大豆二核苷酸 SSR 侧翼序列保守性分析

詹少华[1],盛新颖[2],樊洪泓[2],蔡永萍[2],林 毅[2]

(1. 皖西学院 化学与生命科学系, 安徽 六安 237012 ;2. 安徽农业大学 生命科学学院, 安徽 合肥 230036)

摘 要:从 NCBI 下载了大豆的 EST 和 GSS 序列,以 SSR 两侧序列各 100 nt 的片段作为分析材料,统计 SSR 数目、计算侧翼序列碱基组成,对 SSR 侧翼序列进行了序列比对和数据库 blast 搜索。结果表明:大豆 SSR 分布频率为 1/6.67kb,侧翼序列 GC 含量为 37.83%,二核苷重复是 SSR 的主要类型(65.7%),同一种重复基元的 SSR 存在多类型侧翼序列,同一类型的侧翼序列高度保守,不同类型的侧翼序列之间相似性较小,尤其是非 SSR 区域可以存在与 SSR 侧翼序列相同的序列。由上述结果推断 SSR 引物可能扩增出不含 SSR 的产物,易位可以产生形成新类型 SSR。这一分析结果有助于提高 SSR 引物设计效率和进一步阐明 SSR 的形成机理。

关键词:

中图分类号:S565.1 **文献标识码**:A **文章编号**:1000-9841(2010)02-0195-04

Analysis on the Conservation of Dinucleotide SSR Flanking Sequences in Soybean(*Glycine max*)

ZHAN Shao-hua[1], SHENG Xin-ying[2], FAN Hong-hong[2], CAI Yong-ping[2], LIN Yi[2]

(1. Chemistry and Biology Department, West Anhui University, Lu'an 237012; 2. Life Science School, Anhui Agricultural University, Hefei, 230036, Anhui, China)

Abstract: The soybean EST and GSS sequences were downloaded from NCBI, and the 100 nt SSR flanking sequences were extracted as analyzing material. The number of SSRs and the base composition of the flanking sequences were calculated. Meanwhile, the multiple sequence alignment and the blast scanning were conducted. The results showed distribution frequency of soybean SSRs was 1/6.67kb, GC content of the flanking sequences was 37.83%, and the dinucleotide was major type of the motif. The multiple sequence alignment revealed polymorphic types flanking sequences were in any kind of motif. The same type flanking sequences were highly conserved, but different type flanking sequences had less similarity. Especially, the blast analyzing showed the SSR flanking sequences may exist in no SSR regions. So we claimed no SSR fragments may be amplified by SSR primer pairs, and new type SSR could stem from translocation, which is one kind of chromosomal variations. The analyzing could contribute to efficient design of SSR primers and to further illuminate formative mechanism of simple sequence repeats.

Key words: Soybean; Simple Sequence Repeat; Conservation of Sequence, Flanking Sequence; Primer

真核生物基因组中存在大量长度可以变化的串联重复序列(variable number tandem repeat, VNTR)[1],重复基元数较少(1-6)的 VNTR 被 Jacob 等命名为简单重复序列(simple sequence repeat, SSR)[2],SSR 在基因组中广泛随机分布、多态性好,是一种信息丰富的遗传学标记。1989 年,3 个著名实验室同时报导了根据 SSR 两侧的保守序列设计引物,采用 PCR 方法检测 SSR 的长度多态性[3-5],后来人们发现,SSR 标记还表现为复等位共显性的孟德尔式遗传、重复性好、成本相对较低等优点[6],因此 SSR 分子标记成为当前应用最广泛的分子标记之一[1-9]。

Akkaya 等最早证实了大豆含有高多态性的 SSR 标记[10],目前,SSR 标记在大豆遗传图谱构建、遗传多样性、进化关系、QTL 分析、种质鉴定、分子标记辅育种等方面应用越来越广泛[11-12]。

SSR 侧翼保守序列是设计引物的依据,已经报导的大豆 SSR 引物对超过 900 个[11],目前是通过试验的方法筛选这些引物,对于 SSR 侧翼序列的保守程度、序列长度、在基因组中是否只分布于 SSR 的两侧、是否是所有的 SSR 都具有保守序列、同一种类型的 SSR 是否存在多种保守序列等问题尚缺乏深入研究。1992 年 Akkaya 等分析大豆 SSR 时,Genbank 中只有 141 个大豆序列[10],而目前 Genbank 中大豆核苷酸序列数近 200 万个,随着大豆核苷酸数据库中序列的激增和分析技术的发展,分析这些问题成为可能

