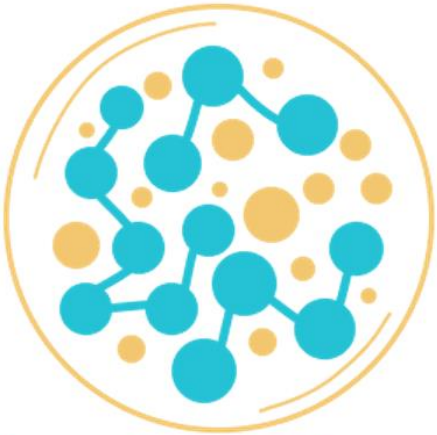
Welcome to Communique's Software!



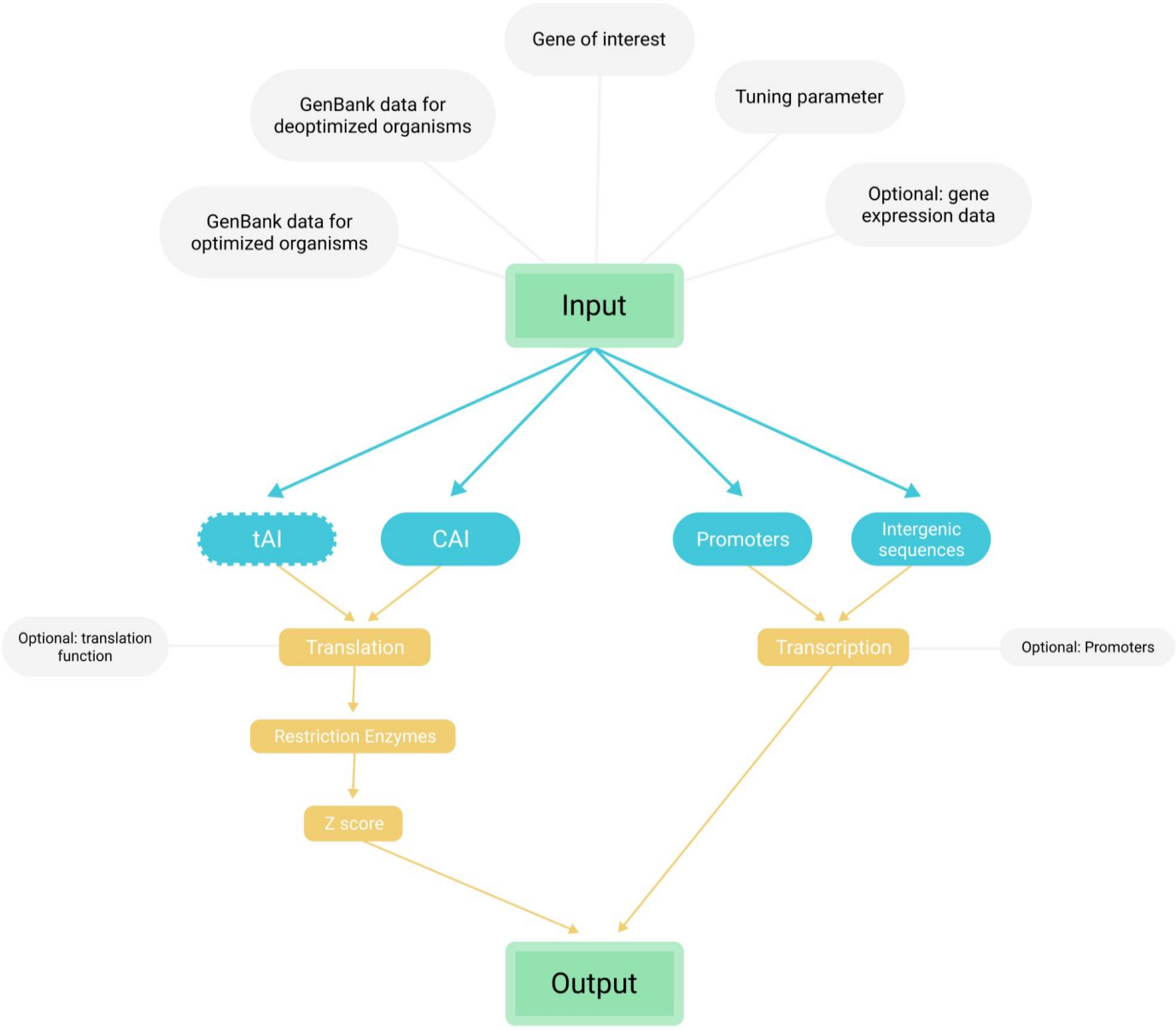
Our software aims to tailor the expression of genetic information to any microbial population, producing a regulated gene ready for transformation along with relevant documentation and predictions, presented in a user-friendly manner. Computational and experimental validations have received results beyond our expectations. Using the "enable" accessibility add-on, appropriate graphic features were added to ensure full accessibility to all users, while implementing the pipeline intuitively using a stepwise interactive framework.

Corresponding to the flexibility of metagenomic data, our software applies to communities with various degrees of characterization; from mere annotated genomes to documented gene expression levels, along with optional advanced features which are fully customizable.

For more details about the software, visit our iGEM wiki.

For any other queries, you can contact us at: igem.tau.21@gmail.com

Introduction



Our software is able to take your genetic information and create a microbiome-specific version of it- you must define two sub-populations:

- **Optimized organisms** - The organisms in the community in which you want to enhance gene expression.
- **Deoptimized organisms** - The organisms in the community in which you want to silence gene expression.

Small evolutionary changes are collected in order to optimize expression in optimized organisms while simultaneously deoptimizing it in deoptimized organisms by engineering the following aspects:

1. **Entry into the cell -** Restriction sites are avoided and placed in strategic positions in order to enhance probability of digestion upon entry to deoptimized organisms.
2. **Transcription -** The endogenous promoters of the microbial community are scanned in search for genetic motifs that promote transcription selectively in the optimized organisms, and not in the deoptimized ones. These motifs are used to select and design synthetic promoters for your gene.
3. **Translation** - The open reading frame (ORF), also called coding sequence (CDS), is synonymously re-coded in order to enhance ribosomal flow and translation efficiency selectively, by using inferred codon usage bias measurements and optimization strategies.

Your output contains all needed information for a full microbiome-specific expression cassette (promoter + open reading frame) which has optimized expression in the optimized organisms and deoptimized expression in the deoptimized organisms, along with statistical analysis of your result.

User Guide

Step 0: Installation

In order to properly install our software for your operating system, please follow the README.txt file that is found in this directory.

Step 1: Intro

In this section, each step of the process of using *Communique*'s software will be demonstrated in detail.

In order to provide increased accessibility to the software, special features may be accessed at the bottom left corner of the screen (Fig. 1, red rectangle). Before you start you can choose different options that will allow you to navigate the software in the most comfortable and optimal way for you.

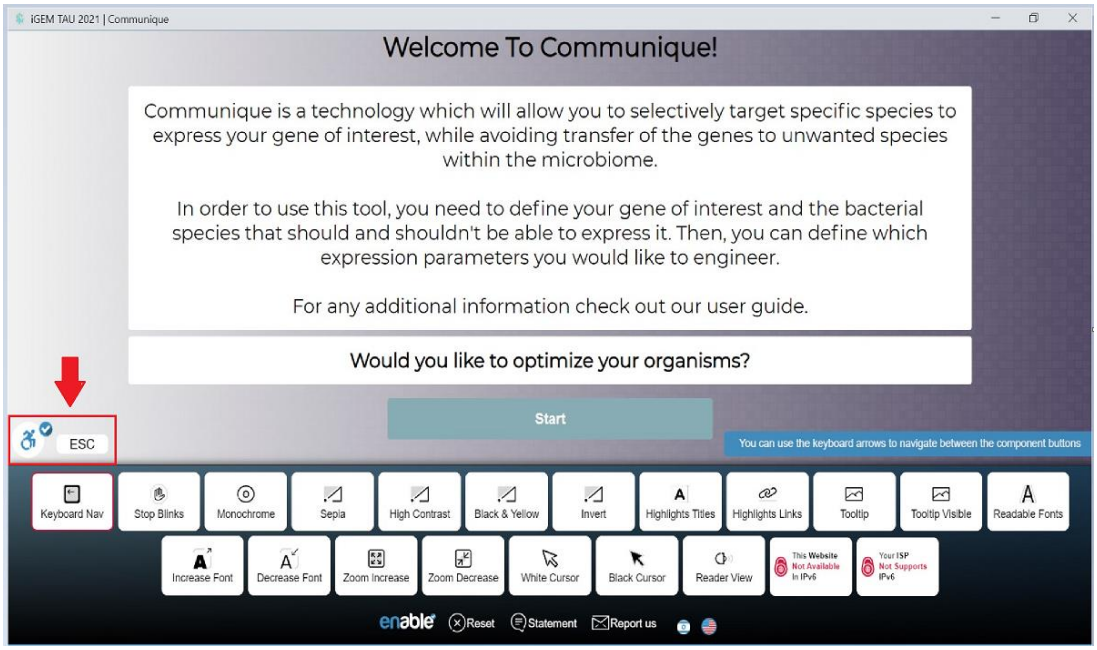


Fig. 1: Start screen and accessibility toolbar (red rectangle)

In order to start the process, please press 'Start' button (Fig. 2, red rectangle).

