# 绪 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奥秘,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。与此同时,来自其它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*)杂志在 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》(*Bioinformatics—Trying to swim in a sea of data*)的文章(Roos,2001)。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?”。近年来高通量测序技术的出现,使数据海洋更添排山倒海之势。生物信息学便是我们能找到的可以畅游数据海洋的一条“轻舟”。生物信息学是一门年轻的学科,它充满挑战和机遇,且引人入胜。

## 第一节 生物信息与生物信息学

### 一、迅速增长的生物信息

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,如分子序列数据(核酸和蛋白质)、蛋白质二级结构和三维结构数据等等(详见第 1-1 章)。由测序仪等产生的大量核酸序列和三维结构数据被存在各类数据库中,这些原始数据构成的数据库就是所谓的初级数据库(primary database);那些由原始数据分析而来的诸如功能区(domain)、二级结构、疏水位点等数据,则组成了所谓的二级数据库(secondary database)。

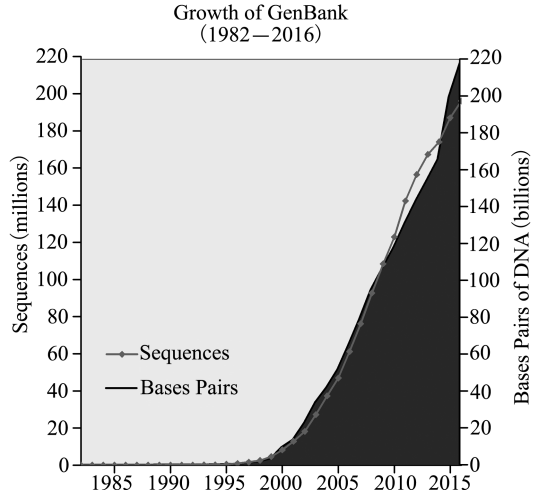


**图 1.1 国际核酸序列数据库 GenBank 记录数量增长情况(截至 2016 年 8 月)(www.ncbi.nlm.nih.gov/genbank/statistics/)**

生物信息的增长是惊人的。近年来,特别是随着高通量测序技术的出现,核酸库的数据每 14 个月左右就要翻一翻。2000 年底,数据库数据达到了创记录的 100 亿个碱基对(103.3 亿,GenBank Release 120, 2000)(图 1.1),而现今已达到 1 815.6 亿个碱基记录,如

果再加上差不多相同数量的基因组测序数据(GenBank 单独列为 WGS 类数据)8 055.5 亿个碱基对,在国际公共核酸序列数据已达到近 9 800 亿个碱基或 1.96 亿条序列数据(表 1.1)。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,遍布世界各地科研实验室或商业服务公司的高通量测序仪(如 Illumina 公司测序仪等)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公开或内部的生物信息库中。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(详见第 1-1 章)。

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、准确解读和充分使用这些信息?

**表 1.1 国际核酸序列数据库 GenBank 各大类核酸碱基数量的增长情况(Benson 等, 2014)**

| Division | Description | Release 197(8/2013) | Annual increase(%)[a] |
|---|---|---|---|
| WGS | Whole-genome shotgun data | 500 420 412 665 | 62.4 |
| TSA | Transcriptome shotgun data | 8 633 123 935 | 49.9 |
| PHG | Phages | 119 812 712 | 42.5 |
| VRL | Viruses | 1 757 202 472 | 22.9 |
| BCT | Bacteria | 10 281 048 518 | 21.8 |
| ENV | Environmental samples | 3 743 277 434 | 10.9 |
| INV | Invertebrates | 2 737 140 646 | 9.8 |
| PAT | Patented sequences | 13 290 161 247 | 9.7 |
| PLN | Plants | 5 963 882 822 | 8.8 |
| GSS | Genome survey sequences | 23 726 384 753 | 8.1 |
| VRT | Other vertebrates | 3 068 956 026 | 6.3 |
| MAM | Other mammals | 911 342 025 | 5.6 |
| HTG | High-throughput genomic | 25 184 819 955 | 3.4 |
| HTC | High-throughput DNA | 656 196 063 | 2.7 |
| UNA | Unannotated | 130 510 | 2.1 |
| EST | Expressed sequence tags | 41 665 629 009 | 1.9 |
| PRI | Primates | 6 425 093 034 | 1.7 |
| SYN | Synthetic | 941 078 074 | 1.4 |
| ROD | Rodents | 4 451 315 297 | 0.4 |
| STS | Sequence tagged sites | 636 326 479 | 0.0 |
| TOTAL | All GenBank sequences | 654 613 333 676 | 45.1 |

