UNIVERSITY OF TARTU
Institute of Computer Science
Computer Science Curriculum

Mark Kerner

Predicting Specific Protein Abundances and Translation Efficiency in Yeast

Bachelor's Thesis (9 ECTS)

Supervisor: Leopold Parts, PhD

Tartu 2020

Predicting Specific Protein Abundances and Translation Efficiency in Yeast

Abstract:

Protein abundances and production rates differ by several orders of magnitude among protein-coding genes. We show that specific protein abundances can be modeled as a function of mRNA abundances, translation efficiency and protein turnover. We then match RNA sequence motifs of 21 RBPs on the untranslated regions of 774 target transcripts. We determine that the RNA sequence motifs on target transcripts untranslated regions do not predict variability in translation efficiency of these transcripts. We use yeast *Saccharomyces cerevisiae* as the model organism.

Keywords: linear regression, sequence motifs, RNA-binding proteins, translation efficiency, protein abundance, Saccharomyces cerevisiae

CERCS: B110 Bioinformatics, medical informatics, biomathematics, biometrics

Kindlate valkude arvukuse ja translatsiooni efektiivsuse ennustamine pärmi näitel

Lühikokkuvõte:

Valkude arvukus ja tootmise määr rakus erinevate valkude vahel erineb mitme suurusjärgu võrra. Me näitame, et kindlate valkude arvukust on võimalik täpselt ennustada sõnumi-RNA (mRNA), translatsiooni efektiivsuse ja valkude lagundamise funktsioonina. Seejärel valime 21 RNA-d siduvat valku (RBP) ja loeme kokku nende järjestusmotiivide esinemise 774 mRNA mitte-kodeeritud piirkondades (UTR). Näitame, et antud RBP-de järjestusmotiivid mRNA-de mitte-kodeeritud piirkondades ei ennusta nende mRNA-de translatsiooni efektiivsuse väärtuste variatsiooni. Kasutame mudelorganismina pärmseent *Saccharomyces cerevisiae*.

Märksõnad: lineaarne regressioon, järjestusmotiivid, RNA-d siduvad valgud, translatsiooni efektiivsus, valkude arvukus, Saccharomyces cerevisiae

CERCS: B110 Bioinformaatika, meditsiiniinformaatika, biomatemaatika, biomeetrika

Contents

1 Introduction

Cells engineered for enhanced production of chemical products is an established and economically important discipline [1]. These cell factories, which serve as the basis of all product-oriented industrial biotechnology, already have an economic footprint of hundreds of billions of USD per year [1]. As an example of their usefulness, microbial cell factories are able to utilize industrial by-products such as carbon dioxide to produce more desirable outcome chemicals [2]. One of the preferred host organisms for cell factories is yeast *Saccharomyces cerevisiae*.

To make these processes cost effective, we need to understand cellular regulatory mechanisms. Protein production is a key factor in cellular regulation, where differences in protein abundances and production rates span several orders of magnitude [3]. Understanding the regulation of protein production would allow to create novel more efficient strains of target host organisms for recycling and sustainable production of chemicals.

Gene expression for protein-coding genes is broadly split into two parts: transcription and translation. In transcription, RNA polymerase synthesizes a messenger RNA (mRNA) product, also called a transcript, from a DNA sequence. In translation, a ribosome synthesizes the final protein product from an mRNA. Multiple protein instances can be synthesized from a single mRNA strand. Also, multiple ribosomes can be bound to the same mRNA at the same time, allowing for multiple protein instances to be synthesized simultaneously [4]. The rate of protein production can vary more than 1000-fold between proteins [3].

One of the key determinants of protein production rate is translation efficiency, which describes how many proteins are synthesized for a transcript in a unit of time, commonly an hour. Codon bias, where synonymous codons, which do not alter the protein produced, but result in higher abundances, and mRNA folding have been identified as key factors in determining translation efficiency [5]. Additionally, RNA binding proteins (RBPs) are reported to regulate translation generally [6] and translation efficiency specifically [7]. RBPs bind with target mRNAs to form mRNA–protein complexes, also called messenger ribonucleoproteins (mRNP). Many RBPs bind to specific RNA sequences, called binding motifs. Several RBPs and their binding motifs have been identified for *S. cerevisiae* [8, 9].

We first set out to determine wether protein abundances can be predicted from mRNA abundances, translation efficiency and protein turnover. We then turn our focus on translation efficiency to determine wether in can be predicted from RNA sequence motifs of specific RBPs. We aggregated binding motifs of 21 RBPs and counted their occurrences on 774 mRNA untranslated regions. We train linear regression models to predict translation efficiency from motif matches and use these models to determine the variability in translation efficiency that can be explained by motif matches. We used yeast *Saccharomyces cerevisiae* as the model organism.

This analysis is based on data from Lahtvee et al. and Giudice et al [3, 9]. Lahtvee et al. provided protein and mRNA abundances, translation efficiency, protein turnover and mRNA sequences. Giudice et al. provided RNA sequence motifs for RBPs.

The contents are divided into two sections. In the first section we describe the data, the motif analysis and the models. In the second section we fit the models to data and evaluate the results. We see that protein abundances can be modeled as a function of mRNA abundances, translation efficiency and protein turnover. We then determine that none of the RBPs' motifs account for variability in translation efficiency.

I would like to thank my supervisor, Leopold, for his patience, encouragement and advice.

2 Materials and methods

2.1 Data

Lahtvee et al. [3] and Giudice et al. [9] provided the primary sources of data.

Lahtvee et al. measured absolute levels of individual transcripts and proteins in *S. cerevisiae* under ten environmental conditions. We used the measurements at the reference condition. They also measured protein turnover for individual proteins. At reference condition, they measured absolute levels of transcripts and proteins for 5354 and 1786 genes respectively. Protein turnover was measured for 1384 proteins. As a function of these measurements, they calculated translation efficiency of 1115 genes under the optimal environmental condition using the equation

$$k_{TL,i} = \frac{C_{prot,i}(k_{deg,i} + \mu)}{C_{mRNA,i}}, \tag{1}$$

where $C_{prot,i}$ and $C_{mRNA,i}$ refer to the measured absolute protein and mRNA abundances at steady state conditions and $k_{deg,i}$ and μ to protein turnover and specific growth rate respectively. Specific growth rate, equal to the fixed dilution rate of chemostats used in the experiments, was 0.1 h^{-1}. Protein turnover was defined as the rate at which a protein is degraded, measured in units of $1/h$ i.e. the fraction of protein abundance degraded in an hour [3]. Translation efficiency was measured in units of proteins synthesized per mRNA per hour.

Lahtvee et al. also provided mRNA 5' UTR and 3' UTR sequences (Figure 1) for 774 genes, which were a subset of the genes that translation efficiencies were provided for.

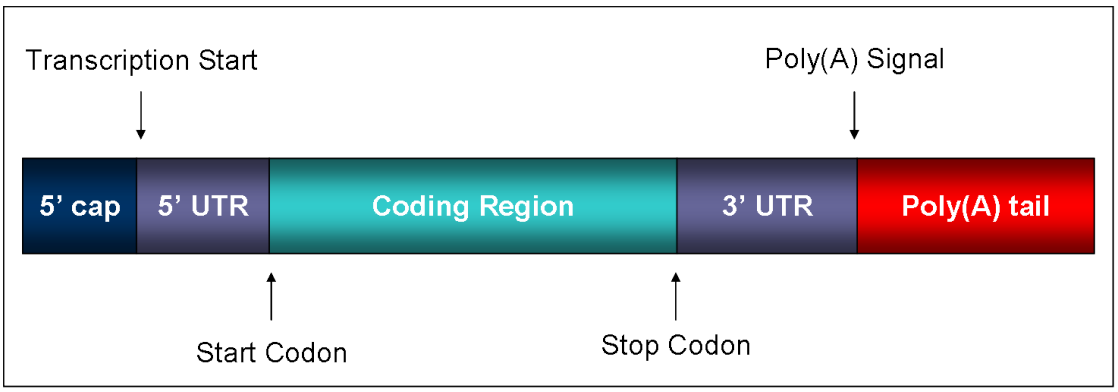

Figure 1: Generic structure of an eukaryotic mRNA [10]

To evaluate the RNA binding proteins effect on translation efficiency, we used the number of matches of its known motifs on target RNA untranslated regions

Gene name	Motif count
PUF4	13
MSS116	9
RNA15	6
PUF3	6
MSL5	5
VTS1	4
EMG1	4
MPT5	4
PRP24	3
PUF2	3
RPB2	2
NPL3	2
DIS3	2
NAB3	2
HRP1	2
PAP1	1
NAB2	1
MTR4	1
RPO21	1
RRP6	1
DBP5	1

Table 1: Number of motif matrices per RNA binding protein provided by Giudice et al. [9]

(UTRs). Giudice et al. [9] provided 370 motifs of 21 different RNA binding proteins (RBP) for yeast *Saccharomyces cerevisiae*. Motifs were represented as position probability matrices (PPM). The 370 motifs listed in database had been generated by 73 unique motif matrices - several consensus sequences can be generated from the same motif matrix if these sequences have equal scores, or probabilities. We used the source 73 motif matrices to find matches on target sequences, not the database entries of precalculated consenus sequences. Therefore, we attained 73 motifs for 21 RNA binding proteins: one RBP could have multiple motif matrices, but no motif matrix was used for more than one RBP, in other words, there was a one-to-many relationship (Table 1).