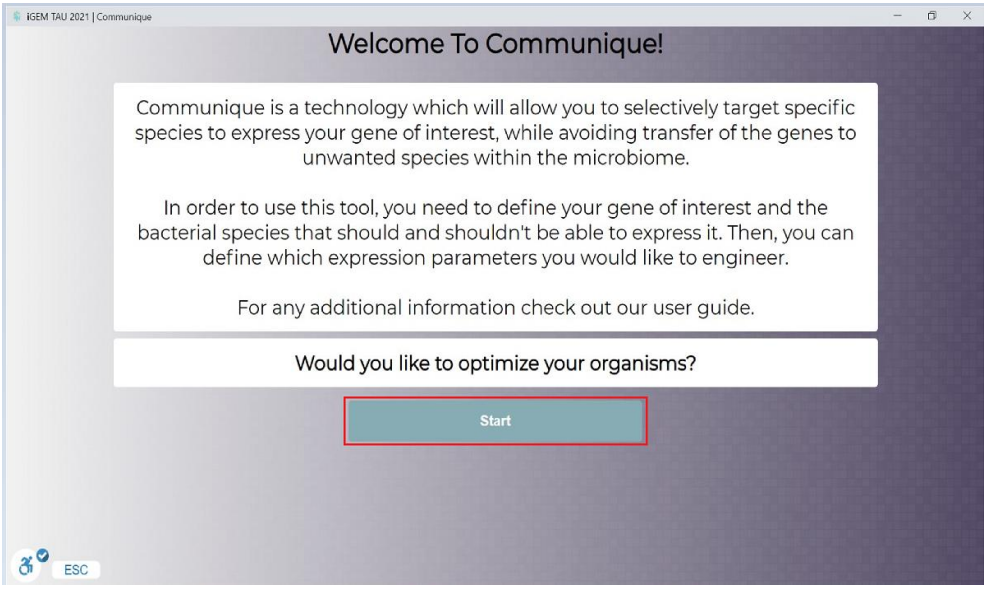


Fig. 2: Start button (red rectangle)

Step 2: Define the microbial population and preferences

Optimized Organisms

Insert the number of organisms you would like to optimize (Fig. 3, red rectangle), and then press '*Submit'* (Fig.3, blue rectangle). At least one organism is required.

In case you want to change the number you of organisms that have already been inserted, you must **first press 'Reset Organisms Number'** (Fig. 3, green rectangle).

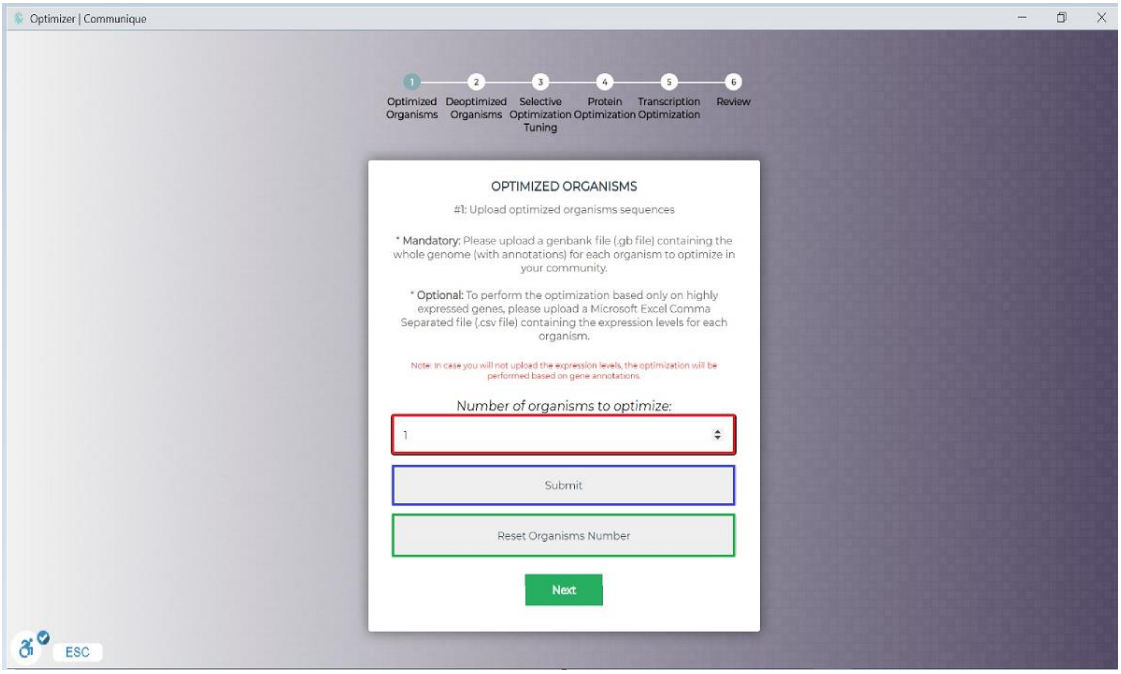


Fig. 3: Optimized organisms screen

After submission, each organism should be provided with a name (Fig. 4, red rectangle), along with a GenBank file containing the **whole genome with annotations** (Fig. 4, blue rectangle). You should be able to download GenBank files for your organisms from the NCBI database (see FAQ for more details). By clicking the '*Choose file'* button, a browser will open, allowing you to upload a file. Please notice that each organism must have a GenBank file uploaded.

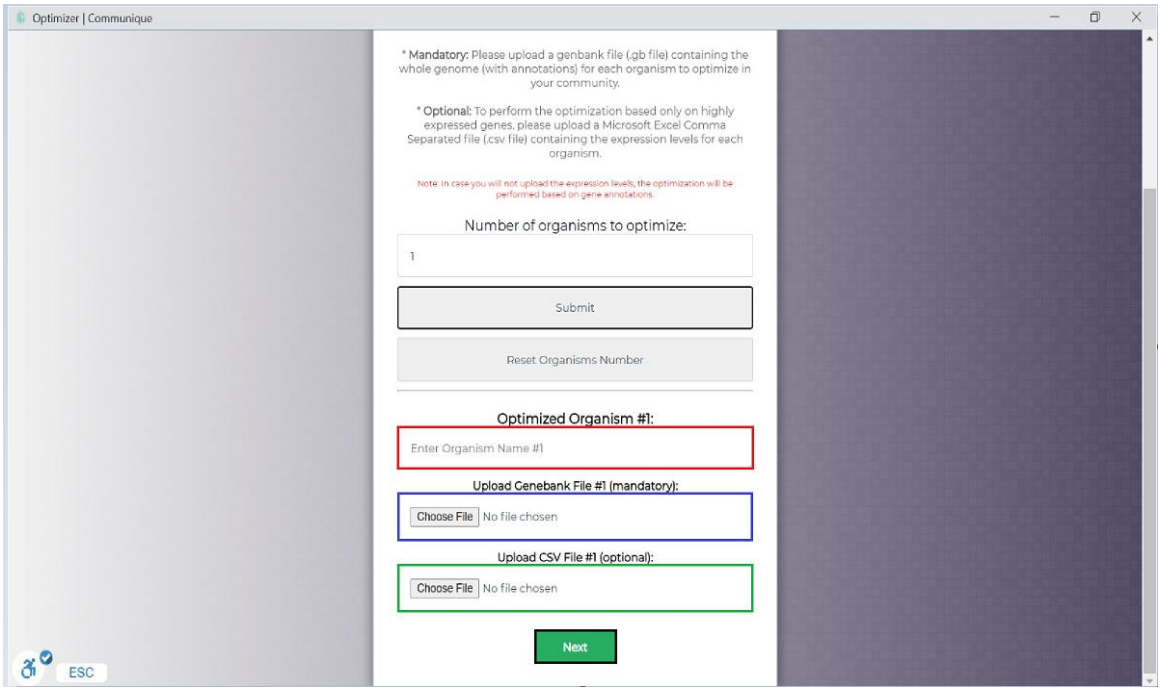


Fig. 4: Provide each organism with a name, GenBank file, and optional .csv file

Advanced option: In order to allow you to fully utilize the characterization of your microbiome, a '.csv' file of expression levels can be inserted (Fig. 4, green rectangle) and used for extraction of the highly expressed genes of your organism for the model. Please note that your file must be in the following format, otherwise the measurements cannot be used (Fig. 5):

- The gene names should be the same as the NCBI gene names of the organism
- Column with the gene names is labeled "gene"
- Column with the gene expression levels is labeled "mRNA_level"

Figure 5 shows the proper format for the expression data of *Bacillus subtilis:*

	A	B	C	D	E	F
1	gene	mRNA_level				
2	dnaA	2.61				
3	dnaN	5.1				
4	yaaA	2.24				
5	recF	8.12				
6	yaaB	7.29				
7	gyrB	10.42				
8	gyrA	9.18				
9	guaB	16.2				
10	dacA	4.3				
11	yaaD	11.94				
12	yaaE	20.23				
13	serS	22.29				
14	dck	5.27				
15	dgk	3.41				
16	tadA	0.41				
17	dnaX	1.75				
18	yaaK	5.44				
19	recR	2.81				
20	yaaL	4.2				

Bacillus_subtilis_mRNA_levels

Fig. 5: .csv example file using mRNA levels of *B. subtilis*

To continue, press *'Next'* (Fig. 4, black rectangle).

