甘薯EST资源的SSR信息分析

黄立飞, 房伯平*, 陈景益, 张雄坚, 罗忠霞
广东省农业科学院作物研究所, 广州 510640

提要: 从NCBI公共数据库下载获得22 371条甘薯EST序列, 去除低质量的和冗余的序列后, 得到总长为5.09×10^3 kb的9 204条唯一序列。从这些序列中搜索到总共436个SSR位点, 平均相距11.68 kb出现一个SSR。这些SSR的出现频率和平均长度分别为4.4%和24.28 bp。在2~6 bp的重复基元中, 六核苷酸重复基元出现频率最高(30.96%), 其次是三核苷酸重复基元(29.59%)和二核苷酸重复基元(24.54%)。出现最多的重复基元是AG/CT (16.28%), 其次是AAG/CTT (11.01%)。
关键词: 甘薯; EST-SSR; 基元

Analysis of SSR Information in EST Resource of Sweet Potato [*Ipomoea batatas* (L.) Lam]

HUANG Li-Fei, FANG Bo-Ping*, CHEN Jing-Yi, ZHANG Xiong-Jian, LUO Zhong-Xia
Crop Research Institute, Guangdong Academy of Agricultural Sciences, Guangzhou 510640, China

Abstract: In this study, 22 371 ESTs of sweet potato (*Ipomoea batatas*) in the database of NCBI were downloaded and some redundants or with low qualities were removed, finally 9 204 unique sequences with total length about 5.09×10^3 kb were obtained. In total, 436 SSRs (4.4%) containing unique sequences were identified. The overall average length of SSRs was 24.28 bp. This was equivalent to 1 SSR per 11.68 kb EST sequence. About 30.96% of the SSRs was hexanucleotides, 29.59% was trinucleotides, 24.54% was dinucleotides, and the remaining 14.91% consisted of tetra- and pentanucleotides. Among the identified SSRs, AG/CT was the most frequent (16.28%) followed by AAG/CTT at 11.01%.
Key words: sweet potato; EST-SSR; motifs

甘薯为同源六倍体, 包括90条染色体, 大约1 050 Mbp DNA, 并且自交不亲和, 这些特性延缓了甘薯分子生物学研究(Arumuganathan和Earle 1991)。近年来包括随机扩增长度多态性DNA (random amplified polymorphic DNA, RAPD)、简单序列重复间区(inter simple sequence repeat, ISSR)和扩增片段长度多态性(amplification fragment length polymorphism, AFLP)等分子标记技术的利用促进甘薯的遗传育种、遗传图谱构建、遗传多态性分析和基因定位等方面研究(Huang和Sun 2000; Gichuki等2003; Cervantes-Flores等2008)。而简单序列重复(simple sequence repeat, SSR)也称作微卫星(microsatellite)标记具有多态性高、多等位性、呈共显性遗传、重复性高、易于用PCR检测和在基因组上分布均匀等特点, 克服了其他标记的缺点, 在甘薯中具有广泛的应用前景(姜树坤等2007; Buteler等1999; Hu等2004)。

但是, 因甘薯基因组开发SSR标记难度大, 开发成本高, 费时费力, 引物数量偏少, 至今报道的甘薯SSR引物仅有100多对, 限制了SSR技术在甘薯中的利用(Buteler等1999; Hu等2004; Ghislan等2005)。采用公共数据库登录的表达序列标签(expressed sequence tag, EST)序列开发EST-SSR是一种相对简便、经济的途径。EST-SSR位点多位于转录区, 非常保守, 是真正与性状连锁的标记; 同时, EST-SSR具有通用性, 可进行比较基因组学研究(金基强等2006)。近年来, 功能基因组学的发展促进了EST测序工作的开展, 使得公共数据库中的EST数量迅速增长。这些大量的而可以共享的EST序列为分子标记的开发和研究提供了丰富的序列资源。这些序列资源已相继在马铃薯、水稻、

