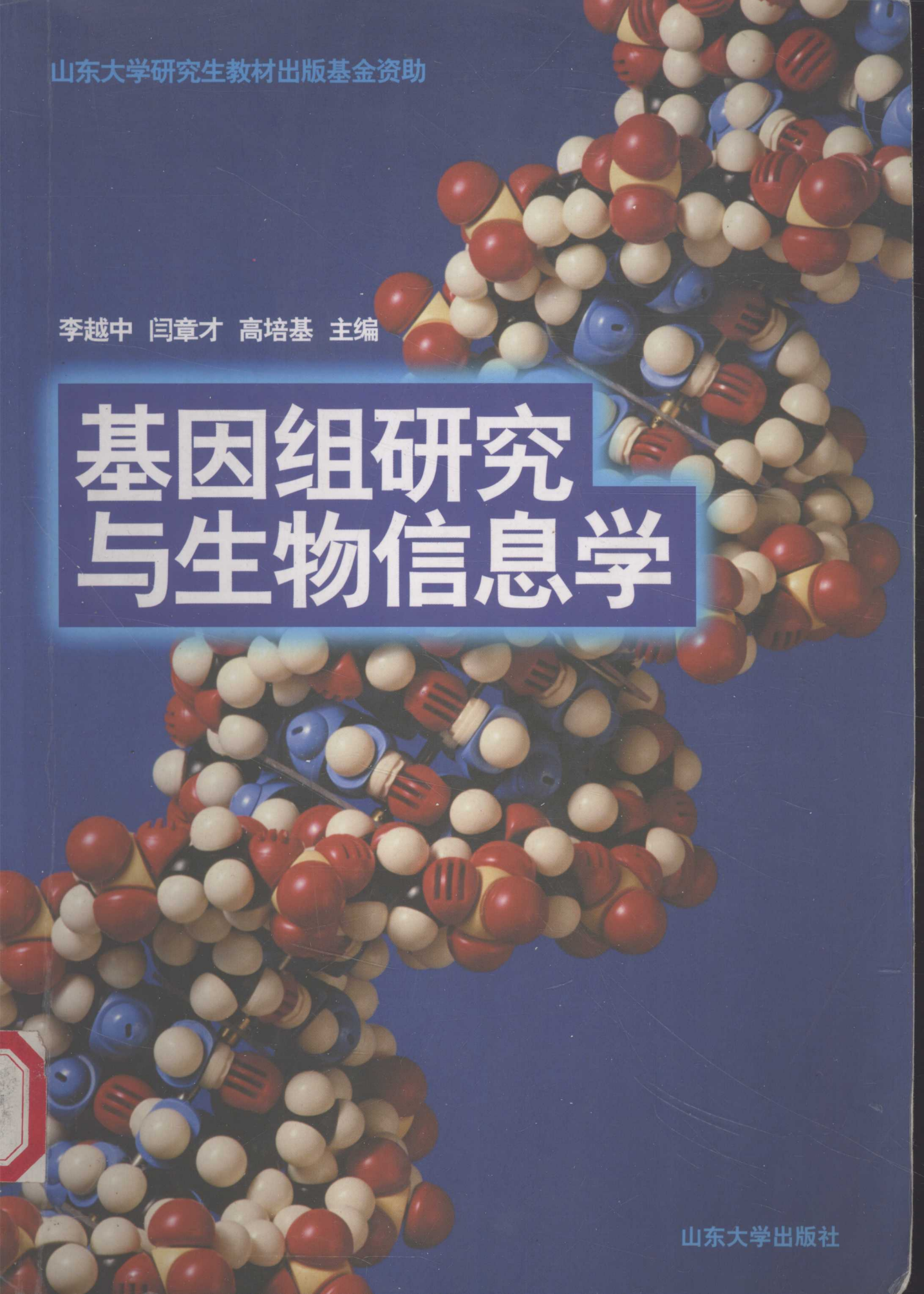
山东大学研究生教材出版基金资助

李越中 闫章才 高培基 主编

# 基因组研究与生物信息学



山东大学出版社

山东大学研究生教材出版基金资助

# 基因组研究与生物信息学

李越中　闫章才　高培基　主编

山东大学出版社

**图书在版编目(CIP)数据**

基因组研究与生物信息学/李越中,闫章才,高培基主编.—济南:山东大学出版社,2001.11(2003.1重印)

ISBN 7-5607-2378-0

Ⅰ.基...