[a]Measured relative to Release 191 (8/2012).

## 二、生物信息学的概念

生物信息学学科是在生物信息急剧膨胀的压力下诞生的。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,而“生物信息学(Bioinformatics)”一词的出现则是在1990年(详见下节)。

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,它通过综合利用分子生物学、遗传学、计算机科学和信息技术而揭示大量且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,在获得基因序列和蛋白质编码区的信息后,进行蛋白质功能、结构模拟和预测等,然后依据特定蛋白质的功能进行必要的药物设计等等一系列应用性研究。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)的描述给出了一个比较清晰的定义:“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.”。据Wikipedia有关“Bioinformatics”词条(http://en.wikipedia.org/wiki/Bioinformatics),生物信息学是统计学和计算机科学在分子生物学领域应用的一门学科。生物信息学最初的使用始于上世纪八十年代的晚期,主要集中在基因组学和遗传学领域,特别是基因组DNA大规模测序出现后。生物信息学的根本目标是增加对生物学过程的认识,具体而言,它更加注重发展和应用有效的计算方法(如模式识别、数据发掘、机器学习算法和可视化技术)来达到这一目标。目前该学科主要的研究领域包括序列联配、基因预测、基因组拼接、药物设计和筛选、蛋白质结构预测、基因表达和蛋白质互作预测、全基因组连锁和进化分析等(“The primary goal of bioinformatics is to increase the understanding of biological processes. What sets it apart from other approaches, however, is its focus on developing and applying computationally intensive techniques (e.g., pattern recognition, data mining, machine learning algorithms, and visualization) to achieve this goal. Major research efforts in the field include sequence alignment, gene finding, genome assembly, drug design, drug discovery, protein structure alignment, protein structure prediction, prediction of gene expression and protein-protein interactions, genome-wide association studies and the modeling of evolution”)。一个生物信息学早期的“路线图”见图1.2。由于高通量测序技术的出现,系统生物学及其他新兴领域的发展,使生物信息学已大大超越了该图的领域范围。

生物信息学最初更多的是关注数据库,有效存储着来自基因组等测序计划完成的序列数据(详见第1-2章)。目前生物信息学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大规模生物数据,如根据基因组DNA序列预测基因序列等。虽然这些预测有时还不非常精准,但这一预测可以作为一盏路灯,指示你应如何开展实验,大大提高分子生物学等研究效率。

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(图1-3展示了NCBI提供的部分在线生物信息学分析工具)。例如,当获得了大量序列数据以后,我们现在已能进行基因家族或同源性分析;进行基因序列的聚类,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等。生物信息学软件很多已成为商

