TITER: predicting translation initiation sites by deep learning

Sai Zhang¹,†, Hailin Hu²,†, Tao Jiang³,4,5, Lei Zhang²,* and Jianyang Zeng¹,*

¹Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing 100084, China, ²School of Medicine, Tsinghua University, Beijing 100084, China, ³Department of Computer Science and Engineering, University of California, Riverside, CA 92521, USA, ⁴MOE Key Lab of Bioinformatics and Bioinformatics Division, TNLIST/Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China and ⁵Institute of Integrative Genome Biology, University of California, Riverside, CA 92521, USA

*To whom correspondence should be addressed.
†The authors wish it to be known that these authors contributed equally.

Abstract

Motivation: Translation initiation is a key step in the regulation of gene expression. In addition to the annotated translation initiation sites (TISs), the translation process may also start at multiple alternative TISs (including both AUG and non-AUG codons), which makes it challenging to predict TISs and study the underlying regulatory mechanisms. Meanwhile, the advent of several high-throughput sequencing techniques for profiling initiating ribosomes at single-nucleotide resolution, e.g. GTI-seq and QTI-seq, provides abundant data for systematically studying the general principles of translation initiation and the development of computational method for TIS identification.

Methods: We have developed a deep learning-based framework, named TITER, for accurately predicting TISs on a genome-wide scale based on QTI-seq data. TITER extracts the sequence features of translation initiation from the surrounding sequence contexts of TISs using a hybrid neural network and further integrates the prior preference of TIS codon composition into a unified prediction framework.

Results: Extensive tests demonstrated that TITER can greatly outperform the state-of-the-art prediction methods in identifying TISs. In addition, TITER was able to identify important sequence signatures for individual types of TIS codons, including a Kozak-sequence-like motif for AUG start codon. Furthermore, the TITER prediction score can be related to the strength of translation initiation in various biological scenarios, including the repressive effect of the upstream open reading frames on gene expression and the mutational effects influencing translation initiation efficiency.

Availability and Implementation: TITER is available as an open-source software and can be downloaded from https://github.com/zhangsaithu/titer.

Contact: lzhang20@mail.tsinghua.edu.cn or zengjy321@tsinghua.edu.cn

Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

Translation initiation plays an important role in mRNA translation, in which the methionyl tRNA unique for initiation (Met-tRNAi) identifies the AUG start codon and triggers the downstream translation process (Hershey et al., 2012; Jackson et al., 2010; Sonenberg and Hinnebusch, 2009). As translation initiation is an essential step in controlling gene expression and protein synthesis, the dysregulation of the initiation process can cause various human diseases, including cancers and metabolic disorders (Hershey et al., 2012; Sonenberg and Hinnebusch, 2009). On the other hand, the mechanisms underlying translation initiation, e.g. the recognition of a translation initiation site (TIS) by the 80S ribosome assembly, are far more complicated than scientists had initially believed (Gao et al., 2015; Lee et al., 2012). In particular, experimental studies have shown that the eukaryotic translation is not always initiated at the canonical AUG start codons (Gao et al., 2015; Kozak, 1989; Lee
In addition to those annotated TISs (aTISs), both upstream (uTISs) and downstream TISs (dTISs) can also occur at non-AUG codons, which yields alternative open reading frames (ORFs) that can be translated into short peptides or affect the expression levels of the main ORFs (Barbosa et al., 2013; Calvo et al., 2009; Hershey et al., 2012; Jackson et al., 2010; Sonenberg and Hinnebusch, 2009). This underscores the necessity of a better understanding of the mechanism of TIS recognition.

Ribosome profiling (ribo-seq), a high-throughput deep sequencing-based technique that measures the ribosome protected fragments in vivo, has become a widely used method to quantify the translation dynamics on a transcriptome (Ingolia et al., 2009, 2012). However, the standard ribo-seq is not suitable for directly detecting TISs (Lee et al., 2012). Based on ribo-seq, additional techniques, e.g. the global translation initiation sequencing (GTI-seq) (Lee et al., 2012) and the quantitative translation initiation sequencing (QTI-seq) (Gao et al., 2015), have been developed for a systematic mapping of the start codon positions at single-nucleotide resolution in vivo. By far, these techniques have provided abundant data for investigating the principles of translation initiation and translational control.

