生物信息学浅说

名家讲演录 续编

Mingjia jiangyanlu xubian

名家讲演录续编

生物信息学浅说

郝柏林　著

上海科技教育出版社

名家讲演录续编

生物信息学浅说

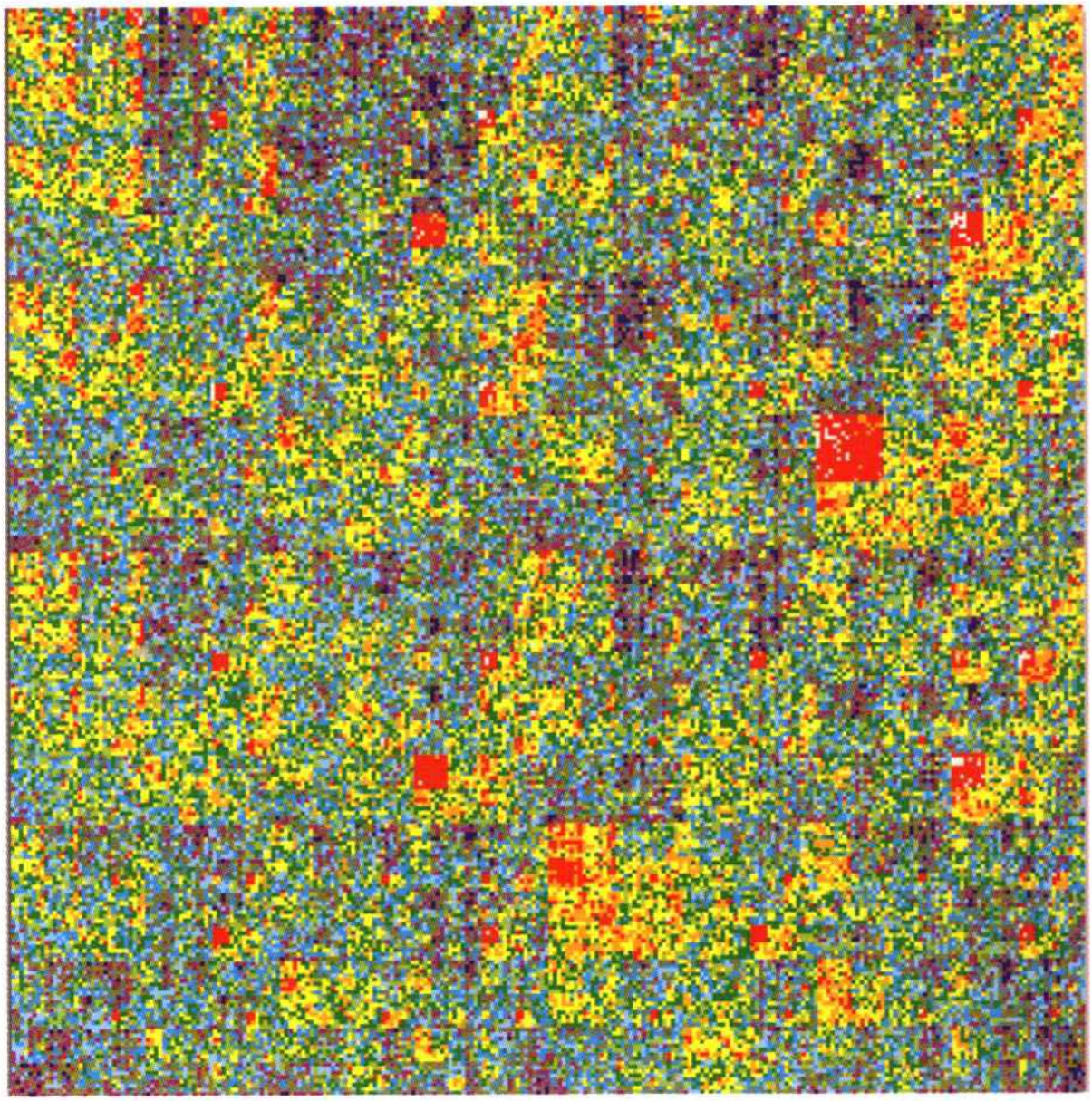
郝柏林　著

策　　划　卞毓麟
责任编辑　卞毓麟
装帧设计　汤世梁

出版　上海科技教育出版社
　　(上海冠生园路 393 号　邮政编码 200235)
发行　上海科技教育出版社
经销　各地新华书店
印刷　常熟市华通印刷有限公司
开本　850×1168　1/64
印张　1.5
插页　2
字数　29 000
印次　2003 年 3 月第 1 版　2003 年 3 月第 1 次印刷
印数　1－3000
书号　ISBN 7－5428－2689－1/N·432
定价　4.50 元

作者简介

