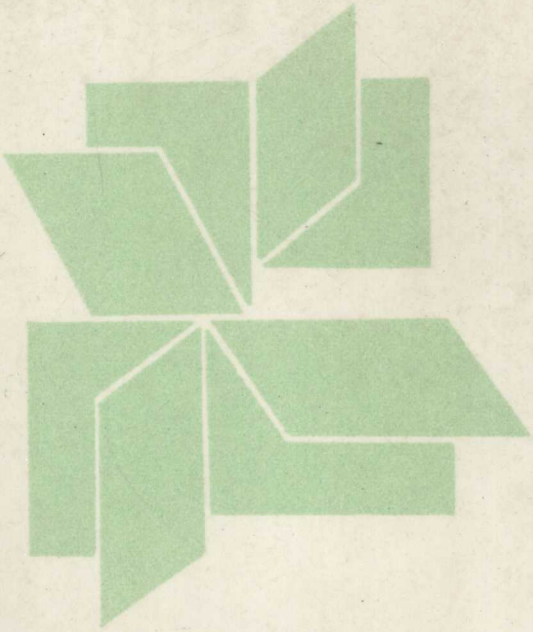
罗辽复 著

理论生物物理学论文集



内蒙古大学出版社

内蒙古大学学术丛书

理论生物物理学论文集

罗辽复　著

内蒙古大学出版社

责编　呼和
封面设计　赵齐坤

理论生物物理学论文集
罗辽复　著

内蒙古大学出版社出版发行
（呼和浩特市大学路1号）
内蒙古自治区新华书店经销
内蒙古地图印刷厂印刷

开本：787×1092/16　印张：29.6875　字数：694千
1995年9月第1版　1995年9月第1次印刷
印数：1—500册

ISBN 7—81015—552—0/0.37
定价：30.00元

理论生物物理学论文集

COLLECTED WORKS ON THEORETICAL BIOPHYSICS

序

本书收集了1983－1995十二年间作者及其合作者在理论生物物理学领域内发表的主要论文。它们散见各处，部分发表于内蒙古大学学报上。今结成此集，以便于读者参考。

理论生物物理学是一个新兴学科，它包括哪些内容，学术界尚无明确的一致的意见。按作者的看法，在现阶段及今后一段时间内，它主要研究的问题有：1°. 与生命起源及早期进化有关的重大理论问题，如手性起源，密码起源及核酸中信息的积累和序的形成等。2°. 由氨基酸和核苷酸序列确定它们的构象（空间结构），并在此基础上从动力学角度研究结构——功能关系。3°. 个体发育及生命活动中的自组织特性。4°. 感官和神经系统的信息处理。我们这几年的工作主要集中于前两方面。

文集中将论文按内容分成A至W共23组。A和B为综述评论。A组主要介绍我们自己的工作。其中A1，A2，A3分别评述了我组在遗传密码、分子序列和构象动力学三个方向上的工作。B组中B1至B5是近几年作者应刊物之约写的介绍生物物理学，特别是理论生物物理学的文章。这几篇文章中的观点后来都收入了《物理学家看生命》的小册子。从C至W的21组分成四部分。第一部分（C－E）讨论与生命起源有关的手性和密码问题；第二部分（F－N）讨论基因序列的进化和表达；第三部分（O－S）讨论从序列到构象及构象动力学；第四部分（T－W）为与生物自组织有关的一些问题。为了读者阅读方便，每部分后加了一个评述性的“编注”。论文基本按原样收入本书，但为了节省篇幅，删除了一些重复。约$\frac{1}{3}$的论文只编入了摘要。个别地方更正了原稿文字中的错误和不妥，有少数几篇论文是初次发表。由于我们组的工作还在继续深入，目前这个文集只是近十年工作的进展报告，希望若干年后有续集问世。

关于生命科学中观念和理论的地位，以及理论工作者如何在生物研究中发挥作用，Crick有过精辟的叙述。“在理论生物学的前进道路上总是充满了陷井。”“优美性以及高度抽象的数学形式所体现的简明性在物理学中是有用的指南，但在生物学中这些智力工具可能是无益的。因此，生物学的理论工作者必须从实验证据中（不管它们是多么含糊不清）取得指导。”所以理论生物物理学应更多地注视从实验资料中总结规律和进行唯象研究。事实上，如果把着眼点转移至此，那么大千世界，芸芸众生，精彩纷呈的生命现象提供了多么广阔的天地供理论家们去驰骋啊！

