RiboSeqDB – a repository of selected human and mouse ribosome footprint and RNA-seq data

Sharipov R.N.[1-4], Yevshin I.S.[1-3], Kondrakhin Y.V.[1-3], Volkova O.A.[5]

[1]Design Technological Institute of Digital Techniques SB RAS, Novosibirsk, Russia

[2]Institute of Systems Biology, Ltd; Novosibirsk, Russia

[3]BIOSOFT.RU, Ltd; Novosibirsk, Russia

[4]Novosibirsk State University, Novosibirsk, Russia

[5]Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia

Abstract

Motivation: At present time, there are many different data sets obtained as the results of ribosome profiling and RNA-seq experiments worldwide. Those data sets, mainly, are collected at public databases like Gene Expression Omnibus (GEO) and Short Read Archive (SRA) that are rather huge and some inconveniences with navigation and search exist there. So, for our work dedicated to analysis of different aspects of genome information realization like preference of different transcription starts, influence of different nucleotide motifs on translation efficacy in human and mouse we looked through and selected a range of datasets on common cell lines. Selected datasets will be stored with the results of our calculations on translation efficacy performed during the project using a BioUML platform.

Results: For storage and analysis of the selected from different sources datasets, a RiboSeqDB database was created. Currently, our database comprising 290 data samples of 21 datasets for human and mouse is being further developed acquiring new features and possibilities for users in navigation, data processing and analysis. Integration of the selected data sets and the results of calculations supported with mathematical tools/instruments, user and API interface of BioUML make RiboSeqDB an important part of original, extensible platform for research in the frames of 'Virtual human' conception.

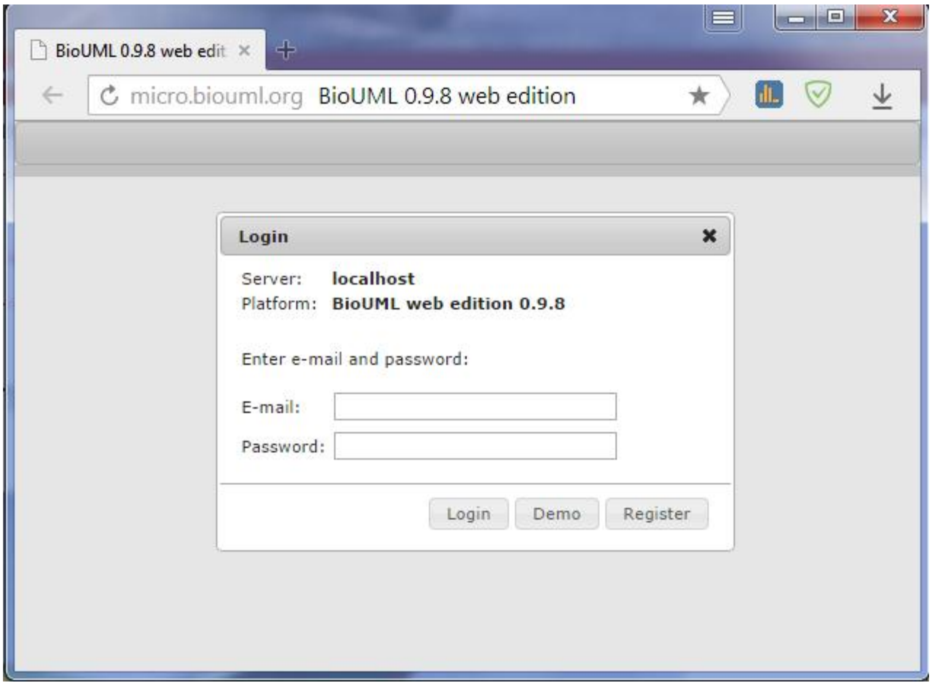
Availability: Demo access at http://micro.biouml.org/bioumlweb/, tab 'Databases'.

