

* 学科发展 *

生物信息学的发展

吴 旻*

(中国医学科学院肿瘤研究所
分子肿瘤学国家重点实验室 北京 100021)

摘要 文章论述了生物信息学的含义、目前取得的成就及探索前沿,就我国发展生物信息学提出了建议。

关键词 生物信息学,人类基因组

1 什么是生物信息学

生物信息学(Bioinformatics)是 80 年代末随着人类基因组计划(Human Genome Project, 简称 HGP)的启动而兴起的一门新的交叉学科,也常被称为基因组信息学。美国每年都要拨出相当大的经费支持生物信息学的发展。日本的经济近年来虽然不甚景气,但在发展科学技术方面,却是雄心勃勃,1998 年用于发展基因组生物信息学技术的经费比 1997 年增加了好几倍(533%)。

广义地说,生物信息学是用数理和信息科学的观点、理论和方法去研究生命现象、组织和分析呈指数增长的生物学数据的一门学科。首先是研究遗传物质的载体 DNA 及其编码的功能大分子蛋白质,以计算机为其主要工具,发展各种软件,对逐日增长的浩如烟海的 DNA 和蛋白质的序列和结构进行收集、整理、储存、发布、提取、加工、分析和研究,目的在于通过这样的分析逐步认识生命的起源、进化、遗传和发育的本质,破译隐藏在 DNA 序列中的遗传语言,揭示人体生理和病理过程的分子基础,为人类疾病的预测、诊断、预防和治疗提供最合理和有效的方法或途径。生物信息学已经成为生物医学、农学、遗传学、细胞生物学等学科发展的强大推动力量,也是药物设计、环境监测的重要组成部分。

2 生物信息学的成就和面临的挑战

在信息科学中,生物信息学还是一名幼童,既不成熟,也不完善,但成长迅速,充满活力,给科学家提供了无限机会。仅就爆炸性增长的 DNA 序列、蛋白质序列和结构的收集、整理、储存、提取、加工这些看似简单的工作,也面临数据增长过快,各种各样数据库的种类愈来愈多的

• 中国科学院院士
收稿日期:1998 年 3 月 15 日

困境,但这也提供了施展创造力的机会。坐在计算机面前敲动鼠标,就可以浏览和造访世界各地的数据库和站点,从英特网上卸下有用的工具,进行分析、对比,不用离开斗室就进入了一个无边无涯的大空间(虚拟实验室),把全世界的有关数据、图象、软件都调到你的计算机里,为你的研究工作提供了前几年连想都不敢想的可能性。然而,数据和站点每时、每日都在迅速增长。人们一天到晚、一年到头在计算机前,也未必能查遍所有的站点,搜集到全部有关的资料,因此,即使是搜寻和采集信息这个看似十分简单的课题,也仍是一个紧迫的研究课题。我们室的生物信息学家在最近8个月中发展了一个生物信息搜索和导航系统。自1997年11月13日上网后,立即受到国内外有关科学工作者的欢迎,就是这个道理。

重要的是要从大量不连贯的信息中发现其中隐藏着的重要信息。基因组信息学的首要任务之一就是发现新的基因和新的功能,比如,人基因组含有30亿对核苷酸,其中大约有10万个决定各种性状和功能的基因。这些基因的定位和分离是当前科学家、医生和企业家们最感兴趣的。连一个小耗子的肥胖基因都能卖上亿的美元。过去几十年中,科学家运用经典的遗传学分析方法,功能克隆、定位克隆等方法,总共定位了大约2000个基因。几年前,美、法、英、加、日等国的104位科学家,联合起来利用当时数据库中的45万个DNA小片段(称EST,表达序列标签)和其它有关信息,在很短时间内(1996)就把16354个人类基因进行了定位。如今库中的EST大约已增加到百万个左右。人类10万个基因中的绝大多数大概都有其EST储存于库中,估计不用太多时间就都能得到定位和分离了。

英国、美国、法国、德国和日本的大约600名科学家在90年代初联合起来,短短几年内即完成了长达12068000个碱基对的酿酒酵母的全基因组测序,并找出5885个编码蛋白质的基因,390个转录rRNA、snRNA和tRNA的基因,说明一个低等的真核生物有6000多个基因就能行使生命的一切主要功能了,而在这个计划开始前的数十年总共才有大约1000个基因通过遗传学分析被鉴定。如此迅速的进展除得益于洲际大协作外,更重要的是由于信息科学的发展和渗透,这种跨越大洋的成百上千人的大协作本身也只能在信息科学和计算机技术发展和普及的今天才有可能实现。

由于测序技术的进步,各国政府和公司资助的测序中心优先对重要病原体和工程菌的基因组进行测序。至今已经有许多种重要微生物的全基因组被测序,表1中列入7种供参考。

表1 7种重要微生物的全基因组测序

名称	基因组大小(Mb,百万碱基)	大致基因数
大肠杆菌	4.6	4288
结核杆菌	4.41	?
枯草杆菌	4.2	—4000
流感嗜血杆菌	1.8	1740
幽门螺杆菌	1.7	1590
梅毒螺旋体	1.1	1234
肺炎支原体	0.8	677

这些细菌基因组的全序列立即成为新的更精确有效的诊断、治疗手段和研制新药物的基础。如近年来卷土重来的结核病不仅危害第三世界的人民,发达国家也不能幸免,英国 Wellcome 药厂资助 Sanger 测序中心完成了结核杆菌的全基因组测序(1998 年初发布)后,立即着手研制基于 DNA 和蛋白质特征的新的诊断方法。

