Generative model for Pseudomonad genomes

Manasa Kesapragada
Applied Mathematics
University of California Santa Cruz
mkesapra@ucsc.edu

R Shane Canon
National Energy Research Scientific Computing Center
Lawrence Berkeley National Laboratory
scanon@lbl.gov

Sean P Jungbluth
Environmental Genomics and Systems Biology Division
Lawrence Berkeley National Laboratory
sjungbluth@lbl.gov

Marcin P Joachimiak
Environmental Genomics and Systems Biology Division
Lawrence Berkeley National Laboratory
mjoachimiak@lbl.gov

Adam P Arkin
Environmental Genomics and Systems Biology Division
Lawrence Berkeley National Laboratory
aparkin@lbl.gov

Paramvir S Dehal
Environmental Genomics and Systems Biology Division
Lawrence Berkeley National Laboratory
psdehal@lbl.gov

Abstract

Recent advances in genomic sequencing have resulted in several thousands of full genomes of pseudomonads, a genera of bacteria important in many science areas ranging from biogeochemical cycling in the environment to bacterial pneumonia in humans. With these high-quality data sets, combined with tens of thousands of somewhat lower quality metagenomically assembled genomes, we create a generative model for pseudomonad genomes. We present a GAN model that generates gene family presence absence lists as a representation of a novel genome. We also demonstrate that the discriminator of this model can be used as a binary classifier to identify incorrect genomes with missing content. In the future, our desired model can be used to generate genomes within a given set of parameters such as, "Generate a genome that is root associated, drought resistant, salt tolerant that will produce this natural product".

1 Introduction

Synthetic biology is benefiting from the massive-scale DNA sequencing of natural (environmental and host-associated) and industrial/laboratory-relevant microbial genomes. The rapidly expanding

Learning Meaningful Representations of Life Workshop at 36th Conference on Neural Information Processing Systems (NeurIPS 2022).

catalog of sequenced microbial genomes provides a wealth of information about gene and trait content that can be used to design and produce synthetic organisms of high biotechnological and medical value. Current research exploring lineage-specific gene family profiles typically uses a pangenomics approach[13], but this workflow is not necessarily the most informative with respect to synthetic organism design, because the rules for transforming a pangenome into a genome are unknown. Despite the widespread interest in accurate design and production of artificial microbial genomes, this remains a grand challenge for synthetic biology disciplines. In this work, we narrow the problem to consider only the gene content.

Generative Adversarial Networks (GANs) are a potentially powerful method to approach artificial microbial genome design. Here we present a generative model trained on gene family presence/absence profiles from the widespread (i.e. large host-range) bacterial Pseudomonas lineage to create artificial gene presence/absence lists for analog genome constructs. From this basic model we can then build toward the generation of complete genome sequences which include additional properties.

2 Methods

Genus level Pseudomonas genome data is sourced and downloaded from National Center for Biotechnology Information (NCBI)[8], licensed under a Creative Commons Attribution-ShareAlike 4.0 International License. The Genome Taxonomy Database (GTDB)[14] is used as a guide to identify Pseudomonad nucleotide sequence genomes of interest. The analysis in this study is done on a random subset of 3,000 of the 12K Pseudomonad genomes from NCBI. Analysis was performed using compute resources from the DOE Systems Biology Knowledge Base(KBase)[1] and National Energy Research Scientific Computing Center (NERSC).

2.1 Pangenome Analysis

The first step towards building the model is to perform a pangenome analysis. A pangenome is the entire set of genes from all strains within a group. It can be understood as the union of all the genomes of a group. The pangenome can be broken down into a "core pangenome" that contains genes present in all genomes, a "shell pangenome" that contains genes present in two or more strains, and a "cloud pangenome" that contains genes only found in a single strain [4, 5, 6]. Pangenome analysis helps in understanding the genetic determinants of biological activity. To perform pangenome analysis we use PPanGGOLiN [3], a Free Software suite used to create and manipulate prokaryotic pangenomes from a set of genomic DNA sequences. We run the ppanggolin workflow by providing a tsv-separated file, a list with first column being a unique genome name and the second column being its path to the associated FASTA file(compressed fna file). The PPanGGoLin analysis on 3000 genomes was run using NERSC Perlmutter HPC system, around 450GB of memory was utilized and took 13 hours to complete. PPanGGOLiN provides multiple outputs to describe a pangenome. For the current study, we use the core, shell, cloud genes lists from the output partition folder and gene_presence_absence.Rtab file.

