* 学科发展*

生物信息学

郝柏林*

(理论物理研究所　北京　100080)

摘要　重点讨论了什么是生物信息学、生物信息学与生物实验的关系以及生物信息学提出的科学问题。

关 键 词　生物信息学

20 世纪的数理科学对无生命物质的结构和运动的研究, 从微观到宇观, 可谓既深且远。生命物质和生命现象正在成为 21 世纪数理科学研究的重要对象。生物数据量的迅猛增长, 既受益于数理科学和计算机科学所提供的方法与手段, 也呼唤着多种学科的共同努力。于是, 生物信息学应运而生。生物信息学是计算机和网络大发展、各种生物数据库迅猛增长的形势下如何组织数据、并从数据中提取生物学新知识的学问。它使生物学家如虎添翼, 而且是数理科学工作者进入生命研究领域的自然插入点之一。

1　什么是生物信息学

20 世纪后半叶分子生物学的长足进展, 把生命活动的物质基础追溯到核酸和蛋白质两大类生物大分子的序列。核酸是遗传信息的携带者, 是由 4 种单体(核苷酸) 聚合成的一维大分子。把核苷酸用字母代表, 遗传信息就编码在 4 种字母写成的课文中。各种蛋白质是生命活动的体现者, 它们是由 20 种氨基酸组成的一维大分子。大肠杆菌的 DNA 包含 460 多万个字母, 编码 4 300 种蛋白质。人的 23 对染色体有 30 多亿个字母, 编码数万种蛋白质。蛋白质要折叠成特定的立体形状, 才能发挥生物活性。核酸和蛋白质的字母序列以及蛋白质中每个原子的坐标, 构成生物数据的主要部分。研究生物大分子的相互作用、它们的演化、变异、生理和病理功能, 也产生着大量数据。

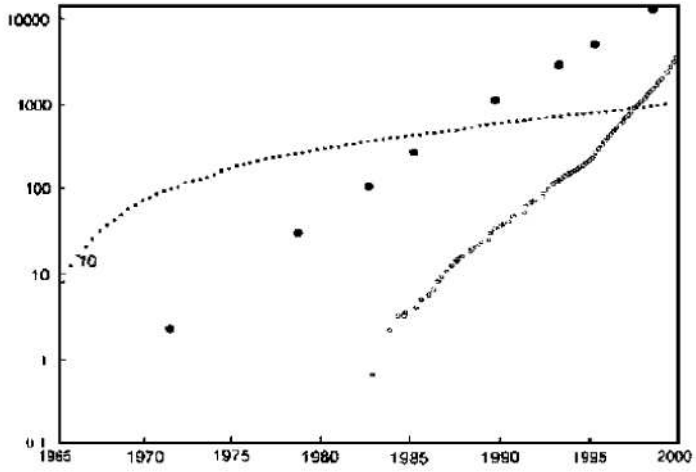
直到不久前, 人类科学实践产生数据量最大的领域, 是高能物理实验和脑神经活动成像, 两者都达到每年 10^{15}字节。现在生物数据的产生率已经达到同样水平, 而且很快要超出前两者。为了说明这种变化, 可以考察下图中画出的三条曲线。缓慢上升、似乎趋近饱和的曲线是 1966 年以来美国国家医学图书馆(National Library of Medicine) 所提供的在线检索服务 MEDLINE 所收录的文章中的一大类, 即“分子生物学和遗传学”论文数目的增长情况。MEDLINE 的选用

---

* 中国科学院院士

收稿日期: 2000 年 6 月 23 日

范围超出医学而囊括了几乎全部重要的生物学期刊。这条曲线大致反映了人类消化理解实验事实和数据, 使之上升为科学知识的过程。从 80 年代初迅速抬头的曲线是美国核酸序列数据库 Gen Bank 中核酸序列数目的增长情况。它清楚地表明, 数据增长越来越快, 传统的研究方式已经来不及迅速消化新数据, 以把后者及时提升为科学知识。



1980 年以来 Gen Bank 中核酸序列总碱基数目(空心圆点、圆点高度须再乘以 10^6), 1966 年以来 MEDLINE 收录分子生物学和遗传学论文数目(实心圆点, 原点高度须再乘以 10^3) 以及 1971 年以来 Intel 公司每个 CPU 芯片上三极管数目(大实心圆点, 圆点高度须再乘以 10^3) 的增长情况