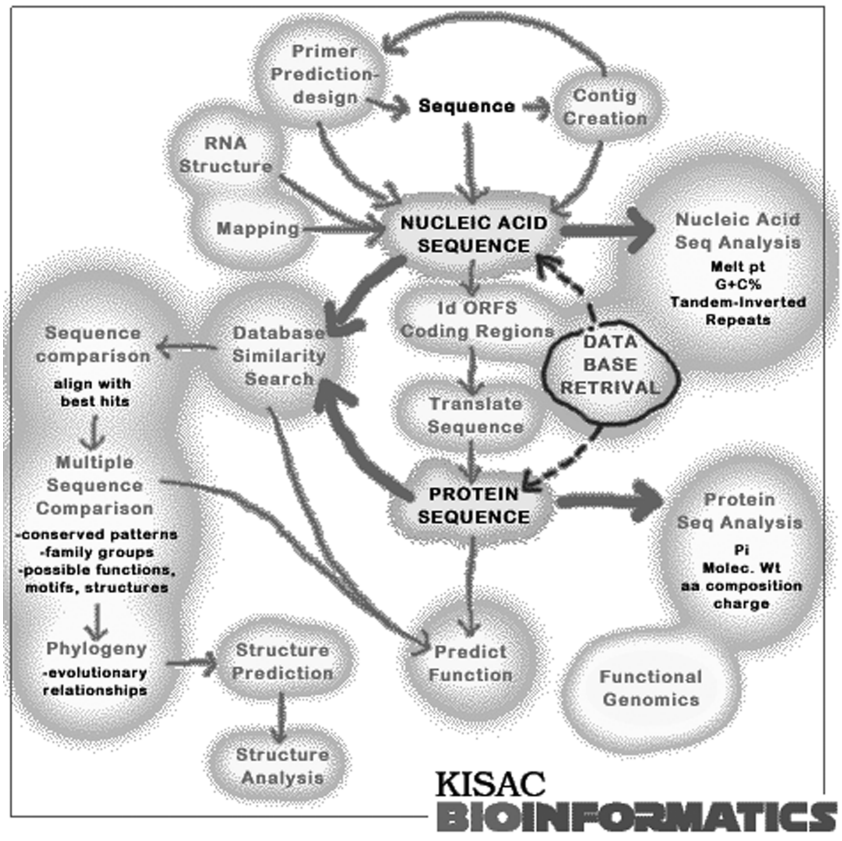


**图 1.2 生物信息学早期的一个“路线图”(取自 http://www.kisac.ki.se/)**
该图给出了生物信息学主要涉及的领域。

业化产品,但很多软件是可以免费获取或利用的。这些分析软件(详见第 4-3 章)已成为生物学最重要的研究工具,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。


**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、ePCR 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 和 Jambeck, 2001)]。同时,生物信息学的分支学科也在大量涌现,如基因组生物信息学、结构生物信息学、化学生物信息学等。

正确认识和理解生物信息学这门新学科非常重要,它有助于对该学科的科学研究和学习。*Bioinformatics* 杂志 2000 年的一篇社论文章(庞洪泉和樊龙江编译,2002),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,就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,而一台大型计算机可能要以百万甚至千万元计算,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,不付钱难以得到;(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,明确拒绝还是接受无效假设(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”。由此把生物信息学仅定位为一门新的应用性学科。正如前面所说,虽然生物信息学是一门新学科,但在 20 世纪 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 生物信息学简史与展望

### 一、生物信息学发展简史

生物信息学的诞生和发展最早可以追溯到 20 世纪 60 年代,两届诺贝尔奖得主波林(L. C. Pauling)分子进化理论的出现(利用蛋白质序列进行进化分析),已预示着生物信息学的来临。1956 年在美国曾召开首次生物学中的信息理论研讨会,也有学者将其作为生物信息学的发源。而“生物信息学(Bioinformatics)”一词的出现则在 1990 年(Claverie, 2000),据

说是由出生于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)。1987 年佛罗里达州立大学任教期间,他认为生物学和信息学结合交叉是未来发展趋势,构思了“Bioinformatics”一词作为这个新领域名字,并于 1990 年组织了第一届生物信息学与基因组研究国际会议(Bioinformatics and Genome Research International Conference)。但据 Wikipedia 有关“Bioinformatics”词条介绍,该词最早由荷兰理论生物学家 Paulien Hogeweg 于 1978 年首先提出的(Hogeweg, P. 1978. Simulating the growth of cellular forms. Simulation 31, 90-96; Hogeweg, P. and Hesper, B. 1978. Interactive instruction on population interactions. Comput. Biol. Med. 8:319-327),他于 1977 年便在他所在的荷兰乌特勒支大学(Utrecht University)建立了理论生物学和生物信息学研究小组。