Background

A RiboSeqDB repository has been developed in the frames of our analysis of transcription and translation efficacy in mouse and human cells. This work is a branch of our activity in progression of the 'Virtual Human' modeling, and the obtained results will become an important part of it. For this purpose, the databases Sequence Read Archive (SRA) [1] and Gene Expression Omnibus (GEO) [2] were scanned and the appropriate data sets satisfying several conditions were chosen: taxon (human or mouse), data type (mRNA-seq or ribosome footprint) and availability of description of experiment – and included in our repository. All available data sets for human and mouse were found, summarized and those utility for further analysis was estimated for every data set. RiboSeqDB does not contain raw or processed data downloaded from SRA and GEO, but contains structured summary, links to the selected data sets and is supported by integration with a BioUML platform [3] that provides web access and wide spectrum of tools for comprehensive analysis.

Database access

RiboSeqDB is available at http://micro.biouml.org/bioumlweb/ via the BioUMLweb platform. [Figure1] demonstrates the login page of BioUMLweb. Access to the database is granted for two categories of users: registered and demo. The 'Demo' button [Figure 1] will allow to reach the data without any registration in read-only mode [Figure 2 and 3]. The 'Register' button will transfer a user to the full registration process to get access basing on provided rights and permissions.



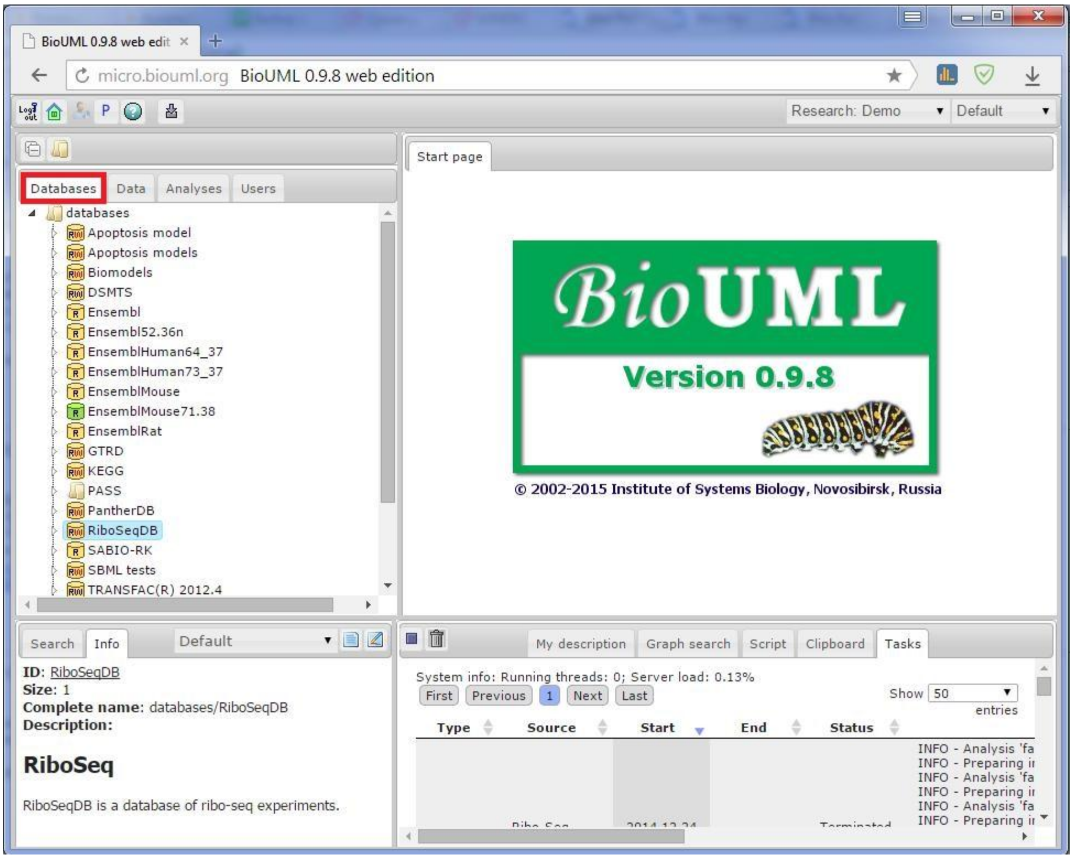
Login

Server: localhost
Platform: BioUML web edition 0.9.8

Enter e-mail and password:

E-mail:
Password:

Figure 1. The login page of BioUMLweb.



