

基于模糊邻近关系的 p53 抑癌基因家族聚类分析

仇建烨, 浦 琰, 朱 平*

(江南大学 理学院, 江苏 无锡 214122)

摘要: 模糊聚类技术在揭示数据的联系及数据间的依赖关系等方面占有重要作用。作者根据模糊等价关系的性质, 利用加权汉明距离法, 建立了序列间的模糊邻近关系, 进而构造模糊等价矩阵进行聚类分析。选取了 18 条人类 p53 及其家族成员 p63、p73 肿瘤蛋白 mRNA 序列, 以序列碱基的比例为指标, 进行聚类分析。发现, 当 $\lambda=0.951\ 59$ 时, 结果分为 3 类, p53、p63、p73 各自成为一类。从聚类角度看 p53、p63、p73 的结构和功能是有区别的。此方法对预测未知基因的结构和功能有一定的生物意义。

关键词: p53;p63;p73;模糊邻近关系;模糊等价矩阵;聚类分析

中图分类号: O 159; Q 74 **文献标志码:** A **文章编号:** 1673—1689(2013)08—0814—07

Clustering Analysis of Family Members of p53 Tumor Suppressor Gene Based on Fuzzy Proximity Relation

QIU Jian-ye, PU Yan, ZHU Ping*

(School of Science, Jiangnan University, Wuxi 214122, China)

Abstract: Fuzzy clustering technology plays an important role in revealing the connection and dependence between the data. Therefore, according to the properties of fuzzy equivalence relation, this paper use the weighted hamming distance method to establish fuzzy proximity relation between sequences, and then make fuzzy equivalence matrix to being clustering analysis. Then select 18 human p53 and its family members p63 and p73 tumor protein mRNA sequences, use the base contents of sequences as index to being clustering analysis. Find that when $\lambda=0.951\ 59$, the result is divided into 3 groups, p53, p63 and p73 each becomes a class. From the perspective of clustering, p53, p63 and p73 have differences in structures and functions. This method has biological significance in predicting the structures and functions of unknown genes.

Keywords: p53, p63, p73, fuzzy proximity relation, fuzzy equivalence matrix, clustering analysis

聚类分析是按确定的标准对客观事物进行分类的数学方法。通过聚类可以区别不同的事物并认

收稿日期: 2012-11-09

基金项目: 国家自然科学基金项目(11271163); 中央高校基本科研业务费专项资金项目(JUSRP51317B)。

* 通信作者: 朱 平(1962—), 女, 安徽巢湖人, 工学博士, 教授, 硕士研究生导师, 主要从事生物数学、计算机理论方面的研究。

E-mail: zhuping@jiangnan.edu.cn

识事物间的相似性。聚类分析在很多领域有着广泛应用,如文献[1]对江西茶鲜叶重金属状况进行了聚类分析。然而由于现实的分类往往伴随着模糊性,所以用模糊理论来进行聚类分析会显得更自然,更符合客观实际,这就是模糊聚类分析。随着模糊理论的发展,模糊聚类分析得到了广泛的应用。模糊聚类方法大致分成了3类:模糊等价矩阵动态聚类分析法^[2]、模糊C-均值聚类算法(FCM)^[3-4]、基于摄动的模糊聚类分析法^[5]。

p53 基因是生物学领域中研究热度最高的基因之一,是一种很重要的抑癌基因,p53 基因是迄今发现与人类肿瘤相关性最高的基因。随着 p53 的深入研究,p53 家族成员自 1997 年来也不断被发现。p63^[6]和 p73^[7]与 p53 基因有较大的同源性,因而被认定为 p53 家族成员。和 p53 一样,p63、p73 也有诱导细胞周期阻滞和凋亡的功能等,但它们也有不同的结构和功能,深入了解它们彼此之间的相似性和差异将对理解肿瘤发生的机制产生重要的影响。

作者利用模糊邻近关系对 18 条人类 p53、p63 和 p73 肿瘤蛋白 mRNA 序列进行分类研究。分析讨论了不同类序列的组成结构和功能的差异。

1 基于模糊邻近关系的聚类分析方法

我们知道,一个合适的分类应当满足:1)自反性:任何一个对象必须和自己在同一类。2)对称性:若对象 u 与对象 v 同类,则与也同类。3)传递性:若对象 u 与对象 v 同类,对象 v 与对象 w 同类,则 u 与 w 也同类。而满足这 3 个条件的关系即是一个等价关系。模糊聚类分析则是根据模糊等价关系进行的。但是由于模糊等价关系中的传递性难于验证,故转而根据模糊邻近关系的传递闭包来进行聚类。

1.1 确定分类对象的特征性指标

选择能够描述 n 个分类对象 $\{u_1, u_2, \dots, u_n\}$ 的 m 个特征性指标 $x_{ij} (i=1, 2, \dots, n; j=1, 2, \dots, m)$ 。要求个指标能够很好地描述被分类对象的特征。

