Tutorial 3 – Selecting the best candidates for microsatellite marker development

1. Introduction

First of all, download the file `tutorial_data.tgz` (http://www.lbm.fmvz.usp.br /trap/tutorials/tutorial_data.tgz) to a directory of your choice. Decompress the file using the following command:

```
tar xzvf tutorial_data.tgz
```

or, alternatively…

```
gzip –d tutorial_data.tgz
```

and then… `tar xvf tutorial_data.tar`

This command will create the `tutorial_data` directory, which contains five subdirectories:

- `config_files` - this directory contains configuration files of three tutorials: `catalogue.cnf`, `annotation.cnf` and `markers.cnf`.
- `data` - this directory contains a multiple sequence FASTA file (`tutorial_data.fasta`) comprising five distinct sequences. Single sequence files of these five sequences are also provided in this directory:
    - `contig00000040.fasta`
    - `contig00000153.fasta`
    - `contig00000278.fasta`
    - `contig00000333.fasta`.
    - `contig00000382.fasta`
- `test` - this will be your working directory for the tutorial and is initially empty.
- `trf` - this directory contains output files previously generated by TRF, using the `tutorial_data.fasta` file as an input.
- `tutorials` – this directory contains three subdirectories (`catalogue_dir`, `annotation_dir` and `markers_dir`). Each one harbors the output files of the corresponding tutorial. We are providing these files just in case you have problems in running TRAP, and want to check how the output files should look for this example data set.

We have previously run TRF version 4.00 on a multiple sequence FASTA file (`tutorial_data.fasta`). As mentioned above, the output files are stored in the `/trf` directory. We used the Linux version of TRF. Command for invoking TRF may vary depending on the platform, version of TRF and configuration of the server. The following command was used in our case:

```
trf400.linux.exe tutorial_data.fasta 2 5 7 80 10 25 1000 –f
```

In this tutorial, we describe how to run TRAP in order to select the best candidates for microsatellite marker development. TRAP will be used to select repeat *loci* according to different user-defined criteria. As a result, TRAP will create TRF-like HTML files for the user to manually inspect the selected *loci*. In addition, TRAP will also create a FASTA file containing nucleotide sequences of the repeat *loci*, which can be conveniently used by any primer design software.

2. Running TRAP on the command line:

To run TRAP, first go to the `test` directory, and then type the following command:

```
trap.pl –i ../trf/tutorial_data.fasta –od markers_dir –of
markers_file –min 2 –max 7 –cpmin 5 –cpmax 20 -fs 100 –id 70 –
ff -trf
```

If everything goes well, you should now find the following directory in this directory:

```
markers_dir
```

Inside the `markers_dir` directory, you will find the following additional files:

```
markers_flanking_regions.txt
markers_file_TRAP_summary_TRF_like.html
```

…and the newly created `html_data/` directory.

3. Understanding TRAP parameters:

A comprehensive explanation on all TRAP parameters is depicted in the *How to run TRAP* document. Please refer to it if you need more information.

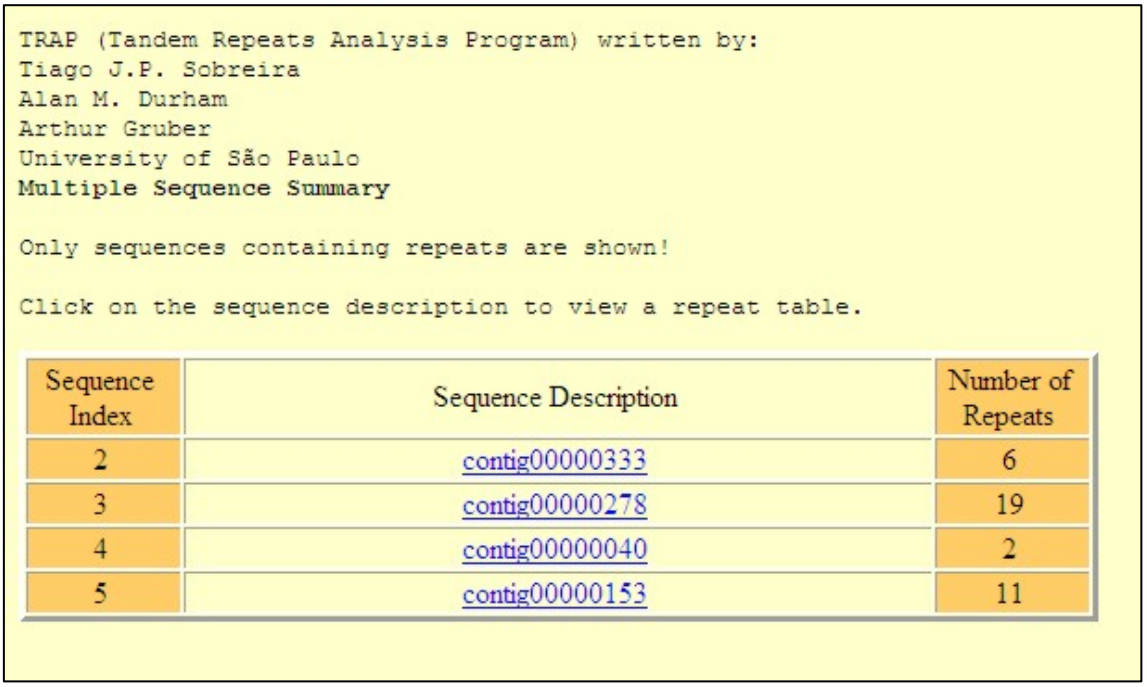
Shortly, the command line used above specifies the following parameters:

- `-i tutorial_data.fasta` – uses files with names containing the prefix `"tutorial_data.fasta"` (see the file names in the `trf` directory) for input.
- `-od markers_dir` – specifies `markers_dir` as the directory that will store TRAP output files.
- `-of markers_file` – specifies `markers_file` as the prefix for all TRAP output file names.
- `-min 5` and `-max 7 -` selects repeats with period sizes in the range of 2 to 7 bp.
- `-cpmin 2` and `cpmax 20` – selects repeat *loci* with copy number in the range of 5 to 20 repeat units.
- `-fs` – defines the minimum size of the 5' and 3' flanking regions of the repeat *locus*.
- `-id 70` – selects repeat *loci* with percentage of matches ≥ 70% between adjacent repeat units overall.
- `-ff` – generates a multiple sequence FASTA file for primer design purposes.
- `-trf` – creates output files that resemble the original TRF output HTML files, but displaying only the repeat *loci* selected by TRAP.

4. Understanding TRAP output files:

The following files are created:

- `markers_file_TRAP_summary_TRF_like.html`
  This is an index HTML file that lists the sequences containing repeat *loci* and respective numbers of *loci* per sequence (see figure below). It presents the same layout as TRF HTML files and allows the user to inspect the candidate *loci* selected by TRAP for microsatellite development.



TRAP (Tandem Repeats Analysis Program) written by:
Tiago J.P. Sobreira
Alan M. Durham
Arthur Gruber
University of São Paulo
Multiple Sequence Summary

Only sequences containing repeats are shown!

Click on the sequence description to view a repeat table.

Sequence Index	Sequence Description	Number of Repeats
2	contig00000333	6
3	contig00000278	19
4	contig00000040	2
5	contig00000153	11

Notice that the original input file (`tutorial_data.fasta`) used in this tutorial contains five sequences (`contig00000040.fasta`, `contig00000153.fasta`, `contig00000278.fasta`, `contig00000333.fasta` and `contig00000382.fasta`. However, only repeat *loci* for four of these sequences were selected by TRAP as candidates for microsatellite development. This means that the original input sequence `contig00000382` does not present any *locus* fulfilling the criteria used in the present selection made by TRAP.

Clicking, for instance, on the contig00000333 link, will pop up a new window displaying the repeat *loci*, with their respective characteristics, including coordinates, period size, copy number, etc. (see figure below).

Indices	Period Size	Copy Number	Consensus Size	Percent Matches	Percent Indels	Score	A	C	G	T
2786--2829	7	6.1	7	95	5	79	14	0	41	43
2943--3019	7	11.3	7	81	4	98	42	38	3	18
3109--3165	7	8.1	7	84	0	86	40	44	4	12
4395--4439	3	14.7	3	91	5	74	36	33	29	0
4594--4609	3	5.3	3	85	0	25	0	38	25	38
5302--5332	4	7.8	4	96	0	41	39	0	0	61

Clicking on any link of the indices row will open up another window with details on the selected repeat *locus*, including the flanking region sequences, repeat consensus sequence, etc. (see figure below).

```
Statistics
Matches: 36,  Mismatches: 0, Indels: 2
        0.95            0.00        0.05

Matches are distributed among these distances:
   7   29  0.81
   8    7  0.19

ACGTcount: A:0.14, C:0.00, G:0.41, T:0.43


Consensus pattern (7 bp):
TAGGGTT

Left flanking sequence: Indices 2286 -- 2785
TTTTTTAATAAAAAAATTAAACTGTATTCCAAAAGAATGGAGGCATGAGATCAATTTCAGATTGC
AATTCAAGATAAACTCATTTTTATTAATTTGCTCAATTTGCTGTTGAGGAATTAAAAATTCAATC
AAACTCTTACATTTAAACAATTTAAAAAATTCAACTTTTATAGTTTTGTAATTCACTTGTCTTTG
TGTCCAAATNAGTTGAAATGAATAAATCAGAGGCAAGCAGCANGCAATTAAACACCAGCAAAGCT
CAATAGTTAAACAGCAGCAGACAGCTGCAGCAGCAGCAGTAGTAGCAGCACCAGTAGCAGCATCA
GCAGCAGTAGTAGTAGTAGTAGTAGCAGCAGCAGCAGTAGTAGCAGCAGCAGTAGCAGCAGGAGC
AGCAGTAGCANGTAGCAGCAACTGTGTAAAACGAATTCCGGAAGGGCCGACTCAATAGGCTGCTG
CAGCAGCAAGTCCGATCTAATACATAAATAATACAAGGCAAGATC


Right flanking sequence: Indices 2830 -- 3329
GTTCTTTGCATTATCCAAAGTGAACAAGAAGAAATCAAAAAAAAAAATTAAAACGAATTCAAATAA
ATCCATTCTTGCTGCAAACAAACCAGCACATTTGCTTTCTAATTAATGCTAAACCTAAACCCTAA
ACCATAAACCCTCAACCCTAAACTCTAAACGCTAAACCCTAAACCCTAATGCTAAACCCTTATGT
TAGCAAACNGGATGGCATCTTTATAGGATAAATANTGAAANTGCATCTGAAGTTGTTTGGTTTTC
CATTTCGCACTTTGAGTGCAAACCCTAACCCGTAAACCCTAAACCCTAAACCCTAAACCCCGAAC
CCTAAACCCTAGCCTCCTAGCCCCTGTAACAGCTCTGCGTGTGTTTCCGAATTCATTCTTTGGCA
CTTCTTTGATTTTGCAATTCAAATCATTTATAATTTCAATTTATTTCAAATATCTTTTATTTTAT
TTATTTTAAAAGAAGTTAAAACCCCGCTGCGGCCGCTGGCTGCCG
```

