Synthetic and Systems Biotechnology 10 (2025) 337–345

Contents lists available at ScienceDirect

Synthetic and Systems Biotechnology

journal homepage: www.keaipublishing.com/en/journals/synthetic-and-systems-biotechnology

Original Research Article

Modeling coding sequence design for virus-based expression in tobacco

Moritz Burghardt [a], Tamir Tuller [a,b,c,*]

[a] Department of Biomedical Engineering, The Iby and Aladar Fleischman Faculty of Engineering, Tel Aviv, Israel
[b] The Segol School of Neuroscience, Tel-Aviv University, Tel Aviv, Israel
[c] Center for Physics and Chemistry of Living Systems, Israel

ARTICLE INFO

Keywords:
ICON
codon usage
heterologous gene expression
transient expression
plants
viral based vectors
synthetic biology

ABSTRACT

Transient expression in Tobacco is a popular way to produce recombinant proteins in plants. The design of various expression vectors, delivered into the plant by *Agrobacterium*, has enabled high production levels of some proteins. To further enhance expression, researchers often adapt the coding sequence of heterologous genes to the host, but this strategy has produced mixed results in Tobacco.

To study the effects of different sequence features on protein yield, we compile a dataset of the yields and coding sequences of previously published expression studies of more than 200 coding sequences.

We evaluate various established gene expression models on a subset of the expression studies. We find that use of tobacco codons is only moderately predictive of protein yield as informative sequence features likely extend over multiple codons. Additionally, we show that codon usage of organisms that use tobacco as a host for expression of their proteins in a similar way as the synthetic system, like viruses and agrobacteria, can be used to predict heterologous expression. Other predictive features are related to tRNA supply and demand, the inclusion of a translational ramp of codons with lower adaptation to the tRNA pool at the beginning of the coding region, and the amino acid composition of the recombinant protein. A model based on all the features achieved a correlation of 0.57 with protein yield.

We believe that our study provides a practical guideline for coding sequence design for efficient expression in tobacco.

1. Introduction

While traditionally, bacterial or mammalian cells have been used as hosts for recombinant expression of biopharmaceuticals, industrial enzymes, and other valuable proteins, plants have become a viable alternative [1]. Advantages of plant-based systems include cost-effectiveness, scalability, and their ability to perform post-translational modifications. Tobacco has established itself as the most used host [2–4].

Transient expression, where the gene of interest is delivered into the grown plant via agroinfiltration (Fig. 1), has been used extensively for achieving expression of the recombinant protein within days [5]. Various expression vectors using genetic elements from plant viruses and agrobacteria have been developed using different strategies for achieving high expression: vectors using the replication machinery of RNA viruses (e.g., magnICON [6], TRBO [7], pEff [8]) or DNA-viruses (e.g., pBY [9]) and non-replicating vectors (e.g., pEAQ, pTRA [10,11]).

While many parameters of these systems have been optimized, designing coding sequences that lead to high expression has remained challenging. Multiple studies have compared yields of wild-type sequences and codon-optimized sequences and found that existing sequence optimization algorithms often do not lead to increased expression and, in some cases, perform worse than the wild-type sequence (Table 1). Surprisingly, some studies enhanced expression in tobacco by using sequences that were codon-optimized for expression in humans. This demonstrates that some coding sequences enhance heterologous expression in tobacco, but established sequence-optimization methods fail to reliably generate them.

Outside of tobacco, various features of the coding sequence, including codon usage [20,21], mRNA folding [22–24] and avoidance/selection of different regulatory patterns [25], have been found to potentially impact all steps of gene expression [26]. Many tools have been developed to optimize one [27,28] or multiple [29,30] of these objectives, but data is lacking regarding the effectiveness of these algorithms [31,32].

To understand which sequence elements lead to high transient

* Corresponding author. Department of Biomedical Engineering, The Iby and Aladar Fleischman Faculty of Engineering, Tel Aviv, Israel.
E-mail address: tamirtul@tauex.tau.ac.il (T. Tuller).

https://doi.org/10.1016/j.synbio.2024.12.002
Received 23 August 2024; Received in revised form 7 December 2024; Accepted 9 December 2024
Available online 11 December 2024

expression in tobacco we have compiled a database of previous experiments with their expression level and coding sequence. We evaluate various gene expression models and infer which sequence features are associated with high protein yield and should therefore be selected in synthetic sequences.

2. Methods

2.1. Selection of heterologous expression studies

It is important to emphasize that since the ultimate objective of heterologous expression is production of the protein, our goal was to model the combined effects of the coding sequence on all steps of gene expression, including transcription, RNA replication, translation and RNA stability. Thus, we aimed at predicting protein levels (and not for example, protein per mRNA, or mRNA levels). While it could be interesting to model mRNA and protein levels separately (or normalize one by the other) to better understand the impact of the coding sequence at each step of gene expression, this is currently not feasible as very few studies reported both metrics.

Studies were selected and processed as follows.

- The gene of interest was transiently expressed through transfection by *Agrobacterium* in *Nicotiana Benthamiana* or *Nicotiana Tabacum*.
- Yield of the recombinant proteins were reported in at least one of three units: expression yield per gram fresh weight (gfw), purified yield per gfw or % of total soluble protein (TSP). To combine data for all three yield metrics we inferred the expression yields per gfw (if not provided) from % TSP/purified yields through linear regressors trained on studies that measured both (Supplementary Fig. S1). To compare data between expression vectors with varying average yields, we normalized each yield by the average for each vector.
- Preferably, coding sequences were provided directly in the studies or through reference to GenBank. When sequences were not specified, we a) chose a representative wildtype coding sequence if the gene was not mentioned to be codon-optimized or b) requested the sequence from the authors if the gene was codon-optimized.
- Additionally, we collected the expression vector and donor organism of the gene for each study. The number of sequences per donor organisms is shown in Supplementary Fig. S2.
- Studies that expressed multiple heterologous sequences and measured the yield of the assembled multimer (e.g., virus-like particles or monoclonal antibodies) were excluded due to uncertainties in the formation efficiency of these complexes (e.g., reported "less than 10 %" for VLPs in Ref. [33]) and the lack of models for predicting the expression level of multiple co-expressed coding sequences.

Table 1
Previous transient expression studies comparing native and optimized sequences. +, ~, - indicate high, average, or low relative expression levels.

