人类基因组突变热点区的简并度特异基因*

刘 强[1)] 徐 军[2)] 陈润生[1)]**

（[1)]中国科学院生物物理研究所，北京 100101；

[2)] *Department of Physiological Science and Laboratory of Neuroendocrinology of The Brain Research Institute, University of California, Los Angeles, CA* 90095, USA）

摘要 突变热点区域是基因突变相对集中的区域，在生物的遗传和变异中有特殊的地位．针对特殊条件下发生突变形成的突变热点区域进行了相关研究．而人类基因组序列的测定和人类基因框架图的绘制，为在全基因组范围内进行突变热点研究提供了条件．分析了人类基因组中 2 831 个基因突变热点区域上简并度的性质，对突变热点区集中在高简并度区或者低简并度区的基因生物学功能进行了分析和分类．研究的焦点集中在某类功能的基因简并度特性一致的情况上．对搜集到的基因简并度特性利用聚类计算进行分析，找到了一些特殊的功能类，属于其中某类功能的基因能够通过聚类分析聚合到一起，从而说明简并度特性也是相近的，这为从基因的简并度特性预测表达物的功能提供了线索．

关键词 突变热点，简并度，聚类分析，人类基因组，单核苷酸多态性（SNP）

学科分类号 Q6

人类全基因组序列测定的完成，以及人类基因序列单核苷酸多态性（SNP）数据库的建立和不断完善，为研究人类基因序列的局域简并度和突变热点关系提供了良好的基础和平台．与以往人类基因序列信息较少时不同，现在我们可以针对基因组进行基因序列的局域简并度和突变热点研究．

突变热点区是在核酸序列或者蛋白质序列上突变集中和频繁发生的区域．这些区域一直是研究的热点，但是已经进行的研究主要集中在对特定条件下发生突变集中区域的研究，即观察和测定在该条件下某特定的核苷酸序列上，哪些位点容易发生突变以及突变的频率和方向[1,2]．由于这些研究都带有特定的目的，并且在染色体上涉及的区域有限，没有在全基因组范围内展开．产生这种情况的原因，一是实验条件的限制，二是过去人类基因组序列信息不完整，研究工作无法在人类全基因组范围内进行．

蛋白质简并度是由基因编码为蛋白质过程中产生的现象．由于每个三联子密码均由 A、C、T、G 4 种碱基组成，而自然界存在组成蛋白质的氨基酸只有 20 种，因此必然存在多种密码子组合对应一个氨基酸残基的情况．蛋白质简并度同其生物学功能[3]、生物进化[4]、转录和翻译[5,6]都有着密切的关系．定义对应一个氨基酸残基密码子的个数为其简并度．20 种氨基酸残基的简并度值列在表 1．由表 1 中可以看出 20 种氨基酸残基的简并度之和为 61，即 64 种密码子排列的可能情况减去 3 种终止子．蛋白质密码子简并度见表 1 所示．

Table 1 Table of degeneracy of protein

A	B	C	A	B	C
Arg	R	6	Gln	Q	2
Leu	L	6	His	H	2
Ser	S	6	Glu	E	2
Thr	T	4	Asp	D	2
Pro	P	4	Tyr	Y	2
Ala	A	4	Cys	C	2
Gly	G	4	Phe	F	2
Val	V	4	Ile	I	3
Lys	K	2	Met	M	1
Asn	N	2	Trp	W	1

A：Amino acid；B：Abbreviate of amino acid；C：Degeneracy.

SNP，又称单核苷酸多态性，现代生物学上认为是由可遗传的突变组成．众多的 SNP 同生物学功能密切相关，目前已进行了大量的研究[7~9]．在这里，我们认为 SNP 位点是突变发生的位置，而 SNP 密集的区域是突变热点区域．由于 SNP 数据

* 中国科学院知识创新工程重大资助项目（KSCX2-2-07 和 KJCXI-08），国家高技术“863”计划资助项目（2002AA231031）和国家重点基础研究发展规划项目（973）（2002CB713805）．

** 通讯联系人．

Tel：010-64888546，Fax：010-64877837

E-mail：crs@sun5.ibp.ac.cn

收稿日期：2004-07-01，接受日期：2004-08-03

库的建立，这里对突变位点的确定可以不受实验条件的限制，利用现有数据即可进行. 人类基因中的SNP位点数据库已经具有相当规模，因此这里利用SNP位点作为标定全基因组范围内突变热点区的根据.

1 材料和方法

1.1 数据的搜集

所有人类基因组数据和SNP位点数据均来自美国国家生物信息中心数据库（NCBI）. 它是现有的关于人类基因组数据库比较完整的数据库. 在人类基因组共约30亿个碱基对中，NCBI标注了基因2万余个. 并非所有的基因都能够进行简并度分析，根据下列要求挑选出合适的基因：a. 给出对应的mRNA序列；b. 分类标注为reviewed；c. 编码区带有SNP位点.

表2中表明了从NCBI中搜集的基因个数及其占全部人类基因的比例.