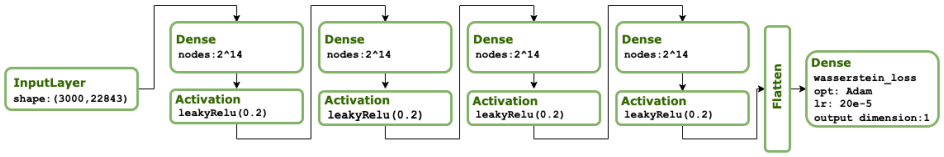
2.2 Data Input

The data input to the model is a presence absence matrix extracted from gene_presence_absence.Rtab file of PPanGGOLiN ouput, where the columns are the genomes used to build the pangenome, the rows are the gene families. There is a 1 if the gene family is present in a genome, and 0 otherwise. For the current analysis we are interested in finding the genes which are present in two or more genomes, and hence we filter the presence absence matrix to contain only the core genes (genes present in all the genomes) and shell genes (genes present in 10-95% of the genomes) and remove the cloud genes (genes less than 10% occurrence). This matrix data to the model can be described as a 2D tensor with genomes and their corresponding genes' present/absent.

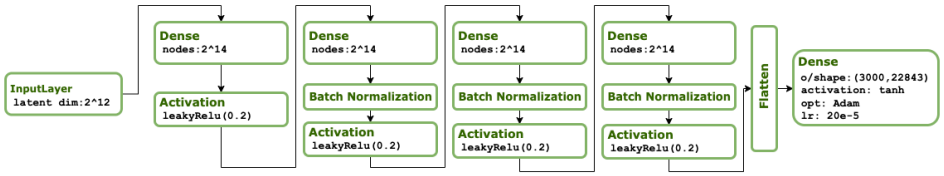
2.3 Model

To build the generative model we use the Wasserstein GAN-Gradient Penalty method. Generative adversarial networks (GANs) [9] are a powerful class of generative modeling subjectively regarded as producing better samples than other methods [10]. GANs have two neural networks playing against each other where one, the generator, learns to generate reasonable data as training and the other, the

discriminator, learns to distinguish the generator's fake data from real data. The generated instances become negative training examples for the discriminator. Instead of using a discriminator to classify or predict the probability of generated data as being real or fake, we use WGAN which changes the discriminator model with a critic that scores the realness or fakeness of a given data using Wasserstein loss [7]. The WGAN gradient penalty method [11] is used in order to ensure model convergence and stability. The network architecture is designed based on the recommendations dealing with optimization using larger matrix data [15] with a sequence of densely connected layers having 2^{14} neurons in each layer, around 70% of the input genes size (22843 core + shell genes), generator input latent vector of size 2^{12} shown in Fig.1. The input dataset is a gene presence-absence matrix of 3000



a) Critic



(b) Generator

Figure 1: WGAN Model for Pseudomonad genomes

genomes with 22843 core and shell genes. We train on randomly selected 80% of the dataset with a batch size of 64. The Adam optimizer with a learning rate of 0.0002, beta_1 of 0.5 and beta_2 of 0.9 are utilized. In the WGAN model, as the critic model must be updated more than the generator model, we update it 3 times more than the generator. The model is trained for 50 epochs on AMD EPYC 7763 CPUs.

