

基于 ORF Finder 方法的植物 ITS 片段结构特点分析

司 源, 郭亦琦, 孔航辉

(浙江大学 生命科学院, 浙江 杭州 310027)

摘要: 非编码区 ITS (The Internal Transcribed Spacer) 片段是目前分子系统学研究中应用最广的分子标记之一。运用生物信息学 ORF finder 技术, 以单子叶网状叶脉植物的 ITS 片段为例进行分析, 探讨了 ITS 内编码区 5.8S 和非编码区 ITS1 和 ITS2 的基本位置。并通过分析单子叶网状叶脉植物和菝葜科部分种类的 ITS 序列特点, 研究了该片段在科、属、种间存在的关系和变异特点。

关键词: ITS 序列; ORF finder; 单子叶网状叶脉植物; 菝葜科

中图分类号: Q94 文献标识码: A 文章编号: 1000-7091(2005)05-0054-03

Analysis of Structure of ITS Sequence on Monocots with Reticulate Vein Basen ORF Finder

SI Yuan, GUO Yi-qi, KONG Hang-hui

(College of Life Sciences, Zhejiang University, Hangzhou 310027, China)

Abstract: ITS (the Internal Transcribed Spacer) is one of the most widely used molecular sequences in molecular systematics. Employing ORF finder of bioinformatics, ITS Sequences of monocots with reticulate vein were analysed, and posit the coding fragment 5.8S and ITS1 and ITS2. By analysing the feature of ITS sequence of monocots with reticulate vein, the differentiation and variation characters of ITS sequences in the families, genera and species studied were discussed.

Key words: ITS Sequence; ORF finder; Reticulate vein monocots; Smilacaceae

生物信息学(Bioinformatics)是一门新兴的交叉学科。广义地说,生物信息学从事与基因组研究相关的生物信息的获取、加工、储存、分配、分析和解释,也就是以基因组 DNA 序列信息分析为基础,寻找到基因组序列中代表蛋白质和 RNA 基因的编码区;同时,阐明基因组中大量存在的非编码区与基因的表达调控关系,破译隐藏在 DNA 序列中的遗传语言规律;在此基础上,归纳、整理与基因组遗传信息释放及其调控相关的转录谱和蛋白质谱的数据,从而认识代谢、发育、分化与进化的规律。本研究主要运用的 ORF(open reading frame finder)系统是 Promotor 生物信息服务平台中的一种,它是一个分析工具,用于在用户或数据库提供的序列中寻找编码框,针对小基因序列(但要大于 50 bp),搜索并报导可能的蛋白质编码区。ORF Finder 把提交的序列分成 6

个亚区(每条链三种,对应三种不同的起始密码子),检测这六个阅读框架,并寻找以启动子和终止子为界限的 DNA 序列,符合这些条件的序列有可能对应一个真正的单一的基因产物^[1]。该方法提供了比用标准序列比对法更有用的信息。

被子植物 ITS(The Internal Transcribed Spacer)序列是核糖体 DNA 中介于 18 S 和 5.8 S 之间(ITS1)以及 5.8 S 和 26 S 之间(ITS2)的非编码转录间隔区。由于 ITS 存在于高度重复的 mtDNA 中,进化速度快并且片段长度不大,通常为 565~700 bp,是多拷贝的中度重复序列。所以扩增时只需用一对引物即可进行全序列测定,使该序列已广泛应用于被子植物等生物的系统进化的研究^{[2],[3]}。但由于该片段的进化速率在不同的类群中不同,在进化慢的类群中由于该片段长度短,提供的信息量有限,有时不能得

收稿日期:2005-06-08

基金项目:国家自然科学基金(30170062)

作者简介:司 源(1985-),女,河北唐山人,本科,主要从事生物进化与系统发育研究;孔航辉为通讯作者。

到期望的分辨率。因此,也要结合其他的信息进行分析。

单子叶植物多为平行叶脉。但也有部分单子叶植物同双子叶植物一样具有网状叶脉。如菝葜科 (Smilacaceae), 薯蓣科 (Dioscoreaceae), 百部科 (Stemonaceae), 以及百合科 (Liliaceae), 天南星科 (Araceae), 泽泻科 (Alismataceae), 兰科 (Orchidaceae) 的部分种。菝葜属 (*Smilax*) 和肖菝葜属 (*Heterosmilax*) 是菝葜属的 2 个主要属, 而菝葜复合种 (*Smilax china* complex) 是菝葜属广布的一个类群。利用序列分析的 ORF finder 方法对单子叶网状脉植物各代表种类的 ITS 片段进行比较研究, 可探讨 ITS 在植物不同水平的结构特征和演化规律。