- `markers_flanking_regions.txt`

  This is a FASTA format text file created for primer design application. For each selected repeat *locus*, TRAP generates a sequence composed by the nucleotide sequence of the 5' flanking region, a stretch of 20 Ns replacing the original repetitive bases, and the 3' flanking region. In this example, the file contains 38 nucleotide sequences with headers listed below:

  - `contig00000333 TRAP 2786 to 2829`
  - `contig00000333 TRAP 2943 to 3019`
  - `contig00000333 TRAP 3109 to 3165`
  - `contig00000333 TRAP 4395 to 4439`
  - `contig00000333 TRAP 4594 to 4609`
  - `contig00000333 TRAP 5302 to 5332`
  - `contig00000278 TRAP 107 to 149`
  - `contig00000278 TRAP 219 to 262`
  - `contig00000278 TRAP 341 to 370`
  - `contig00000278 TRAP 436 to 471`
  - `contig00000278 TRAP 486 to 518`
  - `contig00000278 TRAP 543 to 566`
  - `contig00000278 TRAP 617 to 636`
  - `contig00000278 TRAP 641 to 666`
  - `contig00000278 TRAP 731 to 769`
  - `contig00000278 TRAP 828 to 859`
  - `contig00000278 TRAP 888 to 908`
  - `contig00000278 TRAP 1117 to 1149`
  - `contig00000278 TRAP 1199 to 1250`
  - `contig00000278 TRAP 1338 to 1380`
  - `contig00000278 TRAP 1421 to 1436`
  - `contig00000278 TRAP 1475 to 1515`
  - `contig00000278 TRAP 1586 to 1612`
  - `contig00000278 TRAP 1646 to 1684`
  - `contig00000278 TRAP 2134 to 2155`
  - `contig00000040 TRAP 671 to 696`
  - `contig00000040 TRAP 880 to 899`
  - `contig00000153 TRAP 242 to 276`
  - `contig00000153 TRAP 245 to 270`
  - `contig00000153 TRAP 241 to 284`
  - `contig00000153 TRAP 533 to 565`
  - `contig00000153 TRAP 871 to 918`
  - `contig00000153 TRAP 1194 to 1243`
  - `contig00000153 TRAP 1233 to 1264`
  - `contig00000153 TRAP 1314 to 1357`
  - `contig00000153 TRAP 1652 to 1686`
  - `contig00000153 TRAP 2166 to 2204`
  - `contig00000153 TRAP 2296 to 2339`

The first row of the headers denotes the name of the original nucleotide sequence, followed by an indication that the FASTA was generated by TRAP, and by the coordinates of the respective satellite sequences.

5. Running TRAP with parameters specified in a configuration file:

The same data set and parameters of this tutorial can also be analyzed by TRAP using a configuration file. In this case, the only command line parameter is `-c`, which specifies the configuration file name. If you are in the `test` directory, you should use the following command:

```
trap.pl -c ../config_files/markers.cnf
```

The `markers.cnf` configuration file specifies exactly the same parameters used in the command line described above (see item 2). The configuration file content is listed below:

```
input_prefix = ../trf/tutorial_data.fasta
output_directory = markers_dir
output_file_prefix = markers_file
minimum_period_size = 2
maximum_period_size = 7
minimum_copy_number = 5
maximum_copy_number = 20
define_repeat_sequence =
minimum_flanking_region_size = 100
minimum_match_percentage = 70
table_format =
sort_field =
create_consensus_file = no
create_feature_table = no
create_gff = no
create_flanking_region_file = yes
create_trf_like_file = yes
create_redundancy_report = no
```