Table 2 Number of genes selected from NCBI and the percent of the genes selected in the genes given in NCBI

A	B	C	D	E
1	244	2 430	357	14.7
2	241	1 785	273	15.3
3	199	1 371	241	17.6
4	191	1 064	241	17.5
5	181	1 196	243	20.3
6	170	1 343	325	24.2
7	157	1 282	223	17.4
8	146	928	160	17.2
9	132	1 076	194	18.0
10	134	989	149	15.1
11	136	1 645	304	18.5
12	133	1 272	259	20.4
13	111	496	79	15.9
14	101	937	123	13.1
15	96	854	133	15.6
16	91	1 056	182	17.2
17	84	1 402	360	25.7
18	78	408	67	16.4
19	59	1 577	261	16.6
20	62	736	110	14.9
21	44	300	45	15.0
22	47	697	113	16.2
X	151	1 110	185	16.7
Y	50	161	17	10.6
Total	3 038	26 115	4 589	17.6

A: No. of chromosome; B: Number of pair bases of the chromosome; C: Number of genes given in NCBI; D: Number of genes selected; E: Percent of the genes selected in the genes given in NCBI.

需要说明的是，分类标注为reviewed的基因，其序列的可靠程度要高于标注为predicted或者provisional的基因. 挑选出来的基因共2 831个，占总标注基因数的17.6%. 这是基于对数据可靠性和能否供研究使用所做的必要剔除. 被剔除的基因，要么是通过预测得到，要么没有提供相应的碱基序列，要么是编码区不包含SNP位点，即在本研究中被认为没有突变热点区. 这些缺陷都导致它们无法被纳入研究的范围.

1.2 局域简并度的计算

局域简并度概念是由徐军博士[10]提出来的. 这个概念产生的基础是，一个氨基酸残基位点的局域简并度不仅和它本身位点的简并度有关，还同其附近的氨基酸残基的简并度有关. 因此局域简并度是以某个氨基酸残基位点为中心的一定范围内氨基酸残基简并度的平均值. 这里我们援用这一概念.

要计算局域简并度，首先要确定读框长度L，即在氨基酸残基位点附近多大范围内的其他残基在计算范围之内. 确定L的原则是在给定的L长度下，氨基酸序列的局域简并度离散程度最大. 这里是根据在给定L下，基因氨基酸序列的平均简并度上下2倍标准差范围外的氨基酸位点数量来度量这一离散程度的.

对于某一蛋白质，根据其氨基酸残基序列，按照表1给出的简并度数值，从而组成简并度数列. 然后，根据简并度数列和试验的读框长度L计算局域简并度值，组成局域简并度数列

$$ld_1, ld_2, \cdots\cdots, ld_n$$

n为氨基酸残基数，然后计算局域简并度数列的平均值和标准差.

$$average = \frac{\sum_{i=0}^{n} ld_i}{n}$$

$$std\_ld = \sqrt{\frac{\sum_{i=1}^{n} ld_i - average}{n}}$$

计算简并度数列各元素中落在［$average - 2 \times std\_ld$，$average + 2 \times std\_ld$］区域范围以外的元素个数，并以这一数值作为判定局域简并度序列离散程度大小的依据. 当某一长度的读框对应的离散元素个数最多时，就确定此长度为此基因密码子简并度的读框长度，并且以在读框长度在该基因氨基酸序列上移动计算得到的局域简并度数列，确定为该基因的局域简并度数列. 我们定义局域简并度大

于 *average* + 2 × *std*_ *ld* 的为高局域简并度，小于 *average* + 2 × *std*_ *ld* 的为低局域简并度.

1.3 突变热点区域的确定

对于搜集到的基因序列，根据 NCBI 数据库的信息，可以确定在编码区的 SNP 位点个数，根据它们在 DNA 序列上的位置确定其所在的氨基酸残基. 在该基因的局域简并度曲线图上将对应的位置加以标定，可以看到突变点的分布情况和所在位置的局域简并度值.

1.4 突变热点区域简并度性质的研究

为了定量地研究搜集到的所有基因样本突变热点区域的局域简并度性质，计算相应的参数(表3).

Table 3 Parameter of the local degeneracy of the mutation hotspots of genes selected

P	E
1 A	$average\_\ loc\_\ ld = \frac{\sum_{i=0}^{m} ld_{SNP_i}}{m}$
2 B	$std\_\ ld_{SNP_i} = \frac{ld_{SNP_i} - average\_\ loc\_\ ld}{m}$
3 C	$average\_\ std\_\ loc\_\ ld = \frac{\sum_{i=1}^{m} std\_\ ld_{SNP_i}}{m}$
4 D	$disperse\_\ loc\_\ ld = \sum_{i=1}^{m} \frac{ld_{SNP_i} - average\_\ loc\_\ ld}{m}$
5 E	$std\_\ ld = \sqrt{\frac{\sum_{i=1}^{m} ld_{SNP_i} - average\_\ loc\_\ ld}{m}}$

A: The average local degeneracy of all amino acid residues; B: The standard local degeneracy of the amino acid residues contain SNP; C: The average of standard local degeneracy of the amino acid residues that contain SNP (s); D: The disperse of local degeneracy of the amino acid residues which contains SNP (s); E: The standard error of the local degeneracy of the amino acid residues which contain SNPs; *m*: SNP; P: Parameter; E: Expression.