最后作者要感谢合作者——我的历届硕士生和博士生所作的大量关于序列的统计工作和很多问题上的讨论，它们对于完成本集的论文是十分重要的。作者还要感谢国家和内蒙古自然科学基金的多项支持。

作者于内蒙古大学

1995年8月

理论生物物理学论文集

COLLECTED WORKS ON THEORETICAL BIOPHYSICS

目　　录

第二部分

第三部分

A1

遗传密码的逻辑*

罗辽复

第一部分　遗传密码的突变危险性

《科学》编者按：遗传密码是自然界的一项伟大创造。理论物理学教授罗辽复引进密码字典突变危险性的概念，成功地导出密码的简并规则、简并多重态的分布及排列规则，受到国际同行关注。

遗传密码是自然最伟大的创造之一。在密码中蕴藏了"生命机器"工作的重要原理，包含了生命形成和进化的丰富信息。1943 年，薛定谔(Schrödinger)总结了德尔布鲁克(Delbrück)噬菌体实验，提出遗传物质——基因包含于微观体积(10 倍于原子距离的线度)中，建议把大分子看作一种非周期固体，它含有极大量信息，可作为基因的物质基础。但是，遗传信息存贮在哪一种生物大分子中，早期人们较多地把注意力集中在蛋白质而忽略了核酸。1943 年埃弗里(Avery)等发现有毒性的 S 型肺炎球菌 DNA 和无毒性的 R 型肺炎球菌混合，能使后者转变为 S 型，因而首次证明了 DNA 具有改变细菌遗传的能力。1950 年赫尔希(Hershey)和蔡斯(Chase)在大肠杆菌噬菌体 T_2 的核酸上标记^{32}P，蛋白质外壳上标记^{35}S，感染大肠杆菌后，发现仅^{32}P 注入细菌。用示踪原子跟踪噬菌体感染的过程，确定无疑地证明了 DNA 是遗传物质基础。这个时期一系列重要的实验和理论结果的发表，促使了 DNA 双螺旋模型的诞生。一是查古夫(Charguff)测定了 DNA 的 G 和 C 含量和 A 和 T 含量分别相等，打破了 1:1:1:1 的 4 核苷学说；二是泡令(Pauling)提出了蛋白质由氢键作用形成 α 螺旋的观点；最后也是最重要的是富兰克林(Franklin)和威尔金斯(Wilkins)从 X 射线衍射分析中获得了 DNA 晶体结构资料。在以上工作基础上，1953 年沃森(Watson)和克里克(Crick)提出了 DNA 的双螺旋模型，宣告了一个富有生命力的新科学——分子生物学的诞生。

作为一个兴趣广泛的核物理和天体物理学家，伽莫夫(Gamow)很快注意到 DNA 结构模型可能具有的深刻涵义。1954 年，他提出了由 DNA4 种碱基来编码 20 种氨基酸的遗传密码的思想：蛋白质在 DNA 表面上合成，氨基酸直接位于双螺旋的沟中，在 4 个碱基构成的菱形小孔中放置一个氨基酸("金刚石密码")。后来克里克在评论中说："伽莫夫工作的重要性在于，这是一种真正抽象的密码理论，设有那些冗赘而不必要的化学细节，尽管他的基本观念——认为双链 DNA 是蛋白质合成的模板——显然是完全错误的。"从密码的具体性质看，伽莫夫的密码是交迭的 4 联体，而后来实验证明遗传密码是非交迭的(一个碱基的变化只会引起一个氨基酸的变化)三联密码。因此，伽莫夫密码的具体形式是错误的。

密码的简并规则

密码的破译是通过蛋白质的生物合成工作而获得解决的。1961 年尼伦伯格(Nirenberg)

* 本文第一部分发表于《科学》第 42 卷第 3 期 P. 187－190(1990 年)

	U	G	A	G	
U	Phe	Ser	Tyr	Cys	U
				Trp	C
C	Leu	Pro	His	Arg	A
			Gln		G
A	Ile	Thr	Asn	Ser	U
	Met		Lys	Arg	C
G	Val	Ala	Asp	Gly	A
			Glu		G