Deoptimized Organisms

This section is identical to the previous one, except it is designated for the deoptimized organisms (Fig. 6, Fig. 7).

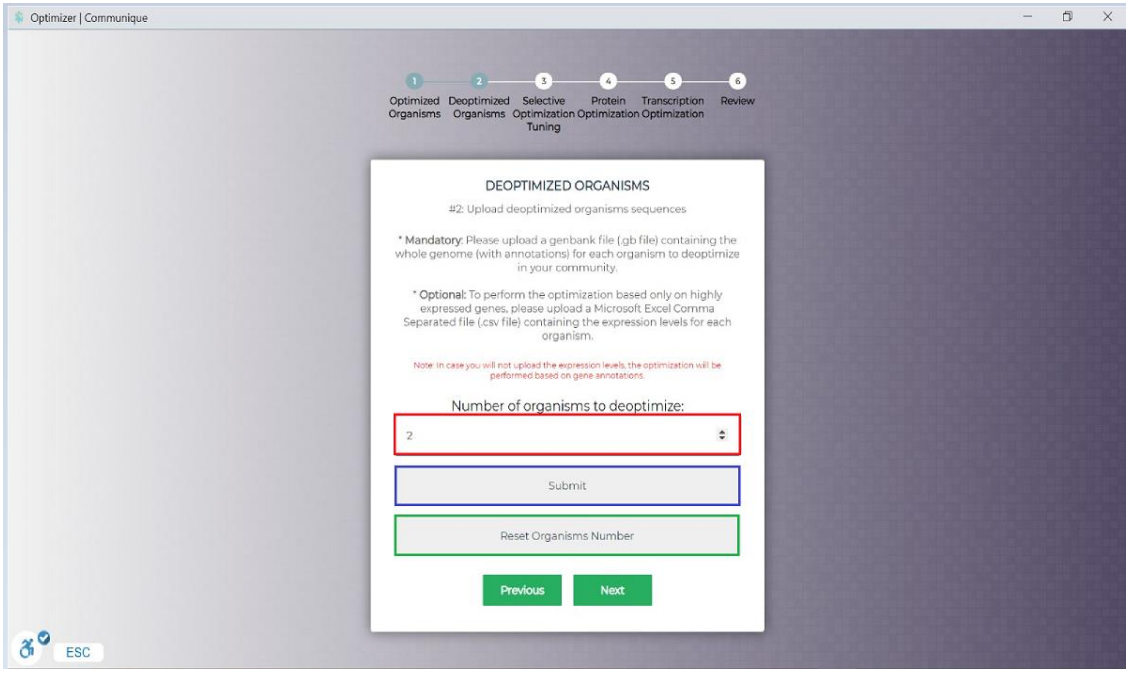


Fig. 6: Deoptimized organisms screen

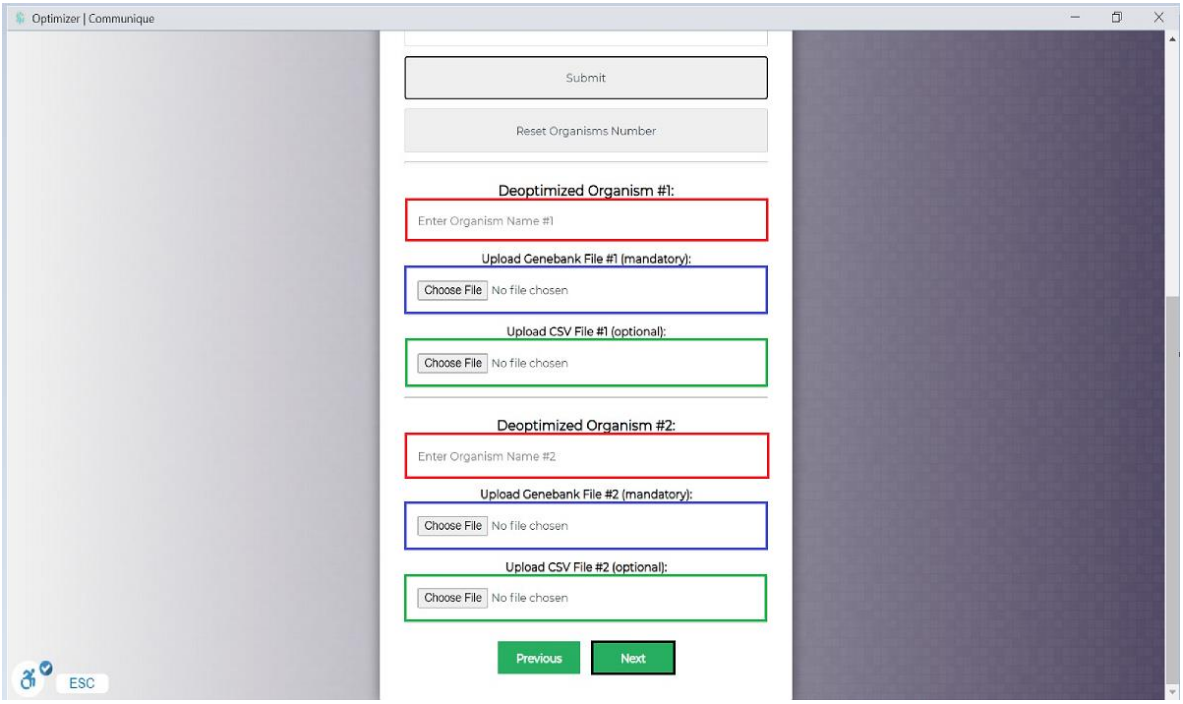


Fig. 7: Name your deoptimized organisms and upload GenBank and optionally .csv files

Selective Optimization Tuning

The **tuning parameter** allows control over the tradeoff between the degree of optimization and deoptimization. This means that you can prioritize optimization of expression in optimized organisms over deoptimization of expression in the deoptimized organisms, and vice versa (for more information, you may refer to question 5 in FAQ).

With your cursor, scroll the bar to choose a value on the scale in the range of 0-100 (Fig. 8, red rectangle) according to your preferences. To continue, press *'Next'* (Fig. 8, black rectangle).

