第 14 章

生物信息学及其在基因组研究中的应用

14.1 引言

生物信息学(bioinformatics)是生物学与计算机科学以及应用数学等学科相互交叉而形成的一门新兴学科。它通过对生物学实验数据的获取、加工、存储、检索与分析,进而达到揭示数据所蕴含的生物学意义的目的(NIH and DOE, 1990)。

数学在生物学中某些领域的应用已经有百年的历史。例如统计学应用于群体遗传学可追溯到 19 世纪后半叶 Galton 的早期工作。但在此后的数 10 年中,这方面的研究并没有产生对大量计算的需求。直到 X 光衍射及核磁共振技术应用于生物大分子结构的研究,计算机成为不可缺少的工具,计算机在生物学中的应用才逐渐形成一门独立的学科。近年来,随着分子生物学,特别是人基因组计划的实施,不断产生出巨量的分子生物学数据,如核苷酸序列(图 14.1)。这些数据有着数量巨大和关系复杂等特

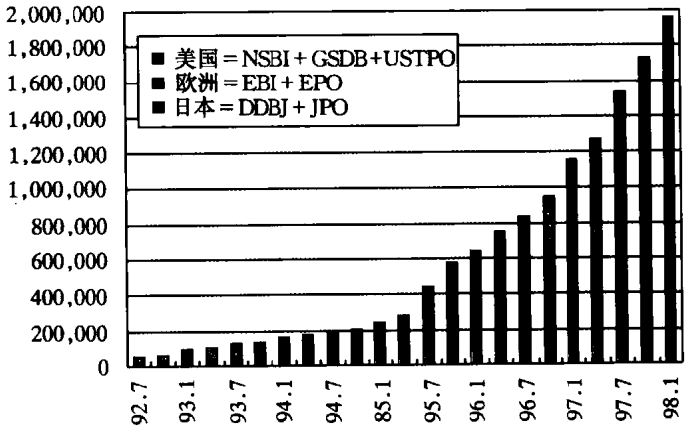

图 14.1

DDBJ/EMBL/GenBank DNA 序列数据库数据增长情况以及美国、欧洲和日本 3 个地区提供的 DNA 序列在数据库总量中占的比例(引自 DDBJ)

征，以至于不利用计算机根本无法实现数据的存储和分析。这样，生物信息学最终形成了一个独立的学科并被推上了生物科学发展的最前沿。

由于当前生物信息学发展的主要推动力来自分子生物学，生物信息学的研究主要集中于核苷酸和氨基酸序列的存储、分类、检索和分析等方面(图 14.2)，所以目前生物信息学可以狭义地定义为：将计算机科学和数学应用于生物大分子信息的获取、加工、存储、分类、检索与分析，以达到理解这些生物大分子信息的生物学意义的交叉学科。

本章介绍生物信息学在分子生物学，特别是在基因组研究中的应用及其发展趋势。

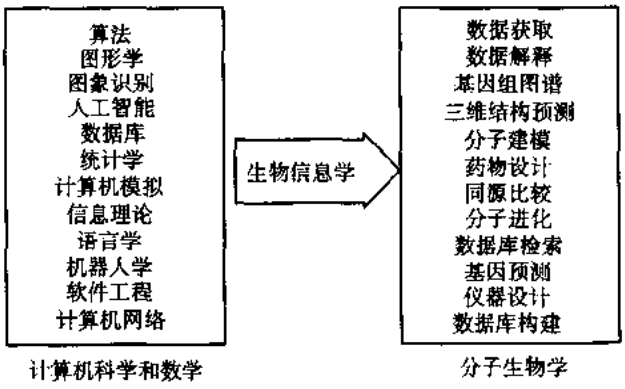

图 14.2　生物信息学与计算机科学以及分子生物学的联系

14.2　生物信息学研究的主要推动力

现代分子生物学的发展，特别是人基因组计划的实施，使生物学家所面对的数据不再是实验记录本上或文献上的几行简单数字，而是公共数据库中数以千兆计的记录。生物信息学就如同一个向导，帮助生物学家从这个生物信息的宝库中寻找他们所需要的生物学知识。在推动生物信息学发展的各种动力中，人类基因组计划和生物医药工业是其中的两个主要力量。

14.2.1　人基因组计划与 EST 战略

美国自从 1990 年开始实施人基因组计划(Human Genome Project，HGP)以来，现在已经进入了第 9 个年头了。HGP 的目标是测出人类和一些模式生物的基因组 DNA 全序列，制作完成全序列的物理图谱，找出全部人基因，以及发展出一系列实现上述目标所需的技术(Rowen L *et al.*，1997)。HGP 计划在 2003 年完成人基因组全序列测定。到 1996 年底，已经全部测序完成并建成物理图谱的生物包括 141 种病毒，51 种细胞器，2 种真细菌(包括大肠杆菌)，1 种古核生物(Archeon)和 1 种真核生物—酵母(Schuler GD *et al.*，1996)。1997 年又有 *Helicobacter pylori*(卷旋杆菌)基因组全序列完成测定。98 年底，*Caenorhabditis elegans*(线虫)的基因组全序列测定也将完成。截止 97 年底，人类基因组的测序工作只完成了 2%。但另一方面，全部人类基因(大约 50 000 ~ 100 000 个)中可能已经有一半(大约 50 000)有了部分测序的表达序列标签(expressed sequence tag，EST)。

EST 是从已建好的 cDNA 库中随机取出一个克隆，从 5′端或 3′端对插入的 cDNA 片

段进行一轮(one pass)自动测序,测得的约 500bp 的一段序列就称为一个 EST。每一个 EST 代表一个表达基因的部分转录片段。由于 EST 测序只测定部分序列,也不需要对克隆排序,因而具有经济和高效的特点。在 EST 测序中平均每个碱基所需费用只有传统 DNA 测序所需的 1/20 (Venter JC, 1993)。同时,由于全部是自动测序,而且从挑克隆、测序、输入数据库全部过程都实现自动化,因而可以节省大量的时间。

通过对 EST 的分析,有助于新基因的发现(Adams MD *et al.*, 1991)、分析基因在不同组织中的表达特异性(Okubo K *et al.*, 1992)和建立 DNA 物理图谱(Boguski MS *et al.*, 1995)。正因为 EST 具有如此的优越性,自从 1991 年 Venter 和他的同事们提出 EST 的概念(Adams MD *et al.*, 1991)以来,EST 测序已经成为许多基因组研究机构的工作重点。由于盈利性研究机构如 Human Genome Science (HGS, Rockville, MD, USA)和 Incyte Pharmaceuticals (Palo Alto, CA, USA)的推波助澜,许多医药公司纷纷投资 EST 测序。特别值得一提的是 Merck 公司(West Point, PA, USA)与 Washington 大学(St. Louis, USA)合作大规模进行 EST 测序,并将序列提供给公共数据库,使 EST 序列数据成为近年来在 HGP 中增加最快的一个部分。得益于大规模的 EST 测序,近几年来在公共数据库中 DNA 序列数据的数量以每年 1.8 倍的速度快速增长,到 1997 年底已经超过 1.2×10^9bp。对如此巨量的数据进行存储、分类、检索、比较等工作,来预测可能的基因和基因产物的结构和功能,没有计算机进行处理那简直是不可想象的。

14.2.2　生物医药工业

生物医药工业也是推动生物信息学发展的一重要动力。近年来生物医药公司纷纷投资生物信息学研究,主要原因在于生物信息学大大加速了新药开发过程。HGP 所推动的大规模 DNA 测序也为生物医药工业提供了大量可用于新药开发的原材料。有些基因产物可以直接作为药物,而有些基因则可以成为药物作用的对象。而生物信息学为分子生物学家提供了大量对基因序列进行分析的工具,可以在以下方面加快新药开发的进程:

(1)资料的获取,包括从数据库中寻找新药开发者感兴趣的基因序列和相关资料文献。

(2)基因功能的预测和基因生理作用的预测。

(3)需要大量信息处理的药物筛选和加工过程(Weinstein JN *et al.*, 1997)。

进一步的情况可以参阅近年的一些综述文章(如 Bains W, 1996)。

生物信息学大大缩短了发现新基因所需的时间。欧美许多医药公司纷纷投资 HGP 特别是 EST 的研究工作,其中不可忽视的一个原因是基因和基因序列可以申请专利保护。至今已有数千宗基因专利注册成功。因为生物信息学能够大大加快传统的基因发现和研究,因而成为各盈利性研究机构和医药公司争夺基因专利的重要工具,同时这一竞争又反过来极大地刺激了生物信息学的发展。

14.3　与 HGP 相关的生物信息学研究

HGP 目的之一,就是找到人类基因组中的所有基因。除功能克隆和定位克隆策略之外,生物信息学为分子生物学家提供了一条寻找和研究新基因的新思路,即从高度自动化

的实验出发，经过数据的获取与处理、序列片段的拼接、可能基因的寻找、基因功能的预测一直到基因的分子进化研究。这个过程的每一个环节，都是生物信息学研究的重要内容。

14.3.1　高度自动化的实验数据的获得、加工和整理

如何将实验室中得到的生物学信息转化为计算机能够处理的数字信息，是生物信息学的一个重要课题。这种转化大量地体现在各种自动化分子生物学仪器应用上，如 DNA 测序仪，PCR 仪等。这类仪器将实验所得的物理化学信号转化为数字信息，并对其作简单分析，再将分析结果用于实验条件的控制，完成高度自动化的实验过程。从事大规模 EST 测序和 DNA 物理图谱构建的实验室都已建立起高度自动化的机器人系统来完成大部分的实验工作。以美国 Whitehead/MIT 的基因组研究中心（Center for Genome Research，Cambridge，MA，USA）为例。该中心从事以 STS（见后）为基础的人基因组物理图谱的绘制工作。他们与 IAS（Intelligent Automation Systems Incorporated，Cambridge，MA，USA）合作，制成了各种自动化仪器，部分仪器达到了工厂水平。这些仪器由电脑控制，完成从自动的 PCR 反应，荧光探针的标记检测一直到数据输入数据库的一系列工作，每一轮能完成 150 000 个反应（Hudson TJ *et al.*，1995）。正是借助这些计算机控制的机器人的出色工作，在全部工作人员只有 24 人的条件下 Whitehead /MIT 只用了 2.5 年就绘制完成有 15 000 个位点，平均间隔 199kb 的人基因组物理图谱。这样高的效率在使用传统技术的条件下是不可想象的。

伴随着实验过程的高度自动化甚至工厂化，从事大规模分子生物学项目的实验室，每天需要存储的数据可以轻易地超过几千兆字节（gigabyte）。这样大的数据量必须用专门的实验室数据管理系统进行处理，以自动完成包括实验进程和数据的记录，常规数据分析，数据质量检测和问题的自动查找，常规的数据说明和数据输入数据库在内的各项工作。由于不同实验室需处理的数据类型各不相同，目前各个实验室都是各自开发自己的系统，还没有成熟的可用于不同实验室的分子生物学数据管理系统。但随着测序逐渐成为实验室的常规工作，对这种系统的需求会越来越大，此类系统的发展将成为大势所趋。

14.3.2　序列片段的拼接

目前 DNA 自动测序仪每个反应只能测序 500bp 左右。如何将这些序列片段拼接成完整的 DNA 顺序就成为接下来的一个重要工作。传统的测序技术通常将克隆进行亚克隆并对亚克隆进行排序。这些工作需要大量的人力物力。现在生物信息学提供了自动而高速地拼接序列的算法，即根据 Lander－Waterman 模型（Lander ES and Waterman MS，1988）利用鸟枪法进行测序，再将大量随机测序的片段用计算机进行自动拼接。这种技术不仅避免了亚克隆排序所需的大量繁琐的工作，还使序列具有一定的冗余性（redundancy，即一定数量的重复）以保证序列中每个碱基的准确性。这种技术已应用于 1.9Mb 的 *Haemophilus influenzae*（流感嗜血杆菌）（Fleischmann RD *et al.*，1995），0.58Mb 的 *Mycoplasmu genitalium*（枝原体）（Fraser CM *et al.*，1995）和 1.66Mb 的 *Methanococcus jannaschii*（甲烷球菌）（Bult CJ *et al.*，1996）的测序工作。并被证明是一种非常高效而廉价的技术（Fraser CM and Fleischmann RD，1997）。

由于重复序列的存在和基因家族等因素的影响，真核生物基因组结构庞大而复杂，因此真核基因组的拼接还无法象在原核生物和病毒基因组中那么有效，还需要进一步改进算法。不过，类似的算法已经用于拼接 EST 顺序（Parsons JD *et al.*，1992；Parsons JD，1995）和降低数据库冗余程度的工作（Grillo G *et al.*，1996），并取得了不错的效果。

序列拼接算法的进一步发展，需要在以下方面进行改进：①将已知的基因组知识应用于拼接算法，以进一步提高拼接真核基因组的有效性。②自动处理自动测序造成的差错，特别是对差错倾向的 EST 顺序更是如此。

14.3.3　基因区域的预测

在完成序列的拼接后，我们得到的是很长的 DNA 序列，甚至可能是整个基因组的序列。这些序列中包含着许多未知的基因，下一步就是将基因区域从这些长序列中找出来。

所谓基因区域的预测，一般是指预测 DNA 顺序中编码蛋白质的部分，即外显子部分。不过目前基因区域的预测已从单纯外显子预测发展到整个基因结构的预测。这些预测综合各种外显子预测的算法和人们对基因结构信号（如 TATA box 和加尾信号）的认识，预测出可能的完整基因。

在介绍算法之前，我们先介绍一下衡量一个算法优劣的标准：敏感性（sensitive）和特异性（specificity）。假设待测序列中有 $M1$ 条序列是基因序列，剩余的 $M2$ 条为非基因序列。我们用程序对待测序列进行预测，N 条序列被预测为基因，其中有 $N1$ 条确实为基因（即 $N1 \subset M1$），其余 $N2$ 条不是基因的一部分（$N2 \subset M2$）。敏感性定义为 $N1/M1$，它表示程序预测的能力。特异性定义为 $N1/N$，它表示预测结果的可信度。敏感性和特异性往往是一对矛盾，一般以敏感性和特异性的平均值作为评判程序优劣的标准。

预测外显子的基本算法，早期有最长 ORF（open reading frame）法。在细菌基因组中，蛋白质编码基因从起始密码 ATG 到终止密码平均有 1 000bp，而长于 300bp 的 ORF 平均每 36kb 才出现一次。所以只要找出序列中最长的 ORF（>300bp）就能相当准确地预测出基因（Claverie JM，1997）。

核苷酸语汇（nucleotide words，即数个连续核苷酸的排列）选用频率的统计差异也被用来区别编码和非编码区域（Claverie JM and Bougueleret L，1986；Bechmann *et al.*，1986）。这种差异可能来自编码和非编码区密码子选用的差异和周期特征的差异，其中一个显著的特征是 6 核苷酸（hexmers）的选用差异（Claverie，1997）。在目前的各种预测程序中这是一种被广泛应用的方法。