图 1 标准密码表。每一个氨基酸(和终止密码)的几个密码子都处在紧邻位置上,它们按一定的简并规则出现。

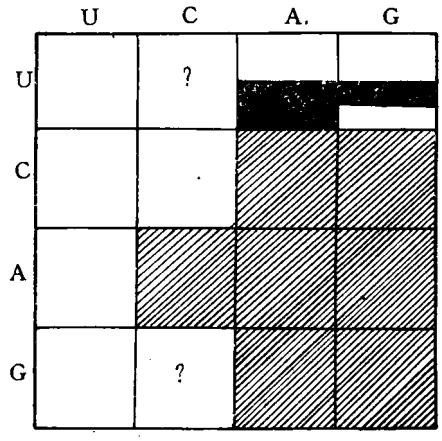


图 2 密码表的亲水畴和疏水畴。阴影区为亲水氨基酸,非阴影区为疏水氨基酸。其中"?"号表示氨基酸分类和反密码子分类不一致。

和马太埃(Matthaei)开始用人工合成的 RNA 作为模板进行无细胞蛋白质合成。他们用 Poly—U 得到苯丙氨酸。以后尼伯伦格(1965 年)和库伦拉(Khorana,1966 年)研究了各种 3—核苷酸 t—RNA 和氨基酸的复合,从而建立了 64 个密码子和 20 种氨基酸的对应关系,完全破译了密码。

64 个密码子对应 20 种氨基酸,其间必存在多个密码子对应一个氨基酸的简并情况。密码的简并规则是怎样的?密码的普适性背后所遵循的逻辑是什么?

密码是一个信息系统,碱基突变是这个系统的固有噪声,它将导致系统的不稳定性。作为长期历史发展的现有密码,应是最能抗干扰的,最稳定的。也就是说,对应于每一种假想的氨基酸—密码子对应关系 C(密码字典 C),可定义一个突变危险性函数 $MD(C)$。从 $MD(C)$的极小化可导出标准字典 C,其中也包括了现有密码的简并规则。当然,由于生物系统的复杂性,它的种属差异和历史性(与历史形成的进化途径有关),一般地说极值只能在放松的统计的意义下理解。我们把这称为近似稳定性原理。所谓突变危险性,它包含突变频率(指改变氨基酸的那部分碱基突变)和突变造成的选择性死亡率两个因素。对于密码简并问题,只须考虑前一个因素。

考虑简并度 j 的密码子多重态,它对应于某种氨基酸或终止密码,突变危险性 $MD(j)$由各种可能的单碱基突变贡献叠加而成。单碱基突变用 4 个参数描述,转换突变系数 u、颠换突变系数 v 和附加的摆动突变系数 W_u、W_v。后者只对密码子的第三个碱基才有,因为克里克摆动也等价于一种突变,而这种摆动发生于第三个碱基上。对于每一种假想的密码子分布,多重态的总 MD 值等于其中各密码子的三个碱基的 u、v、W_u 和 W_v 之和。显然,如果把各个密码子分布在不相关的位置上,它的突变危险性 $MD(j)$就高;反之,如果把它们放在密码表的适当紧邻位置,突变危险性 $MD(j)$就低。因为后一情形中,多重态内各密码子有较多相互突变的机会,而它们对这个多重态的突变危险性 $MD(j)$没有贡献。这样我们就可把多重态的 $MD(j)$看成密码子假想分布的函数,并且把它按由低到高的顺序排列成能级一样的表。最低级所对应的分布是最稳定的,其上有第一激发级、第二激发级等。在确定性理论中,所有多重态都应处于最低级。但是当第一、第二激发级和最低级相距很近(级隙很小),由于涨落的作用,也可使 MD

取略大于极小的值。

根据以上原则对各种简并度的氨基酸(和终止密码)进行计算，我们发现，在参数选取满足

$u>2v$，　$W_s>2u$，

$W_v>u-v$，

$u+W_s>v+W_v$

的条件下，

1) 现有密码字典中 18 种氨基酸的密码子分布皆使 MD 取极小值(最低级)，而且最低级上方的级隙约为极小值的 20%以上；

2) 终止密码子的分布也使 MD 取极小值；

3) 6 重态的最低级、第一激发级、第二激发级之间的级隙很小(约≤15%)，故除了最低级被占据(Leu)外，第一激发级和第二激发级分别为 Arg 和 Ser 所占据。