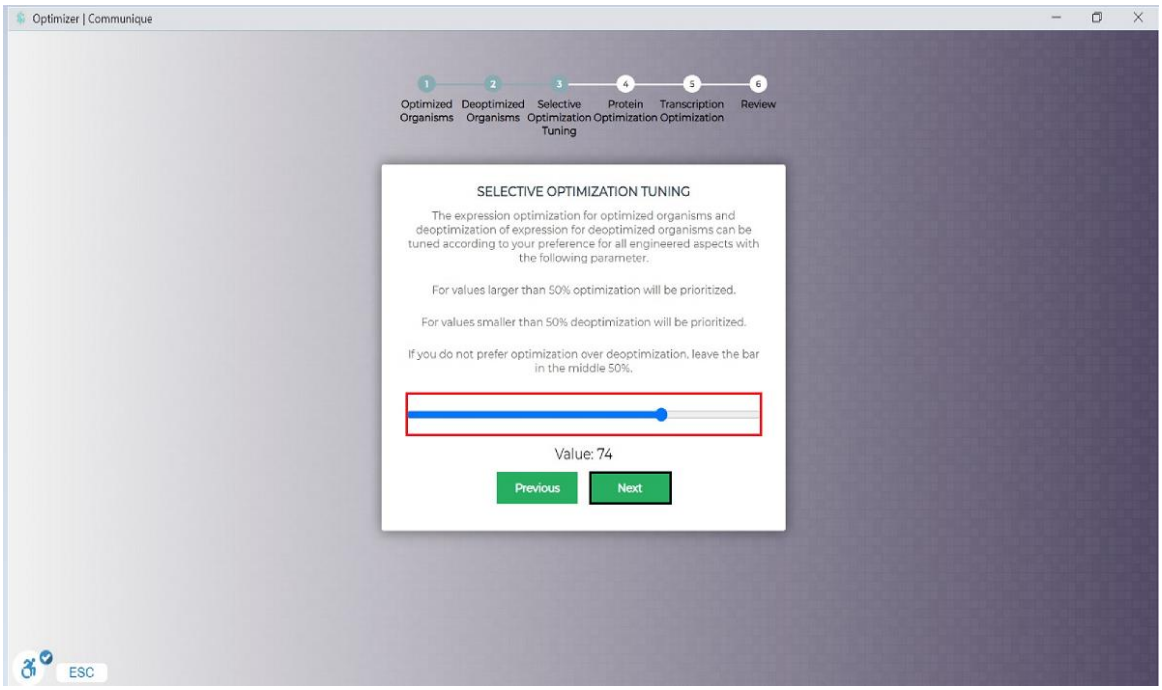


Fig. 8: Selective optimization tuning screen

Step 3: Protein optimization for your gene of interest

Protein Optimization

If you would like to *skip* protein optimization, press the *'Next'* button (Fig. 9, black rectangle).

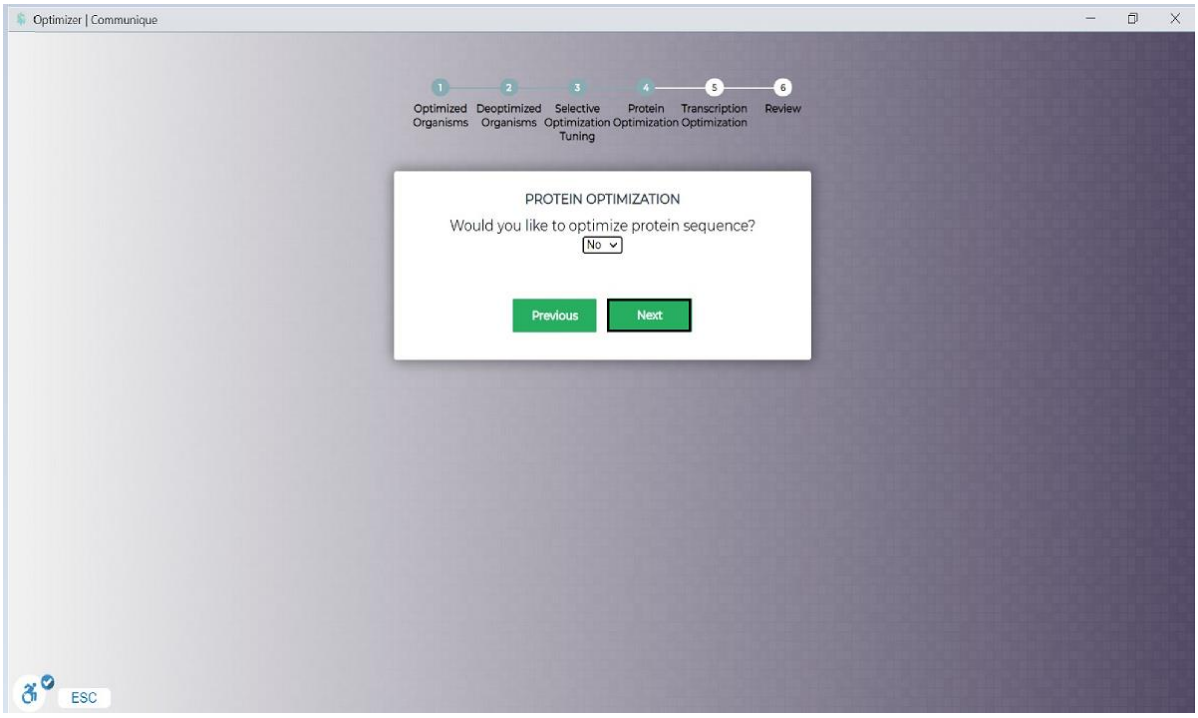


Fig. 9: Protein optimization screen

Otherwise, choose '*yes*' from the dropdown (Fig. 10, red rectangle), insert your protein name (blue rectangle), and upload a FASTA file containing the DNA sequence of the gene you would like to express (green rectangle). You should be able to download FASTA files of your sequences from the NCBI database (see question 2 in FAQ for more details).

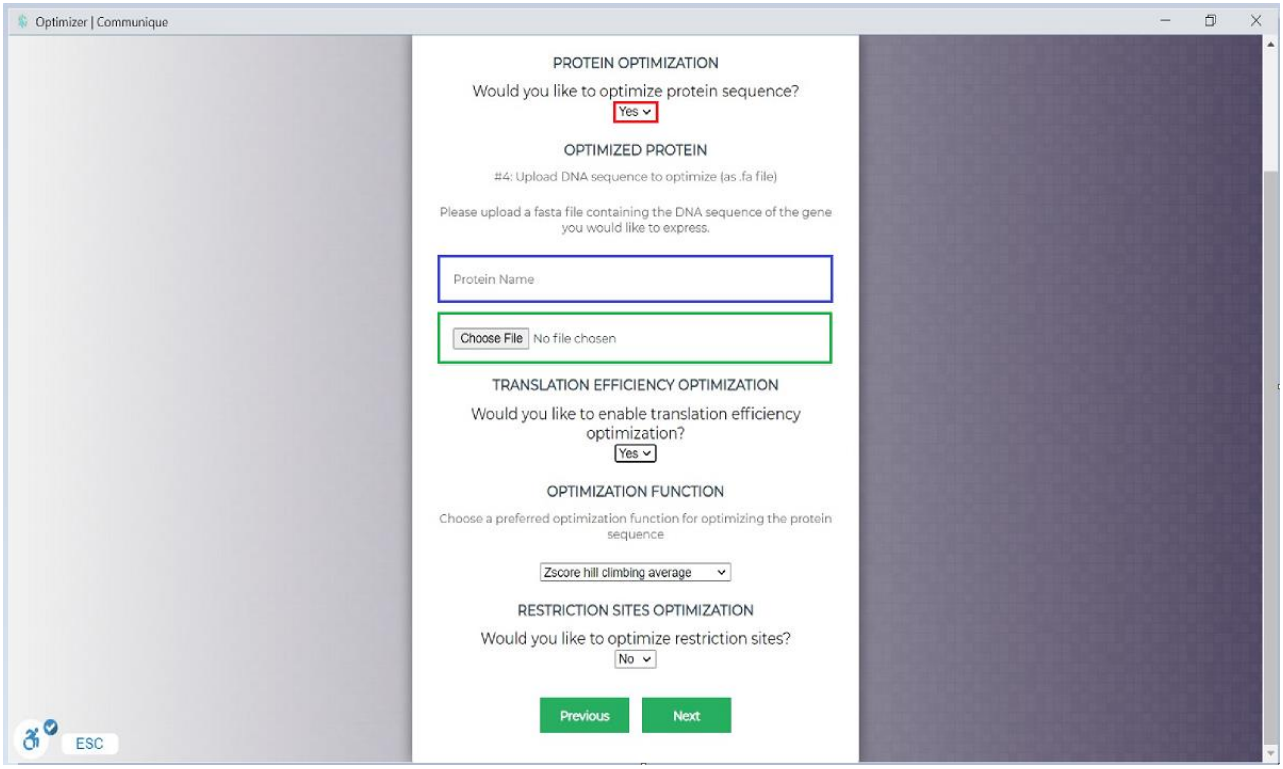


Fig. 10: Insert protein name and upload FASTA file on Protein Optimization screen.

