

DOCUMENTATION ON THE INPUT AND OUTPUT OF FIRE μ SAT₂

Various terminology pertaining to microsatellites, minisatellites, satellites and tandem repeats have been introduced by the academic community. The terminology used by us that is reflected by our input parameters as well as by our output is as follows. A tandem repeat (TR) in a genomic sequence is defined as a string of nucleotides that is characterized by a certain motif which introduces the string, followed by at least one “copy” of the motif. If the copies of the motif are exact, then it is called a perfect tandem repeat (PTR); otherwise it is called an approximate tandem repeat (ATR). An ATR is thus a string of nucleotides repeated consecutively at least twice, where a limited number of small differences between the repeated instances being tolerated, as in the example: **ACGACTACGACGAC**. Reference to a TR should be construed as a reference to either a PTR or an ATR. A TR element (TRE) that matches the identified motif of the TR will be referred to as a PTR element (PTRE). A TRE that does not match the motif is referred to as an ATR element (ATRE). In this example **ACGACTACGACGAC**, **ACG** is considered to be the introductory motif or PTRE. **ACT** is a “copy” of **ACG** where a mutation (mismatch) has occurred - **G** is replaced with **T**. Therefore in the context of the provided example **ACT** is an ATRE that resulted from a mismatch. The motif length is the number of nucleotides the motif consists of. In the case of the example the motif is **ACG** and the motif length is 3. Three types of motif errors are allowed namely deletions, insertions and mismatches. If the motif error = 1 then exactly one mismatch OR one deletion OR one insertion is allowed. If the motif is **ACG** then the ATREs include **ACT** (a mismatch has occurred) **AG** (a deletion has occurred) and **ACGT** (an insertion has occurred). The motif error is formally defined in the papers that are available from the links on our web site.

The Fire μ Sat₂ Windows GUI contains the following edit boxes:

- *Source File*
The user should enter in this edit box, the path to the genetic file (in FASTA format) which is to be scanned for TRs.
- *Output File*
The user should specify the path to a directory where the output file will be placed. The output file can be specified to be written in either a .txt format or a .csv format.
- *Flanking Sequence*
The flanking sequence the user requires should be specified in this edit box. The flanking sequence is a certain number of nucleotides that are output before and after a detected TR.
- *Motif Length*
In the edit box *Motif Length* the length of the motif should be specified. The value can be selected in the combo box on the right of the *Motif Length* edit box.
- *Max Motif Error*
The maximum number of mutations allowed per motif can be selected by clicking on the combo box on the right of the edit box *Max Motif Error*.
- *Max adjacent ATR elements*
The edit box *Max adjacent ATR elements* provides to the users the option of entering a value that indicates the maximum number of ATREs (approximate tandem repeat elements) that are allowed to occur next to

each other. The user can manipulate the number of ATREs that occur adjacently by the spin control on the right of the edit box *Max adjacent ATR elements*. The theoretical detail regarding the maximum number of consecutive ATR elements that are allowed can be found from the links to our published papers provided on our web site.

- *Motif range option*

The motif range option can be activated by clicking on the check box that appears beneath the text box *Motif range option*. If the user activates this option then the *Start motif* and *End motif* turns to black instead of grey as it was before the option had been activated. The motif range option enables the user to specify a range of motifs Fire μ Sat₂ should search for. However, the motifs should all be of the same length that corresponds to the length specified in the *Motif Length* edit box. The start motif should always be lexicographically smaller or equal to the end motif. The system will search for TRs in the corresponding lexicographic range.

If TRs for one specific motif only are needed, then the same string should be given for both the start and end motif.

- *Min required TR elements*

To avoid the output of unwanted data, the user may indicate the minimum number of TREs that has to occur before a TR is output. The edit box *Min required TR elements* can be set by using the spin control on the right of it.

- *Substring Error Options*

The *substring error* (discussed in DocumentationSubstringError.pdf) is computed at appropriate points by Fire μ Sat₂ and then compared against a user-specified threshold value given in the edit box *Max substring error*. During processing the calculated substring error should always be smaller than the indicated substring error. The value of the maximum substring error allowed can be modified by using the spin control on the right of the edit box *Max substring error*.

