Title

ΔG$_{unfold}$ leaderless, a package for high-throughput analysis of translation initiation regions (TIRs) at the transcriptome scale and for leaderless mRNA optimization

Authors

Mohammed-Husain M. Bharmal (orcID: 0000-0002-2365-3034), Jared M. Schrader (orcID:0000-0002-5728-5882)*

Affiliations

Department of Biological Sciences, Wayne State University, Detroit, MI, 48202, USA
*Corresponding Author (Schrader@wayne.edu)

Abstract

Background

Translation initiation is an essential step for fidelity of gene expression, in which the ribosome must bind to the translation initiation region (TIR) and position the initiator tRNA in the P-site (1). For this to occur correctly, the TIR encompassing the ribosome binding site (RBS) needs to be highly accessible (2-5). ΔG_{unfold} is a metric for computing accessibility of the TIR, but there is no automated way to compute it manually with existing software/tools limiting throughput.

Results

ΔG$_{unfold}$ leaderless allows users to automate the ΔG_{unfold} calculation to perform high-throughput analysis. Importantly, *ΔG$_{unfold}$ leaderless* allows calculation of TIRs of both leadered mRNAs and leaderless mRNAs which lack a 5' UTR and which are abundant in bacterial, archaeal, and mitochondrial transcriptomes (4, 6, 7). The ability to analyze leaderless mRNAs also allows one additional feature where users can computationally optimize leaderless mRNA TIRs to maximize their gene expression (8, 9).

Conclusions

The ΔG_{unfold} leaderless package facilitates high-throughput calculations of TIR accessibility, is designed to calculate TIR accessibility for leadered and leaderless mRNA TIRs which are abundant in bacterial/archaeal/organellar transcriptomes and allows optimization of leaderless mRNA TIRs for biotechnology.

Keywords

translation initiation, leaderless mRNA, translation initiation region (TIR), ribosome binding site (RBS)

Background

ΔG_{unfold} is a metric to compute accessibility of translation initiation region (TIR) which is an important determinant of start codon selection (10). Translation initiation is an essential step for fidelity of gene expression, in which the Ribosome must bind to the TIR and position the start codon in the P-site with the initiator tRNA (1). For this to occur, the translation initiation region (TIR) encompassing the ribosome

binding site (RBS) needs to be highly accessible, lacking secondary structure (2-5). In contrast to calculating the minimum free energy (mfe) of the TIR region using nearest neighbor approaches, ΔG_{unfold} provides information about the TIR region boundary by providing the site of ribosome binding, giving an accessibility measure that better correlates with translation efficiency (11-14). ΔG_{unfold} as the metric for computing accessibility has gained appreciation but there is no automated way to compute it with existing softwares/tools. This currently limits the use of ΔG_{unfold} as the low throughput of calculation has made it difficult to perform transcriptome level analyses. Importantly, there are softwares and algorithms available which can predict secondary structure of any mRNA sequence and calculate the minimum free energy(mfe) of the resulting secondary structure. With these tools, one can in principle calculate the ΔG_{unfold} of any structure, however, this requires multiple steps to be executed sequentially and the constrained structure needs to be generated and inputted manually based on the start codon position and TIR location. So, though computation of the ΔG_{unfold} for a single sequence is possible, it is not a practical solution for the entire transcriptome or multiple sequences. Therefore, this *ΔG_{unfold} leaderless* software will make high-throughput transcriptome-level computation of TIR accessibility feasible.

