

# 第一章 生物信息学通论

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了我们面前。于此同时，来自其它生物的基因组信息源源不断从自动测序仪中涌出，堆集如山，浩如烟海。这些海量的生物信息是用特殊的“遗传语言”——DNA 的四个碱基字符(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。

《科学》(*Science*) 在 2001 年 2 月 16 日人类基因组专刊上配发了一篇题为“生物信息学：努力在数据的海洋里畅游”(Roos DS. Bioinformatics—Trying to swim in a sea of data. *Science*, 2001, 291: 1260-1261)的文章。文章写道：“我们身处急速上涨的数据海洋中...，我们如何避免生物信息的没顶之灾呢？”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条“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舰.....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。

## 第一节 生物信息与生物信息学

### 一、迅速膨胀的生物信息

近 20 年来，分子生物学发展的一个显著特点是生物信息的剧烈膨胀，且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)，蛋白质二级结构和三维结构数据、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)。那些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，也应被视为二级数据库。

生物信息的增长是惊人的。近年来，核酸库的数据每 10 个月左右就要翻一翻，2000 年底，数据库数据则达到了创记录的 100 亿个记录，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，每天都有成千上万的数据被源源不断地输入相应的生物信息库中。同时，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。图 1.1 显示出了各种生物信息的同步增长状况。

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？

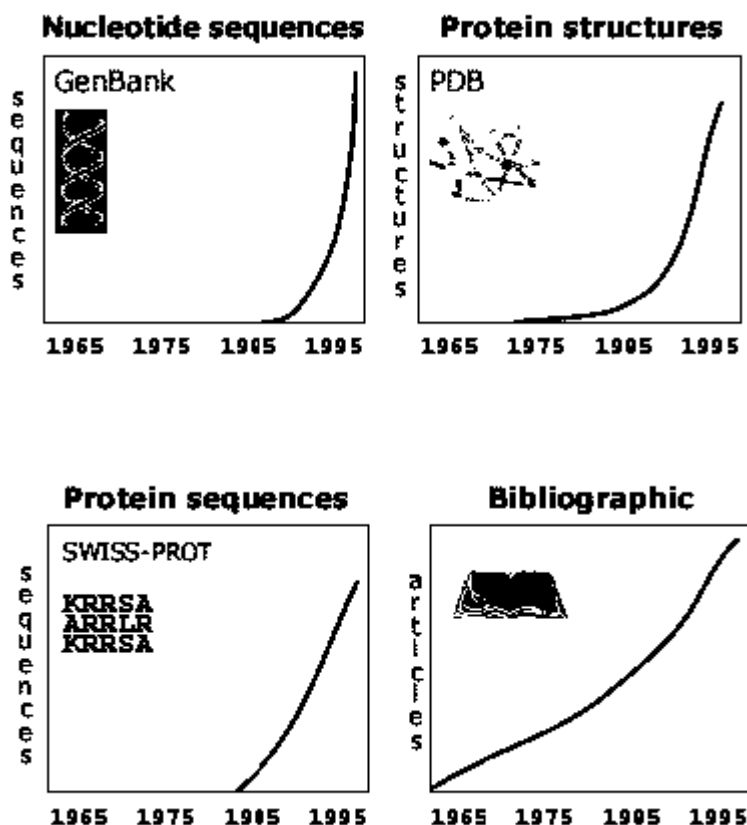


图 1.1 各类生物信息的同步增长状况。图中依次为核酸序列 (GenBank)、蛋白质序列 (PDB)、蛋白质序列 (SWISS-PROT) 和文献数量增长幅度 (引自 NCBI, 2000)。

## 二、生物信息学的概念

生物信息学便是在生物信息的急剧膨胀的压力下诞生了。

一般意义上,生物信息学是研究生物信息的采集、处理、存储、传播、分析和解释等各方面的一门学科,它通过综合利用生物学、计算机科学和信息技术而揭示大量而复杂的生物数据所赋有的生物学奥秘。具体而言,生物信息学作为一门新的学科领域,它是把基因组 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。基因组信息学、蛋白质空间结构模拟以及药物设计构成了生物信息学的 3 个重要组成部分。从生物信息学研究的具体内容上看,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新工具。Claverie (2000)的一段英文描述如下:“Bioinformatics is the science of using information to understand biology. It's the discipline of obtaining information about genomic or protein sequence data. This may involve similarity searches of databases, comparing your unidentified sequence to the sequences in a database, or making predictions about the sequence based on current knowledge of similar sequences.”

