网络时代的生命科学和生物信息技术

罗静初

众所周知，20 世纪 90 年代诞生的国际互连网，以及随之而来的信息高速公路计划，意味着信息时代的到来。21 世纪是生命科学的世纪。信息技术和生物技术，将成为未来经济发展的支柱。集这两大技术于一身的生物信息技术，是当前技术革命的热点；正在逐渐形成的生物信息学，是一门崭新的交叉学科。

1999 年 4 月，国家自然科学基金委员会在北京九华山庄举办了首届“21 世纪核心科学论坛”，议题为“生命科学中的信息科学”。全国各高等院校和科研机构的 41 位专家参加了这次会议，他们分别来自生物、医学、环境、数学、物理、化学、信息、计算机等各个领域，其中包括 14 位院士。与会者“百花齐放”，各抒己见，对物质、能量、信息三者之间关系展开了热烈讨论。尽管“百家争鸣”，众说纷纭，对于“信息与物质、能量同为现实世界中三大基本要素”这一提法，似乎为大多数人接受。的确，在知识经济已经开始占主导地位的 21 世纪，以文字、声音、图形、图像、动画为载体的各种信息在政治决策、经济发展、国家安全、科学研究、教育事业乃至平民百姓的日常生活中正起着越来越大的作用。无论是信息的传输方式、传输速度，或者是信息利用所产生的效果，与“烽火狼烟”、“鸿毛传书”的年代，已经不可同日而语。信息有其固有的属性，既不同于物质，也不同于能量，但却与物质、能量一样，也是一种资源。从某种意义上讲，信息是一种取之不尽、用之不竭的重要资源。用得恰当，可以产生比物资、能量大得多的经济、社会效益。充分利用信息资源，不仅是经济发展的必然需要，也是科学研究的必要手段。计算机是现代化信息储存的主要载体，计算机网络是现代化信息传播的主要介质。2001 年 4 月，新世纪伊始，“网络时代的生命科学和生物技术”学术讨论会在北京召开，标志着计算机网络在我国生命科学研究和生物技术开发中的应用开始普及。几乎同时召开的“首届中国生物信息学大会”，是对我国生物信息学和生物信息技术研究开发队伍的检阅。

90 年代以来，生物信息技术飞速发展，主要得益于基因组计划的实施。几乎与国际互联网同时诞生的人类基因组计划，是迄今为止耗资最为巨大、竞争最为激烈、意义最为深远的大型国际合作研究项目。与人类基因组计划同时起步的模式生物基因组计划，包括小鼠、果蝇、线虫、拟南芥、酵母等，以及由此引发的其它一系列物种基因组测序，如水稻、玉米等农作物基因组，猪、牛、羊、鸡等畜禽基因组，河豚鱼、斑马鱼等鱼类基因组，有的已经完成，有的正在加速进行。据“基因组在线数据库”[GOLD] 2003 年 3 月公布的数据，已经基本完成基因组测序的物种有 132 个，正在进行的有 579 个。“基因组计划进展跟踪”网站 [GenomesMOT] 2002 年 3 月公布的数据表明，16 种古细菌、99 种真细菌、112 种嗜菌体、308 种细胞器、280 种质粒、40 种类病毒、873 种病毒的基因组序列测定已经完成。已完成的细菌和病毒基因组中包括脑炎、肺炎、流感、结核、梅毒、爱兹病等流行病的病原菌或病毒。

人类基因组计划的实施，使我们对于人类自身的研究，进入了一个新阶段。在探索遥远的月球、火星、太阳系乃至整个宇宙奥秘的同时，在掌握了能够毁灭人类自身几百次的核技术以后，人类必须冷静地、认真地考虑涉及自身生存最基本最普遍的问题：健康和环境。据全球最大的基因组测序和生物信息技术企业 Celera 公司称，世界上每分钟有 10 个儿童死于营养不良；每小时有 11 个美国人死于用药不当；艾滋病毒每天在一个病人身上扩增一亿倍。

