

生命科学的新时代

袁越

《三联生活周刊》采编中心,北京 100125

上世纪 80 年代初期,生命科学和计算机科学是最热门的两个学科,吸引了大批科学爱好者投身其中。30 年后的今天,当初齐头并进的两匹骏马已经分出了高下。从 286 开始的个人计算机时代,诞生了一大批响亮的名字:微软、英特尔、苹果、IBM、互联网、谷歌、脸谱、新浪、微博……每一个名字都可以骄傲地自称:我改变了人类的生活方式。反观生命科学,成果却乏善可陈,癌症、艾滋病、帕金森氏症……很多疑难杂症都还没有找到根治的办法。

计算机科学以“摩尔定律”般的速度,将生命科学远远抛在了后面。

要说生命科学没有进步,那也是不公平的。生物学家们控制了绝大部分传染病,降低了心血管疾病的发病率,运用基因工程的方法极大地降低了许多生物制品的成本……更重要的是,我们绘制出了整个人类基因组密码。

最后这项成果,用了 11 年的时间,并于 11 年前宣告成功。那么,11 年后的今天,这项成果带给我们什么实际的好处呢?人类基因组计划的完成对生物学的未来产生了怎样的影响?本文将告诉你答案。

1 人类基因组计划的过去、现在和未来

“11 年前的今天,《科学》和《自然》这两本全世界最具分量的科学期刊几乎同时刊登封面文章,报道了人类基因组计划完成的消息。”美国伊利诺伊大学的威廉·贝克(William Beck)教授开门见山地说:“在我看来,这是与当年沃森和克里克发现 DNA 双螺旋结构具有同等价值的重大发现,甚至可以和达尔文发表《物种起源》相媲美。”

贝克教授是在美国科学促进会(AAAS)年会上说这番话的。AAAS 是全世界最大的非营利性科学组织,现有 265 个分支机构和超过 1000 万名成员,著名的科学专业期刊《科学》(Science)就是由 AAAS 负责出版发行的。AAAS 年会是全世界最大的科学盛会,人类基因组计划无疑是本次大会的热点之一。贝克教授负责组织了一个关于人类基因组的专题讲座,请来几位专家为听众介绍这个领域的现状,并展望未来。

1.1 11 年前的那场革命

让我们先从 1953 年开始讲起。詹姆斯·沃森(James Watson)和弗朗西斯·克里克(Francis Crick)在那一年构建了

DNA 分子的双螺旋结构,组成双螺旋的每条 DNA 单链由 4 种核苷酸按顺序排列而成,分别用 A、T、C、G 这 4 个字母来代表。每 3 个字母对应于 1 个氨基酸,而蛋白质是由 20 种氨基酸首尾相连而成的,也就是说,DNA 的字母顺序直接决定了蛋白质的氨基酸顺序,从而决定了蛋白质的化学性质,DNA 就是以这种间接的方式控制了生命体,并赋予每个生命体独特的个性。

人类基因组计划的构思始于上世纪 80 年代中期。经过短暂的酝酿,美国能源部和卫生部于 1990 年共同投资 30 亿美元成立了“国立人类基因组研究中心”,正式开始了测量人类基因组全序列的工作。有趣的是,该研究中心的前两位主任分别是沃森和克里克,足见这个机构的地位是多么的重要。

该中心原计划花 15 年时间完成测序工作,但到了 2000 年 6 月,当时的美国总统克林顿和英国首相布莱尔就迫不及待地召开联合新闻发布会,向全世界公布了人类基因组草图,比《自然》和《科学》杂志刊登经过同行评议的正规论文提前了 8 个月,可见此事的政治意义有多大。

当时亲自参与测序工作的埃里克·兰德(Eric Lander)教授去年在《自然》杂志上发表过一篇文章,回忆了当时的情况。据兰德教授透露,由于人类基因组存在大量的高度重复顺序,给测序工作带来了意想不到的困难,因此即使是 2001 年《自然》与《科学》杂志公布的数据也只涵盖了人类基因组的 90%左右,很多重复顺序并没有完全测出,并留有大约 25 万个缺口,已测出的那部分顺序错误率也很高。科学家们又经过了 3 年努力,直到 2004 年才又发布了一个更加精确的人类基因组顺序,涵盖了 99.7%的人类基因组,只剩下大约 300 个缺口没有填上,错误率也降到了十万分之一以下。

由此看来,美国能源部和卫生部当年的估计还是很正确的。事实上,1990 年时的 DNA 测序技术还非常原始,需要依靠技术人员亲手制作一块电泳凝胶,每跑一次电泳只能测几百个字母,而人类基因组一共有大约 30 亿个字母,要想把它全部测出来几乎是一件不可能完成的任务。90 年代中期,科学家发明了第二代 DNA 自动测序仪,不但大大提高了 DNA 测序速度,而且把成本也降了下来。到 2001 年人类基因组计划宣告完成时,这种机器的效率已经提高到每台每天测 100 万个字母的水平了。

肯定有很多人还记得 11 年前那场媒体的狂欢。当科学家宣告人类基因组计划终于完成后,全世界的报纸杂志纷纷发表文章,高度赞扬这一事件的伟大意义,并预言大部分疑难杂症都将被攻克,人类的寿命将大大延长,甚至还有人宣布人造超人即将诞生。但是,11 年后的今天,一提起人类基因组计划,几乎所有媒体都是一片哀号。不明真相的读者们肯定纳闷,科学家们当初夸下的海口,为何迟迟没有兑现?

