# Highly accessible translation initiation sites are predictive of successful heterologous protein expression

Bikash K. Bhandari[1,†], Chun Shen Lim[1,†,*], Paul P. Gardner[1,2,*]

[1]Department of Biochemistry, School of Biomedical Sciences, University of Otago, Dunedin, New Zealand
[2]Biomolecular Interaction Centre, University of Canterbury, Christchurch, New Zealand

[†]These authors contributed equally.
*Corresponding authors. Emails: chunshen.lim@otago.ac.nz; paul.gardner@otago.ac.nz

**Abstract:** (149/150 words)
Recombinant protein production in microbial systems is well-established, yet half of these experiments have failed in the expression phase. Failures are expected for 'difficult-to-express' proteins, but for others, codon bias, mRNA folding, avoidance, and G+C content have been suggested to explain observed levels of protein expression. However, determining which of these is the strongest predictor is still an active area of research. We used an ensemble average of energy model for RNA to show that the accessibility of translation initiation sites outperforms other features in predicting the outcomes of 11,430 experiments of recombinant protein production in *Escherichia coli*. We developed TIsigner and showed that synonymous codon changes within the first nine codons are sufficient to improve the accessibility of translation initiation sites. Our software produces scores for both input and optimised sequences, so that success/failure can be predicted and prevented by PCR cloning of optimised sequences.

## Introduction

Recombinant protein expression has numerous applications in biotechnology and biomedical research. Despite extensive refinements in protocols over the past three decades, half of the experiments have failed in the expression phase[1]. Notable problems are the low expression of 'difficult proteins' such as membrane proteins, and the poor growth of the expression hosts, which may relate to the toxicity of heterologous proteins[2] (reviewed in detail elsewhere[3,4]). If these issues are factored out, we expect a strong correlation between mRNA and protein levels. However, this assumption oversimplifies the complexity of translation and turnover of biomolecules because mRNA abundance can only explain up to 40% of the variation in protein abundance[5–11]. Furthermore, the strong promoters used in expression vectors do not always lead to a desirable level of protein expression[12].

For *Escherichia coli*, two main models were proposed to explain the low correlation between mRNA and protein levels, which are based on either codon or mRNA folding analysis. Codon analysis measures a bias in codon usage using codon adaptation index (CAI)[13] or tRNA adaptation index (tAI)[14,15] whereas mRNA folding analysis predicts the presence of

RNA secondary structures and their folding stability. Codon usage bias is thought to correlate with tRNA abundance, translation efficiency and protein production[13–17] but its usefulness has been questioned upon[18–21]. In contrast, many findings support the model based on mRNA folding in which the stability of RNA structures around the Shine-Dalgarno sequence and/or translation initiation sites inversely correlates with protein expression[18,19,21–24]. We recently proposed a third model in which the avoidance of inappropriate interactions between mRNAs and non-coding RNAs has a strong effect on protein expression[25]. The roles of these models in protein expression is still an active area of research.

The common algorithms of gene optimisation samples synonymous protein-coding sequences using 'fitness' models based on CAI, tAI, mRNA folding, and/or G+C content (%)[26–30]. However, these 'fitness' models are usually based on some of the above findings that relied on either endogenous proteins, reporter proteins or a few other proteins with their synonymous variants. It is unclear whether these features are generalisable to explain the expression of various heterologous proteins. To address this question, we studied multiple large datasets across species in order to extract features that allow us to predict the outcomes of 11,430 experiments of recombinant protein expression in *E. coli*. With this information, we propose how such features can be exploited to fine-tune protein expression at a low cost.

## Results

### Accessibility of translation initiation sites strongly correlates with protein abundance

To explore new features that could explain the expression of heterologous proteins, we first examined an *E. coli* expression dataset of green fluorescent protein (GFP) fused in-frame with a library of 96-nt upstream sequences (n=244,000)[21]. We clustered these 96-nt upstream sequences using CD-HIT-EST[31,32], giving rise to 14,425 representative sequences. We calculated the accessibility that represents the opening energy for all possible sub-sequences of these sequences (see Methods). For each sub-sequence region, we examined the correlation between the opening energy and GFP levels. We found that the opening energy of translation initiation sites, in particular from the nucleotide positions −30 to 18 (−30:18), showed a maximum correlation with protein abundance (Fig 1A; $R_s$=−0.65, $P<2.2\times10^{-16}$). This is stronger than the correlation between the minimum free energy −30:30 and protein abundance, which was previously reported as the highest rank feature (Fig 1A; $R_s$=0.51, $P<2.2\times10^{-16}$). The P-values of multiple testing were adjusted using Bonferroni's correction and reported to machine precision. The datasets used and results were summarised in Supplementary Table S1.

We repeated the analysis for a dataset of yellow fluorescent protein (YFP) expression in *Saccharomyces cerevisiae*[23]. This dataset corresponds to a library of 5′UTR variants, in which the 10-nt sequences preceding the YFP translation initiation site were randomly substituted (n=2,041). In this case, the opening energy −7:89 showed a stronger correlation with protein abundance than that of the minimum free energy −15:50 reported previously (Fig 1B; $R_s$=−0.55 versus 0.46).

To examine the usefulness of accessibility in complex eukaryotes, we analysed a dataset of GFP expression in *Mus musculus*[33]. The reporter library was originally designed to measure the strength of translation initiation sequence context, in which the 6- and 2-nt sequences upstream and downstream of the GFP translation initiation site were randomly substituted, respectively (n=65,536). Here the opening energy −8:11 showed a maximum correlation with expressed proteins, which again, is stronger than that of the minimum free energy −30:30 (Fig 1C; $R_s$=−0.28 versus 0.12).

Taken together, our findings suggest that the accessibility of translation initiation sites strongly correlates with protein abundance across species. Interestingly, our findings also suggest that *E. coli* tends to have a longer accessible 5′UTR region than that of *S. cerevisiae* and *H. sapiens* (−30 versus −7 and −8; see Fig 1). This can be explained by the presence of the Shine-Dalgarno sequence[34] at the region −13:−8, which should be accessible to recruit ribosomes.

**Accessibility predicts the outcome of recombinant protein expression**

We investigated how accessibility performs in the real world in prediction of recombinant protein expression. For this purpose, we analysed 11,430 expression experiments in *E. coli* from the 'Protein Structure Initiative:Biology' (PSI:Biology)[35–37]. These PSI:Biology targets were expressed using the pET21_NESG expression vector that harbours the T7lac inducible promoter and a C-terminal His tag[37].

We split the experimental results of the PSI:Biology targets into protein expression 'success' and 'failure' groups (n=8,780 and 2,650, respectively; see Supplementary Fig S2). These PSI:Biology targets spanned more than 189 species and the failures are representative of various problems in heterologous protein expression. Only 1.6% of the experiments belong to homologous protein expression, which is negligible (n=179; see Supplementary Fig S2).