生物信息学最初更多地是关注数据库,那些数据库存储着来自基因组测序计划完成的序列数据。目前生物信息学已今非昔比,它所关注的是各类数据,包括生物大分子的三维结构、代谢途径和基因表达等等。生物信息学最使人们感兴趣的是它利用计

算方法分析生物数据，如根据核酸序列预测蛋白质序列、结构、功能的算法等。虽然这些预测还不是非常精准，但是当可靠的实验数据还无法得到的情况下，这一预测可以作为一盏路灯，指示你应如何开展实验。

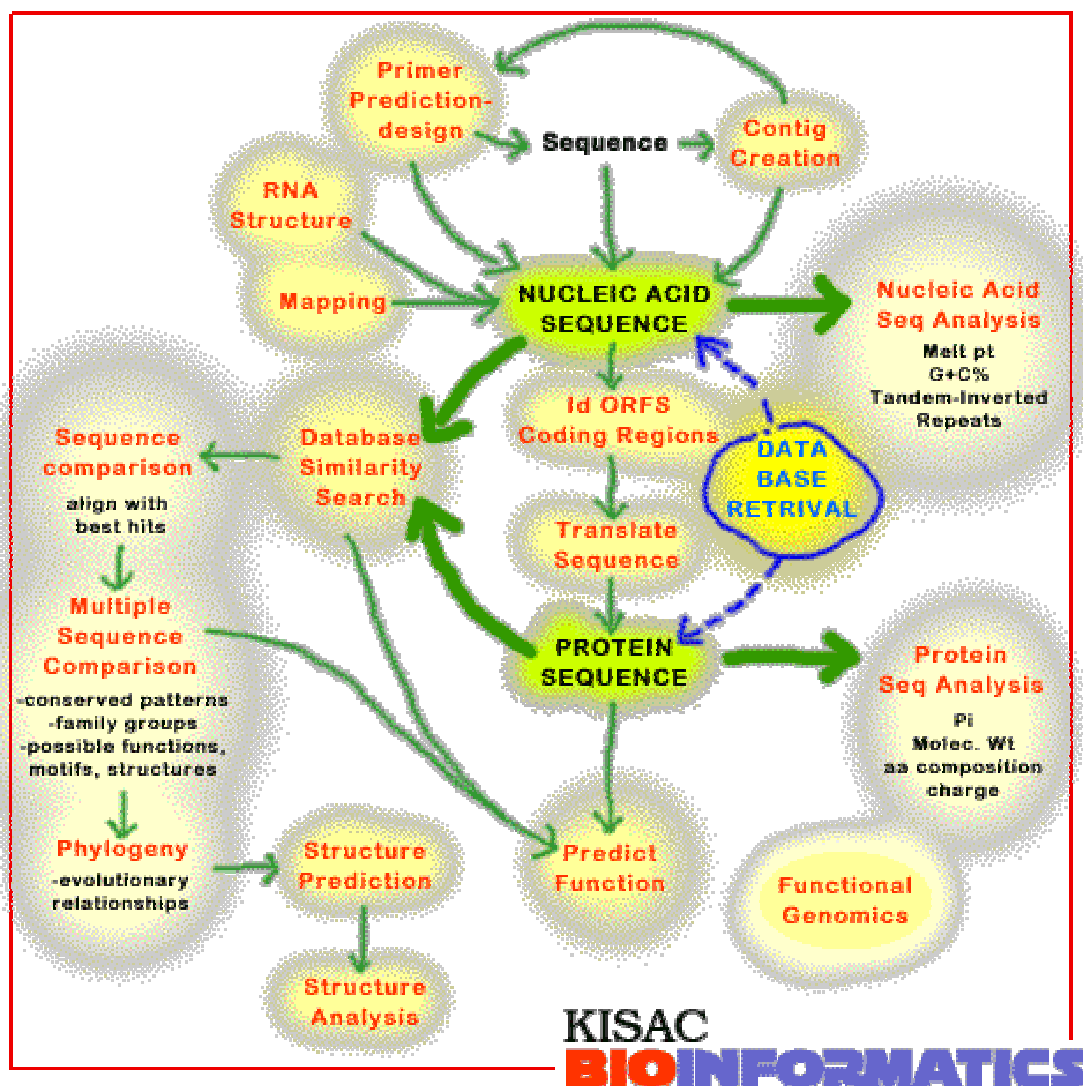


图 1-2 生物信息学“路线图”。取自<http://www.kisac.ki.se/>。

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，波林(Pauling)分子进化理论的出现，已预示着生物信息学的来临。而真正意义上的“生物信息学(Bioinformatics)”一词的出现则是 1990 年(见：“A term coined in 1990 to define the use of computers in sequence analysis” (Claverie, 2000)，据说是由出生在马来西亚的美籍学者林华安(Hwa A. Lim)首次使用的(郝柏林和张淑誉，2002)。