事实上,科学家们并没有夸下海口,盲目乐观的都是那些不明真相的记者们。真正的科学家们都深知生命的复杂程度远超人类的想象,如果把基因组序列当做设计图,科学家只是刚刚看到了图纸的样子而已,还没有学会如何解读它。况且科学家们拿到的只是其中一幢房子的设计图纸,整个人类社区的每一幢房子,甚至每幢房子的每间房间的设计图纸都不一样,任何建筑师都不可能凭借一张通用的草图去修理每一个房间。

话虽如此,科学家们并没有被困难吓住。人类基因组学在这 11 年里取得了长足的进步,下面就让我们来简单梳理一下该领域已经取得的成就。

1.2 人类基因组的秘密

2001 年之前,科学家们只测出了果蝇、线虫、拟南芥,以及包括大肠杆菌在内的 38 种细菌的基因组,人类基因组的长度比它们加起来的总量还要大一个数量级。但是,自 2001 年到现在,科学家们已经测出了小鼠、大鼠、狗、牛、袋鼠、鸟、大猩猩和黑猩猩等一大批高等动物的基因组,甚至还测出了尼安德特人的基因组(生活在距今 20 万至 4 万年前),以及一位死于 5300 年前的“冰人奥兹”(Otzi the Iceman)的基因组。截止到去年为止,已有大约 3 万个人类基因组顺序,以及 4000 种细菌和病毒,250 种真核生物的基因组顺序被测了出来,这就为进化学家们进行横向和纵向的比较提供了充足的素材。

分析结果表明,哺乳动物和鱼类有着共同的祖先,这进一步证明人类祖先确实是从水里登岸的。人类和黑猩猩之间的基因差异出奇的小,而且科学家们已经找到了从猿到人的几个关键基因,比如语言基因和拇指基因(使得人类拇指和其他四指相对,便于使用工具)等,这说明人类的出现并不神秘,不需要上帝的帮助,只要改变几个基因就可以做到。

通过对不同人群基因组的比较,科学家们还修正了人类走出非洲后的迁徙路线。过去的人类学家相信人类离开非洲后走的是一条单方向的路线,通过不断分叉,逐渐遍布整个地球。但新的数据表明,人类在这一过程中经历了很多次回流和杂交,人类 90% 的基因变异发生在人群内部,只有 10% 发生在群间,人种之间的差别远没有过去想象的那么大。这一发现对于人类社会消除种族偏见、打击“优生学”等伪科学起到了决定性的作用。

更让人惊奇的是,最新研究显示,欧洲人和亚洲人很有可能和尼安德特人进行过杂交,并从对方身上获得了 1%—4% 不等的遗传物质,没有走出过非洲的黑人则没有进行过这

种杂交,这件事的文化意义尚待研究。

不久前刚刚完成的“冰人奥兹”基因组测序结果表明,这位死于 5300 年前的欧洲人尚未进化出乳糖酶,说明这个酶的出现发生在畜牧业进入欧洲之后。通过对人类基因组的分析,科学家已经掌握了这个酶在人类演进史上的全部进化过程,并从中推断出了畜牧业在各地的出现时间。与此类似的还有西藏人的耐高原基因等,也都可以帮助历史学家追踪人类进化的脚步。我们有理由相信,在不久的将来,科学家将通过人类基因组序列的分析,构建出每个基因,乃至每个家族的全部进化史,为人类撰写一部更新、更准确的家谱。

不过,大家之所以关注人类基因组计划,更多的原因在于人们非常急切地想要了解自己的身体,并在这一过程中提高医疗健康水平。虽然后者暂时没有看到太多有效的结果,但前者却已有很多让人惊奇的案例。比如,2001 年前科学家们一直高估了人类基因组中的基因数量,他们预测的数值从 3.5 万到 10 万不等。可是等到 2001 年后,科学家们惊讶地发现,人类基因组当中只有不到 2.1 万个基因,只占人类基因组总长度的 1.5% 左右,剩下的除了一些高度重复的“垃圾 DNA”外,都是负责调节基因功能的片段。换句话说,人类基因组中真正有用的 DNA 大部分不是用来编码蛋白质的(这是基因的传统定义),而是影响基因功能的调控因子,后者的作用一直被严重低估了,它们才是决定生命体特征的关键因素。事实上,通过和黑猩猩等哺乳动物的基因组进行对比,科学家发现,从猿到人这一关键步骤中,大部分进化过程发生在调控因子这部分,基因本身进化得很慢。