分子生物学家一直习惯于分离和分析一个一个的基因。比如,p53 基因在恶性肿瘤或 DNA 受到损伤时的改变和作用等等。由于分子生物学技术的进步和数学家、信息科学家和计算机科学家的参与,研究细胞全基因组在生理和各种病理过程中表达的动态变化的可能性已经出现。1997 年美国已拆巨资组织科学家们研究肺癌、乳腺癌、肠癌、卵巢癌和前列腺癌在癌变过程各阶段中细胞全基因组表达的变化。这标志着生物科学正迈向一个新的水平,对于人们认识和掌握生命的奥秘,用于提高健康水平,防治疾病和发展医药工业将产生不可估量的作用。

由于数据库中储存的处于不同进化阶段物种的 DNA 序列和蛋白质信息与日俱增,继结构基因组学和功能基因组学之后又出现了进化基因组学。科学家们对处于不同进化阶段物种的基因组结构和功能进行比较分析,企图最终弄清人类 10 万个基因的起源和进化,结构和功能的演变,发现其间的亲缘关系,像元素周期表那样把基因和蛋白质分类、排序,得到生物学的周期表,根据基因在进化树上的位置,或一小段核苷酸序列,或蛋白质的基序、模块、折叠等,即可预测其来源、结构、功能……。这项浩大的工程显然需要大量生物信息学家经数十年的不懈努力才能完成。

虽然自 HGP 实施以来,测序的效率不断提高,成本亦下降了一半。然而,时间已过半,人类基因组的测序却至今仅完成了大约 6 000 万个碱基对,即全部任务(30 亿碱基对)的 2%。最近美国能源部请了一位理论物理学家评估能源部和国立卫生研究院(NIH)的人类基因组计划,对美国的整个计划执行情况提出不少独到的见解,认为基因组学这个“大科学”比起核和粒子物理学、空间和行星科学、天文学和海洋学等比较成熟的大科学来说,还有不小的差距,对于基因组信息学进行了具体的分析,提出许多有益的建议。

3 建议

生物信息学的特点是投资少,见效快,效益大,适合于我国的现实条件。即从英特网上源源不断地采集数据,进行分析、归类与重组,发现新线索、新现象和新规律,用以指导实验工作的设计,这是一条既快又省的科研路线。可避免不必要的重复,少走弯路,提高我国生物科学的研究水平。关键在于有关学科之间的协作和加速培养一批在数学、物理、信息科学、计算机科学以及分子生物学方面均有造诣的跨学科青年人才。这样的人才在当前全世界都十分缺乏。我们如能充分发挥现有少数人才和单位的潜力,优势互补,协作起来,边做课题边培养研究生,进而在某些有条件的大学里设置生物信息学专业,就能迎接 21 世纪的挑战。

要加快发展我国的生物信息学,还有两件事是必须、也是能够做到的:(1)加宽英特网(internet)的网络频率宽度(bandwidth)。我国的国际通讯出口已经太窄,以致网络传送资料缓慢,无法满足当前的需要,更不用说今后。(2)国内对国际网络通信的收费太高,不仅为美国商业网络通讯费用的 10 倍以上,更远高于科研教育的网络通讯费用。虽然最近给院士们免费上网每月 30 个小时的优惠,但使用最多的还是青年科研人员,实验室往往承受不了上网费用。为

了发展科学、培养人才,建议降低收费(或科研人员可免费上网),不以营利为目的。

致谢 感谢张春霆、姚文萱阅读原稿和提出宝贵意见。

参考文献

- 1 Schuler GD *et al.* A gene map of the human genome. *Science*, 1996, (274): 540—546.
- 2 Goffeau A *et al.* Life with 6000 genes. *Science*, 1996, (274): 546—567.
- 3 Rowen L *et al.* Sequencing the human genome. *Science*, 1997, (278): 605—607.
- 4 Tatusov RL *et al.* A genomic perspective on protein families. *Science*, 1997, (278): 631—637.
- 5 Kunst F *et al.* The complete genome sequence of the Gram—positive bacterium *Bacillus subtilis*. *Nature*, 1997, (390): 249—256.
- 6 Koonin SE. An independent perspective on the Human Genome Project. *Science*, 1998, (279): 36—37.

———— * ————— * ————— * —————

* 简讯 *

中国科学家率先揭示 铍同位素示踪表土季节性迁移

本刊讯 在国家自然科学基金委的资助下,由中国科学院地球化学所环境地球化学国家重点实验室主任万国江研究员主持,以云贵湖泊沉积物中多种放射性元素示踪探讨多种微量元素在水体中的化学迁移变化,从而辨识和提取地球 200 年来环境质量演化的研究,经过 4 年的时间,已经取得重大进展;不仅取得了大量的科学数据,而且出版专著、发表论文 92 篇(部),同时还支持了 4 名博士后、11 名博士生、12 名硕士生的培养工作,这是一项高投入高产出的研究。

通过近年对云贵不同地区不同部位未受翻耕扰动的表土剖面采样的分析表明,铍同位素可用于表土季节性侵蚀的示踪。百花湖沉积物中铍同位素的累计值,揭示出流域内土粒可能搬运进入湖底沉积物中。还反映了宇宙线对大气层作用的季节性变化关系。

该工作在放射性核素和界面地球化学方面达到了国际先进水平。

(益鸣)