

蛋白质内含子演化分析*

谢 君 黄京飞 石秀凡 邵 丹 柳树群 梁宠荣 刘次全**

(中国科学院昆明动物研究所, 昆明 650223)

摘要 蛋白质内含子是能够自我剪切的一段多肽链, 它在生物三大系统中都存在, 但它的分布在各物种间以及蛋白质种类之间极不均衡。蛋白质内含子的演化和扩散也因此引起了人们的极大兴趣。经过系统搜索, 从已知的核酸序列中找到 69 个经典的蛋白质内含子。同源比较和系统树分析表明, 蛋白质内含子的演化应综合先天遗传和后天转移两方面的因素。

关键词 蛋白质内含子, 演化, 亲缘关系, 同源蛋白质内含子

学科分类号 Q71

蛋白质内含子是蛋白质中的一段多肽链, 它通过自我剪切的方式从宿主蛋白质中分离出来, 而两端的外显子以肽键的方式相连^[1]。它既不同于 I 类、II 类 RNA 内含子的自我剪切也不同于翻译后加工中的前体蛋白在酶的作用下的剪切。它不仅丰富了遗传信息翻译后加工的内容, 而且在蛋白质纯化方面也表现出诱人的前景。自从 1990 年发现第一个蛋白质内含子^[2]即引起了生物学家们的注视。

蛋白质内含子在古细菌、真细菌、真菌中均有分布, 而且真核生物的细胞核、细胞质中都存在, 这说明蛋白质内含子分布的广泛性。但迄今发现的蛋白质内含子大部分分布在古细菌中, 而真核中最少。有的物种中蛋白质内含子特别多, 如 *Methanococcus jannaschii* 中, 而大部分物种却不含蛋白质内含子。那么蛋白质内含子是如何演化和扩散的? 有人认为是先天遗传, 并随物种进化, 部分物种的蛋白质内含子丢失; 而另外一些人则认为, 蛋白质内含子是后天插入的, 并能在物种间横向扩散。本文用搜索到的蛋白质内含子进行同源比较和系统树分析, 表明蛋白质内含子的演化可能存在以上两种因素。

1 方 法

我们首先用 BLAST (<http://www.ncbi.nlm.nih.gov/BLAST/>) 中的 TBLAST (蛋白质序列对核苷酸序列搜索) 程序, 在 NR 数据库 (综合了 GenBank^[3], DDBJ^[4], PDB^[5], EMBL^[6] 数据库并剔除其中的重复序列而成的) 中搜索蛋白质内含子的特征序列, 找到 69 个经典蛋白质内含子。用

CLUSTALX^[7] 对蛋白质内含子进行多序列联配, 用 Multiple Sequence Alignment 构建系统树。然后结合同源蛋白质内含子, 对其进行聚类分析。

2 结 果

由于非同源蛋白质内含子的相似性差, 我们使用蛋白质内含子的特征序列进行系统树分析。图 1、图 2 分别列出了蛋白质内含子系统树和它们的外显子的系统树。其中蛋白质内含子的命名参考文献 [1] 的命名规则, 即用前面的三个字母表示蛋白质内含子的所在物种, 后面三个字母表示寄主蛋白。如果寄主蛋白是由几个亚基组成, 则用第 7 个字母表示所在亚基, 如果一个蛋白质中含有几个蛋白质内含子, 则用数字后缀加以区别。如 Mja Pol-1 表示物种 *Methanococcus jannaschii* DNA 多聚酶中的一个蛋白质内含子。同源蛋白质内含子是指插入到同一种蛋白质同一位置的蛋白质内含子。根据文献 [1] 以及蛋白质内含子数据库 (<http://blocks.fhcrc.org/~pietro/inteins/Inteins+hosts.map.html>), 图 1 中属于同源蛋白质内含子的有: Sce VMA 和 Ctr VMA(1); Mtu dnaB, Ssp dnaB 和 Rma dnaB(4); Mle gyrA, Mfl gyrA, Mgo gyrA, Mka gyrA 和 Mma gyrA(2); CIV RIR1, Spd RIR1, Dra RIR1, Pfu RIR1-2, Pho RIR1 和 Pab RIR1-3(5); Pfu RIR1-1 和 Pab RIR1-1(9); Tli pol-1, Tsp pol-2, Psp GBD

* 国家自然科学基金资助项目 (39770418)。

** 通讯联系人。

Tel: 0871-5195183, E-mail: xiesangjun@263.net

收稿日期: 2000-06-26, 接受日期: 2000-08-23

pol1, Psp KOD pol2, Pho pol 和 Mja pol2 (7); Tli pol2 和 Tfu pol2 (8); Tfu pol1, Tsp pol1, Psp pol2 和 Mja pol1 (11); Pho IF2, Mja IF2 和 Pab IF2 (6); Pho RFC, Mja RFC-1 和 Pab RFC-1 (12); Mja RFC-3 和 Pab RFC-2 (3); Pho lon 和 Pab lon (10); Mja Hyp2, Pab Hyp2 和 Pho Hyp2 (13).