The FASTA formatted file should contain the gene name (or other identifier) as the header, followed by the coding sequence, as shown in Figure 11, and must contain a single entry (sequence), as *Communique* can optimize only one protein at a time.

```
>Coding_sequence
ATGGTTTCCAAGGGCGAGGAGGATAACATGGCTATCATTAAAGAGTTCATGCGCTTCAAAGTT
CACATGGAGGGTTCTGTTAACGGTCACGAGTTCGAGATCGAAGGCGAAGGCGAGGGCCGTCCG
TATGAAGGCACCCAGACCGCCAAACTGAAAGTGACTAAAGGCGGCCCGCTGCCTTTTGCGTGG
GACATCCTGAGCCCGCAATTTATGTACGGTTCTAAAGCGTATGTTAAACACCCAGCGGATATC
CCGGACTATCTGAAGCTGTCTTTTCCGGAAGGTTTCAAGTGGGAACGCGTAATGAATTTTGAA
GATGGTGGTGTCGTGACCGTCACTCAGGACTCCTCCCTGCAGGATGGCGAGTTCATCTATAAA
GTTAAACTGCGTGGTACTAATTTTCCATCTGATGGCCCGGTGATGCAGAAAAAGACGATGGGT
TGGGAGGCGTCTAGCGAACGCATGTATCCGGAAGATGGTGCGCTGAAAGGCGAAATTAAACAG
CGCCTGAAACTGAAAGATGGCGGCCATTATGACGCTGAAGTGAAAACCACGTACAAAGCCAAG
AAACCTGTGCAGCTGCCTGGCGCGTACAATGTGAATATTAAACTGGACATCACCTCTCATAAT
GAAGATTATACGATCGTAGAGCAATATGAGCGCGCGGAGGGTCGTCATTCTACCGGTGGCATG
GATGAGCTGTACAAATAA
```

Fig. 11: Example correct .fasta format (using mCherry fluorescencent gene)

If you chose *'yes'*, you may choose at least one of the following optimization options, otherwise this optimization will not be performed:

Translation Efficiency Optimization

During evolution, cellular machinery has adapted to translate certain codons more optimally than others. Different genetic information can be used to infer the profile of preferred codons (also known as the codon usage bias, or CUB) of each organism.

The open reading frame of the optimized gene is re-coded in order to determine the microbiome's codon preferences, thus enhancing the translation efficiency of the optimized organisms while decreasing translation efficiency for deoptimized organisms.

If you would like to optimize translation efficiency, choose 'yes' from the dropdown (Fig. 12, red rectangle), and choose one of the options for optimization function (Fig. 12, blue rectangle).

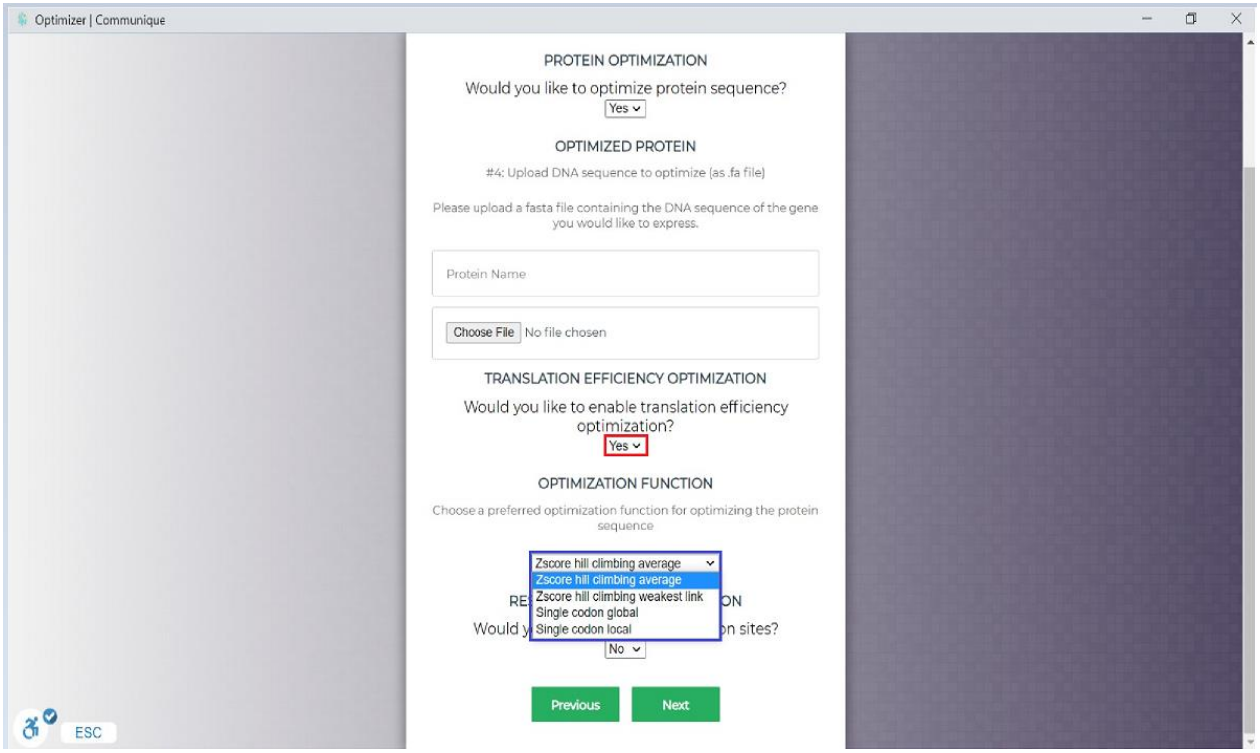


Fig. 12: Translation Efficiency Optimization screen, with option to choose Optimization Function

Restriction Site Optimization

Restriction enzymes are part of the natural immune system of bacteria; they digest foreign DNA by cleaving specific areas determined by the nucleotide sequence. The restriction enzyme footprint of different bacteria is unique, therefore different bacterial organisms contain distinct sets of restriction enzymes. The present model uses the Restriction enzyme database to generate a database of restriction enzymes that are present in the different organisms. Such data is used first and foremost in order to avoid restriction sites of enzymes that are present in the optimized organisms. Moreover, restriction enzymes that are found only in the deoptimized

organisms are examined and the corresponding restriction sites are added to various parts of the gene of interest.

If you would like to optimize restriction sites, choose ‘yes’ in the dropdown (Fig. 13, red rectangle). To continue, press *‘Next’* (Fig. 13, black rectangle).

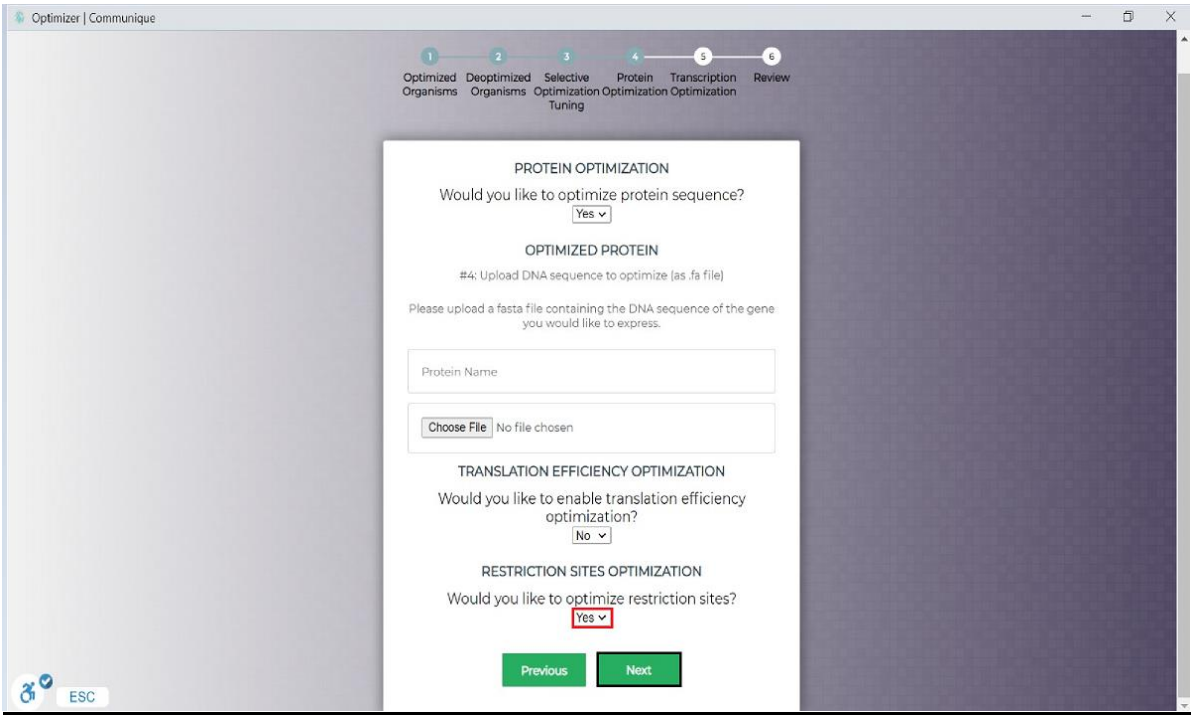


Fig. 13: Choose whether you'd like to optimize Restriction Sites

Step 4: Optimize regulation

Transcription Optimization

Different organisms utilize different transcription factors in order to promote the transcription initiation, as each of these factors recognizes different sets of genomic sequences described as “motifs”.

In this module, the endogenous promoters and intergenic regions of the microbial community are scanned in order to create a set of genetic motifs that promote transcription selectively in the optimized organisms, and not in the deoptimized ones. These motifs are then used to rank optional promoters according to both selectivity and expression abilities, and introduce synthetic changes to them in order to enhance their activity.