1.2 数据规范化

为了消除特性指标数量级不同的影响,需对特性指标进行规范化。这里用最大值规范化,如下:

$$M_j = \max(x_{1j}, x_{2j}, \dots, x_{nj}), j=1, 2, \dots, m,$$

然后作变换: $x'_{ij} = \frac{x_{ij}}{M_j}$,

$$i=1, 2, \dots, n; j=1, 2, \dots, m.$$

1.3 构造模糊邻近矩阵

设 $U=\{u_1, u_2, \dots, u_n\}$ 为有限域,则 u 上的模糊关系可以表示为一个 n 阶模糊矩阵 $R=(r_{ij})_{n \times n}$, 若满足:1)自反性: $r_{ij}=1 (i=1, 2, \dots, n)$; 2)对称性: $r_{ij}=r_{ji} (i, j=1, 2, \dots, n)$; 3)传递性: $R \circ R \subseteq R$, 其中 $R_2 = R \circ R = (r'_{ij})_{n \times n}$, $r'_{ij} = \bigvee_{k=1}^n (r_{ik} \wedge r_{kj})$ 。则称是 $R=(r_{ij})_{n \times n}$ 模糊等价矩阵;若只满足自反性和对称性则称 $R=(r_{ij})_{n \times n}$ 是模糊邻近矩阵。

构造模糊邻近矩阵,就是要得到对象间的相似度。作者考虑到由于各个特性指标在确定对象相似程度时起不同作用,故确定各个指标的权重就有重要的意义。所以先用变异系数法先求得各指标的权值,其步骤如下:

1)求得各指标的均值 $\bar{x}_i (i=1, 2, \dots, m)$ 和标准差 $\sigma_i (i=1, 2, \dots, m)$ 。

2)求得各指标的变异系数:

$$V_i = \frac{\sigma_i}{\bar{x}_i} (i=1, 2, \dots, m).$$

3)对变异系数作归一化处理,便得到各指标的权值:

$$\omega_i = \frac{V_i}{\sum_{i=1}^m V_i} (i=1, 2, \dots, m).$$

在确定权值的基础上定义了加权汉明距离及相似度,距离越大,相似程度越小。因此,我们建立如下的加权汉明距离:

定义 1 用 m 个指标描述对象就得到 m 维指标向量,定义两对象 $u_i=(x_{i1}, x_{i2}, \dots, x_{im})$ 和 $u_j=(x_{j1}, x_{j2}, \dots, x_{jm})$ 之间的加权汉明距离为:

$$d(u_i, u_j) = \sum_{k=1}^m \omega_k |x_{ik} - x_{jk}|, \\ i, j=1, 2, \dots, n$$

两对象之间的相似度为:

$$r_{ij} = 1 - c[d(u_i, u_j)]^\alpha \in [0, 1]$$

这里 c 和 α 为常数, $i, j=1, 2, \dots, n$ 。

由定义 1 即可以得到相应的模糊邻近矩阵 $R=(r_{ij})_{n \times n}$ 。

1.4 聚类

由于由上述方法构造出的对象与对象之间的模糊邻近矩阵 $R=(r_{ij})_{n \times n}$ 满足自反性和对称性,而不一定具有传递性,从而不一定是模糊等价矩阵。故必须由模糊邻近矩阵 R 出发构造一个新的模糊等

价矩阵,然后以此模糊等价矩阵作为基础,进行动态聚类。模糊邻近矩阵 R 的传递闭包 $t(R)$ 就是一个模糊等价矩阵^[8]。

利用平方自合成方法求出模糊邻近矩阵 R 的传递闭包 $t(R)$:

$$R^2 \rightarrow R^4 \rightarrow \Lambda \rightarrow R^2$$

直至出现 $R^2 = R^{2^{k+1}}$, 则 $t(R) = R^2$ 。

设 R 为模糊邻近矩阵, $t(R) = R^k$, R_λ 为 R 的 λ 截矩阵, $t(R)_\lambda$ 为 $t(R)$ 的 λ 截矩阵。若 U 是有限集, 则 $\forall \lambda \in [0, 1]$, 恒有 $t(R_\lambda) = (t(R))_\lambda$ ^[9]。故对于模糊邻近矩阵 R 来说, 关于截关系 R_λ 的相似类可以归并为关于 $(t(R))_\lambda$ 的等价类。因此我们选取阈值 $\lambda \in [0, 1]$, 求出 $t(R)$ 的 λ 截矩阵 $t(R)_\lambda$, 然后按 $t(R)_\lambda$ 进行聚类, 所得到的分类就是在 λ 水平上的等价分类。具体聚类原则如下: 设 $t(R) = (\bar{r}_{ij})_{n \times n}$, $t(R)_\lambda = (\bar{r}_{ij}(\lambda))_{n \times n}$, 则

$$\bar{r}_{ij}(\lambda) = \begin{cases} 1, & \bar{r}_{ij} \geq \lambda \\ 0, & \bar{r}_{ij} < \lambda \end{cases}$$

