关于遗传语言熵减进程的研究

胡 长 安
(湖北医学院)

在十九世纪，关于进化的概念，出现了两种相互对立的论断：一种是以热力学第二定律为根据，认为，在某个遥远年代发生的结构，随后将逐步地趋向于混乱状态．这实质上反映了关于宇宙事物集合中存在的离解因素会导致事物连续解体的进化观念．另一种是关于生物学与社会学中的进化观念，与上述观念相反，这一种进化观念认为，在事物集合中存在着的凝聚作用或自组织作用，将导致事物组织程度的不断增加，导致越来越复杂的新的结构的产生[1]．

这两种学术观点在一些学术领域中的相互渗透，导致人们重新探讨了物理学所持有的经典观念，发展了一种适用于新的问题探讨的数学方法．这促进了某些交叉学科的发展，也给生物学传统的研究方法开辟了新的研究途径．例如，人们利用统计热力学中熵这一严密的科学概念，定量地研究了支配生命成长与发展的核酸分子结构的进化趋向．研究结果不仅符合生物学者从形态结构方面研究得出的进化观点，而且给出了生物学者用其它方法所不可能取得的新的论断，为生物学的研究开创了新局面，本文重点介绍这一研究的观点方法及其初步成果．

自从申农(C. E. Shannon)在通信工程系统方面，利用可由统计热力学中玻耳兹曼熵(即玻耳兹曼关系)导出的公式[2]

$$S = \kappa \ln W = \kappa \ln \frac{N!}{\prod_i N_i!}$$

$$= -\kappa \ln \frac{\prod_i N_i!}{N!}$$

$$= -\kappa \sum_i \ln N_i! + \kappa \ln N!$$

$$= -\kappa \left[\sum_i N_i \ln N_i - \sum_i N_i - N \ln N + N \right]$$

$$= -\kappa \left[\sum_i N_i \ln N_i - \sum_i N_i \ln N \right]$$

$$= -\kappa \sum_i N_i \ln \frac{N_i}{N}$$

$$= -\kappa N \sum \frac{N_i}{N} \ln \frac{N_i}{N}$$

$$= -\kappa N \left[\sum_i P_i \ln P_i \right], \qquad (1)$$

提出一个所谓平均信息熵 H，即

$$H = -\sum_i P_i \ln P_i = \frac{S}{N_\kappa}, \qquad (2)$$

并用它来描述信息源（例如电传打字机字母集合、莫尔斯符号表等）发射消息的不确定性，建立了以统计的数量形式，有效地研究消息在通信工程系统中传递过程的可靠性与经济性的数学理论[3]．之后，人们进一步把通信工程系统中的信息源理解为只不过是自然界和社会中各种随机事件集合的特例，从而把研究问题的观点和方法，由工程系统延拓到其它非物理系统，例如生命系统方面的这类集合．这类集合的统计特征是：集合中的元素 $x_1, \cdots, x_i, \cdots, x_n$ 各自具有的对应概率为 $P(x_1), P(x_2), \cdots, P(x_i), \cdots, P(x_n)$，可以写成

$$[X] = [x_1, x_2, \cdots, x_n],$$

$$[P(x)] = [P(x_1), P(x_2), \cdots, P(x_n)]. \qquad (3)$$

其中概率满足归一条件

$$\sum_{i=1}^{n} P(x_i) = 1. \qquad (4)$$

这一集合的不确定性由公式(2)量度. 通常对(2)式改用以 2 为底的对数,以后我们把它简记为

$$H = -\sum_{i=1}^{n} P(x_i)\log P(x_i). \qquad (5)$$

它的单位是比特/每符号,考虑到集合在各种影响下有着不确定性减少的过程，为了对这些不确定性减少的量进行定量研究，人们按下述方式定义不确定性减少的量为信息量,即

$$I = -\Delta H = -(H_e - H_0) = H_0 - H_e. \qquad (6)$$

其中 H_0 代表原有信息源的信息熵，H_e 代表变化过程发生后新的信息源的信息熵. 例如，由(5)式所描述的信息源发射出一消息(一个已知消息所具有的概率等于 1) 所减少的不确定度，用信息量 I 表示,为

$$\begin{aligned} I &= -(H_e - H_0) \\ &= -\left[\log 1 + \sum_i P(x_i)\log P(x_i)\right] \\ &= -\sum_i P(x_i)\log P(x_i), \qquad (7) \end{aligned}$$

这就是统计热力学中信息论部分的基本公式之一[2]，人们把它用作量度事物有序性的量. 一般，人们把信息熵与信息量用作存在于事物集合中如下的一些相互对立性质的量度，即关于事物集合的随机性与确定性、任意性与组织性、自由与约束、混乱无章与法则以及熵与负熵等对立性质的量度. 在这里，信息量或者简称信息，此词较之我们日常工作中所使用的信息一词,有着更广泛、更深刻的含义：凡是导致事物集合的不确定性、任意性、无组织性、熵等减少的活动过程;或者反之,凡能导致事物集合的肯定性、组织性、法则性与负熵增加的活动过程，对它们活动结果的程度，人们采用信息量这一统一标尺，给以数量多少的量度. 人们就是利用这一基本观点和方法，研究了生物进化过程中，关于不同等级生物内核酸分子结构有序性的变化程度.