这些参数从不同侧面描述了各个基因突变热点区的局域简并度性质. 我们关注的是突变热点存在于高简并度区或者低简并度区的情况. 通过计算，将属于这2种情况之一或者2种情况兼而有之的基因挑选出来，并称之为简并度特异基因.

1.5 聚类计算

利用统计分析软件 SPSS 的 11.0 版本，对表3中计算得到的参数，首先进行归一化处理，消除参数之间量纲带来的差异，然后利用统计方法进行聚类分析. 这里采用的聚类方法是 Hierarchical 即等级树方法. 在计算得到的等级树上可以看到不同功能类基因的分布情况.

1.6 简并度特异基因的分类

对于根据简并度性质挑选出来的这些简并度特异基因，根据 NCBI 数据库对其功能所作的注释分为若干大功能类及小功能类，从而对某一功能类基因的简并度性质作出相应的判断，有条件的可以根据统计学的规律，对人类基因组中相类似功能的基因局域简并度的性质进行预测.

2 结果和讨论

2.1 局域简并度计算结果

将局域简并度计算应用到所有搜集到的2 831个基因上，以多巴胺受体 D1 (DRD1)、组氨酰 tRNA 合成酶 (HARS)、第二类主要组织相溶性复合物 Dr 第三 beta 链 (CDH17) 3个基因为例，他们的简并度参数见表4.

Table 4 Parameter of local degeneracy of DRD1, HAR7 and CDH17

A	B	C	D	E
HARS	510	18	3.544512	0.329047
DRD1	441	52	3.630488	0.269897
CDH17	833	18	3.457516	0.346163

A: Name of gene; B: Number of amino acid; C: Length of reading frame; D: Average of local degeneracy; E: Standard error of local degeneracy.

根据计算局域简并度得到的数值，可以绘出同氨基酸残基相对应的局域简并度曲线图. 图1、图2和图3分别是 DRD1、HARS、CDH7 的局域简并度曲线图.

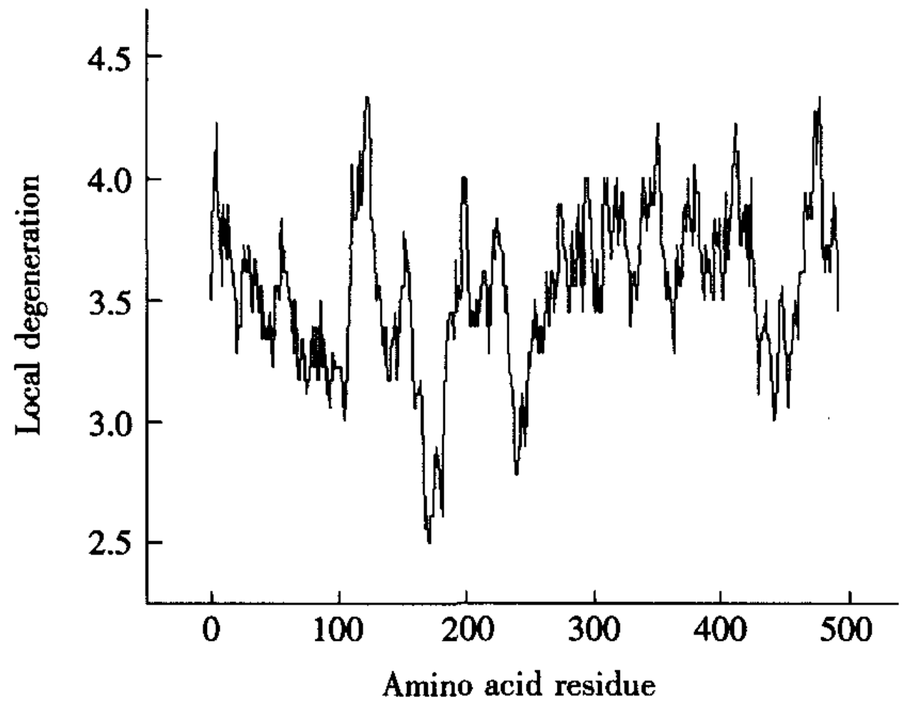

Fig. 1 Figure of local degeneracy of DRD1

Horizon axis represent the amino acid residues, and the longitude axis represent the local degeneracy.