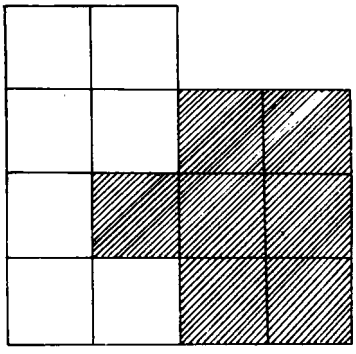


图 3　将 14 个二核苷方块分成相等的两区。这种划分具有突变危险性是极小的，和实验资料也基本一致。

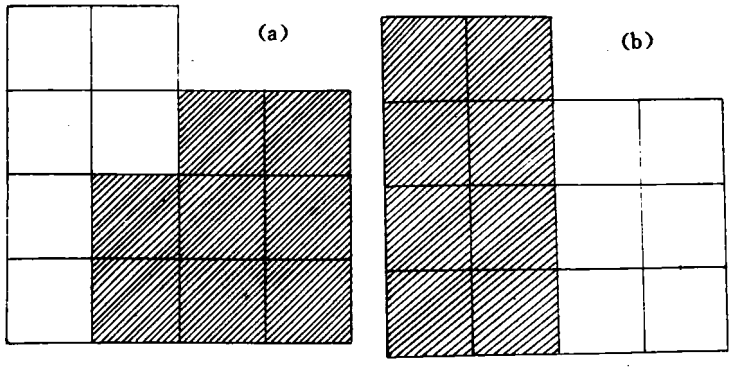


图 4　(a)将 14 个方块按 8+6 分类成亲水区和疏水区。这种划分和实际反密码子分类一致，并具有近似的 MD 最小值。
(b)将 14 个方块按 8+6 分类成亲水区和疏水区。这种划分的 MD 严格极小，不过和实验资料不一致。

简并多重态的分布

上面我们从密码字典的局部(每一多重态)稳定性导出了密码的简并规则。但是，现有标准密码的总体是否稳定？标准密码字典中简并度 $j=1,2,3,4,6$ 的多重态(氨基酸)数分别为

$n_j=2,9,1,5,3$

个。上述问题亦可表达成：能否从密码字典的总体稳定性导出这种分布？事实上，密码字典的突变危险性等于各多重态的突变危险性之和。为了求出分布$\{n_j\}$，必须考虑密码子相互作用。考虑保持氨基酸数和密码子数守恒的过程

$$A_i+A_j\longrightarrow A_k+A_l \qquad (i+j=k+l)。$$

A_i 表示简并度 i 的氨基酸。比较左右两方的 MD 值，MD 下降的过程是稳定性有利的过程；反之，是稳定性不利的过程。因此，密码子相互作用将导致氨基酸在密码子字典中的重新排列，以减小 $MD(C)$值，增加密码的稳定性。计算给定密码子总数 $N(=i+j=k+l)$的一对氨基酸的总 MD 值，结果表明，基态和离开基态很近的第一激发态中，$j=1,2,3,4,6$ 比 $j=5,7$ 更频繁

地出现。这个结果不太依赖于参数的具体取值。这就解释了为什么标准密码中简并度 1,2,3,4,6 的多重态经常出现,而简并度 5,7 的多重态没有发现。看来这是趋于降低 MD 值的密码子相互作用的结果。至于为什么标准密码中没有 $j=8,10$ 的多重态?这可能因为高简并度的同义密码子一般对应于高丰度的氨基酸,而在蛋白质中似乎还不需要简并度为 8,10 这样高丰度的氨基酸。另外,为了扩大蛋白质的信息量,也必须尽量减少高简并度的氨基酸。

除了代表氨基酸的密码子相互作用外,和终止密码子(记作 T_k)相联系的相互作用也是可能的。这类过程有

$$A_i \longrightarrow A_j + T_k \qquad (i=j+k),$$

$$A_i + T_l \longrightarrow A_j + T_k \qquad (i+l=j+k),$$

等等。在线粒体中发现用 AGA/G(通常编码精氨酸)来编码终止密码的非标准运用就可用前一过程来解释,即

$$A_6(\mathrm{Arg}) \longrightarrow A_4 + T_2$$

这个过程中 MD 值减小 $4(W_a-v)$。既然精氨酸的丰度并不需要使用 6 度简并密码子,在进化中它部分突变为终止密码子是可能的。线粒体中还发现用 UGA(通常为终止密码)来编码色氨酸的非标准应用,它可用后一过程来解释,即

