

SATrans manual

1 Introduction

SATrans is a novel bioinformatics tool which was developed to contribute to understand and to biologically interpret RNAseq datasets. This software is ideal for biologists performing transcriptome research because it provides fast and reliable functional annotation of thousands of nucleotide/amino acid sequences on a personal computer or notebook. Moreover, the software allows the functional analysis of gene expression data at the whole transcriptome level based on the GO annotation [1]. It is suitable for any organism, which is of special interest regarding non-model species. We believe that SATrans is user-friendly software because it requires only the basic knowledge of a Linux operating system and provides outputs in the user-friendly form.

1.1 Input files

The software requires two main types of input files, depending on the desired operation. The first required input file is a FASTA formatted file containing the nucleotide/amino acid sequences which should be functionally analyzed – e.g. all the transcripts of the analyzed transcriptome. This file is called a multi-fasta input file (**Figure 1**).

```
>MLOC_54213
IEQIGRNLQLPRPKHSLSLFPSHPRLRSPISEVLASQRPPAHHRFRAQIRPIRAASGHR
SGSVTCGEVVAQTLMQEALACLTSTTMSCKCKGVCMEIGTCCMNVLKWLLAKKALVLMH
PSLDMKHHCMLLLNMTIWSCHMVQKMNIWLCRLSQMMVGMMFQGGPTVMVSVMGKDQV
EQMQIKLKGEWTYLMISCTWYSPSYAGGIYVEQGLPANSGSLLVCMGISGNIWSLGTPEY
LCGTLLMFATVIGMWQISICLVSGVQKAGWLKQGHSGGILGPWGWAGDNWEKHFFGLWLN
AHCMLGQSVMHPLVVAFKGLLIMMDCMNFKLWSVVHSEYLSDATNFEYCLWGELAWLM
>MLOC_55229
IEQIGRNLQLPRPKHSLSLFPSHPRLRSPISEVLASQRPPAHHRFRAQIRPIRAASGHR
SGTDAAYPCSLPCWRVPPGSSVSRPPASLIRFDQIRCNLACRRPQSLEFARGSWRAGVW
VAVGWPMPLAEGVVLGLGFGAVVWGDGGRGKDEGRGKRRGYRAWVYGRGGAGGAGART
VSGAPGLASAASAGAAAGTRLEMVDGLPGMPWRCREFACWLERLFTSCPATQVPWHVA
LVVAFKFFLLIMMDCMNFKLWSVVHSEYLSDATNFEYCLWGELAWLMYHSIVLSCLNWIF
SPAMSFLLTQFVKRQRPVHCWRHWMCHPARVLLMWHCVWMLMHVKIFLLMHLTAPTFLS
```

Figure 1: Example of the multi-fasta file.

The second input file is required only if the “analysis” mode is launched. We call it “differential gene expression input file” and it must have a defined structure, which is typical for the output file of DESeq2 software [2]. It consists of seven columns which are labeled as: 1. gene name, 2. baseMean, 3. log2FoldChange, 4. lfcSE, 5. stat, 6. pvalue and 7. padj (**Figure 2**). In the case of missing columns

or values, the file has to be manually edited. The columns have to be added in and the values have to be filled in (e.g. by value 0). The file has to be presented as tab-delimited or csv format.

	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
MLOC_43997	197,93	4,2900	0,32	13,26	4,12E-40	3,53E-36
MLOC_73656	661,36	3,0999	0,32	9,82	9,28E-23	3,98E-19
MLOC_15785	217,10	2,7941	0,33	8,60	8,16E-18	2,80E-14
MLOC_15784	136,06	2,6496	0,34	7,73	1,04E-14	1,98E-11
MLOC_70443	1773,33	2,5056	0,30	8,24	1,74E-16	4,26E-13
MLOC_36469	999,06	2,4388	0,32	7,67	1,66E-14	2,59E-11
MLOC_15135	50,85	2,4314	0,35	6,90	5,32E-12	5,37E-09
MLOC_17138	37,93	2,4210	0,36	6,78	1,22E-11	1,16E-08
MLOC_67259	1469,60	2,3945	0,31	7,70	1,37E-14	2,35E-11
MLOC_45692	242,53	2,3353	0,30	7,74	9,69E-15	1,98E-11
MLOC_70609	104,38	2,3015	0,34	6,72	1,86E-11	1,67E-08
MLOC_26926	35,44	2,1977	0,36	6,15	7,61E-10	4,84E-07

Figure 2: Example of the output file from the DESeq2, which represents a differential gene expression input file for the SATrans when using the “analysis” mode.

1.2 Quick start

For impatient people, the basic commands are listed here. The user is asked to enter a password for MySQL database system at each launch, what means that the account in MySQL database system must be created and the user privileges have to be granted (see chapter 3.0).

