

Chlamy Sequence Optimizer (CSO)

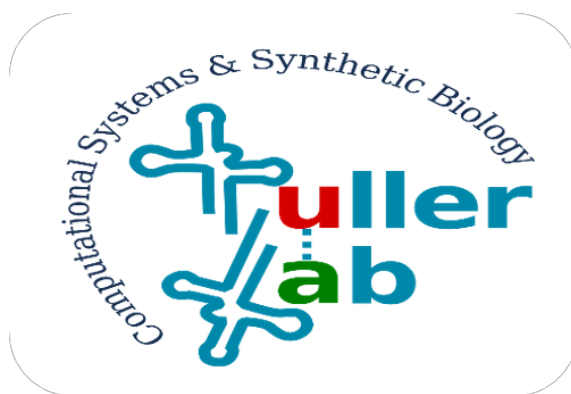
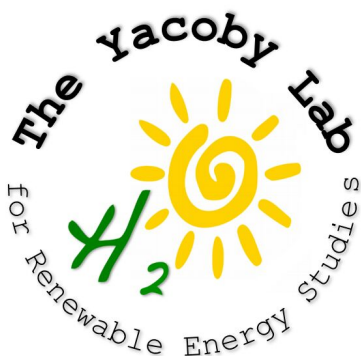
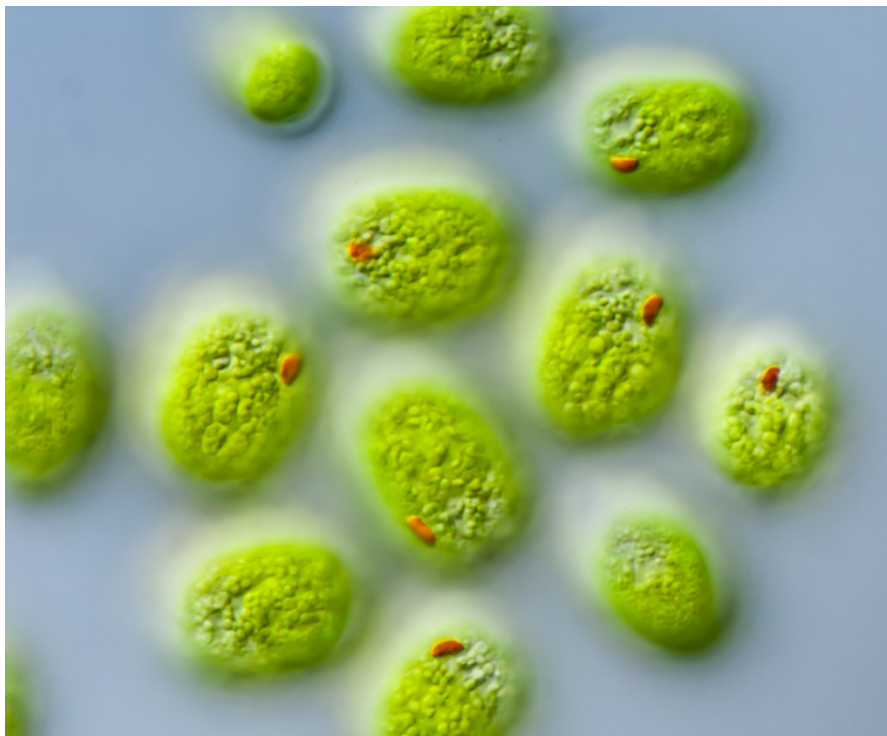




Table of Contents

1. Introduction	3
2. Pipeline	3
3. GUI	4
i. Installation	4
ii. Input.....	5
iii. Process and output	8
4. Analyzing the output	8
Acknowledgments	9



1. Introduction

Chlamy Sequence Optimizer (CSO) is an application built to allow a simple one-stop-shop for designing synthetic DNA coding sequences (CDSs) for expression in the model microalga *Chlamydomonas reinhardtii* (a customizable option for other organisms is available as well). CSO includes algorithms for both nuclear expression and chloroplast expression, and its performance and logic is demonstrated in the following publications:

- 1. Enhancing heterologous expression in *Chlamydomonas reinhardtii* by transcript sequence optimization, Weiner et al., The Plant Journal, 2018. (Nuclear expression)**
- 2. CSO – a sequence optimization software for engineering chloroplast expression in *Chlamydomonas reinhardtii*, Weiner et al., under review (Chloroplast expression)**

This software contains a graphical user interface (GUI) for sequence optimization, which can be downloaded freely from GitHub:

<https://github.com/iddoweiner/Coding-Sequence-optimization-for-Chlamydomonas-reinhardtii>

2. Pipeline

The optimization is composed of three main steps:

- A. Verifying the validity of the input
- B. Executing the selected optimization algorithm
- C. Generating the output

For full pipeline and algorithm description please refer to the papers mentioned above.

3. GUI

i. Installation

The GUI is available for both Windows (recommended) and macOS.

