最邻近序列分析法的贡献

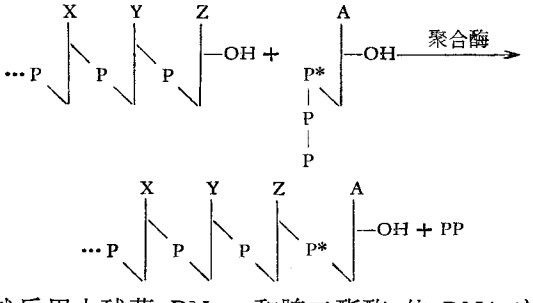
陈 建 华

（上海复旦大学生物系）

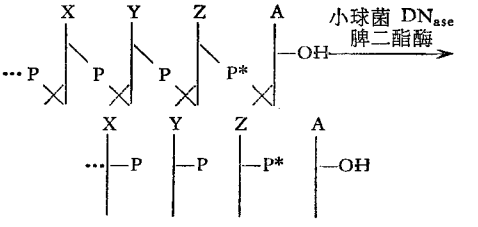
1961 年 J. Josse 等发表了最邻近序列分析法，第一次证明 DNA 的两条链极性相反，在核苷酸顺序分析上迈出了可喜的一步。现在核苷酸序列分析的方法已趋向“自动化”，许多核苷酸顺序已被确定，但这绝不能否定最邻近序列分析法的作用。另一方面，较复杂生物的 DNA 的核苷酸顺序还不容易搞清；即使能够搞清，如果面对那复杂的数据我们找不出它的最本质的特征，那么我们对生命信息的认识就没有多大进步。而从信息论的角度，却可从最邻近序列分析法提供的数据中获得更多的知识。

一、最邻近序列分析法

DNA 复制是以四种 5′三磷酸脱氧核苷（ATP，GTP，CTP，TTP）为原料，以母体 DNA 为模板，在 DNA 多聚酶作用下进行的。如果将其中一种 5′三磷酸脱氧核苷用 P^{32} 标记（如 ATP:P-P-P*  ），则新合成的链中，所有与 A（A. T. G. C 有时表示碱基，有时表示核苷酸，此处 A 指核苷酸）的 5′端连接的磷酸都被 P^{32} 标记了，即



然后用小球菌 DNase 和脾二酯酶，使 DNA 完全水解。由于这两种酶都是在 5′位置上切断磷酸酯键，所以水解产物全为 3′-磷酸脱氧核苷（3′核苷酸），即



于是标记 A 的 P^{32} 就转移到它的“近邻”Z（可以是 A. T. G. C 中的任一种）上。将 3′核苷酸按 A. T. G. C 分离后，分别测量它们的放射活性，从而可以推算各种“近邻”的量。Josse 等对 *M. phlei* 的 DNA 进行最邻近序列分析，关于 A 的数据如表 1。

显然，放射活性的数据与实验条件、操作过程有很大关系。因为 DNA 复制的过程越长，新合成的链越多，掺入 DNA 的 P^{32} 也越多，即放射活性数据的值越大。但不管复制过程长短，DNA 的核苷酸顺序不会改变，所以与 A 邻近的核苷酸 A. T. G. C 的相对值就不会改变。因此，比较有意义的是 T_pA，A_pA，C_pA，G_pA 的数据与总数据（$T_pA + A_pA + C_pA + G_pA$）的比值，它们分别为 $\frac{873}{11703} = 0.075$，$\frac{1710}{11703} + 0.146$，$\frac{4430}{11703} = 0.378$，$\frac{4690}{11703} = 0.401$。这些数据表明：A 的 5′方向近邻 T 占 0.075，A 占 0.146，C 占 0.378，G 占 0.401。

表 1 放射剂量数据

3′核苷酸	序 列	放射活性 c. p. m
T_p	T_pA	873
A_p	A_pA	1710
C_p	C_pA	4430
G_p	G_pA	4690
总计		11703

同样，可用 P^{32} 标记的 TTP，CTP，GTP 做类似实验，得到相应的比值，其结果如表 2。

表 2　各实验中最邻近核苷酸的相对值

	I		II		III		IV
T_pA	0.075	T_pT	0.157	T_pG	0.187	T_pC	0.182
A_pA	0.146	A_pT	0.194	A_pG	0.134	A_pC	0.189
C_pA	0.378	C_pT	0.279	C_pG	0.414	C_pC	0.268
G_pA	0.401	G_cT	0.370	G_pG	0.265	G_pC	0.361
	1.000		1.000		1.000		1.000