Create the database with the name *test* without an input file:

```
perl main.pl --mode create --db_name test --db_user user
```

Create the database using the input FILE.fa, and start an annotation of sequences using a local BLAST with five threads:

```
perl main.pl --mode create --db_name test --db_user user --input FILE.fa --email user@domena.com --local --threads 5 --db_blast /home/directory/nt_NCBI/nt
```

Create the database using the input FILE.fa and start the annotation of sequences using a remote BLAST with five threads – five parallelly annotated sequences:

```
perl main.pl --mode create --db_name test --db_user user --input FILE.fa --email user@domena.com --threads 5 --db_blast nt
```

Make the update of the database by re-annotation of annotated sequences stored already in the database, using the remote BLAST and five threads as the number of parallelly searched sequences:

```
perl main.pl --mode update --db_name test --db_user user --email user@domena.com --threads 5 --db_blast nt
```

Show existing databases with correct structure:

```
perl main.pl --mode show --db_user user
```

Delete the existing database with name *test*:

```
perl main.pl --mode delete --db_name test --db_user user
```

Import GO terms from FILE.txt:

```
perl main.pl --mode import --db_name test --db_user user --input FILE.txt
```

Run the functional analysis of differentially expressed genes from the differential gene expression input file FILE.csv:

```
perl main.pl --mode analysis --db_name test --db_user user --input FILE.csv
```

2 Prerequisites

SATrans was created using a combination of programming and database languages, Perl and SQL, respectively. Perl (<https://perldoc.perl.org/>) implements a user-friendly interface, communication with web services to search in public databases and import/export of data. The SQL database language (<https://dev.mysql.com/downloads/mysql/>) provides operations for MySQL database created by the program during a run. The software is primarily designed for GNU/Linux but can run on other operating systems such as MS Windows or Mac OSX fulfilling requirements for launching.

The software prerequisites are Perl v5.18.2 or higher and MySQL database system v5.5.47 or higher. The program further requires the following Perl modules, which are available on CPAN (Perl Archiving Perl Network, <https://www.cpan.org>):

- threads
- Term::ANSIColor
- LWP
- English
- XML::Simple
- XML::XPath
- File::Basename
- Getopt::Long
- Data::Dumper
- Term::ReadKey
- MySQL::Backup (version 0.04 or higher)

For using the local BLAST search, installation of the NCBI BLAST (<https://www.ncbi.nlm.nih.gov>) is required.

3 Obtaining and installation

The latest source release is available on the GitHub (<https://www.prf.upol.cz/departement-of-molecular-biology/links/>). To install the software from the source package unpack the directory by the command:

```
gzip -d Satrans_v1.3.gz
```

The software can be launched directly from its directory by the command:

```
perl main.pl
```

In this case, the warning about the wrong number of input arguments and the help message is listed. To see the help message only use the command:

```
perl main.pl --help
```

The software requires a creation of the user account in the MySQL database system and obtaining the user privileges. A new user in the MySQL database system can be created by typing the command in the MySQL user interface:

```
CREATE USER 'user'@'localhost' IDENTIFIED BY 'password';
```

The user privileges can be granted to the user by the command:

```
GRANT ALL PRIVILEGES ON *.* TO 'user'@'localhost' WITH GRANT OPTION;
```

The software uses only the long switches (--) instead of the short ones (-).

4 Using SATrans

Usage: perl main.pl [options]*

SATrans can be launched in the several modes which provide different services. Each mode is described in the individual subchapter (4.1 – 4.8). A mode selection depends on the option “mode”. The user is asked for the MySQL database system password at each launching. The typical use cases of modes are listed in the subchapter 1.2.

Global options:

`--help` Prints the help message and exit.

`--input <string>` Path to input file. Input file might be a file containing nucleotide or amino acid sequences in fasta format, or a file containing data for the import into the MySQL database in the tab-delimited form or results file from the DESeq2 analysis in the csv format.

`--email <string>` User has to provide an email address in the standard text format for the InterProScan run.

`--threads <int>` Use this many of sequences which will be analyzed at the same time by the remote BLAST and InterProScan. The recommended value for the remote services is less than 14, more threads could lead to the instability of connection with the remote servers. In case of running the local BLAST (global option `--local`) the value “threads” is the number of CPUs. The default value is 2.

`--mode <string>` Mode in which the program will be launched. The default choice is “analysis”. The other possibilities for this option are create, update, repair, delete, show, export, and import.

`--out <string>` Prefix or path and prefix which SATrans will use to write all output files. The default option is “./out”.

`--out_format` Sets the format of the output files from the analysis mode. The possible options are “csv” and “txt”. The default option is “txt”.

BLAST search options:

`--e_value <double>` Expect value (E value) cutoff for the saving of BLAST hits. The default value is 0.001.