Importantly, *ΔG_{unfold} leaderless* also has TIR optimization program which can be used to optimize expression of proteins in bacteria with predominant non-SD translation initiation mechanism, unlike *E.coli*. *E. coli* has been widely used for expression of recombinant proteins as it is very well characterized and various genetic tools are readily available (15). However, two key challenges for heterologous protein expressions are efficient synthesis (dictated by rates of transcription and translation) and misfolding. Therefore, *E. coli* system cannot be the best choice for all the applications and other hosts are required influenced by the application and availability of tool for their genetic manipulation (15). For instance, translation initiation in *C. crescentus* is not predominantly Shine-Dalgarno (SD) dependent unlike *E. coli*, and only approximately 23% of genes initiate translation via SD mediated mechanism (16-18) with some organisms containing SD sites in as few as 8% (7, 19). Further, RNA-seq based transcription mapping experiments have found that many bacterial mRNAs are “leaderless” and begin directly at the AUG start codon (20-22), and that this type of mRNAs are abundant in pathogens such as *M. tuberculosis* and in the mammalian mitochondria (6). In C. crescentus approximately 17% of its transcriptome is leaderless (16). For leaderless mRNAs, three factors are known to affect translation initiation efficiency: accessibility of TIRs (ΔG_{unfold}), start codon identity and leader length with higher accessibility of TIR, AUG as start codon and any absence of short leaders are most efficient (7, 23-30). Therefore, these properties can be exploited to optimize expression of proteins in such organisms having predominantly leaderless translation initiation machinery. A similar approach has been used in *E.coli*, in which synonymous mutations were rendered in 5’ coding region to elevate the expression of the reporter protein (8). Using this part of the *ΔG_{unfold} leaderless* package, the user can input the coding sequence (CDS) and then the program will alter the start codon to AUG if there is any other codon and make synonymous mutations in the TIR and compute ΔG_{unfold} for all possible combinations and output most preferred sequence in terms of least ΔG_{unfold} or highest accessibility.

Implementation

General requirements

ΔG_{unfold} leaderless is an open-source software available as a linux shell script available through github (https://github.com/schraderlab/dGunfold_program) based upon the RNAfold and dot2ct packages (31, 32). ΔG_{unfold} represents the energy required to unfold the mRNA's ribosome binding site (RBS) (Fig 1A) (10) which can be useful in studies of translation initiation. There are 3 subprograms available based on the need of the user, which are for transcriptome studies, leaderless mRNA TIR optimization, or calculation of custom sequences. Before running ΔG_{unfold} leaderless, the user also needs to install RNAfold (32) and RNAstructure (31) programs as described in the README file. The computational pipeline of the three subprograms are described in the following sections, and detailed instructions to run the software on linux can be found in the README file.

The basic ΔG_{unfold} Computational Pipeline

The core section of this program is ΔG_{unfold} calculation (Fig 1). In order to calculate the ΔG_{unfold} of multiple RNAs, a .txt file with all the RNA sequences and a second .txt file with start codon positions in the same order as the sequences are inputted into the program (Fig 1B). The user can also define the temperature in which they want to calculate ΔG_{unfold}. The ΔG_{unfold} leaderless package will then calculate the minimum free energy of the mRNA structure using RNAfold (ΔG_{mRNA}), and will then constrain the ribosome binding site (RBS) to be single stranded as an approximation for ribosome accessibility, and then calculate the minimum free energy of the constrained mRNA structure (ΔG_{init}) (Fig 1B). With these two measures, the ΔG_{unfold} can be calculated for each RNA (Fig 1B). The RBS region is defined by the size of a ribosome footprint surrounding the start codon (25nt) which is dynamically adapted to smaller sizes for mRNAs with short 5' UTRs, or leaderless mRNAs which completely lack 5' UTRs (Fig 1C). To constrain the structure as single-stranded, all base-pairs within the RBS are constrained to be single stranded before recalculating the ΔG_{init}. The output file contains six columns: RNA sequence, dot-bracket mfe structure, constrained structure, ΔG_{mRNA} values, ΔG_{init} values and ΔG_{unfold} values (Table 1). Users are allowed to additionally control the size of the TIR region for analysis (default = 50nt) which provide additional flexibility to the user.

The ΔG_{unfold} pipeline for transcriptome scale analysis

Bacterial transcriptome architecture maps provide an ideal platform to analyze TIR accessibility on a transcriptome-wide scale. Notably, transcription start site (TSS) data and RNA-seq density data have been used previously to experimentally map bacterial transcriptomes (33). If such TSS data are available, the user can input this together with the genbank file of the respective genome containing the CDS annotations and an operon map to build a transcript architecture model that allows the best starting point for transcriptome analyses. TSSs are assigned to transcripts by searching within 300nts upstream of the start codon of either the CDS (or the leading CDS in an operon). Multiple TSS sites may be assigned to a single operon/gene and in such cases the different mRNA isoforms are included in the transcript architecture map. Once the TSS sites are defined to each operon, the ΔG_{unfold} is calculated for the TIR of each CDS in each mRNA transcriptional units. If no TSS data is provided, the software can approximate each CDS as having an mRNA leader of 25 nts, although this leads to uncertainly about short leadered and leaderless mRNAs. Overall, the ΔG_{unfold} leaderless transcriptome analysis runs within 30 min on a budget desktop cpu for a dataset equivalent to the largest known bacterial transcriptome, suggesting that this software can be run to study any bacterial transcriptome (Fig 2B).