$$A_1 + T_3 \longrightarrow A_2 + T_2$$

其中 MD 值减小 W_a-2W_v。当 $W_a>2W_v$,这也是可能的。

密码表的疏水-亲水畴

以前的讨论中只考虑碱基突变对密码字典稳定性的影响。事实上除了单碱基突变外,氨基酸替代造成的选择死亡也是产生密码不稳定性的重要因素。不妨把略去后一因素的理论称为独立密码子近似。在研究密码子多重态的简并规则和密码表中多重态数分布时,独立密码子近似就够了。但在研究氨基酸在密码表中的具体排列时,还须考虑氨基酸替代造成的疏水性变化引起的附加突变危险性。这里需要区分两种情况,如果氨基酸替代不改变疏水亲水性,突变危险性就小;反之,如果这种氨基酸替代改变疏水亲水性,突变危险性就大。不妨把前者称为组内突变,后者称为组间突变。显然,和一般的组间突变相比,对于组内突变来说,相当于已经进行了 MD 的某种极小化。为了描述这种效应,可令组内突变系数 u 和 v 分别为 $u-y$ 和 $v-z$,而组间突变系数保持 u 和 v 不变。当 $y,z \ll u,v$,$MD(C)$ 可按 y,z 展开,零次项为独立密码子近似,展开到一次,

$$\begin{aligned} MD(C) &= MD(C)\big|_{y=0,z=0} + \quad \text{含 } y,z \text{ 的一次项} \\ &= \sum_j n_j MD(j) + ay + bz, \end{aligned}$$

a 和 b 依赖于字典 C 中氨基酸的排列。因此从 $MD(C)$ 的极小化可求出氨基酸在密码表中的具体排布。

关于氨基酸的疏水-亲水程度,大别为二类:

疏水类:Phe,Leu,Ile,Tyr,Trp(以上强疏水);Met,Val,Pro,Ala,Cys(以上弱疏水)。

亲水类:Thr,His,Gln,Glu,Gly(以上弱亲水);Ser,Asn,Lys,Asp,Arg(以上强亲水)。

若按反密码子 3′-二核苷的疏水性进行分类,也分别为二类:

疏水类:UU,CU,GU,UC(以上强疏水);AU,CC,UA,UG(以上弱疏水)。

亲水类：AC，GC，CA，CG（以上弱亲水）；AA，GA，AG，GG（以上强亲水）。

两种分类除个别位置外，基本上是一致的。关于氨基酸的疏水性，不同作者的分类略有不同。这里根据的是 JT－F. Wong，Microbiol. Sci. 5(1988)174。在密码表上，亲水氨基酸和疏水氨基酸明显地分成两个相连的区域。我们把密码表上这两个区域称为亲水畴和疏水畴，并希望从 *MD*(*C*)的极小化导出这种分布。

64 个密码子在密码表上组成 16 个二核苷方块。如果除去终止密码所在的两个方块 UA 和 UG（设它们已经给定），表中还剩有 14 个二核苷方块。现在的问题是：若将 14 个方块分成两区，一半是亲水，一半是疏水，问密码表应如何分解，才能使 *MD*(*C*)极小？考虑各种假想的分布，我们发现只有当亲水氨基酸和疏水氨基酸分别整齐联在一起时，密码字典的突变危险性才是极小的。这个情况和前面研究的密码简并规则类似。当同类（亲水或疏水）氨基酸安排在相互靠拢的位置上，组内突变就占较大的比重，因而密码字典总体的突变危险性就小。在参数选取满足 $5y>6z$ 条件下，具体计算求得的极小分布具有疏水－亲水畴结构，并和实验资料基本一致。

上面的讨论把 14 个二核苷方块分成相等的两个区域——7＋7 分类。而实际密码的亲水——疏水分布是 8＋6 分类。如果按 8＋6 分类的要求进行计算，同样可以从 *MD*(*C*)的极小得到亲水区和疏水区的畴状结构。不过标准密码字典实际分布的 *MD* 值并不是严格的极小值，而只是近似的极小值。$ay+bz$ 比严格极小值高 8%。为什么实际分布不能使 *MD*(*C*)取严格极值？这和密码的历史形成有关。最早的氨基酸有 Gly，Ala，Ser，Asp，Glu 和 Val。如果一开始二核苷方块 GG，GA，GC，GU 已分别被 Gly，Ala，Asp/Glu，Val 所占据，以后的氨基酸－密码子对应关系就要考虑到这个初始条件。最优化也只能是在给定历史框架下的最优，是近似的最优。因此最后形成的标准密码的 *MD*(*C*)只取近似极小值。