近年来同源比较算法也被应用于预测可能的基因。根据有以下几点：

（1）大约已经有 50% 的基因有了对应的 EST，已知的蛋白质序列也越来越多；

（2）不少原核生物和酵母的全序列已经测定；研究表明有将近一半的脊椎动物基因可以通过 BLASTX（见附录 14.4）在酵母，细菌和线虫的序列数据库中找到相似性相当高的序列（Claverie JM，1993；Green P，1993）；

（3）大多数 EST 都采用每个克隆分别从 5′和 3′测序，克服了早期 EST 只代表 3′外显子的缺点。

许多基因预测的程序都已经整合了同源比较算法（见附录 1.4.2～1.4.4），比如著名

的 GRAIL II 程序(Xu Y *et al.*, 1994)。

近年来隐藏马尔可夫模型(Hidden Markov Model, HMM)异军突起。这种方法将 DNA 序列的形成看作一个随机过程。编码和非编码的 DNA 序列在核苷酸选用频率上有所不同而对应于不同的马尔可夫模型。由于这些马尔可夫模型的统计规律是未知的,而 HMM 能够自动寻找出其隐藏的统计规律,因而称为隐藏马尔可夫模型。对于处理复杂的 DNA 序列,马尔可夫模型还需要学习不同 DNA 结构的信号(Krogh A *et al.*, 1994b)。

前面提到目前的基因预测已经从单纯的编码序列预测发展到整个基因结构的预测。被广泛用来将预测出的各个可能的外显子和内含子而拼接成完整基因的算法是动态规划法(dynamic programming)(Gelfand MS and Roytberg MA, 1993)。这种算法将各种可能的拼接进行计分,从而得出最可能的基因结构。

除上述提到的算法之外,目前被应用于基因预测的算法还有:法则系统(rule-based system)(Guigo R *et al.*, 1992);语言学(linguistic)系统(Dong S and Searls DS, 1994);线性判别分析(Linear Discriminant Analysis, LDA)(Fickett JW and Tung CS, 1992);决策树(decision tree)(Salzberg S, 1995); spliced alignment 算法(Gelfand FS *et al.*, 1996);傅利叶分析(Fourier analysis)(Tiwari S *et al.*, 1997)等。

综合以上算法和人们对基因结构信号知识的基因预测程序已有不少。其中有的对编码顺序的预测准确率高达 90% 以上,并且在敏感性和特异性之间取得了很好的平衡。表 14.1 比较了一些程序的预测能力(Claverie JM, 1997)。表 14.2 列出了这些程序的参考文献和预测比较的参考文献以及程序的网上地址。

表 14.1　一些程序预测能力的比较(Claverie JM, 1997)

程　序	预测对象	敏感性 (% nucl.)	特异性 (% nucl.)	敏感性 (% exact exon)	特异性 (% exact exon)	丢失的外显子(%)	错误的外显子(%)
FGENEH	基因结构	83	93	73	78	15	11
GeneID	基因结构	69	77	42	46	28	24
GeneParser	基因结构	66	79	35	40	29	17
Genie	基因结构	87	88	69	70	10	15
GenLang	基因结构	72	79	51	52	21	21
GENSCAN	基因结构	93	93	78	81	9	5
HEXON	内部外显子	88	80	71	65	10	27
MORGAN	基因结构	83	79	58	51	14	–
MZEF	内部外显子	87	95	78	86	14	7
VEIL	基因结构	83	72	53	49	19	–

注释:
敏感性(% nucl.):% 实际编码序列被成功预测为编码序列
特异性(% nucl.):% 预测为编码的序列实际确实为编码序列
敏感性(% exact exon):% 实际的外显子被准确预测(包括拼接位点)
特异性(% exact exon):% 预测为外显子的序列与实际外显子准确符合
丢失的外显子(%):% 未能预测出的实际外显子
错误的外显子(%):% 预测为外显子的序列实际不是任何外显子的片段

表 14.2　一些程序及其预测比较的参考文献和网上地址

程　序	作　者	比较的参考文献	所用算法	网址
FGENEH	Solovyev *et al.*, 1995	Zhang, 1997	LDA	http://dot.imgen.bcm.tmc.edu:9331/gene-finder/gf.html
GeneID	Guigo *et al.*, 1992	Burset & Guigo, 1996	RB	http://geneid@darwin.bu.edu www.imim.es/GeneIdentification/Geneid/geneid_input.html
GeneParser	Snyder & Stormo, 1993	Burset & Guigo, 1996	DP	http://Beagle.colorado.edu/~eesnyder/GeneParser.html
Genie	Henderson *et al.*, 1997	Henderson *et al.*, 1997	GHMMs, DP	http://www-hgc.lbl.gov/inf/genie.html
GenLang	Dong & Searls, 1994	Burset & Guigo, 1996	LM	http://www.cbil.upenn.edu/~sdong/genlang_home.html
GENSCAN	Burge & Karlin, 1997	Burge & Karlin, 1997	HMMs, DP	http://genomic.stanford.edu/GENSCAN.W.html
HEXON	Solovyev *et al.*, 1994	Zhang, 1997	LDA, DP	http://dot.imgen.bcm.tmc.edu:9331/gene-finder/gf.html
MORGAN	—	—	DT	http://www.cs.jhu.edu/labs/compbio/morgan.html
MZEF	Zhang, 1997	Zhang, 1997		http://Clio.cshl.org/genefinder
VEIL	Krogh *et al.*, 1994a	Krogh *et al.*, 1994a	HMMs, DP	http://www.cs.jhu.edu/labs/compbio/veil.html

注释：DP：Dynamic Programming；DT：Decision Tree；GHMMs：Generalized Hidden Markov Models；HMMs：Hidden Markov Models；LDA：Linear Discriminant Analysis；LM：Linguitic Methods；RB：Rule Based systems

目前的各种算法还存在许多缺陷需要进一步改进，主要有以下两点：

(1)目前算法对基因中的非编码区和基因间序列不加任何区别，所以预测出的基因仍然是不完全的，对 5′和 3′非翻译区(UTR，untranslated region)的预测基本上还是空白。

(2)目前大多数算法都是基于已知基因顺序。如同源比较算法是完全依赖于已知的顺序，而象 HMM 之类的算法都需要对已知的基因结构信号进行学习或训练，由于训练所用的顺序毕竟是非常有限的，所以对那些与学习过的基因结构不太相似的基因，这些算法的预测效果就要大打折扣了。

要解决这两个问题，需要对基因结构进行更加深入的研究，寻找隐藏在基因的不同结构部分的内在统计规律。

14.3.4　基因功能预测

实验手段证实一个预测的新基因后，下一步要做的就是寻找这个基因的功能。生物信息学为此提供了一系列方法，使我们的研究能够有的放矢。

14.3.4.1　序列同源比较

序列同源比较往往是得到新基因后预测其功能的第一步。通过同源比较来预测基因功能是基于这样一个假设：如果基因 A 与基因 B 有相当的同源性，那么基因 A 可能具有类似

基因 B 的功能。利用同源比较算法，将待检测的新基因序列到 DNA 和蛋白质序列数据库中进行同源检索后，我们可以得到一系列与新基因同源性较高的基因或片段。这些基因和片段的已知的功能信息就为进一步研究新基因功能提供了具有相当参考价值的导向。

公共数据库

公共数据库是指利用因特网向公众开放检索的数据库，生物信息学领域的公共数据库主要包括基因序列数据库、蛋白质序列数据库、蛋白质高级结构数据库和各类专门数据库等。

基因序列数据库是公共数据库中数据量最大的一类数据库，包括了大量已知功能的基因序列和更多未知功能的基因组序列。下面介绍目前 3 个最大也是最主要的 DNA 序列数据库：GenBank，EMBL 和 DDBJ。

