The Journal of Multidisciplinary Research (TJMDR)

Content Available at www.saap.org.in ISSN: 2583-0317

QUICK VIEW ON BACTERIAL TRANSLATION ANALYSIS WEB TOOLS

Kona Venkata Sri Krishna[1], Kanigiri Deepthi Sri[2], Mohammed Azharuddin[3], Arrolla Lekha[2]

[1] Jawaharlal Nehru Technological University, Hyderabad, TS, India

[2] K L University, Vijayawada, AP, India

[3] Vignan University, Guntur, AP, India

[4] Sreenidhi Institute of Science and Technology, Hyderabad, TS, India

Received: 04 July 2023 Revised: 22 July 2023 Accepted: 28 June 2023

Abstract

Bacterial translation was learnt by researcher from the past four decades and significant data was generated. Inquisite to understand performance of the bacteria to produce a particular recombinant protein so as to pre-evaluate and make necessary modifications for optimal production is the key interest for researcher and biopharma manufacturers. Over a decade various databases were built and based on this valuable data webtools were developed which enable researcher to tweak the strategy beforehand. Here in this article we outlined various database and webtools based on protein translation which are currently being used.

Keywords: Bacterial translation, webtools, biopharma manufacturers.

***Corresponding Author**

Kona Venkata Sri Krishna

DOI: https://doi.org/10.37022/tjmdr.v3i2.468

Produced and Published by

South Asian Academic Publications

Introduction

Protein translation is the synthesis of proteins from their constituent amino acids. The ribosome is a massive protein-RNA complex that reads codons and binds to messenger RNA (mRNA). The codons are subsequently translated into amino acids, which are then used to construct proteins. Transfer RNA (tRNA) is a minute RNA molecule that transports amino acids to the ribosome. After recognising the codon on the mRNA, the tRNA transfers the relevant amino acid to the ribosome. The tRNA is then freed, enabling it to transport one additional amino acid to the ribosome. This procedure will continue until the protein is synthesised in its entirety(1). Prokaryotic translation analysis using web tools can involve a variety of techniques, including the following:

i. Identification of open reading frames (ORFs): ORFs are regions of DNA that contain the coding sequence for proteins. Web tools can be used to identify ORFs in a prokaryotic genome and predict the corresponding amino acid sequences.

ii. Translation start site prediction: Many web tools can be used to predict the location of the start codon, which is the codon that initiates protein synthesis. This can be useful for identifying potential alternative start sites and for understanding the regulation of protein synthesis.

iii. Codon usage analysis: Some web tools can be used to analyze the frequency of codon usage in a prokaryotic genome, which can provide insights into the evolution of the organism and the regulation of protein synthesis.

iv. Protein structure and function prediction: Once the amino acid sequence has been identified, web tools can be used to predict the three-dimensional structure of the protein and to infer its possible function.

v. Comparative genomics: Web tools can be used to compare prokaryotic genomes, which can be useful for identifying conserved genes and regions, and for understanding the evolution of different organisms.

Web tools can be used for prokaryotic translation analysis, and many other resources that are used frequently specific needs of the researcher are discussed in brief in this paper.

General aspects that are analyzed by webtools to study translation

a. Ribosome profiling

In the studies conducted to determine how mRNA sequences influenced translation rates, the collected data were not properly used. Translation initiation rates cannot be derived from ribosome density until the elongation

rates of each mRNA can be properly predicted or measured. Until then, these measures are meaningless. Because they performed ribosome profiling on modified ribosomes, the scientists were able to see the translation of ribosomes with unique anti-S-D sequences.

b. Shine-Dalgarno structure

Depending on the codon pairing, a Shine-Dalgarno-like structure may be formed. This structure halts the translation process by matching target messenger RNA with 16S ribosomal RNA produced by translated ribosomes. Surprisingly, the research indicates that a strong base pair between 16S ribosomal RNA and an mRNA-Shine-Dalgarno motif is neither necessary nor sufficient for translation initiation.

c. Ribosomal -proteins:

Overall, we were able to demonstrate that R-protein S1 exhibits rna-melting activity at the exit point of a 30S decoding channel and also allows the ribosome to initiate mRNA translation with more flexibility. The mRNA RPSO found in Escherichia coli is one of the most prominent examples. This mRNA encodes the ribosomal protein (r-protein), S15, which is responsible for transporting the pseudoknot structure inside the ribosome. The 30S subunit is responsible for translation, whereas the S15 subunit regulates the pseudoknot structure of r-protein S15. When plasmids are present in recombinant E. coli, the bacteria are referred to as recombinant E. coli, and the concentrations of various host-cell proteins and ribosome components are altered compared to their original forms. The rate-limiting step in protein synthesis is the translation initiation phase, which ensures that the first codon-anticodon connection is generated at the correct location on the 30S ribosomal peptidyl (P) subunit. The irregularities in the local decoding rates that occur at certain mRNA codons are a distinguishing feature of the translation process.

d. Open read-frame

In addition, new research has shown that the open read-frame decoding rate (the time it takes ribosomes to translate an mRNA) is a significant factor in determining the amount of protein produced, but it is not the only one. This is especially true when the codons with the quickest decoding rate are positioned at the 5' end of an mRNA. At the translational level, control may occur in one of two ways: either by directly regulating ribosome progression through mRNAs or by changing mRNA stability via differences in the rate of NoGo decay. Consequently, the concentration of any of these components has no influence on ternary complexes or the action of ribosomes on translation rates.

e. Structure prediction

Additionally, it is often helpful to apply bioinformatics techniques such as JPRED to search for domain borders and predict intrinsically disordered protein regions (IDP). You will never be successful if you try to depict a construct that is insufficiently long and lacks a vital domain component, such as a beta strand. If you try to produce a construct that is too long and includes flexible parts that are susceptible to proteolysis, heterogeneity or the loss of a purification tag will certainly result. Due to their sensitivity to degradation, it is notoriously difficult to synthesise proteins containing a significant number of IDP regions. Regardless, it is essential to keep in mind that many IDP regions might acquire structure via interactions with other molecules. Whether the structure of your protein or a structure that closely resembles it is available, it is important to analyse the accessibility of the N- and C-termini to determine if the insertion of a tag is likely to disrupt the protein's structure. This may be evaluated by determining whether or not the protein termini are buried inside the structure. In this method, it is also feasible to use structure prediction tools such as Phyre 2. MEROPS is a database for locating proteases that act on recombinant proteins intended for bacterial production.

Webtools of Translation

Inferring ribosome pauses and visualizing profiling data: PausePred and Rfeet

Using this web programme, it is possible to infer ribosome pauses from ribosome profiling data (Ribo-seq). Peaks in ribosome footprint density are graded based on their importance relative to the surrounding density. The score enables the comparison of peaks throughout the transcriptome or genome. In addition to the score, PausePred provides the coordinates of the pause, the footprint density at the pause site, and the nucleotide sequence immediately around the pause. Rfeet visualises pauses within Ribo-seq and RNA-seq density graphs generated for certain transcripts or genomic regions. There are existing platforms that have been developed for this purpose, e.g. RiboGalaxy and Plastid, that provide the tools for these steps as well as many additional utilities(2).

Identifying Actively Translated ORFs Using Ribo-Seq: RiboToolkit

It is a web-based tool that allows Ribo-Seq data processing to be centrally managed. Included in this research are data cleaning and quality evaluation, expression analysis based on RPFs, codon occupancy analysis, translation efficiency analysis, differential translation analysis, functional annotation, translation metagenomics analysis, and the identification of actively translated ORFs. In addition, user-friendly online interfaces have been developed to speed the data processing procedure and deliver understandable results. Codon use and ribosome stalling analyses were conducted to identify highly active codons and instances of codon stalling. ORFs that are actively translated may be found more rapidly and efficiently(3).

Web-Based Interactive Chat-Like Tool for Ribo-Seq Data Analysis: RiboChat

It is a web-based, interactive tool that permits direct processing of ribo-seq data in a chat-like format. A user-friendly web interface serves as the front end for a cloud computing service. The object-text detection module is enabled whenever a data analysis query is typed into the

chat window. This allows the module to identify relevant words within the provided text. Using the distinct analytics modules, the identified input characteristics are then ranked to obtain the best possible match. The specialised analytics module is active only when the successful uploading of datasets and setup of parameters have been verified. Overall, it represents a big step forward in the developing trend of next-generation data analytics, enabling the larger research community to analyse translation information concealed inside Ribo-seq data with ease. This is an important development in the expanding area of data analytics for the future generation(4).