ΔG_{unfold} for optimizing leaderless mRNA translation initiation

The lack of secondary structure is a strong determinant of the translation initiation efficiency of leaderless mRNAs (9, 34, 35). As some organisms with abundant leaderless mRNAs are important for biotechnology, we adapted *ΔG_{unfold} leaderless* to generate mutant versions with higher accessibility to optimize recombinant gene expression (Fig 3). The user inputs the sequence of their leaderless mRNA to optimize, and all possible synonymous codon mutations in the RBS are generated and output to the user in a rank ordered list from lowest to highest *ΔG_{unfold}*. The user can then select those variants with optimal properties for optimal expression testing.

Conclusions

ΔG_{unfold} leaderless automates the computation of mRNA TIR accessibility for large number of sequences allowing transcriptome-level analysis. Indeed, runtimes allow users to quickly calculate *ΔG_{unfold}* across a whole transcriptome using a budget desktop computer (Figure 2B). In addition, *ΔG_{unfold} leaderless* can also be used for biotechnology purposes to help optimize the translation initiation of recombinantly expressed leaderless mRNAs for biotechnology purposes. Together, these functions make *ΔG_{unfold} leaderless* useful for the study of TIR accessibility on mRNA translation.

Availability and requirements

Project name: ΔG_{unfold} leaderless
Project home page: https://github.com/schraderlab/dGunfold_program
Operating system(s): Linux/Unix command line
Programming language: Python/shell script
Other requirements: Python 3.0 or higher, Biopython
License: GNU General Public License v3.0
Any restrictions to use by non-academics: e.g. licence needed

List of abbreviations

- TIR – translation initiation region
- RBS – ribosome binding site
- Mfe – minimum free energy
- SD – Shine-Dalgarno

Declarations

Ethics approval and consent to participate

"Not applicable"

Consent for publication

"Not applicable"

Availability of data and materials

All relevant python files and example datasets are freely available in the ΔGunfold leaderless github repository https://github.com/schraderlab/dGunfold_program

Competing interests

"The authors declare that they have no competing interests"

Funding

NIH grant R35GM124733 to JMS and a WSU Rumble fellowship to MHB.

Authors' contributions

MHB wrote software. MHB and JMS tested software and wrote manuscript.

Acknowledgements

The authors thank members of the Schrader lab for critical feedback. The authors thank James Aretakis and Aishwarya Ghosh for beta-testing and feedback on usage.

References

1. Yusupova GZ, Yusupov MM, Cate JH, Noller HF. The path of messenger RNA through the ribosome. Cell. 2001;106(2):233-41.
2. Espah Borujeni A, Channarasappa AS, Salis HM. Translation rate is controlled by coupled trade-offs between site accessibility, selective RNA unfolding and sliding at upstream standby sites. Nucleic Acids Res. 2014;42(4):2646-59.
3. Gu W, Zhou T, Wilke CO. A universal trend of reduced mRNA stability near the translation-initiation site in prokaryotes and eukaryotes. PLoS Comput Biol. 2010;6(2):e1000664.
4. Jones CN, Wilkinson KA, Hung KT, Weeks KM, Spremulli LL. Lack of secondary structure characterizes the 5' ends of mammalian mitochondrial mRNAs. RNA. 2008;14(5):862-71.
5. Nakamoto T. A unified view of the initiation of protein synthesis. Biochem Biophys Res Commun. 2006;341(3):675-8.
6. Montoya J, Ojala D, Attardi G. Distinctive features of the 5'-terminal sequences of the human mitochondrial mRNAs. Nature. 1981;290(5806):465-70.
7. Srivastava A, Gogoi P, Deka B, Goswami S, Kanaujia SP. In silico analysis of 5'-UTRs highlights the prevalence of Shine-Dalgarno and leaderless-dependent mechanisms of translation initiation in bacteria and archaea, respectively. J Theor Biol. 2016;402:54-61.
8. Goltermann L, Borch Jensen M, Bentin T. Tuning protein expression using synonymous codon libraries targeted to the 5' mRNA coding region. Protein Eng Des Sel. 2011;24(1-2):123-9.
9. Kudla G, Murray AW, Tollervey D, Plotkin JB. Coding-sequence determinants of gene expression in Escherichia coli. Science. 2009;324(5924):255-8.
10. Mustoe AM, Corley M, Laederach A, Weeks KM. Messenger RNA Structure Regulates Translation Initiation: A Mechanism Exploited from Bacteria to Humans. Biochemistry. 2018;57(26):3537-9.
11. Boel G, Letso R, Neely H, Price WN, Wong KH, Su M, et al. Codon influence on protein expression in E. coli correlates with mRNA levels. Nature. 2016;529(7586):358-63.
12. Guimaraes JC, Rocha M, Arkin AP. Transcript level and sequence determinants of protein abundance and noise in Escherichia coli. Nucleic Acids Res. 2014;42(8):4791-9.

