Optimization of Violacein Biosynthesis for Enhanced Antimicrobial Properties

Jackson Maines[1], Carlos Gonzalez Rivera[2], Andrew Frank[1,]Andrew Wilson[1,3], Rob Egbert[1]

[1]Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Washington, USA

[2]Electricity Infrastructure and Buildings Division, Pacific Northwest National Laboratory, Richland, Washington, USA

[3]Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA

ABSTRACT. An element of synthetic biology is designing biological systems with novel or enhanced characteristics. Many engineered functions require the coordinated, optimized expression of multiple genes, which can be difficult to achieve in a new genetic context. This study attempts to control the expression of the antimicrobial pigment violacein in *Escherichia coli* (E. coli). Violacein production is determined by a model biosynthetic pathway with multiple pigment outputs controlled by the expression levels of the *vioABEDC* genes. The factors that promote its antimicrobial properties are not well known and the proper balancing of pathway expression and expression-related growth defects could enable the design of cell-based antimicrobial solutions. Our research aims to investigate the relationship between specific gene expression levels within the violacein pathway and the resulting antimicrobial impact of violacein production on a competitor bacterium. We generated a large combinatorial genetic library by tuning the expression levels of the five genes in the pathway using engineered ribosome binding sites. Our results reveal variability in growth across plates, which could point to experimental errors due to timing. Additionally, we find little evidence to suggest different combinations of gene expressions affect growth. These findings highlight the complexity of violacein biosynthesis and the need for further experimentation with more controlled growth conditions. Future attempts to fine-tune gene expression levels, potentially guided by machine learning algorithms, may help lead to optimized violacein production and enhanced antimicrobial properties. A deeper understanding of the factors influencing the violacein pathway could maximize the potential benefits of its antimicrobial properties, with promising applications in areas such as human gut health or soil sciences research.

1 Introduction

The ongoing search for novel antimicrobial agents has led to increased interest in naturally occurring compounds. Violacein, a purple pigment produced by certain bacteria strains, has promising antimicrobial properties[1]. A five gene operon *vioABEDC* expressed a pathway that converts the amino acid tryptophan into violacein or multiple related pigmented compounds through a complex, branched pathway. Fine tuning the expression levels of these five genes could lead to new insights on factors that promote the violacein pathway's antimicrobial properties. Unlocking violacein's full potential could have a diverse impact on pharmaceutical[2], agriculture[2], cancer[3], and material science research[4]. We approached this study from a synthetic biology angle, creating a combinatorial genetic library by uniquely tuning the ribosome binding site (RBS) of each gene within the violacein biosynthetic pathway. By analyzing the relationship between gene expression, bacterial growth, pigment features, and the growth inhibition of competitor microbes by violacein, we aimed to identify optimal gene expression combinations for enhanced biosynthesis. Our findings reveal the complex interaction between gene expression levels and phenotypic outcomes, highlighting the potential for RBS engineering, but also indicating the need for further experimentation under more controlled conditions to better understand the underlying relationships.

2 Methods

Preprocessing the data.

2.1 Genotype Data. Sequencing allows one to observe the positional information of DNA. This is useful in the understanding the function of genes, information about mutations, and other genetic details. With this information, DNA sequencing can be used to help identify desirable traits in an organism. In this research roughly 10,000 samples were sent for sequencing. Each sample has a unique DNA sequence "barcode" to link the RBS variants. We implemented a computational pipeline using Snakemake to process long-read PacBio sequencing data for the analysis of the *vioABEDC* gene cluster[5]. The Snakemake pipeline required repair, as the original implementation developed by a previous employee was non-operational and lacked documentation. We restored functionality to the pipeline, addressing deprecated function calls and implementing new alternatives. We developed documentation that includes implementation guidelines, specifically detailing installation and execution procedures for Windows Subsystem for Linux (WSL) environments. All steps were logged, ensuring reproducibility of the analysis pipeline. The pipeline was designed to process demultiplexed reads, align sequences to a reference, identify sequence variants, and generate consensus sequences. The output was organized hierarchically, with results stored according to barcode identifiers. Consensus sequences were aggregated for every sample and labeled with a circular consensus sequence (CCS) reading. To ensure accurate downstream analysis, we implemented filtering steps on our sequencing data. Given potential issues and variation in sequencing data, we focused on "dominant barcodes" within each sample. A "dominant barcode" can be defined as a sequence reading with one CCS reading greater than 10% of the total CCS reading for that sample.