所幸有一条跨越以上两条曲线、由 8 个数据点构成的第三条线, 它反映出 Intel 公司所生产的大规模集成电路单个 CPU 芯片上的三极管数目的增长速率。正是这一技术进步提供了解决问题的关键手段。目前一个典型的基因测序中心, 每年可以产生 10^{14}字节原始数据。数据的产生、搜集和分析, 都必须依靠计算机和网络, 都必须发展数据库、算法和程序。

应当特别指出, 核酸序列数据库中碱基总数翻番的周期, 现已不到 8 个月, 而且还将加速。这主要是受到人类和其它相关基因组计划的推动。截至 2000 年 6 月 1 日, 已经发表在公开数据库中的完全基因组, 包含 29 个细菌以及酵母、线虫、果蝇等真核生物; 人类的第 21 和 22 号染色体、疟疾原虫的两个染色体也已完成。生物学研究从单个基因和蛋白质, 跨入了高产出、多因子、整体性的"大工业"时代。没有计算机和互联网的长足进步, 这一切都是不可能的。

生物信息学与计算生物学或生物计算有密切关系, 但又不尽相同。目前归入生物信息学领域的大致有以下几个方面:

(1) 各种生物数据库的建立和管理。这是一切生物信息学工作的基础, 通常要有计算机科学背景的专业人员与生物学者密切合作。

(2) 数据库接口和检索工具的研制。数据库的内容来自万千生物学者的日积月累, 最终又为生物学所用, 但不能要求一般生物学工作者具有高深的计算机和网络训练水平。因此, 必须发展查询数据库和向库里提供数据的方便接口。这是专业人员才能胜任的工作, 通常在生物信息中心里进行。

(3) 研究新算法、发展方便适用的程序, 是生物信息学的日常任务。人类基因组计划的实施, 配合大规模的 DNA 自动测序, 对信息的采集和处理提出了空前的要求。从各种图谱的分析, 大量序列片段的拼接组装, 寻找基因和预测结构与功能, 到数据和研究结果的视像化, 无不需要高效率的算法和程序。

(4) 生物信息学最重要的任务, 是从海量数据中提取新知识。这首先是从 DNA 序列中识别编码蛋白质的基因及调控基因表达的各种信号。其次, 从基因组编码序列翻译出的蛋白质序列的数目急剧增加, 根本不可能用实验方法一一确定它们的结构和功能。而从已经积累的

数据和知识出发, 预测蛋白质的结构和功能, 成为常规的研究任务。

(5) DNA 芯片和微阵列的发展, 把一定组织或生物体内万千基因时空表达的研究提上日程。研究基因表达过程中的聚群关系, 从中提取调控网络和代谢途径的知识, 进而从整体上模拟细胞内的全部互相耦合的生化反应, 在亚细胞层次理解生命活动。只有掌握已有数据、发展崭新算法, 才能创造新的知识。这是生物信息学刚刚掀开的新篇章。

其实, 从生物数据库出发, 可以提出和回答许多用实验方法难以解决的问题。这里有发挥创造性的广阔天地。从我们自己近几年的科学实践, 可以举几个例子。许多原核生物完全基因组中, 有明确的缺失或稀少字串。例如至少 10 种细菌不喜欢 ctag 连在一起, 其它细菌则回避另一些短串。这可能反映了演化史中的某种共存关系。字串计数形象化所启发的数学问题, 可用组合学和语言学方法严格地解决。两种极端嗜热真细菌完全基因组的测序结果, 向公认的亲缘关系提出挑战。美国《科学》杂志就此对生命之树的评论, 在一年间从"动摇"变到"拔根"。从完全基因组出发建立的亲缘关系, 则可能有助于澄清问题。我们在研究"组分距离"的过程中注意到, 热球菌属 *Pyrococcus* 的两个完全基因组中重复次数最多的短串 *gttccaataagactaaaa*, 与国际核酸数据库中来自 7 万多物种的 600 多万条 DNA 序列、共 60 多亿碱基比较的结果, 只有同属的三个菌种 *P. abyssi*、*P. horikoshii* 和 *P. furiosus* 共享这段长度为 18 的片段。换言之, 这个短串有可能作为 DNA 水平上属的标记。这类观察引发出新的思考和课题。

2 生物信息学与生物实验

生物信息学的发展, 将造就一批不直接作实验而每天坐在计算机终端前的科学工作者。"生物学是实验科学"这类曾经完全正确、但已不十分符合当今科学实践的提法, 如果不正确理解, 就会在一定时期里挫伤有志于生物信息学的年轻人的积极性, 妨碍他们获得必要的经费支持和晋升。因而在此要专门讲一下生物信息学与生物实验的关系。