13. Li GW, Burkhardt D, Gross C, Weissman JS. Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. Cell. 2014;157(3):624-35.
14. 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. 2010;107(8):3645-50.
15. Rosano GL, Ceccarelli EA. Recombinant protein expression in Escherichia coli: advances and challenges. Front Microbiol. 2014;5:172.
16. Schrader JM, Zhou B, Li GW, Lasker K, Childers WS, Williams B, et al. The coding and noncoding architecture of the Caulobacter crescentus genome. PLoS Genet. 2014;10(7):e1004463.
17. Shine J, Dalgarno L. The 3'-terminal sequence of Escherichia coli 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites. Proc Natl Acad Sci U S A. 1974;71(4):1342-6.
18. Steitz JA, Jakes K. How Ribosomes Select Initiator Regions in Messenger-Rna - Base Pair Formation between 3' Terminus of 16s Ribosomal-Rna and Messenger-Rna during Initiation of Protein-Synthesis in Escherichia-Coli. Proceedings of the National Academy of Sciences of the United States of America. 1975;72(12):4734-8.
19. Chang B, Halgamuge S, Tang SL. Analysis of SD sequences in completed microbial genomes: non-SD-led genes are as common as SD-led genes. Gene. 2006;373:90-9.
20. Beck HJ, Moll I. Leaderless mRNAs in the Spotlight: Ancient but Not Outdated! Microbiol Spectr. 2018;6(4).
21. Cortes T, Schubert OT, Rose G, Arnvig KB, Comas I, Aebersold R, et al. Genome-wide mapping of transcriptional start sites defines an extensive leaderless transcriptome in Mycobacterium tuberculosis. Cell Rep. 2013;5(4):1121-31.
22. Shell SS, Wang J, Lapierre P, Mir M, Chase MR, Pyle MM, et al. Leaderless Transcripts and Small Proteins Are Common Features of the Mycobacterial Translational Landscape. PLoS Genet. 2015;11(11):e1005641.
23. Korman M, Schlussel S, Vishkautzan M, Gur E. Multiple layers of regulation determine the cellular levels of the Pup ligase PafA in Mycobacterium smegmatis. Mol Microbiol. 2019;112(2):620-31.
24. Krishnan KM, Van Etten WJ, 3rd, Janssen GR. Proximity of the start codon to a leaderless mRNA's 5' terminus is a strong positive determinant of ribosome binding and expression in Escherichia coli. J Bacteriol. 2010;192(24):6482-5.
25. Van Etten WJ, Janssen GR. An AUG initiation codon, not codon-anticodon complementarity, is required for the translation of unleadered mRNA in Escherichia coli. Molecular Microbiology. 1998;27(5):987-1001.
26. Jones RL, 3rd, Jaskula JC, Janssen GR. In vivo translational start site selection on leaderless mRNA transcribed from the Streptomyces fradiae aph gene. J Bacteriol. 1992;174(14):4753-60.
27. Chen WC, Yang GP, He Y, Zhang SM, Chen HY, Shen P, et al. Nucleotides Flanking the Start Codon in hsp70 mRNAs with Very Short 5 '-UTRs Greatly Affect Gene Expression in Haloarchaea. Plos One. 2015;10(9).
28. Christian BE, Spremulli LL. Preferential Selection of the 5 '-Terminal Start Codon on Leaderless mRNAs by Mammalian Mitochondrial Ribosomes. Journal of Biological Chemistry. 2010;285(36):28379-86.
29. Hering O, Brenneis M, Beer J, Suess B, Soppa J. A novel mechanism for translation initiation operates in haloarchaea. Mol Microbiol. 2009;71(6):1451-63.
30. Bharmal M-HM, Gega A, Schrader JM. A combination of mRNA features influence the efficiency of leaderless mRNA translation initiation. NAR Genomics and Bioinformatics. 2021:Accepted.
31. Reuter JS, Mathews DH. RNAstructure: software for RNA secondary structure prediction and analysis. BMC Bioinformatics. 2010;11:129.