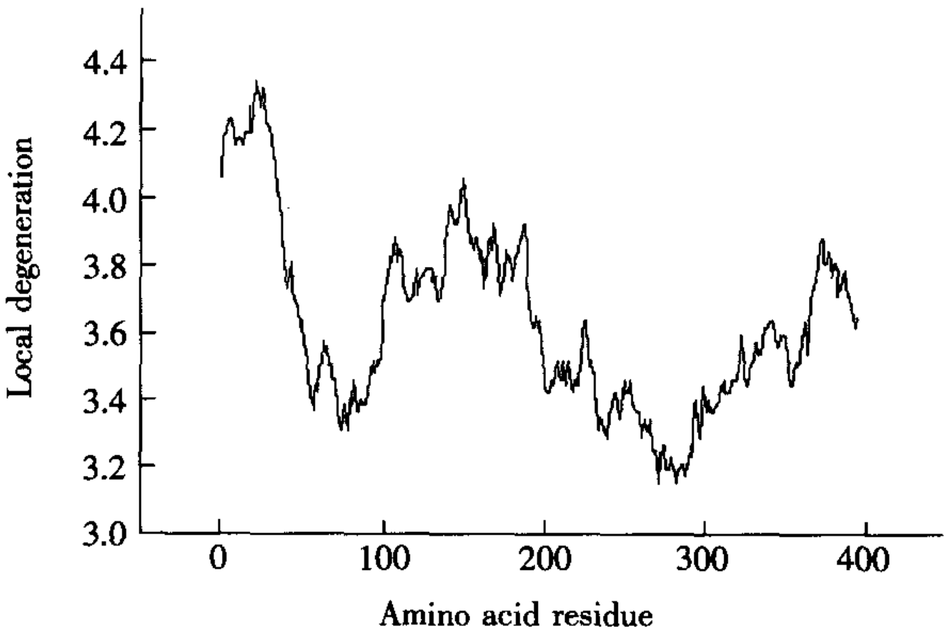

Fig. 2 Figure of local degeneracy of HARS

Horizon axis represent the amino acid residues, and the longitude axis represent the local degeneracy.

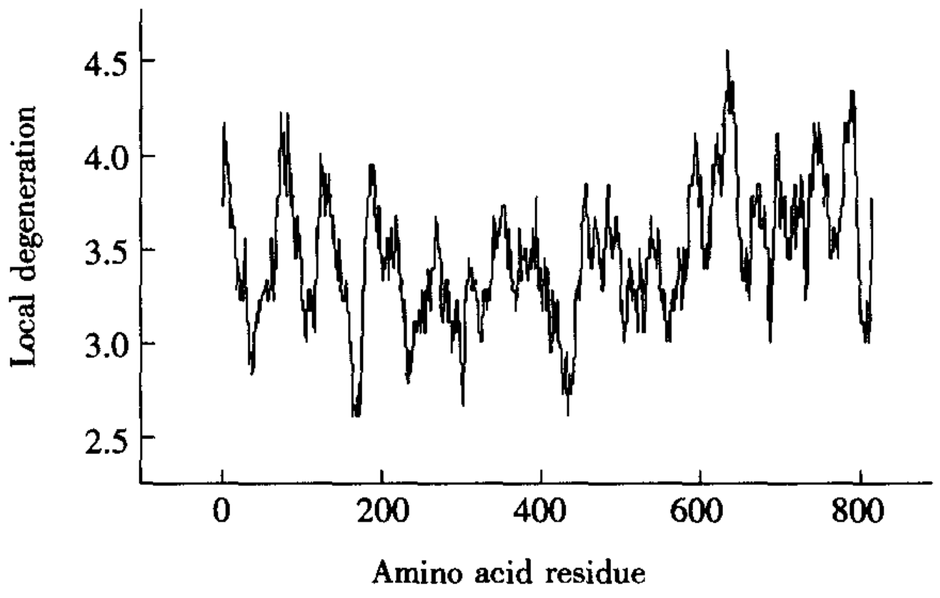

Fig. 3 Figure of local degeneracy of CDH17

Horizon axis represent the amino acid residues, and the longitude axis represent the local degeneracy.

通过计算得到的各个基因平均局域简并度的平均值为 3.56302，这一数值非常接近人类基因组蛋白质密码子的平均简并度 3.54，标准差的平均值是 0.46867.

在图 1、图 2 及图 3 上标定了 DRD1、HARS、CDH7 分别所含 SNP 位点所在的氨基酸残基位置，作出标定 SNP 位点的局域简并度曲线图，如图 4、图 5 和图 6 所示.

根据单个基因局域简并度的性质，可以将其归为 3 种类型：a. 有 SNP 位点落在高简并度区，但是没有 SNP 位点落在低简并度区；b. 有 SNP 位点落在低简并度区，但是没有 SNP 位点落在高简并度区；c. 有 SNP 位点落在高简并度区，并且有 SNP 位点落在低简并度区.

在全部搜集的基因当中，属于上述 3 种情况的有 539 个，它们都属于这里定义的简并度特异基因.

