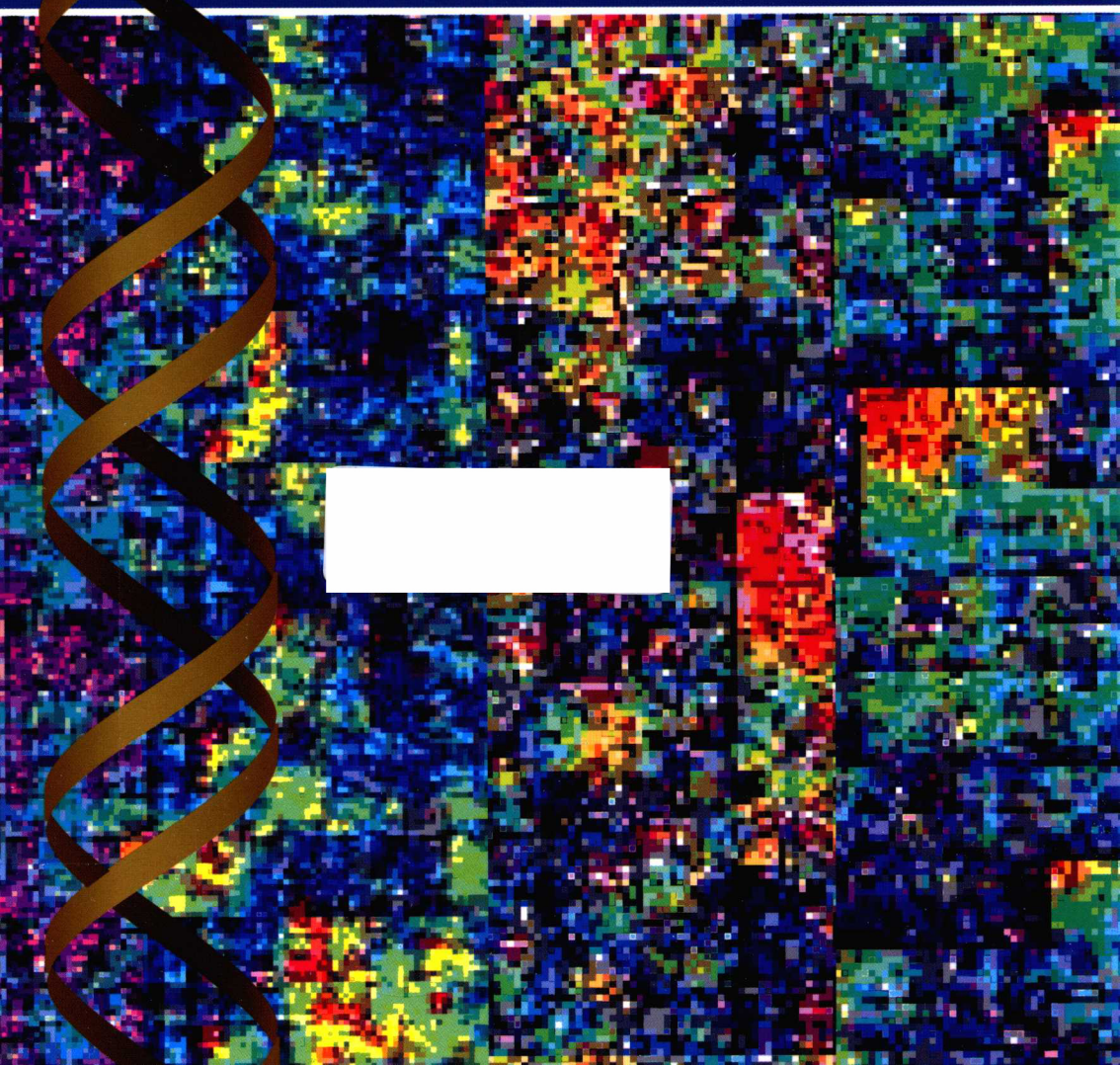
来自基因组的一些数学

郝柏林　著

上海科技教育出版社



来自基因组的一些数学

郝柏林　著

上海科技教育出版社

图书在版编目(CIP)数据

来自基因组的一些数学/郝柏林著. —上海:上海科技教育出版社,2015.12

ISBN 978-7-5428-6331-7

Ⅰ. ①来… Ⅱ. ①郝… Ⅲ. ①基因组—研究 Ⅳ. Q343.2

中国版本图书馆CIP数据核字(2015)第253689号

序言

在20世纪末，物理学对“死物”的研究，从微观粒子的构造和相互作用，到宇宙的演化发展，可谓即深且远。生物是物，生物有理。物理学对“活物”的研究有着长远的历史，而在生物学研究进入分子水平以后，物理学特别是理论物理学有了更为广阔的用武之地。生物有形，生物有数。对活物和生命现象的研究必然用到各种数学工具，甚至带来新的数学问题。