Protein	Vector	native	plant-optimized	human-optimized
HPV L1 protein [11]	pTRA	~	–	+
Amelogenin [12]	pEAQ	n/a	~	~
α-amylase [13]	pCaMterX	~	+	n/a
Enterokinase [14]	magnICON	+	~	n/a
Zera®M2e PB [15]	pTRA	n/a	~	~
Lumbrokinase (PI239) [16]	pBY	~	~	n/a
IFNγ [17]	BaMV	~	~	n/a
Acidic Fibroblast Growth Factor (aFGF) [18]	pBY	~	~	n/a
IgG3/IgM [19]	magnICON	~	+	n/a

2.2. Overview of CDS models and indexes

We evaluated various models of gene expression (Table 2) [38]. Models used can be grouped into two categories: models relying on learning the rules of expression from a set of reference genes (CAI and cARS) and models based on consideration of the biophysical process (tAI, tAI_ramp, nTE and dG_start/dG_ramp/dG_gene). Details on each

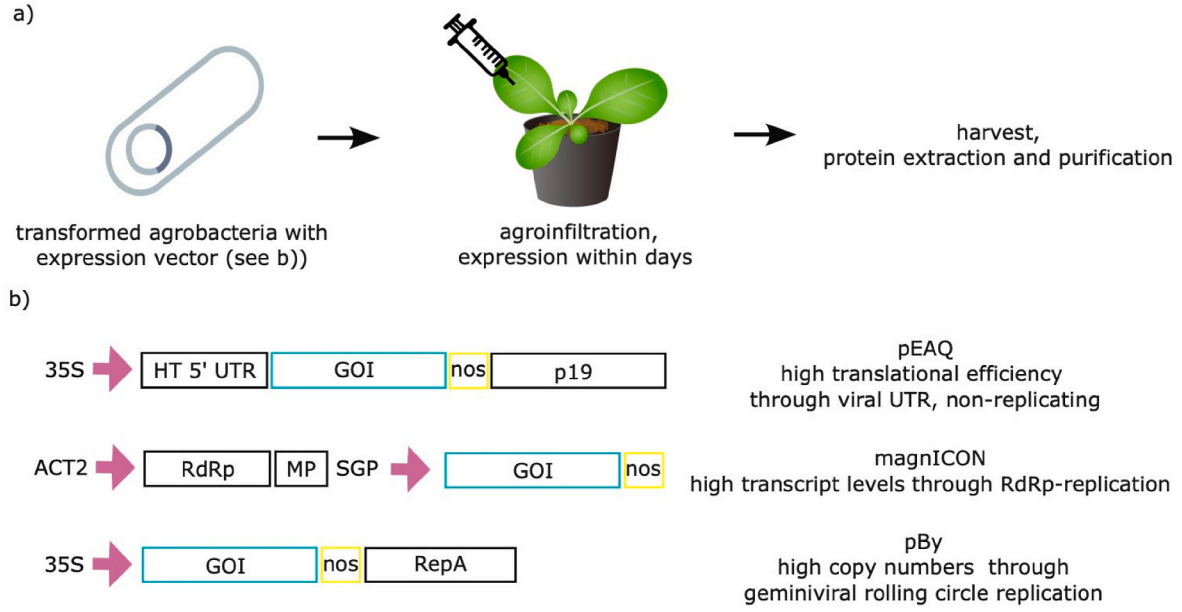

Fig. 1. a) Transformed *Agrobacterium tumefaciens* with expression vectors are delivered into plant leaves. Following delivery to the plant cell nucleus by *A. tumefaciens*, the GOI is expressed by the plant. Protein levels typically peak 3–10 days post-infiltration, after which leaves are harvested and protein is extracted. b) Simplified illustration of three commonly used plant expression vectors and their comparative advantages. 35S, 35S promoter from cauliflower mosaic virus (CPMV); HT 5′ UTR, 5′ UTR from CPMV RNA-2; nos, nopaline synthase terminator from *A. tumefaciens*; p19, RNA silencing suppressor from tomato bushy stunt virus; GOI, gene of interest; Act2, A. thaliana ACT2 promoter; RdRp, turnip veinclearing virus (TVCV) RNA-dependent RNA polymerase; MP, TVCV movement protein; SGP, subgenomic promoter; RepA, the replication proteins from the bean yellow dwarf virus.

Table 2
Overview of gene expression models evaluated in this study.

Feature	Description
Codon adaption index (CAI) [34]	Measures codon usage bias of a gene compared to a set of reference genes.
Chimera Average repetitive substring (cARS) [28]	Measures the appearance of long substrings in a gene that are also found in a set of reference genes.
tRNA adaptation index (tAI) [35]	Adaptation of a coding sequence to the tRNA pool of an organism.
tAI_ramp [36]	tAI calculated using only the first 50 codons normalized by tAI of the entire sequence.
normalized translational efficiency (nTE) [37]	Adaptation of a coding sequence to the supply-demand ratio of each tRNA.
dG_start, dG_ramp, dG_gene [22]	Folding energy of the predicted mRNA structure

model and its use in this study are given below.

2.3. Expression models based on reference genes

The CAI and cARS gene expression models quantify the degree to which a coding sequence follows patterns identified in a set of reference genes. The CAI model does so by calculating a weight for each codon, defined as its frequency in the reference gene set normalized by the frequency of the most frequently used synonymous codon. The CAI of a coding sequence is then given by the geometric mean of the weights of all codons in its sequence [34].

The cARS models expression by identifying the longest substring that originates from each position in the sequence and also appears in the set of reference sequences. The model then calculates the average substring length across all positions. Therefore, cARS is sensitive to sequence features that stretch over more than a single codon and has successfully predicted the expression of endogenous and heterologous genes, often outperforming CAI [28].

For both models, the choice of an appropriate set of reference genes is crucial. To predict transient expression in tobacco, we should choose genes that have a high chance of being highly expressed in the agro-infiltrated plant. Therefore, we considered genes not only from tobacco but also from other organisms that use the plant gene expression machinery similar to the synthetic system.

- Tobacco genes: We used genome assembly GCF_0007151351 for *N. Tabacum* from NCBI. Since selection for high expression is expected to be strongest in highly expressed host genes, reference sets of highly expressed genes are commonly used for learning sequence features to enhance expression. We selected the 100 most highly expressed tobacco genes according to their protein abundance from Pax-DB 5.0 [39]. Protein abundances were mapped to gene sequences according to their UniProt cross-references [40].
- Genes of tobacco viruses: expression of the recombinant protein may affect cellular conditions through metabolic load, allocation of cellular resources (tRNA, ribosomes, etc.) and cellular defense mechanisms such as mRNA silencing. This effect could be especially pronounced for systems that use viral replicons to achieve high transcript numbers. Cellular conditions may be similar to virus-infected cells for which viral genes are selected. Capsid genes particularly tend to be highly expressed [41] and therefore are likely selected for high translational efficiency. Known viruses for *N. Benthamiana* and *N. Tabacum* and their annotated genes were downloaded from Virus-Host DB [42]. The resulting dataset consisted of 664 genes from 176 viruses. To investigate the predictiveness of different virus and gene types we split the data into 4 subsets according to their annotations: capsid genes of RNA viruses (44), capsid genes of DNA viruses (38), and all remaining (non-capsid) genes of RNA viruses (305), remaining genes of DNA viruses (245).
- Following the same logic, we explored learning codon usage rules from genes from *Agrobacterium* T-DNA: 12 Ti-plasmids were downloaded from NCBI (GenBank Acc.: AE007871.2, AF242881.1, DQ058764.1, CP000637.1, KX388535.1, AP002086.1, MK318986.1, KY000031, CP049215.1, CP120216, KY000029, KY000025, CP072164, KY000075). T-DNA regions were identified through the location of T-DNA border consensus sequences [43]. In total 205 T-DNA genes were extracted.

Codon usage weights for each reference sequence set and their correlations are shown in Supplementary Fig. S3.

2.4. Expression models based on biophysical features

Biophysical models estimate the level of expression of a sequence based on our understanding of the gene expression process, in particular gene translation initiation and elongation.

