REPUBLIC OF TURKEY
ACIBADEM MEHMET ALI AYDINLAR UNIVERSITY
INSTITUTE OF HEALTH SCIENCES

A TOOL FOR PREDICTION OF PROTEIN EXPRESSION FROM GENETIC DATA

SILA GERLEVİK

MASTER THESIS

DEPARTMENT of BIOSTATISTICS and BIOINFORMATICS

SUPERVISOR:
Prof. Dr. Osman Uğur Sezerman

ISTANBUL – 2021

REPUBLIC OF TURKEY
ACIBADEM MEHMET ALI AYDINLAR UNIVERSITY
INSTITUTE OF HEALTH SCIENCES

A TOOL FOR PREDICTION OF PROTEIN EXPRESSION FROM GENETIC DATA

SILA GERLEVİK

MASTER THESIS

DEPARTMENT of BIOSTATISTICS and BIOINFORMATICS

SUPERVISOR:
Prof. Dr. Osman Uğur Sezerman

ISTANBUL – 2021

DECLARATION

I declare that; this thesis study is solely my original work. I had no unethical behavior at any stage, from planning to writing; I obtained all the information in this thesis following academic and ethical rules, I have shown all the resources I have used in the bibliography. There is no violation of any patents and copyrights.

…/…/2021

Sıla GERLEVİK

ACKNOWLEDGMENTS

I would like to present my sincere gratitude to Prof. Dr Osman Uğur Sezerman, who is most tolerant advisor that a student can have, for his supervision, patience, support, and trust. I am thankful for the opportunity to work in his lab, and for the point of views he has given me.

I am also very thankful to Prof. Dr. Koray Özduman for the opportunity to work with Acıbadem University Brain Tumor Research Group to expand my knowledge and my passion.

Moreover, I would like to thank to jury members of my thesis, which are, Assoc. Prof. Emel Timuçin, and Assist. Prof. Burcu Bakır Güngör, for their worthy contributions

I am thankful to all my friends with whom we shared the same office for almost 3 years for their mental and guide support and the deepest discussions that shape my personality and mentality. Our comprehensive journal clubs, and scientific boosting have become valuable for me during this period.

I am deeply grateful to my family for her trust in me, friendship, tolerance, patience, support, deep interest, time, and the point of views that they have been earning me. Finally, I want to express the most intense thanks to my dearie, Umut Gerlevik, for always being with me, his effort, support, and especially for his patience and understanding to every idea that pass through my mind.

TABLE OF CONTENTS

LIST OF ABBREVATIONS AND SYMBOLS

CAI	Codon Adaptation Index
CBI	Codon Bias Index
Fop	Frequency of optimal codon
Nc	Number of effective codons
RFM	Ribosome Flow Model
tAI	tRNA Adaptation Index

LIST OF FIGURES

LIST OF TABLES

SUMMARY

Proteins are responsible to function cell properly by playing important roles in physiological process and protein abundance directly affects the enzymatic reactions, protein-protein interaction, pathways. Protein expression is a complex mechanism that is finely regulated and controlled in every step from transcription to translation through all organisms. Absolute protein abundance within cell depends not only mRNA availability but also on tRNA availability, translation rate, co-translation folding, post-translational modifications, and degradation rate. Absolute protein levels within cell are very dynamic and influenced by environmental and biological stimuli and underlying mechanisms of change in protein levels because of these stimuli is not still understood. Current technological developments lead to sequencing-based measure of mRNA levels to perform easily; however, measuring protein abundance are still both complicated and expensive. Protein level within cell is influenced by translation kinetics which depend on features of codon. In this study, we developed machine learning approaches to predict protein abundance level from given coding sequence based on codon and tRNA features by using *Saccharomyces cerevisiae* data. We calculated codon features that enables the frequency information from codon usage and calculated tRNA availability features. We used newly developed Ribosome Flow Model (RFM) to capture information about translation rate depending on codon order of gene. By using these features and codon composition, we developed random forest model, gradient boosting machine, and support vector regression approaches and select the random forest model had better performance among other (R^2=0.56). It gives correlation value between prediction results and experimental yield was 0.79.

Keywords: Codon Usage, Machine Learning Models, Protein Abundance, Translation Rate,

ÖZET

Genetik Verilerden Protein Yoğunluğunu Tahmin Etmek için Bir Araç

Proteinler, fizyolojik süreçte önemli roller oynayarak hücrenin düzgün işleyişinden sorumludur ve protein bolluğu, enzimatik reaksiyonları, protein-protein etkileşimini, yolakları doğrudan etkiler. Protein ekspresyonu, bakterilerden insanlara kadar tüm organizmalarda transkripsiyondan translasyona kadar her adımda hassas bir şekilde düzenlenen ve kontrol edilen karmaşık bir mekanizmadır. Hücre içindeki mutlak protein bolluğu sadece mRNA mevcudiyetine değil, aynı zamanda tRNA mevcudiyetine, çeviri hızına, birlikte çeviri katlanmasına, çeviri sonrası modifikasyonlara ve bozulma hızına da bağlıdır. Hücre içindeki mutlak protein seviyeleri çok dinamiktir ve çevresel ve biyolojik uyaranlardan etkilenir ve bu uyaranlar nedeniyle protein seviyelerindeki değişimin altında yatan mekanizmalar hala anlaşılamamıştır. Mevcut teknolojik gelişmeler, mRNA düzeylerinin dizileme tabanlı ölçümünün kolayca gerçekleştirilmesine yol açmaktadır; bununla birlikte, protein bolluğunu ölçmek için protokoller hala hem karmaşık hem de pahalıdır, hatta kütle spektrometrisinin protein havuzunu yakalamada sınırlamaları vardır. Protein bolluğu seviyesi, kodonun özelliklerine bağlı olan translasyon kinetiğinden etkilenir. Bu çalışmada, Saccharomyces cerevisiae verilerini kullanarak kodon ve tRNA özelliklerine dayalı olarak verilen kodlama dizisinden protein bolluk seviyesini tahmin etmek için makine öğrenme yaklaşımları geliştirdik. Kodon kullanımından frekans bilgisini sağlayan kodon özelliklerini hesapladık; ve tRNA uygunluğu indeksini hesapladık. Genin kodon sırasına bağlı olarak çeviri hızı hakkında bilgi toplamak için yeni geliştirilen Ribozom Akış Modeli'ni (RFM) kullandık. Bu özellikleri ve kodon kompozisyonunu kullanarak rastgele orman modeli, gradyan artırma makinesi ve destek vektör regresyon yaklaşımları geliştirdik ve son olarak performansı iyileştirmek için topluluk yöntemleri uyguladık ve eğitim verileri üzerindeki R^2 değerini yaklaşık %6 artırdık ve tahmin sonuçları ve deneysel sonuçlar arasındaki Pearson korelasyon değerini 0.75 olarak bulduk.

Anahtar Sözcükler: Kodon Kullanımı, Makine Öğrenme Modelleri, Protein Bolluğu,

1. BACKGROUND AND AIM OF THE STUDY

Proteins are accountable for most of biological processes as protein abundance levels within cell directly affect enzymatic reactions, pathway mechanisms, and protein-protein interaction. Absolute protein abundance depends on balance between various process such as transcription and mRNA stability, tRNA availability, ribosome speed, co-translational protein folding, post-translational modifications, and degradation of expressed protein (1). By Mehdi and et al., protein abundance was described the amount of copies of a protein molecule within cell because of the dynamic balance among previously mentioned processes (2).