本书作者在1997年夏天，从统计物理、非线性科学和复杂系统的研究，转入理论生命科学领域，并非一时心血来潮。

1985年中国科学院生物学部举行学部常委扩大会议，讨论生物学的发展战略。本书作者受数学物理学部委托，自始至终参加了这次会议。当时在海外深造的青年学者大部分还没有学成归国，主要是老一代生物学家们在支撑局面。那正是基因组时代揭开大幕的前夕。作者虽然正在全力发展以符号动力学为特色的混沌理论，也深切感到高度复杂和“非线性”的活物质和生命现象，必将成为物理学的重要关注对象。于是在没有设定时间表和特别目标的情形下，开始了转向理论生命科学的从容准备。

1993年中国物理学会委托清华大学组织了一次物理学与生物学的讨论会。物理学会的出版物报道了大部分报告，而只字未提作者强调粗粒化描述的发言。可见“粗粒化”对于物理学会派去的记录人员，当时还是相当陌生的概念。

DNA双螺旋结构和遗传密码的发现，开启了分子生物学时代。生物学已经积累了大量事实和数据，而且每日每时产生着海量新数据。2003年完成的人类基因组计划，花费了约30亿美元测定一个人基因组的约30亿个字母。现在，人们正在接近用1000美元测定一个人基因组的目标。截至2014年年底，全世界已经完成和正在进行的基因组测序计划超过了58,000个，这个数字还在与日俱增。预计几年之内，将会有成百万人类个体的基因组被测序。归根到底，人口、粮食、健康、医药、环境、能源这些全人类面临的重大挑战，都与生物有关，而基本的生物学规律和问题必须从分子水平认

识和解决。

生物医学和生物技术领域的从业人员、研究机构、学术期刊和经费支持都远远多于许多其他方面。然而，管理和资助单位的目标和要求也是明确、具体和紧迫的，那里的科学工作者很少有时间去思考更深层次的问题。对于有志于自然科学基础研究的年轻人，这是时代的机遇。早生20年，没有可能从事这样的工作；晚生20年，重要的问题已经被别人发现和解决。一部分有数理和计算机背景的青年学者，应当抓住时机，义无反顾地进入生命研究领域。

研究机构的设置和大学专业的划分，总是落后于科学发展。从哪一门学科开始并不重要。学科交叉不是议论，而是实践。要自己找切入点，而不是靠别人指点方向。要“眼高手低”：内心深处总是惦记着重大和根本的问题，但从早到晚下功夫掌握事实、搜集数据、静心思考、反复计算和分析，通过粗粒化和视像化来形成概念、提出问题。不管是“用牙啃”，还是“用脚踹”，要想尽一切办法解决问题，而不是把自己欣赏的某种方法或框架强加给大自然。力争“全局在胸”，坚持“单刀直入”。入门是不难的，深入也是办得到的，但是，“找到感觉”是不容易的。物理学早已经从单纯的实验研究发展成为鼎立于实验、理论和计算三大支柱上的成熟的科学。生命科学正在走向成熟的过程中，理论计算和数学方法，注定要发挥日益重要的作用。时不我待，机不可失，有志者奋勇向前！

书中引文按作者姓氏英文或汉语拼音顺序排列在最后，正文中按序号引用。例如，本书部分内容，曾在庆祝杨振宁先生85岁华诞的国际会议上报告[44]。

作者特别感谢各个时期的年轻合作者陈国义、王彬、纪丰民、戚继、俞祖国、王希胤、史晓黎、孙奕钢、高雷、孙健冬、李强、刘劲松、徐昭、周婵、汪浩、虞洪杰、左光宏、张建国、倪培相、刘捷孟、毛洪亮等。作者从苏州大学数学系教授谢惠民、中国科学院理论物理研究所研究员郑伟谋和台湾中央大学教授李弘谦那里学到很多东西；复旦大学理论生命科学研究中心袁力老师和中国科学院文献情报中心魏韧老师在查找资料方面给予了持续帮助；上海科技教育出版社叶剑编辑多年来在书稿写作上表现出极大耐心和支持；理论物理研究所程希有老师在LaTeX排版技术上给予指导。作者在此一并致谢。

