Exploring codon context bias for designing a synthetic thermostable invertase gene in *Escherichia coli*

Han Bin Pek[1#], Maximilian Klement[1,2#], Kok Siong Ang[2#], Bevan K.S. Chung[1], Dave Siak-Wei Ow[1*], Dong-Yup Lee[1,2*]

[1]*Bioprocessing Technology Institute, A*STAR (Agency for Science, Technology and Research), 20 Biopolis Way, #06-01 Centros, Singapore 138668*

[2]*Department of Chemical and Biomolecular Engineering, National University of Singapore, 4 Engineering Drive 4, Singapore 117576*

[#]These authors contributed equally to this work.

*Authors to whom correspondence should be addressed.

Dave Siak-Wei Ow - Phone: +65-6407-0843. Fax: +65-6478-9561. E-mail: dave_ow@bti.a-star.edu.sg

Dong-Yup Lee - Phone: +65-6516-6907. Fax: +65-6779-1936. E-mail: cheld@nus.edu.sg

Abstract

Background: Successful synthetic gene design for recombinant protein expression can be done by matching its translational elongation rate against heterologous host organisms via codon optimization. Amongst the various design parameters considered for gene synthesis, codon context bias has been relatively overlooked compared with individual codon usage which is commonly adopted in most of codon optimization tools. In addition, matching the rates of transcription and translation may lead to enhanced protein folding.

Results: In this study, we evaluated codon context fitness as design criterion for enhancing the expression of thermostable invertase in *Escherichia coli* and explored the relevance of secondary structure regions for folding and expression. We designed three coding sequences by (1) a commercial vendor strategy, (2) codon context for the whole gene, and (3) codon context based on the secondary structure region. Then, the codon optimized sequences were transformed and expressed in *E. coli*. From the resultant enzyme activities measured, codon context fitness proved to have the highest activity as compared to the wild-type control and other criteria.

Conclusions: Codon context was shown to be a more relevant parameter for protein expression in *Escherichia coli* by codon optimization. Thus, we can effectively design synthetic genes within heterologous host organisms.

Keywords: codon optimization; codon context; secondary structure; invertase; gene design; synthetic biology.

1. Background

All amino acids, except methionine and tryptophan, are encoded by two to six synonymous codons. Most notably, the synonymous codons are not equally utilized across different organisms, resulting in the phenomenon of codon usage bias [1, 2]. Furthermore, it was later discovered that such biases are correlated with gene expression levels [3, 4]. This is likely a consequence of the evolution of organisms to utilize synonymous codons with disparate frequencies as a result of random mutations and selection pressure [5]. Thus, natural and synthetic genes to be transfected into heterologous hosts for their expression usually yields poorly-translated proteins. In order to improve such heterologous gene expression, it is required to design customized artificial genes, which can be achieved by matching the codon preference of the host based on several design parameters adjusted for codon optimization [6-11].

The most commonly adopted design parameter is individual codon usage (ICU) which is a result of synonymous codons not being equally utilized across different organisms [12, 13]. In addition to ICU bias, non-random utilization of adjacent codon pairs in organisms has also been reported in several studies [14-16]. This phenomenon is termed codon context, and describes the organization of adjoining codons as a result of potential tRNA-tRNA steric interaction within the ribosomes [10, 17]. Indeed, codon context was shown to correlate with translation elongation rate such that the usage of rare codon pairs decreased protein translation rates [18-21]. Moreover, it is known that proteins fold via particular pathways, influenced by the different regional rates of the secondary structures by which the polypeptides emerge unidirectionally from the ribosome [22]. Thus, variations in elongation rate can considerably influence the folding of the encoded polypeptide [23]. Recently, it was also proposed that these variations in protein elongation rates have evolved to precisely orchestrate the emergence rates of each structural region of the polypeptide for folding or interacting with molecular chaperones [19, 20]. Collectively, a number of studies have reported different codon frequencies within specific structural regions of the encoded protein [23]. Therefore, codon optimization based on the secondary structure regions may enable us to accurately match gene design to the native process of protein translation and folding [13].

In this study, we demonstrated the significance of codon context and investigated the relevance of the translational rate within each secondary structure region for folding and

expression of a thermostable invertase in *E. coli*. Using the wild type (WT) *Thermotoga maritima* invertase sequence as a control, we evaluated the protein activity levels of three coding sequences designed based on (1) a commercial vendor (CV) strategy, (2) codon context (CC) fitness for the whole gene [24], and (3) codon context fitness based on the secondary structure (SS) region [23].

2. Methods

2.1. Implementation of codon optimization frameworks

We conducted codon optimization based on CV, CC, and SS. CV design was performed by a commercial vendor, considering several parameters including codon usage, codon context, RNA secondary structure, and codon-anticodon pair interactions according to their claim, while CC- and SS-based sequences were designed by our previously developed algorithm [24]. As described in the previous work, the CC optimization problem is a combinatorial problem with a nonlinear objective function that is written as:

$$\Psi_{CC} = -\frac{\sum_{k=1}^{3904} \left| q_0^k - q_1^k \right|}{3904}$$

where q^k denotes the occurrence frequency of each possible codon pair (64×61 = 3904), and the subscript denotes the organism (0 for host and 1 for sequence). For an average sized protein of 300 amino acids, the number of possible codon coding sequences is in the order of 10^{100} [25]. Therefore, the genetic algorithm was employed to obtain a good feasible solution within an acceptable amount of time [26].