首先, 作为生物信息学基础和出发点的核酸与蛋白质序列都来自实验。即使是高产出的自动测序机, 也都基于以往的实验成就。同时, 这也表明以往的实验技术已经发展成现代化生产线。不重视从分析数据库获得新知识, 就是忽视大量以往的实验成果。

其次, 在全球每天产生以千万计数的碱基对核酸序列, 从中翻译出以十万计的可能的蛋白序列的时代, 已经根本不可能用实验办法去逐一确定它们的结构和功能。只有根据以往积累的数据和经验, 对大量新序列进行分析筛选, 才能突出应当由实验去决断的问题, 投入极其宝贵的人力物力。这一决策也得借助计算机完成。

第三, 越来越多的物种的基因组将被基本上完全地测定。那种倾毕生精力研究一个基因、一条代谢途径、一种生理周期的时代已经过去。还会有学者这么做, 但他们将只代表一种研究风格, 而不再是学术主流。人们正在阐明细胞内的全部互相耦合的调控网络和代谢网络, 细胞间的全部信号传导过程, 从受精卵到成体的全部生理和病理的基因表达的变化, 等等。这一切都超出手工分析的可能性。

因发明了一种 DNA 快速测序方法而同 F. Sanger 分享 1980 年诺贝尔化学奖的 W. Gilbert 于 1991 年在英国《自然》杂志撰写短文, 针对生物学的研究范式的变化指出, "正在兴起的新的范式在于, 所有的'基因'将被知晓(在可用电子方式从数据库里读取的意义上), 今后生物学研究项目的起点将是理论的。一位科学家将从理论猜测开始, 然后才转向实验去继续或检验该假设。"这一观点正在被越来越多的生物学工作者所认同。

从根本上说, 实验始终起着决定作用。然而, 这并不表明事事取决于实验。许多标准的实验, 已经成为半工业化的日常手续。只有那些有深刻思想的、精心设计的、决定性的新实验, 才同过去一样, 从根本上推动科学发展。回顾物理学在 19 世纪曾是实验科学, 20 世纪上半叶发展成理论和实验密切结合的科学, 20 世纪下半叶成为鼎立在实验、理论和计算三足之上的成熟的发达学科。生物也是物。生物学的发展也会从物理学得到启示。

3　生物信息学提出的科学问题

人类基因组计划将提前完成。全球的众多测序机构, 包括我国建立的可观的测序和处理能力, 将继续运作, 测定越来越多物种的基因组。蛋白质结构的测定, 虽然速度较慢, 但仍将持续增长。比较基因组学和蛋白质组学的研究将主要靠生物信息学方法进行。

我们不仅要注意人类基因组, 还应考虑与农牧业生产和人民健康有关的其它基因组测序与分析。例如, 最近美国孟山都公司宣布完成了水稻基因组的初稿, 对以日本为主的国际水稻基因组计划形成压力。我们如不大力加强自己的水稻基因组工作, 现有差距会继续拉大。其它农作物如大麦、小麦、燕麦、甘蔗、玉米、棉花、大豆、油菜、马铃薯等的基因组计划, 家畜和家禽如鸡、火鸡、猪、绵羊、山羊、马、牛、水牛、家兔、狗、猫的基因组计划正在许多国家进行。全世界从事马基因研究的实验室就有 25 个。关于人类基因是否可以申请专利的争论, 不会扩展到农作物和家畜。我们如不早提对策, 必有吃亏之时。希望农业部门能重视这个问题。

我们不仅要注意核酸序列中的基因部分, 还要发掘隐藏在大量非编码序列中的基因表达的时空调控信息。大量多肽和小蛋白的作用, 目前还不能进入基因组计划, 应当靠生物信息学方法进行预先研究。

除了前面已经提及的数据库和网络组织管理、生物符号序列的分析比较、基因和蛋白质结构与功能的预测等算法及程序设计问题, 生物信息学还向数理科学和计算科学提出不少深刻的研究课题。我们简要地列举几项基本上还没有被写进书籍的“非标准”问题:

(1) 寻求 NP 完备问题的近似解法。许多序列分析比较和分子演化问题导致艰难的识别或优化问题, 其中很多属于 NP 完备范畴, 任何严格求解的方法都会轻易超出现有和未来计算机的能力。放宽数学要求, 把“最优”换成“足够优”, 把“整体优化”换成“局部优化”, 不仅可以大为开拓可解问题的范围, 而且可能更接近自然界的实际情况。

