生物物理学报 第十八卷 第二期 二〇〇二年六月
ACTA BIOPHYSICA SINICA Vol.18 No.2 Jun.2002

遗传密码格式的组合编码数分析

陈惟昌[1]， 陈志义[2]， 陈志华[3]， 王自强[1]
（1. 中日友好临床医学研究所生物物理研究室，北京 100029；2. 中国科学院自动化研究所国家模式识别实验室，北京 100080；3. 中日友好临床医学研究所生物化学及分子生物学研究室，北京 100029）

摘要：用 N 个密码子对 m 个编码对象进行编码的编码格式是 m 元 N 维空间中的一个顶点。64 个密码子对 20 种氨基酸和终止密码子进行编码格式的组合编码数是一个十分巨大的数字。对多元高维编码空间的拓扑特性进行了分析和研究，并由此推导出 $m-N$ 空间的特性三角的排列方式以及给出特性三角公式的数学证明。指出，目前的遗传密码的编码格式是 21 元 64 维编码空间的一个顶点。应用组合数学分析的方法，计算了遗传密码格式的最大组合编码数 $C_M = 4.19 \times 10^{84}$，基因组遗传密码的组合编码数 $C_G = 1.13 \times 10^{80}$ 以及线粒体遗传密码的组合编码数 $C_T = 1.38 \times 10^{79}$ 等。分析结果表明，遗传密码的指定是一个小概率事件，可能来源于 λ 简并后的偶数三联密码配对的组合编码的对称破缺。

关键词：遗传密码；编码对象（码象）和编码元（码子）；组合编码数；多元高维空间（高维栅格空间）；多项式定理；特性三角（广义贾宪与帕斯卡三角）

中图分类号：Q617 **文献标识码**：A **文章编号**：1000-6737(2002)02-0206-07

1961 年，Nirenberg 和 Matthaei[1] 首次发现氨基酸的遗传密码，为基因信息学奠定基础。目前已经知道，DNA 4 个碱基 C，T，A，G 可组成 64 个三联密码子，分别对 20 种氨基酸和终止密码子进行编码。由于 64 大于 21，因此遗传密码子出现简并现象。据估计[2]，类似于目前遗传密码编码格式的可能组合编码数至少为 $10^{71}-10^{84}$，这是一个巨大的天文数字。遗传密码为什么是现在这样的格式，一直是一个谜。Crick[3] 提出“冻结事故理论”（frozen accident theory），他认为 64 个密码子分配给 20 种氨基酸和终止密码子的编码格式，纯属于一次偶然性的事故。我们分析了遗传密码 6 维编码空间的拓扑特性，发现并确立遗传密码子的“拓扑连通性简并法则”[4]，并找出遗传密码子的简并和氨基酸的分子质量、等电点以及残基的化学键结构之间的联系[5]。本文首先对多元高维编码空间的拓扑特征进行了研究。对遗传密码格式的各种可能的组合编码数进行了计算和分析，并比较了基因组遗传密码格式和线粒体遗传密码格式的组合编码数的异同点。对遗传密码的起源问题进行了分析和讨论。

1 编码过程和解码过程

将被编码的对象（码象）（coded image，codim）按照编码的规则，转换成编码单元（密码子，codon）的过程，称为编码（encoding）过程。反之，根据已知的码元，按照译码规则，反过来确定被编码对象的过程称为解码（decoding）过程。可用以下方式表示：

$$\text{codim} \underset{\text{decoding}}{\overset{\text{encoding}}{\rightleftharpoons}} \text{codon}$$

在码象的集合 $M(x)$ 中，x 可以是一组具体的事物或符号，而在码元的集合 $N(y)$ 中，y 通常是一组符号，字母或数字。$M(x) \to N(y)$ 代表编码过程，而 $N(y) \to M(x)$ 代表解码过程。根据编码规则，通常一个码元只能代表一个特定的码象以避免产生歧义性（equivocality）。但一个码象可以有一个以上的码元。所以一般码元的数目应大于或等于码象的数目以保证每一个码象至少有一个码元与之对应。编码过程可看成是从码象集合 M 到码元集合 N 的映射（mapping）[6]。但由于一个码象可以对应于多个码元，所以编码映射是一种多值映射。由于不同的码象 x_1 和 x_2 对应于不同的码元 y_1 和 y_2，所以编码过程 $M \to N$ 是单射（injection）。但由于码象的数目可以小于码元的数目，所以编码过程不是满射。而在解码过程中，因为码元的数目多于码象的数目，

收稿日期：2001-11-26
基金项目：国家自然科学基金资助项目（60171040）
作者简介：陈惟昌，研究员，电话：(010)64221122-4434，E-mail：chenwch@mail.east.net.cn.

