Published online February 21, 2006

Nucleic Acids Research, 2006, Vol. 34, No. 4 ***e31***
doi:10.1093/nar/gnj030

New softwares for automated microsatellite marker development

Wellington Martins, Daniel de Sousa[1], Karina Proite[2], Patrícia Guimarães[2], Marcio Moretzsohn[2] and David Bertioli[3]

Department of Computer Science, Catholic University of Goiás, Brazil, [1]Department of Computer Science, Catholic University of Rio de Janeiro, Brazil, [2]Embrapa Genetic Resources and Biotechnology, Brasília, Brazil and [3]Genomic Sciences and Biotechnology, Catholic University of Brasília, Brazil

Received November 23, 2005; Revised and Accepted January 31, 2006

ABSTRACT

Microsatellites are repeated small sequence motifs that are highly polymorphic and abundant in the genomes of eukaryotes. Often they are the molecular markers of choice. To aid the development of microsatellite markers we have developed a module that integrates a program for the detection of microsatellites (TROLL), with the sequence assembly and analysis software, the Staden Package. The module has easily adjustable parameters for microsatellite lengths and base pair quality control. Starting with large datasets of unassembled sequence data in the form of chromatograms and/or text data, it enables the creation of a compact database consisting of the processed and assembled microsatellite containing sequences. For the final phase of primer design, we developed a program that accepts the multi-sequence 'experiment file' format as input and produces a list of primer pairs for amplification of microsatellite markers. The program can take into account the quality values of consensus bases, improving success rate of primer pairs in PCR. The software is freely available and simple to install in both Windows and Unix-based operating systems. Here we demonstrate the software by developing primer pairs for 427 new candidate markers for peanut.

INTRODUCTION

A molecular marker is an identifiable DNA region whose inheritance can be followed along generations. Molecular markers have become a powerful tool in many areas of biology. Microsatellites, also known as simple sequence repeats (SSR) or short tandem repeats (STR), are repeated sequence motifs, whose unit of repetition is between 1 and 6 bp. They are highly abundant in the genomes of eukaryotes, polymorphic and usually co-dominant and transferable between different mapping populations. Microsatellite markers can also be used in automated genotyping techniques. Thus, they have become one of the most useful molecular markers for a large number of organisms.

Researchers working on the development of microsatellite markers need an efficient way to go from usually hundreds or thousands of trace and/or text sequence files to the identification of new potential markers. However, the softwares currently available for data-mining microsatellites are either restricted to small datasets, not available for Windows, or are not targeted towards marker development. In addition, they do not offer base quality control and with one exception (RepeatMasker when used with Staden), they are not integrated into a sequence analysis software, and as such require already assembled data as input (1–3), http://www.maizemap.org/bioinformatics/SSRFINDER/README.ssrfinder and http://www.repeatmasker.org/.

We have developed a Staden module that integrates the program TROLL (4) to the Staden package. TROLL is an open source program based on the Aho-Corasick algorithm for the detection of perfect microsatellites. The Staden Package is a free, open source, software tool used for sequence assembly and analysis (5) (http://staden.sourceforge.net/). The Staden package, working with the module and TROLL, enables efficient processing of a large number of sequences in chromatogram and/or text form to produce a small database with only microsatellite containing sequences, processed, assembled in contigs and that have not previously been used for marker development. For the primer design we have developed a program that uses Primer3 (6) to design primers flanking the SSR located by the module. It is available as a stand-alone program and through a Web service. This program accepts multiple sequences in the 'experiment file' format of Staden as input and produces a list of primer pairs

*To whom correspondence should be addressed. Tel: +55 62 32271381; Fax: +55 62 32271073; Email: wsmartins@ucg.br
Correspondence may also be addressed to David Bertioli. Tel: +55 61 34484735; Fax: +55 61 33403658; Email: david@pos.ucb.br

for amplification of microsatellite markers. The program can take into account the quality values of consensus bases, improving success rate of primer pairs in PCR. The Staden Package, TROLL and the module are freely available for both Linux and Windows platforms.