对于 $u_i, u_j \in U$, 若 $\bar{r}_{ij}(\lambda) = 1$, 则在 λ 水平上将对象 u_i 和对象 u_j 归为同一类。当 λ 在 $[0, 1]$ 中取不同值时, 相应的分类随之改变, 从而形成一个聚类图。得到的

模糊分类具有动态性, 可根据不同的要求进行分类。

2 p53 基因及其家族成员 p63 和 p73 基因序列聚类分析

2.1 聚类对象

从 GenBank 数据库里选取了人类 p53 肿瘤蛋白及其家族成员 p63, p73 mRNA 序列共 18 条, 编号分别为 NM_001126112.2, NM_001126113.2, NM_001126114.2, NM_001126115.1, NM_001126116.1, NM_001114982.1, NM_001114980.1, NM_001114978.1, NM_001114981.1, NM_001114979.1, NM_001204185.1, NM_001204186.1, NM_001204188.1, NM_001204187.1, NM_001204189.1, NM_001204192.1, NM_001204190.1, NM_001204191.1。以所选序列为对象, 分别给予变量名为 $u_i (i=1, 2, \dots, 18)$ 。

2.2 模糊聚类分析

2.2.1 特征性指标 用 Matlab 编程得到所选序列的碱基 A、G、C、U 的比例 (A%、G%、C%、U%), 见表 1。以 A%、G%、C%、U% 为特性指标, 得到四维指标向量 $u_i = (x_{i1}, x_{i2}, x_{i3}, x_{i4})$ 及特性指标矩阵 $U = [u_1, u_2, \dots, u_{18}]$ 。

表 1 序列碱基比例

Table 1 Base contents of sequences

序列编号	A 比例/%	G 比例/%	C 比例/%	U 比例/%
NM_001126112.2	0.209 81	0.249 61	0.284 00	0.256 57
NM_001126113.2	0.211 62	0.248 21	0.281 40	0.258 77
NM_001126114.2	0.216 23	0.247 06	0.278 27	0.258 44
NM_001126115.1	0.216 64	0.244 83	0.271 69	0.266 84
NM_001126116.1	0.223 38	0.242 10	0.265 81	0.268 72
NM_001114982.1	0.277 27	0.225 38	0.228 03	0.269 32
NM_001114980.1	0.287 42	0.205 45	0.217 37	0.289 76
NM_001114978.1	0.287 19	0.201 53	0.220 36	0.290 92
NM_001114981.1	0.288 94	0.204 87	0.216 82	0.289 38
NM_001114979.1	0.275 26	0.218 12	0.233 10	0.273 52
NM_001204185.1	0.210 59	0.285 91	0.318 68	0.184 82
NM_001204186.1	0.209 77	0.288 35	0.316 64	0.185 25
NM_001204188.1	0.210 85	0.286 89	0.315 25	0.187 01
NM_001204187.1	0.209 53	0.287 31	0.320 10	0.183 06
NM_001204189.1	0.209 31	0.283 64	0.320 91	0.186 14
NM_001204192.1	0.211 12	0.282 84	0.319 30	0.186 74
NM_001204190.1	0.209 08	0.282 73	0.324 26	0.183 92
NM_001204191.1	0.210 41	0.282 26	0.319 43	0.187 90

2.2.2 数据规范化 对特性指标矩阵 U 作最大值规范化,得:

$$U = \begin{pmatrix} 0.726\ 14 & 0.865\ 65 & 0.875\ 84 & 0.881\ 93 \\ 0.732\ 4 & 0.860\ 79 & 0.867\ 82 & 0.889\ 49 \\ 0.748\ 36 & 0.856\ 81 & 0.858\ 17 & 0.888\ 35 \\ 0.749\ 78 & 0.849\ 07 & 0.837\ 88 & 0.917\ 23 \\ 0.773\ 1 & 0.839\ 6 & 0.819\ 74 & 0.923\ 69 \\ 0.959\ 61 & 0.781\ 62 & 0.703\ 23 & 0.925\ 75 \\ 0.994\ 74 & 0.712\ 5 & 0.670\ 36 & 0.996\ 01 \\ 0.993\ 94 & 0.698\ 91 & 0.679\ 58 & 1 \\ 1 & 0.710\ 49 & 0.668\ 66 & 0.994\ 71 \\ 0.952\ 65 & 0.756\ 44 & 0.718\ 87 & 0.940\ 19 \\ 0.728\ 84 & 0.991\ 54 & 0.982\ 79 & 0.635\ 29 \\ 0.726 & 1 & 0.976\ 5 & 0.636\ 77 \\ 0.759\ 74 & 0.994\ 94 & 0.972\ 21 & 0.642\ 82 \\ 0.725\ 17 & 0.996\ 39 & 0.987\ 17 & 0.629\ 25 \\ 0.724\ 41 & 0.983\ 67 & 0.989\ 67 & 0.939\ 83 \\ 0.730\ 67 & 0.980\ 89 & 0.974\ 7 & 0.641\ 89 \\ 0.723\ 61 & 0.980\ 51 & 1 & 0.632\ 2 \\ 0.728\ 51 & 0.978\ 88 & 0.985\ 1 & 0.645\ 88 \end{pmatrix}$$

