第2卷 第2期 林业科学研究 Vol. 2, No. 2
1989年4月 FOREST RESEARCH Apr., 1989

基因分化值和基因调控信息量研究

陶 粮 庞广昌

（中国林业科学研究院林业研究所）

摘要 本文通过对基因调控、表达的分析，探讨了个体分化定量问题，提出用基因分化值 D_g 指标，度量分化细胞之间的内部差异；对调控、表达问题，用信息论的方法加以处理，得到了基因调控熵 H_r，再从 H_r 指标估算出调控信息量。D_g 具有强度性质；H_r 具有容量性质。试验选用同功酶测定这两项分化指标。再用这两项指标进一步讨论分化、进化、生物信息、群体复杂性等问题。

关键词 基因调控；基因表达；基因分化；生物信息；生物进化

分子遗传学是生物科学的一个重要领域。在这个领域里，有关基因分化的研究，已取得了较大的成绩[1]。目前运用基因调控、表达理论虽能对细胞发育过程中的一些分化机制加以说明，但对这一过程中的个体分化状况，仅限于文字描述[2,3]，虽能借助于统计理论对其进行处理，但并未触及分化的本质[4]。在生命过程中，生命信息起着首要的作用。但对生命信息的研究和生物信息量的估算进展缓慢[5-8]。至于通过分析基因调控、表达情况来定量研究分化值及估算基因调控的信息量问题至今未见报道。为此，我们进行了这方面的探讨。

一、基因调控与分化

某物种细胞内一套基因中有 n 个结构基因 g，将这 n 个基因编号，得结构基因位点序列 $g_j(j=1,2\cdots n)$，记为 G：

$$G:\ g_1\ g_2\ g_3\cdots\ g_j\cdots\ g_n$$

这 n 个结构基因受调控后，并不同时全部表达出来，只表达其中一部分，表达序列如下（不表达为空位）：

$$G_1:\ g_1 \quad g_3 \quad g_4\cdots \quad g_j \quad g_{j+2}\cdots \quad g_n$$

$$G_2:\ g_2 \quad g_3 \quad g_4\cdots \quad g_{j+1}\cdots \quad g_{n-1}\ g_n$$

$$G_3:\ g_2 \quad g_4\cdots \quad g_j\cdots \quad g_{n-1}$$

$$\vdots \quad \cdots$$

G_1、G_2、G_3…所属的细胞就成为分化细胞，分化细胞组成了各组织和器官。

本文于1988年5月8日收到。

二、定量化研究分化的方法

基因的各种表达，不仅在数目上有差别，而且在位置上也有差别。为处理问题方便，可以认为在同一组织内细胞相同。如一个个体，有 m 种不同的组织，那么就有 m 种基因位点序列。先将 m 个序列两两进行同位基因比较，将比较值加和，才得出 m 个组织的相互差异。

两组基因序列 $\{G_i\}(i=1,2\cdots m)$、$\{G_i'\}(i=1,2\cdots m')$，如基因数 n、序列数(组织数) m 不同，则 $\{G_i\}$ 中 $G_i(i=1,2\cdots m)$ 之间的差异难以与 $\{G_i'\}$ 中 $G_i'(i=1,2\cdots m')$ 之间差异相比较。基因平均分化值为可比性的差异特征值。再用熵函数描述基因表达。

三、基因的平均分化值

令：$$g_{ij}=\begin{cases}1 & \text{基因表达}\\ 0 & \text{基因不表达}\end{cases}\quad\begin{pmatrix}i=1,2\cdots m\\ j=1,2\cdots n\end{pmatrix}$$

将 G_1、G_2、G_3 … G_m 表示为 n 维空间的向量：

$$G_1=(g_{11}\ g_{12}\cdots\ g_{1j}\cdots\ g_{1n})$$
$$G_2=(g_{21}\ g_{22}\cdots\ g_{2j}\cdots\ g_{2n})$$
$$\cdots\cdots$$
$$G_i=(g_{i1}\ g_{i2}\cdots\ g_{ij}\cdots\ g_{in})$$
$$\cdots\cdots$$
$$G_m=(g_{m1}\ g_{m2}\cdots\ g_{mj}\cdots\ g_{mn})$$