A variety of computational methods have been developed to predict alternative TISs or ORFs (Chew et al., 2016; Hatzigeorgiou, 2002; Li and Jiang, 2005; Zien et al., 2000; Zur and Tuller, 2013). However, most of these methods only focused on the AUG start codon and did not consider the widely observed non-AUG TISs. In addition, few studies have utilized experimental data generated by GTI-seq or QTI-seq in their works, which limits the empirical predictive power of these methods. Recently, a linear regression-based approach, called PreTIS, was proposed to predict non-canonical TISs by incorporating both AUG and its near-cognate codons (i.e. the codons differing from AUG by one nucleotide), in which TISs identified by GTI-seq were used to train the prediction model (Reuter et al., 2016). However, only alternative TISs in the 5' UTR (i.e. uTISs) were considered by PreTIS, and several constraints on the candidate TISs (e.g. the codon position in the transcript and the existence of an orthologous mouse sequence) had to be imposed due to the limitations of their feature engineering. To our best knowledge, our work is the first attempt to predict all possible TISs, including uTISs, aTISs and dTISs, in a unified framework without any restriction to the codon site of interest.

Recently, deep learning has become one of the most effective and powerful prediction methods in machine learning (Hinton and Salakhutdinov, 2006; Hinton et al., 2006). It has been widely used and shown to be able to achieve the state-of-the-art prediction performance on various machine learning tasks, such as speech recognition (Hinton et al., 2012), image classification (Hinton and Salakhutdinov, 2006) and natural language processing (Collobert et al., 2011). In addition, deep learning is gradually gaining its popularity in bioinformatics and has yielded superior performance over conventional learning methods on a variety of biological prediction tasks, such as the predictions of protein-nucleotide binding (Alipanahi et al., 2015; Zhang et al., 2015), functional effects of non-coding sequence variants (Quang and Xie, 2016; Zhou and Troyanskaya, 2015) and ribosome stalling (Zhang et al., 2016).

In this study, we have developed a deep learning-based framework, named TITER (Translation Initiation siTE detector), for accurately predicting TISs based on the available high-throughput sequencing data (Fig. 1). TITER possesses more flexibility than previous methods, and integrates the prior preference of the TIS codon composition as well as their surrounding sequence contexts into a unified framework, in which an ensemble of hybrid deep convolutional and recurrent neural networks is implemented to effectively and robustly capture the sequence features of translation initiation.

Extensive validation tests have shown that TITER can greatly outperform the state-of-the-art computational approaches in detecting TISs. In addition, TITER can successfully identify significant sequence motifs for different TIS codons, including a Kozak-sequence-like motif for the AUG TIS codon. By combining gene expression data with TITER analysis, we found that the predictions of TITER well correlated with translation efficiency (TE) of genes, which basically reconfirmed the repressive effect of uORFs on gene expression. Furthermore, our comparative analyses on several important mutations around TISs showed that the fold changes of TITER prediction scores well conformed to the experimentally verified mutational effects reported in the literature (Calvo et al., 2009; Noderer et al., 2014). These results demonstrated that TITER can offer a powerful tool to model the sequence features of translation initiation and identify potential TISs, which will provide useful insights into understanding the underlying mechanisms of translation initiation.

2 Materials and methods

2.1 Datasets

We mainly used the QTI-seq data collected from the HEK293 cell line (Gao et al., 2015) and the annotated TISs retrieved from Ensembl v84 (Aken et al., 2016) (together denoted by Gao15) to train and test TITER. Specifically, we first retrieved all transcripts, each containing at least one TIS identified by the QTI-seq experiment. Then we combined the QTI-seq identified TISs among these transcripts with the corresponding annotated TISs obtained from the Ensembl database, and regarded them as positive samples. To construct a dataset of negative samples that well reflected the imbalance of positive and negative samples in vivo, for each TIS in the positive dataset, we chose up to 10 codon sites of the same triplet within the same transcript as negative samples. Considering the leaky scanning nature of the translation initiation process (Michel et al., 2014), we searched for the negative samples starting from the 5' end of an mRNA, until ten eligible sites were found. Altogether, the Gao15 dataset consisted of 9776 positive samples and 94 899 negative samples from 4111 transcripts, among which 400 transcripts were reserved for testing and the others were regarded as our training data (denoted by Gao15_test and Gao15_train, respectively).

To further evaluate the prediction performance of TITER in a scenario with an unlimited ratio of positive and negative samples, from the 400 test transcripts, we also constructed another test dataset (denoted by Gao15_test_extended), in which we identified all possible initiation codon sites before the last TIS of each transcript when searching for negative samples. Notably, as QTI-seq identified
versatile TISs in terms of both codon composition and positions, our constructed training and test datasets also reflected this versatility, which thus enabled TITER to capture diverse sequence features of translation initiation that exist in vivo. In particular, we found that in the QTI-seq data derived from Gao et al. (2015), besides the canonical dTISs and the conventional AUG TISs, there also exist a large number of alternative TISs (i.e., uTISs and iTISs) and non-AUG TISs (e.g., CUG, GUG and UUG) (Fig. 2). This observation underscores the necessity of modeling the universal rules of translation initiation.