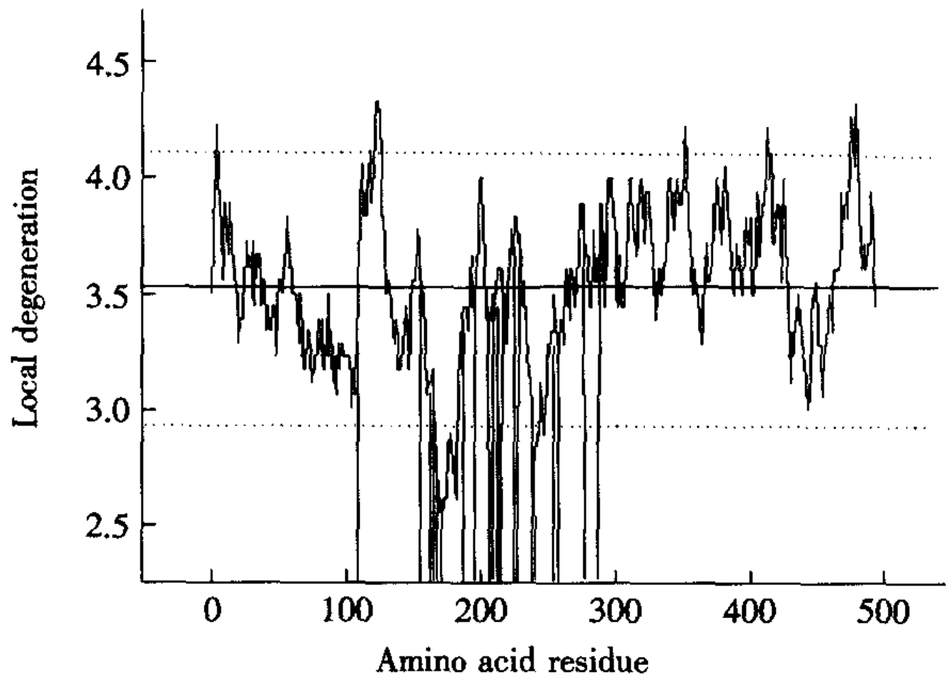

Fig. 4 Figure of local degeneracy of DRD1 marked with SNP spots

Horizon axis represents the amino acid residues, and the longitude axis represents the local degeneracy. Vertical lines represent SNP spots. Part of SNP of DRD1 located in amino acid residue with low local degeneracy and no in high local degeneracy.

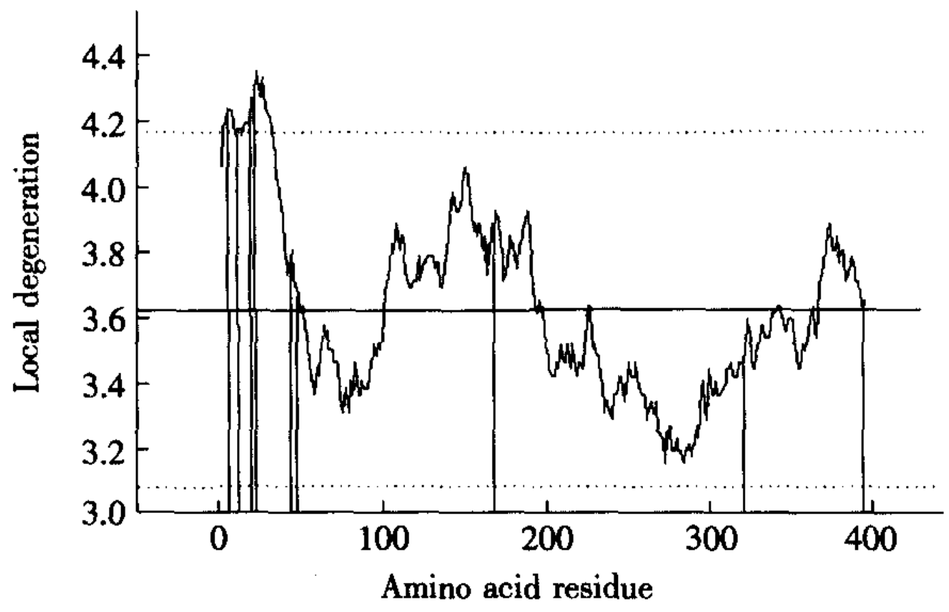

Fig. 5 Figure of local degeneracy of HARS marked with SNP spots

Horizon axis represents the amino acid residues, and the longitude axis represents the local degeneracy. Vertical lines represent SNP spots. Part of SNP of HARS located in amino acid residue with high local degeneracy and no in low local degeneracy.

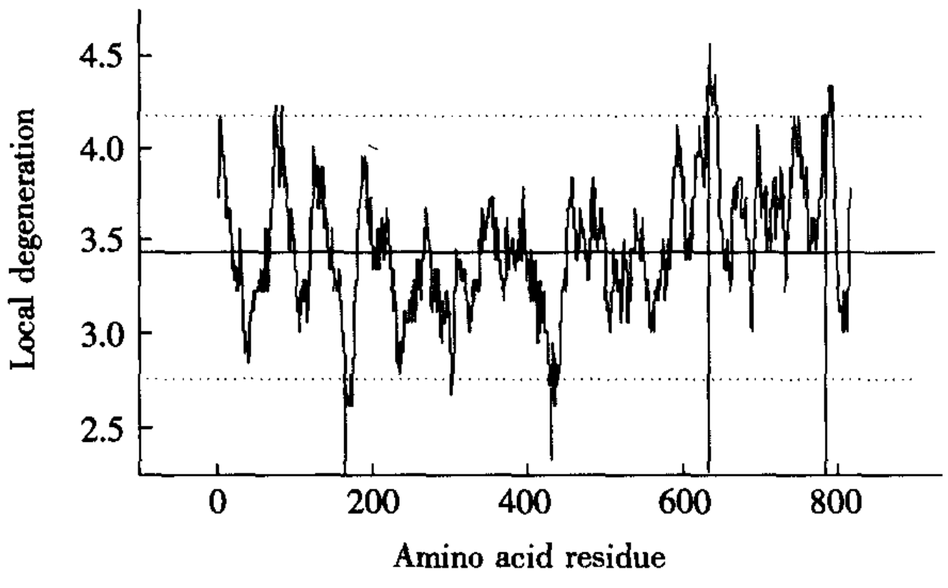

Fig. 6 Figure of local degeneracy of CDH17 marked with SNP spots

Horizon axis represents the amino acid residues, and the longitude axis represents the local degeneracy. Vertical lines represent SNP spots. Part of SNP of CDH17 located in amino acid residues with both high local degeneracy and also those in low local degeneracy.

