APPLICATION NOTE

Genomics & Informatics
Vol. 9(4) 189-193, December 2011

http://dx.doi.org/10.5808/GI.2011.9.4.189
pISSN 1598-866X eISSN 2234-0742

SSR-Primer Generator: A Tool for Finding Simple Sequence Repeats and Designing SSR-Primers

Chang Pyo Hong[1,2], Su Ryun Choi[1] and Yong Pyo Lim[1]*

[1]Department of Horticulture, College of Agriculture and Life Science, Chungnam National University, Daejeon 305-764, Korea, [2]Division of Molecular & Life Sciences, Pohang University of Science & Technology (POSTECH), Pohang 790-784, Korea

Abstract

Simple sequence repeats (SSRs) are ubiquitous short tandem duplications found within eukaryotic genomes. Their length variability and abundance throughout the genome has led them to be widely used as molecular markers for crop-breeding programs, facilitating the use of marker-assisted selection as well as estimation of genetic population structure. Here, we report a software application, "*SSR-Primer Generator*" for SSR discovery, SSR-primer design, and homology-based search of *in silico* amplicons from a DNA sequence dataset. On submission of multiple FASTA-format DNA sequences, those analyses are batch processed in a Java runtime environment (JRE) platform, in a pipeline, and the resulting data are visualized in HTML tabular format. This application will be a useful tool for reducing the time and costs associated with the development and application of SSR markers.

Keywords: EST clustering and assembly, homology-based search, *in silico* amplicon, PCR, simple sequence repeat (SSR), SSR-primer

Availability: SSR-Primer Generator is available at http://168.188.15.158:8080/ssrpg/. A source code for SSR-Primer Generator is also available from the authors upon request (yplim@cnu.ac.kr).

Introduction

Simple sequence repeats (SSRs) are DNA elements found ubiquitously in eukaryotic genomes; in SSRs, 1-6 nucleotides are tandemly repeated. SSR loci have a relatively high mutation rate, resulting in variation in the number of repeat units. The instability of SSRs mainly results from slipped-strand mispairing errors during the DNA replication process (Schlötterer, 2000). In particular, the mutation rate of SSRs has been suggested to increase with the repeat number, with longer SSRs having a mutation bias to become shorter SSRs (Kruglyak *et al.*, 1998).

SSRs are found throughout the genome. They are more abundant in non-coding regions than in coding regions (Li *et al.*, 2004). In particular, SSR density in non-coding regions was highest within the 5'-UTRs of genes, followed by that in 3'-UTRs, introns, and intergenic regions (Hong *et al.*, 2007). In contrast to non-coding regions, coding regions prevalently harbor trinucleotide SSRs, reflecting the codon structure in these regions (Hong *et al.*, 2007). SSRs are also preferentially associated with gene-rich regions, with heterochromatin SSRs mostly associated with retrotransposons (Hong *et al.*, 2007). These data suggest that SSR distribution is non-random. In particular, association of SSRs with genes may play important roles in determining gene function, such as altering transcription, translation, RNA splicing or stability (Li *et al.*, 2004). As an example, the expansion of trinucleotide GAG repeats in the coding region of the huntingtin gene in humans can lead to Huntington's disease (reviewed in Li *et al.*, 2004).

The nature of SSRs gives them a number of advantages over other molecular markers (Jewell *et al.*, 2006): (i) SSRs show high transferability between related species, with high allelic diversity; (ii) multiple SSR alleles may be detected at a single locus by using a simple PCR-based screen; (iii) SSRs are co-dominant; (iv) they are distributed all over the genome, but especially occur in association with genes; and (v) analysis of SSRs may be semi-automated. Thus, SSR markers allow for rapid generation of genetic data from a relatively small amount of sample, with high reliability and reproducibility. They are particularly useful for genetic mapping, linkage and association studies, as well as for phylogenetic and population studies.

Recently, with the increase in the availability of DNA sequence information, there has been an increased need for automated methods to find SSR loci and design primers for amplification of SSR loci from large sequence datasets; this would be a particularly useful tool in crop-breeding programs where it could aid marker-assisted selection as well as estimation of genetic pop-

*Corresponding author: E-mail yplim@cnu.ac.kr
Tel +82-42-821-5739, Fax +82-42-823-1382

Received 11 October 2011, Revised 20 October 2011, Accepted 15 November 2011

ulation structure. Moreover, homology-based search for amplicons including SSRs would be helpful to develop SSR markers linked to genes. Given this need, we have recently developed an application, "SSR-Primer Generator (SSRPG)," which encompasses the following functions: (i) discovery of SSR loci from large amounts of DNA sequence data, including the SSR flanking regions; (ii) design of PCR primers that target amplification of the SSR, (iii) homology-based search for the predicted SSR amplicons, and (iv) visualization of results. Moreover, SSRPG provides a function for producing SSR-primer pairs that are unique in an analyzed dataset and that form stable duplexes with the template DNAs under a given PCR condition. The described application is publicly available at the http://168.188.15.158:8080/ssrpg.