We calculated the opening energy for all possible sub-sequences of the PSI:Biology targets as above (Fig 2). For each sub-sequence region, we used the opening energy levels to predict the expression outcome and computed the prediction accuracy using the area under the receiver operating characteristic curve (AUC; see Fig 2C). A closer look into the correlations and AUC scores calculated for the sub-sequence regions reveals a strong accessibility signal of translation initiation sites (Fig 2B and C, Cambray's GFP and PSI:Biology datasets, respectively). Although the sequences of the Cambray's GFP and PSI:Biology datasets are different, we reasoned that the correlations and AUC scores can be compared by the sub-sequence regions that are in common (see Fig 2A for an example of a sub-sequence region). Based on this idea, we matched the correlations and AUC scores by sub-sequence region and confirmed that sub-sequence regions that have strong correlations are likely to have high AUC scores (Fig 2D). In contrast, the sub-sequence regions that have zero correlations are not useful for predicting the expression outcome (AUC approximately 0.5).

We then asked how accessibility manifests in the endogenous mRNAs of *E. coli*, for which we studied the proteomics dataset of 3,725 proteins consolidated in the PaxDb[38]. As expected, we observed a similar accessibility signal, with the region −25:16 correlated the most with protein abundance (Fig 2E). However, the correlation was rather low (R=−0.17, $P<2.2\times10^{-16}$), which might be due to the limitations of mass spectrometry[39,40]. Furthermore, the endogenous promoters have variable strength, which gives rise to a broad range of mRNA and protein levels[41,42]. Taken together, our results show that the accessibility signal of translation initiation site is surprisingly consistent across various datasets analysed (Supplementary Fig S1 and Fig 2).

**Accessibility outperforms other features in prediction of recombinant protein expression**

To choose an accessibility region for subsequent analyses, we selected the top 200 regions from the above correlation analysis on Cambray's dataset (Fig 2B) and ranked their Gini importance scores in prediction of the outcomes of the PSI:Biology targets. The region −24:24 was ranked first, which is nearly identical to the region −23:24 with the top AUC score (Fig 2C, AUC=0.70). We therefore used the opening energy at the region −24:24 in subsequent analysis.

We asked how the other features perform compared to accessibility in prediction of heterologous protein expression, for which we analysed the same PSI:Biology dataset. We first calculated the minimum free energy and avoidance at the regions −30:30 and 1:30, respectively. These are the local features associated with translation initiation rate. We also calculated CAI[13], tAI[43], codon context (CC)[44], G+C content (%), and Iχnos scores[45]. CC is similar to CAI except it takes codon-pair usage into account, whereas the Iχnos scores are translation elongation rates predicted using a neural network model trained with ribosome profiling data. These are the global features associated with translation elongation rate. The AUC scores for the local features were 0.70, 0.67 and 0.62 for the opening energy, minimum free energy and avoidance, respectively, whereas the global features were 0.58, 0.57, 0.54, 0.54 and 0.51 for Iχnos, G+C content (%), CAI, CC and tAI, respectively (Fig 3A). The local features outperform the global features, suggesting that effects on translation initiation can predict the outcome of heterologous protein expression. Our findings support previous reports that the effects on translation initiation are rate-limiting[18,24] which, interestingly, correlate with the binary outcome of recombinant protein expression (Fig 3B). Importantly, accessibility outperformed all other features.

To identify a good opening energy threshold, we calculated positive likelihood ratios for different opening energy thresholds using the cumulative frequencies of true negative, false negative, true positive and false positive derived from the above ROC analysis (Fig 4, top panel). Meanwhile, we calculated the 95% confidence intervals of these positive likelihood ratios using 10,000 bootstrap replicates. We reasoned that there is an upper and lower bound on translation initiation rate, therefore the relationship between translation initiation rate and accessibility is likely to follow a sigmoidal pattern. We fit the positive likelihood ratios into a four-parametric logistic regression model (Fig 4). As a result, we are 95% confident

that an opening energy of 10 or below at the region −24:24 is about two times more likely belongs to the sequences which are successfully expressed than those that failed.

**Accessibility can be improved using a simulated annealing algorithm**

The above results suggest that accessibility can, in part, explain the low expression problem of heterologous protein expression, we sought to exploit this idea in gene optimisation. We developed a simulated annealing algorithm to maximise the accessibility at the region −24:24 using synonymous codon substitution (see Methods). Previous studies have found that full-length synonymous codon-substituted transgenes may produce unexpected results, in particular a reduction in mRNA level[25,45,46]. Therefore, we sought to determine the minimum number of codons needed for synonymous substitutions in order to achieve near optimum accessibility. For this purpose, we used the PSI:Biology targets that failed to be expressed. As a control, we first applied our simulated annealing algorithm such that synonymous substitutions can happen at any codon of the sequences except the start and stop codons (see Methods). Although full-length synonymous codon substitution was allowed, the changes may not necessarily happen to all codons due to the stochastic nature of our optimisation algorithm. Next, we constrained synonymous codon substitution to the first 14 codons and applied the same procedure (Supplementary Fig S3). Therefore, the changes may only occur at any or all of the first 14 codons. We repeated the same procedure for the first nine and also the first four codons. Thus a total of four series of codon-substituted sequences were generated. We then compared the distributions of opening energy −24:24 for these series using the Kolmogorov-Smirnov statistic ($D_{KS}$; see Fig 5A). The distance between the distributions of the nine and full-length codon-substituted series was significantly different yet sufficiently close ($D_{KS}$=0.09, $P=3.3\times10^{-8}$), suggesting that optimisation of the first nine codons is sufficient in most cases to achieve an optimum accessibility of translation initiation sites. We named our software as Translation Initiation coding region designer (TIsigner), which by default, allows synonymous substitutions up to the first nine codons.

We asked to what extent the existing gene optimisation tools modify the accessibility of translation initiation sites. For this purpose, we first submitted the PSI:Biology targets that failed to be expressed to the ExpOptimizer webserver from NovoPro Bioscience (see Methods). We also optimised the PSI:Biology targets using the standalone version of Codon Optimisation OnLine (COOL)[29]. We found that both tools increase accessibility indirectly even though their algorithms are not designed as such (i.e., the 5′UTR sequence is not taken into account). In fact, a purely random synonymous codon substitution on these PSI:Biology targets using our own script resulted in a similar increase in accessibility (Fig 5B). These results may explain some indirect benefits from the existing gene optimisation tools.

