MetagenomeThreader
a manual

David Schmitz-Hübsch
Stefan Kurtz

Research Group for Genomeinformatics
Center for Bioinformatics
University of Hamburg
Bundesstrasse 43
20146 Hamburg
Germany

26/08/2013

1 Introduction

This document describes the *MetagenomeThreader*, a software tool for the prediction of genes in sequences of metagenome projects. The *MetagenomeThreader* is based on the algorithem from Krause et al [1]. For the prediction of PCSs (Predicted Coding Sequences) the metagenome sequences are blasted against a nucleotide database and the resultig BLAST-hits are used to process a combined score for every position in a DNA-sequence. The combined score reflects the potential that a specific DNA-sequence position is a coding one.

The *MetagenomeThreader* is written in `C` and is based on the *GenomeTools* library [2].
The *MetagenomeThreader* is called as part of the single binary named `gt`.

2 Usage

2.1 The meaning of used type of fonts

Some text is highlighted by different fonts according to the following rules.

- `Typewriter font` is used for the names of software tools.
- `Small typewriter font` is used for file names.
- `Footnote sized typewriter font` with a leading `'-'` is used for program options.
- *small italic font* is used for the argument(s) of an option.

2.2 The functionality of the *MetagenomeThreader*

The *MetagenomeThreader* needs 3 data sets, the metagenome sequences are also required. The BLAST-hits will then be generated through a BLAST of the metagenome-sequences against a nucleotide database such as nt from NCBI. **ATTENTION:** Generating the BLAST-hits is not part of the *MetagenomeThreader* and has to be done preliminary. The resulting BLAST-hits have to be saved in XML-format. If the BLAST-hits are generated, the hit-sequences can be achieved through the *MetagenomeThreader*. In order to generate the hit-sequence file, the *MetagenomeThreader* requires a local copy of the appropriate nucleotide database, such as nt from NCBI, which will be scanned during the *MetagenomeThreader* processing. You can download a nucleotide database from `ftp://ftp.ncbi.nih.gov/blast/db/FASTA` for example.

2.3 The Installation of the *MetagenomeThreader*

The *MetagenomeThreader* will be installed and is ready for use after the installation of the *GenomeTools* (see [2] for installation instructions). There are two different installation options for the *GenomeTools* in relation to the *MetagenomeThreader*. The standard installation installs the *MetagenomeThreader* without the cURL-module. If you prefer to install the *MetagenomeThreader* including the cURL-module, you have to include the the command `make curl="yes"` during the

GenomeTools installation process. A connection to the internet is required to use the curl module. With the curl-module, the MetagenomeThreader will run an online query for missing hit-sequences, otherwise they will be ignored during the processing of the PCS's.

2.4 The Options of the *MetagenomeThreader*

Since *MetagenomeThreader* is part of `gt`, the *MetagenomeThreader* is called as following:

`gt mgth` [*options*] `BLAST-Filename Metagenome-Filename Hit-Sequence-Filename`

`BLAST-Filename` denotes the XML formatted BLAST-hit file, `Metagenome-Filename` the FASTA formatted metagenome-sequences file and the `Hit-Sequence-Filename` the FASTA formatted hit-sequences file. The `BLAST-Filename` has to be created with a BLAST-Program. Because of the dimension of the resulting files in metagenome projects, an online BLAST is neither advisable nor practicable. The hit-sequence file may be empty at the time of the first program call. All three files have to be provided at every program call and can also be zipped files (`.gz`). An overview of all possible options with a short descripton of each option is available in table 1. All options can only be specified once.

Table 1: Overview of the *MetagenomeThreader* Options.

`-s`	specify score for a synonymic base exchange
`-n`	specify score for a nonsynonymic base exchange
`-b`	specify score for a blast-hit end within a query-sequence
`-q`	specify score for a stop-codon within a query-sequence
`-h`	specify score for a stop-codon within a hit-sequence
`-l`	specify score leaving a gene on forward/reverse strand or enter a gene on forward/reverse strand
`-p`	specify maximum span between coding-regions in the same reading frame resume as one prediction
`-f`	specify maximum span between coding-regions in different reading frames resume as coding-regions in the optimal reading-frame
`-o`	specify the database name for the fcgi-database used in the cURL-module
`-k`	specify the database name used for the hit-sequence extraction
`-t`	specify yes|no if a hit-sequence file exists
`-r`	specify 1|2|3 for the format of the output-file
`-a`	specify minimum length of the amino acid sequences in the result
`-d`	specify minimum percent-value for the different kinds in the hit-statistic output
`-e`	specify 1|2|3 for the use of alternative start-codons
`-m`	specify yes|no if the processing has to be based on homology - experimental
`-g`	specify yes|no for the *GenomeTools* test modus - output without the creating date
`-x`	specify yes|no to extend PCSs to max length
`-version`	display version information and exit
`-help`	show all options

2.5 *MetagenomeThreader* Options

`-s` *score*

Specify the score for synonymic base exchanges. *score* has to be specified as a double. If this option is not selected by the user, then *score* is set to 1.00 by default (recommended: positive score).