2.2 Modeling the preference of codon composition of TISs

It has been experimentally validated that there is a significant preference of codon composition among TISs (Gao et al., 2015; Lee et al., 2012). For example, the AUG codon generally plays a dominant role (>50%) at the TISs (Fig. 2a). Interestingly, other codons that differ from AUG at the first nucleotide (nt), including CUG, GUG and UUG, also greatly contribute to translation initiation (~25%; Fig. 2a). Here, we collectively denote these four codons by NUG, in which ‘N’ stands for any possible nucleotide. With this prior knowledge on codon preference, we propose to explicitly model the preference of codon composition of TISs by

\[
\text{CodonScore}(c) = \log\left( \frac{\# \text{ of } c\text{'s at TISs}}{\# \text{ of TISs}} \right) - \log\left( \frac{\# \text{ of } c\text{'s in background}}{\# \text{ of background codons}} \right),
\]

in which \(\text{CodonScore}(c)\) is defined as the preference of codon composition for a certain codon \(c\), denotes the number of the entities in the defined category and the background is defined as the upstream one-third segment of the transcripts, following the same principle as in Gao et al. (2015). In particular, we denote the function between the parentheses on the right-hand side of Equation (1) by \(f(c)\). Then we define \(\log_+(x) = \log(x)\) if \(\log(x) \geq 0\), \(\log_+(x) = -x\) otherwise, where \(x = \min(\log(f(c)))|\log(f(c))| > 0, c\) is a codon\). This definition is introduced as we notice that for those codons that are less similar with the canonical AUG start codon, their abundance in the QTI-seq dataset is actually lower than the background, leading to a negative CodonScore value if we only use the \(\log(x)\) form. Although these codons only account for ~10% in the Gao15 dataset, to avoid the over-devaluation of the CodonScore for these codons due to the nature of logarithm function, we propose to use a linear model to define the CodonScore for these codons. In particular, the top five codons with the highest CodonScore value are AUG, CUG, GUG, UUG and ACG (Supplementary Fig. S1), which is consistent with the statistics shown in Figure 2a.

2.3 Modeling the contextual features of TISs

It has been indicated that the contextual sequences around the TISs can influence the likelihood of translation initiation (Kozak, 1989; Noderer et al., 2014). For example, the upstream and downstream sequences of the canonical AUG TISs exhibit a consensus motif called the Kozak sequence (Kozak, 1989). This observation underlies the rationale of modeling the translation initiation events by encoding the contextual sequence features surrounding TISs.

Here, we develop a deep neural network to systematically model the sequence features of TISs (Fig. 3a). In particular, each TIS is extended both upstream and downstream by 100 nts, which yields the contextual profile of a translation initiation event and is denoted by \(s = (n_1, \ldots, n_{200})\), where \(n_i\) denotes the nucleotide at the \(i\)th position. To characterize the local motifs of the extended sequence, we first encode the nucleotides using the one-hot encoding technique (Pedregosa et al., 2011), that is, a nucleotide of a particular type (A, U, C or G) is encoded by a binary vector of length four, in which the corresponding position is one while the others are zeros, after indexing all four types of nucleotides. Then we employ multiple convolution operators (denoted by conv(\(\cdot\)) to scan the encoded sequence profile and detect the local motifs around each TIS. After that, the pooling operators (denoted by pool(\(\cdot\))) are used to identify the activated motifs and also reduce the dimensions of hidden features. Note that the conventional convolution-pooling structures are order insensitive, as they only detect whether certain motifs exist regardless of their positions or orders. To further characterize the motif order that may also contribute to translation initiation, we also stack a long short-term memory (LSTM) network (denoted by LSTM(\(\cdot\))) upon the convolution-pooling module, which takes the pooled feature vectors as input and models the long-term dependencies between different motifs. Finally, the outputs of the LSTM at all positions are concatenated and fed into a logistic regression layer (denoted by logist(\(\cdot\))) to compute the probability of translation initiation for the input sequence. Indeed, when considering the LSTM outputs at all positions, we also implement the attention mechanism that has been widely used in deep learning (Denil et al., 2012; Larochelle and Hinton, 2010) to leverage the relevance of each position to the final prediction. Altogether, by fully exploiting both sequence composition and motif orders, the hybrid convolutional and recurrent neural network in TITER computes the following score,

\[
\text{ContextScore}(s) = \text{logist}(\text{LSTM(pool(conv(encode(s)))))),
\]

The sequence profiles of non-TISs (i.e. the negative samples) can be modeled in the same manner. Note that a similar hybrid neural network architecture has also been implemented in Quang and Xie (2016) and Hassanazadeh and Wang (2016) for different tasks, i.e. predicting the functional effects of non-coding mutations and the DNA-binding protein targets, respectively.

