Microbial Cell Factories

RESEARCH Open Access

Characterization and optimization of 5´ untranslated region containing poly-adenine tracts in *Kluyveromyces marxianus* using machine-learning model

Junyuan Zeng[1,2], Kunfeng Song[1,2], Jingqi Wang[1,2], Haimei Wen[1,2], Jungang Zhou[1,2], Ting Ni[1,2], Hong Lu[1,2] and Yao Yu[1,2*]

Abstract

Background The 5´ untranslated region (5´ UTR) plays a key role in regulating translation efficiency and mRNA stability, making it a favored target in genetic engineering and synthetic biology. A common feature found in the 5´ UTR is the poly-adenine (poly(A)) tract. However, the effect of 5´ UTR poly(A) on protein production remains controversial. Machine-learning models are powerful tools for explaining the complex contributions of features, but models incorporating features of 5´ UTR poly(A) are currently lacking. Thus, our goal is to construct such a model, using natural 5´ UTRs from *Kluyveromyces marxianus*, a promising cell factory for producing heterologous proteins.

Results We constructed a mini-library consisting of 207 5´ UTRs harboring poly(A) and 34 5´ UTRs without poly(A) from *K. marxianus*. The effects of each 5´ UTR on the production of a GFP reporter were evaluated individually in vivo, and the resulting protein abundance spanned an approximately 450-fold range throughout. The data were used to train a multi-layer perceptron neural network (MLP-NN) model that incorporated the length and position of poly(A) as features. The model exhibited good performance in predicting protein abundance (average $R^2=0.7290$). The model suggests that the length of poly(A) is negatively correlated with protein production, whereas poly(A) located between 10 and 30 nt upstream of the start codon (AUG) exhibits a weak positive effect on protein abundance. Using the model as guidance, the deletion or reduction of poly(A) upstream of 30 nt preceding AUG tended to improve the production of GFP and a feruloyl esterase. Deletions of poly(A) showed inconsistent effects on mRNA levels, suggesting that poly(A) represses protein production either with or without reducing mRNA levels.

Conclusion The effects of poly(A) on protein production depend on its length and position. Integrating poly(A) features into machine-learning models improves simulation accuracy. Deleting or reducing poly(A) upstream of 30 nt preceding AUG tends to enhance protein production. This optimization strategy can be applied to enhance the yield of *K. marxianus* and other microbial cell factories.

Keywords 5´ UTR, Poly(A), *Kluyveromyces marixanus*, Machine-learning model, Heterologous protein expression

*Correspondence:
Yao Yu
yaoyu@fudan.edu.cn

[1] State Key Laboratory of Genetic Engineering, School of Life Sciences, Fudan University, Shanghai, China
[2] Shanghai Engineering Research Center of Industrial Microorganisms, Shanghai 200438, China

Introduction

The 5´ untranslated region (5´ UTR) is the segment of an mRNA that spans from the 5´ end to the position of the start codon (AUG). The 5´ UTR plays key roles in post-transcriptional regulation without altering the protein sequence, making it a favored target in genetic engineering and synthetic biology [1, 2]. The effects of the 5´ UTR on protein production are mediated by the multiple cis-regulatory elements it carries, including the 5´-cap structure [3], the translation initiation context [4, 5], upstream AUGs and upstream ORFs [6–8], internal ribosome entry sites (IRES) [9, 10], nucleotide preferences at positions immediately upstream of AUG [11, 12], secondary structures [13, 14], and G-quadruplexes [15]. These elements primarily regulate translation efficiency and mRNA stability.

Poly-adenine (poly(A)) tract is a common feature in the 5´ UTR. Approximately 28% of genes in *Saccharomyces cerevisiae*, 48% of genes in *Arabidopsis thaliana*, 39% of genes in *Drosophila melanogaster*, 9% of genes in mice, and 11% of genes in humans contain at least one poly(A) longer than 5 nt in the 5´ UTR [16]. Unlike the poly(A) tail added at the 3´ end of mRNA, the sequence encoding 5´ UTR poly(A) is embedded in the primary gene sequence and transcribed as part of the mature mRNA. The frequency of continuous poly$(A)_n$ occurring is $(1/4)^n$. Therefore, the occurrence of 5´ UTR poly(A) is not due to random base combinations but is linked to the function of the 5´ UTR.

The effect of 5´ UTR poly(A) on gene expression remains controversial in different circumstances. 5´ UTR poly(A) forms IRES for cap-independent translation of invasive growth genes in *S. cerevisiae* [17], genes related to pattern-triggered immunity in *Arabidopsis* plants [18], and *GFP* and *Luc* reporter genes in vitro [19, 20]. In a study utilizing a synthetic mRNA library containing over one million 5´ UTR variants, poly(A) was found to enable cap-independent translation in mammalian cells but destabilize mRNAs in the absence of translation in vitro [21]. The repression of translation by 5´ UTR poly(A) was also observed in the auto-regulation of PABP1 (Poly(A) binding protein 1) in mammalian cells [22], and *S. cerevisiae* [23]. The conflicting roles of 5´ UTR poly(A) might be related to its length. A bioinformatic analysis suggests that poly(A) shorter than 12 nt is correlated with improved translation efficiency, while poly(A) longer than 12 nt results in translation repression [24].

Constructing a machine learning model is a promising approach to explaining the complicated roles of 5´ UTR poly(A) in protein production. Several models of the 5´ UTR were constructed using different machine learning strategies, such as partial least squares (PLS) regression [25, 26], random forest [14, 27], support vector machine (SVM) [27], convolutional neural network (CNN) [6, 7, 28, 29], and transformer [30]. These models were successfully applied to predict the protein production driven by the 5´ UTR and to reveal the contribution of 5´ UTR elements to the protein production. To exclude background interference, the construction of a model requires building a randomly or deliberately designed 5´ UTR library that drives the expression of the same reporter gene. However, a similar analysis has not been performed using natural 5´ UTRs containing poly(A), which poses an obstacle to incorporating poly(A) as a novel feature into the machine learning model.

Kluyveromyces marxianus is an unconventional budding yeast species. Due to its long-standing safe association with human food, such as dairy products, grapes, and papaya, *K. marxianus* has been granted GRAS (Generally Regarded As Safe) and QPS (Qualified Presumption of Safety) status in the United States and Europe, respectively [31]. Besides its safety, *K. marxianus* possesses several features beneficial for industrial applications, including fast growth, thermotolerance, a broad spectrum of utilizable carbon sources, and a high capacity for secretion. Therefore, *K. marxianus* is a promising microbial cell factory for producing heterologous proteins, biofuels, and various chemicals [32, 33]. In *K. marxianus*, the deletion of a poly(A) tract inside the *INU1* 5´ UTR increased downstream protein production [34], while abolishing the poly(A) in the *LAC4* 5´ UTR reduced the leaky expression under glucose repression [35]. These results indicate that 5´ UTR poly(A) plays an important role in regulating protein production in *K. marxianus*.

In this study, we constructed a mini-library comprising 207 natural 5´ UTRs harboring poly(A) along with 34 5´ UTRs without poly(A). All the 5´ UTRs are from *K. marxianus*. The effect of each 5´ UTR on protein production was evaluated separately in vivo using a dual fluorescent reporter system. The obtained data were used to construct a multi-layer perceptron neural network (MLP-NN) model, in which poly(A) features were incorporated. The model demonstrated good performance in predicting protein abundance. The model suggests that the length of poly(A) is generally negatively correlated with protein production, while poly(A)s with a distance to AUG between 10 ~ 30 nt exhibit a weak correlation with improved protein production. Consistent with the model's predictions, the deletions of poly(A)s upstream of 30 nt preceding AUG tended to enhance protein production, which was validated through the expression analysis of GFP and a feruloyl esterase (AnFaeA). These results suggest that incorporating poly(A) features into machine-learning models can enhance the accuracy of the prediction. The optimized strategy of the 5´ UTR proposed here could be applied to improve the yield of *K. marxianus* and other microbial cell factories.

Materials and methods

Strains and plasmids

K. marxianus strain, Fim-1ΔU [34], was used as a wild-type strain in this study. All plasmids used in this study are listed in Additional file 2: Table S1. All the primers used are listed in Additional file 2: Table S2. *HXT4* promoter (1351 bp) and 5´ UTR (149 bp) were inserted into an *Xho* I site immediately preceding the open reading frame (ORF) of *GFP* in LHZ676 [36], to obtain LHZ1137. The *HXT4* 5´ UTR of LHZ1137 was replaced by an *Xho* I site to obtain LHZ1138. Different natural 5´ UTRs were amplified using primers OZJY1F/R ~ OZJY240F/R and then inserted into the *Xho* I site of LHZ1138 by Gibson-assembly, resulting in the generation of LHZ1139 ~ LHZ1378. Mutations within 5´ UTRs were introduced by mutagenesis PCR using primers OZJY241F/R to OZJY302F/R, resulting in the generation of LHZ1379 ~ LHZ1440. The cassette of P_{INU1}-SS_{INU1}-*Est1E*-His_6-T_{INU1} of pZP32 [34], was replaced by a cassette of *HXT4* promoter-*Sma* I-$SS_{INU1-P10L}$- *AnFaeA*-*INU1* terminator. The $SS_{INU1-P10L}$ coding an *INU1* signal peptide with a P10L mutation was amplified from pZP33 [34], and the ORF of *AnFaeA* was amplified from LHZ766 [37]. The resulting plasmid was named LHZ1441. The wild-type and mutant 5´ UTRs of *SSH4*, *INU1* and *KLMA_80280* were amplified from LHZ1182, LHZ1438, LHZ1164, LHZ1419, LHZ1158 and LHZ1439. Amplified 5´ UTRs were inserted into the *Sma* I site of LHZ1441 to obtain LHZ1442 ~ LHZ1447. The ORF of *Est1E* in pZP28 [34], was replaced by the ORF of *AnFaeA* amplified from LHZ733 [37], resulting in LHZ1448. Poly(A) inside *INU1* 5´ UTR in LHZ1448 was mutated by mutagenesis PCR using primers OZJY307F/R ~ OZJY320F/R, resulting in the generation of LHZ1449 ~ LHZ1462. The sequences of three backbone plasmids, LHZ1138, LHZ1441 and LHZ1448, are listed in Additional file 2: Table. S3.

Nanopore RNAseq

Three parallel cultures were grown in YPD (2% w/v glucose, 2% w/v peptone; 1% w/v yeast extract) at 30 degrees for 16 h or 72 h. Cells were collected and total RNA was extracted by using ZR Fungal/Bacterial RNA MiniPrep (Zymo Research, R2014). RNA was subjected to Nanopore sequencing at Biomarker Technologies Inc.(Beijing, China). After low-quality reads (length < 500 bp, Q score < 7) were filtered out, 6.5 ~ 8.5 million clean reads were obtained for each sample. The reads were mapped to the FIM-1 reference genome [38], using minimap2, with more than 95% of the reads successfully mapped for each sample. Redundancy was conducted for each sample by filtering sequences with an identity below 0.9 and coverage below 0.85. Alignments showing differences only at the 5´ end were merged to obtain a gff file for each sample. By comparing these files with the FIM-1 reference gff annotation file using bedtools, additional sequences in transcriptions relative to the 5´ end of corresponding coding sequences (CDS) were extracted as the 5´ UTRs and matched to their corresponding genes. A gene may have multiple 5´ UTRs due to differences in transcription start sites and data processing. We identified the most frequent 5´ UTR(s) for each gene. In cases where multiple 5´ UTRs shared the same highest frequency, we selected the longest one. Transcripts per million (TPM) value was used as a measure of gene expression level [39], and was calculated for each gene in each sample. Sequences of 5´ UTRs are listed in Additional file 3 and TPM values of each gene are listed in Additional file 4.

