In Silico Data Mining for Development of *Capsicum* Microsatellites

A.G. İnce, A.N. Onus, S.Y. Elmasulu, M. Bilgen and M. Karaca Akdeniz University, Faculty of Agriculture Antalya 07059 Turkey

Keywords: Data mining, SSR, microsatellites and EST-SSR primers

Abstract

Tandemly repeated DNAs play important roles in evolution and regulatory processes. Tandem Repeats (TRs) exist in coding and non-coding portion of genomes. Simple Sequence Repeats (SSRs) or microsatellites are stretches of DNA consisting of exact and inexact simple tandemly repeated DNA. In this study, distribution and frequency of TRs and SSRs in *Capsicum annuum* Expressed Sequence Tags (ESTs) were investigated. Results demonstrated that TR's and SSR's content of *C. annuum* were quite high in the family Solanaceae. There were considerable variations in the occurrence of the individual SSR motif in *Capsicum annuum.* SSR motifs of mono A/T, di CT/GA, tri CTT/GAA, tetra AAAT/TTTA, penta TTTTC and hexa CTGCTC nucleotides were abundant in *Capsicum.* Exact and inexact SSRs also differed in frequency and distribution in *Capsicum.* Results indicated that frequency and distribution of SSRs differing in tissue/organ ESTs of *Capsicum* indicated that repeats distribution and frequency may not be random in the coding portion of genome differing from the non-coding portion of genomes. Exact and inexact SSR contents of several organs/tissues differed. A total of 300 *Capsicum* EST-SSR primer pairs, of which 124 were tissue/organ specific, were developed.

INTRODUCTION

Pepper (*Capsicum annuum*) is one of the most important genera in the family Solanaceae, along with tomato *(Lycopersicon esculentum),* potato (*Solanum tuberosum*), eggplant *{Solanum melongena*), petunia *{Petunia hybrida),* and tobacco *{Nicotiana tabacum*). Recent studies resulted in exponential growth of published Expressed Sequence Tags (ESTs) for the family Solanaceae. ESTs are single-pass DNA sequences usually about 300-500 nucleotides in length, obtained from mRNA (cDNA) representing genes expressed in a given tissue and/or at a given development stage. One of the useful and interesting applications of ESTs is the study of the gene expression pattern in response to a given organ, tissue or development stage. ESTs have been mined for Single Nucleotide Polymorphism (SNP, Schmid et al., 2003) and SSR (Thiel et ah, 2003). To our knowledge, there is no research in the distribution of repeat containing ESTs among tissue and organs in *Capsicum* and as well as other organisms with the exception of *Arabidopsis* (Bilgen et ah, 2004).

Simple Sequence Repeats (SSRs) or microsatellites are stretches of DNA consisting of exact or inexact simple tandemly repeated short motifs of 1-6 base pairs in length. SSRs are ideal DNA markers because they are highly polymorphic between individuals and highly abundant, dispersed evenly throughout the genomes. SSRs are also inherited in a co-dominant fashion, fast and easy to assay by PCR using two unique primer pairs flanking the tandem repeats. Moreover, they can serve as sequence tagged sites for anchoring in genetic maps (Karaca et ah, 2002). The standard procedure for developing SSRs involves the construction of a small-insert genomic library, the subsequent hybridization with tandemly repeated oligonucleotides and the sequencing of candidate clones; thus, making the process time consuming and labor-intensive. An alternative strategy for development of SSRs arises from increasing information available in genomic DNA and EST databases. Due to the rapid increase of sequence information, the generation of EST-SSR becomes an attractive alternative to complement existing genomic SSR collections (Thiel et al, 2003).

Proc. III[rd] Balkan Symp on Vegetables and Potatoes
Eds.: H.Ö. Sivritepe and N. Sivritepe.
Acta Hort. 729, ISHS 2007

The development of SSR primer pairs can be performed at significantly reduced costs, as EST-SSRs are free by-product of the currently expanding EST databases. Since ESTs represent the transcribed part of the genome, EST-SSR markers led to the direct mapping of the genes. SSRs located in coding regions of important genes that are expressed under various conditions of environment, stress, organ, tissue and development stage would also lead to the development of tissue/organ/development stage specific SSR marker and that would be very valuable to understand the repeat function in gene and mapping for breeding and evolutionary studies. Several different approaches have been employed to identify TRs in various organisms including (Kantety et ah, 2002; Thiel et ah, 2003) . This research was undertaken to: 1) Compare the TRs content of pepper, tomato, potato and tobacco; 2) Determine the distribution and frequency of exact- and inexact- SSRs in pepper and pepper tissues/organs; and 3) Develop SSR primer pairs.

