DOI:10.16262/j.cnki.1000-8217.1999.02.001

·学科进展·

生物信息学——重大科学意义与经济效益兼备的新学科

张春霆*

(天津大学生命科学与生物工程研究院,天津 300072)

[摘 要] 本文就生物信息学的定义,研究的基本问题和可能取得的突破进行了综述,内容包括分子生物学数据库、DNA序列分析、蛋白质结构预测和分子进化等;并着重指出:科学发展史表明,科学数据的巨大积累将导致重大的科学发现。开普勒三大定律和元素周期表的发现和氢原子量子理论的建立无不如此。有理由认为,20世纪末生物学数据的巨大积累也必将导致重大的生物学规律的发现。

[关键词] 生物信息学,人类基因组计划,DNA序列,蛋白质结构,基因与基因组学,蛋白质组学,分子进化,第2套生物学密码,科学发现

1 生物信息学概论

生物信息学,顾名思义是研究生物信息的学科,而生物信息最基本的表达形式乃是一维的分子排列顺序——序列,包括核酸序列和氨基酸序列。但是最基本的仍为DNA序列。历史上,第一个发表的完整的序列是Sanger等人在1956年测定的牛胰岛素B键的氨基酸序列,由51个氨基酸残基组成。Holley等人则于1965年发表了第一个核酸序列——酵母丙氨酸tRNA序列,由77个碱基组成。此后,所发表的序列量增长得非常缓慢,直到1977年,Sanger,Maxam及Gilbert发明了新的DNA测序技术,情况发生了急剧的变化。1990年人类基因组计划启动,主要目标是测出人的基因组中约 3×10^9 碱基对的全部DNA序列。与此同时,一些模式生物的基因组计划也包含在人类基因组计划之中。所谓模式生物系指为了更好地了解人的有关基因的结构与功能等而选取的生物,目的是为了参考和比较,其中包括大肠杆菌、线虫、酵母、拟南芥、果蝇和小鼠等。人类和模式生物基因组计划的快速进展,使得每年所发表的DNA序列量指数式地增长,呈现出"大爆炸"的局面。截止到1998年10月份,已发表了约20亿碱基对的DNA序列,每天增加的量达到 10^6—10^7 碱基对。每个碱基通常用1个字母来表示,若把这20亿个字母印刷成每本1 000页的书,约需要1 000本才能印完。生物信息数据量之大由此可见一斑。

对于这么大量的数据,如何管理、储存和调用是非常困难的,只能借助于计算机数据库管理技术。从80年代初开始分别建立起3个大的国际性的DNA数据库,那就是美国的GenBank,欧洲的EMBL和日本的DDBJ。用户可以通过光盘或其他存储媒介以及通过Internet获取这些序列,包括最新的序列。对于蛋白质的一级结构,即其氨基酸序列,也建立起相应的数据库,著名的有PIR和SWISS-PROT等。迄今为止已有约6 000种蛋白质的结构被阐明,记录这些详尽空间结构的数据库为美国的PDB。此外,还有美国国立图书馆生物信息中心(NCBI)的Entrez: Seqnences数据库,除了序列信息以外,还有文献信息。除这些大型主要数据库以外,还有相对较小的专门性数据库,如GenProEc,为大肠杆菌基因和蛋白质数据库。由于林林总总的数据库太多,总数已达300余种,为了方便研究人员,出现了专门收

* 中国科学院院士.

本文为在第87次香山会议上所作综述报告的基础上经修改而成.
本文于1998年11月3日收到.

集数据库的数据库，如 LIMB(Listing of Molecular Biology Database)。我国学者吴旻在中国医学科学院北京肿瘤研究所，在姚文萱、黄培瑛博士参与下，组织领导建立了"NEE-HOW"大型国际性的生物信息检索系统，其内容已达数百万页的文字。生物学数据库的巨大规模和复杂程度就可想而知了。