`-n` *score*

Specify the score for nonsynonymic base exchanges. *score* has to be specified as a double. If this option is not selected by the user, then *score* is set to -1.00 by default (recommended: negative score).

`-b` *score*

Specify the score for a blast-hit end within a query-sequence. *score* has to be specified as a double. If this option is not selected by the user, then *score* is set to -10.00 by default (recommended: negative score).

`-q` *score*

Specify the score for a stop-codon in the query-sequence. *score* has to be specified as a negative double. If this option is not selected by the user, then *score* is set to -2.00 by default (recommended: negative score).

`-h` *score*

Specify the score for a stop-codon in the hit-sequence. *score* has to be specified as a negative double. If this option is not selected by the user, then *score* is set to -5.00 by default (recommended: negative score).

`-l` *score*

Specify the score for leaving a gene on forward/reverse strand or enter a gene on forward/reverse strand. The argument *score* has to be specified as a negative double. If this option is not selected by the user, the default value is set to -2 (recommended: negative score).

`-p` S_{max}

Specify the maximum span between two coding-regions in the same reading frame in which they resume as one coding-region. The argument S_{max} has to be specified as a positive double. If this option is not selected by the user, S_{max} is set to 400.00 by default.

`-f` S_{max}

Specify the maximum span between coding-regions in different reading frames in which they resume as coding-regions in the optimal reading-frame. The argument S_{max} has to be specified as a positive double. If this option is not selected by the user, S_{max} is set to 200.00 by default.

`-c` DB_{name}

Specify the name of the database used in the cURL-Module. The argument DB_{name} has to be chosen as a valid database which you can find under `http://eutils.ncbi.nlm.nih.gov/entrez/eutils/einfo.fcgi?`. If this option is not selected by the user, DB_{name} is set to *nucleotide* by default.

`-o` *output filename*

Specify the name of the resulting output-file where the predictions will be written to. Each prediction will be represented by an individual block (see chapter 3). If this option is not selected by the user, *output filename* is set to $output$ by default.

`-k` $DB_{Hit-Seq}$

Specify the path to the hit-sequence database. The argument $DB_{Hit-Seq}$ has to be chosen as a valid path to a valid database which you can find under

`ftp://ftp.ncbi.nih.gov/blast/db/FASTA/` for example. If this option is not selected by the user, $DB_{Hit-Seq}$ is set to $nt.gz$ by default.

`-t` *yes|no*

Specify if a hit-sequence file was already created (*yes*) or not created until now (*no*). If this option is not selected by the user, `-t` is set to *no* by default. **ATTENTION:** If there is already a hit-sequence file and you do not use `-t` *yes* the existing hit-sequence file will be deleted.

`-r` *1|2|3*

Specify the format for the output-file. You can choose between .txt (1), .html (2) und .xml (3) format. If this option is not selected by the user, the default value of r is 1.

`-a` AA_{min}

Specify the minimum number of amino-acids in the amino-acid sequences in the output. Shorter amino-acid sequences will not appear in the output. The argument AA_{min} has to be chosen as a value higher or equal than 15. If this option is not selected by the user, the default value of a is 15.

`-d` $Stat_{min}$

Specify the minimum percentage which the species must have to appear in the output-file in the statistic section. The argument *similaritythreshold* has to be chosen from the range [0.0,1.0] and means a percentage. If this option is not selected by the user, the default value of d is 0.0.

`-e` *1|2|3*

Specify the format for the different start-codon modes. You can choose between (1): AUG, (2): AUG, CUG, GUG and (3): AUG, CUG, GUG, UUG. `-e` makes only sense if `-x` is set to *true*. If this option is not selected by the user, the default value of e is 1.