1 材料和方法

表 1 ITS 数据来源及其序列特点
Tab. 1 ITS data source and sequence characters

物种 Species	来源 Locations	长度(bp) Length (ITS1+ 5. 8S+ ITS2)
菝葜(<i>Smilax china</i> L.)	中国, 韩国, 日本的 12 个居群	624
三脉菝葜(<i>Smilax trinervula</i> Miq.)	贵州贵阳	637
小果菝葜(<i>Smilax davidiana</i> A. DC)	浙江杭州	628
薯蓣(<i>Dioscorea opposita</i> Thunb.)	Unpublished	630
对叶百部(<i>Stemona tuberosa</i> Lour.)	Unpublished	733
大百合(<i>Cardiocrinum giganteum</i> L.)	Genebank accession: 3820602	625
掌叶半夏(<i>Pinellia pedatisecta</i> Schott)	Genebank accession: 29422116	721
慈菇(<i>Sagittaria trifolia</i> L.)	Genebank accession: 40353277	751
花叶开唇兰(<i>Anoectochilus formosanus</i>)	Genebank accession: 22651426	760
肖菝葜(<i>Heterosmilax japonica</i> Kunth)	湖南衡山	579

1.2 ORF 分析

进入 NCBI 网站 (<http://www.ncbi.nlm.nih.gov/>), 采用 ORF finder 功能, 将已获得的 ITS 序列输入, 可得到该序列可编码区 5. 8 S。ITS1 和 ITS2 是非编码区, 所以通过 ORF 无法表示出。而编码区 26 S 和 18 S 由于其序列过短或者由于引物干扰等原因采用该方法无法准确测出, 但可采用比对法 (Alignment) 确定 26 S 和 18 S 的位点。

1.3 ITS 序列同源性比较

将各材料的 ITS 序列输入 sequencher 软件 (ver. 4. 0 Gene Code, Ann Arbor, MI), 通过自动排序 (assemble automatically) 进行比对, 手工校正后, 分析种间, 属间, 科间的差异和特点。

2 结果与分析

2.1 ITS 序列结构特点及其同源性

通过 ORF finder 和序列比对方法结合, 可发现

1.1 ITS 数据来源

从本实验室获得菝葜属不同居群的菝葜 (*S. china* L.) 的 ITS 数据 12 份, 菝葜近缘种 2 份, 以及肖菝葜 (*Heterosmilax japonica* L.) 1 份 (表 1)。

从 Genebank 中分别获得其他网状脉单子叶植物薯蓣科 (Dioscoreaceae) 的薯蓣 (*Dioscorea opposita* Thunb.), 百部科 (Stemonaceae) 的对叶百部 (*Stemona tuberosa* Lour.), 百合科 (Liliaceae) 的大百合 (*Cardiocrinum giganteum* L.), 天南星科 (Araceae) 的掌叶半夏 (*Pinellia pedatisecta* Schott), 泽泻科 (Alismataceae) 的慈菇 (*Sagittaria trifolia* Linn.) 以及兰科 (Orchidaceae) 的花叶开唇兰 (*Anoectochilus formosanus* Lindl) 的 ITS 完整片段 (表 1)。

供试材料 ITS 区序列总的特征为: 18 S 结尾序列为 CCTGCGGAAGATATTG, ITS1 片段位于 20~ 282 bp 区, 5. 8 S 结束在 470 bp 左右, ITS2 位于 470~ 730 区, 26 S 起始序列为 GCGACCCCAGGTCAGG。在所有的位点中, 含 143 个变异位点, 集中在 ITS1, ITS2 的两端。

对这些材料 ITS 完整片段 (ITS1+ 5. 8 S+ ITS2) 的结构分析发现: 薯蓣 (薯蓣科) ITS 结构和菝葜科最接近, 因此有的分类学家将薯蓣科和菝葜科归为一目^[5]。对叶百部 (百部科) ITS 片段较长为 733 bp, 5. 8 S 长度为 195 bp, 可划分为 ITS1 1 bp~ 204 bp, ITS2 400 bp~ 733 bp。大百合 (百合科) ITS 片段较短为 626 bp, ITS 1 位于 1 bp~ 223 bp, 5. 8 S 为 102 bp, ITS2 大小区域为 325 bp~ 626 bp。掌叶半夏 (天南星科) ITS 区域总长 721 bp, ITS1 1 bp~ 311 bp, 5. 8 S 长度约为 161 bp, ITS2 474 bp~ 721 bp。慈菇 (泽泻科) ITS 区序列为 720 bp, ITS1 范围在 1 bp~ 274 bp, 5. 8

S 大小为 152 bp, ITS2 片断大小为 426 bp~ 720 bp。花叶开唇兰(兰科) ITS 片短含有部分的 18 S 和 26 S 片断,通过比对和 ORF 结合发现: 1 bp~ 83 bp 为部分 18 S, ITS1 长度为 84 bp~ 232 bp, ITS2 长度是 485 bp~ 737 bp, 5.8 S 为 161 bp, 738 bp~ 760 bp 为部分的 28 S。

菝葜科菝葜的 2 个近缘种, ITS 区序列平均含 630 bp 左右,大多数具有 20~ 35 bp 的空位,分析可见: ITS1 为 1 bp~ 230 bp; 5.8 S 结束在 390 bp 左右; ITS2 位于 390 bp~ 630 bp。肖菝葜 ITS 序列,有 579 bp, 5.8 S 长度为 163 bp, ITS1 长度为 277 bp 左右, ITS2 位于 342 bp~ 579 bp。菝葜(*Smilax china*) 的 12 个居群的 ITS 区平均含 624 bp, ITS1 为 1 bp~ 223 bp, 5.8 S 为 162 bp, ITS2 为 387 bp~ 624 bp。

