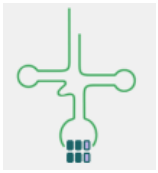


stAIcalc User Guide

1 Introduction

$stAI_{calc}$ is a user interface enabling the calculation of the tRNA Adaptaion Index (tAI, dos Reis, *et al.,* 2004) based on a species-specific inference of the tAI wobble weights. $stAI_{calc}$ is an online tool includes 100 different species from the three domains of life, along with an option to download a standalone software package of the $stAI_{calc}$ that calculates tAI with species-specific weights for new organisms not included in the list.
$stAI_{calc}$ infers the species-specific tAI wobble weights based on the approach described in (Sabi and Tuller, 2014). The main advantage of the approach is the fact that it is based only on the tRNA copy numbers (or tRNA levels if available) and the coding sequences and requires no additional gene expression information.

$stAI_{calc}$ is available online at: http://tau-tai.azurewebsites.net/.
$stAI_{calc}$ is also available for download as a standalone graphical user interface* at: http://tau-tai.azurewebsites.net/DownloadStandAlone.aspx.
The standalone software package requires only the Matlab runtime which is freely available for download within the package.
Installations are available for Windows 64-bit, MACOSX and Linux.

The output of $stAI_{calc}$ is a text file of the tAI values of the input genes and local tAI values for the 61 coding codons, both, calculated based on the species-specific weights.

*For large scale analyses we recommend using the standalone package.

2 stAI$_{calc}$ description

Both, the online tool and the standalone application of stAI$_{calc}$ enable the calculation of the tAI of coding sequences based on species-specific weights related to the efficiencies of the different wobble interactions. By the tAI, the estimated translation efficiency of a translated codon depends on the intracellular concentration of the tRNAs that recognize it and the efficiency of the corresponding codon-anticodon pairing. Specifically, the translational efficiency of the *i*-th codon (out of 61 possible codons) is given by:

$$W_i = \sum_{j=1}^{n_i} (1 - s_{ij}) \cdot tGCN_{ij}$$

Where n_i is the number of tRNA types/anticodons that pair with the *i*-th codon, *tGCNij* is the tRNA gene copy number* (or measured tRNA levels if available); and S_{ij} is a value between 0 and 1 representing a constraint on the pairing between the *i*-th codon and the *j*-th tRNA. The closer to zero the *Sij* weight is, the more efficient the wobble interaction. Normalized weights (w_i) are obtained from the *Wi* weights by dividing each W_i by the maximum value among all codons. Zero W_i weights are set to the mean w_i.

stAI$_{calc}$ is based on the approach suggested in (Sabi and Tuller, 2014) for inferring species-specific S_{ij} weights. The weights were optimized based on a hill-climbing approach that aims at maximizing the correlation between a measure of codon usage bias (inspired by the relative CUB, (Roymondal, et al., 2009)) and the tAI. Specifically, at the first step, an initial guess for the S_{ij} was chosen and the tAI was calculated based on this guess; the S_{ij} vector was iteratively changed until no further improvement in the correlation between tAI and CUB could be achieved.
As the wobble weights are inferred individually for each organism via optimizing the correlation between tAI and CUB, tAI values are expected to correlate with protein levels in organisms where CUB is related to the amount of adaptation to the tRNA pool. This package also provides the final correlation between the optimized tAI and CUB for the user evaluation.

Copy numbers can be retrieved, for example, from the genomic tRNA database at http://gtrnadb.ucsc.edu (Chan and Lowe, 2009)).