Translation elongation requires binding of a complementary tRNA to the mRNA. Codons that can be translated by more abundant tRNAs will on average be translated more quickly. The tAI model provides a value for each gene, quantifying its adaptation to the tRNA pool of an organism by providing a weight for each codon and taking the geometric mean over all codons in the sequence. The weight for each codon is proportional to the sum over the tRNA gene count of all matching tRNAs, accounting for wobble interactions, times their binding affinity to the codon ("s-value") [35]. In our study, tRNA gene counts for *N. tabacum* were taken from GtRNAdb [44]. We included tRNA genes that were tagged as "tertiary filtered" in GtRNAdb, since we found this filter to be misconfigured for plants. Tobacco-specific s-values were calculated using stAIcalc [45]. Additionally, we introduce a new parameter, called tAI_ramp, which we define as tAI calculated over the first 50 codons of the sequence divided by tAI of the entire sequence. The rationale behind this parameter is to measure the presence of a translational ramp, *i.e.*, preference for slow codons at the start of the sequence, a pattern observed in nature for increasing the efficiency of protein production [46].

A variation of tAI, nTE [37], additionally incorporates the demand for each tRNA into the weight for each codon, thereby giving codons whose complementary tRNAs are in high demand a lower weight. We used wildtype samples from GEO acc. GSE236277 [47] as mRNA levels and consequent endogenous tRNA demand. Accounting for tRNA demand of the heterologous gene requires making assumptions about the ratio between endogenous and heterologous tRNA demand, which is unknown. We therefore evaluate nTE for different ratios: 0.1, 0.2, 0.5, 1.0, 2.0 and 5.0, the resulting nTE values are referred to as nTE_0.1, nTE_0.2 etc. The definition of nTE used in this work differs from Ref. [37] in that we did not normalize nTE codon weights by the maximum nTE across codons as scales otherwise cannot be compared between heterologous sequences. For example, if a specific codon was rarely used in a heterologous sequence, this codon was assigned high nTE and normalization by its nTE would cause low nTE for the sequence, regardless of its remaining codon composition.

Folding of the mRNA impacts gene expression both through translation initiation and elongation [22–24]. Mean RNA folding energies were predicted locally over 40 nucleotide windows using ViennaRNA [48] in three regions: around the start codon, where we used the last 10 nucleotides of the 3′-end of the UTR of the respective expression vector and the first 30 nucleotides of the coding sequence, in the "ramp"-region from 30 to 150 nucleotides and the remaining sequence.

2.5. Evaluation of the impact of individual codons and amino acids

We additionally aimed to investigate the effect of individual codons and amino acids on expression. To investigate the impact of individual codons, we calculated a weight for each codon and coding sequence as the frequency of the codon in the coding sequence relative to the most

frequently used synonymous codon. This is equivalent to the codon weights used in the CAI model calculated over the single gene on interest. To investigate the effect of each amino acid, we calculated the frequency of each amino acid in the recombinant protein.

We evaluated features based on their Spearman correlation with protein yield. Additionally, to identify correlations which could not be explained by the previously mentioned models, we calculated partial Spearman correlations using the definition provided in Ref. [49].

2.6. A gradient boosting regressor is used to make predictions based on multiple features

To combine all features shown in Table 2 into a prediction we trained a CatBoost regressor [50], a Gradient boosting (GB) tree model. GB is a machine learning technique that builds a strong model by sequentially adding weak learners where each new model corrects the errors of the previous ones by minimizing a chosen loss function. Gradient boosting tree models have been successfully applied to a wide range of domains [51].

We used the default hyperparameter values. Data was partitioned 100 times randomly into test and training set with 0.25/0.75 split and performance was measured as Spearman correlation between prediction and yield on the test data. To determine feature importances, we used the built-in feature importance calculation of CatBoost.

3. Results

3.1. Many studies have been published without sharing sequence data

We have found 441 articles and theses, describing the transient expression of 625 coding sequences in tobacco (Fig. 2a). Of these sequences, 105 could not be used for gene expression modeling since no yield was published in a standardized format (mg of protein per gfw or % TSP). A further 99 sequences were excluded as they were co-expressed with other sequences to form multimeric complexes. The most common reason for the exclusion of studies was that coding sequences were mentioned to be codon optimized, but not provided in the study. While the sequence optimization algorithm was frequently specified, reliably reconstructing the resulting optimized sequence was not possible because a) some sequence optimization algorithms use randomization, leading to a generation of a different optimized sequence on resubmission of the amino acid sequence [31], b) optimization parameters was omitted or c) the optimization tool was no longer available.

The most used vectors were magnICON (84 sequences, average yield of 1.6 mg/gfw), pEAQ (27 sequences, 0.53 mg/gfw), pTRA (20 sequences, 0.55 mg/gfw) and pBY (12 sequences, 0.6 mg/gfw), see Fig. 2b. Common donor organisms were human, human viruses and bacteria (Supplementary Fig. S2). All examined studies and their sequences are listed in Supplementary Table S1.

3.2. Viral capsid genes predict expression levels better than use of tobacco codons

Gene expression models as listed in Table 2 were evaluated according to their Spearman correlation with recombinant protein yield. To adjust for the varying yields between vectors, we normalized yields by the average for their vector and only used the most used vectors (magnICON, pEAQ, pTRA and pBY). Unless otherwise specified, all results shown use data for all four vectors. Results for each of these vectors separately are given in Supplementary Fig. S4.

Correlations between CAIs calculated using tobacco reference genes and yield were moderate at approximately 0.2 (Fig. 3), considering that CAI-based sequence optimization with endogenous reference genes is the most used sequence-optimization strategy in the dataset. We did find higher correlations to heterologous expression when using cARS models, indicating that tobacco genes include sequence features conducive to high expression that extend over multiple codons, and therefore cannot be detected by CAI.

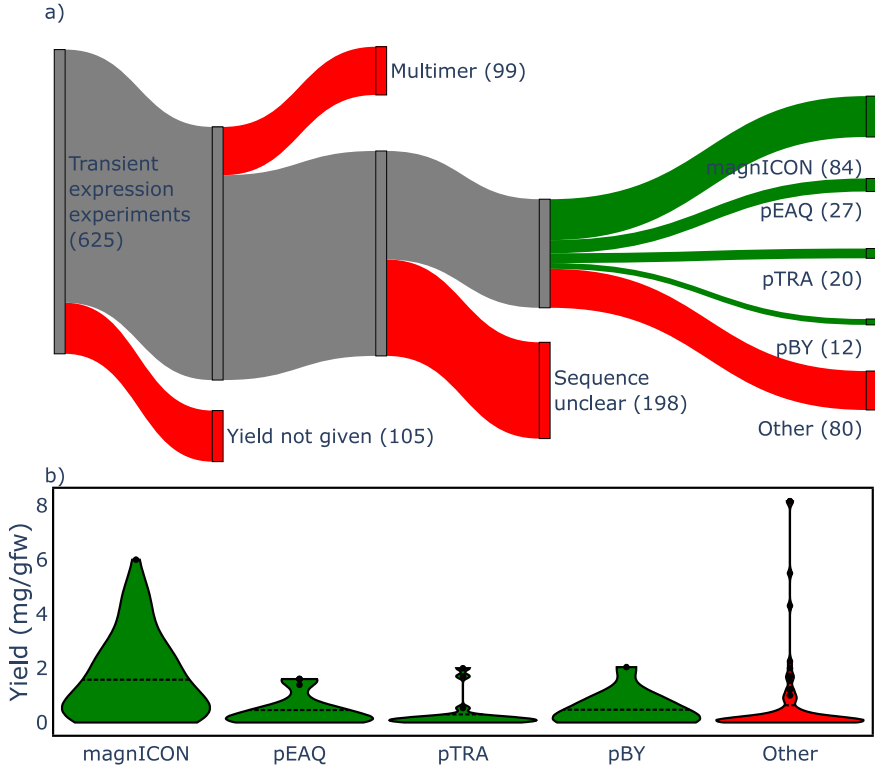

Fig. 2. Top: Examined expression studies and reason for exclusion. For selected studies, the expression vector is shown. Bottom: distribution of yields for the most used expression vectors. Horizontal lines indicate the mean of the distributions.