虽然生物信息学的历史并不长，但正象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，以及获取其它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，但很多软件是可以免费获取的。这些分析软件(见附录)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和生物信息学显示其价值的窗口。

NCBI Tools for data mining

PubMed Entrez BLAST OMIM Taxonomy Structure

Search GenBank for  Go


NCBI  
SITE MAP


BLAST  
standard tool  
for sequence  
analysis

COGs  
Clusters of  
Orthologous  
Groups

ORF finder  
finds open  
reading  
frames

**BLAST** The Basic Local Alignment Search Tool  
(BLAST), for comparing gene and protein sequences against others in public databases, now comes in several flavors including [PSI-BLAST](#), [PHI-BLAST](#), and [BLAST 2 sequences](#). Specialized BLASTs are also available for [human](#), [microbial](#), and [malaria](#) genomes, as well as for [vector contamination](#), [immunoglobulins](#), and [tentative human consensus](#) sequences.

 **Clusters of Orthologous Groups (COGs)** currently covers 21 complete genomes from 17 major phylogenetic lineages. A COG is a cluster of very similar proteins found in at least three species. The presence or absence of a protein in different genomes can tell us about the evolution of the organisms, as well as point to new drug targets.

 **ORF finder** identifies all possible ORFs in a DNA sequence by locating the standard and alternative stop and start codons. The deduced amino acid


 **Electronic PCR** allows you to search your DNA sequence for sequence tagged site (STS), which have been used as landmarks in various types of genomic

图 1.3 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。

生物信息学还有另一个经常被使用的名字：“计算生物学” (computational biology)，此外“计算分子生物学” (computational molecular biology) 和“生物分子信息学” (biomolecular informatics) 等也被使用过。但严格意义上说，计算生物学的范围应更宽泛些 [见“Strictly speaking, bioinformatics is a subset of the large field of computational biology, the application of quantitative analytical techniques in modeling biological system.” (Gibas and Jambeck, 2001)]。

正确认识和理解生物信息学这门新学科非常重要，它有助于该学科的科学学习和研究。《Bioinformatics》杂志的一篇社论文章 (2000, vol 16 no.3，其翻译稿见庞洪泉和樊龙江，生物技术通报，2002，2：47-52)，评析了人们对生物信息学的一些不正确的认识：(1) “人人可以从事生物信息学研究”。这一认识的根源来自对生物信息学的 2 个误解，一是生物信息学研究不需大量经费投入，因为有如此多的数据资源，只要找本生物学教科书，有台电脑并连到国际网上，人人可以从事生物信息学研究；二是生物信息学的软件是免费的。殊不知生物信息的巨量特征目前向计算机提出了严峻的考验，而一台大型新型计算机可能要以千万甚至亿元计算，同时大量先进、最新的生物信息学分析软件包都是商业化产品，不付钱难以到；(2) “你最终还是需要具体的实验”。实验生物学家非常羡慕生物信息学家，认为“他们只是敲敲键盘，然后便是写论文”，他们的研究结果只是一种试验结果的预测，是对实验研究的一种“支持”。在分子生物学研究中，固定的模式应是先有某一假设，然后用某一实验去验证或支持

这一最初的猜测。在生物信息学研究中，也同样进行着这一模式：有一无效假设(例如某一序列在数据库中没有同源序列)，然后进行实验(如搜索数据库)并验证，拒绝或接受无效假设(如该序列的确有或无同源序列)。这是一个标准的假设—实验模式。在其它学科中，计算科学已被作为深入理解科学问题的重要手段，而在生物学领域还没有形成这样的共识；(3)“生物信息学是门新技术，但只是一门技术而已”。由此把生物信息学定位为一门新的应用学科。正如前面所说，虽然生物信息学是一门新学科，但在 60-70 年代，该学科最重要的一些算法便已被提出，生物计算和理论研究便形成雏形。把生物信息学仅仅作为一门应用技术，是从信息学移植来的技术应用于生物学科领域，这是一个致命的误解。生物信息学实际是一门充满丰富知识内涵的学科，它有很多尚待解决的科学问题。这些问题包括生物学方面的(如分子的功能如何进化)和计算方面的(如数据库系统间如何最有效地协同)。生物信息学不仅仅是一个技术平台，它同样需要周详的实验计划和准确的操作，同样需要丰富的想象和一瞬即逝的运气。