在现阶段生物学研究中,人们已经知道,每种生命都具有一种储存与处理遗传信息的系统,以便在它们成长与代代相传过程中,能准确地复制它们自身. 遗传信息就是储存在核酸分子结构之中. 生命系统的基本要求之一是，它应能准确地复制(繁殖)它们自身. 如果这种复制过程不准确,这种生命系统将会消亡. 因此，生命系统中处理遗传信息机制的可靠程度的差别,标志生命系统进化历程的趋向.

生物体中有些大分子例如核酸分子，就具有公式[3]所示的统计特性. 核酸分子有两种，一种叫脱氧核糖核酸,通常用符号 DNA 表示;另一种叫核糖核酸,用符号 RNA 表示. 它们都是由四种不同的核苷酸构成的序列结构. 其中 DNA 是由四种不同的核苷酸构成的、绕某一中心轴形成的螺旋梯状结构，它的四种核苷酸按所含的碱基成分不同而相区别. 这四种碱基叫做腺嘌呤、鸟嘌呤、胞嘧啶和胸腺嘧啶,以后我们分别用符号 A, G, C, T 表示. RNA 与 DNA 稍有不同，它一般是单链式系列结构,它的戊糖是核糖，而在碱基中它以尿嘧啶 (U) 取代胸腺嘧啶 (T),如表 1 所示.

表 1

结构 \ 核酸		RNA	DNA
核苷酸	酸	磷　酸	磷　酸
	戊糖	核　糖	脱氧核糖
	碱基	腺嘌呤(A)	腺嘌呤(A)
		鸟嘌呤(G)	鸟嘌呤(G)
		胞嘧啶(C)	胞嘧啶(C)
		尿嘧啶(U)	胞腺嘧啶(T)

实验表明,在哺乳动物中，DNA 内的碱基对数超过 10^9 对，在细菌中的 DNA 最少也有 10^4 对，而一本较厚的书中字数也不过几十万字,这是一种饶有意义的结构. 所以,即使人们在核苷酸序列中 (以后将按通常所说的碱基系列来表示核苷酸序列) 发现存储有指导蛋白质合成的核苷酸三联体的遗传密码的伟大成就之后,人们将不会停滞在这一研究水平上.

在七十年代前后，生物医学信息论工作者

利用前述观点方法，对核酸分子的碱基系列进行研究时，发现碱基系列中除存储有遗传密码(以后我们称之为初级遗传信息)之外，还叠加有具有语言特征的高级遗传信息（以后我们称为遗传语言）. 亦即人们发现，在 DNA 中，由四种不同碱基集合

$$S = [A, T, C, G] \tag{8}$$

构成的碱基系列，也像由英文字母表(包括代表标点符号的空档)构成的英文文章一样，具有相同的统计结构，即都是一种马尔柯夫链型结构. 在它们的相邻字母（或碱基）之间，存在着一定的统计相依的关系. 具体地说，由字母表

$$[x_1, x_2, \cdots, x_n] \tag{9}$$

构成的一随机系列

$$\cdots x_2x_1x_5x_2x_nx_7x_1x_2x_4\cdots \tag{10}$$

中，在某一字母，例如字母 x_1 之后出现字母 x_i ($i = 1, 2, \cdots, n$) 时，存在着一定的过渡概率(或条件概率)：

$$P(x_i|x_1). \tag{11}$$

这种序列称为一阶马尔柯夫链；而当存在的过渡概率为(例如)

$$P(x_i|x_1x_2\cdots x_m) \tag{12}$$

时，则称相应的序列为 m 级马尔柯夫链. 研究表明，象英文、法文等的文章结构都是具有这种马尔柯夫链型的统计结构[4]. 在七十年代前后，人们研究了六十多种有机体的 DNA (或RNA) 的碱基系列，证明都具有一级马尔柯夫链式结构. 在八十年代前后进一步的研究初步表明，有的病毒的 DNA 碱基系列是二级马尔柯夫链式结构[5,6].

