Briefings in Bioinformatics, 22(5), 2021, 1–11

https://doi.org/10.1093/bib/bbaa412
Problem Solving Protocol

Riboexp: an interpretable reinforcement learning framework for ribosome density modeling

Hailin Hu†, Xianggen Liu†, An Xiao, YangYang Li, Chengdong Zhang, Tao Jiang, Dan Zhao, Sen Song and Jianyang Zeng

Corresponding authors: Jianyang Zeng, Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing, 100084, China. Email: zengjy321@tsinghua.edu.cn and Sen Song, Laboratory for Brain and Intelligence and Department of Biomedical Engineering, Tsinghua University, Beijing, 100084, China; Beijing Innovation Center for Future Chip, Tsinghua University, Beijing, 100084, China. E-mail: songsen@mail.tsinghua.edu.cn.
†These authors contributed equally to this work.

Abstract

Translation elongation is a crucial phase during protein biosynthesis. In this study, we develop a novel deep reinforcement learning-based framework, named Riboexp, to model the determinants of the uneven distribution of ribosomes on mRNA transcripts during translation elongation. In particular, our model employs a policy network to perform a context-dependent feature selection in the setting of ribosome density prediction. Our extensive tests demonstrated that Riboexp can significantly outperform the state-of-the-art methods in predicting ribosome density by up to 5.9% in terms of per-gene Pearson correlation coefficient on the datasets from three species. In addition, Riboexp can indicate more informative sequence features for the prediction task than other commonly used attribution methods in deep learning. In-depth analyses also revealed the meaningful biological insights generated by the Riboexp framework. Moreover, the application of Riboexp in codon optimization resulted in an increase of protein production by around 31% over the previous state-of-the-art method that models ribosome density. These results have established Riboexp as a powerful and useful computational tool in the studies of translation dynamics and protein synthesis. **Availability:** The data and code of this study are available on GitHub: https://github.com/Liuxg16/Riboexp. **Contact:** zengjy321@tsinghua.edu.cn; songsen@tsinghua.edu.cn

Key words: reinforcement learning; ribosome profiling; translation elongation.

Introduction

Translation elongation plays an essential role in protein biogenesis [1, 2] and has been increasingly recognized to associate with human diseases [3, 4]. However, despite numerous research effort [5–13], the underlying mechanisms for the regulation of translation elongation still remain largely elusive. In particular, little research effort has been devoted to modeling the functional roles of the same codon species under different contexts and understanding the complex mechanisms of sequence determinants of local ribosome elongation velocity.

Hailin Hu received his Ph.D. degree in 2019 from School of Medicine at Tsinghua University. His research focuses on bioinformatics and machine learning.
Xianggen Liu is currently a Ph.D. student at Tsinghua University. He mainly focuses on neural networks and machine learning, and their applications in bioinformatics.
An Xiao obtained his Master degree at Tsinghua University. His main research interests include bioinformatics and deep learning.
Yangyang Li is currently a Ph.D. student at Tsinghua University. He mainly focuses on infectious disease and cancer immunology.
Chengdong Zhang received his Ph.D degree at the University of New South Wales. He is the CEO of Beijing Biorise Biotechnology Co., Ltd. His research interests are microbial genetics and enzymology.
Tao Jiang is a distinguished professor at University of California, Riverside. He is an expert in computational molecular biology and machine learning.
Dan Zhao is an assistant research fellow at Tsinghua University. Her research interests mainly include structural biology and epigenetics.
Sen Song received his Ph.D. degree in Biology from Brandeis University. He is a professor at Tsinghua University. His research interests include computational neuroscience, neuroinformatics and machine learning.
Jianyang Zeng is an associate professor at Tsinghua University. His research interests focus on computational biology and machine learning.
Submitted: 17 November 2020; **Received (in revised form):** 11 December 2020

In recent years, the accumulation of ribosome profiling data [14, 15] has provided an unprecedented opportunity for the applications of data-driven methods, especially machine learning-based approaches, in understanding the regulation of translation elongation *in vivo*. In a typical ribosome profiling experiment, the mRNA fragments protected by ribosomes are captured and sequenced in a high-throughput manner. Then, the corresponding A-site codons (i.e. the binding sites for the t-RNA molecules charged with amino acids) of these sequenced fragments are identified with a predefined offset from the ends of the fragments, to derive a quantitative measurement of ribosome density at the codon resolution. Based on these data, several computational models, e.g. ROSE [16], RiboShape [17] and RUST [18], have been proposed to predict the distributions of ribosome footprints along the mRNA transcripts. More recently, Tunney et al. [19] have proposed a new approach, called iχnos, to predict ribosome density from an input mRNA sequence window through a feed forward neural network, which shows superior prediction performance over the previous methods and can be applied to modulate protein expression through sequence design. However, few of these methods explicitly model the relative importance of different sequence features, e.g. codon types, in a given mRNA context. Instead, they tend to provide a global measurement to evaluate the average effect of each feature on local ribosome density.

In this study, we develop a deep reinforcement learning-based framework, named Riboexp, to model ribosome density from mRNA sequences, uncover the corresponding critical sequence features, and rationalize translation elongation dynamics at codon level. From a biological perspective, we assume that the relevance of each sequence element, e.g. codon or nucleotide, to the local translation elongation rate depends on the specific sequence context that it lies in. Therefore, in contrast to the previous prediction methods that feed the whole sequence context into a universal regressor, Riboexp performs a context-dependent feature selection procedure for each sequence input, aiming to select the most relevant sequence features, also called rationales, for this specific learning task. Based on the generated rationales, a predictor is subsequently trained to predict ribosome density. As the data labels do not provide any explicit information on the importance of individual codon positions to translation elongation, the training process of Riboexp is conducted by reinforcement learning in a trial and error manner through the coordination of a carefully designed reward system, which explicitly optimizes both quality and sparsity of the selected rationales.

The advantage of introducing the context-dependent feature selection scheme into our deep learning framework is two-fold. On one hand, the sparsity constraint forces the model to select the most relevant sequence features for prediction and therefore boosts the model performance, which in fact adopts the same philosophy as in current sparsity regularization based statistical learning approaches [20]. On the other hand, the explicit selection of features also greatly enhances the explainability of our deep learning model, enabling the application of Riboexp for mining potentially intriguing biological insights into translation elongation dynamics.

In this work, extensive tests on currently available high-quality ribosome profiling data from three species (i.e. yeast, human and *E. coli*) have demonstrated that Riboexp can significantly outperform the current state-of-the-art method as well as other commonly used deep learning architectures in ribosome density prediction. In addition, we have shown that Riboexp can indicate more informative sequence features for modeling ribosome density. Further, genome-wide analyses of unigram (i.e. single codon) and bigram (i.e. two contiguous codons) profiles of the selected sequence features have also revealed several interesting findings related to translation elongation, further establishing the biological relevance of our model. Last but not least, the advance of our approach in prediction performance can also lead to a more efficient modulation of protein expression, as demonstrated by an increase of nearly 31% in protein production when compared to the previous state-of-the-art method [19]. These results have demonstrated that Riboexp can serve as a powerful tool to explore the sequence features of translation elongation dynamics and analyze various translation elongation-related processes.

Methods

In our study, we present an interpretable framework called Riboexp that models ribosome density by effectively incorporating contextual sequence information that shapes the translation elongation dynamics (Figure 1). The Riboexp framework comprises two deep neural networks, namely the policy network (PolicyNet) [21] and the Predictor. The PolicyNet sequentially makes decisions on whether to select the given codons as rationales and the Predictor predicts the ribosome density based on the selected rationales. Below, we will first define the ribosome density modeling task and then elaborate the components of Riboexp and the optimization objectives.