3 The online tool

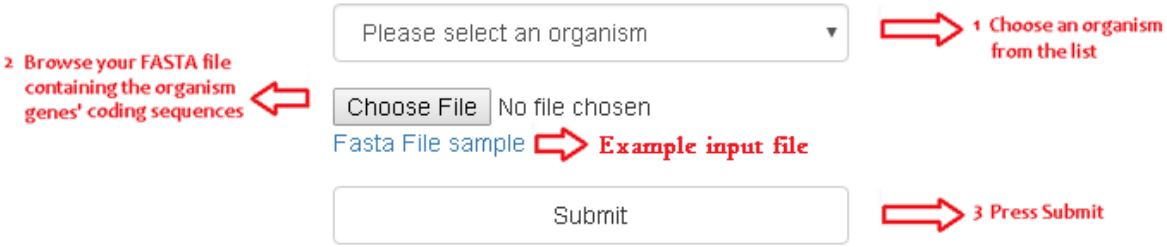
The online tool of $stAI_{calc}$ provides tAI values of genes based on species-specific weights previously inferred in (Sabi and Tuller, 2014). A list of 100 different organisms from the three domains of life available at $stAI_{calc}$ opening page.

3.1 Input

The use of the online $stAI_{calc}$ is straightforward requiring only three simple steps:

1) ***Selecting an organism from the list****:* Organisms' names include the specific genome source and the GenBank record used to infer the S_{ij} in (Sabi and Tuller, 2014).

2) ***Choosing input FASTA file*****:** the user should provide a text file in a FASTA format containing the organism genes' coding sequences. By clicking the "Choose file" tab, a browser is opened allowing the user to upload a file.



The FASTA formatted file should be a text file containing genes names (or other identifier) as the headers followed by the coding sequence (**the coding sequence must start one row after the header** with no empty sequence entered between the header and the coding sequence), for example (*Saccharomyces cerevisiae s288c* gene):

```
>YAL003W
ATGGCATCCACCGATTTCTCCAAGATTGAAACTTTGAAACAATTAAACGCTTCTTTGGCTGACAAGTCATACATTGAAGGTACTGCTGTTTCTCA
AGCTGACGTCACTGTCTTCAAGGCTTTCCAATCTGCTTACCCAGAATTCTCCAGATGGTTCAACCACATCGCTTCCAAGGCCGATGAATTCGAC
TCTTTCCCAGCTGCCTCTGCTGCCGCTGCCGAAGAAGAAGAAGATGACGATGTCGATTTATTCGGTTCCGACGATGAAGAAGCTGACGCTGAA
GCTGAAAAGTTGAAGGCTGAAAGAATTGCCGCATACAACGCTAAGAAGGCTGCTAAGCCAGCTAAGCCAGCTGCTAAGTCCATTGTCACTCT
AGATGTCAAGCCATGGGATGATGAAACCAATTTGGAAGAAATGGTTGCTAACGTCAAGGCCATCGAAATGGAAGGTTTGACCTGGGGTGCTC
ACCAATTTATCCCAATTGGTTTCGGTATCAAGAAGTTGCAAATTAACTGTGTTGTCGAAGATGACAAGGTTTCCTTGGATGACTTGCAACAAAG
CATTGAAGAAGACGAAGACCACGTCCAATCTACCGATATTGCTGCTATGCAAAAATTATAA
```

You can also download the sample input file which appears below the 'choose file' button.

3.2 Output

By clicking the "Submit" button, a new page is opened presenting both, the tAI values of the input coding sequences and the w_i values of each codon.
The following output, for example is shown for the input sequence presented in section 3.1:

Gene Name	tAI
YAL003W	0.4267

Export The Table Data into a CSV File

w(codon) for Organism : Saccharomyces cerevisiae S288c PLN 26-JUN-2012

TTT	TTC	TTA	TTG	TCT	TCC	TCA	TCG	TAT	TAC	TGT	TGC	TGG	CTT	CTC	CTA	CTG	CCT	CCC
0.069	0.518	0.362	0.793	0.569	0.219	0.155	0.170	0.055	0.414	0.028	0.207	0.311	0.007	0.052	0.155	0.118	0.104	0.219

CCA	CCG	CAT	CAC	CAA	CAG	CGT	CGC	CGA	CGG	ATT	ATC	ATA	ATG	ACT	ACC	ACA	ACG	AAT	AAC	AAA
0.518	0.393	0.048	0.362	0.466	0.406	0.311	0.219	0.219	0.052	0.673	0.219	0.104	0.518	0.569	0.219	0.207	0.209	0.069	0.518	0.362