Proteomics analysis

Cells were collected as described in nanopore RNAseq. Cells were resuspended in 8 M guanidine HCl, 100 mM Tris HCl (pH 8.0) and then lysed by glass beads in a FastPrep-24 5G instrument (MP Biomedicals, USA). The supernatant was collected after centrifugation and subjected to the FASP digestion in Microcon PL-10 filters as described previously [40]. Nano LC-MS/MS analysis was performed using an EASY-nLC 1200 system (Thermo Fisher Scientific, USA) coupled to an Orbitrap Fusion Lumos mass spectrometer (Thermo Fisher Scientific, USA). A one-column system was adopted for all analyses. Samples were analyzed on a homemade C18 analytical column (75 µm i.d. × 25 cm, ReproSil-Pur 120 C18-AQ, 1.9 µm (Dr. Maisch GmbH, Germany)) [41]. The results were processed with the UniProt *K. marxianus* protein database (4926 entries, downloaded in 10/27/2020) and using the Mascot (version 2.7.0, Matrix Science) [42]. The mass tolerances were 10 ppm for precursor and fragment mass tolerance of 0.05 Da. Up to two missed cleavages were allowed. The carbamidomethylation on cysteine was set as a fixed modification, and acetylation on the protein N-terminal and oxidation on methionine as variable modifications. The significance threshold was $p < 0.05$, and the minimum number of significant unique sequence was 1. The exponentially modified protein abundance index (emPAI) was found to be approximately proportional to the logarithm of protein concentration [43], and was used for intra-sample comparison of protein abundance [44–48]. The emPAI values were calculated for all proteins and listed in Additional file 5.

Calculation of relative entropy

The relative entropy was used to describe the enrichment and depletion of the base at a specific position [12, 25]. In a given sequence set, the relative entropy of base k at position i is denoted as $E_{set[i,k]}$, which is calculated as follows:

$$E_{set[i,k]} = P_{set[i,k]} \times log_2 \left(\frac{P_{set[i,k]}}{P_{back[i,k]}} \right)$$

$P_{set[i,k]}$ means the probability of the base k at the position i in the sequence set of interest. $P_{back[i,k]}$ means the probability of the base k at the position i in the *background set*, which comprises all other sequences that are not included in the set of interest. In the graph, $E_{set[i,k]}$ is visualized as the height of the logo of base k (A, C, G, or U), where over-representation and under-representation are indicated by positive and negative $E_{set[i,k]}$ values, respectively.

Construction of mini-library and flow cytometry analysis

First, 5´ UTRs with a length of less than 200 nt were selected. Next, 5´ UTRs of genes that ranked in the top 20% based on their values of TPM, emPAI or emPAI/TPM, were selected. Then, 207 5´ UTR containing poly(A) (n≥5) were randomly selected. As controls, 34 5´ UTRs without poly(A) were selected. The length of the longest continuous A tracts (n<5) inside these 34 5´ UTRs was evenly distributed. A total of 241 5´ UTR were selected and cloned into LHZ1138 separately to construct the mini-library containing LHZ1137 and LHZ1139~LHZ1378. Plasmids were transformed into Fim-1ΔU separately and transformants were selected on Synthetic dropout media without uracil (SC-Ura) plates [49]. Transformants were grown in 50 mL SC-Ura liquid medium for 72 h. Cells were washed and resuspended in 50 mM Tris-HCl (pH 7.0). A total of 100,000 cells were subjected to fluorescence-activated cell sorting (FACS) by gallios flow cytometer (Beckman Coulter, USA) and data were analyzed by Flowjo 2.0. The relative protein abundance of GFP was evaluated as the ratio of the average fluorescence intensity of GFP to that of mCherry. The experiment was performed with three biological replicates.

Training and testing of models

For machine learning, a total of 15 features were extracted from each of 241 5´ UTR, in which 12 features were described previously [12]. Three additional features included the length of the 5´ UTR (5´ UTR length), the length of the longest poly(A) in 5´ UTR (poly(A) length), and the distance between the longest poly(A) tract and AUG (poly(A) position). The minimum free energy (MFE) of the entire 5´ UTR along with the first 50 nt in the ORF was calculated using RNAStructure [50]. Models were trained based on MLP-NN using a dataset comprising feature values of 241 5´ UTRs from the mini library, as well as the corresponding relative GFP abundance caused by these 5´ UTRs. The dataset was shown in Additional file 6. Among 241 5´ UTRs, 193 (80%) 5´ UTRs were selected to form the training set, and the other 20% were withheld for the test set. To optimize the hyperparameters of the MLP-NN model, the training set was then divided into 5 subsets for a 5-fold cross-validation. We investigated combinations of hyperparameters, including the number of dense layers and the number of units per layer. Upon reviewing the results, we opted for a setup consisting of 3 dense layers, each housing 300 units. The activation functions of layers were default (relu). The independent test set was used to evaluate the prediction accuracy of the model. The coefficient of determination (R^2) was calculated to represent the prediction capability of the model. All model training and prediction processes were performed in Python 3.8 using TensorFlow 2.12.0. Python codes were available at https://github.com/CODdown/2023--MLP-NN.

Sensitivity analysis of MLP-NN mode

The sensitivity analysis of the MLP-NN model was performed using Shapley Additive Explanation (SHAP), which illustrates the potential influence of features attributed to the model [51]. SHAP sensitivity analysis was performed in Python 3.8 using shap 0.41.0. Python codes were available at https://github.com/CODdown/2023--MLP-NN.

Enzymatic assay of AnFaeA

LHZ1442~LHZ1447 and LHZ1449~LHZ1462 were transformed into Fim-1ΔU separately, and transformants were selected on SC-Ura plates. Transformants were grown for 72 h in SC-Ura+Glutamate medium, in which $(NH4)_2SO_4$ was replaced by 0.1% glutamate since $(NH4)_2SO_4$ inhibits the activity of AnFaeA. Supernatants were subjected to an enzymatic assay of AnFaeA as described previously [34]. The assay was performed with three biological replicates.

Real-time PCR

Transformants were grown in SC-Ura liquid medium or SC-Ura+Glutamate liquid medium for 72 h. Cells were harvested and the total RNA was extracted as described in nanopore RNAseq. RNA was reverse transcribed using HiScript III All-in-one RT SuperMix Perfect for qPCR (Vazyme, R333-01), and cDNA was subjected to real-time PCR using ChamQ Universal SYBR qPCR Master Mix (Vazyme, Q711-02). The mRNA level of *SWC4* served as a control. The experiment was performed with three biological replicates. The real-time PCR data were analyzed following the $2^{-\Delta\Delta Ct}$ method. The relative mRNA level of the target gene was calculated using the following equation:

$$relative\, mRNA\, level_{target} = 2^{-(Ct_{target} - Ct_{SWC4})}$$

where the Ct_{target} is the cycle threshold of the target gene (*GFP* or *AnFaeA*) and the Ct_{SWC4} is the cycle threshold of a housekeeper gene *SWC4*.

Results

5′ UTR poly(A)s are linked with mRNA levels and protein abundance in *K. marxianus*

To characterize the function of 5′ UTR poly(A) in *K. marxianus*, we collected cultures of *K. marxianus* after 16 h and 72 h for nanopore RNAseq and proteomics analysis. The time points of 16 h and 72 h mark crucial stages during fermentation. At 16 h, *K. marxianus* enters the late-exponential stage, and the culture is collected as seed culture for feed-batch fermentation [34]. After 72 h, *K. marxianus* enters the stationary phase. At this stage, fermentation is manually halted, and the culture is collected for expression analysis [34, 52]. Among the various 5′ UTRs of each gene, we selected the most frequently occurring 5′ UTR after redundancy for analysis. If there are multiple 5′ UTRs with the highest frequency, the longest one was chosen. As shown in Fig. 1A, the 5′ UTRs at both time points shared a similar length distribution, with a median length of 100 nt and an enriched peak at 50 nt. The median length of the 5′ UTR in *K. marxianus* was longer than that in *S. cerevisiae* (~50nt) [53], suggesting a different mechanism to control 5′ UTR length among these two species. In this study, continuous adenine tracts equal to or longer than 5 nt within the 5′ UTR were designated as 5′ UTR poly(A)s. At both time points, around 25% of genes in *K. marxianus* harbored at least one 5′ UTR poly(A), a ratio comparable to that of *S. cerevisiae* [16]. The 5′ UTRs containing poly(A) were significantly longer than those without poly(A), suggesting that there is a higher occurrence of poly(A) in long 5′ UTRs (Fig. 1B). As the length of poly(A) increased, the number of 5′ UTRs containing such poly(A) decreased (Fig. 1C). The longest poly(A)s, composed of 23 As, were found in the 5′ UTRs of *NOP15* and *PFK27*. The median distance between the poly(A) and AUG was 76 nt and 86 nt for 16 h and 72 h, respectively, with peaks around 20 nt. In general, the 5′ UTRs and their poly(A) tracts shared similar characteristics at 16 h and 72 h, indicating a consistent regulation of the 5′ UTRs at both time points.

Given the frequent presence of poly(A) close to AUG, we calculated the enrichment and depletion of four bases between 30 nt preceding AUG (-30) and 10 nt after AUG (+10) in genes ranked by the abundance of encoded proteins. The emPAI exhibits a linear correlation with the logarithm of protein concentration [43]. For the 46 proteins in mouse whole cell lysate, the average deviation percentages of emPAI-based abundances from the actual values were within 63% [43]. Abundances of 40 proteins in *Escherichia coli* cytosol measured by the emPAI method correlated well with those determined through isotope dilution of a control lysate (R^2=0.84) [54]. Additionally, emPAI-based protein concentration is automatically available for all proteins identified by MS without any additional experimental setup. Given these advantages, emPAI was employed in this study to indicate protein abundance. In the region around AUG, a Kozak sequence (A A/C A A/C A (AUG) U C/U C) was enriched in the top 20% of genes (Fig. 1E, upper panel). The Kozak sequence of *K. marxianus* closely resembled that of *S. cerevisiae*, indicating the reliability of our analysis [11, 55]. Regarding the upstream sequence, residue A was favored, but G and U were loathed within 30 nt preceding AUG in the top 20% of genes, ranked by the abundance of proteins encoded by genes (Fig. 1E, upper panel). The opposite trend was observed in the bottom 20% of genes (Fig. 1E, lower panel). Notably, the favored As in this region tended to cluster together to form poly(A) tracts around 20 nt preceding AUG, consistent with the peaks of poly(A) distribution shown in Fig. 1D. A similar pattern was observed in genes ranked by mRNA levels, as residue A was favored and tended to cluster around 20 nt before the AUG codon in the top 20% of genes (Fig. 1F, upper panel). In contrast, we did not observe any enrichment of poly(A) tracts between 100 ~ 30 nt preceding AUG in top 20% of genes ranked by either protein abundance or mRNA levels (Additional file 1: Fig S1).