ID: RiboSeqDB
Size: 1
Complete name: databases/RiboSeqDB
Description:

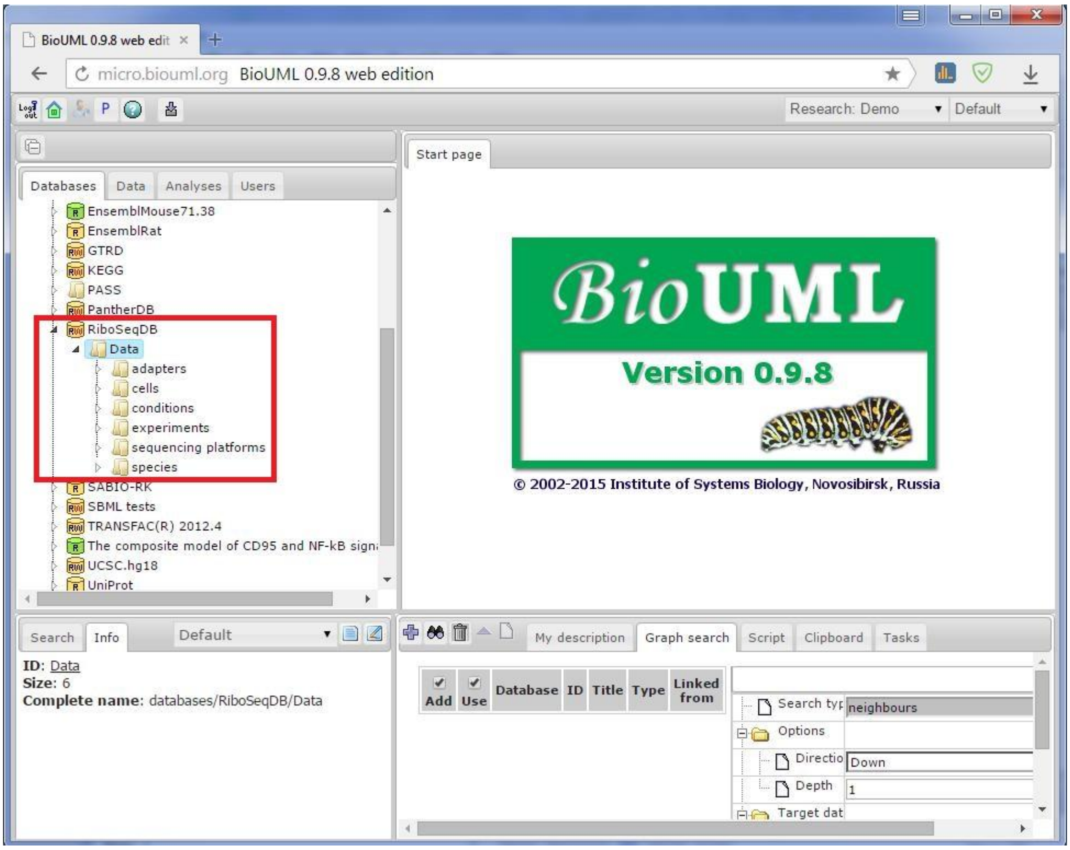
RiboSeq

RiboSeqDB is a database of ribo-seq experiments.

System info: Running threads: 0; Server load: 0.13%

```
INFO - Analysis 'fa
INFO - Preparing i
INFO - Analysis 'fa
INFO - Preparing i
INFO - Analysis 'fa
INFO - Preparing i
```

Figure 2. The inner structure of BioUMLweb and the starting page. A red-marked tab ‘Databases’ contains the list of all databases available in BioUML.



ID: Data
Size: 6
Complete name: databases/RiboSeqDB/Data

Search typ: neighbours
Directio: Down
Depth: 1

Figure 3. The inner collection structure of RiboSeqDB.

Data parameters

All data samples indexed by RiboSeqDB are characterized with a range of parameters (info fields) that were filled in during input to the database [Figure 4].

Name – this field is filled in automatically by the database, this is ID of a sample. All other fields are filled in manually.

Title – a title of an experiment that comprises the inputed data sample.

Description – description of either the data sample or the experiment; a short commentary is also possible.

Species – species of the organism from the experiment. The species is selected from the drop-down menu, and before the input all necessary species have to be added to the data collection 'Species' ([Figure 3], in red rectangle).