MATERIALS AND METHODS

General details of the software

The TROLL module was developed in the Tcl/Tk scripting language, following the Staden Package module development specifications. It works through Pregap4 reading the Staden Experiment Files one by one. These files contain, among other information, the DNA sequences, the corresponding quality values and any vector contamination masking information. All sequences are then concatenated and a special character (#) is used as a separator. TROLL then locates the SSRs, these are then filtered, disregarding those in low quality or vector contaminated regions and the result is written back on the Experiment Files. A file containing the names of experiment files with microsatellites is generated (pregap.SSR.passed), and this can be used for entering into the Gap4 interface of Staden to create a database of entirely microsatellite containing sequences. The tags added to the experiment files may be used by Gap4 to aid in assembly by masking and to visualize the microsatellites. Additionally the tags are used by the primer design program to define the target region around which the Primer3 designs primers (see below).

The following module configuration options are available:

(i) Tag type: used to control Tag type used to mark microsatellite repeats.
(ii) Motif length: used to set the motif lengths (mono, di, tri, tetra, penta and/or hexa) that will be used in the search for SSRs.
(iii) Repeat min: used to set the minimum number of motif repetitions for SSRs of each motif length.
(iv) Motif file: the location of the file containing the motif list required by the TROLL program.
(v) Create SSR.passed file?: for the creation of a file with the names of all sequences containing SSRs. This file is used to enter into Gap4 to create a database of microsatellite containing sequences.
(vi) Create no_SSR_passed file?: used to create a file with the name of all sequences that do not contain SSRs.
(vii) Filter quality: used to set the base quality threshold which will be used when analyzing SSR regions.

The program for primer design is responsible for parsing the experiment format file exported from Gap4 and calling the Primer3 program repeatedly to produce primer pairs flanking the SSRs. For each sequence present in the input file, the program analyses the tags added by the TROLL module (default is 'REPT') and produces a list of the repeats found. Since the input file may contain assembled readings, overlapping repeats are discarded. Based on the list of repeats produced, and on the input parameters chosen by the user (e.g.: GC content, TM, primer length etc.), input Primer3 files are created and processed. The program then analyses Primer3 output to cluster repeats being flanked by the same primer pairs. When all sequences have been analysed, the program outputs a tab spaced file (primers.passed) that includes the contig name, primer names, microsatellite sequence(s), primer sequences, annealing temperatures and the quality of the contig sequences with the primers annealed. In addition, the program provides three other output files, marker_seqs.fasta containing the sequences for which the primer design was successful, not_marker_seqs.fasta with sequences for which primer design was not successful, and primers.failed, which also contains the sequences for which primer design was not successful, but in the experiment file format of Staden. The primers.failed file can be used for another round of analysis using the primer design program with less stringent conditions for the primer design.

Downloading and using the programs

Download and install:

(i) The Staden package from http://staden.sourceforge.net/.
(ii) The TROLL module (TROLL.p4m) from http://finder.sourceforge.net/ (follow the link to Resources). Troll_p4m is distributed as a zip file, containing the module itself, troll binaries (Linux and Windows), a DLL Windows library and installation instructions.
(iii) The primer design program from http://finder.sourceforge.net/ (Follow the link to Resources). The primer design program is available as a Web service, requiring no program installation. Alternatively, the primer design program can be downloaded as a Perl script and installed locally. For the Perl script to work you will need to install Primer3, which is available from http://frodo.wi.mit.edu/primer3/primer3_code.html.
(iv) The Step-by-step protocol for use of the module from http://finder.sourceforge.net/ (Follow the link to Resources). The Step-by-step protocol for use of the module is written for novice users of the Staden Package. However, to use the TROLL module to its full potential, the user needs to be familiar with the Staden. The course documentation that is downloaded with the software is highly recommended (for Windows users in C:\Program files\Staden package\course\).

Examples of marker development from three datasets

TC-repeat enriched genomic library. Peanut (*Arachis hypogaea*) genomic DNA library enriched for TC repeats was made using genomic DNA cut with the restriction enzyme Sau3AI, oligonucleotides and magnetic beads according to a standard protocol (7). Two hundred and eighty genomic clones were sequenced using the forward primer, and plasmids that contained microsatellites were identified using the module. These plasmids were sequenced using the reverse primer, and the resulting 338 chromatograms were processed using the module, as described in the Step-by-Step protocol (Supplementary Data), but also using a naming scheme allowing Staden to identify forward and reverse read pairs from individual clones. After assembly in Gap4, where necessary, forward and reverse reads were joined manually using the 'Find read pairs', 'Find internal joins' and 'Join Contigs' commands. Consensus sequences were exported in experiment file format from the Gap4 interface, and used as input file for the Web interface of the primer design program.