郝柏林，男，1934 年 6 月生，中国科学院院士，第三世界科学院院士。中国科学院理论物理研究所研究员，中国博士后基金会副理事长。1959 年毕业于乌克兰哈尔科夫国立大学。主要研究领域为理论物理、计算物理、非线性科学和理论生命科学。曾任中国科学院物理研究所副所长、理论物理研究所所长，《物理学报》副主编、《中国物理快报》主编，第 19 届国际统计物理大会主席等。屡获中国科学院自然科学奖一等奖、国家自然科学奖二等奖等多种奖励。已出版英文专著《实用符号动力学与混沌》、中文《生物信息学手册》等图书 11 种，主编英文《混沌的方向》丛书和中文《非线性科学丛书》，发表学术论文 130 余篇。已培养多名博士和硕士，并完成了大量学术组织和科学普及工作。



大肠杆菌完全基因组中长度为 8 的短串出现的频度，详见本书正文 68 ~ 71 页

推荐读物

[1] 郝柏林,刘寄星主编. 理论物理与生命科学. 上海:上海科学技术出版社,1997.

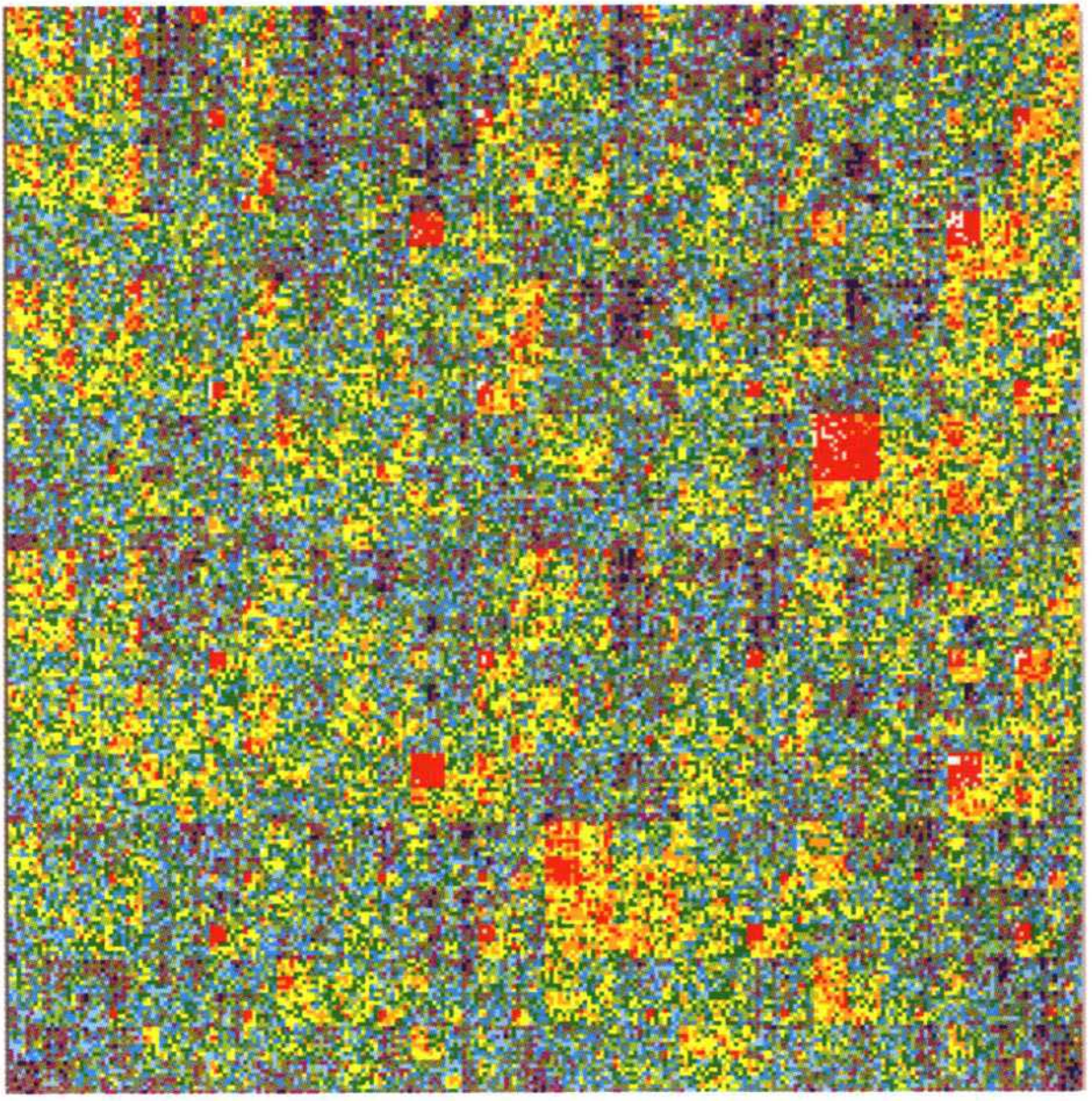
[2] 邹承鲁. 生物学在召唤. 上海:上海科技教育出版社,1999.