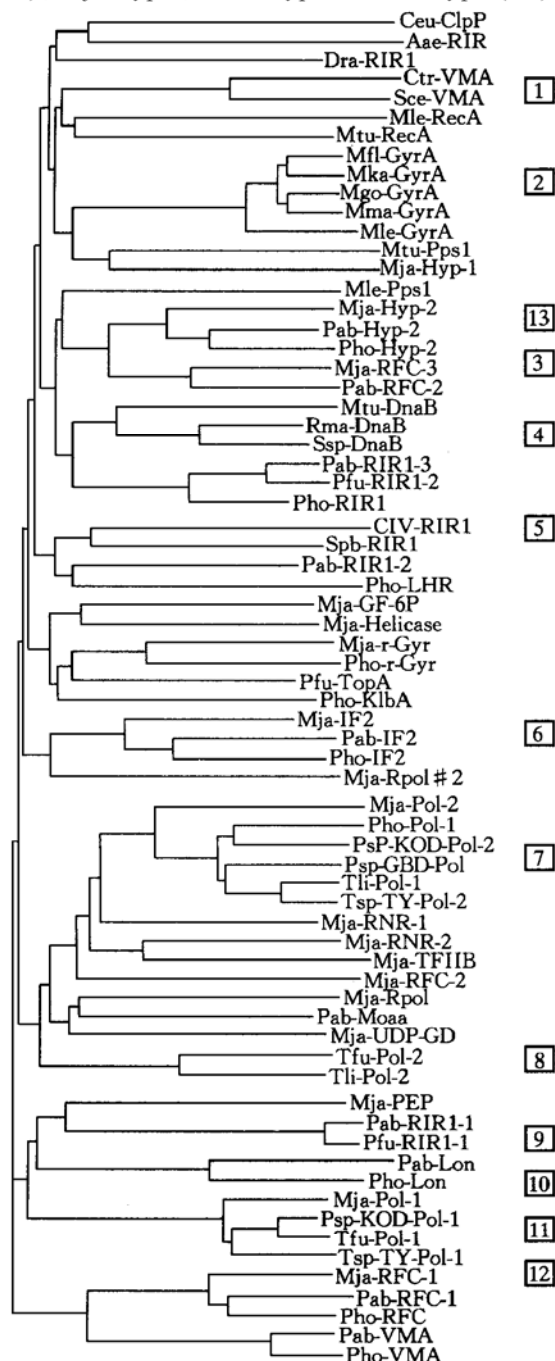


Fig. 1 The phylogenetic tree of inteins

我们将这 69 个蛋白质内含子的外显子分别与其同源蛋白进行比较, 发现在大多数情况下含蛋白质内含子的外显子在亲缘关系上也要近一些, 但也

它们在蛋白质内含子系统树上 (图 1) 和外显子系统树上 (图 2) 都相应聚集在一起. 这暗示这些蛋白质内含子是与外显子一起演化的. 但非同源的蛋白质内含子之间相似性极差, 它们不大可能来源于同一祖先. 而且从蛋白质内含子系统树也看不出其从低级到高级的演化过程.