Definition of the ribosome density modeling task

The traditional ribosome density modeling task is defined as the estimation of the ribosome density for a codon of interest given its contextual sequence [19]. Here, we consider a more difficult task: given a codon contextual sequence $x_1, x_2, \ldots, x_T$ (where T stands for the number of codons) of a target codon at A site and its corresponding experimentally measured ribosome density $y \in \mathbb{R}$, the goal is to not only predict the ribosome density $\hat{y} \in \mathbb{R}$ of the target codon, but also generate a binary integer sequence $s_1, s_2, \ldots, s_T$, indicating which codons along the contextual sequence are relevant for the prediction (1 means relevant and 0 otherwise). More specifically, in our learning task, we aim to minimize the difference between $y \in \mathbb{R}$ and $\hat{y} \in \mathbb{R}$, while achieving a sparse binary integer sequence $\mathbf{s}$ as possible.

Generation of context dependent rationales by the PolicyNet

At each input codon position, the PolicyNet takes an action to decide whether to select the codon information as a feature for the Predictor. Inside the PolicyNet are two modules: (1) a bidirectional recurrent neural network (BiRNN) [22] that integrates features over the entire history by extracting information from both backward and forward states simultaneously; and (2) a recurrent neural network (RNN) with feedback that makes specific decision at each position based on the corresponding hidden states of the BiRNN.

To provide an appropriate representation of the sequence context of translation dynamics, for each given codon at an A site (denoted as position 0), we first extract the codon sequence spanning from positions -5 to +4 as the input to our framework (Supplementary Figure 1). Then, we adopt one-hot encoding of the codon types, nucleotide types and the position information of codons to represent the features of the contextual

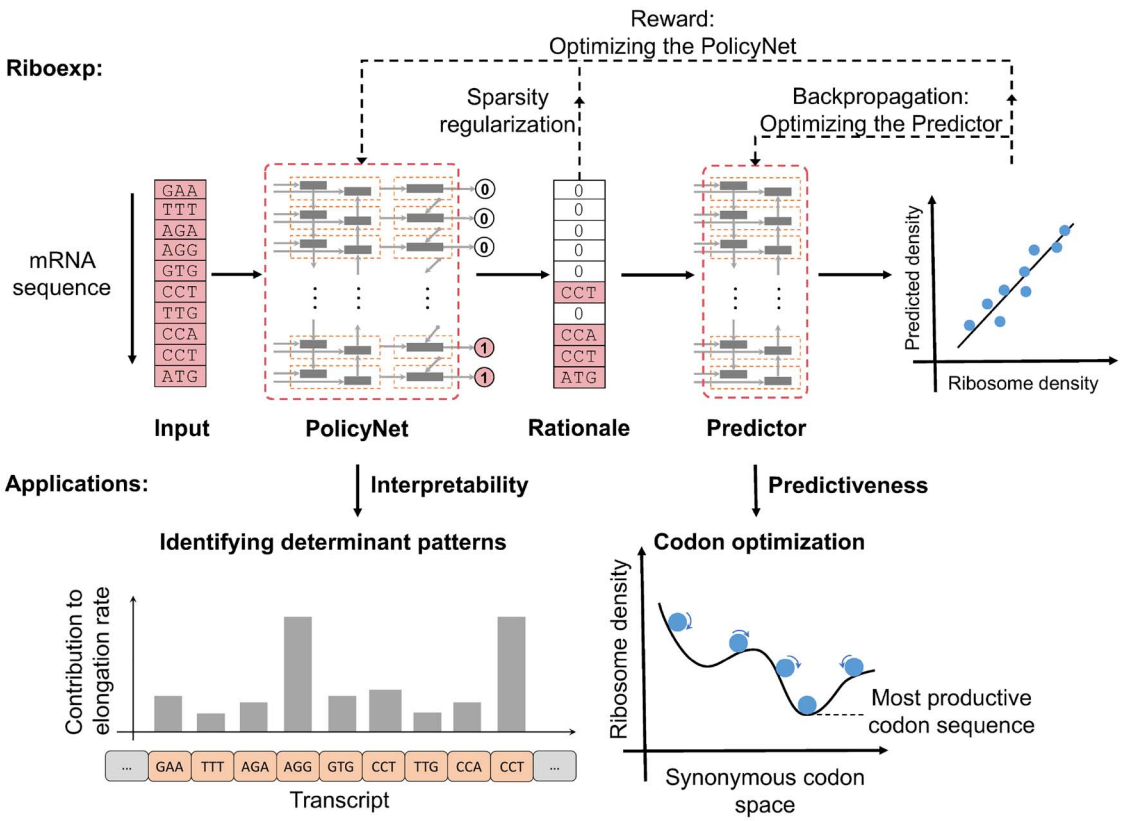


Figure 1. An schematic overview of the Riboexp pipeline and its applications.

sequence $x_1, x_2, \ldots, x_T$, denoted by $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T) \in \mathbb{R}^{T \times D}$ (D indicates dimension size of the features for each codon, Supplementary Notes). Then, a BiRNN takes the encoded features of the input codon sequence as input to capture the context-dependent information. Formally, the BiRNN takes a sequence of codon features recurrently and then concatenates the two hidden states $\overrightarrow{\boldsymbol{h}}_t$ and $\overleftarrow{\boldsymbol{h}}_t$. That is, the hidden state $\hat{\boldsymbol{h}}_t$ of the BiRNN corresponding to the time step t is defined as

$$\hat{\boldsymbol{h}}_t = BiRNN_{\Theta_\alpha}(\boldsymbol{X}, t) \tag{1}$$

$$= [\overrightarrow{\boldsymbol{h}}_t, \overleftarrow{\boldsymbol{h}}_t] = [\overrightarrow{RNN}_{\Theta_\alpha}(\boldsymbol{x}_t, \overrightarrow{\boldsymbol{h}}_{t-1}), \overleftarrow{RNN}_{\Theta_\alpha}(\boldsymbol{x}_t, \overleftarrow{\boldsymbol{h}}_{t+1})] \tag{2}$$

where Θ_α denotes the set of parameters in the PolicyNet, $\overrightarrow{\boldsymbol{h}}_t \in \mathbb{R}^{D_h}$ and $\overleftarrow{\boldsymbol{h}}_t \in \mathbb{R}^{D_h}$ stand for the hidden states of the two recurrent neural networks $\overrightarrow{RNN}_{\Theta_\alpha}$ and $\overleftarrow{RNN}_{\Theta_\alpha}$ for the two opposite directions at the time step t, respectively, and D_h stands for the dimension of individual hidden states. We use the gated recurrent unit (GRU) [23] as our recurrent unit in the BiRNN.

Based on the hidden states of the BiRNN, inspired by Lei et al. [24], we leverage another variant of RNN (i.e. RNN with feedback) to generate actions orderly at each position, indicating whether to select the corresponding codon as a rationale. We use a series of binary variables $\boldsymbol{s} = (s_1, s_2, \ldots, s_T)$ to indicate the rationales of the input codon sequence. We model the above rationale generation procedure as a partially observable Markov decision process [25] that can be solved through reinforcement learning. In other words, the determination of the next action only depends on the current state and the previous action history [21]. More formally, the probability $P(\boldsymbol{s}|\boldsymbol{X})$ for generating the context-dependent rationale $\boldsymbol{s}$ is given by

$$P(\boldsymbol{s}|\boldsymbol{X}) = \prod_{t=1}^{T} P(s_t|\boldsymbol{x}_t, \boldsymbol{s}_{1\sim t-1}), \tag{3}$$

where $\boldsymbol{s}_{1\sim t-1}$ means the variables from s_1 to s_{t-1} and $P(s_t|\boldsymbol{x}_t, \boldsymbol{s}_{1\sim t-1})$ is implemented by an RNN with feedback (Supplementary Notes). Following the common practice in current deep reinforcement learning frameworks [21, 26], during training we sample an action s_t from the predicted policy distribution, while for testing, we choose an action according to the maximum a posteriori probability, i.e. $s_t = argmax\ \boldsymbol{p}(s_t|\boldsymbol{X}, \boldsymbol{s}_{1\sim t-1})$.