Motif matches were counted on all target genes 5' UTR and 3' UTR sequences separately. The motif matches were then aggregated under their respective RNA binding proteins (Table 2). This resulted in two $M \in \mathbb{R}^{774\times21}$ matrices, one for either UTR, where $a_{i,j}$ is RBP j motif count on a target genes i 5' or 3' UTR.

RBP	Matches
NAB3	1758
PAP1	1559
RRP6	1559
MSS116	1017
EMG1	655
MTR4	600
RNA15	570
DIS3	367
DBP5	326
PUF4	289
PRP24	223
RPO21	151
NAB2	123
PUF3	116
MSL5	67
HRP1	58
MPT5	55
RPB2	45
NPL3	44
VTS1	43
PUF2	35

RBP	Matches
NAB3	1934
PAP1	1597
RRP6	1597
MSS116	1045
EMG1	660
MTR4	626
RNA15	525
DIS3	359
DBP5	309
PUF4	293
PRP24	258
RPO21	173
PUF3	134
NAB2	123
MSL5	101
HRP1	72
MPT5	58
RPB2	50
NPL3	49
PUF2	39
VTS1	39

Table 2: Total motif count per RBP on 5' UTRs (left) and 3' UTRs (right).

2.2 Motif analysis

Sequence motifs are short, recurring patterns of nucleotide sequences in DNA and RNA that are presumed to have a biological function. Motifs often indicate sequence specific binding sites for proteins [11]. In this work, we are interested in RNA motifs. These are RNA sequences that indicate binding sites for RNA-binding proteins. The respective regions for binding on the RNA-binding proteins are known as RNA-binding domains.

Some proteins need to bind to their targets in a highly sequence-specific manner and they do not stray from the specificity of their consensus binding site. Other proteins are less sensitive to sequence specificity. The affinity of a binding protein to a specific binding site is typically correlated with how well the site matches the consensus sequence [11].

To capture the variable frequency of bases at each position in a motif sequence, motifs are often represented as a Position Frequency Matrix (PFM) [11]. In a PFM, we record how often each base occurs in known binding sites. PFM can then be

transformed into a Position Probability Matrix (PPM), which we can simply do by converting base counts to probabilities. The latter can in turn be transformed into a Position Weight Matrix (PWM), also known as Position-Specific Scoring Matrix (PSSM), if we assign background weights for the bases. Background weights are typically assigned based on the background genomic sequence, to account for the possible non-equal frequency of bases in the genome. The organism under study in this work, *Saccharomyces cerevisiae*, has a genome with a somewhat biased GC content of 38%, which should be accounted for [11].

2.2.1 Reading motifs

Bio.motifs package from *biopython* tooling was used to read, parse and match motifs [12]. Giudice et al. [9] provided a single file that included position probability matrices (PPMs) for all motifs in the database. Individual matrices were separated by headers, starting with > and included the matrix id and motif length. Each PPM was a matrix $M \in \mathbb{R}^{m \times 4}$, where m is motif length and element $a_{i,j}$ corresponds to the probability of encountering nucleotide j at motif sequence position i. Therefore, PPM rows represented motif sequence positions and columns represented nucleotides, ordered as adenine (A), cytosine (C), guanine (G) and thymine (T) or uracil (U). The sum of probabilities in a row is always 1.0. We used the *Bio.motifs* package to parse each PPM into an object of class *Motif* from the same package. We used the DNA alphabet (ACGT) to read and store the motifs. This suited our purposes, as the UTR sequences provided by [3] were also in the DNA alphabet.

2.2.2 Matching motifs

To count motif matches on sequences, a position specific scoring matrix (PSSM) was generated for each motif. The class *PositionSpecificScoringMatrix* from *Bio.motifs* package was utilized, which can be used to search for matches along a target sequence. A PSSM contains log likelihood scores computed from the PPM and background probabilities. In our case a non-uniform, biased background probability vector b was used to account for the biased GC content of 38% in *Saccharomyces cerevisiae*. Therefore, PSSM elements were calculated by transforming PPM values so that:

$$M_{i,j} = log_2(M_{i,j}/b_j)$$

where $b = [0.31, 0.19, 0.19, 0.31]$.

PSSM searches for matches by scoring all positions on the target sequence. Each position gets a score, therefore a threshold score is necessary to obtain meaningful results. Threshold score can be calculated based on a PSSMs scores distribution. Class *PositionSpecificScoringMatrix* has a method to calculate a distribution of the scores at a given precision. We used the default precision of 10^3 - raising it raises

computational cost non-linearly without significant change in the final threshold score. The resulting distribution object is of class *ScoreDistribution* which has a method for approximating the threshold score which makes the type I error (false positive rate). For every motif, a threshold score was calculated from its PSSM scores distribution for false positive rate of 10^{-4} or 0.01%.

2.3 Linear prediction models

2.3.1 Multiple Linear Regression

Linear regression is a fundamental statistical modelling technique. it is simple, yet powerful for uncovering linear relationships and making predictions based on these linear relationships between the model predictors and the outcome. It can also detect non-linear relationships if the variables are transformed in a way that makes the relationship between the model and outcome linear as a result, assuming that other assumptions still hold.

The main purpose of a linear regression model is to determine how the average value of the continuous variable y varies with predictors $\boldsymbol{x}$. The average values of the outcome are assumed to lie on a regression line, defined by equation

$$E[y|\boldsymbol{x}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p, \tag{2}$$

where $\boldsymbol{x}$ represents the collection of p predictors $x_1, x_2, \ldots, x_p$ in the model, and $\beta_1, \beta_2, \ldots, \beta_p$ are the corresponding regression coefficients. $E[y|\boldsymbol{x}]$ is shorthand for the expected or average value of outcome y at given values of predictors $\boldsymbol{x}$. The intercept, β_0, gives the average value of the outcome, when all the predictors are zero [13].

In (2), the coefficient $\beta_j, j = 1, \ldots, p$ gives the change in $E[y|\boldsymbol{x}]$ for an increase of one unit in predictor x_j, holding other factors in the model constant. The latter meaning that each of the estimates is adjusted for the effects of all the other predictors. If confounding has been convincingly ruled out, then the adjusted coefficient estimates can be interpreted as representing causal effects [13].

The individual observations of the outcome y_i are modelled to vary from the average determined by their predictor values $\boldsymbol{x}$ by an error term ε_i, so that

$$y_i = E[y_i|\boldsymbol{x}_i] + \varepsilon_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi} + \varepsilon_i,$$

where x_{ji} is the value of the predictor variable x_j for observation i [13].

All the statistical assumptions of the linear regression model concern the distribution of ε. Specifically, we assume that $\varepsilon_i \sim i.i.d\, \mathcal{N}(0, \sigma_\epsilon^2)$, meaning that ε is independently and identically distributed and has a normal distribution, mean zero at every value of $\boldsymbol{x}$, constant variance σ_ε^2 at every value of $\boldsymbol{x}$ i.e. homoscedasticity

and values that are statistically independent. In large datasets, the first assumption of normal distribution of ε can sometimes be relaxed. The second assumption of mean zero is important for checking for the linearity of the relationship between numerical predictors and the outcome. Violations of linearity can be fixed by transformations of variables, adding a quadratic term or other methods. When the fourth assumption of statistical independence does not hold, further methods have to be introduced [13].

Unlike the outcome, no distributional assumptions are made about the predictor in a linear regression model. Although no assumptions are made about its distribution, the models tend to perform better when it is relatively variable [13].