Expressed sequence tags (ESTs). Eight thousand eight hundred ESTs from *Arachis stenosperma* in the form of traces were processed through the Staden Pregap4 interface. The BLAST module was run to eliminate published molecular markers for peanut using a BLAST database (8) formatted from 707 known *Arachis* SSR markers [(9–12); M. Moretzsohn and D. Bertioli, unpublished data] (http://www.biomedcentral.com/1471-2229/3/3, http://www.biomedcentral.com/1471-2229/4/11); and the TROLL module was run to identify and select di-, tri-, tetra- and penta-microsatellites with six or more motif repetitions. The resultant pregap.SSR.passed file was entered into the Staden Gap4 interface, using the Gap4 assembler with masking, to create a database of microsatellite containing sequences that have not yet been used for marker development. Consensus sequences were exported in experiment file format from the Gap4 interface, and used as input file for the primer design program.

Text data from GenBank. Three thousand eight hundred and ninety *Arachis* sequences from GenBank in simple text format were processed essentially as described in the Step-by-step protocol for text data with the addition of the use of CrossMatch using the UniVec database to remove contaminating vector sequences, and the use of the BLAST module for the elimination of already known *Arachis* markers (as described for the processing of ESTs).

RESULTS AND DISCUSSION

The number of molecular markers available for *Arachis*, and many other species, is still a limiting factor for genetic mapping. When we had begun working with microsatellite marker development we could not find a software that dealt in an integrated way with the problems that marker development presented, most importantly: the identification of microsatellite containing sequences, the removal of vector sequences and known marker sequences, the processing of sequences into non-redundant contigs, and a large-scale interface with Primer3. To overcome this we developed the softwares published here. Recently we published new microsatellite markers developed using this software in various stages of development, together with the first microsatellite-based genetic map for *Arachis*. In total 271 new microsatellite markers were developed, 69 data-mined from GenBank, 121 from microsatellite enriched genomic libraries and 81 from ESTs (12). The software in the final stage of development described here solves the problems we encountered and allows the efficient data-mining of new markers.

The TROLL module integrates simply and completely into the Staden Package and performed its task extremely fast. By concatenating the input sequences and calling the TROLL program only once, the module finds all SSRs in a single pass, thus maintaining TROLL's linear processing time. TROLL's simplicity and specialized function is targeted exactly at the types of repeats that are most useful for microsatellite marker development: repeats with di-, tri-, tetra- and penta-nucleotide motif sizes. Although TROLL does not find imperfect repeats, we did not find this to be a problem. Long imperfect repeats largely consist of shorter perfect ones, and isolated short imperfect repeats are of very limited use for marker development. The primer design program worked well and, when used with chromatogram data and quality control, effectively eliminated primers being designed to low quality regions of the consensus sequences (for an example output see Table 1). Here we illustrate the use of the module and primer design program with three different datasets, each of which presenting slightly different difficulties for marker development. Staden working with the module proved effective in all cases.

TC-repeat enriched genomic library

Processing the 338 chromatograms produced 43 microsatellite containing contigs. This type of data where every sequence contains a microsatellite, and many of the sequence reads 'die-out' at the end of the microsatellite is difficult to assemble correctly. Incorrect joining of reads that end in microsatellites can easily be a problem. The use of masking in the assembly avoided this problem. The primer design program produced primer pairs for 30 microsatellite markers. The processing of these sequences is included just for demonstration, markers from these sequences are already published (12). Only one of the primer pairs (3.3%) failed to amplify an identifiable PCR product.