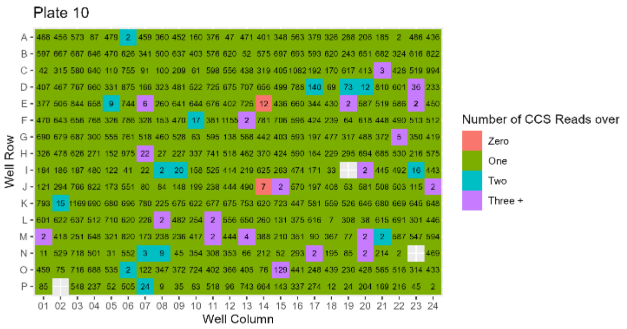

Fig 1. Visualization of CCS reads for a 384-well violacein variant plate. Green squares indicate a sample with only one dominant barcode, each number indicating the number of CCS reads for the dominant barcode. Missing squares had no quality sequencing data.

This filtering step resulted in the validation of approximately 5,000 high-quality wells from the initial pool of 10,000 with one dominant barcode. The other wells were assumed to be low-quality since multiple genotypes may be present. To identify RBS and mutations present in each gene, we performed pairwise alignment with the alignment parameters specified in the following table using bio.align library in Biopython[6].

Match	Mismatch	Open Gap	Extend Gap
+1	-1	-2	-1

Fig 2. Table of alignment scoring system. The Needleman-Wunsch algorithm was used for pairwise alignment and scored with calculated penalties.

The aligned sequences were then analyzed to identify mutations within each gene, which were classified as synonymous, nonsynonymous, or frameshift mutations based on the translated amino acid sequence. To estimate the impact of different RBS sequences we utilized De Novo DNA RBS Calculator[7-13] to predict the translation initiation rate (TIR) for each unique RBS identified for vioA, vioB, vioE, vioD, vioC genes.

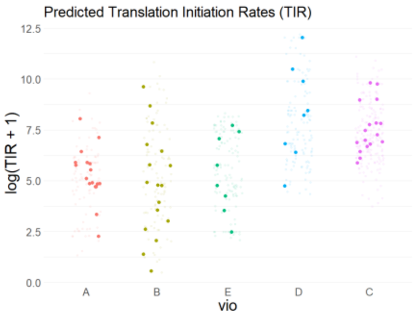

Fig 3. Jitter Plot of Gene Predicted Translation Initiation Rates. Solid circles represent common RBS (n > 100). Predictions done by De Novo DNA RBS Calculator.

Given the number of possible RBS variants for each gene, this resulted in a theoretical 262,144 (16 x 16 x 8 x 8 x 16) unique RBS combinations. This approach allows us in the future to identify which combination of mutations or RBSs impact the phenotypic traits of violacein.

2.2 Growth Data. In microbial growth, we expect 4 regions in the growth curve: lag phase, region before any significant growth; exponential phase, region where the culture is growing exponentially; stationary phase, region where the culture stops growing; and the death phase, when the rate of cell death is higher than that of new cells forming. The end of the lag phase was calculated using gcplyr's function lag_time, which identifies the intersection between a tangent line projected at the point of max growth and the minimum measured optical density (OD) value[14]. As we expect there to be a region of exponential growth, we also expect a region of linear growth in the logarithmic-transformed growth data. This is because of the logarithmic property of the power rule, which is that the log of a power is equal to the exponent times the log of its base. Thus, the exponential region should end at the point where the log of the data stop growing linearly. Using a sliding window approach, we identified this point at which the log-transformed data fell below a defined threshold. The points within the exponential phase were fitted to a two-parameter exponential model (amplitude and rate) to extract exponential growth rates. This method to find exponential growth rate works most accurately for cultures that exhibit a sigmoid shape growth curve but fail to capture accurate growth rates of those that don't. For samples that we couldn't calculate an accurate lag time we labeled them zero. Unless otherwise specified these samples were excluded from subsequent analysis involving OD data.