## 第二节 生物信息学发展简史

表 1.2 列出了生物信息学最近几十年的主要事件。这些事件大多是在“生物信息学”(bioinformatics)一词出现前便发生了。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 的替换矩阵和 Needleman-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 2 项内容上不断改善；(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 系统软件包自 1993 年出现，1995 年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，是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中占有一席之地(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。GenBank 等数据库中数据的增长在近十年来呈直线上升趋势(图 1.1)，这条曲线很容易就使我们联想到生物信息学的发展历程，可以说，这条曲线便是生物信息学近十余年发展的写照。生物信息学在近十余年间经历了长足的发展，并迅速成为生命科学新的生长点。人类基因组计划的实施和生物医药工业的介入是生物信息学迅猛发展的主要推动力。

英国剑桥大学出版社出版的《Bioinformatics》期刊([www.bioinformatics.oupjournal.org](http://www.bioinformatics.oupjournal.org))是目前世界最知名生物信息学的学术期刊之一，它的前身是《Computer Applications in the Bioscience》(CABIOS)，1998 年更名为《Bioinformatics》。该杂志主要发表计算分子生物学、生物数据库和基因组生物信息学方面的文章。另外带有生物信息学字样的杂志还有《Applied Bioinformatics》、《Briefings in Bioinformatics》、《Journal of bioinformatics and computational biology》、《Genomics, proteomics & bioinformatics》、《Proceedings / IEEE

Computer Society Bioinformatics Conference》以及网上生物信息学杂志《BMC Bioinformatics》([www.biomedcentral.com](http://www.biomedcentral.com))等。其它与生物信息学相关的出版物还很多,如《Nucleic Acids Research》、《Genome Research》、《Genomics》、《J. Mol. Biol.》、《BioTechniques》、《BioTechnology Software》等。

表 1.2 生物信息学发展的简史

1962	Pauling 提出分子进化理论
1967	Dayhoff 构建蛋白质序列数据库
1970	Needleman-Wunsch 算法被提出
1977	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; EMBL 创立
1982	-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(Wilbur-Lipman 算法)
1985	快速序列相似性搜索程序 FASTP/FASTN 发布
1988	美国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; 三大核酸数据库(GenBank、EMBL 和 DDBJ)开始国际合作
1990	快速序列相似性搜索程序 BLAST 发布
1991	表达序列标签(EST)概念被提出, 从此开创 EST 测序
1993	英国 Sanger 中心在英国休斯顿建立
1994	欧洲生物信息学研究所(EMBL)在英国 Hinxton 成立
1995	第一个细菌基因组测序完成
1996	酵母基因组测序完成
1997	PSI-BLAST(BLAST 系列程序之一)发布
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	多细胞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

\*主要引自美国国家生物信息中心(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补入。

\*\*以上主要算法的原始文献出处: Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol.* 1970, 48(3):443-53; Staden R. Sequence data handling by computer. *Nucleic Acids Res.* 1977, 4(11):4037-51; Smith TF, Waterman MS. Identification of common molecular subsequences. *J Mol Biol.* 1981, 25;147(1):195-7; Doolittle RF. Similar amino acid sequences: chance or common ancestry? *Science.* 1981, 214(4517):149-59; Wilbur WJ, Lipman DJ. Rapid similarity searches of nucleic acid and protein data banks. *Proc Natl Acad Sci U S A.* 1983, 80(3):726-30; Lipman DJ, Pearson WR. Rapid and sensitive protein similarity searches. *Science.* 1985, 227(4693):1435-41; Karlin S, Altschul SF. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl. Acad. Sci. USA,* 1990, 87:2264-2268.

