Computational and Structural Biotechnology Journal 20 (2022) 2521–2538

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Modeling the effect of rRNA-mRNA interactions and mRNA folding on mRNA translation in chloroplasts

Stav Carmel Ezra, Tamir Tuller *

Department of Biomedical Engineering, Tel Aviv University, Israel

ARTICLE INFO

Article history:
Received 17 January 2022
Received in revised form 15 May 2022
Accepted 15 May 2022
Available online 18 May 2022

Keywords:
Chloroplasts
mRNA translation
Chloroplasts evolution
mRNA folding
rRNA-mRNA interactions
Gene expression optimization
Functional mRNA structures

ABSTRACT

The process of translation initiation in prokaryotes is mediated by the hybridization of the 16S rRNA of the small ribosomal subunit with the mRNA in a short region called the ribosomal binding site. However, translation initiation in chloroplasts, which have evolved from an ancestral bacterium, is not well understood. Some studies suggest that in many cases it differs from translation initiation in bacteria and involves various novel interactions of the mRNA structures with intracellular factors; however currently, there is no generic quantitative model related to these aspects in chloroplasts.

We developed a novel computational pipeline and models that can be used for understanding and modeling translation regulation in chloroplasts. We demonstrate that local folding and co-folding energy of the rRNA and the mRNA correlates with codon usage estimators of expression levels (r = –0.63) and infer predictive models that connect these energies and codon usage to protein levels (with correlation up to 0.71). In addition, we demonstrate that the ends of the transcripts in chloroplasts are populated with various structural elements that may be functional. Furthermore, we report a database of 166 novel structures in the chloroplast transcripts that are predicted to be functional.

We believe that the models reported here improve existing understandings of genomic evolution and the biophysics of translation in chloroplasts; as such, they can aid gene expression engineering in chloroplasts for various biotechnological objectives.

1. Introduction

Chloroplasts are intracellular organelles in plants that are responsible for photosynthesis and carbon fixation; they were generated by an endosymbiosis process in which a eubacterium was engulfed by a common eukaryotic ancestor [1-4]. Chloroplasts contain their own genetic system with a circular double-stranded DNA molecule [5,6]. Some endosymbiont genes were lost, while others were transferred into the host genome as a result of coevolution between the host and the endosymbiont; this led to a significantly smaller chloroplast genome with asize of approximately 120–160 kb and containing and average of 120 genes, most of which encode essential components of the photosynthesis machinery and are therefore essential for chloroplast viability. The size, structure, and genetic content of chloroplast genomes of land plants appear to be relatively conserved. It is reported that ~80% of the genes present in chloroplast genomes of land plants and the most ancient algae, which is a green algae species (*Mesostigma viride*), are shared; this indicates that both gene content and gene order - are generally conserved in chloroplast genomes throughout evolution [7,8]. The majority of chloroplast translation studies have been carried out on land plants ('green' phylogenetic lineages, e.g., tobacco, maize, spinach, and barley), as well as on chlorophytes (green algae, e.g., *Chlamydomonas reinhardtii*), and on *Euglena gracilis* which is not a chlorophyte but a member of the euglenoid algae. However, little is known about the mechanisms of chloroplast translation that developed in other algal lineages due to evolution and diversification over the past 1–2 billion years.

The chloroplast's translational machinery is most closely related to that of eubacteria, but there are some similarities with the nuclear-cytosolic system of eukaryotes [4]. A highly similar composition of the translation machinery between chloroplasts and bacteria indicates the bacterial origin of the chloroplast gene expression mechanism. Chloroplast translation is performed by prokaryotic-type 70S ribosomes, which consist of a small 30S and a large 50S subunits composed of orthologs of *Escherichia coli's* (*E. coli*) reference ribosome rRNAs and proteins [9]. Over the years, the similarity between the translation initiation of prokaryotes and chloroplasts was questioned, and the search for differences

* Corresponding author.
E-mail address: tamirtul@tauex.tau.ac.il (T. Tuller).

https://doi.org/10.1016/j.csbj.2022.05.030
2001-0370/

between the two mechanisms' features has attracted attention. In all systems studied to date, the translation initiation in chloroplasts starts when a complex consisting of the 30S subunit of the ribosome and the initiator tRNA (N formylmethionine), called the preinitiation complex, binds the initiation site in the mRNA [9]. Three protein initiation factors (IF), IF1, IF2, and IF3, which were found as an ortholog of the bacterial IFs, control and accurate initiation process steps [10].

The translation initiation model in prokaryotes can be described by the Shine–Dalgarno (SD) mechanism. According to this model, the 30S subunit of the ribosome binds the mRNA through base-pairing between the SD sequence (another name for the ribosomal binding site) of the mRNA, which is located upstream of the start codon, and the anti-SD (aSD), a conserved sequence found at the 3′-edge of the 16S rRNA of the small subunit of the ribosome [11]. The role of the SD motif in chloroplasts has raised questions and was a source of great research. According to research done in this field to answer these questions, it was found out that there does not seem to be any obvious signal indicating conserved sequences in the SD position in chloroplast mRNA. However, the 16S aSD sequence is highly conserved across chloroplasts [12]. It was discovered that 38% (30 out of 79) of tobacco chloroplasts genes contain no SD-like sequences within 20 nucleotides (nt) upstream from the start codon, and 14% of the genes have the SD-like sequence but not in the expected positions of −18 to −16 upstream of the start codon, therefore there are only 48% genes with SD-like sequence in the expected positions at the 5′ UTRs of their mRNAs [13]. Several studies tried to investigate the role of the SD-aSD interaction in *C. reinhardtii*, *Euglena*, and tobacco chloroplast genes by site-directed replacement mutations or deletion of the SD-like sequences and revealed that some genes require SD-like sequence for translation while others do not. More specifically, in some genes, the altered SD positions had little or no effect on their translation in vivo, genes such as *petD* [14], *atpB*, *atpE*, *rps4*, *rps7* [15] (*C. reinhardtii*), and *rbcL* (*Euglena*) [16], while other genes *rpl2* and *rpls16* (*tobacco*) do not even have such sequences [13]. On the other hand, the altered SD positions had a negative effect on the translation of genes such as *psbA* [17], *psbD, psbC* (*C. reinhardtii*) [18], *atpH* (*Euglena*) [19], *rps14 (tobacco*) [20], indicating that this SD-like element has a positive role in their translation initiation. It has been suggested that the requirement for SD-like sequences may be more important for the translation of highly expressed mRNAs in *C. reinhardtii* [4,18]. The lack of an absolute requirement for the SD-like sequences of several mRNAs indicates that other *cis*-acting elements recruit ribosomes to the start codon position.

Previous research has also shown that chloroplast ribosomal RNA and ribosomal proteins differ from those of *E. coli*; the ratio between ribosomal proteins and rRNA significantly shifted during evolution, favoring ribosomal proteins, which led to modifications in the rRNA domains. As a result, ribosomal proteins in chloroplasts interact differently with rRNAs or other ribosomal proteins and perform structural changes that compensate for altered rRNA domains. These ribosomal changes may result in new contact sites with the mRNA molecule and therefore are hypothesized to affect translational regulation [12,21,22]. Additionally, it was discovered that point mutations are leading to changes in the local structure of ribosomal RNA in chloroplasts; these mutations create significantly folded structures in the positions of the a-SD, which reduce the probability of ribosomes binding to the mRNA [12].

It was discovered that there are protein factors (*trans*-acting factors) that mediate translation initiation by interacting with the mRNA sequence or secondary structures at the 5′UTR located in cis, typically upstream of the reading frame in the mRNA. These factors are gene-specific, and were discovered for specific chloroplast genomes, tables including information regarding some of these protein factors can be found [9,23,24]. Some specific *cis*-elements controlling the translation of individual genes in a particular chloroplast genome were studied and discovered [9]; however, the molecular function of RNA *cis*-elements and proteinaceous *trans*-factors in the regulation of chloroplast translation, in general, is mostly unknown. In addition to primary sequence elements, features of mRNA 2D or 3D structures (or lack of structure) can represent *cis*-elements that influence the translation process [47]; there are some chloroplast genes, for example, with mRNA molecules that have a secondary structure that reveals the ribosomal binding site or the start codon that triggers the translation initiation [26].

Today, there are various pieces of evidence regarding the nature of translation initiation in chloroplasts; however, there is no unified model that can predict translation initiation in chloroplasts. The current study aims to develop such a generic quantitative model.

2. Results

The flow diagram of the study is described in Fig. 1 (all details are in the Materials and Methods section). The analyzed database includes the genome of 4,306 chloroplasts from various species; each chloroplast includes 74 genes on average, so that we analyzed 318,315 genes in total (Fig. 1A). The study is divided into two main parts: the first one includes the development of an energy-based model for translation initiation prediction, and the second part includes the development of a large-scale database with predicted local functional mRNA structure.

The genes in the database were divided into orthologous groups to find the translation regulation characteristics for different gene products. The energy-based model was inferred by using the CAI as a proxy for expression levels (see Materials and Methods section 5.4); for every gene in the database, the CAI was computed and was normalized to be comparable to the CAI of genes from other chloroplasts (Fig. 1B). The energy values in the model were based on the local minimum free energy of a sequence. The local rRNA-mRNA hybridizations (Fig. 1C) and local mRNA folding (Fig. 1D) were calculated for every gene in the database. By these three steps (A-C), an Energy-based Translation Initiation Predictor (*ETIP*) was conducted (Fig. 1H). The local mRNA folding was computed for every gene in every ortholog group (Fig. 1D) and compared to the local mRNA folding generated based on a null model (Fig. 1F). Selection for strong mRNA energy was detected (Fig. 1G) in the multiple sequence alignment (MSA) of every ortholog group (Fig. 1E). After inferring the positions in which there is selection for strong energy in the ortholog group of genes, the local common functional mRNA structures were predicted (Fig. 1I). Combining the energy-based gene expression prediction (Fig. 1H) and the local functional mRNA structures (Fig. 1I) provides predictive biophysical mRNA-rRNA interaction models that can, shed light on novel aspects of chloroplasts translation mechanism and regulation.

2.1. Folding and co-folding between the 5′UTR of the mRNA and the 16S small subunit of the ribosome predict chloroplast gene expression levels

This subsection aims to conduct an energy-based model that will be able to evaluate and predict mRNA translation initiation efficiency. It is expected that there will be a high positive correlation between translation efficiency and protein abundance (PA) values. However, there are no measurements of PA for all the genes in the database, therefore we used normalized CAI scores that are known to be highly correlated with PA [27]. Our biophysical model

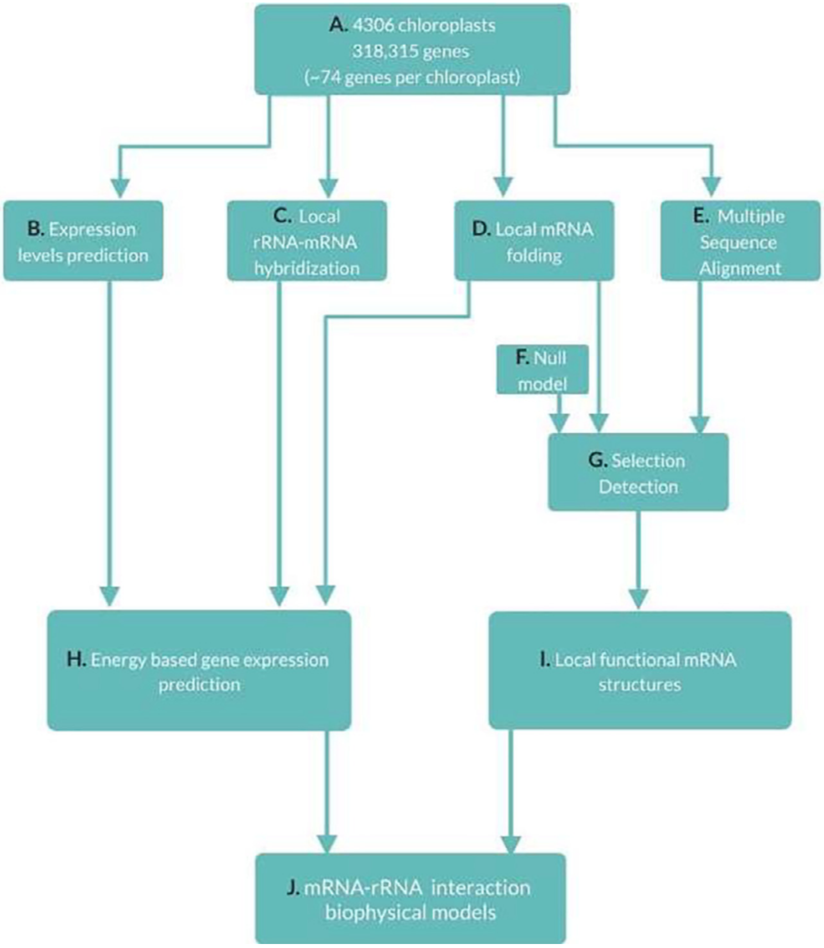

A. 4306 chloroplasts
318,315 genes
(~74 genes per chloroplast)
B. Expression levels prediction
C. Local rRNA-mRNA hybridization
D. Local mRNA folding
E. Multiple Sequence Alignment
F. Null model
G. Selection Detection
H. Energy based gene expression prediction
I. Local functional mRNA structures
J. mRNA-rRNA interaction biophysical models

Fig. 1. The general flow diagram of the study. See details in the main text.

is based on minimum free energy computations of the local mRNA-rRNA hybridization and folding. The model compares the free energy between two states (as can be seen in Fig. 2): 1) before the 16S hybridizes to the mRNA when the mRNA and 16S exhibit self-folding structures; 2) after hybridization when the two sequences bind together and create a new co-folded structure. A greater decrease in the free energy in state 2) compared with state 1) is expected to be related to a more efficient initiation rate as it is related to higher probability of hybridization (for further information, see Materials and Methods sections 5.6 and 5.7).