收稿日期:2009-08-13

基金项目:安徽高校省级自然科学研究重点资助项目(KJ2008A089)。

第一作者简介:詹少华(1963-),男,副教授,博士,研究方向为大豆和棉花分子育种。E-mail:zhansh@wxc.edu.cn。

通讯作者:林毅,教授,博士生导师。E-mail:linyiahau@126.com。

和必要。现以大豆二核苷酸 SSR 为例,分析 SSR 侧翼序列的保守性特征,为有针对性地设计 SSR 引物和分析 SSR 的形成机理提供依据。

1 材料与方法

1.1 序列下载和 SSR 位点的发掘

分别从 NCBI(美国国立生物技术信息中心,http://www. ncbi. nlm. nih. gov)的大豆 EST(Expressed Sequence Tag,表达序列标签)和 GSS(Genome Survey Sequence,基因组勘测序列)数据库查询和随机下载长度不小于600 nt的序列各 5 000 个,应用 SSRIT(http://www. gramene. org/db/searches/ssrtool)在线搜索 SSR,检索标准为重复基元2,3,4,5,6,最小重复次数为5。

1.2 SSR 侧翼序列的处理

选取所有二核苷酸 SSR 起点距 5′端大于 100 nt,终点距 3′端也大于100 nt的序列,截取 SSR 左侧上游 100 nt 序列和右侧下游 100 nt 序列,截取的序列命名原则是:前 2 个字母表示重复基元,第 3 个字母 l 表示左侧序列,r 表示右侧序列,然后是数据编号,例如 acl2 表示 ac 重复基元左侧截取的序列,编号为 2。

1.3 分析软件与数据处理

碱基组成和碱基相似性分析采用软件 BioEdit,序列比对采用软件 ClustalX2. 0. 10 和 NCBI 在线 Blastn。

2 结果与分析

2.1 含 SSR 的序列数及侧翼序列碱基组成

二核苷酸 SSR 一般可以分为 4 类[12],分析序列保守性需要区分序列的方向,所以将二核苷酸 SSR 按照 6 类来统计。总共 10 000 个序列含有 873 个 SSR 位点,大约 12 个序列含有 1 个 SSR 位点,SSR 分布频率总平均为 1/6.67kb,二核苷酸占 65.7%,是 SSR 的主要类型,其中 CG/CG 的 SSR 基元总共只有 1 个。该文分析侧翼序列不少于 100 nt 的 SSR,筛选后的 GSS 序列 99 个,EST 序列 122 个。

表 1 EST 和 GSS 序列含 SSR 的数目

Table 1 SSR number in the EST and GSS sequences

数据库 Database	基元核苷酸数量 Number of nucleotide in motifs										
	2						3	4	5	6	合计 Total
	AT/TA	AC/CA	AG/GA	CT/TC	GT/TG	CG/CG					
GSS	115	29	48	46	16	1	91	3	0	0	349
EST	61	27	98	119	14	0	191	8	4	2	524
合计 Total	176	56	146	165	30	1	282	11	4	2	873

表 2 中分别列出了 EST-SSR 2 个侧翼序列 A + T 和 G + C 的含量,并由此可以计算出大豆 EST-SSR 侧翼序列 A + T 平均为 59.86%,G + C 平均为 40.12%;同理,由表 3 可知大豆 GSS-SSR 侧翼序列 A + T 平均为 63.77%,G + C 平均为 35.54%。在所有被分析的 EST-SSR 和 GSS-SSR 中,A + T 含量 55.64% ~72.71%,总平均为 61.81%,侧翼序列 G + C 含量为 27.29% ~44.36%,总平均为 37.83%。

表 2 大豆 EST-SSR 侧翼序列的碱基组成

Table 2 Base composition of the EST-SSR flanking sequences

基元类型 Motif type	左侧序列 Left side sequences						右侧序列 Right side sequences					
	A	T	G	C	A + T/%	G + C/%	A	T	G	C	A + T/%	G + C/%
AC/CA	321	364	127	288	62.27	37.73	321	291	222	266	55.64	44.36
AG/GA	843	855	592	509	60.64	39.32	844	822	610	524	59.50	40.50
AT/TA	1424	1190	699	785	63.76	36.20	1225	1281	830	761	61.12	38.80
CT/TC	751	903	442	703	59.07	40.89	724	903	538	635	58.11	41.89
GT/TG	122	124	69	85	61.50	38.50	110	118	82	90	57.00	43.00
总计或平均 Total or avarege	3461	3436	1929	2370	61.45	38.53	3224	3415	2282	2276	58.27	41.71