在人类的科学研究史中，这种科学数据的急速和巨大积累还从来没有出现过。如何组织管理和分析利用这些数据，就历史地摆在当代学者的面前。面对这种形势，一门新兴的交叉学科就应运而生了，那就是生物信息学。简单地说，生物信息学的主要任务就是组织和分析这些序列信息。下面给出一个更全面更完整的定义。1995 年 4 月，美国国立卫生研究院(NIH)发表了 1 篇报告，题目为：美国人类基因组计划，第一个五年(1991—1995)。在这篇报告中对生物信息学作了以下定义：生物信息学是包含了生物信息的获取、处理、储存、分发、分析和解释的所有方面的一门学科，它综合运用数学、计算机科学和生物学的各种工具进行研究，目的在于了解大量数据的生物学意义。这个定义对于生物信息学的主要任务、方法和目的作了全面的概括。可以看出，生物信息学是生物学、数学、计算机科学甚至物理、化学等多学科交叉的新学科。这门学科的提出也只是最近几年的事，可见其新了。

2 生物信息学中的基本问题

2.1 序列的比对(Alignment)

生物信息学中最基本的问题之一就是对两个序列进行比较，找出其相似性或不相似性，这往往构成了其他许多研究的基础。两个序列的比较利用比对(Alignment)技术。在比对分析中往往出现这样的情况，在 1 个或 2 个序列的某些位置上插入 1 个或几个空位，则两序列能更好地对齐，得到最大的相似性。问题是在什么位置插入多少空位方能得到最佳对齐呢？这是一个最优化问题。Needleman Wunsch 提出一种动态规划算法，比较成功地解决了以上问题，成为一些比对程序的核心，被广泛应用。但是由于受到计算机的速度和容量的限制，这一算法不能简单地应用到多序列(2 个以上)的比对研究中，还需要另想办法。这方面已有 Feng Doolittle 算法等可用。在很多情况下，两序列相似性并不是很高，但其中某些片段却非常相似。此时，用全序列比对则不易将这些片段找出来。在这种情况下，需要使用局部序列比对技术，其中主要的有 Smith Waterman 算法等。1 个序列测出以后，一般需要对数据库进行相似性检索，以找出与之相似的序列或对该序列新旧性作出判断。这样的检索当然是建立在以上介绍过的算法的基础之上的。国际上著名的相似性检索程序有 BLAST 和 FASTA 等。

2.2 基因识别与 DNA 序列分析

这里指的是蛋白质编码基因的识别，主要是编码 DNA 序列的识别。给定一 DNA 序列，要找出哪一段是编码区。一旦编码区确定，它所编码的蛋白质的氨基酸排列顺序(一级结构)就清楚了，这为预测它的结构与功能奠定了基础。所以，基因识别是生物信息学中的核心问题之一。对于原核生物，问题是要确定所有可能的开放读框(ORF)中，究竟哪一个或哪几个是真正的编码区。这个问题比较容易解决，现有的算法预测编码区的准确度已达到 90% 以上。虽然问题尚未完全解决，但远比真核生物中基因识别问题的情况要好。真核生物基因中编码区往往被许多内含子所打断，分成长短不等的外显子。因此，基因识别就包含了确定外显子与内含子的数目和它们之间的准确边界。这个问题尚未很好解决。现有的算法主要是基于编码区和非编码区(包括外显子和内含子)序列不同的统计学行为，包括碱基出现的概率和条件概率等。人工神经网络技术在这个领域已有广泛的应用。国际上广泛使用的基因识别算法及有关软件有 GRAIL, GenParser, GeneID 等。值得指出的是，作者提出的 DNA 序列的 Z 曲线理论，在基因识别问题中已崭露头角，在国际上占有一席之地。对基因组所有基因的分析逐渐形成了基因组信息学。

DNA 序列分析包含了基因识别问题，但决不止这一些，DNA 序列是一个取之不竭、用之不尽的理论研究源泉。DNA 序列分析最根本的任务是阐明包括词法、语法在内的遗传语言规律。

2.3 蛋白质结构预测

是生物信息学中另一核心问题，意义重大。蛋白质的 3 级(即空间)结构是由一级结构(即氨基酸序列)决定的 Anfinsen 原理，虽然已经确立了 30 余年，但如何由一级结构来预测其 3 级结构这一问题基本上仍未解决。然而这一问题的解决将最终阐明肽链的折叠规律，这将是生物学中一个里程碑式的事件，被称为破译第 2 套生物学密码。