通过序列比对发现,测试类群在属以上水平同源性较小,而且 ITS 长度分化较大,如菝葜科和薯蓣科同源性相对较高但也仅有 60% 左右,百合科和菝葜科 ITS 序列同源性只有 33.48%,其他各科之间的同源性更低。所以可以认为 ITS 片断不适合于科间系统发育分析^[6]。

虽然 ORF finder 系统与人工比对法相比效率可大大提高,但是 ORF finder 方法并不是完全准确的,和人工比对法相比约有 1~ 10 bp 的差异。因为在输入测试序列后,该系统会按照起始子和终止子的位点所在,按照不同的 6 种测试方法进行编码区的寻找,会产生多种结果,要由我们按照 5.8 S 的区域长度范围进行区分。另一方面,对于过短的编码区(小于 50 bp)则无法测试出。但是,对于一未知的物种,我们通过该方法可以得到 ITS1 和 ITS2 的基本位置,以便我们根据已知物种 ITS 片断,推断未知物种的分类信息。

2.2 菝葜科种属间 ITS 序列的差异

菝葜属和肖菝葜的 ITS 区在长度上有所差异,肖菝葜初始端较菝葜属长 5 bp,而菝葜属比肖菝葜属在终止末端多出 35 bp,这都是由 ITS 序列的特性产生的。而两个属在序列差异集中表现在第 5~ 20 位点,第 94~ 127 号位点,第 186~ 191 位点和第 234~ 238 位点。

可见两者在 ITS1 同源性较弱,ITS2 同源性较好,仅差 3 个 bp。

从菝葜复合种(*S. china* complex)的 DNA ITS 树可以发现菝葜 2X 居群和其近缘种三脉菝葜,小果菝葜之间有较高的亲缘关系,其同源性分别为

95.96% 和 97.7%。通过 ITS 比对发现,大部分的碱基差异集中在 8 bp~ 20 bp, 150 bp~ 202 bp,即 ITS1 区域。从序列比对发现,由于 *S. trinervula* 的重复序列导致 *S. china* 和 *S. davidiana* 的碱基空位。进一步分析发现,菝葜(*S. china*) 2X, 4X 和 6X 在 ITS2 区域内同源性很高(91%),而在 ITS1 中则同源性较低(76%),仅集中表现在 1~ 30 bp。这是由于 ITS1 上游的 18 S 的长度在不同居群内是不同的,导致 ITS1 片断起始端位移变化^[7]。庐山居群的 ITS1 最短,韩国居群的 ITS1 最长。将 12 个居群 ITS1 比对后发现,多数居群和湖北 2X 居群有较高的同源性,在 28 bp~ 32 bp, 50 bp~ 62 bp, 104 bp~ 115 bp, 185 bp~ 204 bp 区域内出现了各倍性居群间的分化。这种变异是由于 4X 和 6X 居群重复序列的插入而产生。

通过上述分析,可见 ITS 序列在属以上水平应用前景不大,但是在种内和种间的应用是广阔的。属以下的水平的进化主要依赖于 ITS1 片断的变化,ITS2 相对保守。复合种菝葜(*S. china*)中,在湖北发现的 2X 可能为其原始居群,与 2X 的三脉菝葜,小果菝葜有很近的亲缘关系,有待进一步利用克隆测序寻找其种间关系和起源。

参考文献:

- [1] 齐 眉,于修平,卞继峰. HPV16L1 ORF 的基因克隆及在大肠杆菌中的表达[J]. 山东医科大学学报, 2000, 38 (2): 117~ 119.
- [2] Baldwin B G. Phylogenetic utility of the internal transcribed spaces of nuclear ribosomal DNA implants: An example for the compositas[M]. Mol Phylogenet Evol, 1992, 1: 3~ 16.
- [3] Tsolaki A G, Miller R F, Underwood A P, et al. Genetic diversity at the internal transcribed spacer regions of the rRNA operon among isolates of *Pneumocystis carinii* from AIDS patients with recurrent pneumonia[J]. Journal Infectious Diseases, 1996, 174(1): 141~ 156.
- [4] Swofford D L. PAUP: phylogenetic analysis using parsimony version 4.0 b10. Sinauer, Sunderland, MA, 2002.
- [5] Dahlgren R M T. A system of classification of the angiosperms to be used to demonstrate the distribution of characters[J]. Botaniska notiser, 1975, 128: 119~ 147.
- [6] 赵志礼,徐珞珊,董 辉. Evaluation of ITS Sequence of nrDNA in plant molecular systematic[J]. 植物资源与环境学报, 2000, 9(2): 50~ 54.
- [7] Soltis D E, Soltis P S, Nickrent D L. Angiosperm phylogeny inferred from 18S ribosomal DNA sequences[J]. Ann Missouri Bot Gard, 1997, 84: 1~ 49.