表 3 大豆 GSS-SSR 侧翼序列的碱基组成

Table 3 Base composition of the GSS-SSR flanking sequences

基元类型 Motif type	左侧序列 Left side sequences						右侧序列 Right side sequences					
	A	T	G	C	A + T/%	G + C/%	A	T	G	C	A + T/%	G + C/%
AC/CA	256	286	121	224	60.22	38.33	322	251	93	232	63.67	36.11
AG/GA	648	494	433	323	60.11	39.79	661	446	461	310	58.26	40.58
AT/TA	1430	1509	707	635	68.35	31.21	1650	1585	690	681	68.83	29.17
CT/TC	491	541	286	377	60.71	39.00	481	562	274	374	61.35	38.12
GT/TG	202	307	108	83	72.71	27.29	186	258	145	105	63.43	35.71
总计或平均 Total or avarege	3027	3137	1655	1642	64.42	35.13	3300	3102	1663	1702	63.11	35.94

2.2　同类基元侧翼序列的 ClustalX 比对和相似性分析

由图 1 ClustalX2.0.10 序列比对的结果可以看出,各序列排列并不整齐,表明侧翼保守序列与核心基元之间的距离不完全一致。EST-SSR 的 atl1、atl2 、atl3、atl13 存在高度相似,其中 atl1、atl2 、atl3 的 100 个碱基完全一致,他们与 atl13 相差一个碱基的空位,但是这 4 个序列与其它序列之间的相似性较低,atl11 与 atl4 序列存在 97% 的碱基一致性,atl18 的 100 个碱基中 68 个碱基与 tal13 相应片段的 61 个碱基一致,65 个碱基与 atl20 相应片段的 58 个碱基一致,表明同一种核心基元存在多种保守序

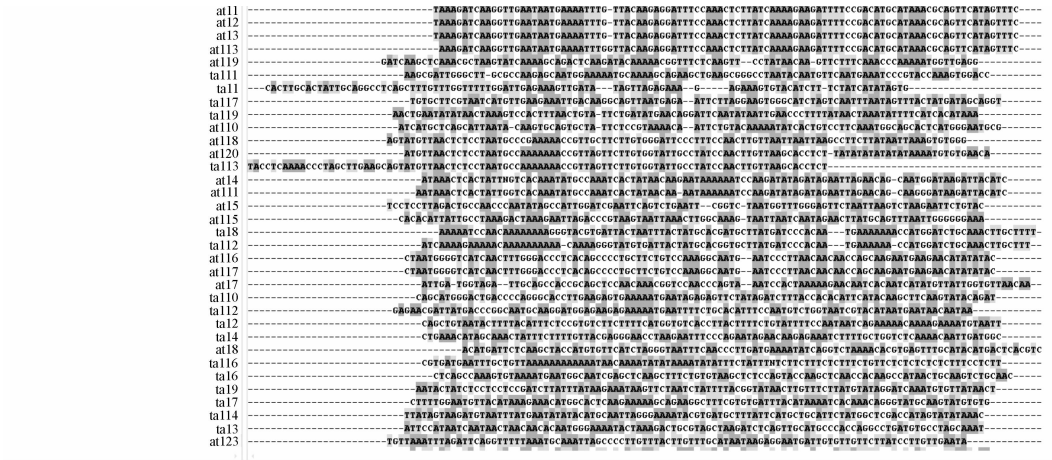

图 1　EST-SSR 基元为 AT/TA 的左侧序列 ClustalX 比对结果

Fig. 1　Multiple sequence alignment for the left sequences of AT/TA motif in the EST-SSR

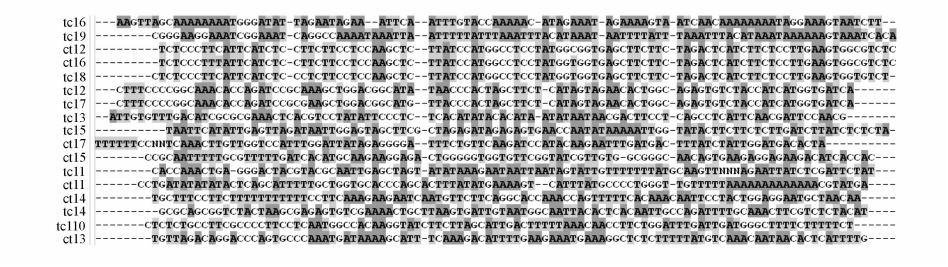

图 2　GSS-SSR 基元为 TC/CT 左侧序列 ClustalX 比对的结果

Fig. 2　Multiple sequence alignment for the left sequences of TC/CT motif in the GSS-SSR

