《走近生物技术》丛书　主编：来茂德　岑沛霖
国家自然科学基金科普基金资助项目

破译生命密码
——基因工程

周雪平　樊龙江　舒庆尧

浙江大學出版社

目　录

生物信息学

数据之海与一叶轻舟

我们处在一个激动人心的时代——基因组时代。科学的进步已使人类可以窥探生命的秘密，甚至包括人类自身。人类基因组在世纪之交被人类自己破译了，这部由30亿个字符组成的人类遗传密码本，已活生生地摆在我们面前。与此同时，来自其他生物的基因组信息源源不断地从自动测序仪中涌出，堆积如山，浩如烟海。这些海量的生物信息是用特殊的“遗传语言”——DNA的4个碱基字符(A、T、G和C)和蛋白质的20个氨基酸字符(A、R、N、D、C、Q、E、G、H、I、L、K、M、F、P、S、T、W、Y和V)——写成的。

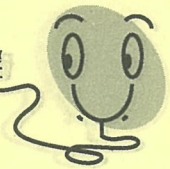
世界权威科学期刊之一《科学》(Science)在2001年2月16日的人类基因组专刊(NO.5507)上发表了一篇题为“生物信息学：努力在数据的海洋里畅游”的文章。文章写道：“我们身处急速上涨的数据海洋中……我们如何避免没顶之灾？”

一叶轻舟也许可以救命！生物信息学便是我们找到的这样一叶“轻舟”，而且我们已在这条轻舟上安装了诸如卫星定位系统等先进的电子设备。也许在不久的将来，人类会造就一艘永不沉没的航空母艇……

生物信息学是一门年青的学科，学科虽然年青，但它充满挑战、机遇，且引人入胜。

生物信息学亮相人间

近20年来，生物信息数量的剧烈膨胀，迅速形成了巨量的生物信息库。这里所指的生物信息包括多种数据类型，如分子序列(核酸和蛋白质)、蛋白质二维结构和三维结构数据和蛋白质疏水性数据，



等等。由实验获得的大量核酸序列和三维结构数据被存在核酸数据库中，这些数据库就是所谓的初级数据库(primary databases)；那些由原始数据分析而来的诸如二级结构、疏水位点和功能区 (domain)数据，则组成了所谓的二级数据库(secondary databases)。由核酸数据库序列翻译而来的蛋白质序列数据组成的蛋白质数据库，也应视为二级数据库。生物信息的增长是惊人的！近年来，核酸库的数据每 10 个月左右就要翻一番，到 2000 年底，达到了创记录的 100 亿个。同时，大量生物(甚至包括我们人类自身)的整个基因组序列被测定完成或正在进行中，遍布世界各地研究实验室的高通量大型测序仪在日夜不停地运转，每天都有成千上万数据在被源源不断地输入相应的生物信息库中；由这些原始数据分析加工而来的蛋白质结构等数据信息也被世界各地的分子生物学、生物信息学等学科领域专家输入二级数据库中。

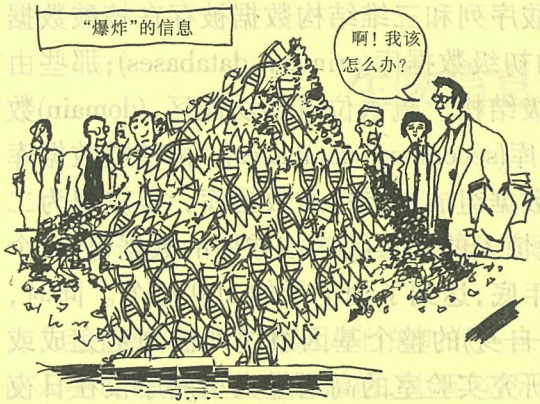
世界上主要的 DNA 和蛋白质序列数据库

数据库（Database）	网址（Address）
DNA	
EMBL	www.ebi.ac.uk/ebi_docs/embl_db/ebi/topembl.html
GenBank	www.ncbi.nlm.nih.gov/web/search/index.html
DDBJ	www.ddbj.nig.ac.jp
蛋白质	
SWISS-PROT	http://expasy.hcuge.ch/sprot/sprot-top.html
TREMBL	http://www.ebi.ac.uk/pub/databases/trembl
PIR	http://www-nbrf.georgetown.edu/pir/

迅速膨胀的生物信息给科学家们提出了一个新问题：如何有效管理、准确解读、充分使用这些信息？

于是，生物信息学在生物信息急剧膨胀的压力下诞生了！

一般意义上，生物信息学(Bioinformatics)是研究生物信息的采集、处理、存储、传播、分析和解释等方面的一门学科，它通过综合利用生物学、计算机科学和信息技术来揭示大量而复杂的生物数据所赋予的生物学奥秘。具体而言，生物信息学是把基因组 DNA 序列信息分析作为源头，在获得蛋白质编码区的信息后进行蛋白质空间结构模拟和预测，然后依据特定蛋白质的功能进行必要的药物设计的



