XSTREAM version 1.73 Documentation

Update December 23, 2011

Aaron Newman and Jim Cooper
Biomolecular Science and Engineering Program
University of California, Santa Barbara

INTRODUCTION

This package contains the command line version of XSTREAM, a software tool designed for rapid identification and architecture modeling of Tandem Repeat (TR) patterns in large protein sequence datasets. XSTREAM can also be used to detect TRs in nucleotide or general alphanumeric sequences. This document describes how to run XSTREAM on a local computer, modify parameter values, and interpret output.

NOTE

For a detailed description of the XSTREAM software tool please refer to: Newman AM, Cooper JB. **XSTREAM: a practical algorithm for identification and architecture modeling of tandem repeats in protein sequences.** *BMC Bioinformatics* 2007 8:382

REQUIREMENTS

Java Virtual Machine version 1.5 or higher
For large sequence files, at least 1 GB RAM

USAGE

XSTREAM processes FASTA formatted sequences only, for example:

```
>identifier
IMATRIMATRIMATRIMATR
```

To run multiple sequences, make a list and save them.

XSTREAM was coded in java and is currently only available as a command line tool. To run XSTREAM with default settings, execute:

```
java -Xmx1000m -Xms1000m -jar xstream.jar myFile.txt
```

Output will be saved in the directory with the XSTREAM executable (to modify, see Parameters below). Since the Java Virtual Machine only allocates 64 Mb by default, you will need to increase the memory available to XSTREAM by using the Xmx,Xms parameters. If you are using 32 bit Windows, the maximum amount of memory you can assign XSTREAM is 1.6 GB, or -Xmx1600m -Xms1600m.

Some effort has been applied toward internationalizing XSTREAM. If XSTREAM crashes immediately, try running it with the JVM argument ‘-Duser.language=en’.

PARAMETERS

To print a list of command line parameters, run XSTREAM without specifying a file or any parameter changes (i.e. no arguments).

The following list briefly describes all parameters. Parameters colored red are described more thoroughly in the XSTREAM paper. Default values are in parentheses.

-**n** change settings to process nucleotide sequence [-i.8 -m10]
-**i** minimum word match (.7)
-**I** minimum consensus match (.8)
-**e** minimum copy number (2)
-**g** maximum gaps (3)
-**D** indel error (.5)
-**m** minimum period (3)
-x maximum period (1/2 of input sequence length)
-**L** minimum TR domain length (10) [e.g., TR is valid with period=2, 5 copies]
-**P** minimum TR coverage [=TR domain length/input sequence length]
-T min. input sequence coverage [=total TR coverage excluding overlaps]
-**B** generate TR block diagram output [PNG format]
-d specify output file directory (current directory) [e.g., -dC:\TRseqs]
-z create excel spreadsheet of TR output
-a insert custom name into output files [e.g., -aMyOutput]
-o disable removal of overlapping TRs
-q disable merging of similar TR domains
-c disable TR consensus clustering, if -n, -c option will instead enable it
-A use substitution alphabet [see page 4 below, e.g. –AC:\sub.txt]
-t multithreading: use more than one CPU (default=1) [e.g., -t4; see pg. 4]
-**f** invoke divide & conquer [done automatically for nucleotide sequences >1Mb], -f[integer], e.g. -f1000 sets D&C fragment length to 1000 (pg. 4)
-G generate html block diagrams hyperlinked to TR multiple alignments
-O change multiple alignment color format to highlight gaps/mismatches
-N disable detection of nested TRs
-V disable automatic invocation of D&C for long nucleotide sequences
-C write output to console
-**W** modify all other parameters in the following way:
-W[short seed length, medium seed length, long seed length, Dynamic Programming (DP) match score, DP miss penalty, DP gap penalty, consecutive gap max (g*), Merging max period breadth, Merging max non-overlap space, Merging non-overlap period fraction, threshold for replacing characters with ‘X’ when merging, Redundancy Removal (RR) α, RR β, RR γ, RR δ, 2-stage period threshold (T), long period filter (t)]
e.g., with default settings: -W[3,5,7,2,-4,-4,3,50,10,.25,.5,.9,.75,.9,.6,10,20]

RECOMMENDED SETTINGS and ESTIMATED P-VALUES
The following tables provide recommended settings and estimated P-values for TRs identified from **protein sequences**:

TR degeneracy	Min Word Match (*i*)	Min Consensus Match (*I*)	Max Gaps (*g*)
High (H)	0.7	0.7	3
Moderate (M)*	0.7	0.8	3
None (N)	1	1	0

