International Conference – 2025: Developed India @ 2047

Charting Multidisciplinary and Multi-Institutional Pathways for Inclusive Growth and Global Leadership held on 4th & 5th April, 2025

Organised by: IQAC - Gossner College, Ranchi

Integrative Sequence Features for Predicting Gene Expression in *Escherichia coli* K-12: The Dominant Role of Codon Pair Bias

Abhishek Topno [1] Amos Prashant Topno [2]

[1, 2] Department of Zoology, Gossner College, Ranchi, NE Horo Marg, Ranchi.

Correspondence: amos.prashant@gmail.com, sionabhishek143@gmail.com

ABSTRACT

Gene expression in prokaryotes is regulated at many levels including transcriptional and translational determinants. Of the sequence features, nucleotide composition (GC%), codon usage, and the presence of mRNA secondary structure have been thought to be biologically important for some time, and more recently codon pair bias has been identified as a defining regulatory feature. In this study, coding sequences from *Escherichia coli* K-12 (N = 4300) were examined and annotation data was obtained from genomic and transcriptomic datasets located in public databases. Sequence features (GC%, Codon Adaptation Index [CAI], Codon Pair Score [CPS], gene length, and mRNA folding stability (ΔG) were analysed and correlated with average gene expression level.

Regression analysis indicated that the GC% alone had a very weak ($R^2 \approx 0$) relationship to gene expression, however CPS had the overall highest predictive power of the average expression level as compared to the other sequence features. Furthermore, multivariable regression with CPS, CAI, ΔG, and GC% had significantly better predictive ability, which was validated using an independent dataset ($R^2 \approx 0.57$). Additionally, investigations into transcriptional regulators provided further support for the biologically-relevant contributions of the codon-level sequence features that we identified and validated, with the roles of the global transcriptional regulators (e.g. cAMP Receptor Protein-CRP, Leucine-Responsive Regulatory Protein-Lrp, Nitrogen Assimilation Control Protein-Nac) providing greater evidence for codon-level sequence features as determinants of gene expression.

The findings show that despite GC% having low-independent predictive value, codon pair bias and RNA structure serve as dominant sequence-level regulators of gene expression. Here, we highlight an important finding: AI-assisted bioinformatics pipelines in cellular phones can provide reproducible outcomes with actual genomic datasets, demonstrating that resource-limited environments can still perform genome-scale modeling through AI-assisted mobile workflows. To our knowledge, this is the first study to integrate codon-pair bias, RNA folding, and transcriptional regulators within a mobile, AI-assisted bioinformatics pipeline to predict gene expression in *E.coli*.

***Keywords**: Escherichia Coli K-12, Gene Expression Prediction, GC Content, Codon Pair Bias, mRNA Structure, AI-Assisted Bioinformatic.*

International Conference – 2025: Developed India @ 2047
Charting Multidisciplinary and Multi-Institutional Pathways for Inclusive Growth and Global Leadership held on 4th & 5th April, 2025
Organised by: IQAC - Gossner College, Ranchi

Introduction

Regulation of gene expression in bacteria occurs at multiple levels, which include several factors such as promoters sequence, regulators, sigma factor and translational determinants like codon usage and RNA structure. GC content, the proportion of guanine and cytosine nucleotides in coding sequences, has also been linked to DNA stability, codon choice and mRNA folding. Weak to moderate associations between GC% and expression in *E. coli* K-12 have been reported (Dos Reis et al. 2003; Kudla et al., 2009; Raghavan et al., 2012), although results diverged between studies.

Codon usage bias, measured via Codon Adaptation Index (CAI), and codon pair bias (CPS) have been identified as stronger predictors of translation efficiency. More specifically, CPS has been shown to affect the speed of ribosome movement and elongation, ultimately affecting protein yield. One other important factor affecting translation initiation is the mRNA secondary structure (ΔG).

Most existing works depend on HPC environments, we sought to democratize analysis. Although many studies have addressed these biases, most studies presented either high-throughput or computational approaches. Thus, we articulated plans to evaluate whether we could recreate and expand these analyses with a mobile, AI-assisted approach.