比较 G_1 与 G_2 的差别，其结果可用绝对值距离表示：

$$\delta G_{12}=\sum_{j=1}^{n}|g_{2j}-g_{1j}|$$

表示向量之间的距离有很多种，这里采用汉明距离较为合理。一般地，G_{i_1} 与 G_{i_2}' 之间的汉明距离为：

$$\delta G_{i_1i_2'}=\sum_{j=1}^{n}|g_{i_2'j}-g_{i_1j}| \tag{1}$$

m 个向量中两两间的汉明距离共有 m^2 个(包括自身与自身比)。如表 1 。

表 1　　m个向量间两两比较的汉明距离

G_i \ $\delta G_{i_1i_2}$	G_1	G_2	…	G_m
G_1	δG_{11}	δG_{12}	…	δG_{1m}
G_2	δG_{21}	δG_{22}	…	δG_{2m}
⋮	⋮	⋮	⋮	⋮
G_m	δG_{m1}	δG_{m2}	…	δG_{mm}

将 m^2 个距离值之和记为 δG^*：

$$\delta G^* = \sum_{(i_1,i_2)} \delta_{i_1 i_2} = \sum_{(i_1>i_2)} \delta G_{i_1 i_2} + \sum_{(i_1<i_2)} \delta G_{i_1 i_2} + \sum_{i=1}^{m} \delta G_{ii}$$

因为 $\delta G_{i_1 i_2} = \delta G_{i_2 i_1}$，$\delta G_{ii} = 0$ $(i = 1, 2 \cdots n)$，所以有：

$$\delta G^* = 2 \sum_{(i_1<i_2)} \delta G_{i_1 i_2} = 2\delta G \tag{2}$$

这里 δG 为 $\sum_{(i_1<i_2)} \delta G_{i_1 i_2}$。$\delta G^*$就是$\{G_i\}$中$G_i(i = 1, 2 \cdots m)$之间的差异，即表达基因在数量和位置上的差异。将 δG^* 除以 m^2 得：

$$\frac{\delta G^*}{m^2} = \frac{2\delta G}{m^2} = \frac{2}{m^2} \sum_{(i_1<i_2)} \delta G_{i_1 i_2} \tag{3}$$

式(3)表示了平均每个组织的分化程度。表示 n 个基因的平均分化程度，将 $\delta G^*/m^2$ 除以 n，即得基因的平均分化值 D_g：

$$D_g = \frac{\delta G^*}{nm^2} = \frac{2}{nm^2} \sum_{(i_1<i_2)} \delta G_{i_1 i_2} = \frac{2}{nm^2} \sum_{(i_1<i_2)} \sum_{j=1}^{n} |g_{i_2} - g_{i_1}| \tag{4}$$

四、分化的基因调控熵和基因调控信息量

在$\{G_i\}(i = 1, 2 \cdots m)$中，第 j 位上基因 g_j 共表达出 $\sum_{i=1}^{m} g_{ij} = A_j$ 个基因，这种表达就可能有 $C_m^{A_j}$ 种方式。当 A_j 接近 0，或接近 m 时，则 $C_m^{A_j}$ 较小，说明了 g_j 在 $\{G_i\}$ 中各个组织 $G_i(i = 1, 2 \cdots m)$ 上表达的差异就小。当 A_j 接近 $m/2$ 时，则 $C_m^{A_j}$ 就大，此时 g_j 在$\{G_i\}$中各个 $G_i(i = 1, 2 \cdots m)$ 上表达的差异就大。所以 $C_m^{A_j}$ 的大小与基因g_j在各种组织中表达的差异有关。