我们可从另一个角度来审视生物信息学的发展历程：美国家生物技术信息中心 (NCBI) 的十年 (1989-1999) 发展史，它是生物信息学近十余年来发展的一个缩影。NCBI 的十年发展史 (图 1.5) 可以说是年年有进步：筹备 (1989)、BLAST 启动 (1990)、Entrez 开始检索 (1991)、GenBank 加盟 (1992)、Entrez 上网和 3-D 分子数据建立 (1993)、NCBI 上网 (1994)、解读序列 (1995)、从序列中发基因 (1996)、PubMed 登网和蛋白质分析 (1997)、GenBank 碱基数据超过 10 亿 (1998)、关注人类基因组 (1999)。GenBank 十年来分子数据的增长曲线也正表明了 NCBI 的十年发展轨迹。



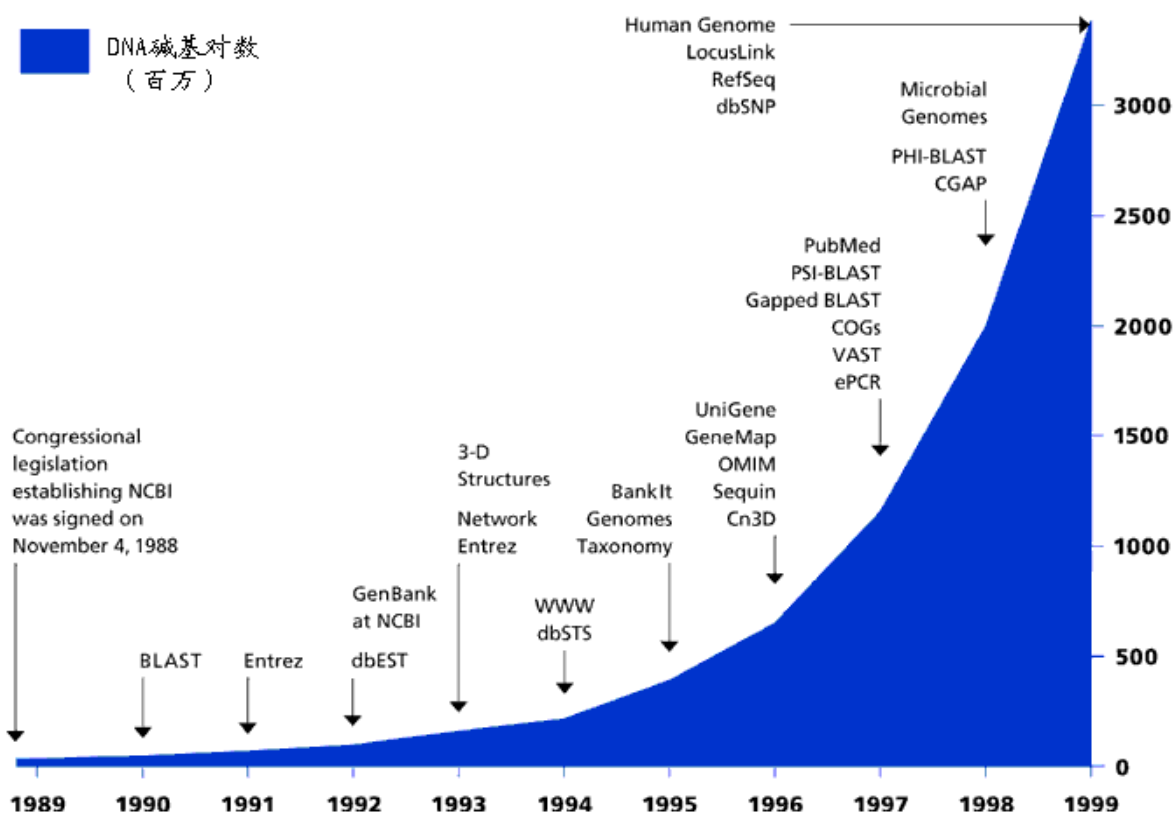
图 1.4 美华盛顿大学 Phil Green 教授。他所研制的自动测序组装系统 Phred-Phrap-Consed 被广泛应用于鸟枪法测序，其中包括人类基因组计划。



图 1.5 *Nature* 和 *Science* 2001 年 2 月 15 日和 16 日人类基因组专刊封面。*Science* 封面中的五位成年人分别为 Celera 公司人类基因组测序计划基因材料的提供者。



图 1.6 美国国家生物技术信息中心(NCBI)十年(1989-1999)发展简史 (NCBI, 1999)