GenBank（Benson DA *et al.*，1998）是美国国立卫生研究所（National Institute of Health，NIH）的基因序列数据库。它的维护单位是美国国家生物技术信息中心（National Center for Biotechnology Information，NCBI）。EMBL 全称是 European Molecular Biology Laboratory（Stoesser G *et al.*，1998）。目前它的维护单位是位于英国的欧洲生物信息学研究所（European Bioinformatics Institute，EBI）。DDBJ 是 DNA Data Bank of Japan 的缩写（Tateno T and Gojobori T，1997），维护单位是日本信息生物学中心（Center for Information Biology，CIB）。NCBI，EBI 和 CIB 相互协作，共同维护这 3 大基因序列数据库。它们每天通过计算机网络互相交换数据，使得 3 个数据库能同时获得最新的数据。此外，他们每年召开两个年会讨论合作事宜（International DNA Data Banks Advisory Meeting 和 International DNA Data Bank Collaborative Meeting）。3 者关系见图 14.3。

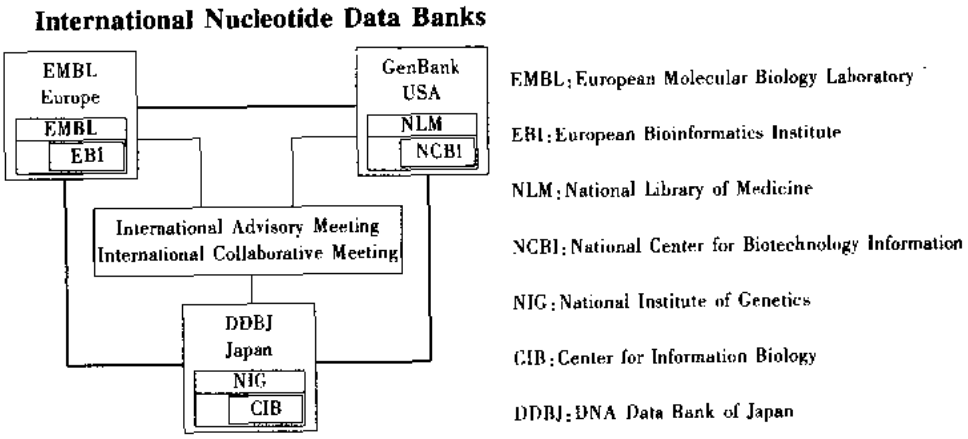

图 14.3　GenBank，EMBL 和 DDBJ 3 者的相互关系（引自 DDBJ）

截至 1998 年 1 月的 DDBJ release 32，GenBank/EMBL/DDBJ 已有 1 956 669 条 DNA 序列（图 14.1），碱基数共 1 300 950 613bp，其中 EST 占 72%。

GenBank/EMBL/DDBJ 都是大致按照生物分类来对序列数据进行分类的。以 GenBank 为例，7 个顶级类别分别为 Archaea（古核生物），Eubacteria（真细菌），Eukaryotae（真核生物），Viroids（类病毒），Viruses（病毒），Other（其他）和 Unclassified（未分类）。其中其他包括各种质粒，IS（插入顺序）等。Unclassified 包括一些未被列入以上分类的生物

的序列数据如 Prion(朊病毒)等。1997 年 11 月的一项统计表明,目前在 GenBank/EMBL/DDBJ 中已有超过 32 000 种生物的基因序列。具体数据见表 14.3。

表 14.3　GenBank/EMBL/DDBJ 中部分生物类群的数据量(数据取自 GenBank)

物种	1996	1997	增加(%)
All	22935	32402	41%
Viruses	1843	2089	13%
Eubacteria	4014	6125	53%
Archaea	230	381	66%
Eukaryota	15657	22493	44%

除按生物种类分类外,GenBank/EMBL/DDBJ 还包含有按照不同的序列数据类型划分的功能数据库,见表 14.4。

表 14.4　部分功能数据库的分类方式和采用此种方式的 DNA 序列数据库(Ouellette BFF and Boguski MS, 1997)

Functional Divisions	Used by Which Database?
EST Expressed sequence tags	DDBJ, EMBL, GenBank
STS Sequence tagged sites	DDBJ, EMBL, GenBank
GSS Genome survey sequences	DDBJ, EMBL, GenBank
HTG High throughput genomic sequences	DDBJ, EMBL, GenBank
PAT Patent sequences	DDBJ, EMBL, GenBank
CON Virtual contigs of segmented Sequences	DDBJ, EMBL, GenBank

下面介绍几个重要的功能数据库:

dbEST 是 EST 序列的数据库。EST 是从 cDNA 库中随机挑取克隆进行部分测序而得到的 DNA 序列。1997 年 10 月 3 日的 dbEST release 100 397 统计,库中有 1 260 382 条 EST,其中人类 EST823 957 条。EST 数据是近年来数量增长最快的一类 DNA 数据。

STS 的数据库是 dbSTS。STS 是 Sequence Tagged Site 的缩写,实际上是一对特定 PCR 引物扩增出来的 60 ~ 1 000bp 的 DNA 序列。具体地说,用这对引物在人基因组中进行 PCR 扩增,可以得到唯一的电泳条带,这个电泳条带所代表的 DNA 片段就称为一个 STS。STS 的数据包括引物序列和 PCR 条件。基于 PCR 反应的 STS 被广泛地用来构建基因组的物理图谱。

dbGSS 是一个新的功能数据库,1996 年 8 月才单独划分出来。GSS (Genome Survey Sequence)的概念类似 EST,只不过它是基因组的片段而不是 cDNA 的片段。GSS 的数据主要来自随机地对基因组片段进行一轮测序,以及外显子捕捉和 Alu PCR 等方法。

dbHTG (High Throughput Genomic Sequence, HTG)是另一个新划分出来的功能数据库。1996 年世界上几个大规模测序中心达成一个协议。这个协议要求各个中心尽可能快地使可用于同源比较的序列(长度 >2kbp 的序列)进入公共数据库,即使数据的整理还未最终完成(比如还未使错误率降到 10^{-4} 以下)。GenBank/EMBL/DDBJ 单独将 HTG 放

在 dbHTG 中,以区别于其他已完成最终整理的序列数据。

除此之外,GenBank/EMBL/DDBJ 还包括 DNA 和蛋白质的非冗余库(non - redundant, NR)。它将 DNA 和蛋白质数据中的冗余信息去除以便于用户进行更有效的同源比较。

数据库还提供同源检索的工具,例如 GenBank 提供一系列 BLAST(见附录 1.4.4)程序,EMBL 提供 FASTA(见附录 1.4.3)和 BLITZ(见附录 1.4.2)程序。DDBJ 提供 BLAST 和 FASTA。

这 3 个数据库同时提供了一系列获取数据的途径,包括 WWW、匿名 FTP、e-mail server 等。此外,GenBank 和 EMBL 还提供集成的检索工具,分别称为 Entrez Browser 和 SRS。它们将 DNA、蛋白质系列,蛋白质 3 级结构,Medline 摘要等相关信息联系在一起。通过它们,用户可以用文献名检索 DNA 或蛋白质序列,或者从基因序列出发检索与此序列有关的文献。GenBank、EMBL、DDBJ 以及 Entrez Browser 和 SRS 的网络地址见附录 1.4.1。

蛋白质序列数据库中的数据要比 DNA 序列库中的数据少得多,但一方面由于大量的蛋白质序列有非常完备的注释,可以为进一步研究新基因的功能提供更多的有用信息,另一方面,蛋白质在进化中的保守程度比 DNA 高,所以,在 DNA 序列库中找不到明显同源的基因序列,可能在蛋白质序列库中找到有功能参考价值的同源蛋白质顺序。下面介绍两大蛋白质序列库:PIR 和 Swiss-Prot。

