王 为,叶泗洪,潘宗瑾,等. 棉花分子标记冗余性检测与评价的方法[J]. 江苏农业学报,2015,31(2):247-252.
doi:10.3969/j.issn.1000-4440.2015.02.004

棉花分子标记冗余性检测与评价的方法

王 为[1,2], 叶泗洪[3], 潘宗瑾[1], 王海洋[1], 高 进[1], 蔡立旺[1], 陈建平[1], 王永慧[1], 潘群斌[1], 王长彪[4]
(1.江苏沿海地区农业科学研究所/农业部沿海盐碱地科学观测实验站,江苏 盐城 224002; 2.中国农业科学院棉花研究所/棉花生物学国家重点实验室,河南 安阳 455000; 3.安徽省农业科学院棉花研究所,安徽 安庆 246003; 4.山西省农业科学院生物技术中心,山西 太原 030031)

摘要: 网上数据库公布的棉花分子标记引物序列存在冗余性,至今也鲜见合适的软件同时分析一对引物的冗余性。为减少不同研究者的重复开发,节约时间和成本,提高对网上序列信息的利用率,借助自主开发的能同时分析一对引物和已释放的引物间是否存在冗余性的软件 SSRD1.0,利用 90 对前人已开发的并用该软件预测过相互间存在冗余性的 SSR 引物,以 TM-1、海 7124 基因组为模板分别进行扩增、电泳、测序。结果表明软件预测和基因型水平检测有 88.8% 的吻合度;在 Identity 阈值设为 50%、70% 时,测序结果表明序列水平和软件预测分别有 75.0%、53.8% 的吻合度。分别从软件预测、基因型、序列 3 个水平进行检测、验证,结果表明这一冗余软件和冗余性预测方法是较为有效、可行的。

关键词: 棉花; 分子标记; SSR; 冗余性软件; 检测; 预测
中图分类号: S562.032 **文献标识码:** A **文章编号:** 1000-4440(2015)02-0247-06

An approach to detecting and evaluating molecular marker redundancy in cotton

WANG Wei[1,2], YE Si-hong[3], PAN Zong-jin[1], WANG Hai-yang[1], GAO Jin[1], CAI Li-wang[1], CHEN Jian-ping[1], WANG Yong-hui[1], PAN Qun-bin[1], WANG Chang-biao[4]
(1. *Institute of Agricultural Sciences in the Coastal District of Jiangsu Province/Observation and Experimental Station of Saline Land in Coastal Area, Ministry of Agriculture, Yancheng* 224002, *China*; 2. *Cotton Research Institute, Chinese Academy of Agricultural Sciences/State Key Laboratory of Cotton Biology, Anyang* 455000, *China*; 3. *Cotton Research Institute, Anhui Academy of Agricultural Sciences, Anqing* 246003, *China*; 4. *Bio-technology Research Center, Shanxi Academy of Agricultural Sciences, Taiyuan* 030031, *China*)

Abstract: Currently there exists redundancy among the sequences of cotton molecular markers which amplifie the same regions on the genome. However, softwares which can be used to simultaneously analyze the redundancy of a pair of primers are not available. To improve the utilization of sequence information, a software named SSRD1.0 was developed to analyze the redundancy between a pair of primers. In addition, 90 pairs of SSR primers in which redundancy was detected by the software were selected to perform amplification, electrophoresis and sequencing using the cotton genotypes of TM-1 and 7124. The consistency were 75.0% and 53.8% between the software prediction and sequence level with a threshold of 50% identity and 70% identity, respectively. Verification from the

收稿日期: 2014-08-02
基金项目: 棉花生物学国家重点实验室开放课题(CB2014A02);中国博士后科学基金项目(2012M520473);国家自然科学基金项目(31460359);江苏省自然科学基金项目(BK20130429、BK20131204);江苏省"六大人才高峰"项目(NY-34);公益性事业单位科研资助项目(KY2013056)
作者简介: 王 为(1980-),男,江苏泗阳人,博士,副研究员,从事棉花分子育种与种质创新研究工作。(E-mail) ww462@126.com。叶泗洪与王为具有同等贡献.。
通讯作者: 王长彪,(E-mail)wcbkls@126.com

levels of software prediction, electrophoresis and sequencing revealed that the software and the redundancy prediction method were effective and feasible.