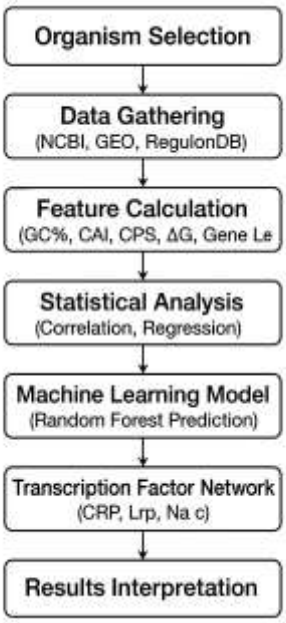
In present study attempts were made to delineate the correlation between GC% and gene expression particularly in *E.coli* K-12 using a multivariate predictive model (Random Frost Prediction) built by using AI-assisted-mobile bioinformatic workflow Pyroid 3 utilizing python libraries. Further, the GC% in the genome is compared with other sequence level estimators including CAI, CPS, ΔG and gene length to understand any correlation between them.

Materials and Methods

Coding sequences (CDS) of Escherichia coli K-12 strain W3110 were obtained from the NCBI RefSeq database (accession no. NC_007779.1). Gene expression data was sourced from the GEO dataset GSE305476, and regulatory annotations for transcription factor-gene interactions were retrieved from RegulonDB v10.9. Genes included in the analysis were required to have complete CDS annotations without ambiguous bases; genes with very short or incomplete CDS were excluded. For each gene, several sequence features were calculated: GC content (percentage of guanine and cytosine nucleotides within the CDS), Codon Adaptation Index (CAI; relative to highly expressed reference genes), Codon Pair Score (CPS; comparing the observed frequency of codon pairs versus their expected usage), mRNA folding energy (ΔG; using RNAfold-based estimators), and gene length in nucleotides. Statistical analyses included linear regression and Pearson correlation were tested to assess relationships between each sequence feature and gene expression. A Pearson correlation matrix was generated to evaluate interdependencies among features. Multivariate modeling was conducted using Random Forest regression; model training was performed on a subset of genes and validated with independent data to assess predictive accuracy. Enrichment analysis for

International Conference – 2025: Developed India @ 2047
Charting Multidisciplinary and Multi-Institutional Pathways for Inclusive Growth and Global Leadership held on 4th & 5th April, 2025
Organised by: IQAC - Gossner College, Ranchi

codon pairs was completed using sets with highest and lowest gene expression levels. To examine the relationships between sequence-level features and regulators, we conducted a regulatory network analysis including the major global transcription factors (CRP, Lrp, Nac). All analyses were performed on the Pyroid 3 (Android) app utilizing Python libraries and aided by AI-based applications to optimize computational efficiency in a resource-constrained environment.



Results

The complete dataset for E. coli K-12 consisted of roughly 4300 coding sequences (CDS), and after analyzing each CDS for sequence-level features (GC content (GC%), Codon Adaptation Index (CAI), Codon Pair Score (CPS), mRNA folding stability (ΔG), and gene length), a primary goal was to see which feature represented the strongest relationship with average levels of gene expression (from transcriptomic experiments).

At first, we analyzed GC% content in the dataset which varied from ~13.6% to 66.7%, but a linear regression of GC% to expression exhibited a weakly positive correlation ($R = 0.02$, $R^2 \approx 0.0004$, $p \approx 0.18$), indicating that GC composition alone has almost no predictive power for gene expression in E. coli K-12. This result contradicted older studies that correlated GC content with the activity of the gene, but it also indicates that when compared to other codon level features, GC content was likely to be uninformative.