郝柏林　　2015年6月26日

于复旦大学理论生命科学研究中心

目录

第一章　DNA测序和基因组时代

“基因组”来自大规模的DNA测序。因此，我们必须从DNA和它的测序讲起。

1.1　发现DNA和它的遗传载体功能

1869年瑞士青年学者米歇尔(Johannes F. Miescher, 1844 – 1895)发现了脱氧核糖核酸分子，即现在大家熟知的DNA。他虽然也曾猜想过DNA可能与遗传有关，但还是倾毕生精力去研究鱼精蛋白。毕竟蛋白质与生命过程的关系已经是当时科学研究的热门，以致1878年恩格斯在《反杜林论》这部著作中就写下了至今还基本正确的语句：“生命是蛋白质的存在方式，这种存在方式本质上就在于这些蛋白质的化学组成部分的不断的自我更新。”

然而，DNA和蛋白质究竟谁是遗传信息的携带者，这个问题曾经处于长期争论之中。直到1944年，美国洛克菲勒大学三位学者设计的决定性实验才证明了DNA作为遗传信息载体的功能[3]，推启开分子生物学时代的大门。

1953年发现DNA的双螺旋结构，随之又破译了DNA如何编码蛋白质的“遗传密码”，并且成功地用化学方法测定了一些蛋白质分子的氨基酸排列。1958年12月桑格(Frederick Sanger, 1918 – 2013)在他的诺贝尔获奖演说[101]中提到的“蛋白质结构”，就是氨基酸的排列顺序，即现在所说蛋白质的“一级序列”，但还不是蛋白质分子的三维空间构造。正是因为知晓了胰岛素的全部氨基酸排列顺序，中国科学家们才能够在1958年提出，并在1965年完成人工全合成牛胰岛素的重大课题。

DNA是由四种核苷酸单体聚合而成的高分子。通常用a、c、g和t四个字母代表腺苷酸、胞苷酸、鸟苷酸和胸苷酸这些“碱基”，而把DNA写成由这四个字母组成的符号序列。DNA序列的长度可以从几百数千，到百万

乃至千万个字母。DNA的第一个字母D代表“脱氧”，说明在作为核苷酸的组成部分的五碳糖上有一个羟基OH脱去了氧O，只剩下H。没有脱氧的核苷酸聚合成核糖核酸，简单地记为RNA。RNA可以用a、c、g和u四个字母组成的符号序列表示，其中代替t的u是尿苷酸。脱氧核糖核酸DNA和核糖核酸RNA都是一维、不分岔、有方向的高分子。DNA通常以稳定的双链形式存在，双链中的a、t和c、g互相以两个或三个氢键联系配成“碱基对”。单链的RNA往往靠局部配对形成种种二级结构，以增加稳定性和完成特定功能。按照DNA序列产生相应的RNA称为“转录”，相反的过程称为“反转录”。这两种存在于自然界中的过程都在被人类认识和掌握之后，成为实验研究和基因工程的手段。

遗传信息保存在DNA序列里。DNA双螺旋中任何一股的信息含量同另外一股等价，因为可以借助$a-t$和$c-g$的配对规则，由一股推出另一股。然而，两股中编码的基因和调控信号是不同的；位于一股某处的基因，在另一股相应位置上就只是个“影子”。如果某一个基因需要“表达”，首先就要把它转录成一段信使RNA (mRNA)；再把mRNA送到细胞质里面大量核糖体之一去翻译成蛋白质。“表达”一词最初指从基因合成蛋白质。现在知道，许多转录产生的RNA也有重要的生物功能，因此也是一种表达方式。核糖体是由许多RNA和蛋白质组成的蛋白质工厂。早在1956年，DNA、RNA和蛋白质的关系就被概括成分子生物学的“中心法则”：DNA制造RNA，RNA制造蛋白质[21, 23]。半个多世纪以来，中心法则已经具有更丰富的内容，这里包含了多项获得诺贝尔奖的科学贡献。

应当指出，围绕中心法则的诸多知识来自对细菌的研究。近些年来对真核生物的研究，揭示了转录、翻译之后对序列的各种修饰，例如DNA序列某些位点的甲基化、蛋白质序列的糖基化和磷酸化等等，都对生物学功能有重大影响。这部分研究被称为epigenomics[1]。总之，中心法则加上外饰基因组学的知识，才能提供对现代分子生物学的较为完整的概念。

