基于内含子和外显子的系统发育分析的比较

王 宁 陈润生*
(中国科学院生物物理研究所, 北京 100101. * 联系人)

摘要 在灵长目动物中分离出7个断裂基因的内含子和外显子序列, 分别进行系统发育分析. 比较结果显示: (1) 基于内含子与基于外显子所得到的进化树的拓扑结构基本一致, 但由于内含子和外显子序列的突变速率不同, 导致进化树的分枝长度不同; (2) 在 *hominoids* 中, 外显子的替代速率下降, 而内含子的替代速率相对稳定; (3) 当选用 *cebus* 作为参考物种时, 基于不同外显子推测出的orangutan的分歧时间相差较大, 而基于不同内含子序列所得的分歧时间比较接近, 且与化石记录相吻合, 这说明内含子序列可作为一个更好的"分子钟"; (4) 对于不同基因, 序列的G+C含量与它的替代速率有一定关系, 当内含子的替代速率大于相应的外显子时, 内含子的G+C含量较低. 所有这些结果显示, 内含子序列可以用来在近亲缘关系的物种中进行进化分析.

关键词 **系统发育分析 内含子 替代速率 分子进化**

在研究分子进化时, 人们通常根据一些相对保守的序列进行分析, 比如rRNA序列、蛋白质的氨基酸序列、编码蛋白质的核酸序列. 这样做的原因一方面是由于这些保守的序列会带来相对可靠的结果; 另一方面是由于在数据库中, 这些序列可以比较方便地得到[1]. 但基于DNA非编码区序列的分子进化分析还少见报道.

通过在近亲缘关系的物种中比较同源基因的内含子序列, 我们发现, 这些序列的相似性仍然很高, 而在亲缘关系较远的物种中, 与相应的外显子相比, 由于承受的进化压力小, 内含子序列的相似比很低. 因此, 在亲缘关系较近的物种中, 内含子序列是否可以与外显子一样用作分子进化分析是本文感兴趣的问题.

自从"分子钟"(molecular clock)的概念被提出以来[2], 一直存在一些争议. 最近的一些文献报道, 在 *hominoids* 中序列的突变速率会下降, 突变速率与物种每一代的生存时间长短有关系[3]. 是否内含子序列的突变速率在 *homimoids* 中也会下降, 除此之外, 基于不同基因的外显子序列所推测出的物种分歧时间常常不同, 是否内含子可以作为一个更好的"分子钟", 这是本文的另一研究目的.

1 数据及方法

( ) 数据搜集. 在GenBank中, 我们检索出灵长目中的所有断裂基因(基因中包含内含子部分), 期望得到一些这样的基因在尽可能多的物种中的分布来做系统发育分析. 同时, 由于序列的长度是重构可靠的系统发育树的关键因素[4], 从中筛选出7个基因, 每个基因的内含子长度大于700 bp. 它们分别是: *Epsilon globin*, *Gamma-1 globin*, *Gamma-2 globin*, *Delta globin*, *Beta globin*, *Immunoglobulin*, 和 *Prepro-insulin*.

( ) 方法. 首先, 我们对每个基因分离出其内含子和外显子序列, 然后对于同一基因在不同物种中的内含子和外显子利用CLUSTAL W程序分别做多序列比较[5]. 在重构进化树时, 采用Neighbor-Joining方法[6], 利用Bootstrap分析方法进行统计可靠性检验[7]. 用Taijima-Nei

的方法[8]来计算 DNA 序列的替代速率.

2 结果

2.1 在不同物种中比较 *Epsilon globin* 外显子和内含子的替代速率

我们在 33 个物种中查到了 *Epsilon globin* 的序列, 表 1 中列出了其中 32 个物种对人的外显子和内含子序列的相似性得分, 以及序列长度的比较. 可以很明显看出, 外显子序列的长度在不同物种中都是一致的(444 bp), 而内含子序列的长度却有所不同(932~ 1 075 bp). 另外, 从序列的相似比可以看出, 不同物种的内含子序列的相似性得分仍然很高, 只比外显子序列的相似比低一些.