[3] 杨焕明,深蓝,张秀清,汪建,刘斯奇,于军. 生命大解密:人类基因组计划. 北京:中国青年出版社,2000.

[4] 郝柏林,张淑誉. 生物信息学手册(第二版). 上海:上海科学技术出版社,2002.

名家讲演录

科学技术与可持续发展	周光召著
跨世纪科学技术发展趋势概述	朱光亚著
超越疑古　走出迷茫	宋　健著
科学的历史经验与未来	路甬祥著
飞速发展中的现代科技	徐匡迪著
中国传统文化里的科学方法	席泽宗著
世纪之交话天文	王绶琯著
生物学在召唤	邹承鲁著
教计算机认字	吴佑寿著
诱人的治癌之道	汤钊猷著
基因和转基因动物	曾溢滔著
探索脑的奥秘	杨雄里著
人类起源之我见	贾兰坡著
人类的过去、现在和未来	吴汝康著
清除邪教再生的土壤	何祚庥著
生物多样性的启迪	孙儒泳著
甲骨百年话沧桑	李学勤著
地理与对称	叶大年著
土壤动物的功绩	尹文英著
造福人类的遗传学	赵寿元著



大肠杆菌完全基因组中长度为 8 的短串出现的频度，详见本书正文 68 ~ 71 页

目 录

序言

2000年6月10日，我在中国科学院第10次院士大会组织的公众学术报告会上作了题为《生物信息学》的演讲。卞毓麟先生建议收入他策划的《名家演讲录》小丛书中，我欣然同意。命笔时感到近两年来的科学实践又给人不少启发。于是离开旧稿，重新写来。读者会发现，这本小册子与发表在《中国科学院院刊》2000年第4期上的讲稿，在文字和内容上已多有不同。能够放下日常的公式推导和

程序设计，静心思考所作学问的前因后果，是写作的收获。1999 年底以来，同北京大学顾孝诚教授联合组织的双周研讨班，2001 年在中国科学院基因组学和生物信息学研究中心与浙江大学共同组织的生物信息学研究生班的讲课，以及我国籼稻基因组框架图完成后寻找基因的实战要求，促使我不断“上网”和学习。特别是与许多青年朋友的切磋砥砺，使我受益匪浅。谨在此一并致谢。