简短的结语

密码的普适性是生命现象中最有兴趣和最令人困惑的问题之一。这种普适性背后所包含的数学关系和物理原理，更是很多研究者关心的问题。弄清密码的逻辑，不仅具有深刻的理论兴趣，而且也可能为基因工程开辟新的途径。因为突变产生的新型氨基酸可能改变密码，从而为人工设计制造新功能蛋白质开辟无限广阔的前景。本文综述了我们最近的工作，通过引进密码字典突变危险性（*MD*）的概念，可以对各种假想的密码表（氨基酸－密码子对应关系）进行计算，求出 *MD*。利用 *MD* 的极小化，成功地导出了密码简并规则和简并多重态的分布及排列规则。

一个基本问题在于是否存在某种极值原理？我们认为极值性反映了稳定性原理。密码的普适性已经说明了稳定性在起作用。关于 *MD* 极小化的具体计算也说明这个原理是存在的。但是这种极小化和稳定性并不排斥历史上的偶然因素（frozen accident），因此是近似的稳定性。显然，近似稳定性原理并不因为它的近似性质而丧失作为一条基本原理的意义。这个情况犹如物理学中大量的对称性是破缺对称性，而不是严格对称性一样。

伟大的数学家希尔伯特曾经说过："认识自然和生命是我们崇高的任务。"本文从一个简单的统一的原则出发，进行了严格的数学推导。对于生命这样复杂的系统，理论计算和实验资料竟能如此好的符合，是颇出乎意料的。这充分说明，数学的、理论物理学的方法可以适用于生物学。普里戈金（Prigogine）在谈现时自然科学的趋势时说："物理学要向复杂的方向发展，而生

物学要向基本的方向发展。"关于生命现象的探索，把它提到数学和物理的深度和基本水平上，就有可能长驱直入。二千多年前，柏拉图在他学派的门上写了"不懂数学者免进"的戒条。今天研究生命科学时，难道不应该从这句富有哲理的话中获得启发吗？

第二部分　密码的对偶性*

密码的对偶性及氨基酸的疏水－亲水性

密码的对偶性及氨基酸的疏水——亲水性　据中国哲学和传统医学理论，生命是两种矛盾因素——阴阳的统一。这种阴阳对偶性不仅在整体的细胞层次上表现出来，也在更基本的分子层次上(氨基酸和核苷酸)表现出来。蛋白质的功能决定于它的构象，而在蛋白质折迭中，氨基酸的极性起着关键作用。球蛋白的结构原则是：亲水性残基呈露于表面而疏水性残基理藏在内部。因此，亲水性和疏水性是反映氨基酸基本特性的一种阴阳对偶性。氨基酸由核酸通过遗传密码指导生成，生命信息凝聚于核酸之链中。作为信息储存基本文字的核酸是否也有这种阴阳对偶性的反映呢？实际上，四种核苷酸按化学结构分成嘌呤和嘧啶，按氢键结合关系分成两种 Watson－Crick 对，因此存在 U↔C、A↔G 的结构不变性和 U↔A、C↔G 的对称关系。为了表达这种对称性，可以假定核酸的四种碱基分别用取阴阳二态的双线(上线和下线)表示。阴态记作－－，阳态记作——。利用上线的二态进行嘌呤嘧啶分类，下线则用于嘌呤或嘧啶类内的再分类。而 W－C 对则发生于阴阳之间。故有表示如下：

U ═　　C ═　　G ═　　A ═

UA 是阴阳对称的，CG 也是阴阳对称的，它们之间可结合成对。我们把碱基的这种数学表示称为核苷酸对偶假设。在此表示中，U 阳性最强，A 阴性最强，C 和 G 处于中间。考虑到上线有较大权重，C 偏阳，G 偏阴，有趣的是碱基的这种阴阳顺序与抗辐射电离的稳定性顺序相一致。