Cell source – a cell line or any other kind of cells that can be selected from the drop-down menu. Cell source has to be preinputed in the data collection 'Cells' ([Figure 3], in red rectangle).

Translation inhibition – an inhibitor of translation that was used in the experiment.

Minimal fragment size – minimal size of the sequenced fragments.

Maximal fragment size – maximal size of the sequenced fragments.

Digestion – an enzyme used for digestion of mRNA.

3' sequence adapter – adapter at the 3' termini of reads used for preparation to sequencing procedure.

Conditions – special conditions (chemical inductors, modifiers, inhibitors, drugs, etc) related to the cell line growth or design of the experiment that can be selected from the drop-down menu and, hence, that have to be preinputed in the data collection 'Conditions' ([Figure 3], in red rectangle).

Sequencing platform – a commercial or self-made sequencing platform that was used to get reads selected from the drop-down menu and, hence, that has to be preinputed in the data collection 'Sequencing platforms' ([Figure 3], in red rectangle).

Sequence data – info on data sample splitted into *Format* (well-known international formats FASTA, FASTQ, SRA and ELAND are available from the drop-down menu) and *URL* (a direct link to the file of raw experimental data).

SRA project id – an identifier of the data sample's project in the SRA database (if available), usually of kind of SRPXXXXXX.

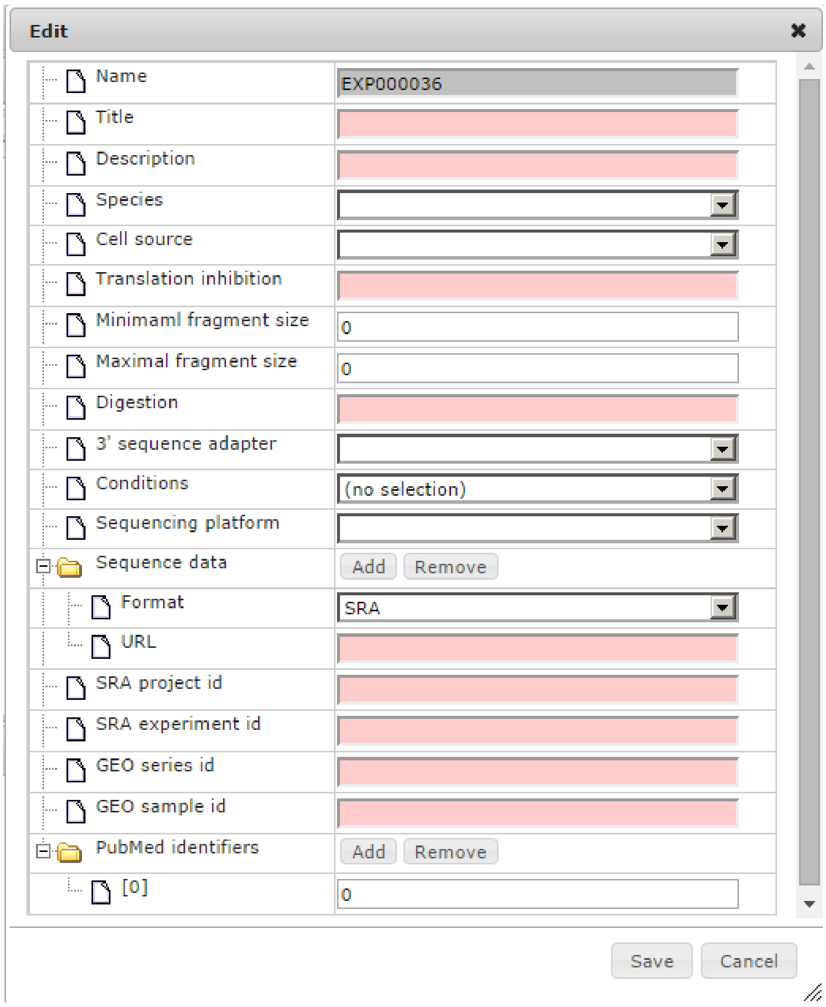
SRA experiment id – an identifier of the data sample in the SRA database (if available), usually kind of SRXAAAAAA.

GEO series id – an identifier of the data series in the GEO database (if available), usually kind of GSEXXXXX.