(2) 面对海量的生物数据, 概率统计方法是基本功。事实上, 频度和关联分析、马可夫链和隐马可夫链、神经网络、贝叶斯统计等, 始终在生物信息学中广为应用。实际的生物序列, 无论 DNA 还是蛋白质, 当然都不是随机的。然而, 如果刻划的角度不妥, 所提取的许多特征量又离开随机序列不远。这表明统计方法不足以充分放大 DNA 序列与随机序列之间以及 DNA 序列之间的差别。必须寻求越出单纯统计方法的新途径。特殊地说, 有没有比随机序列更合适的“参考序列”? 目前对序列分析结果的统计评估, 总是以组分相同的随机序列作为参考。字的组合学中有定理表明, 有限字母集合上的足够长的符号序列具有不可避免的规则性。生物学符号序列是否足够长? 是否应当以具有不可避免的规律性的序列作为参照?

(3) 生物遗传语言和人类自然语言有许多相似之处, 例如多义性、冗余性、容错或纠错性、存在多种方言和个体差异、有长程关联、有某种语法框架但不能完全“生成”等等。同时, 它们又有深刻差别, 例如标点和间隔的不同、两种或多种语言的相互作用、重复序列的数目和功能不同等。经过一定程度的抽象后, 语言学(language 而不是 philology) 的方法应发挥更大作用。

这包括发展统计语言学和代数语言学结合的理论框架,对模糊语言和随机语法的研究等。形式语法很容易推广成模糊语法。对于 N. Chomsky 的串行生成语法,早有文章实现。对于平行生成的 Lindenmayer 系统,最近我组的一位博士后完成了模糊推广。然而,只有能进一步对模糊程度作定量刻划,才有望在生物学中有更多应用。隐马可夫链模型相当随机正规语法,更复杂的层次仍有待开发。广而言之,无论自然语言或遗传语言,都是基于离散的排列组合系统。组合学方法应能在建立生命现象的理论方面发挥更多作用。

(4) 利用 DNA 芯片和微阵列技术,可以获得一定组织或细胞内在生理或病理条件下全部基因表达过程的原始数据,从中重构基因调控网络、代谢网络、信号传导网络和免疫网络,要求发展新的算法。细胞内全部生化反应的模拟,必然要突破均匀分布假设而逐步计入实际的亚细胞结构。原胞自动机模型可能是使用偏微分方程之前的切实可行的步骤。

(5) 许多生物过程的模拟,应当引用有确定论骨架的随机微分方程,而目前这样做的很少。使用较多的方程,属于由终值分布确定的类型(郎之万方程和相应的福克-普朗克方程)。看来,随机微分方程应当有新的提法。

为了更详细地了解生物信息学的内容和发展状况,可以参阅参考文献列举的一些书籍,还可以经常访问重要的生物信息中心的网页,如北京大学生物信息中心的网页 http://www.cbi.pku.edu.cn/,通过此网页可以链接到许多重要的国际生物信息中心。

参考文献

1 Trends Guide to Bioinformatics, 1998.

2 Science, 1999, 284: 1742.

3 B.-L. Hao, H.-C. Lee, S.-Y. Zhang. Fractals related to long DNA sequences and complete genomes. Chaos, Solitons and Fractals, 2000, 11: 825– 836.

4 B.-L. Hao. Fractals from genomes: exact solutions of a biology-inspired problem. Physica, 2000, A282: 225–246.

5 Science, 1998, 280: 672; 1999, 284: 1305.

6 Walter Gilbert. Towards a paradigm shift in biology. Nature, 1990, 349: 99.

7 郝柏林. 建议尽快组建国家级生物医学信息中心. 中国科学院院刊, 2000, (2): 133– 134.

8 Andreas D. Baxevanis, B. F. Francis Ouellette (eds.), Bioinformatics. A Practical Guide to the Analysis of Genes and Proteins, Wiley-Interscience, 1998. xiv+ 370. (汉译本将由清华大学出版社出版)

9 T.K. Attwood, D.J. Parry-Smith. Introduction to Bioinformatics, AWL Press, 1999, xx+ 218. (汉译本将由北京大学出版社出版)

10 S. Misener, S.A. Krawetz (eds.), Bioinformatics. Methods and Protocols, Methods in Molecular Biology, Humana Press, 2000, 132: xi+ 500.

11 郝柏林, 张淑誉. 生物信息学手册, 上海科学技术出版社, 2000(即将出版).