知道了上述事实,对于科学家们研究疾病和遗传的关系有着决定性的影响。科学家早就知道,只有极少数人类遗传病是由某一个单独的基因突变造成的,这类病严格遵从孟德尔遗传规律,因此被叫做“孟德尔病”。除此之外,绝大多数人类疾病都是由很多遗传因素合力造成的,它们中有很多都是在基因调控的层面出了问题。这里面的道理其实很好理解,单个突变造成的孟德尔病在进化上处于绝对劣势,很容易被自然选择所淘汰,但如果某个突变只有微小的影响,甚至在某种情况下是有利的,则很容易遗传下去。这就是为什么大多数看似相同的疾病其实有着不同的遗传因子,很难发明出某个包打天下的特效药,必须根据具体情况区别对待的原因,而这就是个人化治疗的理论基础。

这个思路说起来容易做起来难。2001 年前,科学家只找到了十几个与疾病有关的致病基因,如今这个数字已经达到了 1100 多个,对应于 165 种疾病,这其中绝大部分都是在 2007 年之后发现的,这要得益于总部位于美国加州圣地亚哥市的 Illumina 公司发明的第三代 DNA 测序法,此法一举将以前的 DNA 测序效率提高了 6—8 个数量级,这个提升速度超过了计算机领域的摩尔定律。

此项发明对于生物学研究的影响是极其深远的。以前人们一直拿计算机领域的飞速发展反衬生命科学领域的落后,如今这两个新兴行业第一次联手就打破了摩尔定律,从根本

上改变了生物学家获取数据的渠道和方式。如今任何一位研究基因的生物学家都必须学会如何利用电脑来分析庞大的DNA数据,生物学正在变得越来越像数学。以前那种提出假说,再设计实验证明假说的研究模式也许很快就要过时了,代之以大规模收集数据,从中发现生命规律的全新时代。

癌症的治疗,靠的就是这种新的思路。

2 四个癌症与一个联盟

2.1 癌症与人类基因组

“截至目前,现代医学在抗癌领域的表现是相当令人失望的。”来自加拿大卑诗省癌症研究中心(BC Cancer Agency)的约瑟夫·康纳斯(Joseph Connors)博士开门见山地说,“就拿医学水平相对很高的美国来说,1950到2000这50年里,每年死于心脏病的人占总人口比例从0.59%降到了0.21%,死于脑血管病的比例从0.18%降到了0.047%,死于流感和肺炎的比例从0.048%降到了0.020%,可死于癌症的比例只从0.19%降到了0.18%,可以说几乎没有变化。”

“即便是这0.01%的下降也是有水分的。”康纳斯博士继续说,“分类统计数据显示,死于癌症的美国人比例下降的主要原因是男性吸烟人数剧减,导致肺癌发病率的大幅下降。如果单看癌症治愈率,最近这几十年里几乎没有进步,绝大部分癌症一旦发生转移都无药可治,病人只能期待奇迹发生。”

康纳斯博士是在美国科学促进会(AAAS)年会上说这番话的。康纳斯博士是“癌症与人类基因组”专题讲座的主持人,据他介绍,去年全球新增癌症病人1300万,死亡750万,癌症排名发达国家死亡原因的第一位,在发展中国家虽然暂列第二,落后于传染病,但两者的差距正在迅速缩小。毫无疑问,癌症是公共卫生领域的头号敌人。

癌症的外部诱因千奇百怪,但本质是一样的,都是因为某些基因发生突变导致细胞生长失去控制,这就是为什么当2001年人类基因组计划宣告完成时,很多人认为癌症必将很快被攻克的原因。但是,11年过去了,这个愿望一直没有实现,这是为什么呢?

也许,我们应该换个角度,看看科学家们在这11年里到底都做了些什么。本次专题讲座的主讲人,来自著名的维康信托基金会桑格研究所(Wellcome Trust Sanger Institute)的迈克尔·斯特拉腾(Michael Stratton)教授认为,生物学界在这11年里所取得的重大成就之一就是基因测序的效率提升了好几个数量级。“人类基因组计划开始于1990年,最终花了11年才宣告完成,因为90年代基因测序用的是凝胶法,跑一次胶只能测几百个字母。”斯特拉腾教授说,“后来科学家发明了自动测序仪,到2001年这种机器每台每天就可以测量100万个字母了,而现在每台每天可以测1000亿个字母,其工作效率在这11年里提高了10万倍!”

众所周知,基因是由两条DNA长链按照双螺旋的方式组成的,每条长链由四种核苷酸按照一定顺序排列而成,分别

用A、T、C、G这四个字母来代表。人类基因组大约由30亿个这样的字母组成,如果每页印1000个字母,每本书有1000页的话,需要3000多本书才能将人类基因组序列印完。但如果将其储存在光盘中,那么一张CD就差不多了。事实上,这个对比很能说明问题,随着技术水平的飞速提高,过去曾经被认为不可能完成的任务,在将来很有可能变成家常便饭。

基因测序就是这方面的一个经典案例。

接下来,斯特拉腾教授用四个故事为大家介绍了新的基因测序技术在癌症研究领域的应用。随着基因测序效率的增加和成本的下降,科学家们对癌症的认识达到了前所未有的高度。

2.2 肾透明细胞癌的故事

肾透明细胞癌(Clear Cell Renal Cancer)是肾癌中最常见的一种,仅在美国每年就会新增3万个左右的新病例。这种癌细胞在显微镜下呈透明状,故此得名。

科学家们迫切想要回答的问题是:到底是哪(几)种基因突变导致了肾透明细胞癌?