我们先概略地说明，生物医学信息论工作者怎样利用这样的统计结构，研究调制在碱基系列之上的遗传语言（人们认定是生命过程中的一种程序控制指令)的“语法”严密程度的方法. 对于“语言”的这种统计结构，可以分为“字母”组成方面的统计结构与“字母”系列方面的统计结构. 对于字母组成方面的统计结构的研究方法是这样的：一个打字机的字母表，在未使用之前，各个字母的等概率性将可以反映字母表的最大熵，即完全无序性[7]；而一当它打出一篇足够长的、有一定语法结构的文章之后，这篇文章中字母 x_i 的组成百分比代表了字母表中字母 x_i 的出现概率 $P(x_i)$. 由此，我们可以求出一篇文章具有型如(3)式的统计特征，就可对它按前述方法进行研究. 关于字母序列方面统计结构的研究方法，我们以一级马尔柯夫链为例来说明. 首先，将字母表中的字母，组合成以二个字母为单元(我们称之为二元码组)的新的二元码组集合. 以碱基集合(8)式为例，我们组成一个所谓二重扩展信息源 S^2，即

$$S^2 = [AA, AT, AC, AG, TA, TT, TC, TG, CA, CT, CT, CG, GA, GT, GC, GG], \tag{13}$$

相应各二元码组 x_ix_j (此处 $i, j = 1, 2, 3, 4$)的概率为

$$P(x_ix_j) = P(x_i)P(x_j|x_i). \tag{14}$$

如果 $P(x_j|x_i)$ 与 $P(x_i)$ 为已知，则可求出 $P(x_ix_j)$，因而这又是一个具有型如(3)式的统计集合，又可以按前述观点方法进行研究.

我们以 Phlei 球菌中 DNA 的碱基系列为例，说明这一研究方法及其定量结果. 对于 Phlei 球菌中 DNA 而言，具体字母为

$$x_1 = A, \ x_2 = T, \ x_3 = C, \ x_4 = G,$$

实验表明，它们的概率分别为

$$P(A) = 0.164, \ P(T) = 0.162,$$
$$P(C) = 0.337, \ P(G) = 0.337, \tag{15}$$

而沿 DNA 链中一定方向的四个碱基之间过渡概率的实验值，即关于 $P(A|A)$, $P(T|A)$, $P(C|A), \cdots, P(T|G)$, $P(C|G)$, $P(G|G)$ 的实验值，如表 2 所示.

表 2

$P(x_j\|x_i)$		x_j			
		A	T	C	G
x_i	A	0.146	0.075	0.378	0.491
	T	0.194	0.157	0.279	0.370
	C	0.189	0.182	0.268	0.361
	G	0.134	0.187	0.414	0.265

其中 x_i, x_j, 按前述，依序取 A, T, C, G.

因此，Phlei 球菌中 DNA 组成方面的有

序性程度，可以定量地如下求出．DNA 组成的完全无序性可由最大熵 $H_{max}(S)$ 表示，它由集合中元素的等概率分布给出[7]，即

$$H_{max}(S) = -\sum_{i=1}^{4} P(x_i)\log P(x_i)$$

$$= -\sum_{i=1}^{4} \frac{1}{4}\log\frac{1}{4}$$

$$= \log 4 = 2 \text{ 比特/碱基．} \quad (16)$$

而实际的信息熵值，则按(15)式给出的实验值，求出如下：

$$H(S) = -\sum_{i=1}^{4} P(x_i)\log P(x_i)$$

$$= -(0.164\log 0.164 + 0.162\log 0.162 + 0.337\log 0.337 + 0.337\log 0.337)$$

$$= 1.910 \text{ 比特/碱基．} \quad (17)$$

可以看出，实际的信息熵值偏离最大熵值．根据(6)式，在组成方面的有序性，可以由相对于完全无序性的偏离量

$$D_1 = H_{max}(S) - H(S)$$

$$= 2 - 1.910$$

$$= 0.090 \text{ 比特/碱基} \quad (18)$$

给出，其中 D_1 称为一类偏离．

另一方面，关于信息源中各元素之间由完全自由，偏离到有一定程度的相互约束这类有序性的求法，我们也以 Phlei 球菌中 DNA 碱基的二重扩展信息源(13)为例，定量说明如下．设碱基系列中各个碱基为完全相互独立时，则必有

$$P(x_ix_j) = P(x_i)P(x_j|x_i) = P(x_j)P(x_i),$$

例如 $P(AA) = P(A)P(A)$，$P(AT) = P(A)P(T),\cdots$．利用(15)式中数据，我们求出碱基间为相互独立时、相应二重扩展信息源中各二元码组的概率 $P(x_ix_j)$ 值，即 $P(AA) = P(A)P(A)$，$P(AT) = P(A)P(T),\cdots$，等的数值，列于表 3．