Table 1. Example output from the primer design program formatted and simplified

Contig name	Microsatellite	Product size	FWD primer	T_m	REV primer	T_m
AS1RM8P1E08	(CGG)6	218	GAACGCGTCAAAGAACATAACA	60.173	TTAACAACAGCAACAGCTTCGT	59.984
AS2RM11P1H01	(AG)7	352	AAGAAACGACGGTGAATCTGTT	60.040	GCCTTTTCTTCTCTCTCCCTCT	59.621
AS1RM15P1H08	(TC)6	266	GTAGTGGCAGAGCCTGTCTTCT	60.081	GTGGCCAATCTGTAAATCCAAT	60.082
AS1RM15P1H08	(CAG)6	309	TCTTCGCCTTGGTATATTGGAT	59.829	CAGGGGTGTAAGTGCAGTGTAA	60.088
AS1RM14P1B10	(CAA)8	316	ATCGAATCCCAACTCACTCACT	60.001	GCTTGGGTTCTTGTGGAGTAAC	60.040
AS1RN11P1C04	(TA)6	372	CGGGGAAATTCTTTTATATTTCA	58.501	ATTTTGTAAAGGATTCGCTCGT	59.188
AS1RM11P1A04	(TTC)7	181	GAGACCTTATGGTGGATCTTGG	59.827	AACGCTCCGTACAAGAGAAGAG	60.075
AS2RM7P1E10	(AATA)6	170	CACGTGAAAGTTCATATCGTGTC	59.553	ATTCATGATTCCTTACGCGACT	59.993
AS1RM9P1B07	(ATC)8	322	TGGGGAAATCTTAACACAAAGG	60.210	GCAAGAAGGAGTTGAAGAAGGA	59.998
AS2RM7P1A09	(AT)8	357	AATCAACTTTGTGGACTTGTAA	54.102	AAGCATAGGGAAAATGAAAT	53.159
AS2RM11P1H07	(GAA)6	220	CGATAAAGTCTCAGGTGAGCAA	59.517	AGAGGGAGTGGTGAATGAATGT	59.862
AS2RM6P1F03	(CT)7	135	TCCTTCAAGCTCGTCTTCTACC	60.019	GAAATCTGACGCAATGTTCAAG	59.752
AS1RN26P1G09	(TTA)20	234	GTTATTGCTGTGGTGAAGAACG	59.677	TTGACTTTCATATGCACCTCTCA	59.761
AS2RM9P1B06	(AT)8	195	TAAAGTTGATCTTGAATCTTCC	52.967	CGCGATCATTTATTTAACTT	52.482
AS2RM1P1A05	(CTT)8	209	TCTTCCAATTCTCTCTGCCTTC	59.962	ATGGAGTGACCAACATAAACCC	59.983

Raw output from the program is a tab-spaced file with, in addition to the information above, primer names and base qualities of the contig regions that bind to the primers.

ESTs

Of the 8800 ESTs, 282 were above the specified threshold of similarity to known *Arachis* SSR markers and were eliminated by the BLAST module. The TROLL module detected 213 other chromatograms with microsatellites. Using the Gap4 interface these were assembled into 142 contigs. Although the original dataset is large, the resultant database is small, and is operated efficiently on a standard personal computer. The primer design program produced primer pairs for 75 new microsatellite markers. These primer pairs have not been fully tested yet, but in our previous work (12) only 6 of 81 EST primer pairs (7.4%) failed to amplify an identifiable PCR product.

Text data from GenBank

A total of 563 of the 3890 sequences from GenBank were above the specified threshold of similarity to known *Arachis* markers and were eliminated by the BLAST module, TROLL detected 482 other sequences with microsatellites and these were assembled into 463 contigs. Although counter-intuitive, it is possible to use quality control even with text data. Staden assigns a nominal quality value of two for bases of text data, but when text data overlaps within a contig, a higher value of confidence is given. This can be used, for instance, to design primers only using consensus sequences supported by at least two reads. However, since the GenBank data we processed has low redundancy, primers were designed without quality control and primer pairs for 352 markers were designed. It is to be expected that data from GenBank may give a lower success rate of primers compared to chromatogram data. Some sequences may have been rejected for marker development by the original authors because the repeat type has relatively low polymorphism, or because of low sequence quality. Nevertheless, in our previous work, data-mining from GenBank was a productive and low cost option for marker development. Of the 69 markers data-mined, only 9 (13%) did not amplify any scorable fragments (12).