我们知道，脱氧核糖核酸 DNA 和蛋白质是两种主要的生物大分子，前者是遗传密码的携带者，是信息的载体。后者是功能单位，是一切生命活动的基础。完成人类基因组 30 亿个碱基对的全序列测定，只是人类基因组计划的第一步。搞清楚人类基因组全套遗传密码的含义，则是人类基因组计划的最终目标。由此引发的结构基因组、功能基因组、药物基因组、转录组、蛋白组计划已经开始实施。以基因芯片技术为代表的一系列生物技术则是功能基因组、药物基因组研究和开发的基础。基因芯片技术正在日趋成熟，已经开始从实验室研究走向规模生产。基因芯片技术在疾病诊断、药物开发等方面有着巨大前景。随着基因组计划的实施，核酸和蛋白质序列数据迅速增长。由基因芯片应用产生的数据，更是难以估量。

面对如此浩瀚的数据海洋，传统的生物学研究模式已经不能适应，必须借助于生物信息技术和生物信息学手段。要给生物信息学下一个严格的定义，似乎并不容易。对目前流行的生物信息学和生物信息技术的粗浅理解大体如下。以核酸、蛋白质等生物大分子数据为主要对象，以数理科学、信息科学和计算机科学为主要手段，以计算机网络为主要研究环境，以计算机软件为主要研究工具，构建各种类型的专用、专门、专业数据库，研究开发面向生物学家的新一代计算机软件，对浩如烟海的原始数据进行存储、管理、注释、加工，使之成为具有明确生物意义的生物信息，并通过对生物信息的查询、搜索、比较、分析，从中获取基因编码、基因调控、核酸和蛋白质结构功能及其相互关系等理性知识。在大量信息和知识的基础上，探索生命起源、生物进化以及细胞、器官和个体的发生、发育、病变、衰亡等生命科学中重大问题，搞清它们的基本规律和时空联系，建立“生物学周期表”。

不言而喻，生物信息技术需要使用数理统计、模式识别、动态规划、密码解读、语意解析、信令传递、神经网络、遗传算法以及隐马氏模型等各种工具。生物信息学涉及到分子生物学、遗传学、细胞生物学、发育生物学、生物化学、生理学、免疫学、药物学、农业生物学、环境生物学等生命科学中的许多分支，同时必须有数学、物理学、化学、信息科学、计算机科学等自然和工程科学多个学科的参与，是一门新型交叉学科。它对 21 世纪生命科学乃至其它自然科学，具有不可估量的奠基和推动作用。生物信息技术所要处理的对象是数据，生物信息学所要研究的对象也是数据，即从大量数据中提取信息、从大量信息中获取知识、将大量知识转化为技术，将大量知识上升为理论。“数据－信息－知识－技术”，是生物信息技术开发的基本过程；“数据－信息－知识－理论”，是生物信息学研究的基本模式。

生物医学文献 PubMed 检索结果表明，“生物信息学”(bioinformatics)作为专用术语，开始出现于 1990 年[Benson 等，1990]。1993 年，仅有 3 篇文献；而 10 年来，有关生物信息学的文献数目急速增长（表 1）。

表 1　生物信息学相关文献增长表

年份	1993	1994	1995	1996	1997	1998	1999	2000	2001
文献数	3	9	10	74	82	291	350	696	963

*表中列出以“Bioinformatics”为关键词，对美国国家生物技术信息中心 NCBI 生物医学文献摘要数据库检索结果，检索时间为 2002 年 8 月 3 日。

尽管生物信息学这一专用术语出现于 90 年代，但生物信息技术的手段，特别是计算机在分子生物学中的应用可以追溯到 70 年代。利用计算机收集和存储生物大分子数据，是计算机在分子生物学中应用的一个重要组成部分。1977 年，美国布鲁克海文国家实验室开始利用计算机搜集蛋白质三维空间结构原子坐标数据，并以磁带作为介质定期向世界各地的用户发布[Bernstein 等, 1977]。1979 年，美国洛斯阿拉莫斯国家实验室 Goad 领导的小组开始

筹建核酸序列数据库 GenBank；欧洲分子生物学实验室也与 1981 年开始建立核酸序列数据库 EMBL。蛋白质序列数据库 PIR、SwissProt 也于 80 年代中期相继建立。