2.2.3 模糊邻近矩阵 由变异系数法得到各指标的权值,表 2 为序列各碱基比例对应的权值。文献[10]运用序列间的距离对序列做同源性分析,进而构建系统进化树。可见用距离法能够很好的表达序列间的相似程度。这里令 c 和 α 都为 1,通过定义 1

计算得到序列的相似度,计算得到模糊邻近矩阵 R ,见表 3。

2.2.4 聚类 由于求模糊邻近矩阵 R 的传递闭包 $t(R)$ 计算量大,故用 Matlab 编程得到模糊等价矩阵 $t(R)=R^{16}$,见表 4。

选取阈值 λ , 求出 $t(R)$ 的 λ 截矩阵 $t(R)_\lambda$, 然后按 $t(R)_\lambda$ 对对象进行聚类,结果见表 5。

具体的聚类结果如聚类图见图 1。聚类图给出了各 λ 值对应的分类,形成动态聚类,便于全面了解对象聚类。

3 结果与分析

当 $\lambda=0.951\ 59$ 时,分为 3 类,第 1 类 u_1-u_5 为 p53 序列,第 2 类 u_6-u_{10} 为 p63 序列,第 3 类 $u_{10}-u_{18}$ 为 p73 序列。虽然 p53 与 p63 和 p73 的同源比较高,如在 DNA 结合结构域 p53 与 p63 的同源性达到 60%,与 p73 的同源性达到了 63%。但从聚类的角度看,不在一类的基因在结构和功能上还是有差异的。

表 2 各指标的权值

Table 2 Weights of each index

指标	x_1	x_2	x_3	x_4	总和
均值	0.802 632	0.879 928	0.865 461	0.803 404	—
标准差	0.114 458	0.111 839	0.127 542	0.155 933	—
变异系数	0.142 6	0.127 1	0.147 4	0.194 1	0.611 2
权重	0.23	0.21	0.24	0.32	1

表 3 模糊邻近矩阵

Table 3 Fuzzy proximity matrix

1	0.993 2	0.986 74	0.970 67	0.956 9	0.873 21	0.820 24	0.818 51	0.818 62	0.868 65	0.868 35	0.869 14	0.872 38	0.864 74	0.870 03	0.871 81	0.865 59	0.873 99
0.993 2	1	0.992 81	0.977 48	0.963 71	0.880 01	0.827 04	0.825 31	0.825 42	0.875 46	0.862 79	0.862 34	0.867 23	0.857 94	0.86	0.867 1	0.858 78	0.868 14
0.986 74	0.992 81	1	0.983 94	0.970 16	0.886 47	0.833 5	0.831 77	0.831 88	0.881 91	0.856 33	0.855 88	0.860 77	0.851 48	0.856 76	0.860 64	0.852 32	0.861 68
0.970 67	0.977 48	0.983 94	1	0.986 23	0.902 53	0.849 57	0.847 83	0.847 94	0.897 98	0.840 27	0.839 82	0.844 71	0.835 42	0.840 7	0.844 58	0.836 26	0.845 61
0.956 9	0.963 71	0.970 16	0.986 23	1	0.916 31	0.863 34	0.861 6	0.861 71	0.911 75	0.826 49	0.826 05	0.830 93	0.821 65	0.826 93	0.830 8	0.822 49	0.831 84
0.873 21	0.880 01	0.886 47	0.902 53	0.916 31	1	0.947 03	0.945 3	0.945 41	0.984 74	0.742 8	0.742 35	0.747 24	0.737 95	0.743 23	0.747 11	0.738 79	0.748 51
0.820 24	0.827 04	0.833 5	0.849 57	0.863 34	0.947 03	1	0.993 47	0.997 54	0.951 59	0.689 83	0.689 38	0.694 27	0.684 98	0.690 27	0.694 14	0.685 83	0.695 18
0.818 51	0.825 31	0.831 77	0.847 83	0.861 6	0.945 3	0.993 47	1	0.991 86	0.949 85	0.688 1	0.687 65	0.692 54	0.683 25	0.688 53	0.692 41	0.684 09	0.693 45
0.818 62	0.825 42	0.831 88	0.847 94	0.861 71	0.945 41	0.997 54	0.991 86	1	0.949 96	0.688 21	0.687 76	0.692 65	0.683 36	0.688 64	0.692 52	0.684 2	0.693 56
0.868 65	0.875 46	0.881 91	0.897 98	0.911 75	0.984 74	0.951 59	0.949 85	0.949 96	1	0.738 24	0.737 8	0.742 69	0.733 4	0.738 68	0.742 55	0.734 24	0.743 59
0.868 35	0.862 79	0.856 33	0.840 27	0.826 49	0.742 8	0.689 83	0.688 1	0.688 21	0.738 24	1	0.995 59	0.994 13	0.995 15	0.994 22	0.994 77	0.991 36	0.993 25
0.869 14	0.862 34	0.855 88	0.839 82	0.826 05	0.742 35	0.689 38	0.687 65	0.687 76	0.737 8	0.995 59	1	0.995 51	0.994 08	0.992 06	0.991 31	0.988 25	0.990 08
0.872 38	0.867 23	0.860 77	0.844 71	0.830 93	0.747 24	0.694 27	0.692 54	0.692 65	0.742 69	0.994 13	0.995 11	1	0.990 71	0.991 26	0.993 54	0.985 49	0.992 2
0.864 74	0.857 94	0.851 48	0.835 42	0.821 65	0.737 95	0.684 98	0.683 25	0.683 36	0.733 4	0.995 15	0.994 08	0.990 71	1	0.993 17	0.990 84	0.992 28	0.989 81
0.870 03	0.863 22	0.856 76	0.840 7	0.826 93	0.743 23	0.690 27	0.688 53	0.688 64	0.738 68	0.994 22	0.992 06	0.991 26	0.993 17	1	0.996 12	0.994 23	0.995 09
0.871 82	0.867 1	0.860 64	0.844 58	0.830 8	0.747 11	0.694 14	0.692 41	0.692 52	0.742 55	0.994 77	0.991 31	0.993 54	0.990 84	0.996 12	1	0.991 52	0.997 64
0.865 59	0.858 78	0.852 32	0.836 26	0.822 49	0.738 79	0.685 83	0.684 09	0.684 2	0.734 24	0.991 36	0.988 25	0.985 49	0.992 28	0.994 23	0.991 52	1	0.990 65
0.873 99	0.868 14	0.861 68	0.845 61	0.831 84	0.748 15	0.695 18	0.693 45	0.693 56	0.743 59	0.993 25	0.990 08	0.992 2	0.989 81	0.995 09	0.997 64	0.990 65	1

