Tutorial 2 – Identifying microsatellite *loci* and generating feature table files for automated annotation

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 identify and automatically annotate microsatellite *loci*, according to user-defined criteria. TRAP will generate feature table files that can be opened in annotation editors like Artemis. You can download Artemis from the Sanger Institute web site at the address http://www.sanger.ac.uk/Software/Artemis/.

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 annotation_dir –of
annotation_file –min 2 –max 7 –cpmin 2 –id 80 –ft -gff
```

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

```
annotation_dir/feature_table
annotation_dir/feature_table
```

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

```
annotation_file_tutorial_data.fasta.s2_contig00000333.tab
annotation_file_tutorial_data.fasta.s3_contig00000278.tab
annotation_file_tutorial_data.fasta.s4_contig00000040.tab
annotation_file_tutorial_data.fasta.s5_contig00000153.tab
annotation_file_index.csv
```

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

```
annotation_file_tutorial_data.fasta.s2_contig00000333.gff
annotation_file_tutorial_data.fasta.s3_contig00000278.gff
annotation_file_tutorial_data.fasta.s4_contig00000040.gff
annotation_file_tutorial_data.fasta.s5_contig00000153.gff
annotation_file_index.csv
```

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 annotation_dir` – specifies `annotation_dir` as the directory that will store TRAP output files.
- `-of annotation_file` – specifies `annotation_file` as the prefix for all TRAP output file names.
- `-min 2` and `-max 7 -` selects repeats with period sizes within 2 to 7 bp.
- `-cpmin` 2 – selects repeat *loci* with copy number ≥ 2.
- `-id 80` – selects repeat *loci* with percentage of matches ≥ 80% between adjacent repeat units overall.
- `-ft` – creates feature table files for those input sequences containing repeat *loci* that selected by TRAP. The tab files, together with the corresponding FASTA-format sequences can be loaded onto Artemis.

- `-gff` – creates GFF files for those input sequences containing repeat *loci* that selected by TRAP. The GFF files, together with the corresponding FASTA-format sequences can be loaded on an annotation editor such as Apollo and Artemis.

4. Understanding TRAP output files:

The following files are created:

- `annotation_file_index.csv`

This file describes the correspondence between the `tab` file names and the sequence names. Sequence names are extracted from the FASTA headers of the multiple sequence FASTA file used as an input for TRF (note: CSV files can be opened by any spreadsheet program).

- `annotation_file_tutorial_data.fasta.s2_contig00000333.tab`
- `annotation_file_tutorial_data.fasta.s3_contig00000278.tab`
- `annotation_file_tutorial_data.fasta.s4_contig00000040.tab`
- `annotation_file_tutorial_data.fasta.s5_contig00000153.tab`

These `tab` files contain annotation data that can be directly submitted to NCBI/EMBL/DDBJ data banks or used as input visualized on annotation editors like Artemis.

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, you should use only the command line parameter `–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/annotation.cnf
```

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

```
input_prefix = ../trf/tutorial_data.fasta
output_directory = annotation_dir
output_file_prefix = annotation_file
minimum_period_size = 2
maximum_period_size = 7
minimum_copy_number = 2
maximum_copy_number =
define_repeat_sequence =
minimum_flanking_region_size =
minimum_match_percentage = 80
table_format =
sort_field =
create_consensus_file = no
create_feature_table = yes
create_gff = yes
create_flanking_region_file = no
create_trf_like_file = no
create_redundancy_report = no
```