1.2 DNA的测序技术

作为由四种单体a、c、g和t组成的高分子，只有把单体的排列顺序测定出来，才能得到关于DNA结构的初步知识。用化学方法测定DNA中的核苷酸顺序曾经是颇为艰苦的实验室工作。在1950年代后期，测定含有几个核

[1]目前的中译“表观基因组学”颇为误导，或许“外饰基因组学”更为确切。

苷酸的DNA片段，就足以构成一篇博士论文。直到1977年，DNA大规模测序的两种基本方法，才先后发表在美国《国家科学院报告》这同一个期刊上。这两种方法，一是化学降解法[85]，即用专门的化学反应在特定的核苷酸字母处把已经合成的DNA序列“咬断”；二是聚合终止法[102]，即让DNA的聚合过程在特定的字母处停止。两者都导致长长短短的以同一个特定字母结束的寡核苷酸串，再用“跑凝胶”等办法测出串的长度，即特定字母的位置。这些方法同毛细管凝胶技术结合，最终导致了自动化的测序设备。两种方法的主要发明者，分享了1980年度的半个诺贝尔化学奖。

一个生物细胞里的DNA被提取、扩增(某些新一代测序技术不需扩增)之后，被分成大量小片段即“读段”(reads)去测定核苷酸顺序。读段的总长度会达到基因组长度的许多倍，这倍数称为覆盖度，通常记为X。X可以从几倍到成百上千倍，因所用技术和测序目的而不同。大量读段要用计算机拼接成更长的段落，最好的情况下可以拼接成单个的染色体。有了足够长的连续的DNA段落之后，就可以开始寻找基因和预测对基因表达的调控信号。这些步骤都涉及许多数学和统计方法，我们基本上不在本书里讲述。只在第三章里会结合泊松分布讲一下对基因组测序有重要应用的Lander–Waterman公式，在第七章里根据我们自己的经验叙述在基因组里寻找基因的有关知识。

从1977到1995的18年中，测定了约800个“基因组”。不过，这些基因组都不属于独立生活的生物体，而是来自病毒和噬菌体(噬菌体是细菌的“病毒”)这些寄生生物。直到1995年才首次发表了两种独立生活的细菌的全基因组，这两种细菌是生殖道支原体(*Mycoplasma genitalium*)[31]和流感嗜血菌(*Haemophilus influenzae*)[26]。所谓“独立生活”，有明确的含义。若凡是维持生命活动所需的大分子和构成这些大分子的单体，都在细胞内的生物化学工厂中自己制造，那这种生物就是独立生活的。因此，寄生在别的生物中的各种细菌，多数也属于独立生活的物种。

1.3 人类基因组计划

早在1985年美国能源部就在一批科学家的促进下，提出了人类基因组的测序计划。这个计划从1990年启动，准备花费30亿美元来测定构成人类基因组的30亿个字母。事实上，测序的原始材料来自5个人类个体，他们是分属于高加索、亚洲、非洲和西班牙等族群的两男三女。

中国在人类基因组计划中承担并完成了1%的测序任务。这在很大程度上是一批自称“华大基因”的年轻人的功劳。原来人类基因组计划在国际上酝酿日益成熟时，国内一些有远见的学者也想努力参与。但是测序技术所涉及的庞大资金，使有关部门望而却步。然而，国家自然科学基金委在学部委员吴旻等推动下，还是在1993年拨款300万元人民币，启动了我国的人类基因组工作。当时设想的重点，是开展与少数民族基因多样性有关的研究。1996年科技部主导的863计划，扩大了对人类基因组计划的支持，先后在北京和上海成立了人类基因组研究的北方和南方中心。不过，那时并没有把直接参与国际人类基因组计划，作为主要目标。原来863计划的早期指导思想，就是“跟踪”国际上高新技术的发展。

1990年代中、后期从海外归来的一批青年学者，对于我国参加国际人类基因组计划，抱着更为积极的态度。他们在中国科学院遗传研究所内成立了一个小组，自行购置了一台测序仪，从测定云南腾冲嗜热菌(*Thermoanaerobacter tengcongensis*)[2] [4]的基因组开始实战练兵。1999年9月底，国际人类基因组计划在英国剑桥召开例行的第5次战略会议，检查各个实验室已经接近尾声的测序工作。当时以观察员身份参加这次会议的一位中国学者[3] 看到我国即将错过参与这一伟大计划的末班车，就挺身而出，在用腾冲嗜热菌的部分测序结果演示了自己的实际能力后，争得了1%的测序任务，而且选择了估计基因含量比较丰富的第3号染色体短臂上的一段。