To further identify the link between poly(A) tracts and protein abundance, we calculated the percentage of genes containing 5′ UTR poly(A) in groups of genes distinguished by a ratio between protein abundance and mRNA levels. This ratio served as a rough indicator of translation efficiency. Regarding the 5′ UTR poly(A) located within 30 nt preceding AUG, the percentage of genes containing this 5′ UTR poly(A) was positively correlated with the protein/mRNA ratio (Fig. 1G), suggesting that poly(A)s located close to AUG increase translation efficiency. However, the percentage of genes containing 5′ UTR poly(A) at any position was negatively correlated with the protein/mRNA ratio (Fig. 1H), suggesting that 5′ UTR poly(A)s generally inhibit translation. Consistent results were obtained when analyzing the relationship between poly(A) and protein/mRNA ratio in ungrouped genes. The average protein/mRNA ratio in genes with 5′ UTR poly(A) was significantly lower than that in genes without 5′ UTR poly(A) (Additional file 1: Fig S2A). The average protein/mRNA ratio in genes containing 5′ UTR poly(A) located within 30 nt preceding AUG was significantly higher than that in genes lacking this 5′ UTR poly(A) (Additional file 1: Fig S2B). The contradictory results between total poly(A) and poly(A) close to AUG suggest that the effect of poly(A) on protein production is position-dependent.

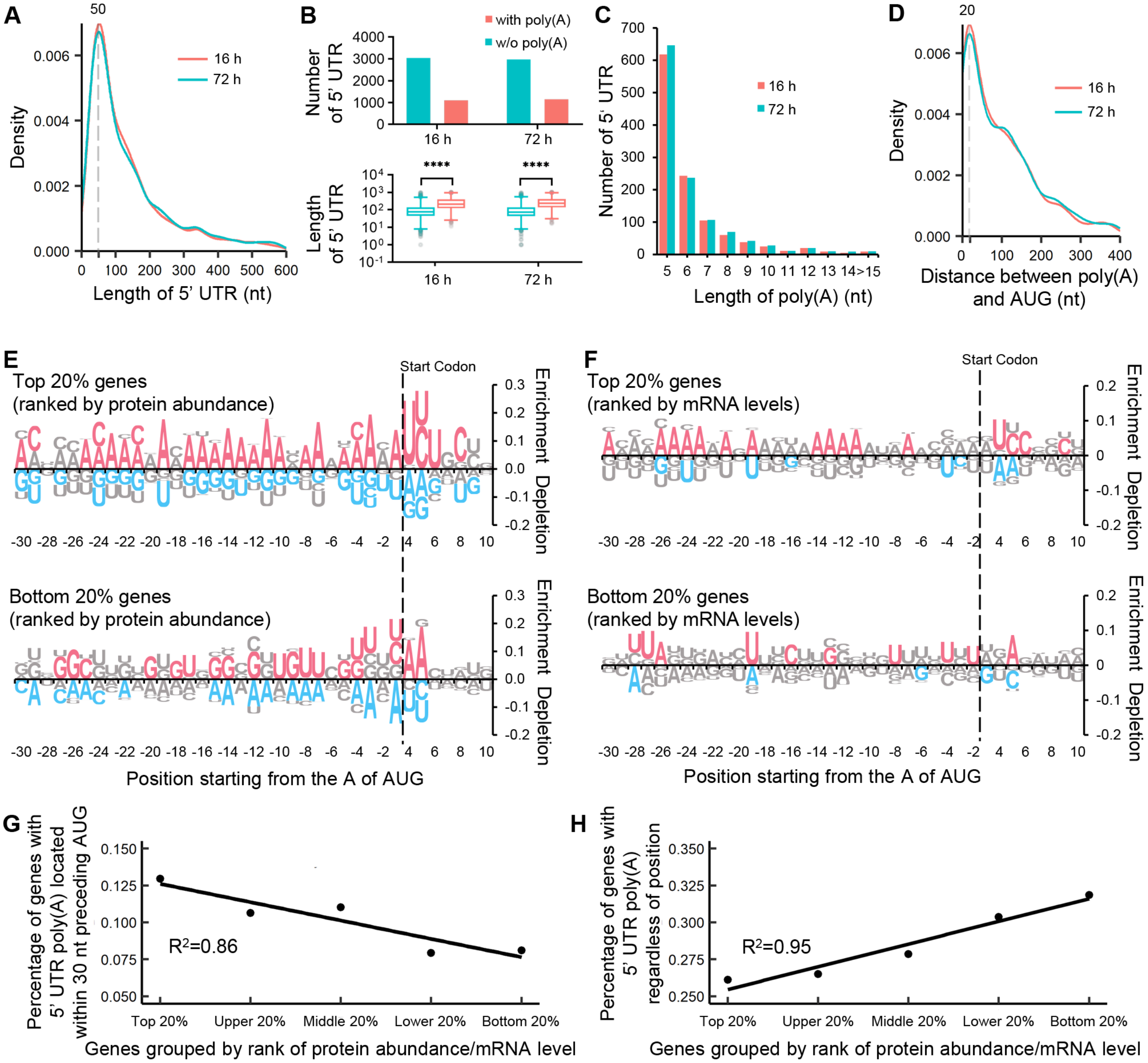


Fig. 1 5´ UTR poly(A)s are linked with mRNA levels and protein abundance. (**A**) Distributions of 5´ UTR lengths in the *K. marxianus*. Cells were collected after 16 h and 72 h of growth, and subjected to the nanopore sequencing. A total of 4228 5´ UTRs from the 16 h sample and 4210 5´ UTRs from the 72 h sample were analyzed. A peak around 50 nt was indicated. (**B**) Number and length distributions of 5´ UTRs with or without poly(A) in 16 h and 72 h samples. The significance was assessed by a two-tails t-test. **** $p < 0.0001$. (**C**) Number of 5´ UTR containing poly(A) of various lengths. (**D**) Distribution of distance between 5´ UTR poly(A) and start codon (AUG). A peak around 20 nt was indicated. (**E**, **F**) Enrichment and depletion of four bases between 30 nt preceding and 7 nt after AUG (-30 ~ + 10) in different groups of genes. The genes were grouped based on the abundance of the encoded proteins (**E**) or the mRNA levels (**F**). The significance was assessed using a two-tailed Fisher's exact test. Red or blue logos represented $p < 0.05$, while gray logos represented $p > 0.05$. (**G**, **H**) Correlation between the percentage of genes containing 5´ UTR poly(A) and the genes grouped by the ratio between protein abundance and mRNA level. Protein abundance and mRNA level were represented by emPAI and TPM values, respectively. The genes containing 5´ UTR poly(A) located within 30 nt preceding AUG were shown on the left (**G**), while those containing 5´ UTR poly(A) at any position were shown on the right (**H**)

Evaluate effects of 5´ UTR poly(A)s on protein production by a dual-reporter system

Evaluating the effects of 5´ UTR poly(A) on protein production based on transcriptomic and proteomic data was interfered with by gene context, including the promoter, ORF, and terminator. To reduce this interference, we constructed a dual-reporter system (Fig. 2A). We screened for natural 5´ UTRs with high abundance and a length of less than 200 nt. Among these 5´ UTRs, we selected a total of 207 5´ UTRs containing poly(A) as well as 34 5´ UTRs without poly(A) as controls. Each 5´ UTR was then cloned separately into LHZ1138 between

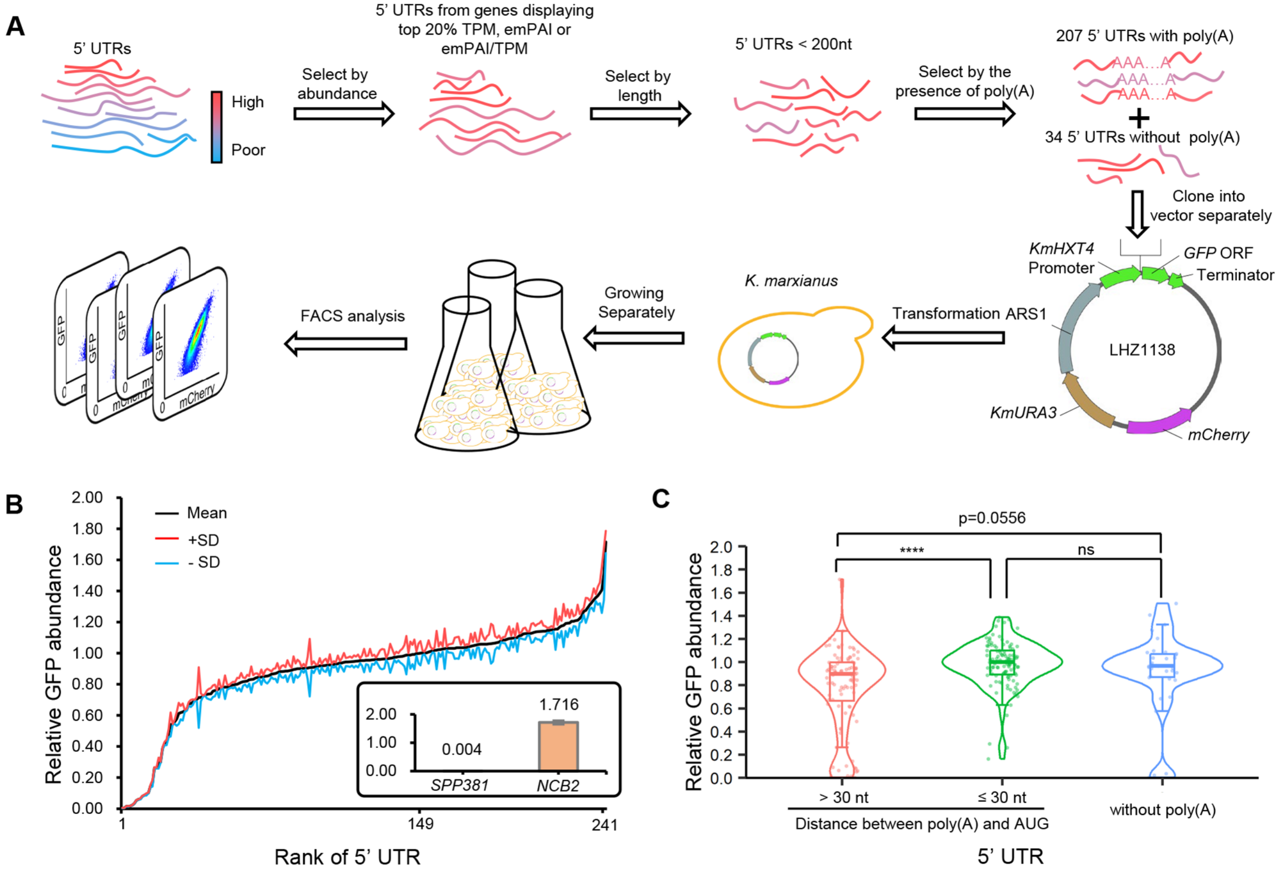


Fig. 2 Evaluation of 5´ UTR poly(A)s through a dual-reporter system. (**A**) Flow chart of the dual-reporter system used to evaluate the effect of 5´ UTRs on protein production. Natural 5´ UTRs from abundant genes were randomly selected and cloned into separate vectors. The constructed plasmids were then transformed into *K. marxianus*, and the hosts were cultured separately for 72 h before FACS analysis. (**B**) Relative GFP abundance caused by 5´ UTRs. The relative GFP abundance was calculated as GFP/mCherry, where the relative GFP abundance caused by *HXT4* 5´ UTR was designated as 1. The mean relative GFP abundance of each 5´ UTR was ranked in ascending order, with the standard deviations (SD) shown (n = 3). The insert displayed 5´ UTRs that caused the lowest and highest relative GFP abundance. (**C**) Comparison of the relative GFP abundance caused by 5´ UTRs with poly(A) located at different positions. The significance was assessed using a two-tailed t-test. ****$p < 0.0001$; ns, $p > 0.05$

a strong *HXT4* promoter and the ORF of *GFP*, with the vector containing a centromeric sequence (ARS1) to maintain a single copy in vivo and a cassette to express mCherry constitutively. The constructed plasmids were transformed into *K. marxianus*, cultured separately, and subjected to FACS. The abundance of GFP was normalized by the amount of mCherry.