Protein levels in cells is very dynamic and affected by environmental and biological process; thus, reliable quantification of absolute protein levels is necessary to understand cellular function and its effect on cellular phenotypes completely (3). Due to complexity of the rules and control of proteome, quantification of protein level change in cell resulted in biological stimuli does not completely understood up to date (4). This lack of information about change in protein levels can cause several problems like low production yield in recombinant protein production and prevention of diagnosis in disease caused by synonymous mutation, which is mutation that do not change the amino acid order in protein. Understanding the determinants and regulations in protein expression mechanisms of the organism can provide the detection of protein concentration and allow the efficient recombinant protein production within given host and identification of biological mechanisms behind phenotypic variation and disease. Thus, predictive model for protein abundance can lead benefit for the heterologous protein expression, understanding biological mechanisms, and diagnosis for diseases (5,6).

Although currently, mass spectrometry technologies and proteomic analysis enables the detection of lots of proteins in different organisms, still great portion of

proteins cannot be detected because of limitations of physiochemical and experimental conditions and large cost (7). The protein abundance can be primarily determined by combining expression and translation factors like mRNA abundance, tRNA availability, translation rate, protein turnover rate, codon usage bias (CUB) (8). Since solely mRNA abundance is not sufficient to detect protein abundance, various features of mRNA have been proposed to understand effects of codon features on protein expression, which is most likely related to codon usage of organisms like codon adaptation index, frequent of optimal codon, number of effective codon, G+C content, and tRNA adaptation index (9). Although global protein abundance analysis is significant for system biology and biotechnology, it is a complicated work to use bulk data and information for understanding the expression capabilities (4).

With the help of these information, the purpose of this thesis is a development an algorithm to predict protein abundance from the given coding sequence for the host organisms by using the ribosome flow model outputs, several explanatory indices of codon usage and tRNA based indices.

2. INTRODUCTION

2.1 Biological Background

2.1.1 Protein expression and genetic code

Efficient process of protein expression is required to function cell properly. Expression of genes and translation process of genetic information into proteins by ribosomes is the main process in living cells (10,11). Transition of genetic information to protein by mRNA transcribed from DNA is described as translation process. mRNA template includes particular combination of 61 trinucleotide codons that translate into 20 amino acids (12). The inherent redundancy of genetic code is referred as that all amino acids, (except Tryptophan and Methionine) can be encoded up to six different codons, which is termed as synonymous codons (13). During translation elongation, decoding by ribosome, codons interacted with the anticodon of its matching (cognate) trinucleotide sequences in transfer RNA (tRNA), to insert the amino acids carried by the tRNA. Translation elongation rate is impacted by the synonymous codons, which correspond to same amino acids but diverge in the relative usage in genome, in abundance of their decoding tRNAs, and in necessity of some codons for wobble interaction. Wobble interaction is described as non-Watson-Crick base pairing of the third base in codon and first base in tRNA anticodon (14,15). Among lots of organisms synonymous codons are used with different frequencies, which is called codon usage bias (CUB) (16,17). Some codons paired by abundant tRNA are efficiently translated, called optimal codons, and are found in highly expressed genes in yeast and *Escherichia coli*, also their codon bias is higher (16,18). The influence of CUB on translation efficiency is emphasized by correlating the codon usage and translation efficiency in genome-wide and several studies in which genes' codon replaced with optimal or suboptimal codons influences protein expression as expected (19–21).

Frequently used codons as described by codon usage bias patterns are classified as “fast” codons whereas rare codons are classified as “slow” codons (22).

Although studies for more than 30 years, underlying mechanisms of codon-mediated effects on protein expression are not fully solved. The main ideas on translation elongation rate at different codons have been suggested to depend on tRNA concentration and slower translation elongation rate, because of the gathering of small influences of single codons paired by rare tRNA, decrease the translation efficiency (23,24). This hypothesis has confirmed with several experimental studies for yeast and microorganisms (25–27), which indicates that highly expressed genes utilize a subset of optimal codons based on their particular important iso-acceptor tRNA levels. However, this hypothesis has not confirmed for unicellular organisms and higher eukaryotes; such as humans in which their codon usage was described only by genomic GC content, and fly or worm in which selection determine their codon usage partly (28).

2.1.2 Translation elongation rate determinants

Translation elongation is a complex progress, in which is required mRNAs, tRNAs, ribosomes, and many trans-acting factors and regulators to work together functionally. Ribosomal subunits (40S and 60S) are assembled at mRNA start codon with binding of initiator methionyl-tRNA at the P site of 80s ribosome complex by facilitating with initiation factors (eIFs) proteins. During elongation process, 80S complex moves through mRNA template by three nucleotides at a time, enlarging the polypeptide, in coordination with aminoacyl-tRNAs (aa-tRNA), GTP carrying correct (cognate) anticodon for codon in mRNA, and several elongation factor (eEFs) (called as ternary complex) in the ribosomal A site (12,29). Three steps in elongation process are tRNA selection, peptidyl transfer, and translocation. In tRNA selection, ternary complex firstly binds to ribosome and then codon recognition and elongation factor disassociate by GTPase activation caused to GTP hydrolysis and then in the P site

amino acid portion of aa-tRNA get closer to the peptidyl tRNA where peptidyl transfer to take place. After then, the ribosome translocate forward one codon, when catalysing of elongation factor G and GTP hydrolysis, so tRNA also move towards E site first and then move towards P sites (12). This elongation cycle continues as next codon in the ribosomal A site waits for the new cognate, near-cognate, or non-cognate tRNA arrival, which termed according to their none, single or multiple base mismatches in given codon respectively (30). Studies show that translation elongation rate is significantly related to tRNA selection in A site of ribosome where cognate anticodon recognition occur faster and near-cognate anticodon is rapidly rejected (30–32).

It is appeared that codon usage biases relate with tRNA copy numbers of organism itself. Codons, recognized slower when their corresponding tRNAs are found at low concentration within cells, termed as rare codons and they can affect the translation elongation speed (33). Some studies shows that codon usage frequencies regulate ribosome traffic on mRNA and rare codons leads the ribosome to pause and to accumulate the nascent polypeptide during translation. In addition to this, previous studies in wide range from fungi to animals indicates that rare codons decrease the translation rate whereas preferred codons rise the speed of translation elongation rate (34–36).

2.1.3 Co-translational folding

Co-translational folding is the process that nascent protein emerging from ribosome starts acquired its tertiary structure by folding while it is synthesized by the ribosome. During translation process, folding process takes place simultaneously so that codon translation timing can influence the protein product. Thus, studies showed that slowing of translation lead to increase co-translational folding since slower codon translation rates allow a polypeptide to take more time for folding in translation process (22,37–39). Thus, common idea is that slowing of translation at the certain locations incline to rise the probability of domain-wise folding in protein whereas

speeding of translation incline to decrease it, when competing processes like early termination of translation or missed amino acid do not occur (40–42).

Codon usage enables the translation speed to variate during mRNA translation since codon bias often change throughout the gene coding region; thus, codon usage- and order-dependent translation process would influence the time that allows different co-translation folding events (33). Studies of heterologous overexpression of proteins in *E. coli* cells indicate that using rare codons instead of common codons can cause small decreases of protein activity and solubility (38). Significantly, the effect of codon usage on co-translational folding was shown in study which mammalian gamma-B crytallin is expressed in *E.* coli by tracking of fluorescence intensity changes (43). It has been proposed that sequenced-based manipulation can be increase the protein folding in heterologous expression. The effect of local differences in translation rates on co-translational folding can be studied by using two different principles like computational researches to correlate between codon combination of mRNA and structural features of the proteins or like biochemical studies to detect influence of synonymous mutations on protein function (12).

2.2 Computational background

2.2.1 History of quantification of protein expression

Since mRNA measurement is performed easily, first analyses of protein abundance deduced from global mRNA quantification by microarray technologies. However, due to effects of post-translational, translational and degradation mechanisms, for more reliable results, protein abundance level require the direct measurement like mass spectrometry-based shotgun proteomics experiment, two dimensional electrophoresis, and high-throughput cell imaging (2,3). Although mass spectrometry technologies and proteomic analysis enables the detection of lots of

proteins, still lots of proteins cannot be detected because of limitations of physiochemical and experimental conditions and large cost (44). Currently, ribosome profiling method enables the relative time of ribosome at each codon in gene before moving next, which theoretically strong method to detect influence of a single codon on translation elongation rate; however, there are several technical obstacles to manage this promise (18).

