结构的规律,进一步了解其功能。然后可利用蛋白质工程技术,人工构建具有特殊功能的蛋白质,如构建耐温的酶,在温度调控下加速其反应过程。

这个大课题所需的经费是一个庞大的数字。据有关专家估算,使用美国目前的技术测定一个碱基约需1美元,DNA双链全长的测定就需要60亿美元的预算。在日本,由于实现了技术半自动化,每个碱基只需0.17美元,DNA双链全长的测定需10亿美元预算。若能实现全过程技术的自动化,每个碱基测定的费用有可能降至1美分,则全部经费需6 000万美元。科学家呼吁各国政府、财团、工商界、富豪的资助,共同完成这项直观的了解人类本身本质的伟大科学工程。

· 简讯 ·

北京MIT校友会和美国MIT校友会本部

1987年10月31日在欧美同学会总部召开了北京麻省理工学院校友会成立大会。在京的30多名MIT校友参加了大会,他们中既有年过八旬,20年代就留学MIT的老学友,也有新近留学刚归来的中青年校友。会上讨论并通过了北京麻省理工学院校友会章程。根据章程,凡在北京居住的、曾在美国麻省理工学院毕业、肄业、任访问学者或研究人员、教职工的校友,均可成为会员;在京工作的外籍校长可成为海外会员。会上推选清华大学前副校长赵访熊教授为会长,吴仲华、梁守槃、侯祥麟、陆元九、崔国文、陈东威、张钦楠、陈德蔚为杨友麒为理事。

北京麻省理工学院校友会是欧美同学会下属的一个分会。正在与美国麻省理工学院校友会本部联系成为该会的一个分会。

美国MIT校友会本部现有会员七万余人。近60个分会遍布世界各地。会员每年都为母校捐助可观的校友基金。校友会出版《技术评论(Technology Review)》杂志外,还经常在坎布里奇(Cambridge, MIT所在城市)和美国各大城市举行各种会议、班级团聚和研讨会等。

(杨雄风)