Fig. 3 includes examples of local self-folding structures of mRNA and 16S rRNA and their co-folded structure, for two different genes of *Abelia sanguinea's* chloroplast.

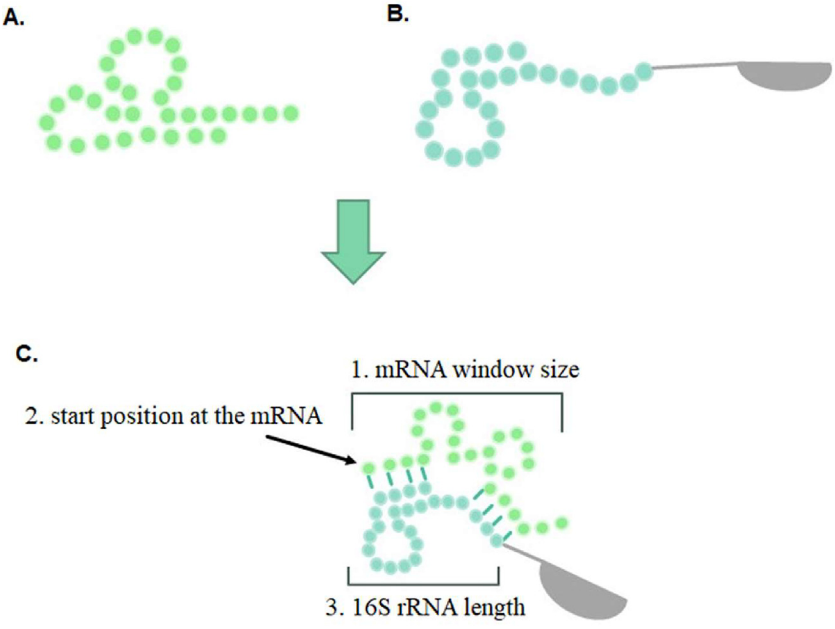
Since translation efficiency is associated with higher protein levels, we expect a negative Spearman correlation between the predictions of our model and PAs, and between the predictions and CAI scores. We aimed to investigate the typical properties of the local structures (four parameters) that comprise the free energy model: 1) the mRNA window length, 2) the position upon the 5′UTR of the mRNA where the local structure starts, 3) the 16S rRNA sequence length from the 3′ edge, and 4) the *ETIP* constant that determines the subtraction of the self-foldings from the co-folding energy (Fig. 2).

The energy model relies on finding the optimized parameters out of a set of values such that the energy values calculated will optimally predict the CAI scores. It is expected that the parameters that optimize the correlation will have meaning in terms of translation mechanisms and will imply or reveal mRNA functional structures and properties of the mRNA – rRNA interactions that correspond to translation regulation and reflect translation efficiency.

According to the literature, transaltion of different genes is generally regulated in different ways; therefore it is expected that different orthologous groups will have different optimal parameters of the energy-based model.

The different stages of the optimization process are described in a flow diagram in Fig. 3, which describes Fig. 1H of the project's global flow diagram in more detail.

As a first step, all the genes in the database were divided into training and test datasets (Fig. 3H1). Then a hill-climbing optimization algorithm was applied on the training dataset to find the optimal parameters that predict the codon usage levels for every ortholog group (Fig. 3H4). By assuming that there is a finite number of regulation strategies in chloroplasts we added constraints to the objective function, such that instead of examining all the values in every parameter set, we took a subset of values of size *X* (we checked different *X* values) that must include all the possible parameters of the model for all the orthologous groups. This approach simplified the model and reduced overfitting. For every *X*-value, optimization was performed (Fig. 3H5) with multiple different initiation points, by randomly selecting a new different subset of values to check (Fig. 3H6). We also compared the optimization with a case in which the parameters weren't limited at all. Every initial point for every *X* reached an optimal correlation with optimal energy parameters for every group and then the correlations were validated with the test set and were compared to



$$minusfold_i = cofold_{mRNA_i\&16S_j} - \overbrace{C}^{\text{4. Constant}} \cdot (fold_{16S_j} + fold_{mRNA_i})$$

Fig. 2. Illustrations of the local self-folding and co-fold of the mRNA and rRNA. A. The local self-folding of the mRNA. B. The local self-folding of the 16S rRNA from the 3′ edge. C. The local interaction between the mRNA and the 16S rRNA, with the four typical properties of the local structures and hybridization: 1. mRNA window size – the length of the mRNA portion that interacts with the 16S sequence of the ribosome, 2. The start position at the mRNA - the position, relative to the start codon, of the first nucleotide of the mRNA window sequence, 3. 16S rRNA length – the length of the 16S portion from its 3′end that interacts with the mRNA, 4. The model's constant that determines the subtraction of the self-foldings from the co-folding energy and corrects second-order aspects of translation regulation.

the null model (Fig. 3H7). For further details, see Materials and Methods section 5.6.

Fig. 4A shows the optimal correlation between the free energy model and CAI for every *X*.

The correlations are all negative, although for ease of viewing they are presented in figure 4A as absolute values. All the optimal correlations are in the expected direction and are significant: above 0.61 with P-value (Pv) < 10^{-324}. The strongest correlation was obtained for *X* = *5*, which is also a local maximum between the correlations of *X* = *3,7*. Fig. 4B presents the distances between the optimal correlation of the real and the null models, divided by the null model's standard deviation (STD). We selected the optimal solution from the minimal *X* that got high correlation in addition to high difference from the null model, *X* = *5* answers these conditions with a correlation of r = −0.63 with Pv < 10^{-324}, and differs from the correlation of the null model by 400 STD. As elaborated in Materials and Methods section 5.9 we included two types of null models: in one, the permutations were less global than the other. The results using the less global null model are presented in Supplementary S1 section 2. Supplementary S1 also includes the results related to all the three types of energy models we conducted (see Fig. 2). The scatter plot of the optimal correlation between the Z-scored CAI values and the energy values are shown in Fig. 4C with approximately 16,000 points (see Supplementary S1 Fig. 3 for the scatter plots of the optimal correlations for all the three energy models). The correlations of every chloroplast's genome in the database were calculated separately such that for every gene of a certain chloroplast the energy was calculated according to the optimized parameters of the ortholog group it belongs to. The genomes correlations can be seen in in Fig. 4D and in Supplementary S1 Fig. 5. The genome's optimal correlations are in the expected direction with a median correlation of r = −0.64, the Pv of the genome with the median correlation is Pv=$6 \cdot 10^{-10}$; these results demonstrate that the energy model conducted can be used as a gene expression predictor for every gene of every chloroplast's genome.

2.2. Different gene families in chloroplasts have different translation initiation mechanisms, most of which do not rely on the Shine-Dalgarno interaction

The optimized parameters for every ortholog group were taken from the optimal correlation of *X* = *5* (see Fig. 5). It can be observed that the optimized parameters of the null model distribute uniformly, and all the parameters' values of the real model differ from the null model which gives confidence that the model is meaningful. As mentioned above, we also performed randomization which maintains various aspects of the real data; in this case the inferred values of the real model still significantly differ from the null model. The optimized parameters of the real and null models for all three types of energy models can be seen in Supplementary S1 Fig. 4, and Supplementary S4 includes tables with the optimized parameters for every gene product and for every energy model.

In addition, it can be seen that there are parameters that optimize the prediction for a majority of the orthologous groups. We call groups that share the same optimized parameters that also

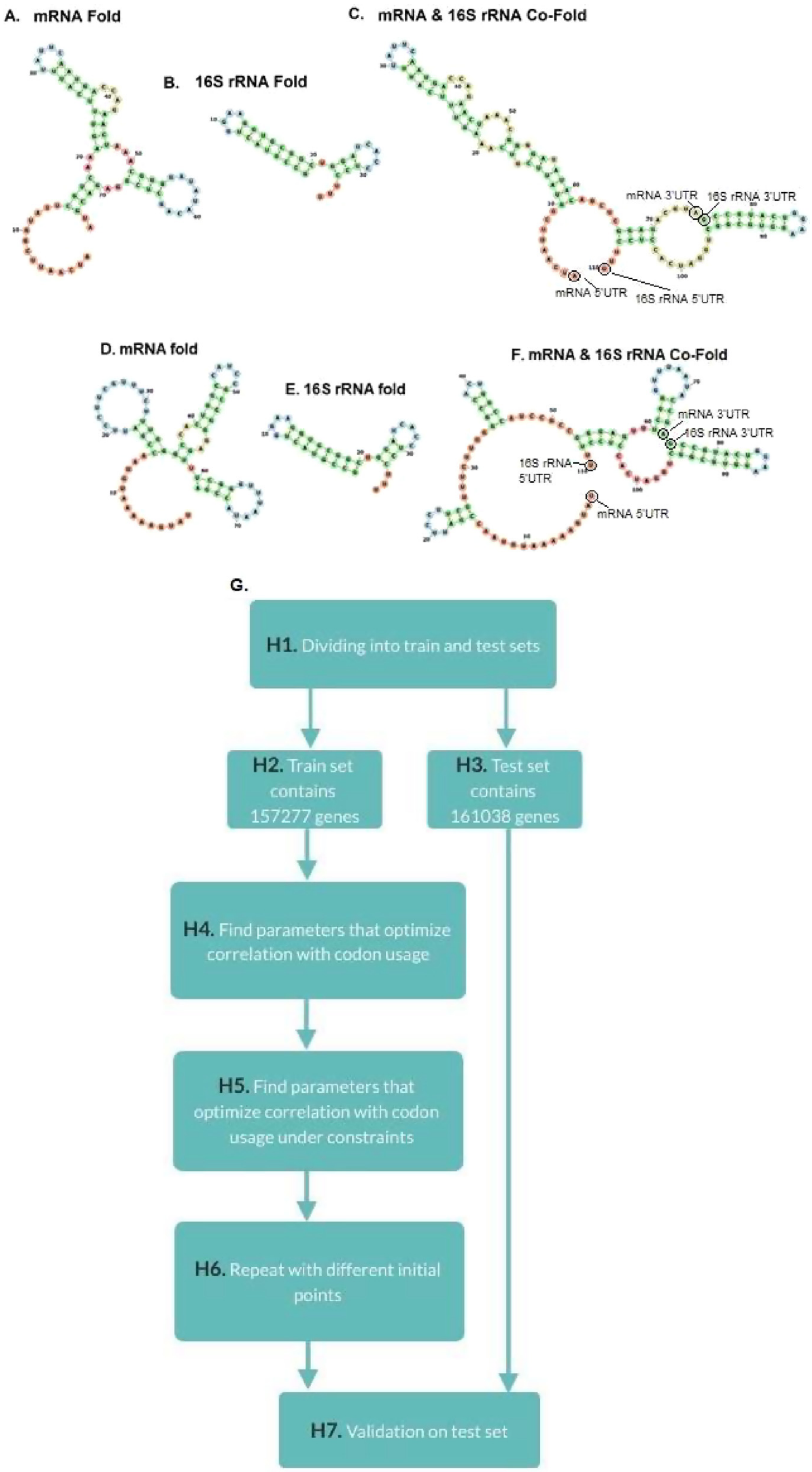

A. mRNA Fold
B. 16S rRNA Fold
C. mRNA & 16S rRNA Co-Fold
mRNA 3'UTR
16S rRNA 3'UTR
mRNA 5'UTR
16S rRNA 5'UTR
D. mRNA fold
E. 16S rRNA fold
F. mRNA & 16S rRNA Co-Fold
mRNA 3'UTR
16S rRNA 3'UTR
16S rRNA 5'UTR
mRNA 5'UTR
G.
H1. Dividing into train and test sets
H2. Train set contains 157277 genes
H3. Test set contains 161038 genes
H4. Find parameters that optimize correlation with codon usage
H5. Find parameters that optimize correlation with codon usage under constraints
H6. Repeat with different initial points
H7. Validation on test set