For Windows: download and run the **MyAppInstaller_web.exe**

For Mac: download the folder **MyAppInstaller_web.app/Contents**, move it to the applications folder and run it.

* Important Note: While downloading and / or running these files, you may come across security alerts protecting against installation of applications downloaded from the web. In this case you can modify your security preferences to accept this software. You can be assured it is safe 😊

In order for the software to work, MATLAB runtime – which can be downloaded freely – is required. If you do not already have MATLAB runtime, the installation pipeline will ask permission to **automatically download it from the web (recommended)**. Alternatively, you can download it yourself at:

<https://www.mathworks.com/products/compiler/matlab-runtime.html>

Once installed, windows users can simply open the application manually and start using it.

Mac users will need to open the application from the terminal in order to include the MATLAB runtime package in their path:

- Open the terminal
- Type the following in a continuous line:
 - Full path to the shell script: **run_Chlamy_sequence_optimizer.sh**,
 - one space,
 - Full path to the directory in which MATLAB runtime is installed

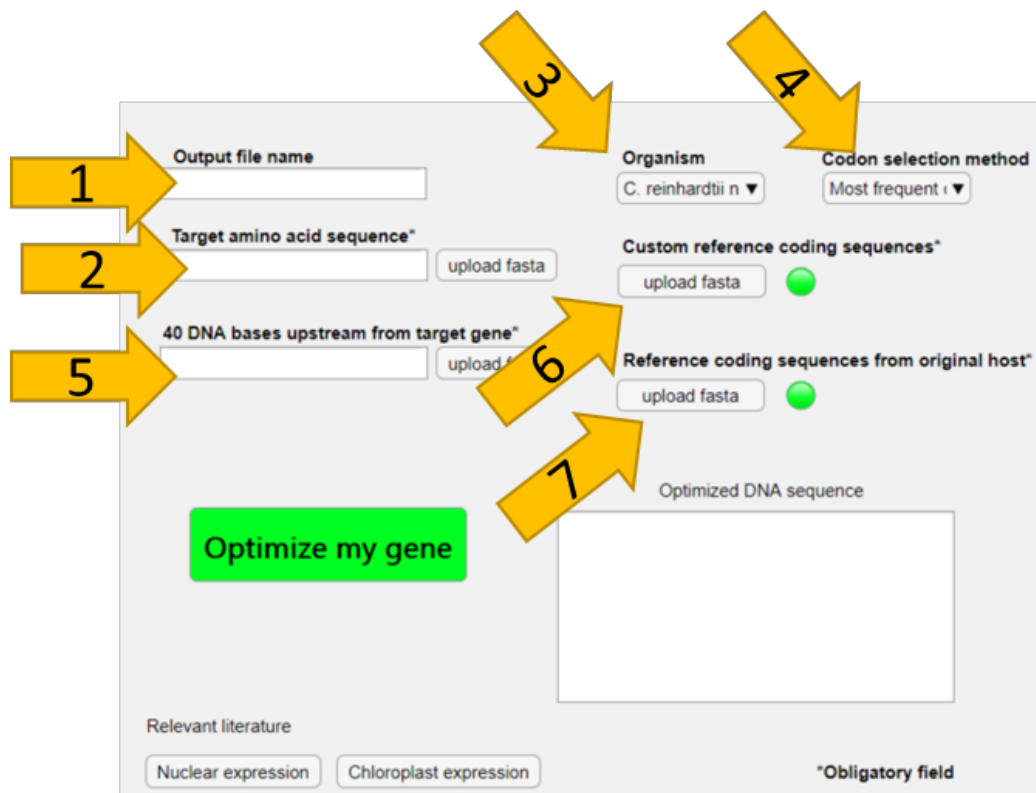
For example:

```
~/Desktop/Applications/Tel_Aviv_University/Chlamy_sequence_optimizer/application/run_Chla  
my_sequence_optimizer.sh ~/Desktop/Applications/MATLAB/MATLAB_Runtime/v93
```

Executing this script will open the GUI with full MATLAB runtime permissions (note that simply opening the application manually might not work because MATLAB runtime will not be available).

ii. Input

CSO requires user input; the exact input required varies depending on the optimization algorithm used. In the figure provided below, all inputs were numbered 1-7 to facilitate explanations.



The screenshot shows the Chlamy Sequence Optimizer GUI. Seven yellow arrows with numbers 1 through 7 point to specific input fields:

- 1: Output file name
- 2: Target amino acid sequence*
- 3: Organism (dropdown menu showing C. reinhardtii n)
- 4: Codon selection method (dropdown menu showing Most frequent)
- 5: 40 DNA bases upstream from target gene*
- 6: Reference coding sequences from original host* (upload fasta button)
- 7: Optimized DNA sequence (output area)

Other visible elements include:

- upload fasta button for Target amino acid sequence*
- Custom reference coding sequences* (upload fasta button and green status indicator)
- Relevant literature (Nuclear expression, Chloroplast expression buttons)
- *Obligatory field

(1) Output file name

CSO creates an output FASTA file with your optimized DNA sequence, together with some other information about the optimization. In this field you can enter the desired name of this file (*e.g.* My optimized gene). There is no need to write the file extension.