GEO sample id – an identifier of the data sample in the GEO database (if available), usually kind of GSMXXXXXX.

PubMed Identifiers – an identifier of the article indexed in PubMed that describes the experiment comprising inputed data sample.

Either missing fields or fields demanding more precise definitions are postponed and can be filled in later.

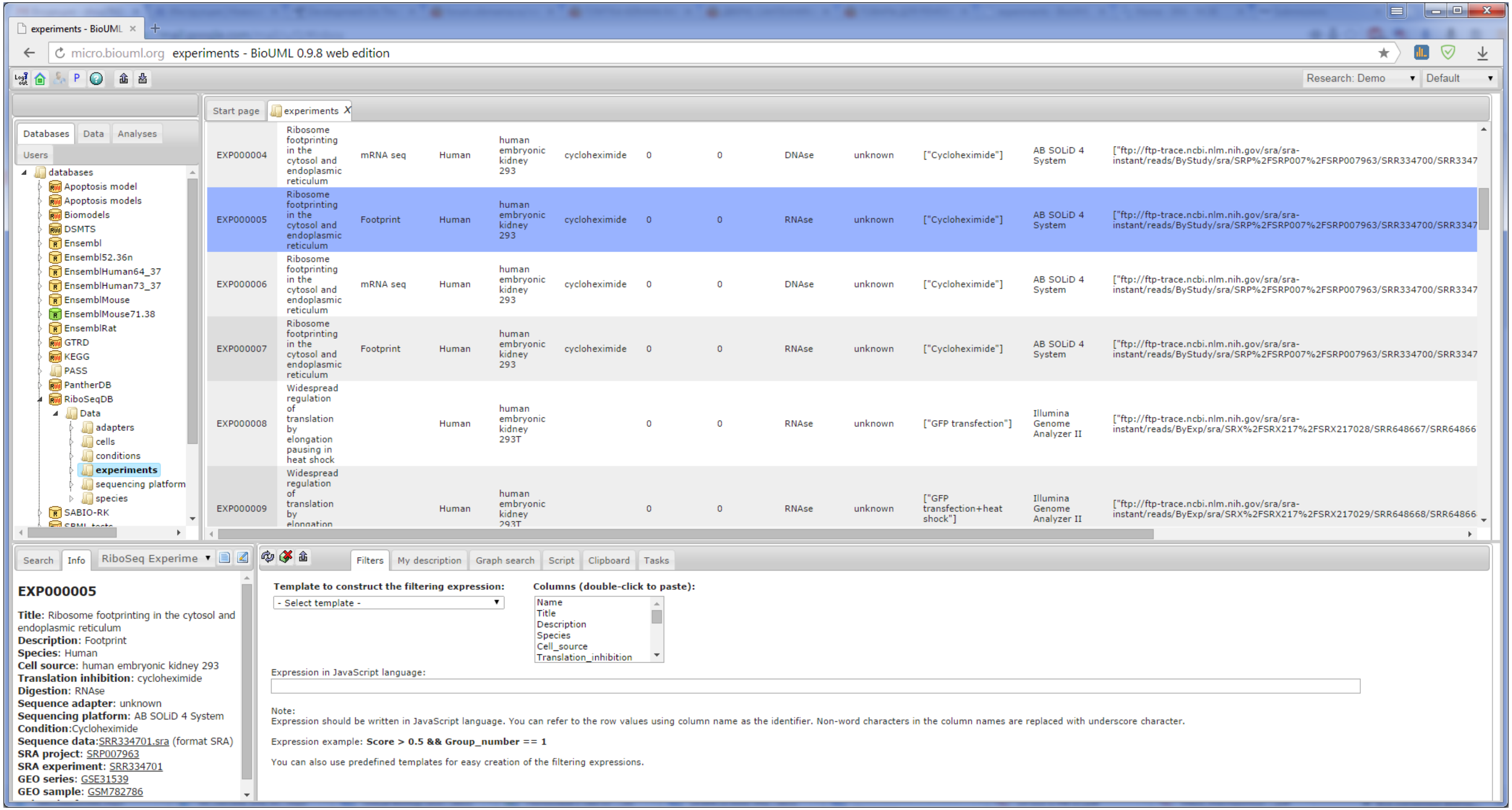


Name	EXP000036
Title	
Description	
Species	
Cell source	
Translation inhibition	
Minimaml fragment size	0
Maximal fragment size	0
Digestion	
3' sequence adapter	
Conditions	(no selection)
Sequencing platform	
Sequence data	Add Remove
Format	SRA
URL	
SRA project id	
SRA experiment id	
GEO series id	
GEO sample id	
PubMed identifiers	Add Remove
[0]	0

Figure 4. A card with parameters that should be filled in for every data sample in RiboSeqDB.