表 2 中 16 种顺序之间的相互关系的确定有两种方法：（1）测定 *M. phlei* 的 DNA 的 A. T. G C 组成比(即百分含量)，Josse 测得它们为 0.162，0.165，0.338，0.335。然后将 0.162 乘以表 2 的第 I 列数据，0.165 乘以第 II 列数据……。即得到这 16 种顺序的频度，即最邻近序列频度。（2）设它们的百分含量分别为 a、t、g、c，根据已有数据算出它们的值。对于 T_pA，A 的 5′方向近邻是 T，同样也可认为 T 的 3′方向近邻是 A。因为 A 的百分含量为 a，所以 T_pA 含量为 $0.075a$，同理 T_pT 含量为 $0.157t$，T_pG 含量为 $0.187g$，T_pC 含量为 $0.182c$，所有这些值相加即为 T 的总含量 t。即：

$$0.075a + 0.157t + 0.187g + 0.182c = t$$

同理可得其它三个等式：

$$0.146a + 0.194t + 0.134g + 0.189c = a$$

$$0.378a + 0.279t + 0.414g + 0.268c = c$$

$$0.401a + 0.370t + 0.265g + 0.361c = g$$

解上四元一次方程组得：

$a = 0.164$，$t = 0.162$，$g = 0.337$，$c = 0.337$

由以上数据求频度，得到结果如表 3。

表 3　*M. phlei* DNA 的最邻近序列频度

	T_p	A_p	C_p	G_p	
$dATP^{32}$	T_pA 0.012	A_pA 0.024	C_pA 0.063	G_pA 0.065	0.164
$dTTP^{32}$	T_pT 0.026	A_pT 0.031	C_pT 0.045	G_pT 0.060	0.162
$dGTP^{32}$	T_pG 0.063	A_pG 0.045	C_pG 0.139	G_pG 0.090	0.337
$dCTP^{32}$	T_pC 0.061	A_pC 0.064	C_pC 0.090	G_pC 0.122	0.337
	0.162	0.164	0.337	0.337	1.000

二、DNA 信息的冗余结构

各种生物的 DNA 有其独特的碱基组成，A. T. G. C 的含量并不是均等的。各种 DNA 的 16 种最邻近序列频度也有其独特而非随机的款式。在没有更详细的生物 DNA“指纹”的情况下，碱基组成和最邻近序列款式能粗略地刻划生命信息的特征。但从信息论的角度来看，这些数据还有其更深刻的内容——反映了遗传信息的冗余度结构。

冗余度是信息论中一个很重要的概念。其定义是：

$$R = 1 - \frac{H}{H_{\max}} \tag{1}$$

其中 H 代表某一信源的信息量，$H_{\max}$ 为该信源可能达到的最大信息量。冗余即多余。冗余度则表示该信源中无用消息所占的比例。如由 A. T. G. C 四个符号构成的信源其 $H_{\max} = 2$，而在 *M. Phlei* 中 $H = 1.911$，$R = 0.045$。即在此信源中有 4.5% 的消息是多余的。信息论中，冗余并不意味着完全无用。信源具有冗余度能保证其信息有相应的可靠性，在传输过程中具一定的保真度。如天气预报常常要播三遍，其中二遍毫无新消息，似乎完全是多余的，但它能使收报者准确无误地收到消息。

生命的遗传信息不是核苷酸，也不是能翻译成氨基酸的三联体，而是由数目庞大的核苷酸构成的有序元素组— DNA（或 RNA）。以 n 个有序元素组为元构成的空间称 n 拍（n-tuple）空间：

$$S_n = \{x_{i_\sigma}, \sigma = 1, 2\cdots n, i_\sigma = 1, 2, 3, 4\} \tag{2}$$

构成 DNA 的核苷酸只有四种，所以 i_σ 只有四种取值。如果每一核苷酸出现的概率与前面出现哪种核苷酸无关，则 n 拍元 $(x_{i_1}x_{i_2}\cdots x_{i_n})$ 出现的概率 P_n 为：

$$P_n = P_{i_1} \cdot P_{i_2} \cdots \cdot P_{i_n} \tag{3}$$

但若 $(x_{i_1}x_{i_2}\cdots x_{i_n})$ 中第 σ 个核苷酸出现的概率与前面 m 个核苷酸如何出现有关，那么我们

把 n 拍元看成是一个具 m 重记忆的马尔柯夫信源的输出。其出现的概率 $P_n^{(m)}$ 为:

$$P_n^{(m)} = P_{i_1} \cdot P_{i_1 i_2} \cdots P_{i1\cdots i(m+1)} \cdots P_{i(n-m)\cdots i(n-1)in} \tag{4}$$

这通常是 n 个概率的乘积，其中 $P_{i_1\cdots iK}$ 是当 $x_{i_1}\cdots x_{i(K-1)}$ 出现时 x_{iK} 出现的概率，其下标最多可有 $m+1$ 个。而 S_n 的熵为:

$$H_n^{(m)} = -\sum_{i_\sigma=1}^{4} P_n^{(m)} \log_2 P_n^{(m)} \tag{5}$$

将(4)代入(5)式得:

$$H_n^{(m)} = H_M^{(0)} + H_M^{(1)} + H_M^{(2)} + \cdots + H_M^{(m-1)} + (n-m)H_M^{(m)} \tag{6}$$