在回答这个问题前,先来介绍一下致癌基因的知识。早在几十年前科学家们就知道,基因突变并不平等,而是有高低贵贱之分。具体说,有的基因突变能够直接导致细胞发生癌变,这种突变被叫做驱动突变(Driver Mutation);另外一些突变则要么对细胞无影响,要么直接导致细胞死亡,不会发生癌变,这样的突变叫做伴随突变(Passenger Mutation)。

在过去的30年里,通过对驱动突变的收集整理,科学家们发现了430个基因与癌症直接相关,如果这些基因在适当场合发生了“适当”的突变,就会让该细胞变成癌细胞。已知人类基因组中包含大约2.1万个基因,也就是说,大约有2%的人类基因可以被称为致癌基因。

再回到开头的问题:到底哪些突变会导致肾透明细胞癌呢?科学家们找来101个肾透明细胞癌患者,把每个人癌细胞中的2.1万个基因序列通通测了一遍,从中寻找驱动突变的踪迹。显然,这个办法在10年前几乎是不可想象的,但因为测序技术水平的飞速发展,这个看似很“笨”的办法摇身一变,成了最合理的手段。

结果显示,在这101名患者中,VHL突变出现了56次,PBRM1突变出现了44次,这两个驱动突变显然是肾透明细胞癌的主因。除此之外,还发现了11个驱动突变,但出现的次数较少,最多的只有7次,少的只出现了1次。虽然次数少,但它们的存在本身就说明肾透明细胞癌的发病机理可以是多种多样的。

从分布情况看,有将近一半的患者有不止一个驱动突变,最多的同时出现了三个!对于这样的患者,如果只服用针对某一种致癌基因的药物,就不管用了。

除此之外,另有22名肾透明细胞癌患者的癌细胞内没有发现任何一个已知的驱动突变,说明还有很多未知的致癌基因等待科学家们去发现。

有了这样一张驱动突变分布表,科学家就知道,应该先

把精力放在 VHL 和 PBRM1 这两个致癌基因上,找出对付它们的手段,这就等于成功了一半。更重要的是,这项研究再一次证明,同样一种癌症(肾透明细胞癌)可能有着完全不同的病因(致癌基因),以原产地为分类标准的癌症命名法(比如肺癌、肾癌等)已经过时了,未来的癌症一定是以基因突变的类型而划分的,癌症的治疗也一定是从基因着手,先测出患者究竟哪(几)个致癌基因发生了变异,再对症下药。

接下来一个很自然的问题就是:驱动突变究竟是怎么发生的?请看下一个故事。

2.3 恶性黑色素瘤的故事

恶性黑色素瘤(Malignant Melanoma)是皮肤或者其他器官的黑色素细胞发生癌变而产生的恶性肿瘤,仅在美国每年就有 5 万新增病例,约 7000 人死亡。已知有四类人容易得这种病,分别是痣多的人、常晒太阳的人、有家族遗传史的人以及肤色较浅的人。所以白人当中得这个病的较多,中国人则相对较少。

欧美国家对此病很重视,早在 30 多年前,就从一名 45 岁的白人男性患者身上取下一些癌细胞,在实验室条件下培养成功,使之成为研究恶性黑色素瘤的实验材料,这就是癌症实验室常用的 COLO-829 细胞系。虽然这名患者在取样一年后因化疗失败而去世,但他的细胞后代已经遍布全世界,继续为科研做贡献,这也算是另一种形式的不朽吧。

当科学家们终于有能力测出人类基因组全部顺序后,首先想到的就是对常用的癌细胞系进行测序研究,COLO-829 是最先完成基因全测序的细胞系之一。当科学家们把测序结果和正常细胞进行对比后发现,这株癌细胞的基因组全乱套了,一共有 66 个插入/缺失型突变,以及 37 个基因重组型突变。更引人注意的是,这株癌细胞的基因组内一共发生了 3.3345 万个单核苷酸突变,这种突变俗称点突变,就是基因序列中的一个字母被另一个字母所代替。点突变的发生频率较高,但大部分点突变都是中性的,对细胞的生活没有影响。话虽如此,这个细胞系居然有 3 万多个点突变,总数还是很惊人的。

那么,这些突变都是如何造成的呢?科学家们按照点突变发生的种类和位置进行了统计,发现其中有将近 2 万个突变是由 C 变成了 T。更有趣的是,这个位点的上游、紧挨着的那个核苷酸几乎全都是 T,也就是说,原基因组的 TC 被改成了 TT。