表 1 灵长目不同物种与人类相比 *Epsilon globin* 基因外显子和内含子的序列相似性得分及序列长度比较

物种名	外显子		内含子	
	相似比/%	长度/bp	相似比/%	长度/bp
alouatta seniculus	95.7	444	85.5	932
alouatta belzebul	95.5	444	85.5	932
aotus azarai	96.6	444	85.1	940
ateles belzebuth	95.5	444	86.3	940
eulemur macaco	90.8	444	76.8	963
brachyteles arachnoides	95.9	444	86.6	945
cacajao calvus	96.2	444	85.7	951
callicebus torquatus	96.4	444	85.1	949
callimico goeldii	96.2	444	85.5	941
cebus albifrons	95.0	444	84.9	934
pan troglodytes	100.0	444	98.3	982
chiropotes satanas	96.2	444	85.8	950
daubentonia madagascariensis	91.9	444	75.9	963
fat-tailed dwarf lemur	91.4	444	76.4	970
gray mouse lemur	91.2	444	74.1	956
galago crassicaudatus	89.4	444	70.3	934
ateles geoffroyi	95.7	444	86.2	942
gorilla gorilla	100.0	444	98.1	980
callithrix jacchus	96.4	444	84.8	940
lagothrix lagotricha	95.7	444	85.9	942
leontopithecus rosalia	96.6	444	85.6	940
macaca mulatta	98.2	444	95.2	979
callicebus moloch	95.7	444	86.0	947
orangutan	98.9	444	95.5	985
otolemur crassicaudatus	89.4	444	70.3	936
pan paniscus	98.0	444	97.7	987
pithecia irrorara	96.4	444	86.7	950
cebuella pygmaea	95.9	444	84.8	940
saguinus midae	95.9	444	84.7	938
saimiri sciureus	96.4	444	75.3	1 075
tarsius syrichta	89.9	444	70.6	943

我们分别计算了 *Epsilon globin* 在 32 个物种和人之间基于内含子、外显子以及外显子第 3 碱基的替代速率之间的关系, 并显示在图 1 中. 可以看出, 内含子的替代速率总是最高的, 外显子的替代速率最低, 而外显子的第 3 碱基的替代速率位于中间; 同时还可以看出, 外显子替

代速率高的物种, 内含子的替代速率也很高.