表 1.2 列出了生物信息学发展过程中的主要事件,这些事件大多是在“生物信息学”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 4 个主要阶段:(1)理论基础形成期(20 世纪 60-70 年代):以 Dayhoff 的替换矩阵和 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内容: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不断改善;(2)学科成熟期(20 世纪 80 年代):以分子数据库和 BLAST 等数据库序列搜索程序为代表。1982 年三大核苷酸序列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 等为代表的数据库工具软件和相应的新算法被大量提出和研制,极大地改善了我们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(上世纪 90 年代-2005):以基因组测序及其拼接与分析技术为代表。基因组测序计划,特别是人类基因组计划的实施,产生以亿计的分子数据;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碱基识别、拼装和编辑等,为当时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、各国竞相发展的热点学科。如同 GenBank 数据库中数据增长的直线上升趋势(图 1.1),它同样是生物信息学发展的写照。生物信息学在这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;(4)高通量测序技术时期(2006—):以第二代测序技术 Solexa(后来的 Illumina 测序技术)和第三代测序技术及其相关数据分析技术为代表。高通量测序技术彻底改变了生物信息学研究对象(序列)的产生数量、成本、特征和应用领域等,它带来了一系列生物信息学方法的变革和创新,例如基因组拼接方法等。该技术使特定群体(如特定人群、不同作物等)在基因组水平遗传变异的检测成为可能,基于如此大规模基因组水平的遗传变异数据(如 SNP)可以根本改变我们许多研究思路和水平,例如个性化医疗成为可能,使基于生物信息学的遗传诊断更加便捷和准确(所谓精准医疗);作物基因组设计育种或基因组选择育种成为可能。

英国剑桥大学出版社出版的 *Bioinformatics* 期刊(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* 以及网络生物信息学杂志 *BMC Bioinformatics*(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*、*Genome Biology* 等。

**表 1.2 生物信息学学科发展的主要事件**

| 时间 | 事件 |
|---|---|
| 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 | 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(Wilber-Lipman 算法) |
| 1985 | 快速序列相似性搜索程度 FASTP/FASTN 发布 |
| 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 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)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 |
| 2001 | 人类基因组草图公布 |
| 2002 | 植物拟南芥等基因组序列公布 |
| 2004 | 蛋白质组学兴起 |
| 2005 | 第二代高通量测序技术出现 |
| 2008 | 基于高通量测序的转录组测序技术 RNA-SEQ 等出现 |
| 2009 | 第三代高通量测序技术出现 |

*引自美国家生物信息中心(NCBI)Education-Bioinformatics Milestone(2000),原文截至1999年果蝇基因组测序完成,2000年以后内容为本书作者补入。

**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 USA. 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

## 二、生物信息学技术的应用

虽然作为一个年轻学科,生物信息学对整个生物学发展产生了巨大推动作用。*Nature*总结了生物学领域引用率最高的100篇论文,生物信息学(包括系统进化方面)共有10篇入选,其中1篇甚至进入前十(表1.3)。

**表1.3 生物学学科历史上引用率最高的前100篇论文(Richard等,2014):生物信息学入选的10篇论文及其相关生物信息学工具**

| 工具/方法 | 排位名次 | 发表年份 | 论 文 |
|---|---|---|---|
| ClustalW | 10,28 | 1994<br>1997 | Thompson, J. D., Higgins, D. G. & Gibson, T. J. Clustal W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res.<br>Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F. & Higgins, D. G. The CLUSTAL_X Windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. |
| BLAST | 12,14 | 1990<br>1997 | Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. J. Mol. Biol.<br>Altschul, S. F.Madden, T. L. Schaffer, A. A. Zhang, J. H. Zhang, Z. Miller, W. Lipman, D. J. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res. |
| Neibour-joining method | 20 | 1987 | Saitou, N. & Nei, M.The neighbor-joining method: A new method for reconstructing phylogenetic trees. Mol. Biol. Evol. |