Key words: cotton; molecular marker; SSR; redundancy software; detection; prediction

生物信息学技术革命使得基因组学、蛋白组学等飞速发展。近年来尤其是棉花基因组 A 组[1]、D 组[2]测序完成后,棉花公共数据库呈指数式增长,其中以三大核酸序列数据库 NCBI 的 GenBank 数据库、EBI 维护的 EMBL 数据库、日本国立遗传学研究所的 DDBJ(DNA Data Bank of Japan)数据库为代表。如何合理、高效利用这些数据并应用到基因组学研究中是一个迫切需要解决的问题。其中之一就是利用这些基因组、转录组测序序列信息开发分子标记,如 SSR 标记等,然而引物冗余性(相似性)是标记开发过程面临的一个重要问题,但相关研究报道较少。CMD 数据库(http://www.cottonmarker.org/)上公布的引物存在冗余性(冗余性引物至少占 14.28%),至今也鲜见合适的软件同时分析一对引物的冗余性,造成不同的研究者合成冗余引物、重复扩增,浪费时间和成本,降低了对网上公共序列信息的利用率。所谓冗余性引物是指 2 对引物中的 1 条或 2 条序列相似性较高,导致重复扩增,引物功效降低,也称相似性引物。引物相似有 2 种情况,一种是正式匹配即一对引物和另外一对引物正向序列匹配,另一种是反式匹配就是一对引物的正向和另外一对引物的反向序列匹配[3]。

为系统、集成研究棉花序列资源,开发非冗余的功能标记及进行相关应用研究,并为基因组测序、转录组测序产生的海量信息积累技术资料,本研究利用前期自主开发的能同时分析一对引物和已释放的引物间是否冗余的软件 SSRD1.0[3],对 90 对来自 CMD 网站相互间存在冗余性的 SSR 引物,分别从软件预测、基因型、序列 3 个水平进行检测、验证这一冗余软件和这一冗余性预测方法的有效性。

1 材料与方法

1.1 棉花材料

陆地棉标准系 TM-1 、海岛棉海 7124。

1.2 引物来源

90 对 SSR 引物均选自 CMD 网站(http://www.cottonmarker.org/),已用自主开发的引物冗余性预测软件 SSRD1.0(国家版权局登记号:2011SR001433,软件具体开发方法请见文献[3])检测过都存在引物间冗余性。参照 NCBI 网站(http://www.ncbi.nlm.nih.gov/),引物相似性 Identity 阈值设为 50%、70%来分析(CMD 网站引物相似性阈值为 81%,Blenda 等[4]引物相似性阈值为 90%。本研究为近一步增加筛选力度,减少冗余性引物发生,提高引物利用效果,两对引物之间序列相似性在 70%、50%时就认为存在冗余性)。39 组是 2 对引物间存在冗余,2 组是 3 对间存在冗余性(BNL3512、BNL3547、BNL4060 以及 MGHES-2、MGHES-3、NAU5454),还有 1 组是 6 对间存在冗余性(MUCS141、MUCS152、MUCS410、MUCS422、MUSS563、MUSS598)。

1.3 试剂药品

引物由上海英骏生物技术有限公司合成, *Taq* DNA 聚合酶和 dNTP 均购自河南普金生物技术有限公司。

1.4 DNA 提取、PCR 扩增和电泳检测

采用改良的 CTAB 法[5]提取各材料基因组 DNA。PCR 反应配方:总体积为 10.0 μl,10×Reaction buffer(含 Mg^{2+})1.0 μl,dNTPs(10 mmol/L)0.5 μl,每对引物的正、反向引物(10 μmol/L)各 1.0 μl,*Taq* DNA 聚合酶(2.5 U/μl)0.2 μl,模板 DNA(50 ng/μl)1.0 μl,ddH_2O 5.3 μl。PCR 反应程序为:95 ℃预变性 2 min;94 ℃变性 40 s,57 ℃退火 45 s,72 ℃延伸 60 s,共 30 个循环;72 ℃延伸 7 min;15 ℃保存至结束。PAGE 电泳方法:8%的聚丙烯酰胺凝胶电泳检测,采用 BIO-RAD 公司 PowerPac HCTM 电泳仪,北京六一仪器厂 DYCZ-30 电泳槽装置。电泳缓冲液为 1×TBE,在扩增产物中加入 1.5 μl 溴酚蓝上样缓冲液混均匀,取 1.8 μl 加入点样孔,190 V 恒压电泳 45 min。银染分析:参照张军等[6]、Bassam 等[7]的方法。