Fitting the regression model to sample data amounts to finding estimates $\hat{\beta}$. Linear regression uses the ordinary least squares (OLS) estimation to find $\hat{\beta}$ values. When assigning the regression line, ordinary least squares tries to minimize the sum of squared vertical distances (differences) of data points to the regression line. This makes OLS easy to compute. Additionally, OLS estimates have desirable statistical properties. They are minimally variable and well behaved in large samples even when the distributional assumptions concerning ε are not precisely met. A drawback of OLS is its sensitivity to outliers, which arises from squaring of vertical differences [13].

The OLS estimates $\hat{\beta}$ determine the fitted value $\hat{y}$ for every data point:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \ldots + \hat{\beta}_p x_{pi}$$

Residuals are the difference between observed and fitted values, defined as:

$$r_i = y_i - \hat{y}_i,$$

and function as the sample analogue of ε. This makes them important in fitting the model, in estimating the variability of the parameter estimates, and in checking model assumptions and fit [13].

Central to OLS are various sums of squares. First is the total sum of squares (TSS):

$$TSS = \sum_{i=i}^{n} (y_i - \bar{y})^2,$$

where $\bar{y}$ is the sample average of the outcome y. This captures the total variability of the outcome about its mean. The TSS is split into two parts. The first is the model sum of squares (MSS):

$$MSS = \sum_{i=i}^{n} (\hat{y}_i - \bar{y})^2.$$

This captures the variability of the outcome about its mean that can be accounted for by the model. The second part is the residual sum of squares (RSS):

$$RSS = \sum_{i=i}^{n} (y_i - \hat{y}_i)^2.$$

RSS is part of the TSS that cannot be accounted for by the model. By definition, RSS is minimized by the fitted regression line. The identity $TSS = MSS + RSS$ holds true.

The sums of squares determine the coefficient of determination, R^2, such that

$$R^2 = \frac{MSS}{TSS}.$$

Coefficient of determination can be interpreted as the proportion of the variability of the outcome that is accounted for by the model.

2.3.2 Evaluating models

If the model is evaluated on the same data that was used for training it, the estimated error can be too optimistic and is unlikely to generalize to out-of-sample data. This is called overfitting. Therefore it is a conventional practice to hold out a part of the dataset, before training and selecting models, to later evaluate their performance on what we call out-of-sample data. We call the part of data held out for evaluating models a testing set and the part used for training the models a training set. However, if we use the testing set to iteratively evaluate and improve the models, then the models would get optimized for the testing set and the reported results would again be too optimistic. Therefore the testing set must only be used for a final evaluation of the model to get an estimate of how well the model is able to generalize to new cases.

To evaluate models during training, we used a procedure called k-fold cross-validation. In k-fold cross-validation the training set is split into k subsets. Each subset goes through a procedure where it is held out as a testing set, the remaining data is used to train the model and the trained model is then evaluated on the held out subset. This results in k evaluations. We used the mean of these k evaluations as our final estimate of a model's performance.

We used two metrics for evaluation: the coefficient of determination (R^2) and root-mean-squared error (RMSE).

Coefficient of determination is described in section 2.3.1 on multiple linear regression. There is a difference, in that the R^2 that we use for evaluation is calculated on out-of-sample data, whereas that section describes calculating it on fitted values.

RMSE is the root of mean squared error (MSE), where mean squared error is the mean of squared residuals, such that

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=i}^{n}(y_i - \hat{y}_i)^2}.$$

RMSE can be interpreted as the standard deviation of the residuals [13]. As such, it is comparable to the standard deviation of the outcome. For a good predictive model we can expect a RMSE much lower than the outcome standard deviation in the sample on which the predictions were made.

2.4 Prediction model assumptions

2.4.1 Normality and log-transformation of variables

In a multiple linear regression model, the error term ε is assumed to have a normal distribution. The sample analogue for the error term ε is the distribution of residuals. Although it is possible for the residuals to be normally distributed when the outcome y is not, and conversely, then transforming the outcome y to get a more normal distribution with reduced skewness is often successful for also reducing the skewness of the distribution of residuals [13]. As mentioned in subsection 2.3.1, the models also tend to perform better when the distributions of the predictors are relatively variable.

Transformations affect the distribution of a variable by emphasizing and de-emphasizing differences in certain ranges of values of the variable. One such transformation is to replace y with $\log(y)$. Log transformation emphasizes differences at the lower end of the scale and de-emphasizes those at the higher end [13].

The distribution of translation efficiency is highly right-skewed, taking a form comparable to an exponential distribution (Figure 2). We log-transformed translation efficiency values to emphasize differences at the lower end and de-emphasize differences in the higher end. The resulting log-transformed values of translation efficiency are approximately normally distributed, with only slight deviations at the ends (Figure 2). Log-transformed translation efficiency was used as both a predictor (subsection 3.1) and an outcome (subsection 3.2).

When the outcome is log-transformed, predictor coefficient estimates are interpreted differently. With natural log transformation of the outcome,

$$100(e^{\hat{\beta}} - 1),$$

where $\hat{\beta}$ is the coefficient estimate, is interpretable as percentage increase in the average value of the outcome per unit increase in the predictor [13]. For example, if

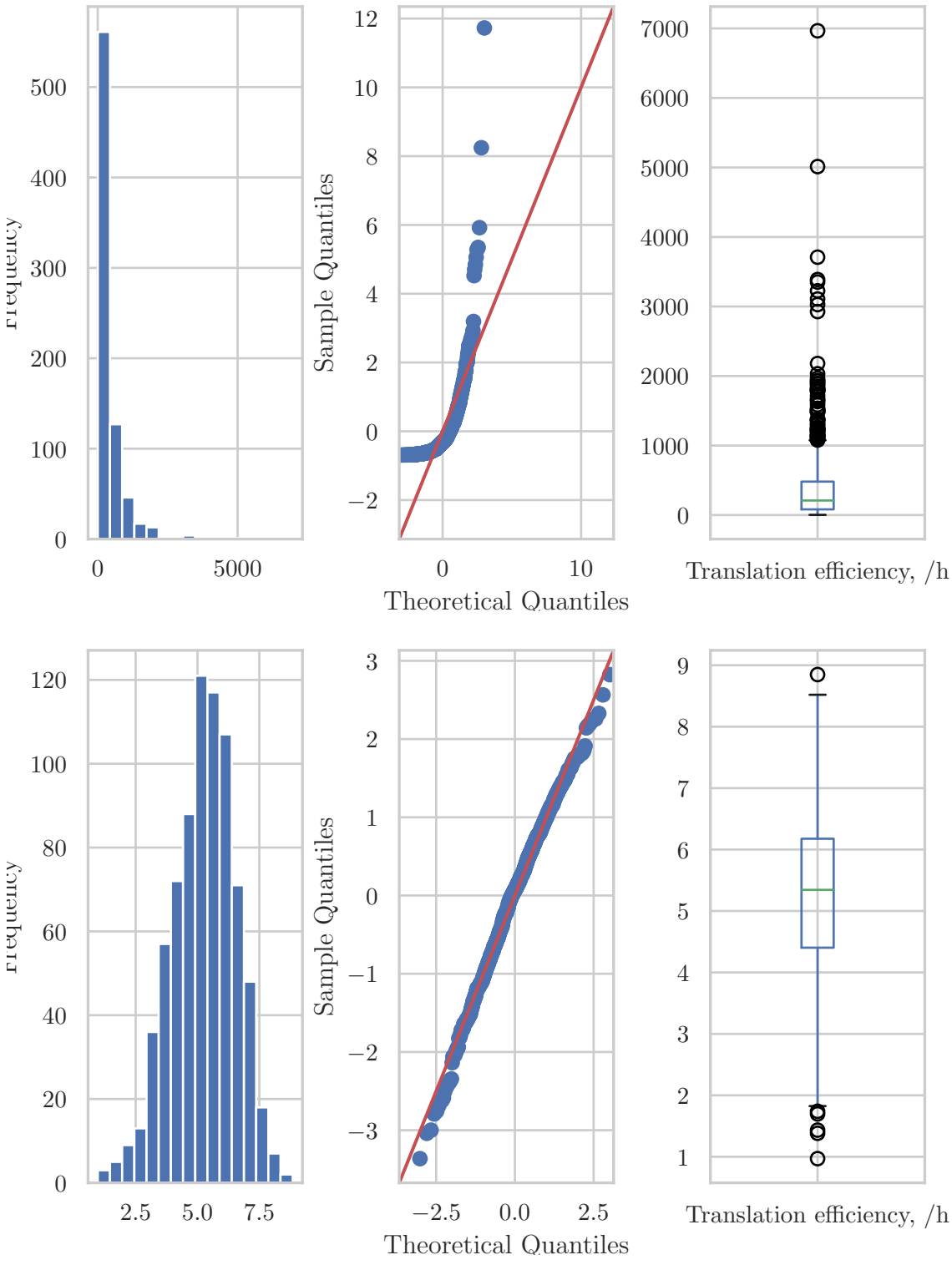

Figure 2: Distributions: translation efficiency (up) vs. log-transformed translation efficiency (bottom)

we train a linear regression model with some RBPs motif count as a predictor and log-transformed translation efficiency as outcome, then if the coefficient estimate for the motif count predictor variable is $\hat{\beta} = 0.1$, then the model predicts a $100(e^{0.1} - 1) = 10.52\%$ increase in translation efficiency for every added motif of that RBP.