Once the action sequence $\boldsymbol{s}$ is generated, the corresponding rationales are determined. We combine the actions and the input codon sequence to derive a rationale matrix $\boldsymbol{R} \in \mathbb{R}^{T \times D}$, that is,

$$\boldsymbol{R}_t = s_t \boldsymbol{x}_t. \tag{4}$$

Estimating ribosome density with rationales by the Predictor

The Predictor models the distribution of ribosomes along the transcripts as a regression function given the selected rationales. The Predictor possesses three submodules: (1) A BiRNN (i.e. $BiRNN^c_{\Theta_\beta}$) that learns the interdependencies and patterns of rationales from the input rationale matrix $\boldsymbol{R}$; (2) a max-pooling layer that selects the maximum value for each dimension of hidden states over all the time steps to perform dimension reduction; and (3) a fully connected layer that predicts the ribosome density based on the extracted features. The detailed

computations of the predicted ribosome density $\hat{y}$, denoted by $Predictor_{\Theta_\beta}(\mathbf{R})$, are given by

$$\mathbf{c}_k = BiRNN^c_{\Theta_\beta}(\mathbf{R}, k), \tag{5}$$

$$u_i = max\{\mathbf{c}_{1,i}, \mathbf{c}_{2,i}, \ldots, \mathbf{c}_{T,i}\}, \tag{6}$$

$$\hat{y} = Predictor_{\Theta_\beta}(\mathbf{R}) = \boldsymbol{w}_\beta \boldsymbol{u} + b_\beta, \tag{7}$$

where $\mathbf{c}_k \in \mathbb{R}^{2D_c}$ stands for the vector of hidden states at step k output by $BiRNN^c_{\Theta_\beta}$, D_c stands for the hidden size, $\boldsymbol{w}_\beta \in \mathbb{R}^{1\times 2D_c}$ and $b_\beta \in \mathbb{R}$ stand for the weights and the bias term for the fully connected layer, respectively, and Θ_β stands for the set of parameters in the Predictor. If the RNA folding energy is provided, we derive the RNA folding energy (denoted as $E_{\text{fold}} \in \mathbb{R}^{T\times 3}$) corresponding to every 3T-nucleotide (T-codon) sequence window along each mRNA transcript, following the same protocol used in the previous studies [19, 27]. We then use another BiRNN module similar to $BiRNN^c_{\Theta_\beta}$ to extract the features and feed them to the last layer of Predictor (see Supplementary Notes 1.4 for more details).

Optimization objectives

Our framework follows the basic architecture of an actor-critic algorithm [28] in reinforcement learning, in which one component (i.e. the PolicyNet) is designed to generate action sequences and the other (i.e. the Predictor) produces rewards according to the previously generated action sequences to prompt superior actions in the next trials. During the training process, the Predictor is optimized through the standard back-propagation strategy, while the PolicyNet is trained through the reinforcement learning algorithm to maximize the expected rewards.

Formally, the Predictor can be optimized via minimizing the expected mean square error over the distribution of the selected rationales, i.e.

$$\min_{\Theta_\beta} \sum_{\mathbf{X},y\in\mathcal{D}} \mathbb{E}_{\mathbf{R}\sim p(\cdot|\mathbf{X})}\, \mathcal{L}(\mathbf{X}, \mathbf{R}, y) = \tag{8}$$

$$\min_{\Theta_\beta} \sum_{\mathbf{X},y\in\mathcal{D}} \mathbb{E}_{\mathbf{R}\sim p(\cdot|\mathbf{X})} \left[(Predictor_{\Theta_\beta}(\mathbf{R}) - y)^2\right], \tag{9}$$

where $\mathcal{D}$ stands for the collection of all training instances, $\mathcal{L}$ stands for the loss function.

On the other hand, the training of the PolicyNet is challenged by the lack of direct supervision information on the importance of individual codon positions from the training samples. To handle this issue, we apply a reinforcement learning strategy to train the PolicyNet. Intuitively, we seek for a reward function to encourage the PolicyNet to generate sparse and the most relevant rationales, which are expected to represent the most relevant biological determinants of translation elongation dynamics. In particular, we leverage the $L1$ norm to constrain the feature vector $\mathbf{s}$ for the sake of sparsity and use the mean square error to assess the quality of the rationales. Therefore, the reward function is defined as

$$R_{final}(\mathbf{X}, \mathbf{R}, y) = -\mathcal{L}(\mathbf{X}, \mathbf{R}, y) - \lambda||\mathbf{s}||_1, \tag{10}$$

where $||\cdot||_1$ stands for the $L1$ norm, and λ stands for the penalty factor of sparsity. Similar to Ranzato et al. [29], we use the REINFORCE algorithm [30] to enforce the PolicyNet to generate action sequences (i.e. $\mathbf{s}$) toward the maximum of the expected rewards and subtract the rewards by a baseline term (i.e. the average over multiple samples) to reduce the variance. But different from these methods, our framework does not utilize the stepwise supervision, i.e. we do not have feature importance annotation during the training process. More optimization details are provided in Supplementary Notes.

Results

Riboexp accurately predicts ribosome density

To evaluate the performance of Riboexp in predicting ribosome density, we first trained and tested our model using a high-quality yeast ribosome profiling dataset [31]. We chose the top 500 genes (Supplementary Figure 2) with the highest ribosome footprint density values as in Tunney et al. [19], and further conducted a series of 3-fold cross validation tests to evaluate the prediction performance of each method. In each fold, we randomly selected two-thirds of the genes as the training set and one-third of the genes as the test set.

We first compared our method with a recently published state-of-the-art method, namely iχnos [19], which adopted a feed-forward neural network to predict ribosome density from local sequence features (Supplementary Notes). Expectedly, with the more elaborate feature selection and modeling schemes, Riboexp achieved significant improvement in the regression performance of ribosome density, with an increase of 5.9% in terms of the average per-gene Pearson correlation coefficient on the yeast dataset where the RNA folding energies were utilized ($P<0.01$ by paired sample t-tests with Bonferroni correction, Table 1). In addition, we also showed that the Riboexp framework can outperform several other deep learning-based approaches, including the convolutional neural network (denoted by pre-Riboexp (CNN)), as well as the bidirectional gated recurrent neural network (BiGRU) with the attention mechanism (denoted by pre-Riboexp (Att)), which employs a soft attention mechanism to derive the contribution of each input codon and incorporates it into the prediction process (Supplementary Notes). The best hyperparameters of all methods were calibrated through a grid search procedure (Supplementary Table 1). Notably, although all of these new deep learning-based approaches showed an increase in prediction performance, Riboexp still achieved the best performance ($P<0.01$ by paired sample t-test with Bonferroni correction, Table 1). We conjectured that the superiority of Riboexp in prediction performance mainly stemmed from two aspects. On one hand, Riboexp exploited a powerful neural network (i.e. BiGRU) that has sufficient expressive capacity to learn the deep representations of the input. On the other hand, the PolicyNet allowed an efficient filtering of the input information for the prediction by a binary feature selection scheme, which can restrict the overfitting of the Predictor, a phenomenon commonly encountered when modeling the noisy ribosome profiling data [32].

As the aforementioned experiments were conducted mainly on the top 500 genes with the highest ribosome densities, to ensure that the prediction performance of our model can also be generalized to most of the genes *in vivo*, we also examined the prediction performance on all the other 4342 genes that were also detected in the ribosome profiling experiment [12] and had at least 200 read counts (denoted as the remaining genes). Consistent with the previous report [19], all the methods showed a decrease in prediction performance (Table 1), mainly due to the decreased data quality caused by lower transcript

Table 1. Comparison of prediction performance of Riboexp with that of different baseline methods in terms of average per-gene Pearson correlation coefficient on the datasets from three species. For each dataset, we performed ten 3-fold cross-validation tests on the high-density genes to obtain the average performance (n=30). For the test on the remaining genes, the final predictions of test samples were obtained by averaging the three models trained from individual folds. Since RNA folding energies of genes were not provided in the human and *E. coli* datasets, the structural information (abbreviated as struc. info.) was not included for the tests on these two datasets.