The list of the inputed data can be obtained by double-click on the folder ‘Experiments’ from the ‘Data’ folder of the RiboSeqDB database.

[Figure 5] demonstrates the opened ‘Experiments’ table with the inputed data.



EXP000004	Ribosome footprinting in the cytosol and endoplasmic reticulum	mRNA seq	Human	human embryonic kidney 293	cycloheximide	0	0	DNAse	unknown	["Cycloheximide"]	AB SOLiD 4 System	["ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByStudy/sra/SRP%2FSRP007%2FSRP007963/SRR334700/SRR3347
EXP000005	Ribosome footprinting in the cytosol and endoplasmic reticulum	Footprint	Human	human embryonic kidney 293	cycloheximide	0	0	RNAse	unknown	["Cycloheximide"]	AB SOLiD 4 System	["ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByStudy/sra/SRP%2FSRP007%2FSRP007963/SRR334700/SRR3347
EXP000006	Ribosome footprinting in the cytosol and endoplasmic reticulum	mRNA seq	Human	human embryonic kidney 293	cycloheximide	0	0	DNAse	unknown	["Cycloheximide"]	AB SOLiD 4 System	["ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByStudy/sra/SRP%2FSRP007%2FSRP007963/SRR334700/SRR3347
EXP000007	Ribosome footprinting in the cytosol and endoplasmic reticulum	Footprint	Human	human embryonic kidney 293	cycloheximide	0	0	RNAse	unknown	["Cycloheximide"]	AB SOLiD 4 System	["ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByStudy/sra/SRP%2FSRP007%2FSRP007963/SRR334700/SRR3347
EXP000008	Widespread regulation of translation by elongation pausing in heat shock		Human	human embryonic kidney 293T		0	0	RNAse	unknown	["GFP transfection"]	Illumina Genome Analyzer II	["ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX%2FSRX217%2FSRX217028/SRR648667/SRR64866
EXP000009	Widespread regulation of translation by elongation		Human	human embryonic kidney 293T		0	0	RNAse	unknown	["GFP transfection+heat shock"]	Illumina Genome Analyzer II	["ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX%2FSRX217%2FSRX217029/SRR648668/SRR64866

EXP000005

Title: Ribosome footprinting in the cytosol and endoplasmic reticulum
Description: Footprint
Species: Human
Cell source: human embryonic kidney 293
Translation inhibition: cycloheximide
Digestion: RNAse
Sequence adapter: unknown
Sequencing platform: AB SOLiD 4 System
Condition:Cycloheximide
Sequence data:SRR334701.sra (format SRA)
SRA project: SRP007963
SRA experiment: SRR334701
GEO series: GSE31539
GEO sample: GSM782786

Template to construct the filtering expression: - Select template -

Columns (double-click to paste):
Name
Title
Description
Species
Cell_source
Translation_inhibition

Expression in JavaScript language:

Note:
Expression should be written in JavaScript language. You can refer to the row values using column name as the identifier. Non-word characters in the column names are replaced with underscore character.

Expression example: **Score > 0.5 && Group_number == 1**

You can also use predefined templates for easy creation of the filtering expressions.

Figure 5. A part of the ‘Experiments’ table of RiboSeqDB. The EXP000005 card is selected (data summary is represented in the left bottom pane).

RiboSeq data processing

A special workflow was developed to preprocess RiboSeq data collected in the database (http://micro.biouml.org/bioumlweb/#anonymous=true&de=analyses/Workflows/Ribo-Seq%20preprocessing; [Figure 6]).