When both predictor and outcome are transformed using natural logs, then

$$100(e^{\hat{\beta} \ln 1.01} - 1)$$

can be interpreted as the percentage increase in the average value of the outcome per 1% increase in the predictor [13]. For example, if we train a linear regression model with log-transformed translation efficiency as predictor and log-transformed protein abundance as outcome, then if the coefficient estimate for the log-transformed translation efficiency is $\hat{\beta} = 0.1$, then the model predicts a $100(e^{0.1 \ln 1.01} - 1) = 0.10\%$ increase for every 1% increase in the predictor.

The interpretation on coefficient estimates is based on the interpretation of the differences between log-transformed translation efficiency values. While the absolute predictions on log-transformed translation efficiency can be transformed back to be on a non-logarithm scale, the differences don't behave like that and can't be transformed back to the original scale of translation efficiency (h^{-1}). Instead, they have to be interpreted as percentage change in translation efficiency.

2.4.2 Confounding and collinearity

As individual correlations of RBPs' motif count with translation efficiency are low (Table 3), care must be taken to correctly model the effect of these predictors on translation efficiency. We first describe how confounding, and then collinearity, can introduce bias into parameter coefficient estimates.

Vittinghoff et al. [13] define confounding as present when the difference in mean values of the outcome between populations defined by a potentially causal variable of interest is not equal to its causal effect on that outcome. As a consequence, the regression coefficient estimate for the causal variable given by fitting a linear model to a sample of data from the combined population will be biased for the causal effect. In that scenario, the association between the variable of interest and outcome is confounded by another effect i.e. a confounder. [13].

Causality, in turn, or a causal effect of an exposure on a continuous variable is defined as the difference between population means in the presence as compared to the absence of exposure, when the actual and counterfactual outcomes are observed for every member of the population, holding all other variables constant. If the means differ, then the exposure is a causal determinant of the outcome. Note that the causal effect is defined on the difference of population means. This does not rule

RBP	Pearson's r
NAB3	-0.0849
RNA15	-0.0676
NPL3	-0.0668
DBP5	-0.0617
PRP24	-0.0558
HRP1	-0.0521
NAB2	-0.0484
MSS116	-0.0434
EMG1	-0.0361
RPB2	-0.0294
MTR4	-0.0262
RRP6	-0.0261
PAP1	-0.0261
DIS3	-0.0154
RPO21	-0.0079
MSL5	0.0125
PUF4	0.0182
MPT5	0.0503
PUF3	0.0532
PUF2	0.0867
VTS1	0.0924

RBP	Pearson's r
RPB2	-0.0901
MSL5	-0.0711
NAB2	-0.0419
HRP1	-0.0384
PUF2	-0.0092
PUF4	-0.0051
RPO21	-0.0050
NPL3	-0.0026
PUF3	-0.0020
RRP6	-0.0019
PAP1	-0.0019
MPT5	0.0013
NAB3	0.0114
DBP5	0.0144
PRP24	0.0149
VTS1	0.0213
RNA15	0.0229
DIS3	0.0253
MTR4	0.0268
MSS116	0.0287
EMG1	0.0620

Table 3: RBPs motif count on 5' UTRs (left) and 3' UTRs (right) correlated to log-transformed translation-efficiencies.

	Pearson's r
Length CDS	-0.44
Length 5' UTR	-0.08
Length 3' UTR	0.01
Total length	-0.43

Table 4: UTR lengths correlated with log-transformed translation efficiency

out variation in the causal effects of exposure at the individual level, as individual causal effects can exist in the absence of an overall population causal effect. Also, the causal effect of a predictor on an outcome need not be a direct causal effect and can exist anywhere on the causal pathway, mediated by the factors closer to the direct cause [13].

In reality, however, it is generally not possible to observe counterfactuals, so for each individual the outcome is observable for only one of the conditions. Multiple

regression model can control for confounding by holding other causal determinants of the outcome constant. If all causal determinants of the outcome Y are included as the model predictors, then the coefficient estimates for these predictors can be interpreted as causal effects of these predictors.

We can identify potential confounders on a priori grounds. The potential confounder must be associated with both the predictor of interest and outcome [13]. UTR length is one such confounder, as it affects both translation efficiency and motif count - we describe this in-depth below. Furthermore, the distribution of nucleotides in UTR could affect both RBP motif counts and translation efficiency, for example through mRNA folding [5]. However, ruling out the latter is beyond the scope of this thesis. There could be additional confounders, which we have not identified. Therefore we have to be conservative when assigning causal interpretations to RBPs motif counts coefficients.

The number of different positions for placing a motif is given by subtraction of lengths $l_{UTR} - l_{motif} + 1$, as such, the space of possible positions on an UTR increases linearly with its length. This introduces a correlation between the number of motif matches and UTR length. Correlation between total matches on UTRs across all RBPs and UTR lengths are $r = 0.67$ and $r = 0.64$ for 5' and 3' UTR respectively. Additionally, UTR lengths have some correlation with translation efficiency (Table 4). As a result, UTR length may confound the association between translation efficiency and a RBPs motif count. Therefore, the estimated coefficient for a RBPs motif count would be biased, if we don't adjust for UTR length. Although UTR lengths have weak correlations with translation efficiency, these can become significant if there are many predictors of interest with small effects as is the case with RBP features (Table 3, 4) [13].

Confounding can either attenuate or increase the coefficient estimates of target variables. At one extreme, the effect of a factor of interest may be completely confounded by a second variable. At the other extreme, the unadjusted analysis might produce no association between the predictor and outcome as the predictor will be negatively confounded by another predictor. One way negative confounding can happen is when the two predictors are positively correlated, but have regression coefficients with the opposite sign [13].

Negative confounding might explain why most RBPs correlate negatively with translation efficiency on 5' UTR (Table 3). 5' UTR length is negatively correlated with translation efficiency with $r = -0.08$ - if motif counts have small effects, this small negative correlation can be enough for UTR length to completely confound the effect of motif counts on translation efficiency on RBPs with small negative regression coefficients or produce no association on RBPs with small positive regression coefficients. More positive correlations can be seen in 3' UTR correlations, which could be explained by a slightly positive correlation $r = 0.01$ between 3'

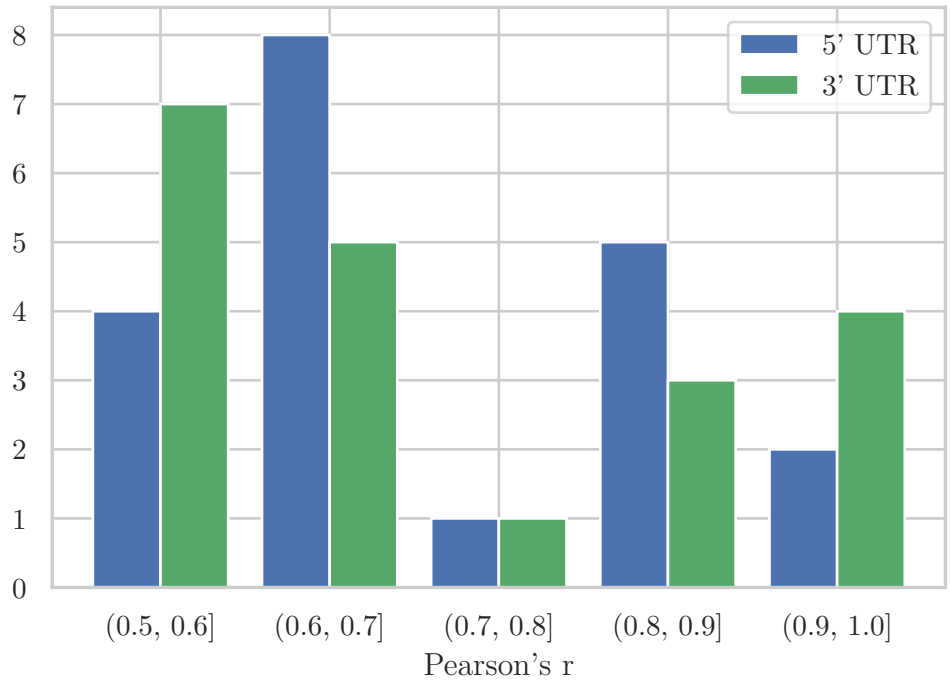

Figure 3: Distribution of correlations between RBPs' motif counts higher than $r = 0.5$ out of a total of $\binom{21}{2} = 210$ pair-wise combinations

UTR length and translation efficiency (Table 3). To adjust for UTR length, it must be included as an independent variable in the linear regression model alongside motif count.

