金桥会议传真

第 2 期
（总第 2 期）

主办单位：上海产业技术研究院
上海浦东金桥技术开发区管委会
上海市中国工程院院士咨询与学术活动中心　　2013年8月25日

编者按:在金桥会议“大数据产业技术和商业模式”主题分会上，上海产业技术研究院李亦学研究员作了题为“生物医学大数据：挑战和机遇”演讲，经汇总整理形成如下咨询建议，供参阅。

大数据与生物医学

一、行业发展态势

在大数据的时代，生物医学领域将会是最为重要的一个大数据应用行业。生物医学领域最大的变革是我们即将进入个性化或精准医疗时代，而支撑个性化医疗和个性化用药的基础正是生物医学领域的大数据。每个人的基因组大小为 30 亿个碱基，个体之间的遗传差异信息为 300 万个碱基位点。不考虑其它分子水平的信息，仅基因组水平的个体化数据就是上百万到上亿字符的信息。如此众多的个体化分子水平差异数据，让我们可以对每个个体或每类疾病表型进行精确分型，从而实现全方位的分子检测和个性化医疗。特别值得注意的是，这些信息不仅仅停留在基础研究阶段，有相当一部分已经进入临床应用。生物医学数据的爆发，对信息管理提出严峻的技术挑战，同时也意味着巨大的商业机会。

由于分子检测技术的突破，生物医学领域的数据增长速度已经突破了计算性能发展速度的摩尔定律。其中，个体遗传信息的增长速度差不多是 4 个月翻倍，而计算能力则是 18 个月翻倍。此外，生物医学文献的增长速度也非常快，这些信息的积累为分子表型数据的临床转化奠定了理论基础。海量数据的飞速增长意味着计算机存储、运算和生物医学数据产出之间存在很大的缺口。当前互联网状

态下，生物医学数据传输，存储和管理等环节都存在一定程度的技术障碍，迫切需要信息技术的突破，以及符合生物医学领域行业特点的数据压缩、传输、存储和管理技术的研发。

分子检测技术和大数据相关技术的发展使得基于生物医学大数据的个性化医疗成为现实。以测序仪为代表的分子检测技术，已经成为生物医药领域常规的通用的研究工具，正在逐步进入医院，成为临床检测的常规技术手段。测序仪技术的发展可以看作一项革命性技术，是生物医学大数据行业的基石。测序仪的革命性，体现在价格更便宜，测试时间更短，通量更高。因为新一代测序技术的发展，让全基因组测序从首个个人基因组的 30 亿美元，20 年时间投入，变成一千美金，一天就能完成。除了遗传水平的数据，分子表型的数据还可以包括蛋白质组，代谢组方面的数据。这样，我们可以从分子水平全方位的描述一个个体的状态。再考虑生物医学数据的异构性质，从分子、细胞、组织与系统到行为疾病，从遗传学，生理学到成像与临床试验，高度复杂的数据，能够同步记录 1000 个细胞甚至更多。与此同时，可以采集到不同层次高度动态的数据。因此，必然导致数据的爆炸性增长，生物医学数据的规模从 GB，TB 级别逐步增长到 PB，EB 级别。

如何从生物医学大数据变成临床应用需要一个艰苦的过程，也是一个当前最大的挑战。生物医学数据增长非常快，而从数据到信息到指导临床功效则进展非常缓慢。美国在 1972 年提出消灭肿瘤计划，但 1982 年就宣布这个计划失败，后来又提出通过人类基因组计划解决肿瘤问题。基因组图谱解析已经超过 30 年，对晚期肿瘤病人治疗方法的改进只令存活时间延长了三个半月。为了加速生物医学数据到临床诊疗的转换，我们需要利用云计算技术，实现数据传输，分析，共享和分析；开发异构源数据整合和互操作技术；最后实现可视化，用直观的手段展示复杂的数据，便于人们理解。

