Target Sequence Analysis Tool (TSAT)

User Guide

Contents

Source code & Software

The Source code for TSAT can be found on Github:

https://github.com/taltendorf/Target-Sequence-Analysis-Tool-TSAT-

Alternatively, the source code can be found under the following DOI:

10.5281/zenodo.10342035

A released executable version of TSAT can be downloaded here:

https://github.com/taltendorf/Target-Sequence-Analysis-Tool-TSAT-/releases/tag/v1.0

Objective of TSAT

TSAT is a computational tool for the rapid analysis of NGS data obtained from Phage display and Mirror Image Phage display selections. It is able to extract desired genetic information, translate the information from DNA into protein and in an optional second step filter the obtained information for sequences with a higher probability to bind targets while excluding sequences that may not be target specific due to unspecific binding, amplification advantages or a bias in the library distribution. It was developed using the modules Tkinter, Bio, base64, re, collections, sqlite3, time, datetime, sys, threading and was written in Python version 3.6.5.

Prerequisite

TSAT accepts data in the form of FASTQ files encoded with UTF-8. To directly investigate the processed information a secondary program to visualize databases will be needed (e.g., db-browser for sqlite). Extraction of sequence information and translation into protein can be done without secondary programms. For the filtering of unwanted sequences as well as generation of different scores TSAT

needs at least one library file, one target selection as well as one empty selection of the same round as input data.

Startup

TSAT can be started by running the executable or source code with a python interpreter.

How to use

Upon start TSAT needs information input from the user to function correctly. Following is a step-by-step guide on how to fill in all the information.

Database creation

In a first step the user needs to create a new database or connect to an existing one. It is recommended to create a new database for every run of TSAT to avoid loss or corruption of data. By clicking the "create new database" button the user is able to create and name a new database file (.db).

Data insertion

The data that TSAT will analyze must be provided by the user using the "select file" buttons. The user is asked to provide a FASTQ file. Only one file can be given per path. If the user wants to use the optional filtering it is important that the corresponding file is inserted into each path. TS stands for Target selection. This selection should be done against the target and is divided into 1-3 according to the increasing selection rounds. ES stands for empty selection. This selection is not against the target and should be done in parallel to the Target selection. It is used to identify sequences that are unlikely to be associated with the target but are present due unspecific binding. Again, this is divided into 1-3 according to the selection round. DC stands for Direct control. This selection was conducted against the target in the previous round and is subsequently performed without a target. DC selections are divided into 2 and 3 as the first round has to be the TS1 round. The library file is the final path option and represents the phage library state before a selection was performed. This file is used to identify sequences that are elevated in their frequency due to an amplification advantage or bias of the library. After file selection the path to the file is presented in the window. In case of a mix up it is possible to change the selected file by selecting another. After the files have been selected the user has to check which files were provided by marking the checkboxes to the right of the select file buttons. This is vital as TSAT won't operate on files that were not confirmed here. A user can use this feature to include or exclude a certain selection round for a second analysis without having to re-enter general Information. If a second analysis is performed, it is recommended to create a new database and connect to this newly created database. TSAT does not overwrite information already present in the database.

Translation direction

Here the user has to choose one of four possibilities for how the translation of excised DNA should be performed. Depending on how the phage library is set up the possibilities are Forward, Reverse, Forward complement and Reverse complement.

Framing region input

TSAT identifies desired nucleotide information via a regular expression function. This identification is dependent on framing regions surrounding the desired nucleotide information as they are part of the search pattern given to the function. The information on the search pattern has to be provided but may be saved in a database accessible with SQL for fast accessibility during future reuses. The search pattern is expected in the following format XXX(.+?)XXX where (.+?) represents the unknown region of interest and XXX stands for the framing nucleotides, which have to be provided by the user. During the operation the framing regions denoted by XXX define when the desired information starts and when it