其中:

$$H_M^{(K)} = -\sum_{i_\sigma=1}^{4} P_{i_1} \cdot P_{i_1 i_2} \cdots P_{i1\cdots i(m+1)} \log_2 P_{i1\cdots iK+1}, \quad (K = 0, 1\cdots m) \tag{7}$$

由(6)式可知 $\lim\limits_{n\to\infty} \dfrac{H_n^{(m)}}{n} = H_M^{(m)}$，所以，若 $m \ll n$ 则有:

$$\frac{H_n^{(m)}}{n} = H_M^{(m)} \tag{8}$$

定义:

$$D_{K+1} = H_M^{(K-1)} - H_M^{(K)}, \quad (K = 1, 2\cdots m) \tag{9}$$

且:

$$D_1 = \log_2 4 - H_M^{(0)} \tag{10}$$

将(9)，(10)相加得:

$$\begin{aligned} D_1 + D_2 + \cdots + D_{m+1} &= \log_2 4 - H_M^{(m)} \\ &= \log_2 4 - \frac{H_n^{(m)}}{n} \\ &= \log_2 4 \left[1 - \frac{H_n^{(m)}}{n \log_2 4}\right] \end{aligned} \tag{11}$$

(11)式方括号中的量就是 Shannon 定义的冗余度。于是 Gatlin 提出：冗余度不是一个标量，而是具有一定结构的 $m+1$ 维矢量。即:

$$R = \frac{1}{\log_2 4}(D_1 + D_2 + \cdots + D_{m+1}) \tag{12}$$

为计算 R，可先由(7)式算出 $H_M^{(K)}$。有:

$$H_M^{(0)} = -\sum_{i_\sigma=1}^{4} P_{i_1} P_{i_1 i_2} \cdots P_{i1\cdots i(m+1)} \log_2 P_{i_1}$$

$$= -\sum_{i_1=1}^{4} P(x_{i_1}) \log_2 P(x_{i_1})$$

$$= H(x_{i_1}) \tag{13}$$

$$H_M^{(1)} = -\sum_{i_\sigma=1}^{4} P_{i_1} P_{i_1 i_2} \cdots P_{i1\cdots i(m+1)} \log_2 P_{i_1 i_2}$$

$$= -\sum_{i_\sigma=1}^{4} P_{i_1} P_{i_1 i_2} \cdots P_{i1\cdots i(m+1)} \log_2 \frac{P(x_{i_1} x_{i_2})}{P(x_{i_1})}$$

$$= H(x_{i_1} x_{i_2}) - H(x_{i_1}) \tag{14}$$

$$\therefore \quad D_1 = \log_2 4 - H(x_{i_1}) \tag{15}$$

$$D_2 = H_M^{(0)} - H_M^{(1)} = 2H(x_{i_1}) - H(x_{i_1} x_{i_2}) \tag{16}$$

由 DNA 碱基组成数据 $P(x_{i_1})$ 能得到 $H(x_{i_1})$，由最邻近序列分析数据可得到 $H(x_{i_1} x_{i_2})$，从而可算出冗余结构中的 D_1 和 D_2。

此外，Gatlin 还定义了一个量 $RD2$

$$RD2 = \frac{D_2}{D_1 + D_2} \tag{17}$$

它表示在冗余结构中 D_2 所占的比例。

根据 J. Josse 等和 M. N. Swartz 等的实验数据，笔者按上法计算得到如表 4 那样的结果。在 R-RD2 图上标出相应点，即得到图 1。

在冗余度 R 中，D_1 表示由于核苷酸出现几率不均等而产生的与最随机状态的偏离。所以，D_1 增加表明某些核苷酸出现的机会大于另一些核苷酸。这将使得消息多样性潜力迅速下降，其极端情况是某一核苷酸出现的几率为 1，