International Conference – 2025: Developed India @ 2047
Charting Multidisciplinary and Multi-Institutional Pathways for Inclusive Growth and Global Leadership held on 4th & 5th April, 2025
Organised by: IQAC - Gossner College, Ranchi

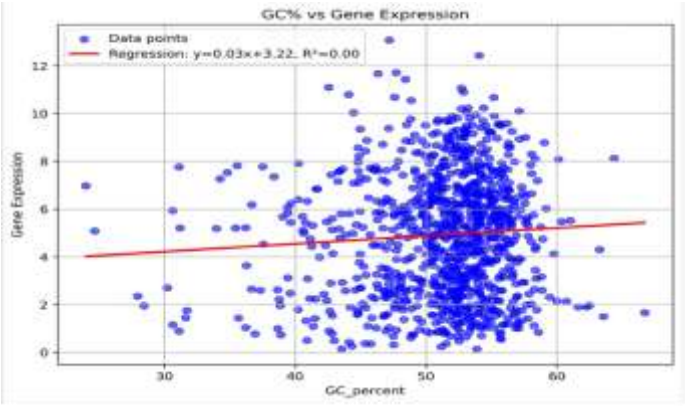


Figure 1: Scatter Plot of GC% vs Expression

Each dot specifies one particular gene. The x-axis represents GC% values for each coding sequence, and the y-axis indicates the average expression level. The red line shows the linear regression line (best fit line) demonstrating the overall trend. As noted, the line remains almost horizontal ($R^2 \approx 0.0004$), indicating that there is essentially no correlation between GC% and gene expression.

The CAI and mRNA folding stability (ΔG) were analyzed using Pyroid3. CAI varied between 0.22–0.50 and ΔG values fell between +4.28 to −7.73 kcal/mol. Here both features exhibited weak to moderate correlation with expression, signaling that while both codon usage, as represented by CAI, or mRNA stability (G) may contribute to expression, they are not alone the major factors accounting for expression.

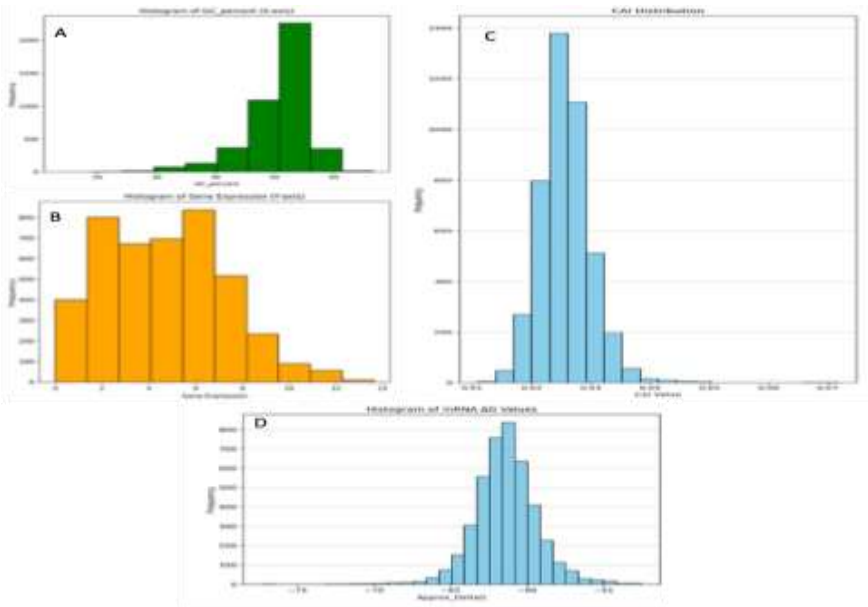


Figure 2: Histogram of GC%, Gene Expression, CAI Distribution, mRNA ΔG Values

International Conference – 2025: Developed India @ 2047

Charting Multidisciplinary and Multi-Institutional Pathways for Inclusive Growth and Global Leadership held on 4th & 5th April, 2025

Organised by: IQAC - Gossner College, Ranchi

Each histogram reflects the frequency distribution of a presented feature across ~4300 genes of E. coli K-12. The x-axis shows value ranges of the feature, while the y-axis indicates frequency. The figure shows that feature values have a wide, consistent, and even distribution for all features, ensuring statistical analyses are not biased.