Ⅱ.①李... ②闫... ③高...

Ⅲ.①基因组—研究生—教材②生物信息论—研究生—教材

Ⅳ.①Q343.2②Q811.4

中国版本图书馆CIP数据核字(2001)第083488号

山东大学出版社出版发行

(山东省济南市山大南路27号 邮政编码:250100)

山东省新华书店经销

安丘市百花印刷包装有限责任公司印刷

787×1092毫米 1/16 20.25印张 468千字

2001年11月第1版 2003年1月第2次印刷

印数:1201—2200册

定价:40.00元

**主　编**　李越中　闫章才　高培基

**参加编著人员**(按姓氏音序排列)

高培基　胡　玮　李越中

邱智军　王禄山　闫章才

# 前　言

揭开生命的奥秘是人类长期追求的目标.

生命是什么？生命是如何起源、繁衍、进化的？生命又是如何调控其复杂、多样、不可思议的功能的？令人惊叹的生物世界的多样性是由仅约20种氨基酸组成的蛋白质结构多样性及其生物学功能的多样性实现的.而蛋白质的多样性则由四种核苷酸,A,T,C,G的线性排列顺序的多样性所控制.这种可以通过繁殖而遗传的多样性储存在各种生物细胞的DNA中.纵观现代生物学的研究史,生命的遗传物质一直是我们探索生命奥秘的兴奋点.从20世纪40年代证明DNA是生命性状特征延续的遗传载体;50年代发现DNA的反向双螺旋结构;60年代经典遗传学的发展和发现DNA的遗传密码;70年代发现生命的中心法则;80年代分子生物学——即DNA的遗传操作;到90年代全面开展基因组DNA全序列的测定.我们可以看到,DNA——遗传本质的研究始终贯穿了生命科学研究的主干线.

长期以来,生物学一直是一门实验科学.科学家们努力地企图使生物学的规律能够用数字来表征,用公式来推导,由理论假设来指引实验的进行.20世纪90年代以来,全球生物科学研究由于人类基因组计划的实施而发生了一场真正的革命.诞生了基于全基因组水平的生物学实验.有关基因组的研究已经成为目前最为热门的科学话题.

各领域科学技术的进步极大地促进了生命科学的发展.同时,也产生了大量的生物学数据.如何从这浩如烟海的数据中发掘生命的信息,已经成为生物科学未来研究的主攻方向.并由此诞生了研究生命的新的学科——生物信息学.这是不同于以往生物学研究的一门新学科.它以生物学实验室研究的数据为材料,利用不断强大的网络和计算机技术,以对数据的软件分析和计算为手段,在芯片上完成对生命本质的研究探索.并借此指导生物学实验的进行.

我们从1996年开始收集有关生物信息学的资料,并为研究生开设了"核酸与蛋白质序列分析"的选修课,酝酿开设《生物信息学》课程。通过1999和2000年两年为研究生开设的《生物信息学》课程的教学,并结合纤维素酶、木霉菌等的研究进行了功能性位点搜索,酶活性中心序列比较等研究(已刊登在*J. Biol. Chem.*, *J. Protein Chem.*等刊物上),我们不但对生物信息学有了更深的认识,同时也为这本教材积累了有关的内容。由于生物信息学是由诸多学科知识汇集成的一门新兴的研究生命的学科,而有关的教课老师均为研究生物学的。因此,我们在教学材料的编写和教学中不可避免地遇到了各种各样的困难。在1999年刚刚开始《生物信息学》教学时,手边缺乏完整可利用的教科书,当时的主要参考资料是1995年出版的*Methods in Enzymology*(266卷,*Computer Methods For Macromolecular Sequence Analysis*)中介绍的生物信息学内容,以及大量的英文综述期