`--align_length <int>` Sequence length cutoff value for the saving of BLAST hits. The default value is 20 nucleotides/amino acids.

`--max_number_hit <int>` Maximum number of BLAST hits per sequence which will be saved. Default value is 20.

`--blast_mode <string>` Selected type of blast used by BLAST. Possible values are blastn, blastp, blastx, tblastn, tblastx. The default value is blastn.

`--local` Sets up the local BLAST search.

Options for differential gene expression data analysis:

`--log2f_cut <double>` Log2FoldChange cutoff value (absolute value from the DESeq2 csv file) for the functional analysis of the differentially expressed genes. The default value is 2.0 or -2.0 for upregulated or downregulated genes, respectively.

`--cut_hist <double>` Cut-off value (>0) for histogram output file. Setups the size of a range of log2FoldChange values in which a number of genes with the assigned GO term is counted. The default value is 0.5.

`--pAdj_cut <double>` PAdj cutoff value (from the DESeq2 csv file) for the functional analysis of the differentially expressed genes. The default value is 0.01.

Database options:

`--db_blast <string>` Name of the selected BLAST database. For the remote BLAST, the option is nt or nr. For the local BLAST, the option is the path to BLAST preformatted database. The default value is “nt”.

`--db_host <string>` Name of the host (usually a server or computer) of the MySQL database system. The default value is “localhost”.

`--db_name <string>` Name of the MySQL database which will be created or used by SATrans.

`--db_user <string>` Name of the MySQL database user.

4.1 Mode create

After launching this mode, the MySQL database is created together with all the database tables. The BLAST and InterProScan tools are searching the selected databases and the results are stored in the database. If no input file is provided, the database is created, but no BLAST and InterProScan is performed. BLAST search might be remote or local, depending on the option `--local`. If this option is used, the local BLAST is provided and CPU's usage depends on the value of the option `--threads`.

In this mode, the following parameters can be used: `align_length`, `db_blast`, `db_host`, `db_name`, `db_user`, `email`, `e_value`, `input`, `local`, `max_number_hit` and `threads`. The parameters are described above.

4.2 Mode update

The mode provides updating of an existing database, for example by adding new sequences. If the input file is provided when launching this mode, all the sequences from the file are inserted into a selected database and only these new sequences are functionally annotated by BLAST and InterProScan. In the case of duplicate sequence name, the warning message is listed and the sequence is not functionally annotated again. If no input file is provided, only the sequences with no functional annotation (stored in the MySQL database) are reannotated by BLAST and InterProScan.

In this mode, the following parameters can be used: align_length, db_blast, db_host, db_name, db_user, email, e_value, blast_mode, input, local, max_number_hit and threads. The parameters are described above.

4.3 Mode repair

This mode provides a repair of an existing MySQL database if the errors are not repaired by the SATrans automatically. In this mode, all the sequences, which are stored in the MySQL database and still not processed by SATrans, will be functionally annotated by BLAST and InterProScan.

In this mode, the following parameters can be used: align_length, db_blast, blast_mode, db_host, db_name, db_user, email, e_value, input, local, max_number_hit and threads. The parameters are described above.

4.4 Mode analysis

The analysis mode arranges data from the differential gene expression file according to the user's settings, i.e. cut-off values (options log2Fold_cut, pAdj_cut). The differential gene expression input file must have defined format (output format of DESeq2 [2] in the tab-delimited or csv format).

The mode merges the functional annotation results with the expression data and performs the analysis of these data with respect to the GO annotation. The appearance of each assigned GO term in between the differentially expressed genes (DEGs, for all DEGs, and also up-regulated and down-regulated separately) is counted and compared to the appearance of the GO term in the whole data set (multi-fasta input file, transcriptome). The significance of the GO term appearance in between the DEGs is tested by the Fischer exact test [3]. The test is based on the two-dimensional contingency table (**Table 1**) and provides the information about the impact of the tested condition or genotype on the molecular function or biological process, and helps to interpret the data biologically. More details about the test can be found at https://www.pathwaycommons.org/guide/primers/statistics/fishers_exact_test/.

	Number of differentially expressed genes	Number of NO differentially expressed genes	Sum
Number of genes annotated by the GO term	a	b	a+b
Number of genes unannotated by the GO term	c	d	c+d
Sum	a+c	b+d	a+b+c+d

Table 1: Contingency table for Fischer exact test.

Another analysis performed in the frame of this mode is counting the number of the DEGs which log2FoldChange values belong into the defined ranges of the values (option cut_hist). The results are provided for each assigned GO term, thus for the each GO term the histogram of DEGs's log2FoldChange values can be assembled.

The mode provides three types of results which are described in the subchapter 5.1.

In this mode the following parameters can be used: cut_hist, db_host, db_name, db_user, log2f_cut, pAdj_cut, input, out_format and out_pref. The parameters are described above.