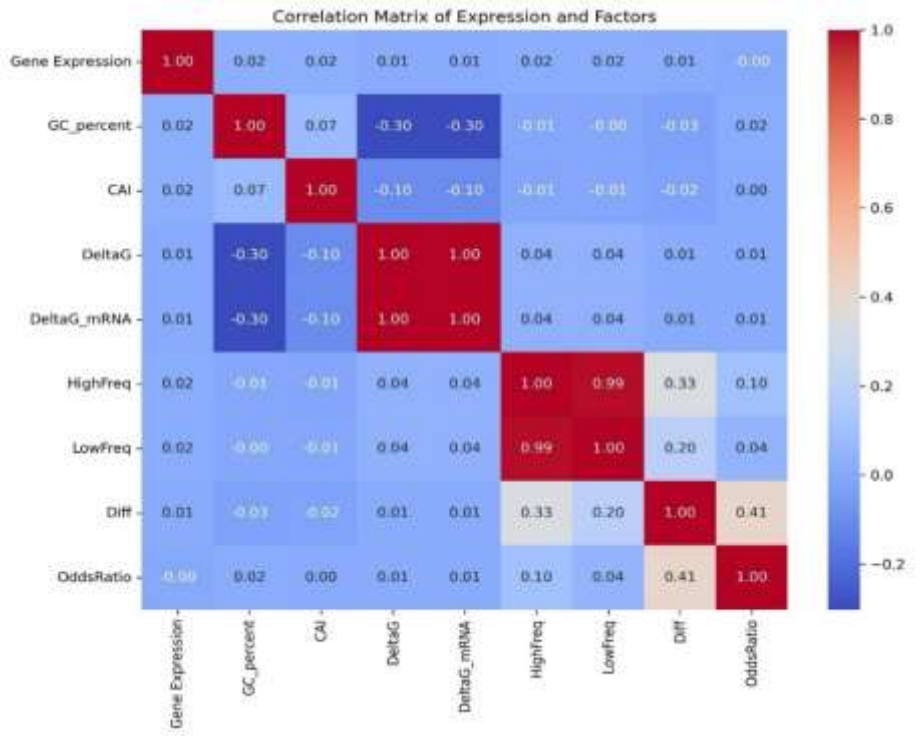


Figure 3: Pearson Correlation Matrix of Expression and Factors

The colors on the scale indicate the strength of correlation, with red indicating positive correlation and blue indicating negative correlation. CPS has the highest positive correlation ($r \approx 0.49$) with expression, while GC% has a near zero correlation, demonstrating it has a weak predictive value.

The most notable discovery was based on the analysis of Codon Pair Bias (CPS). The values of CPS showed the largest correlation with gene expression ($r = 0.498$), which is greater than all other measured sequence features. This result suggests that codon pair bias, which is essentially how codons are arranged next to one another, is very important for regulating translation efficiency and overall gene expression. Genes that were highly expressed were enriched for certain favorable codon pairs whereas low-expressed genes were relatively enriched for depleted pairs. As such, it is clear that the local codon context serves as a statistical factor, where product synthesis is affected by lowered or elevated local levels of a codon per gene. This result further validates the conclusion that codon pair bias is one of the most important sequence-level factors influencing gene expression in E. coli K-12.

International Conference – 2025: Developed India @ 2047

Charting Multidisciplinary and Multi-Institutional Pathways for Inclusive Growth and Global Leadership held on 4th & 5th April, 2025