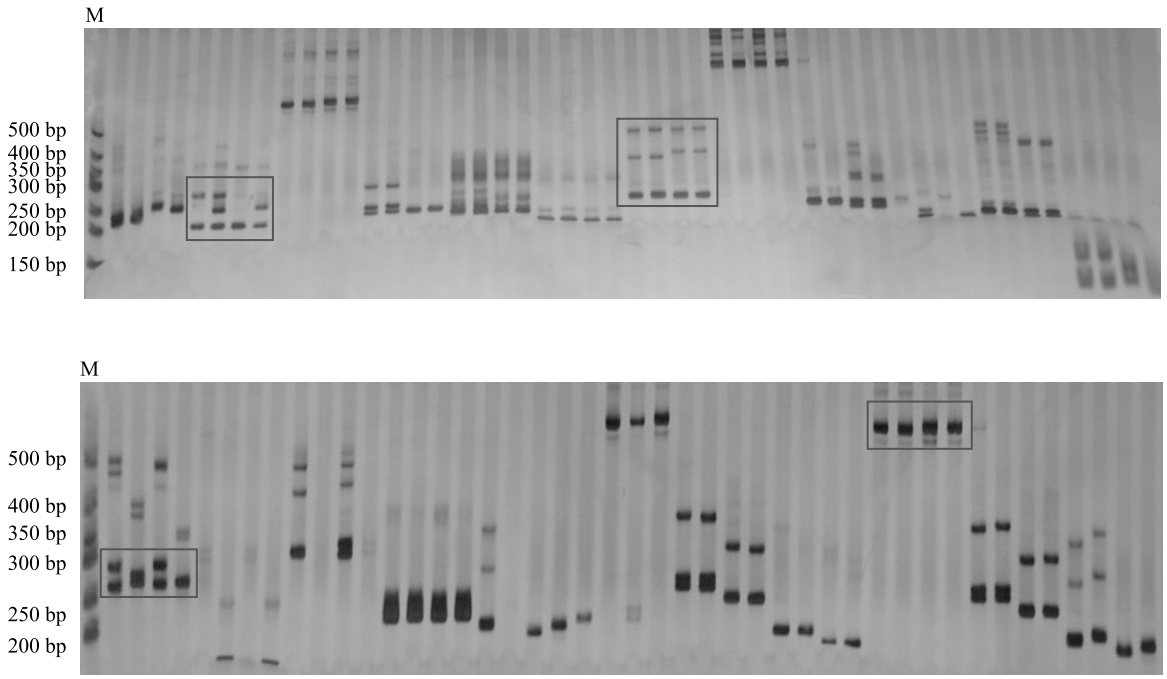
1.5 引物冗余性研究及测序

2 对(或 3 对、6 对)冗余引物同时以一个模板分别扩增(TM-1 或海 7124),看电泳带型是否一致,一致表示软件预测的初步可靠。为了进一步验证,再将电泳条带一致的送去测序(为了更全面评价,

将条带不一致所对应的 PCR 产物也送去测序),根据测序序列的相似性情况判断电泳结果,最后根据电泳和测序结果一起验证软件功效。PCR 反应以 TM-1 为模板,同时以海 7124 为模板重复 1 次,作对照。PCR 反应用 50 μl 体系(10 μl 体系各组分扩大 5 倍),PCR 反应 5 μl 用来检测,45 μl 送去测序,测序时提供了单向引物(每对 10 μl)。测序由上海生物工程有限公司完成。用 Chromas 软件观看序列峰图,用 DNAman 软件进行序列比对。