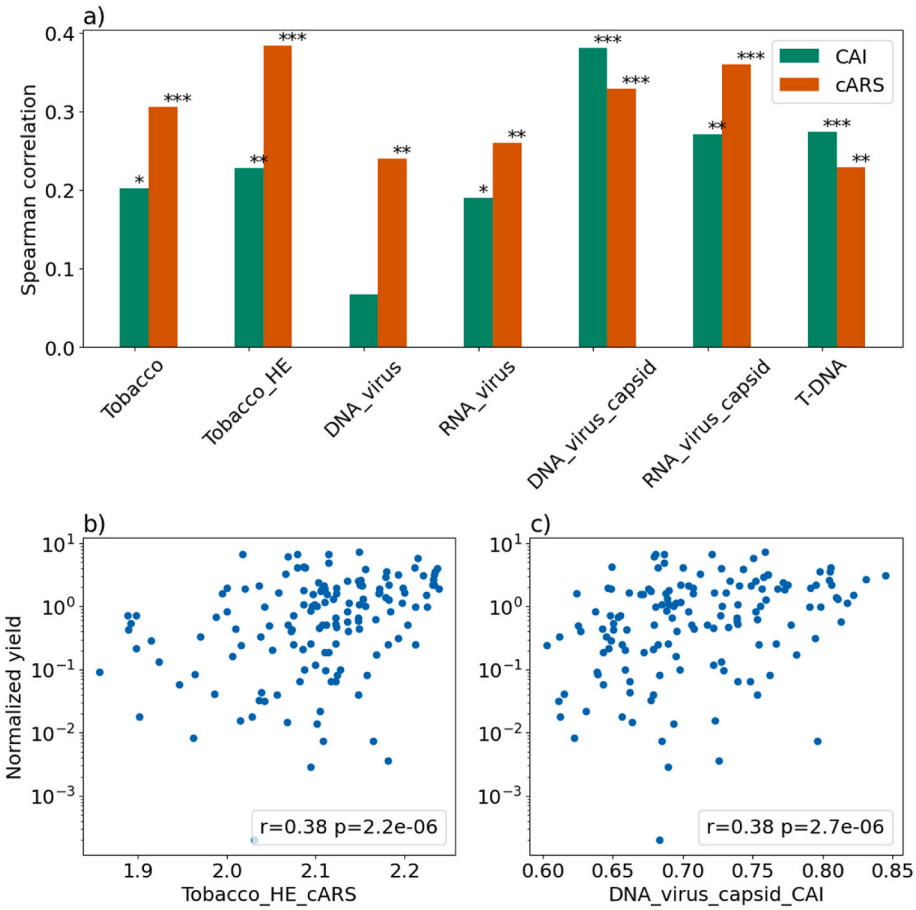

Fig. 3. a): Spearman correlations between protein yields and CAI and cARS gene expression models based on different reference sequence sets. *, ** and *** indicate statistical significance at the 5 %, 1 % and 0.1 % level, respectively. Tobacco: all tobacco genes, Tobacco_HE: top 100 most highly expressed tobacco proteins, DNA/RNA_virus, all genes excluding capsids of tobacco DNA/RNA viruses DNA/RNA_virus_capsid: capsids tobacco of DNA/RNA viruses, respectively, T-DNA: genes from *Agrobacterium* T-DNA b) and c): Scatter plots between yield and the two highest-scoring features, cARS based on highly expressed tobacco genes and CAI based on capsid genes of DNA viruses.

Gene expression models trained on viral genes were most predictive when selecting capsid genes, supporting the hypothesis that these genes in particular are selected for high translational efficiency in tobacco. This is observed for both DNA and RNA viruses. We also show that T-DNA genes contain codons that are preferred for high heterologous expression. We hereby show that learning preferred sequence features from organisms that use the host similarly to the synthetic system is a viable alternative to learning from the host.

3.3. Inclusion of a translational ramp enhances expression

The tAI gene expression model was a significant predictor of recombinant protein yield (Fig. 4). Additionally, we found that sequences with a low tAI_ramp value, *i.e.*, that used codons with low translational efficiency at the start of the sequence compared to the rest of the sequence, were significantly more highly expressed. As previously discussed in Refs. [24,46], codons with low tRNA supply are expected to be translated more slowly and can be selected at the start of coding sequences to improve global translational efficiency by avoiding ribosomal queuing at later parts of the sequence. We expect that this effect to be particularly relevant at high transcript levels, as achieved by self-replicating vectors, as gene expression is more likely to be rate-limited by the number of productive ribosomes. As an additional determinant of translation speed we evaluated mRNA folding strength in the ramp region (nucleotides 30–150), but did not find a significant correlation with yield. mRNA folding around the start codon and the remaining parts of the sequence were also not significantly correlated with yield.

Accounting for the tRNA demand of the heterologous sequence, as done in the nTE model, improved correlations compared tAI model, suggesting that the production of the heterologous protein consumes significant cellular resources and impacts the tRNA pool. Spearman correlation with yield was maximal when setting heterologous tRNA demand to >2 times endogenous demand.

3.4. Expression modeling for replicating and non-replicating vectors is similar, but not identical

We considered whether relevant sequence features may differ between expression vectors due to the differing steps of expression in each system as shown in Fig. 1b. Due to the small number of datapoints for each vector other than magnICON, for this we grouped together sequences for the two non-replicating vectors, pTRA and pEAQ, which we assume to behave similarly. We then calculated and compared Spearman correlations between protein yield and sequence features for sequences expressed using magnICON and non-replicating vectors (Fig. 5).