续表

| 工具/方法 | 排位名次 | 发表年份 | 论 文 |
|---|---|---|---|
| Bootstrap | 41 | 1985 | Felsenstein, J. Confidence limits on phylogenies: an approach using the bootstrap. Evolution. |
| MEGA | 45 | 2007 | Tamura, K., Dudley, J., Nei, M. & Kumar, S. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol. Biol. Evol. |
| UWGCG | 75 | 1984 | Devereux, J., Haeberli, P. & Smithies, O. A comprehensive set of sequence-analysis programs for the VAX. Nucleic Acids Res. |
| ModelTest | 76 | 1998 | Posada, D. & Crandall, K. A. MODELTEST: Testing the model of DNA. Bioinformatics |
| MrBayes | 100 | 2003 | Ronquist, F. & Huelsenbeck, J. P. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics |

生物信息学家们除了潜心研发新算法新软件,同时也在努力使他们的方法“平民化”,使广大生物学研究者能自己使用这些方法。他们主要在两个方面做了努力:一是生物信息学方法的程序化,研发适用于 PC 计算机操作系统(如 Window 系统)的生物信息学软件;二是使他们的方法网络化,提供一种所谓网络在线服务的生物信息学分析平台,生物学家只要进行网上递交,而不需要操心后台计算机问题。目前这种“平民化”方法已成为生物信息学家与生物学家建立联系的最主要桥梁。根据欧洲生物信息学研究所的分类,将生物信息学网络分析平台分为三类: SSS (Sequence Search Services),MSA (Multiple Sequence Alignment)和 BSA (Biological Sequence Analysis),即序列搜索、多序列联配和序列分析三类。同时,除了单一分析界面,现在还可以为特定目的构建一个生物信息学分析流程系统(bioinformatics workflow management systems)。这一趋势在著名生物信息学相关刊物《核酸研究》(NAR)得到了充分体现:该刊从 1993 年开始每年第一期均将刊出所谓分子数据库专刊(Database Issue),专门介绍世界范围内提供网络访问的主要公开分子数据库,其中包括已有的数据库的更新情况和新出现的数据库。这是生物信息学家构建的与生物学家建立联系的第一座桥梁;2004 年,该刊再次创立生物信息学软件网络服务专刊(Web Server Issue),于每年七月第一期刊出。该专刊主要介绍提供网络访问和在线分析的主要生物信息学方法。这是生物信息学家构建的与生物学家建立联系的第二座桥梁。

那么利用这些生物信息学家构建好的方法或平台,生物学研究者利用 PC 机能做什么分析呢? 当然可以做 DNA、RNA 和蛋白质序列分析,包括观察序列构成、蛋白质三维结构、RNA 二级结构、DNA 的回文结构预测等;可以做 DNA 编码区分析,获得编码的蛋白质序列;可以做基因组水平的分析等。但绝大多数人使用生物信息学方法或数据库,集中在数据库搜索(包括 PubMed/Medline 文献搜索、DNA 和蛋白质序列检索)、利用 BLAST 比较数据库序列与自己获得的序列、利用 ClastalW 进行多序列连配等。

生物学家们有时需要做出一个重要的判断:何时需要寻求生物信息学专业人员的帮助,而不是一味地花费大量时间进行简单重复工作,或尝试无谓的复杂生物信息学分析。何时

需要生物信息学专业人员？当你想做超过 100 条以上序列分析、当你想利用需要 LINUX 系统的软件，当你想利用高通量测序数据进行分析，当你需要处理大规模数据时（如芯片数据），当你的数据结构或完整性有问题，当你需进行复杂数据分析（如需要许多假设或先决条件或统计测验等）等等，我们建议你寻求生物信息学专家建议或帮助。