为了推动生物医学大数据的管理，在科学和技术层面，应当搭建可互操作的数据库，支持生物信息学研究；开发生物信息学数据分析工具,为研究者提供不断升级的数据检索和分析工具；创建数据分析与管理工具的开发中心，服务于国家和产业；构建强有力的资源和基础设施，云计算等等。在社会层面，应当制定规范，推广教育培训，形成标准、词汇与知识本体。

中国科技部在过去的 10 多年来投入了大量资金在生物医学领域研究，产生了海量的数据，怎样动态高效地对这些数据进行二次挖掘，还有待推动。例如，科学院启动中国人群肝癌的个性化图谱群及分子分型模式项目，总投资额达到 10 亿。整个项目涉及上万个样本，将产生 4PB 数据，从基因组信息到蛋白质、代谢信息，基因组测序与数据验证、转录组测序到数据验证，最后目标是找到肝癌、糖尿病等疾病的生物标志物、治疗靶点，并进行预警分析。目前在公共或私人领域中，已经有超过350000个私人人类个体基因组被测序，产生了大约1000PB数据，这是难以同时解决的巨大数据量。中英临床试验样本银行也是一个正在进行的项目，将产生 PB 乃至 EB 级的 DNA 及相关数据，一旦测序成本降到每个个体 1000 美元，项目将会大规模地进行。在国际上，美国也启动了类似的项目，但其中关键的数据并不释放到公共资源中。因此，此类项目的实施对于国家战略有非常重要的意义。

我国已经成为世界领先的不断产生大量生物与生物医学数据的国家，然而，目前还没有一个类似 NCBI、DDBJ、EBI 的国家级生物数据库中心或联盟，以促进生物数据量数据挖掘，以及国内和国际共享。因此，我们目前只是一个数据资源大国，而不是强国。数据共享推进研究的一个典型案例是“腓骨肌萎缩症（CMT）”的研究。CMT 是一种遗传性神经系统疾病，患者最初会感到四肢无力，随后逐步恶化，最终可能终身依赖轮椅。它是单个基因缺陷引发的疾病，中国人群和欧美人群的致病基因不同。科学家们采用传统研究方式历经二十多年未能解析 CMT 的发病机制。采用数据驱动的知识发现策略，基于 12 人的家系样本，进行整个家族的全基因组测序，产生了 360GB 的人类基因组数据，通过比对非常直接地获得了致病基因。这个案例的意义在于，在大数据时代，我们可以通过简单的方式解决一些我们以前无法解决的棘手问题。

二、市场化运作的关键点

1、产业关键技术和环节

- 生物医学相关的云计算技术；
- 生物医学大数据相关的数据产出技术；
- 生物医学相关的应用开发和增值服务；
- 生物医学行业监管和行业标准

2、商业模式

目前大数据在医疗市场的商业化方向主要包括医疗设备和基因检测。

开发生产可穿戴设备。该设备能够检测睡眠质量、体重等各种生命体征的设备，看似简单，但背后有大数据理念的支撑——它可以把这些信息实时传送到网络上，一方面可以返回针对个人数据的分析结果以及保健指导，另一方面可以方便地实现个人与众多汇总数据的比较，更直观地了解自身的健康水平。以大数据概念的电子秤为例，在市场上可以卖到 99 美元的价格。这个领域存在非常大的市场空间。

提供基因检测服务。根据特定的疾病和某些基因之间的关系，进行疾病的诊断、风险预测，以及用药指导等等。乳腺癌相关基因 BRCA1/2 基因专利解禁后，基因检测的市场前景更加被看好。

3、值得投资的行业和需突破的方向

- 生物医学云计算支撑技术，重点是生物医学数据传输，安全和管理技术行业；
- 分子检测设备技术，重点是测序仪技术，便携式生物特征检测设备行业；
- 依据生物医学大数据开发的医学行业应用，重点是制药，个性化医疗行业；
- 生物医学行业政策制定和研究。

（汇总整理：李亦学　李园园）

（交流资料　仅供参考）

“金桥产业技术创新会议”秘书组
地址：上海科苑路 1278 号　邮编：201203 邮箱：jqcz@sast.org.cn