取 $C_m^{A_j}$ 对数，(对数的底可任意，一般取 2)记为 S_j：

$$S_j = \log_2 C_m^{A_j} \tag{5}$$

假定 $C_m^{A_j}$ 种表达中，各种表达都是等概的，则每种表达出现的概率为 $1/C_m^{A_j}$，将式(5)变换得：

$$S_j = \log_2 C_m^{A_j} = -\log_2 \frac{1}{C_m^{A_j}} = -\sum_{k=1}^{A_j} \frac{1}{A_j} \log_2 \frac{1}{C_m^{A_j}} \tag{6}$$

这样，就得到了基因 g_j 的分化基因调控熵(以下简称基因调控熵或调控熵)，S_j 就是调控基因在不同组织中所表现出来的熵值。将 n 个基因的调控熵加和：

$$S_r = \sum_{j=1}^{n} S_j = \sum_{j=1}^{n} \log_2 C_m^{A_j} \tag{7}$$

S_r 即为基因 $g_j(j = 1, 2 \cdots n)$ 在$\{G_i\}(i = 1, 2 \cdots m)$中的总调控熵，它标志一个个体$\{G_i\}(i = 1, 2 \cdots m)$中基因表达状况。

某个个体有基因调控熵S_r，如果做基因表达试验，例如同功酶试验，可以从试验中完全了解到基因的表达，也就完全知道了调控情况。从信息理论可知，在这试验中所获得的该个

体基因调控的信息量 I_r 为：

$$I_r = S_r \tag{8}$$

所以基因调控熵也称为基因调控信息熵。显然为估算生物的信息量提供了一条途径。

五、基因平均分化值与基因调控熵的关系

从式(7)中可以看出，S_r 值随 m、n 的增大而增大。只有当 m 和 n 是个体的真实数时，S_r 值才是这个个体基因的真实调控值，所以 S_r 值不易求得。但对两个个体取较多且相同的 m 和 n 数时，所得的两个 S_r 值也可进行比较。D_g 值具有强度性质，取足够大的 m 和 n 可测出，即能标志基因的调控状况。D_g 和 S_r 都用来描述基因的调控情况，所以它们之间必有内在的相互联系。D_g 与 S_r 之间的关系可用方程组表示：

$$\begin{cases} n\dfrac{m^2}{2}D_g = \sum\limits_{j=1}^{n}\delta G_j \\ S_r = \sum\limits_{j=1}^{n}\log_2\dfrac{m!}{A_j(m-A_j)!} \end{cases} \tag{9}$$

其中 $$\delta G_j = \sum_{(i_1<i_2)} |g_{i_2 j} - g_{i_1 j}|$$

六、分化值和调控熵值讨论

1．如基因全不表达，$g_{ij}=0$，则 $\delta G=0$，所以 $D_g=0$。又因 $\sum\limits_{i=1}^{m} g_{ij}=0$，所以

$$S_r = \sum_{j=1}^{n} S_j = \sum_{j=1}^{n} \log_2 C_m^0 = \sum_{j=1}^{n} \log_2 1 = 0$$

这是一种极端情况，表明该个体已停止生命活动。

2．如基因全表达，$g_{ij}=1(i=1,2\cdots m, j=1,2\cdots n)$，则 $\delta G=0$，所以 $D_g=0$。又因 $\sum\limits_{i=1}^{m} g_{ij} = m(j=1,2\cdots n)$，

所以 $$S_r = \sum_{j=1}^{n} S_j = \sum_{j=1}^{n} \log_2 C_m^m = \sum_{j=1}^{n} \log_2 1 = 0$$