DNA 和蛋白质序列比对，是生物信息学最基本、最常用的算法之一。70 年代初出现的点阵法（Dot Matrix），可以说是序列比对算法的雏型［Gibbs 和 McIntyre, 1970］；几乎同时出现的序列全局比对动态规划算法，至今依然是序列比对算法的基础[Needleman 和 Wunsch, 1970]。80 年代初，序列局部比对算法出现[Smith and Waterman, 1981a, 1981b]。1988 年，Pearson 和 Lipman 发表了序列比对改进算法，奠定了数据库快速搜索的基础，并于 1990 年分别以 FASTA [Pearson, 1990]和 BLAST [Altschul]计算机程序实现，使数据库快速搜索成为可能。这两个程序一开始就免费提供使用，可以下载并安装在本地计算机上，是目前最为流行的生物信息学软件。特别是美国国家生物信息技术中心（National Center for Biotechnology Information，NCBI）和欧洲生物信息学研究所（European Bioinformatics Institute, EBI）提供的 BLAST 数据库搜索服务，成了分子生物学研究的常用工具，BLAST 一文的引用高达 10,000 多次。在蛋白质结构预测方面，二级结构的预测开始于 1978 年[Chou 和 Fasman, 1978；Garnier 等，1978]，而基于同源模板的三级结构预测甚至更早就有人开始尝试[Kretsinger, 1975]。

基于微型计算机的序列分析计算机软件于 80 年代初期开始使用，后来发展成 DNAStar、PCGene、MacVector 等商业软件。国内开发的 GoldKey 软件，也基于 PC 微机。80 年代中期，基于小型计算机系统的大型序列分析软件包 GCG 走向市场，后来成为 UNIX 平台下的主流产品。90 年代，计算机图形学和计算机硬件技术的发展，产生了以 SGI 公司的计算机图形工作站为基础的蛋白质三维空间结构显示和分子模拟、分子设计软件。数据共享和软件共享是国际生物信息研究开发的特点。除了上述数据库搜索软件 BLAST 和 FASTA 外，多序列比对软件 CLUSTALW、基因组序列装配软件 PHRED/PHRAP、基因识别软件 GenSCAN、系统树构建软件 Phylip 等，都是共享软件。特别需要一提的是 STADEN 和 EMBOSS 两个免费软件包。STADEN 由英国医学研究委员会资助开发，于 90 年代初开始使用，起初主要用于序列装配，现在已经包括许多 DNA 和蛋白质序列分析工具。EMBOSS 是欧洲分子学网络组织于 1999 年开始启动的国际合作项目，目前由英国基因组园区的 Bleasby 和 Rice 主持。EMBOSS 是目前内容最为丰富、功能最为齐全的生物信息软件包，并有便于生物学家使用的 Web 接口 JEMBOSS、PISE、wEMBOSS 均可免费下载。

因特网诞生以前，期刊杂志是信息交流的主要渠道。《计算机在生物学中应用》(Computer Application to Biosciences，CABIOS)创刊于 1985 年，主要刊登计算机在分子生物学中应用的文章，并于 1998 年改名为《生物信息学》 (Bioinformatics) ，是当前生物信息领域内主要杂志之一。 《计算生物学杂志》(Journal of Computational Biology)则侧重刊登有关算法方面的文章。1999 年，由欧洲分子生物学网络组织主办的《生物信息学简报》(Briefings in Bioinformatics)开始出版。2002 年 7 月，《生物信息学和计算生物学》(Journal of Bioinformatics and Computataional Biology, JBCB)创刊。除了上述专刊外，有关计算机在分子生物学中应用的文章，也发表在各种生物学杂志上，如《核酸研究》(Nucleic Acids Research)、《分子生物学杂志》(J Mol. Biol)等，其中以《核酸研究》杂志最为集中，1982、1984、1986 年第 1 期专集刊登分子生物学数据库以及 DNA 和蛋白质序列分析等方面文章。中断了 10 年后，1996 年起恢复了这一传统，每年第 1 期专门刊登有关文章，其中主要为数据库介绍（表 2），2003 年 7 月起，每年出版一期生物信息 Web 服务器专刊。

表 2　《核酸研究》杂志刊登生物信息学相关文献统计

年份	1982	1984	1986	1996	1997	1998	1999	2000	2001	2002
文献数	38	85	69	58	64	93	105	115	97	113