表 4 模糊邻近矩阵的传递闭包

Table 4 Transitive closure of fuzzy proximity matrix

$$t(R) = \begin{pmatrix} 1 & 0.993\ 2 & 0.992\ 81 & 0.983\ 94 & 0.983\ 94 & 0.916\ 31 & 0.916\ 31 & 0.916\ 31 & 0.916\ 31 & 0.916\ 31 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 \\ 0.993\ 2 & 1 & 0.992\ 81 & 0.983\ 94 & 0.983\ 94 & 0.916\ 31 & 0.916\ 31 & 0.916\ 31 & 0.916\ 31 & 0.916\ 31 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 \\ 0.992\ 81 & 0.992\ 81 & 1 & 0.983\ 94 & 0.983\ 94 & 0.916\ 31 & 0.916\ 31 & 0.916\ 31 & 0.916\ 31 & 0.916\ 31 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 \\ 0.983\ 94 & 0.983\ 94 & 0.983\ 94 & 1 & 0.986\ 23 & 0.916\ 31 & 0.916\ 31 & 0.916\ 31 & 0.916\ 31 & 0.916\ 31 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 \\ 0.983\ 94 & 0.983\ 94 & 0.983\ 94 & 0.986\ 23 & 1 & 0.916\ 31 & 0.916\ 31 & 0.916\ 31 & 0.916\ 31 & 0.916\ 31 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 \\ 0.916\ 31 & 0.916\ 31 & 0.916\ 31 & 0.916\ 31 & 0.916\ 31 & 1 & 0.951\ 59 & 0.951\ 59 & 0.951\ 59 & 0.984\ 74 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 \\ 0.916\ 31 & 0.916\ 31 & 0.916\ 31 & 0.916\ 31 & 0.916\ 31 & 0.951\ 59 & 1 & 0.993\ 47 & 0.997\ 54 & 0.951\ 59 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 \\ 0.916\ 31 & 0.916\ 31 & 0.916\ 31 & 0.916\ 31 & 0.916\ 31 & 0.951\ 59 & 0.993\ 47 & 1 & 0.993\ 47 & 0.951\ 59 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 \\ 0.916\ 31 & 0.916\ 31 & 0.916\ 31 & 0.916\ 31 & 0.916\ 31 & 0.951\ 59 & 0.997\ 54 & 0.993\ 47 & 1 & 0.951\ 59 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 \\ 0.916\ 31 & 0.916\ 31 & 0.916\ 31 & 0.916\ 31 & 0.916\ 31 & 0.984\ 74 & 0.951\ 59 & 0.951\ 59 & 0.951\ 59 & 1 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 \\ 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 1 & 0.995\ 99 & 0.995\ 11 & 0.995\ 15 & 0.994\ 77 & 0.994\ 77 & 0.994\ 23 & 0.994\ 77 & 0.994\ 77 \\ 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.995\ 59 & 1 & 0.995\ 11 & 0.995\ 15 & 0.994\ 77 & 0.994\ 77 & 0.994\ 23 & 0.994\ 77 & 0.994\ 77 \\ 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.995\ 11 & 0.995\ 11 & 1 & 0.995\ 11 & 0.994\ 77 & 0.994\ 77 & 0.994\ 23 & 0.994\ 77 & 0.994\ 77 \\ 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.995\ 15 & 0.995\ 15 & 0.995\ 11 & 1 & 0.994\ 77 & 0.994\ 77 & 0.994\ 23 & 0.994\ 77 & 0.994\ 77 \\ 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.994\ 77 & 0.994\ 77 & 0.994\ 77 & 0.994\ 77 & 1 & 0.996\ 12 & 0.994\ 23 & 0.996\ 12 & 0.996\ 12 \\ 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.994\ 77 & 0.994\ 77 & 0.994\ 77 & 0.994\ 77 & 0.994\ 77 & 1 & 0.994\ 23 & 0.997\ 64 & 0.994\ 23 \\ 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.994\ 23 & 0.994\ 23 & 0.994\ 23 & 0.994\ 23 & 0.994\ 23 & 0.994\ 23 & 1 & 0.994\ 23 & 0.994\ 23 \\ 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.873\ 99 & 0.994\ 77 & 0.994\ 77 & 0.994\ 77 & 0.994\ 77 & 0.994\ 77 & 0.996\ 12 & 0.997\ 64 & 0.994\ 23 & 1 \end{pmatrix}$$