32. Gruber AR, Lorenz R, Bernhart SH, Neubock R, Hofacker IL. The Vienna RNA websuite. Nucleic Acids Res. 2008;36(Web Server issue):W70-4.
33. Bharmal MH, Aretakis JR, Schrader JM. An Improved Caulobacter crescentus Operon Annotation Based on Transcriptome Data. Microbiol Resour Announc. 2020;9(44).
34. de Smit MH, van Duin J. Secondary structure of the ribosome binding site determines translational efficiency: a quantitative analysis. Proc Natl Acad Sci U S A. 1990;87(19):7668-72.
35. Scharff LB, Childs L, Walther D, Bock R. Local Absence of Secondary Structure Permits Translation of mRNAs that Lack Ribosome-Binding Sites. Plos Genetics. 2011;7(6).

Table and Figure legends

Figure 1: ΔG_{unfold} core computation

A) Graphical representation showing the unfolding of the ribosome binding site (RBS) in the translation initiation region (TIR), facilitating binding of the ribosome to the RBS. Also, showing the calculation for ΔG_{unfold}. B) Flowchart of execution of ΔG_{unfold} leaderless program with user input sequences. The core component of this pipeline is shown within the dotted box. An example output file is included in Table 1. C) Algorithm for accounting for short/leaderless mRNAs. Flowchart showing RBS length selection for constraining based on the length of 5' UTR.

Figure 2: Transcriptome-wide ΔG_{unfold} computation

A) Flowchart of execution of ΔG_{unfold} leaderless program for the entire transcriptome. If TSS data is available, the exact transcriptional units defined by transcription start site data can be used (left), or if not available, each transcriptional unit will be calculated as if it contained a leader of >=25 nts. An example output file is included in Table 2. B) Scatter plot showing the time required to execute the core component of the program (Y-axis) for different number of sequences (X-axis). The maximum size of the X-axis was chosen to be equivalent to the largest bacterial genome identified. Analysis was performed using an AMD Ryzen 3200g desktop cpu.

Figure 3: Translation initiation region (TIR) accessibility optimization program

Flowchart of execution of ΔG_{unfold} leaderless program for the TIR accessibility optimization. Optimization is performed using synonymous codon mutations in the RBS and providing the user a rank-order list of all variants based on their ΔG_{unfold} (Table 3).

Table 1: Output file format of ΔG_{unfold} leaderless program with user input sequences.

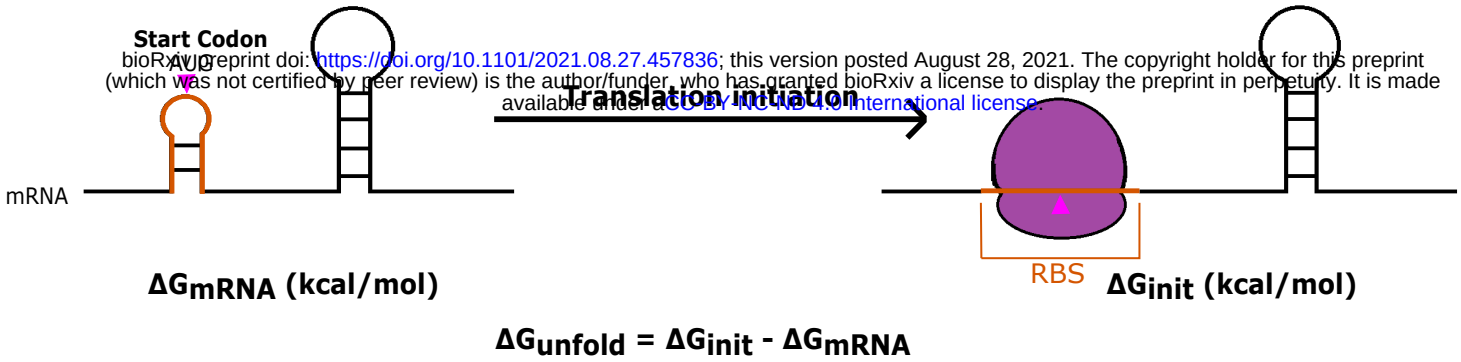
(XLSX).