刊文献，我们希望通过我们的介绍，使对此感兴趣的同学能够了解并通过自己的努力，进入这一生物学研究的新领域。此后，我们获得了 Andreas D. Baxevanis & B. F. Francis Ouellette 撰写的 *Bioinformatics*(1998)，这是一部非常好的生物信息学指导书(已由李衍达和孙之荣主编翻译出版，清华大学出版社，2000 年 8 月)，在我们 2000 年的教学中重点参考了这本书。本教材的编写重点参考了上述两部书，同时在我们的研究工作基础上，对多个生物信息学分析软件作了编译介绍，此外，还整理参考了大量的中英文期刊综述文献，如 Collins 等描述的 HGP 目标(*Science*, 1998, 282:682)，邹承鲁先生关于第二遗传密码的阐述(《科学通报》, 2000, 45:1681)，贺福初先生对蛋白质组的综述(《科学通报》, 1999, 44:113)，赵国屏先生对微生物基因组的展望(《生命科学》, 1998, 10:65)等等，在此不一一列举。此外，由于有许多文献在整理时，已不能确记出处，谨在此表示感谢！

由于作者对数学和计算机科学知识的贫瘠，不能深入地探讨生物信息学各种软件程序的编写原理，因此在教材的编写以及教学中，主要是结合基因组学的研究对生物信息学进行了介绍。希望能为有志于生物信息学研究和基因组学研究的生物学领域和其他领域的同学以及研究人员提供参考借鉴。

感谢山东大学研究生院对本教材出版的资助。

感谢山东大学出版社为本书的顺利出版给予的大力支持。

谨以此书献给我们的老师王祖农先生八十五华诞！

李越中　闫章才　高培基

2001 年 5 月于山东大学

# 目　录

# 第一章

# 信息学与生物信息

## 1.1 信息学

### 1.1.1 信　息

人们在日常生活工作中经常使用信息这个术语，比如当你接到一封信、阅读一份文献、采集一份标本、作一番调查……信息作为一个通俗的概念，人们容易理解.

在科学领域中，“信息”的概念则是非常多义和模糊的.这是一门学科的幼年期的常见现象.三个世纪前，能量同样是一个模糊概念，科学家直觉地承认它在物理过程中的重大意义，但是缺乏数学的严格性.今天，能量已被我们充分理解，我们承认它是实在的和基本的物理量.但信息则至今令人迷惑不解，部分原因是它以各种不同的姿态在众多的科学领域中出现.在相对论中，信息传播不允许比光还快；在量子力学中，系统的状态用它的最大信息量来描述；在热力学中，信息的落降伴随着熵值的增升；而在生物学中，基因是包含执行某种任务所需要的信息的指令……

信息研究的一个里程碑是二战期间美国电气工程师 Shannon 对噪音无线电频道的分析.但对信息是否永恒地保存在各种物理过程中仍没有取得共识.当恒星坍缩，形成黑洞，并随后“蒸发”时，信息发生了什么变化？信息是不可逆地消失了，还是以某种方式转移了？

地球上的生命种类繁多、性状表现复杂，其本身就是一个巨大的信息源.当我们环顾生命世界，可以发现形态各异的生物个体；感觉各种显现的颜色、散发的气味、发出的声音、表现的动作……时时刻刻都在活动的生命向我们展示其信息的复杂性.生物体内的新陈代谢，也在不断地产生各种信息，并且在信息的控制与调节下实现其正常的生理功能.在动物世界，传递、加工和处理信息的专有器官——神经系统的信息传递一直是生物学研究的重要课题；而在人类大脑，思维作为处理信息的最复杂形式尚待彻底探索.在我们生命的延续中，遗传信息以 DNA 中核苷酸为代码的信息传递，其信息量比人类任何通信传输系统都更庞大、更复杂.在一个生态环境中，各种生物之间相对平衡，并不断地发展演替，各种生物群落内、群落间也包含信息.当今生态环境学科中提出的生物多样性保护，按照信息观点理解，其实质也是生物信息多样性的保护.以至信息生态学的名词已被提出

来.总之,无论是从微观到宏观,从个体到群体,从生理到生态,生命现象到处都向我们展示出生物信息具有丰富的内容和广阔的研究前景.