## Discussion

Our findings show that the accessibility of translation initiation sites is the best predictor of heterologous protein expression in *E. coli,* as originally proposed in the 1970s/80s[47]. Increasing the accessibility of the 5′ region, including the Shine-Dalgarno sequence, facilitates the recruitment of ribosomes and therefore increases the translation initiation rate and protein level. In a landmark study, Salis et al. designed a total of 132 synthetic ribosome binding sites using minimum free energy models[27]. They found that weakly structured ribosome binding sites result in high red fluorescent protein levels. This was supported by recent studies using the endogenous *folA* and *adk* genes[48] and a dual-reporter system in *E. coli*[49]. These studies, and many others, support our finding that optimisation of the accessibility of translation initiation sites is a key to improve heterologous protein production.

Previous studies have used minimum free energy models to define the accessibility of a region of interest[27,48,49]. However, we have discovered that the opening energy is a better choice for modelling accessibility (see Fig 1A for example). Opening energy is an ensemble average of energy that takes into account of suboptimal RNA structures that are not reported by minimum free energy models by default[50,51]. Currently, the modelling of accessibility using opening energy is largely used for the prediction of RNA-RNA intermolecular interactions, for example, as implemented in RNAup and IntaRNA[52,53]. Our study has shown that this approach can be used to identify the key accessibility regions that are consistent across multiple large expression datasets.

The strengths of our approach (implemented in the TIsigner webservice and software tool) are four-fold. Firstly, the likelihood of success or failure can be assessed prior to running an experiment. Users can compare the opening energy calculated for the input and optimised sequences and the distributions of the 'success' and 'failure' of the PSI:Biology targets. We also introduced a scoring scheme to score the input and optimised sequences based upon how likely they are to be expressed (Fig 4; also see Methods). Secondly, optimised sequences can have up to the first nine codons substituted (by default), meaning that gene optimisation using a standard PCR cloning method is feasible. For cloning, we propose a nested PCR approach, in which the final PCR reaction utilises a forward primer designed according to the optimised sequence[54] (Fig 5C). Thirdly, the cost of gene optimisation can be reduced dramatically as gene synthesis is replaced with PCR using our approach. This enables high-throughput protein expression screening using the optimised sequences, generated at a low cost. Finally, tunable expression is possible, i.e. high, intermediate or even low expression 5′ codon sequences can be designed, allowing for more control over heterologous protein production. Although our study focuses largely on the expression of recombinant proteins without an N-Terminal fusion tag, our findings might give meaningful insights to other systems.

## Methods

### Sequence features analysis

Minimum free energy, opening energy and avoidance were calculated using RNAfold, RNAplfold and RNAup from ViennaRNA package (version 2.4.11), respectively[50–52,55–58]. RNAfold was run with default parameters. For RNAplfold, sub-sequences were generated from the input sequences to calculate opening energy (using the parameters -W 210 -u 210). For RNAup, we examined the stochastic interactions between the region 1:30 of each mRNA and 54 non-coding RNAs (using the parameters -b -o). RNAup reports the total interaction between two RNAs as the sum of energy required to open accessible sites in the interacting molecules $\Delta G_u$ and the energy gained by subsequent hybridisation $\Delta G_h$ [50]. For the interactions between each mRNA and 54 non-coding RNAs, we chose the most stable mRNA:ncRNA pair to report an inappropriate mRNA:ncRNA interaction, i.e. the pair with the strongest hybridisation energy, $(\Delta G_h)_{min}$.

CAI, tAI and CC were calculated using the reference weights from Sharp and Li[13], Tuller et al.[43] and Ang et al.[44], respectively. Translation elongation rate was predicted using Iχnos[45] trained with ribosome profiling data (SRR7759806 and SRR7759807)[59]. See Supplementary Table S1 for the datasets used in this study.

### TIsigner

Finding a synonymous sequence with a maximum accessibility is a combinatorial problem that spans a vast search space. For example, for a protein-coding sequence of nine codons, assuming an average of 3 synonymous codons per amino acid, we can expect a total of 19,682 unique synonymous coding sequences. This number increases rapidly with increasing number of codons. Heuristic optimisation approaches are preferred in such situations because the search space can be explored more efficiently to obtain nearly optimal solutions.

To optimise the accessibility of a given sequence, TIsigner uses a simulated annealing algorithm[60–63], a heuristic optimisation technique based on the thermodynamics of a system settling into a low energy state after cooling. A simulated annealing algorithm has been used to solve several combinatorial optimisation problems in bioinformatics. For example, we previously applied this algorithm to align and predict non-coding RNAs from multiple sequences[64]. Other studies use this algorithm to find consensus sequences[62] and optimise the ribosome binding sites[27] and mRNA folding[65] using minimum free energy models.

According to statistical mechanics, the probability $p_i$ of a system occupying energy state $E_i$, with temperature $T$, follows a Boltzmann distribution of the form $e^{-E_i/T}$, which gives a set of probability mass functions along every point $i$ in the solution space. Using a Markov chain sampling, these probabilities are sampled such that each point has a lower temperature then the previous one. As the system is cooled from high to low temperatures ($T \rightarrow 0$), the samples converge to a minimum of $E$, which in many cases might be the global minimum[62]. A frequently used Markov chain sampling technique is Metropolis-Hastings

algorithm in which a 'bad' move $E_2$ from initial state $E_1$ such that $E_2 > E_1$, is accepted if $R(0,1) \geq p_2/p_1$, where $R(0,1)$ is a uniformly random number between 0 and 1.

In our implementation, each iteration consists of a move that may involve multiple synonymous codon substitutions. The algorithm begins at a high temperature where the first move is drastic, synonymous substitutions occur in all replaceable codons. At the end of the first iteration, a new sequence is accepted if the opening energy is smaller than that of the input sequence. However, if the opening energy of a new sequence is greater than that of the input sequence, acceptance depends on the Metropolis-Hastings criteria. The accepted sequence is used for the next iteration, which repeats the above process. As the temperature cools, the moves get milder with fewer synonymous codon changes (Supplementary Fig S3). Simulated annealing stops upon reaching a near optimum solution.

For the web version of TIsigner, the default number of replaceable codons is restricted to the first nine codons. However, this default setting can be reset to range from the first four to nine codons, or the full length of the coding sequence. Furthermore, TIsigner runs multiple simulated annealing instances, in parallel, to obtain multiple possible sequence solutions. There is a possibility to select tunable expression levels when the T7lac promoter is selected (as the expression scores were calculated based on the PSI:Biology dataset; see below). Among the solutions, the sequence that matches most closely to the users' selected target expression score is chosen as the optimum. The option for tunable expression is not available for custom UTRs, the sequence with minimum opening energy is chosen as the optimum.