这位观察员回国后，各个方面作出不同的反应。笔者亲自听到一位在科学领导层中位置颇为显著的院士说：“这么大的事情，不请示就承担下来!”于是几天后，在见到这位研究员时提问：“这么大的事情，你请示了没有?”答复是，写过多次报告，都被“留中不发”，没有下文。

所幸中国科学院当时的领导采取了积极态度，拨出一笔专款，支持了测序工作。这样，在2000年6月26日美国总统克林顿在白宫和英国首相布莱尔在白厅同时宣布人类基因组草图基本完成时，中国也是6个正式参加国和20个实验室之一。在英国《自然》杂志发表的论文[69]中，有5位中国科学家署名。相形之下，虽然有些俄国科学家以个人身份在某些实验室参加了这项工作，俄国却不是国际人类基因组计划的成员国。

[2] 2004年改变属名，现在的拉丁名字是*Caldanaerobacter tengcongensis*。

[3] 杨焕明，当时是中国科学院遗传研究所研究员，2005年当选中国科学院院士。

1.4 新一代测序技术

最初的人类基因组计划，就曾估计到测序技术的进步必然会加速计划的实现。事实上，早期的做法是先测定各种“物理图谱”、“遗传图谱”等，在对基因分布有一定了解的基础上分段完成测序。那时的测序设备可以一次测出500 – 700个字母的读段，再把大量读段拼接成更长的DNA序列。这种做法有利于多个单位进行合作，并且能保证较高的测序质量。这也是国际人类基因组测序计划所采取的战略。然而，一位名叫文特尔(J. Craig Venter, 1946 –)的科学家成为半路杀出的程咬金，他建议把整个基因组拿来，随机打断成大量短段，分别进行测序，然后靠强大的计算机进行拼接。这种全基因组“鸟枪法”或“霰弹法”测序，虽然在一开始遭到质疑和批评，但很快成为测序技术的主流，以致人类基因组草图就是用两种方法同时发表的[69, 117]。

测序技术的创新从未停步。关于第二代、第三代测序方法的建议层出不穷，以致很难真正划分代次。最早的自动测序基于特异性的化学反应，其输出必然是待测DNA经过大量扩增以后的平均结果。加速甚至摆脱扩增步骤，是一个重要的发展方向。发展过程中曾经一度以牺牲读段长度为代价，同时靠高度并行产生的大通量做补偿。可以反映单个DNA及其环境的单分子测序和单细胞测序也已经应运而生。与早期的“化学测序”对比，后来的发展引入越来越多的物理手段，例如利用“零模波导”使荧光精确报告单个核苷酸的变化。我们如果在这里列举任何测序设备的实例，在本书到达读者手中时，都会是过时的历史记录。虽然如此，还是可以指出一篇2012年的论文[96]，给那个时期的测序水平留下一些指标。应当说，以1000美元的代价测定一个人的基因组，已经不是遥不可及的目标。2013年夏天已经清楚，几年之内就会有100万个人类个体的基因组被测序。这预示着，对人类的生理和病理研究，必然在分子水平上有更深刻的结果。

对于大规模的基因组测序，我国学术界曾经有人说过“测序不是科学”的贬词。诚然，大规模的DNA测序设备已经属于高新技术领域，但它们的基本原理无不基于以往的科学研究成果。更重要的事实是，今后越来越多的生物学研究，要从基因组测序开始。可以说，没有测序，就没有生物科学。早在1991年，两种快速测序方法的发明者之一吉尔伯特(W. Gilbert, 1932 –)就为英国《自然》杂志撰写短文[37]，针对生物学研究范式的变化指出，“正在兴起的新的范式在于，所有的新的‘基因’将被知晓(在可以

用电子方式从数据库里读取的意义下)，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。”

快速度、高通量的DNA测序把生物学推进到大科学时代，培育出一批从事大科学的研究集体和组织者。那种倾毕生精力研究一个基因、一条代谢途径、一种生理过程的时代已经过去。还会有学者这么做，但是他们将只代表一种研究风格，而不再是学术主流。对于在生物学研究中使用数学方法和计算手段，也有类似的观察。如果我们的学术评价和业绩考核体系，不适应科学进步的历史步伐，我国生物学的发展必定要蒙受负面影响。