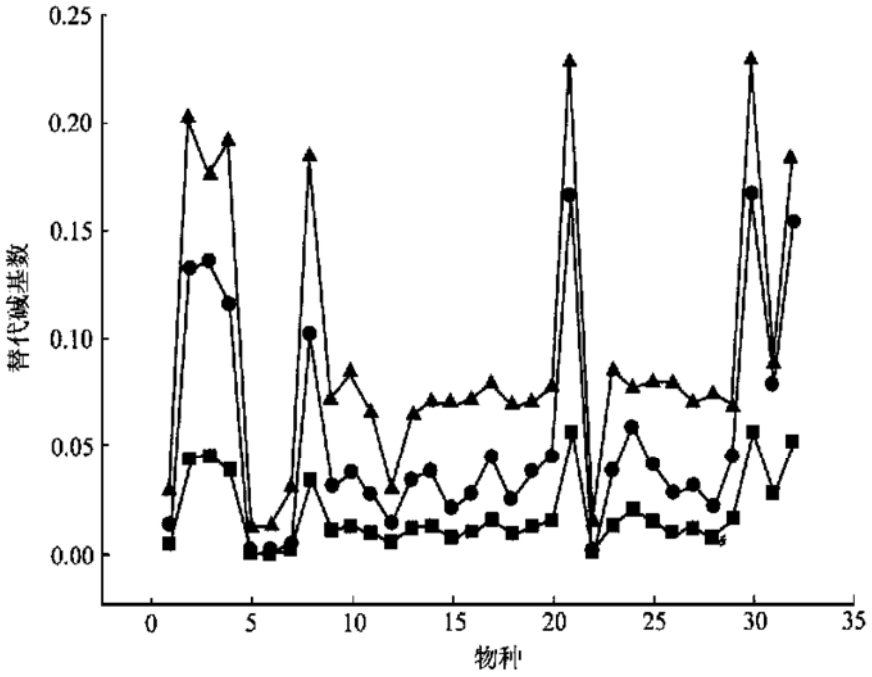

图 1 *Epsilon globin* 基因在灵长目 32 个物种中相对于人的替代数的比较

▲示内含子序列, ●示由外显子的第 3 碱基构成的序列, ■示外显子序列

2. 2 基于内含子和外显子序列重构进化树的拓扑结构和分枝长度比较

为了避免物种树与基因树的差别, 我们在 *Epsilon globin* 中挑选出 8 个进化关系已知的物种(根据化石的记录) , 基于它们的外显子和内含子序列重构进化树, 结果(图 2 和 3) 可以明显地看出, 基于这两部分序列重构出的进化树的拓扑结构完全一致. 基于其他基因的外显子和内含子序列, 我们做了同样的重构分析, 并把基于内含子序列重构出的进化树列于图 4. 在这里如果忽略人(human) 、黑猩猩(chimpanzee) 及大猩猩(gorilla) 三者之间的亲缘关系, 除了在 *Gamma-1 globin* 基因之外的其他 6 个基因中, 基于内含子与基于外显子的拓扑结构(未列出)