Fig. 3. Examples of the local self-folding and co-fold of the mRNA (length of 75 nt, one nt upstream of the start codon) and 16S rRNA (length of 35 nt from the 3′ end), of Abelia Sanguinea's chloroplast, and the flow diagram of the energy-based gene expression prediction. A. mRNA folding of rpl20 gene (ribosomal protein L20). B. 16S rRNA folding. C. Co-fold of A. and B. D. mRNA folding of atpF gene (ATP synthase CF0 subunit I). E. 16S rRNA folding. F. Co-fold of D. and E. G. Flow diagram of the energy-based gene expression prediction.

belong to more than 10% of the groups "typical groups"; we call the remaining groups "non-typical groups". For mRNA window length, the values that appear in more than 10% of the groups are: lengths of *85* nt (74%) and *35* nt (18%). As for the 16S window length, the typical parameter's values are: *22* nt (72%) and *41* nt (17%). The typical values of the parameter related to the position of the win-

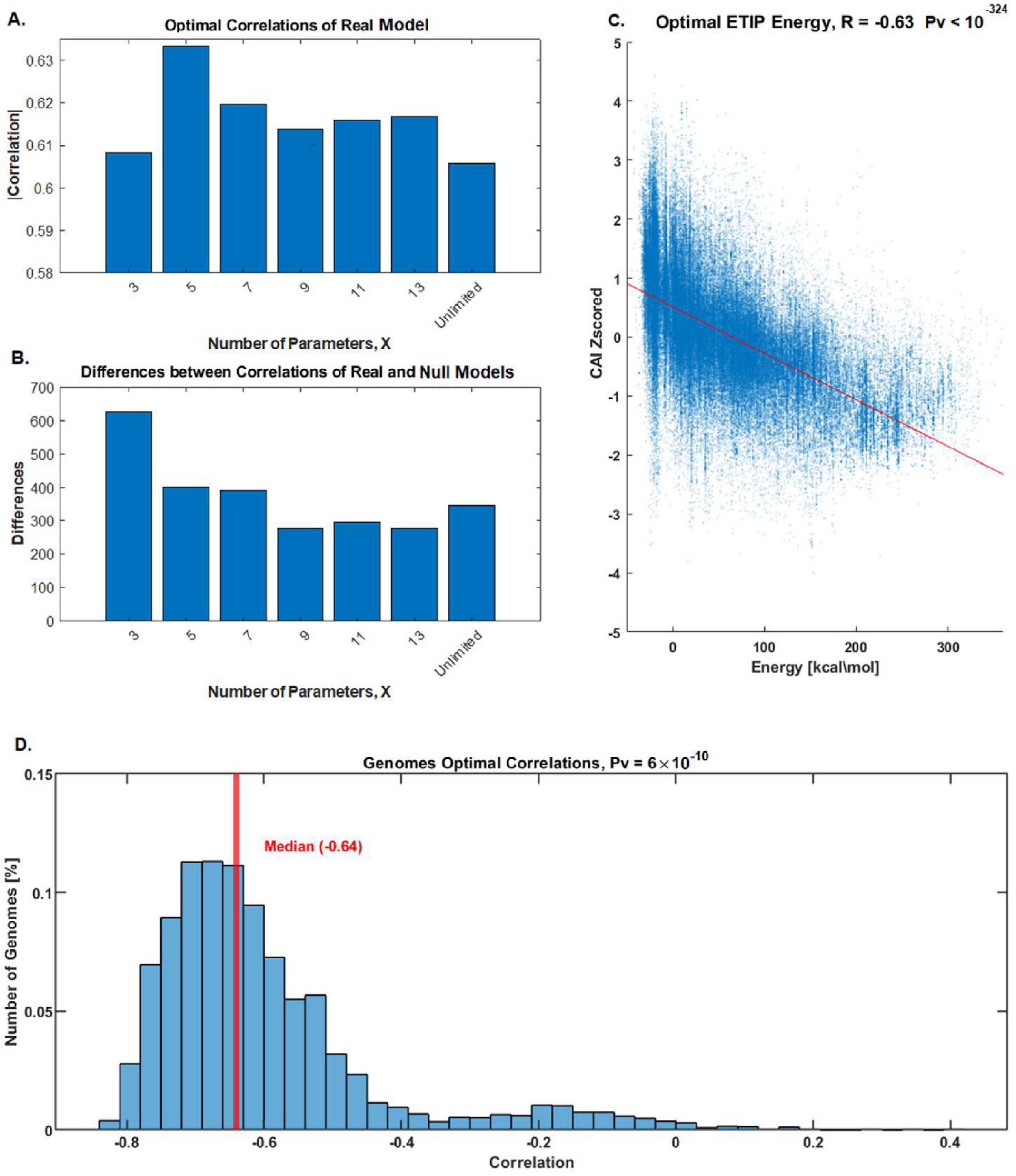


Fig. 4. Optimal Spearman correlations of ETIP model. A. Optimal Spearman correlations of real genomes, and for every X (number of parameters that limits the degree of freedom). B. Distances between optimal correlations of real and random genomes divided by the std of the random correlations for every X. C. Optimal correlation between ETIP energies (x-axis) and Z-scored CAI scores (y-axis) when the dots correspond to the energy and Z-scored CAI of all the genes in the test set. D. All genomes optimal correlations, Pv of the genome with the median correlation.

dow at the 5'UTR are: *26* nt (78%) and *7* nt (13%) upstream of the start codon; these three parameters have two peaks, one with ~75% and the second one with ~16% of the groups, that together sum up to ~91%. However, the constant parameter has three peaks, at 7 (53%), 0 (20%), and 9 (13%), which in total covers 87% of the groups. When considering all the typical values mentioned above, we conclude that there are 49 (64%) groups that are considered typical (i.e., groups that all their parameters are typical), and the rest (36%) are non-typical groups.

We also studied the sets of all four parameters mentioned above. There are three sets of parameters that repeat in a high number of groups; the sets are used by 34%, 14%, and 12% of the typical groups respectively. The parameters of these sets are presented in Table 1. Notably, a 16S rRNA window length of *22* nt and a position of *26* nt upstream of the mRNA start codon are shared by all the three sets (58% of the genes in the typical groups); in addition, 46% of the typical groups have an mRNA window length of *85* nt. In the case of the ETIP constant, the values *7* and *9* (which together account for 44% of the typical groups) are close to each other and support the conjecture that the self-folding influences can have a high effect on the translation efficiency. However, the ETIP constant was optimal at *0* for 14% of the typical groups, in these cases the hybridization between the mRNA and the 16S is likely to be more important than the self-folding for the predictive power of the model mechanism. It could be concluded that the first set in Table 1 includes the optimized parameters which probably have an important role in translation initiation regulation that affect the translation efficiency in most genes.

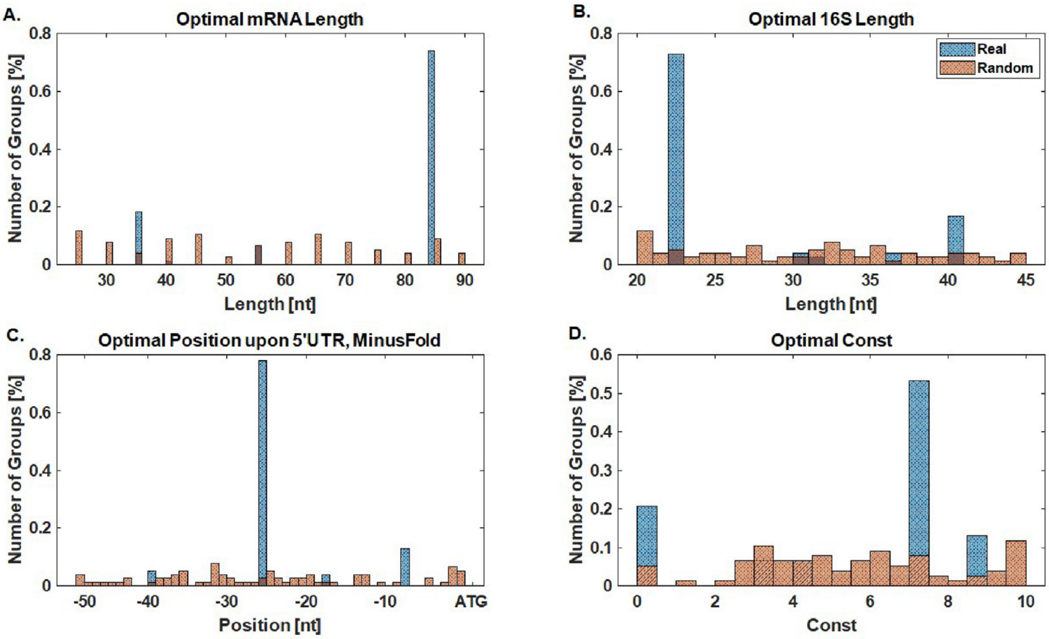


Fig. 5. All the groups' optimal energy parameters of ETIP model. A. Window length of mRNA. B. Window length of 16S. C. Position upon 5′UTR. D. ETIP Constant.

Table 1
Typical parameters sets.

Number of shared groups [%]	mRNA window length [nt]	16S window length [nt]	Position at 5′UTR upstream of the mRNA start codon [nt]	*ETIP* Constant
32	85	22	26	7
14	85	22	26	0
12	35	22	26	9

2.3. Genes of C. reinhardtii that belong to the typical translation initiation regulation groups do not rely on the Shine-Dalgarno interaction

In this subsection, we aimed to better understand the genes related to the typical and non-typical groups. According to *C. reinhardtii's* PA, the PA of the non-typical groups tend to be higher than the PA of other groups with Pv = 0.02, in addition, the typical groups tend to be lowly expressed with Pv = 0.03. The distribution of PA is presented in Fig. 6A.

We calculated the average aSD-SD energy in *C. reinhardtii* for the typical and non-typical groups: the average energy of the non-typical groups is significantly stronger (-1.489) than the average energy of the typical groups (-0.997), with Pv = 0.018 as shown in Fig. 6B. The average position of the SD sequence in the typical groups is 35–30 nt upstream of the start codon, whereas for the non-typical groups it is 16–8 nt upstream of the start codon, in accordance with the typical position of SD in prokaryotes (see Fig. 6C, Pv < 10^{-324}).

2.4. High correlation of a model based on codon usage and the ETIP with energy measurements of protein abundance and ribosomal profiling values

First, a regression was conducted for predicting the PA values of *C. reinhardtii's* genes, once with the CAI scores only and once with the CAI scores and the *ETIP* values. The Spearman correlation between the observed PA values and the predicted ones was calculated. The correlation of the predicted PAs with a regression model based only on the CAI scores is r = 0.65 (Pv = $0.31 \cdot 10^{-6}$), whereas the correlation with the predicted PAs by the regression model based on the CAI scores and the ETIP values is r = 0.71 (Pv = $0.73 \cdot 10^{-8}$). The results show that the energy-based model improves the PA predictions. In order to examine the significance of additional information of the energy model towards the PA values, the partial correlation was calculated and resulted with r = −0.399 and Pv = 0.003 which supports the conjecture that the energy-based model is useful for predicting PAs; moreover, with 95% confidence the coefficient of the energy in the regression is not zero supporting the conjecture that it significantly improves predictions.