Dataset	Genes	Numbers	Methods					
			Riboshape	RUST	iχ nos	pre-Riboexp (CNN)	pre-Riboexp (Att)	Riboexp
Yeast (w/ struc. info.)	High-density genes	500	-	-	0.567 ± 0.001	0.599 ± 0.002	0.581 ± 0.002	**0.626** ± 0.002
	Remaining genes	4342	-	-	0.459 ± 0.002	0.471 ± 0.003	0.462 ± 0.002	**0.498** ± 0.003
Yeast (w/o struc. info.)	High-density genes	500	0.367 ± 0.002	0.483	0.552 ± 0.002	0.588 ± 0.003	0.565 ± 0.002	**0.604** ± 0.004
	Remaining genes	4342	0.363 ± 0.002	0.427	0.454 ± 0.002	0.457 ± 0.001	0.469 ± 0.003	**0.493** ± 0.002
Human	High-density genes	500	0.220 ± 0.004	0.252	0.445 ± 0.004	0.462 ± 0.004	0.438 ± 0.004	**0.468** ± 0.004
	Remaining genes	1081	0.190 ± 0.003	0.201	0.376 ± 0.002	0.404 ± 0.002	0.405 ± 0.003	**0.412** ± 0.004
E. coli	High-density genes	500	0.487 ± 0.003	-	0.741 ± 0.002	0.749 ± 0.004	0.751 ± 0.004	**0.770** ± 0.003
	Remaining genes	1375	0.448 ± 0.002	-	0.564 ± 0.004	0.585 ± 0.002	0.589 ± 0.003	**0.611** ± 0.002

abundance. Nevertheless, we found that even on this dataset, Riboexp still yielded significantly better regression performance than other deep learning-based methods and iχnos ($P<0.01$ by paired sample t-test with Bonferroni correction), indicating its better generalizability. In addition, we compared the distributions of per-gene Pearson correlation coefficients generated by different approaches and further confirmed the advantage of Riboexp (Supplementary Figure 3).

As an ablation study, we investigated the effect of RNA structural features in the ribosome density prediction and found that the performance of all the models was consistently decreased when the structural information was removed from the models' input. This observation indicated that RNA structural features are informative for ribosome density modeling, which also coincided with the previous reports [33]. Under this setting, we also compared our model with two previous non-neural-network based methods, namely RUST [18] and Riboshape [17], which do not use structural information as input. Our results showed that Riboexp still significantly outperformed iχnos and other baseline models for the case without using structural information.

In addition to the aforementioned yeast dataset, we also evaluated the performance of each model on two additional datasets derived from different species, namely human HEK293T cell [34] and *E. coli* cells [35]. The 3-fold cross-validation tests (Table 1) further demonstrated the superiority of our framework. As Riboexp employs a more complicated computation paradigm than previous models, we also tested its computation demand and confirmed the feasibility of our approach using ordinary computation resource (Supplementary Notes, Supplementary Table 2).

Riboexp outperforms other methods in feature attribution

As the introduction of the reinforcement learning-based feature selection scheme significantly improved the performance of ribosome density prediction, we next tried to examine the quality of selected features, i.e. rationales, in our framework. We also compared our feature selection strategy with two families of commonly used attribution methods. The first family is the back-propagation-based methods, represented by integrated gradients [36], which do not require the modification of the original network and only rely on the back-propagation procedure to pass the gradients or activation signals in a readily trained neural network. Here, the integrated gradients were computed from the well-trained BiGRU. The second family is the attention mechanism [37–40], which was implemented by an attention network, i.e. pre-Riboexp (att). The details of all the compared feature selection methods can be found in Supplementary Notes. Intuitively, an attribution method that is recognized to outperform another is expected to generate features that can yield better predictive power in the corresponding learning task. Such an evaluation strategy has been widely used in machine learning fields [24, 41]. Therefore, in our computational experiment, we first selected the features generated by different attribution methods based on their importance scores assigned to individual input positions. Then those unselected positions were masked as zero, which was equivalent to removing the input information in such positions. Subsequently, the features selected by different attribution methods were used to train, respectively, a BiGRU for the evaluation of the prediction performance resulting from each set of the selected features (Supplementary Notes).

Expectedly, all three attribution methods showed a great improvement over the random selection, indicating their ability in acquiring important positions for predicting ribosome density. In particular, we found that for different feature lengths tested, the features selected by Riboexp provided the best prediction performance measured in terms of both MSE and Pearson correlation coefficient (Figure 2, Supplementary Table 3). Similar results were also obtained using a simple multiple layer perceptron (MLP) regressor (Supplementary Figure 4). These results clearly demonstrated the success of our reinforcement learning-based strategy in selecting the most relevant features for accurately predicting ribosome density.

Riboexp indicates the roles of different codon species in translation elongation

Next, we attempted to analyze how Riboexp learns the biological rules that govern the regulation of translation elongation dynamics. Due to the sparsity constraint introduced in the feature selection stage, our Riboexp model selected part of the input codons as rationales for ribosome density regression. For each codon type, we calculated its frequency of being selected as rationales and derived the selected fraction by normalizing the frequency with its total number of the occurrences. Following the general terminology of natural language processing [43], we termed each single codon as a 'unigram' token. Interestingly,

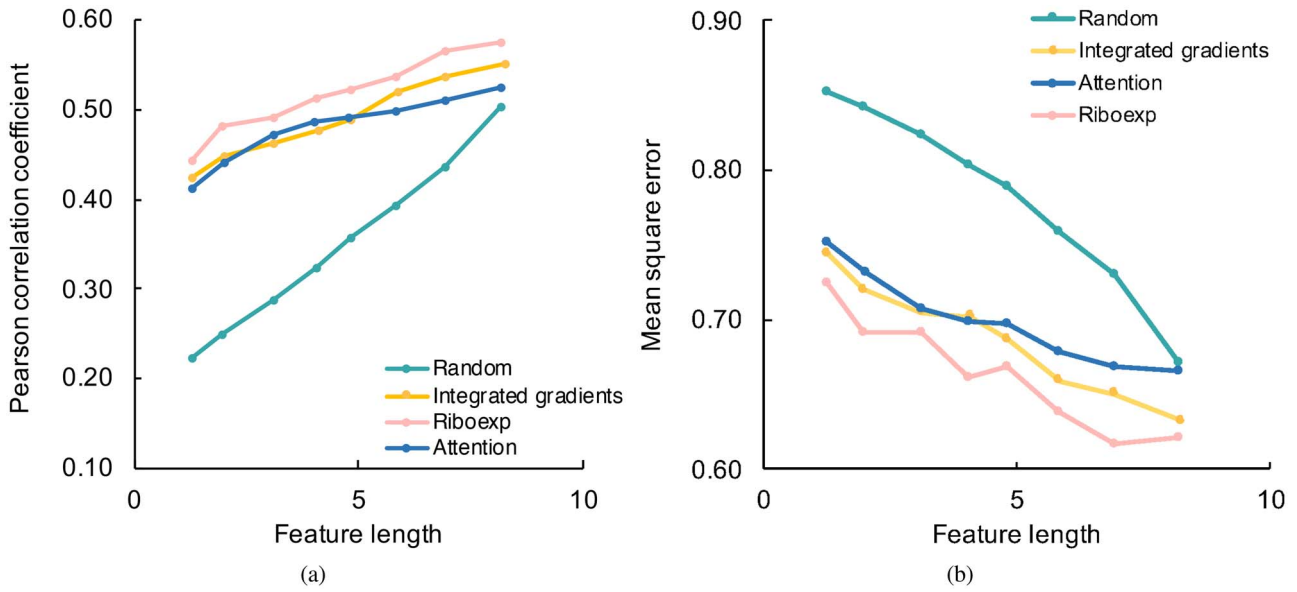


Figure 2. Comparison of different feature attribution methods in predicting ribosome density. By adjusting the sparsity factor for Riboexp or the threshold parameters in other approaches, features were selected with different lengths and used to train a BiGRU regressor to evaluate the resulting prediction performance in terms of mean square error (a) and Pearson correlation coefficient (b). This analysis was conducted on the high-density genes with randomly split training and test data in the yeast dataset.

we found that the fractions of being selected as rationales are associated with codon identity and display distinct distributions among the 61 types of amino-acid coding codons (Figure 3(a)). In particular, those codons encoding amino acids with charged property, e.g. arginine, and special conformations, i.e. proline and glycine, are more frequently selected as rationales, which is consistent with the previous reports [9, 12, 32, 33]. In contrast, the codons encoding hydrophobic amino acids are less involved in the final prediction, demonstrating the effects of amino acid species on rationale selection. We also noticed that the extremely rare codon encoding arginine (CGG) was always selected as rationales by the PolicyNet, indicating a particular role of this codon species in translation elongation dynamics.