从当前生命科学的研究来看,核酸和蛋白质等生命大分子的信息学依然是我们研究的中心.正如核酸和蛋白质向我们展示的,生命的秘密不在于其组成的化学物质的多样性,而在于其整体的逻辑结构和分子的有序排列所表达的生物学意义的多样性.有人比喻,生命是一部建立在生物分子元器件基础上的超级计算机,完成 DNA、RNA、蛋白质以及其他生物分子所编辑的生物指令的信息处理系统.

### 1.1.2 信息的度量

为了利用数学工具来研究信息,需要对信息进行度量,从而引进信息量的概念.信息体现在事物的种种表现状态中,这些表现状态我们称之为信息符号,譬如文字的各种字母符号,生物的各种不同种类、生物体表现性状的各种不同类型、组成蛋白质的多种氨基酸、组成核酸的核苷酸……信息符号常用一个字母来表示,体现信息的全体信息符号的集合称为状态空间.一般说来符号愈多体现的信息量愈大.正如你收到一份电报,电报很长,大量的文字符号可以传递大量的信息.但是信息的度量却不是完全建立在出现信息符号的多少上的.如我们接到一份完全由同一个文字符号组成的电报,这样的电报符号再多,也不可能为你带来许多信息.

信息量的增加,以系统中熵的减少为代价,如我们定义信息量

$$I = -K\log W \tag{1}$$

$$I = K\log\Gamma \tag{2}$$

在这两个公式中,系数 $K$ 及对数的底尚未确定.系数 $K$ 可以用最方便的方法加以选定.如令 $K=1$ 和以 2 为对数的底,则

$$\begin{aligned} I &= -\log_2 W \\ &= \log_2\Gamma \end{aligned} \tag{3}$$

系统的熵 $S$ 同实现该系统某一给定状态的方式数目 $\Gamma$ 可以联系

$$\begin{aligned} S &= k\ln\Gamma \\ &= lk\ln 2 \\ &= \frac{1}{\log_2 e}kl \end{aligned} \tag{4}$$

或把数值代入,从而 $I$ 和 $S$ 可以联系得

$$\begin{aligned} S &\approx 10^{-16} I \text{ erg/K} \\ &\approx 2.3\times 10^{-24} I \text{ cal/K} \end{aligned} \quad \text{式中 } I \text{ 的单位是 bit} \tag{5}$$

在一个绝热孤立的系统,如果它的熵保持不变,这种系统的信息是无法得到的.如液体在容器中处于冰冻状态,熵就会减小,但信息应当增加,因为在液体中混乱分布的分子,现在都以确定的方式居留于晶体的格点上.

我们可以说,熵是系统信息量的量度,信息和熵的等价性在某种意义上说类似于爱因斯坦定律:$E = mc^2$.这个公式反映了质量和能量的等价性,即物理量用不同的单位来量度.我们可以写出守恒定律:

$$S + I = 常数 \tag{6}$$

熵的增大,意味着信息的减小.

因此我们又可以用熵的单位(e.u.)来表示 $I$,亦即可以用(卡/度)或(比特)来表示 $I$.

根据(5)式,很多很多比特只不过等价于很少很少的熵,用熵的单位来衡量比特的值是很大很大的.有时候文献中把信息流和能量流进行对比.实际上信息只可能在具体的物理过程中传递,而这些过程的能量和熵的变化都很少.

我们也遇到了细胞和生物体的"逆熵"概念.逆熵这个术语通常理解为表示活系统极其有序的程度.莫诺写道:"生命系统同无生命晶体的差别只是生物体更加有序而已."但这种说法是错误的,让我们考虑一下 Blumenfeld 所做的简单计算.

我们估算一下从细胞形成生物体之后结构的有序程度.人体大约含有 $10^{13}$个细胞,如果这些细胞各不相同(实际上并非如此),那么对于细胞的惟一分布,我们可以得到

$$\begin{aligned} I &= \log_2(10^{18}!) \\ &\approx 10^{13}\log_2 10^{13} \\ &\approx 10^{14}\text{bit} \end{aligned}$$

这个数值相当于 $10^{-9}$cal/K.