Next, a similar process was conducted to predict the ribosomal profiling values of *C. reinhardtii's* genes [28]. In this case the correlation of the predicted ribo-seq values with a regression model based only on the CAI scores is r = 0.60 (Pv = $0.26 \cdot 10^{-4}$), whereas the correlation with the predicted ribo-seq values by the regression model based on the CAI scores and the ETIP values is r = 0.66 (Pv = $0.35 \cdot 10^{-4}$). The partial correlation resulted in r = −0.33 and Pv = 0.035, and the coefficient of the energy in the regression is not zero (also with 95% confidence). The scatter plots of the predicted and real PAs and ribo-seq values are presented in Fig. 7.

2.5. Chloroplast genes tend to have strong structures upstream of the start codon and downstream of the stop codon

Here, we aimed to infer functional local mRNA structures in different chloroplast genes. The functionality of structures can determine their interactions with the rRNA, protein factors, micro-mRNAs, and other components, which can affect various gene expression mechanisms (for instance translation regulation, mRNA stability, mRNA transcription, mRNA transport). They can also

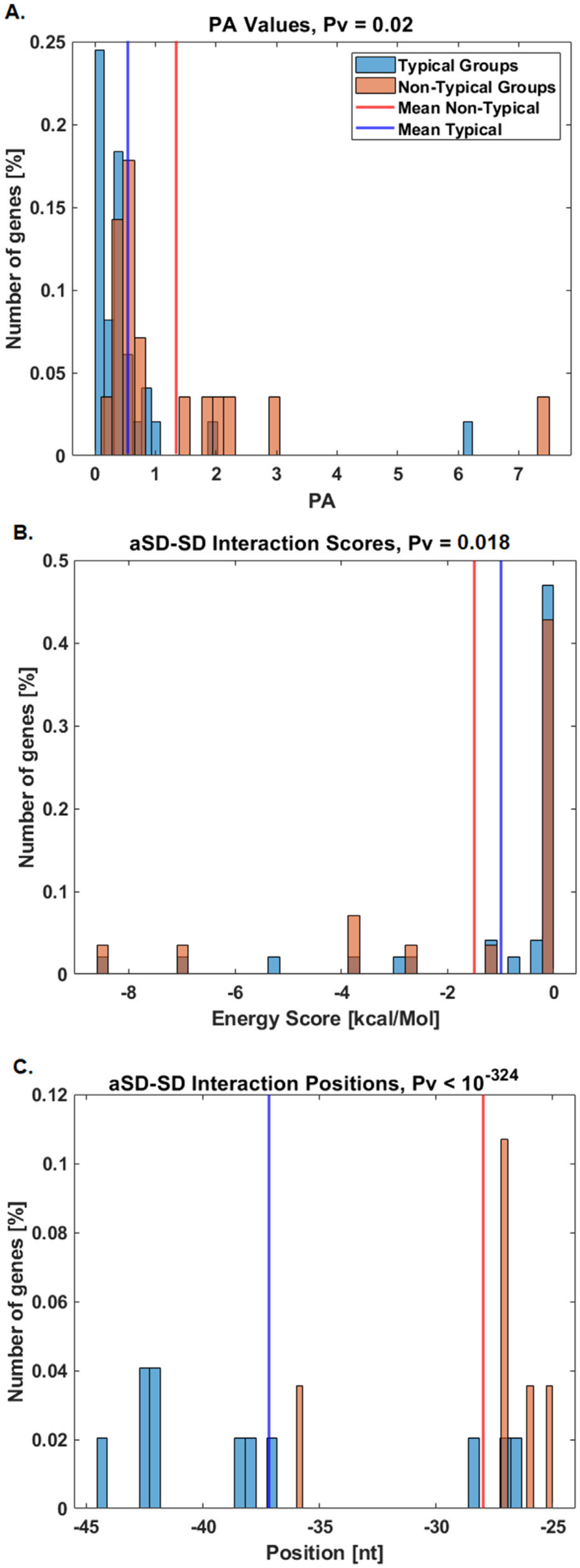


Fig. 6. Histograms of *C. reinhardti* genes divided into typical and non-typical groups. A. PA. B. Scores of aSD-SD interaction. C. Positions of aSD-SD interaction.

affect translation by changing the distance between regulatory sequence motifs and the start codon or by interacting with regulatory sequence motifs.

To this end, folding energies of the mRNA were calculated for every mRNA in the real and random groups and the selection for strong folding was determined by calculating Z-scores and empiric Pvs for every position in the mRNA. The regions of interest to examine the significance of the Z-score values were limited to the positions in the alignment in which more than 50% of the genes in the ortholog group contribute a nucleotide (i.e. the majority of the values in the position are not indels; see Supplementary S1 Fig. 1).

Figs. 8A-8B includes the Pvs related to strong/weak folding energy in comparison to the null model for different positions along the mRNAs. Figs. 8C-8F presents the number of groups that have significant Z-score values normalized by the number of groups with a nucleotide in the positions for the real groups and in comparison to the null model.

There are a few major findings related to this analysis: first, there are many groups with a significant negative Z-score downstream of the stop codon, as shown in Fig. 8B and Fig. 8F. It can be seen that at the positions of 8–20 nt and 140–195 nt downstream of the stop codon there are ~45–48% of groups with significant negative Z-score which differs by more than 45 STD from the average random Z-scores ($Pv < 10^{-324}$). This result suggests that the mRNA undergoes selection to be folded in the positions downstream of the stop codon, possibly to improve the efficiency of the termination by the release factors (RFs). Second, as can be seen in Figs. 8 and 8E, there is a significant negative Z-score at positions 70–50 nt, 185–170 nt and 280 nt upstream of the start codon where the number of groups with such score is close to 40%, higher by more than 30 STD from the average random Z-scores ($Pv < 10^{-324}$); this may suggest that the mRNA tends to undergo selection to be strongly folded upstream of the start codon. It is possible that there are factors that interact with the mRNA in these positions via these structures to promote initiation (as was suggested in [9,23,24]). In addition as can be seen in Fig. 8A and 8C, the mRNA tends to be open at the positions of 30–1 nt upstream of the start codon; this may promote efficient recognition of the start codon by the initiation complex (a signal that also appears in many nuclear genomes [29]). Specifically, ~20–25% of the groups have this signal, which is also stronger by approximately 20 STD than the average random Z-scores ($Pv < 10^{-324}$). In Fig. 8D it can be seen that at the 3′ end there is no significant tendency for any position to be weakly folded in comparison to the null model.

2.6. Conserved potentially functional mRNA structures at the ends of the coding regions

mRNA molecules are populated with functional local structures that can affect gene expression regulation in various ways such as: 1) The binding of the RNA binding proteins (e.g., to the RNA loop); note that the existence of a structure can decrease the distance between the binding motif and the start codon and thus improve the translation efficiency. 2) Via base pairing the structures can prevent the interactions of RNA binding proteins with unwanted binding motifs. 3) The structure can improve the stability of the mRNA by blocking exonucleases.

We expect that functional structures will tend to be conserved throughout all the genes in the ortholog group and structures that are not conserved will probably not be functional. In this study, we tried to detect the functional consensus secondary structure for every ortholog group. Such structures can be used to inform modeling and engineering of gene expression in chloroplasts.

In order to predict the functional secondary structures, positions with significant strong energy folding were discovered by comparing the local self-folding energies of every position at the

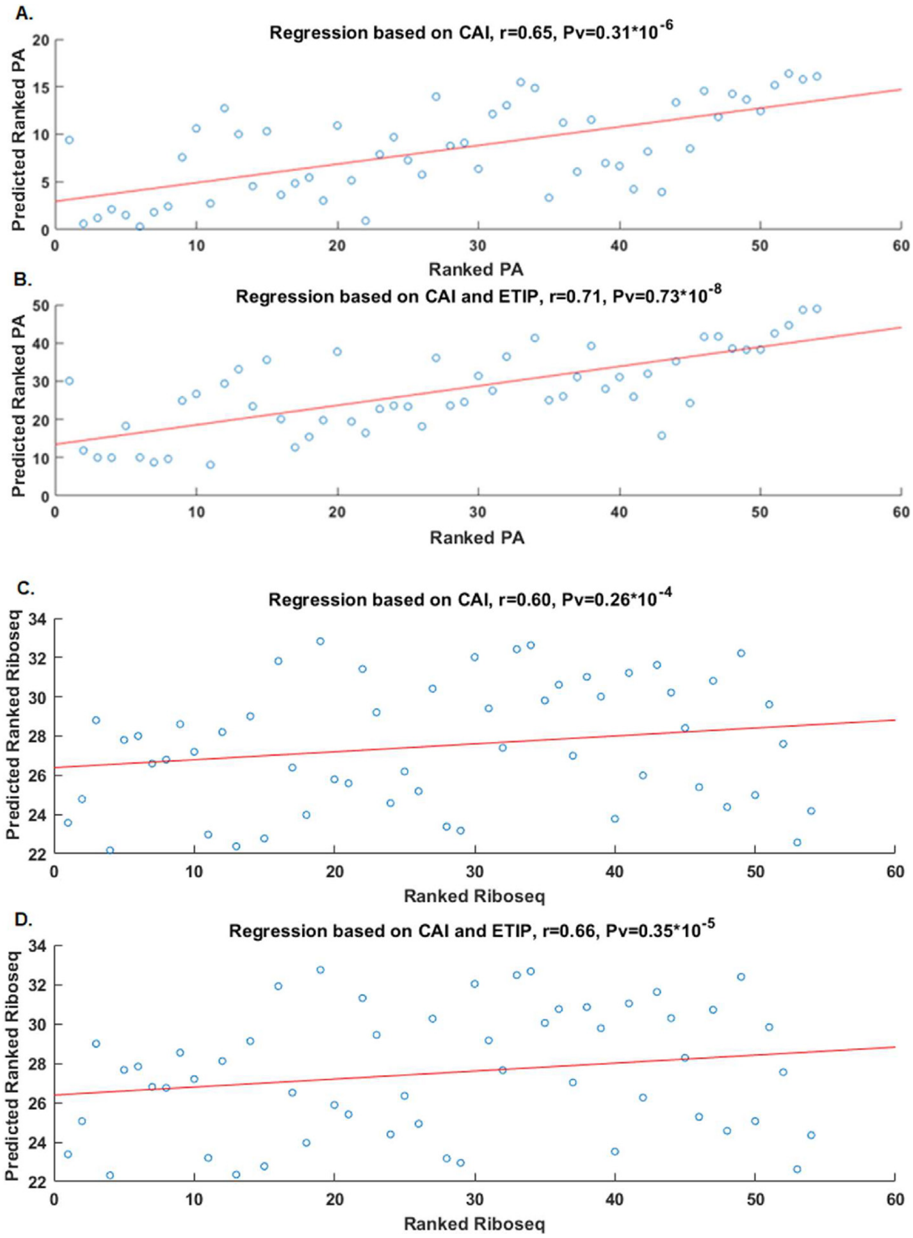


Fig. 7. Spearman Correlations of predicted values of *C. reinhardtii* A. Predicted PA by CAI scores. B. Predicted PA based on CAI scores and ETIP energy values. C. Predicted ribo-seq values by CAI scores. B. Predicted ribo-seq values based on CAI scores and ETIP energy values.

MSA of the real mRNAs to the folding energies obtained by the null model in the same position. At the next step, the consensus secondary structures were detected for these positions on the MSA for every ortholog group by finding the structures based on a dynamic programming algorithm that searches for conserved minimum energy structure over the entire alignment (more details in the Materials and Methods section 5.14 and 5.15).

The information that we provide regarding every structure includes the consensus structure, the energy of the structure in kcal/mol, the expected frequency of the structure among large set of identical mRNAs and the expected structure's diversity related to the current position (i.e., how many different structures we expected to see in this position). The results of this section include many consensus secondary structures related to various orthologous groups which are expected to be functional.