Furthermore, there are significant correlations between RBPs' motif counts themselves, which can indicate collinearity. Out of $\binom{21}{2} = 210$ possible RBP pairs, 20 pairs have correlations higher than $r = 0.5$ and 7 pairs higher than $r = 0.8$ for both 5' and 3' UTR (Figure 3). This can be intuitively explained by two factors. As UTR length increases, the space for motif matches increases for all motifs, resulting in more counts on longer UTRs and fewer counts on shorter UTRs, for all motifs. This is also described by motif counts correlation with UTR lengths as described previously. Also, different RBPs' motifs can be similar - they can share a high percentage of nucleotides, subsequences etc. - therefore they could have similar counts. Two RBPs - PAP1 and RRP6 - share an identical motif, also the only motif for both, therefore their motif counts are equal across all target genes and their pairwise correlation for motif count is $r = 1.0$ on both UTRs. Collinearity

violates linear regression's assumption of independence, therefore care must be taken if multiple RBPs' motif counts are used as predictors or when interpreting such models.

3 Results

We start off by modelling protein abundance as a function of mRNA abundance, translation efficiency and protein turnover. We then direct our focus on translation efficiency to model it as a function of a RBPs motif count on a gene. By comparing models of different RBPs, we set out to determine which have most control over translation efficiency. Linear regression was used for all the models.

3.1 Predicting protein abundances

First, we model protein abundance as a function of mRNA abundance, translation efficiency and protein turnover. We modeled it using linear regression, with mRNA abundance, translation efficiency and protein turnover as the predictors and protein abundance as the outcome.

By taking the intersection of absolute protein and mRNA abundances, protein turnover and translation efficiencies provided by Lahtvee et al. [3], we gathered a total of 1092 genes that had all of the four values at the reference condition. We held out 20% of the data for testing, resulting in 874 instances for the training set and 218 instances for the testing set. As all the values had highly right-skewed distributions, we log-transformed them to achieve more normal distributions (Figure 4).

it is important to note that translation efficiency was not measured, but calculated from protein and mRNA abundances and protein turnover (1). We can rearrange (1), such that

$$C_{prot,i} = \frac{k_{TL,i}\ C_{mRNA,i}}{k_{deg,i} + \mu}. \tag{3}$$

Therefore, we are trying to emulate this formula using linear regression, omitting the μ for constant growth rate.

The model performed exceptionally well on both training and test data. On training data, the 5-fold cross-validation mean for R^2 was 0.99, meaning that the three predictors account for almost all of the variability in the outcome protein abundance. This makes sense if we consider the equation (3). The 5-fold cross-validation mean for RMSE on training data was 1266.93. This is an order of magnitude lower than the standard deviation of protein abundances of 17282.96 in training data. On testing data we repeated similar results with R^2 equal to 0.99 and RMSE equal to 1075.54. The standard deviation of protein abundances in testing data was 12108.50.

As both predictors and outcome were log-transformed, the coefficient estimates must be interpreted in changes in percentages of both predictors and outcome, as described in section 2.4.1. Both mRNA abundance and translation efficiency predict

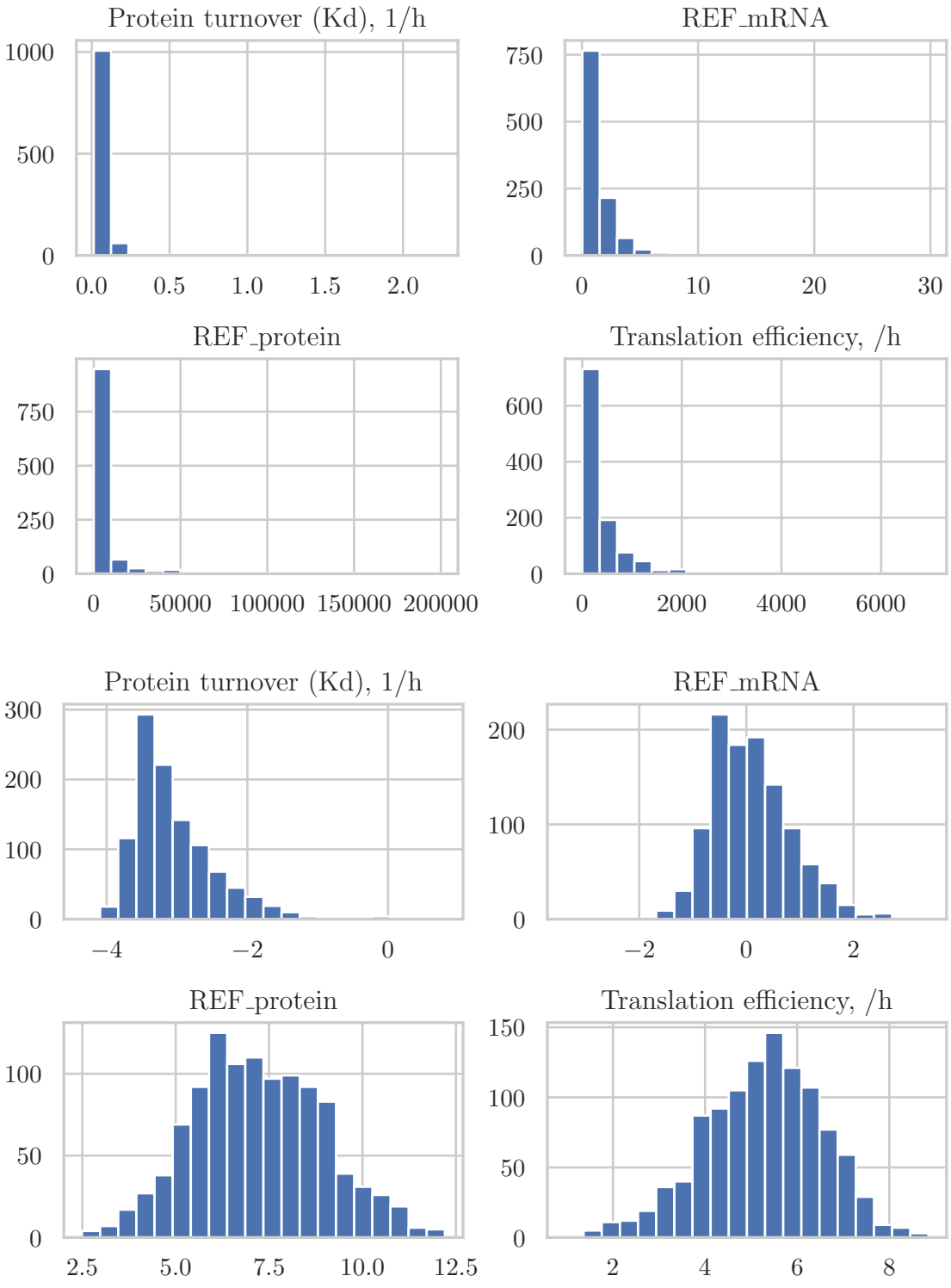

Figure 4: Distributions of values (up) and their log-transformations (down)