MATERIALS AND METHODS

A total of 23,165,289 ESTs from GenBank (ftp://ftp.ncbi.nih.gov/repository /dbEST/) were scanned and a total of 397,548 ESTs from Capsicum, Lvcopersicon. Nicotiana and Solanum spp. were processed to identify four different repeat types. Tandem repeats (TRs) consisted of exact and inexact repeats up to 1000 nucleotides in length. Exact SSRs were considered to contain motifs that are between one and six nucleotides in length. The minimum motif length criteria were defined as being ten repeats for mononucleotides, six repeats for dinucleotides and five repeats for all higher-order motif length, according to Thiel et al. (2003). Inexact repeats and compound repeats were determined according to Bilgen et al. (2004). To identify unique keywords representing only one particular tissue/organ we used a homemade program (unpublished) that searched for annotated ESTs and picked the keywords used in this analysis. A total of 5 organs/tissues were used as keywords for *Capsicum.* Simple regression coefficient analysis was utilized to investigate whether the repeat motifs of organs/tissues were dependent on the EST numbers studied. Using this information we tried to estimate expected repeat frequency for each organ/tissue (Bilgen et al., 2004). Flanking SSR primer pairs were designed using PRIMER3 software (Rozen and Skaletsky, 1998).

RESULTS AND DISCUSSION

Analyses revealed that distribution of ESTs and EST-TRs in some species family Solanaceae differed. *Solanum tuberosum* had the highest number of ESTs in GenBank, followed *Lycopersicon esculentum* (Table 1). *Capsicum annuum* ESTs contained the highest amount of tandem repeats (15.64%) in family of Solanaceae studied. The longest tandem repeat of 150 nucleotides was found in *Solanum tuberosum* ESTs. Most of the TR- ESTs with unit length a multiple of 3 (9-12-15, etc) bps were the most frequent among the repeat units searched, probably reflecting that they are part of coding sequences. Tandem repeats ranging from mono to one thousand nucleotides were present in every 7.06 kb for tomato, 1,87 kb for tobacco, 5.2 kb for potato and 2.95 kb for *Capsicum annuum* ESTs indicating that pepper is rich in tandem repeats containing genes. It seems to us that genome size was not correlated with the tandem repeat amounts in ESTs since the genome size of *Capsicum annuum* is smaller than *Nicotiana tabacum,* but its TR content was much higher than tobacco.

A total of 30,149 *Capsicum annuum* ESTs were screened, 7,272 EST contained exact-SSRs (20.42%). For mononucleotides, poly A/T tract was found to be more abundant compared with poly G/C. Of the di-, tri- tetra-, penta- and hexanucleotide repeats, CT/GA, CTT/GAA, AAAT/TTTA, TTTTC and CTGCTC had the highest frequency (Table 2). Our findings in the present study agreed with previous studies (Thiel et al, 2003; Bilgen et al., 2004). Frequency of inexact-SSRs was 14.42%. The distribution of exact SSR amounts differed from inexact-SSRs but the frequency of individual repeat types were similar with one exception. In the trimeric repeats AGA/TCT were the most frequent one in inexact-SSRs instead of CTT/GAA found in exact-SSRs. This was probably due to the exact module of TRA misses mismatch-containing tandem repeats but inexact module detects both mismatch-containing and exact tandem repeats (Bilgen et all, Exact compound repeats and

inexact compound repeats, if they frequently occur; would affect the overall average, were surprisingly rare in *Capsicum.* Thirty-three exact compound repeats and only one inexact compound repeats were found where two repeat regions were immediately adjacent. These constitute only 0.45% of the SSRs found in *Capsicum* making little difference to the average distribution. In plants, it is known that compound repeats appear to comprise about 10% of SSRs (Bull et ah, 1999). Since compound SSRs show significant variation, they are very valuable tools in plant breeding studies.

We studied 5 different tissues/organs of *Capsicum annuum.* Analysis of root ESTs (890), leaf (5,698), anther (1,398), flower bud (2,016) and fruit (7,478) indicated that leaf and flower bud contained the highest amounts of exact- and inexact SSRs. However, based on the correlation analyses we found that exact and inexact-SSR amounts of anther, were due to the higher ESTs studied. Indeed anther should have had higher exact- and inexact-SSRs than observed (Fig. 1). Thus, the exact- and inexact-SSR content of flower bud was higher among the five tissues/organs studied. Exact- and inexact-SSR amounts of leaf and fruit were significantly differed. Leaf ESTs contained less amounts of inexact- SSRs while fruit ESTs contained less amounts of exact-SSRs.