AAG	AGT	AGC	AGA	AGG	GTT	GTC	GTA	GTG	GCT	GCC	GCA	GCG	GAT	GAC	GAA	GAG	GGT	GGC	GGA	GGG
1	0.014	0.104	0.569	0.485	0.725	0.219	0.104	0.182	0.569	0.219	0.259	0.197	0.104	0.776	0.725	0.654	0.111	0.828	0.155	0.222

Export The Table Data into a CSV File

Output interpretation: The tAI value calculated for a gene is based on the geometrical mean of the w_i of the codons composing it. The tAI of a gene represents the amount of adaptation of the coding sequence to the intracellular tRNA pool of the organism (dos Reis, et al., 2004).

The values appear below the w(codon) are the w_i weights, these are based on the concentration of the tRNAs that recognize the codon and the efficiency of the corresponding codon-anticodon pairing (as inferred species-specifically). The w_i value of the *i*-th codon is equivalent to the nominal speed it takes for the ribosome to translate it. These values are normalized by the maximum w_i, thus, the codon owning a value of 1 is the fastest codon in the sequence (with respect to tRNA concentration and pairing efficiency).

4 The standalone application

The standalone package of $stAI_{calc}$ is a user interface enabling the calculation of tAI with species-specific weights (Sabi and Tuller, 2014) for any given organism for whom coding sequences and tRNA gene copy numbers/tRNA levels exist.

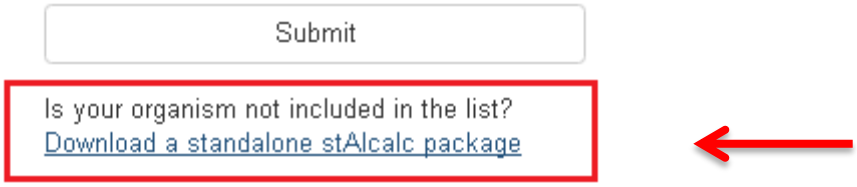
The package can be used on computers that do not have MATLAB installed.

The package is suitable for large scale analyses. It can be used also for organisms that are not included in the list provided at the online tool.

4.2 Installation guide

To install the package please follow these 3 steps:

1) **Download the package to your computer**: The package is available for download below the submit button of the online tool:



2) **Install the Matlab runtime***: You can either:

Option 1:* Run the installation file ***MyAppInstaller_web* which is included in the package (**recommended**).

**Option 2:* Navigate to http://www.mathworks.com/products/compiler/mcr/index.html and download the appropriate runtime version*:

Version 9.0 (R2015b) for **Windows** 64-bit and **MACOSX**.

Version 9.0.1 (R2016a) for **Linux**.

3) **Run the application:** If you are using **windows**, simply run the executable file ***stAIcalc.exe***.