We allow users to select desirable target expression scores for the experiments using the T7lac inducible promoter. To implement this criterion, the posterior probabilities of success for input and optimised sequences are evaluated using the following equations from Bayesian statistics:

$$positive\ posterior\ odds = prior\ odds \times fitted\ positive\ likelihood\ ratio \quad (1)$$

$$positive\ posterior\ probability = \frac{positive\ posterior\ odds}{(1 + positive\ posterior\ odds)} \quad (2)$$

The fitted positive likelihood ratios in equation (1) were obtained from the following 4-parametric logistic regression equation:

$$fitted\ positive\ likelihood\ ratio = d + \frac{a-d}{1+\left(\frac{positive\ likelihood\ ratio}{c}\right)^b} \quad (3)$$

with parameters a, b, c, and d. The prior probability was set to 0.49, which is the proportion of 'Expressed' (n=21,046) divided by 'Cloned' (n=42,774) of the PSI:Biology targets reported as of 28 June 2017[1]. Posterior probabilities were scaled as percentages to score the input and optimised sequences.

The presence of terminator-like elements[66] in the protein-coding region may result in expression of truncated mRNAs due to early transcription termination. Therefore, we implemented an optional check for putative terminators in the input and optimised

sequences by cmsearch (INFERNAL version 1.1.2)[67] using the covariance models of terminators from RMfam[68,69]. We also allow users to filter the output sequences for the presence of restriction sites. Restriction modification sites (AarI, BsaI, and BsmBI) are avoided by default.

**Sequence optimisation**

We submitted the PSI:Biology targets that failed to be expressed (n=2,650) to the ExpOptimizer webserver from NovoPro Bioscience (https://www.novoprolabs.com/tools/codon-optimization). A total of 2,573 sequences were optimised. The target sequences were also optimised using a local version of COOL[29] and TIsigner using default settings. We also ran a random synonymous codon substitution as a control for these 2,573 sequences.

**Statistical analysis**

AUC and Gini importance scores were calculated using scikit-learn (version 0.20.2)[70]. The 95% confidence intervals for AUC scores were calculated using DeLong's method[71]. Spearman's correlation coefficients and Kolmogorov-Smirnov statistics were calculated using Pandas (version 0.23.4)[72] and scipy (version 1.2.1)[73,74], respectively. Positive likelihood ratios with 95% confidence intervals were calculated using bootLR package[75,76]. The P-values of multiple testing were adjusted using Bonferroni's correction and reported to machine precision. Plots were generated using Matplotlib (version 3.0.2)[77] and Seaborn (version 0.9.0)[78].

**Code and data availability**

Our code and data can be found in our GitHub repository (https://github.com//Gardner-BinfLab/TIsigner_paper_2019). These include the scripts and Jupyter notebooks to reproduce our results and figures. TIsigner is written in Python 3.6 and the source code is available on (https://github.com/Gardner-BinfLab/TIsigner). The public web version of this tool runs at https://tisigner.otago.ac.nz.

**Acknowledgements**

We thank Professor Ivo Hofacker for fruitful discussions at the Benasque RNA Meeting, and Dr Ronny Lorenz for helpful discussions about RNAplfold. We are grateful to Dr Craig van Dolleweerd and members of the Biomolecular Interaction Centre at the University of Canterbury for supporting this research. This work was supported by the Ministry of Business, Innovation and Employment, New Zealand (MBIE grant: UOOX1709).

## References


1. Home : Metrics. *PSI* Available at: http://targetdb.rcsb.org/metrics/. (Accessed: 14th June 2019)
2. Kimelman, A. *et al.* A vast collection of microbial genes that are toxic to bacteria. *Genome Res.* **22**, 802–809 (2012).
3. Berlec, A. & Strukelj, B. Current state and recent advances in biopharmaceutical production in Escherichia coli, yeasts and mammalian cells. *J. Ind. Microbiol. Biotechnol.* **40**, 257–274 (2013).
4. Rosano, G. L. & Ceccarelli, E. A. Recombinant protein expression in Escherichia coli: advances and challenges. *Front. Microbiol.* **5**, 172 (2014).
5. Abreu, R. de S., de Sousa Abreu, R., Penalva, L. O., Marcotte, E. M. & Vogel, C. Global signatures of protein and mRNA expression levels. *Molecular BioSystems* (2009). doi:10.1039/b908315d
6. Hanson, G. & Coller, J. Codon optimality, bias and usage in translation and mRNA decay. *Nat. Rev. Mol. Cell Biol.* **19**, 20–30 (2018).
7. Lim, C. S., Wardell, S. J. T., Kleffmann, T. & Brown, C. M. The exon–intron gene structure upstream of the initiation codon predicts translation efficiency. *Nucleic Acids Research* **46**, 4575–4591 (2018).
8. Stevens, S. G. & Brown, C. M. In silico estimation of translation efficiency in human cell lines: potential evidence for widespread translational control. *PLoS One* **8**, e57625 (2013).
9. Schwanhäusser, B. *et al.* Global quantification of mammalian gene expression control. *Nature* **473**, 337–342 (2011).
10. Bernstein, J. A., Khodursky, A. B., Lin, P.-H., Lin-Chao, S. & Cohen, S. N. Global analysis of mRNA decay and abundance in Escherichia coli at single-gene resolution using two-color fluorescent DNA microarrays. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 9697–9702 (2002).
11. Taniguchi, Y. *et al.* Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science* **329**, 533–538 (2010).
12. Tegel, H., Ottosson, J. & Hober, S. Enhancing the protein production levels in Escherichia coli with a strong promoter. *FEBS J.* **278**, 729–739 (2011).
13. Sharp, P. M. & Li, W. H. The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* **15**, 1281–1295 (1987).
14. Reis, M. d. & d. Reis, M. Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Research* **32**, 5036–5044 (2004).
15. Sabi, R. & Tuller, T. Modelling the Efficiency of Codon–tRNA Interactions Based on Codon Usage Bias. *DNA Research* **21**, 511–526 (2014).
16. Gutman, G. A. & Hatfield, G. W. Nonrandom utilization of codon pairs in Escherichia coli. *Proc. Natl. Acad. Sci. U. S. A.* **86**, 3699–3703 (1989).
17. Brule, C. E. & Grayhack, E. J. Synonymous Codons: Choose Wisely for Expression. *Trends Genet.* **33**, 283–297 (2017).
18. Kudla, G., Murray, A. W., Tollervey, D. & Plotkin, J. B. Coding-sequence determinants of gene expression in Escherichia coli. *Science* **324**, 255–258 (2009).
19. Plotkin, J. B. & Kudla, G. Synonymous but not the same: the causes and consequences of codon bias. *Nature Reviews Genetics* **12**, 32–42 (2011).