这个发现就好比是福尔摩斯在破译“跳舞小人密码”时找到的那个只有 5 个字母的单词,整个密码系统从此便迎刃而解了。原来,科学家早就知道,晒太阳之所以能导致基因变异,原因就在于太阳光中的紫外线能够让 DNA 长链中的 T 和 C 发生化学反应,生成一个新的化学键。这个额外的化学键改变了 DNA 长链的三维结构,在 DNA 双螺旋上形成一个鼓包。在细胞分裂时,DNA 合成酶遇到这个鼓包便过不去了,它的解决办法就是把后面那个 C 替换成 T,也就是说,TC 变为 TT 是紫外线诱导基因变异的一个标志性结果。

故事讲到这里,真相大白了。COLO-829 细胞系有超过一半的点突变都很可能是由于紫外线照射而引起的,说明这位男性患者生前一定非常爱晒日光浴,而这个习惯很可能最终导致了死亡。当然上面这个结论并不难得出,只要去查一下他的生活史就可以了。这个例子只是证明科学家有能力仅仅通过基因组测序分析,便从源头上搞清整个癌症的发病过程。如果将来再遇到比这个病例更难的情况,完全可以照此办理。

在这个故事里,基因测序就像是考古学家手里的那把小铲子,借助它可以考证出癌细胞的全部生活史。那么,科学家知道这个有什么用呢?请看下一个故事。

2.4 原发性胰腺导管癌的故事

我们常说的胰腺癌主要指的就是原发性胰腺导管癌(Primary Ductal Carcinoma of the Pancreas),这种癌和已故苹果公司总裁乔布斯得的那种胰腺癌不一样,非常厉害,恶性程度很高。

胰腺癌之所以厉害,一个原因是早期很难发现,另一个原因就是极易转移。民间有个说法,不怕得癌,就怕扩散,这话是有道理的。很多癌症本身是可治的,一旦发生转移,基本上就没治了。

斯特拉腾教授为我们展示了一张患病肝脏的照片,上面布满了一团团像菜花一样的白斑,它们都是转移至此的胰腺导管癌细胞。那么,这些癌细胞团是如何转移至此的呢?这个问题无外乎有两种答案:第一,它们是由一个个转移至此的癌细胞单独形成的;第二,先有一个癌细胞转移至此安营扎寨,再慢慢扩散开来。

要想回答这个问题,任何常规手段均无法做到,必须通过基因分析法。过去基因测序能力有限,科学家们只能分析点突变,结果仍然说不清。现在有了更为强大的测序方法,科学家们便选择其中 8 个面积较大的“菜花”,分别提取其中的癌细胞进行全基因组测序,终于搞清了胰腺导管癌的整体扩散路径。

分析表明,这位病人肝脏上面的这 8 个癌细胞团经历了 38 个比较重大的基因重组,也就是染色体被打断,然后再重新拼接起来。而这 38 个基因重组全都是一样的,说明前文所说的第二个假说是正确的,即先是有个胰腺导管癌细胞转移至肝脏,然后再扩散开来。

事实上,这位病人的胰腺导管癌细胞已经转移到肺部了,于是科学家又从病人的肺部取出一批癌细胞,进行了同样的分析,发现它们全都一样,但和肝脏上的“菜花”有少许不同,这个结果说明癌细胞先是转移到了肝脏,再从这里转移到了肺部。换句话说,用这种方法可以弄清癌细胞的全部转移路径。

这样刨根问底的分析,必须以强大的基因测序技术为后盾,否则是不可想象的。这个技术不但为我们描绘了癌细胞的转移路径,也为我们揭示了恶性肿瘤的成长史。科学家们从同一个癌症患者身上收集到若干个位于不同部位的癌

细胞进行基因分析,发现癌细胞不是一下子就“变坏”的,而是有个渐进的过程。先是一个基因发生突变,导致细胞生长失去了控制,然后在此基础上又有一个基因发生突变,导致细胞具备了扩散到肝脏的能力,再发生一个基因突变,使它又具备了扩散到肺部的能力……如此类推,最终这个癌细胞依靠一个个基因突变,逐渐扩散到了病人全身。

整个过程和生物进化非常相似,都是突变加上选择,最终导致更加适应环境的新物种的产生。我们都知道,进化的力量是无比强大的,这就是为什么癌细胞的生命力如此顽强的真正原因。

那么,我们该如何去面对这样一个凶恶的敌人呢?请看下一个故事。

2.5 威罗菲尼的故事

除了动手术外,对付癌症曾经只有两个办法:化疗和放射性疗法。两种疗法的基本思路都是一样的,简而言之就是“宁可错杀一千也不放过一个”。事实证明,这个办法效果有限,而且副作用太强,很多病人承受不了。

本文的前三个故事为我们描述了科学家是如何利用强大的基因测序工具研究癌症的,这个研究过程如今已经结出一颗硕果,这就是去年10月刚刚被美国FDA批准的新一代抗癌药威罗菲尼(Vemurafenib)。这个药和过去那种广谱的化疗药物完全不同,而是一种专门对付某一特定类型癌症的“窄谱”抗癌药。