生物信息学家们面对的是堆积如山的DNA片段

一门新的学科。基因组信息学、蛋白质空间结构模拟以及药物设计，成了生物信息学的3个重要组成内容。从生物信息学研究手段的具体内容上看，生物信息学应包括这3个主要部分：①新算法和统计学方法研究；②各类数据的分析和解释；③研制有效利用和管理数据的新工具。

生物信息学的诞生和发展最早可以追溯到上个世纪的60年代，鲍林 (Pauling) 分子进化理论的出现已预示着生物信息学的诞生，而

美国国家生物技术信息中心(NCBI)网站数据分析工具网页。图中包括BLAST、COG、ORF finder、Electronic PCR等工具软件。

“生物信息学”一词的出现，则是近10年的事情。虽然生物信息学的历史并不长，但正像生物信息的迅猛发展一样，生物信息学已发展了大量独具学科特色的分析方法和分析软件。例如，当获得了大量序列数据以后，我们现在已能进行序列家族或同源性分析；进行序列的聚类，建立进化树并确定序列间的进化关系；进行代谢途径相关基因的同源性分析，以及获取其他生物代谢途径的相关信息等。分析软件更是层出不穷，通过网络就可以搜索到大量的相关信息，这些软件很多已成为商业化产品，并且很多软件是可以免费获取的。这些分析软件已成为生物信息学最重要的研究手段，是生物学家获取信息的重要途径，也是显示生物信息学价值的窗口。

生物信息学发展简史

纵观生物信息学的发展历史，可将它分为3个主要阶段：①萌芽期(60—70年代)。以Dayhoff替换矩阵和Neelleman-Wunsch算法为代表，它们实际组成了生物信息学的一个最基本的内容和思路：序列比较。它们的出现，代表了生物信息学的诞生(虽然“生物信息学”一词很晚才出现)，以后的发展基本是在这两项内容上的不断改善。②形成期(80年代)。以分子数据库和BLAST等相似性搜索程序为代表。1982年，三大分子数据库的国际合作使数据共享成为可能，同时为了有效管理与日俱增的数据，以BLAST、FASTA等为代表的工具软件和相应的新算法大量被提出和研制，极大地改善了人类管理和利用分子数据的能力。在这一阶段，生物信息学作为一个新兴学科已经形成，并确立了自身学科的特征和地位。③高速发展期(90年代-至今)。以基因组测序与分析为代表。基因组计划，特别是人类基因组计划的实施，分子数据以亿计。基因组水平上的分析使生物信息学的优势得以充分表现，基因组信息学成为生物信息学中发展最快的学科前沿。Phred-Phrap-Consed系统软件包于1993年出现后仅两年中，已广泛应用于鸟枪法测序中序列的碱基识别、拼装和编辑等，成为目前人类基因组等测序计划的主要应用软件，与BLAST一起在人类基因组计划的研究历史上占有一席之地(见Science 2001年2月16日

人类基因组专刊“A history of Human Genome Project”一文)。在此阶段,生物信息学已成为举世瞩目、竞相发展的热点学科。

以下所列的是生物信息学最近几十年的主要事件*,可谓是生物信息学发展的简史,而这些事件大多是在“生物信息学”一词出现前发生的。

1962 Pauling 提出分子进化理论

1967 Dayhoff 构建蛋白质序列替换矩阵

1970 Needleman-Wunsch 算法被提出

1977 DNA 测序,Staden 利用计算机软件分析 DNA 序列

1981 Smith-Waterman 算法出现

1981 序列模序(motif)的概念被提出(Doolittle)

1982 GenBank 数据库(Release3)公开。三大核酸数据库(GenBank、EMBL 和 DDBJ)开始国际合作

1982 λ-噬菌体基因组被测序

1983 Wilbur 和 Lipman 提出序列数据库的搜索算法

1985 快速序列相似性搜索程序 FASTP/FASTN 面市

1988 美国家生物技术信息中心(NCBI)创立

1988 欧洲分子生物学网络 EMBnet 创立

1990 快速序列相似性搜索程序 BLAST 面市

1991 表达序列标签(EST)概念被提出,从此开创 EST 测序

1993 英国 Sanger 中心迁址英国 Hinxton

1994 欧洲生物信息学研究所在英国 Hinxton 成立

1995 第一个细菌基因组测序完成

1996 酶母基因组测序完成

1997 PSI-BLAST(BLAST 系列程序之一)面市

1998 PhilGreen 等人研制的自动测序组装系统 Phred-Phrap-Consed 系统正式发布

1998 线虫基因组测序完成

1999 果蝇基因组测序完成

2000 人类基因组测序基本完成

2001 人类基因组初步分析结果公布

* 以上资料主要引自美国家生物信息中心 (NCBI)Education-Bioinformatics Milestone (2000),原文截止至 1999 年果蝇基因组测序完成。有关人类基因组、PhilGreen 等的自动测序组装系统和三大核酸数据库的合并等内容,为作者所补充。