Identifying Translation Features: ORFik:

It is an easy-to-use application programming interface (API) and collection of tools for researching translation and its modification. It develops a system that integrates data from several sources and expands the capabilities of GenomicRanges so that it can now analyse the transcriptome in addition to the genome. ORFik simplifies the processes required to process, analyse, and visualise the several steps of translation, with a particular emphasis on the start and elongation phases. It may accept high-throughput sequencing data from ribosome profiling in order to quantify ribosome elongation and RCP-seq or TCP-seq data in order to quantify ribosome scanning. ORFik may also use CAGE data to precisely characterise 5' UTRs and RNA-seq data to compute translation based on RNA abundance. These choices are both available. ORFik can annotate translated regions like proteins or upstream open reading frames, and it can support and compute more than thirty distinct translation characteristics and metrics compiled from the scientific literature (uORFs). We demonstrate how ORFik may be used to rapidly annotate the dynamics of 5' UTRs in diverse tissues, locate their uORFs, and characterise their scanning and translation in the downstream protein-coding regions of the genome(5).

Optimisation of Translation Initiation Sites on mRNA Using TIsigner

TIsigner allows you to influence protein synthesis by optimising the accessibility of translation initiation sites on mRNA.With the use of RNAplfold, the sections used to calculate accessibility (opening energy) are particular to the expression hosts. E. Because E. Since E. coli is the most common protein expression host, the default settings are optimised for this organism. E. coli bacteria have a T7 lac promoter system. In this instance, just the protein-coding sequence is necessary, and the default 5'-UTR (5' untranslated region) sequence is the most common shortened variant of the T7 promoter. With regard to the E. With the T7 lac promoter system, the default behaviour is to choose as the initial solution the optimised sequence that is closest to the desired expression level(6).

Tool to Obtain Solubility-Enhancing Tags by SoDoPE

SWI is the foundation of our interactive solubility analysis and optimization application, SoDoPE. When you submit the form, a query is sent to the HMMER web service to get domain annotation.The user may choose a point on the globe, and the likelihood of solubility, flexibility, and GRAVY (grand average of hydropathicity) for that area are shown in real time. Several solubility-enhancing tags may be compared and shown relative to user-specified locations using a bar plot. These tags include thioredoxin (TRX), maltose binding protein (MBP), small ubiquitin-related modifier (SUMO), and glutathione S-transferase (GST). Users may also insert a desired fusion sequence, which may be specified in nucleotide or protein sequence format(7).

Codon Tuning Strategies in the Production of Recombinant Proteins: ExpressInHost

It is a piece of open-source software that may be used to enhance the production of recombinant proteins. In this context, codon tuning strategies are based on codon use bias, the amount of guanine-cytosine nucleotides present, mRNA secondary structure, and sequence repeats. Numerous factors may alter the rate of translation along mRNAs, which is not constant. Despite the fact that it has been demonstrated that the secondary structure of messenger RNA (mRNA) is not a crucial factor, These tools are intended to complement the commercially available GenSmart Design and GeneWiz software packages. In the creation of recombinant proteins, three distinct codon-tweaking approaches are used. These procedures are referred to as Mode 1: Direct Mapping, Mode 2: Optimization and Conservation I, and Mode 3: Optimization and Conservation II. Mode 1: Direct mapping is the first of these strategies. In addition to considering tRNA abundances in the host organism, the three approaches listed above are used to optimise heterologous protein synthesis. Other popular tools often lack these approaches, which are necessary for preserving essential translation sites that aid in the correct folding of co-translational intermediates(8).