2.4 Model training and model selection

Given the training samples \(\{(s_i, y_i)\}\), the loss function of our deep learning framework is defined as the sum of the negative log likelihoods, i.e.

\[
\text{loss} = - \sum_i \log(y_i \cdot \text{ContextScore}(s_i) + (1 - y_i) \cdot (1 \cdot \text{ContextScore}(s_i))),
\]

where \(y_i\) indicates whether \(s_i\) is the sequence profile of a TIS or not. We use the standard error backpropagation algorithm (Rumelhart et al., 1986) and the stochastic gradient descent method (Bengio, 2012) to train the hybrid neural network and search for the network...
weights that minimize the loss function of Equation (3). Several regularization techniques, including the max-norm constraints on weights (Srebro et al., 2005), dropout (Srivastava et al., 2014) and early stopping (Bengio, 2012), are also employed to optimize the training process and address the overfitting problem.

Our hybrid deep neural network architecture and various optimization techniques used in the training process have introduced several hyperparameters, e.g. the kernel size, kernel number and max-norm of weights, that need to be determined. A proper hyperparameter calibration procedure can help yield better solutions to the optimization problem in Equation (3). Here, we use the tree-structured Parzen estimator (TPE) approach (Bergstra et al., 2011) to calibrate the hyperparameters in our model, including kernel size, kernel number and max-norm of weights in the convolution layer, pooling length of the max-pooling layer, output dimension of the LSTM layer, dropout rate, and the optimizer algorithm. In particular, we first use all the positive training samples and an equal number of randomly selected negative training samples to optimize the hyperparameters based on TPE (with 100 evaluations), and then choose the hyperparameters that achieve the minimum loss to further train our final models (Supplementary Table S1).

We realize that the task of predicting TISs is an imbalanced classification problem (i.e. much more negative samples than positive ones), for which the standard training procedure designed for balanced samples cannot be applied directly. On the other hand, making use of more negative samples in the training process can lead to a more robust model with less variance in prediction (Wallace et al., 2011). To tackle this imbalance problem, here we employ a bootstrapping-based technique (Fig. 3b) derived from the theory established in Wallace et al. (2011). Briefly, we first construct several groups of bootstrap samples (denoted by $S_i$) from the original imbalanced population $S$, by randomly selecting samples with replacement. Then for each group $S_i$, we balance the samples by downsampling, i.e. randomly selecting an equal number of positive and negative samples from $S_i$, which yields a balanced dataset $B$. After that, a hybrid deep neural network $N_i$ is trained based on each dataset $B_i$ independently, resulting in an ensemble of binary classifiers ($N_i$). Given an input sequence $s$, its final prediction score is averaged over the prediction scores output by all classifiers, i.e.

$$
\text{ContextScore}(s) = \frac{1}{n} \sum_{i=1}^{n} N_i(s),
$$

in which $n$ is the total number of the constructed balanced datasets in $\{B_i\}$ (which is also equivalent to the total number of trained classifiers in $\{N_i\}$). We apply this bootstrapping-based technique to our training dataset and train 32 independent deep neural networks. After that, their prediction scores are averaged and used as the final estimated probability of translation initiation for the given input sequence profile.

Owing to the nature of non-convex optimization, random weight initialization can influence the search results of the gradient descent algorithm (Bengio, 2012). This initialization bias may also introduce variance to our modeling and further affect the prediction performance. In TITER, the aforementioned bootstrapping-based technique can alleviate such an initialization bias in addition to the sample bias, as the network weights have been initialized independently for each balanced sample group before the training process.

The hybrid deep neural network of TITER has been implemented using the Keras library (https://keras.io/), and the Tesla K20c GPUs have been used to speed up the training process. The TPE algorithm for hyperparameter calibration has been implemented based on Hyperas (https://github.com/maxpumperla/hyperas/), a Python library for optimizing hyperparameters of the models implemented based on Keras.