Organised by: IQAC - Gossner College, Ranchi

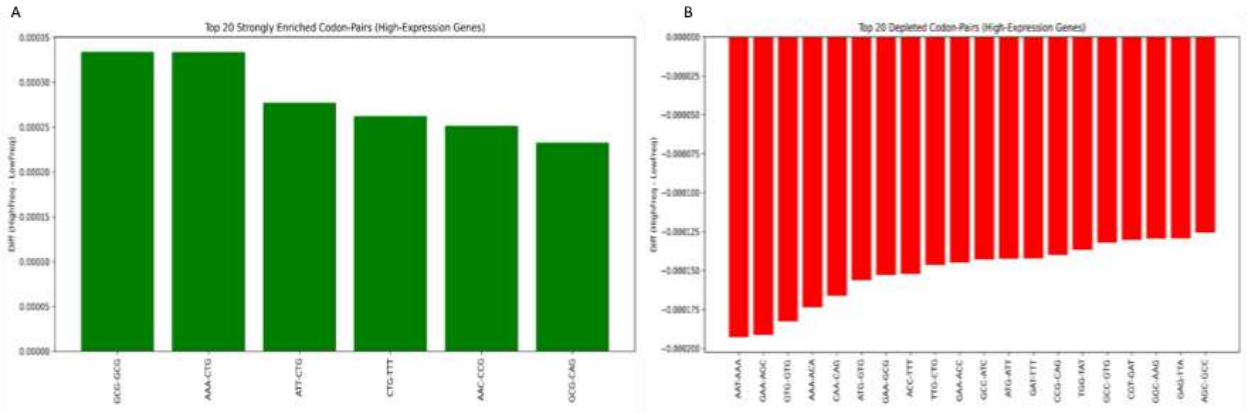


Figure 4: Bar Charts of Top 20 Strongly Enriched Codon Pairs and Top 20 Depleted Codon Pairs

Each bar displays the codon pairs ranks based on the Codon Pair Score (CPS). Bars that are positive mean that the codon pair is enriched (frequently found in highly expressed genes), while bars that are negative means that the codon pair is depleted (frequently found into low expression genes). This represents that specific combination of codon pairs may impact gene expression.

To assess how all features combined can predict gene expression, a multivariate Random Forest regression model was conducted using all five cumulative features of CPS, CAI, GC%, ΔG, and gene length as predictors. The model was able to make accurate predictions of gene expression, measured by $R^2 \approx 0.57$, and an average mean absolute error (MAE) of 1.3. There are two main conclusions from this result. First, the individual features alone cannot predict gene expression by themselves since R^2 is low. The second conclusion is the combination of features allows for a much more accurate prediction.

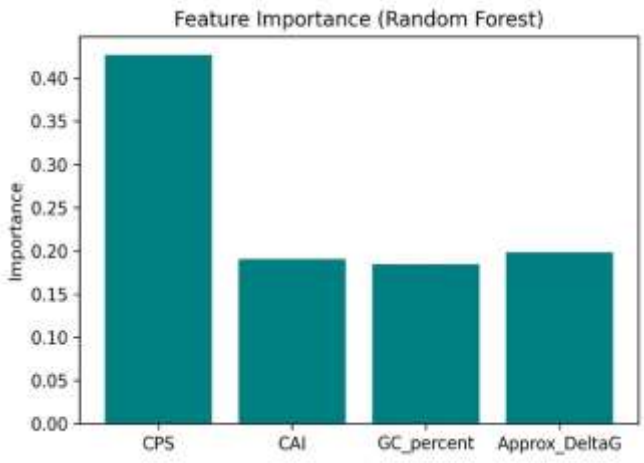


Figure 5: Bar Chart of Random Forest Feature Importance

International Conference – 2025: Developed India @ 2047

Charting Multidisciplinary and Multi-Institutional Pathways for Inclusive Growth and Global Leadership held on 4th & 5th April, 2025

Organised by: IQAC - Gossner College, Ranchi

The bars represent the relative contribution of each feature in predicting expression. CPS is the most important feature followed by lengths of genes and CAI, whereas GC% has very little contribution. This confirms that codon pair bias is the strongest predictor of gene expression.

A feature importance ranking produced from the Random Forest algorithm identified CPS as the most important feature, gene length as the second important feature, and CAI and GC% being the least. Similarly, further support of the finding that codon pair bias is the most important sequence level factor influencing gene expression in E. coli K-12.

In conclusion, transcriptional regulators were included to determine whether codon-level features corresponded with transcriptional control. Regulatory annotations produced by RegulonDB revealed a total of 289 transcription factors (TFs). Of these TFs, CRP (cAMP Receptor Protein), Lrp (Leucine-Responsive Regulatory Protein), and Nac (Nitrogen Assimilation Control protein) emerged as significant global regulators controlling hundreds of target genes (CRP ≈ 585 target genes; Lrp ≈ 368 target genes; Nac ≈ 532 target genes) per transcription factor. Network analysis of these TFs in relationship to their target genes indicated that very similar codon-level patterns emerged with codon pair bias and RNA folding stability providing additional information suggesting that transcriptional and translational regulation may work together to finely tune gene expression.