Several models were significantly correlated with yield for both expression systems: DNA_virus_capsid_CAI, T-DNA_CAI, Tobacco_HE_cARS and nTE. For all features shown in Table 2, the Spearman correlation between Spearman correlations for magnICON and non-replicating was 0.65 ($p < 0.01$). CDS Features, for which the

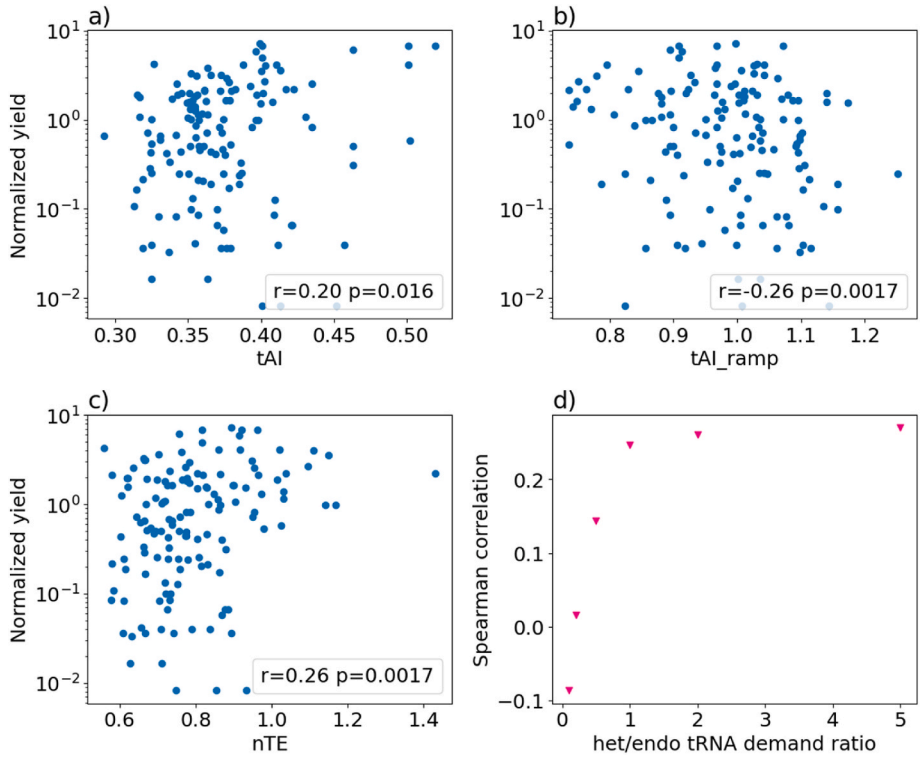

Fig. 4. Yield vs tAI (a), tAI_ramp (b) and Yield vs nTE (c), for a heterologous/endogenous tRNA demand ratio of 2. d): Spearman correlations between yield and nTE evaluated using different ratios between heterologous and endogenous tRNA demand Spearman correlation.

difference in predictiveness between the two systems was largest were cARS based on viral capsids and tAI_ramp, which were both more predictive in magnICON. Thus, the two types of vectors share common relevant features; however, since less than half ($r^2 < 0.5$) of the variance of features ranking by one system can be explained by the second one, there likely exist sequence aspects which are different.

3.5. Usage of some amino acids, cysteine in particular, significantly inhibits yield

To further understand the impact of individual codons and amino acids on expression, we evaluated the predictiveness of codon and amino acid usage for each codon and amino acid separately. We did this by calculating codon weights for each heterologous gene sequence separately and correlating each codon weight to the protein abundance expressed by the sequence. We also evaluated correlations between the frequency of each amino acid in each sequence and yield. To find correlations that were not already explained by the previously discussed features, we calculated partial Spearman correlations, controlling for DNA_virus_capsid_CAI, tAI_ramp, Tobacco_HE_cARS (Table 3). While many codon weights are significantly correlated with expression, partial correlations for all but one codon (GCT) are insignificant, suggesting that these correlations can be explained by the gene expression models explained above.

Additionally, we find that the amino acid composition of the recombinant protein strongly impacts its yield. We find that high usage of lysine was associated with high expression, while arginine and cysteine is deleterious for expression, also after controlling for the above-mentioned features. With a Spearman correlation of −0.49, the frequency of cysteine is the most strongly correlated feature with yield in this study.

3.6. A model combining multiple features improves yield prediction

To combine multiple features into a single prediction we trained CatBoost regressors, one including all features defined in Table 2 and one additionally including amino acid frequencies due to their significant impact on protein yield. Over 100 test-train splits, the models achieved average test set Spearman correlations of 0.41 and 0.57 with and without amino acid frequencies, respectively (Fig. 6a and b. A scatter plot showing yield predictions for a representative (*i.e.*, one that achieves Spearman correlation close to the average) test-train split in Supplementary Fig. S5.

Many features that were highly ranked by CatBoost (Fig. 6c and d were features previously discussed for their significant correlation with yield, for example tAI_ramp, cARS based on viral capsids, and the frequency of cysteine and lysine. Interestingly, folding-related features scored relatively highly, despite their insignificant correlation with yield (Fig. 6); this suggests that these features are important but the relation is not a simple monotonic one. Also, CatBoost assigned higher importance to Tobacco_cARS than Tobacco_HE_cARS, *i.e.*, cARS calculated using all tobacco genes as reference genes instead of only the 100 most highly expressed ones, despite the latter one being more correlated with yield, as shown in Fig. 3. Therefore, there might also be important information included in less highly expressed tobacco genes.

4. Discussion

While there are previous studies about the effect of sequence features on heterologous protein expression (e.g., Refs. [20–23]), most of them focus on microorganisms and none of them studied plant vectors. Here, we compiled a dataset of coding sequences transiently expressed in tobacco and their expression yield. We found several features that were significant predictors of yield. Central findings of our analysis are that engineered sequences should.

- imitate the codon usage of viral capsid genes
- include long substrings that appear in highly expressed tobacco genes
- include a translational ramp