\*图中涂黑部分表示 GenBank 数据库 DNA 碱基数据增长情况(单位：百万)；

\*\*图中各年主要事件说明：

1989：NCBI 被国会批准于 1988 年 11 月成立；

1990：BLAST 搜索程序研制完成；

1991：Entrez 检索系统(光盘)建立；

1992：GenBank 划归 NCBI，NCBI 建立 EST(表达序列检签)数据库(dbEST)；

1993：Entrez 检索网络系统建立，同时 Entrez 中增加三维大分子结构数据内容；

1994：NCBI 网站建立，STS(序列标签位点)数据库(dbSTS)在 NCBI 建立；

1995：向 GenBank 发送 DNA 序列系统 BankIt 面市，随着人类基因组计划的开展和数据库数据的膨胀，NCBI 分别建立基因组数据库和分类浏览器；

1996：为了帮助从序列中发现基因，UniGene、GeneMap(人类基因转录图谱)、OMIM(Online Mendelian Inheritance in Man)、Cn3D 数据库建立，序列发送新系统 Sequin 面市；

1997：文献检索库 PubMed 上网 新的搜索程序 PSI-BLAST(Position-Specific Iterated BLAST)和 Gapped BLAST(允许空位)研制完成，载体搜索工具 VAST 和 PCR 分析软件 ePCR 面市，COG(Clusters of Orthologous Groups)开始用于蛋白质序列的直系同源分析；

1998：20 种微生物基因组数据被公开，PHI-BLAST(Pattern Hit Initialed BLAST)完成，癌症基因组结构计划(CGAP)开始实施；

1999：完成一系列用于人类基因组分析工具和资源：LocusLink、RefSeq 和 dbSNP。

Ouzounis 和 Valencia(2003) (见 Christos A. Ouzounis and Alfonso Valencia. Early bioinformatics: the birth of a discipline ----- a personal view. *Bioinformatics*. 2003, 19(17): 2176-2190) 总结了截止到 10 年前 (上世纪 90 年代初) 生物信息学发展的重要研究成果, 其中还列出了所谓“TOP 20 PAPERS”(表 1.3)。当然这只是他们的一家之言, 难免有自己的偏好, 仅供参考。例如著名的 Smith-Waterman 算法(1981)就没有被列入。

表 1.3 早期影响生物信息学发展的 20 篇经典文献。取自 Ouzounis and Valencia (2003)。

Publication	Comments
Zuckerandl and Pauling, 1965b	First use of molecular sequences for evolutionary studies
Fitch and Margoliash, 1967	Use of molecular sequences to build trees
Needleman and Wunsch, 1970	First implementation of dynamic programming for protein sequence comparison
Lee and Richards, 1971	Calculation of accessibility on protein structures
Chou and Fasman, 1974	First secondary structure prediction method
Tanaka and Scheraga, 1975	Simulation of protein folding
Dayhoff, 1978	First collection of protein sequences
Hagler and Honig, 1978	One of the first explicit attempts to simulate protein folding
Doolittle, 1981	Seminal paper examining divergence and convergence in protein evolution
Felsenstein, 1981	One of the first statistical treatments of evolutionary tree construction
Richardson, 1981a	The most comprehensive description of protein structure to that date
Kabsch and Sander, 1984	Discovery with profound implications for model building by homology and structure prediction
Novotny <i>et al.</i> , 1984	The inability of distinguishing correct from incorrect structures threw back structure prediction approaches for a long while
Chothia and Lesk, 1986	Examination of divergence between sequence and structure
Doolittle, 1986	Influential book on sequence analysis
Feng and Doolittle, 1987	The first approach for an efficient multiple sequence alignment procedure, later implemented in CLUSTAL
Lathrop <i>et al.</i> , 1987	One of the first applications of Artificial Intelligence in protein structure analysis and prediction
Ponder and Richards, 1987	The very first threading approach, using sequence enumeration
Altschul <i>et al.</i> , 1990	The implementation of a sequence matching algorithm based on Karlin's statistical work
Bowie <i>et al.</i> , 1991	The first implementation of protein structure prediction using threading

### 第三节 基因组时代：生物信息学的应用与展望

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。当基因组测序计划持续开展, 研究重点已逐步从数据的积累转向数据的解释。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、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学科, 在许多方面影响着医学、生物技术和人类社会。现在作为一名分子生物学者, 不具备一些基本的生物信息学技能已几乎难以胜