2　结果与分析

挑选 90 对用 SSRD1.0 软件检测过均存在冗余性的引物,然后分别以模板 TM-1、海 7124 扩增(代表 2 个重复),共 180 个 PCR 反应,部分引物对扩增电泳结果见图 1。图 1 中左方框表示 2 对冗余性引物对 2 个模板扩增的电泳条带不一致(上图为 NAU1243、NAU2455 引物对;下图为 MUSS010、MUSS306 引物对),右方框则表示扩增电泳条带一致(上图为 NAU1195、NAU3982 引物对;下图为 NU1199、NAU1243 引物对)。电泳检测结果有 3 种:条带一样、不一样、未扩增出结果。电泳结果:90 对引物,重复 2 次,5 对没有成功扩增,9.5 对不一致(2 个模板有 1 个模板有结果,算 0.5 对),88.8% 条带一致,即表明软件预测和基因型水平检测有 88.8% 的吻合度。其中 TM-1、海 7124 2 个模板重复性很好,基本一致,仅 NAU1119、MUS558 引物对之间的重复不太好(序列相似性分别为 81.36%、11.74%)。



方框中左边 2 个泳道是以 TM-1 为模板,右边 2 个泳道是以海 7124 为模板。上图左方框中扩增引物对为 NAU1243、NAU2455,右方框为 NAU1195、NAU3982;下图左方框中扩增引物对为 MUSS010、MUSS306,右方框为 NU1199、NAU1243。

图 1　部分冗余引物对 TM-1 及海 7124 的扩增结果

Fig. 1　Amplification of redundant primers using TM-1 and Hai7124 as templates

对电泳带型一致的再进行测序验证,为了更全面评价,将条带不一致所对应的 PCR 产物也进行测序。测序结果:90 对引物扩增 2 个模板,共 180 个 PCR 反应,有 24 个由于浓度低不能成功测序,有 16 个目的条带有 2 条或多条,即重叠现象,无法完成测序,其他正常完成测序,部分序列相似性如表 1 所示。在 2 个重复间结果绝大部分一致,也有不一致的如 NAU1119、MUS558 引物,说明 TM-1 和海 7124 在该位点基因组有差异。在序列水平表明带型一致的 2 条或 3 条或 6 条相似性在 11.47% 至 94.9% 间变化(表 1),参照 NCBI 网站序列 Identity 阈值分 50%、70% 来分析。Identity 阈值定为 50.0%,相似

性 75.0%(39/52);Identity 阈值定为 70.0%,相似性 53.8%(28/52),分别表明序列水平和软件预测有 75.0%、53.8% 的吻合度。以上分别从软件预测、基因型、序列 3 个水平说明这一冗余软件和这一冗余性预测方法是较为有效、可行的,在标记评价、利用及基因组研究中具有重要意义。

表 1 部分冗余引物对的扩增产物序列相似性

Table 1 Sequence similarity of amplified products using some primer pairs

引物名称	序列相似性(%)	
	TM-1	海 7124
Gh356	–	91.18
Gh209		
Gh511	61.79	77.57
Gh320		
MUCS408	89.02	–
MUCS139		
MUCS409	52.07	51.45
MUCS140		
MUSB0273	89.23	91.47
MUSB0269		
MUSS456	81.36	77.40
MUSS255		
NAU1119	81.36	11.74
MUSS558		
NAU1243	95.46	95.29
NAU1199		
NAU2201	83.33	86.62
HAU064		
NAU2455	84.97	87.57
NAU1243		
NAU2946		
MGHES-2	92.90	93.19
MGHES-3		
NAU5454		
MUCS141	63.35	64.33
MUCS152		
MUCS410		
MUCS422		
MUSS563		
MUSS598		
TMB0604	92.75	94.90
TMB0534		
TMB2367		

–表示没有测序结果。

部分 PCR 扩增产物序列 FASTA 格式如下:

>G01031269 (111) 55 sequence exported from chromatogram file

GAATGTATATATATAATAATTAGTCACCAATATTCTAACTATATCTGTACAAACATTATTTATATACTTATATATTTCACGGGACTTCCCTATTTCTTATAAATTGTTATGAATTAATAAAAAAGGATGAAAAACAATTGCTTGATTGCTTGA