A Web-Based Tool for Plasmid Analysis: PLACNETw

PLACNETw, a web-based application based on the PLACNET system, offers an interactive graphical user interface, automates BLAST searches, and extracts decision-making-relevant data. It enables a domain expert to see scaffold graphs and related information about contigs and reference sequences, making interactive pruning methods on a personal computer possible without the need for additional software. Following the pruning phase, each plasmid transforms into a unique connected component subgraph.The user will be prompted immediately to download the analysis results. PLACNET is a graph-based tool for reconstructing plasmids using pair-end information from next-generation sequencing. Assembled contigs and reference genomes are the two kinds of nodes found in PLACNET graphs, as well as the two types of edges (scaffold links and homology to references). PLACNET requires manual graph pruning; however, people with little bioinformatics knowledge may find this procedure difficult.In PLACNET, the graph is

manually modified, often known as "pruning," resulting in increased precision and sensitivity. Both Plasmid Finder and ACLAME are analytical tools designed to assist with plasmid research. While both are capable of identifying and evaluating individual plasmid contigs, neither was meant to assemble them. By building a network of contig linkages, PLACNET facilitates the display and study of plasmids in whole genome sequencing studies. It also permits the identification of plasmids, making it a helpful instrument for conducting broad genetic investigations of plasmid populations. Two levels comprise the PLACNET Web application: the server layer and the client layer. It uses bash and python on the server layer to access PLACNET and blast queries, and as a result, it provides the client layer with a collection of files. Scaffold detection and read assembly are included.where the read assembly is carried out using the Velvet Optimizer script to determine the ideal read length (screening the last 32 bp of the maximum read length). The subsequent values will be used by PLACNET's Bowtie2 algorithm to search for potential scaffold links between assembled contigs. NCBI's collection of whole genomes and plasmids is used to compare contigs to a frequently updated reference database. This comparison is conducted to identify pertinent references. The programme Prodigal is used to identify potential coding sequences, which are then linked to plasmid protein reference libraries in search of relaxases, replication initiation proteins, and incompatibility groups(9).

Predictive Modeling of the Production of Recombinant Proteins in the Periplasm: PERISCOPE-Opt

Typically, optimising the fermentation process for the production of recombinant proteins (RPP) requires a substantial financial investment. Machine learning (ML) strategies are efficient in reducing the number of tests conducted and have several applications in RPP. These ML-based approaches, on the other hand, are primarily concerned with amino acid sequence characteristics, hence neglecting the potential impact of fermentation process factors. Combining features derived from fermentation process conditions with those derived from amino acid sequence, the current study develops an ML-based model that predicts the maximal protein yields and corresponding fermentation conditions for the expression of a target recombinant protein in the periplasm of Escherichia coli. This is accomplished by combining parameters acquired from fermentation process settings and amino acid sequences. The first phase of the strategy entailed utilising two independent sets of XGBoost classifiers to detect whether the target protein expression levels were high (more than 50 mg/L), medium (between 0.5 and 50 mg/L), or low (less than 0.5 mg/L). The second step of the framework included three regression models, two of which were SVMs and one of which was a random forest. These models were used to anticipate the expression yields for each expression-level class. The predictor attained an overall average accuracy of 75% and a Pearson coefficient correlation of 0.91 for the instances that it accurately detected, according to the results of numerous separate tests. As a result, our model provides a trustworthy alternative to completing a large number of tests using the trial-and-error technique to determine the appropriate fermentation conditions and yield for RPP. Moreover, it is implemented as an open-access web server called Periscope-Opt(10).

Predicting and regulating the initiation of translation : RBS Calculator

The ribosome binding site (RBS) calculator is a tool for predicting and regulating the initiation of translation and protein synthesis in bacteria. The approach can precisely predict the rate of translation initiation for each start codon inside an mRNA transcript. Optimizing a synthetic RBS sequence is another method for reaching a certain translation start rate. It is possible to rationally regulate the translation rate of a protein coding sequence across a 100,000-fold range using the RBS Calculator. First, an overview of the RBS Calculator's potential biotechnology applications, such as the optimization of synthetic metabolic pathways and genetic circuits, will be presented. Then, we go into the concepts, methodologies, and algorithms that support the thermodynamic model and optimization procedure of the RBS Calculator. In the final section of this paper, we present a method for accurately measuring the amounts of steady-state fluorescent protein expression.These strategies and suggestions will aid in your comprehension of the RBS Calculator and its applications(11).