收稿 2008-09-26 修定 2008-11-11
资助 国家高技术研究发展计划("863" 计划) (2006AA100107)、国家科技基础条件平台项目(2005DKA2100205)、广东省自然科学基金项目(8151064001000033)、广东省科技计划项目(2008B020200003)。
* 通讯作者(E-mail: bpfang01@163.com; Tel: 020-85514242)。

大麦、小麦、木薯、油菜、人参、茶树和东方牡蛎等多种植物开发了EST-SSR标记(Holton等2002; Thiel等2003; Feingold等2005; Peng等2005; Rota等2005; Wang和Guo 2007; 金基强等2006; 李小白等2007; 杨成君和王军2008; 彭丁文等2008)。

截至2008年9月为止, 在NCBI数据库中已登录了22 371条甘薯EST, 但是目前还没有用这些EST大规模开发SSR的报道。本文对现有甘薯EST中的SSR信息进行了分析, 以了解甘薯EST-SSR的发生频率和特点, 为进一步建立EST-SSR标记, 进行分子生物学研究建立基础。

材料与方法

甘薯[*Ipomoea batatas* (L.) Lam] EST来自NCBI (美国国家生物技术信息中心)数据库(http://www.ncbi.nlm.nih.gov/sites/entrez/), 共计22 371条。采用EST-trimmer软件(http://pgrc.ipk-gatersleben.de/misa/download/est-trimmer.pl)除去5'端或3'端50 bp的ploy T或ploy A以及那些长度小于100 bp的EST序列; 对于长度大于700 bp的EST则保留其5'端700 bp。

预处理后的EST, 通过软件CAP3 (http://seq.cs.iastate.edu/capdownload.html)进行片段重叠群分析和聚类, 拼接时设定的初始装配参数为默认值(Huang和Madan 1999)。

用MISA软件(http://pgrc.ipk-gatersleben.de/misa/)对聚类后的EST进行SSR搜索。筛选标准为: 搜索的长度标准为二核苷酸、三核苷酸、四核苷酸、五核苷酸、六核苷酸的最少重复次数10、6、5、4、3次以上。同时, 也筛选中间被少数碱基(间隔小于或等于100 bp)打断的不完全重复的SSR。

用SSR出现频率和SSR平均分布距离来描述EST-SSR, 同时定义出现频率前2位的重复基元作为优势重复基元。计算公式为: (1) SSR出现频率, $f_c=c/n\times100\%$, c为搜索到的SSR数量, n为无冗余EST数量; (2) SSR平均分布距离, $f_N=N/c$, N为无冗余EST数量的总碱基数。

实验结果

1 EST序列的获得与组装

截至2008年9月1日为止, 在NCBI数据库中已登录了22 371条甘薯的EST序列, 这些EST序列分别来源于11个不同类型的甘薯cDNA文库。构建这些cDNA文库的材料以块根为主, 其EST的数量最多, 达到了21 979条, 而来自叶片的有178条, 试管苗214条。

全部EST序列经过EST-trimmer软件处理后, 用CAP3软件进行组装。组装后产生了9 204个唯一序列, 序列总长为5 092.82 kb, 其中包括3 059个重叠群和6 145个单一序列。

2 EST-SSR的出现频率

在搜索的9 204个唯一序列中, 总共发现了分布于407个唯一序列的436个SSR位点, 占全部唯一序列的4.4%。在含有SSR的407条唯一序列中没有发现包含3个以上SSR位点的唯一序列, 其中包含1个SSR位点的有378条, 包含2个SSR位点有29条。从分布情况来看, 甘薯EST-SSR平均每11.68 kb就出现1个SSR, 但不同重复类型间差异很大(表1)。

3 甘薯**EST-SSR**的特性

在搜索出的甘薯EST-SSR中, 总共观察到134种重复基元(表2)。其中二、三、四、五、六核苷酸重复基元分别有3、10、12、24、85种。从出现的频率来看, 二核苷酸重复基元中的优势重复基元为AG/CT和AT/AT, 出现频率分别为