每个细胞大约含有 $10^5$ 个生物聚合物分子.如果这些分子各不相同,而且在细胞中是惟一一种分布,我们得到

$$\begin{aligned} I &= \log_2(10^8!) \\ &\approx 10^8\log_2 10^8 \\ &\approx 2.6\times 10^9\text{bit} \end{aligned}$$

对于全部细胞来说,所得到的信息量为

$$\begin{aligned} I &\approx 10^{13}\times 2.6\times 10^9 \\ &\approx 2.6\times 10^{22}\text{bit} \end{aligned}$$

即相当于 $6\times 10^{-2}$cal/K.

人体约含有 7000g 蛋白质和 150gDNA 物质,相当于 $3\times 10^{25}$个氨基酸残基和$3\times 10^{28}$个核苷酸.使所有这些单位环节作惟一确定的分布,蛋白质相当于 $1.3\times 10^{26}$bit 的信息;DNA 相当于 $6\times 10^{23}$bit 的信息,分别等于 300cal/K 和 1.4cal/K.因此构成生物体过程中熵的减少不超过 301.5cal/K,相当于 170g 水蒸气凝结为水的熵的变化,这是一个很小的量.从这方面来看,生物体缺乏任何特殊的有序性,生物体的序的数量级与一块同质量岩石的有序程度并无多大差别.但是晶体同生物体有重大差别,尽管它们的信息量可能相等,但信息的特征不同,晶体含有重复的、多余的信息.也就是说,晶体的结构是周期性的,晶格的基本晶胞重复很多很多次.相反,生物体是具有大量非多余信息的非周期性"晶体".

这里还没有完全说明晶体和生物体的差别.生物体是开放系统,亦即是根据平衡的弱和强相互作用进行运转的一台"化工机器",正是这种强、弱的相互作用才产生直接的联络和反馈.系统具有复杂的、有功能的结构,系统的特性取决于每个部件的位置和状态.因此,生命系统不同于气体或周期性晶体系统.细胞和生物体都是动态系统.仅用熵来描述是不够的,因为尽管这种描述是合理的,但它毕竟无法解释动态系统中的活生生的情况.

布卢门费尔德曾提出一个简单实例：引擎包括气缸和活塞，这两样东西都用金属制成，这些零部件的熵可以计算.如果我们把活塞从汽缸中拿掉，尽管熵不发生什么变化，但是引擎却不能运转了.

利用对比的方法，我们可以计算细胞生物聚合物的信息量及相应的熵.但是，这种计算并不能使我们理解生物分子的特性.真正重要的并不是 DNA 中的信息量，而是 DNA 短文中所包含的蛋白质合成的指令以及其他的指令.换句话说，对生物学而言，重要的并不是信息的“含量”，而是信息的“含义”，亦即在编译指令过程中信息的价值.当然，热力学规律对无生命系统和生命系统都是适用的，但是理解生命现象还需要诸如目前正在发展的系统论和自动控制论等其他的物理学.尽管尚未建立不同于现有物理学原理的新原理，但对“机器”系统论的研究却是新的，而且这种研究有利于对特殊动态系统（化学的分子系统）的物理概念和物理形式进行推广.

动力学和统计学的关系极其复杂，蒸汽机中统计的部分是水蒸气，动力学部分是金属的机器零部件，这两部分在结构和功能上都是彼此分开的.生物系统中，动力学部分和统计学部分是统一在一起的.生物体是一种非周期性的“晶体”，是由不同成分所组成的非均匀、但有序的系统.这种定义也同生物体的各个功能部件（器官、组织、细胞和单个球蛋白）的状态有关.从物理学角度研究生物的复杂问题，一种方法即是利用物理化学的数学模型，并把各种方法统一到理论生物学中，这一领域的工作正在进行，并且已经开始得到极其重要而有意义的结果.

## 1.2 生物信息

### 1.2.1 生物信息的基本特征——遗传密码

没有什么比基因代码能更好地解释生命的杰出计算技能了.从化学的观点看来，核酸和蛋白质这两类生物大分子之间几乎没有什么关联.但在生物体内，脱氧核酸 DNA 和核酸 RNA 储存信息，并巧妙地形成信息通道，转移信息到蛋白质；蛋白质则执行信息的功能，并进一步转移信息.这些分子共同演绎生命的种种奇迹.但令人惊奇的是，DNA 的化学组成仅是 4 种脱氧核苷酸（腺嘌呤、鸟嘌呤、胸腺嘧啶和胞嘧啶，A, G, T 和 C）不同排列的长链，RNA 是 4 种核苷酸（A, G, C 和尿嘧啶 U）连成，而蛋白质则是由 20 种氨基酸组成的长链.这些物质本身对完成生命的奇迹是无能为力的.

