科学杂志文章!

后基因组时代的思考：两种大科学的抉择

吴家睿

从科学研究的历史来看，形成一门学科并非一件容易的事。但在人类基因组计划实施的短短几年间，以××组学（-omics）构成的学科如雨后春笋般，迅速地在生命科学界蔓延。最早出现的是与DNA相关的“基因组学”（genomics），随后又产生了许多与各种生物大分子或小分子相关的“组学”，如蛋白质组学（proteomics）、转录组学（transcriptomics）、代谢组学（metabolomics）等。复合名词则更是不可胜数，以基因组学为例，在文献中就可以看到结构基因组学、功能基因组学、癌症基因组学、药物基因组学、毒理基因组学、环境基因组学和营养基因组学等。这些术语的出现，从积极的方面来看，表现了生命科学的活力和迅速发展的势头。从消极的方面来看，则暗示了一种浮躁和轻率。本文试图对后基因组时代出现的这诸多的“组学”进行一番梳理，并对这些新兴学科进行反思和讨论。

工程型与概念型大科学

人类基因组计划常被人们誉为生命科学的“登月计划”。这一比喻应该说是很恰当的，不仅说明这两者都是大科学，有大量人力物力的投入，而且表明两者都拥有一个清晰、具体的目的。对于前者而言，是测出人类基因组所含的32亿个碱基对；对于后者来说，则是让人类跨越38.4万千米的空间距离，登上月球。换句话说，这两个计划都属于科学工程。凡是工程都具有这样一个特点：目的明确，可进行评估和度量。比如要建造一幢楼房或架设一座桥梁，显然我们是有着明确的目的，而且可以对工程的实施进度和完成情况进行具体的和定量的评判。尽管“登月计划”和人类基因组测序工作要远比盖房子复杂和困难，但本质上都符合工程的范畴。根据这一标准，笔者把生命科学领域中研究目的可以被明确界定和度量的大科学，如测定物种基因组全序列的基因组学，称为“工程型大科学”。

生命科学领域中还存在另外一类大科学，例如“相互作用组学”（interactomics）、药物基因组学、环境基因组学等。它们与工程型大科学有着很大的区别，因为其研究目的不是明确可辨的，通常也难以对其进行具体的评估。这类大科学通常围绕着某种概念来进行研究，例如相互作用组学是以“相互作用”这一概念为主导，环境基因组学则以“环境”这一概念为核心。但是，在“相互作用”和“环境”指导下的研究内容是模糊的，研究边界也是变化的。此外，这类大科学不同于工程型大科学的另一特点是，研究永无止境，没有结束的客观依据或判定标准。人类基因组序列一旦测完，就可宣称人类基因组计划结束。但是根据什么来判断酵母相互作用组的研究工作完成与否呢？笔者把这类没有明确目的和判定尺度的大科学研究称为“概念型大科学”。

当然，对这两种类型的大科学的区别有时是很微妙的。美国国立癌症研究所在1997年发起了一个“癌基因组解剖学计划”（Cancer Genome Anatomy Project，CGAP），其目的是要收集和分析与癌症有关的遗传和基因组数据。这个计划内的两个子计划——哺乳动物基因收集（Mammalian Gene Collection）和癌细胞染色体畸变计划（Cancer Chromosome Aberration Project），则分别属于工程型和概念型大科学。因为前者有可以判据的目标——收集所有人和小鼠的基因表达产物（全长cDNA），后者却无法判定其目标的实现——收集所有癌细胞的染色体畸变类型。从这个意义上说，代谢组学或蛋白质组学都更接近概念型大科学，因为没有标尺测量它们的完成情况。

两 难 的 抉 择

迄今为止，在生物学的大科学研究领域，基因组学最为成功，从低等微生物到高等动植物中的许多物种的基因组都已被破译。基因组学的成功理所当然，因为它是典型的工程型大科学。此外，基因组学成功的另外一个原因是对技术的强烈依赖性。只要技术可行，目的就能达到。在1980年代初提出测定人类基因组的想法时，许多科学家都认为这是一个不可能实现的计划，因为当时的测序能力极低，一年不过数万个碱基。随着DNA自动测序技术的出现和发展，测序能力迅速提高，在1998年已达到年测序能力9000万碱基，2003年估计会达到每年5亿碱基。这种对技术的依赖也正是工程型大科学的一个主要特征。所以，如果要准备开展一个大科学项目，那么应该着眼于那些研究目的明确、技术方法可行的工程型大科学项目。