表 4　EST-SSR 侧翼序列的相似性

Table 4　Similarity of the EST-SSR flanking sequences

侧翼序列类型 Flanking sequences type	atl1	atl2	atl3	atl4	atl11	atl13	atl16
atl2	1.000						
atl3	1.000	1.000					
atl4	0.252	0.252	0.252				
atl11	0.250	0.250	0.250	0.960			
atl13	0.980	0.980	0.980	0.252	0.250		
atl16	0.254	0.254	0.254	0.269	0.278	0.254	
atl17	0.254	0.254	0.254	0.269	0.278	0.254	1.000

列,且保守序列的长短不一,这一点从图 2 的 GSS-SSR 也得到证明,例如 ctl2、ctl6、tcl8 之间以及 tcl2,tcl7 之间存在高度保守序列。同时也发现 tcl10 等序列与其他被分析序列不存在保守序列。表 4 定量地表明了 SSR 侧翼序列的上述特点,其他侧翼序列比对都得到类似结果。

2.3　侧翼序列的 Blast 分析

同一个重复基元的两侧序列成为 1 对侧翼序列,EST 序列中,acl2 和 acr2 是同一个重复基元 AC 的侧翼序列,以 EST-SSR 的 acl2 和 acr2 2 个序列为例,通过 NCBI 大豆 EST 数据库在线 Blast 查询(2009 年 8 月 9 日),与 acl2 高度相似的序列 36 个,与 acr2 高度相似的序列 68 个,比较二者的查询结果,其中 32 条记录为相同序列,该 32 条序列查找到 64 个 SSR 位点,其中每个序列都至少包含 1 个基元为 AC 或 GT 的 SSR,重复基元为 GT 的 SSR,其互补链中存在(AC)n 的 SSR,GSS 的 acl2 与大豆 GSS 数据库 Blast 比对,共得到 27 条高度相似序列,GSS 的 acr2 与大豆 GSS 数据库 Blast 比对,查询到 32 条高度相似序列,2 者共有的序列 13 条,该 13 条序列有 6 条含有(AC)n 或(GT)n 的 SSR,表 5 是该 6 条序列与 acl2、acr2 的 Blast 结果。

表 5 GSS-SSR 的侧翼序列与大豆 GSS 数据库 Blast 分析结果

Table 5 Blast between the flanking sequences of the GSS-SSR and the soybean GSS database

记录号 Accession	AC 基元左侧序列 acl2 Left side sequences acl2 of motif AC					AC 基元右侧序列 acr2 Right side sequences acr2 of motif AC				
	最大得分 Max score	总分 Total score	查询覆盖率 Query Coverge/%	期望值 Evalue	最大匹配 Max ident/%	最大得分 Max score	总分 Total score	查询覆盖率 Query Coverge/%	期望值 Evalue	最大匹配 Max ident/%
BH610110	185	185	100	2e-46	100	185	185	100	2e-46	100
CG813598	185	185	100	2e-46	100	185	185	100	2e-46	100
CG821161	185	185	100	2e-46	100	185	185	100	2e-46	100
ED732909	169	169	100	2e-41	97	100	100	63	8e-21	95
ED754777	163	163	100	1e-39	96	152	152	94	2e-36	95
ED775287	152	152	100	2e-36	94	135	135	93	2e-31	93

3 讨论

从分析结果来看,对同一重复基元的 SSR 而言,存在多种侧翼序列,根据序列的保守性可以将这些序列分成若干类型,同一类型的侧翼序列高度保守,相似性可以达到 100%,但是不同类型的侧翼序列之间保守性较低,相似性一般介于 20% ~30% 之间,保守序列的长度也可以发生变化,而且保守序列与 SSR 的起点或终点的距离也可以不同。采用成对的侧翼序列与大豆核苷酸数据库 Blast 分析,发现非 SSR 区域可以同时存在这样的 2 个侧翼序列,如果依据这样的侧翼序列设计引物,PCR 扩增的产物可以不包含 SSR,例如 GSS-SSR 的 acl2 和 acr2 可以同时与 CL895026、CZ509004、CZ522300、ED685085、ED717916、ED736050、ED743287 存在高度相似性。SSR 引物可能扩增出非 SSR 产物,SSR 分子标记就不仅仅是简单重复序列的长度多态性,在利用 SSR 进行基因定位时,也要考虑到产物的假阳性可能。