3 Results

To evaluate the performance of the models, we first test the generator for the percentage of core genes present in their gene presence absence lists. These gene presence absence lists of the genomes have a median of 73% of the required core genes which can be shown in a clustermap in Fig. 2, with number of genomes on the y-axis and core genes on the x-axis. Next, we test the critic model on unseen test genomes, which are the 20% data split before training. The critic model predicted 598 genomes as real among 600 test genomes showing 99% accuracy. When we evaluate the critic using synthetic incorrect genomes created by removing the core genes, the model was only able to rightly predict 1 as incorrect out of the 4 incorrect genomes. This result on the incorrect genomes is obvious as the critic was trained only on the fake data generated by the generator but was not pre-trained with the fake data from the input source. Hence, we plan to further train the model with a combination of real and incorrect genomes and test the critic on a larger number of incorrect genomes. From these observations we can say that the model has performed fairly well on generating gene presence absence matrices with 73% accuracy and the critic has identified 99% of the real genomes.

4 Conclusion

In this work, we present three results: the calculated Pseudomonad bacteria pangenome, a generative model to identify incorrect genomes, and generated gene family presence/absence lists for artificial genomes. This research contributes to finding incomplete gene complements of low-quality genomes (e.g. metagenomically assembled genomes) as well as the study of gene functions and positional information by introducing a novel method for tackling this problem. Despite restricting the input

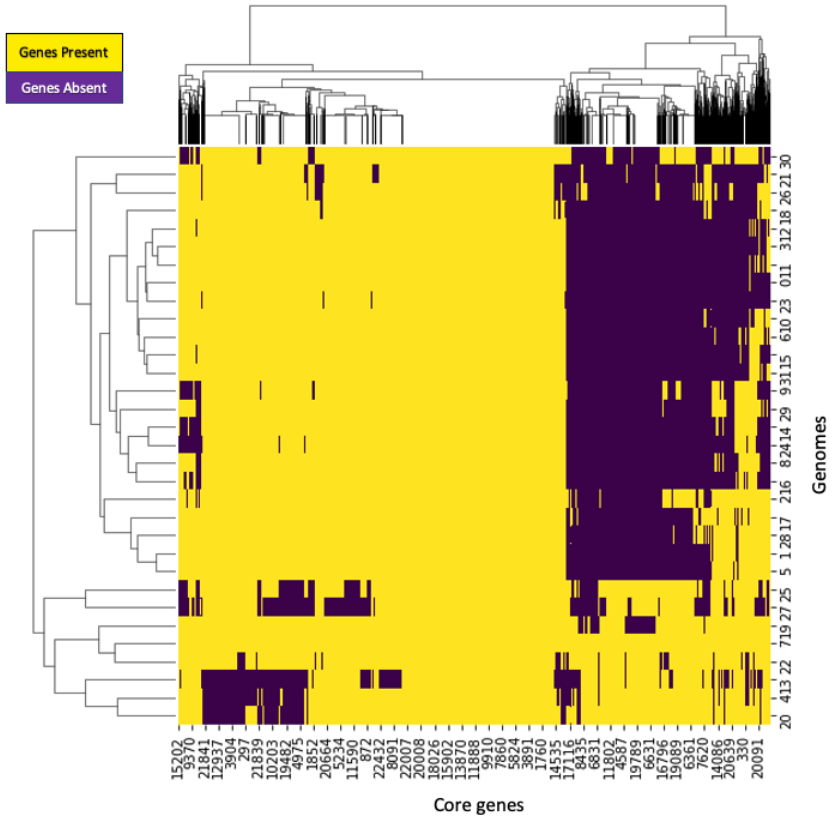


Figure 2: Generator ouptut evaluation

features to only gene presence/absence, the model is apparently able to learn enough of the gene covariance structure to make artificial gene lists. The model can accurately predict real genomes, which has immediate applications in the assessment of genomes constructed from metagenomic sequencing (MAGs) which can often be incomplete or chimeric[12, 2]. Given the performance of our current model using only gene presence/absence, we believe that incorporation of additional biologically-significant gene features will significantly improve the model. Ultimately, our goal is to create a model capable of generating genomes with desired characteristics, this will require augmenting our gene level features, such as functions and position; multi gene modules or operons and pathways; and organism level trait annotations, will make the generative model significantly more informative. However, even the most basic of these features, gene function and position, are not straightforward to assign, extract or encode into the feature tables. Gene functional annotation is imprecise and because bacterial genomes are often highly fragmented, even the ordering of genes on the chromosome is not known. Higher level features, such as organismal traits and traits that emerge under specific environmental interactions are very heterogenous and incomplete from most genomes, which introduce additional challenges which may require development of new methods. Despite these challenges, we believe that this work represents a solid foundation to build towards designer genomes.