Work Flow of SSRPG

SSRPG is a user-friendly web-application for (i) finding SSRs, (ii) designing SSR-primers, and (iii) homology-based search for their *in silico* amplicons in a batch process (Fig. 1).

SSR search

SSRs are searched from an input, which should be given in the form of multiple FASTA format DNA sequences, by SPUTNIK. The resulting output is a combined result table providing summary information of the SSRs identified, and that includes the repeat types identified, the motifs involved, their positions, length, repeat scores, and the flanking regions. After a search for SSRs, sequences containing the SSRs and their flanking regions and their summary information are used as queries (sequence, SSR-target region, primer product size range, etc.) that are required for the creation of a PRIMER3 input file.

SSR-primer design

A PRIMER3 input file is created into which the following parameters are entered: (i) common parameters for all sequences, including required primer size (20 bases), primer T_m (59°C), primer GC% (maximum: 60%; minimum: 40%), Max T_m difference (integer: 1), Max complementarity (PRIMER_SELF_ANY=5), Max 3' complementarity (PRIMER_SELF_END=2), Max Poly-X (integer: 3), CG clamp (integer: 0), Max 3' stability (integer: 9), and Max Ns accepted (integer: 0); and (ii) parameters for individual sequences, including primer sequence ID, nucleotide sequence containing the SSR derived from the first step (SSR search), the SSR target region, and the product size range (minimum: 150 bases). These parameters have been set as strict criteria to ensure robust PCR amplification in *Brassica* species.

SSR-primers are designed from the complete input by PRIMER3, and the resulting output (i.e., SSR ID, PCR product size, primer ID, primer sequence, primer length, primer T_m, primer GC%) is parsed. To increase the stability of primers as well as PCR efficiency, primer pairs with simple repetitive DNA sequences are filtered from the full list of primer sets, and similarly, cross-homologous primer pairs showing high sequence similarity (i.e., ≥90%) between different primer pairs are also filtered out. However, if one of such a pair of primers is unique, the pair is not removed from the full list of primer sets, because it may bring about specific amplification in a given template.

Homology-based search for in silico amplicons

The resulting *in silico* amplicons (SSR PCR products) can be selectively searched against NCBI non-redundant protein sequence database by using BLASTX, with adjustment of cutoff *E*-value (i.e., 1*E*-20), and best BLAST hit for each amplicon is parsed (i.e., score,

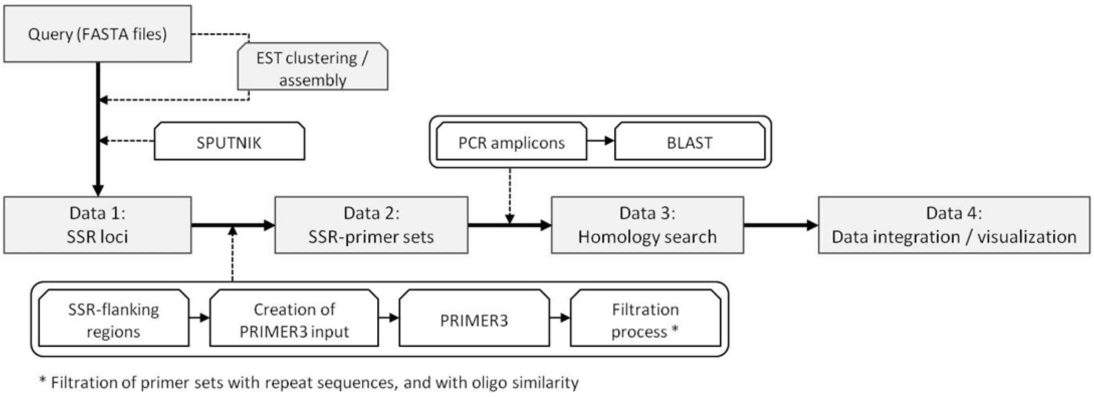
Query (FASTA files)
EST clustering / assembly
SPUTNIK
PCR amplicons
BLAST
Data 1: SSR loci
Data 2: SSR-primer sets
Data 3: Homology search
Data 4: Data integration / visualization
SSR-flanking regions
Creation of PRIMER3 input
PRIMER3
Filtration process *
* Filtration of primer sets with repeat sequences, and with oligo similarity

Fig. 1. Work flow of SSR-Primer Generator.