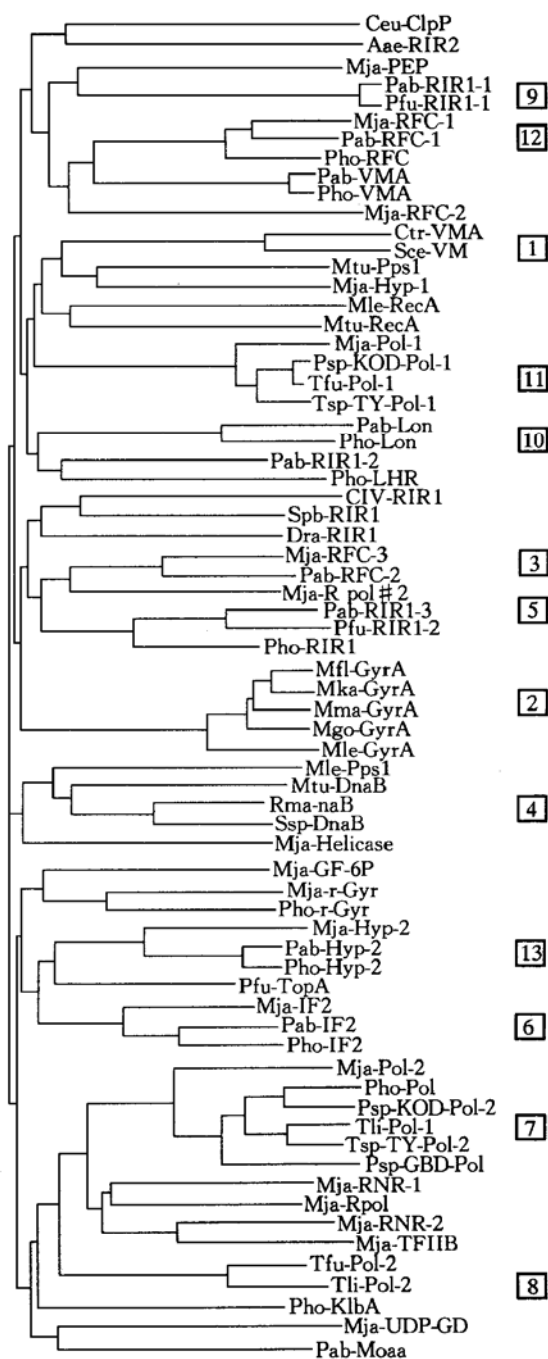


Fig. 2 The phylogenetic tree of exons

有例外, 如 DNA 多聚酶中和 DNA 螺旋酶中的含蛋白质内含子的外显子亲缘关系就不是最近的. 表 1 列出的是 9 种 DNA 多聚酶, 其中 *Pyrococcus*

horikoshii, *Pyrococcus* sp, *Thermococcus litoralis*, *T. Fumicolans*, *Thermococcus* sp, *Methanococcus jannaschii* 的 DNA 多聚酶中含有蛋白质内含子 (Tli pol 1, Tsp pol 2, Psp GBD pol 1, Psp KOD pol 2, Pho pol 和 Mja pol 2), 而 *Pyrococcus abyssi*, *Pyrococcus woesei*, *Pyrococcus furiosus* DNA 多聚酶中不含这一同源蛋白质内含子. 如果蛋白质内含子先天就有, 并与外显子一起演化, 那么含有蛋白

质内含子的 DNA 多聚酶应该亲缘关系更近, 但从表 1 看, 并非如此. 特别是含有蛋白质内含子的 *Methanococcus jannaschii* 的 DNA 多聚酶在亲缘关系上更远离其他 8 个物种, 而不含蛋白质内含子的 *Pyrococcus abyssi* 与含蛋白质内含子的 *Pyrococcus horikoshii* 的距离比其他含有蛋白质内含子的 4 个物种都要近. 这种现象是蛋白质内含子与蛋白质外显子一起演化假说无法解释的.

Table 1 Comparison of sequence similarity of 9 DNA polymerases

Pho pol 1	0.00000	0.61423	0.29224	0.08518	0.23547	0.19603	0.25209	0.30419	0.94236
Psp KOD pol		0.00000	0.68754	0.59796	0.70489	0.66567	0.68847	0.67521	0.47445
Tli POL-1			0.00000	0.29980	0.36659	0.31919	0.30670	0.06149	0.91346
Pal pol				0.00000	0.22938	0.18987	0.26824	0.30267	0.95487
Pwo pol					0.00000	0.03215	0.33603	0.35756	0.97310
Pfu pol						0.00000	0.29251	0.31401	0.94136
Tfu pol							0.00000	0.31898	0.99856
Tsp T Y pol								0.00000	0.91627
Mja pol 2									0.00000