2.3 Agar Plate Data. To assess the antimicrobial activity of the violacein expression variants, we performed an agar plate-based assay. Samples were first cultured overnight at 37°C, then spotted onto thirty 384-well agar plate containing *Bacillus subtilis*, a bacterium sensitive to violacein. After incubation, top-down photographs of the plate were analyzed using image processing techniques. Specifically, we used Grounded-Segment-Anything (Grounded-SAM), a combination of Grounded Dino from IDEA Research and Segment Anything from Facebook Research[15], to automate the measurement of key phenotypic traits. This approach allows us to measure parameters, such as location on the plate, size of the culture, pigment features, and the area of inhibition, the area of the zone in which bacterial growth is inhibited. Combining these three sources of data together we created a multimodal dataset, which we can further analyze.

3 Results

During initial data exploration, a subset of the samples exhibited growth defects characterized by unrealistic lag times or negative growth rates. Upon further investigation of these growth defective samples, we found no significant associations between the TIR of *vioABEDC* genes and the growth defects. We therefore attributed these defects to experimental errors and excluded them from the analysis of the growth data. In the analysis of lag time in non-defective samples we observed a bimodal distribution in both induced and uninduced conditions.

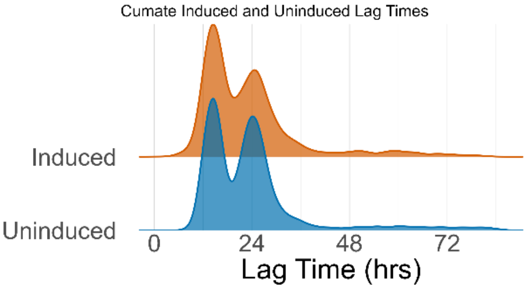

Fig 4. Density plots of lag times with and without cumate, a chemical inducer of the synthetic violacein pathway. Statistics performed by paired Wilcoxon signed-rank test, comparing uninduced less than or equal to induced. p = < 2.2e-16

We suggest that this bimodal distribution is a result of varying durations of growth. As listed in the table below, plates varied in growth from less than 24 hours to more than 72 hours. We recommend performing the growth assays again with a subset of the strains from wells with single dominant barcodes.

Group	1	2	3	4	5	6
Plates	3-7	8-10	11-15	16-20	21-25	26-30
Duration (hours)	~66	~24	~24	~90	~48	~66

Fig 5. Table of plates grown together in microplate reader and the duration of growth of each group.

We additionally examined the relationship between gene TIR and growth rate. Our initial analysis suggested a potential connection between the *vioE* RBS and growth rate. However, Kurkal-Wallis rank sum tests revealed no statistically significant association in either induced (p = 0.2345) or uninduced (p = 0.2129) groups. Furthermore, we explored potential interactions between TIR levels of different *vioABEDC* genes but found no statistically significant relationships, with the interaction between *vioED* being the most significant with p-values greater than .1 for both the induced and uninduced groups.

4 Conclusion

Despite inconsistency in our initial growth data, this study provides a foundation for understanding the relationship between the *vioABEDC* genes and the production of violacein. Future work in the revision of lab procedures and recultivation of problematic samples should be prioritized to apply more robust statistical analysis. The new lab procedures should include a strict duration of growth times for every plate and equal incubation times. Our Grounded-SAM model to extract area of inhibition needs future work to consistently segment every culture. The computer vision model's generalizability should be evaluated using newly cultivated agar plate samples to assess its accuracy. Using area of inhibition and pigment data collected by Grounded-SAM we can further link genotype to phenotype characteristics. By integrating phenotypic traits with novel computation modeling, we can begin to predict the phenotypic outputs of specific gene expression combinations. This approach may ultimately yield new insights in the optimization of violacein biosynthesis for diverse biological applications.

5 Acknowledgements

I would like to thank Rob Egbert taking me as an intern even though I don't understand biology. Additionally, I want to thank Soujanya Akella for taking time out of her day to answer my random biology questions, and Andrew Frank for racking his memory about the experimental design of the poolGLASS project. Lastly, thank you to the SynBio team and others that did the experimentation for this project.

This work was supported in part by the U.S. Department of Energy, Office of Science, Office of Workforce Development for Teachers and Scientists (WDTS) under the Science Undergraduate Laboratory Internships Program (SULI).