故解码过程是满射（surjection），即 $N \to M$ 的映射充满集合 M。如果码象的数目和码元的数目相同而且呈一一对应关系，则称为一一对应编码（one to one coding），亦称双射（bijection）。例如用 4 个字母 C、T、A、G 对胞嘧啶、胸腺嘧啶、腺嘌呤、鸟嘌呤 4 种核苷酸进行编码即是一一对应编码，称为核苷酸的字符编码。如用 4 个二进制数 00、01、10、11 对 4 种核苷酸编码则称为核苷酸的数字编码[7]。通常用 n 个码元对 n 个码象进行一一对应编码时，其可能的组合格式的编码数为 $n!$，但当码元的数目 n 大于码象的数目 m 时，可以出现多个码元对应于同一个码象的情况，称为简并编码（degenerate coding）。遗传密码即是用 64 个码元（三联密码子）对 21 个码象（20 种氨基酸和终止密码）进行编码的过程，可作为简并编码格式的例子。

2　m 元 N 维空间的拓扑特性

用 N 个码元对 m 个码象进行编码时（$m \leqslant N$），可以考虑把 m 个码象比喻为 m 个不同的容器，在 m 个不同容器中放入 N 个不同粒子时对每个粒子而言，共有 m 个容器可供选择，N 个粒子共有 m^N 种选择，故 N 个不同粒子放入 m 个不同容器的组合数为 m^N，组成一个 m 元 N 维编码空间。m 元 N 维空间的顶点数目为 m^N，每一个顶点代表一种编码格式，可用一个 N 位的 m 进制数表示。在研究遗传密码格式的组合编码数之前，需要对多元高维空间的拓扑特性进行深入分析。

2.1　m 元 N 维空间两顶点之间的特性距离

2.1.1　两个 N 位 m 进制数的按位对称差（bitwise symmetrical difference）的定义

设有 m 元 N 维空间中的两个顶点 P 及 P'，其中　$P = x_N x_{N-1} \cdots x_2 x_1$，

$P' = x_N' x_{N-1}' \cdots x_2' x_1'$

定义 $\Delta x_i = | x_i - x_i' |$ 为顶点 P 与 P' 的第 i 位对称差（差的绝对值），则定义顶点 P 与顶点 P' 的特性距离（characteristic distance）d_c 为：

$$d_c(P, P') = \sum_{i=1}^{N} \Delta x_i \qquad (1)$$

显然，当 $m = 2$ 时，特性距离 d_c 即是汉明距离 d_h，故汉明距离可看成是特性距离在 $m = 2$ 时的特殊情形。

特性距离 d_c 满足以下三个距离定义的标准：

(a)　$d_c(P, P) = 0$　　（自反性）

(b)　$d_c(P_1, P_2) = d_c(P_2, P_1)$　　（对称性）

(c)　$d_c(P_1, P_3) \leqslant d_c(P_1, P_2) + d_c(P_2, P_3)$（三角不等式）

例如，4 元 5 维空间的两个顶点：$P_1 = 30121$，$P_2 = 13210$，则 $d_c(P_1, P_2) = 8$。

2.1.2　m 元 N 维空间（简称 $m - N$ 空间）顶点 V 的特性值

定义 $m - N$ 空间顶点 $P = x_N x_{N-1} \cdots x_2 x_1$ 的特性值为 P 与原点 $(0, 0, \cdots 0)$ 的特性距离，即 $d_c(P) = d_c(0, P)$，显然 $V_c(P) = \sum_{i=1}^{N} x_i$，例如，在 4 元 5 维空间中，$V_c(30121) = 7$。当 $m = 2$ 时，顶点的特性值即顶点的汉明值[6]。

2.2　$m - N$ 空间顶点的对偶运算

定义位值 x 的对偶运算为：$(\sim x) = p - x$，$(p = m - 1)$ 如下：

x：　0，　1，　……　$p-1$，　p

$\sim x$：　p，　$p-1$，　……　1，　0

则顶点 P 的对偶顶点 $(\sim P)$ 为各位 x 的对偶，例如，若 $P = 30121$ 是一个 4 元 5 维空间的一个顶点，则 $(\sim P) = 03212$，其中 $(\sim 0) = 3$，$(\sim 1) = 2$，$(\sim 2) = 1$，$(\sim 3) = 0$。

2.3　$m - N$ 空间的作图法

将原点 $(0, 0, \cdots 0)$ 放在最左侧，极点 $(p, p, \cdots p)$ 对称地放在最右侧，其余各顶点按特性值的大小从左向右排列成特性柱（characteristic column），每一个特性柱的顶点从小到大，由下向上排列。再将相邻的两个顶点（二者的特性距离等于 1）用边联结，即构成一个 $m - N$ 空间（亦称高维栅格空间）。现以 $m = 4$，$p = 3$，$N = 0, 1, 2, 3$ 为例，说明 $m - N$ 空间的作图过程。