一个生物信息学研究者需要怎样的基本条件呢？Gibas 和 Jambeck 在他们的 *Developing Bioinformatics Computer Skills* 书中大致给出了如下标准：

- 应该具备分子生物学的核心知识，否则你会经常碰壁；
- 你当然要对分子生物学的中心法则知道得一清二楚；
- 你应该对至少 1-2 个用于序列分析或模型的主要分子生物学软件了如指掌；
- 你可以在用计算机命令行环境下轻松工作；
- 你应能用 C/C++计算机语言或 PERL 或 Python 脚本语言进行编程。

## 三、生物信息学学科展望

蛋白质、DNA 和 RNA 序列的计算分析在不断发生着变化。生物学实验新技术，如测序技术使实验数据急剧增长，当基因组测序计划持续开展，生物信息学研究重点已逐步从数据的积累转向数据的解释。用于基因组拼接、序列相似性搜索、DNA 序列编码区识别、分子结构与功能预测、进化过程的构建等等计算工具已成为生物信息学重要组成部分。这些工具有助于我们了解生命本质和进化过程。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、农学和人类社会。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。


**图 1.4 生物信息学家们面对的是堆集如山的 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（所谓基因组分析第二阶段）？

我们处在一个基因组和大数据时代。许多新技术，例如高通量测序技术（第二代和第三

代)、人工智能技术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。我们目前面临着一个大数据时代:数据的产生越来越快速,越来越便宜,但数据产生后数据处理的能力严重滞后。一个明显的例子是人类基因组测序。目前测定一个人类基因组可以在 1000 美元甚至更低成本获得,大量基因组被测定(如图 1.5);同时今后的发展人类基因组测序会变成医院等常规遗传诊断手段,这样产生的数据将是海量的。这就给生物信息学分析提出了更加严峻的挑战:如何高速和准确的分析数据并为诊断提供信息。

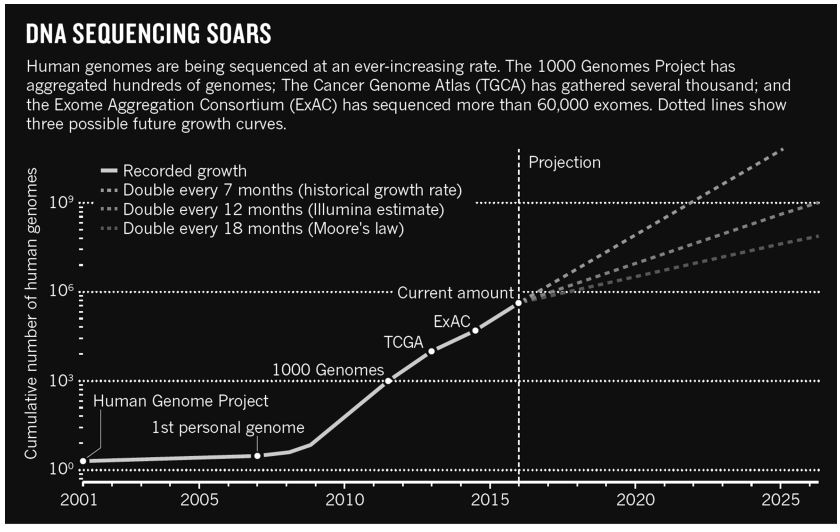


**图 1.5 人类基因组测序数量的增长(引自 Eisenstein,2015)**

目前人类基因组被测序数量已达到 $10^6$水平,同时图中给出了今后增长的三个预测曲线。

生物信息学还面临许多其他困难,这些困难是在大规模生物学科技项目中所有生物学家都将碰到的。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。生物信息学者对合作对方生物学问题的无知可能导致误解。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。