Basically, the genetic algorithm is a heuristics-based search algorithm that borrows concepts from evolutionary biology (“survival of the fittest”) and genetics (mutation, recombination). The algorithm, like other evolutionary algorithms, utilizes a population of solution candidates (top 5% expressing genes in genome) which are encoded as strings called “chromosomes” with an associated objective function value (“fitness”). The algorithm then iteratively searches the solution space by modifying the population of candidates. The search procedure is described as follows:

1. Randomly initialize a population of candidate solutions within the search space.
2. Evaluate the fitness of each candidate.
3. Rank the candidates according to fitness and check for search termination criterion.
4. If termination criterion is not satisfied, discard the bottom 50% population and use the remaining (i.e. top 50%) to generate new candidates.
5. The new candidates are generated from the fittest by the recombination and mutation operators, and then used to refill the population of candidates.
6. The process from 2 to 5 repeats until the termination criterion is met.

The search can be terminated when the best solution obtained exceeds a targeted value, or when the number of iterations exceeds a pre-set limit.

For the codon optimization, it is natural to use the RNA coding sequence as the “chromosome”, where each character triplet (codon) encodes an amino acid. The recombination and mutation operators operate on the codons with reference to the standard codon translation table, ensuring that the encoded protein remains unchanged. In this work, the search was set to terminate if the change in fitness was less than 0.5% for across 100 iterations, or 10000 iterations had passed.

Similarly, the procedure for SS design follows CC-based codon optimization algorithm, with modified target sequences on the basis of secondary structures of genes considered. Given the secondary structure information, we separately take codon context of the sequences corresponding to alpha helices, beta sheets, and the remaining structural regions. Therefore, the objective function is rewritten as:

$$\Psi_{SS} = -\frac{\sum_{k=1}^{3904}\left|q_{\alpha,0}^{k} - q_{\alpha,1}^{k}\right|}{3904} - \frac{\sum_{k=1}^{3904}\left|q_{\beta,0}^{k} - q_{\beta,1}^{k}\right|}{3904} - \frac{\sum_{k=1}^{3904}\left|q_{X,0}^{k} - q_{X,1}^{k}\right|}{3904}$$

where q^k denotes the occurrence frequency of each possible codon pair (64×61 = 3904). The subscripts, from left to right, denote secondary structure type (α for alpha helices, β for beta sheets, and X for other structures) and organism (0 for host and 1 for sequence). The computed score for each secondary structure is normalized by the number of possible codon pairs. The objective function value is negative as we wish to minimize the codon context difference between the expression host and the heterologous protein sequence.

2.2. Gene synthesis and cloning

Three variants of the wild-type invertase from *Thermotoga maritima* enzyme (UniProt ID: O33833.1) were designed with respect to CV, CC, and SS optimization based on the highly expressed genes found in *Escherichia coli*. CV was provided by commercial vendor while CC and SS were designed based on the codon context fitness of the sequences of whole gene and each secondary region, respectively, as described in Section 2.1. The genes were then synthesized by a commercial vendor, amplified via PCR with the respective primer pairs (Table 1), and the resulting bands were spin-column purified (Qiagen). pET28a (Novagen) was used as the expression plasmid vector, and was linearized with NcoI and XhoI. The

linearized vector was gel-purified. The In-Fusion HD cloning kit (Clontech) was used for insertion of the various invertase sequences into the linearized vector, as per manufacturer's protocol. Recombinant pET28a-invertase plasmids were sequenced (Axil Scientific) to ensure no mutations were present.

2.3. Expression and isolation of invertase in *E. coli*

The various pET28a-invertase recombinant plasmids were transformed into *E. coli* BL21 (DE3) (Invitrogen) as per manufacturer's protocol. A single colony was picked and incubated overnight at 37 ^{o}C with shaking in 5 ml of 2YT media supplemented with glucose (per litre: 16 g tryptone; 10 g yeast extract; 5 g NaCl; 1 g glucose), and containing kanamycin at a concentration of 50 µg/ml. The overnight culture was used to inoculate a larger culture volume of 2YT/kanamycin to an OD_{600} of 0.05, and incubated at 37 ^{o}C with shaking. Upon reaching an OD_{600} of 0.5, IPTG was added to a final concentration of 0.5 mM to induce invertase expression. Protein induction was carried out for 5 hours before harvesting. One OD_{600} of cells were collected for subsequent analysis.

The harvested cells were centrifuged for 5 minutes at 12,000g, and the resulting cell pellet was washed with 0.5 ml of LEW buffer (per litre: 50 mM NaH_2PO_4; 300 mM NaCl; pH 8.0). The cell pellet was resuspended in 0.5 ml of LEW buffer, and sonicated six times for 25s each, at a power output of 4 W. The suspension was centrifuged at 12,000g for 20 minutes at 4 ^{o}C. The supernatant (soluble fraction) was collected, and the pellet was washed with 0.5ml of LEW buffer. The washed pellet was then thoroughly resuspended in 0.5 ml of LEW buffer containing 8 M urea, before being centrifuged at 12,000g for 20 minutes at 4 ^{o}C. The supernatant was also collected (insoluble fraction) for further analysis.

2.4. Gel electrophoresis and Western blot analysis

The soluble and insoluble fractions were run on a 4-12% reducing gradient gel (Invitrogen), and transferred to a nitrocellulose membrane was achieved with a Trans-Blot SD Semi-Dry Electrophoretic Transfer Cell (Bio-Rad). The membrane was probed with anti-His-HRP antibodies (Novagen). All protocols were performed as per manufacturers' protocols.