3 讨 论

Gimble^[8]观察到, 当不含蛋白质内含子的等位基因通过有性生殖、接合、转导等进入酵母细胞中时, 这一不含蛋白质内含子的等位基因被蛋白质内含子的自导引核酸内切酶切开, 通过同源重组, 使蛋白质内含子基因转移. Fsihi^[9]发现编码 *gyrA* 中的蛋白质内含子的密码子使用频率不同于外显子的密码子使用频率, 这也暗示了 *gyrA* 蛋白质内含子是来自不同的物种. 以上 DNA 多聚酶中蛋白质内含子分布情况又给蛋白质内含子在不同物种间的横向传递提供了证据. 另外根据我们的分析结果, 有些蛋白质内含子的序列相似性非常低, 有的蛋白质内含子不可能来源于同一祖先. 最后, 如果蛋白质内含子是来源于同一祖先, 那么它们系统发生树上应该有一个从古细菌到真细菌然后再到真核生物的进化过程, 但从图 1 来看, 蛋白质内含子并没有这种依生物三大系统而分类聚集的现象, 这也暗示蛋白质内含子可能是在不同物种间横向传递的. 但从本文的蛋白质内含子和外显子的系统分析来看, 蛋白质内含子也存在与外显子一起演化的情况.

综合以上情况, 作者认为蛋白质内含子的演化可能存在先天遗传和后天插入两个方面因素. 最先

也许是由不同蛋白质内含子插入不同的物种, 如非同源蛋白质内含子也许是由不同的蛋白质内含子演化来的. 后来这些蛋白质内含子一边与外显子一起演化, 如大部分的同源蛋白质内含子. 同时又象分子寄体一样有可能插入其他的物种, 如 DNA 多聚酶中的蛋白质内含子. 基因重组可能是同源蛋白质内含子扩散的主要方式, 但这种方式扩散的条件是形成双倍体. 特别是当单倍体生物中的蛋白质内含子只有在偶尔形成双倍体时, 蛋白质内含子以同源重组的方式扩散才有可能. 至于非同源蛋白质内含子是如何插入异源物种, 是不是也象病毒一样侵入到其他物种的基因中呢? 现在还没有直接的实验证据.

参 考 文 献

- 1 Perler F B, Olsen G J, Adam E. Compilation and analysis of intein sequences. *Nucleic Acids Res*, 1997, **25** (6): 1087~ 1093
- 2 Kane P M, Yamashiro C T, Wolczyk D F, *et al.* Protein splicing converts the yeast TFP1 gene product to the 69-kD subunit of the vacuolar H⁺-adenosine triphosphatase. *Science*, 1990, **250** (4981): 651~ 657
- 3 Benson D A, Boguski M S, Lipman D J, *et al.* GenBank. *Nucl Acids Res*, 1999, **27** (1): 12~ 17
- 4 Yoshio T, Satoru M, Motonori O, *et al.* DNA data bank of Japan (DDBJ) in collaboration with mass sequencing teams. *Nucl Acids Res*, 2000, **28** (1): 24~ 26

- 5 Helen M B, Westbrook J, Feng Z, *et al.* The protein data bank. Nucl Acids Res, 2000, **28** (1): 235~ 242
- 6 Wendy B, Broek A, Camon E, *et al.* The EMBL nucleotide sequence database. Nucl Acids Res, 2000, **28** (1): 19~ 23
- 7 Thompson J D, Gibson T J, Plewniak F, *et al.* The clustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucl Acids Res, 1997, **25** (24): 4876~ 4882
- 8 Gimble F S, Thorner J. Homing of a DNA endonuclease gene by meiotic gene conversion in *Saccharomyces cerevisiae*. Nature, 1992, **357** (6376): 301~ 306
- 9 Fsihi H, Vincent V, Cole S T. Homing events in the *gyrA* gene of some mycobacteria. Proc Natl Acad Sci USA, 1996, **93** (8): 3410~ 3415

Analysis of the Intein' s Evolution^{*}

XIE Jun, HUANG Jing-Fei, SHI Xiu-Fan, SHAO Dan, LIU Shu-Qun, LIANG Chong-Rong, LIU Ci-Quan^{**}

(Kunming Institute of Zoology, The Chinese Academy of Sciences, Kunming 650223, China)

Abstract Inteins are internal segment peptides which can self-splicing at the protein level. Although inteins are found in all three domains of life, they are not evenly distributed among species and proteins. So the evolution of intein have been capturing much attention. 69 classical inteins were found through systematically searching nucleic acids database. The comparison of homologous protein and phylogenetic tree of inteins suggest that the evolution of inteins should combine two causes: lateral transmission and inheritance.

Key words intein, evolution, phylogeny, allelic intein

^{*} This work was supported by a grant from National Natural Science Foundation of China (39770418).

^{**} Corresponding author. Tel: 86-871-5195183, E-mail: xiesangjun@263.net

Received: June 26, 2000 Accepted: August 23, 2000