AutoSSR:大批量、高效开发和分析SSR基元的软件

王长彪[1,2]，郭旺珍[1*]，张天真[1]，李燕娥[2]，刘惠民[3]
(1.南京农业大学作物遗传与种质创新国家重点实验室,南京 210095;
2.山西省农业科学院棉花研究所,运城 044000;3.山西省农业科学研究院,太原 030006)

摘要:为从大量的序列中快速发掘和利用简单序列重复(微卫星,SSRs),开发了一个软件工具——AutoSSR。该软件能够从一个包括大量FASTA格式的序列文件中发掘SSR,并分析不同SSRs类型的分布状况。与已开发的类似软件相比,该软件可以不受序列数量和不同系统平台的限制,高效、快速完成SSR信息的搜索和分析。利用该软件对GenBank上释放的来源于棉花、苹果和水稻的DNA序列进行了SSR发掘,得到了36000个SSRs。其中,对来源于棉花的SSRs信息,进一步利用Primer 3软件开发了位点特异性引物,为棉花基因组分析奠定了基础。

关键词:简单序列重复;表达序列标签;发掘;软件
中图分类号:S562 **文献标识码**:A
文章编号:1002-7807(2009)03-0213-05

AutoSSR: An Improved Automatic Software for SSR Analysis from Large-scale EST Sequences

WANG Chang-biao[1,2], GUO Wang-zhen[1*], ZHANG Tian-zhen[1], LI Yan-e[2], LIU Hui-min[3]
(1. *State Key Laboratory of Crop Genetics & Germplasm Enhancement, Cotton Research Institute, Nanjing Agricultural University, Nanjing* 210095, *China*; 2. *Cotton Research Institute, Shanxi Academy of Agricultural Sciences, Yuncheng, Shanxi* 044000, *China*; 3. *Shanxi Academy of Agricultural Sciences, Taiyuan* 030006, *China*)

Abstract: Microsatellites, also known as simple sequence repeats (SSRs), are repeated small sequence motifs that are highly polymorphic and abundant in the genomes of eukaryotes. We have created a software tool, AutoSSR, for quickly detecting simple sequence repeats from a large number of sequences. The software can detect SSRs from one file, including large number of FASTA-formatted sequences, and analyze the distributions of different SSR types. Importantly, this tool is not restricted to sequence numbers and different system platforms with high efficiency and vast amounts of information. We have applied this tool for the discovery of SSRs within DNA sequences publicly released in GenBank in *Gossypium*, *Malus Pumila* and *Oryza sativa* L., in which we detected over 36000 SSRs. Of the identified SSRs, we parsed information *Gossypium* to Primer 3 software for designing locus-specific primers and applications in genomics analysis.

Key words: simple sequence repeats; expressed sequence tags; distribution; software
CLC:S562 **Document Code**:A
Article ID:1002-7807(2009)03-0213-05

Received Date: 2008-09-11 **Author**:WANG Chang-biao (1975-), assistant researcher, wcbksl@126.com;
* corresponding author, moelab@njau.edu.cn

Sponsors:Grants Visiting Scholar Project from the National Lab in Nanjing Agricultural University (ZW2007005); the National High-tech Program (2006AA10Z111); the Program for New Century Excellent Talents in University (NCET-04-0500); the 111 Program for Ministry of Education (B08025)

Microsatellites, otherwise known as simple sequence repeats (SSRs), are among the most useful genetic markers in biology. SSRs are tandem repeats of short (1 to 6 bp) DNA sequences, and exist throughout the entire genomes of eukaryotic organisms in both non-coding and coding regions. The distinguishing features of SSR loci include their high information content, co-dominant inheritance patterns, consistent distribution along chromosomes, reproducibility and locus specificity[1-3]. Furthermore, SSRs demonstrate a high degree of transferability among related species, making them excellent markers for comparative genetic and genomic analyses[4].

With the advantage of SSR hyper-variability among related organisms, the informative and excellent markers have been widely applied to resolving many research areas including high-density genetic mapping[5], molecular tagging of genes[6], genotype identification[7], analysis of genetic diversity[8-9] and marker assisted selection (MAS) breeding[10]. Thus, SSRs have become one of the most important molecular marker types for use in the analysis of many organisms at the genome level[11-13].