*表中列出《核酸研究》杂志刊登有关生物大分子数据库、序列分析等算法专刊的年份和文献数。

1983 年出版的“生物科学中的计算”(Geisow 和 Barrett，1983)是较早出现的有关生物计算的专著。1989 年，Fasman 主编的“蛋白质结构预测和蛋白质构像原理”(Fasman, 1989)较为详尽地介绍了蛋白质二级结构和三级结构预测的方法。90 年代中期以来，有关生物信息学的书籍不断涌现，至今已陆续出版了 50 多本。其中有些是很好的教科书，如 2002 年 Mount 编写的“生物信息学——序列和基因组分析”[Mount，2002]。国内最近 2 年也开始出现翻译、影印或自编的生物信息学书籍。

除了文献杂志和书刊外，学术会议也是信息交流的重要渠道。“国际分子生物学智能系统大会”（International Conference on Intelligent System for Molecular Biology，ISMB）是目前生物信息学领域内规模最大的国际会议，现由国际计算生物学学会（International Society for Computational Biology，ISCB）组织。1993 年以来每年举办一次，通常于夏天召开。除美国外，英国、希腊、德国、加拿大、丹麦先后组织过 ISMB，2003 年将在澳大利亚举行。“国际计算分子生物学研究年会”（Annual International Conference on Research in Computational Molecular Biology，RECOMB）始于 1997 年，是生物信息学领域内另一重要国际会议，一般于每年 4 月召开，已举行过 6 次，其中三次在美国，另外三次分别在日本、加拿大和法国。2003 年将在德国举行。“太平洋计算生物学讨论会”（Pacific Symposium on Biocomputing，PSB）则于每年 1 月初在夏威夷召开，2003 年为第 7 次。此外，德国、英国、日本等国也已经举办了多次国际性计算生物学或生物信息学年会。泰国于 2002 年主办了国际生物信息学大会。我国于 2001 年、2002 年在北京召开了第 1 届、第 2 届全国生物信息学大会。尽管起步时间落后于德国 10 多年，但与会人数远远超过德国生物信息学大会。除国内代表外，来自美国、英国、加拿大、法国、西班牙、香港、台湾的 10 多位代表参加了参加第 2 届年会。

综上所述，70 年代初到 80 年代底的 20 多年中，计算分子生物学或生物计算，从无到有，逐步发展，为生物信息技术和生物信息学的诞生孕育了必要条件。80 年代中后期，随着计算机技术特别是微型计算机技术日趋成熟和计算机网络的出现，计算机在分子生物学中的应用开始普及。美国国家生物技术信息中心 NCBI 于 1988 年成立，欧洲分子生物学网络组织(European Molecular Biology Network，EMBnet)也在同一年诞生，应该说不是偶然的巧合。1990 年开始的人类基因组计划和几乎同时诞生的国际互联网，则是生物信息技术和生物信息学诞生的强有力催化剂。基于互联网的 Web 技术，则更使这一新技术、新学科的发展如虎添翼。美国国家生物技术信息中心 NCBI、欧洲生物信息学研究所 EBI、瑞士蛋白质数据库专家系统（Expert for Protein Analysis System，ExPASy）等国际著名生物信息中心提供了大量生物信息资源。国内，北京大学生物信息中心(CBI)于 1996 年加入 EMBnet，成为该组织的中国节点。上海生命科学院生物信息中心(BioSino)也于 2000 年成立。天津、广州、西安等地许多大专院校和科研单位近年来建立了生物信息网站，而国内计算生物学研究则早在 80 年代末、90 年代初就已经开始。

不同领域、不同个人对于生物信息学的理解有所不同。目前比较一致的看法，主要是指

对基因组计划产生的大量分子生物学数据的研究，包括近几年出现的 DNA 芯片数据和蛋白组数据。若把生物学研究对象分为生态、种群、个体、器官、细胞、分子若干个水平，那么，目前生物信息技术和生物信息学的研究主要集中在分子水平上，确切地说，应该称“分子生物信息技术”或“分子生物信息学”。其研究范围十分广泛，大体包括以下方面：基因组序列装配、注释、分析，基因识别和基因组结构预测，基因转录表达调控，非编码序列特征和功能，疾病相关基因和其他功能基因的鉴定和分析，基于基因功能的药物靶标寻找和药物设计，RNA 二级结构预测，RNA 可变剪接和转录组分析，蛋白质序列特征分析，生物大分子相互作用和信号转导、代谢网络，蛋白质折叠机制和结构分类，蛋白质结构预测和分子设计，全基因组比较和系统发育、分子进化，等等。