Due to technical obstacles, limited number of detection and large cost of experimentally detection of protein abundance levels, scientists have tried mathematical methods and statistical approaches to predict protein abundance level with given codon and tRNA information. According to the correlation between absolute protein levels and CUB, by using machine learning methods, prediction tools for non-model species or experimentally undetected proteins can be doable with metrics of codon usage bias (45,46). Since codon usage and tRNA pool are highly organism specific, there would be limitations to produce protein if the RNA utilizes codon with low concentrate tRNA; thus, expression levels and features of codon encoding employed in training of model must be calculated for the target organisms (46). To capture information about codon selection, features were calculated such as Codon Adaptation Index (CAI), Codon Bias Index (CBI), effective codon number (Nc), frequency of optimal codon (Fop) and tRNA Adaptation Index (tAI). CAI measure the synonymous codon usage bias of gene and it resembles to synonymous codon frequency of a reference set of gene of organism. CBI is a measurement to detect gene utilizing a subset of optimal codons. Nc is a measure for overall codon bias and Fop is the ratio of optimal codons to synonymous codons. tAI is a measurement of adaptation of coding sequence based on tRNA copy number of an organism (21,47–50).

Currently, a number of methods exists to use for prediction of protein abundance level, in which most of them use protein abundance within the cell by using mRNA expression and codon features as key predictive features (2). Tuller et al., combine mRNA level, tAI, and evolutionary rate of the transcripts to predict protein abundance

with linear regression (21). Futcher et al., shows good relation of protein abundance with mRNA abundance and codon bias after applying log-transform (51), Huang et al, estimated translation rate according to the sequence features of mRNA and functional features of mRNA by applying maximum relevance and minimum redundancy method and by selecting feature they improve the prediction model and they achieved to estimate the translation rate as low or high (52). Dos Reis et and colleagues developed a statistical model to measure translation selection in any given genome and they optimized the Wright's Nc (28). These models did not cope with data with missing values so Mehdi et al., introduced Bayesian network model combining transcriptomic and proteomic data by using tRNA adaptation index, mRNA level, protein interaction with mRNA, mRNA folding energy and half-life, to use for condition-specific data (2). Welch et al., constructed prediction model with partial least square (PLS) by using codon bias calculated from codon frequencies (53). Supek and Smuc built a prediction model for expression with free energy of protein folding and codon bias of codon frequencies with support vector regression (SVR) (54). Fernandes and Vinga built SVR and PLS model at the same data used by Welch et al., and Supek and Smuc to increase performance of prediction of the protein level by adding extra features and ensemble averaging (46).

2.2.2 Machine learning Models

Random forest (random decision forest) is a machine learning method which handles by building a multitude decision tree during running, in which for classification output is the class selective by most trees while for regression, the mean or average prediction for individual trees is the output. Random forest can be used for both classification and regression (55). Gradient boosting is a method for regression and classification as well. However, it develops a model in a stage-wise fashion to create model with ensemble of weak decision trees, where it constructs the model and generalizes them by optimizing of a differentiable loss function. It is often outperforms random forest. (56). Support vector regression works with approximation function, which is described by the set of weights for the input variables, that deviates the most

ε from the training samples. It uses a form of regularization to generalize the large dimensional input features by maintaining the small weights. Because all points will not be inside ε tolerance band, slack variables that refer to distance from each point to the ε band around the approximating function need to be considered. This function is calculated by solving an optimization problem to describe the cost of a point that is outside the ε band. Training points generate the approximating function is defined as support vectors. Support vector regression uses kernel function which enables to construct non-linear functions by transforming data into higher dimensional feature space. There are several kernel functions, and in this study, radial basis function was used. Kernel functions are used to do calculations in any d-dimensional space where d is higher than 1; Radial Basis Function kernel uses exponent which gives the polynomial equation to infinite power, which would give a curve fitting any complex dataset (46).

3. MATERIALS AND METHODS

3.1 Dataset

The protein abundance values were acquired from a unified dataset of *Saccharomyces cerevisiae* from YeastMine dataset which were populated by *Saccharomyces* Genome Database (57). This dataset was filtered to have protein abundance values that were for about 4000 proteins in which includes only genes with verified ORFs to make sure the encoding codes to start with a start codon and a stop with stop codon as well as divided by three. This unified protein abundance dataset was created by normalizing and scaling all 21 yeast proteome datasets to the most intuitive protein abundance unit as molecules per cell (3). They extract raw protein abundance value from 21 global quantitative studies of the yeast proteome, which some of them recorded in arbitrary units (a.u.) and compared values from each study with each other. They normalized the protein abundance values and converted all measurements of protein abundance into molecules per cell. They curated data with gene systematic name, proteins median abundance.

3.2. Features

Features was calculated to construct prediction model for protein abundance by using genetic data of *S. cerevisiae*. Features can be classified in three main group like base composition of coding sequence, codon usage indices, and tRNA-based indices. Here we explained all indices in detail and as shown in Table 1 all indices were summarized.

3.2.1. Composition of coding sequence

These features are organisms independent and rely solely on nucleotides in the given sequences and gives information about nucleotide composition of coding sequence. G+C content of gene (GC) was calculated by ratio of the G and C bases towards coding sequence (58). G+C content at third position of synonymous codon (GC3) is a fraction of codons, which are synonymous at the third codon position, which have either guanine or cytosine at the third position. Base composition at silent site (A3s, T3s, G3s, C3s) was four separate features that described the found of each base at synonymous third codon position separately. Even though G3s and C3s related with GC3s content, this index does not directly comparable. It describes the found of each base at synonymous third codon position whereas in the GC3s, each synonymous amino acid has at least one synonym with G or C in the third position. Therefore, A3s is the frequency that codons have an A at their synonymous third position, relative to the amino acids that could have with A in the synonymous third codon position.

3.2.2 Codon usage indices

Codon usage is different and multivariate for organisms, and the frequency of each codon within a coding sequence has some part of explanation of selection of codons, spreading over 61 separate features. To extract this information, multiple codon indices were calculated through the years, attempting to summarize, simplify, and explain the bias within codon usage of gene. These indices are usually dependent to organisms, requiring the prior knowledge about the preferred codons of an organism to determine the main trends in variation of the data. By using the codon usage frequency of organism, codon indices, which can summarize the codon selection information of target gene, can be calculated. Here we describe some codon indices which can be utilized to analyze codon sequence.

3.2.2.1 Effective Number Codons

Effective number of codons (Nc) is a simple measurement that quantifies deviation of codon usage of a genes from random usage of synonymous codons. Nc can be easily quantified by using only codon usage table and it is not dependent of gene length and amino acid composition (28,50). Its minimum value can be 20, while only per amino acid used solely one codon, and its maximum value 61, while all codons are used equally; thus, Nc is a intuitively meaningful measurement for the codon preferences of a gene (50).

3.2.2.2 Codon adaptation Index

Codon adaption index (CAI) (47) is most widespread method to analyse the codon usage bias. CAI measures of relative adaptiveness of the codon usage of a gene towards the codon usage of highly expressed genes of that organism. CAI measure the deviation according to reference set of genes unlike the effective number of codons (Nc) that measure the deviation of codon selection from uniform bias.

CAI is described as geometric mean of the relative adaptiveness score (ω) of each codon, which is the monitored frequency of each codon to the usage ratio of the mostly used codon for the same amino acid, over the length of the gene sequence.