To better characterize the influence of each codon species to the predicted ribosome density, we further performed a quantitative analysis in which we masked a codon in the input sequence for the Predictor and calculated the difference in the predicted ribosome density with and without this specific codon information as its contribution to the prediction (Figure 3(b)). Here, the predicted ribosome density with the masked input served as a reference score and the increase beyond this score thus represented the contribution of this masked codon in the original sequence. We mainly focused this analysis on the role of each codon located at the A site of an elongating ribosome, as it had been shown to act as the most important position determining the translation speed [12]. Note that as the non-rationale codons were already masked as zeroes in the input to the Predictor, our prediction was performed using only the codons corresponding to the selected rationales. Expectedly, we noticed a relatively large contribution of CGG (arginine), CGA (arginine) and CCG (proline) to the accumulation of ribosome density, largely due to their amino acid properties and rarity among the synonymous codons [19]. Inspired by this observation, we further examined the correlation between codon rarity and the mean contributions of individual codon species to the ribosome density at A site (Figure 3(c)), which may help explain the discrepancy in the effects of different synonymous codons of the same amino acid. Intriguingly, we found that the mean contribution of each codon species to the predicted ribosome density was negatively correlated with its codon usage (Pearson correlation coefficient -0.64), measured by the codon adaptation index (CAI [44]). Similarly, it also showed a negative correlation with the tRNA adaptation index (tAI) as defined in Weinberg et al. [12] in terms of Pearson correlation coefficient (-0.44, Supplementary Figure 5), which was consistent with the previous reports [45, 46].

Riboexp discovers important bigram patterns of sequence determinants in translation elongation

Our previous analysis on unigram rationale patterns can verify the biological relevance of our method. Here, we further extended the above analysis from the unigram setting to a bigram one (i.e. a token consisting of two contiguous codons) to explore more complex context-dependent biological features selected by our model. In this way, we aimed to understand how the combination and the order of codons can affect the translation elongation process. Since a particular codon may be repeatedly selected when Riboexp predicted ribosome density for each codon position along a transcript with a sliding window, here we used the frequency of a codon being selected as a rationale by Riboexp to estimate its overall contribution to the entire transcript. Then we considered a bigram to be a rationale of the transcript when the frequencies of both codons in the bigram token being selected as rationales were higher than the average value (Supplementary Figure 6). For a particular bigram of codons, we used the frequency ratio of being selected as rationales versus not being selected as rationales to indicate its enrichment (Supplementary Notes). The genome-wide landscape of the bigram enrichment is illustrated in Figure 3(d). To confirm that our method did not simply capture the codon usage bias in the genome, we also plotted the profile of bigram codon usage in Figure 3(e) for comparison, which was defined according to the production of the reciprocal of relative synonymous codon

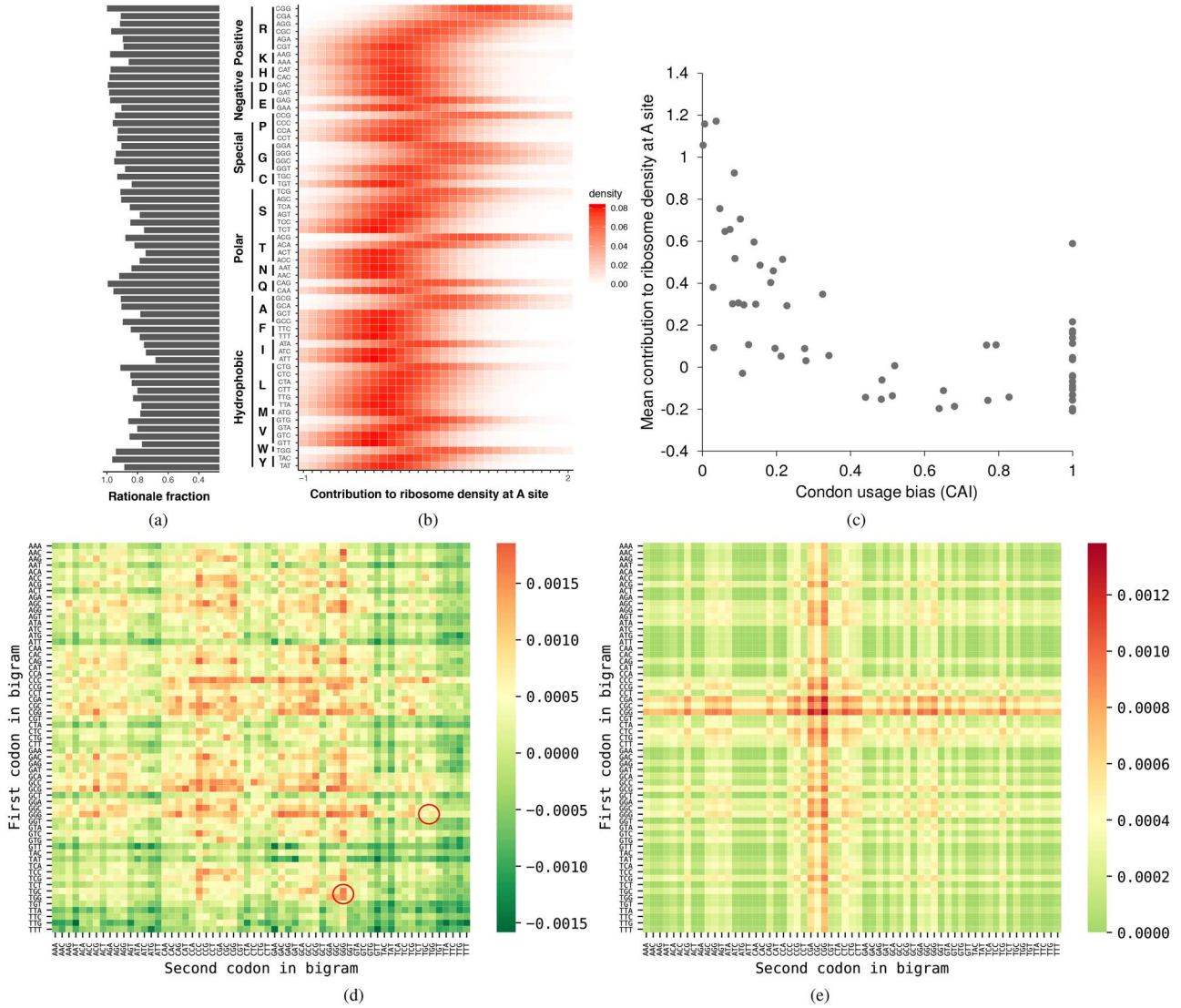
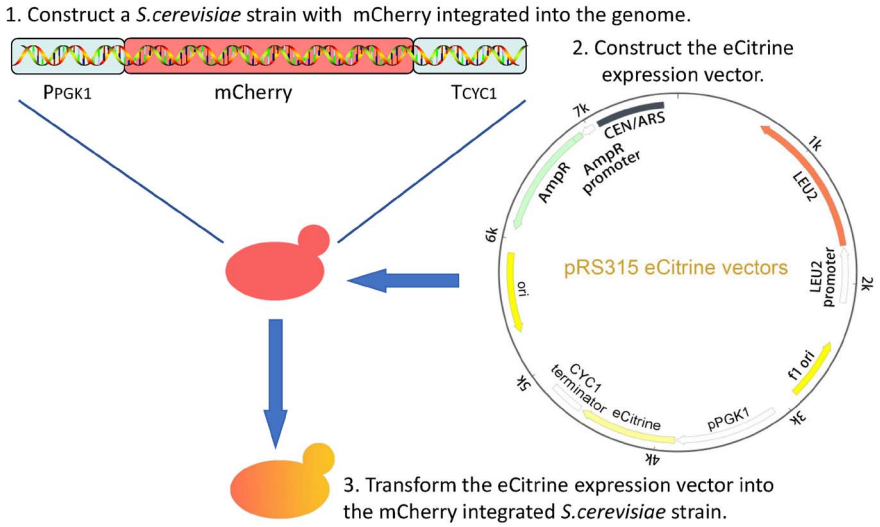