6. Opening the annotation files generated by TRAP on Artemis annotation tool

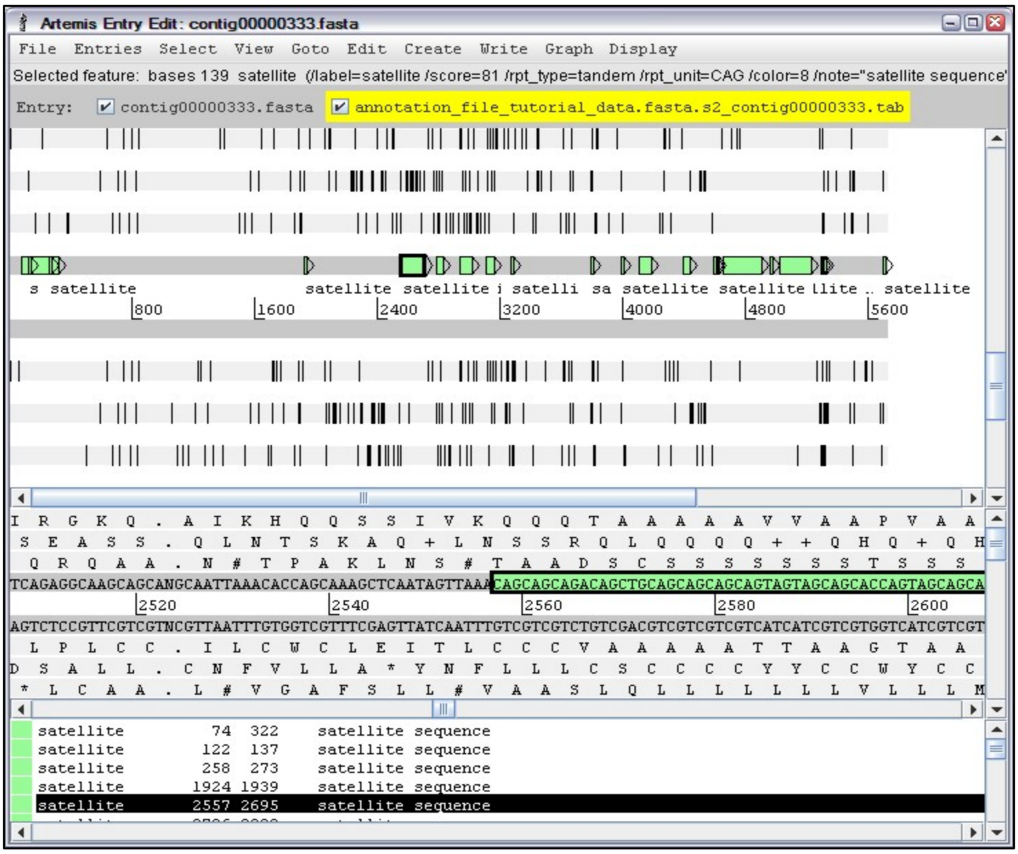
In order to visualize the automated annotation generated by TRAP, you should load both a sequence in FASTA format and the corresponding annotation tab files. For this purpose, we have generated individual FASTA files from the original multiple sequence file (`tutorial_data`):

```
contig00000040.fasta
contig00000278.fasta
contig00000382.fasta
contig00000153.fasta
contig00000333.fasta
```

These files are stored in the `tutorial_data/data` directory. Please notice that there are five FASTA sequence files, but only four annotation `tab` files. The sequence of `contig00000382.fasta` did not present any repeat *locus* in conformity with the user-defined parameters of this example. For this reason, no corresponding `tab` file was generated.

To visualize the annotation files on Artemis proceed with the following steps:

1. Invoke Artemis with the command `art`.
2. Use the command `File Open contig00000333.fasta.`
3. Once the sequence is loaded and displayed on Artemis, use the command `File Read An Entry...` and load the file `annotation_file_tutorial_data.fasta.s2_contig00000333.tab`
4. You will be able to visualize the sequence and the corresponding annotation of the tandem repeats. See figure below:


Artemis Entry Edit: contig00000333.fasta
File Entries Select View Goto Edit Create Write Graph Display
s satellite
satellite satellite i satelli sa satellite satellite llite .. satellite
800
1600
2400
3200
4000
4800
5600
2520
2540
2560
2580
2600