### 2.5 Integrating the preference of codon composition and the contextual features of TISs

As the neural network in the binary classification scenario mainly discriminates the positive and negative samples based on the contextual sequence features, it may be less sensitive to the difference in the preference of codon composition of TISs. To improve the sensitivity of our method and to make a tradeoff between these two complementary information, here we integrate both CodonScore and ContextScore to derive the final score representing the likelihood of translation initiation, i.e.

$$
\text{TISScore}(s) = \text{CodonScore}(s) \times \text{ContextScore}(s),
$$

where $s$ denotes the sequence profile of a codon site of interest, and a high TISScore is outputted only if both ContextScore and CodonScore are large enough. Basically, the final score is obtained by weighting between the contributions of the prior preference of codon composition and the sequence contexts to translation initiation.

### 3 Results

#### 3.1 TITER accurately predicts TISs

Here, we performed extensive tests to show that TITER greatly outperformed the state-of-the-art methods in predicting TISs, including WRENT (Chew et al., 2016) and PreTIS (Reuter et al., 2016). Note
that WRENT cannot compute the initiation scores for those non-AUG codons, while PreTIS only focuses on the 5' UTR for the near-cognate codons (i.e. the codons differing from AUG by one nucleotide).

After hyperparameter calibration, we first tested our method using a five-fold cross-validation procedure and evaluated its prediction performance based on both the areas under the receiver-operating characteristic (AUROC) and the precision recall (AUPR) curves. We found that without using the recurrent layer or the bootstrapping-based technique in our framework, the prediction performance dropped dramatically compared to that of the proposed model (Supplementary Fig. S2), especially for the AUPR score (with drop in AUPR by \~10%), demonstrating the important contributions of these techniques to reduce the modeling variance and boost the prediction accuracy. Also, our model yielded better prediction performance on the AUG codons than non-AUG ones (Supplementary Fig. S2). This indicated that the sequence features of AUG TISs may provide more predictive information in our framework than those of non-AUG TISs. In addition, we observed that even for the AUG sites, our method greatly outperformed WRENT, with increases in AUROC and AUPR scores by 20.9 and 47.0%, respectively.

We also performed additional tests on an independent dataset involving 400 transcripts derived from the Gao15 dataset (see Section 2). We first validated the prediction performance of TITER on the Gao15_test dataset, in which the positive and negative samples were selected in the same way as in the construction of our training data, resulting in 935 TISs and 9098 non-TISs. Test results revealed comparable or even superior prediction performance on this dataset compared to the previous cross-validation results (Fig. 4a and b, Supplementary Fig. S2), which further validated that our model did not suffer from the overfitting problem. Similarly, our method also achieved better prediction performance on AUG sites than non-AUG ones, and greatly outperformed WRENT on AUG codons with increases in AUROC and AUPR scores by 17.4 and 46.1%, respectively (Fig. 4a and b). Furthermore, we tested TITER on a more natural and realistic setting, in which we considered all negative samples (i.e. non-TISs) surrounding the positive ones (i.e. TISs) in a transcript. By mainly focusing on those codons with relatively high probability of translation initiation (i.e. CodonScore > 1, including AUG, CUG, GUG, UUG and ACG), we selected all the non-TISs of the same triplet before the last TISs for each transcript as negative samples to construct the Gao15_test_extended dataset, which resulted in 767 positive and 9914 negative samples in total. Test results showed that TITER also yielded excellent prediction performance with AUROC and AUPR scores of 89.1 and 61.8%, respectively (Fig. 4a and b). Specifically, the prediction performance was improved after incorporating the prior knowledge on the preference of being a TIS for that particular codon. By fixing the weights of a trained neural network, we can optimize the above problem using the gradient descent technique.

To further validate the generalization of our framework across different datasets and organisms, we additionally evaluated the prediction performance of TITER on an additional dataset of mouse (denoted by Gao15_mouse), which was also derived from Gao et al. (2015). All the data preprocessing procedures were the same as those for the human data (i.e. the Gao15 dataset). In particular, we randomly selected 360 transcripts as the test data and used the remaining transcripts as our training data. The final prediction performance was evaluated on the test data. Based on the same hyperparameter values calibrated on the Gao15 dataset (Supplementary Table S1), we found that the prediction performance of TITER on the mouse data was comparable or even superior to that on the human data, demonstrating the generalization capability of our framework (Supplementary Fig. S4).