According to our analysis, some of the orthologous groups have more than one typical local structure, most of the groups have 0–2 structures at the 5′UTR (i.e. structures that begin at the 5′UTR and may also include part of the ORF) and 0–1 structures at the 3′UTR (i.e. structures that end at the 3′UTR but may begin at the ORF), as can be seen in Fig. 9A and B. The lengths of the structures at the 5′UTR are between 50 and 350 nt while most of them have a length of ~100 or ~200 nt (Fig. 9C), and most structures at the 3′UTR are in the range of 200–450 nt (Fig. 9D). The lengths of the structures and their geometry may be related to the lengths and properties of the factor that binds to these structures that contributes to the translation initiation or termination. The frequencies of the structures both at the 5′UTR and 3′UTR are very high and close to 100% (Fig. 9E and F; i.e. almost 100% of the mRNA copies are expected to have the predicted structures). The diversities of the structures at the 5′UTR mostly ranges between 8 and 24 (Fig. 9G) and similarly, at the 3′UTR the diversities range between 1 and 28 (Fig. 9H); low diversity means that the molecule has fewer options for local probable structures to fold into, therefore the

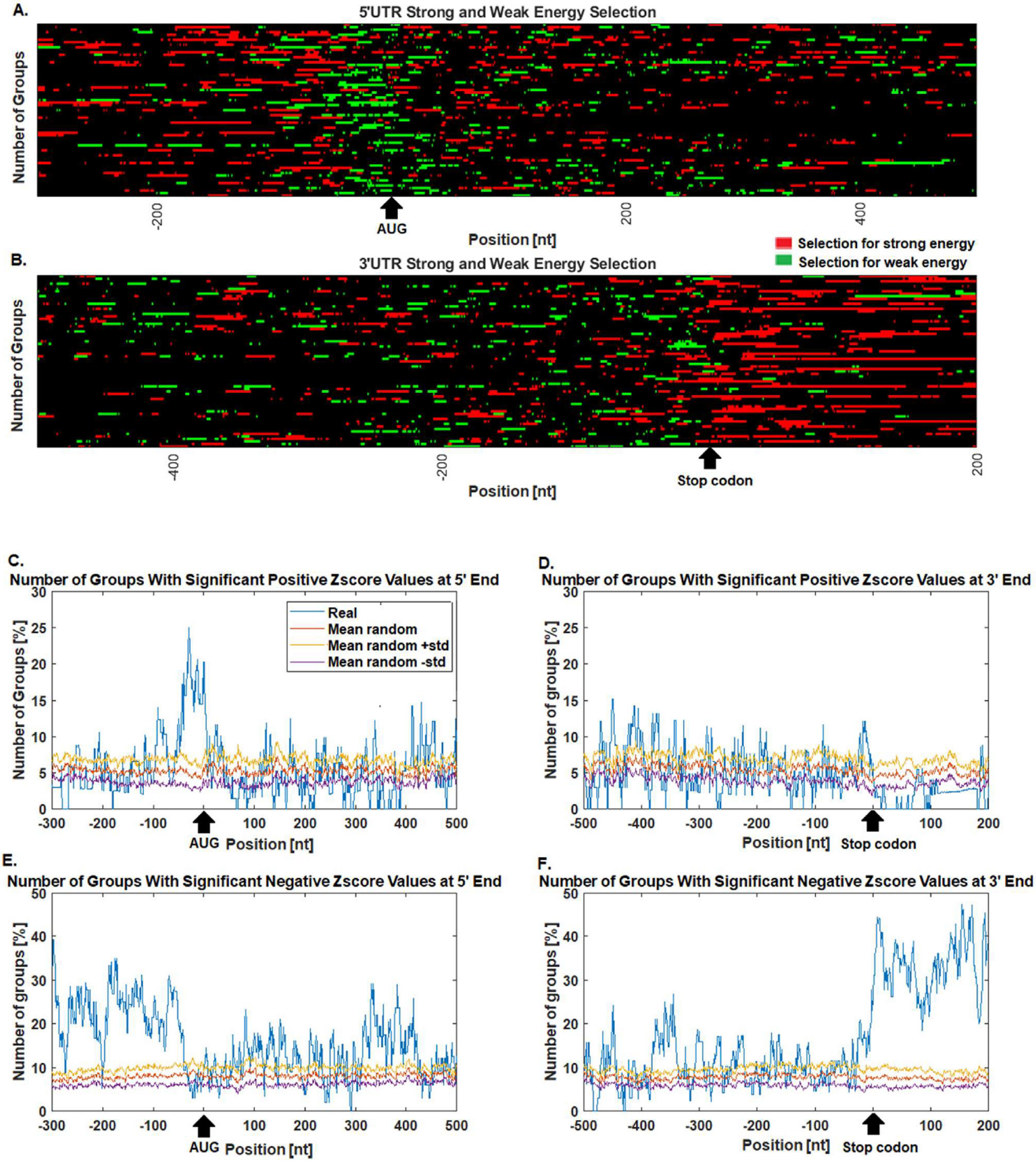


Fig. 8. P values and Z-scores related to a selection for/against strong energy (i.e. stronger/weaker folding in comparison to the null model) according to positions along the mRNA. A.-B. Heat maps of the significant P values for selection for (red)/against (green) strong energy according in different position along the mRNA. A. Alignment to the 5′UTR. B. Alignment to the 3′UTR. C.-F. Number of groups in percent of orthologous groups (y-axis) that has significant Z-score values in different positions (x-axis) along the mRNA of the real groups (blue) compared to the null model (orange). C. Significant positive Z-score values; alignment to the 5′UTR, D. Significant positive Z-score values; alignment to the 3′UTR. E. Significant negative Z-score; alignment to the 5′UTR. F. Significant negative Z-score; alignment to the 3′UTR. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

probability of the molecule being folded in the detected structure is higher. The energies of the structures at the 5′UTR and 3′UTR ranges between −20 to −2 kcal/mol (Fig. 9I and 9J).

2.7. A database of novel conserved secondary structures in chloroplasts

We detected 96 conserved structures at 5′UTR and 70 conserved structures at 3′UTR that we have compiled into a database. All the structures and information regarding their position on the mRNA, length, energy, frequency, diversity, as well as their images are reported in Supplementary S3 and S5.

An example of consensus structures can be seen in Fig. 10A-D. Fig. 10A is the consensus structure that appears in the gene psbC at the 5′UTR, its product is photosystem II CP43 chlorophyll apoprotein. The start position of the structure is nucleotide 806 upstream of the start codon, in the length of 75 nt, the energy is

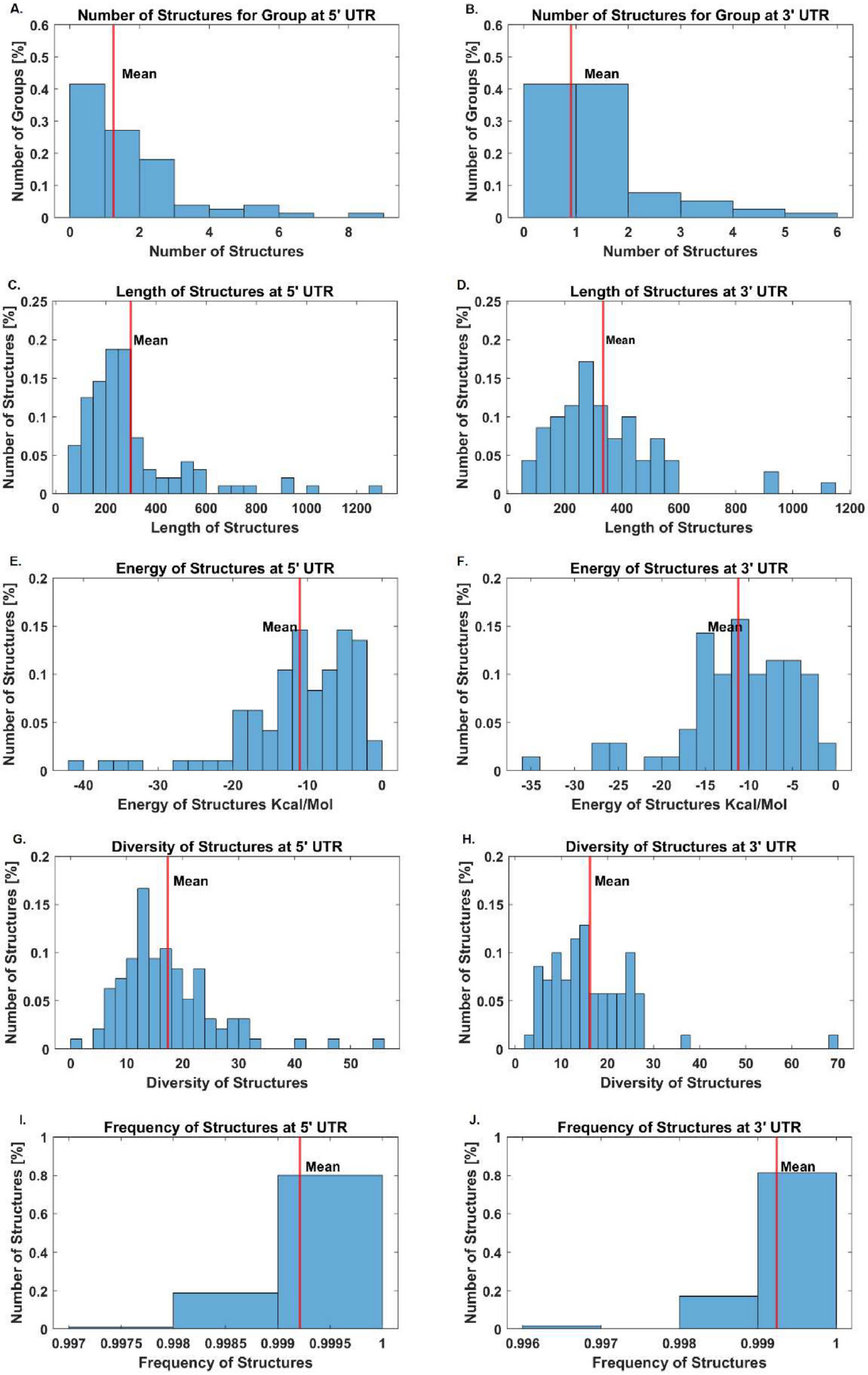


Fig. 9. The general statistics related to the detected conserved structures. A. Number of structures at the 5′UTR. B. Number of structures at the 3′UTR. C. Structures' lengths at the 5′UTR. D. Structures' lengths at 3′UTR, E. Structures' frequencies at the 5′UTR. F. Structures' frequencies at the 3′UTR. G. Structures' diversities at the 5′UTR. H. Structures' diversities at the 3′UTR. I. Structures' energies at the 5′UTR. J. Structures' energies at the 3′UTR.

−11.9 kcal/mol, the frequency is 0.9994 which means that 99% of the sequence's copies would have the specific structure and is considered very high, and the diversity is 12.65 which means that there are 12.65 different structures out of the sequence's copies. Fig. 10B includes the consensus structure that appears in the gene infA at the 5′UTR, its product is translation initiation factor IF-1. The start position of the structure is at nucleotide 244 upstream of the start codon, length of 288 nt, energy of −18.6 kcal/mol. The frequency is 0.9988 (very high), and diversity is 20.5. Fig. 10C is the consensus structure that appears in the gene atpB

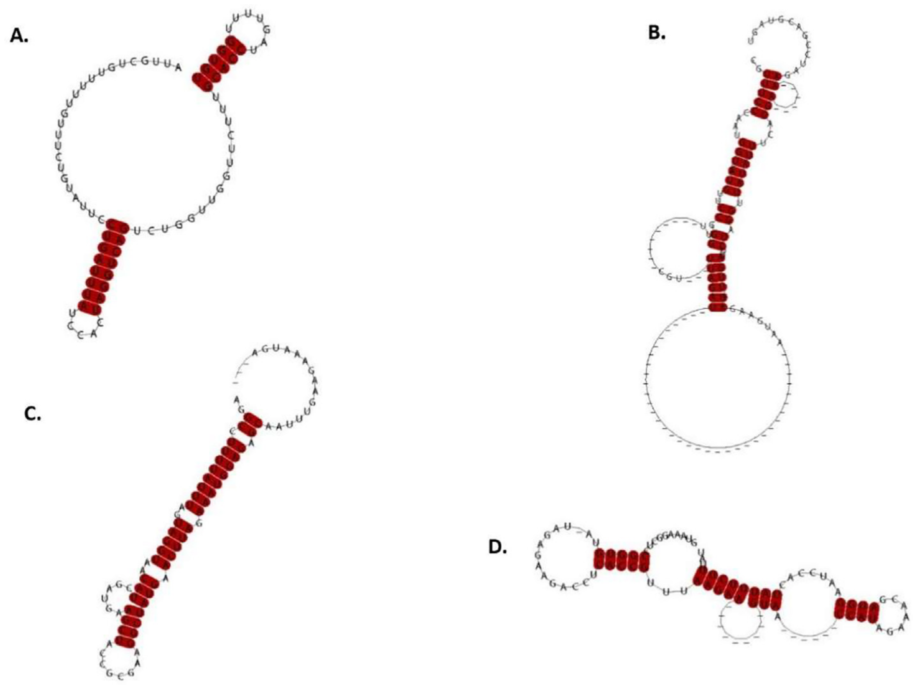


Fig. 10. Examples of the consensus secondary structures. A. Consensus secondary structure in the mRNA of the gene psbC at the 5′UTR; its length is 75 nt and it is located 806 nt upstream of the start codon. B. Consensus secondary structure of gene infA at the 5′UTR; its length is 288 nt and it is located, 244 nt upstream of the start codon. C. Consensus secondary structure of gene atpB at the 3′UTR; its length is 85 nt and it is located, 82 nt upstream of the stop codon. D. Consensus secondary structure of gene rpl20 at the 5′UTR; its length is of 94 nt and it is located 880 nt upstream of the start codon.

at the 3′UTR, the product of this gene is ATP synthase CF1 subunit beta. The start position of the structure is at nucleotide 82 upstream of the stop codon, and is 85 nt long, with an energy of -10.9 kcal/mol. The frequency is 0.9992 (very high), and diversity is 24.47. Fig. 10D includes the consensus structure of that gene rpl20 at the 5′UTR, the product of this gene is the ribosomal protein L20. The start position of the structure is at nucleotide 880 upstream of the start codon, with a length of 94 nt, and energy of -10.3 kcal/mol. The frequency related to this structure is 0.9996 (very high), and related diversity is 13.19.