此药的发明正是基因分析的结果。原来,科学家们经过10多年的研究,发现了一个强致癌基因BRAF,大约70%的恶性黑色素瘤、30%的甲状腺癌和10%的直肠癌是由于这个基因发生了突变而造成的。进一步研究显示,BRAF基因的正常功能就是将外来的生长信号传递到细胞内部,指导细胞开始分裂。一旦发生驱动突变,该基因便永久性地处于“开”的状态,于是这个细胞便一直以为自己收到了生长信号,开始了无休止的细胞分裂。

一旦搞清了机理,下一步就容易多了。科学家通过药物筛查,找到了一种化合物,能够把BRAF基因重新“关”上。临床试验表明,如果病人的恶性黑色素瘤确实是由BRAF突变所引起的,那么这种药的疗效非常显著,用药几周后大部分黑色素瘤就会自行消失。

显然,病人在用药前必须先测DNA,确定是BRAF突变才能用药。事实上,这就是最近炒得沸沸扬扬的“个人化治疗”的一个典型案例,也是人类基因组计划带给广大癌症患者最大的福音。可惜的是,威罗菲尼尚不完善,绝大部分病人体内的恶性黑色素瘤最终都会发生基因变异,对威罗菲尼产生抗性。要想彻底解决这个问题,也只有依靠基因测序技术。科学家们正在进行一项大规模研究,希望能找出癌细胞产生抗性的原因,然后改进威罗菲尼,阻止或者延缓抗性的出现。

这样的研究无论是难度还是规模都是史无前例的,这就要求科学家们必须团结起来,加强国际间的合作,于是就有了国际癌症基因组联盟。

2.6 国际癌症基因组联盟

基因测序技术的进步为癌症研究领域提供了超大量的数据,这些数据需要整合起来进行研究。于是,2008年4月29日,国际癌症基因组联盟(The International Cancer Genome Consortium,简称ICGC)在英国伦敦正式成立了。该组织计划投资10亿美元,用10年的时间把最常见的50种癌症的致病基因及其突变类型全部找出来,绘制成一张详尽的癌症变异基因全图谱。

目前该联盟一共有12名成员,分别负责一种或几种癌症的研究。中国也加入了该联盟,被分到的任务是研究胃癌。

据斯特拉腾教授介绍,该联盟定下的目标是在余下的7年时间里把目前已知的2.5万种癌细胞的基因组全部测出来。如果这一宏大目标最终得以实现的话,科学家就掌握了绝大部分癌症的生活史,这将极大地提升科学家对癌症的认识,为抗癌药物的研发打下基础。

癌症是一种复杂的疾病,攻克癌症没有捷径可走,只有从源头做起,脚踏实地一步一步来。让我们耐心等待。

3 从基因到生命

3.1 光有基因组是不够的

2000年6月,美国总统克林顿和英国首相布莱尔在白宫联合召开新闻发布会,克林顿称这项成果让人类“终于学会了上帝造人时使用的语言”,布莱尔则宣布“癌症即将成为过去”。2001年2月15日,《自然》和《科学》杂志同时宣布人类基因组计划完成后,很少有人注意到,2001年2月21日,美国《商业周刊》刊登了一幅漫画,标题叫做《寻找生命中真正重要的东西》。漫画中,一个圆胖子把代表DNA双螺旋的瘦高个一把推开,自己站在了聚光灯下。圆胖子下面标注着一个英文单词——Protein(蛋白质)。

“我对这幅漫画印象很深,它告诉我们,人类基因组计划的完成只是一个阶段性胜利,我们距离揭开生命奥秘还早着呢。”美国密歇根大学生物信息学教授、前美国科学促进会(AAAS)主席吉尔·奥曼(Gil Omenn)教授对大家说,“从基因组到疾病,中间还隔着很长的距离,到底谁能来填补这个空白呢?”

今年温哥华AAAS年会,奥曼教授负责主持一个关于人类基因组研究的讲座。主讲人克里斯·奥弗洛(Chris Overall)是一位来自加拿大英属哥伦比亚大学(University of British Columbia)的生物化学教授,他先用几个数据说明基因组为什么不能解释所有的问题:“人类一共有20244个基因,果蝇有15016个基因,线虫有20470个基因,光凭基因数量,很难想象这三种动物在进化树上的地位差得如此之远。”

众所周知,基因的功能是编码蛋白质,而蛋白质是所有生命活动的基础,无论是身体结构还是生命功能都是由蛋白质来实现的。人、果蝇和线虫无论是体积还是复杂程度都有着天壤之别,为什么在基因数量上没有体现出这种差异呢?答案就在于基因和蛋白质之间并不是一一对应的关系,它们

之间还隔着好几步呢。

从科学上讲,蛋白质的生产过程需要经历三个步骤,每一步都可以改变最终结果。首先,DNA被“转录”成相应顺序的RNA,后者是蛋白质合成的直接模板,这个过程可以发生剪切和重组等诸多变化。其次,RNA被“翻译”成蛋白质,这个过程也不简单,RNA和蛋白质之间并不是正比关系,尤其是丰度较低的蛋白质更是如此。第三,“翻译”后的蛋白质还要经过进一步“修饰”过程才能被派上用场,从某种程度上说,这个修饰过程是最重要的,存在着大量调控的空间。