2.4　$m - N$ 空间的拓扑特点

(a)　m^N 个顶点中的两两共轭顶点 p 和 $(\sim p)$，以中心点呈对称排列。例如在 4 元 3 维空间中 003 与 330，111 与 222 等均呈中心对称排列（图 1）。

(b)　$m - N$ 空间顶点的阶数和度数。在 $m - N$ 空间的顶点中，0 与 p 位于边的端点，称为外点数值，其余 $1, 2, \cdots p-1$ 在边的内部，称为内点。

(c)　每一个 $m - N$ 空间的顶点共有 N 位数，在 N 位数中如有 t 位为内点数，则称该顶点为 t 阶顶点，该顶点共有 $N + t$ 条邻边，即其顶点是 $N + t$ 度（degree）顶点。0 阶顶点各位的数值为 0 及 p 组成，没有内点数值位。其邻边数为 N，即端顶点之度数为 N。如顶点的 N 位的数值全部为内点的数值，称为 N 阶顶点，其度数为 $2N$，共有 $2N$ 条邻边。例如，

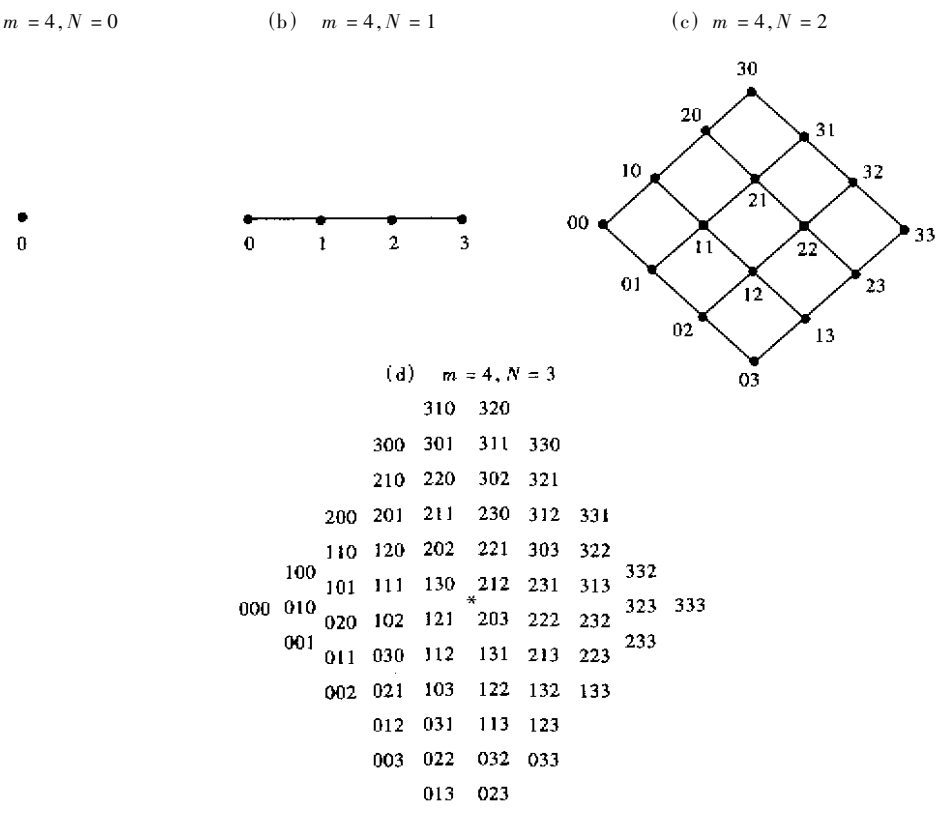

Fig.1 Examples of $m-N$ space

在 4 元 2 维空间中,0 和 3 是外点,1 和 2 是内点,顶点 00、03、30、33 为 0 阶 2 度顶点,各有 2 条邻边,01、02、31、32 等顶点是 1 阶 3 度顶点,各有 3 条邻边,11、12 等顶点是 2 阶 4 度顶点,各有 4 条邻边。

(d) m 元 N 维空间中,t 阶顶点的数目为:$C_N^t \cdot 2^{N-t} \cdot (m-2)^t$,例如 在 4 元 2 维空间中 0 阶的顶点数为 4,1 阶的顶点数为 8,2 阶的顶点数为 4。这是因为,$m-N$ 空间的 2^N 个外点,组成 2 元 N 维空间,共有 $C_N^t \cdot 2^{N-t}$ 个 t 维子空间,则在 t 维空间中的内点总数为 $C_N^t \cdot 2^{N-t} \cdot (m-2)^t$。故得 $m-N$ 空间中 m^N 个顶点按阶数 t 的展开式:

$$m^N = [2+(m-2)]^N = \sum_{t=0}^{N} C_N^t \cdot 2^{N-t} \cdot (m-2)^t。$$

式中 t 为顶点的阶数(相当于 2 元 N 维空间中子空间的维数)。

(e) $m-N$ 空间中全部子空间的总数 $W(m,N)$。

$m-N$ 空间零维子空间顶点的数目为 m^N。

$m-N$ 空间一维子空间的数目为:

$C_N^1 \cdot m^{N-1} \cdot p$,$(p=m-1)$。

$m-N$ 空间 t 维子空间的数目为:

$C_N^t \cdot m^{N-t} \cdot p^t$。

故 $m-N$ 空间子空间的总数为:

$$W(m,N) = \sum_{t=0}^{N} C_N^t m^{N-t} \cdot p^t = (m+p)^N = (2m-1)^N \quad (2)$$

$W(m,N)$ 的数目如表 1:

Table 1 Number of subspace $W(m,N)$ in $m-N$ space

m \ N	1	2	3	4	5	6
1	1	1	1	1	1	1
2	3	9	27	81	243	729
3	5	25	125	625	3125	15625
4	7	49	343	2401	16807	117649
5	9	81	729	6561	59049	531441
6	11	121	1331	14641	161051	1771561

当 $m=2$ 时,$W(2,N)=3^N$,与 2 元 N 维空间的子空间总数一致。$m-N$ 维空间子空间的编码,可用 N 位的 $0,1,\cdots p$ 以及 $\lambda_1,\lambda_2,\cdots\lambda_p$ 进行编码表示。其中 λ_i 是联结 x_{i-1} 到 x_i 的边。例如:4 元 3 维空间中,$\lambda_3\ 2\ \lambda_1$ 代表其中一个 2 维子空间平面,此平面的 4 个顶点是:220,221,320,321。

2.5 m^N 的特性项展开式

m 元 N 维空间的 m^N 个顶点,共有 $N(m-1)+1=Np+1$ 项特性值。具有相同特性值的顶点的

数目，${}_m\mathrm{T}_i^N$，称为特性项（i 为特性值）。各类 $m-N$ 空间的特性项的数值如表 2。

Table 2　The characteristic terms of $m-N$ space

(a) $m=1, N=0\rightarrow 6\cdots$

1
1
1
1
1
1
1

(b) $m=2, N=0\rightarrow 6\cdots$

1
1, 1
1, 2, 1
1, 3, 3, 1
1, 4, 6, 4, 1
1, 5, 10, 10, 5, 1
1, 6, 15, 20, 15, 6, 1
……

(c) $m=3, N=0\rightarrow 6\cdots$

1
1, 1, 1
1, 2, 3, 2, 1
1, 3, 6, 7, 6, 3, 1
1, 4, 10, 16, 19, 16, 10, 4, 1
1, 5, 15, 30, 45, 51, 45, 30, 15, 5, 1
1, 6, 21, 50, 90, 121, 141, 121, 90, 50, 21, 6, 1
……

(d) $m=4, N=0\rightarrow 6\cdots$

1
1, 1, 1, 1
1, 2, 3, 4, 3, 2, 1
1, 3, 6, 10, 12, 12, 10, 6, 3, 1
1, 4, 10, 20, 31, 40, 44, 40, 31, 20, 10, 4, 1
1, 5, 15, 35, 65, 101, 135, 155, 155, 135, 101, 65, 35, 15, 5, 1
1, 6, 21, 56, 120, 216, 336, 456, 546, 580, 546, 456, 336, 216, 120, 56, 21, 6, 1
……

(e) $m=5, N=0\rightarrow 4\cdots$

1
1, 1, 1, 1, 1
1, 2, 3, 4, 5, 4, 3, 2, 1
1, 3, 6, 10, 15, 18, 19, 18, 15, 10, 6, 3, 1
1, 4, 10, 20, 35, 52, 68, 80, 85, 80, 68, 52, 35, 20, 10, 4, 1
……

(f) $m=6, N=0\rightarrow 3\cdots$

1
1, 1, 1, 1, 1, 1
1, 2, 3, 4, 5, 6, 5, 4, 3, 2, 1
1, 3, 6, 10, 15, 21, 25, 27, 27, 25, 21, 15, 10, 6, 3, 1
……

性三角，有以下性质：

(a) $\sum_{i=0}^{k} {}_m\mathrm{T}_i^N = m^N$，$k=N(m-1)=Np$，$p=m-1$，即特性项展开式共有 $k+1$ 项，此 $k+1$ 项之和为 m^N。

(b) ${}_m\mathrm{T}_i^N = {}_m\mathrm{T}_{k-i}^N$，即在特性项展开式中，与两端等距离的两项相等。

(c) 当 k 为偶数时，特性项展开式的中间项（第 $(k/2)+1$ 项）之值最大，当 k 为奇数时，特性项展开式的中间两项，即第 $(k+1)/2$ 和第 $(k+3)/2$ 之值最大。

(d) ${}_m\mathrm{T}_0^N = {}_m\mathrm{T}_k^N = 1$，${}_m\mathrm{T}_1^N = {}_m\mathrm{T}_{k-1}^N = N$