2.5. Enzyme activity assay for invertase

The protocol for determining the enzyme activity of the expressed recombinant invertase was adapted from Liebl *et. al*. [27]. Briefly, 1 µg of protein (as determined by Bradford assay) in 40 µl of either LEW buffer (soluble fraction) or 8 M urea (insoluble fraction) was added to 0.5 ml of Reaction buffer (per litre: 50 mM sodium acetate; 120 mM sucrose, pH 5.5, adjusted at 75 ^{o}C), and the solution was incubated at 75 ^{o}C for 15 minutes in a heat block with shaking. The mixture was cooled to 4 ^{o}C on ice, and 750 µl of Stop solution (per 100 ml: 1 g 3,5-dinitrosalicyclic acid; 20 ml 2 N NaOH; 30 g Rochelle salt; 0.05 g sodium sulphite) was added. The mixture was further incubated at 95 ^{o}C for 15 minutes in a heat block with shaking, after which it was cooled immediately to 4 ^{o}C on ice and centrifuged briefly. The absorbance at 575 nm was measured to determine the amount of invertase activity. The volume of crude protein samples used was proportional to the volume of culture, factoring in the metabolic burden of protein expression as seen by the differing biomass yield.

3. Results and Discussion

3.1. Codon context fitness to improve protein expression in heterologous hosts

Among several design criteria for codon optimization, recently codon context-based fitness has been shown to significantly improve protein expression in yeast and mammalian systems [16, 24]. Therefore, we have applied same criterion to the microbial production system *E. coli*, in order to evaluate its efficacy and relevance. Arguably, the protein translation rates should match the protein folding rates, especially when molecular chaperones are not co-expressed [28]. As such, we may increase the fraction of soluble proteins and at the same time reduce the formation of inclusion bodies. Thus, the relative importance of protein secondary structure for designing sequences of high expression protein can be explored by implementing an SS based strategy (Methods); the codon pair fitness of the secondary structure regions of a given gene sequence can be adjusted by matching with those of the host organism (Figure 1). To do so, each type of secondary structures, i.e., alpha helices, beta sheets, and others remaining regions were extracted from host genome information based on structural data in the Protein Data Bank; accordingly their codon pair patterns were established as reference inputs to the algorithm [29]. Similarly, secondary structure regions of the target gene were collected based on the structural model and matched to the host reference inputs by the codon optimization, resulting in SS-based sequence. The current strategy was applied to invertase from *Thermotoga maritima*, which has extensive amounts of beta sheets in its secondary structure.

3.2. Sequence alignment and protein expression

Invertase is an enzyme used in the food industry to break down sucrose into glucose and fructose. The wild-type invertase sequence derived from the hyperthermophilic *Thermotoga maritima* is characterized by a bimodular arrangement with extensive beta sheets [30]. Initially, the enzyme sequence (WT) was codon optimized for expression in *E. coli* using three different criteria: (1) CV, (2) CC, and (3) SS. It should be noted that the resulting amino acid sequences of the various codon optimized invertase' were identical to the native enzyme. The difference among the nucleotide sequences (WT, CV, CC and SS) are highlighted in Supp. Figure 1.

Subsequently, the WT, CV, CC, and SS sequences were transformed and expressed in *E. coli* shake flask culture, and the optical density at 600nm (OD_{600}) was measured at regular

intervals. Figure 2A illustrates that cells expressing the codon-optimized invertase (CV, CC, and SS) have similar growth curves, with WT having a slightly lower final biomass. Cells expressing the WT coding sequence exhibited a shorter exponential phase, with a subsequent linear growth rate, reaching a lower final OD_{600} as compared to the other coding sequences. This is most likely explained by the higher metabolic burden imposed on the cell from the non-codon-optimized invertase, which may deplete cellular pools of rare tRNA [12].

Invertase was expressed in the cytoplasm and extracted by cell lysis using sonication. The crude samples were separated into soluble and insoluble fractions, and examined for protein expression by Western blot (Figure 2B). A no vector (BL21) and no insert (pET28a) control were run to confirm that the protein is only expressed in the engineered vector upon induction. A clear visible band representing invertase (~50 kDa) was seen on the Western blot. Visual comparison of the soluble protein bands between the different samples revealed a stronger signal for CC and SS based fitness, both of which were optimized based on the codon pair frequency, as compared to WT and CV [24]. Here we assume that SS based fitness may have improved activity since the adjustment of codon context based on the secondary structure should result in a higher percentage of folded proteins.

3.3. Comparison of codon-optimized invertase using an enzyme activity assay

We used an enzyme activity assay which exploited invertase's ability to hydrolyse sucrose to form glucose and fructose. Glucose further undergoes a reaction where it reduces 3,5-dinitrosalicylic acid to 3-amino,5-nitrosalicylic acid, the presence of which can be detected by measuring the absorbance at 575 nm. Thus, based on the stoichiometric relationship, the amount of absorbance of 3-amino,5-nitrosalicylic acid is directly proportional to the level of invertase activity present in the sample (Figure 3A). Absorbance at 575 nm (after correcting for background) was plotted in Figure 3B, and the WT and CV measurements were seen to be similar, indicating that no significant improvement was observed when the gene was optimized based on the CV. Expectedly, the CC optimized invertase clearly showed an increased activity of 39.8% and 41.8% greater than the WT and CV, respectively. However, SS optimized invertase, which was assumed to have improved activity compared to the CC optimized invertase, was seen to have the lowest activity ($p<0.05$).

To interpret the enzyme activity result (Figure 3B), CV, CC, and SS sequences were mapped into the optimized codon sequences plot in terms of both individual codon usage and codon context scores in Figure 4. It should be noted that the solutions on the Pareto front (black circles), generated by the multi-objective optimization algorithm, are not dominated by any other solutions. One of the solutions is the CC sequence which was obtained by optimizing the single objective of codon context, and therefore has the best fitness with respect to this design parameter. From the plot, it is clear that the CV sequence is positioned relatively far from the Pareto front, indicating that it was not optimized with respect to either individual codon usage or codon context. This was not surprising as the sequence was designed by the vendor based on multiple design parameters. Similarly, the SS sequence has a lower codon context fitness compared to the CC sequence, although it is still higher than the CV sequence.