To facilitate the comparison between TITER and PreTIS, we also constructed a separated uTISs test dataset (denoted by Calviello16). Specifically, this dataset included AUG TISs identified by Ribotester (Calviello et al., 2016), a statistical method to define the ORFs through the three-nucleotide periodicity of the ribosome profiling data. The TISs from the transcripts and their isoforms used for training either TITER or PreTIS were excluded, resulting in 383 available transcripts. Furthermore, since the feature engineering of PreTIS required the transcript harboring the TIS of interest to have an orthologous mouse sequence and to be at least 99 nts downstream from its transcript start site, we also excluded all the samples that failed to meet these requirements (~164 transcripts), leaving 253 transcripts available for the test. The positive and negative samples were labeled using the same method as described in Reuter et al. (2016), resulting in an imbalanced dataset containing 227 positive and 864 negative samples. Tests on the Calviello16 dataset showed that TITER still greatly outperformed both WRENT and PreTIS in this independent dataset, with increases in AUROC by 19.0 and 3.9%, respectively, and in AUPR by 17.6 and 3.8%, respectively (Fig. 4c and d). Note that here the comparison only focused on the AUG codons since Ribotester only identified the ORFs starting with AUG. In addition, TITER yielded excellent prediction performance on the Gao15 dataset when considering uTISs (still outperforming PreTIS) and dTISs separately (Supplementary Fig. S5).

3.2 TITER captures the sequence motifs of different TIS codons

The hybrid deep neural network of TITER can be easily extended to generate the sequence motifs of different TIS codons. Based on the similar idea to that used in Simonyan et al. (2013), here we generated the sequence motifs of a particular TIS codon by optimizing the following problem:

$$\arg\min_{s_l,s_r,\epsilon} (\text{ContextScore}(s_l \oplus \epsilon \oplus s_r) - 1)^2,$$

in which \(s_l\), \(s_r\), and \(\epsilon\) stand for the left sequence, the right sequence and the TIS codon of interest, respectively, and \(\oplus\) represents the sequence concatenation operation. Basically, the above formulation minimizes the difference between the predicted and the positive labels, and finds the optimal sequence with the highest probability of being a TIS for that particular codon. By fixing the weights of a trained neural network, we can optimize the above problem using the gradient descent technique.

Here, we mainly focused on the motif of a region covering both upstream and downstream 10 nts from the TIS codon of interest. In particular, for the canonical AUG start codon, our generated motif was quite similar to the Kozak sequence, i.e. (gcc)gccRcAUGG, which was previously validated to be consensus for tTISs (Kozak, 1989) (Fig. 5a). In the Kozak sequence, a lower case indicates a commonly occurred base, a upper case denotes a highly conserved base, and the sequence in a bracket is of uncertain significance. However, simply depending on the Kozak sequence or similar methods like WRENT cannot accurately identify the experimentally-verified AUG TISs (Fig. 4), as the prior knowledge on the sequence motif surrounding the annotated AUG TISs may lead to a biased estimation, which may also explain the slight difference between our generated sequence motif and the Kozak sequence (Fig. 5a). Note that the Kozak sequence and the position weight matrix (PWM)-based
Deep learning for predicting translation initiation

3.3 Prediction of TITER correlates with TE

Previous studies have shown that translation initiation at aTISs and uTISs can play important functional roles in regulating gene expression (Calvo et al., 2009; Chew et al., 2016; Ferreira et al., 2013; Hinnebusch et al., 2016). In particular, it has been believed that the strength of translation initiation signals at aTISs can positively correlate with TE, while the occurrence of uTISs may repress the expression of the main ORFs (Calvo et al., 2009; Chew et al., 2016; Ferreira et al., 2013; Hershey et al., 2012; Hinnebusch et al., 2016; Jackson et al., 2010; Sonenberg and Hinnebusch, 2009). Here, we were particularly interested in investigating the contributions of the predicted TISScore values at aTISs and uTISs to the TE of the main ORFs.

Here, we define the score of TE as the logarithm of the protein expression level divided by the corresponding mRNA expression level. Specifically, the tandem mass spectrometry (MS/MS) data and the mRNA-seq data of HEK293 cell line were obtained from the previously published studies (Geiger et al., 2012; Nam et al., 2014) and were used to derive the levels of protein and mRNA expression, respectively. The iBAQ normalized intensity and the RPKM value were averaged among different replicates for protein and mRNA expression, respectively. We only considered those proteins that were detected in at least two out of three replicates in the MS/MS data. The Uniprot IDs of the genes measured by mass spectrometry were matched to the Ensembl transcript IDs by the Uniprot Retrieve/ID mapping interface (Consortium, 2015). As the current tandem mass spectrum technique cannot accurately distinguish isoforms, our analysis was carried out on the gene level as in Calvo et al. (2009).