(e) ${}_m\mathrm{T}_i^N = \sum_{j=0}^{m-1} {}_m\mathrm{T}_{i-j}^{N-1}$。当 $m=2$ 时，$C_N^i = C_{N-1}^i + C_{N-1}^{i-1}$，此即贾宪定理。

(f) ${}_m\mathrm{T}_i^N = 0$，当 $i<0$ 或 $i>k$ 时。

2.5.2　特性三角与贾宪三角

当 $m=2$ 时，特性三角即著名的贾宪三角[8]。公元 1050 年左右，我国古代著名数学家贾宪在“开方作法本源图”中，作出幂数为六的贾宪三角，比西方的帕斯卡（Pascal）三角（1650 年）约早 600 年。我们推导出 m 为任意正整数时的特性三角，可称之为广义贾宪三角（extended Jia-Xian triangle）。

2.5.3　特征三角定理 ${}_m\mathrm{T}_i^N = \sum_{j=0}^{m-1} {}_m\mathrm{T}_{i-j}^{N-1}$ 的证明

已知：当 $N=1$ 时，$m^1 = 1+1+\cdots+1 = m$ 成立。

当 $N=2$ 时，$m^2 = 1+2+\cdots+m+\cdots+2+1 = (2(m(m-1)/2)+m = m^2$ 亦成立。现设 m^{N-1} 次方时，特性三角定理成立，往证 m^N 成立。

令 $p=m-1$，$Np=k$，$(N-1)p=k'$，

由于 $m^N = m^{N-1}\times m = m^{N-1}\times(1+1+\cdots+1)$

$= ({}_m\mathrm{T}_0^{N-1} + {}_m\mathrm{T}_1^{N-1} + \cdots + {}_m\mathrm{T}_i^{N-1} + \cdots + {}_m\mathrm{T}_{k'-1}^{N-1} + {}_m\mathrm{T}_{k'}^{N-1}) \times (1+1+\cdots+1)$，（括号内共有 m 个 1）。

将上式展开相加得：

$$
\begin{aligned}
&{}_m\mathrm{T}_0^{N-1}+{}_m\mathrm{T}_1^{N-1}+\cdots+{}_m\mathrm{T}_i^{N-1}+\cdots+{}_m\mathrm{T}_k^{N-1}\\
&\quad {}_m\mathrm{T}_0^{N-1}+\cdots+{}_m\mathrm{T}_{i-1}^{N-1}+\cdots+{}_m\mathrm{T}_{k-1}^{N-1}+{}_m\mathrm{T}_k^{N-1}\\
&\quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\\
&\quad \cdots\cdots+{}_m\mathrm{T}_{i-p}^{N-1}+\cdots\cdots\cdots\cdots+{}_m\mathrm{T}_k^{N-1}\\
\hline
&{}_m\mathrm{T}_0^N+{}_m\mathrm{T}_1^N+\cdots\cdots+{}_m\mathrm{T}_i^N+\cdots\cdots\cdots\cdots\cdots+{}_m\mathrm{T}_k^N
\end{aligned}
$$

上式共有 m 行，每下一行向右错后一项，将此 m 行展开式逐项对齐相加可得：

$$
{}_m\mathrm{T}_0^N={}_m\mathrm{T}_0^{N-1}=1,
$$

$$
{}_m\mathrm{T}_1^N={}_m\mathrm{T}_1^{N-1}+{}_m\mathrm{T}_0^{N-1}=(N-1)+1=N,
$$

$$
{}_m\mathrm{T}_i^N=\sum_{j=0}^{p}\mathrm{T}_{i-j}^{N-1}
$$

$$
{}_m\mathrm{T}_k^N=\mathrm{T}_k^{N-1}=1
$$

亦即 $${}_m\mathrm{T}_j^N=\sum_{j=0}^{m-1}{}_m\mathrm{T}_{i-j}^{N-1} \tag{3}$$

是即所证。

2.5.4 多项式定理

$$
m^N=\sum_{i=0}^{p}{}_m\mathrm{T}_j^N;\ (X_0+X_1+\cdots+X_p)^N={}_m\mathrm{T}_0^N\cdot x_0^N+{}_m\mathrm{T}_1^N\cdot x_0^{N-1}\cdot x_1+\cdots+{}_m\mathrm{T}_i^N[X^N(i)]+\cdots+{}_m\mathrm{T}_{k-1}^N\cdot x_p^{N-1}\cdot x_{p-1}+{}_m\mathrm{T}_k^N\cdot x_p^N \tag{4}
$$

$[X^N(i)]$ 为特性值为 i 的齐次项集合，故 x_0，x_1，$\cdots x_p$ 各次幂，可按 $m-N$ 空间顶点方式呈对称排列。此即多项式定理。