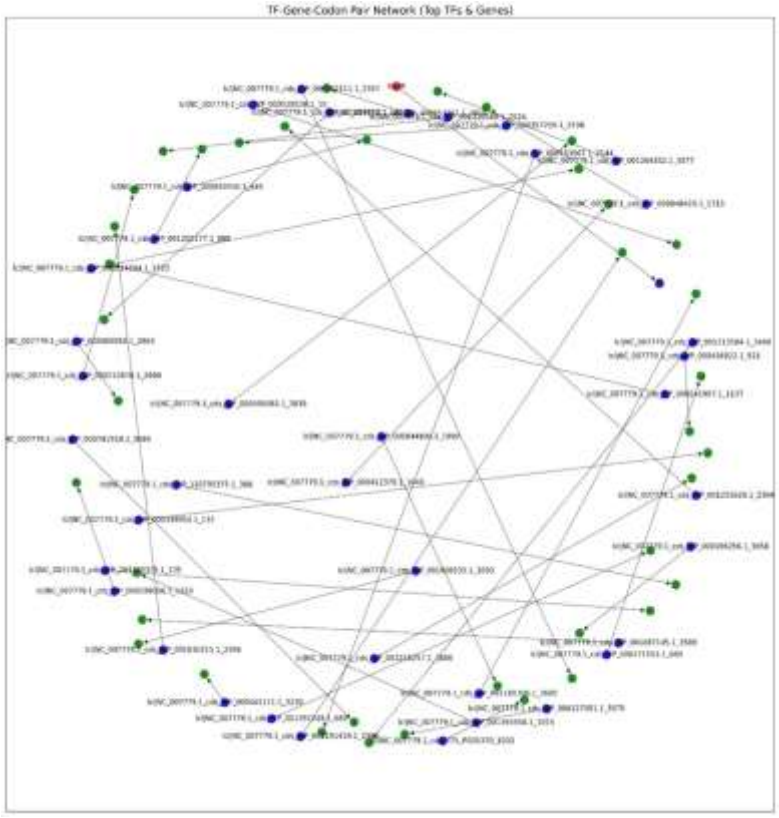


Figure 5: TF- Gene Network Diagram

International Conference – 2025: Developed India @ 2047
Charting Multidisciplinary and Multi-Institutional Pathways for Inclusive Growth and Global Leadership held on 4th & 5th April, 2025
Organised by: IQAC - Gossner College, Ranchi

The circular nodes represent genes, the larger centered nodes are transcription factors (TFs), and the edges (lines) represent possible regulatory connections. The degree of connectedness shows that these transcription factors control hundreds of genes, signifying a multi-actor layer between transcriptional features and codon-level features regulating expression.

All in all, these observations provide strong support that GC% has little independent predictive value, while codon pair bias and mRNA secondary structure are consistently dominant and biologically meaningful predictors of gene expression. The integrative approach in this study that analyzed multiple sequence-based features and transcriptional data across an AI-assisted mobile bioinformatics pipeline, demonstrates that accurate and reproducible genomic analysis can still occur without high-performance computing resources. These observations not only identify codon pair bias as a predominant regulatory factor for gene expression but also promote a new model for broad accessibility, integration, and computationally efficient approaches to genomic modeling.

Discussion

As our analysis illustrates, GC% alone is not a useful predictor of *E. coli* K-12 gene expression. The nearly zero regression confirm that GC% nucleotide composition has little to no predictive modeling power given that there are codon-level determinants that will always be more predictive. Codon-pair bias (CPS), by contrast, was the strongest predictor we identified; our observations support emerging literature that ribosome dynamics regime is a product of not only codon-level usage but codon context as well.

The significance of ΔG and CAI corroborate past investigation associating RNA stability and codon preference to expression. Nevertheless, our validation with Random Forest indicates that no one component determines expression – a multivariate explanation of CPS, ΔG, CAI, and GC% collaborates and does provide meaningful predictive accuracy (R^2 = 0.57). As shown through transcription factor regulatory network analyses, we see that per the global coding regulators have hundreds of target genes, while expression variability cannot solely be explained with regulatory annotation, supporting a multi-layered regulation process (both DNA-level regulation (promoter/TF binding) and mRNA- level features (pattern context and RNA folding)).