Regardless of the SSR type, exact or inexact-SSRs, for all organs/tissues, the frequency of SSRs decreased with increasing repeat lengths with the exception of root, anther and flower bud trimeric repeats. Trimeric repeats of these organs/tissues were higher than dimeric repeats. Results of the present study also indicated that SSR motif types were not randomly distributed in *Capsicum annuum* organs/tissues (Table 3). Similar results about unevenly distributed individual SSR motif type occurrence in organ/tissue ESTs have also been reported in *Arabidopsis* spp. (Bilgen et al, 2004). Trimeric motifs, TAA, TAG and TGA rarely appeared, probably because they code for stop codons that have a direct effect on protein synthesis (Chin, 1996).

For the development of SSR primer pairs, we designed primers for not all the identified EST-SSRs. Primilary studies indicated that at least one thousand primer pairs could be generated from *Capsicum annuum* ESTs. For the remaining SSRs, primer pairs could not be designed for one of the following reasons: (1) EST sequences containing SSRs were very short; (2) SSRs were too close to the cloning site of the EST; (3) the flanking sequences were not unique; and (4) calculated annealing temperature of the primer pairs differed more than 1°C. The occurrences of SSRs or TRs in transcribed genes (ESTs) accepted by the scientific community and the use of EST derived SSRs in differential display study have also been reported (Saha et al., 2003). In this study, we developed a total of 300 *Capsicum* EST-SSR primer pairs, of which 124 were tissue/organ specific. Developed SSRs primer pairs are available upon request from corresponding author of this manuscript. These primer pairs could be used in differential display studies in *Capsicum* for characterization of repeat containing genes. The construction of a molecular genetic linkage map using ESTs-derived primer pairs of pepper plant will provide a bridge to the utilization of the genomic information and genetic resources of tomato, potato, and eggplant studies.

Literature Cited

Arumuganathan, K. and Earle, E.D. 1991. Nuclear DNA Content of Some Important Plant Species. Plant Molecular Biology Reporter. 9:211-215.

Bilgen, M., Karaca, M., Onus, A.N. and Ince, A.G. 2004. A software program combining sequence motif searches with keywords for finding repeats containing DNA sequences. Bioinformatics 2004; doi:10.1093.

Bull, L.N., Pabon-Pena, C.R. and Freimer, N.B. 1999. Compound microsatellite repeats: practical and theoretical features. Genome Research.9:830-838.

Chin, E.C.L. 1996. Maize simple repetitive DNA sequences: abundance and allele variation. Genome.39:866-873.

Kantety, R.Y., La Rota, M., Matthews, D.E. and Sorrells, M.E. 2002. Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum and wheat. Plant Molecular Biology. 48:501-510.

Karaca, M., Saha, S., Jenkins, J.N., Zipf, A., Kohel, R. and Stelly, D.M. 2002. Simple Sequence Repeat (SSR) markers linked to the *Ligon Lintless (Li\)* mutant in cotton. J.

Heredity. 93:221-224.
Parisi, V., Fonzo, V. D. and Aluf.-Pentini, F. 2003. STRING: finding tandem repeats in DNA sequences. Bioinformatics.19:1733-1738.
Rozen, S. and Skaletsky, H.J. 1998 Primer3 Code available at http://www.genome.wi.mit.edu/genome_soflware/other/primer3.html
Saha, S., Karaca, M., Jenkins, J.N., Zipf, A.E., Reddy, O.U.K., Pepper, A.E. and Kantety, R. 2003. Simple sequence repeats as useful resources to study transcribed genes of cotton. Euphytica. 130:355-364.
Schmid, K.J., Sorensen, T.R., Stracke, R., Torjek, O., Altmann, T., Mitchell-Olds, T. and Weisshaar, B. 2003. Large-scale identification and analysis of genome-wide single-nucleotide polymorphisms for mapping in *Arabidopsis thaliana.* Genome Research. 13:1250-1257.
Thiel, T., Michalek, V. and 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.

Tables

Table 1. Distribution of tandem repeats in some species of family Solanaceae.