换一种通俗说法,如果把基因看做建筑图纸,建筑商先要组建一系列包工队,分别负责建造墙壁、房梁和屋顶,再由装修工人负责做最后的润色。每个工人都有权对图纸做出修改,如果不了解修改的过程,仅凭那张原始建筑图是无法确定最后出来的结果到底是什么样的。

加拿大英属哥伦比亚大学的麦克·海顿(Michael Hayden)教授用一个案例说明光靠基因组完全无法解释生命的复杂性:“2008年《自然杂志遗传学分册》(Nature Genetics)上有篇文章,通过对1.6万人的基因组进行研究,找到了12个与身高有关的基因。但是经过仔细分析,科学家发现这12个基因加起来对一个人身高的影响也只有2%左右,也就是说,假设一个人幸运地拥有了全部这12个基因,他也只能比平均值高1英寸。”

正因为上述原因,在人类基因组计划完成之前,科学家就已经开始未雨绸缪,着手研究从基因到生命之间的那几个步骤了。他们的研究方法和基因组一样,都是穷尽法,即把所有可能性都找出来,再从中寻找规律。

3.2 从转录组学到蛋白质组学

既然DNA需要先转录成RNA才能起作用,那就干脆把所有的RNA顺序全部测出来好了,这个领域叫做转录组学(Transcriptomics),这个词是模仿基因组学(Genomics)而生造出来的。美国伊利诺伊大学的威廉·贝克(William Beck)教授在本次AAAS大会上向听众介绍了转录组学已经取得的成就,按照他的估计,人类一共有大约10万个转录子(蛋白质合成的最小单元),是基因数量的5倍。换句话说,平均每张建筑图纸都可以通过剪切和拼接等办法,从而变成5张不同的新图纸。

科学家在研究RNA的过程中发现,人体细胞中还存在着大量不直接生产蛋白质的siRNA,这种小分子单链RNA通常由20—25个字母组成,它们能够和相应的DNA或者RNA结合,从而影响基因的表达。这个发现曾经掀起过一阵风暴,引来很多大制药厂参与豪赌,试图用siRNA来生产新药。比如,2007年默克制药公司花了11亿美元将“斯纳医疗”(Sirma Therapeutics)公司买下,用于开发siRNA药物。同年,罗氏制药公司(Roche)和武田制药公司(Takeda)则花费4.31亿美元买下了阿尔尼拉姆(Alnylam)公司,以此方式进入这个领域。诺华制药公司也通过收购股权的方式和阿尔尼拉姆公司有合作。但是,由于siRNA分子极不稳定,很难将其运送到指

定地点,这次豪赌惨遭失败,到2010年时包括诺华、罗氏和辉瑞在内的许多大制药公司纷纷撤销了siRNA部门,基本上退出了这一领域,大笔前期投资也因此而打了水漂。

这件事从一个侧面说明,新药研发目前是一项风险极大的行当。

转录组学研究的是蛋白质生产过程的中间步骤,为什么不干脆跳过这一步,直接研究蛋白质组学(Proteomics)呢?事实上,科学家很早就明白这件事早晚非做不可,Proteomics这个词早在1997年就被造出来了。但是,蛋白质组学远比基因组学要困难得多,原因在于一个人的基因组在每个细胞里基本上都是一样的,可蛋白质的组成情况在每个器官、每个组织、每个细胞,甚至同一个细胞的不同时间段内都不一样,蛋白质图谱必须按照采样地点和时间的不同分别绘制,蛋白质研究者所面临的是一个无比艰巨的任务。

来自加拿大英属哥伦比亚大学的蛋白学家克里斯·欧福罗(Chris Overall)教授向听众解释了问题的复杂性:“蛋白质是由氨基酸首尾相连组成的,蛋白酶可以在这条氨基酸链上的任意位置切一刀,将其变成两个功能完全不同的新蛋白质。也可以从一端开始,依次移除若干个氨基酸,整个蛋白质的性质就会发生天翻地覆的变化。事实上,很多癌症就是这么来的。”

根据瑞士生物信息学院(Swiss Institute of Bioinformatics)的统计,人体内至少存在500万种不同的蛋白质,这就相当于每个新合成的蛋白质可以有50—100种不同的修饰方式。换句话说,每幢房子可以有50—100种不同的装修设计,最后出来的结果全都不一样。

来自加拿大多伦多大学的科学家安德鲁·艾米利(Andrew Emili)教授用一个例子向听众解释了蛋白质组学的存在意义:“人和黑猩猩之间的差别远大于小鼠和大鼠之间的差别,这说明人和黑猩猩蛋白质的差别比小鼠和大鼠蛋白质的差别要大。但是,如果只比较基因组,人和黑猩猩之间的差别比小鼠和大鼠之间的差别居然还要小上10倍,可见光靠基因组分析是远远不够的。”