1947 年，斯特恩提出“遗传信息”的概念，并且认为遗传信息是以核蛋白分子上的“调制”（modulation）形成而记录的.斯特恩把遗传信息的记录（储存）比喻为唱片上的印纹.特别有意义的是，他指出遗传信息的再现或复制有直接和间接两种方式：直接复制即细胞分裂周期中核蛋白分子的增殖，就好像是从一张唱片复制另一张相同的唱片；而间接复制则是通过个体发育而表达遗传性状，有如唱片上的信息记录可通过唱机演放出来.斯特恩说：“染色体核蛋白上的基因调制在其表面发生的化学反应而表现特定的生物学作用.”这句话隐含了遗传物质是作为模板（后来人们知道是通过转录 mRNA）而起作用的概念.

1953 年，Watson 和 Crick 根据对 DNA 晶体的 X 光衍射图谱的仔细分析、计算和研究，提出著名的 DNA 分子双螺旋结构模型，并初步解释了它的复制机制.后来的实验研

究表明,DNA 分子是以半保留方式进行复制的,即母体 DNA 分子相互缠绕的两条单链在解旋酶的作用下解旋,各自复制一条子链,从而形成两个 DNA 双链分子.于是母体 DNA 分子中储存的遗传信息就传递给子体 DNA 分子了.

实验研究也表明,遗传物质 DNA 表达为遗传性状的过程是通过遗传信息的转录和翻译而实现的.在转录过程中,DNA 分子作为合成 mRNA(信使核糖核酸)分子的模板.mRNA 分子从 DNA 那儿转录来的遗传信息,再通过翻译为蛋白质,变换为蛋白质分子的信息和表型信息.

DNA 和 mRNA 分子都由 4 种核苷酸组成,而组成蛋白质分子的单体——氨基酸却有 20 种.4 种核苷酸是如何决定蛋白质分子中的每一种氨基酸的?1954 年,著名的美籍俄裔物理学家盖莫夫通过排列组合的方法提出三联体密码子假说.60 年代初期,用人工合成特定核苷酸成分的 RNA 作模板,控制多肽的合成,从而证明了核苷酸与氨基酸对应关系的遗传密码子假说,解读了大自然的"天书".

遗传密码是一种通用密码.从病毒、细菌一直到人体细胞,都采用同样的遗传密码.但在真核生物中线粒体的遗传密码与细胞核的通用密码略有差异(见表 1-1).

**表 1-1 线粒体遗传密码与通用密码的差别**

| 三联体密码子 | 通用密码编码的氨基酸 | 人和牛线粒体密码编码的氨基酸 | 红色面包霉线粒体密码编码的氨基酸 | 酵母线粒体密码编码的氨基酸 |
|---|---|---|---|---|
| UGA | 终止信号 | 色氨酸 | 色氨酸 | 色氨酸 |
| AUA | 异亮氨酸 | 甲硫氨酸 | 异亮氨酸 | 甲硫氨酸 |
| CUA | 亮氨酸 | 亮氨酸 | 亮氨酸 | 苏氨酸 |

### 1.2.2 生命中信息的传递

从 DNA 应用的四字母表翻译成蛋白质应用的 20 字母系统,所有的地球生命都是用相同的代码.实现生命的关键问题是,精巧独到的代码系统是怎样产生的?笨拙的原子怎样自发地写出自己的软件?得到和启动第一个生命细胞所需的非常古怪的信息形态是从哪里来的?没有人知道答案,科学家在这个问题上的争论按传统分成两大阵营.一边认为一切是由偶然发生的——生命是化学上的异常侥幸——这是玛纳德的观点.要计算在无序的化学混合物中把适当的分子恰好调整成所需要的精致排列是可能的,但几率极低.如果生命如我们认为的那样产生于偶然,则在可观察的宇宙中仅仅碰到一次.