5 Author Contributions

MK: Formal Analysis; Software; Validation; Visualization; Writing - Original Draft. PSD: Conceptualization; Methodology; Supervision; Writing - review editing. APA: Supervision; Funding Acquisition; Methodology. MPJ: Writing - Review Editing; Resources. SPJ: Writing - Review Editing; Data Curation. RSC: Resources; Supervision. PSD and APA: Corresponding authors.

6 Acknowledgements

This work was supported by the Office of Science, Office of Biological and Environmental Research, of the US Department of Energy under Award Numbers DE-AC02-05CH11231, DE-AC02-06CH11357, DE-AC05-00OR22725, and DE-AC02-98CH10886, as part of the DOE Systems Biology Knowledgebase. This research used resources of the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility located at Lawrence Berkeley National Laboratory, operated under Contract No. DE-AC02-05CH11231.

References

[1] Arkin AP et al. "KBase: The United States Department of Energy Systems Biology Knowledgebase." In: *Nature Biotechnology* (2018). DOI: 10.1038/nbt.4163.

[2] Chivian D. et al. "Metagenome-assembled genome extraction and analysis from microbiomes using KBase". In: *Nat Protoc* (2022). DOI: 10.1038/s41596-022-00747-x.

[3] Gautreau G et al. "PPanGGOLiN: Depicting microbial diversity via a partitioned pangenome graph." In: *PLOS Computational Biology 16(3): e1007732.* (2021). DOI: 10.1371/journal.pcbi.1007732.

[4] Medini D et al. "The microbial pan-genome". In: *Current Opinion in Genetics Development* 15 (2005). DOI: 10.1016/j.gde.2005.09.006.

[5] Vernikos G et al. "Ten years of pan-genome analyses". In: *Current Opinion in Microbiology* 23 (2015), pp. 148–154. DOI: 10.1016/j.mib.2014.11.016.

[6] Wolf Y.I et al. "Updated clusters of orthologous genes for Archaea: a complex ancestor of the Archaea and the byways of horizontal gene transfer." In: *Biology direct* 7 46. (2012). DOI: 10.1186/1745-6150-7-46.

[7] Martin Arjovsky, Soumith Chintala, and Léon Bottou. "Wasserstein Generative Adversarial Networks". In: *Proceedings of the 34th International Conference on Machine Learning*. Vol. 70. 2017.

[8] National Center for Biotechnology Information Bethesda (MD): National Library of Medicine (US). "National Center for Biotechnology Information (NCBI)". In: (1988). DOI: https://www.ncbi.nlm.nih.gov/.

[9] Ian Goodfellow et al. "Generative Adversarial Nets". In: *Advances in Neural Information Processing Systems*. Vol. 27. 2014. URL: https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf.

[10] Ian J. Goodfellow. "NIPS 2016 Tutorial: Generative Adversarial Networks". In: *ArXiv* abs/1701.00160 (2017). DOI: 10.48550/arXiv.1701.00160.

[11] Ishaan Gulrajani et al. "Improved Training of Wasserstein GANs". In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper/2017/file/892c3b1c6dccd52936e27cbd0ff683d6-Paper.pdf.

[12] Mobberley JM. Nelson WC Tully BJ. "Biases in genome reconstruction from metagenomic data". In: *PeerJ* (2020). DOI: 10.7717/peerj.10119.

[13] Udwary D.W. Mouncey N.J. Otani H. "Comparative and pangenomic analysis of the genus Streptomyces". In: *Scientific Reports* 12 (2022). DOI: 10.1038/s41598-022-21731-1.

[14] Donovan Parks et al. "GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy". In: *Nucleic Acids Research* 50 (2021). DOI: 10.1093/nar/gkab776.

[15] Vincent Vanhoucke, Andrew Senior, and Mark Z. Mao. "Improving the speed of neural networks on CPUs". In: *Deep Learning and Unsupervised Feature Learning Workshop, NIPS 2011*.