mRNA Markup Workflow

An implementation of Bioextract mRNA Markup Workflow in Galaxy system

Małgorzata Komór

Overview

The mRNA Markup workflow, created within the BioExtract Server (www.bioextract.org), was implemented on Galaxy platform (http://galaxyproject.org/). Its aim is to analyze and annotate a set of transcripts in order to distinguish sequence artifacts, potential chimeras, likely full-length protein-coding mRNAs, miRNAs and potentially novel transcripts which can be used in further analysis.

Using NCBI BLAST+, MuSeqBox (Multi-query sequence BLAST output examination with MuSeqBox developed by the Brendel Group), and Python scripts, the workflow passes through the following steps.

0. Input sequence submission
1. Eliminate Vector Contamination
2. Eliminate bacterial contamination
3. find matches in a reference protein database
    - 3.1 Identify potential full-length coding sequences
    - 3.2 Identify potential chimeric sequences
4. Find matches in a Comprehensive Protein database
5. Find matches in Protein Domain Database
6. Produce summary report

Detailed description of the procedure is available under following link (http://bioservices.usd.edu/mrnamarkup.html#description)

Tools

mdnMK.xml

This tool uses makeblastdb from NCBI Blast+ (http://blast.ncbi.nlm.nih.gov/) in order to make a nucleotide database from a fasta formatted file with sequences.

mdpMK.xml

This tool uses makeblastdb from NCBI Blast+ (http://blast.ncbi.nlm.nih.gov/) in order to make a protein database from a fasta formatted file with sequences.

blastnMK.xml

This tool uses blastn from NCBI Blast+ (http://blast.ncbi.nlm.nih.gov/) in order to search a nucleotide database using a nucleotide query. The tool is run with additional options: **-show_gis -evalue 1e-20,** as it is on Bioextract Server. It must be provided with fasta file with nucleotide sequences as a query and a database, which can be either cached or obtained from Galaxy history.

blastxMK.xml

This tool uses blastx from NCBI Blast+ (http://blast.ncbi.nlm.nih.gov/) in order to search a protein database using a translated query. The tool is run with additional options: **-show_gis -evalue 1e-20 –seg yes,** as it is on Bioextract Server. It is also possible to add additional parameters: number of displayed descriptions and number of displayed alignments, which are also used in the original workflow. The tool must be provided with fasta file with nucleotide sequences as a query and a database, which can be either cached or obtained from Galaxy history.

rpstblastnMK.xml

This tool uses rpstblastn from NCBI Blast+ (http://blast.ncbi.nlm.nih.gov/) in order to search a profile database using a nucleotide query. The tool is run with additional options: **-show_gis -evalue 1e-10 –seq yes,** as it is on Bioextract Server. It is also possible to add additional parameters: number of displayed descriptions and number of displayed alignments, which are also used in the original workflow. The tool must be provided with fasta file with nucleotide sequences as a query and a database, which can be either cached or obtained from Galaxy history.

museqboxMK.xml

This tool examines the BLAST output using MuSeqBox program (http://www.plantgdb.org/MuSeqBox/help.php) and saves it in tabular form. It can be run with the following options: -M to indicate potentially chimeric sequences, -F for full-length coding sequences, -q to print queries that with no blast hits, -n to select the first n

hits, -s to select the best 'nhsps' HSPs from each hit. The tool must be provided with Blast or MuSeqBox output file.

msbpMK.xml

This tool passes the MuSeqBox output file and fasta formatted blast input file to tool partMK.py. Both scripts form MuSeqBox Partition procedure.

partMK.py

This script examines the MuSeqBox output file and divides its sequences into two files, one with sequences with blast hits and the other one with sequences with no blast hits. Both files have fasta format.

reportMK.xml

This tool passes the MuSeqBox Partition output files from each step to the tool reportMK.py.

reportMK.py

This script produces the final report for mRNA markup workflow. It takes MuSeqBox Partition output file from each step and counts number of sequences in each file. It produces a summary text file with number of sequences of each type.