PIR (Protein Information Resource) (Barker DG *et al.*, 1998)这个蛋白质数据库的前身是 Dayhoff MO 在 60 年代早期建立的 Atlas of Protein Sequence and Structure。从 1988 年开始,PIR 由 PIR-International 负责维护。成员包括美国 National Biomedical Research Foundation (NBRF), 德国 Martinsried Institute for Biochemistry (MIPS), 日本 Japan International Protein Information Database (JIPID)。

Swiss-Prot (Bairoch A and Apweiler R, 1998)是 EMBL 和瑞士日内瓦大学合作建立的一个蛋白质数据库,现在由 EBI 和日内瓦大学共同维护。至 1996 年 10 月,Swiss-Prot 拥有大约 60 000 条蛋白质序列,氨基酸共 21 000 000 个。

PIR 和 Swiss-Prot 都提供蛋白质系列数据的同源检索服务以及蛋白质序列的分类和注释。注释的内容包括蛋白质的功能、翻译后加工、结构域特征、二级结构、三级结构、同源性、疾病相关信息等。除了收集蛋白质序列外,它们还将 GenBank/EMBL/DDBJ 的 DNA 序列中的编码部分"翻译"成氨基酸顺序,作为蛋白质序列的补充。

PIR 和 Swiss-Prot 的网络服务包括 WWW,匿名 FTP,还提供数据库的 CD-ROM。PIR 和 Swiss-Prot 的网络地址见附录 1.4.1。

蛋白质高级结构数据库提供的基于结构的同源比较(structure - structure alignment),可以为研究新基因的高级结构和分子生物学机制提供有价值的信息。目前最著名的蛋白质高级结构数据库 PDB(Protein Data Bank)是生物大分子三级结构的数据库。包括了原子标记、文献引用、一级和二级结构信息,以及晶体结构和核磁共振的数据。

随着测定生物大分子三级结构技术的进步,PDB 中的数据量也在加速增长(表 14.5)。通过 PDB 进行同源比较而预测基因或蛋白的可能高级结构和功能,已经成为功

能预测的一个重要途径。

表 14.5　PDB 蛋白质数据库历年数据增加情况(数据取自 PDB)

年　份	存入量	释放量
1973	10	10
1974	7	2
1975	18	21
1976	47	27
1977	27	38
1978	26	26
1979	32	31
1980	20	30
1981	39	26
1982	59	47
1983	27	50
1984	36	29
1985	27	30
1986	28	29
1987	64	28
1988	129	79
1989	192	89
1990	306	164
1991	512	205
1992	635	226
1993	922	849
1994	1111	1392
1995	1221	1002
1996	1448	1224
1997	1848	1640

PDB 以 CD-ROM 的形式每 3 个月出版一次。同时也提供 WWW 和匿名 FTP 服务。PDB 的网络地址见附录 1.4.1。

除了上述介绍的六个重要的数据库外,世界各地还分布着数百个公共数据库,其中大部分是专门数据库,如蛋白质家族,Motif 数据库,生理反应数据库,还有专门针对某种生物的数据库如水稻数据库等。附录 1.4.1 中列出了一部分数据库的网址,以供参考。

公共数据库与因特网相连,为世界各地的科学家提供快速高效的服务,因而成为获取分子生物学数据的最佳媒介。然而随着各种数据库的不断增加和数据库数据量的近乎指数式的增长,种类繁多而内容庞杂的数据给数据库的维护和检索带来了困难,今后数据库

发展的重要趋势有以下几点：

（1）进一步提高已有数据库的质量，主要是提高时效性，使用户能随时获得最新的资料；提高数据的质量，降低数据冗余，增加数据说明，提供尽可能详尽的注释。

（2）进一步增加专门数据库，以满足不同用户对不同类型数据的需求，同时也提高检索效率。

（3）加强数据库的整合。将位于世界各地，由不同科学家维护的数据库的不同类型的信息，通过计算机网络有机地联系起来，使得用户在对多个数据库内容进行检索时，如同面对同一个数据库一样。这同时也免去了不同数据库数据格式转换给用户检索带来的困难。这方面的工作已经得到各方面科学家的重视，并作了有益的尝试（Karp PD，1995）。一些整合的数据库浏览系统已经开始在因特网上提供服务，比如前面提到的 Entrez Browser 和 SRS。

同源比较算法

将新发现的一段核苷酸序列或氨基酸序列同数据库中的序列进行同源性比较是进行数据库检索，预测新基因功能的重要手段。同源比较分为整体对齐（global alignmen）和局部对齐（local alignment）两大类。整体对齐对两个序列全长的相似性作出判断。而局部对齐则着眼于两个序列是否有局部序列的相似。由于数据库中许多基因的序列是不完整的，而分子生物学家对序列中的保守顺序比对非保守顺序更感兴趣。所以数据库检索的同源比较算法以局部对齐为主体。

目前最流行的可用于局部对齐的算法有 Smith-Waterman 算法、FASTA 和 BLAST。对它们的简要介绍见附录 1.4.2～1.4.4。

严格的动态规划算法，如 Smith－Waterman 算法计算两个序列的最大可能的相似性，可以处理碱基替换和间隔（gap，包括缺失与插入）等问题，因而具有最高的敏感性，可以找到相似性较差的同源序列。但这种算法计算量相当大，对于一对长度分别为 m 和 n 个元素的序列的 alignment 需要正比于 m × n 次（O（mn）次）比较。所以，用此算法进行大量的数据库检索，只有在超级计算机或大型并行计算机上才能实现。

FASTA 算法及其早期版本 FASTP 在速度上大大提高，而只牺牲了一点敏感性。在一些实际的检索比较中，FASTA 的结果与 Smith-Waterman 算法的结果的相关系数（r）达到 0.85～0.99（Brutlag DL *et al.*，1993），而 FASTA 的速度可达到 Smith-Waterman 算法的 20 倍，如果只运行前 3 步，即以 *initn*（见附录 1.4.3）为最终分值的话，FASTA 速度是 Smith-Waterman 的 60 倍（Pearson WR，1991）。

BLAST 算法（Altschul SF *et al.*，1990）在速度上则更进一步，可以达到 FASTA 的几倍到几十倍。但是由于 BLAST 不考虑间隔的问题，敏感性比 FASTA 更差。Gapped BLAST 和 PSI-BLAST（Altschul SF *et al.*，1997）是 BLAST 算法的改进，它不但可以处理间隔问题，大大提高敏感性，同时提高了程序的速度，达到原来 BLAST 算法速度的 3 倍。但是，BLAST 算法的一个重要优势在于可以在检索中估计比较的统计显著性（significance），这是 gapped BLAST 和 PSI-BLAST 所不能取代的。

实际进行同源比较时，可以在速度和敏感性上进行权衡，使用适当的算法。

14.3.4.2　同源比较的发展方向

用于将序列在序列数据库中进行同源比较的 3 种流行的算法：Smith-Waterman 算法，FASTA 和 BLAST 算法各自的优缺点前面已经作了介绍。这些算法虽然已经相当成熟，但也不是没有需要改进的地方。面对飞速增加的数据库数据，如何同时获得高敏感性和高速度仍然是一个课题。

同源比较算法中另一个需要继续发展的方面是同源比较算法中使用的计分矩阵（见附录 1.4.3）的完善，特别是间隔的计分方法的研究。研究证明，使用更好的计分矩阵能够使算法的敏感性显著提高（Pearson WR，1991）。

需要解决的另一个问题是目前数据库中部分数据的冗余度太高。特别是 EST 库，某些基因甚至有数千条 EST 与之对应。所以对数据库进行同源检索所得到的结果可能是一大堆无用的信息淹没了有用的信息。这个问题可以通过屏蔽掉检索序列中的重复顺序（Claverie JM and States DJ，1993）或清除数据库中冗余数据的方法（Grillo G *et al.*，1996）得到部分的解决。