表1 SSR在甘薯唯一序列的出现频率

Table 1 Occurrence frequency of SSRs in a set of sweet potato unique sequences

重复类型	SSR数目	占全部SSR比例/%	出现频率/%	平均分布距离/kb
二核苷酸	107	24.54	1.16	47.60
三核苷酸	129	29.59	1.40	39.48
四核苷酸	27	6.19	0.29	188.62
五核苷酸	38	8.71	0.41	134.02
六核苷酸	135	30.96	1.47	37.72
总计	436	100.00	4.74	11.68

表2 重复基元类型的优势重复基元的比较分析

Table 2 Comparative analysis of the predominant motifs in different motif types

重复基元	重复基元数目	优势重复基元	SSR数量	出现频率/%
二核苷酸	3	AG/CT	71	16.28
		AT/AT	33	7.57
三核苷酸	10	AAG/CTT	48	11.01
		AAT/ATT	24	5.50
四核苷酸	12	AAAG/CTTT	5	1.15
		AAAT/ATTT	7	1.61
五核苷酸	24	AAAAC/GTTTT	5	1.15
		AATAT/ATATT	4	0.92
六核苷酸	85	AAAAAG/CTTTTT	7	1.61
		AAAAAT/ATTTTT	7	1.61
总计	134	12	211	48.39

16.28%和7.57%; 三核苷酸重复基元中的优势重复基元为AAG/CTT和AAT/ATT, 出现频率分别为11.01%和5.50%; 四核苷酸重复基元中的优势重复基元AAAG/CTTT和AAAT/ATTT, 出现频率分别为1.15%和1.61%; 五核苷酸重复基元中的优势重复基元为AAAAC/GTTTT和AATAT/ATATT, 出现频率分别为1.15%和0.92%; 而在六核苷酸重复基元中的优势重复基元为AAAAAG/CTTTTT和AAAAAT/ATTTTT, 出现频率均为1.61%。由上述结果说明, 不同重复基元中出现频率最高的是AG/CT, 其次是AAG/CTT。10个优势重复基元占到整个SSR出现频率的48.39%, 其中二、三核苷酸的优势重复基元占到整个优势重复基元的83.41%。

从图1可以看出, 大部分甘薯EST-SSR的位点长度都集中在18~20 bp范围内, 其次是分布在

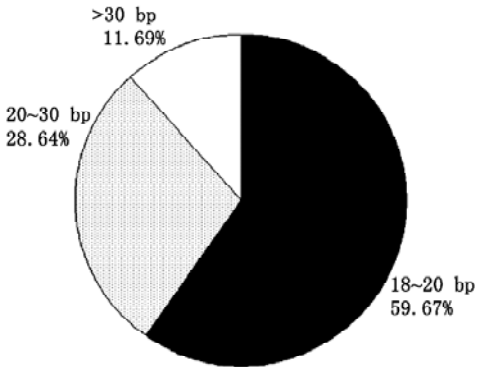

图1 甘薯EST-SSR位点长度的分布

Fig.1 The length distribution of EST-SSR loci in the sweet potato

20~30 bp范围内, 长度大于30 bp的位点只占很少一部分。具体而言, 甘薯EST-SSR位点长度分布在18~20 bp范围内的约有250个, 占到整个EST-SSR的59.67%; 在20~30 bp范围内有120个, 占到整个EST-SSR的28.64%; 而长度大于30 bp的则不到50个, 百分率仅为11.69%。EST-SSR位点的平均长度为24.28 bp。

讨 论

本文对22 371条甘薯EST组装后进行了搜索, 发现436个SSR, 其出现频率为4.4%, 平均分布距离为11.68 kb。从目前的报道来看, 这个平均分布距离高于茶树(2.61 kb) (金基强等2006)、油菜(4.34 kb) (李小白等2007)、人参(5.7 kb) (杨成君和王军2008)、木薯(6.02 kb) (彭丁文等2008), 低于水稻(11.81 kb)、小麦(17.42 kb)、大豆(23.80 kb)和玉米(28.32 kb) (Gao等2003)等植物, 表明甘薯EST-SSR丰富度处于中等水平。