If you would like to *skip* transcription optimization, press “Next” (Fig. 14, black rectangle).

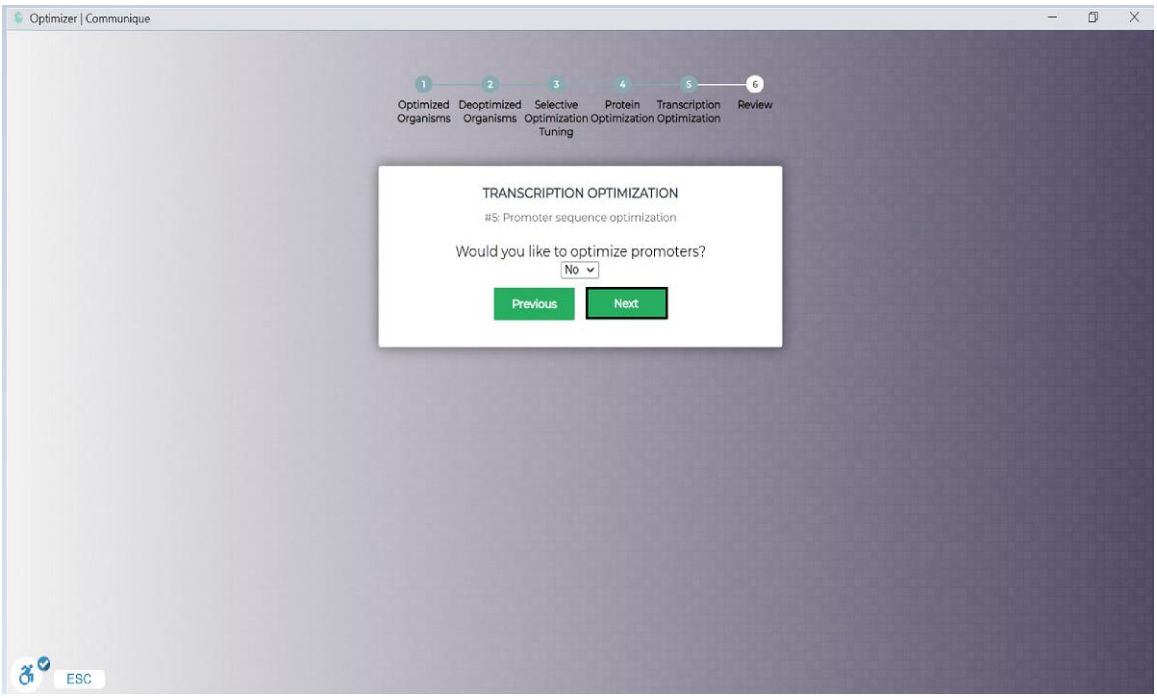


Fig. 14: Transcription optimization screen

Otherwise, choose ‘yes’ from the dropdown (Fig. 15, red rectangle).

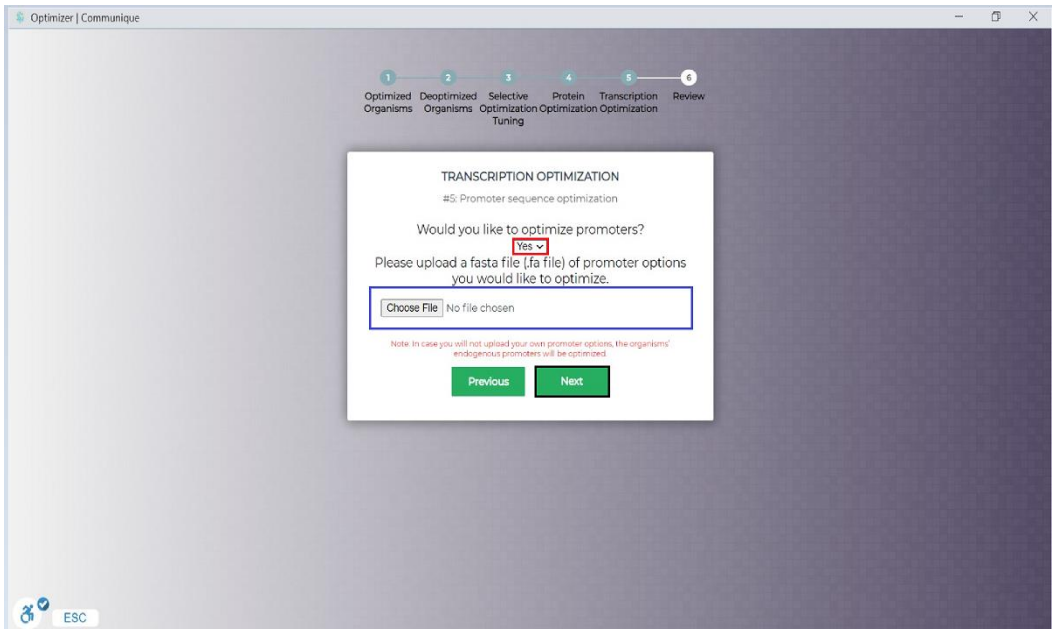


Fig. 15: Transcription Optimization screen with "yes" chosen

Advanced option: you can optimize promoters from an external list (also suitable in case of synthetic promoters), instead of the default optimization which is performed on the organisms' endogenous promoters. In this case, you are required to upload a FASTA file (Fig. 15, blue rectangle) containing promoters names (or another identifier) as the headers, followed by their sequence, as shown in Figure 16:

```
>tufa
CCGGGCACCATCAAGCCGCACACCAAGTTCGAATCTGAAGTGTACATTCTGTCCAAAGATGAAGGCGG
>rpss
AATAATTTTAGAGGATAAGCCATGCCACGTTCTCTCAAGAAAGGTCCTTTTATTGACCTGCACTTGCT
>ompa
CTGGATCCGAAAGACGGTTCCGTAGTTGTTCTGGGTTACACCGACCGCATCGGTTCTGACGCTTACAA
>ompf
AATAAATTTACAAACACCAGCGGCTTCGCCAACAAAACGCAAGACGTTCTGTTAGTTGCGCAATACCA
>rple
CAGGGCTATCCGATCGGCTGTAAAGTAACTCTGCGTGGCGAACGCATGTGGGAGTTCTTTGAGCGCCT
>rplv
TCTTCTGCTCAGAAGGTTCGCCTTGTTGCTGACCTGATTCGCGGTAAGAAAGTGTCGCAGGCTCTGGA
>ompt
AAATACAGCGGCTGGGTGGAATCATCTGATAACGATGAACACTATGACCCGGGAAAAAGAATCACTTA
>pyrb
CACAGCTCTATTGAAGAAGTGATGGCGGAAGTAGACATCCTGTACATGACCCGCGTGCAAAAAGAGCG
>rplw
ATTCGTGAAGAACGTCTGCTGAAGGTGCTGCGTGCACCGCACGTTTCTGAAAAAGCGTCTACTGCGAT
>rplm
TACACTCCGCACGTAGATACCGGTGATTACATCATCGTTCTGAACGCTGACAAAGTTGCTGTAACCGG
```

Fig. 16: Example of an appropriate FASTA file for external promoters

This FASTA file should contain optional promoters fit for your desired expression pattern (multiple entries/sequences), from which a final promoter will be selected.

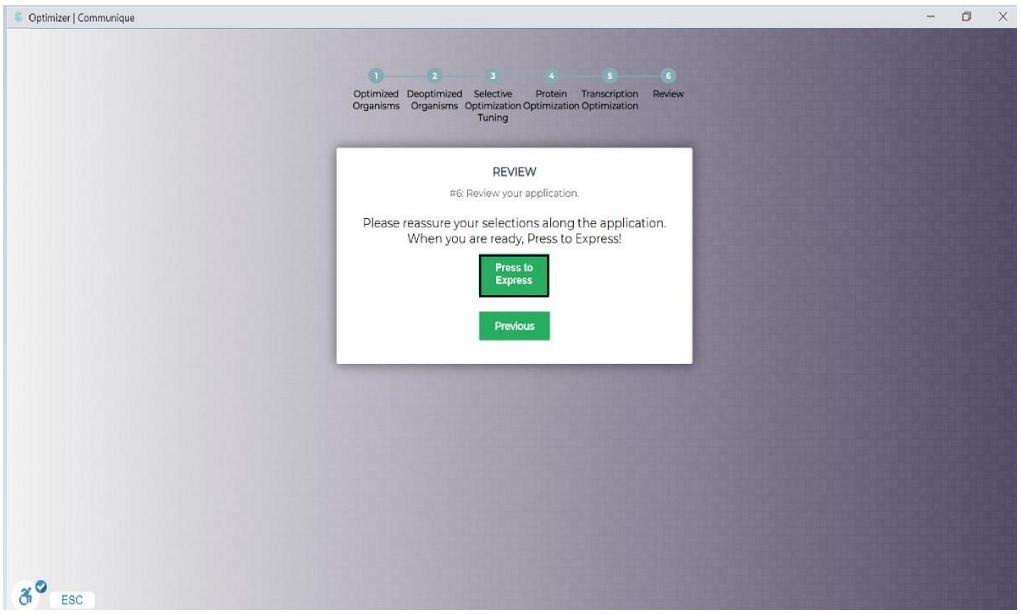
To continue, press *'Next'* (Fig. 15, black rectangle).