Jeffreys 等认为 SSR 是由于高等生物减数分裂过程中经过不对等交换或者 DNA 的滑动复制产生的[13-14],但是这只能解释高度保守的 SSR 侧翼为什么存在重复基元的长度多态性。同一重复基元的 SSR 具有多种侧翼序列以及基因组中非 SSR 区域存在 SSR 侧翼序列,这表明 SSR 还可以通过易位等方式产生。

碱基组成是设计引物的重要参数,从表 2 和表 3 可以看出大豆 SSR 的侧翼序列 AT 比例多数在 60% ~65% 之间,常规 SSR 引物设计,缺乏对 SSR 侧翼序列保守性的分析,这样设计的引物需要通过试验来筛选,此工作耗时又增加成本,通过 Blast 分析,如果能找到较多与侧翼序列高度相似的序列,优先选择这些序列设计引物,成功率有可能显著提高。对于已经设计的引物,通过同源序列相似性分析以及高度相似序列的多少作为初选手段,是否可以减少使用引物的盲目性,有待于进一步证明。

参考文献

[1] Nakamura Y, Leppert M, O'Connell P, et al. Variable number tandem repeat (VNTR) markers for human gene mapping [J]. Science, 1987, 235:1616-1622.

[2] Jacob H J, Lindpaintner K, Lincoln S E, et al. Genetic mapping of a gene causing hypertension in the stroke-prone spontaneously hypertensive rat [J]. Cell, 1991,67:213-224.

[3] Litt M, Luty J A. A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene [J]. American Journal of Human Genetics,1989, 44:397-401.

[4] Weber J L, May P E. Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction [J]. American Journal of Human Genetics, 1989, 44: 388-396.

[5] Tautz D. Hypervariability of simple sequences as a general source of polymorphic DNA markers [J]. Nucleic Acids Researcn. 1989, 17:6463-6471.

[6] Powell W, Machray G C, Provan J. Polymorphism revealed by simple sequence repeats [J]. Trends in Plant Science, 1996(1): 215-222.

[7] 杨喆,刘丽君,高明杰,等.大豆高蛋白基因分子标记及其在大豆育种中的应用[J].大豆科学,2008,27(2):186-189.(Yang Z, Liu L J, Gao M J, et al. QTL Tagging for high protein gene and using molecular marker assistant selection in soybean breeding [J]. Soybean Science, 2008, 27(2):186-189.)

[8] 王俊,王林林,刘章雄,等.大豆地方品种遗传结构及其保存研究[J].大豆科学,2008,27(3):361-365.(Wang J, Wang L L, Liu Z X, et al. Genetic structure and conservation of soybean landraces[J]. Soybean Science, 2008, 27(3):361-365.)

[9] 汤复跃,周立人,程潇,等.大豆 M 型细胞质雄性不育恢复基因 SSR 标记初步定位[J].大豆科学,2008,27(3):383-386 (Tang F Y, Zhou L R, Cheng X, et al. SSR Marker location for fertility restorer gene of M-cytoplasmic male sterility in soybean [J]. Soybean Science, 2008, 27(3):383-386.)

[10] Akkaya M S, Shoem aker R C, Specht J E, et al. Length polymorphism of simple sequence repeat DNA in soybean [J]. Genetic, 1992, 132:1131-1139.

[11] 王彪,邱丽娟.大豆 SSR 技术研究进展[J].植物学通报,2002, 19(1):44-48.(Wang B, Qiu L J. Current advance of simple sequence repeats in soybean [J]. Chinese Bulletin of Botany, 2002, 19(1):44-48.)

[12] 詹少华,盛新颖,樊洪泓,等.大豆 EST 序列长度与 SSR 特性的关系[J].大豆科学,2009,28(2):204-208.(Zhan S H, Sheng X Y, Fan H H, et al. Relationship between the length of soybean ESTs sequence and characters of EST-SSR[J]. Soybean Science, 2009,28(2):204-208.)

[13] Jeffreys A J, Wilson V, Thein S L. Hypervariable "minisatellite" regions in human DNA [J]. Nature, 1985, 314: 67-73.

[14] Crow J F, William F D. Anecdotal, historical and critical commentarieson genetics: Unequal crossing over then and now [J]. Genetics, 1988, 120: 1-6.