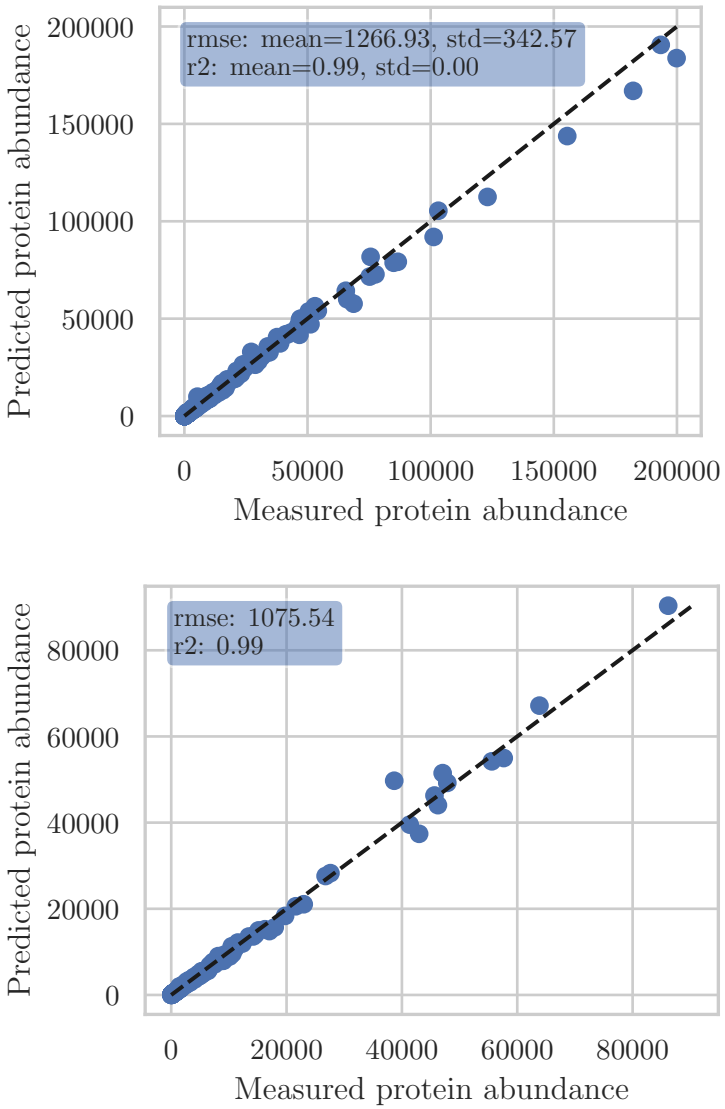

Figure 5: Measured vs. predicted: 5-fold cross-validation results on training data (up) and final evaluation on testing data (down). The values have been transformed back to the original scale.

close to 1% increase in the protein abundance, for their respective increases by 1% (Table 5). For its increase by 1%, protein turnover predicts a −0.47% decrease in protein abundance. it is important to note, however, that this model does not reflect how the protein abundance changes with different values of predictors

	intercept	C_{mRNA}	k_{TL}	k_{deg}
coefficient estimate	0.4958	0.9721	0.9856	-0.4764
% increase in outcome for 1% increase in predictor	–	0.97	0.99	-0.47

Table 5: Protein prediction model coefficients and intercept. Interpretation of coefficients.

for a *particular gene*, which would instead be subject to cellular mechanisms of translation [14]. It only models how protein abundances change when we compare the predictors of one gene to another at the same steady-state condition.

3.2 Predicting translation efficiency

In the previous subsection we modeled protein abundances as a function of mRNA abundance, translation efficiency and protein turnover and saw that these account for 99% of the variability in protein abundance values. We now turn our focus on translation efficiency, as one of the determinants of protein abundance, and investigate wether RBP motifs impose control over it. Specifically, we set out to measure the effects of RBPs' motif count on translation efficiency. We hypothesized that for some RBPs, their motif count on target genes could predict some part of the variability in translation efficiency universally across all of the subjected target genes.

To avoid running into issues caused by collinearity between RBPs' counts, we modeled the effect of each RBPs motif count independently using linear regression. We created two models for each RBP by modelling the effect of their 5' UTR and 3' UTR motif counts. We used RBPs motif count on an UTR and UTR length as predictors, and log-transformed translation efficiency as outcome. UTR length was added as a predictor to adjust for its confounding effect on translation efficiency. As we had motifs for 21 RBPs and created two models for each, we created 42 models in total. These models demonstrate the effect of a particular RBPs motif count on a particular UTR, on translation efficiency.

We aggregated motif counts of 21 RBPs on 774 target genes 5' UTR and 3' UTR, along with the length of these UTRs. The data was split into training and testing sets of 620 and 154 target genes respectively. We used 5-fold cross-validation to validate each model on training data. We scored models by their root-mean-squared error and their R^2 score.

Standard deviation of the log-transformed translation efficiency in the training set was 1.2638. The best model's (PUF2 on 5' UTR) cross-validation mean of root-mean-squared error was 1.2577 with a slightly negative R^2 score of -0.0110 (Table 6). Therefore none of the models had statistical value as a predictive model

	$intercept$	n_{RBP}	l_{UTR}	$RMSE_{CV_{mean}}$	$R^2_{CV_{mean}}$
PUF2	5.3807	0.5751	-0.0010	1.2577	-0.0110
VTS1	5.3831	0.4287	-0.0010	1.2586	-0.0117
PUF3	5.3837	0.2276	-0.0011	1.2591	-0.0129
MPT5	5.3884	0.3503	-0.0011	1.2607	-0.0157
HRP1	5.3879	-0.2198	-0.0008	1.2619	-0.0173
PRP24	5.3903	-0.1059	-0.0007	1.2620	-0.0174
NAB3	5.3898	-0.0211	-0.0006	1.2630	-0.0190
NPL3	5.3880	-0.1493	-0.0008	1.2636	-0.0199
PUF4	5.3858	0.0551	-0.0010	1.2639	-0.0203
RRP6	5.3910	-0.0097	-0.0008	1.2642	-0.0208
PAP1	5.3910	-0.0097	-0.0008	1.2642	-0.0208
MSL5	5.3893	0.1059	-0.0010	1.2643	-0.0209
EMG1	5.3874	0.0135	-0.0010	1.2644	-0.0210
RNA15	5.3892	-0.0227	-0.0008	1.2644	-0.0211
MSS116	5.3901	-0.0106	-0.0008	1.2656	-0.0232
MTR4	5.3930	-0.0209	-0.0008	1.2662	-0.0240
DIS3	5.3891	-0.0067	-0.0009	1.2665	-0.0245
DBP5	5.3884	-0.0341	-0.0008	1.2675	-0.0262
NAB2	5.3928	-0.1333	-0.0008	1.2677	-0.0265
RPB2	5.3894	0.1280	-0.0009	1.2695	-0.0293
RPO21	5.3905	0.0756	-0.0010	1.2701	-0.0300

	$intercept$	n_{RBP}	l_{UTR}	$RMSE_{CV_{mean}}$	$R^2_{CV_{mean}}$
NAB2	5.2756	-0.2372	0.0001	1.2688	-0.0281
EMG1	5.2768	0.0701	-0.0005	1.2698	-0.0305
HRP1	5.2737	-0.0912	-0.0001	1.2701	-0.0307
RRP6	5.2757	-0.0309	0.0003	1.2704	-0.0311
PAP1	5.2757	-0.0309	0.0003	1.2704	-0.0311
MSL5	5.2823	-0.1914	-0.0000	1.2707	-0.0316
PRP24	5.2800	-0.0241	-0.0001	1.2714	-0.0328
RPB2	5.2751	-0.3579	0.0000	1.2716	-0.0331
PUF4	5.2801	0.0267	-0.0002	1.2716	-0.0330
MTR4	5.2765	-0.0170	-0.0001	1.2717	-0.0331
NAB3	5.2784	0.0014	-0.0002	1.2725	-0.0346
MPT5	5.2781	0.0178	-0.0002	1.2726	-0.0347
VTS1	5.2769	0.1688	-0.0002	1.2727	-0.0354
RPO21	5.2778	-0.0080	-0.0001	1.2728	-0.0350
NPL3	5.2786	0.0610	-0.0002	1.2730	-0.0355
PUF3	5.2781	-0.0037	-0.0001	1.2732	-0.0356
PUF2	5.2783	0.0025	-0.0002	1.2735	-0.0360
DIS3	5.2783	-0.0048	-0.0001	1.2743	-0.0374
MSS116	5.2783	0.0095	-0.0002	1.2756	-0.0397
DBP5	5.2778	0.0140	-0.0002	1.2771	-0.0420
RNA15	5.2771	0.0152	-0.0002	1.2773	-0.0422

Table 6: Models with single RBPs motif count on 5' UTR (top) and 3' UTRs (bottom). Ordered by the root-mean-squared error on testing set.

or for statistical inference by coefficient estimate analysis.