From the Western blot results, the CC and SS sequences showed higher expression levels compared to the CV and WT sequences; this may be correlated with the codon context fitness scores. In addition, the highest enzyme activity of the CC optimized invertase confirmed that codon context is one of most relevant parameters for synthetic gene design in *E. coli* expression system. However, the comparatively lower enzymatic activity measured for the SS sequence raises new question of the validity of secondary structure based design parameter for codon optimization and its relation with protein folding. One possible reason to explain such a lower activity in the SS sequence is the effect of individual codon usage bias. From Figure 4, the SS sequence has the lowest score for individual codon usage when compared to the CV and CC sequences. Although we have previously shown that codon context is more effective and relevant than individual codon usage for codon optimization [25], there may be a threshold that separate the regions where the importance of codon context or individual codon usage dominate the other. In the current set of experiments, the individual codon usage did not affect protein expression levels as seen in the Western blot. Instead, the use of rare codons could have increased translational pauses to an extent that ribosomes detached, leading to misfolding and loss of enzyme activity [23]. Clearly, further investigation should be required to deconvolute the effects of codon context and individual codon usage on protein expression and folding.

4. Conclusion

A non-native gene to the *E. coli* host coding for invertase derived from *Thermotoga maritima* was optimized based on CV, CC, and SS, and expressed to compare the activity of the enzymes. Amongst the optimization criteria considered, CC was found to generate the sequence with the highest enzyme activity per culture volume. Furthermore, SS, which adopts the secondary structure as the key element of gene design, showed a decrease in enzyme activity compared to the wild-type. Our study further illustrates the importance of CC fitness as a design parameter for optimizing the sequence towards improved heterologous protein expression.

Acknowledgements

We would like to acknowledge Rachel Chia and Eileen Lee for experimental support. This work was supported by the National University of Singapore, Biomedical Research Council of A*STAR (Agency for Science, Technology and Research), a Singapore International Graduate Award (SINGA) scholarship award from A*STAR Graduate Academy (A*GA), Singapore, and a grant from the Next-Generation BioGreen 21 Program (SSAC, No. PJ009520), Rural Development Administration, Republic of Korea.

Author Contributions

HBP carried out the experiments and wrote the manuscript. MK contributed to the analysis of the results and wrote the manuscript. KSA supported the algorithm development and synthetic gene design, and wrote the manuscript. BKSC developed the algorithm and designed the synthetic genes. DSWO helped design the experiments, provided guidance, and edited the manuscript. DYL coordinated the project and edited and approved the manuscript.

References

1. Grantham R, Gautier C, Gouy M, Mercier R, Pave A: **Codon catalog usage and the genome hypothesis.** *Nucleic acids research* 1980, **8:**r49-r62.
2. Plotkin JB, Dushoff J, Fraser HB: **Detecting selection using a single genome sequence of M. tuberculosis and P. falciparum.** *Nature* 2004, **428:**942-945.
3. Gouy M, Gautier C: **Codon usage in bacteria: correlation with gene expressivity.** *Nucleic acids research* 1982, **10:**7055-7074.
4. Duret L, Mouchiroud D: **Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis.** *Proceedings of the National Academy of Sciences of the United States of America* 1999, **96:**4482-4487.
5. Hershberg R, Petrov DA: **Selection on codon bias.** *Annual review of genetics* 2008, **42:**287-299.
6. Fuglsang A: **Codon optimizer: a freeware tool for codon optimization.** *Protein expression and purification* 2003, **31:**247-249.
7. Villalobos A, Ness JE, Gustafsson C, Minshull J, Govindarajan S: **Gene Designer: a synthetic biology tool for constructing artificial DNA segments.** *BMC bioinformatics* 2006, **7:**285.

8. Puigbo P, Guzman E, Romeu A, Garcia-Vallve S: **OPTIMIZER: a web server for optimizing the codon usage of DNA sequences.** *Nucleic acids research* 2007, **35:**W126-131.
9. Bentele K, Saffert P, Rauscher R, Ignatova Z, Bluthgen N: **Efficient translation initiation dictates codon usage at gene start.** *Molecular systems biology* 2013, **9:**675.
10. Smith D, Yarus M: **tRNA-tRNA interactions within cellular ribosomes.** *Proceedings of the National Academy of Sciences of the United States of America* 1989, **86:**4397-4401.
11. Sorensen MA, Pedersen S: **Absolute in vivo translation rates of individual codons in Escherichia coli. The two glutamic acid codons GAA and GAG are translated with a threefold difference in rate.** *Journal of molecular biology* 1991, **222:**265-280.
12. Gustafsson C, Govindarajan S, Minshull J: **Codon bias and heterologous protein expression.** *Trends in biotechnology* 2004, **22:**346-353.
13. Siller E, DeZwaan DC, Anderson JF, Freeman BC, Barral JM: **Slowing bacterial translation speed enhances eukaryotic protein folding efficiency.** *Journal of molecular biology* 2010, **396:**1310-1318.
14. Yarus M, Folley LS: **Sense codons are found in specific contexts.** *Journal of molecular biology* 1985, **182:**529-540.
15. Tats A, Tenson T, Remm M: **Preferred and avoided codon pairs in three domains of life.** *BMC genomics* 2008, **9:**463.
16. Lanza AM, Curran KA, Rey LG, Alper HS: **A condition-specific codon optimization approach for improved heterologous gene expression in Saccharomyces cerevisiae.** *BMC systems biology* 2014, **8:**33.
17. Gutman GA, Hatfield GW: **Nonrandom utilization of codon pairs in Escherichia coli.** *Proceedings of the National Academy of Sciences of the United States of America* 1989, **86:**3699-3703.
18. Coleman JR, Papamichail D, Skiena S, Futcher B, Wimmer E, Mueller S: **Virus attenuation by genome-scale changes in codon pair bias.** *Science* 2008, **320:**1784-1787.
19. Deane CM, Saunders R: **The imprint of codons on protein structure.** *Biotechnology journal* 2011, **6:**641-649.