E-value, description, and alignment result). This analysis provides information whether the SSR amplicons are associated with protein-coding regions. Selectively, the use of nucleotide sequence database, including promoters, 5'-/3'-UTRs, exons, and introns, with BLASTN enable to analyze the genomic feature for amplicons. However, this step will be available with an administration authorization after local installation of SSRPG because of dependence on server performance. A source code for SSR-Primer Generator is also available from the authors upon request (yplim@cnu.ac.kr).

Additional options

To produce SSR-primers from unique, contiguous transcript sequence data, ESTs can be clustered and assembled by TGICL (Pertea *et al.*, 2003), and the resulting contig and singleton sequences can then be used as a query for the analysis process in SSRPG.

As another option, redundancy between different primer datasets that have been generated in different work flows can be assessed. This process would prevent duplication in primer orders when primers are derived from the same DNA region.

Data visualization

After completion of each process, the corresponding data on SSRs, SSR-primers, and BLAST search are visualized in HTML format and can be downloaded as a tab-delimited text file. In particular, because the results are cross-linked to each other, co-visualization of different data is allowed in HTML format. All results produced from each process are also stored into a MySQL database.

Implementation

SSRPG is supported on Linux and on Java Virtual Machine (JVM) operating systems. The programs (i.e., SPUTNIK, PRIMER3, BLAST, and Python scripts for data parsing and filtering), which make up the analysis pipeline shown in Fig. 1, are batch-processed with supporting of JSP programming. Data processing and input and output services are implemented as servlets running in an Apache Tomcat container on a Webserver. A MySQL database is also used to manage data that are produced during corresponding analysis processes.

Features of SSRPG

SSRPG provides some advantages for discovering SSRs and designing SSR-primers: (i) It uses multiple FASTA format DNA sequences; (ii) it provides ease of use through access to the web; (iii) it allows batch processing of the analysis pipeline; (iv) it allows extra-processing for increasing the stability of primers (filtering of primer sets with repeat sequences and with cross-homology between different data sets); and (v) it allows cross-linking between different results in HTML-based tabular format (viz., data integration). Moreover, homology-based search for *in silico* amplicons would be helpful to develop SSR markers linked to genes. Such an approach also enables one to identify marker transferability between closely related genomes by comparing amplicons to other related genomic sequences.

Performance Testing of SSRPG

For performance testing of SSRPG, 197,047 genomic survey sequences (GSSs) and 148,391 expressed sequence tag (EST) sequences of *Brassica rapa* ssp. *pekinensis* were processed in SSRPG. In particular, ESTs were initially clustered and assembled by TGICL, and a total of 16,834 unique sequences were analyzed. Based on the type of primary SSR motif, a total of 29,245 and 4,101 SSRs were found in the GSS and EST datasets, respectively, and of those SSRs found, 10% for GSS and 17% for EST were finally designed as SSR-primer pairs (Table 1). Because of the use of stringent parameters for primer design as well as the filtration process that removes the primer pairs having simple repetitive DNA sequences and/or showing cross-homology with high sequence similarity, those SSR-primers will enable robust PCR amplification. To test the utility of SSR-primers designed, we performed the PCR assay with 150 and 114 sampled GSS- and EST-derived primer sets, respectively. The resulting assay revealed a high success rate of PCR amplification (98.7% for GSS and 95.6% for EST) (Table 1). Moreover, clear single amplicon was observed in the assay regardless of repeat types (data not shown). This result supports that SSRPG produces SSR-primers with high specificity and stability for PCR. Additionally, in the primer design, SSRs with C or G-containing motif (i.e., AC, AAG, ATC, AAC, AGG, AGC, ACT, CCG, AAAG motifs) showed relatively high success rate comparing to those with AT-rich motif (i.e., AT, AAT, and AAAT motifs) (Table 1). This suggests that the degree of GC content as well as low complexity in SSRs and their flanking regions affect the determination of target regions for SSR-primer design.