目前, 由于EST-SSR的频率一般是通过搜索数据库中的EST序列而估算出的, 各个研究者在不同的植物上对搜索SSR重复类型和长度等标准的不同SSR筛选标准不同, 导致EST-SSR的出现频率也会有所变化, 例如金基强等(2006)对茶树的EST-SSR筛选要求二核苷酸重复最少为7次重复, 李小白等(2007)对油菜的筛选标准为二核苷酸重复的次数在6次或6次以上, 而本文则要求二核苷酸至少要重复10次或10次以上, 这样严格的筛选标准, 必然导致SSR出现频率降低。另一方面, 也与所分析数据的多少以及对数据的分析程序密切相

关。本文对数据库中下载的所有EST序列进行了预处理, 并且去冗余、聚类和组装, 这样得到的9 204个唯一序列, 搜索到436个SSR位点, 而如果不对这些EST序列进行处理, 按照相同的搜索标准竟然能搜索到1 151个SSR位点。

大多数植物的EST-SSR主要为二、三核苷酸重复类型, 但出现频率最高的重复基元类型则有所差异(Wang和Guo 2007; 彭丁文等2008; 杨成君和王军2008)。本文结果显示, 甘薯EST-SSR主要集中在二、三、六核苷酸, 其中六核苷酸比例最大, 占总EST-SSR的30.96%, 其次为三核苷酸重复, 再次为二核苷酸重复, 这是已报道的其他物种所没有的。即使在相同的搜索标准下, Gao等(2003)曾报道小麦、水稻、玉米和大豆都以三核苷酸重复为主, 没有出现以六核苷酸重复为主的特征, 这可能是甘薯与这些物种间真实的差异。Hu等(2004)分析4 153条甘薯EST的结果也表明, 二、三、六核苷酸重复的SSR最多, 但三核苷酸重复比例最大, 其次为二核苷酸和六核苷酸, 这些差异可能与其搜索标准、较少数量的EST序列以及序列的冗余性密切相关。本文中, 在二核苷酸重复基元中又以AG/CT占有绝对优势, 这与报道的油菜(李小白等2007)、木薯(彭丁文等2008)、小麦、玉米、高粱和水稻(Kantety等2002)等作物情况相同, 而三核苷酸重复则以AAG/CTT为主, 与大豆、水稻(Gao等2003)、木薯(彭丁文等2008)类似。范三红等(2003)认为这些占优势的重复基元可能与其编码相应蛋白质的使用频率较高有关。

Temnykh等(2001)认为, 当SSR长度在20及20 bp以上时, 在不同品种间显示出较高的多态性, 长度低于20 bp多态性就会降低。而在本文中EST-SSR搜索就严格限定了EST-SSR的长度, 最低限度为18 bp, 这样得到的EST-SSR全部为多态潜能高的SSR。Hu等(2004)的研究表明, 甘薯二、三、六核苷酸重复的SSR的多态性较高, 而在本文中这3类重复基元占的比重特别大, 所以作者所获得的EST-SSR大多数是属于多态性潜能高的SSR。因此, 本文搜索出的甘薯EST-SSR应具有较高的利用价值。

从整个分析结果看, 甘薯EST-SSR出现频率较高, 而且类型十分丰富。因此, 本文的结果为新的EST-SSR标记的开发、基因定位、遗传图谱绘制和功能推测, 对甘薯分子标记的理论研究和育种利用具有一定的参考意义。

参考文献

范三红, 郭蔼光, 单丽伟, 胡小平(2003). 拟南芥基因密码子偏爱性分析. 生物化学与生物物理进展, 30: 221~225

姜树坤, 王政海, 钟鸣, 张丽, 徐正进, 刘少霞(2007). 辽宁省近15年的部分水稻主栽品种的简单重复序列(SSR)多态性分析. 植物生理学通讯, 43 (1): 69~72