生物信息技术开发和生物信息学研究，大体可以分为三个层次：第一层次是广大生物学家，主要是利用计算机作为工具，充分利用国际互联网上丰富的生物信息资源，进行实验设计、数据分析和结果处理。这里所说的生物学家，包括医学生物学、农业生物学、环境生物学、药物学等领域的研究开发人员。说得更确切一些，主要是指生物信息技术的应用。第二层次是熟悉计算机技术的生物学家或具有一定生物学基础的计算机或其他领域的研究开发人员，主要包括数据库构建和维护、软件整合和开发、数据采掘和知识发现，应该属于生物信息技术的应用和开发。第三个层次则主要是数学、物理学、统计学、信息科学等领域对生物感兴趣的研究人员，主要从事理论模型和新算法研究，可以认为主要是生物信息学研究。上述三个层次有着不可分割的关系，可形象地比作“一线、二线、三线”。处于“一线”的生物学家，从人数上说，是整个队伍的主体。他们熟悉自己多年研究的对象，从单个基因、蛋白的序列、结构、功能到基因家族，乃至整个基因组，了解相关的细胞学、遗传学、生物化学、生理学、药理学、组织学、免疫学等各方面的背景知识。而“二线”的研究开发恰恰可以为他们提供各种手段和工具。尽管“三线”研究人数相对较少，其重要性不可忽视。一个新模型或新算法的提出，往往可以产生革命性的变化。除了较好的数理功底外，生物学背景知识是从事“三线”研究的重要基础。

上述三个层次的研究都离不开计算机网络，无论是数据和软件的获取、文献的查询、结果的分析，都需要通过计算机网络。构建使用方便、操作简单、功能齐全的用户接口和内容丰富、更新及时、管理完善的数据库系统，对于第一层次的生物学家来说，将极大地提高工作效率。DNA 序列测定的发明者之一，诺贝尔奖获得者 Gilbert 在 1991 年发表的 Nature 短评“生物学研究模式的改变”中指出，“必须把我们的个人计算机接到全世界范围的计算机网络中去，以便随时跟踪日新月异的数据库资源，及时和同行进行学术交流”[Gilbert, 1991]。10 年后的今天，有关生物信息学的网站已遍布全球，以关键词“Bioinformatics”通过搜索引擎 Google 检索得到的相关网页数多达百万。然而，目前常用的生物信息软件，无论来自于学术单位的免费软件，或者价格昂贵的商业软件，都远远不能满足上述要求。仅就数据格式而言，同一个 DNA 或蛋白质序列，在不同数据库中有不同存储格式；用于不同程序时，又有不同输入输出格式；用户必须首先熟悉这些格式之间的转换。而用于序列分析的软件，多达几百种，对于不熟悉它们的用途和使用方法的用户，需要首先学会如何使用这些软件，如何分析程序运行所得结果，不仅费时费力，有时还经常出错。尽管已经有了许多基于 Web 的分析工具，但依然没有脱离以单独的计算方法为基础、以单个计算机程序为中心、以单一计算处理结果为目标的基本布局，有的以生物学家难以理解的单调的文字输出结果。

Nature 杂志于 2002 年 5 月刊登了美国冷泉港实验室 Stein 的文章 指出，只有建立基于 Web 的统一服务模式才能使生物数据充分得以应用。显而易见，利用软件工程、数据库管理、

网络技术等最新计算机和信息技术，开发基于 Web 的生物信息网上实验室(WebLab)，对基因图谱、核酸和蛋白质序列、蛋白质功能、基因表达、代谢途径等各种一次、二次数据库，以及数据库检索、数据库搜索、基因组序列装配、基因识别、引物设计、序列特性分析、亲缘关系分析、分子模型、结构预测、药物设计、药物筛选等各种软件进行系统整合，按生物学家熟悉的实验流程方式，建立一系列可视化、点击式、“傻瓜型”网上实验平台，通过因特网为生物学、医学、农学、环境科学、药物学等领域，以及医院、生物技术和制药公司等行业从事生命科学和生物技术研究、开发、应用的生物学家提供功能齐全、使用方便、数据丰富、结果可靠的分子生物学、基因组学、遗传学等常用实验设计、数据分析、结果处理的综合型、智能型、实用型生物信息分析、应用系统。只有这样，才能真正实现 EMBOSS 项目负责人 Bleasby 所说的“Half day on the Web，saves you half month in the lab”。