3.3.2.3 Frequency of optimal codons

Frequency of optimal codons (Fop) is a simple index that measure of species-specific bias towards translational optimal codons in a species. It was defined as ratio

between the frequency of optimal codons and the total number of synonymous codons (27).

3.2.2.4 Codon bias index

Codon bias index (CBI) (26) is directional the codon bias, which measures the extent to which a gene uses a preferred codons defined for an organism. In a gene with extreme codon bias meaning that only the preferred codons are used for all triplets in gene, CBI will be valued as 1 whereas in a gene with random codon usage, CBI will be valued as 0. Also, CBI can be a negative value, which indicates that the number of optimal codons is less than in a random selection. CBI valued importantly lower than zero indicates bias to rare non-preferred codons.

3.2.3 tRNA based indices

3.2.3.1 tRNA adaptation index

tRNA adaptation index (tAI) stand for the adaptation of coding sequence with respect to tRNA copy number of an organism. tAI provides to extract the influence of tRNA abundance and availability along with mRNA sequence on the translation rate (59). tRNA copy number across some genomes is highly associated with tRNA amount within the cell, and tRNA abundance is significantly assumed as a driving force on translation selection so that measurement of tRNA usage of gene would enable indirect way for detection of translational selection (28,60). tRNA adaptation index is described as geometric mean of relative adaptiveness of codon to gene's genomic tRNA pool.

3.2.3.3 Ribosome flow model

Ribosome flow model (RFM) is a probabilistic model for the translation elongation process. It is like tAI because of their codon adaptation to the tRNA pool. However, unlike tAI, RFM is sensitive to codon order and ribosome jamming effect. Because of the stochastic nature of translation, RFM aim to get the influence of the order and composition of the codons, and size of ribosomes on translation rate (61).

RFM has two parameters, which are the initiation rate λ and the number of codons C that is the "size" of the ribosome. mRNA molecules are splitted like coarse-grained into sites of C codons, which is suggested as 25 codons size by authors. It is possible the estimate translation rate of single codons with respect to tAI or similar metrices; thus, RFM uses a relative adaptiveness value to calculate the probability of tRNA that will be paired to its codon.

Table 1. Summary of features that were used in this study.

Features	Definition
G+C content	Ratio of G an C bases in a gene
G+C content at 3rd synonymous position (GC3s)	Fraction of codons, which are synonymous at the third codon position, have either guanine or cytosine at the third position
Base composition at silent site (A3s, T3s, G3s, C3s)	Four separate features that described the found of each base at synonymous third codon position separately
Codon Adaptation Index (CAI)	Geometric mean of the ratio of the observed frequency of codon to the frequency of its most abundant synonymous codon
Codon Bias Index (CBI)	Frequency of optimal codons to random usage of synonymous codons
Frequency of optimal codons (Fop)	Ratio of optimal codons to synonymous codons that are predefined for an organism.
Number of Effective codon (Nc)	Deviation of the codon selection of a gene from the random usage of synonymous codons
Lenght of silent size	Frequency of synonymous codon
tRNA Adaptation Index (tAI)	The amount of adaptation of a gene to its genomic tRNA pool
tAI bottleneck value	The minimum value of tAI calculated regionally within gene
Ribosome Flow Model output (Translation rate)	Translation rate is calculated by considering the affinity between tRNA species and codons, the effect of codon order, and composition on translation rates, and ribosome jamming

3.2 Implementations of features

To evaluate whether machine learning can predict absolute protein abundance level or capture any underlying pattern of translation process by using codon usage indices, set of codon usage features were calculated. Codon usage metrices that are calculated individually for given ORF for each protein were used. To construct the model, aforementioned indices were calculated by using various tools such as CodonW (62) for codon indices and raw features of gene, stAIcalc (49) for tRNA based indices, and RFMapp (61) for translation rate from RFM output. All analysis was done with R 4.0.0 via RStudio.

3.3 Regression model training and prediction

Different machine learning models and different machine learning libraries were used to construct regression model such as support vector regression in “e1071” (63), random forest , gradient boost machine in “gbm” (64). Dataset was randomly split into two subsets for training as 75% of data and testing for 25% of data. To select the hyperparameters, randomized approach in which each parameter was sampled from a subset of possible values and values were selected according to models with best evaluation metrics while applying to test data. Repeated cross-validation was applied as 5 repeats for 10 folds. Models were built with “Caret” library on the same training data, with the same resampling parameters. Then model with best performance among others were chosen to make prediction on test data. All machine learning models were built by R 4.0.0 via RStudio.

4. RESULTS

4.1 Predictive results

5 repeated 10-cross validation results for different models shown in Table 2. RMSE values of regression models were close each other, where random forest (RF) had lower value as 0.0260, followed by radial basis function-based support vector regression (RBF-SVR) having 0.0269, and gradient boosting machine (GBM) as 0.028. Random forest had higher R^2 value as 0.517, followed by RBF-SVR with 0.463, and GBM with 0.454.

Table 2. 5 repeated 10-fold cross validation RMSE and Rsquared values on training data.

	RMSE	Rsquared
GBM	0.0285	0.454
RBF-SVR	0.0269	0.463
RF	0.0260	0.517

Test data results of models were shown in Table 3. RMSE values were 0.0291, 0.0314, and 0.0342 for RF, GBM and RBF-SVR respectively, and Rsquared values were 0.564, 0.440, and 0.361 for RF, GBM, and RBF-SVR respectively. Results of RF model were more reliable on test data rather than the GBM, where lowest RMSE and highest Rsquared. RBF-SVR was worse model on test data unlike on the training data where GBM had lowest performance. On test data GBM showed better performance than RBF-SVR.

Table 3. RMSE and R^2 values of models on test data.

	RMSE	Rsquared
GBM	0.0314	0.440
RBF-SVR	0.0342	0.361
RF	0.0291	0.564

According to the performances on both training and test data, random forest showed better performance than other models so that prediction was performed with random forest model and correlation between predicted results and experimental protein yields was shown in Figure 1. Log transformation was performed on both predicted and experimental protein abundance values and Pearson correlation coefficient was calculated as 0.79 ($p < 0.0001$).

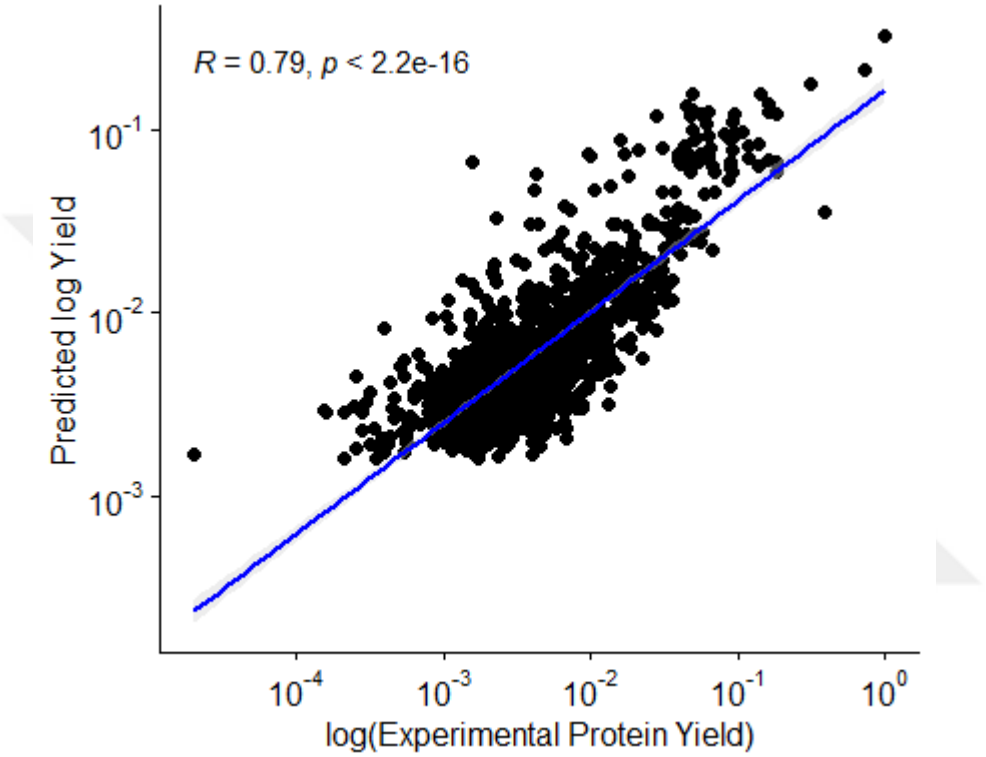

Figure 1. Scatter plot of experimental protein yield versus predicted protein abundance levels with Pearson correlation value and significant p value ($p < 0.0001$).