表 5 聚类结果

Table 5 Results of clustering

阈值 λ	分类数	分类结果
1	18	$\{u_1\}, \{u_2\}, \{u_3\}, \{u_4\}, \{u_5\}, \{u_6\}, \{u_7\}, \{u_8\}, \{u_9\}, \{u_{10}\}, \{u_{11}\}, \{u_{12}\}, \{u_{13}\}, \{u_{14}\}, \{u_{15}\}, \{u_{16}\}, \{u_{17}\}, \{u_{18}\}$
0.997 64	17	$\{u_1\}, \{u_2\}, \{u_3\}, \{u_4\}, \{u_5\}, \{u_6\}, \{u_7\}, \{u_8\}, \{u_9\}, \{u_{10}\}, \{u_{11}\}, \{u_{12}\}, \{u_{13}\}, \{u_{14}\}, \{u_{15}\}, \{u_{16}\}, \{u_{17}\},$
IV	IV	IV
0.916 31	2	$\{u_1, u_2, u_3, u_4, u_5, u_6, u_7, u_8, u_9, u_{10}\}, \{u_{11}, u_{12}, u_{13}, u_{14}, u_{15}, u_{16}, u_{17}, u_{18}\}$
0.873 99	1	$\{u_1, u_2, u_3, u_4, u_5, u_6, u_7, u_8, u_9, u_{10}, u_{11}, u_{12}, u_{13}, u_{14}, u_{15}, u_{16}, u_{17}, u_{18}\}$

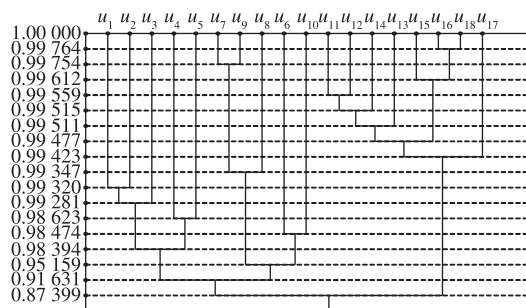


图 1 序列聚类图

Fig. 1 Diagram of sequences clustering

3.1 聚类结果中各类序列的碱基比例比较

u_1-u_5 基因序列中腺嘌呤 A 的平均比例为 21.553 6%, 鸟嘌呤 G 的平均比例为 24.636 2%, 胞嘧啶 C 的平均比例为 27.6234%, 尿嘧啶 U 的平均比例为 26.186 8%, GC 比例为 52.259 6%。

u_6-u_{10} 基因序列中腺嘌呤 A 的平均比例为 28.321 6%, 鸟嘌呤 G 的平均比例为 21.107%, 胞嘧啶 C 的平均比例为 22.313 6%, 尿嘧啶 U 的平均比例为 28.258%, GC 比例为 43.4206%。

$u_{11}-u_{18}$ 基因序列中腺嘌呤 A 的平均比例为 21.008 3%, 鸟嘌呤 G 的平均比例为 28.499 1%, 胞嘧啶 C 的平均比例为 31.932 1%, 尿嘧啶 U 的平均

比例为 18.560 5%, GC 比例为 60.431 2%。

可见第 2 类、第 3 类 GC 比例与 AU 比例分界比第 1 类明显。第 1 类、第 3 类 GC 比例比 AU 比例高, 而第 2 类 GC 比例比 AU 比例低。这与各类基因的功能有一定的关系。

