COSRaptor: a software for large-scale COS identification and polymorphic microsatellite in several plant species and cultivars

Grillo G[1], Licciulli F[1] ,Catalano D[2]

Motivation

Detection and study of genetics variation can assist us to comprehend the molecular basis of variations in many biological phenomena in plants. Genetic or DNA based marker techniques such as RFLP (restriction fragment length polymorphism), SSR (simple sequences repeats), AFLP (amplified fragment length polymorphism) are routinely used in many fields of interest in plant science. The repetitive DNA sequences as simple sequences repeat (SSR) or sequences tandem repeat (STR) are widely spread in the eukaryotic1 and prokaryotic genomes2,3. SSR, in plants, are ubiquitous in transcribed sequences typically locus specific co-dominant, in same case the SSR are multi-allelic highly polymorphic. Public EST datasets are rich resources to found orthologous–specific EST-SSR markers for genotyping application in numerous species in flowering and crop plant. In this work we have developed a bioinformatics system to found the "Conserved Orthologous Sequences4" (COS) in ESTs and to assess SSR polymorphism in different species, variety and/or ecotype. These polymorphic SSRs in COS regions are useful for germplasm characterization and breeding applications. The procedure for individuation of the COS is scalable and reusable to assess various molecular markers predicted by in-silico analysis (ie RFLP or SNP).

Methods

To estimate the SSR polymorphism we used two different EST datasets belonging to the Asteraceae family, the former composed by four different species (Cynara scolimus, Centaurea solstitialis, Centuarea maculosa and Carthamus tinctorius) and the latter composed by the ESTs of five different Helianthus annus cultivars (Emil, Psc8, RHA280, RHA801, HA89). We have used a web agent embedded in Perl script to retrieve from the NCBI EST Database the ESTs of Helianthus annus and to collect the sequences in five "cultivar specific" datasets; then for each cultivar dataset we used an in-house modified version of Misa perl script to found the SSRs on the collected ESTs. We have created a bioinformatics framework composed by: a Blast engine to found significative homologous regions in two species/cultivars and a dynamic algorithm to extend the homologous regions, detected above, to all species/cultivars. The algorithm identifies the longest COS

[1] Istituto di Tecnologie Biomediche (ITB) - CNR, Via Amendola 122/D, Bari, Italy [2] Istituto di Genetica Vegetale (IGV), Via Amendola 165/A, Bari, Italy

region shared by a cluster of homologous ESTs for each combination of selected species/cultivars. Bearing in mind the large amount of data to handle we developed a relational database. The database was used to store and manage the data produced during the analysis steps (EST sequences, blast results, SSR analysis data), to support the algorithm to generate the COS region and to localize the SSRs in the COS.

Results

We detected 16,072 SSRs on 20,977 ESTs in the first dataset (four species) and 20,852 SSR on 22,630 ESTs in Helianthus annus cultivars by Misa software. We studied the diversity and "used sequence space" of the different SSR in these two datasets and we observed more variability in the Helianthus annus cultivars than in the other four Asteraceae species; in fact we found 2,074 and 6,612 different SSRs in the species and in the cultivars datasets respectively, while 334 are the SSRs shared between the two datasets. The total number of COS regions detected, in the four species, are 54,451, reduced to 27,101, considering COS region longer than 500 pb (higher than the estimated average of EST lengths).

We mapped the SSRs on the detected COS obtaining 7,367 (in the four species) and 1,116 (in Helianthus annus cultivars) COS regions with almost one SSR. Preliminary analysis shows that the 6% and 10 % in species and cultivars respectively are "polymorphic COS". These in-silico predicted polymorphic SSRs will be validated by in-vivo experiments.

Contact e-mail

domenico.catalano@igv.cnr.it

Supplementary information

1) Distinct patterns of SSR distribution in the Arabidopsis thaliana and rice genomes. Mark J Lawson and Liqing Zhang. Genome Biol. 2006; 7(2): R14.

2) Short-sequence DNA repeats in prokaryotic genomes. van Belkum A, Scherer S, van Alphen L, Verbrugh H. Microbiol Mol Biol Rev. 1998 Jun;62(2):275-93.

3) Simple sequence repeats in prokaryotic genomes. Jan Mrázek, Xiangxue Guo, and Apurva Shah. Proc Natl Acad Sci U S A. 2007 May 15; 104(20): 8472–8477.

4) An SSR-based genetic linkage map of the model grass Brachypodium distchyon. Garvin DF, McKenzie N, Vogel JP, Mockler TC, Blankenheim ZJ, Wright J, Cheema JJ, Dicks J, Huo N, Hayden DM, Gu Y, Tobias C, Chang JH, Chu A, Trick M, Michael TP, Bevan MW, Snape JW. Genome 2010 Jan; 53(1):1-13.