金基强, 李素芳, 龚晓春, 卢美贞, 姚艳玲, 忻雅, 崔海瑞(2006). 茶树EST资源中的SSR信息分析. 科技通报, 22 (4): 471~476

李小白, 张明龙, 崔海瑞(2007). 油菜EST资源的SSR信息分析. 中国油料作物学报, 29 (1): 20~25

彭丁文, 郑柳城, 朱宏波(2008). 木薯EST资源的SSR信息分析. 中国农学通报, 24 (2): 433~436

杨成君, 王军(2008). 人参EST资源的SSR信息分析. 植物生理学通讯, 44 (1): 69~73

Arumuganathan K, Earle ED (1991). Nuclear DNA content of some important plant species. Plant Mol Biol Rep, 9: 208~218

Buteler MI, Jarret RL, LaBonte DR (1999). Sequence characterization of microsatellites in diploid and polyploid *Ipomoea*. Theor Appl Genet, 99: 123~132

Cervantes-Flores JC, Yencho GC, Kriegner A, Pecota KV, Faulk MA, Mwanga ROM, Sosinski BR (2008). Development of a genetic linkage map and identification of homologous linkage groups in sweetpotato using multiple-dose AFLP markers. Mol Breed, 21: 511~532

Feingold S, Lloyd J, Norero N, Bonierbale M, Lorenzen J (2005). Mapping and characterization of new EST-derived microsatellites for potato (*Solanum tuberosum* L.). Theor Appl Genet, 111: 456~466

Gao LF, Tang JF, Li HW, Jia JZ (2003). Analysis of microsatellites in major crops assessed by computational and experimental approaches. Mol Breed, 12 (3): 245~261

Ghislan M, Gruneberg W, Benavides J, Mwanga R (2005). Application of molecular markers for gene pool division and heterosis estimation under drought stress conditions in sweet potato. In: Proceedings of GCP 2005 Annual Research Meeting: 2005 Competitive and Commissioned Project Mid-Year Reports, 59~63

Gichuki ST, Berenyi M, Zhang DP, Hermann M, Schmidt J, Glossl J, Burg K (2003). Genetic diversity in sweetpotato [*Ipomoea batatas* (L.) Lam.] in relationship to geographic sources as assessed with RAPD markers. Genet Resour Crop Evol, 50: 429~437

Holton TA, Christopher JT, McClure L, Harker N, Henry RJ (2002). Identification and mapping of polymorphic SSR markers from expressed gene sequences of barley and wheat.

Mol Breed, 9: 63~71

Hu JJ, Nakatan M, Mizuno K, Fujimura T (2004). Development and characterization of microsatellite markers in sweetpotato. Breed Sci, 54: 177~188

Huang JC, Sun M (2000). Genetic diversity and relationships of sweetpotato and its wild relatives in *Ipomoea* series *Batatas* (Convolvulaceace) as revealed by inter-simple sequence repeat (ISSR) and restriction analysis of chloroplast DNA. Theor Appl Genet, 100 (7): 1050~1060

Huang XQ, Madan A (1999). CAP3: a DNA sequence assembly program. Genome Res, 9: 868~877

Kantety RV, Rota ML, Matthews DE, Sorrells ME (2002). Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum and wheat. Plant Mol Biol, 48: 501~510

Peng JH, Nore L, Lapitan V (2005). Characterization of EST-derived microsatellites in the wheat genome and development of eSSR markers. Funct Integr Genomics, 5: 80~96

Rota LR, Kantety RV, Yu JK, Sorrells ME (2005). Nonrandom distribution and frequencies of genomic and EST-derived microsatellite markers in rice, wheat, and barley. BMC Genomics, 6: 23~34

Temnykh S, DeClerck G, Lukashova A, Lipovich L, Cartinhour S, McCouch S (2001). Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): frequency, length variation, transposon associations, and genetic marker potential. Genome Res, 11: 1441~1452

Thiel T, Michalek W, Varshney RK, Graner A (2003). Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). Theor Appl Genet, 106: 411~422

Wang YP, Guo XM (2007). Development and characterization of EST-SSR markers in the eastern oyster *Crassostrea virginica*. Marine Biotechnol, 9: 500~511