Source	Genome Size (Mbp/lC)[1]	Total EST length (Mbp)	Total EST No.	EST-TRs No.	%	TR/kbp EST
Lycopersicon esculentum	907-1000	76.3	151792	10290	6.78	7.06
Nicotiana tabacum	4221-4646	5.14	10426	195	1.87	25.21
Solanum tuberosum	1597-1862	96.9	158454	8244	5.20	10.52
Capsicum annuum	2702-3420	13.93	30149	4716	15.64	2.95

[1] Arumuganathan and Earle (1991).

Table 2. Frequency and distribution of SSRs in *Capsicum annuum.*

	Exact-SSR No		%	InExact-SSRNo		%
Mono-	5596	A/T	76.95	3650	A/T	79.37
Di-	809	CT/GA	11.12	353	CT/GA	7.68
Tri-	774	CTT/GAA	10.64	341	AGA/TCT	7.41
Tetra-	25	AAAT/TTTA	0.34	31	AAAT/TTTA	0.67
Penta-	2	TTTTC	0.03	15	TTTTC	0.33
Hexa-	33	CTGCTC	0.45	208	CTGCTC	4.52
CRs		33	0.45		1	0.02
Total		7272	20.42		4599	14.5

CRs: Compound Repeat

Table 3. Frequency, type and distribution of SSRs in *Capsicum annuum* tissues/organs.

Tissues	Type	Mono-	Di-	Tri-	Tetra-	Penta-	Hexa-
	Exact-	A/T	AG/TC	AGA	-	-	AAACCA
	SSRs	100%	40%	16.66%	-	-	100%
JLvOOX	Inexact-	A	AG/TC	AGA	-	-	AAACCA
	SSRs	100%	60%	27.27%	-	-	20%
	Exact-	A/T	CT/GA	CCA/GGT	AATT	-	ACACAG
	SSRs	98.7%	46.07%	13.51%	50%	-	33.33%
T pof JLvÇcil	Inexact-	A/T	CT/GA	AGA/TCT	AATT	ACCCC	CCACGA
	SSRs	99.9%	61.81%	23.53%	50%	33.33%	31.43%
	Exact-	A/T	AG/TC	ACA	ATTT	-	-
	SSRs	98.84%	50%	13.04%	100%	-	-
AmllCr	Inexact-	A/T	TC	CAG	ATTT	-	CTCTTC
	SSRs	100%	42.85%	20%	50%	-	33.33%
	Exact-	A/T	CT/GA	AGA	TCTT	-	-
Flower	SSRs	99.45%	52%	16.07%	100%	-	-
bud	Inexact-	A/T	AG	AGA	TCTT	GAATT	AAACCA
	SSRs	99.9%	60%	25%	100%	100%	11.1%
	Exact-	A/T	CT/GA	CTT/GAA	TA AT	TCTCT	AAAACA
	SSRs	98.04	48.42%	12.96%	30%	100%	33.33%
	Inexact-	A/T	CT/GA	AGA/TCT	AAAT	CCTCT	CTGCTC
	SSRs	99.48%	58.62%	13.59%	28.57%	25%	23.08%

Figures

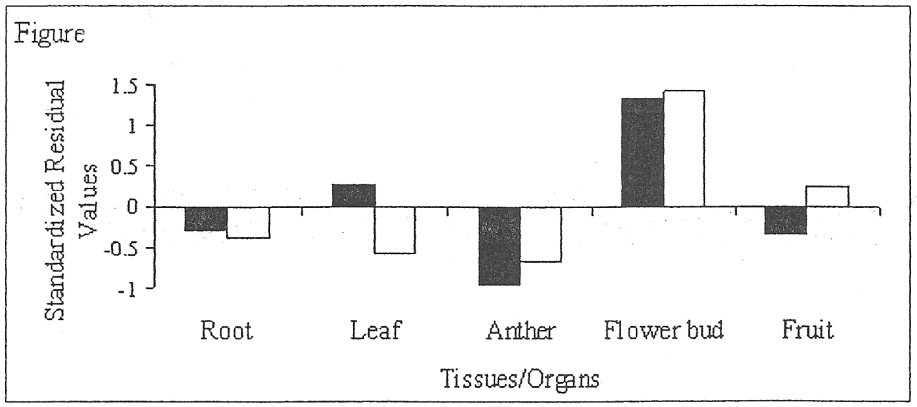

Fig. 1. Standardized residual values of Exact and InExact SSRs found in *Capsicum annuum.* Black columns are exact-SSRs and white ones are inexact-SSRs.