20. Boël, G. *et al.* Codon influence on protein expression in E. coli correlates with mRNA levels. *Nature* **529**, 358–363 (2016).
21. Cambray, G., Guimaraes, J. C. & Arkin, A. P. Evaluation of 244,000 synthetic sequences reveals design principles to optimize translation in Escherichia coli. *Nat. Biotechnol.* **36**, 1005–1015 (2018).
22. de Smit, M. H. & van Duin, J. Secondary structure of the ribosome binding site determines translational efficiency: a quantitative analysis. *Proc. Natl. Acad. Sci. U. S. A.* **87**, 7668–7672 (1990).
23. Dvir, S. *et al.* Deciphering the rules by which 5'-UTR sequences affect protein expression in yeast. *Proc. Natl. Acad. Sci. U. S. A.* **110**, E2792–801 (2013).
24. Tuller, T. & Zur, H. Multiple roles of the coding sequence 5′ end in gene expression regulation. *Nucleic Acids Research* **43**, 13–28 (2015).
25. Umu, S. U., Poole, A. M., Dobson, R. C. & Gardner, P. P. Avoidance of stochastic RNA interactions can be harnessed to control protein expression levels in bacteria and archaea. *Elife* **5**, (2016).
26. Villalobos, A., Ness, J. E., Gustafsson, C., Minshull, J. & Govindarajan, S. Gene Designer: a synthetic biology tool for constructing artificial DNA segments. *BMC Bioinformatics* **7**, 285 (2006).
27. Salis, H. M., Mirsky, E. A. & Voigt, C. A. Automated design of synthetic ribosome binding sites to control protein expression. *Nature Biotechnology* **27**, 946–950 (2009).
28. Raab, D., Graf, M., Notka, F., Schödl, T. & Wagner, R. The GeneOptimizer Algorithm: using a sliding window approach to cope with the vast sequence space in multiparameter DNA sequence optimization. *Syst. Synth. Biol.* **4**, 215–225 (2010).
29. Chung, B. K.-S. & Lee, D.-Y. Computational codon optimization of synthetic gene for protein expression. *BMC Syst. Biol.* **6**, 134 (2012).
30. Terai, G., Kamegai, S. & Asai, K. CDSfold: an algorithm for designing a protein-coding sequence with the most stable secondary structure. *Bioinformatics* **32**, 828–834 (2016).
31. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
32. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
33. Noderer, W. L. *et al.* Quantitative analysis of mammalian translation initiation sites by FACS-seq. *Mol. Syst. Biol.* **10**, 748 (2014).
34. 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.* **71**, 1342–1346 (1974).
35. Chen, L., Oughtred, R., Berman, H. M. & Westbrook, J. TargetDB: a target registration database for structural genomics projects. *Bioinformatics* **20**, 2860–2862 (2004).
36. Seiler, C. Y. *et al.* DNASU plasmid and PSI:Biology-Materials repositories: resources to accelerate biological research. *Nucleic Acids Res.* **42**, D1253–60 (2014).
37. Acton, T. B. *et al.* Robotic cloning and Protein Production Platform of the Northeast Structural Genomics Consortium. *Methods Enzymol.* **394**, 210–243 (2005).
38. Wang, M., Herrmann, C. J., Simonovic, M., Szklarczyk, D. & von Mering, C. Version 4.0 of PaxDb: Protein abundance data, integrated across model organisms, tissues, and cell-lines. *Proteomics* **15**, 3163–3168 (2015).
39. Tabb, D. L. *et al.* Repeatability and reproducibility in proteomic identifications by liquid

chromatography- tandem mass spectrometry. *J. Proteome Res.* **9**, 761–776 (2009).
40. Nilsson, T. *et al.* Mass spectrometry in high-throughput proteomics: ready for the big time. *Nat. Methods* **7**, 681–685 (2010).
41. Deuschle, U., Kammerer, W., Gentz, R. & Bujard, H. Promoters of Escherichia coli: a hierarchy of in vivo strength indicates alternate structures. *EMBO J.* **5**, 2987–2994 (1986).
42. Delvigne, F. *et al.* Taking control over microbial populations: Current approaches for exploiting biological noise in bioprocesses. *Biotechnol. J.* **12**, (2017).
43. Tuller, T., Waldman, Y. Y., Kupiec, M. & Ruppin, E. Translation efficiency is determined by both codon bias and folding energy. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 3645–3650 (2010).
44. Ang, K. S., Kyriakopoulos, S., Li, W. & Lee, D.-Y. Multi-omics data driven analysis establishes reference codon biases for synthetic gene design in microbial and mammalian cells. *Methods* **102**, 26–35 (2016).
45. Tunney, R. *et al.* Accurate design of translational output by a neural network model of ribosome distribution. *Nat. Struct. Mol. Biol.* **25**, 577–582 (2018).
46. Ben-Yehezkel, T. *et al.* Rationally designed, heterologous S. cerevisiae transcripts expose novel expression determinants. *RNA Biol.* **12**, 972–984 (2015).
47. Pelletier, J. & Sonenberg, N. The involvement of mRNA secondary structure in protein synthesis. *Biochem. Cell Biol.* **65**, 576–581 (1987).
48. Bhattacharyya, S. *et al.* Accessibility of the Shine-Dalgarno Sequence Dictates N-Terminal Codon Bias in E. coli. *Mol. Cell* **70**, 894–905.e5 (2018).
49. Nieuwkoop, T., Claassens, N. J. & van der Oost, J. Improved protein production and codon optimization analyses in Escherichia coli by bicistronic design. *Microb. Biotechnol.* **12**, 173–179 (2019).
50. Mückstein, U. *et al.* Thermodynamics of RNA–RNA binding. *Bioinformatics* **22**, 1177–1182 (2006).
51. Bernhart, S. H., Mückstein, U. & Hofacker, I. L. RNA Accessibility in cubic time. *Algorithms Mol. Biol.* **6**, 3 (2011).
52. Lorenz, R. *et al.* ViennaRNA Package 2.0. *Algorithms Mol. Biol.* **6**, 26 (2011).
53. Mann, M., Wright, P. R. & Backofen, R. IntaRNA 2.0: enhanced and customizable prediction of RNA–RNA interactions. *Nucleic Acids Res.* **45**, W435–W439 (2017).
54. Sambrook, J. & Russell, D. W. *Molecular cloning: a laboratory manual. Vol. 3*. (CSHL Press, 2001).
55. Hofacker, I. L. *et al.* Fast folding and comparison of RNA secondary structures. *Monatshefte für Chemie / Chemical Monthly* **125**, 167–188 (1994).
56. Bernhart, S., Hofacker, I. L. & Stadler, P. F. Local Base Pairing Probabilities in Large RNAs. *Bioinformatics*
57. Bompfünewerer, A. F. *et al.* Variations on RNA folding and alignment: lessons from Benasque. *J. Math. Biol.* **56**, 129–144 (2008).
58. Lorenz, R., Hofacker, I. L. & Stadler, P. F. RNA folding with hard and soft constraints. *Algorithms Mol. Biol.* **11**, 8 (2016).
59. Mohammad, F., Green, R. & Buskirk, A. R. A systematically-revised ribosome profiling method for bacteria reveals pauses at single-codon resolution. *Elife* **8**, (2019).
60. Kirkpatrick, S., Gelatt, C. D. & Vecchi, M. P. Optimization by Simulated Annealing. *Science* **220**, 671–680 (1983).