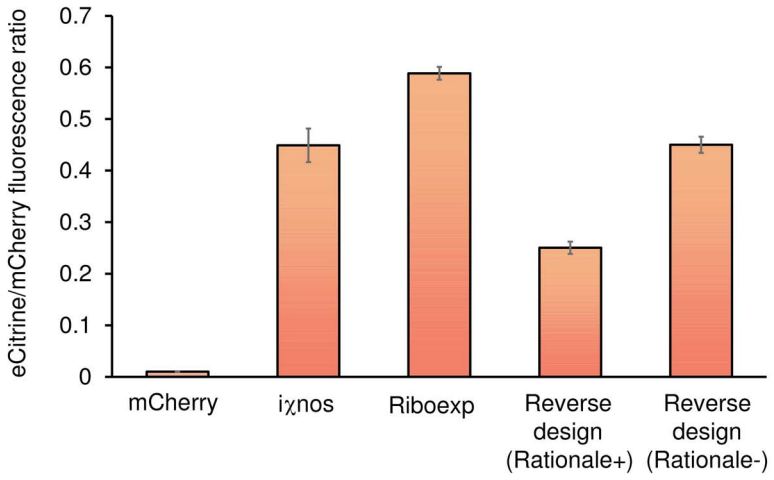
Figure 3. Genome-wide analyses of unigram and bigram profiles of the sequence features selected by Riboexp for modeling translation elongation dynamics. (a) The fraction of being selected as rationales by Riboexp for each codon species. (b) The distribution of the contribution of each codon species to the predicted ribosome density at A site. The contribution was measured by the difference between the predicted ribosome densities with and without this specific codon information and the distribution was smoothed through a kernel density estimation approach [42]. (c) The mean contributions of individual codon species to the predicted ribosome density at A site are associated with codon rarity. The codon usage was measured by the codon adaptation index (CAI). The analysis was performed on the high-density genes using randomly selected test data. (d) The bigram enrichment analysis based on the rationale patterns generated by Riboexp. A case of asymmetric bigram enrichment pattern was circled in red color. (e) The bigram usage measured by the production of the reciprocal of RSCU of the two contiguous codons.

usage (RSCU) of the two contiguous codons [47]. We found that the bigram enrichment profile of rationales (Figure 3(d)) showed a much more informative and scattering distribution than that of bigram codon usage (Figure 3(e)), indicating that our model was able to capture a much more complex pattern.

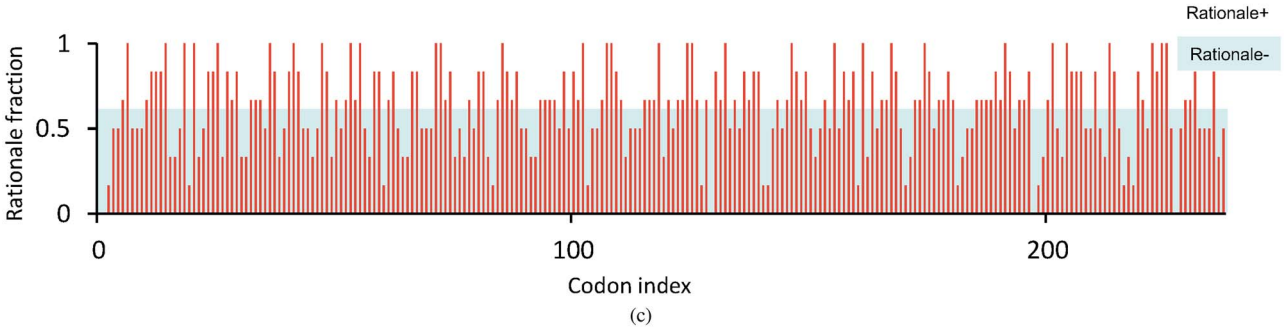
When analyzing the importance of bigram codons, a natural question is how the order of the two codons may affect their roles in translation elongation. Therefore, we also checked the symmetry of the observed bigram patterns in Figure 3(d). Generally speaking, we observed that most bigrams were symmetric, meaning that a bigram codonA-codonB has a similar enrichment value with the bigram codonB-codonA, where codonA and codonB stand for two arbitrary codons. This observation indicated that, for the majority of the bigram codons, the order between codonA and codonB may only play a moderate role in identifying the bigram importance (Pearson correlation coefficient of 0.85 between the matrix of the bigram enrichment and its transposed matrix). A representative example of this kind was CCC CGC, in which the pairs of corresponding amino acids proline and arginine displayed high enrichment scores, regardless of their order. On the other hand, there were still a small number of exceptions (i.e. asymmetries) in which the order of two adjacent codons within a bigram mattered. For example, the only codon that encodes tryptophan (i.e. TGG) showed a significant bigram enrichment when followed by the rare glycine codon GGG, but the enrichment was largely decreased when the order of these two codons was reversed (i.e. GGG TGG). Similar asymmetric bigram enrichment patterns can also be observed in other bigrams involving the glycine codon (such as AAC GGG, AGC GGG and TGC GGG), which suggested that the role that



(a)



(b)



(c)

Figure 4. Codon optimization performed by Riboexp. (a) A dual fluorescence yeast reporter system that was used to validate the codon optimization results. (b) The eCitrine/mCherry mean fluorescence intensity (MFI) ratio was measured by flow cytometry. Three biological replicates were used for each measurement. 'iχnos' and 'Riboexp' represent the expression levels of the optimal eCitrine sequences designed by iχnos [19] and Riboexp, respectively. 'Reverse design (Rationale+)' and 'Reverse design (Rationale-)' represent the expression levels of the eCitrine sequences obtained from the reverse design process on the Rationale+ and Rationale- codons, respectively. (c) The fraction for each codon position to be selected as rationales, after considering all the sequence context windows. The first half of codons possessing the highest rationale fraction is denoted as Rationale+ and the other half is denoted as Rationale- (cyan mask).

glycine plays in the translation elongation process may vary significantly according to different codon orders. Also, we conducted an ablation study to further investigate the role that the order of codons played on translation elongation (Supplementary Table 4). We found that, although the order of codons only showed a moderate influence in learning the interpretability (i.e. the indicated importance), the order of codons was still critical to predicting ribosome density. To our best knowledge, our work is the first attempt to study these asymmetric patterns in translation elongation at bigram level as well as the

genome-wide landscape of the bigram profiles. Along with this direction, more complex biological rules can be explored in the future.

Riboexp generates productive codon sequences for protein expression

It has been widely accepted that synonymous codon usage can strongly affect protein expression levels. As Riboexp significantly outperformed other models in quantifying local ribosome density, we next sought to explore its potential in generating more productive codon sequences for protein expression with optimal combination of synonymous codons. Here, we refined the previously proposed dynamic programming (DP) codon optimization algorithm [19] by imposing a restraint that only the codons selected as rationales were involved in the calculation of ribosome density (Supplementary Notes and Supplementary Figure 7(a)), which thus directed the optimization process to mainly consider the most relevant codon positions. We denoted this algorithm as the rationalized dynamic programming (RDP) to avoid confusion. The purpose of the RDP algorithm was to minimize the overall translation time (defined by the sum of the ribosome densities predicted by Riboexp), which was equivalent to maximizing the ribosome read-through rates and thus boosting protein production.