4.5 Delete mode

This mode provides deletion of an existing MySQL database. The user is prompted to confirm deletion of the database during the deletion process.

In this mode, the following parameters can be used: db_host, db_name, and db_user. The parameters are described above.

4.6 Show mode

After launching this mode, the list of databases which can be used by the SATrans is provided. The output provides also the information about the number of records in the database tables.

In this mode, the following parameters can be used: db_host, db_name, and db_user. The parameters are described above.

4.7 Export mode

This mode serves for the export of the MySQL database tables into the txt files. The tables are exported in the tab-delimited format, having the same structure as in the database.

The tables which might be exported in this mode are called: "1_Sequence" (containing data about sequences which are stored in database), "2_Hits" (containing results of the BLAST),

“3_InterProScan_data” (containing results of the InterProScan), 6_GO_parse (containing obtained GO terms at the lowest GO level) and “8_GO_analysis” (containing information about all GO terms for each gene which has been functionally annotated, as well as the information from “4_term” table about the semantics of the GO term). The files are described in the subchapter 5.2.

In this mode, the following parameters can be used: db_host, db_name, and db_user. The parameters are described above.

4.8 Import mode

This mode provides the possibility to import GO terms into an existing MySQL database. The input file must have a defined structure and tab-delimited format, containing two columns. The first column contains the name of the sequence/gene and the second one contains the GO number in the defined form (GO:0000651).

In this mode, the following parameters can be used: input, db_host, db_name, and db_user. The parameters are described above.

5 SATrans output files

The SATrans produces several types of output files in the directory in which it was launched. There are the output files which are the results of the “analysis” mode and the output files which are the results of the “export” mode. All the output files are described in the following subchapters.

5.1 Analysis mode output files

5.1.1 Annotation file

The annotation file contains the functional annotation of each DEGs. A name of the file consists of user-defined prefix (default is “out”) and fixed suffix (_Annotation.txt). The file contains results from the BLAST and InterProScan search, as well as expression data imported from the differential gene expression (DESeq2) input file. The file consists of the following columns:

- 1) **Seq_name** – Name of the gene/sequence.
- 2) **baseMean** – The baseMean value for the gene/sequence. The value is imported from the differential gene expression input file, thus originally calculated by DESeq2.
- 3) **Log2FoldChange** – Log2FoldChange value for the gene/sequence. The value is imported from the differential gene expression input file, thus originally calculated by DESeq2.

- 4) **pValue** – pValue for the gene/sequence, which determines if the gene is or is not significantly differentially expressed. The value is imported from the differential gene expression input file, thus originally calculated by DESeq2.
- 5) **pAdj** – Adjusted pValue for the gene/sequence, which determines if the gene is or is not significantly differentially expressed. The value is imported from the differential gene expression input file, thus originally calculated by DESeq2.
- 6) **Blast annotation** – Functional annotation of the gene/sequence based on the best BLAST hit.
- 7) **Blast accession** – The accession number of the best BLAST hit.
- 8) **IPR_PFAM_Group** – The specification of the PFAM group from the result of the InterProScan search.
- 9) **IPR_PFAM_Description** – The description of the InterProScan based PFAM annotation.
- 10) **GO_terms_list** – List of the assigned GO terms.

5.1.2 GO analysis output file csv/txt

This is the main analysis output file created for all DEGs, but also separately for up- and down-regulated genes. The name of the file consists of user-defined prefix (default is “out”) and fixed suffix (_GO_analysis_all.txt/csv, _GO_analysis_downregulated.txt/csv or _GO_analysis_upregulated.txt/csv). The file consists of the following columns:

- 1) **GO_Level** – Level number in the tree structure of the GO terms. Level one represents Ontology source.
- 2) **Ontology source** – Denotes which of the three sub-ontologies Molecular Function (MF), Biological Processes (BP), Cellular Component (CC) or Undefined (UN) – the term belongs.
- 3) **GO_id** – GO term number identifier in the defined structure (GO:0000644).
- 4) **Description of GO** – A text description of the GO term.
- 5) **Number of GO annotated sequences** – Number of all the genes/sequences from the MySQL database with the assigned GO term.
- 6) **DEGs number** – Number of the differentially expressed genes/sequences with the assigned GO term.
- 7) **Percent** – Percentage of the DEGs in relation to all genes/sequences with the assigned GO term.

- 8) **Mean log2FoldChange** – Mean of log2FoldChange values for the DEGs with the assigned GO term. This column is not present in the file created for all DEGs.
- 9) **Extreme value** – The lowest log2FoldChange value among the down-regulated genes, or the highest log2FoldChange among the up-regulated genes with the assigned GO term. This column is not present in the file created for all DEGs.
- 10) **Max histogram range** – Range of the log2FoldChange values where the most of the DEGs with the assigned GO term belongs, thus their log2FoldChange value is within this range. This column is not present in the file created for all DEGs.
- 11) **Max histogram number** – The number of the DEGs belonging to the max histogram range (described above). This column is not present in the file created for all DEGs.
- 12) **p-value** – Fischer exact test p-value. The test calculates the significant differential expression of the DEGs's GO term in relation to all genes with the assigned GO term – differentially as well as no differentially expressed.