Specifically, for a Uniprot ID that was mapped to multiple transcripts, we selected the transcript with the largest expression value measured by mRNA-seq. The final dataset of matched proteins and mRNAs contained 5752 genes in total. Consistent with the previous report (Lundberg et al., 2010), we also observed a certain level of correlation between protein and mRNA expression levels (Spearman’s correlation coefficient $r = 0.67$), validating the quality of this dataset to some extent. For uTISs, we followed the same definition of uORF as described in Calvo et al. (2009), in which a uORF is defined as a continuous segment of codons that has a start codon in the 5’ UTR, a stop codon before the end of the main ORF and a minimum length of nine nts (including the stop codon).

We first employed a linear regression model (implemented based on the scikit-learn library (Pedregosa et al., 2011)) to predict the TE value from the TISScore of aTIS for each transcript. We applied a 10-fold cross-validation procedure to evaluate the correlation between the predicted and the experimentally derived values. Test results showed that the predicted TE values based on the TISScore values of aTISs computed by TITER well correlated with the experimentally derived TE values (Spearman’s correlation coefficient $r = 0.234$; Table 1), indicating that TISScore can provide predictive power to estimate the strength of translation initiation. Note that...
there are a variety of biological processes between translation initiation and the final translated protein products that were not included in our modeling, e.g. translation elongation and protein folding, which may explain the weakness of the observed correlation in this test. We further integrated the TISScore values of those uTISs that were confidently predicted to form a uORF and with CodonScore >1 and ContextScore >0.95 into the regression model. Interestingly, after integrating the TISScore values of uTISs, the correlation between the predicted and experimentally derived TE values increased from 0.234 to 0.245 (Table 1). In particular, the feature coefficients in the regression model for the TISScore values of eligible uTISs were negative, indicating the repressive effect of uORF on the protein expression. Such an increase in correlation was limited, which was consistent with the previous report that the repressive effect of uORF on protein expression level is relatively weak in human (Chew et al., 2016). To further validate the effectiveness of the TISScore, we also performed the same prediction task by replacing the TISScore values of uTISs with the number of AUG codons in the 5’ UTR. We found that with this feature replacement, the difference in correlation was almost negligible (Table 1), which thus demonstrated the necessity of considering the contextual sequence features of TISs in the prediction model. We also considered the length of 5’ UTR (which is presumably related to the number of undetected uORFs) in the regression model to test whether those undetected uORFs are influencing TE other than the detected TISs. Test results showed that this additional feature did not affect the prediction performance largely (Table 1), implying that very unlikely our method can suffer from the false negative issue.

### 3.4 TITER quantifies mutational effects on translation initiation

Previously, several studies have identified mutations that are putative to affect TE through the alteration in the sequence contexts of TISs (Kozak, 2002; Noderer et al., 2014; Wolt et al., 2011) or the introduction of extra uORFs (Barbosa et al., 2013; Calvo et al., 2009; Hinnebusch et al., 2016). As TITER has been shown to be able to accurately predict *bona fide* TISs in human transcriptome, we further explored its ability to quantify the mutational effects that may be related to the physiological and pathological conditions. In particular, we selected two sets of mutations that have been validated through different quantitative reporter assays in Noderer et al. (2014) and Calvo et al. (2009), respectively, and then evaluated how the changes of the predicted scores associated with the mutations can reflect their functional effects in *vivo*.

Through flow cytometry, Noderer et al. (2014) quantitatively measured the effects of seven mutations derived from the COSMIC database (Forbes et al., 2015) and observed consistent effects with other known tumor expression patterns (Fig. 6a). Specifically, they employed a dual fluorescence vector with a GFP reporter under the control of a specific TIS context as well as an independent IRES-driven RFP reporter as the internal standard, and the final result was reported with the GFP/RFP ratio. As the experiments were performed by expressing the reporter from a plasmid, we fed the plasmid sequences containing each TIS into the TITER framework, and calculated the changes of the predicted scores (i.e. ContextScore) along with the mutations. As expected, the changes of ContextScore values were in good agreement with the changes of the experimentally measured GFP/RFP ratios (Fig. 6b; Pearson’s correlation coefficient $r=0.83$), indicating that TITER was able to capture the TIS contexts that are related to TE, even though the mutation information was not included in our training data.