ends. The (.+?) modifier defines that the program should match all characters that are between the two framing regions, then stop after the first match and return the result. This approach was used to increase the analysis speed of TSAT and to ensure that each genetic sequence can only provide one returned peptide. An example for a valid search pattern would be GATTCCAGG(.+?)TACGACCCG. All provided data will be searched with the same search pattern by the regular expression function; it is not possible to mix search patterns during an ongoing analysis. In order to use the manual inputs, the user has to check the checkbox for Manual input and provide the framing regions in the entry box to the left. Alternatively, the user may create a new database containing frequently used framing regions. This is done by using the connect/create button under “select them from the database”. Here the user can connect to an existing database (if no database is available the user can create a new empty database by using the create new database button on the top left corner of the screen) with framing region. If a connection has been established successfully the user will see two to three new buttons appear. If no framing regions are present in the database only two buttons will appear. Add new framing regions entry will open a new window where the user can insert the required data in the provided entry boxes and insert the data into the empty database. Be careful to enter all information before inserting this information into the database as every entry will be counted as a separate entity. The ID will later be visible for the user to select from the database so be sure to name the entry that it is recognizable for later use. The framing regions have to be inserted into the second entry box. Identical to the manual insertion the format is XXX(.+?)XXX. The third entry box may be used to enter an Author name. It will be saved in the database as the person who inserted the data. This information will not be displayed anywhere and can only be accessed by direct interaction with the database. The translation direction (Forward, Reverse, Forward complement or Reverse complement) may be inserted here but is not yet used by TSAT. This may change in a future update. After insertion of a new set of framing regions the user has to click the refresh button in order to reconnect to the database and display the newly entered information. A dropdown menu will appear containing the ID of every entry in this database. By selecting an ID, the corresponding framing regions will be used by TSAT during its operation. If this method is used the manual input checkbox mustn’t be check.

User

Here the user may insert his name. This data will be saved in the database. This step is optional.

Selection

Here a name for the selection may be given. This data will be saved in the database. This step is optional.

Start data extraction and translation

After all the required information has been entered and selected the operation may be started by using the Go button at the bottom of the program. TSAT will analyze the provided information and provide an overview over the progress in the current status box.

Filtering of processed data

The second functionality of TSAT is to filter the processed data in order to identify sequences that have a higher probability of being target specific and to exclude sequences that are thought to have a high frequency due to amplification advantage, starting bias or non-specific binding. To use this feature, one can either connect to a database containing at least one target selection, one empty selection and a library file or directly use the feature after finishing an extraction a translation with at least the mentioned files. To start the filtering, check the boxes corresponding to the files that are contained in the connected database under “please select your tables”. The filtering is started by pressing the Sqlite Operation button at the bottom of TSAT. The filtering may take a while depending on the operating

system. During the filtering TSAT will create two new scores to present two key features of every sequence, the empty and enrichment score. The enrichment score represents the frequency of a sequence during its last target selection round divided by the frequency of the same sequence in the library. This score is used to identify sequences that are only present as a result of a biased sequence distribution within the used phage library. The empty score represents the frequency of a sequence during its last target selection round divided by the frequency of the same sequence in the last empty selection round. This score is used to identify sequences that are present as a result of binding to other components of the selection setup (e.g., plate, selection medium or blocking agent). TSAT excludes sequences that are smaller than 8 amino acids. Additionally, every sequence is analyzed for its enrichment from target round to target round as well as compared to the empty selection and direct control. Sequences that do not increase from one target round to the next are excluded. Sequences that have a higher frequency in the corresponding empty selection and sequences whose frequency increases in a direct control compared to the previous target selection are removed. A new file is created after the filtering is completed. The user can decide where the file should be saved and might rename it. The file will be saved as a .FASTA file. All sequences that were not discarded are saved in FASTA format. Sequences are sorted in descending order according to their respective Empty score.

Citation

If you used TSAT in your research and used the obtained data in a publication please cite the usage of this computational tool. TSAT was published under the following DOI:

TSAT is currently not published. This will be updated as soon as it is.

Q&A

Q: I see a star shape in my protein code, what is that?

A: The star shape is used if a stop codon was translated. This might happen when you use strains that substitute a certain stop codon with another amino Acid (e.g., amber suppressor).

Q: How can I review the translated data that is stored in the database?

A: TSAT cannot display the translated data that is stored in the database. For this please use a different software that is able access and display such data. As an example, you could use DB browser for SQLite

Q: How does TSAT handle sequences that are not present in the empty selection or library?

A: In this case the sequence will be handled as if it was present 0.5 times in the selection or library.

Q: Why do I need the library file for filtering?

A: The library is used to check the state before a selection occurred. In the filtering it serves as a “target selection 0”.

Q: Why do I see the same sequence twice or more often?

A: This happens when there are multiple codons that code for the same amino acid. TSAT orders at the level of DNA meaning those “duplicated” sequences all have a different genetic code.