3 组合编码数计算公式的推导

3.1 一一对应编码

设有 N 个码元对 N 个码像进行编码，如果不加限制，每个码元都有 N 种选择，N 个码元共有 N^N 个组合编码数。但如限制每一个码像只能有 *1* 个码元，则第一个码元可有 N 种选择，第二个码元只有 $N-1$ 种选择，第 N 个码元只有 *1* 种选择。所以在一一对应编码时，共有 $N!$ 种组合编码数。

3.2 简并编码

当码元数 N 多于码像数 m 时，出现简并编码，例如当码元数为 4 个 (A, B, C, D) 对 2 个码像 $(1, 2)$ 进行编码时，如果不加限制，共有 $2^4=16$ 个组合编码数，可以分为以下 3 种情况：

(a) 一个码像有 3 个码元，另一个码像有一个码元。先设码像 1 有 3 个码元，码像 2 有 1 个码元，此时的组合编码数为：

$$
D_N=\mathrm{C}_4^3\cdot\mathrm{C}_1^1=\frac{4!}{3!1!}=4
$$

D_N 为码元的组合数。但码像之间可以互换，即

$$
D_M=\frac{2!}{1!1!}=2
$$

D_M 为码象的组合数。故此一情况的总组合编码数为：

$D_T=D_N\cdot D_M=4\times 2=8$。其编码方式如下所示。

$[(A, B, C)_1;(D)_2];[(D)_1;(A, B, C)_2];$
$[(A, B, D)_1;(C)_2];[(C)_1;(A, B, D)_2];$
$[(A, C, D)_1;(B)_2];[(B)_1;(A, C, D)_2];$
$[(B, C, D)_1;(A)_2];[(A)_1;(B, C, D)_2]$

(b) 2 个码像各有 2 个码元，则

$$
D_N=\frac{4!}{2!2!}=6,
$$

$$
D_M=\frac{2!}{2!}=1,
$$

$D_T=D_N\cdot D_M=6\times 1=6$。其编码方式如下所示。

$[(A, B)_1;(C, D)_2];[(C, D)_1;(A, B)_2];$
$[(A, C)_1;(B, D)_2];[(B, D)_1;(A, C)_2];$
$[(B, C)_1;(A, D)_2];[(A, D)_1;(B, C)_2]$

(c) 一个码像有 4 个码元，另一个码像没有码元，则

$$
D_N=\frac{4!}{4!}=1
$$

$$
D_M=\frac{2!}{1!1!}=2
$$

$D_T=1\times 2=2$。

其编码方式如下所示。

$[(A, B, C, D)_1;()_2];\ [()_1;(A, B, C, D)_2]$

(d) 三种情况总计：$8+6+2=16=2^4$。

3.3 组合编码公式

(a) 设用 N 个码元，对 m 个码像进行编码，并设 $N>m$，其中第 1 个码像有 d_1 个码元，第 2 个码像有 d_2 个码元，… 第 m 个码像有 d_m 个码元。则 $d_1+d_2+\cdots+d_m=N$，并设 $d_1, d_2, \cdots d_m$ 两两各不相等。则第一个码元有 $\mathrm{C}_N^{d_1}$ 种选择，第二个码元有 $\mathrm{C}_{N-d_1}^{d_2}$ 种选择，… 第 m 个码元有 $C_{d_m}^{d_m}$ 种选择，则

$$
D_N=\mathrm{C}_N^{d_1}\cdot\mathrm{C}_{N-d_1}^{d_2}\cdot\cdots\cdot\mathrm{C}_{d_m}^{d_m}=\frac{N!}{d_1!d_2!\cdots d_m!} \tag{5}
$$

D_N 为码元组合数；

$$
D_M=\frac{m!}{1!1!\cdots 1!}=m! \tag{6}
$$

D_M 为码象组合数；

$$
D_t=D_N\cdot D_M=\frac{N!}{d_1!d_2!\cdots d_m!}\cdot m! \tag{7}
$$

D_t 为总的组合编码数。

(b) 设用 N 个码元,对 m 个码像进行编码,共有 r 种编码方式,其中 d_1 简并度(即每一码像有 d_1 个码元) 共有 m_1 个码像, d_2 简并度共有 m_2 个码像, …d_r 简并度共有 m_r 个码像, 则有: $m_1 + m_2 + \cdots + m_r = m$, $m_1 d_1 + m_2 d_2 + \cdots + m_r d_r = N$, 其中 $d_1, d_2, \cdots d_r$ 两两各不相等。则,

$$D_N = \frac{N\ !}{(d_1\ !)^{m_1}(d_2\ !)^{m_2}\cdots(d_m\ !)^{m_r}} \tag{8}$$

$$D_M = \frac{m\ !}{m_1\ !m_2\ !\cdots m_r\ !} \tag{9}$$

$$D_r = D_N \cdot D_M = \frac{N\ !}{(d_1\ !)^{m_1}(d_2\ !)^{m_2}\cdots(d_r\ !)^{m_r}} \cdot \frac{m\ !}{m_1\ !m_2\ !\cdots m_r\ !} \tag{10}$$