5.1.3 Histogram file txt

This output file is created separately for up- and down-regulated genes. The name of the file consists of a user-defined prefix (default is “out”) and fixed suffix (_Histogram_downregulated.txt, _Histogram_upregulated.txt). The file contains the list of all GO terms which have been assigned to the DEGs and provides the number of DEGs which log2FoldChange values belong to the defined log2FoldChange value range. The option cut_hist defines the range setup as follows:

<min(log2FoldChange); min(log2FoldChange) + cut_hist>,

< min(log2FoldChange) + cut_hist; min(log2FoldChange) + 2*cut_hist>,

< max(log2FoldChange) - cut_hist; max(log2FoldChange)>.

The file consists of the following columns:

- 1) **GO_id** – GO term number identifier in the defined structure (GO:0000644).
- 2) **Min** - The lower limit of the log2FoldChange value.
- 3) **Max** - The upper limit of the log2FoldChange value.
- 4) **Number of genes** – The number of DEGs with the assigned GO number, which log2FoldChange value belongs to the range defined by Min and Max.

5.2 Export mode output files

5.2.1 Sequence_export.fa file

The file contains fasta formatted sequences included in the database.). The header of the sequence consists of the name of the sequence and the length value. The table is created from database table 1_Sequence.

5.2.2 Annotation_export.txt file

The annotation file contains the functional annotation of each sequence in the input database. The file contains results from the BLAST and InterProScan search. The file consists of the following columns:

- 1) **Seq_name** – Name of the gene/sequence.
- 2) **Blast annotation** – Functional annotation of the gene/sequence based on the best BLAST hit.
- 3) **Blast accession** – The accession number of the best BLAST hit.
- 4) **IPR_PFAM_Group** – The specification of the PFAM group from the result of the InterProScan search.
- 5) **IPR_PFAM_Description** – The description of the InterProScan based PFAM annotation.
- 6) **GO_terms_list** – List of the assigned GO terms.

5.2.3 Hits_export.txt file

The table is identical to the database table 2_Hits (see below) and consists of the following columns:

- 1) **Seq_name** – Name of the gene/sequence.
- 2) **Hit_definition** – Description/title of the matched database sequence.
- 3) **Hit_accession** – Accession number of the matched database sequence.
- 4) **Hit_length** – Length of the matched database sequence.
- 5) **Bit_score** – Alignment bit-score.
- 6) **Score** – Alignment raw score.
- 7) **Evalue** – Alignment Expect value.
- 8) **Query_from** – Start of alignment in the query sequence.
- 9) **Query_to** – End of alignment in the query sequence.
- 10) **Identity** – Number of identical matches.

11) Positive – Number of positive-scoring matches.

12) Align_length – Alignment length in the query sequence.

5.2.4 GO_parse_export.txt file

The table is identical to the database table 6_GO_parse (see below) and contains only one GO term per gene/sequence. It is the most specific, the lowest level term. The table consists of the following columns:

- 1) **Seq_name** – Name of the gene/sequence.
- 2) **Entry_ac** – Accession number of the InterProScan entry. Exported from the table 3_InterProScan_data, value 3_entry_ac.
- 3) **GO_number** – GO term number identifier in the defined structure. Example: GO:0000644.

5.2.5 GO_analysis_export.txt file

The table is identical to the database table 8_GO_analysis. The table contains detailed results of GO annotation analysis for each sequence and consists of the following columns:

- 1) **Level** – same as GO_level in the chapter 5.1.1.
- 2) **Group** – same as Ontology_Source in the chapter 5.1.1.
- 3) **Seq_name** – Name of the gene/sequence.
- 4) **Term_id** - Unique serial number (exported from the database table 4_term, see below).
Example: 192
- 5) **GO_number** - GO term number identifier in the defined structure. Example: GO:0000644
- 6) **Term_name** - Character varying a textual label for the term. Each term has a single such label. The name should be unique within an ontology. Example: "cysteine biosynthetic process"

5.2.6 InterProScan_data_export.txt file

The table is identical to the database table 3_InterProScan_data (see below) and consists of the following columns:

- 1) **Seq_name** – Name of the gene/sequence.
- 2) **Database** – Name of the sub-database form the InterProScan.
- 3) **Score** – Alignment score.