综上所述，网络时代生命科学研究的模式，在很大程度、很多方面已经不同于传统的生物学研究模式。高通量、大规模、整体性、系统性，是其突出的特点。生物信息技术和生物信息学，是这一新的研究模式的必备工具。数学、物理学等学科和生物学的结合，是生物学中新发现和技术创新的基础。150 年前，孟德尔仔细观察了豌豆的各种性状，通过精心设计的杂交实验，并利用经典的统计学方法，发现了遗传学基本定律。50 年前，沃森和克里克从他人的 DNA 分子 X-射线衍射实验结果中得到启发，经过计算，并利用模型和推理方法，提出了 DNA 分子的双螺旋模型。数学、物理学、信息科学、计算机科学等学科与生命科学结合，是生物学研究的新特点。尽管观察、假设、实验、模型、推理等依然是生物学研究的常用方法，面对浩瀚如海的数据，基于高性能计算机和计算机网络的数据处理和计算在生命科学研究和生物技术开发中所起的作用将越来越大。可以预言，未来年代中，实验生物学、理论生物学和计算生物学三者将在生物学研究中共同发挥应有的作用。如果 100 多年前有人说，“没有数学的科学是不完善的科学”；那么，今天我们有理由说，“不使用计算机的生物学家是落伍的生物学家”。

必须指出，以上讨论的生物信息技术和生物信息学，以及它在网络时代的生命科学中的作用和地位，只包括整个生命科学研究领域中的一小部分，还没有涉及到环境、生态等宏观生物学，更没有涉及行为、感觉、记忆、语言、思维、情绪等与高级神经系统特别是人脑功能有关的复杂生命过程。生物系统是一个极为复杂的系统，生命过程是一个极为复杂的过程。大自然以 30 亿个字母写成的人类基因组天书，需要我们去破译；困扰人类健康的病魔，需要我们去征服；日益恶化的自然环境，需要我们去改善；生命现象、生命起源、生物进化、生命本质等许多奥秘，还需要几代、几十代、几百代的科学家通力合作、共同努力，才能逐步揭开。尽管生物信息学还只是伊呀学语的婴儿，生物信息技术依然是蹒跚学步的幼童。但是，犹如一切新技术、新学科一样，其生命力不可估量。有志于生物信息技术和生物信息学的年轻朋友，让我们共同迎接挑战，开拓未来，与这一年轻的学科共同成长！

后记

本文 2002 年 7 月交稿，2003 年 3 月修定，2004 年 10 月收到科学出版社寄来的校样。近两年来，生命科学领域有了不少新进展，而生物信息学也有许多新的研究方向，如代谢和调控网络，蛋白组数据分析、基因和蛋白芯片数据处理等，本文均未涉及。应编辑部要求，所有文献引用全部删除，但文中依然保留作者和日期，读者可通过 PubMed 查到。文中“Half day on the Web, saves you half month in the lab”很难译成恰切的中文，只好原样保留。好在有兴趣阅读本文的读者，大概都有一定的英文基础，不难理解这一段话的含义。

网上发表此文的说明

2016 年 7 月 25-29 日，北京大学继续教育学院举办第二期香港青年才俊英才培训班，希望为参加培训班的 40 多名非生物专业学员介绍一下生物信息学这一新兴学科的产生、现状和前景。本文收录在科学出版社于 2005 年 1 月出版的《21 世纪 100 个交叉科学难题》一书中，对生物信息学的产生及其交叉学科的特点提出了一孔之见。本文完稿于 13 年前，而最近十多年来，生物信息学研究领域又有了长足的进展，文中有些资料已经过时，有些观点有失偏颇，恳请读者指正。

罗静初

2016 年 7 月 22 日