Step 5: review and results

Review

If you don't want to change anything, you're ready to "Press to Express" (black rectangle) (Fig. 17)

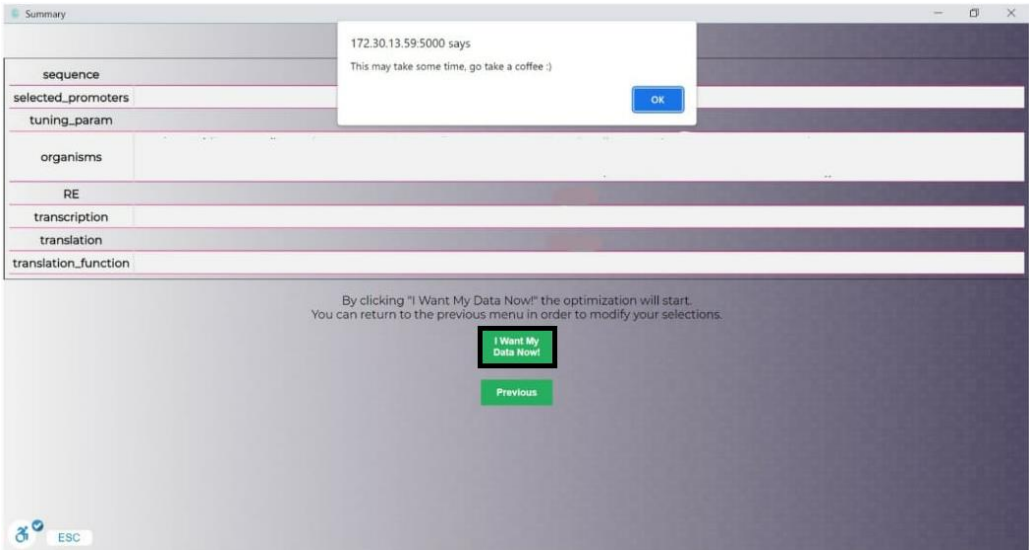
(**Figure 17** - Click the "Press to Express" in order to complete optimization).



Summary of Application

Make sure you inserted the correct data and did not forget anything, then press **"I Want My Data Now!"** (black rectangle) (Fig. 18), and take a break until your results will be ready.

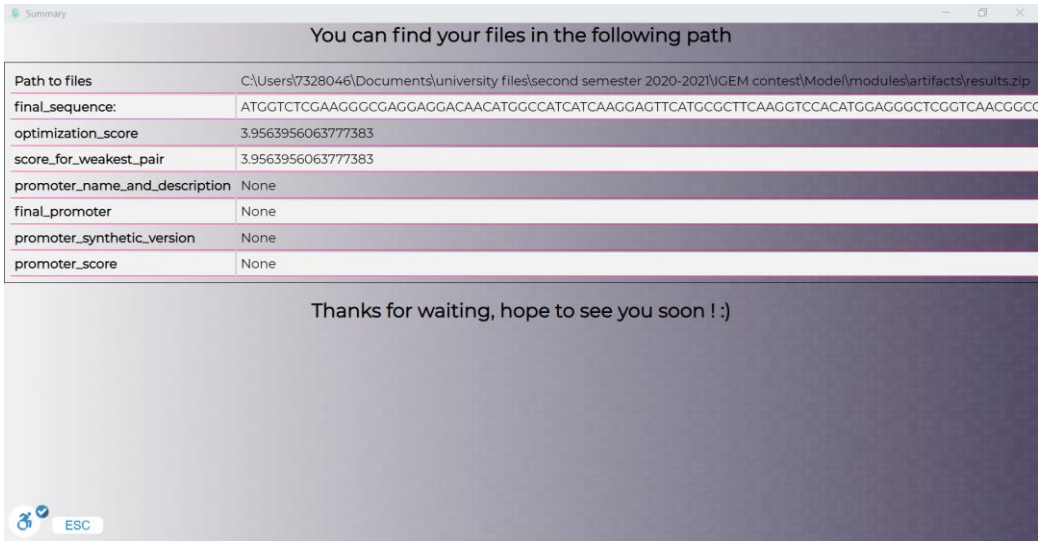
(**Figure 18** - User input summary screen).



Optimization results

The following results will be printed on your screen (Fig. 19):

(**Figure 19** - Optimization results screen).



Output summary

1. Coding sequence optimization (blue rectangle): the optimized DNA sequence of the desired gene and two optimization statistics:
    - Weakest link score
    - Optimization index score
2. Transcription optimization (green rectangle): The most selective promoter (including its description) is returned along with:
    - A synthetic version of it, which is likely to be even more selective
    - The E-value of the promoter
3. Path to the results.zip file (red rectangle): the results zip contains:
    - produced FASTA files: one for the open reading frame, one for the promoters (only if the modules were optimized)
    - Log files: contain all appropriate information regarding the analysis and optimization process for each module separately:
        i. ‘user input.log’: Genomic composition analysis
        ii. ‘coding sequence log’: characterization of the restriction enzymes which recognize the new sequence, the final codon usage bias measurement used for the calculation
        iii. ‘promoters.log’ : information about the discovered motifs sets and their significance.
        iv. ‘output.log’: summary of the output

      Along with additional statistical calculations for all optimization steps.

FAQ

1. How can I download the correct genome for my bacterial species?

Go to the NCBI website. In this example we will download the genome of *E. coli*.

In the search line, choose the genome database on the left side, and write your species on the right side.

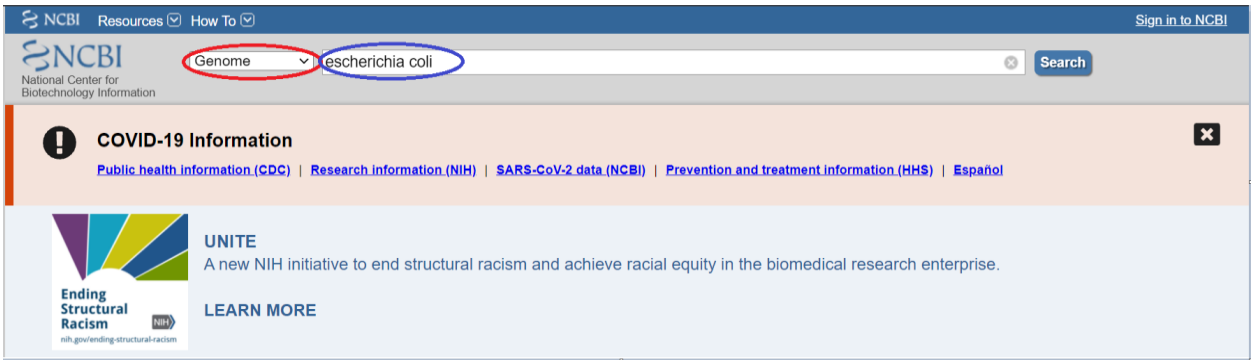
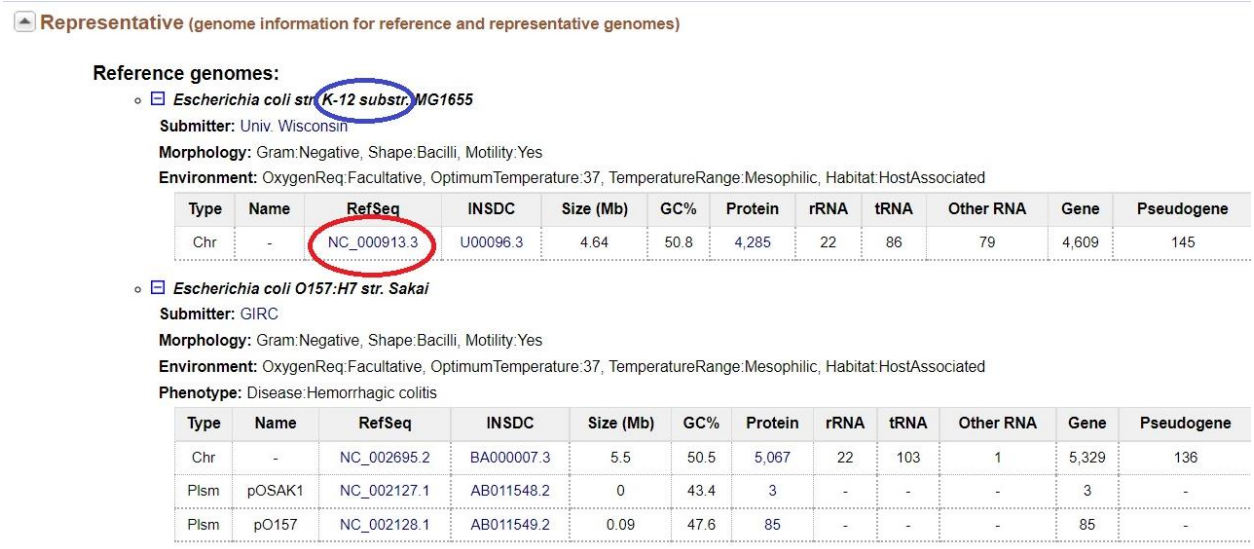


Figure 16: *The search line on the NCBI website. Database option are circled in red (should be 'genome'), the species circled in blue (in our case Escherichia coli)*

Scroll down to the reference genomes table and choose the RefSeq of your substring.