这是一种所有结构基因全表达，不受任何控制的情况，此种现象的出现是不可能的。

3．如基因表达相同，$G_1=G_2=\cdots G_i=\cdots G_m\neq 0$ 则 $\delta G=0$，所以 $D_g=0$，又因 $\sum\limits_{i=1}^{m} g_{ij}=0$ 或 $m(j=1,2\cdots n)$，所以

$$S_r = \sum_{j=1}^{n} S_j = \sum_{j=1}^{n} \log_2 C_m^{0\text{或}m} = 0$$

此时表示分裂后的细胞与原细胞相同，没有分化出不同的组织。

4．只要有两个以上序列的基因表达不同，$G_{i_1}\neq G_{i_2}(i_1\neq i_2)$，则 $\delta G>0$，所以 $D_g>0$。又因 $\sum_{i=1}^{m} g_{i,j}=A_j(j=1,2\cdots n)$，所以　$S_r=\sum_{j=1}^{n} S_j=\sum_{j=1}^{n}\log C_m^{A_j}>0$，表示该生命体已分化。$G_g$、$S_r$ 值愈大，分化程度愈大。

5．所有组织的同位基因有一半表达，$\sum_{i=1}^{m} g_{i,j}=m/2(j=1,2\cdots n)$，可证得：

$$\delta G=\left(\sum_{i=1}^{m} g_{i,j}\right)\times\left(m-\sum_{i=1}^{m} g_{i,j}\right)=\frac{m}{2}\left(m-\frac{m}{2}\right)=\left(\frac{m}{2}\right)^2=\delta G_{j,\max}$$

所以

$$\delta G_{\max}=\sum_{j=1}^{n}\delta G_{j,\max}=n\left(\frac{m}{2}\right)^2$$

$$D_{g,\max}=\frac{2\delta G_{\max}}{nm^2}=\frac{2n}{nm^2}\left(\frac{m}{2}\right)^2=\frac{1}{2}$$

上式表明个体的最大分化值为 1/2，与基因数 n、组织数 m 无关。

当 $A_j=\sum_{i=1}^{m} g_{i,j}=m/2$ 时，$C_m^{A_j}$ 取得值最大(为了讨论问题方便，假定 m 为偶数)，则一个基因数为 n、组织数为 m 的个体得到最大调控熵：

$$S_{r,\max}=\sum_{j=1}^{n} S_{j,\max}=\sum_{j=1}^{n}\log C_m^{m/2}=n\log_2 C_m^{m/2}$$

上式说明该个体基因调控熵达到了极限，如要继续增加调控信息量，必须增加组织数或基因数。

以植物为例，大约 $n=3\times10^4$，如 $m=40$，那么它的最大调控熵为：

$$S_{r,\max}=3\times10^4\log_2 C_{40}^{20}=1.110\,1\times10^6\ (\text{bit})$$

七、分化值 D_g 的测定及举例

测定分化值 D_g，最好是基因数和组织数 m(各种组织有特定基因表达序列)为真实值，然而真核细胞生命体的基因数约为 3 万个[9]；由于各种组织众多，要把这么大的 n 和 m 全部测出来是不可能的，也是不必要的。只能随机地抽取一部分组织和基因作试样。抽样要尽可能有代表性，最好多抽一些。

要测定基因表达状况，应测定基因表达的产物，而不是它的本身。基因指导蛋白质的合成，基因的表达能在蛋白质上表现出来。试验选用同功酶为试材是合适的。其优点如下：①易测定；②操作方便；③用不同的底物染色可测到不同系列的基因；④灵敏度高[10]；⑤种

类多；⑥代表性强。

严格来说，应在各种组织上取样，如在器官上取样，测出来的基因表达状况，是数种不同组织的综合表现，所测得的D_g值是一个比真实值要小的近似值。当然这种近似仍可以用于比较，所以取样时，应在各个个体相同的器官上取样，才能得到相对可比较的D_g值。

对一植株(采自河南南沼县的油松)，在幼叶G_1、老叶G_2、根G_3、嫩茎G_4上取样。可得四个酶谱：G_1、G_2、G_3、G_4。从这四个酶谱上可得出一个基本酶谱G_0，用$G_i(i=1,2,3,4)$与G_0相比较，当G_i与G_0上相应的酶带出现时，记为1；如不出现，则记为0，如表2。