4.2 Factors affecting protein expression and abundance

Each feature was correlated against protein abundance value to find the feature that correlated with protein expression level as shown in Figure 2. Spearman correlation was used because it is not use assumption about the distributions of features. In figures and inside the text, m_1, m_3, m_6, m_9, m_12, m_15, m_25 referred to tAI bottleneck metrices, where numbers stand for sliding window size while

calculation for geometric means of relative adaptiveness of tRNA through the gene (*m* = {1, 3, 6, 9, 12, 15, 25}) and L_sym referred to frequency of synonymous codons; moreover A3s, T3s, C3s, G3s referred to base compositions of each nucleotide in third synonymous codon position of a gene while GC3s referred to fraction of gene, which had synonymous codon, had either G or C at third synonymous codon position.In Figure 2, MedianPA referred to median protein abundance of genes. tAI value had highest correlation value with the protein abundance along with the RFM value (TranslationRate) as 0.661 and 0.656, respectively. Codon indices had very low correlation with protein expression levels, where CAI had 0.067, CBI had 0.193, Fop had 0.115, and Nc had negative correlation value as -0.304. GC value had 0.267 correlation value while A3s and L_sym (frequency of synonymous codons) had negative correlation value as -0.221 and -0.206 respectively with protein abundance. According to the results features about length and composition about coding sequence (L_sym, A3s, G3s, T3s) were negatively correlated with protein abundance levels.

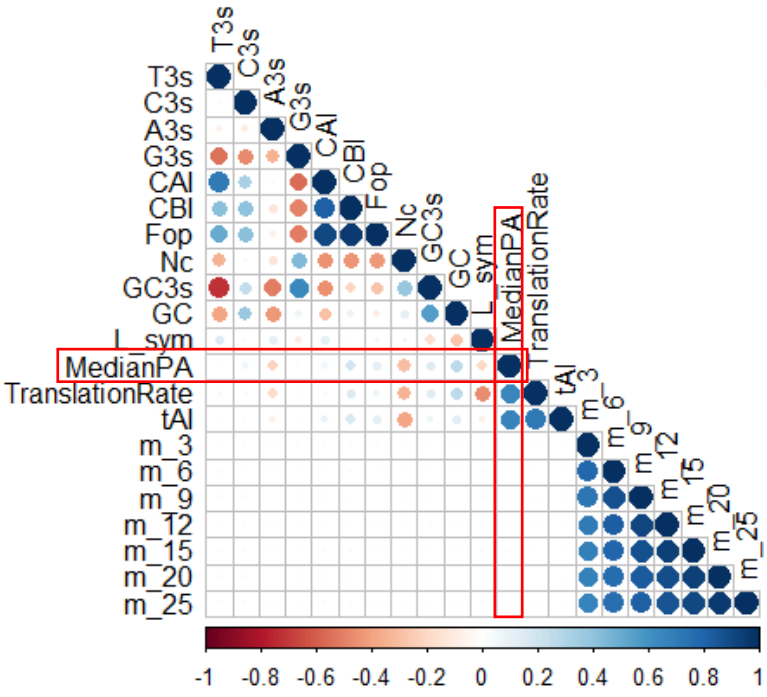

Figure 2. Correlation between protein abundance and features in training data

As shown in Figure 2, protein abundance columns were selected, tAI and RFM output (translation rate) were more correlated than codon usage indices. Feature

importance based on the machine learning models were also calculated separately. As shown in Figure 3, based on random forest model L_sym (length of silent sizes), tAI, and RFM output (TranslationRate) were relatively most important ones, followed by CAI, GC3s, T3s, m_25, m_6, m_3, Fop, m_9 and CBI and G3s, respectively.

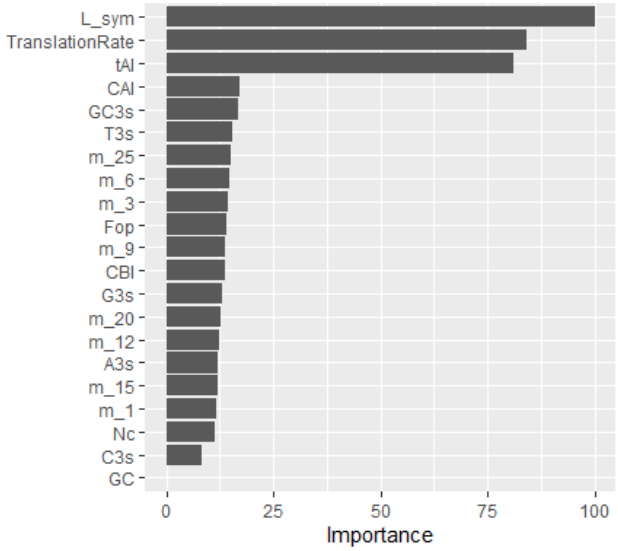

Figure 3. Feature importance based on random forest on training data

Also, as shown in Figure 4, according to the RBF-SVR model, tAI and RFM value were relatively most important ones, followed by Nc, CAI, CBI, C3s, and Fop respectively.

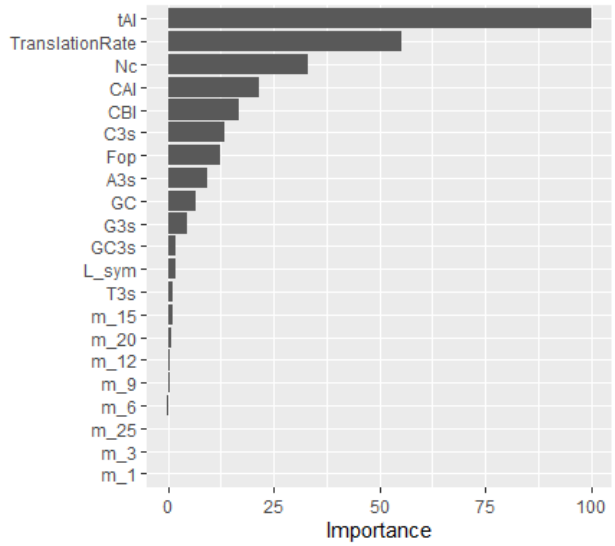

Figure 4. Feature importance according to RBF-SVR model

As shown in Figure 5, based on GBM model, again tAI was the relatively most important feature but it was followed by synonymous length of gene (L_sym) rather than RFM value unlike the other models, and it was followed by GC, RFM value, CAI, and tAI bottleneck value with sliding window size as 9. Important features in other models like CBI, Fop, Nc, and C3s had relatively very low importance in GBM model.