Regression Method to Evaluate the Effect of Genetic Variants on Gene Translation: PGExpress

To get a deeper knowledge of the link between genotype and phenotype, it is crucial that we precisely characterise the translational process. The development of efficient methods for assessing translation efficiency and/or protein expression from nucleotide sequences is a basic challenge in computational biology. Particularly, predicting the influence of genetic variants on gene expression would allow the optimization of certain pathways and functions for the development of innovative biological systems. PGExpress is a novel regression approach that can accurately estimate the log2-fold change in the translation efficiency of an mRNA sequence in E. coli. The PGExpress algorithm requires 12 inputs, which correspond to the expected secondary structure of RNA and anti-Shine-Dalgarno hybridization free energies. To train the method, 1,772 sequence variants (WT-High) representing 137 essential E. coli genes were used. Genes from coli were utilised. After the GFP superfolder, we analysed 13 different sequence variants of the first 33 nucleotides encoding identical amino acids for each gene. Each variant of a gene is represented by a unique sequencing block. Among these sequence blocks are the ribosome binding site (RBS), the first 33 nucleotides of the coding region (C33), the rest of the coding region (CC), and their related combinations. Our gradient-boosting-based

algorithm, PGExpress, was trained on the WT-High dataset using ten iterations of a gene-based cross-validation procedure. PGExpress performed well in this test, with a correlation value of 0.60 and a root-mean-square error (RMSE) of 1.3. PGExpress got an overall accuracy of 0.74, a Matthews correlation coefficient of 0.48, and an area under the receiver operating characteristic curve (AUC) of 0.81 when the regression work was recast as a classification problem. PGExpress trumps the RBSCalculator when it comes to predicting the log2-fold change in translational efficiency and its variance on the WT-High dataset. true about the regression task. By evaluating five freshly produced mRNA sequence variants in-house, we were able to demonstrate the efficiency of our methodology. The predicted expression levels of new variants are consistent with the results of our E. coli study and investigations(12).

Bacterial Transcription Using Biophysical Models: Promoter Calculator

It is capable of predicting the rates of site-specific transcription initiation across each of the 70 promoter sequences. The model was first trained and validated by measuring transcription rates and locating transcription start sites (TSSs) on hundreds of in vitro-generated promoter sequences. This was followed by further validation using thousands of identified promoters inside cells. The model as a whole has 346 visible parameters used to determine the intensities of interactions at the 10 hexamer, 10 extended motif, 35 hexamer, upstream element (UP) element, discriminator, spacer region, and ITR, with verified predictions over 22,132 distinct promoter sequences. The notion was subsequently implemented via the construction of synthetic promoter sequences with specified transcription start rates and the identification of sources of cryptic transcription in engineered genetic systems. This model gives a deeper understanding of how canonical and non-canonical patterns influence transcription rates throughout all potential DNA regions, resulting in the transcriptional profile of a genetic system. We designed a cost-effective (and humanly understandable) strategy. It was developed on a biophysical model of bacterial transcription initiation, and its validity has been verified across thousands of promoters with widely different sequences. Thus, customised genetic systems may now govern transcription start rates at specific places. Our model illustrates how many weak interactions, like the protein's promiscuous activity on DNA sequences devoid of matching binding motifs, contribute to RNAP/70 transcriptional control. Our bottom-up model-building strategy is easily scalable to other RNAPs and complexes, with the eventual goal of developing a universal system-wide language for engineering gene control in synthetic genetic systems. It also illustrates how developments in machine learning may be utilised to enhance rather than replace existing thermodynamic formalisms(13).

Conclusion

Protein synthesis is necessary for the normal function of every cell. Proteins are involved in almost every biological process, from DNA replication to cell signal transmission. We show that a computer model can accurately forecast the numbers of ribosomes, tRNA, mRNA, and elongation factors in Escherichia coli depending on its growth rate. If you wish to construct a protein with a lot of IDP regions, you need to think about whether or not they are prone to degradation. Proteins may create structures as a consequence of interactions with other molecules, such as other proteins and ribosomes. The translation initiation step, which ensures that the first codon-anticodon interaction is made at the appropriate spot on the ribosomal peptidyl (P)-subunit, is the rate-limiting step in protein synthesis. At the translational level, control may occur in one of two ways: either by directly altering ribosome progress through mRNAs or by modifying the mRNA's stability. Using the Rfeet tool in the context of ribo- and RNA-seq density plots, pauses may be seen. Codon use and ribosome stalling analyses were conducted to identify highly active codons and instances of codon stalling. A user-friendly web interface serves as the front end for a cloud computing service.