>G01031276 (125) 63 sequence exported from chromatogram file

GGACATGTATGCGAGTCTACAAGCCACCAGGACTGAGTTGGAATTGATCTGACCCGAGCGAGACCGTGGCTGATGAGTTCTCTTGCAATGCCAATGAGAATCTAGAACATGAGATTTGAAGAAGAAGAAGAAGAAGAAGAAGAAGAGAGTTGAAGAACATGAAAGCCAAA

3　讨 论

3.1　引物冗余性原因

由于不同的标记开发者利用不同的软件、基于不同的标准可能会导致冗余引物的出现,从而导致重复合成引物、重复扩增、功效减低、费工费力,影响对网上生物数据的利用效率。如发掘 SSR 位点的软件就有 SSRIT、MISASSRmine[8] 等;SSR 的查找标准就有“南农”、“华农”标准之分:“南农” 标准中二至六核苷酸重复次数分别为 9、6、5、4、3,复合型的 SSR 整体长度不小于 24 bp[9-10];而“华农”的查找标准为二至六核苷酸重复次数分别为 7、5、4、4、4[11-12]。

另一个冗余性的原因可能是棉属不同棉种的基因组内涉及许多类型的重复基因,即序列的相似性,如直系同源基因(Orthologous loci),同一基因由于生殖隔离等原因分布在 2 个物种内(如亚洲棉 A 基因组和陆地棉、海岛棉中的 A 亚组);旁系同源基因(Paralogous loci),同一基因由于基因复制(Gene duplication)或基因组倍增(Genome duplication)等分子事件在同一物种内分离成结构不同的功能基因,但有共同起源关系;部分同源位点(Homologous loci)是 A 基因组和 D 基因组的部分同源(Homoelogous)等[3,13]。

3.2　引物冗余性阈值的设定

参照 NCBI 网站,引物相似性 Identity 阈值设为 50%、70%,而 CMD 网站引物相似性阈值为 81%,Blenda 等[4] 引物相似性阈值为 90%。本研究为近一步增加筛选力度、减少冗余性引物发生、提高引物利用效果,将两对引物之间序列相似性在 70%、50% 时就认为存在冗余性。

3.3　冗余性检测软件、评价方法适用范围

第二代分子标记 SSR 曾经广泛地被应用[14-17],由于其诸多优点[18-19],可以推测该类标记还将继续存在一段时间,有一定利用价值。本试验所采用的冗余引物是 SSR 标记,而作为第三代分子标记 SNP[20-22],未来在棉花研究上将越来越受关注[23-24],而其冗余性问题下一步可以考虑用本研究方法去验证和分析。

本研究测序采用 Sanger 测序法,从引物 3′端之后第 1 个碱基开始测序(没有测序的引物序列)。测序时,由于荧光染料的干扰,测序结果在引物 3′端后面的 10 至 30 个可能导致误读,可以根据已知序列信息及测序彩图作出判断。为减少假阳性、提高准确性,设定海岛棉和陆地棉 2 个基因组 DNA 作为重复,针对每一对引物来讲,只有一个基因组 DNA 产生扩增产物时定义为 0.5 对引物,同时对电泳产物全部进行测序。

本研究基于自主开发的冗余性预测软件提出冗余性检测评价方法,用 90 对前人已开发的用该软件预测过相互间存在冗余性的 SSR 引物,以 TM-1、海 7124 基因组为模板分别进行扩增、电泳、测序,结果表明软件预测和基因型水平检测有 88.8% 的吻合度;在 Identity 阈值设分别为 50%、70% 时,测序结果表明序列水平和软件预测有 75%、53.8% 的吻合度。分别从软件预测、基因型、序列 3 个水平进行检测、验证,结果表明这一冗余软件和冗余性预测方法是较为有效、可行的。

参考文献:

[1] LI F G, FAN G Y, WANG K B, et al. Genome sequence of the cultivated cotton *Gossypium arboretum* [J]. Nature Genetics, 2014, 46(6):567-572.

[2] WANG K B, Wang Z W, LI F G, et al. The draft genome of a diploid cotton *Gossypium raimondii* [J]. Nature Genetics, 2012, 44(10):1098-1103.

[3] 王　为,王长彪,刘　方,等. 棉花非冗余性 EST-SSR 新标记的开发及其评价[J]. 作物学报, 2012, 38(8): 1443-1451.