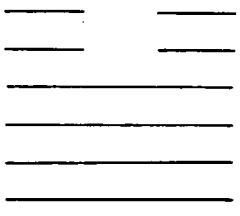


图 1

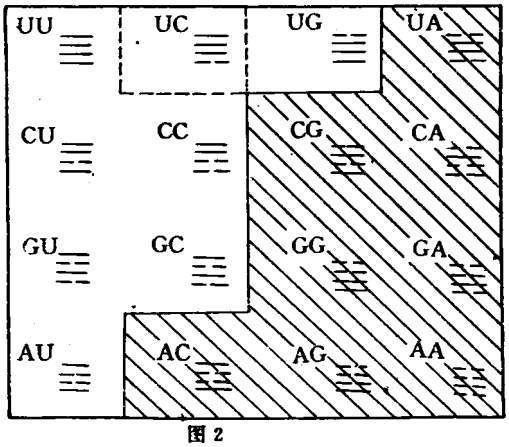


图 2

遗传密码是核苷三联体，因此每一密码子须用六线图表示。我们规定，第一个核苷对应的双线画在中间(即 3,4 线)，第二个核苷的双线画在其上下侧(即 2,5 线)，第三个核苷双线画在最外侧(即 1,6 线)。例如，色氨酸 UGG 的图见图 1。鉴于密码的前二字母在决定氨基酸性质方面更重要，上述六线图中中间四线是基本的，把中间四线画在密码子字典上，如图 2 所示。

---

* 第二部分以《遗传密码结构》为题发表于《百科知识》1992 年第 2 期 P. 63

现在来讨论氨基酸的疏水－亲水性——这是生命的普适阴阳对偶性在生物大分子水平的表现。基于核苷酸的阴阳对偶性分析和图 2，我们自然地推测：密码四线图(图 2)中包含的阳线(——)越多，则疏水性越强；反之，阴线(－－)越多，则亲水性越强。如两种线数相等，阴线偏上者为亲水，阳线偏上者为疏水。这样就决定了图 2 每一方块的疏水性。唯一不定性发生于 AU 和 UA。若进一步考虑密码子第二位比第一位具有更大的决定性，则 UA 应属于亲水，AU 应属于疏水。于是图 2 便分成两区域：一个是亲水(用斜线表示)，一个是疏水。

从实验角度讲，氨基酸的疏水性标度有很多不同的确定方法，大别为两类：一是通过氨基酸在水中自由能的测量来定义疏水性，二是通过统计不同残基埋在分子内部的比例来确定疏水性。不同方法确定出的结果往往颇有出入。从理论上讲，疏水性和多种量子力学作用力有关，也包含熵效应的因素。但归根到底归结为氨基酸残基的化学结构。因此，可这样区分疏水和亲水：如侧链末端原子为 NH 或 OH，则氨基酸为亲水的；若末端为 CH 或 SH，则氨基酸为疏水的；如末端为环，只要环中有 NH 或 OH 就是亲水的，否则是疏水的。Gly 情况特殊，其识别部位是肽端的 NH。据此，亲水类有 Arg，Lys，Asp，Glu，Asn，Gln，His，Tyr，Ser，Thr，Gly；疏水类有 Ile，Val，Leu，Phe，Met，Ala，Trp，Cys，Pro。这个分类和新近较普遍采用的几种综合疏水性标度都较接近。将此实验分类和图 1 理论结果比较，符合很好。唯一疑点在丝氨酸，Ser 是亲水氨基酸，但其简并密码子的一部分 UC 所对应的 tRNA 反密码子是疏水的(图 2 中用虚线框住)，这和理论分类中 UC 和 AG 属于不同畴也是一致的。附带指出，从上面的分析途径还可导出一条简单的规律：密码子和它的反义密码子具有相反的疏水性。如 UUC(Phe)是疏水的，它的反义密码子 AAG(Lys)是亲水的。正反义密码子具有相反阴阳性的一项应用是：用和癌基因反义的寡核苷来控制癌变。