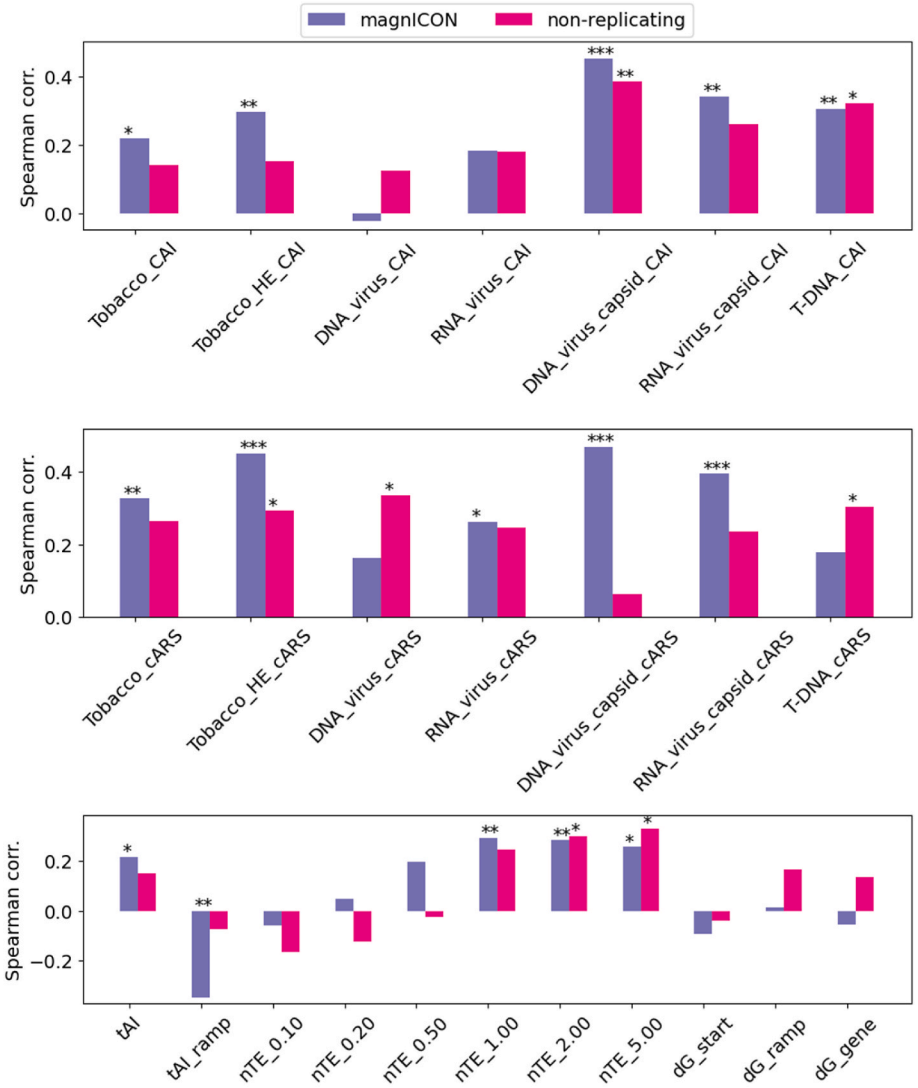

Fig. 5. Spearman correlations between features and yield for sequences expressed in magnICON vs non-replicating (pEAQ/pTRA) vectors. First row: CAI features, second row: cARS features, third row: biophysical features.

Combining all features gave a Spearman correlation of 0.57. We note that correlations obtained in this work are likely lower bounds for real correlations due to the significant noise associated with the combination of studies with varying experimental conditions.

By predicting protein yield, we model the combined effects of the coding sequence on all steps of gene expression, including transcription, RNA replication in the case of replicating vectors, translation, mRNA stability and co-translational folding of the protein and the resulting effects on its stability. Some of the applied models are designed to describe one specific step (e.g., tAI/nTE describing translation elongation), while models based on reference genes (CAI and cARS) can capture multiple steps if there is selection for them in the reference gene set. While tAI and nTE were significantly predictive of yield (Fig. 4), highlighting the importance of optimization for translation elongation, they were outperformed by cARS (Fig. 3). This suggests that other steps of gene expression are significantly affected by patterns in the coding sequence which extend over more than one codon. Further understanding the effects at each step would require additional data, for example transcript levels and ribosomal occupancies, which is available for few and none of the examined studies, respectively.

Regarding translational ramps, it is important to mention that the relationship between tAI_ramp and protein yield likely is non-linear and correlation with yield should become positive for very small tAI_ramp values where extremely low elongation rates in the ramp region become the bottleneck of heterologous protein production. We did not observe such non-linearities within the dataset, where most tAI_ramp values fell between 0.8 and 1.2 (Fig. 4), *i.e.*, the translational efficiency in the ramp region was 80–120 % of the remaining sequence. This suggests that the relationship reverses at some tAI_ramp value at or smaller than 0.8.

In the examined studies, optimized sequences were predominantly designed using commercial codon optimization algorithms. While their exact implementations are not public, to our knowledge these algorithms optimize single-codon-objectives and therefore likely miss higher-dimensional information encoded in longer substrings that are captured by the Chimera models. We are also not aware of these algorithms optimizing global translation dynamics through the inclusion of translational ramps. Therefore, we expect that significant improvement over existing sequences can be achieved through optimization of the objectives above.

The number of datapoints in this study was severely limited by unpublished sequences. Given the mixed results of existing sequence optimization efforts as shown in Table 1 and implied lack of understanding of the characteristics of an optimized coding sequence, we recommend that researchers share their optimized sequences. Additionally, sharing coding sequences is needed to ensure reproducibility.

In our study we combined data of RNA-virus-based, DNA-virus-based

Table 3
Spearman correlations of yield with codon weights and amino acid frequencies. Columns: aa, Amino acid; preferred_codon, codon encoding the respective amino acid whose weight has the highest spearman correlation to yield, r_codon/partial_r_codon, (partial) Spearman correlation between yield and the weight of the preferred codon; r_aa/partial_r_aa, (partial) Spearman correlation between yield and the frequency of the amino acid. *, ** and *** indicate statistical significance at the 5 %, 1 % and 0.1 % level, respectively.

aa	preferred_codon	r_codon	partial_r_codon	r_aa	partial_r_aa
A	GCT	0.23**	−0.25**	0.07	−0.15
C	TGT	0.05	0.09	−0.49***	−0.33***
D	GAT	0.22**	−0.12	0.22**	0.1
E	GAG	0.17*	0.03	0.19*	0.13
F	TTC	0.1	0.01	−0.02	0.04
G	GGT	0.24**	−0.04	0.15	0.01
H	CAC	0.19*	0.14	0.04	0.17*
I	ATC	0.26**	0.14	0.01	−0.04
K	AAG	0.19*	−0.02	0.29***	0.17*
L	CTT	0.2*	−0.09	−0.18*	0.02
M	ATG	n/a	n/a	−0.05	0.06
N	AAC	0.15	0.01	0.06	−0.18*
P	CCT	0.33***	0.05	−0.01	0.04
Q	CAG	0.22**	0.17	−0.14	−0.2*
R	AGA	0.11	0.09	−0.35***	−0.19*
S	TCC	0.14	0.15	−0.1	−0.05
T	ACC	0.15	0.08	−0.09	−0.01
V	GTG	0.22**	0.06	0.22**	0.14
W	TGG	n/a	n/a	−0.22**	−0.14
Y	TAC	0.14	0.03	−0.06	−0.08

and non-replicating vectors, however, in principle it is possible that the differing steps of expression for each system (Fig. 1b) manifest in varying preferred sequences features. Due to the small number of sequences for many vectors we were only able to compare magnICON and non-replicating vectors and overall found similar preference for many sequence features in both systems (Fig. 5). Interestingly, features describing the similarity of the CDS to viral capsids were more predictive in the RNA virus-based magnICON system, possibly due to selection of these sequences for conditions that are similar in the synthetic system. Additionally, the inclusion of a translational ramp was conducive of high expression in magnICON, but not in non-replicating vectors. A possible explanation for this could be that the high transcript levels produced by replicating vectors starve the free ribosome pool, which makes efficient ribosome allocation as achieved by the avoidance of ribosomal queuing through translational ramps more important. It would be interesting to make a similar comparison for DNA virus-based vectors like pBY, which is currently not feasible due to the small number of published sequences for this system.

Finally, we found that strong predictors of the yield of the recombinant protein were related to its amino acid composition. Frequent arginine and cysteine residues were associated with low expression. A possible explanation of the negative effect of arginine-usage on yield could lie in its positive charge, causing translational pausing through interactions with the ribosomal exit tunnel [52]. The same effect, however, is not observed for lysine (also positively charged) possibly since on average the adaptation of lysine codons to the tRNA pool is higher than for arginine – tAI weights for codons encoding lysine are 0.53–0.69 compared to 0.15–0.29 for arginine (Supplementary Fig. S3). The large negative impact of cysteine residues hints at incorrect disulfide bridge formation and resulting misfolding and degradation. Additional yield improvements could therefore come through replacement of certain amino acids residues but would require validation of proper structure and function of the engineered protein, as done in e.g., Ref. [53].

CRediT authorship contribution statement

Moritz Burghardt: Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Data curation, Conceptualization. **Tamir Tuller:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Methodology, Investigation, Funding acquisition, Data curation, Conceptualization.