4　遗传密码的组合编码数

4.1　最大可能组合编码数 C_M

遗传密码是用 64 个三联密码子(码元)对 21 种码像 (20 种氨基酸和终止密码 X) 进行编码, 64 个码元对 21 种码像的最大可能编码数 $C_M = 21^{64} = 4.1883 \times 10^{84}$,此值与 Bertman 估计的 $10^{71} - 10^{84}$ 中的最大值属于相同数量级, 但 C_M 中包括了一些码像可能没有码元编码的特殊情况。

4.2　基因组遗传密码的组合编码数 C_G

在基因组遗传密码中, 单一码元的码像有 2 个(M,W),有 2 个码元的码像有 9 个(H,Q,F,Y,C,N,K,D,E);有 3 个码元的码像有 2 个(I,X),有 4 个码元的码像有 5 个(P,T,V,A,G),有 6 个码元的码像有 3 个(L,S,R),代入组合编码公式得:

$$C_G = \frac{64!}{(1!)^2(2!)^9(3!)^2(4!)^5(6!)^3} \cdot \frac{21!}{2!9!2!5!3!}$$
$$= 2.3040 \times 10^{69} \times 4.8886 \times 10^{10} = 1.1263 \times 10^{80}$$

4.3　线粒体遗传密码的组合编码数 C_T

在线粒体遗传密码中[9],没有单一码元的码像, 也没有 3 个码元的码像。有 2 个码元的码像 12 个(H,Q,F,Y,C,N,K,D,E,M,W,I),有 4 个码元的码像有 7 个(P,T,V,A,G,R,X),有 6 个码元的码像有 2 个(L,S)。代入组合编码公式得:

$$C_T = \frac{64\ !}{(2\ !)^{12}(4!)^7(6!)^2} \cdot \frac{21!}{12!7!2!}$$
$$= 1.3029 \times 10^{70} \times 1.0582 \times 10^{9} = 1.3787 \times 10^{79}$$

$C_G / C_T = 8.1693$,计算结果表明 C_G 比 C_T 约大 8 倍,相差约 1 个数量级。

5　遗传密码起源的分析和讨论

基因组遗传密码的组合编码数为 1.13×10^{80}, 此一数值大得惊人。可以认为,遗传密码的指定,是一个接近零的小概率事件, 亦即遗传密码的指定是各种可能组合编码数的 1.13×10^{80} 分之一。Crick 认为遗传密码的指定是一次偶发的冻结事件, 可能是由于这一概率太小的缘故。但目前, 我们已有两组氨基酸遗传密码,一是基因组遗传密码,一是线粒体遗传密码。二者共有 16 个氨基酸(码像)的编码完全相同,只有 4 个氨基酸(I,M,R,W)和终止密码 X 不同。这些都可以用编码平面的对称破缺加以解释[10]。我们认为,基因组遗传密码和线粒体遗传密码二者之间并不是完全无关的独立的偶发事件, 二者之间有着密切的联系。由于线粒体在进化上出现较早, 可以认为线粒体的遗传密码亦出现较早。线粒体遗传密码的特点是只有 2,4,6 偶数简并度而没有 1,3 奇数简并度,而且其起始密码子有 2 个,终止密码子有 4 个,故其对称程度较高。根据以上分析, 偶数简并度遗传密码的出现, 应是进化历程的早发事件。偶数简并度的码元,其第 3 位碱基通过嘧啶及嘌呤的聚类, 例如, 天冬氨酸 D 的编码为 GA(C 与 T) 可聚类为 GAY(Y 为嘧啶), 而谷氨酸 E 的编码为 GA(A 和 G) 可聚类为 GAR(R 为嘌呤), 即可将 64 个三联密码子简并为 32 个三联密码配对 (codon pairs)。用 32 个密码配对同时对 21 个码像进行编码,可以大大压缩编码空间的维数,从而大大减少遗传密码的组合编码数。32 个码元对 21 个码像的最大可能的组合编码数 $C_M' = 21^{32} = 2.0465 \times 10^{42}$, 大大小于 $21^{64}(4.1883 \times 10^{84})$ 这一数字。如以 32 个码元按线粒体遗传密码的组合方式进行编码, 其组合编码数 C_T' 为:

$$C_T' = \frac{32\ !}{(1!)^{12}(2\ !)^7(3!)^3} \cdot \frac{21!}{12!7!2!} = 6.0426 \times 10^{40}$$