A total of 241 5´ UTRs resulted in a broad range of relative GFP abundance (Fig. 2B). Compared to the natural 5´ UTR of *HXT4*, the lowest relative GFP abundance caused by the 5´ UTR of *SPP381* was 0.004 (Fig. 2B, insert). Despite mRNA and protein levels of *SPP381* being abundant, the low relative GFP abundance caused by its 5´ UTR indicates that the strong effect of gene context bypasses the effect of 5´ UTR. In contrast, the highest relative abundance caused by 5´ UTR of *NCB2* was 1.76 (Fig. 2B, insert). The relative GFP abundance caused by 5´ UTRs spanned an approximately 450-fold range, indicating the great potential of 5´ UTRs in regulating the production of proteins.

The mean relative abundance caused by 5´ UTRs that contain poly(A) within 30 nt preceding AUG was significantly higher than those contain poly(A) beyond 30 nt ($p<0.0001$) (Fig. 2C). The result suggest the effect of 5´ UTR poly(A) on protein production varies depending on its position around −30 nt, which was consistent with data in Fig. 1G. Meanwhile, the mean relative abundance of 5´ UTRs that contain poly(A) beyond 30 nt was slightly lower than that of 5´ UTRs without poly(A) ($p=0.0556$), suggesting poly(A)s located distantly to AUG show negative effects on protein production. 5´ UTRs without poly(A) did not show a significant difference compared to 5´ UTRs with poly(A) smaller than 30 nt.

Machine-learning model reveals the negative and position-dependent effect of 5´ UTR poly(A) on protein production

The data obtained through the dual-reporter system was used to train a predictive model of 5´ UTR, which included the twelve features used in the training of previous models [12, 26]. These features included out-of-frame upstream AUG and upstream ORF (oofuAUG), MFE, and nucleotide preferences at the position immediately upstream of AUG. In addition, three new features were incorporated, including 5´ UTR length, poly(A) length, and poly(A) position. To include the data of 34 5´ UTRs without a poly(A) longer than 5 nt, the longest continuous As in these 5´ UTRs were used as poly(A)s to calculate relevant features. Therefore, a total of 241 5´ UTRs with 15 features each, along with relative GFP abundance caused by each 5´ UTR, constituted the dataset for model construction (Additional file 6). The dataset was randomly divided into two subsets. One subset, consisting of 15 feature values and the corresponding relative GFP abundance values of 193 5´ UTRs, was utilized as a training set to calibrate the model based on MLP-NN after optimizing hyper-parameters using a 5-fold cross-validation. Following the calibration, the remaining data of 48 5´ UTRs was employed as a test set to evaluate the predictive performance of the model. Model prediction reveals that a linear fit between the observed abundance and predicted abundance in the test set yielded an R^2 of 0.7595 (Fig. 3A), indicating that the MLP-NN model can successfully predict the protein abundance caused by 5´ UTRs. We conducted four additional train-test splits to perform training and validation, resulting in the acquisition of 4 additional MLP-NN models. The average R^2 for test set predictions of 5 models, including one original and four additional models, was 0.7290 (Additional file 1: Fig S3A). When new features (5´ UTR length, poly(A) length, and poly(A) position) were excluded from five distinct training-test splits, the average R^2 decreased to 0.6403 (Additional file 1: Fig S3B), suggesting incorporation of these three features into the training process improves the model's performance. In addition, we also built models based on SVM and random forest. However, both models resulted in lower R^2 values for the test set (Additional file 1: Fig S4). These results suggest that, compared with the SVM and random forest models, the MLP-NN model is more effective at accurately capturing the association between 5´ UTR features and protein abundance. To assess our model's competency in predicting protein production in other yeasts, we utilized a library comprising half a million 50-nt-long random 5´ UTRs previously tested in *S. cerevisiae* [7]. The impact of each 5´ UTR on *HIS3* production was assessed by measuring the enrichment of cells harboring the 5´ UTR after cultivation in selection media [7]. From this library, we selected 700 5´ UTRs containing poly(A) and 115 5´ UTRs lacking poly(A). The ratio of 700:115 was comparable to that in our 5´ UTR library. Employing the MLP-NN model, we predicted the enrichment of each 5´ UTR, yielding an R^2 of 0.503 between predicted and measured enrichments (Additional File 1: Fig S5). The relatively low R^2 could be attributed to differences between *S. cerevisiae* and *K. marxianus*. Additionally, our model might have underperformed in predicting enrichment, which served as an indirect indicator of protein abundance [7].

Based on the MLP-NN model, we conducted a sensitivity analysis to evaluate the impacts of features on protein abundance. SHAP (SHapley Additive exPlanations) values were calculated for each feature, and negative and positive SHAP values indicate negative and positive impacts on protein abundance, respectively, while the absolute value reflects the magnitude of the impact. Based on the rank of absolute SHAP values, 5´ UTR length was found to be the most influential feature on protein abundance (Fig. 3B). The feature of oofuAUG ranked as the second most influential feature (Fig. 3B), which is expected given that out-of-frame upstream AUGs and upstream ORFs have been shown to significantly impact translation efficiency [6–8]. Another important mRNA feature, MFE, ranked as the third most influential feature (Fig. 3B). The poly(A) position ranked fourth, indicating that the distance between poly(A) and AUG has an important impact on protein abundance (Fig. 3B). The impact of poly(A) length was relatively small, ranking 11th out of the total 15 features (Fig. 3B).

To visualize the relationship between feature values and SHAP values, red and blue dots were used to mark high and low feature values, respectively (Fig. 3C). High values of 5´ UTR length were found to be strongly correlated with negative SHAP values (Fig. 3C). The result indicates that the length of the 5´ UTR generally has a negative impact on protein abundance, possibly because longer 5´ UTRs have a greater tendency to form secondary structures that hinder ribosomal scanning [13]. In contrast, less negative MFE indicates a less stable secondary structure [56], leading to positive SHAP values. As expected, increased values of oofuAUG were strongly correlated with negative SHAP values, indicating the strong negative effects of out-of-frame upstream AUGs and upstream ORFs on translation [6–8]. Overall, most correlations between feature values and SHAP values were consistent with previous studies [12], indicating that the regulatory mechanisms of 5´ UTR features are conserved in *K. marxianus*. The relationships between poly(A) features and SHAP values were examined in detail. The length of poly(A) was found to be negatively related to the SHAP value (Fig. 3D). In most cases, poly(A) longer than 5 nt was associated with a negative SHAP value. The result indicates that poly(A) generally

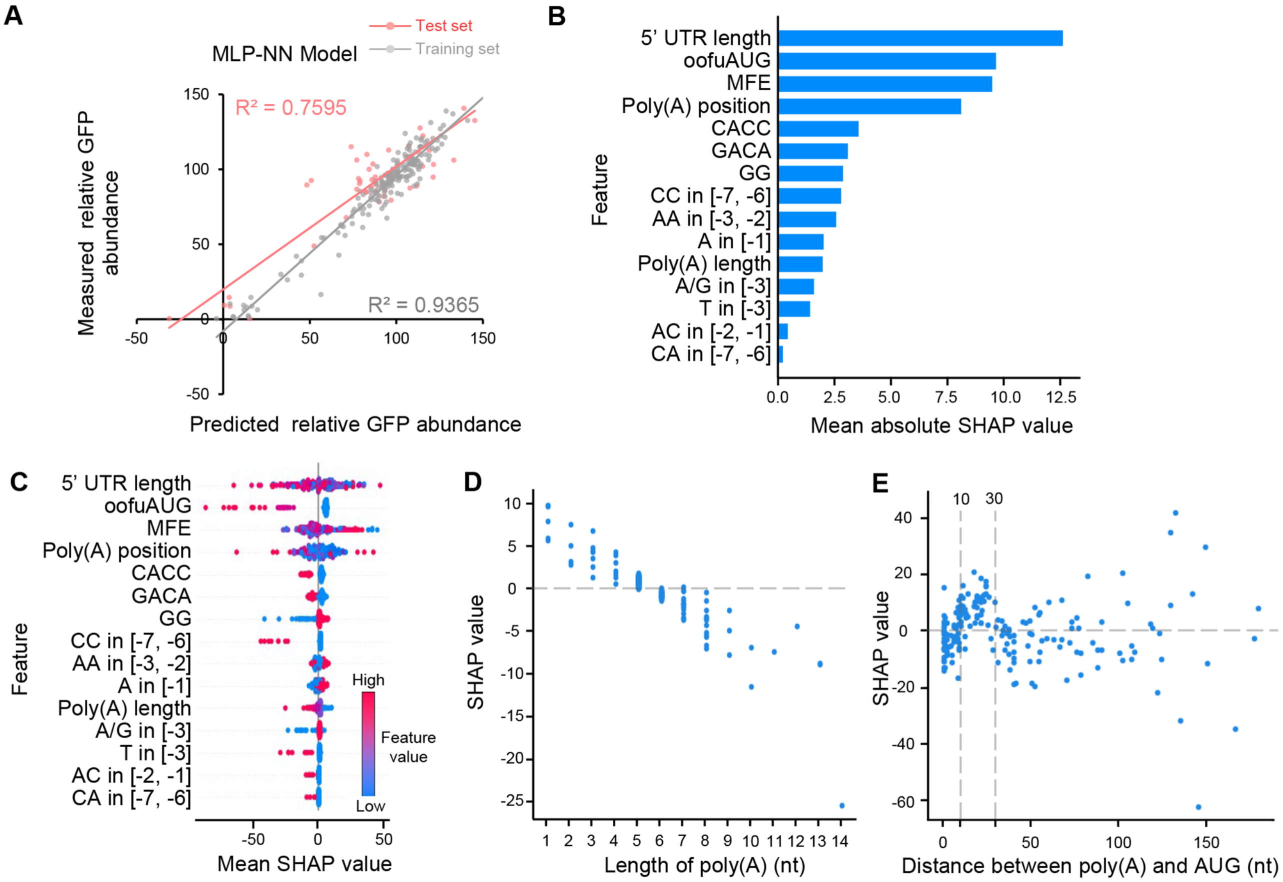
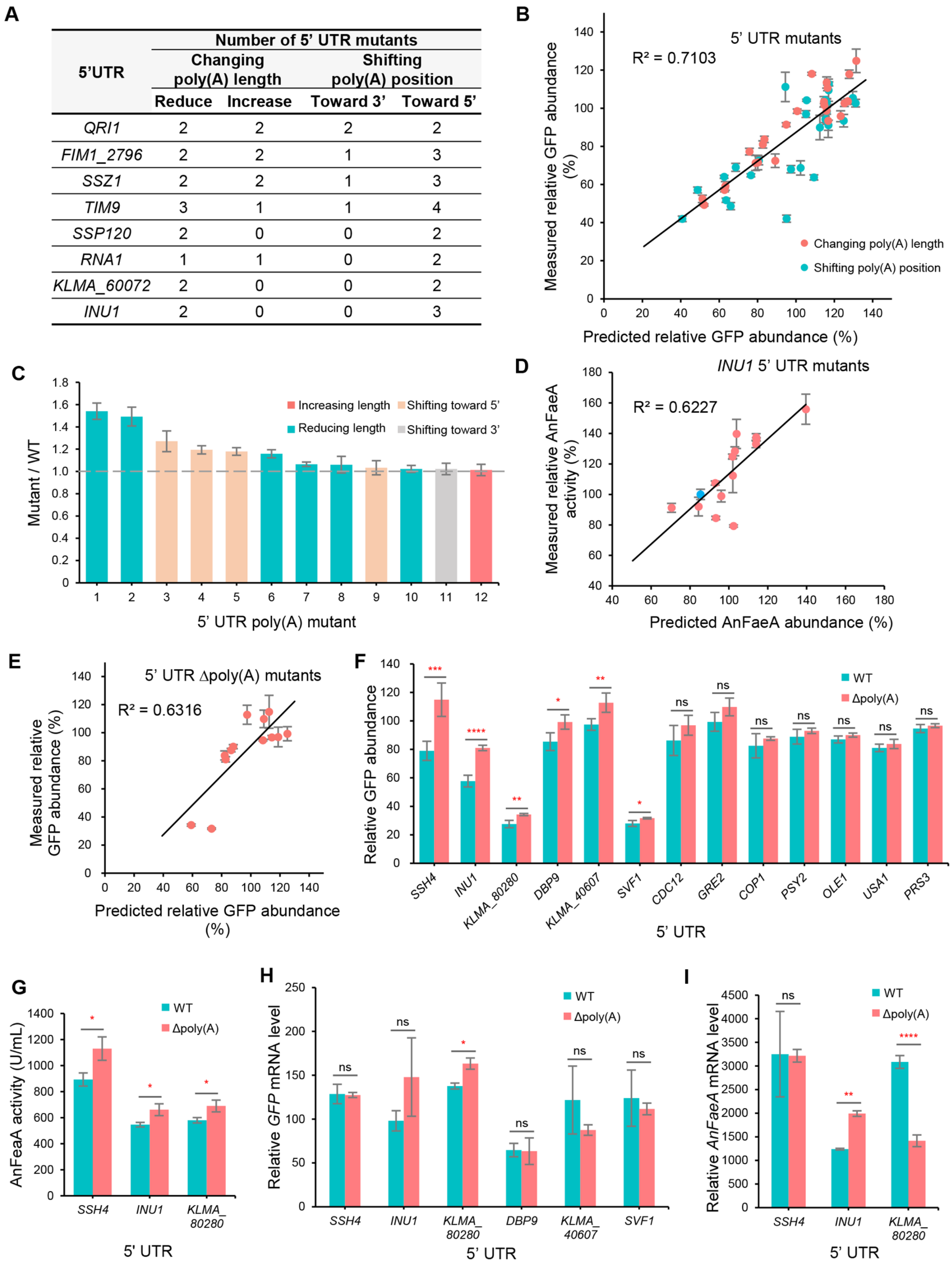