20. Spencer PS, Barral JM: **Genetic code redundancy and its influence on the encoded polypeptides.** *Computational and structural biotechnology journal* 2012, **1:**e201204006.
21. Gromadski KB, Rodnina MV: **Kinetic determinants of high-fidelity tRNA discrimination on the ribosome.** *Molecular cell* 2004, **13:**191-200.
22. Fedorov AN, Baldwin TO: **Cotranslational protein folding.** *The Journal of biological chemistry* 1997, **272:**32715-32718.
23. Spencer PS, Siller E, Anderson JF, Barral JM: **Silent substitutions predictably alter translation elongation rates and protein folding efficiencies.** *Journal of molecular biology* 2012, **422:**328-335.
24. Chung BK, Yusufi FN, Mariati, Yang Y, Lee DY: **Enhanced expression of codon optimized interferon gamma in CHO cells.** *Journal of biotechnology* 2013, **167:**326-333.
25. Chung BK, Lee DY: **Computational codon optimization of synthetic gene for protein expression.** *BMC systems biology* 2012, **6:**134.
26. DE G: **Genetic algorithms in search, optimization, and machine learning.** *Addison-Wesley Pub Co* 1989.
27. Liebl W, Brem D, Gotschlich A: **Analysis of the gene for beta-fructosidase (invertase, inulinase) of the hyperthermophilic bacterium Thermotoga maritima, and characterisation of the enzyme expressed in Escherichia coli.** *Applied microbiology and biotechnology* 1998, **50:**55-64.
28. Ow DS, Lim DY, Nissom PM, Camattari A, Wong VV: **Co-expression of Skp and FkpA chaperones improves cell viability and alters the global expression of stress response genes during scFvD1.3 production.** *Microbial cell factories* 2010, **9:**22.
29. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank.** *Nucleic acids research* 2000, **28:**235-242.
30. Alberto F, Bignon C, Sulzenbacher G, Henrissat B, Czjzek M: **The three-dimensional structure of invertase (beta-fructosidase) from Thermotoga maritima reveals a bimodular arrangement and an evolutionary relationship between retaining and inverting glycosidases.** *The Journal of biological chemistry* 2004, **279:**18903-18910.
31. Bernstein FC, Koetzle TF, Williams GJ, Meyer EF, Jr., Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M: **The Protein Data Bank: a computer-based**

archival file for macromolecular structures. *Journal of molecular biology* 1977, **112:**535-542.

Figure Legends

Figure 1. Schematic illustrating the process of codon context based fitness of the secondary structure regions. The hosts' genes are categorized according to their secondary structure, and the codon context profile for each secondary structure is derived (A, B, X). A principal component analysis plot of codon context shows the differences between the secondary structure regions. Subsequently, the target protein, invertase, with the protein shown and its secondary structure regions identified as obtained from the Protein Data Bank (PDB ID: 1UYP; http://www.rcsb.org/pdb/) [30, 31], is optimized based on the codon context profile derived for each secondary structure from the hosts' genes.

Figure 2. Growth curve of invertase expressing *E. coli* and Western analysis and enzyme activity assay. (A) OD at 600nm was plotted against time for each strain. Cells expressing the CV, CC, and SS optimized gene sequences exhibited similar growth characteristics, while the WT control had a linear growth curve, reaching a similar final OD after 8 hours. Induction with IPTG was performed 2 hours after inoculation, as indicated by the arrow. (B) Protein expression was qualitatively analysed by Western blot. The overexpression of the protein at approximately 50 kDa matches is confirmed by detection of the target protein.

Figure 3. Enzyme activity assay of wild-type and codon optimized invertase expressed in *E. coli*. (A) Stoichiometric reaction of the enzyme activity assay to measure invertase activity. Sucrose is hydrolysed to form glucose and fructose; the glucose further reduces 3,5-dinitrosalicylic acid to 3-amino,5-nitrosalicylic acid, which can be detected at 575 nm. (B) Absorbance at 575 nm per cell culture volume was potted for each coding sequence. CC optimized illustrated the highest level of enzyme activity, with CV and WT having comparable levels. SS optimized displayed the lowest level of enzyme activity. Error bars represent one standard deviation. Statistical significance was determined by t-test as a comparison to WT (*: $p<0.05$; **: $p<0.001$).

Figure 4. Individual codon usage and codon context characteristics of wild type and optimized sequences. Figure shows a comparison of ICU and CC distributions in wild type, CVO, SSO, and CCO sequences. The resulting differences in host specific fitness are depicted in the CC & ICU fitness plot. The Pareto front (denoted by black circles) depicts the trade-off between individual codon usage and codon context fitness. The shading of plot points (squares) indicates the codon adaptation index, the geometric mean of the weight of each codon for the coding sequence.

Table 1. List of primers used.