ORFik can annotate regions that have been translated, such as proteins or upstream open reading frames, and it can support and compute more than thirty distinct translation characteristics and metrics. SWI is the foundation of our interactive solubility analysis and optimization application, SoDoPE. Codon tuning strategies are based on codon usage bias, guanine-cytosine nucleotide abundance, mRNA secondary structure, and sequence repeats. PLACNET is a graph-based tool for reconstructing plasmids using pair-end information from next-generation sequencing. PLACNET is a web application that constructs a network of contig linkages to display and analyse plasmids in whole genome sequencing research.

In PLACNET, the graph is manually modified, often known as "pruning," resulting in increased precision and sensitivity. The ribosome binding site (RBS) calculator may be used to predict and control translation initiation and protein synthesis in bacteria. Optimizing a synthetic RBS sequence is another method for reaching a certain translation start rate. In this study, we introduce PGExpress, a novel regression approach that can accurately predict the log2-fold change in the translation efficiency of an mRNA sequence in E. coli. The approach was trained using 1,772 sequence variants, which represented 137 essential genes. We designed a cost-effective (and humanly understandable) strategy. It was constructed based on a biophysical model of bacterial transcription initiation, and its validity has been validated across hundreds of promoters with vastly different sequences. Thus, customised genetic systems may now govern transcription start rates at specific places.

References

1. Krishna S, Yim DG, Lakshmanan V, Tirumalai V, Koh JL, Park JE, et al. Dynamic expression of tRNA-derived small RNAs define cellular states. 2019;20(7):e47789.
2. Kumari R, Michel AM, Baranov PV. PausePred and Rfeet: webtools for inferring ribosome pauses and visualizing footprint density from ribosome profiling data. RNA (New York, NY). 2018;24(10):1297-304.
3. Liu Q, Shvarts T, Sliz P, Gregory RIJNar. RiboToolkit: an integrated platform for analysis and annotation of ribosome profiling data to decode mRNA translation at codon resolution. 2020;48(W1):W218-W29.
4. Xie M, Yang L, Chen G, Wang Y, Xie Z, Wang HJBiB. RiboChat: a chat-style web interface for analysis and annotation of ribosome profiling data. 2022;23(2):bbab559.
5. Tjeldnes H, Labun K, Torres Cleuren Y, Chyżyńska K, Świrski M, Valen EJBb. ORFik: a comprehensive R toolkit for the analysis of translation. 2021;22(1):1-16.
6. Bhandari BK, Lim CS, Gardner PPJNAR. TISIGNER. com: web services for improving recombinant protein production. 2021;49(W1):W654-W61.
7. 7. Bhandari BK. Computational tools for improving recombinant protein production: University of Otago; 2021.
8. Stansfield I, Romano MCJapa. ExpressInHost: A codon tuning tool for the expression of recombinant proteins in host microorganisms. 2021.
9. de Toro M, Lanza VF, Vielva L, Redondo-Salvo S, de la Cruz F. Plasmid reconstruction from Next-Gen data: a detailed protocol for the use of PLACNETw for the reconstruction of plasmids from WGS datasets. Horizontal Gene Transfer: Springer; 2020. p. 323-39.
10. Packiam KAR, Ooi CW, Li F, Mei S, Tey BT, Ong HF, et al. PERISCOPE-Opt: Machine learning-based prediction of optimal fermentation conditions and yields of recombinant periplasmic protein expressed in Escherichia coli. 2022.
11. Salis HM. The ribosome binding site calculator. Methods in enzymology. 2011;498:19-42.
12. Zhao L, Abedpour N, Blum C, Kolkhof P, Beller M, Kollmann M, et al., editors. Predicting gene expression level in E. coli from mRNA sequence information. 2019 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB); 2019: IEEE.
13. LaFleur TL, Hossain A, Salis HMJNc. Automated model-predictive design of synthetic promoters to control transcriptional profiles in bacteria. 2022;13(1):1-15.