To demonstrate the superiority of our Riboexp method for the aforementioned optimization task, we constructed a dual fluorescence yeast system (Figure 4(a)) to validate the protein expression results. In order to simplify the procedure of constructing the dual fluorescent yeast system, unlike the previous technique [19] which required mating between the haploid yeast strains, our new system integrated one fluorescence reporter gene into a haploid yeast host strain, and then used a single copy vector to carry the other fluorescence reporter gene. To achieve a dual expression system, we first constructed a baker yeast strain with the mCherry module integrated into the genome using a pAbAi vector. Then, this strain was further transformed with a non-integrating pRS315 vector containing the eCitrine module, which was maintained as a single copy by a CEN6 centromere concatenated to an autonomously replicating sequence (Supplementary Notes).

Intriguingly, by minimizing the total ribosome density predicted by Riboexp, we obtained an optimal sequence that highly expressed eCitrine, showing a 31% higher expression level when compared to the best sequence designed by the previous method [19], as determined by the eCitrine/mCherry mean fluorescence intensity (MFI) ratio in flow cytometry (Figure 4(b), Supplementary Figure 7(b) and Supplementary Table 5). This result indeed confirmed that the superior prediction performance of Riboexp over iχnos (Table 1) can be converted to a large advantage in designing efficient codon sequences for protein production. Considering that the optimal sequence designed by iχnos has been validated to be much better than the randomly generated synonymous eCitrine sequences [19], the improvement of our optimal sequence designed by Riboexp over the randomly generated ones can also be well established.

Next, we also attempted to examine whether the codons that were repeatedly selected as rationales in different sequence contexts were actually more related to the protein expression levels. To achieve this goal, we first calculated the frequency of each codon being selected as a rationale by our model in different sequence windows on the optimized sequence. Then, the first half (N = 120) of codons with the highest rationale frequencies was defined as Rationale+ and the remaining was defined as Rationale- (Figure 4(c)). Starting from the most productive eCitrine sequence, we again applied the RDP algorithm but changing the objective to maximize the overall translation time, which was therefore named as a reverse design process. Here, the effects of varying different codons on protein expression can be reflected by the change of the measured fluorescence signals. By restricting the RDP process to either Rationale+ or Rationale- codons, we found that changing the Rationale+ codons had nearly doubled the effects on protein expression compared to varying the Rationale- codons (Figure 4(b)), indicating a strong relevance of rationales selected by Riboexp in the modulation of protein expression.

Discussion

Technical bias

Computational analysis on ribosome profiling data plays a crucial role in understanding the sequence features that modulate translation elongation dynamics. Recently, the studies of technical bias and shortcomings [32, 48] of ribosome profiling have attracted wide interests in the computational biology community [49–52]. Generally speaking, there are two major sources in the experimental protocols that may introduce bias to the ribosome profiling data. The first one is ribosome harvesting, during which the introduction of translation inhibitors or isolating steps may cause artificial shifts of the elongating ribosome profiles. The second one is sequencing library construction, in which ligase preparation or read stability may bias the abundance of the sequencing results. However, despite the widely acknowledged existence of technical bias, modeling and filtering them by computational methods are generally hard to achieve, since it generally lacks proper supervision to separate bias from true signals. For example, Gerashchenko et al. [53, 54] verified several potential origins of technical bias, and Diament et al. [55] performed a series of evaluations on the magnitudes across a range of studies, but they did not mention an effective way to remove the bias computationally. Without any doubt, understanding the problems related to technical bias would also be beneficial for deciphering the biological determinants of translation elongation dynamics. Nevertheless, the previous work [19] and our current study have demonstrated that the computational models trained based on current ribosome profiling data can provide sufficient guidance to fully exploit the regulatory features of translation elongation and thus improve the codon optimization results.

Future directions

Our current study serves as the *first* attempt to apply reinforcement learning (RL) to study the translation elongation process. Considering that mRNA transcripts are discrete by nature, we believe that RL-based approaches (excelling in producing discrete actions) will be further explored in the related computational biology fields to provide more detailed rationales or better explainability for specific demands. For example, we can extend the current reward function in our framework to further explore the possibility of generating the rationales with more specific properties, e.g. positively or negatively contributing rationales, and thus enable our model to provide more detailed mechanistic insights into the translation elongation process. In addition, we will also consider more perspectives, e.g. RNA stability, to further polish the codon optimization process driven by our computational model. Moreover, we expect that our framework

can be further applicable to the identification of key sequence elements for various learning tasks such as protein design [56] and synthetic biology [57] in the future.

Conclusion

In this work, empowered by the deep reinforcement learning strategy, we have proposed a novel context-dependent feature selection approach to obtain the most relevant sequence features within the input sequence of each sample, thus providing a more accurate framework for modeling the ribosome density on mRNA transcripts. We showed that Riboexp can provide useful biological insights into understanding the determinants of the uneven distribution of ribosomes during translation elongation. In addition, Riboexp can greatly improve protein production compared with the previous state-of-the-art method.

Key Points

- Riboexp provides a new perspective for modeling ribosome density. It is a novel computational approach to explicitly model the determinants of the translation elongation dynamics.
- Riboexp can significantly outperform current state-of-the-art methods in ribosome density prediction.
- The strong interpretability enables Riboexp to reveal interesting findings related to translation elongation.
- The application of Riboexp in codon optimization can largely improve protein production compared with the previous state-of-the-art method.

Supplementary data

Supplementary data are available online at *Briefings in Bioinformatics*.

Funding

This work was supported in part by the National Natural Science Foundation of China (61872216, 81630103, 31900862 and 31871071), the Turing AI Institute of Nanjing, the Zhongguancun Haihua Institute for Frontier Information Technology and Beijing Brain Science Special Project (No. Z181100001518006). Sen Song acknowledges the support of key scientific technological innovation research project by Ministry of Education and a grant from Institute Guo Qiang, Tsinghua University.