61. Ingber, L. Adaptive simulated annealing (ASA): Lessons learned. (2000).
62. Keith, J. M. *et al.* A simulated annealing algorithm for finding consensus sequences. *Bioinformatics* **18**, 1494–1499 (2002).
63. Brownlee, J. *Clever Algorithms: Nature-inspired Programming Recipes*. (Jason Brownlee, 2011).
64. Lindgreen, S., Gardner, P. P. & Krogh, A. MASTR: multiple alignment and structure prediction of non-coding RNAs using simulated annealing. *Bioinformatics* **23**, 3304–3311 (2007).
65. Gaspar, P., Moura, G., Santos, M. A. S. & Oliveira, J. L. mRNA secondary structure optimization using a correlated stem-loop prediction. *Nucleic Acids Res.* **41**, e73 (2013).
66. Chen, Y.-J. *et al.* Characterization of 582 natural and synthetic terminators and quantification of their design constraints. *Nat. Methods* **10**, 659–664 (2013).
67. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935 (2013).
68. Gardner, P. P. & Eldai, H. Annotating RNA motifs in sequences and alignments. *Nucleic Acids Res.* **43**, 691–698 (2015).
69. Kalvari, I. *et al.* Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Res.* **46**, D335–D342 (2018).
70. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
71. DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **44**, 837–845 (1988).
72. McKinney, W. Data Structures for Statistical Computing in Python. in *Proceedings of the 9th Python in Science Conference* 51–56 (2010).
73. Oliphant, T. E. Python for Scientific Computing. *Computing in Science Engineering* **9**, 10–20 (2007).
74. Millman, K. J. & Aivazis, M. Python for Scientists and Engineers. *Computing in Science Engineering* **13**, 9–12 (2011).
75. Marill, K. A., Chang, Y., Wong, K. F. & Friedman, A. B. Estimating negative likelihood ratio confidence when test sensitivity is 100%: A bootstrapping approach. *Stat. Methods Med. Res.* **26**, 1936–1948 (2017).
76. R Core Team. *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, 2019).
77. Matplotlib: A 2D Graphics Environment - IEEE Journals & Magazine. Available at: https://doi.org/10.1109/MCSE.2007.55. (Accessed: 17th June 2019)
78. Waskom, M. *et al.* mwaskom/seaborn: v0.9.0 (July 2018). (2018). doi:10.5281/zenodo.1313201

## Figures

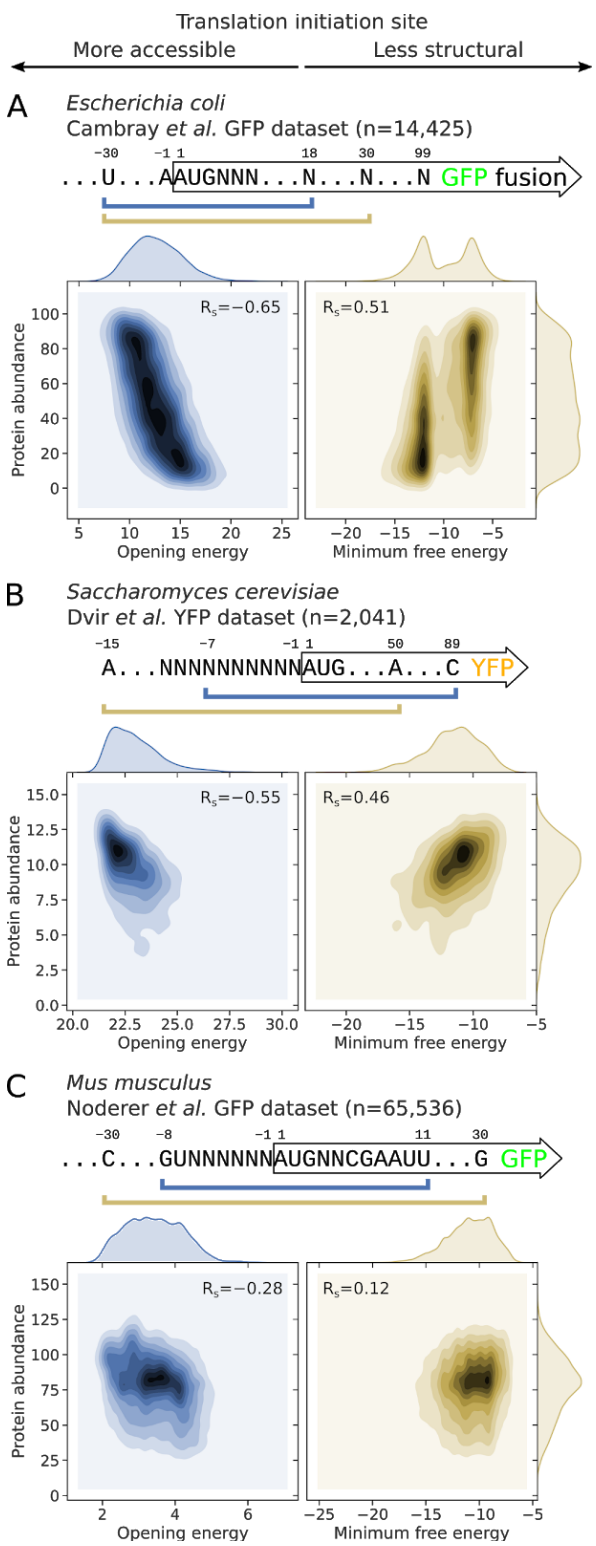


**Fig 1. Correlations between the opening energy of translation initiation sites and protein abundance are stronger than that of minimum free energy. (A)** For *E. coli*, the opening energy at the region −30:18 shows the strongest correlation with protein abundance (also see Fig 2B or Supplementary Fig S1A, sub-sequence l=48 at position i=18). For this analysis, we used a representative GFP expression dataset from Cambray et al. (2018). The reporter library consists of GFP fused in-frame with a library of 96-nt upstream sequences (n=14,425). The minimum free energy −30:30 shown was determined by Cambray et al. (right panel). **(B)** For *S. cerevisiae*, the opening energy −7:89 shows the strongest correlation with protein abundance (also see Supplementary Fig S1B, sub-sequence l=96 at position i= 89). For this analysis, we used the YFP expression dataset from Dvir et al. (2013). The YFP reporter library consists of 2,041 random decameric nucleotides inserted at the upstream of YFP start codon. The minimum free energy −15:50 was previously shown to correlate the best with protein abundance (right panel). **(C)** For *M. musculus*, the opening energy −8:11 shows the strongest correlation with protein abundance (also see Supplementary Fig S1C, sub-sequence l=19 at position i=11). For this analysis, we used the GFP expression dataset from Noderer et al. (2014). The GFP reporter library consists of 65,536 random hexameric and dimeric nucleotides inserted at the upstream and downstream of GFP start codon, respectively. The minimum free energy −30:30 was shown (right panel). $R_s$, Spearman's rho. Bonferroni adjusted P-values are statistically significant ($<2.2\times10^{-16}$) for the correlations between opening energy and protein abundance shown in the left panels.