表 3

$P(x_ix_j)$	A	T	C	G
A	0.0262	0.0267	0.0543	0.0548
T	0.0267	0.0272	0.0553	0.0548
C	0.0543	0.0553	0.1122	0.1132
G	0.0548	0.0558	0.1132	0.1142

由表中数据，可以求出相应的信息熵 $H^{I}(S^2)$（其中上角标 I 表示各元素之间为相互独立）值为

$$H^{I}(S^2) = -(0.0262\log 0.0262 + 0.0267\log 0.0267 + \cdots + 0.1132\log 0.1132 + 0.1142\log 0.1142)$$

$$= 3.819 \text{ 比特/二元码组．} \quad (19)$$

考虑到 Phlei 球菌中 DNA 碱基系列为一级马尔柯夫链，因而二元码组的概率应为 $P(x_ix_j) = P(x_i)P(x_j|x_i)$，例如 $P(AA) = P(A)P(A|A)$，$P(AT) = P(A)P(T|A),\cdots$，所以相应概率应以(15)式和表 2 中所示数据代入求出．表 4 列出所求结果．

表 4

$P(x_ix_j)$	A	T	C	G
A	0.024	0.031	0.064	0.045
T	0.012	0.026	0.061	0.063
C	0.063	0.045	0.090	0.139
G	0.065	0.060	0.122	0.090

由此我们可以求出二重信息源 S^2 中二个碱基之间存在约束时的信息熵 $H^{D}(S^2)$（其中上角标 D 表示各元素之间是统计相依）值，即

$$H^{D}(S^2) = -(0.024\log 0.024 + 0.031\log 0.031 + \cdots + 0.122\log 0.122 + 0.091\log 0.091)$$

$$= 3.792 \text{ 比特/二元码组．} \quad (20)$$

一般地说，$H^{I}(S^2)$ 大于 $H^{D}(S^2)$[6]．因此，根据(6)式，我们以系列中元素之间相互独立的二重扩展信息源的熵为标准，元素之间彼此相关的系列所具有的有序性程度，可以由偏离量

$$D_2 = H^{I}(S^2) - H^{D}(S^2)$$

$$= 3.819 - 3.792$$

$$= 0.027 \text{ 比特/二元码组} \quad (21)$$

表示，其中 D_2 称为二类偏离．一类偏离 D_1 与二类偏离 D_2 通称为信息密度，它们在研究遗传

（下转第 128 页）

（上接第 85 页）

语言结构的抗干扰能力方面有重要意义[6].

与我们讨论的中心问题有关的是．人们利用上述计算方法，求出了六十多种生物中 DNA 的 D_2 值．计算结果的对比表明，脊椎动物中的 DNA 的信息指标 D_2 值，一般地高于低等有机体的 D_2 值，这是与生物学的进化观点一致的．即对于较高级的动物，它们的遗传语言具有较严格的“语言法则”，并具有能足够表达较复杂指令系统的丰富词汇[6]．因此，D_2 是生命系统遗传机能与生命过程程序控制指令系统进化程度的良好信息指标，人们称之为进化指标．它明确地以统计物理学中熵的概念，表明了宇宙中存在着一些事物集合，它们的熵值随着历史的进程而趋向于逐渐减少．

利用统计物理学中熵的概念研究生物分子水平结构这一初步成果，不仅给生物医学指出了新的有效研究方法和途径，开创了新的领域和局面，而更重要的是，这一研究，进一步揭去了生命系统一层神秘的罩纱，缩小了所谓生命系统具有“特殊性”这一界限不定的“禁地”，给物理学工作者对生命的发生、发展、成长与持续过程的研究，展现出一幅生动的物理图象，提供了一个可以有所作为的用武之地．未来将会表明，在认识生命系统的过程中，物理数学将不断发展它们自身，显示出它们在生命系统的科学研究中，存在着难以局限的无尽潜力．

参 考 文 献

[1] P. Glansdorff, I. Prigogine, Thermodynamic Theory of Structure, Stability and Fluctuations, John Wiley and Sons Ltd., London, (1971), 287.

[2] Joachim E. Lay, Thermodynamics, Charles Merrill Books, Inc., Columbus, Ohio, (1963), 304.

[3] C. E. Shannon, *Bell Syst. Techn. Journal,* **27** (1948), 379—423; 623—656.

[4] Abramson Norman, Information Theory and Coding, McGraw-Hill, New York, (1963), 33.

[5] Granero-Porati et al., *J. Theor. Biol.,* **86** (1980), 401.

[6] L. L. Gatlin, Information Theory and Living Systems, Columbia University Press, New York, (1972), 35.

[7] Fazlloluh M. Reza, An Introduction to Information Theory, McGraw-Hill, New York, (1961), 77.