Primer	Nucleotide sequence (5'-3')
Inv-His-F	AGGAGATATACCATGCACCACCACCACCACCAC
Inv-WT-R	GGTGGTGGTGCTCGAGTCATTACAACCATATGTTCTCGAGTTC
Inv-CVO-R *	GGTGGTGGTGCTCGAGTCATTACAGCCAGATGTTTTCCAGTTC
Inv-CCO-R *	GGTGGTGGTGCTCGAGTCATTACAGCCAGATGTTTTCCAGTTC
Inv-SSO-R	GGTGGTGGTGCTCGAGTTATCATAACCAGATATTTTCCAGTTCG

* Same sequence

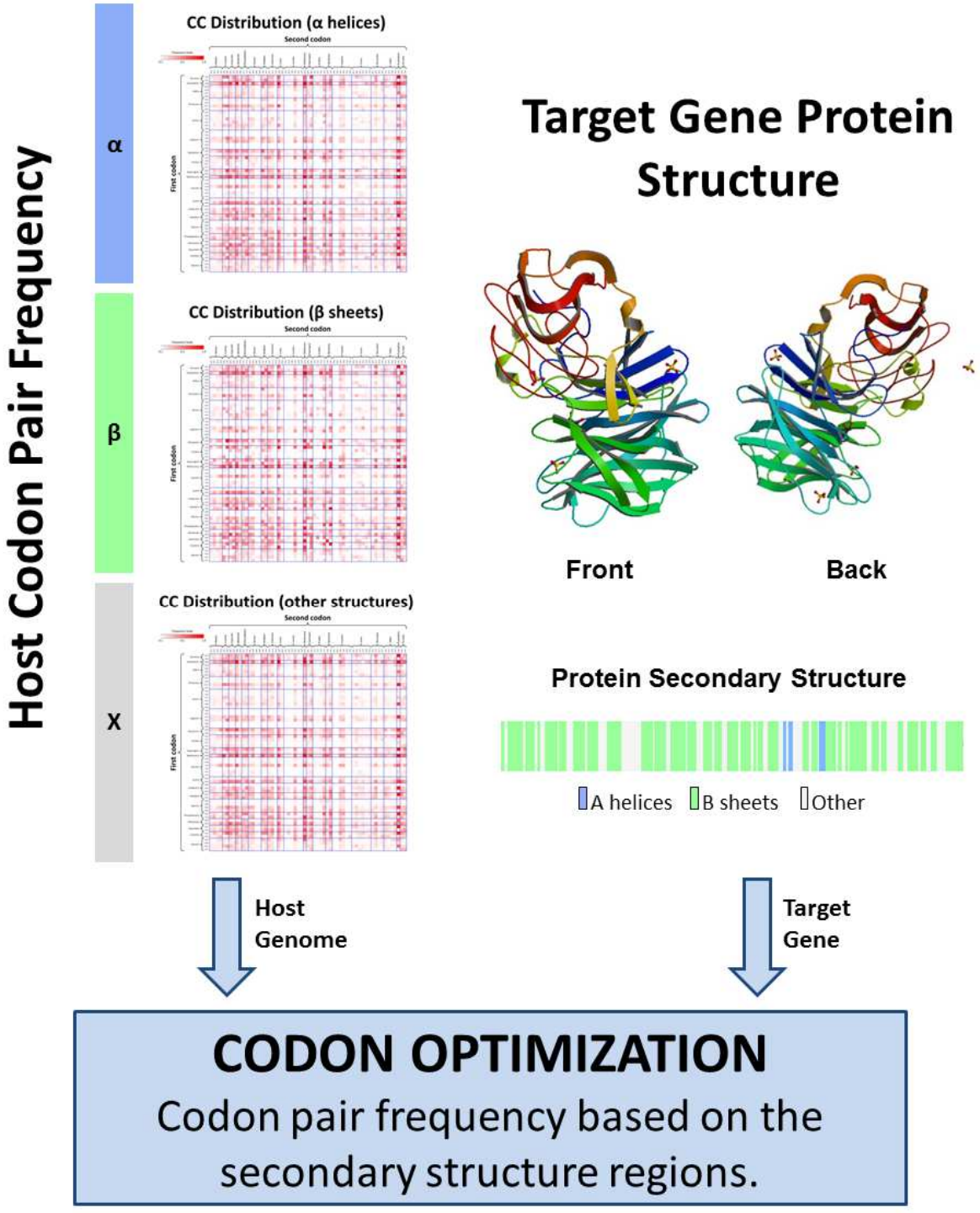
Host Codon Pair Frequency
α
β
X
CC Distribution (α helices)
CC Distribution (β sheets)
CC Distribution (other structures)
Second codon
First codon
Target Gene Protein Structure
Front
Back
Protein Secondary Structure
A helices
B sheets
Other
Host Genome
Target Gene
CODON OPTIMIZATION
Codon pair frequency based on the secondary structure regions.