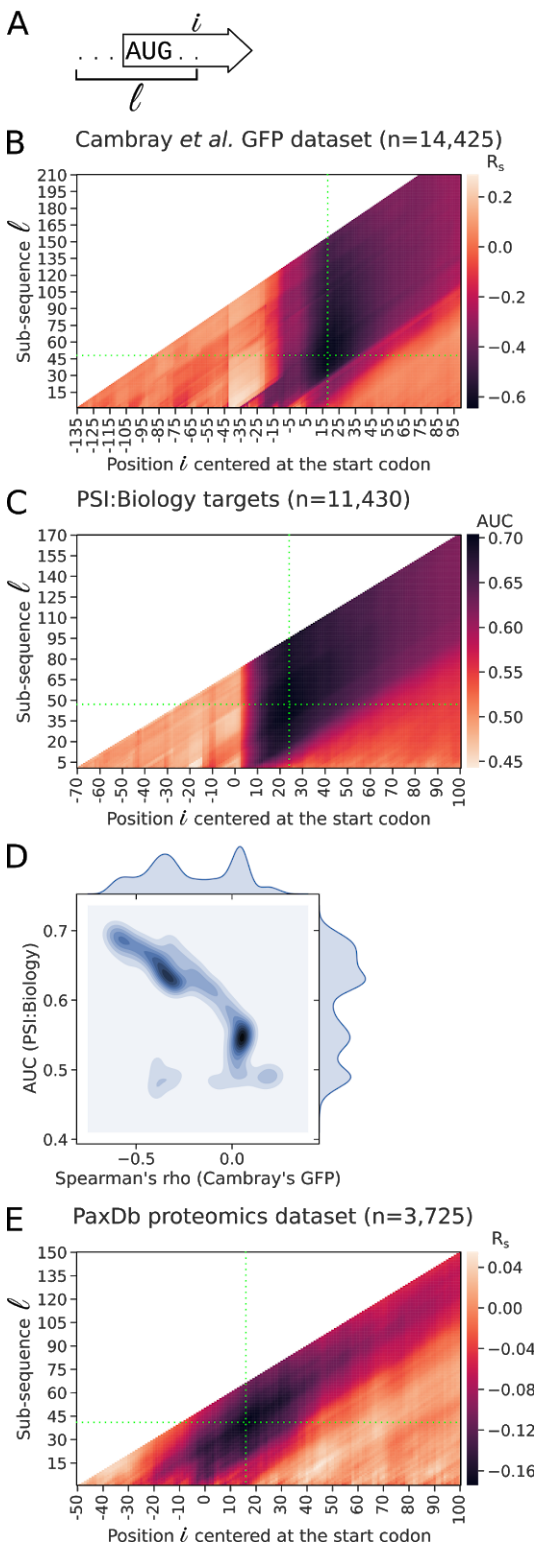


**Fig 2. Strong correlations between the opening energy and protein abundance are predictive of the outcomes of recombinant protein expression in *E. coli*. (A)** Schematic representation of a transcript sub-sequence l at position i for the calculation of opening energy. For example, sub-sequence l=10 at position i=10 corresponds to the region 1:10. **(B)** Correlation between the opening energy for the sub-sequences of GFP transcripts and protein abundance. The opening energy at the region −30 to 18 nt (sub-sequence l=48 at position i=18, green crosshair) shows the strongest correlation with protein abundance [$R_s$=−0.65; n=14,425, GFP expression dataset of Cambray et al. (2018)]. For this dataset, the reporter plasmid used is pGC4750, in which the promoter and ribosomal binding site are oFAB1806 inducible promoter and oFAB1173/BCD7, respectively. **(C)** Prediction accuracy of the expression outcomes of the PSI:Biology targets using opening energy (n=11,430). The opening energy at the region −23:24 (sub-sequence l=47 at position i=24, green crosshair) shows the highest prediction accuracy score (AUC=0.70). For this dataset, the expression vector used is pET21_NESG, in which the promoter and fusion tag are T7lac and C-terminal His tag, respectively. **(D)** Comparison between the correlations and AUC scores by sub-sequence region taken from the above analyses. The sub-sequence regions that have strong correlations are likely to have high AUC scores, whereas the sub-sequence regions that have no correlations are likely not useful in prediction of the expression outcome. **(E)** Correlation between the opening energy for the sub-sequences of *E. coli* transcripts and protein abundance. The transcripts used for this analysis are protein-coding sequences concatenated with 50 and 10 nt located upstream and downstream, respectively. The opening energy at the region −25:16 (sub-sequence l=41 at position i=16, green crosshair) shows the strongest correlation with protein abundance ($R_s$=−0.17; n=3,725, PaxDb integrated proteomics dataset). $R_s$, Spearman's rho.

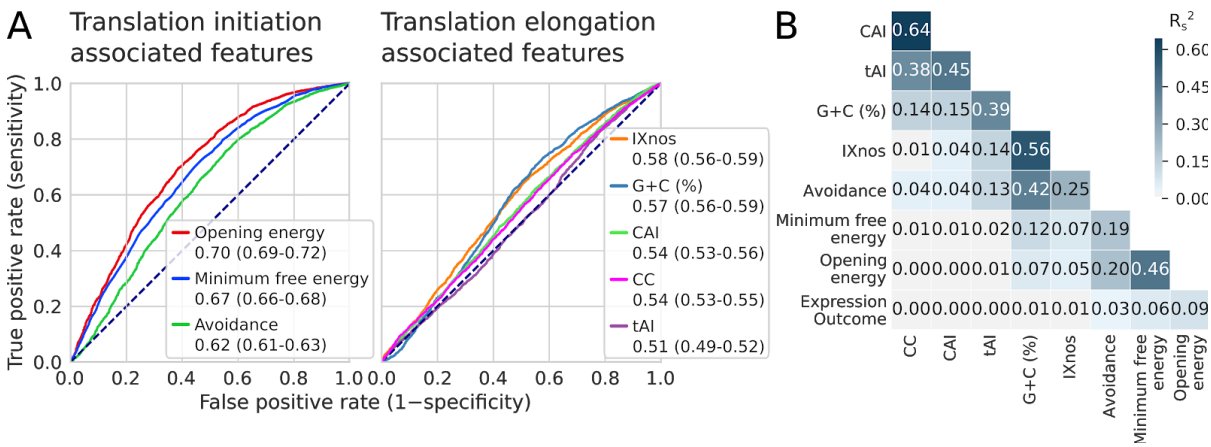


**Fig 3. Accessibility is a strong predictor of heterologous protein expression.** (**A)** ROC analysis for prediction of the expression outcomes of the PSI:Biology targets (n=8,780 and 2,650, 'success' and 'failure' groups, respectively). The features associated with translation initiation rate analysed are the opening energy −24:24, minimum free energy −30:30 and avoidance 1:30 (left panel). The feature associated with translation elongation rate are tRNA adaptation index (tAI), codon context (CC), codon adaptation index (CAI), G+C content (%) and Iχnos (right panel). The Iχnos scores are translation elongation rates predicted using a neural network model trained with ribosome profiling data. The AUC scores with 95% confidence intervals are shown. **(B)** Relationships between the features and expression outcome represented as squared Spearman's correlations ($R_s^2$). The opening energy −24:24 is the best feature in explaining the expression outcome.