生物信息学离开不了大规模序列数据,序列数据的共享性非常重要。测序仪每天产生的大量初级数据(primary data),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等,往往对生物信息学者的介入和数据的应用产生直接影响。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,所以希望国内外一些政府资助的大规模基因组水平的研究计划(如表达分析和蛋白质组学研究)都能及时共享数据,毕竟这是纳税人的钱资助的项目。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,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、中国科学院上海生命科学院生物信息中心、生物物理研究所、清华大学、天津大学、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,可被广泛用于基因组拼接及其新基因的发现等,以达到抢占新基因专利制高点的目的。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。

面对生物大数据,一些生物信息学关键技术急需解决或优化(杨焕明,2012):

1. 大基因组 *de novo* 组装算法与软件

基因组组装是进行基因组分析的第一步,也是目前影响基因组学发展的最大阻碍之一。由于测序技术的不断更新,数据量的爆炸直接对组装的算法提出更高要求。同时,一些物种基因组很大且复杂,其重复序列和高杂合度等特征直接导致其基因组组装非常困难。

2. 大基因组注释核心技术

目前主要包括四个研究方向:重复序列的识别、非编码 RNA 基因的预测、蛋白质编码基因和结构的预测以及基因功能的注释。

3. 比较基因组与进化分析核心技术

通过对动植物基因组数据的比较基因组学分析,可以识别物种间共有保守基因和物种内特有保守基因;系统发生、正选择基因鉴定、染色体进化等分析技术。

4. 大基因组重测序数据分析核心技术

遗传多态性包括单核苷酸多态性(SNP)、插入删除变异(InDel)、结构性变异(SV)、拷贝数变异(CNV)等检测方法的开发和优化。该技术是动植物基因组的遗传变异及进化研究的基础,可以为分子育种提供指导。

5. RNA 分析技术

一个基因组内编码和非编码 RNA 数量巨大。转录组在有无基因组参考序列情况下,其分析组装算法和数据利用效果完全不同。同时,由于非编码小 RNA 和长 RNA 特异的作用机制,使其分析方法研究尤其重要。

合成生物学的发展引人注目。美国生物学家克雷格·文特尔领导的研究小组在 2010 年 5 月 20 日出版的美国 *Science* 杂志上,宣布他们创造了一个人造生命,更准确地说,他们利用实验室里现成的化学试剂,制造出了含有约 1 000 个基因的基因组。近日,在美国波士顿哈佛大学医学院 25 名科学家,提出在未来 10 年内合成一条完整人类基因组的计划,这意味着人造生物的诞生和合成生物时代已经来临。合成生物的起点是利用生物信息学方法设计生物基因组结构或构成,然后再进行实际的基因组合和实验(张春霆,2009)。目前合成生物

学的核心技术引入了生物信息学方法,如基因线路(Genetic circuit)和合成基因组等。

此外,从技术层面上,除了高通量大数据分析问题外,生物信息学还呈现如下几个发展趋势:针对复杂性状的基因网络分析问题;在表型和分子数据之间建立准确关联;人工智能在序列数据分析和诊断中的应用;大数据技术的引人和应用等等。