Figure 1

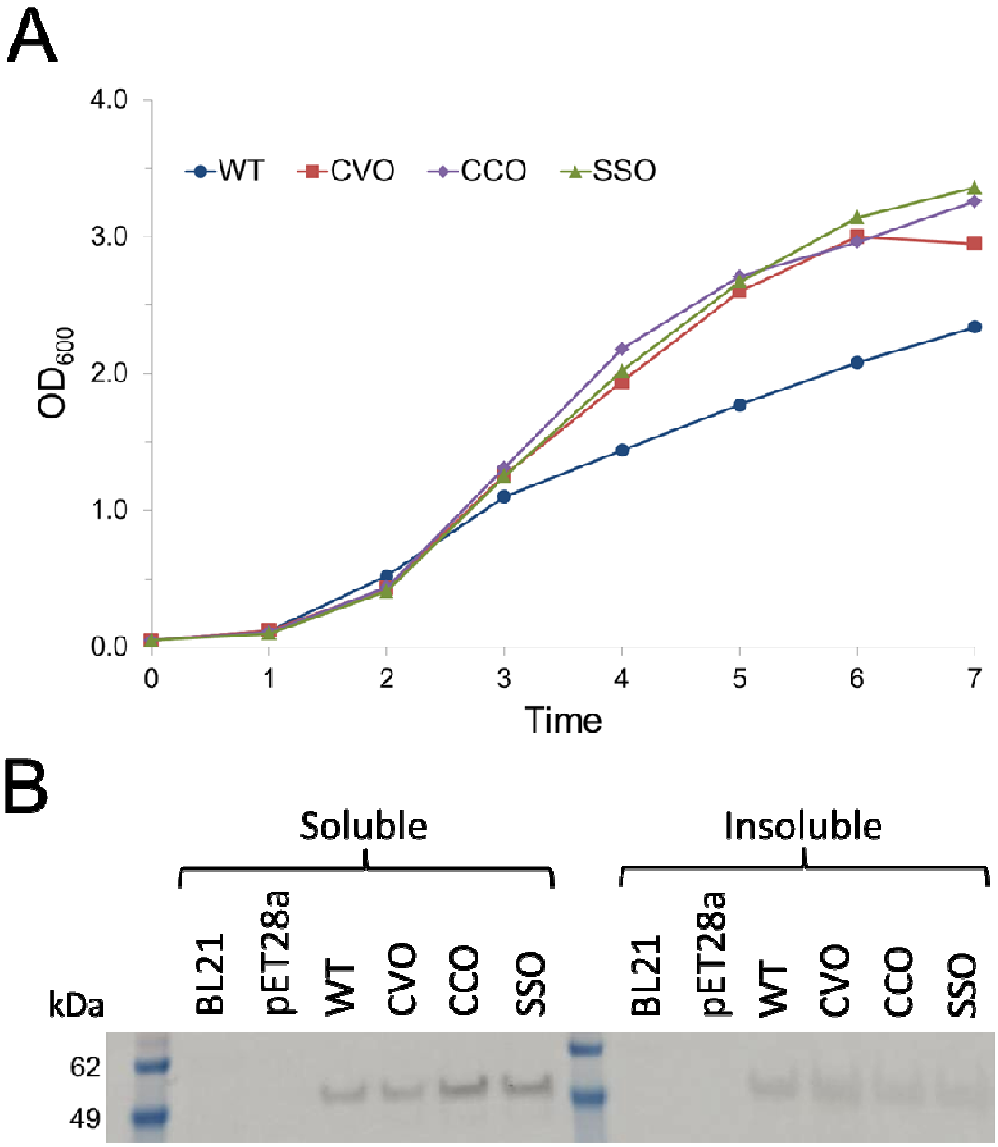
A
4.0
3.0
2.0
1.0
0.0
OD600
WT
CVO
CCO
SSO
0
1
2
3
4
5
6
7
Time
B
Soluble
Insoluble
BL21
pET28a
WT
CVO
CCO
SSO
BL21
pET28a
WT
CVO
CCO
SSO
kDa
62
49

Figure 2

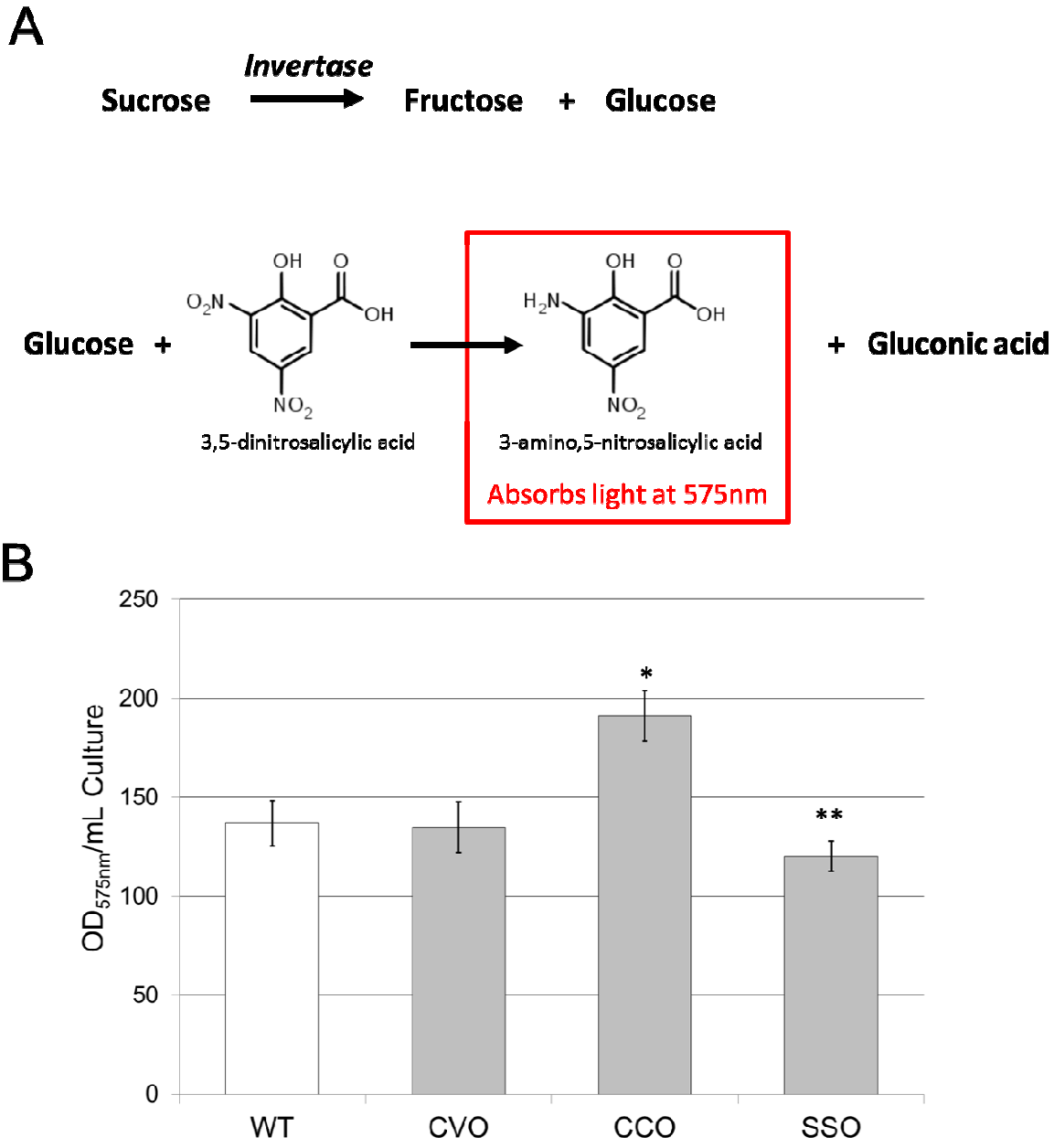
A
Invertase
Sucrose
Fructose + Glucose
Glucose +
+ Gluconic acid
3,5-dinitrosalicylic acid
3-amino,5-nitrosalicylic acid
Absorbs light at 575nm
B
OD575nm/mL Culture
250
200
150
100
50
0
*
**
WT
CVO
CCO
SSO