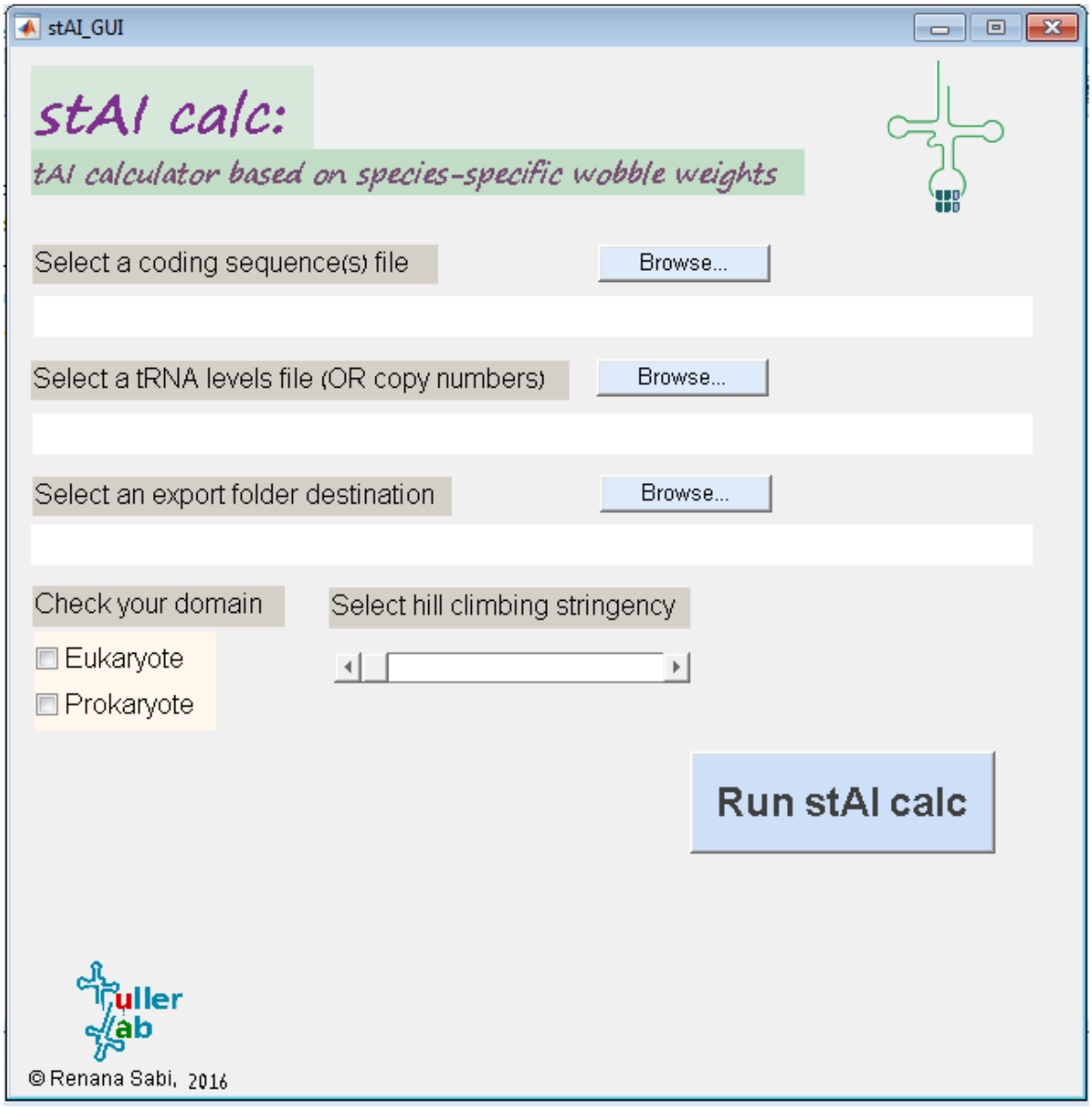
If you are using **Linux** or **MACOSX**, you should include the path where the runtime is installed. Open the command prompt and type ./run_stAIcalc.sh <mcr_directory> <argument_list> where <mcr_directory> is the directory where the appropriate MATLAB Runtime is installed* and <argument_list> is all the arguments you want to pass to your application. *For example*, if you have version 9.0.1 of the MATLAB Runtime installed in /mathworks/home/application/v901, run the shell script as: ./run_stAIcalc.sh /mathworks/home/application/v901.

*Version 9.0 for win-64bit and MACOSX, version 9.0.1 for Linux. If you already have the full MATLAB installed on your computer, please check that you have the appropriate runtime version. If you do - you can find the runtime location by using the 'mcrinstaller ' function from within the Matlab command window. If you have other version of the runtime, please run the installation file *MyAppInstaller_web* which is included in the package.

4.3 Input

The $stAI_{calc}$ GUI contains three input fields, a checkbox indicating the domain of the query organism and one slider indicating the desired hill climbing stringency (see details in section 4.1.5).

Sample input files of both coding sequences and tRNA copy numbers located in a folder named sample_input_files)



4.3.1 Organism Coding Sequence

The first input is a text file in a FASTA format containing the organism coding sequence(s). As in 3.1, the FASTA formatted file should be a text file containing genes names (or other identifier) as the headers followed by the coding sequence.