3. Discussion

Today there is no generic large-scale computational model of translation initiation in chloroplasts that can be used for all genes and all organisms. Therefore, our general aim of this study was to develop novel quantitative models that connect mRNA translation to mRNA and rRNA local folding in chloroplasts; these models are expected to help in understanding the evolution and biophysics of chloroplasts. We studied translation mechanisms in chloroplast genomes by conducting an energy-based model (*ETIP*) that will efficiently predict protein levels in different orthologous groups of chloroplast genes. Based on the *ETIP* we studied the local folding of the mRNA and the local interactions between the mRNA and the 16S rRNA of the small ribosomal subunit. In addition, we identified functional secondary structures that have a wide consensus in genes that belong to the same ortholog group of different chloroplasts genomes. We validated our models via the analysis of PA values and ribosomal profiling values of genes in the chloroplasts of *C. reinhardtii*, a green unicellular alga that is widely used as a model system for studying fundamental aspects of chloroplasts and is also widely used as a model in biotechnology [29]. These results demonstrate the biotechnological promise of our models. In addition, we created, for the first time, a database of 77 orthologous groups out of 4,306 different chloroplast genomes, which can aid future research of chloroplast genomes.

Our predictive energy-based model of translation efficiency is based on the local folding of the mRNA and the local mRNA-rRNA hybridization; it has different parameters for different orthologous groups. We inferred for each ortholog group the typical local energy parameters that are expected to determine translation initiation regulation. The optimized correlation across all genomes between our energy model and the Z-scored CAI values is $r = -0.63$ with $Pv < 10^{-324}$.

The model also efficiently predicts the Z-scored CAI values when a correlation is computed for each genome separately (median correlation of $r = -0.64$ and $Pv = 6 \cdot 10^{-10}$for the genome with the median correlation) which supports the conjecture that this model is generic and universal for all genes in all chloroplasts. Moreover, CAI scores are known to be highly correlated with PA values and we showed that adding the energy-based model to the CAI scores can further improve the correlation with PA values. In all cases we provide comparisons to the null model that support the conclusion that the results are meaningful.

Our model is composed of four variables that describe the local structure and provide a constant parameter which corrects for second-order aspects of translation regulation that are not directly considered in the model. Although it has been reported that PPR proteins support translation by interacting with the 5′UTR of some chloroplasts' UTRs [9,23,24], with the current available data it is impossible to specifically add one (or more) proteins (such as PPR) to our models due to the following reasons: 1) First, there is no quantitative data that measures how PPR proteins interact with mRNA. Note that these interactions are gene and organism-specific. Without such data we cannot infer parameters of a relevant model. 2) The aim of the model we developed here is to provide generic predictive power without getting into the specific details of the many factors involved in translation (since we do not have data related to all of them). Thus, all these aspects are inferred indirectly via the C constant. 3) There are many additional factors (not only PPR) related to translation which would be necessary to include in a non-generic model; such a model would have with nonoptimal performances.

The parameters related to the *ETIP* vary among the different orthologous groups, suggesting that indeed different gene families in chloroplasts use different translation initiation mechanisms. Recently it was discovered that in some chloroplasts translation initiation regulation of some genes is based on SD interaction while in other genes it is not. Our model confirms that the SD interaction is not a ubiquitous mechanism of translation initiation in chloroplasts although it does facilitate initiation in some genes; furthermore, our model suggests that typical translation regulation in chloroplasts is not SD dependent.

Our analysis suggests that 64% of gene families have a 'typical' translation initiation mechanism with optimal parameters consisting of an mRNA window size of *85* nt, a start position of *26* nt upstream of the start codon, a 16S rRNA window size of *22* nt from its 3′ end, and an *ETIP* constant of *7*. The mRNA-rRNA interaction at this position of the mRNA probably has an important role in the translation initiation that should be further investigated to elucidate the exact biophysical mechanisms. As described in the introduction [12], it was discovered that the structure of the 16S rRNA in the region of the anti-SD sequence tends to have point mutations that lead to a stronger structure in these positions; in accordance with these findings, the high optimized constant of the *ETIP* model shows that the self-folded structures of the mRNA and the 16S rRNA are dominant in the initiation regulation.

In accordance with prior studies [4,13-20], genes that were found not to be dependent on the SD (e.g. *petD*, *atpB*, *atpE*, *rps4*, *rps7*, *rbcL*, *rpl2* and *rpls16*) were found in the typical groups while genes that require the aSD-SD interaction during translation initiation (e.g. *psbA*, *psbD*, *psbC*, *atpH* and *rps14*) were found in the non-typical groups. It was also previously suggested that highly expressed genes in *C. reinhardtii* rely on SD interactions in their translation initiation regulation. We provide evidence to support this, with C. *reinhardtii's* non-typical genes being significantly more highly expressed (Pv = 0.01) and showing stronger SD interactions than in the other gene groups with SD sequences positions similar to those in prokaryotes (i.e. 16–8 nt upstream of the start codon). Conversely, the typical groups identified in this study are lowly expressed, their aSD-SD interactions tend to be weaker, and their SD sequences tend to be positioned 35–30 nt upstream of the start codon). Thus, our study supports the hypothesis that typical translation regulation in chloroplasts tends not to rely on SD motifs although it is still used in some genes.

Future work should be done in this field in order to improve our predictive model: a direct high-quality measurement of translation over all chloroplasts could improve the quality of our models. Since such data is not available, we used CAI as a proxy which is known to be related to translation efficiency and other gene expression steps [30]. In addition, we analyzed the PA and ribo-seq measurements of values *C. reinhardtii*'s (in Results section 2.4) since currently this is the only microalgae with such large scale measurements.

Our model could also be improved and further validated by generating libraries of heterologous genes that are designed based on our model and measuring their expression levels (as was done for other organisms [31,33]). Such an experiment would provide insights into the strength of the effect of each parameter of the model on protein levels and the causality of the reported relationships.

In this study, we investigated the positions with significant strong/weak folding selection across the mRNAs in different orthologous groups. We observed novel patterns of selection for strong mRNA folding at the mRNA ends that may be related to unique chloroplast regulatory aspects; the positions with selection for strong folding tend to appear upstream of the start codon and downstream of the stop codon.

We created a database containing 166 predicted functional mRNA structures that are specific to different orthologous groups in chloroplasts, this database is open-source and can be used for studying and engineering chloroplast genes.

As mentioned in the introduction, there are many previous studies that describe various translation factors which function via the interaction with specific RNA structures in chloroplast genes [9,23,24]; these cases are usually presented schematically without a generic quantitative model. The potentially functional secondary structures that we report here may explain some of these. In order to further investigate these structures and the related factors and understand their functionality, it would be helpful to perform experiments that include rRNA-protein cross-linking followed by capturing the different potential factors with antibodies, as is frequently done for the study of RNA binding proteins (e.g. see [33,35]). For example it is possible to examine the interactions between protein PPR10 and atpH of the maize chloroplast; as was previously discovered the role of this protein is to stabilize the atpH gene's mRNA that affects the translation and makes it stronger [9]. Furthermore, some experiments could be conducted in order to check the importance and functionality of the structures by introducing mutations that will affect their folding and studying the effect of these changes on the expression levels of the relevant genes.

4. Conclusions

In this study we conducted a predictive energy-based model of translation initiation (*ETIP*) in chloroplasts that considers the local folding and co-folding energy of the rRNA and the mRNA. A model which combines the *ETIP* with measures of codon usage is expected to yield a correlation of up to 0.71 with protein levels and 0.66 with ribosomal profiling measures. This model can be used to engineer genes in the chloroplast with desired expression levels.

We found the local energy parameters for our model that influence the translation regulation for every ortholog group and demonstrated that there are some different gene families in chloroplasts that use different parameters and thus probably have different translation mechanisms. In agreement with previous studies [4,13-20], our model predicts that in most of the genes in the chloroplasts, translation initiation does not rely only on an aSD-SD interaction; in this study, we provide more details of the alternative translation initiation models.

We observed novel patterns of selection for strong mRNA folding at the ends of the transcripts that may be related to unique chloroplast regulatory aspects. In addition, we created a database of 166 predicted functional mRNA structures that are specific to different orthologous groups in chloroplasts that can be also used for modeling and engineering gene expressions in chloroplasts.

5. Materials and methods

5.1. The analyzed organisms

All the chloroplasts' genomes (4603) were downloaded from NCBI according to their accession number which were extracted from NCBI website (https://www.ncbi.nlm.nih.gov/).

Information regarding the genomes downloaded from NCBI including the genomes name and accession number can be found in Supplementary S2.

5.2. Genes' regions, 5′UTR, ORF, 3′UTR

The ORF for every gene was taken according to the positions in the genome listed in the information from NCBI. The UTRs in almost all the genomes that were analyzed in this study are not annotated. Since it is not clear where the 5′UTR of every gene begins we took the nucleotides from the end of the ORF of the previous gene in the chromosome until the last nucleotide before the start codon as the 5′UTR raw data, and the nucleotides of the 3′UTR of every gene was taken from the end of the ORF until the start codon of the following gene in the chromosome (with median predicted 5′UTR length which is 334 nucleotides and median 3′UTR which is 262). However, we performed a procedure to infer the relevant features of the model (i.e., the length and position of the window upstream of the ORF) by fitting the model to gene expression (as elaborated in Materials and Methods section 5.7). Indeed, the inferred relevant UTR was usually short (a few dozen nucleotides) according to the results. The fact that the model has high predictive power supports the conjecture that our inference is meaningful.

5.3. Protein abundance and ribosomal profiling values

The PA values that were used in this study are based on the normalized average PA from the following two sources:

1) *C. reinhardtii* PA values that were downloaded from [35]; in this case, for each gene the PA was calculated by averaging over all the hours of the day.

2) PA values that were downloaded from PaxDB and includes whole organism PA of *C. reinhardtii* GPM,Aug 2014) [36].

The PA values from every source were normalized such that both sources will have the same average of 1 by dividing the average PA.

C. reinhardtii ribo-seq values were downloaded from [28].

5.4. CAI calculation

CAI, codon adaptation index, is a computational method of predicting the expression level of a gene based on its codon sequence [30].

The steps for calculating this index are:

a) Calculating weights for every codon *i* according to the frequency of this codon in a reference set (the highly expressed genes of the current genome) divided by the maximum frequency of the codon that encodes for the same amino acid, according to Eq. (1):

$$w_i = \frac{x_i}{\max(x)} \tag{1}$$

where w_i refers to the weight of codon *i*, x_i refers to the frequency of codon *i* in the reference set, max(x) refers to the maximal synonymous codon's repetitions.

b) Calculating CAI for every gene in length of *L* codons is by the geometrical mean of all the codons' weights in the sequence according to Eq. (2):

$$CAI_{gene} = \left(\prod_{i=1}^{L} w_i\right)^{\frac{1}{L}} \tag{2}$$

DCBS refers to the directional codon bias score, a measure of the strength of the codon usage bias (CUB) of the gene [37].