3.2 聚类结果中各类序列的定位和一些功能比较

p53 基因定位于染色体 17p13.1, p63 基因定位于染色体 3q27-29, p73 定位于染色体 1p36.2-1p36.3。第 1 类序列 u_1-u_5 定位于人类染色体 17p13.1, 第 2 类序列 u_6-u_{10} 定位于人类染色体 3q28, 第 3 类序列 $u_{11}-u_{18}$ 定位于人类染色体 1p36.3。

各类基因的功能的相同与不同之处见表 6(引自文献[11])。

表 6 p53, p63 和 p73 的功能比较

Table 6 Comparisons of function between p53, p63 and p73

	p53	p63	p73
发育过程中起关键作用	-	+	+
抑制肿瘤功能	+	?	+/-
转录因子活性	+	+	+
诱导细胞凋亡功能	+	+	+
被肿瘤病毒抑制	+	-	-
被胞外刺激信号活化	+	?	some
被 MDM2 抑制	+	?	-

从表 6 可以看出,由于与 p53 具有同源性很高的 DNA 结合位点,可以确定 p63 和 p73 有转录因子活性的作用,能够诱导细胞凋亡^[12]。

p63 和 p73 在细胞发育过程中起关键作用,而 p53 没有此功能;p53 会被肿瘤病毒抑制,而 p63 和 p73 不会被肿瘤病毒蛋白抑制;p53 能被细胞外刺激信号活化,p63 有这功能则未知,而有一些 p73 可以;p53 可以被 MDM2 抑制,p63 能否被 MDM2 抑制还未确定,p73 不能被 MDM2 抑制;而在抑制肿瘤方面,p53 是抑癌基因,对于 p63 是否有抑制肿瘤的功能还未被认定,p73 有一定抑制肿瘤功能,但能否被定义为抑癌基因还需要进一步研究。

从组成结构和功能看,此分类较为合理,我们可以在已知一些基因结构和功能的情况下,运用此

模糊聚类的方法对其它一些未知基因功能进行预测,有一定生物学意义。

4 结 语

如今癌症对人类的健康威胁很大。在人类 50% 以上的肿瘤组织中均发现了 p53 基因的突变,如大肠癌、乳腺癌、胃癌、肺癌等^[13-16],一旦 p53 发生突变就可能致癌,所以对它的深入研究尤为重要。而作为 p53 家族成员,p63 和 p73 同样有转录因子活性的作用且能够诱导细胞凋亡,然而是否具有抑制肿瘤功能还未确定,所以对于 p63 和 p73 更广泛与深入的研究可以帮助理解其在发育及肿瘤中的作用,有利于加深对整个 p53 家族的了解,为肿瘤的基因治疗和抗肿瘤药物的研发提供技术支持。

参考文献:

- [1] 熊春红,曹芳,涂北平,等. 江西茶鲜叶中重金属状况多变量分析[J]. 食品与生物技术学报,2012,31(1):55-60.
XIONG Chun-hong, CHAO Fang, TU Bei-ping, et al. Multivariate analysis of heavy metals in fresh tea leaves from Jiangxi province[J]. **Journal of Food Science and Biotechnology**, 2012, 31(1):55-60. (in Chinese)
- [2] Liang G S, Chou T Y, Han T C. Cluster analysis based on fuzzy equivalence relation [J]. **European Journal of Operational Research**, 2005, 166(1):160-171.
- [3] Hung C C, Kulkarni S, Kuo B C. A new weighted fuzzy C-Means clustering algorithm for remotely sensed image classification[J]. **IEEE Journal of Selected Topics in Signal Processing**, 2011, 5(3):543-553.
- [4] Kang J Y, Min L Q, Luan Q X, et al. Novel modified fuzzy c-means algorithm with applications [J]. **Digital Signal Processing**, 2009, 19(2):309-319.
- [5] He Q, Li H X, Shi Z Z, et al. Fuzzy clustering method based on perturbation[J]. **Computers & Mathematics with Applications**, 2003, 46(5-6):929-946.
- [6] Yang A, Kaghad M, Wang Y, et al. p63, a p53 homolog at 3q27-29, encodes multiple products with transactivating, death-inducing, and dominant-negative activities[J]. **Molecular Cell**, 1998, 2(3):305-316.
- [7] Jost C A, Marin M C, Kaelin W G J. p73 is a simian [correction of human] p53-related protein that can induce apoptosis[J]. **Nature**, 1997, 389(6647):191-194.
- [8] 罗承忠. 模糊集引论(上册)[M]. 北京:北京师范大学出版社,1989.
- [9] 唐旭清,方雪松,朱平. 基于模糊邻近关系的结构聚类[J]. 系统工程理论与实践,2010,30(11):1986-1996.
TANG Xu-qing, FANG Xue-song, ZHU Ping. Structural clusters based on fuzzy proximity relations [J]. **Systems Engineering - Theory & Practice**, 2010, 30(11):1986-1996. (in Chinese)
- [10] 王伟,仇建烨,朱平. 基于进化树理论的甲型流感病毒血凝素同源性及其进化分析[J]. 工程数学学报,2012,29(4):507-514.
WANG Wei, QIU Jian-ye, ZHU Ping. Influenza virus A hemagglutinins homology and evolution analysis based on the evolutionary tree theory[J]. **Chinese Journal of Engineering Mathematics**, 2012, 29(4):507-514. (in Chinese)
- [11] Lohrum M A, Vousden K H. Regulation and function of the p53-related proteins; same family, different rules [J]. **Trends in Cell Biology**, 2000, 10(5):197-202.
- [12] Allocati N, Dillio C, De L V. P63/p73 in the control of cell cycle and cell death [J]. **Experimental Cell Research**, 2012, 318(11):1285-1290.
- [13] Lopez L, Oliveira L P, Tucci P, et al. Different mutation profiles associated to P53 accumulation in colorectal cancer [J]. **Gene**, 2012, 499(1):81-87.
- [14] Dookeran K A, Dignam J J, Ferrer K, et al. P53 as a marker of prognosis in African-American women with breast cancer[J].