Fig. 3 Construction and analysis of a machine-learning model that predicts the GFP abundance by features of 5´ UTR. (**A**) Validation of the MLP-NN model. The plot compared the measured versus the predicted relative GFP abundance, with R^2 for the train and test sets included. (**B**) 5´ UTR features ranked by their mean absolute SHAP values. Mean SHAP values were obtained by performing sensitivity analysis on the MLP-NN model. Description of the features: 5´ UTR length, length of 5´ UTR; oofuAUG, number of out-of-frame upstream AUGs and upstream ORFs; MFE, minimum free energy; poly(A) position, the distance between the longest poly(A) tract and AUG; CACC, the presence of at least one CACC motif in the 5´ UTR; GACA, the presence of at least one GACA motif in the 5´ UTR; GG, the presence of at least one GG motif in the 5´ UTR; CC in [-7, -6], the presence of the motif CC at position [-7, -6] relative to the position of AUG; AA in [-3, -2], the presence of the motif AA at position [-3, -2]; A in [-1], the presence of the A at position – 1; poly(A) length, length of the longest poly(A) in 5´ UTR; A/G in [-3], the presence of the A or G at position – 3; T in [-3], the presence of the T at position – 3; AC in [-2, -1], the presence of the motif AC at position [-2, -1]; CA in [-7, -6], the presence of the motif CA at position [-7, -6]. (**C**) The relationship between the values of 5´ UTR features and SHAP values. Red and blue dots indicated high and low feature values, respectively. (**D**) A negative correlation between SHAP value and the length of poly(A). (**E**) The relationship between SHAP value and the position of poly(A). The majority of SHAP values were positive at distances between 10 and 30 nt, as indicated

acts as a negative regulator of protein production, with longer poly(A) resulting in a greater negative effect on protein abundance. In Fig. 3E, SHAP values quantified the impact of poly(A) at each position on the model's output, which might uncover differences obscured by the comparison of mean GFP production shown in Fig. 2C. There was no linear correlation between the poly(A) positions and SHAP values, but poly(A)s with a distance to AUG between 10 ~ 30 nt showed a correlation with positive SHAP values (Fig. 3E). A similar relationship between SHAP values and poly(A) features was detected in the sensitivity analysis using four models trained with additional training-test splits, reflecting the reliability of both the model construction and sensitivity analysis (Additional file 1: Fig S6). The results suggest that poly(A)s in this region may improve protein production. It was consistent with the opposite behaviors between the total 5´ UTR poly(A)s and 5´ UTR poly(A)s located close to AUG (Fig. 1G, H), indicating that the effect of poly(A) on protein production is dependent on its position.

Optimization of 5´ UTR poly(A) to improve protein production with guidance of the machine-learning model

The MLP-NN model was trained using data from natural 5´ UTRs. To validate the predictive accuracy of the model on non-natural sequences, 5´ UTRs from 8 genes were randomly selected and their poly(A) tracts were altered in length or position, resulting in a total of 50



5'UTR	Number of 5' UTR mutants			
	Changing poly(A) length		Shifting poly(A) position	
	Reduce	Increase	Toward 3'	Toward 5'
QRI1	2	2	2	2
FIM1_2796	2	2	1	3
SSZ1	2	2	1	3
TIM9	3	1	1	4
SSP120	2	0	0	2
RNA1	1	1	0	2
KLMA_60072	2	0	0	2
INU1	2	0	0	3

Fig. 4 (See legend on next page.)

5´ UTR mutants (Fig. 4A). The MLP-NN model predicted the relative GFP abundance caused by the 5´ UTR mutants and these predictions were compared with the experimental measurements. The results showed that the model's predictions on 5´ UTR mutants were accurate, with an R^2 value of 0.7103 (Fig. 4B). When ranking the ratio between the relative GFP abundance caused by the 5´ UTR mutants and that caused by the corresponding

(See figure on previous page.)

Fig. 4 Increasing protein production by reducing or deleting 5′ UTR poly(A) with the guidance of the machine-learning model. (**A**) Summary of 5′ UTR mutants. The mutants were divided into two groups: one with changes in poly(A) length, and the other with shifts of the poly(A) position. (**B**) Plot representing the measured versus predicted relative GFP abundance of 5′ UTR mutants in (**A**). The value of R^2 was shown in the plot. Standard deviations (SD) of measured abundance were shown (n = 3). Mutants that changed the poly(A) length and position were labelled in red and green, respectively. (**C**) Ratio between relative GFP abundance caused by mutant 5′ UTRs and that caused by wild-type 5′ UTRs. The mutants were ranked in descending order based on the ratios. Four types of mutants were distinguished by different colors. Mean ± SD was shown (n = 3). (**D**) Plot representing the measured AnFaeA activity versus predicted AnFaeA production in poly(A) mutants of *INU1* 5′ UTR. The value of R^2 was shown in the plot. SD of measured abundance were shown (n = 3). The point representing the wild-type 5′ UTR of *INU1* was colored blue. (**E**) Plot representing the measured versus predicted relative GFP abundance caused by 5′ UTR Δpoly(A) mutants. The value of R^2 was shown in the plot. Mean ± SD of measured abundance was shown (n = 3). (**F**) Comparison between the relative GFP abundance caused by 5′ UTR Δpoly(A) mutants and that caused by the wild-type 5′ UTRs. Mean ± SD was shown (n = 3). The significance was assessed by a two-tailed t-test. **** $p < 0.0001$; *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$; ns > 0.05. (**G**) Comparison between the AnFaeA activity caused by 5′ UTR Δpoly(A) mutants and that caused by the wild-type 5′ UTRs. Enzymatic activity was measured after culturing for 72 h. Mean ± SD was shown (n = 3). The significance was assessed by a two-tailed t-test. * $p < 0.05$. (**H**, **I**) Comparison between relative mRNA levels of *GFP* (**H**) or *AnFaeA* (**I**) expressed by 5′ UTR Δpoly(A) mutants and those expressed by wild-type 5′ UTRs. mRNA was extracted from samples described in (**E**) and (**F**) and subjected to qPCR analysis. The mRNA levels of *GFP* or *AnFaeA* were normalized with the mRNA level of *SWC4*. Mean ± SD was shown (n = 3). The significance was assessed by a two-tailed t-test. **** $p < 0.0001$; ** $p < 0.01$; * $p < 0.05$; ns > 0.05

wild-type 5′ UTRs, 6 of the top 12 mutants with the largest fold changes contained a reduced length of poly(A) (Fig. 4C). These results suggest that reducing the poly(A) tracts tends to improve protein production, which is consistent with the observation that the length of poly(A) is negatively related to protein production (Fig. 3D). To validate the accuracy of the model's predictions regarding the production of proteins other than GFP, alterations were made to 5′ UTR poly(A) in a multiple-copy plasmid [34], which drives secretory expression of a feruloyl esterase (AnFaeA) through the promoter and 5′ UTR of *INU1*. A total of 14 5′ UTR mutants were constructed. The secretory activities of AnFaeA caused by these *INU1* 5′ UTR mutants were measured and compared with that caused by wild-type *INU1* 5′ UTR. The measured relative activities of AnFaeA exhibited a linear correlation with the predicted values (R^2=0.6227) (Fig. 4D). Therefore, these results suggest that the model is capable of predicting the production of different proteins in *K. marxianus*.

Compared to reducing the length of the 5′ UTR poly(A), deletion of the full poly(A) tract is more applicable in sequence engineering and might lead to a more dramatic effect on protein production. To avoid interfering with the potential positive effect of poly(A) with a distance to AUG between 10 ~ 30 nt (Fig. 3E), we proposed a strategy to delete poly(A) upstream of 30 nt preceding AUG. Thirteen 5′ UTRs containing poly(A) upstream of -30 nt were selected and deletions of poly(A)s in these 5′ UTRs were predicted to increase protein production by the MLP-NN model. These 5′ UTR Δpoly(A) mutants were constructed and measured. The measured relative GFP abundance caused by the 5′ UTR Δpoly(A) mutants showed a decent linear fit with predicted abundance (R^2=0.6316), again proving the accuracy of the model's prediction (Fig. 4E). Among the 13 5′ UTR Δpoly(A) mutants, 6 mutants caused a significant increase in the GFP abundance compared to the wild-type 5′ UTRs, while the remaining 7 mutants showed the same GFP abundance as the wild-type 5′ UTRs (Fig. 4F). Therefore, approximately 50% of the poly(A) mutants increased protein production. To verify the applicability of this strategy in different sequence contexts, the top three 5′ UTRs (*SSH4*, *INU1*, *KLMA_80280*) showing the highest fold increase in GFP abundance after deleting poly(A) were constructed on a multiple-copy vector to drive the secretory expression of AnFaeA. As shown in Fig. 4G, the deletions of poly(A)s in the 5′ UTRs of *SSH4*, *INU1*, and *KLMA_80280* significantly improved the production of AnFaeA compared to the wild-type 5′ UTRs. The results indicate that deletions of the 5′ UTR poly(A)s upstream of 30 nt preceding AUG tend to improve protein production in different ORF contexts.