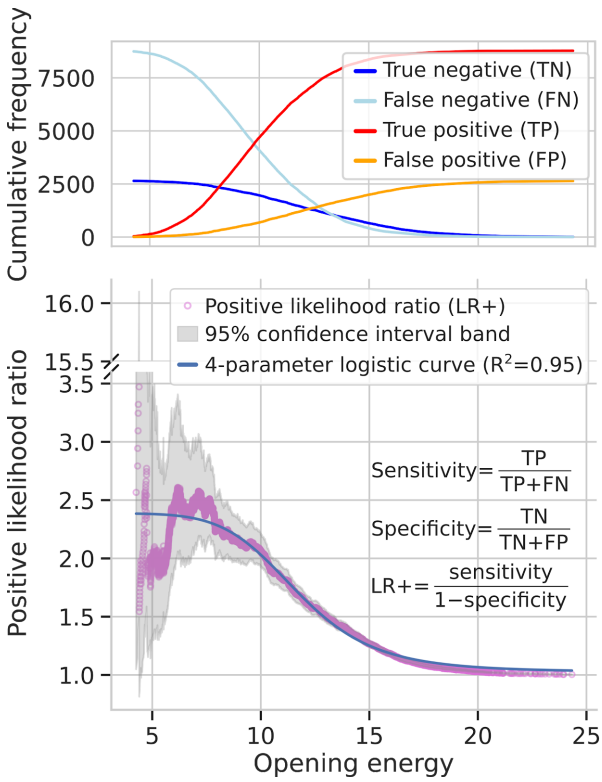


**Fig 4. Opening energy of 10 or below at the region −24:24 is about two times more likely to come from the target genes that are successfully expressed than those that failed (with 95% confidence).** Cumulative frequency distributions of the true positive and false positive (less than type), and true negative and false negative (more than type) derived from the ROC analysis in Fig 2A (left panel, opening energy −24:24). These values were used to estimate positive likelihood ratios with 95% confidence intervals using 10,000 bootstrap replicates. The estimated ratios and/or confidence intervals are inaccurate at low numbers of true positives or true negatives. Therefore, a four-parameter logistic curve was fitted to the positive likelihood ratios. Fitted values are useful to estimate the posterior probability of protein expression.

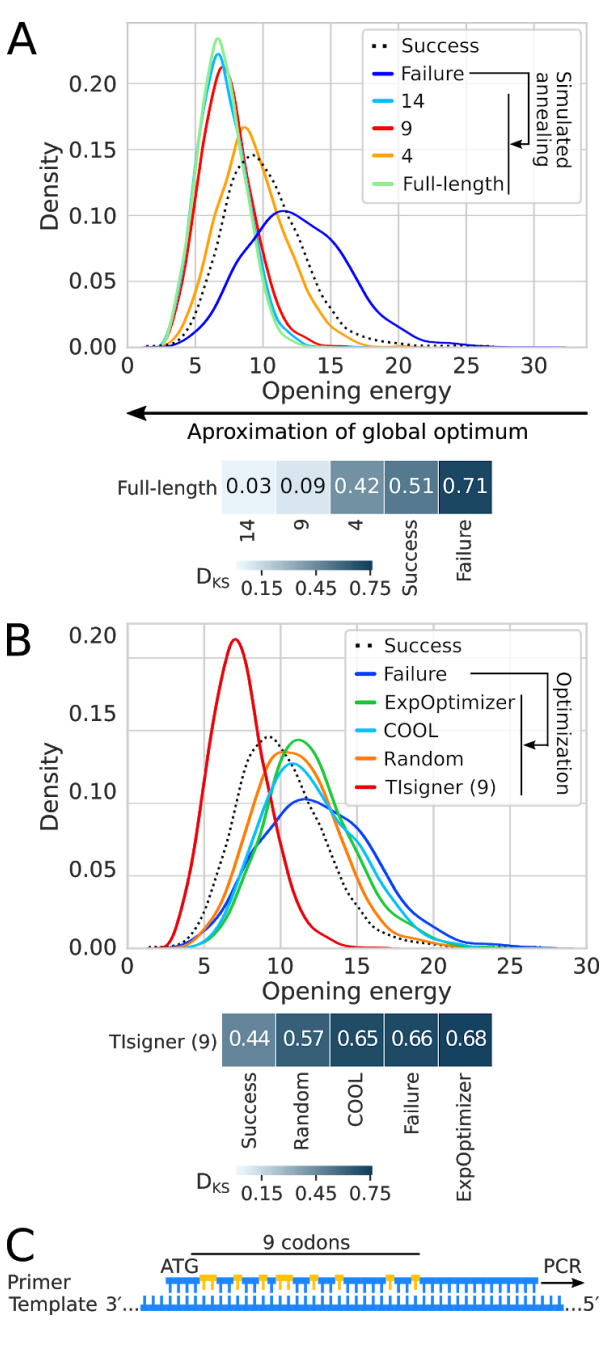


**Fig 5. Accessibility of translation initiation sites can be increased by synonymous codon substitution within the first nine codons using simulated annealing. (A)** Accessibility of translation initiation sites increases with increasing number of the first N replaceable codons. The PSI:Biology targets that failed to be expressed were optimised using simulated annealing (n=2,650). The Kolmogorov-Smirnov distance between the distributions of '9' and 'full-length' was significantly different but sufficiently close ($D_{KS}$=0.09, $P<10^{-7}$), indicating that optimisation of the first nine codons can achieve nearly optimum accessibility. For comparison, the distribution of the PSI:Biology targets that were successfully expressed are shown (n=8,780). **(B)** Accessibility of translation initiation sites can be increased indirectly using the existing gene optimisation tools and random synonymous codon substitution. 'TIsigner (9)' refers to the default settings of our tool, which allows synonymous substitutions up to the first nine codons (as above). **(C)** Accessibility of translation initiation sites can be optimised using PCR cloning. The forward primer should be designed according to TIsiger optimised sequences. For example, using a nested PCR approach, the optimised sequence can be produced using the forward primer designed with appropriate mismatches (gold bulges) to amplify the amplicon from the initial PCR reaction.