表2 器官酶谱(染色底物：H_2O_2)

器官 \ 酶带	基本酶谱				
	g_1	g_2	g_3	g_4	g_5
G_1(幼叶)	0	1	0	0	0
G_2(老叶)	0	1	1	0	1
G_3(根)	1	1	1	1	1
G_4(嫩茎)	1	1	1	0	0

或用矩阵表示：

$$\{g_{ij}\}=\begin{Bmatrix}0&1&0&0&0\\0&1&1&0&1\\1&1&1&1&1\\1&1&1&0&0\end{Bmatrix}$$

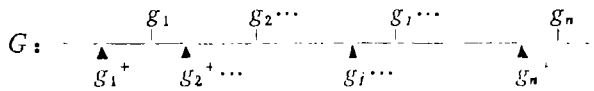
由式(4)得：$D_g=\frac{2\delta G}{nm^2}=\frac{2\times14}{5\times4^2}=0.3500$

为获得较准确的D_g值，可适当多取一些器官和多种染色底物。

八、讨 论

1．分化值和调控熵值的本质 分化指数是通过基因产物——蛋白质的出现状况测出来的，它不涉及蛋白质本身的结构和功能，只与结构基因受控状况有关。

设g_j^+为结构基因g_j的调节基因，则基因表达序列为(示意图)：



从式(4，7)可知，如$g_j(j=1,2\cdots n)$在$\{G_i\}(i=1,2\cdots m)$中都表达或都不表达，则$D_g=0$和$S_r=0$，也就是说调节基因不起作用(这里不包括量上的调节)。D_g、S_r值愈大，说明基因受控愈复杂。可见D_g、S_r值是调节作用的变量与结构基因无关，只是借用结构基因表现出来。

2．分化与生长发育 细胞的分生、分化，导致了生物体的生长和发育。从单个细胞或组织生长发育至器官健全的生物体，分化程度越来越大，D_g、S_r值亦随之变化。所以，分化值D_g、调控熵S_r表明了生物体的生长发育状况。但其中D_g、S_r值随生物体生长发育遵循的变化规律，还待今后进一步的探索，有待进行大量的试验。

3．分化与进化 千差万别的生物体能够生存，是因为它们都适应周围的环境。如果对它们进行分类的话，则可以形成从简单的适应至复杂的适应一个系列。适应环境的复杂生物，显然是从简单的发展而来，进化就是这种从简单到复杂的发展过程。

人比大肠杆菌进化，决不是说人这个物种比大肠杆菌更适应，而是人具有比大肠杆菌更复杂的系统来适应环境。

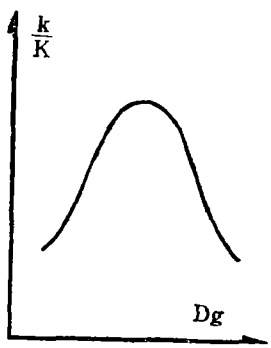
关于生命的复杂程度，尚未见到一个令人满意的定量化描述形成[11]。也就是说无法回答

"人比狗进化多少"这样的问题。通过以上的分析和例子，我们认为 D_g 值对这个问题是一个较为合理的指标。对不同生物的最大(或同一个发育阶段)的分化值进行比较，当然可以衡量它们的进化程度。