A previous study showed that the presence of poly(A) in the 5′ UTR decreased mRNA levels [57]. However, we found that deletions of poly(A)s showed inconsistent effects on mRNA levels. Among the 7 mutants that caused an increase in GFP abundance, only the Δpoly(A) mutant of *KLMA_80280* 5′ UTR significantly increased the *GFP* mRNA level, while the mRNA levels expressed by the other Δpoly(A) mutants were not significantly different from those expressed by the wild-type 5′ UTRs (Fig. 4H). Among the three mutants that caused an increase in AnFaeA abundance, the Δpoly(A) mutant of *INU1* 5′ UTR increased the *AnFaeA* mRNA level, while the Δpoly(A) mutant of *KLMA_80280* 5′ UTR decreased the mRNA level (Fig. 4I). Therefore, the increased protein abundance observed in the 5′ UTR Δpoly(A) mutants was not solely due to increased mRNA levels. The results suggest that poly(A) represses protein production, either with or without reducing mRNA levels.

Discussion

In prior studies, researchers synthesized randomly designed short 5′ UTRs (less than 100 nt) to build libraries and determined the impact of each 5′ UTR on reporter protein abundance [6, 7, 12, 25, 27]. A parallel assay was commonly performed during this step, wherein cells containing different 5′ UTRs were grown and sequenced together [6, 7, 12, 27]. Subsequently, predictive models were constructed by incorporating the features of the 5′ UTR [12, 25, 26]. In our study, we followed a similar strategy with some modifications. Firstly, we synthesized a library of natural 5′ UTRs with lengths ranging from 12 to 197 nt. Despite longer sequences increasing the complexity of machine learning, analyzing native sequences provides valuable information about natural regulation. Secondly, we evaluated each 5′ UTR's impact on protein production separately in vivo, reducing potential interference from other cells during parallel reporter assays. Lastly, we incorporated the length and position of poly(A) into machine learning, which had not been done before. Our model accurately predicted protein production induced by the 5′ UTR (R^2=0.7595), suggesting that poly(A) features are effective in constructing high-quality models. To further enhance the model's performance, the dataset of 5′ UTRs needs to be expanded to augment the diversity of selected features, encompassing poly(A) features. Constructing a larger library comprising natural 5′ UTRs from *K. marxianus* using high-throughput synthesis techniques is a potential avenue [58, 59]. However, this task is challenging considering that the reported libraries typically consisted of 5′ UTRs with lengths of 100 nt or less [6, 7, 12, 25, 27]. Moreover, it's crucial to introduce parallel assays to measure the impacts of 5′ UTRs on protein expression within a larger library.

The SHAP sensitivity analysis based on the MLP-NN model reveals a general negative correlation between poly(A) length and protein production, suggesting that poly(A) primarily functions as a negative regulator of protein production. Deletions of poly(A)s showed inconsistent effects on mRNA levels (Fig. 4H, I), suggesting that poly(A) represses protein production either with or without reducing mRNA levels. Consistent with this hypothesis, it was shown that the 5′ UTR poly(A) of *PABP1* can independently repress mRNA translation and reduce mRNA abundance [57]. To repress translation, the 5′ UTR poly(A) recruits PABP1 and prevents the 40S ribosomal subunit from moving to the initiation codon [22]. A similar mechanism may be employed by the 5′ UTR poly(A) in *K. marxianus*. Meanwhile, 5′ UTR poly(A) may reduce mRNA levels by decreasing its stability. The negative effect of poly(A) on mRNA stability is likely proportional to its length, since longer 5′ UTR poly(A) sequences were found to result in shorter mRNA half-lives in vitro [21]. This may partially explain the negative correlation between poly(A) length and protein production. The Ccr4–Not and Pan2–Pan3 complexes, which are responsible for the 3′-end poly(A) tail shortening, have been proposed to mediate the degradation of mRNAs containing 5′ UTR poly(A) [21]. These complexes are conserved from yeast to humans [60, 61], suggesting that they likely play similar roles in degrading 5′ UTR poly(A) sequences in *K. marxianus*.

It is noteworthy that the Δpoly(A) mutant of *KLMA_80280* 5′ UTR reduced the *AnFaeA* mRNA level (Fig. 4I). The poly(A) in *KLMA_80280* 5′ UTR is precisely located at the 5′ end of the mRNA, and the sequence around the transcriptional start site (CCA[+1]AAAA) matches the consensus sequence of the transcriptional initiator in *Schizosaccharomyces pombe* ((C/T)(C/T)(A/G)[+1]N(A/C)(A/C)) [62], where (A/G)[+1] represents the transcription start site. The transcriptional initiator is a core promoter element found near the transcription start site on the DNA, playing a role in directing transcription initiation [62]. Therefore, the deletion of the poly(A) tract might impair the transcription of *KLMA_80280* 5′ UTR, leading to decreased *AnFaeA* mRNA level. In contrast to the decrease in *AnFaeA* mRNA level, the Δpoly(A) mutant of *KLMA_80280* 5′ UTR slightly increased the *GFP* mRNA level (Fig. 4H). Similar inconsistent effects on mRNA levels were observed in the Δpoly(A) mutant of *INU1* 5′ UTR, which caused a slight increase in the *AnFaeA* mRNA level but no alteration in the *GFP* mRNA level (Fig. 4H, I). These inconsistent effects of poly(A) deletion on mRNA levels might be attributed to the distinct ORFs and 3′ UTRs present in the *GFP* and *AnFaeA* expression cassettes. As integral components of mature mRNA, the 5′ UTR, ORF, and 3′ UTR determine the translation process and secondary structure of mRNA, directly influencing mRNA decay and stability [63–68]. Hence, in different sequence contexts, deletions of 5′ UTR poly(A) might have varying effects on mRNA stability and levels.

The analysis of SHAP values also indicates a weak correlation between improved protein production and 5′ UTR poly(A) located between 10–30 nt upstream of AUG, suggesting a position-specific effect of 5′ UTR poly(A). This finding aligns with the observation that poly(A)s with a distance of 30 nt or less from AUG were enriched in genes with high translation efficiency (Fig. 1G). Poly(A) can form an IRES for cap-independent translation [17, 18, 21]. In yeast, the IRES is typically located immediately upstream of the AUG [69]. Hence, compared with poly(A) located further from the AUG, poly(A) close to the AUG has a higher probability of forming an IRES and enhancing translation. Cap-independent translation induced by nearby poly(A) may counteract the negative effect of poly(A), leading to a net positive effect on protein production. The effects of 5′ UTR poly(A) on protein production are summarized in Fig. 5.

The MLP-NN model was effectively employed to guide the optimization of natural 5′ UTR containing poly(A) (R^2=0.6227). Reducing the length or removing poly(A) located upstream of 30 nt preceding AUG appeared to

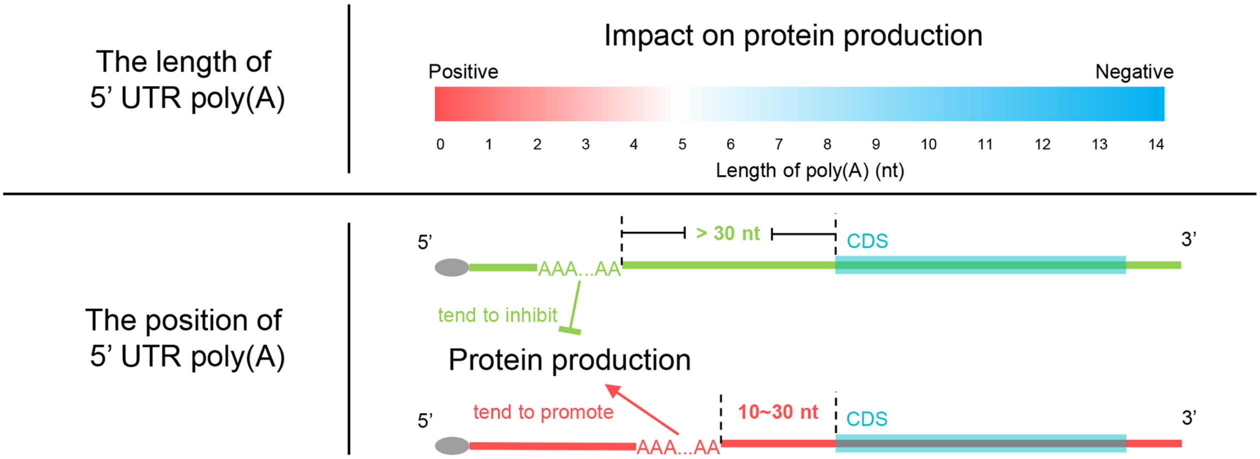


Fig. 5 Impact of length and position of 5´ UTR poly(A) on protein production

be an effective way to enhance protein production. This optimization strategy was successfully applied to GFP and AnFaeA (Fig. 4F, G), indicating its effectiveness across different transcriptional contexts. Traditionally, the natural promoter and 5´ UTR from the same gene were used together to drive the expression of the gene of interest. Several popular 5´ UTRs used in microbial cell factories, such as the 5´ UTRs of *TDH1* and *GAL1* in *S. cerevisiae* [70], and the 5´ UTR of *AOX1* in *Pichia pastoris* [71], contain poly(A) tracts beyond or around 30 nt preceding AUG. Manipulating poly(A)s in these 5´ UTRs might offer a new approach to enhancing yield.

Conclusion

An MLP-NN model was trained using features encompassing 5´ UTR poly(A) length and position, which demonstrated good performance in predicting protein production. The model showed that poly(A) with a distance to AUG between 10~30 nt is slightly correlated with improved protein production. 5´ UTR poly(A) upstream of 30 nt is negatively associated with protein production. Moreover, the negative effect of poly(A) on protein production increases with tract length. With the guidance of the machine model, reducing or removing poly(A) upstream of 30 nt preceding AUG is an effective strategy for improving protein production. This approach could be applied to enhance the yield of *K. marxianus* and other microbial cell factories.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12934-023-02271-3.

