

中国生物信息学机构建设刍议

周 雁* 黄谷扬

(中国科学院基因组信息学中心/华大基因研究中心 北京 101300 复旦大学 上海 200433)

关键词 生物信息学, 研究机构, 建议

1 技术与生物学

在生物技术的后基因组时代, 开发新药是其灵山真经。人类基因组学首先和终极的目标就是在 21 世纪产生预防性的、预测性的、个体化的新药。

生物技术, 诸如生物信息、基因芯片以及蛋白质组的发展已在基因组学中起了重要作用, 并成为基因组学迅猛发展的推动力。然而, 此类初始性技术的开发费用昂贵, 获利甚难, 生物信息学亦仅是基因组学研究的一项初始性技术。进一步说, 技术总是从一个专业领域发展成为完整平台的一部分。分子生物学在生物技术领域的作用即是一例。因此, 技术开发需要一个不断创新的基础环境支持, 而世界上只有很少几个地方具有这样的环境。

中国并没有包括风险资金和创业上市板块的金融机制来支撑新兴技术, 我们也缺乏像硅谷那样充足的、训练有素的技术人才库和创新发明力量。最重要的是, 在中国没有一个足够大的生物制药研发市场来消化如生物信息或基因芯片等高新技术。

另一方面, 中国尚无一个从资金到员工充分的、为生物制药学的发展所需要的成长环境。因此, 中国的生物技术应该集中于生物研究及其制药应用, 而不是初始性技术的开发。根据这个原则, 我们不能也不应该纯粹为了技术来发展技术, 而是任何技术的发展, 包括生物信息都必须为生物制药的研究开发活动服务。

2 科研与产业

从 20 世纪 90 年代初生物信息学行业问世以

来, 还没有“纯粹”的生物信息学软件供应业者赢利过, 甚至连大型数据库提供者如 Incyte 和 Celera 也失去了优势。单纯的生物信息产业正在衰退, 其原因在于市场太小(全世界市场总额不足 1 亿美元), 技术和客户要求变化太快, 开发成本太高。在中国, 生物信息商业化的前景更不容乐观, 目前的国内市场割裂且不完整, 每年市场总额低于人民币 1 000 万元; 亦无现实机会将产品推到国际市场。

上述以生物学为中心的原则也要求中国的生物信息学必须向为生物学研究服务的方向发展, 包括数据分析、in silico 发现、数据模拟, 而不是商业产品开发。因此, 中国大多数的生物信息学中心应为由政府投资以及产业赞助的非赢利研究机构。这些中心在将来商机出现时可甩出或合资形成商业实体。为了维持这些生物信息学机构的中立, 保留他们在公共区域, 他们的软件及数据库就必须对学术机构及公共大众免费开放使用, 而其它服务可收取成本价。

只要这些中心是在公共领域, 互联网的性质会使任何“地方化”的中心孤立和落伍。为了提高知名度和使用度, 中心向全世界提供公共数据库和软件算法是划算的。

3 集中与分散

在此生物学服务导向原则下, 生物信息学的应用要求分散使用的软件和服务来推动生物学研究在不同地方的发展。而且分散化的生物信息学研究 and 开发环境将会比集中化的环境能更快地培育

* 杭州华大基因研究中心生物信息部主管, 复旦大学生命科学院遗传研究所博士生
收稿日期: 2002 年 1 月 21 日

生物信息学产业市场。

然而,生物信息系统的发展和相关人才的培养需要并受益于规模化,这在技术、资金及生物信息人才有限的中国显得尤为迫切。将要成立的一些生物信息中心应整合资源,统一人员培训、软件和算法的开发和引进,为学术界和产业界提供软件和服务。

基于生物信息学的性质,新整合的中心一开始并不需要完全集中于一地,当然资金和管理仍须协调好,这样会减少运行的最初成本。

4 数据库:一个,很多,还是不需?

几乎每个关于生物信息的提议都离不开数据库。然而一个数据库并非一堆不同数据的集合。一套数据是否需要数据库应以成本利益分析以及要求评估的结果为准来考虑。美国国立生物技术信息中心(NCBI)的 GenBank 序列仍存为无数据库的文本文件,而许多商业公司试图建立一个整合的生物数据库的努力尚无令人满意的结果。

数据库和数据存储系统(带应用软件)造价高昂,维护和更新费用更高。我们现在仍缺乏对不同种类数据库之间科学关系的完全理解(如遗传标记与基因开放阅读框的关系,或蛋白质模体和基因表达丰度之间的关系),这使得把所有的基因组数据塞入一个单一的数据库变得困难和不实际。新生物数据库的快速涌现使得大数据库内容和检索的更新越来越困难,数据的导入和导出也越来越缓慢。

除非绝对必要,我们不需建任何数据库。如果多个数据库的联合系统还能起作用,我们也不需建立整合的单一大数据库。大量的数据可以文本文件格式存储,而将索引存入数据库。生物学上关联甚少的不同种类的数据,则可分别储存于独立的数据库中,并通过连接数据表进行即时查询。

5 购买,自建还是利用?