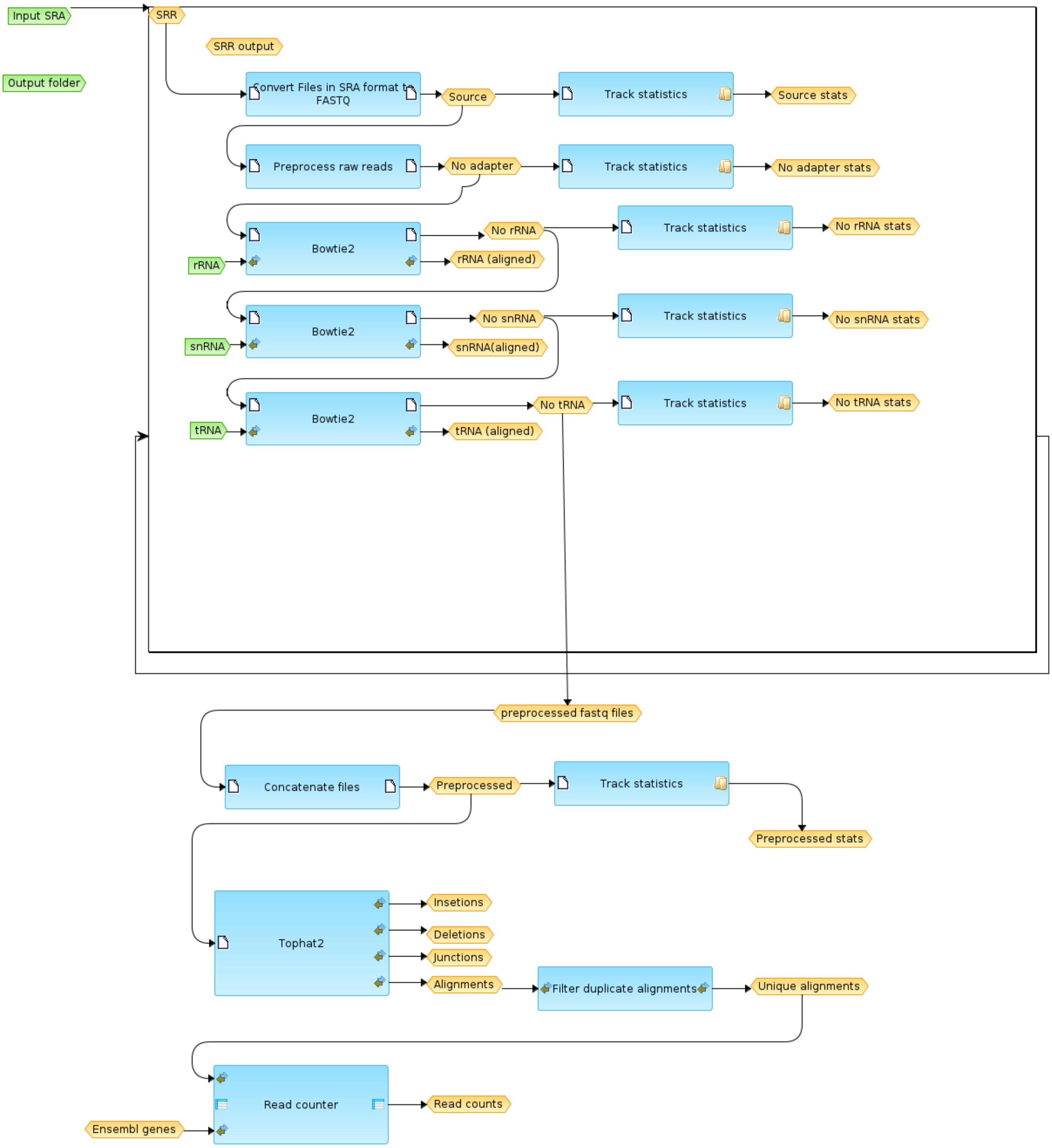


Figure 6. A workflow for RiboSeq data preprocessing implemented in BioUML.

As the input, sequenced data format SRA is acceptable (SRRs from the SRA database). 3' adapters are trimmed in the reads basing on the sequencing platform and experiment design information. Reads less than 9 bp are filtered, as well as the reads corresponding to known rRNAs, tRNAs and snRNAs using well-known sequence aligner Bowtie2. Track statistics is gathered at the every step of filtering procedure.