TR Significance	Min Copy No. (*e*)	Min Period (*m*)	Min Domain Length (*L*)	P-value (N)	P-value (M)	P-value (H)
Very High	3	3	20	$<10^{-5}$	$<10^{-4}$	$<10^{-4}$
High*	2	2	10	$<10^{-4}$	<0.02	<0.02
Moderate	2	1	5	<0.1	<0.1	<0.1

*=default settings

P-value cutoffs are color-coded according to the TR degeneracy level that they represent. We calculated P-value cutoffs for TR significance by first randomly shuffling characters within every protein sequence from the large and diverse NCBI non-redundant (NR) protein dataset (~7M sequences, downloaded April 2008) XSTREAM was run on all random sequences using the three different colored degeneracy settings shown above (High, Moderate, None). P-values were estimated by taking the number of protein sequences identified by XSTREAM for different ranges of TR periods and copy numbers, and dividing that count by the total number of randomized proteins in the NR dataset. This analysis was repeated for each parameter set using two independently randomized versions of NR, and the results were completely reproducible (R^2 ~ 1).

ADDITIONAL USEFUL PARAMETERS

Tandem Repeat Proteins

Parameters for detecting proteins with high TR content:

-P By setting minimum TR coverage high, one can restrict the output to tandem repeat proteins (TRPs), for example all proteins containing at least one TR covering at least 2/3 of the protein sequence.

-T Similar to P, but combines the protein coverage of all TRs in a given sequence, which is useful for identifying multi-domain TRPs.

Substitution Alphabet

Included in the package is a sample substitution alphabet text file (sub.txt) that defines character equivalence classes. This powerful feature forces XSTREAM to treat similar amino acids (e.g. isoleucine and leucine) as the same character during TR detection. To define a substitution matrix, simply place equivalent characters on the same line in a text file. Then, import the substitution alphabet using the command -AC:\yourFile.txt, for example.

Command Line Examples

1) Find TRs with ≥30% sequence coverage, ≥10 copies, domain length ≥20, and save output in C:\TRs

```
java -Xmx1000m -Xms1000m -jar XSTREAM.jar myFile.txt -P.3 -e10 -L20 -dC:\TRs
```

2) Find nucleotide TRs with maximum period=1000, use D&C with fragment length=100000, enable consensus clustering, and add “yeast” identifier to output files

```
java -Xmx1000m -Xms1000m -jar XSTREAM.jar myFile.txt –n –x1000 –f100000 -c -ayeast
```

Multithreading

XSTREAM is coded to take advantage of parallel processing. To use multiple CPUs, simply invoke XSTREAM with the parameter –tX, where X= number of processors and X is an even number.

Divide and Conquer

For long input sequences such as eukaryotic chromosomes, XSTREAM uses Divide and Conquer (D&C) to save memory. The length of the overlap between sequence fragments is the maximum period. Use of D&C is highly recommended for nucleotide sequences, and is thus invoked by default.

Graphical Output and Spreadsheet

To fully utilize the output of XSTREAM, one can generate colorful block diagram schematics (-B option), which depict TR architectures in protein sequences. All protein sequences are normalized to the same length. In addition, the spreadsheet output (-z) provides a convenient format for further analyses of the results.

HTML OUTPUT CLARIFICATION

For a general html output description, please see the paper.

1) Output 1 (TR Statistics):

i) Repeat Statistics Table:

Repeat Count = # unique TRs if consensus clustering was run
otherwise, = total TRs

TR Seq Count = # TR-containing input sequences (e.g. proteins)

ii) Tandem Repeats Table:

Repeat Column = TR ids; If consensus clustering was run (default): unique TRs = numerical ids, TRs belonging to same class are below numerical id and are denoted by the dash character (-).

iii) Running Time (botton of page) =

Total TR detection, processing, and consensus clustering time (excluding time required to write output).

If *benchmarking* XSTREAM, it is recommended to disable consensus clustering as this procedure is not a necessary part of TR identification and will result in less accurate running time output.

2) Output 2 (Multiple Alignments):

If consensus clustering was run, TR multiple alignments of the same class are listed next to one another. For each non-singleton TR class, a table is given underneath the Repeat id showing the number of input sequences and multiple alignments (instances) within the TR class.

TROUBLESHOOTING

Although error checking has not been thoroughly implemented, XSTREAM has been extensively tested. If XSTREAM crashes, please double-check your command line arguments, as any mistakes here will likely result in an error. If you believe XSTREAM has legitimately crashed on your input sequence(s), we would like to know. Please email the XSTREAM team at xstreamweb@gmail.com with your feedback and any bugs you happen to encounter. Thanks, and have fun!