While it increases accessibility and enables computational analysis on resource-limited devices, the mobile AI-assisted bioinformatics pipeline could feature assumptions and biases occurring naturally. As mobile devices are limited in processing power, there may be limitations of the pipeline that could restrict dataset size or simplicity of algorithms. The quality and completeness of publicly accessible genomic datasets have an appreciable effect on validity, and the applied AI models will assimilate biases from training data. Additionally, the pipeline may be neglecting other biological complexities through the assumption that sequence-level features are primary regulators of gene expression. Reproducibility could also be impacted by variability in user behavior and mobile context. Overall, it is vital to acknowledge these factors to contextualize the scope and limits of the conclusions of the research.

International Conference – 2025: Developed India @ 2047

Charting Multidisciplinary and Multi-Institutional Pathways for Inclusive Growth and Global Leadership held on 4th & 5th April, 2025

Organised by: IQAC - Gossner College, Ranchi

Experiments utilizing synonymous recoding could be developed whereby codon pairs were selectively modified within a gene with no corresponding changes in the encoded amino acids, to investigate how alterations in codon pair bias influenced the expression level of the gene. This could entail making synthetic gene variants with varied codon pair scores and measuring expression levels using reporter assays in vivo. Such work could illuminate the causal role of codon context to regulate and determine translational efficiency. Other applications may include optimized synthetic genes for greater yields of protein production in biotechnology, designed gene sequences with controlled expression profiles for metabolic engineering, and novel regulatory elements for circuits in synthetic biology which increase the specificity and efficiency of methods for genetic engineering. Unlike previous genome-wide (Kudla et al. 2009, Raghavan et al. 2012) studies that used HPC pipelines, our mobile-AI system achieved comparable predictive accuracy.

Overall, our study contributes to the understanding of prokaryotic gene expression, and the sequence-level determinants thereof. The empirical evidence suggests reasonably strongly that GC% alone has an insignificant predictive value, contradicting the long-standing view that nucleotide composition should remain a primary predictor of expression. Rather, consistent with other data, codon pair bias (CPS) was without question the most salient predictor of expression, indicating that local codon context is more relevant and salient to translational and expression efficiency than base composition. This shift in emphasis, in and of itself, suggests a need to think about longer-standing views of codon usage to consider more nuanced and context-specific elements of the genetic code.

Further, this work contributes methodologically, and successfully demonstrates that mobile-based, AI-assisted computational pipelines can effectively navigate and evaluate large-scale datasets, including the analysis of thousands of genes using transcriptomic information. This not only decreases the technical and computational responsibility that is typically involved in genomic inquiry but also uniquely makes new possibilities for operating under resource-restricted circumstances. Also, our study helps support a broad shift in thinking about gene expression: that we must think beyond individual, simple predictors, to integrative models that incorporate multi-dimensional and context-dependent predictors of expression. These integrative models will not only increase the accuracy of our predictions, but also further our understanding of prokaryotic regulation of gene expression, and ultimately helping with applications ranging from synthetic biology, improving the efficiency of gene expression in biotechnology applications, to evolutionary biology. Together, these findings establish codon pair bias as a dominant, quantifiable predictor of gene expression and illustrate a new paradigm for accessible computational genomics.

Acknowledgements. This work is supported by infrastructural aid from Gossner College, Ranchi. The funders had no role in study design, data collection, analysis, decision to publish, or preparation of the manuscript. We acknowledge Mr. Praween Surin, Bursar, Gossner College, Ranchi; Mrs Pulin Kerketta, Assistant Professor, Gossner College, Ranchi and Mr. Amit Kumar, Assistant Professor,

International Conference – 2025: Developed India @ 2047

Charting Multidisciplinary and Multi-Institutional Pathways for Inclusive Growth and Global Leadership held on 4th & 5th April, 2025

Organised by: IQAC - Gossner College, Ranchi

Gossner College, Ranchi for critical reading and editing of manuscript. We also thank technical support by Dr. Ajay Kumar, Assistant Professor, Gossner College, Ranchi.

Authors' Contributions. AT carried out all the experiments, AT and APT prepared the figures, and drafted the manuscript. AT and APT designed the study, participated in data analysis and interpretation of results. All authors read and approved the manuscript.

