mRNA Markup

User Manual

Antoni Szych, University of Warsaw
antoni.szych@students.mimuw.edu.pl

1. What is mRNA Markup?

The mRNA Markup application is used for primary analysis of a set of transcripts, generated from assembly of RNA sequences. The input transcript set is partitioned in many ways, including into contaminants (sequencing artifacts), potential chimeras, likely full-length protein-coding mRNAs, miRNAs, and potentially novel transcripts for further analysis with other programs.
This application uses MuSeqBox (developed by the Brendel Group).

2. Prerequisites

The following should be present prior to running the application:

- Windows or Linux based system, 32 or 64 bit
- Python language interpreter, version 2.7 or above
- Biopython module for Python
- NCBI BLAST+ application suite, version 2.25 or above, accesible in the application directory or via the PATH system variable
- NCBI BLAST+ compatible databases (already in database format, not in fasta files!):
    - Vector Database: for example UniVec
    - Bacterial Database: for example E. Coli from NCBI Nucleotide
    - Reference Database: for example ATpepTAIR10 (A set of Arabidopsis protein sequences available at http://www.plantgdb.org/XGDB/phplib/download.php?GDB=At)
    - All Protein Database: for example UniRef90-Viridiplantae (http://www.uniprot.org/uniref/?query=identity:0.9+taxonomy:33090&format=*&compress=yes)
    - Protein Domain Database: for example NCBI's CDD

3. Usage

1. Unpack the archive containing the application (if compressed).
2. Be sure that you have all the databases prepared and present in the application directory.
3. Chmod `mRNA_markup.py` and `MuSeqBox` for execution.
4. Type `./mRNA_markup.py sequence_file.fasta evalue bacterial_db vector_db domain_db refprot_db protein_db` and press Enter. For example: `./mRNA_markup.py mRNA_in.fasta 1e-20 bacteriaDB uniVec CDD atpeptair10 uniprot90` (If you are in the Windows environment, omit "`./`")

4. Output files

The program produces following output files:

`summary.txt` - statistics for partitioning of all input sequences
`VC-mRNA.fas` - vector-contaminated sequences
`not_VC-mRNA.fas` - not vector-contaminated sequences
`BC-mRNA.fas` - bacterial-contaminated sequences
`not_BC-mRNA.fas` - not bacterial-contaminated sequences
`RA-mRNA.fas` - sequences matching ReferenceDB
`not_RA-mRNA.fas` - sequences not matching ReferenceDB
`FL-mRNA.fas` - full-length coding Sequences
`not_FL-mRNA.fas` - not full-length coding Sequences
`PC-mRNA.fas` - potential chimeric sequences
`not_PC-mRNA.fas` - not potential chimeric sequences
`AA-mRNA.fas` - sequences matching AllProteinDB
`not_AA-mRNA.fas` - sequences not matching AllProteinDB
`CD-mRNA.fas` - sequences matching ProteinDomainDB
`not_CD-mRNA.fas` - remaining sequences

Temporary files, useful for debugging:

`step1.xml`
`step2.xml`
`step3.xml`
`step4.xml`
`step5.xml`
`fullcds.txt`
`RA.msb`
`PC.msb`

5. Licence

6. Contact

Author can be contacted via email antoni.szych@students.mimuw.edu.pl