- 4) **Evalue** – Alignment Expect value.
- 5) **Signature_name** – Name of the InterProScan signature.
- 6) **Signature_desc** – Description of the InterProScan signature.
- 7) **Signature_ac** – Accession number of the InterProScan signature.
- 8) **Entry_type** – Type of the entry which tells you what you can infer when a protein matches the entry. The entry types are: H – Homologous Superfamily, F – Family, D – Domain, R – Repeat, S – Site
- 9) **Entry_name** – Name of the InterProScan entry.
- 10) **Entry_desc** – Description of the InterProScan entry.
- 11) **Entry_ac** – Accession number of the InterProScan entry.

5.3 Stat.log file

The file summarizes information concerning the functional annotation of the sequences stored in the database. The file is created by SATrans in both “analysis” and “export” mode. In the case of “analysis” mode, the file also consists of separate information only about the differentially expressed genes. The file consists of the following information in individual lines:

- 1) A number of sequences/DEGs which are stored in the database.
- 2) A number of sequences/DEGs with BLAST annotation / Number of DEGs
- 3) A number of sequences/DEGs with InterProScan annotation.
- 4) A number of sequences/DEGs with GENE3D annotation from InterProScan search.
- 5) A number of sequences/DEGs with PANTHER annotation from InterProScan search.
- 6) A number of sequences/DEGs with PFAM annotation from InterProScan search.
- 7) A number of sequences/DEGs with GENE3D annotation from InterProScan search.
- 8) A number of sequences/DEGs with PHOBIUS annotation from InterProScan search.
- 9) A number of sequences/DEGs with PRINTS annotation from InterProScan search.
- 10) A number of sequences/DEGs with PROSITE_PATTERNS annotation from InterProScan search.
- 11) A number of sequences/DEGs with SIGNALP_EUK annotation from InterProScan search.

- 12) A number of sequences/DEGs with SIGNALP_GRAM_NEGATIVE annotation from InterProScan search.
- 13) A number of sequences/DEGs with SIGNALP_GRAM_POSITIVE annotation from InterProScan search.
- 14) A number of sequences/DEGs with SMART annotation from InterProScan search.
- 15) A number of sequences/DEGs with SUPERFAMILY annotation from InterProScan search.
- 16) A number of sequences/DEGs with COILS annotation from InterProScan search.
- 17) A number of sequences/DEGs with HAMAP annotation from InterProScan search.
- 18) A number of sequences/DEGs with PROSITE_PROFILES annotation from InterProScan search.
- 19) A number of sequences/DEGs with TIGRFAM annotation from InterProScan search.
- 20) A number of sequences/DEGs with TMHMM annotation from InterProScan search.
- 21) A number of sequences/DEGs with PIRSF annotation from InterProScan search.
- 22) A number of sequences/DEGs with PRODOM annotation from InterProScan search.
- 23) A number of sequences/DEGs with NO annotation.

6.0 Database structure

The database created by SATrans includes 9 database tables (**Figure 3**), which are identical to the “export” mode output tables. The data are inserted into these tables from a MySQL database managed by the Gene Ontology Consortium (<http://www.geneontology.org/>). The table 1_Sequence contains basic information concerning the sequences stored in the database such as name, length of sequence and status of the annotation process. In the tables 2_Hits and 3_InterProScan_data are stored the results of functional annotation performed by BLAST and InterProScan, respectively. Tables named 4_term and 5_term2term contain information about the assigned GO terms and their description and the relationships between the terms. Information about DEGs is stored in the table 7_DEGs. Both annotation tools (BLAST and InterProScan) record the GO terms which are stored in table 6_GO_parse. Table 8_GO_analysis contains basic information about the GO annotation of each analyzed sequence and table 9_GO_results stores the results of the GO term analysis with respect to the DEGs performed by the „analysis“ mode.