核苷酸序列与蛋白质序列进行同源比较比单纯的 DNA 对 DNA 或蛋白质对蛋白质的比较更为复杂。而 BLAST 等算法都是简单地将 DNA 序列“翻译”成氨基酸序列后与蛋白质序列进行比较的，所以 DNA 序列中的内含子部分和各种原因造成的读框移动会使算法的敏感性大大地降低。为了解决这个问题，States DJ 和 Botstein D（1991）提出了一个解决内含子问题的新模型，Huang X 和 Zhang J（1996）进一步改进了这个模型，使之能够解决读框移动的问题。但是这个算法的计算量仍然很大，有待进一步改进。

14.3.4.3　寻找蛋白质家族保守顺序

通过同源检索，我们可能推测待检的新基因是某个蛋白质家族的新成员，下一步就是寻找新基因中包含的该蛋白质家族的保守序列，这样也就为进一步深入研究其功能作好了准备。

多序列同源比较，或称为多序列对齐（multiple-sequence alignment），是将多个序列进行同源比较以发现其共同的结构特征的方法，被广泛用来寻找基因家族或蛋白质家族中的保守部分。Feng-Doolittle 算法（Feng DF and Doolittle RF，1987）是较常用的多序列对齐算法。其他的新算法包括 HMM 方法（Sonnhammer ELL *et al.*，1997），Gibbs sampling（Lawrence CE *et al.*，1993）以及处理多结构域蛋白质家族的算法（Adams RM *et al.*，1996）。由于保守部分往往与家族成员的功能密切相关，所以通过这些方法建立蛋白质家族数据库，能够帮助科学家更好地认识基因的功能。最具代表性的蛋白质家族保守序列的数据库有 PRINTS（Attwood TK and Beck ME，1994；Attwood TK *et al.*，1998）、BLOCKS（Henikoff S and Henikoff JR，1994）、Pfam（Sonnhammer ELL *et al.*，1997）、Sbase（Murvai J *et al.*，1996）和 Prosite（Bairoch A *et al.*，1996）等。这些数据库可以帮助我们把新基因所属的蛋白质家族及其保守部分找出来，并提供这个家族其他成员的结构和功能信息。

14.3.4.4　蛋白质结构的预测

有时一个可能的新基因通过同源检索找不到任何同源基因。这种序列就称为“孤

儿"基因(orphan gene)。生物信息学也提供一些预测孤儿基因功能的方法。这就是通过基于结构的同源比较(structure-structure alignment)寻找结构同源的基因或直接预测其高级结构来推测其可能的功能。有许多蛋白质高级结构数据库提供结构同源比较的检索。如前面介绍的 PDB 和 FSSP(见附录 1.4.1)。另一方面,直接预测基因产物的高级结构的算法现在已经有不少,然而,由于蛋白质的折叠结构实在太复杂,使得计算最佳构象非常困难。如果结构生物学在这方面的研究能够有所突破,无疑将大大推动基因功能的预测。

14.3.5　分子进化的研究

通过上述种种方法我们可以预测出一个新基因的可能具有的功能。然而预测新基因只是生物信息学研究的一个方面,这门学科的根本目标是探究隐藏在生物数据后面的生物学知识。对于基因组研究来说,一个重要的研究方向就是分子序列的进化。通过比较不同生物基因组中各种结构成分的异同,可以大大加深我们对生物进化的认识。这方面的研究已逐步形成一个称为比较基因组学(comparative genomics)的新学科(见第 8 章)。从各种基因结构与成分的进化,密码子使用的进化,到进化树的构建,各种理论上和实验上的课题都等待生物信息学家的研究。

14.4　生物信息学的发展展望

作为计算机科学和数学应用于分子生物学而形成的交叉学科,生物信息学已经成为基因组研究中强有力的必不可少的研究手段。为了能够更好地服务于基因组研究,生物信息学在将来的发展中需要做以下几方面的努力(NCHGR, 1994):

(1)理论研究。任何学科的发展都离不开基础理论的研究,生物信息学也不例外。它对许多学科都提出了巨大的挑战。这些学科包括分子进化遗传学、群体遗传学、统计生物学、基因组学以及计算机科学和应用数学的相关学科。如果基础理论研究得不到应有的发展,生物信息学的发展将受到严重的阻碍。

(2)软件的重用和说明。现在虽然已经开发出大量的软件工具,但是大多数软件缺乏技术细节的描述,使得新软件编制时不能很好地利用已有的软件资源,不得不从头开始,造成各种软件都有自己的输入输出格式,相互之间互不通用。同时,大量软件的出现带来一个新问题,即生物学家面对数量众多的软件无从选择。这两个问题的解决需要对各种软件的功能特性和技术细节进行详尽的介绍,并进行比较。这样的话,新软件的编制者可以避免一些编程的重复劳动,甚至直接利用已有的程序模块(Fischman J, 1996),并且可以编制已有软件输出格式的接口,统一输入输出的格式,用户也可以方便地选择适合的软件。

(3)集成数据库。数据库整合的重要性在前面已有介绍,这里不再重复。

(4)生物数据的质量监控。监控已有的生物数据究竟具有多大的可信度,对于生物物理图谱的构建工作将有十分重大的意义。

(5) 加强生物学家和计算机科学家以及数学家之间的沟通。长期以来 3 类科学家都

是埋头于各自的研究领域，而不关心其他学科的发展和要求。这种状况在我国尤为突出。生物信息学的发展要求 3 者之间加强沟通。其意义不仅在于推动生物信息学自身的发展，而且将形成促进整个生物学发展的强大动力。

由于生物信息学在科学上和商业上都具有非同一般的重要性，各国政府机构和商业组织都纷纷投资生物信息学的研究。以欧洲为例，欧共体投资 9 000 000 德国马克（折合 5 000 000 美元）在英国剑桥建立了专门从事生物信息学研究的欧洲生物信息学研究所（EBI）。英国生物技术和生物科学研究会（Biotechnology and Biological Sciences Research Council，BBSRC）从 1994 年开始着手培训熟悉生物信息学的生物学家，并每年投入 1 000 000英镑资助生物信息学研究（Gavaghan H，1997）。德国科研部（BMBF）在 1993 年至 1996 年间投入 37 000 000 德国马克（合 20 000 000 美元）作为生物信息学的研究经费，并将再投资 30 000 000 德国马克给从事生物信息学研究的研究所（Strobl M，1997）。

在我国，生物信息学随着人类基因组研究的展开才刚刚起步，但已显露出蓬勃发展的势头。许多科研单位已经开始或准备开始从事这方面的研究工作。北京大学研究建立起一个 EMBL 的镜像数据库（即完整地将 EMBL 的数据库移植过来），并提供部分的检索服务（http://www.ipc.pku.edu.cn/mirror/mirror.html；http://www.ebi.pku.edu.cn）。在复旦大学遗传学研究所，为克隆新基因而建立的一整套生物信息系统也已初具规模。中科院上海生化所、生物物理所等单位在结构生物学和基因预测研究方面也有相当的基础。

生物信息学作为基因组研究的有力武器，被广泛地用来加快新基因的寻找过程，以达到将“有用”新基因抢先注册专利的目的。在这场世界范围内的竞争中，中国科学家以及科研资金投向的决策部门如何结合我国科研水平的现状、优势领域等客观情况将有限的投资投入以求获得最大可能的科学研究以及商业回报，是一个无法回避的新课题。