To aid the development of new potential microsatellite markers from thousands of sequences, researchers require an easy-to-use and efficient software tool. The most well-known programs that mine microsatellites from a DNA sequence are SPUTNIK (http://abajian. net/sputnik/), SSRIT[14] and TROLL[15]. However, they are either restricted to small datasets, not available for windows or possess limited efficiency in marker identification capabilities. Furthermore, they do not show the distribution of SSRs. In order to overcome the above shortcomings, herein we have created a software tool, AutoSSR, for discovering SSRs with high efficiency within bulk FASTA-formatted sequence data, and displaying the various SSR motif distribution patterns. We believe that our tool will aid researchers through SSR identification and analysis to streamline their current marker identification efficiency.

1 Materials and methods

1.1 Program options

AutoSSR is a stand-alone tool that may be run on the command line. The input is multiple file in FASTA format containing the sequence (s). In order to specify the search criteria, an additional file, AutoSSR. ini, containing the microsatellite search parameters is required, which has the following structure: following a text string beginning with 'def', pairs of numbers are expected, whereas defines the first number as the motif size and the second number as the lower threshold of repeats for the specific motif. The default minimum repeat number value is that of a dinucleotide such as 9, trinucleotide 6, tetranucleotide 5, pentanucleotide 4, and hexanucleotide 3. Following a text string beginning with 'int' a single number defines the maximal number of bases between two adjacent microsatellites to specify the compound microsatellite type. The default value is 3.

The users can change these parameters for different SSR detection methods according to their study demand.

1.2 Implementation

AutoSSR can run in different platforms such as DOS, Windows, and Linux/Unix. The program run requires the following steps: (1) Change the parameter of motif size, number of repeat for specific motif and the number of bases between two adjacent microsatellites in "AutoSSR. ini" file. (2) Save the sequence file and the program in the same folder. (3) Run the program in different platforms. DOS: input the program name(AutoSSR. exe); Windows: double click the program file (AutoSSR. exe); Linux/Unix: perl AutoSSR. pl.

1.3 Program output

AutoSSR uses a recursive algorithm to search for repeated patterns of nucleotides of lengths ranging between 2 and 6 bp. The motif

shifting criterion has been considered in this program[15]. All SSR types are classed into two subtypes, perfect and compound SSRs[16]. The output of the AutoSSR program has two tab-delimited text files and one PNG file. One of the text files, named sequence-filename. ssr, is the SSR identification result detected from bulk sequence data, while the other, named sequence-filename. statistics, holds the statistics of different motif distributions.

File "filename. ssr" includes sequence ID (ID), SSR number (SSR nr.), SSR type, SSR, SSR size, start, end and sequence length (seq_length). SSR type has two subtype options, called perfect (p) and compound (c) SSRs. The perfect SSRs are subdivided into 5 categories, p2, p3, p4, p5 and p6, according to their motif lengths. The results can be parsed to Primer3 software[17] to aid in developing the locus specific SSR primer pairs. Two additional SSR categories, compound (c) and compound overlap SSRs (co), are sorted within the compound SSR subtype. The filename. statistics file includes the total number of sequences examined, total size of examined sequences (bp, base pairs), total number of identified SSRs, number of SSR-containing sequences, number of sequences containing more than 1 SSR, number of SSRs present in compound formation, distribution to different repeat type classes, frequency of identified SSR motifs and frequency of classified repeat types (considering sequence complementary). The PNG is an image format file of different SSR distribution (default more than 100 SSR). An example is shown in the screen shot illustrated in Fig. 1.

A

ID	SSR nr.	SSR type	SSR	size	start	end	seq_length
AC202832	1	p2	(TA)12	24	7085	7108	94452
AC202832	2	p6	(TGTGAT)3	18	58873	58890	94452
AC202832	3	p2	(AT)10	20	65851	65870	94452
AC202832	4	p3	(TAT)7	21	94265	94285	94452
AC202831	1	p2	(CT)21	42	3338	3379	97917
AC202831	2	p2	(AT)23	46	6192	6237	97917
AC202001	3	p4	(TTAT)5	20	15996	16015	97917
AC202831	4	p6	(TGATTT)3	18	16808	16825	97917
AC202831	5	p6	(TTTAAT)3	18	27139	27156	97917
AC202831	6	p5	(AAAAT)4	20	54403	54422	97917
AC202831	7	p6	(ATTTTT)5	30	71768	71797	97917
AC202830	1	p2	(AT)11	22	996	1017	101659
AC202830	2	p5	(ATCGA)5	25	12327	12351	101659
AC202830	3	p6	(TATTCA)3	18	20941	20958	101659
AC202830	4	p2	(AT)10	20	30174	30193	101659
AC202830	5	p2	(TA)22	44	32245	32288	101659
AC202830	6	p6	(ATATAC)3	18	44616	44633	101659
AC202830	7	p2	(AT)16	32	47049	47080	101659