蛋白质 3 级结构预测分为 2 条途径：演绎法和归纳法。前者是基于一个基本假设，即蛋白质的天然构象是一个能量最低状态。原则上说，只要求出

这一能量最低状态,蛋白质的3级结构就能准确预测出来。然而这一问题在数学上表现为极多变量、高度非线性函数的最优化问题,至今未能解决。后者,即归纳法是从现在已知其结构的大约6 000余种蛋白质中总结规律,进而预测其他未知蛋白质的结构。这一途径非常有希望。在3级结构预测中的同源模建和Threading方法都属于这一类。2级结构、2级结构组成和结构类预测也都应用归纳法,取得了相当的成功。不妨一提的是,作者在球蛋白质2级结构组成和结构类预测研究中取得了当前国际上最好的结果或最好的结果之一,得到较广泛的引用。我国学者王志新对蛋白质折叠类型总数的精确计算,引起国际上的广泛注意。

研究一个基因组中所编码的所有蛋白质的结构、功能及相互关系,称为蛋白质组学(Proteome)。它不仅是分子生物学实验问题,也是一个生物信息学问题。

2.4 分子进化

通过比较不同物种基因组中DNA或氨基酸序列的异同来研究生物的进化,称为分子进化,为当前生物信息学中的热门课题之一。可以在DNA或氨基酸序列的水平上比较研究。由于蛋白质的结构比序列更为保守,因而通过比较蛋白质空间结构上的异同来研究分子进化,往往能得到更多的信息,是当前较热的课题。美国学者Dayhoff是分子进化研究的开创人之一,她发表的一系列论文成为分子进化的经典性工作。例如,细胞色素C是一种含铁的,在真核细胞的生物氧化过程中起传递电子作用的蛋白质,100个以上种属的细胞色素C的氨基酸序列已经测定。用Dayhoff方法对它们进行比较而建立起的进化关系,与用考古学等其他方法所建立的大体上一致。令人高兴地指出,我国学者贺福初、吴祖泽通过序列比较发现了协同进化现象。

3 重大的科学意义与巨大的经济效益

在当今的生物信息学领域流传着一个新名词,叫着KDDM(Knowledge Discovery and Data Mining),意指通过分析数据来发现新的知识。从某种意义上来说整个生物信息学研究就是KDDM。生物信息学是当前知识创新的重要战场,世界各工业科技先进国家无不投入大量人力物力从公开发表或私自拥有的生物学数据中挖掘新知识。形象地说,这种挖掘的目的是为了一手"挖"诺贝尔奖(科学意义),另一手"挖"美元(经济效益)。遗传密码是从DNA序列中发现的第1套生物学密码,第2套,第3套……,还有待于发现。当然,分子生物学实验研究将在这些发现中起重要甚至关键性作用,但生物信息学的贡献是必不可少的。所挖掘出的新知识将可直接或间接地导致经济效益。例如,从公开发表的EST(表达顺序标签)库中,把一些cDNA片段利用某种算法建立起完整的基因cDNA,从而构建所谓2级数据库,甚至3级数据库,可直接获利。至于在分析基因所编码蛋白质空间结构与功能基础上所开展的药物设计,可能发现新的药物,有的药物之效益需用亿美元为单位来衡量。甚至1个数学方法,如前文提到的Smith-Waterman算法,也可以赚钱。已有3家美国公司生产能运行S-M算法的专用计算机,售价10万到数十万美元。当然,这里所说的经济效益基本上仍属于潜在的。指望今天对生物信息研究进行投资,明天就要求有所经济回报的做法,是不适当的。