在克隆新基因的思路方面，我们觉得我国不应该照搬国外克隆新基因所用的方法，即我们在前面介绍的从 EST 测序入手的一整套克隆思路，而应该走生物信息学和定位克隆相结合的道路。具体地说就是一方面进行各种遗传疾病家系的采集，从家系分析入手，寻找致病基因在染色体上的位置，然后对这个区域进行测序，再利用生物信息学的手段预测候选基因和它的功能并用实验加以证实；另一方面直接从现有公共数据库中的 EST 出发，用生物信息学的方法寻找可能有研究价值的新基因，并用实验方法来研究证实。我们认为这种双管齐下克隆新基因的方法可能更适合我国人类基因组研究在财力、物力和研究人才资源等方面的客观条件。这主要反映在以下几点：

（1）中国作为一个发展中国家，在一些相对落后的地区还存在着一些遗传上与外界相对隔离的群体。这是研究遗传病家系的极好材料，一个隔离群体的少量样本就可以得到在开放群体中大量样本才能提供的致病基因的信息，而在美国等发达国家这样的隔离群体几乎已经找不到了。

（2）从家系以及遗传隔离的自然群体分析入手，直接针对致病基因进行探索。这样每克隆到一个新基因，都是一个功能明确的基因。而沙里淘金式的 EST 战略则需要在大量未知功能的新基因中寻找少数的“有用”基因。所以，以定位克隆策略为基础的基因克隆研究应该至少放在与 EST 策略同等的地位。

（3）公共数据库中的信息只要联上因特网就能访问，数据获取非常方便。EST 数据

的增长速度远远快于用实验手段研究基因的速度，所以很有可能有相当数量的 EST 所代表的"有用"新基因还没人研究。

所以与其与美国等发达国家拼资金拼技术，不如充分利用我国丰富的家系资源和公共数据库中的免费资源，将有限的资金用在具有明确科学、经济和社会效益的研究方向。

在生物信息系统的构建方面，应该避免重复投资。国家应当集中创建一两个具有一定规模的生物信息中心，建立面向全国的生物学数据库检索和数据分析系统。这个系统的建立可以分两步走。第 1 步我们要将国外公共数据库中的内容和相关软件收集和集中起来，提供检索和下载。第 2 步是将这些资源有机地组合，建立一个统一的生物信息平台。通过这个平台用户可以将各种格式的数据提交给设在生物信息中心的服务器，在服务器上进行一系列的检索和数据分析。用户不必关心各种数据库和软件的输入输出格式，只需一个简单的客户端软件甚至只需一个 WWW 浏览器就能完成全部工作。整个生物信息平台不仅是一个集成的数据库，而且是一个集成的软件工具(Fischman J, 1996)。

相信在 HGP 和即将开始的中国人基因组研究计划中，生物信息学将发挥越来越大的作用，并推动生物学进入一个全新的境界。

(刘晓明　张荣梅　罗泽伟)

参考文献

[1] Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merri CR, Wu A, Olde B, Moreno RF *et al*. 1991. Complementary DNA sequencing: expressed sequence tags and human genome project. Science, 252(5013):1651—1656

[2] Adams RM, Das S, Smith TF. 1996. Multiple domain protein diagnostic patterns. Protein Sci, 5(7):1240—1249

[3] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol, 215(3):403—410

[4] Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res, 25(17):3389—402

[5] Attwood TK, Beck ME, Flower DR, Scordis P, Selley JN. 1998. The PRINTS protein fingerprint database in its fifth year. Nucleic Acids Res, 26(1):304—308

[6] Attwood TK, Beck ME. 1994. PRINTS——a protein motif fingerprint database. Protein Eng, 7(7):841—848

[7] Bains W. 1996. Company strategies for using bioinformatics. Trends Biotechnol, 14(8):312—317

[8] Bairoch A, Bucher P, Hofmann K. 1996. The PROSITE database, its status in 1995. Nucleic Acids Res, 24(1):189—196

[9] Bairoch A, Apweiler R. 1998. The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1998. Nucleic Acids Res, 26(1):38—42

[10] Barker WC, Garavelli JS, Haft DH, Hunt LT, Marzec CR, Orcutt BC, Srinivasarao GY, Yeh LSL, Ledley RS, Mewes HW, Pfeiffer F, Tsugita A. 1998. The PIR-International Protein Sequence Database. Nucleic Acids Res, 26(1):27—32

[11] Beckmann JS, Brendel V, Trifonov EN. 1986. Intervening sequences exhibit distinct vocabulary. J Biomol Struct Dyn, 4(3):391—400

[12] Benson DA, Boguski MS, Lipman DJ, Ostell J, Ouellette BF. 1998. GenBank. Nucleic Acids Res, 26(1):1—7

[13] Boguski MS, Schuler GD. 1995. ESTablishing a human transcript map. Nat Genet, 10(4):369—371

[14] Brutlag DL, Dautricourt, J-P, Diaz R, Fier J, Moxon B, Stamm R. 1993. BLAZE™: An implementation of the Smith-Waterman sequence comparison algorithm on a massively parallel computer. Computers Chem, 17(2):203—207

[15] Bult CJ, White O, Olsen GJ, Zhou L, Fleischmann RD, Sutton GG, Blake JA, FitzGerald LM, Clayton RA, Gocayne JD, Kerlavage AR, Dougherty BA, Tomb JF, Adams MD, Reich CI, Overbeek R, Kirkness EF, Weinstock KG, Merrick JM, Glodek A, Scott JL, Geoghagen NSM, Venter JC. 1996. Complete genome sequence of the methanogenic archaeon, Methanococcus jannaschii. Science, 273(5278):1058—1073

[16] Burge C, Karlin S. 1997. Prediction of complete gene structures in human genomic DNA. J Mol Biol, 268(1):78—94

[17] Burset M, Guigo R. 1996. Evaluation of gene structure prediction programs. Genomics, 34(3):353—367

[18] Claverie JM. 1993. Database of ancient sequences. Nature, 364(6432):19—20

[19] Claverie JM. 1997. Computational methods for the identification of genes in vertebrate genomic sequences. Hum Mol Genet, 6(10):1735—1744

[20] Claverie JM, Bougueleret L. 1986. Heuristic informational analysis of sequences. Nucleic Acids Res, 14(1):179—196

[21] Claverie, J-M and States, DJ. 1993. Information enhancement methods for large scale sequence analysis. Computers Chem, 17(2):191—201

[22] Dong S, Searls DB. 1994. Gene structure prediction by linguistic methods. Genomics, 23(3):540—551

[23] Feng DF, Doolittle RF. 1987. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. J Mol Evol, 25(4):351—360

[24] Fickett JW, Tung CS. 1992. Assessment of protein coding measures. Nucleic Acids Res, 20(24):6441—6450

[25] Fischman J. 1996. Bioinformatics. Working the Web with a virtual lab and some Java. Science, 273(5275):591—593

[26] Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM *et al*. 1995. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science, 269(5223):496—512

[27] Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, Fleischmann RD, Bult CJ, Kerlavage AR, Sutton G, Kelley JM *et al*. 1995. The minimal gene complement of Mycoplasma genitalium. Science, 270(5235):397—403

[28] Fraser CM, Fleischmann RD. 1997. Strategies for whole microbial genome sequencing and analysis. Electrophoresis, 18(8):1207—1216

[29] Gershon D, Sobral BW, Horton B, Wickware P, Gavaghan H, Strobl M. 1997. Bioinformatics in a post-genomics age. Nature, 389(6649):417—422

[30] Gelfand MS, Mironov AA., Pevzner PA. 1996. Gene recognition via spliced sequence alignment. Proc Natl Acad Sci USA, 93(17):9061—9066

[31] Gelfand MS, Roytberg MA. 1993. Prediction of the exon-intron structure by a dynamic programming approach. Biosystems, 30(1—3):173—182

[32] Gershon D, Sobral BW, Horton B, Wickware P, Gavaghan H, Strobl M. 1997. Bioinformatics in a post-genomics age. Nature, 389(6649):417—422