另外，根据胚胎学证据[13]，生物的发育过程重演了生物的进化历程，所以跟踪分化值在不同发育阶段的变化情况，也应能导出生物进化的历程。

一般来说，同一物种的各个个体的 n、m 相同，所以不同的个体可用 D_g 值表示其复杂程度。利用种群群体平均分化值：

$$\overline{D}_g=\frac{1}{K}\sum_{k=1}^{K}D_{g,k}\quad (K\text{为种群个体抽样个数}),$$



分化值分布示意图

可导出选择压力对进化所起的作用。设有一个生物种群，它们的分化值呈常态分布(见图)。对它们施加不同的选择压会产生什么结果呢？淘汰掉分化值大的还是小的个体呢？这与分化值大小无关。可以通过对处于不同选择压下的相同种群(如最适地方和不适合地方)的种群平均分化值 $\overline{D}_g$ 进行比较，就可得出上述结论。例如，山西关帝山地区为油松的中心产区[14]；河南南沼县的油松长势也不错，而辽宁建平县和青海互助县的油松处于较劣的环境，其种源的选择压力大于上述两个地区。这四个地区油松种群平均分化值 $\overline{D}_g$ 表明选择压力可能使 $\overline{D}_g$ 变小(表 3)。

表 3　油松种群 $\overline{D}_g$ 值测定

地　　区	$\overline{D}_g$
辽宁建平县	0.255 2
青海互助县	0.244 0
山西关帝山	0.326 4
河南南沼县	0.286 4

4. 分化值概念的扩展——群体复杂度　分化值 $\overline{D}_g$ 描述了同一生物体不同组织之间的差异情况，如果把同一生物体上的每一个细胞都作为一个个体来看待的话，那么这个生物体就相当于一个由无数单细胞组成的群体。因此，基因表达上不同的细胞也就相当于遗传上不同的个体。由此可见分化值的计算方法也适合于对群体复杂性的数量化描述。对群体中不同个体不同组织的等位基因两两进行比较，取其差异值之和，可导出群体复杂度的表达式，由此证明它与群体的进化历史和选择压力有关。

参 考 文 献

[1] 李振刚，1985，分子遗传学，安徽科技出版社。

[2] 李士鹏，1987，LDH同功酶的发生遗传学探讨，生物科学动态，2：14。

[3] 李振刚，1985，发育中的基因控制理论，生物科学动态，6：1。

[4] K·马瑟，1977(冯午，1981)，生统遗传学导论，农业出版社。

[5] C. I. J. M. Stuart, 1985, Bio-informational Equirvalence, J. Theor. Biol., 113: 611—636.

[6] 权文富等，1985，DNA的信息研究与设想，生物科学参考资料，第十九集，科学出版社。

[7] Andrzej Konopka, 1984, Is the Information Content of DNA Evolutionarily Significant?, J. Theor. Biol., 107: 697—704.

[8] Volkenstein, M. V.,1982, Physics and Biology, Academic Press, New York.
[9] Ⅱ·史密斯, 1977, (李镛泾等, 1986), 植物细胞分子生物学, 科学出版社。
[10] Steven D. Tanksley, 1983. Isozymes in Plant Genetics and Breeding, Part A, Elsevier.
[11] Bendall, D. S.,1983, Evolution from Molecules to Men, Cambridge University Press.
[12] 李难, 1982, 生物进化论, 人民教育出版社。
[13] 徐化成等, 1981, 油松天然林的地理分布和种源区的划分, 林业科学, 3:258。

STUDY ON THE GENE DIFFERENTIATION VALUE AND GENE REGULATORY INFORMATION

Tao Liang Pang Guangchang

(The Research Institute of Forestry CAF)

Abstract This paper provides a method of measuring genetic differentiation by analyzing gene regulation and expression. A Genetic Differentiation Value (D_g) was found, which could be used to determine the difference among different tissues. Using the method of the information theory, the authors found a function, that is the Gene Regulation Entropy (H_r), which could describe the extent of differentiation. The quantity of regulation information could be estimated from H_r. D_g has the property of intensity and H_r has the property of content. The value of D_g and H_r could be measured by isozymes or other proteins. In the end of the paper, how to use D_g and H_r in differentiation, biological evolution, biological information, the complexity of population and practical application in forest population genetics have been discussed.

Key words gene regulation; gene expression; gene differentiation; biological information; biological evolution