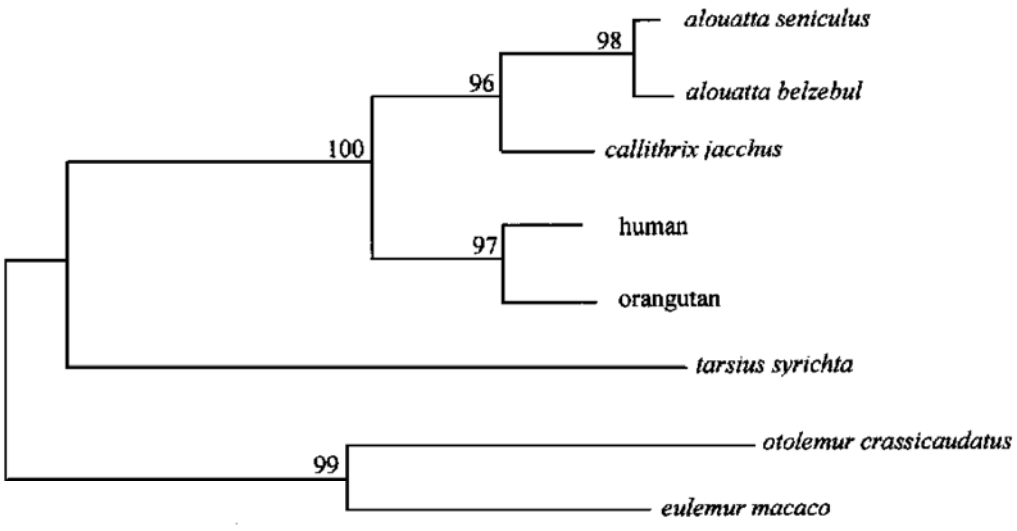

图 2 基于 *Epsilon globin* 外显子序列的系统发育分析

节点旁的数字代表 Bootstrap 分析 500 次后的结果, 分枝长度每——代表 0. 001 074

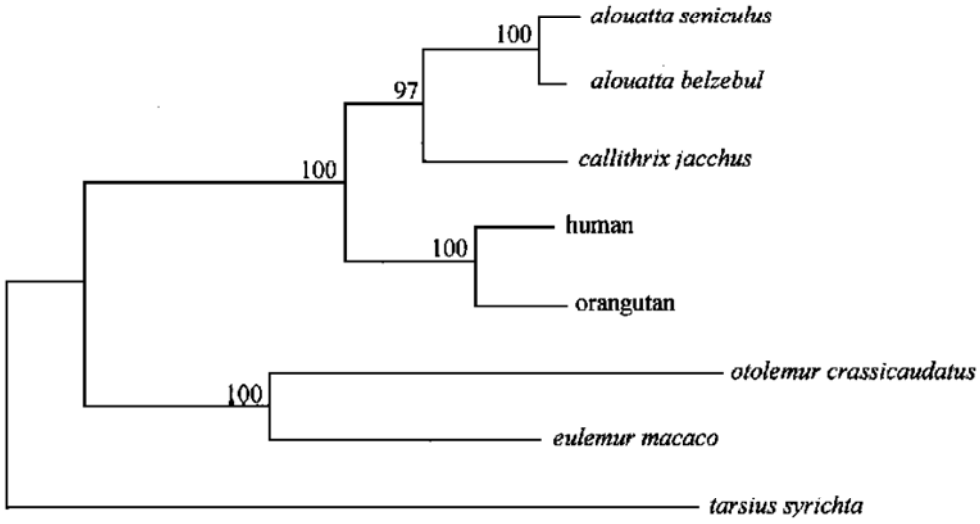

图 3 基于 *Epsilon globin* 内含子序列的系统发育分析

节点旁的数字代表 Bootstrap 分析 500 次后的可靠性结果, 分枝长度每——代表 0. 003 457

都是一致的. 通过这些结果可以明显看出, 内含子序列的确可以提供有用的进化信息; 也就是说, 在比较近亲缘物种之间的进化关系时, 从外显子中可以得到关于物种的进化信息, 在内含子中同样可以得到.

由于目前关于 human, chimpanzee 及 gorilla 三者之间的亲缘关系仍无定论, 我们的分析结果显示: 在 *Epsilon globin* 和 *gamma-1 globin* 中, 分别基于内含子和外显子序列进行重构可以得到一致的结论, 即在 *Epsilon globin* 中, human 与 chimpanzee 亲缘关系更近, 在 *gamma-1 globin* 中, human 与 gorilla 关系更近; 而对于其他基因, 基于外显子和内含子这两部分序列的结果不同. 值得一提的是, 在这几个基因中, 基于内含子序列的结果都显示出 gorilla 与 chimpanzee 的亲缘关系更近. 虽然在我们分析的大多数情况下, 基于内含子与基于外显子的拓扑结构是一致的, 但它们的分枝长度不同, 这主要是由于这两部分序列在进化过程中的替代速率不同.

2. 3 内含子和外显子的替代速率比较

将不同基因内含子和外显子的每个碱基的替代数列于表 2, 可以明显看出, 除了在 *Immunoglobulin* 基因中, human 和 old world monkey 之间内含子的替代速率低于外显子之外, 在其他 6 个基因中每对物种都显示出内含子的替代速率大于外显子, 尤其在 human 与 orangutan, human 与 old world monkey 这两对物种中十分显著.

Porter 等人[9]提出, *hominoids* 在从 old world monkey 中分离出来后, 出现替代速率下降的趋势; 而 Wilson 等人却认为这种下降是由于错误地估计了人、猿的分歧时间而导致的, 他们通过相对速率检验认为并不存在在 *hominoids* 中速率下降的现象. 但是最近的文献报道再一次为速率下降这个假设提供了证据[10, 11], 他们在比较两个物种间的替代速率的变化时, 选出另一物种作为参考, 提出了 *hominoids* 的突变速率低于 old world monkey, 人的突变速率低于猿.