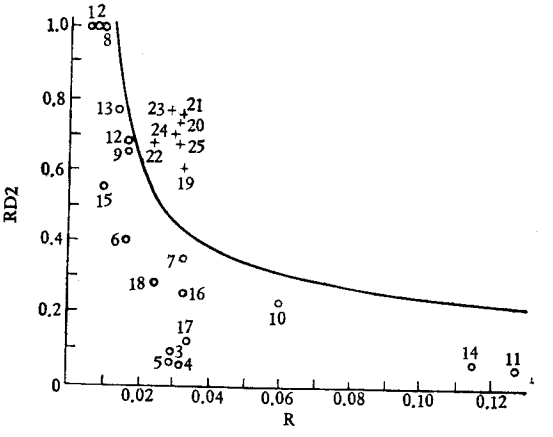


图 1　从最邻近序列分析数据得出的 25 种生物的 DNA 冗余结构矢量在 R-RD2 中的分布

表 4 各种生物 DNA 的信息冗余结构

序号	DNA来源	D_1	D_2	R	$RD2$
1	λ+	0	0.011	0.006	1
2	λdg	0	0.014	0.007	1
3	T_2	0.053	0.006	0.029	0.096
4	T_4	0.059	0.005	0.032	0.071
5	T_6	0.054	0.005	0.029	0.084
6	ϕ×174	0.018	0.012	0.015	0.407
7	*H. influenzae*	0.040	0.023	0.032	0.366
8	*E. coli*	0	0.016	0.008	1
9	*B. subtilis*	0.010	0.019	0.015	0.656
10	*M. phlei*	0.089	0.027	0.058	0.235
11	*M. lysodeikticus*	0.129	0.012	0.126	0.049
12	*A. aerogenes*	0.009	0.020	0.015	0.690
13	*Chlamydomonas*	0.006	0.010	0.012	0.769
14	*Tetrahymena pyriformis*	0.211	0.015	0.113	0.065
15	Whear germ	0.007	0.010	0.009	0.564
16	*Echinus esculenta* (sea urchin)	0.048	0.016	0.032	0.244
17	*Paracentrotus lividus* (sea urchin)	0.058	0.008	0.033	0.115
18	海星睾丸	0.031	0.013	0.022	0.288
19	人脾	0.025	0.038	0.032	0.599
20	兔肝	0.017	0.044	0.031	0.723
21	鸡红细胞	0.017	0.047	0.032	0.731
22	鲑肝	0.014	0.029	0.022	0.675
23	小牛胸腺	0.014	0.039	0.027	0.734
24	牛精子	0.017	0.038	0.028	0.698
25	鼠肝	0.021	0.041	0.031	0.663

其他核苷酸为零，没有消息变化，达到信息论中的绝对零值。

D_2 是 $(x_{i_1}x_{i_2})$ 矢量中两核苷酸相互不独立与相互独立状态的偏离。D_2 的增加也会降低信源的消息多样性潜力，但它的效果比 D_1 缓慢。然而，D_1 和 D_2 的增加都对由总 R 值来量度的消息的可靠性和保真度有贡献。

生物总是从简单到复杂，从低级向高级进化的。DNA 含量的逐渐增加，是生物逐渐复杂化的一个标志。但，事实上许多两栖类的 DNA 含量比哺乳动物的还要高，可见生物之间的复杂性差异不仅仅表现在信息量方面，更重要的可能还表现在信息结构方面。复杂生物必须能稳定地保存自己的种族，同时又能适应千变万化的自然环境。而生物若要消息既具有高度可靠性和保真度，又要消息多样性潜力最大，它的遗传信息的结构必须：(1) R 要有适当大的值，(2) R 中的 D_2 的比例应尽可能高一些。因此 R 和 $RD2$ 两种值综合起来，可以作为生物高级性和复杂性的尺度，可以按这两种值将简单生物和复杂生物划分开来。例如，脊椎动物在 R-RD2 图上几乎都处于 $R \cong 0.02 \sim 0.04$ 比特，$RD2 \cong 0.6 \sim 0.8$ 比特的区域内。非脊椎动物则不在此区域，它们或者 R 值较小，达不到一定的冗余度；或者 $RD2$ 小，表现出结构简单。脊椎动物和非脊椎动物之间有一条较明确的界线，这条界线也就是 Von. Neumann 所谓的“复杂性位垒”。

参 考 文 献

[1] Guschlbauer, W.: *Nucleic Acid Structure*, Springer-Verlag, New York Inc, 1976.

[2] Weil, J. H.: *Biochimie Générale* (3e éd.), Masson Paris, 1979.

[3] Lehninger, A. L.: *Biochemistry* (Sec. ed.), Worth Publishers, Inc, 1977.

[4] 喜安善市、室贺三郎著(李文清译)：信息论，上海科技出版社，1962 年。

[5] Feinstein, A.: *Foundation of Information Theory*, McGraw-Hill Co., New York, 1958.

[6] Gatlin, L. L.: *Molecular Anthropology*, Plenum Press, New York, 81—87, 1976.

[7] Josse, J., et al.: *J. Biol. Chem.*, **236**, 864, 1961.

[8] Swartz, M. N., et al.: *J. Biol. Chem.*, **237**, 1961, 1962.

[9] Shannon, C. E.: *The Mathematical Theory of Communication*, University of Illinois Press, Urbana, Ill., 1949.

[10] Von Neumann, J.: *Theory of Self-reproducing Automata* (A. W. Burkes ed.), University of Illinois Press, Urbana Ill, 1966.

[本文于1981年10月30日收到]