任。实验室的每一项技术，从简单的克隆、PCR 到基因表达分析都需要在计算机上进行数据的处理，这些工作均需要理解 DNA 和蛋白质分析工具的基本算法。



生物信息学家们面对的是堆积如山的 DNA 片段。这是在人类基因组序列 2001 年完成后出现的一幅漫画：如何真正破译人类自身的庞大的基因组？

我们处在一个基因组时代。许多新技术，如用于大规模测序工程的毛细电泳 (capillary electrophoresis)，基因芯片制造的光刻技术 (photolithography) 和机器人技术 (robotics technology) 等应用于基因组研究，使我们能在以前不可能达到的尺度和角度上观察生物学现象：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。这些新技术的一个共同特点是产生大量的数据。例如 GenBank 数据库已拥有了超过  $10^{10}$  个 DNA 序列数据，并以每年翻一翻的速度增长。那些分析基因表达模式、蛋白质结构、蛋白质间互作等的新技术又会产生更多的数据。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。

生物信息学面临着越来越多的困难，许多困难是在我们面对大规模科技工程时，所有生物学家都将碰到的问题。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。这一问题将使那些可以培养下一代生物信息学者的人才匮乏。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并无什么实践意义。如何找到共同感兴趣的问题是生物信息学的重要目标。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。

一个生物信息学研究者需要怎样的基本条件呢？Gibas and Jambeck 在他们的《Developing Bioinformatics Computer Skills》(C. Gibas and P. Jambeck, O'REILLY, 2001) 书中大致给出了如下标准：

- You should have a fairly deep background in some aspect of molecular biology. ...but without a core of knowledge of molecular biology you will, as one person told us, “run into brick walls too often.”
- You must absolutely understand the central dogma of molecular biology.
- You should have substantial experience with at least one or two major molecular biology software packages, either for sequence analysis or molecular modeling.
- You should be comfortable working in a command-line computing environment.
- You should have experience with programming in a computer language such as C/C++, as well as in a scripting language such as Perl or Python.

生物信息学作为一个组合学科，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，它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 将作用有限。

生物信息学研究的一个核心问题是数据库的开发：如何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、文献记录等数据。其次是从诸如组装完成的核酸或蛋白质序列中识别模式的算法、用于相似性比较或系统发育构建的序列列线 (alignment)、线性序列或高维结构的模序 (motif) 识别和基因表达的共有模式等等。

如上所述，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：初级数据 (primary data) 的组成、谁应拥有这些数据、应什么时候和如何公开、对数据的进一步使用可否设置限制等。目前已经隐现的两方面问题可能阻碍生物信息学研究的进展，即 (1) 数据公开前的使用问题和 (2) 对已公开数据的保存限制。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) 采用了一种数据正式公布前即上网释放的政策，许多其它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。但是，这种利他主义的数据释放政策需要一些保护，如对产生初级数据的人应能使之得到应有的认可。有人最近提出用类似于“私人通信”的方式来处理这些尚未正式公布的数据，这样可以在一定程度上保护这些数据的知识产权。生物信息学研究面对的第二个问题并不是对数据使用的限制而是对下游研究的限制，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并可进行何种分析。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。不妨想象一下，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。5 年前，百慕大协定 (Bermuda Conventions) 为基因组序列的释放建立了一个很好的标准；鉴于数据释放和使用政策对生命科学研究的深远影响，我们有必要认真考虑为接下来的 5 年制定些什么标准。在后基因组时代 (postgenomic era)，人们期待在对生物发育机理、代谢过程和疾病认识方面有所突破。可以肯定地预言，生物信息学研究将对我们的一些认识产生根本性改变，如基因表达调控、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。捆住手脚，要在数据的海洋中畅游是很困难的。

在中国，生物信息学随着人类和水稻等基因组研究的展开已显露出蓬勃发展的势头。许多大学和科研院所已经投入大量人力开设生物信息学专业、建立生物信息学研

研究所（中心）并从事这方面的研究工作，例如北京大学生物信息中心 (<http://www.ipc.pku.edu.cn/>)、中国科学院上海生命科学院生物信息中心 ([www.biosino.org.cn](http://www.biosino.org.cn))、清华大学、天津大学、内蒙古大学、复旦大学以及浙江大学生物信息学研究所 (<http://ibi.zju.edu.cn>) 等等。生物信息学作为基因研究的有力武器，被广泛用于新基因的发现，以达到将有用新基因抢先注册专利的目的。在这场抢基因的国际竞争中，如何结合我国科研、开发状况，重点投入以求得局部优势和商业回报，是中国科学家和相关部门必须面对的新课题。