生物信息学专业领域的就业形势一片光明。最近 *Science* 专门对生物信息职业前景进行了调查(Levine,2014)。调查报道表明,随着产业界和学术圈对于生物信息学认知上的转变和大数据的扩张,促成了生物信息学领域工作机会的增长。以前科学家和公司往往会将生物信息学作为一种工具,但这门学科已经进化,具备自己的研究领域,"生物信息学家现在是创新的马达"。因此,正如猎头公司 Klein Hersh 国际的研发高级总监 Jared Kaleck 指出的一样,当前生物信息科学家有很多机会可以在生物技术、大型制药行业中寻找到生物信息、大数据分析方面的工作。生物信息学岗位在不同的公司其组织安排会有所不同。在制药企业和大型生物技术公司,大数据科学家可能会发现自己处于完全不同类型的组织架构中,例如,所有的大数据科学家和生物信息学家都集中在核心团队工作,另外一种状况是生物信息学家的岗位是分散的,分布在不同的部门或领域。行业对目前生物信息学从业者也有明确的要求。专家一致认为,最成功(或获得理想岗位)的生物信息学家往往具有大量的生物信息学技能,但最重要的一点总是对生命科学知识的掌握,也称作该行业的"专业知识"。实际上,"你对生物学的理解越深,你越能在这个领域的工作中游刃有余"。人事部门会专门寻找在多个生命科学领域拥有博士学位的科学家,包括分子和细胞生物学、化学、遗传学、免疫学和流行病学。除此之外,产业界的大数据工作也要求额外的关键技能,如文本挖掘、本体论、数据集成、机器学习和信息架构。当然,一些公司要求从业者具备优异的"量化能力",包括一系列的统计能力以及包罗万象的计算能力。这些能力的基础是编程能力,如 C++或 Java 的编码,或 PERL 或 Python 的脚本编写;能够控制操作系统如 UNIX 和 Linux,并具备 Hadoop 和 NoSQL 数据库等常用工具的知识。如果能够具备数据可视化和建立有效用户界面的经验,以及对于硬件一定熟悉度,则会增加你的"销路"。除了解决科学问题的能力,生物信息学家必须具备沟通能力。礼来公司高级分析研究员 Stephen Ruberg 表示,"生物信息学是团队作战",因此要求项目管理、团队建设和沟通的经验。实际上,"能够与其他科学家沟通才是我们最注重的技能"。此外,灵活度以及能够迅速适应也是至关重要的。"这是一个快节奏的环境,你必须要有不断使用新工具的心态,要不两年内你就要被淘汰了"。

## 第三节 本书的组织与使用

本书的定位是作为生物信息学专业学生的入门教材和非生物信息学专业学生和科研工作者的基本教材。因此,本书的建议阅读对象为本科和研究生阶段学生和从事生物学及其相关专业领域(如医学、农学等)研究与开发工作者。

本书共分为四部分(四篇):生物信息学基础(第一篇)、高通量测序数据分析(第二篇)、生物信息学外延与交叉(第三篇)和生物信息学资源与实践(第四篇)。作为生物信息学的基础篇(第一篇),包括 8 章内容,涵盖序列数据类型与产生、分子数据库、序列联配算法、基因预测方法、系统发生树构建和蛋白质结构预测等。同时,该篇还包含了一个生物信息学计算机基础,包括操作系统、主要编程语言等。该篇内容主要目的是使学习者掌握生物信息学

的基本概念、主要方法和编程语言等;第二篇为高通量序列数据分析篇,主要针对目前第二和三代测序技术产生的核苷酸序列数据,包括7章内容,涵盖高通量测序数据为基础的基因组拼接、基因组变异、转录组、非编码RNA、甲基化和宏基因组等分析原理和技术。同时也包括基于质谱的蛋白质组学数据分析方法;第三篇为生物信息学外延与交叉篇,总4章。生物信息学本身是一门交叉学科,其自身引入了许多其他学科方法用于序列数据分析,同时,生物信息学学科范围也有不同的界定,有些内容与其他学科有重叠。本篇介绍了与生物信息学紧密相关的四个生物学学科(系统生物学、群体遗传学、数量遗传学和合成生物学)。最后一篇为生物信息学资源与实践篇,包括4章,主要罗列了生物信息学主要相关术语、重要数据库和生物信息学软件工具资源,并结合实验课程内容,设计了8个序列数据相关问题及其生物信息学解决方案。

上述第一和第四篇内容建议作为本科教学的基本内容,第二和三篇可以作为研究生和生物信息学专业学生教学内容。


**图1.6 这幅漫画会使生物信息学初学者会心一笑。希望本书可以帮助你"站起来"!**
(画中EXPASY为著名在线蛋白质序列数据资源与分析平台)