Table 2: Output file formats of ΔG_{unfold} leaderless program for transcriptome analysis.

(XLSX).

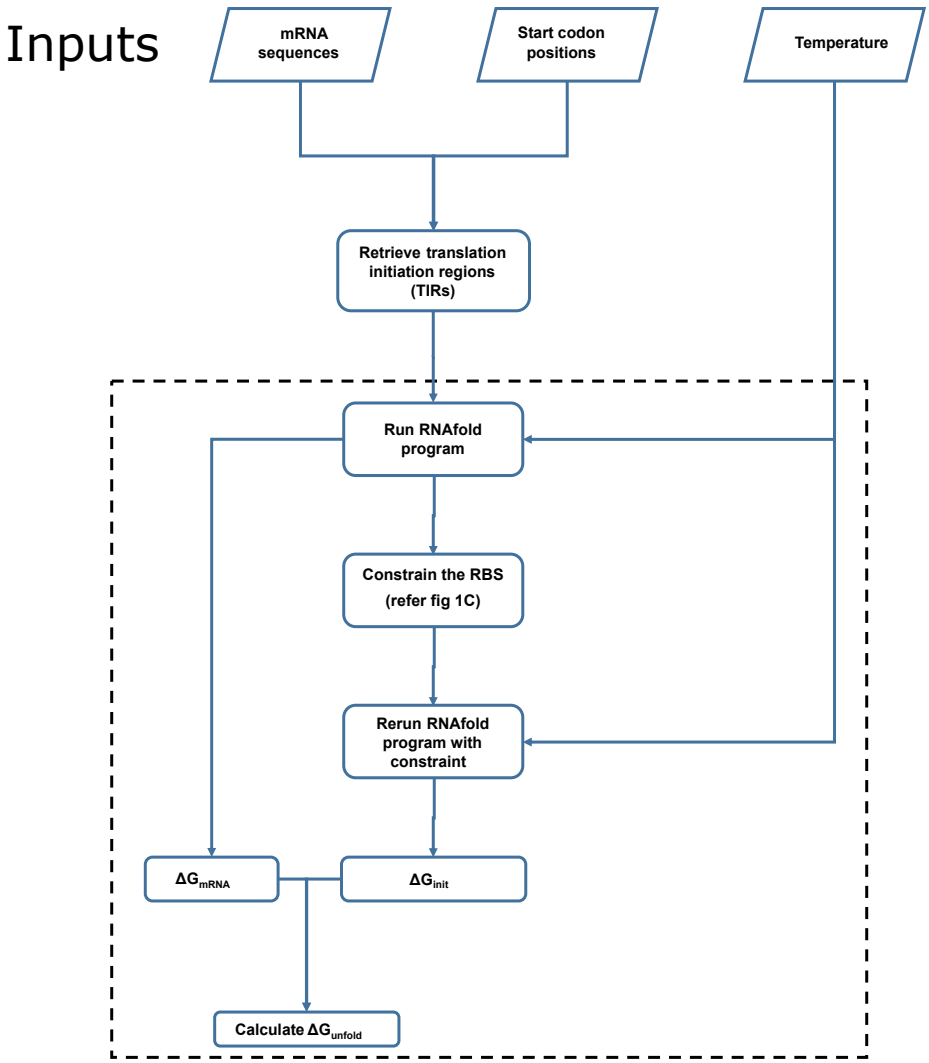
Table 3: Output file format of ΔG_{unfold} leaderless program for TIR accessibility optimization.
(XLSX).

A.