Annals of Surgical Oncology, 2010, 17(5): 1398–1405.

- [15] Ji F, Jin X, Jiao C H, et al. FAT10 level in human gastric cancer and its relation with mutant p53 level, lymph node metastasis and TNM staging[J]. *World Journal of Gastroenterology*, 2009, 15(18): 2228–2233.
- [16] Feldser D M, Kostova K K, Winslow M M, et al. Stage-specific sensitivity to p53 restoration during lung cancer progression[J]. *Nature*, 2010, 468(7323): 572–575.

科技信息

英国食品标准局拟修订果汁及果浆条例

据英国食品标准局消息, 6月7日英国食品标准局发布通报称, 拟修订果汁及果浆条例(北爱尔兰), 目前正就此征求意见, 修订后的果汁和果浆条例计划在2013年实施。

[信息来源] 食品伙伴网. 英国食品标准局拟修订果汁及果浆条例 [EB/OL]. (2013-6-14). <http://www.foodmate.net/news/daodu/2013/06/234776.html>.

<http://www.foodmate.net/news/daodu/2013/05/232706.html>

欧盟拟禁止 2,5-二甲基-3-乙酰基噻吩用于食品

5月22日英国食品标准局发布通报称, 欧盟成员国支持禁用 2,5-二甲基-3-乙酰基噻吩的决定, 该决定有望获得欧盟委员会批准。

欧盟食品安全局经评定认为, 2,5-二甲基-3-乙酰基噻吩(3-acetyl-2,5-dimethylthiophene)具有基因毒性, 会破坏细胞的遗传物质 DNA, 关系到人类健康安全。基因毒性物质不应该添加到食品链。

2,5-二甲基-3-乙酰基噻吩用于给食物增加一种烧焦的果仁风味, 它仅由少量的制造商生产, 整体使用很低(在欧盟每年总使用报告为 2.3 千克)。

[信息来源] Food Standards Agency. EU votes to ban 3-acetyl-2,5-dimethylthiophene [EB/OL]. (2013-5-22). <http://www.food.gov.uk/news-updates/news/2013/may/euban>.

研究发现塑化剂 DEHP 主要通过食品进入人体

据德国联邦风险研究所(bfr)消息, 2013年5月7日德国联邦风险评估研究所公布了一项有关 DEHP 的研究, DEHP 主要通过污染食品进而影响人体健康。

DEHP 是最常用的塑化剂, 主要用于增强塑料的柔韧性, 目前 DEHP 在环境及人类尿液中均有发现。与其他邻苯二甲酸酯一样, DEHP 具有生殖毒性, 有充分的证据证明 DEHP 具有激素类效应, 可以对人类生育能力或者对胎儿发育造成影响, 食品是 DEHP 的主要的暴露源。

研究人员研究了邻苯二甲酸二异辛酯(DEHP)在德国人群中污染情况, 暴露来源、暴露途径, 以及 DEHP、邻苯二甲酸二异壬酯(DINP)的毒物动力学研究结果。

研究涵盖了 20 年来德国儿童、青少年及成年人消费行为数据和食品、玩具及含合成材料的消费品(鞋子、化妆品、纺织品、室内灰尘、车内空气)等 37 类产品中 DEHP 暴露数据。

研究发现, 大多数情况下人群 DEHP 摄入总量较低, 但是特定人群具有 DEHP 大量暴露的潜在可能。根据最新的食品消费数据及食品、消费品中 DEHP 含量数据, DEHP 最高暴露量评估结果与人体中 DEHP 代谢产物含量测定结果一致。

研究还发现, 与未加工、未包装的食品相比, 含脂肪调味品(如蛋黄酱)、含油方便食品(如罐装的鱼类、蔬菜, 含油的罐头)中 DEHP 含量较高。食品中塑化剂污染主要是由加工过程及包装材料污染造成。

[信息来源] 食品伙伴网. 研究发现塑化剂 DEHP 主要通过食品进入人体 [EB/OL]. (2013-6-17). <http://www.foodmate.net/news/keji/2013/06/234906.html>.