2.2 聚类计算结果

首先进行总体聚类分析，即以539个简并度特异基因为聚类样本，以SNP位点所在氨基酸残基，如表3中所列的局域简并度参数为聚类参数，绘出相应的等级树．在该等级树中，若根据聚类关系分为4个主干支，基因能够全部聚合在一个主干支中的功能类有18类，如表5所示．包括表达为角蛋白、钙通道蛋白、醛酮还原酶等的基因．实际上，其中很多功能类被聚合在更小的范围内．

Table 5 Function classes and the genes of them that are classified to a common branch in cluster analysis

	A	B	C
1	A kinase (PRKA) anchor protein	AKAP13, AKAP2	2
2	Desmoglein	DSG1, DSG2	2
3	Keratin	KRT14, KRT17, KRT6A, KRT6B	4
4	Laminin	LAMB1, LAMB3, LAMC1	3
5	Williams-Beuren Syndrome	WBSCR20B, WBSCR20C	2
6	Major histocompatibility complex class Ⅰ	HLA-A, HLA-B	2
7	Excision repair cross-complementing rodent repair deficiency, complementation group 2	ERCC5, ERCC6	2
8	Bone morphogenetic protein receptor	BMPR1A, BMPR2	2
9	Aldo-keto reductase family 1	AKR1C1, AKR1C2, AKR1C3, AKR1C4	4
10	Apolipoprotein	APOA4, APOE	2
11	*Calcium channel, voltage-dependen*	CACNA1A, CACNG1	2
12	Chorionic gonadotropin	CGB, CGB5, CGB7, CGB8	4
13	Polymerase (RNA) Ⅱ (DNA directed) polypeptide	POLR2A, POLR2J	
14	Cadherin	CDH1, CDH17, CDH2, CDH3, CDH4	5
15	Fibulin	FBLN1, FBLN2	2
16	Tripartite motif-containing	TRIM10, TRIM36, TRIM38	3
17	Rhesus blood group	RHCE, RHD	2
18	RAD21 homolog	RAD21, RAD9	2

A: Function class; B: Genes of the function; C: Number of genes of the function class.

然后进行分染色体聚类分析，考虑到各个染色体之间遗传特性的差异，进化速率的不同，以及SNP位点分布的差异，我们将各个染色体包含的基因，根据局域简并度的性质做聚类分析，得到相应的等级树．在各个染色体中，同一功能类的基因经常能被聚合到一起的功能类，包括表达为核糖体蛋白、激酶、各种受体、转移酶以及细胞色素等基因．

根据聚类结果分析，不论是在总体聚类还是分染色体聚类当中，能够在等级树当中被聚合在一起，说明同一功能类的基因在简并度性质上具有相似性．

2.3 简并度特异基因分类

将539个简并度特异基因编码的蛋白质按照生物学功能分类．这里分为4大类，如表6中所列．应该说这里的分类并不很严格，只是为了方便分析简并度特异基因生物学功能所做的大致划分．

Table 6 Protein function classes

F	A	B	C	Total
Ⅰ	68	94	7	169
Ⅱ	13	26	5	44
Ⅲ	15	26	2	43
Ⅳ	71	190	22	283
Total	167	336	36	539

Ⅰ: Protein related to comprise and structure of cell; Ⅱ: Protein related to disease and immunity; Ⅲ: Receptor; Ⅳ: Protein related to regulation; F: Great function class; A: Number of genes in the great function class with SNP spots of a single gene located in low degeneracy region; B: Number of genes in the great function class with SNP spots of a single gene located in high degeneracy region; C: Number of genes in the great function class with SNP spots of a single gene located in both low and high degeneracy region.

将每一大类功能蛋白质，按照具体的功能细分为若干小类.

其中有 24 类，对于每一功能小类的基因突变热点只出现在高简并度区，比如表达为各种抗原、角蛋白、微管蛋白、钙蛋白酶等基因. 功能小类及基因列于表 7 中.

Table 7 Function classes and genes of them that contain mutation hotspots only in high local degeneracy region