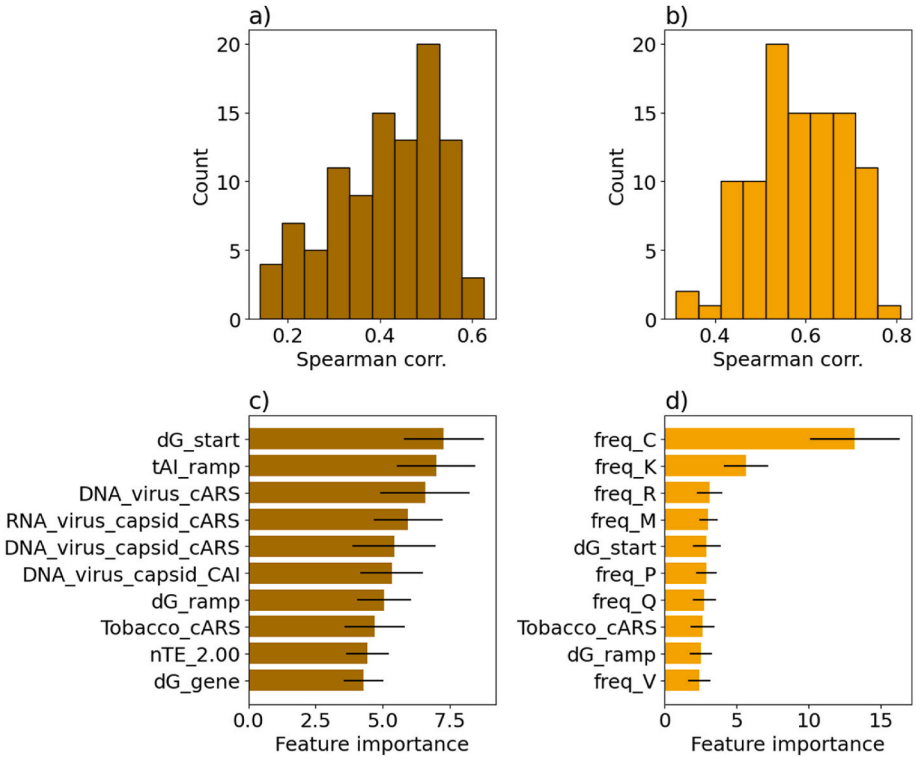

Fig. 6. a) and b): Distribution of Spearman correlations between CatBoost predictions and yield on the test over 100 test-train splits for models with and without amino acid frequencies, respectively. c) and d): average feature importances. Error bars represent standard deviations over the different test-train splits.

Declaration of competing interest

The authors do not have any conflict of interest.

Acknowledgements

This study was supported in part by a fellowship from the Edmond J. Safra Center for Bioinformatics at Tel-Aviv University and the Israeli cultivated meat consortium.

We thank all researchers who shared their coding sequences. We also thank Shimshi Atar for helpful discussions.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.synbio.2024.12.002.