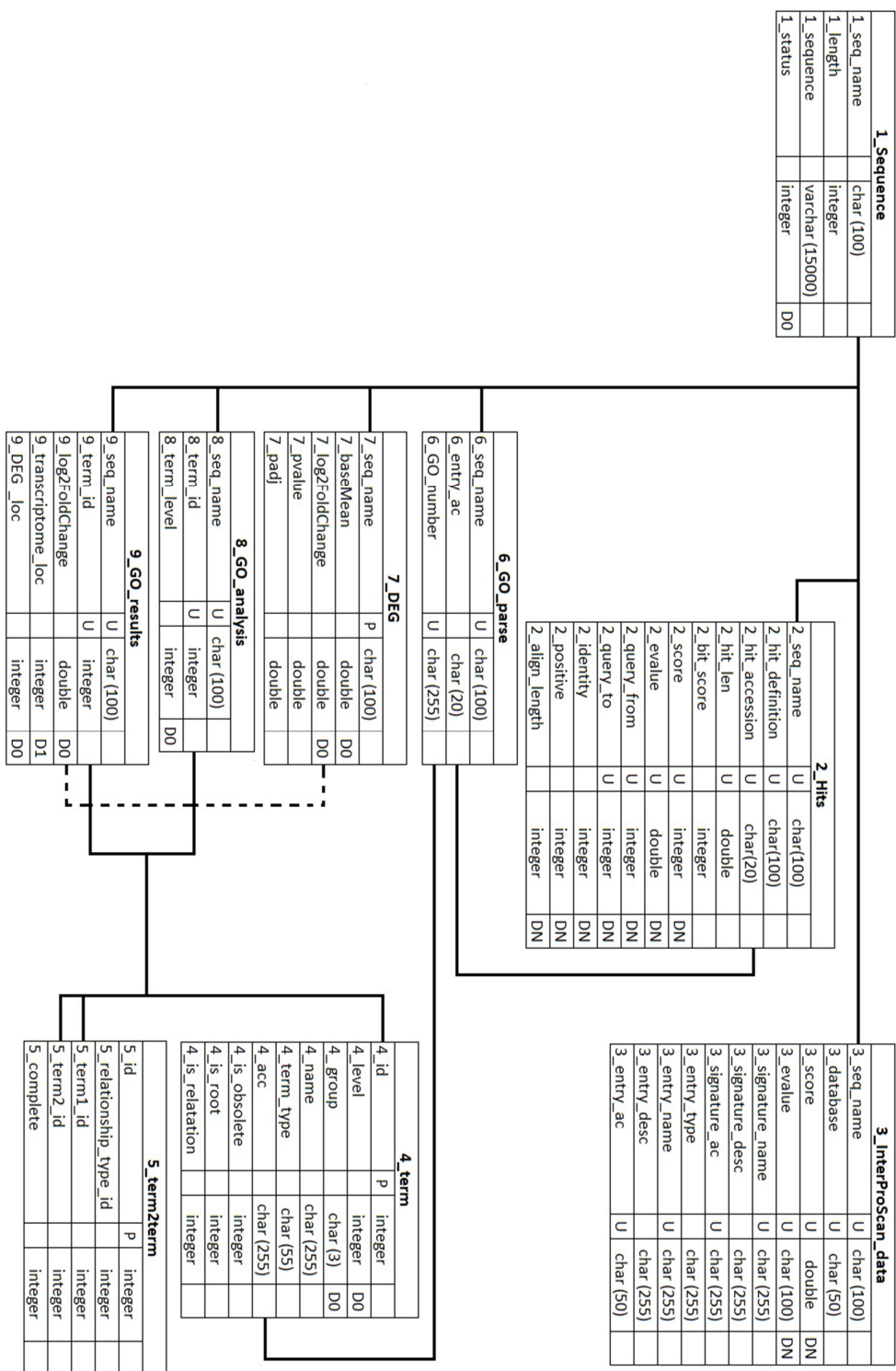


Figure 3: The structure of the database which is created by the SATrans. P primary key; U unique; DN default null; D0 default value 0; D1 default value 1.

6.1 Description of database tables

6.1.1 Table 1_Sequence

Name of the column	Format of value	Description of column
1_seq_name	char (100)	Name of the sequence.
1_length	integer	Length of the sequence.
1_sequence	varchar (15000)	Sequence.
1_status	integer	Status of the annotation process. If 0 or 1, the annotation of the sequence is not finished yet. If the value is 2, the annotation was successfully finished.

6.1.2 Table 2_Hits.

Name of the column	Format of value	Description of column
2_seq_name	char (100)	Name of the sequence.
2_hit_definition	char (100)	Description/title of the matched database sequence.
2_hit_accession	char (20)	Accession number of the matched database sequence.
2_hit_len	double	Length of the matched database sequence.
2_bit_score	integer	Alignment bit-score.
2_score	integer	Alignment raw score.
2_evalue	integer	Alignment Expect value.
2_query_from	integer	Start of alignment in the query sequence.
2_query_to	integer	End of alignment in the query sequence.

2_identity	integer	A number of identical matches.
2_positive	integer	A number of positive-scoring matches.
2_align_length	integer	Alignment length in the query sequence.

6.1.3 Table 3_InterProScan_data

Name of the column	Format of value	Description of column
3_seq_name	char (100)	Name of the sequence.
3_database	char (50)	Name of the sub-database form the InterProScan.
3_score	double	Alignment score.
3_evalue	char (100)	Alignment Expect value.
3_signature_name	char (255)	Name of the InterProScan signature.
3_signature_desc	char (255)	Description of the InterProScan signature.
3_signature_ac	char (255)	Accession number of the InterProScan signature.
3_entry_type	char (255)	Type of the entry which tells you what you can infer when a protein matches the entry. The entry types are: H – Homologous Superfamily, F – Family, D – Domain, R – Repeat, S - Site
3_entry_name	char (255)	Name of the InterProScan entry.
3_entry_desc	char (255)	Description of the InterProScan entry.