Example databases and the new primers designed are in Supplementary Data. In total, 427 pairs of primers for new candidate markers for *Arachis* are presented.

In summary, advances in genetic mapping are reliant upon the development of sufficient numbers of molecular markers, and large scale marker development depends upon efficient processing of the data. The Staden software with the TROLL module and primer design program provides a unified and efficient interface for this.

Availability

TROLL, the module and primer design program, and instructions for installation and use are available free and can be downloaded from the Resources section of http://finder.sourceforge.net/. A Web version of the primer design program is also available from the same link.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank the Conselho Nacional de Pesquisa e Desenvolvimento (Reference: 400617/2003-0), the Pró-Reitoria de Pós-Graduação e Pesquisa—Universidade Católica de Goiás, the Universidade Católica de Brasília, the European Union INCO Programme (ARAMAP reference: ICA4-2001-10072), and the Generation Challenge Program (Project number 31) for financial support. The authors also thank anonymous referees for their comments. Funding to pay the Open Access publication charges for this article was provided by Generation Challenge Programme.

Conflict of interest statement. None declared.

REFERENCES

1. Benson,G. (1999) Tandem repeats finder: a program to analyse DNA sequences. *Nucleic Acids Res*., **27**, 573–580.
2. Kolpakov,R., Bana,G. and Kucherov,G. (2003) mreps: efficient and flexible detection of tandem repeats in DNA. *Nucleic Acids Res*., **31**, 3672–3678.
3. Kurtz,S., Choudhuri,J., Ohlebusch,E., Schleiermacher,C., Stoye,J. and Giegerich,R. (2001) REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res*., **29**, 4633–4642.
4. Castelo,A.T., Martins,W.S. and Gao,G.R. (2002) Tandem repeat occurrence locator. *Bioinformatics*, **8**, 634–636.
5. Staden,R., Beal,K.F. and Bonfield,J.K. (1998) The Staden Package. In Misener,S. and Krawetz,S.A. (eds), *Computer Methods in Molecular Biology, Bioinformatics Methods and Protocols*. The Humana Press Inc., Totowa, NJ, Vol.132, pp.115–130.
6. Rozen,S. and Skaletsky,H. (2000) Primer3 on the WWW for general users and for biologist programmers. In Krawetz,S. and Misener,S. (eds), *Bioinformatics Methods and Protocols: Methods in Molecular Biology*. The Humana Press Inc., Totowa, NJ, pp. 365–386.
7. Rafalski,J.A., Vogel,J.M., Morgante,M., Powel,W., Andre,C. and Tingey,S.V. (1996) Generating and using DNA markers in plants. In Birren,B. and Lai,E. (eds), *Analysis of non-mammalian genomes: a practical guide*. Academic, NY, pp. 75–134.
8. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol*., **215**, 403–410.
9. Ferguson,M.E., Burow,M.D., Schulze,S.R., Bramel,P.J., Paterson,A.H., Kresovich,S. and Mitchell,S. (2004) Microsatellite identification and characterization in peanut (*A.hypogaea* L.). *Theor. Appl. Genet*., **108**, 1064–1070.
10. He,G., Meng,R., Newman,M., Gao,G., Pittman,R.N. and Prakash,C.S. (2003) Microsatellites as DNA markers in cultivated peanut (*A.hypogaea* L.). *BMC Plant Biol*., **3**, 3.
11. Moretzsohn,M.C., Hopkins,M.S., Mitchell,S.E., Kresovich,S., Valls,J.F.M. and Ferreira,M.E. (2004) Genetic diversity of peanut (*Arachis hypogaea* L.) and its wild relatives based on the analysis of hypervariable regions of the genome. *BMC Plant Biol*., **4**, 11.
12. Moretzsohn,M.C., Leoi,L., Proite,K., Guimarães,P.M., Leal-Bertioli,S.C.M., Gimenes,M.A., Martins,W.S., Valls,J.F.M., Grattapaglia,D. and Bertioli,D.J. (2005) A microsatellite based, gene-rich linkage map for the AA genome of *Arachis* (Fabaceae). *Theor. Appl. Genet*., **111**, 1060–1071.