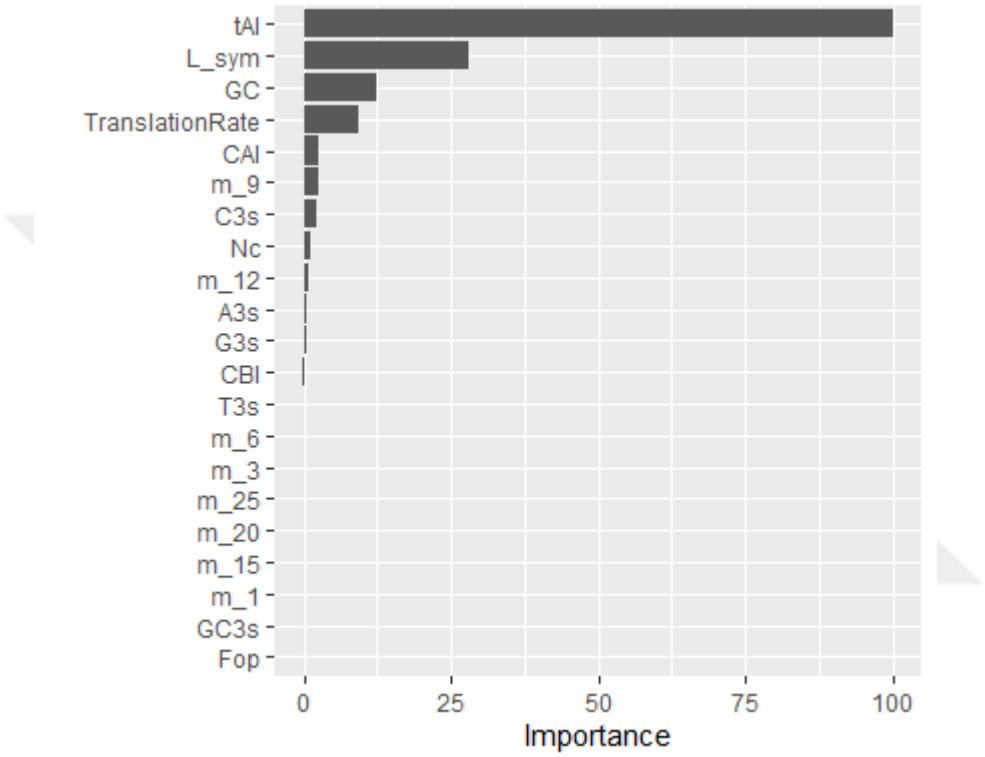

Figure 5. Feature importance based on gradient boosting machine model.

After ensemble averaging, tAI and RFM value were relatively most important ones, followed by L_sym, Nc, CAI, GC, C3s, CBI, and Fop respectively. According to these models, tAI, L_sym, and TranslationRate (RFM output) were clearly showed as important features.

5. DISCUSSION AND CONCLUSION

In this study, we developed a machine learning approaches to predict the protein abundance level by using codon usage features, tRNA related features, and RFM value. We tried three different models by using random forest, gradient boosting, and support vector regression on the same data. Features were calculated with CodonW programme and with RFMapp for translation rate feature. Several studies in literature tried to improve prediction of protein abundance for different organisms with different features. Most studies used mRNA expression level as main predictive feature for protein abundance determination within the cell. Tuller, et al constructed a predictor for *S. cerevisiae* with mRNA levels data, tAI, and the evolutionary rate; this predictor attains a correlation value as 0.76 with experimental data by using linear algorithm had 0.55 prediction accuracy (21). Our prediction results of protein abundance correlated with experimental protein yield with 0.79 value by using random forest regression. Fernandes and Vinga tried to improve machine learning performance on two different data and method (one from Welch et al (53) and other from Supek and Smuc (54)). They improved the models predictive ability by adding two different input features and applying ensemble averaging to the SVR or Partial Least Square model (46). We also tried the greedy ensemble on three models (which was not recorded in this study); however, performance of model did not improve significantly so that we decided to use random forest model had lowest RMSE and highest R^2 values. To cope up with the missing data in biological data, Mehdi and his colleagues used Bayesian network model relied on mRNA level, mRNA-protein interaction, mRNA folding energy and half-life, and tRNA adaptation index and they showed that estimated protein yield was more dynamic than observed mRNA expression (2). This information shows that solely mRNA levels could not be the feature for prediction of protein abundance accurately; thus, in this study, we tried to quantify and to extract the information about codon, tRNA, and ribosome flow for given coding sequences.

Unlike the other methods in literature, we used Ribosome Flow Model output which gives the translation rate value based on association between codon-anticodon pairing by considering the ribosome jamming on open reading frame. This feature was clearly important for all three models and affected model performance. Ribosome flow model considers the codon order, size of ribosomes, and their order and its output includes steady-state occupancy probabilities of ribosomes at each sites (number of codons given by free parameters *C*) and steady-state translation rates (61). Other features we used unlike the other models was frequency of synonymous mutation in given codon sequence (L_sym), which was shown within top important features in two models (RF and GBM). Previously, synonymous codons were considered as silent, which did not affect the protein expression importantly; however, lastly evidences shows that codon usage through synonymous codons regulates and controls protein structure and gene expression by influencing on co-translational protein folding, translational efficiency and accuracy, mRNA stability and transcription (22). Studies with ribosome profiling have showed that ribosome occupancy time is different for the synonymous codons so that synonymous codons affect differently the speed of translation and translation rate which leads the co-translational folding kinetics of a protein (43). Some synonymous codons were labelled as rare codons and they slow down or pause the translation procedures at some important domain of protein to fine-tune the co-translational folding process, which is highly conservative and optimized during evolution (22). Based on this information, as expected, length of synonymous codons was most important features in our random forest model which had best performance on both training data and test data, which is indicated that frequency of synonymous codon within the encoding sequence was remarkable feature to estimate protein yield of given coding sequence.

Several features were used to quantify the effects of such different factors to the expression levels of protein product in this study. Commonly, most preferred metrices like codon adaptation index (CAI) and frequency of optimal codons (Fop) appeared to be little to no association to protein abundance in our data. On the other hand, tRNA adaptation index (tAI), and metric that calculated via Ribosome Flow Model (RFM) termed as translation rate referred as best measures of protein production rate in our

data. Since parameters required for tAI and RFM can be calculated much more easily unlike the CAI, which requires the determination of the highly expressed genes for the organism, these metrices can be used for relatively unknown species and new host for expression.

To sum up, we propose a model can predict the protein abundance levels from given codon sequence for different hosts. This prediction tool can be used for simplification of technical aspects of protein expression, can be used for filtering the infeasible targets, selection of host for a target protein, and can be used for optimization of codon by estimating the limitations of translation efficiency. Also, such a prediction tool can be used for the understanding of underlying mechanisms of disease caused by silent mutations or used for understanding of changes in protein production because of the translational mechanisms.

6. REFERENCES

1. Wu L, Candille SI, Choi Y, Xie D, Jiang L, Li-Pook-Than J, et al. Variation and genetic control of protein abundance in humans. Nature. 2013;499(7456):79–82.

2. Mehdi AM, Patrick R, Bailey TL, Boden M. Predicting the dynamics of protein abundance. Mol Cell Proteomics [Internet]. 2014 [cited 2021 Jun 15];13(5):1330–40. Available from: /pmc/articles/PMC4014288/

3. Ho B, Baryshnikova A, Brown Correspondence GW, Ch C, Sh OOO, Oh HO, et al. Unification of Protein Abundance Datasets Yields a Quantitative Saccharomyces cerevisiae Proteome Differential regulation Mass Spectrometry GFP-Microscopy TAP-Immunoblot Protein abundance in molecules per cell RNA-seq Stress protein abundance Ribosome profiling Ribosome Protein R N A Tag-affected proteins Stress abundance changes Article Unification of Protein Abundance Datasets Yields a Quantitative Saccharomyces cerevisiae Proteome. Cell Syst [Internet]. 2018 [cited 2021 Jun 15];6. Available from: https://doi.org/10.1016/j.cels.2017.12.004

4. Newman JRS, Ghaemmaghami S, Ihmels J, Breslow DK, Noble M, DeRisi JL, et al. Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. Nature [Internet]. 2006 Jun 15 [cited 2021 Jun 15];441(7095):840–6. Available from: https://www.nature.com/articles/nature04785

5. Gustafsson C, Minshull J, Govindarajan S, Ness J, Villalobos A, Welch M. Engineering genes for predictable protein expression [Internet]. Vol. 83, Protein Expression and Purification. NIH Public Access; 2012 [cited 2021 Jun 15]. p. 37–46. Available from: /pmc/articles/PMC3746766/

6. Li H, Siddiqui O, Zhang H, Guan Y. Joint learning improves protein abundance prediction in cancers. BMC Biol [Internet]. 2019 Dec 23 [cited 2021 Jun 15];17(1):1–14. Available from: https://doi.org/10.1186/s12915-019-0730-9

7. Vitrinel B, Koh HWL, Kar FM, Maity S, Rendleman J, Choi H, et al. Exploiting interdata relationships in next-generation proteomics analysis. Mol Cell Proteomics. 2019 Aug 9;18(8):S5–14.