在这里, 我们进一步比较了外显子和内含子的替代速率. 首先, 两个物种的分歧时间需要已知. 在通常情况下, 物种的分歧时间是给出一个范围, 所以在这里, 我们根据文献报道的范围及其平均分歧时间来计算物种的替代速率. 从表 3 中明显看出, 外显子的替代速率的确在 *homoinids* 中降低了(human 与 chimpanzee 之间的替代速率是 0. 72, human 与 orangutan 之间的替

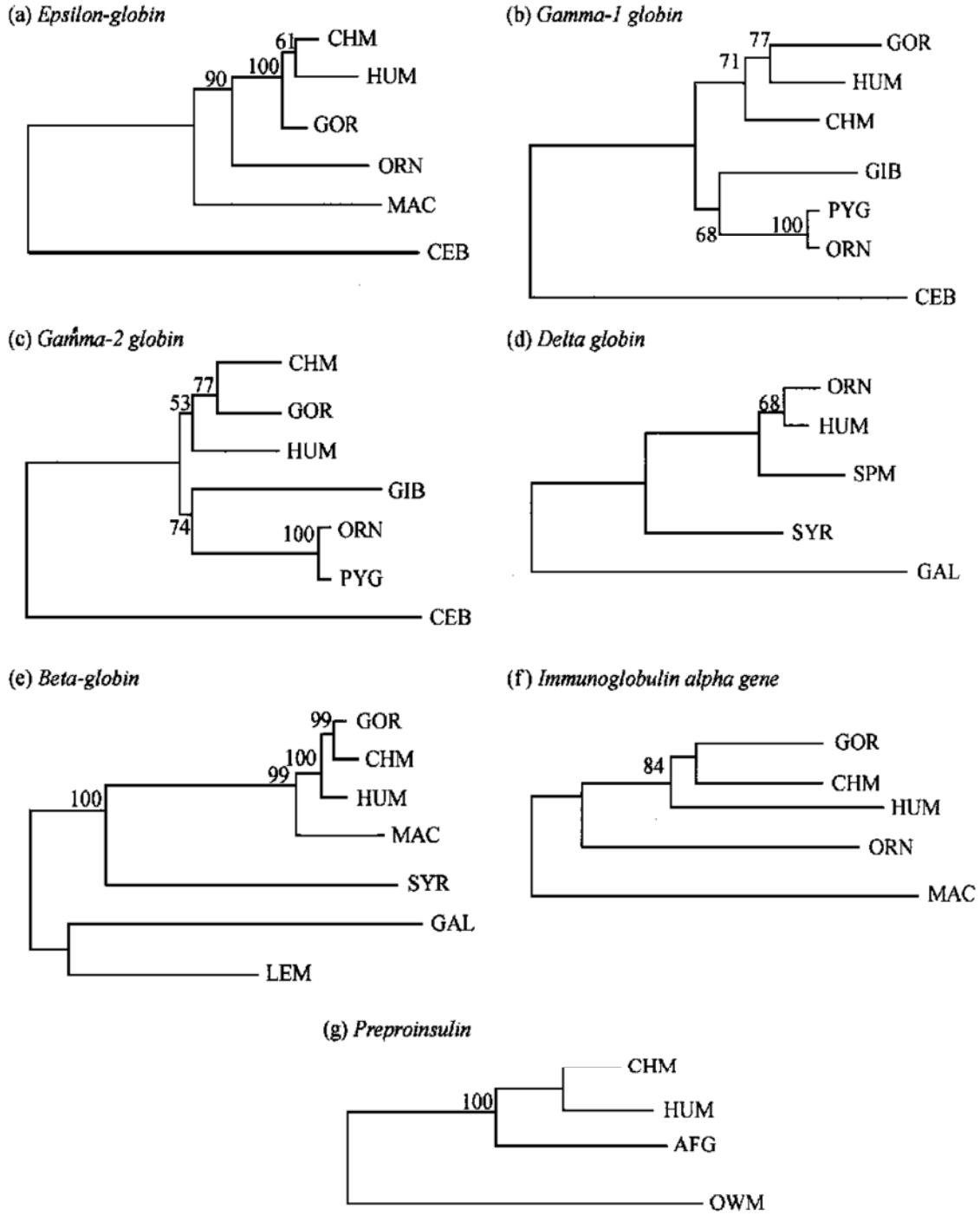

图 4 基于 7 个基因的内含子序列所进行的系统发育分析结果

Neighbor-Joining 方法用来重构进化树, 节点旁边的数字代表 Bootstrap 分析 500 次后的统计可靠性结果. 物种的缩写如下: HUM: *homo sapiens*, CHM: *pan troglodytes* (chimpanzee), GOR: *gorilla gorilla*, ORN: *pango pygmaeus* (orangutan), MAC: *macaca mulatta*, CEB: *ecbus albifrons*, GIB: *hylobates lar* (gibbon), PYG: *pongo pygmaeus*, SPM: *aletes geoffroyi* (spider monkey), SYR: *tarsius syrichta*, GAL: *galago crassicaudatus*, LEM: *eulemur macaco* (lemur), AFG: *cerocopithecus aethiops* (African green monkey), OWM: *aotus trivirgatus* (owl monkey)

代速率是 0.92, human 与 old world monkey 之间是 1.08). 而对于内含子来说, 在 *hominoids* 中平均替代速率不但没有下降, 反而上升了(human 与 chimpanzee 之间的替代速率是 1.74, human 与 orangutan 之间是 1.86, 而 human 与 old world monkey 之间的替代速率是 1.32).

表 2 不同基因中物种之间的每 100 个碱基的替代数

		ε[a]	γ-1	γ-2	δ	β	*Ig*	*Pre-in*