References

[1] Daniell H, Streatfield SJ, Wycoff K. Medical molecular farming: production of antibodies, biopharmaceuticals and edible vaccines in plants. Trends Plant Sci 2001;6(5):219–26.
[2] Burnett MJB, Burnett AC. Therapeutic recombinant protein production in plants: challenges and opportunities. Plants, People, Planet 2020;2(2):121–32.
[3] Shanmugaraj B, Bulaon CJI, Phoolcharoen W. Plant molecular farming: a viable platform for recombinant biopharmaceutical production. Plants 2020;9(7):842. 2020, Vol. 9, Page 842.
[4] Fischer R, Buyel JF. Molecular farming – the slope of enlightenment. Biotechnol Adv 2020;40:107519.
[5] Peyret H, Lomonossoff GP. When plant virology met Agrobacterium: the rise of the deconstructed clones. Plant Biotechnol J 2015;13(8):1121–35.
[6] Marillonnet S, Thoeringer C, Kandzia R, Klimyuk V, Gleba Y. Systemic Agrobacterium tumefaciens–mediated transfection of viral replicons for efficient transient expression in plants. Nat Biotechnol 2005;23(6):718–23.
[7] Lindbo JA. TRBO: a high-efficiency tobacco mosaic virus RNA-based overexpression vector. Plant Physiol 2007;145(4):1232–40.
[8] Mardanova ES, Blokhina EA, Tsybalova LM, Peyret H, Lomonossoff GP, Ravin NV. Efficient transient expression of recombinant proteins in plants by the novel pEff vector based on the genome of potato virus X. Front Plant Sci 2017;8:247.
[9] Chen Q, He J, Phoolcharoen W, Mason HS. Geminiviral vectors based on bean yellow dwarf virus for production of vaccine antigens and monoclonal antibodies in plants. Hum Vaccine 2011;7(3):331–8.
[10] Sainsbury F, Thuenemann EC, Lomonossoff GP. pEAQ: versatile expression vectors for easy and quick transient expression of heterologous proteins in plants. Plant Biotechnol J 2009;7(7):682–93.
[11] Maclean J, Koekemoer M, Olivier AJ, Stewart D, Hitzeroth II, Rademacher T, Fischer R, Williamson A-L, Rybicki EP. Optimization of human papillomavirus type 16 (HPV-16) L1 expression in plants: comparison of the suitability of different HPV-16 L1 gene variants and different cell-compartment localization. J Gen Virol 2007;88(5):1460–9.
[12] Pegoraro M, Matić S, Pergolizzi B, Iannarelli L, Rossi AM, Morra M, Noris E. Cloning and expression analysis of human amelogenin in Nicotiana benthamiana plants by means of a transient expression system. Mol Biotechnol 2017;59:425–34.
[13] Zhu H, Reynolds LB, Menassa R. A hyper-thermostable α-amylase from Pyrococcus furiosus accumulates in Nicotiana tabacum as functional aggregates. BMC Biotechnol 2017;17:1–11.
[14] Maema OG, others. Expression feasibility of recombinant enterokinase in nicotiana benthamiana. 2014.
[15] Mbewana S, Mortimer E, Pèra FFPG, Hitzeroth II, Rybicki EP. Production of H5N1 influenza virus matrix protein 2 ectodomain protein bodies in tobacco plants and in insect cells as a candidate universal influenza vaccine. Front Bioeng Biotechnol 2015;3:197.
[16] Dickey A, Wang N, Cooper E, Tull L, Breedlove D, Mason H, Liu D, Wang KY, others. Transient expression of lumbrokinase (PI239) in tobacco (Nicotiana tabacum) using a geminivirus-based single replicon system dissolves fibrin and blood clots. Evid base Compl Alternative Med 2017;2017.
[17] Jiang M-C, Hu C-C, Lin N-S, Hsu Y-H. Production of human IFNγ protein in Nicotiana benthamiana plant through an enhanced expression system based on bamboo mosaic virus. Viruses 2019;11(6):509.
[18] Ha J-H, Kim H-N, Moon K-B, Jeon J-H, Jung D-H, Kim S-J, Mason HS, Shin S-Y, Kim H-S, Park K-M. Recombinant human acidic fibroblast growth factor (aFGF) expressed in Nicotiana benthamiana potentially inhibits skin photoaging. Planta Med 2017;83(10):862–9.
[19] Sun, L., Kallolimath, S., Palt, R., Eminger, F., Strasser, R., and Steinkellner, H. Codon optimization regulates IgG3 and IgM expression and glycosylation in N. benthamiana. Front Bioeng Biotechnol, 11, 1320586.
[20] Welch M, Govindarajan S, Ness JE, Villalobos A, Gurney A, Minshull J, Gustafsson C. Design parameters to control synthetic gene expression in Escherichia coli. PLoS One 2009;4(9):e7002.
[21] Burgess-Brown NA, Sharma S, Sobott F, Loenarz C, Oppermann U, Gileadi O. Codon optimization can improve expression of human genes in Escherichia coli: a multi-gene study. Protein Expr Purif 2008;59(1):94–102.
[22] Kudla G, Murray AW, Tollervey D, Plotkin JB. Coding-sequence determinants of gene expression in Escherichia coli. Science 2009;324(5924):255–8. 1979.
[23] Ben-Yehezkel T, Atar S, Zur H, Diament A, Goz E, Marx T, Cohen R, Dana A, Feldman A, Shapiro E, Tuller T. Rationally designed, heterologous S. cerevisiae transcripts expose novel expression determinants. RNA Biol 2015;12(9):972–84.
[24] Tuller T, Waldman YY, Kupiec M, Ruppin E. Translation efficiency is determined by both codon bias and folding energy. Proc Natl Acad Sci USA 2010;107(8):3645–50.
[25] Li GW, Oh E, Weissman JS. The anti-Shine–Dalgarno sequence drives translational pausing and codon choice in bacteria. Nature 2012;484(7395):538–41.
[26] Bergman S, Tuller T. Widespread non-modular overlapping codes in the coding regions. Phys Biol 2020;17(3).
[27] Grote A, Hiller K, Scheer M, Münch R, Nörtemann B, Hempel DC, Jahn D. JCat: a novel tool to adapt codon usage of a target gene to its potential expression host. Nucleic Acids Res 2005;33:W526–31.
[28] Diament A, Weiner I, Shahar N, Landman S, Feldman Y, Atar S, Avitan M, Schweitzer S, Yacoby I, Tuller T. ChimeraUGEM: unsupervised gene expression modeling in any given organism. Bioinformatics 2019;35(18):3365–71.
[29] Zulkower V, Rosser S. DNA Chisel, a versatile sequence optimizer. Bioinformatics 2020;36(17):4508–9.
[30] Taneda A, Asai K. COSMO: a dynamic programming algorithm for multicriteria codon optimization. Comput Struct Biotechnol J 2020;18:1811–8.
[31] Ranaghan MJ, Li JJ, Laprise DM, Garvie CW. Assessing optimal: inequalities in codon optimization algorithms. BMC Biol 2021;19(1):1–13.
[32] Webster GR, Teh AY-H, Ma JK-C. Synthetic gene design—the rationale for codon optimization and implications for molecular pharming in plants. Biotechnol Bioeng 2017;114(3):492–502.
[33] Pang EL, Peyret H, Ramirez A, Loh H-S, Lai K-S, Fang C-M, Rosenberg WM, Lomonossoff GP. Epitope presentation of dengue viral envelope glycoprotein domain III on hepatitis B core protein virus-like particles produced in Nicotiana benthamiana. Front Plant Sci 2019;10:455.
[34] Sharp PM, Li WH. The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res 1987;15(3):1281–95.
[35] Reis M dos, Savva R, Wernisch L. Solving the riddle of codon usage preferences: a test for translational selection. Nucleic Acids Res 2004;32(17):5036–44.
[36] Tuller T, Zur H. Multiple roles of the coding sequence 5′ end in gene expression regulation. Nucleic Acids Res 2015;43(1):13.
[37] Pechmann S, Frydman J. Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. Nat Struct Mol Biol 2012;20(2): 237–43. 2012 20:2.
[38] Bahiri-Elitzur S, Tuller T. Codon-based indices for modeling gene expression and transcript evolution. Comput Struct Biotechnol J 2021;19:2646.
[39] Huang Q, Szklarczyk D, Wang M, Simonovic M, von Mering C. PaxDb 5.0: Curated Protein Quantification Data Suggests Adaptive Proteome Changes in Yeasts. Mol Cell Proteom 2023;2(10):10064.
[40] Breuza L, Poux S, Estreicher A, Famiglietti ML, Magrane M, Tognolli M, Bridge A, Baratin D, Redaschi N, Consortium U. The UniProtKB guide to the human proteome. Database 2016;2016:bav120.
[41] Dawson WO, Lehto KM. Regulation of tobamovirus gene expression. Adv Virus Res 1990;38:307–42.
[42] Mihara T, Nishimura Y, Shimizu Y, Nishiyama H, Yoshikawa G, Uehara H, Hingamp P, Goto S, Ogata H. Linking virus genomes with host taxonomy. Viruses 2016;8(3):66.
[43] Jouanin L, Bouchez D, Drong RF, Tepfer D, Slightom JL. Analysis of TR-DNA/plant junctions in the genome of a Convolvulus arvensis clone transformed by Agrobacterium rhizogenes strain A4. Plant Mol Biol 1989;12:75–85.
[44] Chan PP, Lowe TM. GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes. Nucleic Acids Res 2016;44(D1):D184–9.
[45] Sabi R, Volvovitch Daniel R, Tuller T. stAIcalc: tRNA adaptation index calculator based on species-specific weights. Bioinformatics 2017;33(4):589–91.
[46] Tuller T, Carmi A, Vestsigian K, Navon S, Dorfan Y, Zaborske J, Pan T, Dahan O, Furman I, Pilpel Y. An evolutionarily conserved mechanism for controlling the efficiency of protein translation. Cell 2010;141(2):344–54.
[47] Shi R, Kernodle SP, Steede TM, Lewis RS. Modified physiology of burley tobacco plants genetically engineered to express Yb1, a functional EGY enzyme. Planta 2023;258(4):1–11.
[48] Lorenz R, Bernhart SH, Höner zu Siederdissen C, Tafer H, Flamm C, Stadler PF, Hofacker IL. ViennaRNA package 2.0. Algorithm Mol Biol 2011;6(1):1–14.
[49] Kim S. Ppcor: an R package for a fast calculation to semi-partial correlation coefficients. Commun Stat Appl Methods 2015;22(6):665.
[50] Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features. Adv Neural Inf Process Syst 2018;31.
[51] Bentéjac C, Csörgő A, Martínez-Muñoz G. A comparative analysis of gradient boosting algorithms. Artif Intell Rev 2021;54(3):1937–67.
[52] Lu J, Deutsch C. Electrostatics in the ribosomal tunnel modulate chain elongation rates. J Mol Biol 2008;384(1):73–86.
[53] Matoba N, Kajiura H, Petruccelli S, Strasser R, Y-j S, Shin Y-J, König-Beihammer J, Vavra U, Schwestka J, Kienzl NF, Klausberger M, Laurent E, Grünwald-Gruber C, Vierlinger K, Hofner M, Margolin E, Weinhäusel A, Stöger E, Mach L. N-glycosylation of the SARS-CoV-2 receptor binding domain is important for functional expression in plants. Front Plant Sci 2021;12:689104.