相反,生物决定论者认为偶然性是第二位的,分子有序的强制形态是"自然法则"的结果.举例说,美国生源说先驱福克斯宣称,化学令氨基酸精确地、井井有条地结合,使它具有生物学的功能.果真如此,这就像自然界的化学物质具有一种倾向性或者协同作用,激发产生生命的物质.难道物理学和化学的定律中包含着生命的蓝图?生命的关键性信息是怎样被译成这些定律的代码的?

提出这个问题,我们需要对支持生命的信息的本性作更认真的思考.有一个重要蹬观察结果,即富有信息的结构趋向于缺乏形式.这个性质在称为算法信息论的数学分支中作了最清楚的说明.算法信息论把它作为计算机程序的输出来寻求信息组合的量化.试考虑

二进制数列 1010101010101010…它可以用简单的指令“Print 10 n times”把它产生出来.输入指令远比输出数列短得多.它反映这样的事实:输出包含一种重复的形式,这种形式很容易简洁地描述出来.因此,输出只有很少的信息量.

反之,一个明显无序的数列(如 1101010010010010111)不能浓缩为一简单的指令,所以它具有高信息量.如果要 DNA 有效地储存信息,它最好在系列中不包含太多的形式,因为形式代表信息的冗赘.生物化学家证实了这种期望值,有序化的基因组多半看上去像四个组分字母的“无序搀和物”.

基因组序列杂乱无章的本性与生物决定论相悖.物理定律常能预言规则结构而不能预言无序结构.比如,晶体是一种有周期结构的有规则原子阵列,好像上面所提的重复二进制数列,因此几乎没有信息.晶体结构建造在物理定律之中,如同它们的周期形式是由数学对称性所决定的一样,是这些定律所固有的.但是氨基酸的“无序”或者 DNA 的“无序”,却不能建造在物理定律“之中”.

同样,生命的出现也不能建造在化学定律“之中”.这一事实的直接证明来自对 DNA 结构的测定.DNA 双螺旋阶梯中的每一梯级由两个键段组成,它们像锁扣和螺栓一样耦合在一起.化学测定最终确定了将键段连接在一起的性质,也确定了把它们连接到梯子各端的力;然而在逐次梯级之间并没有化学键.化学并不介意于梯级的秩序,生命则可以在“一闪念”间改变它们.正像计算机指令键盘中字母键同纸和墨水的化学成分和性质无关一样.因此 DNA 的制造信息的“字母”与核酸的化学性质无关.正是这种生命的能力使它解脱化学组成和化学性质的束缚.而生物决定论则意味着有一件只会阻止而不会加强生物创造力的拘束衣.

如果生命代表着一种从化学组成和化学性质的逃脱,我们解释生命现象就不能求助于化学组成和化学性质.那么哪里还可能找到生命的解释呢? 我们认为生命是复杂的信息过程,所以到信息论和复合(complexity)论中去寻找答案才有意义! 生物信息没有译成物理学和化学定律的代码(至少现在如此),那么它来自何处? 科学家一致的看法是:生命信息不是自发出现的(也许除了大爆炸?),生命系统中的信息一定是用某种方法在它们的环境中生成的.虽然没有什么已知物理法则能够从“无”中产生信息,然而可以有某种原理来解释信息是如何从环境中获得并积累在宏观分子之中的.

### 1.2.3 密码子与氨基酸在概率上的相关性

我们以遗传密码为例,介绍并讨论盖莫夫提出的三联体密码子假说所依据的排列组合的计算.

1. 排列问题

(1) 有重复的排列.DNA 分子的四种核苷酸 A, T, G, C 中,每次取三个排列成三联体密码子,排列可有重复(如 AAA, AAG, CCC, CTT, 等等),那么共可排列成多少种密码子?

从 $m$ 种不同的元素里,每次取 $n$ 个元素进行可重复的排列,总共有 $m^n$ 种可能的排列方式.所以,A, T, G, C 排成有重复核苷酸的三联体密码子的种数共可有 $m^n = 4^3 = 64$ (种).

(2) $n > m$ 的有重复的排列.由于元素可以重复,故在此情形下 $n$ 可大于 $m$.由四种