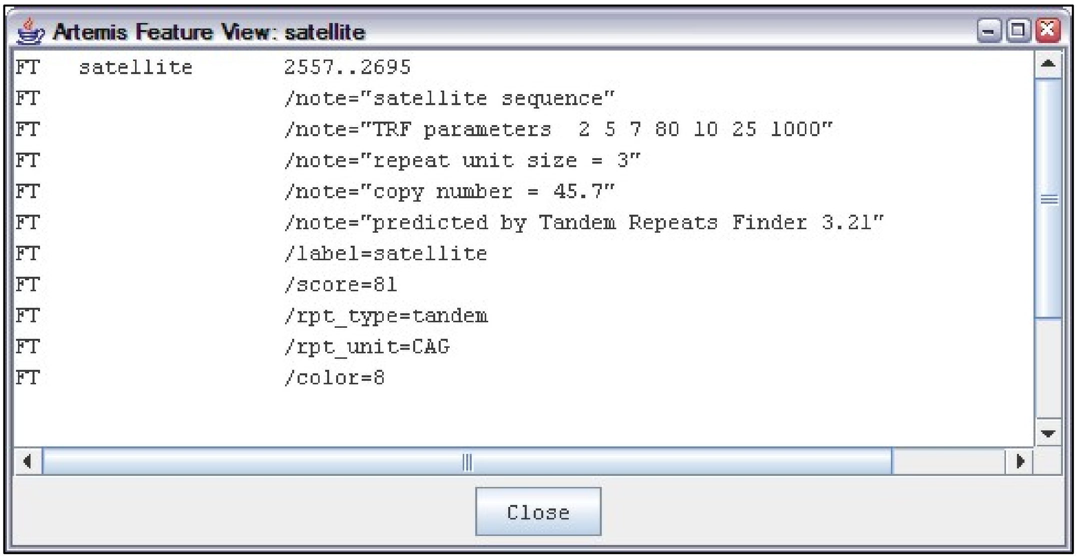
Selected feature: bases 139 satellite (/label=satellite /score=81 /rpt_type=tandem /rpt_unit=CAG /color=8 /note="satellite sequence'

Entry: ☑ contig00000333.fasta ☑ annotation_file_tutorial_data.fasta.s2_contig00000333.tab

```
I  R  G  K  Q  .  A  I  K  H  Q  Q  S  S  I  V  K  Q  Q  Q  T  A  A  A  A  A  V  V  A  A  P  V  A  A
 S  E  A  S  S  .  Q  L  N  T  S  K  A  Q  +  L  N  S  S  R  Q  L  Q  Q  Q  Q  +  +  Q  H  Q  +  Q  H
  Q  R  Q  A  A  .  N  #  T  P  A  K  L  N  S  #  T  A  A  D  S  C  S  S  S  S  S  S  S  T  S  S  S
TCAGAGGCAAGCAGCANGCAATTAAACACCAGCAAAGCTCAATAGTTAAACAGCAGCAGACAGCTGCAGCAGCAGCAGTAGTAGCAGCACCAGTAGCAGCA
AGTCTCCGTTCGTCGTNCGTTAATTTGTGGTCGTTTCGAGTTATCAATTTGTCGTCGTCTGTCGACGTCGTCGTCGTCATCATCGTCGTGGTCATCGTCGT
  L  P  L  C  C  .  I  L  C  W  C  L  E  I  T  L  C  C  C  V  A  A  A  A  A  T  T  A  A  G  T  A  A
D  S  A  L  L  .  C  N  F  V  L  L  A  *  Y  N  F  L  L  L  C  S  C  C  C  C  Y  Y  C  C  W  Y  C  C
 *  L  C  A  A  .  L  #  V  G  A  F  S  L  L  #  V  A  A  S  L  Q  L  L  L  L  L  L  L  V  L  L  L  M
```

```
satellite        74  322    satellite sequence
satellite       122  137    satellite sequence
satellite       258  273    satellite sequence
satellite      1924 1939    satellite sequence
satellite      2557 2695    satellite sequence
```

5. Select the satellite sequence comprised between coordinates 2557 and 2695 by clicking with the mouse on the corresponding feature box.
6. Now click on the `View` command of the main menu bar and select `View Selected Features`.
7. A window will pop up showing the automatic annotation generated by TRAP for this sequence.



```
FT   satellite       2557..2695
FT                   /note="satellite sequence"
FT                   /note="TRF parameters  2 5 7 80 10 25 1000"
FT                   /note="repeat unit size = 3"
FT                   /note="copy number = 45.7"
FT                   /note="predicted by Tandem Repeats Finder 3.21"
FT                   /label=satellite
FT                   /score=81
FT                   /rpt_type=tandem
FT                   /rpt_unit=CAG
FT                   /color=8
```