“要想全面了解蛋白质的功能,仅仅绘制出蛋白质图谱也是远远不够的。”艾米利教授补充道,“蛋白质之间存在各种相互作用,信息就是通过这些相互作用在蛋白质之间传递的,所以蛋白质组学的研究范围比基因组学更广,除了要研究蛋白质的种类、数量、转换率(蛋白质合成和分解的速率之比)、运输路径、翻译后修饰等性质之外,还要研究蛋白质和蛋白质之间的关系、蛋白复合体的功能属性,甚至蛋白质和药物之间的化学反应特性。”

艾米利教授为听众展示了一张反映酵母细胞蛋白质互相作用情况的“联络图”,上头密密麻麻的点和线看得人眼花缭乱。

那么,如此复杂的蛋白质组学什么时候才能绘制完成呢?或者换个问法:人类什么时候才能体验到蛋白质组学带来的好处呢?答案出乎所有人的预料。

3.3 从预防医学到治未病

间皮瘤(Mesothelioma)是一种较为罕见的肺部恶性肿瘤,绝大部分间皮瘤是由于病人吸入石棉而引起的。这种癌症的潜伏期很长,那么有没有办法预测出哪一位石棉工人将会得这种癌呢?

科学家尚未搞清间皮瘤的发病机理,暂时还不能从分子水平上对间皮瘤做出可靠的预后。但是,美国科罗拉多大学的科学家拉里·格德(Larry Gold)博士决定运用蛋白质组学的思路,在未搞清发病机理的前提下实现这一目标。他找来 357 个间皮瘤样本,与接触了石棉但没有发病的样本做对比,分析两者在蛋白质层面的差别,希望从中发现规律。

这个对比不是用常见的抗体法做的,因为单克隆抗体的专一性还是不够高,不能满足这个实验的要求。格德教授发明了一种“低解离率修饰的配体”(Slow Off-rate Modified Aptamers,简称 SOMAmer),用来代替抗体的功能。简单说,这个 SOMAmer 就是一种寡核苷酸,本身带有 DNA 和蛋白质的双重性能,既可以和蛋白质发生特异性结合,又可以采用已经成型的 DNA 序列分析法,准确而又快速地做出鉴定。更可贵的是,SOMAmer 和蛋白质之间的结合特异性极高,每一个 SOMAmer 只和一种蛋白质相结合,目前尚未发现例外。

有了这个新工具,格德教授就可以实现自己的理想了。他对比了间皮瘤和石棉接触者的血清样本,分别绘制出了两种样本的蛋白质图谱,找出两者的区别,再用这个区别为标准,对陌生的血清样本做鉴定,准确率和特异性都达到了惊人的 98%。

格德教授还用类似的方法对非小细胞肺癌(Non-Small Cell Lung Cancer)、胰腺癌、肾癌、阿尔兹海默氏病、帕金森氏

病、心脏病和糖尿病等常见疾病做了对比,希望能找出一种诊断方法,能够提早预测出疾病的发生,这样做的好处是显而易见的,如果能进一步提高准确性和灵敏度的话,其商业前景一片光明。

格德教授成立了一家名为 SomaLogic 的生物技术公司,正在进行临床试验,争取早日获得 FDA 批准,将这项技术商业化。他的近期理想是运用这项技术预测疾病的发生,为个人化治疗(Personalized Medicine)出把力。他的远期理想是扩大应用范围,将这项技术用于个人化健康(Personalized Wellness)领域,真正做到“治未病”。

必须再次强调,这项技术的核心不是生物学,而是信息科学。格德教授不需要知道每一种病的发病机理,他只要通过这项技术获得海量的数据,然后用计算机分析这些数据,从中发现规律就可以了。也就是说,这项技术背后的思路和基因组学、转录组学完全一样,都是对传统生物学研究方法的一次革命。

4 结语

生命科学发展到今天,虽然取得了无数令人惊讶的成就,但在某些疾病的治疗方面应该说不尽如人意的,与同时起家的信息科学相比,差异更加明显。新一代生物学家决定另辟蹊径,与计算机科学联手,建立了生命信息科学(Bioinformatics)这门新学科,基因组学、转录组学和蛋白质组学属于生命信息科学旗下的三个分支。这个结合的意义极为深远,它彻底改变了生物学的研究思路和方法,让我们有理由期待下一场新技术革命的爆发。

(责任编辑 马骁骁)

·学术动态·

“第七届矩阵与算子学术研讨会”征文

由哈尔滨工程大学主办的“第七届矩阵与算子学术研讨会”拟于 2012 年 7 月 13 日在哈尔滨市召开。

征文范围:(一)矩阵和算子不变量保持问题;(二)矩阵与算子广义逆;(三)图谱理论;(四)符号矩阵;(五)矩阵与组合;(六)矩阵计算;(七)环上矩阵;(八)代数图论;(九)矩阵与算子在控制等工程中的应用;(十)数学教育。

全文截稿日期:2012 年 4 月 30 日。

联系电话:0451-82519384、15904606024。

电子信箱:linghuanzhang@hrbeu.edu.cn。

通信地址:哈尔滨工程大学理学院(150001)。

大会网站:<http://conference.hrbeu.edu.cn/conference/index.asp>。