Additional file 1: **Fig S1**: Enrichment and depletion of four bases between 100 nt and 30 nt preceding AUG (-100~-30) in different groups of genes. The genes were grouped based on the abundance of the encoded proteins (**A**, **B**) or the level of the produced mRNAs (**C**, **D**), where the top 20% (**A**, **C**) and bottom 20% (**B**, **D**) were selected to calculate the relative entropy of four bases in this region. The significance was assessed using a two-tailed Fisher’s exact test. Logos colored in red or blue represented $p < 0.05$, while gray logos represented $p > 0.05$. **Fig S2**: Comparison of protein abundance/mRNA level of different gene groups. Genes were categorized into two groups based on either the presence or absence of 5´ UTR poly(A) (**A**), or the presence or absence of 5´ UTR poly(A) with a distance of 30 nt or less from AUG (**B**). The significance was determined using a two-tailed t-test. ** $p < 0.01$. * $p < 0.05$. **Fig S3**: Validation of constructed MLP-NN models after five training-test splits using two types of feature selection. (**A**) A total of 15 features, including 5´ UTR length, poly(A) length and poly(A) position were included. A total of 5 different models were constructed using different training-test splits, and the last model was shown in Fig. 3A as a representative. The average coefficient of determination (R^2) for predicting the test sets was 0.7290. (**B**) A total of 12 features were included, while features of 5´ UTR length, poly(A) length and poly(A) position were excluded. A total of 5 different models were constructed using different training-test splits. The average R^2 for predicting the test sets was 0.6403. **Fig S4**: Validation of the random forest model (**A**) and the support vector machine model (**B**). The plot compared measured versus the predicted relative GFP abundance, with R^2 for the train and test sets included. **Fig S5**: Validation of the MLP-NN model’s ability to predict protein production in *S. cerevisiae*. A 5´ UTR library consisting of half a million 50-nt sequences was constructed previously in *S. cerevisiae* [7]. The impact of each 5´ UTR on *HIS3* production was assessed by measuring the enrichment of cells harboring the 5´ UTR after cultivation in selection media. From this library, a total of 700 5´ UTRs with poly(A) and 115 5´ UTRs without poly(A) were selected. Fifteen features were extracted from the 5´ UTRs, and the MLP-NN model was employed to predict enrichments of 5´ UTRs. The predicted enrichments were compared with measured enrichments, resulting in an R^2 of 0.503. **Fig S6**: The relationship between SHAP values and poly(A) features from models constructed using four additional training-test splits

Additional file **2**: **Table S1** List of plasmids. **Table S2** List of primers. **Table S3** Sequences of LHZ1138, LHZ1441 and LHZ1448

Additional file **3**: Sequences of 5´ UTRs

Additional file **4**: List of TPM values

Additional file **5**: List of emPAI values

Additional file **6**: Dataset for model training

Acknowledgements
We express our gratitude to Professor Qiang Huang from Fudan University for his valuable assistance in constructing the machine-learning models, and to Ling Huang from the proteomics platform at the State Key Laboratory of

Genetic Engineering for conducting the proteomics analysis. This research was financially supported by the National Natural Science Foundation of China [31970549, 92249302], Science and Technology Research Program of Shanghai [19DZ2282100], Open Fund of State Key Laboratory of Genetic Engineering [SKLGE-2318].

Author contributions

YY and HL planned the study design and supervised the experimental work. JZeng, JW and JZhou performed the experimental work. JZeng, KS, HW and TN analyzed the data. YY and JZeng wrote the manuscript. All authors read and approved the final manuscript.

Data Availability

The datasets supporting the conclusions of this article are included within the article and its Additional files.

Declarations

Competing interests

The authors declare no competing interests.

Received: 15 October 2023 / Accepted: 12 December 2023
Published online: 03 January 2024

References

1. De Nijs Y, De Maeseneire SL, Soetaert WK. 5´ untranslated regions: the next regulatory sequence in yeast synthetic biology. Biol Rev Camb Philos Soc. 2020;95(2):517–29. https://doi.org/10.1111/brv.12575
2. Hinnebusch AG, Ivanov IP, Sonenberg N. Translational control by 5´-untranslated regions of eukaryotic mRNAs. Science. 2016;352(6292):1413–6. https://doi.org/10.1126/science.aad9868
3. Shatsky IN, Terenin IM, Smirnova VV, Andreev DE. Cap-Independent translation: what's in a name? Trends Biochem Sci. 2018;43(11):882–95. https://doi.org/10.1016/j.tibs.2018.04.011
4. Kim Y, Lee G, Jeon E, Sohn EJ, Lee Y, Kang H, Lee DW, Kim DH, Hwang I. The immediate upstream region of the 5´-UTR from the AUG start codon has a pronounced effect on the translational efficiency in *Arabidopsis thaliana*. Nucleic Acids Res. 2014;42(1):485–98. https://doi.org/10.1093/nar/gkt864
5. Xu L, Liu P, Dai Z, Fan F, Zhang X. Fine-tuning the expression of pathway gene in yeast using a regulatory library formed by fusing a synthetic minimal promoter with different Kozak variants. Microb Cell Fact. 2021;20(1):148. https://doi.org/10.1186/s12934-021-01641-z
6. Sample PJ, Wang B, Reid DW, Presnyak V, McFadyen IJ, Morris DR, Seelig G. Human 5´ UTR design and variant effect prediction from a massively parallel translation assay. Nat Biotechnol. 2019;37(7):803–09. https://doi.org/10.1038/s41587-019-0164-5
7. Cuperus JT, Groves B, Kuchina A, Rosenberg AB, Jojic N, Fields S, Seelig G. Deep learning of the regulatory grammar of yeast 5´ untranslated regions from 500,000 random sequences. Genome Res. 2017;27(12):2015–24. https://doi.org/10.1101/gr.224964.117
8. Johnstone TG, Bazzini AA, Giraldez AJ. Upstream ORFs are prevalent translational repressors in vertebrates. EMBO J. 2016;35(7):706–23. https://doi.org/10.15252/embj.201592759
9. Spriggs KA, Stoneley M, Bushell M, Willis AE. Re-programming of translation following cell stress allows IRES-mediated translation to predominate. Biol Cell. 2008;100(1):27–38. https://doi.org/10.1042/BC20070098
10. King HA, Cobbold LC, Willis AE. The role of IRES trans-acting factors in regulating translation initiation. Biochem Soc Trans. 2010;38(6):1581–6. https://doi.org/10.1042/BST0381581
11. Li J, Liang Q, Song W, Marchisio MA. Nucleotides upstream of the Kozak sequence strongly influence gene expression in the yeast S. Cerevisiae. J Biol Eng. 2017;11:25. https://doi.org/10.1186/s13036-017-0068-1
12. Dvir S, Velten L, Sharon E, Zeevi D, Carey LB, Weinberger A, Segal E. Deciphering the rules by which 5´-UTR sequences affect protein expression in yeast. Proc Natl Acad Sci U S A. 2013;110(30):E2792–801. https://doi.org/10.1073/pnas.1222534110
13. Eisenhut P, Mebrahtu A, Moradi Barzadd M, Thalen N, Klanert G, Weinguny M, Sandegren A, Su C, Hatton D, Borth N, Rockberg J. Systematic use of synthetic 5´-UTR RNA structures to tune protein translation improves yield and quality of complex proteins in mammalian cell factories. Nucleic Acids Res. 2020;48(20):e119. https://doi.org/10.1093/nar/gkaa847
14. Calviello L, Venkataramanan S, Rogowski KJ, Wyler E, Wilkins K, Tejura M, Thai B, Krol J, Filipowicz W, Landthaler M, Floor SN. DDX3 depletion represses translation of mRNAs with complex 5´ UTRs. Nucleic Acids Res. 2021;49(9):5336–50. https://doi.org/10.1093/nar/gkab287
15. Hansel-Hertsch R, Beraldi D, Lensing SV, Marsico G, Zyner K, Parry A, Di Antonio M, Pike J, Kimura H, Narita M, et al. G-quadruplex structures mark human regulatory chromatin. Nat Genet. 2016;48(10):1267–72. https://doi.org/10.1038/ng.3662
16. Lo Giudice C, Zambelli F, Chiara M, Pavesi G, Tangaro MA, Picardi E, Pesole G. UTRdb 2.0: a comprehensive, expert curated catalog of eukaryotic mRNAs untranslated regions. Nucleic Acids Res. 2023;51(D1):D337–D44. https://doi.org/10.1093/nar/gkac1016
17. Gilbert WV, Zhou K, Butler TK, Doudna JA. Cap-independent translation is required for starvation-induced differentiation in yeast. Science. 2007;317(5842):1224–7. https://doi.org/10.1126/science.1144467
18. Wang J, Zhang X, Greene GH, Xu G, Dong X. PABP/purine-rich motif as an initiation module for cap-independent translation in pattern-triggered immunity. Cell. 2022;185(17):3186–200. https://doi.org/10.1016/j.cell.2022.06.037. e17.
19. Gudkov AT, Ozerova MV, Shiryaev VM, Spirin AS. 5´-poly(A) sequence as an effective leader for translation in eukaryotic cell-free systems. Biotechnol Bioeng. 2005;91(4):468–73. https://doi.org/10.1002/bit.20525
20. Shirokikh NE, Spirin AS. Poly(A) leader of eukaryotic mRNA bypasses the dependence of translation on initiation factors. Proc Natl Acad Sci U S A. 2008;105(31):10738–43. https://doi.org/10.1073/pnas.0804940105
21. Jia L, Mao Y, Ji Q, Dersh D, Yewdell JW, Qian SB. Decoding mRNA translatability and stability from the 5´ UTR. Nat Struct Mol Biol. 2020;27(9):814–21. https://doi.org/10.1038/s41594-020-0465-x
22. Bag J. Feedback inhibition of poly(A)-binding protein mRNA translation. A possible mechanism of translation arrest by stalled 40 S ribosomal subunits. J Biol Chem. 2001;276(50):47352–60. https://doi.org/10.1074/jbc.M107676200
23. Brandariz-Nunez A, Zeng F, Lam QN, Jin H. Sbp1 modulates the translation of Pab1 mRNA in a poly(A)- and RGG-dependent manner. RNA. 2018;24(1):43–55. https://doi.org/10.1261/rna.062547.117
24. Xia X, MacKay V, Yao X, Wu J, Miura F, Ito T, Morris DR. Translation initiation: a regulatory role for poly(A) tracts in front of the AUG codon in *Saccharomyces cerevisiae*. Genetics. 2011;189(2):469–78. https://doi.org/10.1534/genetics.111.132068
25. Ding W, Cheng J, Guo D, Mao L, Li J, Lu L, Zhang Y, Yang J, Jiang H. Engineering the 5´ UTR-Mediated regulation of protein abundance in yeast using nucleotide sequence Activity relationships. ACS Synth Biol. 2018;7(12):2709–14. https://doi.org/10.1021/acssynbio.8b00127
26. Decoene T, Peters G, De Maeseneire SL, De Mey M. Toward predictable 5´UTRs in *Saccharomyces cerevisiae*: development of a yUTR Calculator. ACS Synth Biol. 2018;7(2):622–34. https://doi.org/10.1021/acssynbio.7b00366
27. Cao J, Novoa EM, Zhang Z, Chen WCW, Liu D, Choi GCG, Wong ASL, Wehrspaun C, Kellis M, Lu TK. High-throughput 5´ UTR engineering for enhanced protein production in non-viral gene therapies. Nat Commun. 2021;12(1):4138. https://doi.org/10.1038/s41467-021-24436-7
28. Washburn JD, Mejia-Guerra MK, Ramstein G, Kremling KA, Valluru R, Buckler ES, Wang H. Evolutionarily informed deep learning methods for predicting relative transcript abundance from DNA sequence. Proc Natl Acad Sci U S A. 2019;116(12):5542–49. https://doi.org/10.1073/pnas.1814551116
29. Zrimec J, Borlin CS, Buric F, Muhammad AS, Chen R, Siewers V, Verendel V, Nielsen J, Topel M, Zelezniak A. Deep learning suggests that gene expression is encoded in all parts of a co-evolving interacting gene regulatory structure. Nat Commun. 2020;11(1):6141. https://doi.org/10.1038/s41467-020-19921-4
30. Vaishnav ED, de Boer CG, Molinet J, Yassour M, Fan L, Adiconis X, Thompson DA, Levin JZ, Cubillos FA, Regev A. The evolution, evolvability and engineering of gene regulatory DNA. Nature. 2022;603(7901):455–63. https://doi.org/10.1038/s41586-022-04506-6
31. Lane MM, Morrissey JP. *Kluyveromyces marxianus*: a yeast emerging from its sister's shadow. Fungal Biology Reviews. 2010;24(1–2):17–26. https://doi.org/10.1016/j.fbr.2010.01.001
32. Leonel LV, Arruda PV, Chandel AK, Felipe MGA, Sene L. *Kluyveromyces marxianus*: a potential biocatalyst of renewable chemicals and lignocellulosic ethanol production. Crit Rev Biotechnol. 2021;41(8):1131–52. https://doi.org/10.1080/07388551.2021.1917505