*This file will be created by default in the directory where the app is running. If you wish to change this, you can give your file a name specifying the full path to where you want it to be saved (e.g. C:\Users\user\Desktop\My optimized gene).

- If a file with the name you specified already exists, the new data will be appended to it.
- If you leave this field empty, the default name of the output file will be 'optimization_output'.

(2) Target amino acid sequence – obligatory field in all cases

This is your gene of interest, for which the sequence optimization will be carried out.

- Use 1-letter amino acid symbols (<http://www.fao.org/docrep/004/Y2775E/y2775e0e.htm>).
- Include the START (typically 'M') and STOP ('*') codons.
- You can either upload a FASTA file containing your sequence, or type / paste it in the field.
- Both lower and upper case letters are acceptable.
- The optimization will not begin if this field contains invalid AA characters, or if it is left empty.

(3) Organism – the default selection is *C. reinhardtii* nucleus

Select your required host. The two basic options are either ***C. reinhardtii* nucleus** (def) or ***C. reinhardtii* chloroplast**. Selecting between these two options will let CSO know which set of reference sequences to load and use. The third option is **custom**; selecting this option means that you want to use the CSO algorithms to design a synthetic gene for expression in a different host (e.g. a different alga/plant, bacteria, mammalian cells, etc.). If you select this option, input field (6) will appear and you will be required to provide a set of coding sequences from your host of choice that will be used as the algorithms' reference sequences. We would like to note that this use-case will give you the general versions of the algorithms, without the *C. reinhardtii* specific optimization embedded into the two other use-cases. Still, performing sequence optimization in this manner is expected to yield decent results on average.

(4) Codon selection method – the default selection is *most frequent codons*

Select your algorithm of choice. The default option (most frequent codons) will simply give you the same codon (the “best”, or most frequent one) per each amino acid. This will result in a monotone (or “boring”)



sequence with a codon adaptation index value of exactly 1, given the CSO reference set. The other options are:

A. *Optimal codons and folding*

B. *ChimerMap*

C. *Mimic original host*

For algorithm descriptions see the papers referenced from the buttons on the bottom left corner (relevant literature).

(5) 40 bases upstream from target gene – obligatory field only if it appears, otherwise it is hidden

This can be considered as the 5' UTR region of your target gene. The sequence in this region is taken into consideration while selecting codons for your target gene.

- You can either upload a FASTA file containing your sequence or type / paste it in the field.
- Both lower and upper case letters are acceptable.
- If you provide more than 40 nucleotides, only the last 40 will be considered.
- The optimization will not begin if this field contains invalid DNA characters, or if the sequence provided is shorter than 40.

(6) Custom reference coding sequences – obligatory if 'custom' organism was chosen

Upload a FASTA file of coding sequences taken from the organism in which the user wishes to express his gene.

(7) Reference coding sequences from original host – obligatory if 'mimic' algorithm was chosen

Upload a FASTA file of coding sequences taken from the original organism that your gene of interest comes from.

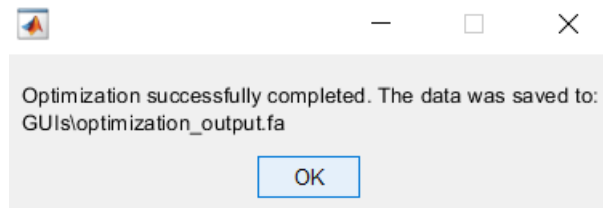
NOTE – in order for the *Mimic original host* algorithm to work, one of these CDSs must code for the AA sequence you have specified in field #2.

iii. Process and output

Once the input fields are properly filled, hit the green 'Optimize my gene' button. A wait-bar will appear and show you the progress of your run. Besides the output file (see section 4) which will be created at the end of your run, you will also be able to see your optimized DNA sequence on the GUI screen.

4. Analyzing the output

At the end of each run, CSO creates an output file containing important information regarding your optimization. If you have not specified a name for the output file, it will be given a default name. In any case, this file will be saved to the current working file where the software is located. At the end of the optimization, a message specifying the filename and its directory will appear:



The output file is saved in FASTA format. It can easily be read by any text editor, such as Notepad++:

<https://notepad-plus-plus.org/>

Besides the optimized DNA sequence, the output file also contains other information regarding your optimization (the total length of the gene, its GC content (0-1), the mean mRNA folding energy of the final sequence in the START area [if relevant], your specified 40 NT upstream from the target gene [if relevant], and the optimization time).



Acknowledgments

We would like to thank Mr. Gerd Guenther, who took the fascinating *Chlamydomonas* picture used as a cover picture for this application.

His work is available at: https://www.allposters.com/-st/Gerd-Guenther-Posters_c180530_.htm

Written by Iddo Weiner, Ido.nadav@gmail.com