	A	B	C	D
1	Ⅱ	Antigen	LMAN1, ANXA13, BPAG1, BST1, DSG3, MOX2	6
2	Ⅰ	Heterogeneous nuclear ribonucleoprotein	HNRPA3, HNRPH1	2
3	Ⅰ	Desmoglein	DSG1, DSG2	2
4	Ⅰ	Keratin	KRT14, KRT17, KRT6A, KRT6B	4
5	Ⅰ	Laminin	LAMB1, LAMB3, LAMC1	3
6	Ⅰ	Mitochondrial ribosomal protein	MRPL28, MRPL38, MRPL40	3
7	Ⅰ	Sorting nexin	SNAG1, SNX2	2
8	Ⅰ	Tubulin	TUBA1, TUBA3	2
9	Ⅱ	Major histocompatibility complex, class I	HLA-A, HLA-B	2
10	Ⅱ	Glutamate receptor	GRIA4, GRIN1, GRM4, GRM5	4
11	Ⅱ	Interleukin recepto	IL17R, IL2RA, IL4R	3
12	Ⅴ	Myosin,	MYH7, MYL1	2
13	Ⅴ	Tumor necrosis factor receptor Superfamily	TNFRSF10B, TNFRSF11B, TNFRSF1A	3
14	Ⅴ	Metalloproteinase	ADAM12, ADAM20, ADAM21, ADAMTS2, ADAMTS5, ADAMTS7	6
15	Ⅴ	Aldehyde dehydrogenase 3 family	ALDH3A2, ALDH3B2	2
16	Ⅴ	Calpain,	CAPN1, CAPN2	2
17	Ⅴ	Chemokine	CCL23, CCL25	2
18	Ⅴ	Cyclin-dependent kinase inhibitor	CDKN1A, CDKN3	2
19	Ⅴ	Chorionic gonadotropin, beta polypeptide	CGB, CGB5, CGB7, CGB8	4
20	Ⅴ	DEAD/H (Asp-Glu-Ala-Asp/His) box Polypeptide	DDX19, DDX4	2
21	Ⅴ	Glutathione S-transferase	GSTA1, GSTA2	2
22	Ⅴ	Matrix metalloproteinase	MMP23A, MMP23B	2
23	Ⅴ	Polymerase (RNA) Ⅱ (DNA directed) Polypeptide	POLR2A, POLR2J	2
24	Ⅴ	Tumor necrosis factor	TNF, TNFSF11	2

A: Great function class; B: Function class; C: Genes of the function class; D: Number of genes of the function class. The tabs of great function class of this table are corresponding to those in Table 6.

对于另外 5 个功能小类的基因，例如激酶锚蛋白、醛脱氢酶家族Ⅰ等，每一功能小类的基因突变热点只出现在低简并度区，这些功能小类及基因列于表 8.

Table 8 Function classes and genes of them that contain mutation hotspots only in low local degeneracy region

	A	B	C	D
1	I	A kinase anchor protein	AKAP13, AKAP2	2
2	I	Williams-Beuren syndrome	WBSCR20B, WBSCR20C	2
3	V	Aldo-keto reductase family 1	AKR1C1, AKR1C2, AKR1C4, AKR1C3,	4
4	V	UDP-Gal: betaGlcNAc beta	B3GALT5, B4GALT1	2

A: Great function class; B: Function class; C: Genes of the function class; D: Number of genes of the function class. The tabs of great function class of this table are corresponding to those in Table 6.

3 讨 论

突变热点区域的密码子简并度特性是描述生物遗传特性的参数. 基因序列的突变和密码子的简并都是伴随遗传产生的，这一点可以从突变产生的原理和各个物种遗传序列密码子简并度的比较看出.

由聚类结果可以看出，确实可根据突变热点区域的简并度参数将某些相近功能的基因聚在一起，说明突变热点区域的简并度参数确实从一个方面代表了基因的性质.

由简并度特异性基因的分类来看，确实存在若干类基因，它们的简并度特性具有一致性，即只有处于高简并度区的 SNP 位点而不存在处于低简并度区的 SNP 位点，属于相反情况. 从而说明，简并度的特性同生物学功能存在某种程度的联系.

从简并度特异基因功能分类表（表6）和高低简并度特异基因功能分类表（表7 和表8）可以看到，SNP 位点集中于高简并度区的功能类和基因数量明显多于集中于低简并度区的情况. 这应该也是进化的结果. 人类基因组编码的蛋白质密码子平均简并度在各个物种之中已经是相当高了，甚至在已经考察的物种中还没有这一数值超过人类的情况发生. 这里计算的高低局域简并度，从概念上说是各个基因中局域简并度比较的相对值，仍然出现突变热点集中在高简并度区的情况，说明人类进化的程度是相当高的.

可以看到，有的功能类基因，虽然在总体聚类中并不能聚在一起，但是在分染色体的聚类中却可以聚为一类. 究其原因可能有两个，一是与各个染色体的进化速率不同有关，而简并度又是进化的产物，并且同进化过程密切相关，所以可以从一定程度上解释这一现象；二是在总体聚类中干扰的基因数量大幅度增加，减小了同一功能类基因被聚类在一起的概率. 当然也有些功能类的基因在总体聚类中可以聚在一起，但是在分染色体的聚类分析中却能聚合到一起，原因可能是本来此类功能的基因样本数较少，分到各个染色体后只剩一个，找不到同类基因，也就无从进行功能分析.

同时，在所有挑选的基因中，具有简并度特异性的基因是少数，大约只占 20%. 如果放在全部人类基因组中看，具有简并度特异性的基因只占 2%. 由于这些比例比较低，也就是考察的样本在总的样本中所占比例较少，使得由计算得到的关于某一类基因简并度特性的结论，很难根据统计学的规律推广到这一类功能的所有基因上. 造成这一现象是由于确实具有简并度特异性的基因不多，同时由于现有的 SNP 数据库并不完整，使本来可以纳入简并度研究的 SNP 位点没有找到而无法进行. 这一点可以从每天都有新的 SNP 数据被发现，并被提交到 NCBI 等 SNP 数据库中以供研究使用的情况看出. 所以随着 SNP 数据库的不断完善，关于 SNP 位点简并度的研究还可以继续下去.