DCBS is calculated with the following equations:

the directional codon bias (DCB) of a codon triplet *xyz*:

$$d_{xyz} = \max\left(\frac{f(x,y,z)}{f_1(x)\cdot f_2(y)\cdot f_3(z)}, \frac{f_1(x)\cdot f_2(y)\cdot f_3(z)}{f(x,y,z)}\right) \tag{3}$$

The DCBS of a gene of length *L* codons:

$$DCBS = \frac{\sum_{i=1}^{L} d_{xyz}}{L} \tag{4}$$

CAI was calculated for every gene in every chloroplast genome according to the following steps:

a) DCBS was calculated for every gene, 20% of the genes with the highest DCBS were taken as the reference set for CAI calculations.
b) CAI scores were calculated for every gene.
c) The CAI scores were normalized for every chloroplast by replacing them with a Z-score according to Eq. (3):

$$Zscore = \frac{x_i - mean(x)}{std(x)} \tag{5}$$

where *x* represents the entire group of values (i.e. CAI values related to all the genes in the genome), and x_i is one value from the group (i.e. the CAI of one gene).

5.5. Orthologous groups

In order to create orthologous groups of chloroplast genes we first grouped together genes from different chloroplasts with the same gene product according to the proposed chloroplast gene names [38–41].

The homology between every couple of genes in every group was estimated by BLAST [41].

Pairs of genes that did not meet with the following conditions were filtered:

- E value higher than 10^{-5} (considered as not significant enough based on empirical studies [42]).
- Identity percent lower than 50%.

After this elimination step, a graph was calculated for every group by the remaining pair of genes that are considered to be sufficiently similar. In this graph, nodes are genes in the group and an edge between a couple of genes means that these genes are sufficiently similar according to the conditions explained above. The number of edges that are related to each gene in the graph was calculated, and genes that are connected to less than 40% of the rest of the group members were eliminated from the group. Lastly, genes with lengths significantly different than the mean gene length of the group (probably because of false sequence annotations) were eliminated:

- Gene length shorter than 65% of the group's mean gene length.
- Gene length longer than 140% of the group's mean gene length.

As a result, we generated 77 orthologous groups from the genes in the database with an average of 4,325 genes per group.

Based on this procedure, a gene is added to an orthologous group if its similarity to the rest of the group is above a certain threshold. This threshold (an edge in the graph) is sensitive enough to detect homology even for pair of organisms that are not evolutionary close. Thus since the threshold is absolute and not relative, we do not expect to miss orthologs due to the distribution of organisms in the databases.

5.6. Minimum free energy calculations

The minimum free energies of different sequences (of mRNA and 16S rRNA) were calculated using ViennaRNA package which predicts the secondary structure of RNA sequences and provides the minimum free energy of the thermodynamic ensemble [43].

The analyses along this study used two types of energy calculations by ViennaRNA:

a) RNAfold – calculates the minimum free energy of an RNA sequence, used to calculate the following energies:
   I. mRNA fold - minimum free energy and secondary structure of an mRNA sequence.
   II. 16S fold - minimum free energy and secondary structure of a 16S rRNA sequence.
b) RNAcofold – predicts the secondary structure upon a dimer formation, used to calculate the following energy:
   III. Co-fold of mRNA and 16S - minimum free energy of the hybridization between the mRNA sequence and the 16S rRNA sequence.

The lower and more negative the minimum free energy is, the stronger the structure is folded.

5.7. Energy-based model

We conducted a minimum free energy-based model related to energies of the local mRNA-rRNA hybridization. The model is based on the biophysical model in which at first the mRNA and the 16S are folded in self-folding structures with energy calculated by RNAfold from the previous section, and the second stage is that the two sequences, the mRNA and the16S rRNA, bind together and create a new folded structure, the hybridization energy is calculated by RNAcofold as described previously.

The mRNA window inferred in the model is the portion of the 5′ of the mRNA (which can include both parts of the 5′UTR and parts of the ORF) representing a fragment that (most) effect mRNA translation. This window is described by its length in nucleotides and the position in the transcript where this window starts; thus, the “start position” refers to the position, relative to the start codon, of the first nucleotide of the mRNA window sequence that interacts with the 16S sequence of the ribosome. For example, if the start position of the mRNA window is −15 with a window size of 35 nt, it means that the mRNA window that interacts with the 16S window according to the model is 35 nucleotides long, and it starts at the position of 15 nucleotides upstream of the start codon.

The model, which we called “Energy based Translation Initiation Predictor” (*ETIP*), estimates whether the hybridization energy between the local sequence of the mRNA and the 16S rRNA is stronger than the folding energies of the mRNA and the 16S rRNA separately. If it is, then the probability that the sequences will bind together in order to initiate translation will be higher and the translation will be more efficient. The model actually estimates how much the stability of the mRNA and the 16S rRNA was improved by their hybridization. The energy model is calculated according to Eq. (4):

$$(4)ETIP_i = cofold_{mRNA_i 16S_j} - C * \left(fold_{16S_j} + fold_{mRNA_i}\right) \tag{4}$$

When *i* refers to the *i*’th gene in the database, $16S_j$ is the 16S rRNA sequence of chloroplast *j*.

The aim of the correction factor *C* is to deal with and correct second-order aspects of translation regulation that are not directly considered in the model.

The purpose of this energy-based model is to predict the gene translation initiation efficiency that is expected to be highly correlated with the PA value and therefore with the CAI scores of a gene.

In order to predict the Z-scores of the CAI scores we investigated what are the typical properties of the local structures of the mRNA and its interactions with the 16S rRNA, which optimizes the correlation between the energy value of a gene’s local structures and its CAI score, hence we characterized the structures by four parameters:

1) *mRNA window length*
2) *Start position* of the mRNA window on the 5′UTR of the gene
3) *16S rRNA window length* from the 3′ end
4) *The correction factor, C,* of the *ETIP* model.

We expect that genes with different products will have different translation initiation mechanisms and it will tend to be conserved among the different genes in a group; thus, our model constrains all genes from a certain group to have identical parameters.

For comparison, we also conducted simpler energy-based models: one is based solely on the local folding energy of the mRNA which is predicted by ViennaRNA RNAfold algorithm, and the second one is based on the co-folding energy of the hybridization between the local mRNA sequence and the local 16S rRNA sequence predicted by ViennaRNA RNAcofold algorithm. The mRNA folding model is based on the optimization of parameters 1 and 2, and the co-fold model is based on the optimization of parameters 1,2, and 3. We show that *ETIP* outperformed the simpler models in terms of the correlation with the Z-scored CAI. The optimal correlations and the distance from the null model of all the three energy models (mRNA fold, co-fold, *ETIP*) can be seen in Supplementary S1 Figs. 2–4.

5.8. Optimization process for the energy-based model

The optimization process was conducted by hill-climbing which is an optimization algorithm that makes local steps that improve the objective function until reaching optimization. We randomly divided all the genes in the database into 50% training and 50% test sets such that every set included an equal amount of genes from every ortholog group. The objective function, in this case, is the Spearman correlation between the energies’ values and the Z-scored CAI of the genes in the train set, which is expected to be highly negative, since the more negative the *ETIP* energy result is, the higher the probability of the hybridization occurring, resulting in more efficient translation initiation and therefore higher PA values and CAI scores.

In the optimization process, the aim is to choose the values of the local structure’s four energy parameters that optimize the correlation. Every parameter has a set of values that can be checked and can be selected in the process:

1) mRNA window length, *25–90* nt in steps of 5 nt.
2) 16S rRNA window length from its 3′ end, *20–45* nt in steps of 1 nt.
3) Start position of the mRNA window on the 5′UTR of the gene, *50–0* nt upstream of the start codon, *0* means the start codon itself, in steps of 1 nt.
4) Const of *ETIP*, *0–10* in steps of 0.5.

When the mRNA window size (parameter 1) is bigger than the start position (parameter 3), the sequence of the mRNA’s local structure includes the start codon of the gene.

As already explained, the solution of the process will be such that every group has a set of four parameters, one for every parameter type; such a set is considered as the optimized parameters set for the group.

In addition, we added constraints to the hill-climbing algorithm such that the chosen parameters are not from the entire range described above but are from a limited sub-set in size of *X* that was sampled from this range. We added these constraints to simplify the model and to reduce overfitting; these constraints are also based on the assumption that there is a finite (relatively small) number of regulation strategies in chloroplasts that tend to appear in many genes.

The process was performed for *X = 3, 5, 7, 9, 11, 13* and we also ran the optimization without limiting the possible parameter values (i.e. each parameter can be selected from the entire range mentioned above). Eventually, the results were validated by the test set and were compared to the null model.

5.9. Null model for evaluating the energy-based translation initiation predictor

Two types of null models were conducted. The first one is based on shuffling the Z-score related to the CAI values between the genes of an ortholog group, this operation maintains all the fundamental properties of the real ortholog group (e.g. the amino acid sequence, codon usage, GC content, and evolutionary conservation).

The second null model includes a more global shuffling: the Z-score related to the CAI between all the genes in the database was shuffled. The results of the first, less global, null model appears in Supplementary S1 section 2, and the results of the second one are presented in the main text.

5.10. A regressor for predicting the PA value and ribosomal profiling values of C. reinhardtii's genes

To show that the ETIP model adds predictive information over the CAI values, a regressor was inferred in order to predict the PA values and the ribosomal profiling values of *C. reinhardtii's* genes based on the combinations of CAI scores and the *ETIP* values; its performance was compared to prediction based only on CAI. The predictor was based on ranked values of all the variables and it was evaluated by computing Spearman correlation. In addition, we computed partial correlations to show that *ETIP* has significant correlation with PA values and the ribosomal profiling when controlling for the CAI values.

5.11. P-values

All empiric Pvs in this study were calculated as the fraction of null model randomization with higher/lower value than the real model. Pv lower than 0.05 was considered significant.

5.12. P-value of highly expressed genes in the non-typical groups, and lowly expressed genes in the typical groups, for genes of C. reinhardtii

In order to find out if the *C. reinhardtii* genes that are in the typical/non-typical groups tend to be highly or lowly expressed, the average PA of the genes in the typical/non-typical groups were compared via a permutation test to the average PA of 100 sampled *C. reinhardtii's* gene with similar size to the typical/non-typical groups.

5.13. aSD-SD interaction PSSM in C. reinhardtii

In order to receive the positions where the aSD-SD interaction is most likely to appear in the 5′UTR of the mRNA in *C. reinhardtii's* genes, the hybridization energy between the mRNA and the aSD of prokaryotes ('UCCUCC') was calculated with ViennaRNA RNAcofold algorithm, using a 6 nt sliding window moving along the sequence of the mRNA in 1 nt steps until the start codon. As a result, the position with the lowest interaction score (co-fold energy) was received for every gene. In order to find out if the *C. reinhardtii* genes in the typical/non-typical groups tend to have a strong/weak aSD-SD interaction, the average interaction score of the genes in the typical/non-typical groups were compared via a permutation test to the average interaction score of 100 sampled *C. reinhardtii* genes with similar size to the typical/non-typical groups.

5.14. Significant strong/weak folding selection in different positions upon mRNA

In order to investigate the positions with significant strong/weak folding selection upon the mRNA of genes in an ortholog group, folding energies of the mRNA were calculated by RNAfold of ViennaRNA package, with a 39 nt sliding window moving in 1 nt steps such that the energies are calculated for every position at the mRNA, divided into the 5′UTR (positions at the 5′UTR, from the start of 5′UTR until the start codon; the last window includes 38 nt of the ORF) and 3′UTR (positions at the 3′UTR, from the stop codon until the end of the 3′UTR; the first window includes 38 nt of the ORF).

For every group, the average folding energy values were calculated for every position of the mRNA, and Z-score and empiric Pv were calculated for every position with 50 null models, with an alignment to the start codon or the stop codon.

The Z-score was calculated according to Eq. (3) and Pv was estimated empirically based on the null model mentioned above. A significant Pv was considered lower than 0.05.

In order to define the threshold for an unusual Z-score value in a certain position, we compared it to the position surrounding it with the following procedure:

The difference between the Z-score of the position and the average Z-scores of the surrounding of 100 nt (100 nt to the left, and 100 nt to the right) was calculated according to Eq. (5):

$$Diff = Zscore_i - \frac{\left(\sum_{i-100}^{i+100} Zscore_j\right)}{200} \qquad (5)$$

The threshold of this measure (*Diff*) was computed based on the distribution of values of this measure for orthologous groups generated by the null model; the top and bottom 5 percentile (corresponding to the *Diff* value of -1.5 and 1.5) were used as the significant Z-score threshold.