Note that the package includes a sample file named: ***sample_cds_fasta.txt***
The file is located within the folder ***sample_input_files***.

4.3.2 Organism tRNA levels

The second input is a text file containing a list of anticodons (representing the different tRNA types) and their corresponding tRNA levels . If tRNA levels are not available, tRNA gene copy number can be used instead (can be retrieved for example from: http://gtrnadb.ucsc.edu). The text file should be formatted as follows:

Anticodons	tGCN
AAA	0
GAA	10
TAA	7
.	.
.	.
.	.

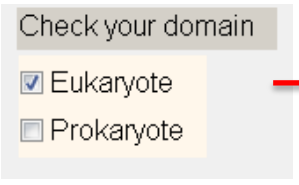
First row must include names for each column; the first for anticodons (in any desired order) and the second column for the tRNA levels (or copy numbers).

Each row (starting from the second one) represents an anticodon/tRNA. The first column contains strings and the second one contains numbers.

Note that the package includes a sample file named: ***sample_trna_txt.txt***
The file is located within the folder ***sample_input_files***.

4.3.3 Organism domain

Since there is an additional wobble interaction unique for prokaryotes (bacteria and archaea) the user should check the domain of the input organism.



If the two domains are mistakenly checked or if none is selected. Pressing "Run stAI calc" results in an error that is presented in red on the GUI.

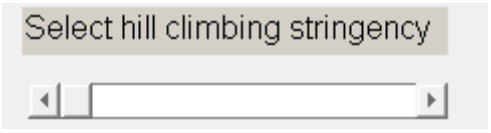
4.3.4 Output folder

The user should specify the folder path into which $stAI_{calc}$ saves the output text files (tAI and w_i values, see section 4.3).

4.3.5 Hill climbing stringency

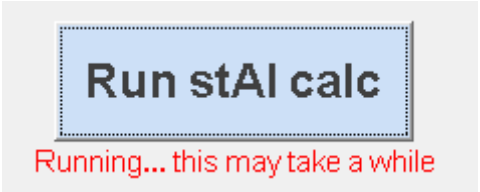
This input parameter defines the required stringency of the hill climbing (HC) approach used to optimize the s-values of the tAI. The parameter is equivalent to the number of initial guesses for the s-values used for the optimization. Higher stringency results in a better sampling of the search space, thus, has higher chances to achieve better solution. The disadvantage however, is the running time which increases with the increase in hill climbing stringency.

This parameter is presented in the GUI as a slider in which the left corner (default) represents the lowest stringency (one initial guess) and the right one represents the highest (*i.e.*, the maximum number of initial guess as described in (Sabi and Tuller, 2014).



4.4 Submission

After all inputs were defined, the stAIcalc application can be executed by pressing "Run stAI calc". If no errors are printed in red on the GUI, the optimization starts. The running time of the tool runs depends on the chosen HC stringency.



4.5 Output

4.5.1 Text files

Two output files are provided by $stAI_{calc}$ application:

- *output_tAI_file.txt:* contains the names of the input genes and their resultant tAI values based on the species-specific inferred weights.
- *output_tAI_file.txt:* contains the w_i weights for each codon.

4.5.2 Correlation between tAI and CUB

As described previously, the optimization is based on maximizing the correlation between the tAI and a measure of codon usage bias (CUB). The final correlation obtained by the optimization is presented on the GUI, for example:

5 References

Chan, P.P. and Lowe, T.M. (2009) GtRNAdb: a database of transfer RNA genes detected in genomic sequence, *Nucleic acids research*, **37**, D93-D97.
dos Reis, M., Savva, R. and Wernisch, L. (2004) Solving the riddle of codon usage preferences: a test for translational selection, *Nucleic acids research*, **32**, 5036-5044.
Roymondal, U., Das, S. and Sahoo, S. (2009) Predicting gene expression level from relative codon usage bias: an application to Escherichia coli genome, *DNA research*, **16**, 13-30.
Sabi, R. and Tuller, T. (2014) Modelling the Efficiency of Codon–tRNA Interactions Based on Codon Usage Bias, *DNA Research*, **21**, 511-526.