8. Ferreira M, Ventorim R, Almeida E, Silveira S, Silveira W. Protein Abundance Prediction Through Machine Learning Methods. bioRxiv [Internet]. 2020 Sep 19 [cited 2021 Jun 16];2020.09.17.302182. Available from: https://doi.org/10.1101/2020.09.17.302182

9. Bhandari BK, Lim CS, Gardner PP. TISIGNER.com: web services for improving recombinant protein production. Nucleic Acids Res [Internet]. 2021 Mar 21 [cited 2021 Jun 16];(1). Available from: https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkab175/6179363

10. Gingold H, Pilpel Y. Determinants of translation efficiency and accuracy. Molecular Systems Biology. 2011.

11. Cannarrozzi G, Schraudolph NN, Faty M, von Rohr P, Friberg MT, Roth AC, et al. A role for codon order in translation dynamics. Cell. 2010;

12. Spencer PS, Barral JM. Genetic code redundancy and its influence on the encoded polypeptides. Comput Struct Biotechnol J. 2012;1(1):e201204006.

13. Varenne S, Buc J, Lloubes R, Lazdunski C. Translation is a non-uniform process. Effect of tRNA availability on the rate of elongation of nascent polypeptide chains. J Mol Biol. 1984;

14. Grantham R, Gautier C, Gouy M, Mercier R, Pavé A. Codon catalog usage and the genome hypothesis. Nucleic Acids Res [Internet]. 1980 Jan 11 [cited 2021 Jun 17];8(1):197. Available from: /pmc/articles/PMC327256/?report=abstract

15. Sharp PM, Tuohy TMF, Mosurski KR. Codon usage in yeast: Cluster analysis clearly differentiates highly and lowly expressed genes. Nucleic Acids Res [Internet]. 1986 Jul 11 [cited 2021 Jun 17];14(13):5125–43. Available from: /pmc/articles/PMC311530/?report=abstract

16. Plotkin JB, Kudla G. Synonymous but not the same. Natl Rev Genet. 2011;

17. Hershberg R, Petrov DA. Selection on codon bias. Annual Review of Genetics. 2008.

18. Brule CE, Grayhack EJ. Synonymous Codons: Choose Wisely for Expression [Internet]. Vol. 33, Trends in Genetics. Elsevier Ltd; 2017 [cited 2021 Jun 17]. p. 283–97. Available from: /pmc/articles/PMC5409834/

19. Brockmann R, Beyer A, Heinisch JJ, Wilhelm T. Posttranscriptional expression regulation: What determines translation rates? PLoS Comput Biol [Internet]. 2007 Mar [cited 2021 Jun 17];3(3):0531–9. Available from: /pmc/articles/PMC1829480/

20. Tuller T, Waldman YY, Kupiec M, Ruppin E. Translation efficiency is determined by both codon bias and folding energy. Proc Natl Acad Sci U S A [Internet]. 2010 Feb 23 [cited 2021 Jun 17];107(8):3645–50. Available from: /pmc/articles/PMC2840511/

21. Tuller T, Kupiec M, Ruppin E. Determinants of protein abundance and translation efficiency in S. cerevisiae. PLoS Comput Biol [Internet]. 2007 Dec [cited 2021 Jun 17];3(12):2510–9. Available from: /pmc/articles/PMC2230678/

22. Spencer PS, Siller E, Anderson JF, Barral JM. Silent substitutions predictably alter translation elongation rates and protein folding efficiencies. J Mol Biol. 2012 Sep 21;422(3):328–35.

23. Novoa EM, Pavon-Eternod M, Pan T, Ribas De Pouplana L. A role for tRNA modifications in genome structure and codon usage. Cell [Internet]. 2012 Mar 30 [cited 2021 Jun 18];149(1):202–13. Available from: http://www.cell.com/article/S0092867412002127/fulltext

24. Sharp PM, Cowe E, Higgins DG, Shields DC, Wolfe KH, Wright F. Codon usage patterns in Escherichia coli, Bacillus subtilis, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Drosophila melanogaster and Homo sapiens; a review of the considerable within-species diversity. Nucleic Acids Res [Internet]. 1988 Sep 12 [cited 2021 Jun 18];16(17):8207–11. Available from: /pmc/articles/PMC338553/?report=abstract

25. Gouy M, Gautier C. Codon usage in bacteria: Correlation with gene expressivity. Nucleic Acids Res. 1982;

26. Bennetzen JL, Hall BD. Codon selection in yeast. J Biol Chem. 1982;

27. Ikemura T. Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: A proposal for a synonymous codon choice that is optimal for the E. coli translational system. J Mol Biol. 1981;

28. dos Reis M, Savva R, Wernisch L. Solving the riddle of codon usage preferences: A test for translational selection. Nucleic Acids Res [Internet]. 2004 [cited 2021 Jun 3];32(17):5036–44. Available from: /pmc/articles/PMC521650/

29. Neelagandan N, Lamberti I, Carvalho HJF, Gobet C, Naef F. What determines eukaryotic translation elongation: recent molecular and quantitative analyses of protein synthesis. Open Biol [Internet]. 2020 Dec [cited 2021 Jun 18];10(12):200292. Available from: https://royalsocietypublishing.org/doi/abs/10.1098/rsob.200292

30. Fluitt A, Pienaar E, Viljoen H. Ribosome kinetics and aa-tRNA competition determine rate and fidelity of peptide synthesis. Comput Biol Chem [Internet]. 2007 Oct [cited 2021 Jun 19];31(5–6):335–46. Available from: /pmc/articles/PMC2727733/

31. Plant EP, Nguyen P, Russ JR, Pittman YR, Nguyen T, Quesinberry JT, et al. Differentiating between Near- and Non-Cognate Codons in Saccharomyces cerevisiae. PLoS One [Internet]. 2007 [cited 2021 Jun 19];2(6):e517. Available from: www.plosone.org

32. Chu D, Barnes DJ, Von Der Haar T. The role of tRNA and ribosome competition in coupling the expression of different mRNAs in Saccharomyces cerevisiae. Nucleic Acids Res [Internet]. 2011 [cited 2021 Jun 19];39(15):6705–14. Available from: /pmc/articles/PMC3159466/

33. Liu Y. A code within the genetic code: Codon usage regulates co-translational protein folding [Internet]. Vol. 18, Cell Communication and Signaling. BioMed Central Ltd; 2020 [cited 2021 Jun 19]. p. 1–9. Available from: https://doi.org/10.1186/s12964-020-00642-6

34. Zhao F, Yu CH, Liu Y. Codon usage regulates protein structure and function by affecting translation elongation speed in Drosophila cells. Nucleic Acids Res [Internet]. 2017 Aug 21 [cited 2021 Jun 19];45(14):8484–92. Available from: https://pubmed.ncbi.nlm.nih.gov/28582582/