`-m` *yes|no*

Specify the processing mode. If `-m` is set to *yes*, the specified scores are used if there are equal amino acids, otherwise they are used if the amino acids are different. The `-m` has only experimental status. Results are not checked. If this option is not selected by the user, the default value of m is *no*.

`-g` *yes|no*

Specify if the testmode should be used to make the results comparable in the *GenomeTools* test suite. This option is not relevant for normal use of the *MetagenomeThreader*. If this option is not selected by the user, the default value of g is *no*.

`-x` *yes|no*

Specify if the extended mode should be switched on. If `-x` is set to *yes* the processed PCSs will be extended to its max length in both directions. If this option is not selected by the user, the default value of x is no.

`-help`

MetagenomeThreader will show a summary of all options on `stdout` and terminate with exit code 0.

`-version`

Shows the version of *GenomeTools*.

3 Example of a Result from the *MetagenomeThreader*

3.1 Header of the *MetagenomeThreader* Result

MetagenomeThreader	
Result 29.11.2007	
Parameter settings	
Synonymic Value:	1.0000
Non-Synonymic Value:	-1.0000
Blast-Hit-End Value:	-10.0000
Query Stop-Codon Value:	-2.0000
Hit Stop-Codon Value:	-5.0000
Frameshift-Span:	200.0000
Prediction-Span:	400.0000
Leavegene-Value:	-2.0000
cURL-DB:	nucleotide
Output-Filename:	Metagenome
Output-Fileformat:	2
(1/2/3)	
Hitfile:	1
(yes=1/no=0)	
Min-Protein-Length	50
(>=15)	
Min-Result-Percentage:	0.0000
Extended-Modus	1
(yes=1/no=0)	
Homology-Modus	0
(yes=1/no=0)	
Codon-Modus	1
(1/2/3)	

Content:
- date of program execution
- used program parameters

3.2 Sequence-Section of the *MetagenomeThreader* result

Query-DNA-Entry Section A

Query-DNA-Def: read_11|beg|21|length|987|forward|NC_000961|read_12|chimeric|false|gi
Query-DNA-Sequence
tataaaatagccaaatctgagcctaaaagcccttggatatccatgatttagaacgaaccatcccctcttattcaggagaagttttctgaactcttcaacatcctcga...

Coding-DNA-Entry-Section B

Coding-DNA
tataaaatagccaaatctgagcctaaaagcccttggatatccatgatttagaacgaaccatcccctcttattcaggagaagttttctgaactcttcaacatcctcga...
Protein-Sequence
MLLREVTREERKNFYTNEWKVKDIPDFIVKTLELREFGFDHSGEGPSDRKNQYTDIRDLEDYIRATA...
Hit-Information Section
gi-nr: gi18892016 gi-def: Pyrococcus furiosus DSM 3638, section 12 of 173... from: 6558 to: 7299
gi-nr: gi9453868 gi-def: Pyrococcus furiosus priA gene for DNA primase... from: 1 to: 642
gi-nr: gi3342818 gi-def: Thermophilic archaeon Bonch-Osmolovskaya primase... from: 21 to: 341

Content:
- A: Output of the metagenome sequence and the metagenome sequence definition
- B: for every PCS
  - PCS DNA-sequence
  - PCS amino acid sequence
  - for the PCS processing used hit-definition

3.3 Statistic-Section of the *MetagenomeThreader* result

0.7764 Salmonella enterica subsp. enterica serovar Choleraesuis str. SC-B67
0.7764 Salmonella typhimurium LT2, section 22 of 220 of the complete genome
14.2553 Vibrio cholerae O395 chromosome 2, complete genome
7.1809 Pyrococcus furiosus DSM 3638, section 12 of 173 of the complete genome
0.2029 Escheria coli K12 MG1655, complete genome

Content:
- percentage of hit-sequence length vs. length of all hit-sequences used for the PCS prediction
- BLAST-hit definitions

In the HTML-format the GI-number of the BLAST-hits is a hyperlink to the linked NCBI entry.

ATTENTION: It is recommended that the hit-sequence file is stored at two different locations because if an existing hit-sequence file is used and the *MetagenomeThreader* is called without `-t yes`, the hit-sequence file will be overwritten and a new scan of the nucleotide database and / or an online

query of the hit-sequences has to be performed.

References

[1] L. Krause *et al*. Finding novel genes in bacterial communities isolated from the environment. *Bioinformatics*, 22:e281-e289, 2006.

[2] G. Gremme. The GENOMETOOLS genome analysis system. `http://genometools.org`.