C_T' 亦大大小于 C_T (1.3787×10^{79})。故此可以认为,遗传密码最先出现的是用 32 个密码配对同时对 21 个码像进行偶数简并编码的线粒体遗传密码, 其后由于两个简并平面 1001λλ 平面 (I 和 M 组成的编码平面) 和 011λ1λ 平面 (W 和 X 组成的编码平面) 的对称破缺,出现 1,3 奇数简并度,包括 1 个起始密码子和 3 个终止密码子, 从而演化出基因组遗传密码的编码方式[11]。但由于 $C_T'(6.0426 \times 10^{40})$ 也是一个很大的数目, 因此在线粒体遗传密码之前是否还有更早的遗传编码格式或是直接来源于偶发冻结事件,值得进一步深入研究。

6　小结

综上所述, 应用 m 元 N 维编码空间拓扑特性

的分析方法，对遗传密码的组合编码格式，进行了研究，并由此推导出遗传密码的组合编码格式是在 21 元 64 维编码空间中的一个特定顶点的结论。应用组合数学的方法计算出基因组遗传密码以及线粒体遗传密码的组合编码数。通过对基因组遗传密码以及线粒体遗传密码的分析，认为基因组遗传密码的 1，3 两种奇数简并度，可能来源于线粒体遗传密码的对称破缺。在线粒体遗传密码中，第三位变偶位的密码子只能识别嘧啶 Y 及嘌呤 R 而不能识别酮基 K 和氨基 M，这一现象是否与变偶位密码子的“质量识别”特性有关，拟在另文进行探讨。

参考文献：

[1] Nirenberg MW, Matthaei H. The dependence of cell-free synthesis in E. coli naturally occuring or synthetic polyribosomas[J]. *Proc Natl Acad Sci*, 1961,47:1588-1602.

[2] Hornos JE, Hornos YM. Algebraic model for the evolution of the genetic code[J]. *Physical Review Letters*, 1993,71(26):4401-4404.

[3] Crick HHC. Codon anticodon pairing-The wobble hypothesis[J]. *J Mol Biol*, 1966,19:548-555.

[4] 陈惟昌，陈志华，陈志义．遗传密码的简并及其高维空间的拓扑结构[J]．自然科学进展，1999,9(2):175-178.

[5] 陈志华，陈惟昌，邱红霞，等．氨基酸的分子结构与遗传密码简并及二维集合分类[J]．生物物理学报，2001,17(1):187-194.

[6] 吴品三．近世代数[M]．北京：高等教育出版社，1987.7-14.

[7] 陈惟昌，陈志华，陈志义，等．遗传密码和 DNA 序列的高维空间数字编码[J]．生物物理学报，2000,16(4):760-768.

[8] 谷超豪．数学词典[M]．上海：上海辞书出版社，1995.654.

[9] 王文清，周成，刘枫，等．遗传密码表与《易经》[J]．北京大学学报(自然科学版)，1998,34(4):471-480.

[10] 陈惟昌，陈志华，邱红霞，等．遗传密码的高维空间对称性[J]．生物物理学报，2001,17(2):337-343.

[11] 陈惟昌，陈志华，王自强，等．线粒体遗传密码及基因组遗传密码的对称分析[J]．生物物理学报，2002,18(1):87-94.

ANALYSIS OF COMBINATORIAL CODING NUMBERS OF THE GENETIC CODE PATTERNS

CHEN Wei-chang[1], CHEN Zhi-yi[2], CHEN Zhi-hua[3], WANG Zi-qiang[1]

(1. Department of Biophysics, China Japan Friendship Institute of Medical Sciences, Beijing 100029, China;
2. National Laboratory of Pattern Recognition, The Chinese Academy of Sciences, Beijing 100080, China;
3. Department of Biochemistry and Molecular Biology, China Japan Friendship Institute of Medical Sciences, Beijing 100029, China)

Abstract: The coding pattern which uses N codons to encode m objects is a vertex in N dimension space of m elements. The combinatorial coding number of 64 codons to encode 20 amino acids and the terminate code is very huge. The topological properties of the polynomial high dimension spaces (the $m-N$ spaces) were first analyzed and the characteristic triangles (Chen Weichang Triangles) of the $m-N$ spaces were obtained. A mathematical proof of the characteristic triangles was also given. Obviously, the coding pattern of the genetic code is a vertex in a 64 dimension space of 21 elements. Using the combinatorial mathematical method, the following combinatorial numbers of genetic coding patterns had been calculated: the maximum combinatorial number of genetic coding patterns C_m ($C_m = 4.19 \times 10^{84}$); the combinatorial number of genomic coding pattens C_G ($C_G = 1.13 \times 10^{80}$), the combinatorial number of mitochondrial coding patterns C_T ($C_T = 1.38 \times 10^{79}$). It is suggested that the determination of the genetic code is an event of extremely small probability. The origin of the genomic genetic codes might be from the symmetry breaking of the triplets pairs of the mitochondrial codes.

Key Words: Genetic code; Coded object (codim) and coding element (codon); Combinatorial coding numbers; Polynomial high dimension space (hypergrid space); Polynomial theorem; Chen Weichang triangle (extended Jia Xian and Pascal triangle)