[4] BLENDA A, FANG D D, RAMI J F, et al. A high density consensus genetic map of tetraploid cotton that integrates multiple

component maps through molecular marker redundancy check[J]. Plos One, 2012 , 7(9): e45739.

[5] 宋国立,崔荣霞,王坤波,等. 改良 CTAB 法快速提取棉花 DNA[J]. 棉花学报,1998, 10(5):273-275.

[6] 张 军,武耀廷,郭旺珍,等. 棉花微卫星标记的 PAGE/银染快速检测[J]. 棉花学报,2000,12(5):267-269.

[7] BASSAM B J, CAETANO-ANOLES G, GRESSHOFF P M. Fast and sensitive silver staining of DNA in polyacrylamide gels[J]. Anal Biochem, 1991, 196: 80-83.

[8] 来德勇. 陆地棉花发育 EST 测序分析及其相关基因筛选与微卫星标记开发[D]. 北京:中国农业科学院,2011.

[9] WANG C B, GUO W Z, CAI C P, et al. Characterization, development and exploitation of EST-derived microsatellites in *Gossypium raimondii* Ulbrich[J]. Chin Sci Bull, 2006, 21(3): 316-320.

[10] 吕远大,蔡彩平,王 磊,等. 海岛棉 EST-SSRs 分布特征及新标记的开发与利用[J]. 科学通报, 2010,55 (19):1886-1890.

[11] 余 渝,王志伟,冯常辉,等. 草棉 EST-SSRs 的遗传评价[J]. 作物学报,2008, 34(12):2085-2091.

[12] 俞 渝. 棉花种间群体配子重组率差异、偏分离研究与高密度分子标记遗传图谱构建[D]. 武汉:华中农业大学, 2010.

[13] ZHU H Y, HAN X Y, LÜ J H, et al. Structure, expression differentiation and evolution of duplicated fiber developmental genes in *Gossypium barbadense* and *G. hirsutum*[J]. BMC Plant Biol, 2011, 11: 40.

[14] 王长彪. 与棉纤维发育相关的 EST 生物信息学分析[D]. 南京:南京农业大学, 2007.

[15] 魏利斌,张海洋,郑永战,等. 芝麻 EST-SSR 标记的开发和初步研究[J]. 作物学报, 2008, 34(12): 2077-2084.

[16] 徐照龙,易金鑫,余桂红,等. 藜科 6 种耐盐植物遗传多样性的 EST-SSR 分析[J]. 植物遗传资源学报, 2011, 12(1):113-120.

[17] 张艳欣,林忠旭,李 武,等. 海岛棉 EST-SSR 引物的开发与应用研究[J]. 科学通报, 2007, 52(15):1779-1787.

[18] GUO W Z, CAI C P, WANG C B, et al. A microsatellite-based, gene-rich linkage map reveals genome structure, function, and evolution in *Gossypium*[J]. Genetics, 2007,176:527-541.

[19] 张培培,王夏青,余 杨,等. 首批海岛棉基因组来源的微卫星标记的分离、评价和定位[J]. 作物学报, 2009, 35(6): 1013-1020.

[20] CHUANFU A, SUKUMAR S, JOHNIE N, et al. Transcriptome profiling, sequence characterization and SNP-based chromosomal assignment of the EXPANSIN genes in cotton[J]. Mol Genet Genomics, 2007,278:539-553.

[21] CHUANFU A, SUKUMAR S, JOHNIE N, et al. Cotton (*Gossypium* spp.) R2R3-MYB transcription factors SNP identication, phylogenomic characterization, chromosome localization, and linkage mapping[J]. Theor Appl Genet,2008, 116:1015-1026.

[22] ROBERT L B, DAVID B H, SCOTT M Y, et al. Development and mapping of SNP assays in allotetraploid cotton[J]. Theor Appl Genet,2012, 124:1201-1214.

[23] 吴 玲,付凤玲,李晚忱,等. 利用生物信息学方法进行基于表达序列标签的玉米单核苷酸多态性标记的开发[J]. 核农学报,2010,24(5):968-972.

[24] 唐富福,徐非非,包劲松,等. 全基因组关联分析在水稻遗传育种中的应用[J]. 核农学报,2013,27(5):598-606.

(责任编辑:张震林)