Filtered/preprocessed FASTQ then is concatenated and transferred to Tophat2 – “a fast splice junction mapper for RNA-Seq reads. It aligns RNA-Seq reads to mammalian-sized genomes using the ultra high-throughput short read aligner Bowtie, and then analyzes the mapping results to identify splice junctions between exons.” [4]. Reads that cannot be mapped unambiguously are removed from the process. Unique alignments are transferred to Read counter supported with annotation data from the Ensembl database, then number of reads corresponding to genes is calculated and the preprocessing procedure, thus, is finished. The final step is calculation of reads mapped to transcripts (http://micro.biouml.org/bioumlweb/#de=analyses/Methods/BSA/Count%20reads%20in%20transcripts; then select ‘Demo’). This number of the mapped reads is taken into account for estimation of translation level of mRNAs.

Application of workflows allows to perform processing automatically of big volumes of experimental data and by the same manner, applying corrections if it is necessary. This is important by several causes:

- same order of processing leads to non-biased results induced by application of different methods to raw data;
- speed up of the process and easily reproduction of the whole chain of manipulations with data due to the logging in journal;
- more convenient conditions for debugging of the data processing pipeline;
- easy modification of the list of the allied methods and processing parameters (thresholds, etc).

RiboSeqDB statistics

Currently, RiboSeqDB contains data corresponding to the following cell lines:

Experiments – 290 samples in 21 datasets in total.

Cell lines and cell types – 14:

- 3T3 (mouse embryonic fibroblasts),
- BJ fibroblasts (human),
- Brain neurons (mouse),
- C4-4 (human ovarian cancer),
- HCMV-infected human foreskin fibroblasts,
- HEK293 (human embryonic kidney),
- HEK293T (human embryonic kidney transfected),
- HeLa (human cervical cancer),
- MEF (mouse embryonic fibroblasts),
- mESC (mouse embryonic stem cells),
- mESC A3-1 (mouse embryonic stem cells A3-1),

- Neutrophils (mouse),
- PC3 (human prostate cancer),
- PEO1 (human ovarian cancer).

Sequencing platforms – 5:

- AB SOLiD 4 System,
- Illumina Genome Analyzer II,
- Illumina Genome Analyzer IIx,
- Illumina HiSeq 2000,
- Illumina MiSeq 2000.

Future prospects

Current state of RiboSeqDB is not final. The next stage of work will be bridging all the gaps using literature data and sequencing protocols. RiboSeqDB is the repository, both, for storage and fast application of data freely available from the public sources and their integration with the results of analysis performed in the frames of our research projects. This database is a part of the BioUML scientific platform that provides extended functionality for users covering wide spectrum of possible tasks starting from the formal description of the research object and modeling and ending with processing of the data obtained with modern high-throughput '-omic' methods. In complex with BioUML, RiboSeqDB will provide users with reliable informatics support, fast and convenient navigation, data search, supply with a big enough set of tools for analysis and become a safe harbour for data storage and exchange. Next year it is planning to implement a workflow for processing of raw ribosome footprint data that will allow a user to reach new horizons and get new advantages in analysis of genome information realization in cells meanwhile approaching to 'Virtual Human' implementation.

Literature

1.

Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Edgar R, Federhen S, Feolo M, Geer LY, Helmberg W, Kapustin Y, Khovayko O, Landsman D, Lipman DJ, Madden TL, Maglott DR, Miller V, Ostell J, Pruitt KD, Schuler GD, Shumway M, Sequeira E, Sherry ST, Sirotkin K, Souvorov A, Starchenko G, Tatusov RL, Tatusova TA, Wagner L, Yaschenko E.: Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2008 Jan; 36 (Database issue): D13-21. Epub 2007 Nov 27. PMID: 18045790

2.

Barrett T1, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, Yefanov A, Lee H, Zhang N, Robertson CL, Serova N, Davis S, Soboleva A. NCBI GEO: archive for functional genomics data sets--update. Nucleic Acids Res. 2013 Jan;41(Database issue):D991-5. doi: 10.1093/nar/gks1193. Epub 2012 Nov 27. PMID: 23193258.

3.

http://wiki.biouml.org/index.php/BioUML_wiki

4.

http://ccb.jhu.edu/software/tophat/index.shtml