hum/ chm[c]	E[b]	0.00	0.68	0.45		0.00	3.09	1.83
	I	1.24	3.29	3.03		1.68	3.29	2.10
	R	–	4.84	6.73		–	1.06	1.15
hum/ orn	E	1.14	1.14	2.29	1.14		5.32	
	I	3.82	4.68	4.45	3.57		5.77	
	R	3.35	4.11	1.94	3.13		1.09	
hum/ owm	E	1.83				5.17	8.92	5.67
	I	4.92				6.63	6.68	8.20
	R	2.69				1.28	0.75	1.45
hum/ nwm	E	5.17	7.84	7.34	7.64			10.60
	I	11.90	12.46	12.99	9.40			18.83
	R	2.30	1.59	1.77	1.23			1.78
hum/ gal	E	11.53			12.28	10.91		
	I	36.48			56.49	43.81		
	R	3.16			4.60	4.02		
hum/ syr	E	10.99			12.02	12.34		
	I	29.10			39.53	33.23		
	R	3.30			3.29	2.69		

a) 从左至右分别是下列基因的缩写: *epsilon globin*, *gamma-1 globin*, *gamma-2 globin*, *delta globin*, *beta globin*, *Immunoglobulin alpha* 和*preproinsulin*; b) E, I, 和 R 分别代表外显子、内含子和内含子相对外显子的替代数的比率; c) 以下物种的缩写分别为: hum: human being; chm: chimpanzee; orn: orangutan; owm: old world monkey, 代表 macaca 和 African green monkey; nwm: new world monkey, 代表 cebus and owl monkey; gal: *galago crassicaudatus*; syr: *tarsius syrichta*

表 3 human 与 chimpanzee, human 与 orangutan, human 与 old world monkey 在不同基因外显子和内含子的替代速率