Conclusion

SSRPG is a user-friendly web application for (i) SSR discovery, (ii) SSR-primer design, and (iii) homology

Table 1. SSR discovery and SSR-primer design from GSSs and ESTs of *Brassica rapa*

Repeat type	Type of primary SSR motif[a]	No. of identified motif		No. of SSR-primer pair						PCR assay			
				Before filtration		After filtration		Frequency of SSR-primer[b]		Tested primer pair (No.)		Amplification success (No.)	
		GSS	EST	GSS	EST	GSS	EST	GSS	EST	GSS	EST	GSS	EST
Mononucleotide	A	9,570	N/A[c]	1,949	N/A	942	N/A	0.10	N/A	N/A	N/A	N/A	N/A
	C	518	N/A	106	N/A	57	N/A	0.11	N/A	N/A	N/A	N/A	N/A
	Subtotal no.	10,088		2,055		999							
Dinucleotide	AG	3,855	897	890	123	390	94	0.10	0.11	10	17	10	17
	AT	4,169	136	506	24	268	20	0.06	0.15	N/A	5	N/A	5
	AC	797	99	183	16	76	12	0.10	0.12	10	6	10	5
	Subtotal no.	8,821	1,132	1,579	163	734	126						
Trinucleotide	AAG	3,246	835	789	198	366	157	0.11	0.19	10	20	10	20
	AAT	1,255	87	204	11	107	8	0.09	0.09	10	6	10	6
	ATC	1,343	413	333	96	166	78	0.12	0.19	10	8	9	8
	AAC	1,203	247	281	67	135	51	0.11	0.21	10	7	10	7
	AGG	1,043	470	282	146	130	109	0.13	0.23	10	14	10	14
	ACC	630	293	152	66	70	49	0.11	0.17	10	7	10	6
	ACG	256	103	56	16	25	14	0.10	0.14	10	6	10	5
	AGC	421	279	102	70	59	55	0.13	0.20	10	5	10	3
	ACT	177	55	41	18	22	14	0.12	0.25	10	5	10	5
	CCG	314	136	86	44	47	38	0.15	0.28	10	5	10	5
	Subtotal no.	9,888	2,918	2,326	732	1,127	573						
Tetranucleotide	AAAT	238	14	40	1	19	1	0.08	0.07				
	AAAC	100	15	23	1	14	1	0.14	0.07	10	N/A	9	N/A
	AAAG	110	22	27	4	16	1	0.59	0.05	10	1	10	1
	Subtotal no.	448	51	90	6	49	3			10	2	10	2
	Total no.	29,245	4,101	6,050	901	2,909	702	0.10	0.17	150	114	148 (98.7%)	109 (95.6%)

[a]Of the entire motif identified, primary motifs were analyzed. [b]Frequency, SSR motif no. corresponding to SSR-primers designed after filtration/No. of identified SSR motifs. [c]N/A, not applicable.

search of *in silico* amplicons from a large amount of DNA sequences in a batch process. This application will greatly reduce the time and costs associated with the development and application of SSR markers in endeavors such as genetic mapping, association mapping, comparative analysis (viz., the study of synteny), and phylogenetics. Moreover, this will be a useful tool to provide public SSR marker resources, which would, in turn, promote sharing of the associated analysis data.

Acknowledgements

Dr. CP Hong was supported by the Korea Research Foundation Grant funded by the Korean Government (KRF-2008-357-F00014). Dr. YP Lim and Dr. SR Choi were supported by grants for the Technology Development Program for Agriculture and Forestry (Grant no. 607003-05), Ministry of Food, Agriculture, Forestry, and Fisheries, Republic of Korea, and by a grant from the Next-Generation BioGreen 21 Program (Plant Molecular Breeding Center No. 20110701-305-595-001-04-00), Rural Development Administration, Republic of Korea.

References

Hong, C.P., Piao, Z.Y., Kang, T.W., Batley, J., Yang, T.J., Hur, Y.K., Bhak, J., Park, B.S., Edwards, D., and Lim, Y.P. (2007). Genomic distribution of simple sequence repeats in Brassica rapa. *Mol. Cells.* 23, 349-356.

Jewell, E., Robinson, A., Savage, D., Erwin, T., Love, C.G., Lim, G.A., Li, X., Batley, J., Spangenberg, G.C., and Edwards, D. (2006). SSRPrimer and SSR Taxonomy Tree: Biome SSR discovery. *Nucleic Acids Res.* 34, W656-W659.

Kruglyak, S., Durrett, R.T., Schug, M.D., and Aquadro, C.F. (1998). Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. *Proc. Natl. Acad. Sci. U.S.A.* 95, 10774-10778.

Li, Y.C., Korol, A.B., Fahima, T., and Nevo, E. (2004). Microsatellites within genes: structure, function, and evolution. *Mol. Biol. Evol.* 21, 991-1007.

Pertea, G., Huang, X., Liang, F., Antonescu, V., Sultana, R., Karamycheva, S., Lee, Y., White, J., Cheung, F., Parvizi, B., Tsai, J., and Quackenbush, J. (2003). TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics* 19, 651-652.

Schlötterer, C. (2000). Evolutionary dynamics of microsatellite DNA. *Chromosoma* 109, 365-371.