4 百年一遇的科学发现良机

科学发现需要一定的条件,其中重要的一条是机遇。科学发展史表明:巨大的科学数据的积累将导致重大的科学发现。可以试举几例。17世纪初,德国科学家开普勒从其老师,丹麦天文学家第谷手中接过了其穷毕生精力所观测到的700多颗天体的运行数据,经过分析挖掘,终于发现了开普勒三大定律,为牛顿万有引力定律的发现开辟了道路。另一例子是元素周期表的发现。19世纪中叶,在已经发现了63种元素和数以万计化合物及有关数据的基础上,俄国化学家门捷列夫发现了元素周期表,为现代化学奠定了基础。第3个例子是20世纪初,氢原子和其他原子光谱学数据的大量积累,促使丹麦物理学家玻尔于1913年提出了氢原子的量子理论,为量子力学的建立奠定了基础。还可以举出若干例。历史的经验值得注意,温故而知新。有理由认为,20世纪末21世纪初生物学数据的巨大积累,也必将导致重大的科学发现。这些发现有的已崭露头角,如蛋白质中肽链的折叠规律(第2套生物学密码),DNA序列中有关基因表达调控的密码等;有的现在还想象不出来。但在是否有重大发现这一点上是不容置疑的。问题是,中国人在其中有何作为?

一个国家一个民族对某项科学发现作出贡献,除了要有科学发展的机遇外,还得有一定的社会发展为条件。在经历了上千年封建锁国和上百年列强侵略欺凌的贫穷落后的中国,指望我们的先人成为上述3项科学发现的首创者是不现实的,因为缺乏

必要的社会条件。而今天情况不一样了，经过 20 年的改革开放，中国走上了振兴的道路。党的科教兴国政策开创了中国科学的春天。生命科学发展的高峰与中国社会发展的高峰或发展曲线的"上跳沿"不期而遇。这是百年一遇的科学发现的良机。如果我们坐失良机，使我们的后人在读生物学教科书时，与我们这代人学习量子力学一样，只见外国人的名字，不见中国人的踪迹，那么，我们就会受到后人的指责，后人也不会原谅我们，我们生物信息学工作者在这个问题上也将成为愧对历史，愧对后人的一代。然而我想不会是这样！那么中国生物信息学工作者到底应该怎么办？在此生物信息学发展的关键时刻，我提出这个问题，谨供我国有关学者和科技领导人考虑。

BIOINFORMATICS——A NEW DISCIPLINE HAVING BOTH SIGNIFICANT SCIENTIFIC MEANING AND ECONOMIC BENEFIT

Zhang Chunting

(*Institute of Life Science and Biotechnology, Tianjin University, Tianjin* 300072)

Abstract In this paper the definition, the basic problems and the possible breakthrough of Bioinformatics, including biological databases, analysis of DNA sequences, and prediction of protein structures and molecular evolution, are reviewed. The paper is points out with emphasis that the history of science development demonstrates that the vast accumulation of scientific data will lead to great scientific discoveries. The discovery of Kepler three laws and the periodic table of chemical elements as well as the establishment of the quantum theory of hydrogen atom are typical examples. It is reasonable to infer that the vast accumulation of biological data by the end of this century would lead to the discovery of great biological laws.

Key words Bioinformatics, Human Genome Project, DNA sequence, protein structures, gene and genomics, proteome, molecular evolution

·资料·信息·

国家自然科学基金财政拨款增长与资助概况

自 1986 年以来，国家自然科学基金财政拨款逐年增长（图 1），平均年增长率为 21.83%。受国家自然科学基金资助项目的平均资助强度由 1986 的 2.77 万元/项增长至 1998 年的 12.3 万元/项（图 2），平均年增长率为 28.67%。国家自然科学基金资助项目的负责人年龄也明显年轻化（图 3），如 35 岁以下项目负责人由 1986 年的 1.28% 上升至 1998 年的 31.37%，45 岁以下项目负责人由 1986 年的 12.15% 上升至 1998 年的 59.10%。1986 年以来国家自然科学基金财政拨款增长与资助情况详见表 1。 （下转 106 页）

表 1 1986—1998 年国家自然科学基金财政拨款增长与资助情况一览表

年 度	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998
年度拨款（100 万元）	80.1	105.5	124	142	158.9	184.6	243 7	315.6	418.7	520.2	647.9	777.3	888.6
平均资助强度（万元）	2.77	3.08	3.34	3.28	3 44	3 70	5.10	6.00	7.60	8.70	10.00	11.40	12.30
35 岁以下负责人(%)	1 28	5.22	9.23	12.59	16.85	18 35	22.83	27.10	26 90	29.70	31 20	35 80	31 70
45 岁以下负责人(%)	12.15	13.96	20.71	21.27	25.91	27 30	31.80	37.20	39.60	45.30	48.80	55 10	59.10