但是，生命科学领域的工程型大科学也有其先天不足。首先，这类大科学不是针对具体的生物学问题来进行的，也不能回答或解决具体的生物学问题。其研究的最终结果只是为生物学问题的研究准备一个数据库，提供一种进行大规模、高通量研究的基础。例如芽殖酵母（S. cerevisiae）基因组全序列的测定，一方面给出了所有基因的信息，另一方面让基因芯片分析、蛋白质相互作用组研究成为可能。如果这些数据没有被用于进一步的功能性研究，其价值将会大打折扣。

科学研究的标准之一是可重复性，不同的实验室得到同样的结果才是真实可信的。但是，在工程型大科学中，这

一标准就难以贯彻了。很少有人愿意把一个已经测完的基因组，再投入大量的人力和物力去重测一遍。虽然人们会制定一套标准来防止错误，如美国国立卫生研究院和能源部设立了测序合格的三个标准，但显然还会有不少错误的信息存在于数据库中。不久前，美国人类基因组计划的专家分析了国际人类基因组计划（HGP）公布的人类基因组序列，以及美国塞莱拉（Celera）公司采用鸟枪法测定的人类基因组序列，认为塞莱拉公司并没有做什么事，只是把公共数据库里的数据重新拼装而已。塞莱拉公司的专家则否认这一指控。这一案例表明，即使对同一个基因组分别测序，两个数据库的差异也是不容易说清楚的。蛋白质组数据的问题更为严重，因为实验条件的微小差别，都可以导致不同的蛋白质表达谱。国际蛋白质组织在2002年4月，专门成立了旨在建立统一标准的蛋白质组学标准计划（Proteomics Standards Initiative）。工程型大科学的这种不可重复性对研究者和科研管理者都是一个挑战：怎样评判这类生物学大工程的质量？

回过头来看生命科学领域的概念型大科学，它们显然有着诱人的另外一面：这类研究的内容或目的通常是与生物学现象或问题紧密相关的。例如，癌细胞染色体畸变计划的实施，有助于了解癌变机理和对肿瘤的诊断。此外，这类大科学与常规实验室研究有着紧密的联系，各种小型实验室的研究力量都能够加入到这种类型的研究工作中。而工程型大科学常常局限于一些大型的研究实体，如在美国，公共的测序工作主要是由三个测序中心承担。目前，概念型大科学种类和项目要远远多于工程型大科学，原因在于公众和政府更愿意把钱投到有实际意义的研究中，科学家更容易参与到与生物学问题和现象相关的研究中。

对于概念型大科学而言，研究的核心理念必然涉及到概念和假设。如果起始的概念和假设是正确的，那么研究工作就是有意义的。反之，研究工作的价值就很成问题。而工程型大科学就不存在这种潜在的危险。美国国立卫生研究院在2002年10月宣布，启动一项被称为“单型作图”（Haplotype Map）的计划，在3年时间内投入1亿美金，构建人类基因组的单型图谱。“单型”（haplotype）是一个从基因组研究中形成的新概念：基因组的DNA序列并不是随机的排列在一起，而是由被称为“单型”的基本结构单元所组成。启动“单型作图”计划的假设是，“单型”在不同人种是不一样的，而且单型与疾病有着密切的关系。但是，有许多科学家反对这一计划，认为这个概念和假设都尚未被证实，很有可能不是这么回事。由此可见，从事概念型大科学的风险要远大于工程型大科学。

借用一下数学术语，工程型大科学是收敛的，有一个终点；而概念型大科学是发散的，难以界定其最终的研究目的，因此概念型大科学要取得工程型大科学那样的成功常常是很困难的。美国在1970年代初曾掀起过一场攻克癌症的战争。当时是由总统挂帅，国会立法，实施“国家癌症计划”。然而，30多年过去了，尽管投入了远远超过人类基因组计划的人力和物力，却没有取得人们所预期的成果，因为癌症的复杂性远远超过了人们在计划启动时对癌症的理解。今天，在后基因组时代出现的这些形形色色的概念型大科学，究竟有多少成果可以收获还是很难估计的。

从上述讨论中可以看到，生命科学领域中的这两类大科学有某种互补性，工程型大科学的短处正好是概念性大科学的长处，反之亦然。古人曾说过，鱼与熊掌不可兼而得之。对于这两类大科学来说，是否也是只能选取其中之一种，还是有某种方式可以同时兼顾？以笔者看来，系统生物学是一种能够把这两类不同的大科学进行整合的途径。工程型大科学实际上就是所谓“发现的科学”，以构造数据库为主要任务；概念型大科学本质上属于“假设驱动的科学”，以研究生物学问题为主线。而系统生物学的特点，正是整合“发现的科学”和“假设驱动的科学”（系统生物学的详细介绍见本刊2002年第6期第22页）。