References

1. Ingolia NT. Ribosome footprint profiling of translation throughout the genome. *Cell* 2016; **165**(1): 22–33.
2. Dever TE, Dinman JD, Green R. Translation elongation and recoding in eukaryotes. *Cold Spring Harb Perspect Biol* 2018;a032649.
3. Kimchi-Sarfaty C, Oh JM, Kim I-W, et al. A" silent" polymorphism in the MDR1 gene changes substrate specificity. *Science* 2007; **315**(5811): 525–8.
4. Ishimura R, Nagy G, Dotu I, et al. Ribosome stalling induced by mutation of a CNS-specific tRNA causes neurodegeneration. *Science* 2014; **345**(6195): 455–9.
5. Brar GA. Beyond the triplet code: context cues transform translation. *Cell* 2016; **167**(7): 1681–92.
6. Tuller T, Waldman YY, Kupiec M, et al. Translation efficiency is determined by both codon bias and folding energy. *Proc Natl Acad Sci* 2010; **107**(8): 3645–50.
7. Stadler M, Fire A. Wobble base-pairing slows in vivo translation elongation in metazoans. *RNA* 2011; **17**(12): 2063–73.
8. Qian W, Yang J-R, Pearson NM, et al. Balanced codon usage optimizes eukaryotic translational efficiency. *PLoS Genet* 2012; **8**(3):e1002603.
9. Charneski CA. Laurence D. Hurst. Positively charged residues are the major determinants of ribosomal velocity. *PLoS Biol* March 2013; **11**(3): 1–20.
10. Dana A, Tuller T. The effect of tRNA levels on decoding times of mRNA codons. *Nucleic Acids Res* 2014;42(14):917181.
11. Pop C, Rouskin S, Ingolia NT, et al. Causal signals between codon bias, mRNA structure, and the efficiency of translation and elongation. *Mol Syst Biol* 2014; **10**(12):770.
12. Weinberg DE, Shah P, Eichhorn SW, et al. Improved ribosome-footprint and mRNA measurements provide insights into dynamics and regulation of yeast translation. *Cell Rep* 2016; **14**(7): 1787–99.
13. Duc KD, Song YS. The impact of ribosomal interference, codon usage, and exit tunnel interactions on translation elongation rate variation. *PLoS Genet* 2018; **14**(1):e1007166.
14. Ingolia NT, Ghaemmaghami S, Newman JRS, et al. Genome-wide analysis *in vivo* of translation with nucleotide resolution using ribosome profiling. *Science* 2009; **324**(5924): 218–23.
15. Ingolia NT, Hussmann JA, Weissman JS. Ribosome profiling: global views of translation. *Cold Spring Harb Perspect Biol* 2018;a032698.
16. Zhang S, Hu H, Zhou J, et al. Analysis of ribosome stalling and translation elongation dynamics by deep learning. *Cell systems* 2017; **5**(3): 212–20.
17. Liu T-Y, Song YS. Prediction of ribosome footprint profile shapes from transcript sequences. *Bioinformatics* 2016; **32**(12): i183–91.
18. Connor PBFO, Andreev DE, Pavel VB. Comparative survey of the relative impact of mRNA features on local ribosome profiling read density. *Nat Commun* 2016; **7**: 12915.
19. Tunney RJ, McGlincy NJ, Graham ME, et al. Accurate design of translational output by a neural network model of ribosome distribution. *Nat Struct Mol Biol* 2018;25(7):577–82.
20. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc B Methodol* 1996;267–88.
21. Sutton RS. Andrew G Barto. *Reinforcement learning: An introduction. MIT press* 2018.
22. Graves A, Mohamed A r, Hinton GE. Speech recognition with deep recurrent neural networks. In: *Acoustics, speech and signal processing (icassp), 2013 IEEE international conference on*. IEEE, 2013, 6645–9.
23. Cho K, van Merrienboer B, Gulcehre C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *In EMNLP* 2014;1724.
24. Lei T, Barzilay R, Jaakkola T. Rationalizing neural predictions. *In EMNLP, pages* 2016;107–17.
25. Astrom KJ. Optimal control of markov processes with incomplete state information. *Journal of Mathematical Analysis and Applications* 1965; **10**(1): 174–205.
26. Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning. *Nature* 2015; **518**(7540): 529.

27. Lorenz R, Bernhart SH, Siederdissen CHZ, et al. Viennarna package 2.0. *Algorithms for Molecular Biology* 2011; **6**(1):26.
28. Barto AG, Sutton RS, Anderson CW. Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Trans Syst Man Cybern* 1983;(5): 834–46.
29. Ranzato MA, Chopra S, Auli M, et al. Sequence level training with recurrent neural networks. *In ICLR* 2016.
30. Williams RJ. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning* 1992; **8**:229–56.
31. Subtelny AO, Eichhorn SW, Chen GR, et al. Poly (a)-tail profiling reveals an embryonic switch in translational control. *Nature* 2014; **508**(7494): 66.
32. Artieri CG. Hunter B. Fraser. Accounting for biases in riboprofiling data indicates a major role for proline in stalling translation. *Genome Res* 2014;24(12):2011–21.
33. Tuller T, Veksler-Lublinsky I, Gazit N, et al. Composite effects of gene determinants on the translation speed and density of ribosomes. *Genome Biol* 2011; **12**(11): 1–18.
34. Iwasaki S, Floor SN, Ingolia NT. Rocaglates convert dead-box protein eif4a into a sequence-selective translational repressor. *Nature* 2016; **534**(7608): 558.
35. Mohammad F, Green R, Allen R. Buskirk. A systematically-revised ribosome profiling method for bacteria reveals pauses at single-codon resolution. *Elife* 2019; **8**:e42591.
36. Sundararajan M, Taly A, Yan Q. *Axiomatic attribution for deep networks* In *ICML*, 2017.
37. Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.
38. Singh R, Lanchantin J, Sekhon A, et al. Attend and predict: Understanding gene regulation by selective attention on chromatin. In: *Advances in Neural Information Processing Systems*, 2017, 6788–98.
39. Hu H, Xiao A, Zhang S, et al. DeepHINT: understanding HIV-1 integration via deep learning with attention. *Bioinformatics* 2019;35(10):1660–7.
40. Luo Y, Ma J, Liu Y, et al. Deciphering signaling specificity with interpretable deep neural networks. In: *Research in Computational Molecular Biology: 22nd Annual Conference, RECOMB 2018, Paris, France, April 21–24, 2018, Proceedings*. Springer International Publishing, 2018.
41. Hooker S, Erhan D, Kindermans P-J, et al. A benchmark for interpretability methods in deep neural networks. In: *Advances in Neural Information Processing Systems*, 2019, 9737–48.
42. Sheather SJ, Jones MC. A reliable data-based bandwidth selection method for kernel density estimation. *J R Stat Soc B Methodol* 1991; **53**(3): 683–90.
43. Cavnar WB, Trenkle JM, et al. N-gram-based text categorization. In: *Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval, volume 161175*. *Citeseer*, 1994.
44. Sharp PM, Li W-H. The codon adaptation index – a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res* 1987; **15**(3): 1281–95.
45. Dana A, Tuller T. The effect of trna levels on decoding times of mrna codons. *Nucleic Acids Res* 2014; **42**(14): 9171–81.
46. Sabi RENANA, Tuller TAMIR. Modelling the efficiency of codon–trna interactions based on codon usage bias. *DNA Res* 2014; **21**(5): 511–26.
47. Sharp PM, Tuohy TMF, Mosurski KR. Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Res* 1986; **14**(13): 5125–43.
48. Hussmann JA, Patchett S, Johnson A, et al. Press. Understanding biases in ribosome profiling experiments reveals signatures of translation dynamics in yeast. *PLoS Genet* December 2015; **11**(12): 1–25.
49. Wang H, McManus J, Kingsford C. Accurate recovery of ribosome positions reveals slow translation of wobble-pairing codons in yeast. In: Singh M (ed). *Research in Computational Molecular Biology*, 2016, 37–52.
50. Fang H, Huang Y-F, Radhakrishnan A, et al. Scikit-ribo enables accurate estimation and robust modeling of translation dynamics at codon resolution. *Cell systems* 2018; **6**(2): 180–91.
51. Zhao D, Baez W, Fredrick K, et al. Riboprop: a probabilistic ribosome positioning algorithm for ribosome profiling. *Bioinformatics* 2019;35(9):1486–93.
52. Lauria F, Tebaldi T, Bernabò P, et al. Ribowaltz: optimization of ribosome p-site positioning in ribosome profiling data. *PLoS Comput Biol* 2018; **14**(8):e1006169.
53. Gerashchenko MV, Gladyshev VN. Translation inhibitors cause abnormalities in ribosome profiling experiments. *Nucleic Acids Res* 2014; **42**(17): e134–4.
54. Gerashchenko MV. Vadim N Gladyshev. *Ribonuclease selection for ribosome profiling Nucleic Acids Research* 2017; **45**(2): e6–6.
55. Diament A, Tuller T. Estimation of ribosome profiling performance and reproducibility at various levels of resolution. *Biol Direct* 2016; **11**(1):24.
56. Natan E, Endoh T, Haim-Vilmovsky L, et al. Cotranslational protein assembly imposes evolutionary constraints on homomeric proteins. *Nat Struct Mol Biol* 2018; **25**(3): 279.
57. Ciryam P, Morimoto RI, Vendruscolo M, et al. In vivo translation rates can substantially delay the cotranslational folding of the escherichia coli cytosolic proteome. *Proc Natl Acad Sci* 2013; **110**(2): E132–40.