第二章　粗粒化和视像化

2.1　粗粒化与符号描述

现代科学描述自然界时，不可能在从微观到宏观甚至宇观的一切层次上同时进行细致刻画。人们必须瞄准一定的层次，忽略更细小层次上的结构和运动，代之以平均后的“参数”，同时也必须把更大尺度的影响，处理成某种背景。例如，考虑一粒落入水中的花粉，用显微镜观察它的运动轨迹。在视野中看到的花粉，它同大量水分子不断碰撞，做不规则的断断续续的折线运动。每段折线上花粉受到的阻力可以用摩擦系数描述。摩擦系数可以由实验测定。当然，原则上也可以从分子碰撞机制出发，计算摩擦系数；那就要转入下一个描述层次，动用物理学武库中的不同兵器。长长短短的折线变化，则是由分子碰撞的随机性涨落决定；涨落的大小与温度有关，温度代表着更大层次上的环境。

粗粒化的观察伴随着符号描述。试想我们看到6个小写英文字母

$$b,\ t,\ u,\ d,\ c,\ s,$$

从事粒子物理的学者会立即认出来，这是6种“夸克”的名字，它们各有一定的电荷、质量、自旋和其他量子数。对于更多的科学工作者，

$$p,\quad n,\quad e$$

这些字母代表着质子、中子、电子。人们并不关心质子或中子由哪3个夸克组成，只要知道它们的质量、电荷、自旋、磁矩就成了。

化学家们更习惯于

$$\mathrm{H},\ \mathrm{C},\ \mathrm{N},\ \mathrm{O},\ \mathrm{P},\ \mathrm{S},\cdots$$

这些大写字母，它们分别代表着氢、碳、氮、氧、磷、硫……这些原子，它们各自有一定的原子序数、原子量、化学价、亲和力和离子半径等。用

原子符号可以书写出各种分子，例如

$$H_2O,\ CO_2,\ NO,\cdots,$$

它们具有一定的分子量，是透明液体或无色无嗅气体，等等。然而，在遇到还不算太大的核苷酸和氨基酸分子时，如果每次把几十个原子的符号和化学键连接都写出来，则既不方便也无必要。人们用

$$a,\ c,\ g,\ t$$

四个字母代表四种核苷酸，只要知道它们在配对组成双螺旋时，a和t由两个氢键相连，属于“弱”耦合，而c和g由三个氢键维系，属于“强”耦合。这里的“强”和“弱”，与粒子物理中的相应概念差了许多个数量级。

与此类似，构成蛋白质的氨基酸通常用20个大写英文字母代表：

$$A,\ C,\ D,\ E,\ \cdots,W.$$

它们各有一定的物理化学特性。由成百上千个氨基酸组成的某种蛋白质，在研究生物化学反应路径或调控网络时，又可以用一个符号表示，例如同乳腺癌有关的蛋白BRCA1和BRCA2(把5个字母看成一个符号)。

还可以继续列举类似的例子。总之，粗粒化通常伴随着使用符号。自然科学中所使用的许多符号，都代表着粗粒化的结果。粗粒化是引入近似的结果。然而，恰当的粗粒化可以帮助导致严格的结论。历史上最具有启发意义的粗粒化实例，乃是伽利略的比萨斜塔实验。如果在1589年伽利略就拥有20世纪末的激光测量技术和计算机控制的数据采集系统，那他从比萨斜塔上同时松手放下的两个重量不同的物体，绝对不会在空气中同时落到地面。正是因为没有拥有相应的观测精度，伽利略才发现了精确的自由落体定律。自然科学工作者的“艺术”修养高低，往往表现在善于实行粗粒化的程度。

粗粒化的过程，伴随着使用符号。在许多情况下，这些符号还进一步组成符号序列。DNA和蛋白质都是粗粒化导致的符号序列。符号和符号序列还蕴含着已知和未知的信息。粗粒化把我们带近依靠符号序列传输信息的经典领域：信息论。现代信息论的一篇奠基性论文，虽然只字未提生物，却对描述生命现象的数学理论具有根本指导意义。

2.2 香农信息论第三定理

香农(Claude Shannon, 1916 – 2001)在1948年发表的一篇论文，标题