3_entry_ac	char (50)	Accession number of the InterProScan entry.
------------	-----------	---

6.1.4 Table 4_term

Name of the column	Format of value	Description of column
4_id	integer	Unique serial number.
4_level	integer	The level number within the directed acyclic graph. Level 1 represents the most general categories and provides the most coverage, whereas Level 5 provides more specific information and less coverage.
4_ontology_source	char (3)	Denotes which of the three sub-ontologies – Molecular Function (MF), Biological Processes (BP), Cellular Component (CC) or Undefined (UN) – the term belongs.
4_name	char (255)	Character varying a textual label for the term. Each term has a single such label. The name should be unique within an ontology, the uniqueness recommendation is relaxed in the case of obsolete terms, which are also housed in this table: there can be many "ex-terms" with the same name. (Example: "cysteine biosynthetic process")
4_term_type	char (55)	The ontology or namespace to which this term belongs (OBO-Format: *namespace* tag) (Example: biological_process) (Note: the column name is somewhat misleading, but is retained for historical reasons. It would be better named "namespace" or "ontology") The namespace for GO terms will always be molecular_function, biological_process or cellular_component. The relations defined in the main GO obo file (from which this table is populated) go into the gene_ontology namespace, with the exception of is_a, which has namespace "relationship" (taken from the obo

		relation ontology). is_a is a built-in relation as far as obo is concerned, so it does not go in the gene_ontology namespace
4_acc	char(255)	Unique identifier for this term. This should be in OBO bipartite ID format and should be unique within OBO, but this is not enforced at the schema level. (Example: GO:0019344)
4_is_obsolete	integer	Equals 1 if this row corresponds to an obsoleted "ex-term". Valid values: 0 or 1.
4_is_root	integer	Equals 1 if this term is the root term in the ontology graph.
4_is_relation	integer	Equals 1 if this term is a relation (relationship type).

6.1.5 Table 5_term2term

Name of the column	Format of value	Description of column
5_id	integer	Unique serial number.
5_relationship_type_id	integer	References an entry in the term table corresponding to the relation that holds between term2 and (Example: a reference to a row "part_of" in the term table).
5_term1_id	integer	The "parent" node of the edge. For example, in the edge corresponding to "nucleus part_of cell", (all nuclei are part_of some cell) term1_id is "cell".
5_term2_id	integer	The "child" node of the edge. For example, in the edge corresponding to "nucleus part_of cell", (all nuclei are part_of some cell) term2_id is "nucleus".
5_complete	integer	Equals 1 if this edge comprises an element of the *complete definition*, a set of necessary and sufficient conditions. Note that this field is always =0 for current publically deployed instantiations of the GO database,

		but is currently used in experimental instantiations for housing so-called "cross-products".
--	--	--

6.1.6 Table 6_GO_parse

Name of the column	Format of value	Description of column
6_seq_name	char (100)	Name of the sequence.
6_entry_ac	char (20)	Accession number of the InterProScan entry. Exported from the table 3_InterProScan_data, value 3_entry_ac.
6_GO_number	char (255)	GO term number identifier in the defined structure (GO:0000644).

6.1.7 Table 7_DEGs

Name of the column	Format of value	Description of column
7_seq_name	char (100)	Name of the sequence.
7_baseMean	double	baseMean value from the differential gene expression input file used in the “analysis” mode.
7_log2FoldChange	double	log2FoldChange value from the differential gene expression input file used in the “analysis” mode.
7_pvalue	double	pValue from the differential gene expression input file used in the “analysis” mode.

7_padj	double	pAdj value from the differential gene expression input file used in the “analysis” mode.
--------	--------	--

6.1.8 Table 8_GO_analysis

Name of the column	Format of value	Description of column
8_Seq_name	char (100)	Name of the sequence.
8_Term_id	integer	Unique serial number exported from table 4_term.
8_Level	integer	Level of GO term in a tree structure.

6.1.9 Table 9_GO_results

Name of the column	Format of value	Description of column
9_seq_name	char (100)	Name of the query sequence.
9_Term_id	integer	Unique serial number exported from table 4_term.
9_log2FoldChange	double	log2FoldChange value from the differential gene expression input file used in the “analysis” mode.
9_transcriptome_loc	integer	A number of all sequences from the multi-fasta input file annotated by the GO term.
9_DEGs_loc	integer	Number of DEGs annotated by the GO term.

7.0 References

- [1] Ashburner,M *et al.* (2000) Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, 25, 25–29.
- [2] Love,M.I. *et al.* (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, 15, 550.
- [3] Fisher,R.A. (1935) The logic of inductive inference. *J. Roy. Statist. Soc.*, 98, 39–82.