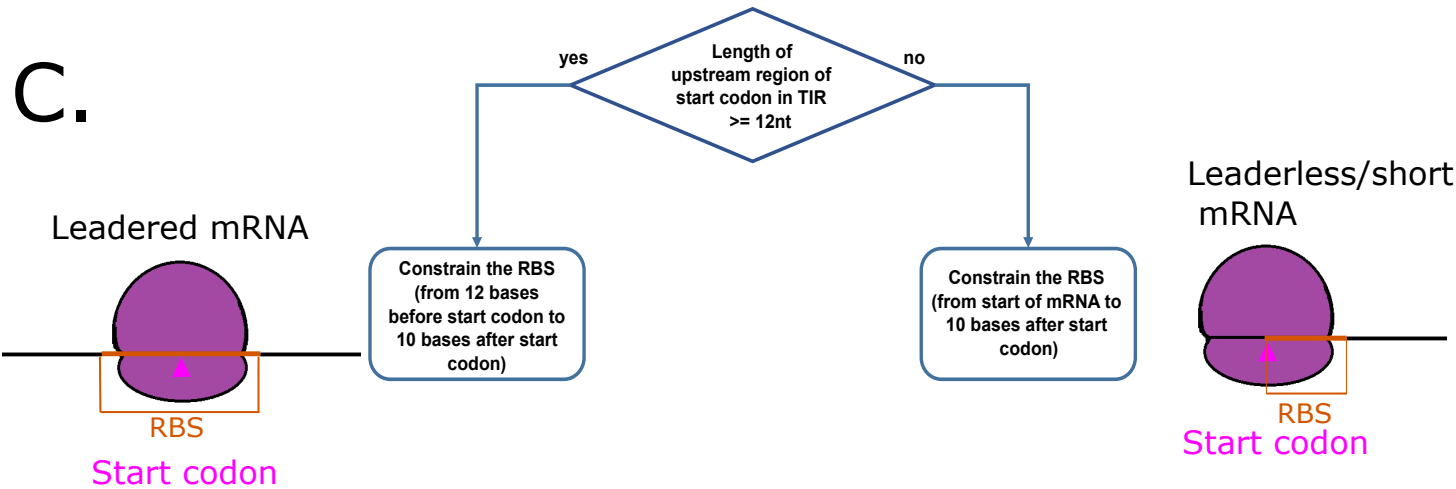


$$\Delta G_{unfold} = \Delta G_{init} - \Delta G_{mRNA}$$

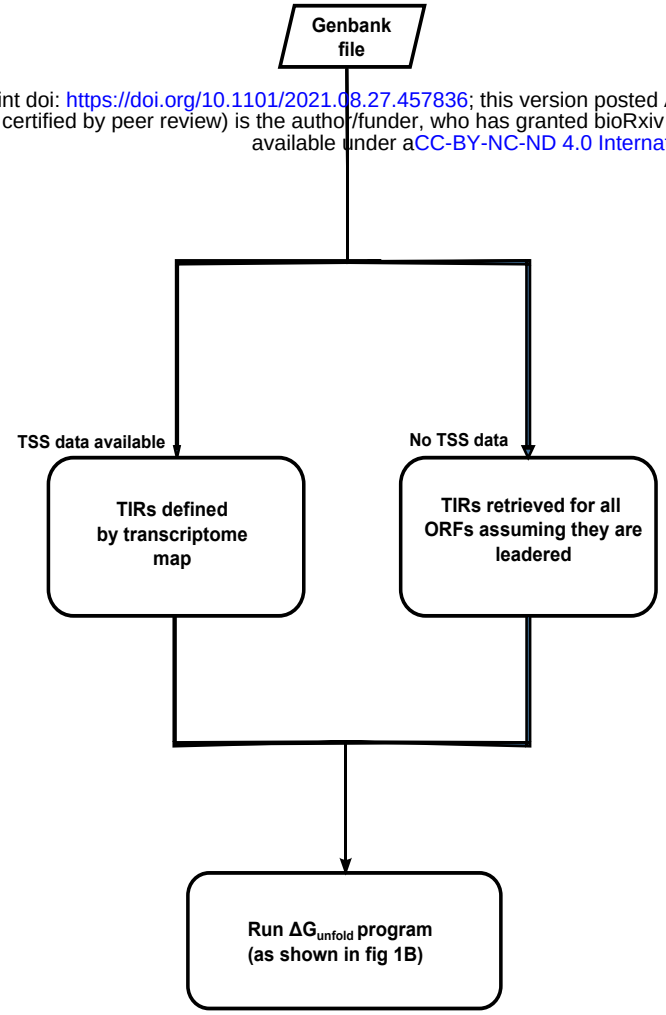
B.



C.



A.



B.