To better understand the result, we sampled the best model (PUF2 motif count on 5' UTR) and set out to analyze wether the residuals, as the sample analogue of error term ε, meet the assumptions of linear regression. We based our analysis on the methods described by Vittinghoff et al. [13]. We trained the model on the entire training set and calculated its residuals by subtracting measured values from

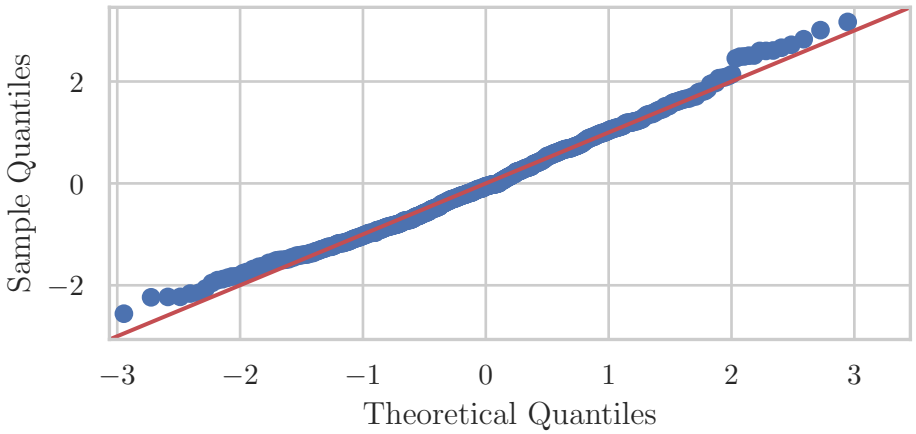

Figure 6: PUF2 5' UTR model residuals distribution versus standard normal distribution

fitted values.

First, we compared the distribution of residuals to a standard normal distribution (Figure 6). There we can see that it follows the distribution well, with only slight deviations at the ends. Therefore the assumption of a normal distribution is met.

We then plotted the model residuals to fitted values and predictors (Figure 7). None of the plots demonstrate significant divergence from mean zero i.e. the residuals seem to be distributed roughly equally on both sides of 0 on the vertical axis. Therefore the assumption of mean zero of residuals across observations i.e. linearity is met. When analyzing the vertical spread of residuals, we can see that the plot for fitted values shows a somewhat wider range in the middle, but this can be due to higher number of observations in that range and thus there could be some larger residuals only by chance. The plot for motif counts shows a narrower range where the motif count is 1. Also, the plot for 5' UTR displays a narrowing range from left to right. However, for both of these cases, we apply the same explanation we did for the plot of fitted values. Accordingly, we conclude that the vertical spread of residuals is similar across the ranges of fitted values and predictors and the assumption of homoscedasticity is met. Finally, we assume that the residuals are independent as we expect mRNA sequences and lengths to be independent between target genes.

As the model assumptions are well-met, it is unlikely that further transformations of data, non-linear or otherwise, are going to improve the linear relationships between predictors and outcome, and the fit of this model.

We then aggregated RBPs motif counts on both UTRs under a single model,

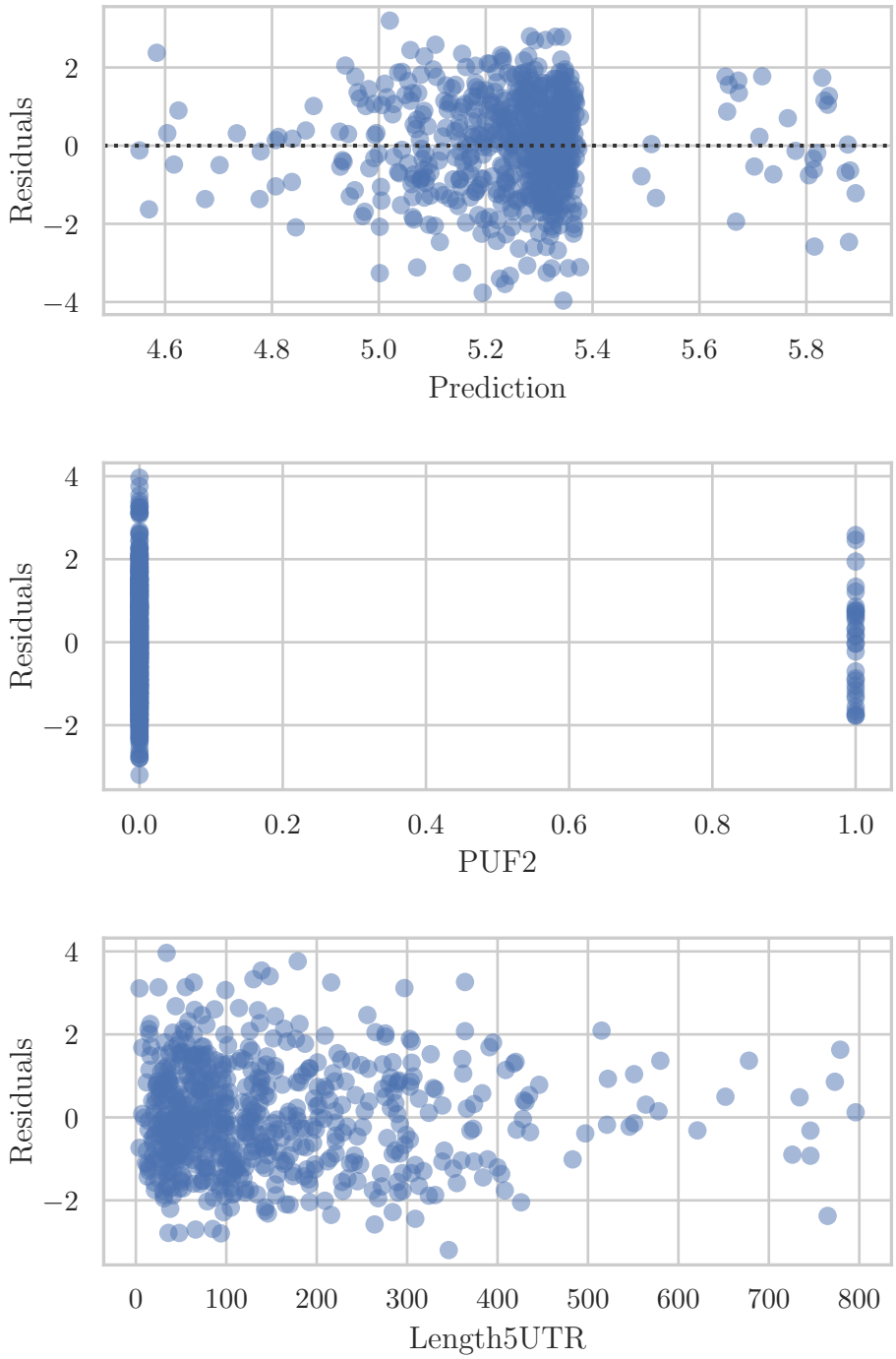

Figure 7: PUF2 5' UTR model residuals vs. fitted values and predictors.

	intercept	n_{RBP}^{5UTR}	n_{RBP}^{3UTR}	l_{5UTR}	l_{3UTR}	$mean_{RMSE_{CV}}$	$mean_{R^2_{CV}}$
PUF2	5.3971	0.5754	0.0119	-0.0010	-0.0001	1.2649	-0.0226
VTS1	5.3965	0.4166	0.1168	-0.0010	-0.0001	1.2655	-0.0232
PUF3	5.3992	0.2275	0.0091	-0.0011	-0.0001	1.2659	-0.0239
HRP1	5.3944	-0.2271	-0.0876	-0.0007	-0.0000	1.2663	-0.0242
PRP24	5.4096	-0.1065	-0.0112	-0.0007	-0.0001	1.2669	-0.0253
MPT5	5.4011	0.3493	0.0314	-0.0011	-0.0001	1.2673	-0.0262
EMG1	5.4019	0.0138	0.0718	-0.0010	-0.0005	1.2678	-0.0269
RRP6	5.4016	-0.0077	-0.0303	-0.0008	0.0003	1.2681	-0.0271
PAP1	5.4016	-0.0077	-0.0303	-0.0008	0.0003	1.2681	-0.0271
MSL5	5.4025	0.1087	-0.1824	-0.0009	0.0000	1.2689	-0.0284
NAB3	5.4072	-0.0213	-0.0002	-0.0006	-0.0001	1.2690	-0.0286
PUF4	5.4026	0.0552	0.0227	-0.0010	-0.0002	1.2691	-0.0286
NAB2	5.4048	-0.1240	-0.2314	-0.0008	0.0002	1.2699	-0.0297
NPL3	5.4036	-0.1469	0.0549	-0.0008	-0.0001	1.2702	-0.0306
MTR4	5.4037	-0.0203	-0.0185	-0.0008	0.0000	1.2713	-0.0323
DIS3	5.4045	-0.0064	-0.0090	-0.0009	-0.0001	1.2740	-0.0367
MSS116	5.4033	-0.0105	0.0082	-0.0008	-0.0002	1.2748	-0.0381
RPB2	5.3998	0.1572	-0.3610	-0.0009	0.0001	1.2751	-0.0387
RNA15	5.4042	-0.0230	0.0138	-0.0008	-0.0002	1.2752	-0.0386
RPO21	5.4064	0.0780	-0.0137	-0.0010	-0.0001	1.2767	-0.0407
DBP5	5.4030	-0.0345	0.0131	-0.0008	-0.0001	1.2782	-0.0435

Table 7: Models with single RBPs motif count on 5' UTR and 3' UTRs as separate independent variables in the same model. Ordered by the cross-validation mean of root-mean-squared error.

still separately for each RBP. This resulted in a total of four predictors: RBPs motif count on both UTRs and both UTRs length. This did not improve the results and none of the models were able to get a positive R^2 score (Table 7). The best model was achieved by the same RBP, but with worse scores than the best single UTR model.