[33] Goad WB, Kanehisa MI. 1982. Pattern recognition in nucleic acid sequences. I. A general method for finding local homologies and symmetries. Nucleic Acids Res, 10(1):247—263

[34] Gotoh O. 1982. An improved algorithm for matching biological sequences. J Mol Biol, 162(3):705—708

[35] Green P, Lipman D, Hillier L, Waterston R, States D, Claverie JM. 1993. Ancient conserved regions in new gene sequences and the protein databases. Science, 259(5102):1711—1716

[36] Grillo C, Attimonelli M, Liuni S, Pesole G. 1996. CLEANUP: a fast computer program for removing redundancies from nucleotide sequence databases. Comput Appl Biosci, 12(1):1—8

[37] Guigo R, Knudsen S, Drake N, Smith T. 1992. Prediction of gene structure. J Mol Biol, 226(1):141—157

[38] Henderson J, Salzberg S, Fasman KH. 1997. Finding genes in DNA with a Hidden Markov Model. J Comput Biol, 4(2):127—141

[39] Henikoff S, Henikoff JG. 1994. Protein family classification based on searching a database of blocks. Genomics, 19(1):97—107

[40] Huang, X. and Zhang, J. 1996. Methods for comparing a DNA sequence with a protein sequence. Computer Applications in the Biosciences, 12(6):497—506

[41] Hudson TJ, Stein LD, Gerety SS, Ma J, Castle AB, Silva J, Slonim DK, Baptista R, Kruglyak L, Xu SH *et al*. 1995. An STS-based map of the human genome. Science, 270(5244):1945—1954

[42] Karp, PD. (ed.) 1995. Proceedings of the 1995 Meeting on the Interconnection of Molecular Biology Database
http://www.ai.sri.com/~pkarp/mimdbd/95/abstracts.html

[43] Krogh A, Mian IS, Haussler D. 1994a. A hidden Markov model that finds genes in *E. coli* DNA. Nucleic Acids Res, 22(22):4768—4778

[44] Krogh A, Brown M, Mian IS, Sjolander K, Haussler D. 1994b. Hidden Markov models in computational biology. Applications to protein modeling. J Mol Biol, 235(5):1501—1531

[45] Lander ES, Waterman MS. 1988. Genomic mapping by fingerprinting random clones: a mathematical analysis. Genomics, 2(3):231—239

[46] Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC. 1993. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. Science, 262(5131):208—214

[47] Lipman DJ, Pearson WR. 1985. Rapid and sensitive protein similarity searches. Science, 227(4693):1435—1441

[48] Murvai J, Gabrielian A, Fabian P, Hatsagi Z, Degtyarenko K, Hegyi H, Pongor S. 1996. The SBASE protein domain library, Release 4.0: a collection of annotated protein sequence segments. Nucleic Acids Res, 24(1):210—213

[49] NCHGR (National Advisory Council for Human Genome Research). 1994. Report: NCHGR CESTEC directors' meeting on genome informatics
http://www.gdb.org/Dan/nchgr/report.html

[50] Needleman SB, Wunsch CD. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins J Mol Biol, 48(3):443—453

[51] NIH and DOE. 1990. In Understanding Our Genetic Inheritace. The US Human Genome Project: The First Five Years. FY 1991—1995. NIH Publication, No. 90—1590

[52] Okubo K, Hori N, Matoba R, Niiyama T, Fukushima A, Kojima Y, Matsubara K. 1992. Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression. Nat Genet, 2(3):173—179

[53] Ouellette BF, Boguski MS. 1997. Database divisions and homology search files: a guide for the perplexed. Genome Res, 7(10):952—955

[54] Parsons JD, Brenner S, Bishop MJ. 1992. Clustering cDNA sequences. Comput Appl Biosci, 8(5):461—466

[55] Parsons JD. 1995. Improved tools for DNA comparison and clustering. Comput Appl Biosci, 11(6):603—613

[56] Pearson WR. 1991. Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA. Genomics, 1991 Nov;11(3):635—650

[57] Pearson WR, Lipman DJ. 1988. Improved tools for biological sequence comparison. Proc Natl Acad Sci USA, 85(8):2444—2448

[58] Rowen L, Mahairas G, Hood L. 1997. Sequencing the human genome. Science, 278(5338):605—607

[59] Salzberg S. 1995. Locating protein coding regions in human DNA using a decision tree algorithm. J Comput Biol, 2(3):473—485

[60] Schuler GD, Boguski MS, Stewart EA, Stein LD, Gyapay G, Rice K, White RE, Rodriguez-Tome P, Aggarwal A, Bajorek E, Bentolila S, Birren BB, Butler A, Castle AB, Chiannilkulchai N, Chu A, Clee C, Cowles S, Day PJ, Dibling T, Drouot N, Dunham I, Duprat S, East C, Hudson TJ *et al*. 1996. A gene map of the human genome. Science, 274(5287):540—546

[61] Sellers, P. H. 1974. On the theory and computation of evolutionary distances. J Appl Math (siam), 26:787—793

[62] Smith TF, Waterman MS. 1981. Identification of common molecular subsequences. J Mol Biol, 147(1):195—197

[63] Solovyev VV, Salamov AA, Lawrence CB. 1994. Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames. Nucleic Acids Res, 22(24):5156—5163

[64] Solovyev VV, Salamov AA, Lawrence CB. 1995. Identification of human gene structure using linear discriminant functions and dynamic programming. In Proc. Third International Conference on Intelligent Systems for Molecular Biology. Ismb, 3:367—375

[65] Sonnhammer EL, Eddy SR, Durbin R. 1997. Pfam: a comprehensive database of protein domain families based on seed alignments. Proteins, 28(3):405—420

[66] States DJ, Botstein D. 1991. Molecular sequence accuracy and the analysis of protein coding regions. Proc Natl Acad Sci USA, 88(13):5518—5522

[67] Stoesser G, Moseley MA, Sleep J, McGowran M, Garcia-Pastor M, Sterk P. 1998. The EMBL nucleotide sequence database. Nucleic Acids Res, 26(1):8—15

[68] Snyder EE, Stormo GD. 1993. Identification of coding regions in genomic DNA sequences: an application of dynamic programming and neural networks. Nucleic Acids Res, 21(3):607—613

[69] Tateno Y, Gojobori T. 1997. DNA Data Bank of Japan in the age of information biology. Nucleic Acids Res, 25(1): 14—17

[70] Tiwari S, Ramachandran S, Bhattacharya A, Bhattacharya S, Ramaswamy R. 1997. Prediction of probable genes by Fourier analysis of genomic sequences. Comput Appl Biosci, 13(3):263—70

[71] Venter JC. 1993. Identification of new human receptor and transporter genes by high throughput cDNA (EST) sequencing. J Pharm Pharmacol 45 Suppl, 1:355—360

[72] Weinstein JN, Myers TG, O'Connor PM, Friend SH, Fornace AJ Jr, Kohn KW, Fojo T, Bates SE, Rubinstein LV, Anderson NL, Buolamwini JK, van Osdol WW, Monks AP, Scudiero DA, Sausville EA, Zaharevitz DW, Bunow B, Viswanadhan VN, Johnson GS, Wittes RE, Paull KD. 1997. An information-intensive approach to the molecular pharmacology of cancer. Science, 275(5298):343—349

[73] Xu Y, Mural R, Shah M, Uberbacher E. 1994. Recognizing exons in genomic sequence using GRAIL II. Genet Eng (NY), 16:241—253

[74] Zhang MQ. 1997. Identification of protein coding regions in the human genome by quadratic discriminant analysis. Proc Natl Acad Sci USA, 94(2):565—568