5.15. Detecting the conserved, potentially functional secondary structures of the mRNA

It is known that mRNA molecules tend to include local functional structures that, among others, can regulate gene expression. We expect that functional structures will be relatively conserved in comparison to non-functional ones. In order to detect the consensus secondary structures that are potentially functional in the different orthologous groups we performed the following steps: first, the MSA of the mRNAs (nucleotide 3′/5′ UTR MSA and amino acid ORF MSA) was computed by Clustal Omega [44]. Next, the folding energies and Z-score were calculated similarly to the previous section on the MSAs, for positions at the 5′UTR (in this case the folding energies for every position are aligned to the start codon) and for positions at the 3′UTR (in this case the folding energies for every position are aligned to the stop codon).

Positions on the mRNA with significant negative Z-scores compared to the surrounding Z-scores (which are considered significant strong energy) were entered into RNAalifold- a ViennaRNA tool which predicts a consensus secondary structure of a set of aligned sequences; this tool finds a structure in this region that is conserved in all the genes in the MSA and reaches the minimum free energy using dynamic programming [45,47]. The output of the RNAalifold algorithm is the consensus secondary structure, its free energy, the predicted frequency of the structure, and the predicted diversity of the structure. The structures are divided into two regions in the gene: the 5′ end of the transcript (the 5′UTR and the beginning of the ORF) and the 3′ end (the 3′UTR and the end of the ORF).

5.16. Null model for detecting the conserved functional secondary structures

In this sub-section we describe how we computed a null model that maintains various fundamental properties of the mRNA MSAs such as the GC content, codon distribution, the encoded proteins, and the sequence distances (and thus the evolutionary distances among the sequences in the MSA) induced by the MSA. Among others the randomization controls for the distribution of organisms in the dataset.

ORF MSA randomization included swapping the codons (while ignoring positions with indels) of the same AAs between two columns of the MSA that are similar in more than 95% of the AAs, and while considering only columns that have no more than 15% indels. UTR MSA randomization included swapping the nucleotides between two columns that have no more than 20% indels. We performed in each case *10n* columns swapping, when *n* is the length of the MSA (in AAs for the ORF MSA, and in nucleotides for the UTR MSA).

5.17. Source code

All the code generated in this study appears in https://www.cs.tau.ac.il/~tamirtul/ChloroplasTrans/.

6. Author Statement

SCE and TT analyzed the data and wrote the paper.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2022.05.030.

References

[1] Bhattacharya D, Medlin L. Update on evolution algal phylogeny and the origin of land plants 1. n.d.
[2] Whatley JM. The endosymbiotic origin of chloroplasts. Int Rev Cytol 1993;144:259–99. https://doi.org/10.1016/S0074-7696(08)61517-X.
[3] Whatley JM, Whatley FR. Chloroplast evolution. vol. 87. 1981.
[4] Zerges W. Translation in chloroplasts. 2000.
[5] Sugiura M. The chloroplast genome. vol. 19. 1992.
[6] Bedbrook JR, Kolodner R. The structure of chloroplast DNA. 30. 1979.
[7] Jiao Y, Guo H. Prehistory of the angiosperms: characterization of the ancient genomes. Adv Bot Res 2014;69:223–45. https://doi.org/10.1016/B978-0-12-417163-3.00009-3.
[8] Martin W, Rujan T, Richly E, Hansen A, Cornelsen S, Lins T, et al. Evolutionary analysis of Arabidopsis, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. Proc Natl Acad Sci U S A 2002;99:12246–51. https://doi.org/10.1073/pnas.182432999.
[9] Zoschke R, Bock R. Chloroplast translation: Structural and functional organization, operational control, and regulation. Plant Cell 2018;30:745–70. https://doi.org/10.1105/tpc.18.00016.
[10] Zheng M, Liu X, Liang S, Fu S, Qi Y, Zhao J, et al. Chloroplast translation initiation factors regulate leaf variegation and development. Plant Physiol 2016;172:1117–30. https://doi.org/10.1104/pp.15.02040.
[11] Sites RB. 3'-Terminal 1974;71:1342–6.
[12] Weiner I, Shahar N, Marco P, Yacoby I, Tuller T. Solving the riddle of the evolution of Shine-Dalgarno based translation in chloroplasts. Mol Biol Evol 2019;36:2854–60. https://doi.org/10.1093/molbev/msz210.
[13] Sugiura M, Hirose T, Sugita M. Evolution and mechanism of translation in chloroplasts. Annu Rev Genet 1998;32:437–59. https://doi.org/10.1146/annurev.genet.32.1.437.
[14] Sakamoto W, Chen X, Kindle KL, Stern DB. Function of the Chlamydomonas reinhardtii petd 5' untranslated region in regulating the accumulation of subunit IV of the cytochrome b6/f complex. Plant J 1994;6:503–12. https://doi.org/10.1046/j.1365-313x.1994.6040503.x.
[15] Fargo DC, Zhang M, Gillham NW, Boynton JE. Shine-Dalgarno-like sequences are not required for translation of chloroplast mRNAs in Chlamydomonas reinhardtii chloroplasts or in Escherichia coli. Mol Gen Genet 1998;257:271–82. https://doi.org/10.1007/s004380050648.
[16] Koo JS, Spremulli LL. Analysis of the translational initiation region on the Euglena gracilis chloroplast ribulose-bisphosphate carboxylase/oxygenase (rbcL) messenger RNA. J Biol Chem 1994;269:7494–500. https://doi.org/10.1016/s0021-9258(17)37313-1.
[17] Mayfield SP, Cohen A, Danon A, Yohn CB. Translation of the psbA mRNA of Chlamydomonas reinhardtii requires a structured RNA element contained within the 5' untranslated region. J Cell Biol 1994;127:1537–45. https://doi.org/10.1083/jcb.127.6.1537.
[18] Nickelsen J, Fleischmann M, Boudreau E, Rahire M, Rochaix J-D. Identification of cis-acting RNA leader elements required for chloroplast psbD gene expression in Chlamydomonas. Plant Cell 1999;11:957–70. https://doi.org/10.1105/tpc.11.5.957.
[19] Betts L, Spremulli LL. Analysis of the role of the Shine-Dalgarno sequence and mRNA secondary structure on the efficiency of translational initiation in the Euglena gracilis chloroplast atpH mRNA. J Biol Chem 1994;269:26456–63.
[20] Hirose T, Kusumegi T, Sugiura M. Translation of tobacco chloroplast rps14 mRNA depends on a Shine-Dalgarno-like sequence in the 5'-untranslated region but not on internal RNA editing in the coding region. FEBS Lett 1998;430:257–60. https://doi.org/10.1016/s0014-5793(98)00673-5.
[21] Bieri P, Leibundgut M, Saurer M, Boehringer D, Ban N. The complete structure of the chloroplast 70S ribosome in complex with translation factor pY. EMBO J 2017;36:475–86. 10.15252/embj.201695959.
[22] Yamaguchi K, Subramanian AR. The plastid ribosomal proteins. Identification of all the proteins in the 50 S subunit of an organelle ribosome (chloroplast). J Biol Chem 2000;275:28466–82. https://doi.org/10.1074/jbc.M005012200.
[23] Daniell H, Chase C. Molecular biology and biotechnology of plant organelles: chloroplasts and mitochondria. Springer; 2004.
[24] Marín-Navarro J, Manuell AL, Wu J, Mayfield SP. Chloroplast translation regulation. Photosynth Res 2007;94:359–74. https://doi.org/10.1007/s11120-007-9183-z.
[25] Mauger David M, Siegfried Nathan A, Weeks Kevin M. The genetic code as expressed through relationships between mRNA structure and protein function. FEBS Letters 2013;8:587–1180. https://doi.org/10.1016/j.febslet.2013.03.002.
[26] Prikryl J, Rojas M, Schuster G, Barkan A. Mechanism of RNA stabilization and translational activation by a pentatricopeptide repeat protein. Proc Natl Acad Sci U S A 2011;108:415–20. https://doi.org/10.1073/pnas.1012076108.
[27] Bahiri-Elitzur S, Tuller T. Codon-based indices for modeling gene expression and transcript evolution. Comput Struct Biotechnol J 2021;19:2646–63. https://doi.org/10.1016/j.csbj.2021.04.042.
[28] Zoschke R, Watkins KP, Barkan A. A rapid ribosome profiling method elucidates chloroplast ribosome behavior in vivo. Plant Cell 2013;25:2265–75. https://doi.org/10.1105/tpc.113.111567.
[29] Peeri M, Tuller T. High-resolution modeling of the selection on local mRNA folding strength in coding sequences across the tree of life. Genome Biol 2020;21:63. https://doi.org/10.1186/s13059-020-01971-y.
[30] Rochaix JD. Chlamydomonas reinhardtii as the photosynthetic yeast. Annu Rev Genet 1995;29:209–30. https://doi.org/10.1146/annurev.ge.29.120195.001233.
[31] Sharp PM, Li W-H. The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res 1987;15:1281–95. https://doi.org/10.1093/nar/15.3.1281.
[32] Ben-Yehezkel T, Atar S, Zur H, Diament A, Goz E, Marx T, et al. Rationally designed, heterologous S. cerevisiae transcripts expose novel expression determinants. RNA Biol 2015;12:972–84. https://doi.org/10.1080/15476286.2015.1071762.
[33] Goodman DB, Church GM, Kosuri S. Causes and effects of N-terminal codon bias in bacterial genes. Science 2013;342:475–9. https://doi.org/10.1126/science.1241934.
[34] Han T, Kim JK. Mapping the transcriptome-wide landscape of RBP binding sites using gPAR-CLIP-seq: experimental procedures. Methods Mol Biol 2016;1361:77–90. https://doi.org/10.1007/978-1-4939-3079-1_5.
[35] Hafner M, Katsantoni M, Köster T, Marks J, Mukherjee J, Staiger D, et al. CLIP and complementary methods. Nat Rev Methods Primers 2021;1:20. https://doi.org/10.1038/s43586-021-00018-1.

[36] Strenkert D, Schmollinger S, Gallaher SD, Salomé PA, Purvine SO, Nicora CD, et al. Multiomics resolution of molecular events during a day in the life of Chlamydomonas. Proc Natl Acad Sci U S A 2019;116:2374–83. https://doi.org/10.1073/pnas.1815238116.

[37] Wang H, Gau B, Slade WO, Juergens M, Li P, Hicks LM. The global phosphoproteome of Chlamydomonas reinhardtii reveals complex organellar phosphorylation in the flagella and thylakoid membrane. Mol Cell Proteomics 2014;13:2337–53. https://doi.org/10.1074/mcp.M114.038281.

[38] Renana S, Tamir T. Modelling the efficiency of codon-tRNA interactions based on codon usage bias. DNA Res 2014;21:511–25. https://doi.org/10.1093/dnares/dsu017.

[39] Hallick RB, Bottomley W. Proposals for the naming of chloroplast genes. Plant Mol Biol Reporter 1983;1:38–43. https://doi.org/10.1007/BF02712675.

[40] Hallick RB, Bairoch A. Proposals for the naming of chloroplast genes. III. Nomenclature for open reading frames encoded in chloroplast genomes. Plant Mol Biol Rep 1994;12:30–1. https://doi.org/10.1007/BF02671562.

[41] Hallick RB. Proposals for the naming of chloroplast genes. II. Update to the nomenclature of genes for thylakoid membrane polypeptides. Plant Mol Biol Rep 1989;7:266–75. https://doi.org/10.1007/BF02668635.

[42] Korf I, Yandell M, Bedell J. Blast. O'Reilly Media, Inc. 2003.

[43] Li L, Stoeckert Jr CJ, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 2003;13:2178–89. https://doi.org/10.1101/gr.1224503.

[44] Lorenz R, Bernhart SH, Zu Siederdissen CH, Tafer H, Flamm C, Stadler PF, et al. ViennaRNA Package 2.0. Algorithms Mol Biol 2011;6:1–14.

[45] Sievers F, Higgins DG. Clustal omega. Curr Protoc Bioinf 2014;48:3–13.

[46] Hofacker IL. RNA consensus structure prediction with RNAalifold. Methods Mol Biol 2007;395:527–43. https://doi.org/10.1385/1-59745-514-8:527.

[47] Bernhart SH, Hofacker IL, Will S, Gruber AR, Stadler PF. RNAalifold: improved consensus structure prediction for RNA alignments. BMC Bioinf 2008;9:1–13.