基因	外显子替代速率($\times 10^{-9}$)	内含子替代速率($\times 10^{-9}$)
human/ chimpanzee		
epsilon-globin	0.00(0.00~ 0.00)	0.89(0.62~ 1.24)
gamma-1 globin	0.49(0.34~ 0.68)	2.35(1.65~ 3.29)
gamma-2 globin	0.32(0.23~ 0.45)	2.16(1.52~ 3.03)
beta-globin	0.00(0.00~ 0.00)	1.20(0.84~ 1.68)
Immunoglobulin	2.21(1.55~ 3.09)	2.35(1.65~ 3.29)
prepro-insulin	1.31(0.92~ 1.83)	1.50(1.05~ 2.10)
平均	0.72(0.51~ 1.01)	1.74(1.22~ 2.44)
human/ orangutan		
epsilon globin	0.48(0.36~ 0.57)	1.59(1.19~ 1.91)
gamma-1 globin	0.48(0.36~ 0.57)	1.95(1.46~ 2.34)
gamma-2 globin	0.95(0.72~ 1.15)	1.85(1.39~ 2.23)
delta-globin	0.48(0.36~ 0.57)	1.49(1.12~ 1.79)
Immunoglobulin	2.22(1.66~ 2.66)	2.40(1.80~ 2.89)
平均	0.92(0.69~ 1.10)	1.86(1.39~ 2.23)
human/ old world monkey		
epsilon globin	0.37(0.31~ 0.46)	0.98(0.82~ 1.23)
beta-globin	1.03(0.86~ 1.29)	1.33(1.11~ 1.66)
Immunoglobulin	1.78(1.49~ 2.23)	1.34(1.11~ 1.67)
Prepro-insulin	1.13(0.95~ 1.42)	1.64(1.37~ 2.05)
平均	1.08(0.90~ 1.35)	1.32(1.10~ 1.65)

根据化石的记录, 如果将 human 与 cebus (new world monkey) 的分歧时间定为 38 Ma, 根据分子钟的假设, 在不同基因中, 外显子所推测出的 orangutan 的分歧时间为 5. 8~ 11. 78 Ma, 而基于内含子所推测出的 orangutan 的分歧时间为 12. 16~ 14. 44 Ma. 根据化石的记录, orangutan 的分歧时间为 10~ 16 Ma. 这个结果再次证明外显子的替代速率确实是在 *hominoids* 中降低了, 而且, 我们还注意到, 基于内含子所推测的分歧时间与化石记录是比较一致的.

2. 4 G+ C 含量和替代速率的关系

在表 4 中, 我们计算了内含子序列的 G+ C 含量, 并将它与外显子进行比较. 可以明显看出, 除了 *Immunoglobulin* 基因之外, 内含子的 G+ C 含量都比相应外显子低. 在 G+ C 含量与序列的替代速率两者间存在以下关系: 当序列的 G+ C 含量低时, 它的替代速率高. 这与一些文献中认为非编码区中的 A+ T 含量高的结果一致[12].

表 4 不同基因中外显子和内含子的 G+ C 含量

基因	外显子/ %	内含子/ %
Epsilon globin	52. 3	35. 6
Beta globin	56. 1	33. 9
Delta globin	55. 5	35. 5
Gamma-1 globin	53. 7	44. 2
Gamma-2 globin	54. 0	44. 0
Immunoglobulin	63. 5	67. 5
Prepro-insulin	66. 0	63. 9

3 讨论

本文针对 7 个断裂基因在灵长目物种中进行了基于内含子和外显子序列的系统发育分析的比较, 结果显示内含子序列可以像外显子一样为研究近亲缘物种之间的进化关系提供有用的信息, 而且在 *hominoids* 中, 内含子不存在替代速率下降的现象, 这意味着与外显子序列相比较, 内含子序列是一个更好的“分子钟”. 随着数据库中非编码区序列的逐渐增多, 可以进行更深一步的研究.

在计算突变速率时, 本文只考虑了替代速率, 而忽略了序列中的插入和缺失现象, 主要原因是由于核酸的替代速率通常会比插入和缺失速率高 10 倍以上[13]. 这样做会不可避免地给结果带来一些影响, 尤其是对于非编码区序列. 我们正在探索一些好的方法在计算突变速率时加进插入和缺失的因素.

致谢 作者非常感谢 Pennsylvania State University 的 M. Nei 教授寄来 MEGA 软件包. 本工作为国家自然科学基金(批准号: 39392900) 资助项目.