B

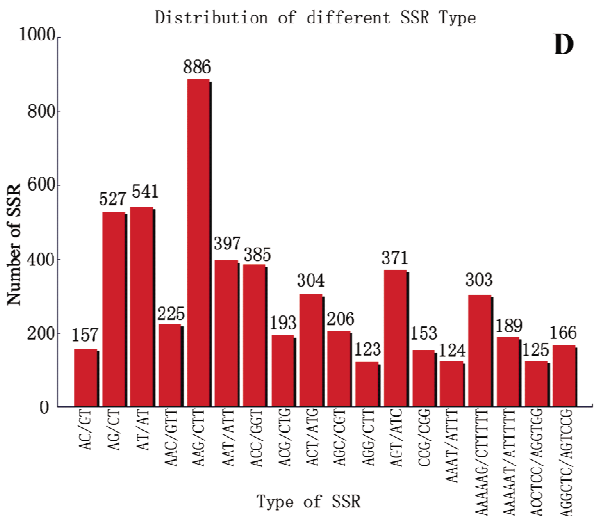
```
RESULTS OF MICROSATELLITE SEARCH
--------------------------------

Total number of sequences examined:              280383
Total size of examined sequences (bp):           182954914
Total number of identified SSRs:                 9072
Number of SSR containing sequences:              8564
Number of sequences containing more than 1 SSR:  461
Number of SSRs present in compound formation:    93

Distribution to different repeat type classes
---------------------------------------------

Motif size      Number of SSRs
2               1227
3               3243
4               440
5               496
6               3666
```

C

Frequency of classified repeat types (considering sequence complementary)

Repeats	3	4	5	6	7	8	9	10	11	12
AC/GT	-	-	-	-	-	-	42	43	50	6
AG/CT	-	-	-	-	-	-	122	82	47	53
AT/AT	-	-	-	-	-	-	161	121	74	33
CG/CG	-	-	-	-	-	-	2			
AAC/GTT	-	-	-	133	78	11	1	1	1	
AAG/CTT	-	-	-	437	211	102	56	28	12	19
AAT/ATT	-	-	-	189	73	24	19	11	11	12
ACC/GGT	-	-	-	226	97	27	7	7	10	5
ACG/CTG	-	-	-	140	38	9	2	2	1	
ACT/ATG	-	-	-	160	82	28	17	6	3	2
AGC/CGT	-	-	-	92	60	26	19	4	1	1
AGG/CCT	-	-	-	63	44	11	4		1	
AGT/ATC	-	-	-	198	84	28	28	9	8	3
CCG/CGG	-	-	-	78	60	6	8	1		
AAAC/GTTT	-	-	-	38	10		1			
AAAG/CTTT	-	-	-	40	25			1		
AAAT/ATTT	-	-	-	88	35	1				
AACC/GGTT	-	-	-	2	1					
AACG/CTTG	-	-	-		1					

D



Three result files are filename. ssr (A), filename. statistics (B and C) and SSR_bar. png(D).

Fig. 1　An example for running AutoSSR software

2 Results and Discussion

The AutoSSR tool was executed on Intel Pentium Ⅳ 64-bit 400 MHz with 1 Gb RAM, running in Windows XP. A FASTA file containing 280383 *Gossypium* spp. (Cotton) EST (expressed sequence tag) sequences (183 Mb, millions of base pairs) was processed in 7 min and produced 9072 SSRs. Another FASTA file containing 257996 *Malus Pumila* Mill. (Apple) EST (expressed sequence tag) sequences (123 Mb) was processed in 4 min and detected 12737 SSRs. 62827 *Oryza sativa* L. (Rice) genomic sequences (103 Mb) in a FASTA file were processed in 3 min and mined 14663 SSRs. These results demonstrate that the program execution time has a linearity correlation with the file size.

The most-well known programs, SPUTNIK, SSRIT, and TROLL can also mine SSRs from a FASTA formatted-file including multiple sequences. We have tested their execution time under the same situation. Table 1 provides data on an execution time comparison among SPUTNIK, SSRIT, TROLL and AutoSSR. These data were compiled using the same level of optimization. The input DNA sequence was 183 Mb (EST sequences of *Gossypium* spp.). The search was conducted for SSRs of lengths greater than 18 bp, and all motifs 2 to 6 bp long. The total number of motifs searched was 9072. The execution times of the different software programs was TROLL> SPUTNIK> AutoSSR > SSRIT. Even though SSRIT is time-saved among all tested programs, it does not consider motif shifting; an oversight that can leads to SSRs redundancy.