In line with the guidelines suggested by [?], the value of the substring error depends, *inter alia* on penalties (or weights) allocated by the user to mismatches, deletions and insertions respectively. Underneath the edit box *Max substring error* are three additional edit boxes where the values of the different penalties can be adjusted according to the requirements of the user. The *Mismatch penalty* edit box, the *Deletion penalty* edit box as well as the *Insertion penalty* edit box have combo boxes on their right that serve to modify the three different penalties respectively. The value of σ should always be smaller than the threshold value specified by the user.

The user may rely on system default values for the penalties. These appear automatically in the three penalty edit boxes as follows:

- *Mismatch penalty*: 0.5.
- *Delete penalty*: 1.0.
- *Insert penalty*: 1.0.

A penalty weight of 0 may be chosen for one or more of the mutation types, in which case no penalty is assigned to ATREs that derive from that mutation type. The range of the penalty values is ≥ 0 and ≤ 1 .

motif	pos	len	TR	n_ptre	n_atre	n_m	n_d	n_i
GAT	6495	15	GATAATTATAATTAT	1	4	4	0	0
GCT	4507	15	GCTCCTGCAGCTGCC	2	3	3	0	0
GGT	7921	18	GGTGCTGGTCTGGTAGT	3	3	3	0	0
GTA	6138	29	GTAGGTAGTATTATACTAGGTAAGTAGGA	2	7	3	1	3
GTA	6142	25	GTAGTATTATACTAGGTAAGTAGGA	2	6	3	1	2

TABLE 1. Sample output generated by Fire μ Sat₂

- *Execute*

Finally the *Execute* button appears at the bottom of the window. When the user is satisfied that the input is correct, then the execute button should be selected. Fire μ Sat₂ will then create an output file of either type .txt or type .csv in the directory specified by the user. This output is discussed in the next section.

0.1. **The output of Fire μ Sat₂.** The output of Fire μ Sat₂ is a text file in comma separated value format. This implies that each data object is separated by an *coln* character(s) (line feed and a carriage return character). The data fields, within each data object, are separated by commas. The file format as described corresponds to prerequisites of .csv extension files.

In general a spreadsheet program displays the data of the relevant file in a sheet in which data objects are in separate rows. The corresponding comma-delimited fields of the respective data objects are displayed in separate columns. The heading of each column may constitute a brief description of its contents. The data of corresponding fields of each data object is written in rectangular cells underneath each other filling out the contents of the respective columns. In Table 1 an example is given of five lines of output generated by Fire μ Sat₂ in .csv file format.

Fire μ Sat₂ generates nine different columns as output. The nine different columns have appropriate headers. The headers and a brief explanation of the contents follows below:

- *Column 1: motif*

The header of the first column is *motif*. The output in the first column of each of the five rows is alphabetic character string to indicate the introductory motif of the detected TRs.

- *Column 2: pos*

The header of the second column is *pos*. This is an integer value giving the offset index position of the respective detected TRs.

- *Column 3: len*

The header, *len*, of the third column refers to the word length—i.e. the lengths of the respective detected TRs.

- *Column 4: TR*

The fourth column has the header *TR* and contains the concatenation of alphabetical characters that constitute the detected TR.

- *Column 5: n_ptre*

The header of the fifth column *n_ptre* refers to the number of TREs that are exact copies of the introductory motif within the detected TR.

- *Column 6: n_atre*

The sixth column gives the number of ATREs that were counted within the detected TR.

- *Column 7: n_m*
The seventh column gives the number of mismatches that occurred within the detected TR.
- *Column 8: n_d*
The eight column contains the number of TREs considered to be deletions in the detected TR.
- *Column 9: n_i*
The last column, column 9, indicates the number of TREs considered to be insertions.

REFERENCES

1. G. Benson, *Tandem repeats finder*, Nucleic acids research **27** (1999), no. 2, 573 – 580.

DNA-ALGO GROUP, FASTAR-RESEARCH GROUP, DEPARTMENT OF COMPUTER SCIENCE, SCHOOL OF INFORMATION TECHNOLOGY, UNIVERSITY OF PRETORIA, SOUTH-AFRICA.
E-mail address: `driddc@unisa.ac.za`