基因组时代的宠儿:生物信息学展望

蛋白质、DNA 和 RNA 序列的计算分析,在上世纪 80 年代末已发生了根本性变化。高效实验新技术,特别是测序技术是这一变化的推动力。这些新技术使实验数据急剧增长。随着基因组测序计划持续开展,用于序列分类、相似性搜索、DNA 序列编码区识别、分子结构与功能预测和进化过程的构建等方面的计算工具,已成为研究工作的重要组成部分。这些工具有助于我们了解生命本质和进化过程,同时对新药和新疗法的发现具有重要意义。生物信息学已成为介于生物学和计算机科学学科前沿的重要交叉学科,在许多方面影响着医学、生物技术和人类社会。

我们处在一个基因组时代。许多新技术,如用于大规模测序的毛细管电泳、基因芯片技术等,已应用于基因组研究,使我们能达到在以前达不到的尺度和角度观察生物学现象——某一基因组的所有基因、某一个细胞中的所有转录产物、某一组织中的所有代谢过程,等等。这些新技术的一个共同特点是产生大量的数据。如何管理这些数据、解读它们,并使各领域的生物学家们能容易地使用它们,是生物信息学面临的巨大挑战。

生物信息学也面临着越来越多的困难。对初学者而言,很少有人能在计算机科学和生物学研究两方面同时拥有扎实的背景;同时,对对方研究问题的无知可能导致误解。例如,编写用于拼接 EST 重叠群的程序对于生物学者来说是非常重要的,但对于计算机科学家来说,这没有任何新意;同样,证明在一定条件下不可能构建一个整体最佳系统发育树(phylogenetic tree)可能是计算机科学的一个重要命题,但对于生物学家来说却无实践意义。如何找到共同感兴趣的问题,是生物信息学的重要目标。然而,所谓"真正"的生物学研究,已越来越多地在计算机前完成;同时,越来越多的计算机科学的课题将来自生物学问题。

生物信息学需要有多方面的数据资源,这无疑又增加了生物信息学面临的困难。没有这些数据资源和以新方式组合这些数据的能

力,生物信息学学科领域范围将受到极大限制。例如,基因相似性搜索程序 BLAST 的广泛应用除了得益于它的算法外,还得益于那些公共数据库,如 GenBank、EMBL 和 DDBJ。没有这些数据库供查询,BLAST 的作用将受到限制。

生物信息学研究的一个核心问题是数据库的开发,以整合和最有效地查询来自诸如基因组 DNA 序列、mRNA 表达的空间和时间模式(spatial and temporal pattern)、蛋白质结构、免疫反应和文献记录等数据。

要做到这一点,数据的共享性和应用性非常重要,这引起人们对数据释放(公开)政策的关注。认识到数据尽早释放对许多研究具有重要意义。人类基因组计划(Human Genome Project,HGP)采用的是一种数据正式公布前即上网释放的政策,许多其他基因组计划目前也采用了相同的做法。由于生物信息学强烈依赖于各种来源的数据资源,所以希望一些基因组水平的研究计划(如表达分析和蛋白质组学研究)也能采取相同的政策。生物信息学研究面对的另一个问题并不是对数据使用的限制而是对其下游研究的限制,如将一些数据并入新的或已有的数据库中。这一问题对于生物信息学研究更为关键,因为这不仅涉及何时可以进行生物信息学分析,而且涉及到可进行何种分析。塞莱拉(Celera)公司最近公布的人类基因组初步分析结果便集中引发了这一问题。该公司测得的原始数据(即初级数据)仅由这家私人公司释放,并对这些数据的进一步存储和加工设定了限制。不妨想像一下,基因组学研究处于这样一种境地——公共数据库(GenBank/EMBL/DDBJ)没有相应数据,由于所有权的限制使数据拼接无法进行,这该是多么令人沮丧!因此,有必要为基因组序列的释放建立一个很好的标准。在后基因组时代(postgenomic era),人们期待在对生物发育机理、代谢过程和疾病的认识方面有所突破。可以肯定地预言,生物信息学研究将根本性改变我们的一些认识,如在基因表达调控、蛋白质结构预测、比较进化学和药物开发等领域。只有在数据共享的情况下,基因组水平的研究才有可能进行;而捆住手脚,要在数据的海洋中畅游是很困难的。

在我国,生物信息学随着人类基因组研究的展开虽刚刚起步,但

已显露出蓬勃发展的势头，许多大学和科研院所已经开始或准备从事这方面的研究工作。北京大学的生物信息中心已建立了 EMBL 的镜像(mirror)数据库(http://www.ipc.pku.edu.cn/mirror/mirror.html)，并已能提供部分检索服务；中国科学院上海生命科学院建设了一个大型生物信息服务器(www.biosino.org.cn)……我国已召开了第一届生物信息学学术大会，而第二届基因组学术会议也即将召开，这些都预示着我国生物信息学研究的兴起。

（樊龙江）