密码的对偶性表明核苷酸和氨基酸具有相同的特性——它们都可用"阴阳"来描述。这提供了一条把二者对应起来联系起来的纽带。由此可以解释遗传密码的起源。

当代生命科学正面临巨大飞跃，一个显著的特点是理性的数学的精神之渗入。近代自然科学的先驱者伽里略在论述"宇宙这部宏大的书"时说，"它是用数学语言写成的……没有它们，一个人只能在黑暗的迷宫里徘徊。"这席话对于今天我们阅读生命之书不是同样具有指导意义吗？通过这篇短文，读者还看到了当人们一旦读懂这些数学语言，复杂的结构会变得那么简单明了。正如伽里略所说："大自然作任何事情，都采取最简单的方法。"它在创造遗传密码时，不是也遵循同一原则吗？

参考文献

[1] Luo LF Origins of life 18:65(1988) [本文集 D6]

[2] Luo LF Origins of life 19:621(1989) [本文集 D7]

[3] 罗辽复 内蒙古大学学报 23:395(1992) [本文集 D8]

[4] 罗辽复 李前忠 科学通报 30:1056(1985) [本文集 D3]

[5] 罗辽复 内蒙古大学学报 17:513(1986) [本文集 D5]

A2

生命信息的宝库*

罗辽复

《科学》编者按：核酸序列与蛋白质序列包含了基因表达和调控、蛋白质结构和功能的大量信息。生命的形成、发育和演化，以及新陈代谢、世代更迭等等都包含在这本"天书"之中。

随着50～70年代序列分析技术(氨基酸测序、蛋白质电泳、DNA测序、重组DNA、限制酶方法等)的发展，蛋白质和核酸序列的资料迅速积累。1972年戴霍夫(Dayhoff)等首先收集了蛋白质序列资料，至1978年共搴集约1300个序列共20万个氨基酸。关于核酸，1965～1978的13年中共测定发表了12000个核苷酸。经过4年至1982年底，蛋白质资料已达2000序列33万个残基，而核酸资料则达到809序列100万个碱基。往后速度愈来愈快，1987年全世界基因文库中储存的核酸资料已超过1000万个基因，目前正以700万个碱基的速度进一步积累着。

氨基酸的序列决定了蛋白质的一级结构、高级结构和功能。核苷酸的序列又指导着氨基酸序列的形成，决定着基因的表达和调控。它们决定了包括生命形成和演化以及个体发育、新陈代谢等在内的一切生命活动的基本方面。并且这每一种由长达百、千、万个单位组成的序列都是稳定的，一个微小的变更(例如碱基取代、缺失和插入)都可能引起生命活动的严重障碍。由此可见，这些丰富的序列资料向人类提供了生命系统最重要的信息，展示在我们面前的是一本厚厚的生命之书。然而直到今天，人们只是发现了这本天书的若干片断，还没有完全找到它，离开读懂和理解它就更远了。大自然是如何写下这本生命之书的，我们几乎还是一无所知。40多年前，原子物理学家薛定谔(Schrödinger)提出了"生命是什么?"这个著名问题后，预见到生命信息贮存在类似于非周期固体的大分子中，这个观点开辟了遗传学的信息主义学派。分子生物学的发展告诉我们，生命是高度组织起来的物质，是物质的最高级运动形式。然而生命不仅是物质和运动，生命还是信息。必须从信息的角度去探究生命的底蕴。核酸和蛋白质序列是生命信息的宝库，如何开发这个宝库，是当代自然科学面临的一大挑战。关于这些序列资料在自然科学史上的地位，不妨打一个比方。众所周知，19世纪末原子分子光谱曾经对近代物理学起了多么大的推动作用，难道人们不正是从氢原子特异的谱线系的研究中逐步领悟到原子构造的秘密，并建立起微观运动的规律——量子力学吗?对于19世纪末的自然科学家来说，可见光范围的10万条原子谱线，光谱中那些确定的数量关系是那样的奇异，好象原子在给人类打电报，而人们却不理解它的意义。今天类似的情况又重复出现。核酸、蛋白质序列中包含的极为丰富的信息内含，对于任何一个试图探究生命奥秘的人，不也感到同样的困惑和奇异吗？有理由相信，生命之书开始被读懂之日，自然科学会经历怎样的飞跃！

氨基酸和核苷酸序列的研究，不仅对于生命科学极端重要，而且对于物理学和一般自然科学也非常重要。探索复杂性是当代物理学和自然科学的一大前沿。人们企图建立一个复杂系统的普遍理论。但是这个问题要深入下去，还离不开对具体系统的分析。自然界之复杂无过于

---

* 本文发表于《科学》第42卷第4期P271－274(1990年)。此处作了一些补充。