35. Yang Q, Yu CH, Zhao F, Dang Y, Wu C, Xie P, et al. ERF1 mediates codon usage effects on

mRNA translation efficiency through premature termination at rare codons. Nucleic Acids Res [Internet]. 2019 Sep 26 [cited 2021 Jun 19];47(17):9243–58. Available from: https://pubmed.ncbi.nlm.nih.gov/31410471/

36. Yu CH, Dang Y, Zhou Z, Wu C, Zhao F, Sachs MS, et al. Codon Usage Influences the Local Rate of Translation Elongation to Regulate Co-translational Protein Folding. Mol Cell [Internet]. 2015 Sep 3 [cited 2021 Jun 19];59(5):744–54. Available from: https://pubmed.ncbi.nlm.nih.gov/26321254/

37. Zhou T, Weems M, Wilke CO. Translationally optimal codons associate with structurally sensitive sites in proteins. Mol Biol Evol. 2009;

38. Komar AA, Lesnik T, Reiss C. Synonymous codon substitutions affect ribosome traffic and protein folding during in vitro translation. FEBS Lett [Internet]. 1999 Dec 3 [cited 2021 Jun 19];462(3):387–91. Available from: https://febs.onlinelibrary.wiley.com/doi/full/10.1016/S0014-5793%2899%2901566-5

39. O'brien EP, Vendruscolo M, Dobson CM. Kinetic modelling indicates that fast-translating codons can coordinate cotranslational protein folding by avoiding misfolded intermediates. Nat Commun [Internet]. 2014 Jan 7 [cited 2021 Jun 19];5(1):1–11. Available from: www.nature.com/naturecommunications

40. Drummond DA, Wilke CO. Mistranslation-Induced Protein Misfolding as a Dominant Constraint on Coding-Sequence Evolution. Cell. 2008 Jul 25;134(2):341–52.

41. Buchan JR, Stansfield I. Halting a cellular production line: responses to ribosomal pausing during translation. Biol Cell [Internet]. 2007 Sep 1 [cited 2021 Jun 19];99(9):475–87. Available from: www.biolcell.org

42. Shoemaker CJ, Green R. Translation drives mRNA quality control [Internet]. Vol. 19, Nature Structural and Molecular Biology. Nature Publishing Group; 2012 [cited 2021 Jun 19]. p. 594–601. Available from: https://www.nature.com/articles/nsmb.2301

43. Buhr F, Jha S, Thommen M, Mittelstaet J, Kutz F, Schwalbe H, et al. Synonymous Codons Direct Cotranslational Folding toward Different Protein Conformations. Mol Cell [Internet]. 2016 Feb 4 [cited 2021 Jun 20];61(3):341–51. Available from: https://pubmed.ncbi.nlm.nih.gov/26849192/

44. Otto A, Becher D, Schmidt F. Quantitative proteomics in the field of microbiology [Internet]. Vol. 14, Proteomics. Wiley-VCH Verlag; 2014 [cited 2021 Jun 15]. p. 547–65. Available from: https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/full/10.1002/pmic.201300403

45. M F, R V, E A, S S, W S. Protein Abundance Prediction Through Machine Learning Methods. 2020 Sep 19 [cited 2021 Jun 20]; Available from: https://europepmc.org/article/ppr/ppr215241

46. Fernandes A, Vinga S. Improving protein expression prediction using extra features and ensemble

averaging. PLoS One [Internet]. 2016 Mar 1 [cited 2021 Jun 20];11(3):e0150369. Available from: http://sels.tecnico.ulisboa.

47. Sharp PM, Li WH. The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. 1987;15(3):1281–95.

48. Ikemura T. Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: A proposal for a synonymous codon choice that is optimal for the E. coli translational system. J Mol Biol. 1981;151(3):389–409.

49. Sabi R, Daniel RV, Tuller T. StAIcalc: TRNA adaptation index calculator based on species-specific weights. Bioinformatics [Internet]. 2017 Feb 15 [cited 2021 Jun 20];33(4):589–91. Available from: http://gtrnadb.ucsc.edu

50. Wright F. The "effective number of codons" used in a gene. Gene. 1990 Mar 1;87(1):23–9.

51. Futcher B, Latter GI, Monardo P, McLaughlin CS, Garrels JI. A Sampling of the Yeast Proteome. Mol Cell Biol [Internet]. 1999 Nov [cited 2021 Jun 21];19(11):7357–68. Available from: /pmc/articles/PMC84729/

52. Huang T, Wan S, Xu Z, Zheng Y, Feng KY, Li HP, et al. Analysis and prediction of translation rate based on sequence and functional features of the mRNA. PLoS One [Internet]. 2011 [cited 2021 Jun 21];6(1). Available from: https://pubmed.ncbi.nlm.nih.gov/21253596/

53. Welch M, Govindarajan S, Ness JE, Villalobos A, Gurney A, Minshull J, et al. Design parameters to control synthetic gene expression in Eschorichia coli. PLoS One [Internet]. 2009 Sep 14 [cited 2021 Jun 21];4(9). Available from: https://pubmed.ncbi.nlm.nih.gov/19759823/

54. Supek F, Šmuc T. On relevance of codon usage to expression of synthetic and natural genes in Escherichia coli. Genetics [Internet]. 2010 Jul [cited 2021 Jun 21];185(3):1129–34. Available from: https://pubmed.ncbi.nlm.nih.gov/20421604/

55. Ho TK. The random subspace method for constructing decision forests. IEEE Trans Pattern Anal Mach Intell. 1998;20(8):832–44.

56. Madeh Piryonesi S, El-Diraby TE. Using Machine Learning to Examine Impact of Type of Performance Indicator on Flexible Pavement Deterioration Modeling. J Infrastruct Syst [Internet]. 2021 Jun 10 [cited 2021 Jun 21];27(2):04021005. Available from: https://ascelibrary.org/doi/abs/10.1061/%28ASCE%29IS.1943-555X.0000602

57. Balakrishnan R, Park J, Karra K, Hitz BC, Binkley G, Hong EL, et al. YeastMine-An integrated data warehouse for Saccharomyces cerevisiae data as a multipurpose tool-kit. Database [Internet]. 2012 [cited 2021 Jun 20];2012. Available from: https://pubmed.ncbi.nlm.nih.gov/22434830/

58. Pozzoli U, Menozzi G, Fumagalli M, Cereda M, Comi GP, Cagliani R, et al. Both selective and neutral processes drive GC content evolution in the human genome. BMC Evol Biol [Internet].

2008 [cited 2021 Jun 20];8(1):99. Available from: /pmc/articles/PMC2292697/

59. Man O, Sussman JL, Pilpel Y. Examination of the tRNA adaptation index as a predictor of protein expression levels. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) [Internet]. Springer, Berlin, Heidelberg; 2007 [cited 2021 Jun 20]. p. 107–18. Available from: https://link.springer.com/chapter/10.1007/978-3-540-48540-7_10

60. Percudani R, Pavesi A, Ottonello S. Transfer RNA gene redundancy and translational selection in Saccharomyces cerevisiae. J Mol Biol. 1997 May 2;268(2):322–30.

61. Zur H, Tuller T. Rfmapp: Ribosome flow model application. Bioinformatics [Internet]. 2012 Jun 15 [cited 2021 Jun 20];28(12):1663–4. Available from: http://www.cs.tau.ac.il/~tamirtul/RFM_Installers/install.htm

62. Peden JF. Analysis of codon usage. Citeseer. 2000;

63. Meyer, D. and Wien FT. Support Vector Machines. Interface to Libsvm Packag e1071. 2015;

64. Ridgeway G. gbm: Generalized Boosted Regression Models [Internet]. R Package Version 2.1. 2013. Available from: https://cran.r-project.org/web/packages/gbm/gbm.pdf

7. CURRICULUM VITAE