此外，也可以将这些由人类基因组得到的关于简并度特性的方法应用在其他物种，如大鼠等模式生物上进行类似计算，以验证其在该种生物基因组上的正确性，并且比较与人类基因组结论的异同.

致谢 感谢中国科学院生物物理研究所凌伦奖副研究员、邓巍助理研究员、陈燕俊博士研究生、何杰博士研究生、张治华博士研究生的大力支持.

参 考 文 献

1 Kidwell M G, Holyoake A J. Transposon-induced hotspots for genomics instability. Genome Research, 2001, **11**: 1321 ~ 1322

2 Rogozin I B, Pavlov Y I, Bebenek K, *et al*. Somatic mutation hotspots correlate with DNA polymerase eta error spectrum. Nature Immunology, 2001, **2** (6): 530 ~ 536

3 Kilpatrick D R, Nottay B, Yang C F, *et al*. Serotype-specific identification of polioviruses by PCR using primers containing mixed-base or deoxyinosine residues at positions of codon degeneracy. J Clint Microbiology, 1998, **36** (2): 352 ~ 357

4 Otsuka J, Kawai Y, Sugaya N. The influence of selection on the evolutionary distance estimated from the base changes observed

between homologous nucleotide sequences. J Theory Biology, 2001, **213** (2): 129 ~ 144

5 Deana A, Ehrlich R, Reiss C. Silent mutations in the *Escherichia coli* ompA leader peptide region strongly affect transcription and translation *in vivo*. Nucleic Acids Research, 1998, **26** (20): 4778 ~ 4782

6 Toha J, Soto M A. Neural network in the transcription- translation process in eukaryotic cells and viruses, a comparison. Virus Genes, 1995, **10** (3): 211 ~ 215

7 Lilleberg S L. In-depth mutation and SNP discovery using DHPLC gene scanning. Curr Opin Drug Discov Devel, 2003, **6** (2): 237 ~ 252

8 Fridman C, Ojopi E P, Gregorio S P, *et al*. Association of a new polymorphism in ALOX12 gene with bipolar disorder. 35: Europe Arch Psychiatry Clint Neuroscience, 2003: **253** (1): 40 ~ 43

9 Kim L H, Lee H S, Kim Y J, *et al*. Identification of novel SNPs in the interleukin 6 receptor gene (IL6R). Human Mutation, 2003, **21** (4): 450 ~ 451

10 Xu J, Chen R, Xiao Z X. Analysis of potential functional region using local degeneracy: mutational hotspots in human factor Ⅸ are localized in high-degeneracy regions. Analysis Biochemistry, 1994, **223** (1): 71 ~ 73

Analysis of Degeneracy Special Gene in Human Genome*

LIU Qiang[1], XU Jun[2], CHEN Run-Sheng[1]**

([1] *Bioinformatics Laboratory, Institute of Biophysics, The Chinese Academy of Sciences, Beijing* 100101, China;
[2] *Department of Physiological Science and Laboratory of Neuroendocrinology of The Brain Research Institute, University of California, Los Angeles, CA* 90095, USA)

Abstract This is a research of the relationship between the characteristics of local degeneracy on mutational hotspots regains on genes of human genome. There includes the introduction of related research background and a draft of the calculation methods of local degeneracy and clustering. SNP spots are regarded as the foundation to decide mutational hotspots. 2 831 genes of human genome are selected and analyzed by their characteristics of local degeneracy on hotspots regains. Furthermore, genes whose hotspots are aggregated in relatively high local degeneracy regains and their low counterpart are analyzed and classified by their biological function. The selected genes are clustered by their parameters of local degeneracy. Some function classes whose selected genes are clustered together through clustering are found. For certain function classes, all selected genes of each class have similar parameters of local degeneracy. It appears that the local degeneracy of amino acid residues that contain SNPs of certain gene depends on the function of the proteins these genes produce. The fundamental reason for this phenomenon may be that by their nature, SNPs connect local degeneracy in mutational hotspot regions and biological function. Depending on whether they are degenerate SNPs affect the functions of proteins, and each no degenerate SNP may be a potential source of phenotypic diversity. Human genes appear to select high local degeneracy at the amino acid residues that contain SNPs. This may be the result of natural selection. This result shows that their trait of local degeneracy also like each other, which is a good clue to predict gene function by their local degeneracy.

Key words mutational-hotspots, local-degeneracy, cluster analysis, human genome, SNP

* This work was supported by grants from The National Knowledge Innovation Program of The Chinese Academy of Sciences (KSCX2-2-07 and KJCX1-08), The State 863 High Technology R&D Project of China (2002AA231031) and The Special Funds for Major State Basic Research of China (2002CB713805).

** Corresponding author. Tel: 86-10-64888546, Fax: 86-10-64877837, E-mail: crs@sun5.ibp.ac.cn

Received: July 1, 2004 Accepted: August 3, 2004