Table 1 Execution time for AutoSSR, SPUTNIK, SSRIT and TROLL software packages

Programs	Time
AutoSSR	7 min
SPUTNIK	16 min
SSRIT	5 min
TROLL	20 min

Notes: The test load is the *Gossypium* spp. (cotton) EST sequences (183 Mb); Platform: Pentium Ⅳ 400 MHz, 1 Gb RAM, Windows XP; Average time for 20 executions, negligible standard deviation.

AutoSSR performed better overall than the other available tools. Apart from speed, the advantages of using AutoSSR include a convenient statistics output format that enables easy analyis of SSR distributions in the input sequence data, and the ability to parse data to Primer3 for locus-specific primer design. In our lab, 9072 SSRs mined from *Gossypium* EST sequences using AutoSSR software have been parsed to Primer3, and 3250 PCR amplification primers were developed and publicly released at the CMD web page (http://www.genome.clemson.edu/projects/cotton).

The shortcoming of AutoSSR software is that the command line interface may seem unfriendly to some users. However, this feature can be mastered easily. The software package can be downloaded from State Key Laboratory of Crop Genetics & Germplasm Enhancement in Nanjing Agricultural University (http://www.njaustatelab.net/ShowCenter/download.jsp).

Acknowledgement

We are grateful to Dr. Zhang Da-yong and Dr. Wang Hai-hai for beneficial discussions.

References:

[1] KASHI Y, King D, Soller M. Simple sequence repeats as a source of quantitative genetic variation[J]. Trends Genet, 1997, 13:74-78.

[2] RöDER M S, Korzun V, Wendehake K, et al. A microsatellite map of wheat[J]. Genetics, 1998, 149: 2007-2023.

[3] RöDER M S, Korzun V, Gill B S, et al. The physical mapping of microsatellite markers in wheat[J]. Theor Appl Genet, 1998, 41:278-283.

[4] GUO W, Wang W, Zhou B, et al. Cross-species transferability of *G. arboreum*-derived EST-SSRs in the diploid species of *Gossypium*[J]. Theor Appl Genet, 2006, 112:1573-1581.

[5] GUPTA K, Balyan S, Edwards J, et al. Genetic mapping of 66 new microsatellite (SSR) loci in bread wheat[J]. Theor Appl Genet, 2002, 105:413-422.

[6] MOLNAR S J, Rai S, Charette M, et al. Simple sequence repeat (SSR) markers linked to *E*1, *E*3, *E*4,

and E7 maturity genes in soybean[J]. Genome, 2003, 46:1024-1036.

[7] ELLWOOD S R, D'Souza N K, Kamphuis L G, et al. SSR analysis of the Medicago truncatula SARDI core collection reveals substantial diversity and unusual genotype dispersal throughout the Mediterranean basin[J]. Theor Appl Gene, 2006, 112:977-983.

[8] BARKLEY N A, Roose M L, Krueger R R, et al. Assessing genetic diversity and population structure in a citrus germplasm collection utilizing simple sequence repeat markers (SSRs) [J]. Theor Appl Genet, 2006, 112:1519-1531.

[9] HOKANSON S C, Szewc-McFadden A K, Lamboy W F, et al. Microsatellite (SSR) markers reveal genetic identities, genetic diversity and relationships in a Malus × domestica borkh. core subset collection [J]. Theor Appl Genet, 1998, 97:671-683.

[10] FU Y B, Peterson G W, Yu J K, et al. Impact of plant breeding on genetic diversity of the Canadian hard red spring wheat germplasm as revealed by EST-derived SSR markers[J]. Theor Appl Genet, 2006, 112: 1239-1247.

[11] GUPTA P K, Varshney R K. The development and use of microsatellite markers for genetic analysis and plant breeding with emphasis on bread wheat[J]. Theor Appl Genet, 2000, 113:163-185.

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

[13] TAUTZ D. Hypervariability of simple sequences as a general source for polymorphic DNA markers[J]. Nucleic Acids Research, 1989, 17:6463-6471.

[14] TEMNYKB S, DeClerck G, Lukashova A, et al. Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): frequency, length variation, transposon associations, and genetic marker potential[J]. Genome Research, 2001, 11: 1441-1452.

[15] CASTELO A T, Martins W, Gao G R. TROLL—tandem repeat occurrence locator[J]. Bioinformatics, 2002, 18: 634-636.

[16] WEBER J L. Informativeness of human (dC-dA)n. (dG-dT)n polymorphisms[J]. Genomics, 1990, 7: 524-530.

[17] ROZEN S, Skaletsky H. Primer 3 on the WWW for general users and for biologist programmers[J]. Methods in Molecular Biology, 2000, 132:365-386.