Figure 3

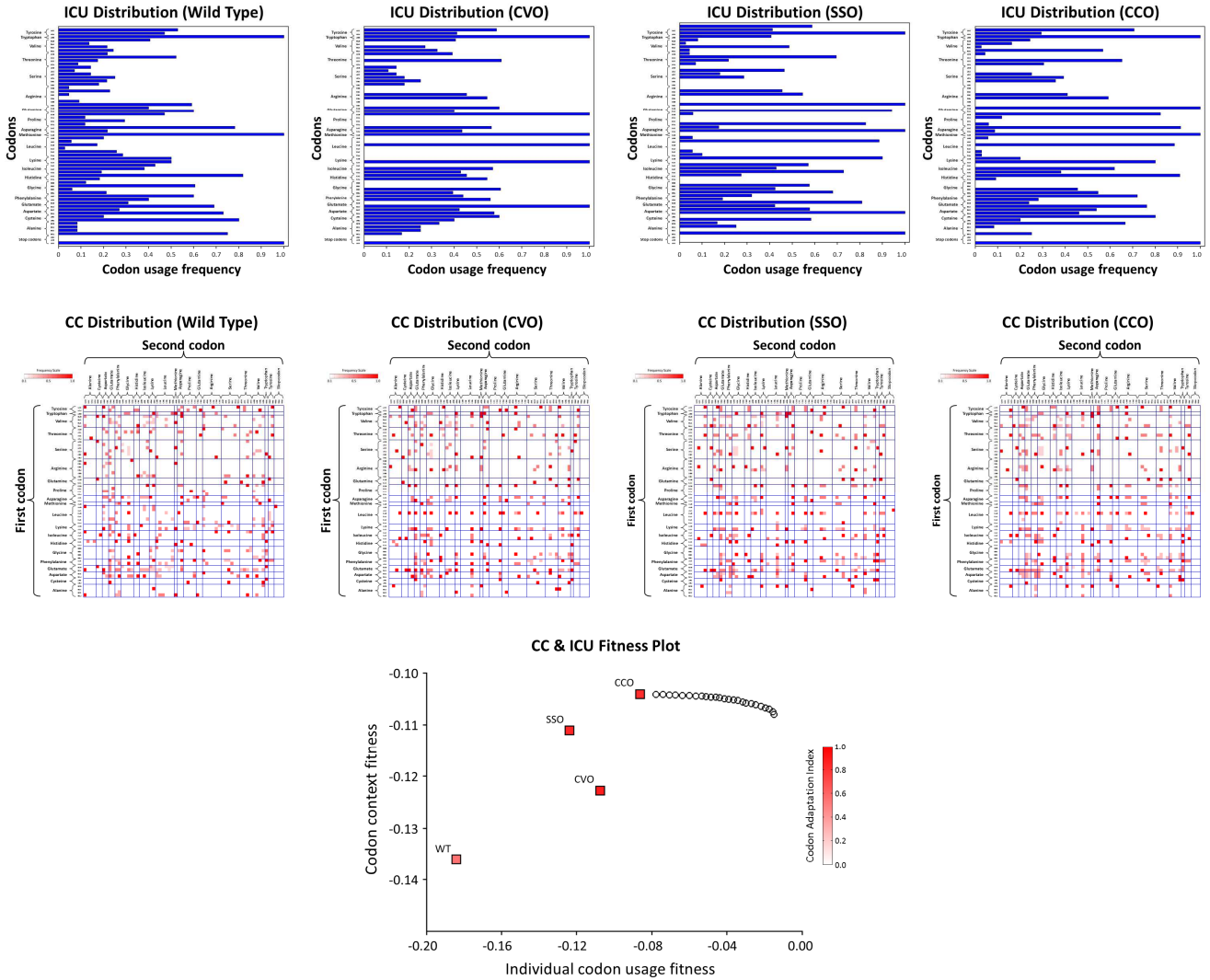
ICU Distribution (Wild Type)
ICU Distribution (CVO)
ICU Distribution (SSO)
ICU Distribution (CCO)
Codons
Codon usage frequency
CC Distribution (Wild Type)
CC Distribution (CVO)
CC Distribution (SSO)
CC Distribution (CCO)
Second codon
First codon
CC & ICU Fitness Plot
Codon context fitness
Individual codon usage fitness
Codon Adaptation Index
WT
CVO
SSO
CCO

Figure 4