大多数的学者,尤其是中国的学者,在生物信息学工作中有着一种全面“自家香”(非自创不可)的倾向,这种想法违背了软件开发的规则。就如我们已不再需要开发自己的 DNA 测序仪及测序试剂

盒一样,我们也不需要去开发大多数已商业化的供日常使用的生物信息学工具。

生物学家一个通病就是一般意识不到开发一个内外通用、随学科发展而更新变化的软件包需要花费多少的资金和时间。这些资金部分用于软件工程师的高成本(包括大流动量)和硬件频繁更新,生物信息学机构的耗资并不亚于生物技术机构。生物学家很少意识到内部开发往往要比购买市场上的现成产品更浪费财力。

根据前述的以生物学为主的原则、生物学的需要和成本效益的分析结果,我们可决定一个技术或软件包是需要内部开发、直接购买、还是在更改和整合之后采用。判断标准可遵循一个“70%法则”:若一个商业产品符合我们 70% 以上的需要就可购买;若只符合部分需要但仍颇有用,我们可以一个合理的价格来修改定制。我们应交由外部承包一切对核心任务不重要的东西,在内部只开发最重要的、专有的、能使我们的工作增值的软件。

6 使命和目标

我们的生物信息中心必须为中国的科研机构提供生物信息学服务和培训,促进基因组学和生物制药业在中国的发展,并向公共领域开放数据库和算法。我们的生物信息中心应成为:

- 开放非专有的软件、算法和数据库的资源中心;
- 提供软件开发、数据分析和咨询/合同服务的服务中心;
- 为大学及研究所和产业界授课、组织专训及演示产品的培训中心;
- 推进在遗传学、基因组学、发现药靶、in silico 药物开发等学科中的数据导向的研究的协调中心;
- 开发新的算法、软件和数据库,发表学术论文及建立知识产权的研究中心;
- 开发生物信息学产品和服务的企业孵化器。

生物信息学是一个从遗传学到蛋白质组学的宽广的技术领域。我们的生物信息中心必须集中于自己的核心能力和资源。我们认为,重点应在 in silico 药靶和药物的开发,依托中国的生物医学导向

的基因组学研究。

7 任务和项目(三年内)

(1) 建立在 Ensembl(<http://www.ensembl.org/>) 基础上的整合的、评注完全的基因组浏览器;

(2) 包括 SNP 基因型、连锁分析、相关性研究和疾病相关标记数据库的遗传分析软件包;

(3) 在 Snomed 和 ICD9/10 临床标准基础上的流行病学和病理学的分类目录数据库系统;

(4) 连接各种遗传学、基因组学、蛋白质组学和临床的数据库的地域性的数据库;

(5) 在 DNA 聚类分析基础上的基因表达和动态(主要是交替剪接和 SNP)的识别软件和数据库系统;

(6) 在 GO 和 KEGG 基础上的基因功能和蛋白质路径分类以及注解系统;

(7) 为发现、确认和优化药物靶分子用的基因分析流水线软件;

(8) 为寻找先导化合物用的蛋白质折叠、小分子接埠和虚拟筛选软件包;

(9) 包括 ADME 分析的药物动力学和毒性模拟系统;

(10) 通过客户端-服务器为研究人员提供在线基因分析;

(11) 由客座教师授课,每年为本科生和研究生开设不同层次的培训课程;

(12) 为地方生物信息学研究人员举办不定期、经常性的演讲和专训;

(13) 区域性、全国性和国际性的生物信息学研讨会和学术大会;

(14) 产品演示服务:为国内外的生物信息学产品和服务业者提供窗口;

(15) 为中国的生物信息学研究人员提供合作和咨询服务;

(16) 为中国的生物信息学研究人员提供深层数据分析服务;

(17) 为中国和亚洲的生物信息学界提供专有数据的数据库服务;

我们的软件开发原则是“越简单越好”。生物学和技术都在快速发展,因此,软件系统也必须适用性强、模块化、低成本。

8 挑战和要求

8.1 长期持续的资助

这是一个在中国科研机构中普遍存在的、影响发展的问題。获取启动资金相对容易,长期维持稳定的资助则很难。政府的科研经费不足以支持大型的研究中心,任何一家生物信息学中心的成功至少需要有三年的起始资助。

8.2 人员和管理

在生物信息学人员雇用上,质量向来比数量更重要。一个优秀的软件工程师抵过 10 个平庸的程序员。在中国生物信息学领域普遍存在另外一个失误,就是缺乏专业软件工程师,而初出茅庐、刚修了几门电脑课、并无软件开发经验的生物学毕业生过剩。虽然招聘有经验的软件专业人员比较困难和昂贵,但生物信息学中心必须愿为这个收益很快的项目投资。

中心内同时进行的多个项目使得建立由技术(数据库、应用程序、数据分析和系统管理)上的垂直管理和项目上的水平管理组成的矩形管理系统成为必要。这对于管理人员是很大的挑战,也是我们一定要聘用有经验的专业人员的另外一个原因。