郝柏林　2002 年 5 月 6 日
复旦大学理论生命科学研究中心
中国科学院理论物理研究所

恩格斯在1885年论及各门科学使用数学的情况时，曾经指出，数学的应用“在生物学中=0”。[①] 诚然，人类作为大自然的产物和一部分，从来离不开自己生存环境中的万物。人们给各种鸟兽鱼虫、草木蔬果起名字，描述它们的性状和对人类的利与害。最早的生物学是纯粹的描述科学，其核心就是生物分类学。分类基于物种之间的差异，却同时也使人类逐步认识到各种生物的共同点。动

---

① 恩格斯. 自然辩证法. 于光远等译编. 北京：人民出版社，1984. 172

物和植物都具有“生命”,就是一个伟大的发现,但生命的起源和本质,可能是永恒的研究课题。人们早就知道“种瓜得瓜,种豆得豆”,各种生物的性状都世代相传,但并不明白其原因和机理。生物是物,化学和物理学的发展为生物学的进步提供了工具和思想。

光学显微镜的发明,导致了生物细胞和微生物的发现。电子显微镜和其他显微技术的进步,使人们看到细胞内部的许多构造和活动。从研究原子、分子到由小分子聚合成的大分子,人们终于认识到一切生命活动都基于核酸和蛋白质这两大类生物大分子。X射线衍射法提供了确定晶体结构的手段,许多蛋白质、蛋白质和核酸的复合体,以及蛋白质和蛋白质的复合体,被成功地结晶和分析。生物学知识的主体,从宏观性状的描述集合,转向分子和结构数据的积累。今天,生物学

已经成为人类科学活动中产生数据最多的领域之一。

目前在数据量上可以与生物学相比拟的，只有高能物理实验和脑神经活动的三维成像，两者都达到每年产生 10^{15} 字节的水平。生物学数据的年产量也在同一水平。10^{15} 这样的数字是怎样一个概念呢？自从我们所处的这部分宇宙从“大爆炸”产生以来，大约过去了 10^{10} 年。即使换算成秒，不过是 10^{17} 秒。不难估算，自从地球上有人类以来，所有的人所讲过的总“字”数，不会超过 10^{20}。对付这样的“大数”，必须使用人类自学会用火以来的最伟大的发明，那就是电子计算机。

我们说电子计算机是人类有史以来最伟大的发明，是因为它代表了人类一切活动中最迅速的进步。从 1943 年研制的每秒钟可以完成 330 次乘法运算的 ENIAC 计算机，到

前两年我国自行研制的每秒完成4000亿次浮点运算的曙光3000计算机(它还不是国内和国际上最快的计算机),计算速度提高了10^9倍。试问还有哪一个领域,有如此迅猛的进步?从步行到乘协和喷气式飞机,速度提高不超过1000倍!事实上,海量生物数据也是在计算机协助下产生的。人类社会的全部生产和科学活动,包括生物学和生物技术,都注定要靠电子计算机发展前进。

然而,从科学发展和世界各国的人力物力投入看,高能物理学等领域在近期内不会有戏剧性的突变,生物学很快就要在数据量上独占鳌头。生物学数据主要是核酸序列、蛋白质序列和各种生物大分子的结构数据;日积月累的大量科学文献也是一种数据。这海量数据的产生和消化,都离不开计算技术和互联网络的飞速发展。看来,无论感情上

是否愿意和思想上有无准备，生物学者们都必须学会在数据的汪洋大海中遨游。从现在起，不出十年，许多临床医师和农林牧业工作者也要同生物数据库和生物软件打交道。生物学和生物技术的发展正在成为跨学科的事业。生物信息学就在这样的历史背景下应运而生。

一、什么是生物信息学

“生物信息学”是20世纪最后10年中才出现的一个新词，是英文 bioinformatics 的直接对应。这个名词容易引起一些望文生义的误解。生物如何在彼此之间并且与外界交换和处理信息，是人们研究多年的传统领域，也有学者称之为生物信息学。不过，只要翻开近几年出版的数十种生物信息学书籍和期刊，就可以看出目前通用的“生物信息学”一词的涵义与传统的理解颇为不同。现代生物