Representative (genome information for reference and representative genomes)

Reference genomes:

- ***Escherichia coli str. K-12 substr. MG1655***

Submitter: Univ. Wisconsin

Morphology: Gram:Negative, Shape:Bacilli, Motility:Yes

Environment: OxygenReq:Facultative, OptimumTemperature:37, TemperatureRange:Mesophilic, Habitat:HostAssociated

Type	Name	RefSeq	INSDC	Size (Mb)	GC%	Protein	rRNA	tRNA	Other RNA	Gene	Pseudogene
Chr	-	NC_000913.3	U00096.3	4.64	50.8	4,285	22	86	79	4,609	145

- ***Escherichia coli O157:H7 str. Sakai***

Submitter: GIRC

Morphology: Gram:Negative, Shape:Bacilli, Motility:Yes

Environment: OxygenReq:Facultative, OptimumTemperature:37, TemperatureRange:Mesophilic, Habitat:HostAssociated

Phenotype: Disease:Hemorrhagic colitis

Type	Name	RefSeq	INSDC	Size (Mb)	GC%	Protein	rRNA	tRNA	Other RNA	Gene	Pseudogene
Chr	-	NC_002695.2	BA000007.3	5.5	50.5	5,067	22	103	1	5,329	136
Plsm	pOSAK1	NC_002127.1	AB011548.2	0	43.4	3	-	-	-	3	-
Plsm	pO157	NC_002128.1	AB011549.2	0.09	47.6	85	-	-	-	85	-

Figure 17: *the reference genomes table of E. coli. The RefSeq of the substring (which in our case K-12, circled in blue) circled in red*

In the page that is opened, press on “send to” button and then choose “Complete Record”, “File” and “GenBank (full)” (See figure below). Note! You have to choose the full option in order to get all the data. If you wish to download the data in FASTA format, choose FASTA instead of GenBank (full).

Finally, click on create file and download your data.

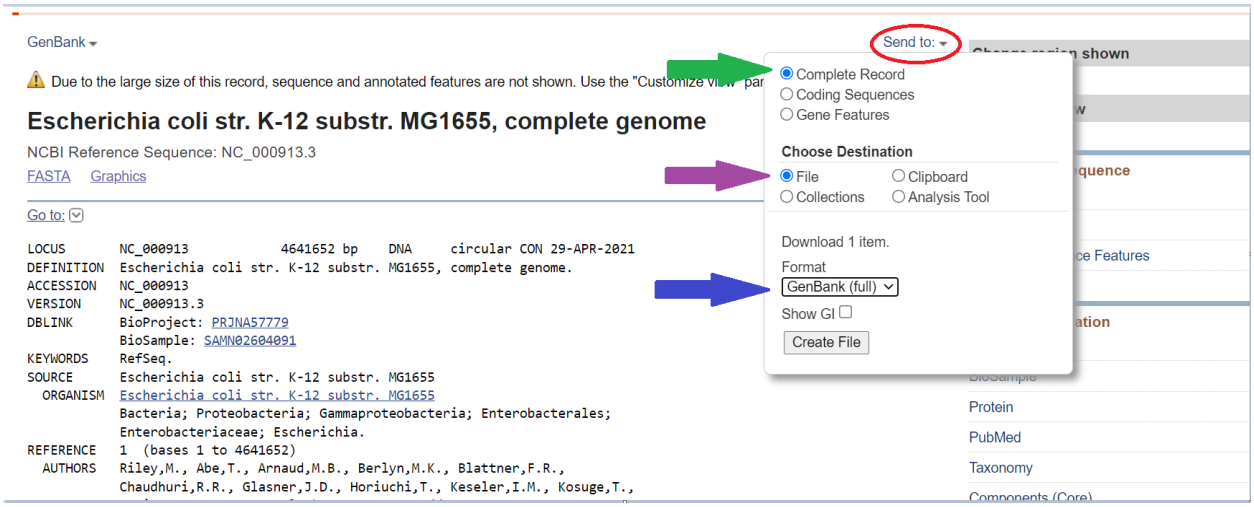


Figure 18: *Send to button circled in red, complete record option marked with a green arrow, File option marked with a purple arrow and GenBank (full) option marked with a blue arrow.*

2. What is the difference between the translation optimization functions?

The optimization strategies can be divided into 2 main approaches:

- Single codon optimization: optimize the codons based only on the profiles themselves, without respect to the initial sequence or the translation efficiency scores that the endogenous proteins receive. Local optimization refers includes normalization according to the synonymous scores at each position, and global optimization includes profile normalization according to scores from all codons.
- Whole proteome optimization: this optimization takes into account the significance of each change to the entire protein in to context of each individual organism- i.e., if an organism tends to have proteins with very similar codon usage bias scores, even a small improvement might be significant. This causes dependencies between the different codons and is solved using hill climbing to optimize the weakest link or the average score.

For small microbiomes, the weakest link optimization is recommended, and for larger ones the average hill climbing optimization tends to perform better.

3. How do I interpret my results

Protein Optimization Scores:

- Weakest Link Score - In this optimization, any positive score indicates that the margins of the two groups are separated, so that even the organism that was least able to be optimized is still better than all of the deoptimized organisms. Moreover, when looking at the clusters of the two organisms - this score indicates the two clusters are fully separated. Larger values indicate better separation between the two groups, values larger than one are generally regarded as a good result.

- Average Difference Score - This index was calculated for different groups: for the whole community, for the optimized organisms, and for the deoptimized organisms.

  For the whole community, a positive score means the optimization worked well (higher/lower CUB values with respect to optimized/deoptimized organisms correspondingly, in comparison to the original values and considering the value of the tuning parameter). Here, larger values indicate a better optimization process.

  For the optimized / deoptimized organisms, the higher the absolute value of the index, the more selective the optimization.

Transcription Optimization Score:

- E-value: short for Expect Value, the E-value is one of the most common indices used for the significance of biological sequence alignments in the context of a database. It represents how likely it is to observe the same results (or better) by chance, or alternatively: how many alignments with the same score (or better) can be expected in a random database of the same size. This means that for a large enough database, the smaller the E-value, the more significant and trustworthy the alignment. It is worth mentioning that the E-value is not a probability (does not range between 0 and 1), and is only bounded from below by 0.

4. Does the number of organisms I select affect my results?

The factor that determines the quality of the optimization is not the direct size of the microbiome, but the genetic variation between the optimized and deoptimized group- if the optimized and deoptimized groups are evolutionarily (and functionally) different, the performance remains nearly the same for various group sizes.

5. **How exactly does the tuning parameter control the tradeoff between optimization and deoptimization?**

The tuning parameter effects two optimization modules- transcription and translation.

In the transcription optimization, the selective motif set is built from two types of motifs- motifs related to selectivity and motifs related to optimal transcription. The tuning parameter controls the thresholds effecting the entrance of these two types into the final motif set.

However, the effect on translation efficiency depends on the translation optimization function- in the single codon approach, it is used to weigh the different profiles in selection of the most optimal codon for each position. In the whole microbiome optimizations, the tuning parameter effects the optimization through the statistical measure that is optimized.

Copyright Notice

The Communique software was created by the 2021 iGEM team from Tel Aviv University.

The software and its code are the intellectual property of the iGEM TAU group and are to be used for non-commercial purposes only.

Copyright notice for MEME Suite:

MEME Suite

Authors: Timothy L. Bailey, William Noble

Permission to use, copy, modify, and distribute any part of this software for educational, research and non-profit purposes, without fee, and without a written agreement is hereby granted, provided that the above copyright notice, this paragraph and the following three paragraphs appear in all copies.