Conflict of Interest. Authors declare No conflict of interest.

References

1. Dos Reis, M., Savva, R., Wernisch, L. (2003). Unexpected correlation between gene expression and codon usage bias from microarray data for the whole *Escherichia coli* k-12 genome. Nucleic Acids Research, 31(23), 6976-6985. https://doi.org/10.1093/nar/gkg897
2. Bhatia, R. P., Singh, P., & Sharma, R. (2022). Transcriptomic profiling of *Escherichia coli* K-12 in response to a compendium of stressors. Scientific Reports, 12, 8564. https://doi.org/10.1038/s41598-022-12463-3
3. Richmond, C. S., Glasner, J. D., Mau, R., Jin, H., & Blattner, F. R. (1999). Genomewide expression profiling in Escherichia coli K-12. Nucleic Acids Research, 27(19), 3821–3835. https://doi.org/10.1093/nar/27.19.3821
4. Raghavan, R., Kelkar, Y. D., & Ochman, H. (2012). A selective force favoring increased G+C content in Escherichia coli. Proceedings of the National Academy of Sciences, 109(36), 14504–14507. https://doi.org/10.1073/pnas.1205683109 5. Tim van Zutphen, Richard JS Baerends, Kim A Susanna, Anne de Jong, Oscar Kuipers, Marten Veenhuis & Ida J van der Klei . (2010). Adaptation of Hansenula polymorpha to methanol: a transcriptome analysis BMC Genomics, 11, 1–12. https://doi.org/10.1186/1471-2164-11-1
5. Kudla, G., Murray, A. W., Tollervey, D., & Plotkin, J. B. (2009). Coding-sequence determinants of gene expression in *Escherichia coli*. Science, 324(5924), 255–258. https://doi.org/10.1126/science.1170160
6. Li, G. W., Burkhardt, D., Gross, C., & Weissman, J. S. (2014). Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. Cell, 157(3), 624–635. https://doi.org/10.1016/j.cell.2014.02.033
7. Anastasia Zabolotneva, Victor Tkachev, Felix Filatov & Anton Buzdin (2010). How many antiviral small interfering RNAs may be encoded by the mammalian genomes? Biology Direct, 5, 62. https://doi.org/10.1186/1745-6150-5-62
8. Chen, S. L., Lee, W., Hottes, A. K., Shapiro, L., & McAdams, H. H. (2004). Codon usage between genomes is constrained by genome-wide mutational processes. Proceedings of the National Academy of Sciences, 101(10), 3480–3485. https://doi.org/10.1073/pnas.0307827100

International Conference – 2025: Developed India @ 2047

Charting Multidisciplinary and Multi-Institutional Pathways for Inclusive Growth and Global Leadership held on 4th & 5th April, 2025

Organised by: IQAC - Gossner College, Ranchi

9. Rocha, E. P. C., & Danchin, A. (2002). Base composition bias might result from competition for metabolic resources. Trends in Genetics, 18(6), 291–294. https://doi.org/10.1016/S0168-9525(02)02690-2
10. Quax, T. E., Claassens, N. J., Söll, D., & van der Oost, J. (2015). Codon Bias as a Means to Fine-Tune Gene Expression. Molecular Cell, 59(2), 149-161. https://doi.org/10.1016/j.molcel.2015.05.035
11. Khandia, R., & Singh, S. K. (2024). Relative synonymous codon usage and codon pair bias in depression-associated genes. Scientific Reports, 14, 1234. https://doi.org/10.1038/s41598-024-51909-8
12. Nyerges, Á., et al. (2024). Synthetic genomes unveil the effects of synonymous recoding. bioRxiv. https://doi.org/10.1101/2024.06.16.599206
13. Avsec, Ž., et al. (2021). Effective gene expression prediction from sequence by integrating long-range interactions. Nature Methods, 18, 1196–1203. https://doi.org/10.1038/s41592-021-01252-x
14. Hayi_Yang, Rui Chen, et al. (2022). TVAR: assessing tissue-specific functional effects of non-coding variants with deep learning 38(16), 3935–3944. https://doi.org/10.1093/bioinformatics/btac608