参 考 文 献

1 Russo C, Takezaki N, Nei M. Efficiencies of different genes and different tree-building methods in recovering a known vertebrate phylogy. Mol Biol Evol, 1996, 13: 525~ 536

2 Zuckerkandl E, Pauling L. Evolutionary divergence and convergence in proteins. In: Bryson V, Vogel H J eds. Evolving Genes and

Proteins. New York: Academic, 1965. 97~ 166
3 Li W-H, Tanimura M. The molecular clock runs more slowly in man than in apes and monkeys. Nature, 1987, 326: 93~ 96
4 Nei M. Phylogenetic analysis in molecular evolutionary genetics. Annu Rev Genet, 1996, 30: 371~ 403
5 Higgins D G, Bleasby A J, Fuchs R. Clustal V: improved software for multiple sequence alignment. Comput Appl Biosci, 1992, 8: 189~ 191
6 Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol, 1987, 4: 406~ 425
7 Joseph Felsenstein. Confidence limits on phylogenies: an approach using the bootstrap. Evolution, 1985, 39(4): 783~ 791
8 Taijima F, Nei M. Estimation of evolutionary distance between nucleotide sequences. Mol Biol Evol, 1984, 13: 269~ 285
9 Porter C A, Sampaio I, Schneider H, et al. Evidence on primate phylogeny from ε*globin* gene sequences and flanking regions. J Mol Evol, 1995, 40: 30~ 55
10 Hasegawa M, Kishino H, Yano T. Man's place in Hominoidea as inferred from molecular clocks of DNA. J Mol Evol, 1987, 26: 132~ 147
11 Maeda N, Wu C I, Bliska J, et al. Molecular evolution of intergenic DNA in higher primates: pattern of DNA changes, molecular clock, and evolution of repetitive sequences. Mol Biol Evol, 1988, 5: 1~ 20
12 Li W H, Graur D. Fundamentals of molecular evolution. Massachusetts: IAC Publishers, 1991
13 Saitou N, Ueda S. Evolutionary rates of insertion and deletion in noncoding nucleotide sequences of primates. Mol Biol Evol, 1994, 11: 504~ 512

(1998-12-28 收稿, 1999-04-07 收修改稿)

人 APP_{770} 基因在小鼠中的过量表达可导致 Aβ 的堆积和 ChAT 活性的显著减少

李 辉
(中国科学院动物研究所, 北京 100080)

摘要 利用转基因技术, 使人野生型 APP_{770} 基因在小鼠脑组织中过量表达. 鼠龄为5个月的转基因鼠主要在皮层和海马部位出现类似阿尔茨海默症的β-淀粉样肽(Aβ)的堆积, 其数目随鼠龄的增长而增加, 鼠龄为10个月的转基因小鼠在脑组织出现大量的Aβ的堆积. 同时, 10个月鼠龄的转基因小鼠的胆碱能神经系统发生了改变, 表现为皮层和海马部位的乙酰胆碱转移酶(choline acetyltransferase, ChAT)活性显著降低.

关键词 阿尔茨海默症 β-淀粉样肽 乙酰胆碱转移酶 转基因

阿尔茨海默症(Alzheimer's disease)是一种渐进性的神经退变性疾病, 其临床症状主要表现为记忆与认知能力的逐步丧失, 显著的病理症状表现为在大脑某些与认知功能相关的部位(皮层、海马等)形成细胞外的老年斑、细胞内的神经纤维性缠结以及神经元的死亡[1]. 此外, 在阿尔茨海默症的患者中, 与学习记忆功能有关的胆碱能神经系统也发生了改变[2]. 典型的老年斑通常包含一个核心结构, 其主要成分为Aβ, 是淀粉样前体蛋白(amyloid precursor protein, APP)的代谢产物. 根据氨基酸组成长度的不同, APP主要分为 APP_{770}, APP_{751} 和 APP_{695} 三种[3].