We conclude that, based on this analysis, no RBPs motif count on mRNA untranslated regions accounts for variability in translation efficiency values.

4 Discussion

Understanding what controls variability of protein abundances is important if we were to influence it. We demonstrated that all of the variability in protein abundances can be explained by mRNA abundance, translation efficiency and protein turnover. Therefore, influencing protein levels comes down to influencing these three targets. To influence these, it is again important to first understand what controls them. We demonstrated that for 21 RNA binding proteins, the presence of their motifs on untranslated regions does not account for variability in translation efficiency.

To discuss further, RBPs motif matches on a target mRNA could be an unreliable proxy for actual RBP binding events on that mRNA, which we could then expect to be an actual regulator of translation. If that is the case, then the relationship between motif matches and actual RBP binding events could be too noisy to find any statistical effects of RBP binding. Many RBPs indeed lack canonical RNA-binding domains, which could further support this scenario [6, 15]. However, we could expect such proteins to be more generally applied and therefore not explain the differences in translation efficiency, unless they require specific higher structures. Indeed, Ray et al. report that vast majority of RBPs bind to target sequences in single-stranded RNA and none absolutely require a specific secondary structure, and additionally obtained dozens of significant associations between the presence of motif matches and specific regulatory outcomes [8]. This makes a strong case for motif matches being a reliable proxy for RBP binding events. In that scenario, we can conclude that for the 21 RBPs under analysis in this work, their binding to untranslated regions of target RNAs does not control the translation efficiency of these RNAs.

An additional caveat would be that RBPs often work in concert to regulate translation [7]. These dynamic interactions could not be captured by modelling RBP motif matches separately. Such interactions must be carefully modeled and were beyond the scope of this thesis.

In future analyses, further RBPs could be targeted, possibly using prior knowledge of their function in translational regulation for selection. A priori knowledge could help to select more relevant RBPs and correctly model the interaction between them. Ribosome densities, as a determinant of translation efficiency, could be targeted as an alternative outcome [16]. More complex models, such as deep neural networks, could be employed to find associations between RNA sequences and outcome translation efficiency or protein abundances.

References

[1] Anne Mathilde Davy, Helene Faustrup Kildegaard, and Mikael Rørdam Andersen. "Cell Factory Engineering". In: *Cell Systems* 4.3 (Mar. 2017), pp. 262–275. ISSN: 24054720. DOI: 10.1016/j.cels.2017.02.010. URL: http://dx.doi.org/10.1016/j.cels.2017.02.010.

[2] Niv Antonovsky et al. "Sugar Synthesis from CO2 in Escherichia coli". In: *Cell* 166.1 (June 2016), pp. 115–125. ISSN: 10974172. DOI: 10.1016/j.cell.2016.05.064.

[3] Petri Jaan Lahtvee et al. "Absolute Quantification of Protein and mRNA Abundances Demonstrate Variability in Gene-Specific Translation Efficiency in Yeast". In: *Cell Systems* 4.5 (2017), 495–504.e5. ISSN: 24054720. DOI: 10.1016/j.cels.2017.03.003.

[4] Bruce Alberts et al. *Molecular Biology of the Cell*. Ed. by John Wilson and Tim Hunt. Vol. 0. 0. Garland Science, Aug. 2017, p. 0. ISBN: 9781315735368. DOI: 10.1201/9781315735368. URL: https://www.taylorfrancis.com/books/9781317563754.

[5] Hila Gingold and Yitzhak Pilpel. "Determinants of translation efficiency and accuracy". In: *Molecular Systems Biology* 7 (2011), p. 481. ISSN: 17444292. DOI: 10.1038/msb.2011.14. URL: www.molecularsystemsbiology.com.

[6] Gideon Dreyfuss, V. Narry Kim, and Naoyuki Kataoka. "Messenger-RNA-binding proteins and the messages they carry". In: *Nature Reviews Molecular Cell Biology* 3.3 (2002), pp. 195–205. ISSN: 14710072. DOI: 10.1038/nrm760.

[7] J. J. David Ho et al. "A network of RNA-binding proteins controls translation efficiency to activate anaerobic metabolism". In: *Nature Communications* 11.1 (May 2020), p. 2677. ISSN: 2041-1723. DOI: 10.1038/s41467-020-16504-1. URL: https://doi.org/10.1038/s41467-020-16504-1%20http://www.nature.com/articles/s41467-020-16504-1.

[8] Debashish Ray et al. "A compendium of RNA-binding motifs for decoding gene regulation". In: *Nature* 499.7457 (2013), pp. 172–177. ISSN: 00280836. DOI: 10.1038/nature12311. URL: http://dx.doi.org/10.1038/nature12311.

[9] Girolamo Giudice et al. "ATtRACT-a database of RNA-binding proteins and associated motifs". In: *Database* 2016 (2016), p. 35. ISSN: 17580463. DOI: 10.1093/database/baw035. URL: http://www.highcharts.com.

[10] Plociam / CC BY-SA (http://creativecommons.org/licenses/by-sa/3.0/). *The structure of a mature eukaryotic mRNA. A fully processed mRNA includes the 5' cap, 5' UTR, coding region, 3' UTR, and poly(A) tail.* 2005. URL: https://upload.wikimedia.org/wikipedia/commons/d/d3/Mature_mRNA.png.

[11] Patrik D'Haeseleer. "What are DNA sequence motifs?" In: *Nature Biotechnology* 24.4 (Apr. 2006), pp. 423–425. ISSN: 10870156. DOI: 10.1038/nbt0406-423.

[12] Peter J.A. Cock et al. "Biopython: Freely available Python tools for computational molecular biology and bioinformatics". In: *Bioinformatics* 25.11 (2009), pp. 1422–1423. ISSN: 13674803. DOI: 10.1093/bioinformatics/btp163. URL: www.python.org.

[13] Eric Vittinghoff et al. *Regression Methods in Biostatistics*. Statistics for Biology and Health. Boston, MA: Springer US, 2012. ISBN: 978-1-4614-1352-3. DOI: 10.1007/978-1-4614-1353-0. URL: http://link.springer.com/10.1007/978-1-4614-1353-0.

[14] Gábor Csárdi et al. "Accounting for Experimental Noise Reveals That mRNA Levels, Amplified by Post-Transcriptional Processes, Largely Determine Steady-State Protein Levels in Yeast". In: *PLoS Genetics* 11.5 (2015). ISSN: 15537404. DOI: 10.1371/journal.pgen.1005206.

[15] Matthias W. Hentze et al. "A brave new world of RNA-binding proteins". In: *Nature Reviews Molecular Cell Biology* 19.5 (2018), pp. 327–341. ISSN: 14710080. DOI: 10.1038/nrm.2017.130. URL: http://dx.doi.org/10.1038/nrm.2017.130.

[16] Nicholas T. Ingolia. "Ribosome Footprint Profiling of Translation throughout the Genome". In: *Cell* 165.1 (Mar. 2016), pp. 22–33. ISSN: 10974172. DOI: 10.1016/j.cell.2016.02.066.

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, **Mark Kerner**,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose

   Predicting Specific Protein Abundances and Translation Efficiency in Yeast,

   mille juhendaja(d) on Leopold Parts,

   reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.

2. Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi DSpace kaudu Creative Commonsi litsentsiga CC BY NC ND 3.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni.

3. Olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile.

4. Kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Mark Kerner
10.08.2020