In another study, Calvo et al. (2009) carried out a series of luciferase assays to demonstrate the effects of the additionally introduced non-overlapping uORFs on protein expression of clotting factor XII (FXII). Their study involved six sequence variants that were associated with the non-overlapping uORFs (Fig. 6c). Similar to the above analysis, the sequence contexts of individual TISs, including both aTISs and uTISs, were fed into the TITER framework as input. We observed a good correlation even when only considering the changes of the prediction scores for aTISs (Pearson’s correlation coefficient $r=0.75$). Notably, the correlation was further improved (Pearson’s correlation coefficient $r=0.85$), when the ContextScore values of uTISs were included through a linear leaky scanning model proposed by Ferreira et al. (2013) (Fig. 6d), i.e.

$$\text{ContextScore}_r = (1 - k \times \text{ContextScore}_a) \times \text{ContextScore}_r$$

where ContextScore stands for the calibrated ContextScore, ‘a’ and ‘u’ denote the ‘annotated’ and ‘upstream’ TISs, respectively, and k is the model parameter. Note that here we set $k=0.86$ based on the previous result of the synthetic reporter assay reported in Ferreira et al. (2013). Together with our previous genome-wide analysis, these results demonstrated the ability of TITER to quantitatively evaluate the repressive effects of uTIS/uORF on protein expression.

To further confirm that the signal differences presented above were not due to the experimental bias and truly reflected the physiological or pathological effects, we also performed two additional analyses where the input sequences were changed to the corresponding sequences of the real transcripts. As expected, the changes of ContextScore for the transcript sequences still maintained a good consistency with the changes of experimental signals in both studies, with only a moderate fluctuation in the correlation, which thus further validated the biological relevance of our results (Supplementary Fig. S3).

### 4 Discussion

The prediction of TISs has long been considered an important task in the studies of gene expression regulation (Jackson et al., 2010;
Deep learning for predicting translation initiation

Hershey et al., 2012; Sonenberg and Hinnebusch, 2009). For years, progress has been made on the identification of the contextual sequence features that can prompt translation initiation, with the focus on the AUG start codon of the main ORFs (Chew et al., 2016; Hatzigeorgiou, 2002; Kozak, 1989; Zien et al., 2000; Zur and Tuller, 2013). However, the existence of alternative initiation codons and multiple initiation positions in vivo further complicate this prediction problem. Herein, we have developed a deep learning based framework, called TITER, that can automatically extract intrinsic sequence features from the experimentally identified TISs. By integrating both the preference of codon composition and the contextual sequence information, our unified framework was able to accurately predict various types of TISs. The subsequent motif analysis has expanded our current understanding of the sequence contexts of favored TIS codons. Furthermore, additional analyses of gene expression and mutations showed that TITER can be applied to accurately estimate the probability of translation initiation in various biological scenarios.

Recently, Reuter et al. proposed a linear model, called PreTIS, that was trained based on ribosome profiling data and can predict both AUG and non-AUG TISs in the 5' UTR (Reuter et al., 2016). However, the application of this method is limited by the codon position (i.e. at least 99 nts downstream from the transcript start site and in the 5' UTR) and the existence of mouse orthologous. Since TITER does not rely on any explicit feature engineering, it possesses the generality of using any input sequence for prediction. Moreover, the extensive tests have shown that TITER can greatly outperform PreTIS, further demonstrating the superiority of our method.

Previously, a number of techniques based on ribosome profiling have been developed to identify and characterize the TISs. In particular, Lee et al. (2012) developed the GTF-seq technique, which employed an initiation-specific small molecule ribosome inhibitor, called lactimidomycin, to capture TISs with both AUG and alternative codons on a genome-wide scale. As an updated version of this method, Gao et al. proposed a dual inhibition technique, called QTI-seq, to address the limitation of the previous method regarding the amplification of ribosome signals on TISs and the inflation of signals at the 5' ends of transcripts, which achieved a quantitative profiling of initiating ribosomes (Gao et al., 2015). Moreover, they also applied stringent mapping protocol and statistical test to increase the fidelity of the identified TISs. Therefore, in our study we mainly chose the dataset generated from Gao et al. (2015) to facilitate the training and test of TITER. With the careful selection of data source and the accurate prediction performance on various test settings, we believe that TITER will be useful for the community to investigate probable TISs and further expand our understanding of the mechanisms underlying translation initiation.

Funding
This work was supported in part by the National Basic Research Program of China Grant 2011CB00300 and 2011CB00301, the National Natural Science Foundation of China Grant 61033001, 6136136003 and 61472205, the US National Science Foundation Grant DBI-1262107 and IIS-1646333, the China’s Youth 1000-Talent Program, and the Beijing Advanced Innovation Center for Structural Biology.

Conflict of Interest: none declared.

References