References

1. Chauhan, A., Mathkor, D.M., Joshi, H., Chauhan, R., Sharma, U., Sharma, V., Kumar, M., Saini, R.V., Saini, A.K., Tuli, H.S., Kaur, D. and Haque, S. (2025), Mechanistic Insight of Pharmacological Aspects of Violacein: Recent Trends and Advancements. Journal of Biochemical and Molecular Toxicology, 39: e70114. https://doi.org/10.1002/jbt.70114
2. Ahmed A, Ahmad A, Li R, Al-Ansi W, Fatima M, Mushtaq BS, Basharat S, Li Y, Bai Z. Recent Advances in Synthetic, Industrial and Biological Applications of Violacein and Its Heterologous Production. J Microbiol Biotechnol. 2021 Nov 28;31(11):1465-1480. doi: 10.4014/jmb.2107.07045. PMID: 34584039; PM CID: PMC9705886.
3. Durán N, Nakazato G, Durán M, Berti IR, Castro GR, Stanisic D, Brocchi M, Fávaro WJ, Ferreira-Halder CV, Justo GZ, Tasic L. Multi-target drug with potential applications: violacein in the spotlight. World J Microbiol Biotechnol. 2021 Aug 16;37(9):151. doi: 10.1007/s11274-021-03120-4. PMID: 34398340.
4. Park, H., Park, S., Yang, Y. H., & Choi, K. Y. (2021). Microbial synthesis of violacein pigment and its potential applications. *Critical Reviews in Biotechnology*, *41*(6), 879–901. https://doi.org/10.1080/07388551.2021.1892579
5. Mölder, F., Jablonski, K.P., Letcher, B., Hall, M.B., Tomkins-Tinch, C.H., Sochat, V., Forster, J., Lee, S., Twardziok, S.O., Kanitz, A., Wilm, A., Holtgrewe, M., Rahmann, S., Nahnsen, S., Köster, J., 2021. Sustainable data analysis with Snakemake. F1000Res 10, 33.
6. Peter J. A. Cock, Tiago Antao, Jeffrey T. Chang, Brad A. Chapman, Cymon J. Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, Michiel J. L. de Hoon, Biopython: freely available Python tools for computational molecular biology and bioinformatics, *Bioinformatics*, Volume 25, Issue 11, June 2009, Pages 1422–

1423, https://doi.org/10.1093/bioinformatics/btp163

7. An Automated Model Test System for Systematic Development and Improvement of Gene Expression Models
Alexander C. Reis and Howard M. Salis
ACS Synthetic Biology 2020 *9* (11), 3145-3156
DOI: 10.1021/acssynbio.0c00394
8. Systematic Quantification of Sequence and Structural Determinants Controlling mRNA stability in Bacterial Operons
Daniel P. Cetnar and Howard M. Salis
ACS Synthetic Biology 2021 *10* (2), 318-332
DOI: 10.1021/acssynbio.0c00471
9. Amin Espah Borujeni, Daniel Cetnar, Iman Farasat, Ashlee Smith, Natasha Lundgren, Howard M. Salis, Precise quantification of translation inhibition by mRNA structures that overlap with the ribosomal footprint in N-terminal coding sequences, *Nucleic Acids Research*, Volume 45, Issue 9, 19 May 2017, Pages 5437–5448, https://doi.org/10.1093/nar/gkx061
10. Translation Initiation is Controlled by RNA Folding Kinetics via a Ribosome Drafting Mechanism
11. Amin Espah Borujeni and Howard M Salis, Journal of the American Chemical Society 2016 138(22), 7016-7023, Translation Initiation is Controlled by RNA Folding Kinetics via a Ribosome Drafting Mechanisms DOI: 10.1021/jacs.6b01253
12. Amin Espah Borujeni, Anirudh S. Channarasappa, Howard M. Salis, Translation rate is controlled by coupled trade-offs between site accessibility, selective RNA unfolding and sliding at upstream standby sites, *Nucleic Acids Research*, Volume 42, Issue 4, 1 February 2014, Pages 2646–2659, https://doi.org/10.1093/nar/gkt1139
13. Salis, H., Mirsky, E. & Voigt, C. Automated design of synthetic ribosome binding sites to control protein expression. *Nat Biotechnol* 27, 946–950 (2009). https://doi.org/10.1038/nbt.1568
14. Blazanin M (2024). “gcplyr: an R package for microbial growth curve data analysis.” *BMC Bioinformatics*, 25(232). doi:10.1186/s12859-024-05817-3, version 1.11.0.
15. Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, & Lei Zhang. (2024). Grounded SAM: Assembling Open-World Models for Diverse Visual Tasks.