33. Baptista M, Domingues L. *Kluyveromyces marxianus* as a microbial cell factory for lignocellulosic biomass valorisation. Biotechnol Adv. 2022;60:108027. https://doi.org/10.1016/j.biotechadv.2022.108027
34. Zhou J, Zhu P, Hu X, Lu H, Yu Y. Improved secretory expression of lignocellulolytic enzymes in *Kluyveromyces marxianus* by promoter and signal sequence engineering. Biotechnol Biofuels. 2018;11:235. https://doi.org/10.1186/s13068-018-1232-7
35. Liu B, Wu P, Zhou J, Yin A, Yu Y, Lu H. Characterization and optimization of the LAC4 upstream region for low-leakage expression in *Kluyveromyces marxianus*. Yeast. 2022;39(4):283–96. https://doi.org/10.1002/yea.3682
36. Wu P, Zhou J, Yu Y, Lu H. Characterization of essential elements for improved episomal expressions in *Kluyveromyces marxianus*. Biotechnol J. 2022;17(4):e2100382. https://doi.org/10.1002/biot.202100382
37. Shi T, Zhou J, Xue A, Lu H, He Y, Yu Y. Characterization and modulation of endoplasmic reticulum stress response target genes in *Kluyveromyces marxianus* to improve secretory expressions of heterologous proteins. Biotechnol Biofuels. 2021;14(1):236. https://doi.org/10.1186/s13068-021-02086-7
38. Yu Y, Mo W, Ren H, Yang X, Lu W, Luo T, Zeng J, Zhou J, Qi J, Lu H. Comparative genomic and transcriptomic analysis reveals specific features of Gene Regulation in *Kluyveromyces marxianus*. Front Microbiol. 2021;12:598060. https://doi.org/10.3389/fmicb.2021.598060
39. Zhao S, Ye Z, Stanton R. Misuse of RPKM or TPM normalization when comparing across samples and sequencing protocols. RNA. 2020;26(8):903–09. https://doi.org/10.1261/rna.074922.120
40. Wisniewski JR, Zougman A, Nagaraj N, Mann M. Universal sample preparation method for proteome analysis. Nat Methods. 2009;6(5):359–62. https://doi.org/10.1038/nmeth.1322
41. Kovalchuk SI, Jensen ON, Rogowska-Wrzesinska A, FlashPack. Fast and simple Preparation of Ultrahigh-performance Capillary columns for LC-MS. Mol Cell Proteomics. 2019;18(2):383–90. https://doi.org/10.1074/mcp.TIR118.000953
42. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS. Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis. 1999;20(18):3551–67. https://doi.org/10.1002/(SICI)1522-2683(19991201)20:18<3551::AID-ELPS3551>3.0.CO;2-2
43. Ishihama Y, Oda Y, Tabata T, Sato T, Nagasu T, Rappsilber J, Mann M. Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. Mol Cell Proteomics. 2005;4(9):1265–72. https://doi.org/10.1074/mcp.M500061-MCP200
44. Bradshaw E, Saalbach G, McArthur M. Proteomic survey of the *Streptomyces coelicolor* nucleoid. J Proteom. 2013;83:37–46. https://doi.org/10.1016/j.jprot.2013.02.033
45. Carvalhais V, Franca A, Pier GB, Vilanova M, Cerca N, Vitorino R. Comparative proteomic and transcriptomic profile of *Staphylococcus epidermidis* biofilms grown in glucose-enriched medium. Talanta. 2015;132:705–12. https://doi.org/10.1016/j.talanta.2014.10.012
46. Ma G, Wang P, Yang Y, Wang W, Ma J, Zhou L, Ouyang J, Li R, Zhang S. emPAI-assisted strategy enhances screening and assessment of *Mycobacterium tuberculosis* Infection serological markers. Microb Biotechnol. 2021;14(4):1827–38. https://doi.org/10.1111/1751-7915.13829
47. Miksik I, Ergang P, Pacha J. Proteomic analysis of chicken eggshell cuticle membrane layer. Anal Bioanal Chem. 2014;406(29):7633–40. https://doi.org/10.1007/s00216-014-8213-x
48. Muccilli V, Saletti R, Cunsolo V, Ho J, Gili E, Conte E, Sichili S, Vancheri C, Foti S. Protein profile of exhaled breath condensate determined by high resolution mass spectrometry. J Pharm Biomed Anal. 2015;105:134–49. https://doi.org/10.1016/j.jpba.2014.11.050
49. Burke D, Dawson D, Stearns T, Cold Spring Harbor Laboratory. Methods in yeast genetics: a Cold Spring Harbor Laboratory course manual. 2000 ed. Plainview, N.Y.: Cold Spring Harbor Laboratory Press; 2000.
50. Ali SE, Mittal A, Mathews DH. RNA secondary structure analysis using RNAstructure. Curr Protoc. 2023;3(7):e846. https://doi.org/10.1002/cpz1.846
51. Lundberg SM, Lee SI. A Unified Approach to Interpreting Model Predictions. In 31st Annual Conference on Neural Information Processing Systems (NIPS); Dec 04–09; Long Beach, CA. Neural Information Processing Systems (Nips); 2017.
52. Saini P, Beniwal A, Vij S. Comparative analysis of oxidative stress during aging of *Kluyveromyces marxianus* in Synthetic and Whey Media. Appl Biochem Biotechnol. 2017;183(1):348–61. https://doi.org/10.1007/s12010-017-2449-9
53. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science. 2008;320(5881):1344–9. https://doi.org/10.1126/science.1158441
54. Ishihama Y, Schmidt T, Rappsilber J, Mann M, Hartl FU, Kerner MJ, Frishman D. Protein abundance profiling of the *Escherichia coli* cytosol. BMC Genomics. 2008;9:102. https://doi.org/10.1186/1471-2164-9-102
55. Hamilton R, Watanabe CK, de Boer HA. Compilation and comparison of the sequence context around the AUG startcodons in *Saccharomyces cerevisiae* mRNAs. Nucleic Acids Res. 1987;15(8):3581–93. https://doi.org/10.1093/nar/15.8.3581
56. Churkin A, Weinbrand L, Barash D. Free energy minimization to predict RNA secondary structures and computational RNA design. Methods Mol Biol. 2015;1269:3–16. https://doi.org/10.1007/978-1-4939-2291-8_1
57. Melo EO, de Melo Neto OP, Martins de Sa C. Adenosine-rich elements present in the 5´-untranslated region of PABP mRNA can selectively reduce the abundance and translation of CAT mRNAs in vivo. FEBS Lett. 2003;546(2–3):329–34. https://doi.org/10.1016/s0014-5793(03)00620-3
58. LeProust EM, Peck BJ, Spirin K, McCuen HB, Moore B, Namsaraev E, Caruthers MH. Synthesis of high-quality libraries of long (150mer) oligonucleotides by a novel depurination controlled process. Nucleic Acids Res. 2010;38(8):2522–40. https://doi.org/10.1093/nar/gkq163
59. Verardo D, Adelizzi B, Rodriguez-Pinzon DA, Moghaddam N, Thomee E, Loman T, Godron X, Horgan A. Multiplex enzymatic synthesis of DNA with single-base resolution. Sci Adv. 2023;9(27):eadi0263. https://doi.org/10.1126/sciadv.adi0263
60. Passmore LA, Coller J. Roles of mRNA poly(A) tails in regulation of eukaryotic gene expression. Nat Rev Mol Cell Biol. 2022;23(2):93–106. https://doi.org/10.1038/s41580-021-00417-y
61. Collart MA, Panasenko OO. The Ccr4–not complex. Gene. 2012;492(1):42–53. https://doi.org/10.1016/j.gene.2011.09.033
62. Rojas DA, Urbina F, Valenzuela-Perez L, Leiva L, Miralles VJ, Maldonado E. Initiator-Directed transcription: fission yeast nmtl initiator directs preinitiation complex formation and transcriptional initiation. Genes (Basel). 2022;13(2). https://doi.org/10.3390/genes13020256
63. Ryczek N, Lys A, Makalowska I. The Functional Meaning of 5´UTR in Protein-Coding Genes. Int J Mol Sci. 2023;24(3). https://doi.org/10.3390/ijms24032976
64. Heck AM, Wilusz J. The interplay between the RNA decay and Translation Machinery in Eukaryotes. Cold Spring Harb Perspect Biol. 2018;10(5). https://doi.org/10.1101/cshperspect.a032839
65. Bae H, Coller J. Codon optimality-mediated mRNA degradation: linking translational elongation to mRNA stability. Mol Cell. 2022;82(8):1467–76. https://doi.org/10.1016/j.molcel.2022.03.032
66. Mayr C. What Are 3' UTRs Doing? Cold Spring Harb Perspect Biol. 2019;11(10). https://doi.org/10.1101/cshperspect.a034728
67. Mauger DM, Cabral BJ, Presnyak V, Su SV, Reid DW, Goodman B, Link K, Khatwani N, Reynders J, Moore MJ, McFadyen IJ. mRNA structure regulates protein expression through changes in functional half-life. Proc Natl Acad Sci U S A. 2019;116(48):24075–83. https://doi.org/10.1073/pnas.1908052116
68. Fischer JW, Busa VF, Shao Y, Leung AKL, Structure-Mediated RNA. Decay by UPF1 and G3BP1. Mol Cell. 2020;78(1):70–84e6. https://doi.org/10.1016/j.molcel.2020.01.021
69. Xia X, Holcik M. Strong eukaryotic IRESs have weak secondary structure. PLoS ONE. 2009;4(1):e4136. https://doi.org/10.1371/journal.pone.0004136
70. Pelechano V, Wei W, Steinmetz LM. Extensive transcriptional heterogeneity revealed by isoform profiling. Nature. 2013;497(7447):127–31. https://doi.org/10.1038/nature12121
71. Staley CA, Huang A, Nattestad M, Oshiro KT, Ray LE, Mulye T, Li ZH, Le T, Stephens JJ, Gomez SR, et al. Analysis of the 5´ untranslated region (5´UTR) of the alcohol oxidase 1 (AOX1) gene in recombinant protein expression in *Pichia pastoris*. Gene. 2012;496(2):118–27. https://doi.org/10.1016/j.gene.2012.01.006

Publisher's Note