Research Plan for CSE3000 Research Project

Synthetic data generation for the optimization of strains in metabolic engineering using generative models

Marcin Jarosz

November 19, 2023

Background of the research

Enzymes catalyze virtually all cellular reactions along metabolic pathways [1]. Metabolic engineering involves the precise manipulation of those pathways to achieve specific system behaviors, such us higher product flux, typically for the production of economically significant substances like fuels, essential chemicals, or pharmaceuticals [2]. To give some examples, metabolic engineering has been used in the optimization of lycopene biosynthesis in E Coli [3] and xylose utilisation in S. cerevisiae [4].

A challenge in metabolic engineering lies in the expenses associated with modifying a strain to yield a sufficient output for economic viability. The high costs are attributed to various factors, with one significant aspect being the expenses related to obtaining the data essential for guiding the engineering process, as physical acquisition of genes from large number of organisms can be either costly (e.g. require specialized equipment) or experimentally complicated (experiments can take long). For that reason, machine learning models that generate synthetic data are of increasing interest.

In order to address this issue, various machine learning techniques and models have been proposed to optimize the multi-step pathways, some showing promising results [5]. However, generative machine learning models have not been exhaustively tried or benchmarked. Generative models are models which attempt to model the underlying distribution of data, and can then be used to generate new data from the same probability distribution. The available models are for example variational autoencoders, diffusion models, generative adversarial networks (GANs), and many more [6].

In this project, a probabilistic PCA and Generative Adversarial Network will be tried in order to determine if they can be utilized in the field of metabolic engineering to reduce the costs of obtaining necessary data.

Research Question

1. How can Generative Adversarial Networks be utilized to generate synthetic data for optimizing strains in metabolic engineering, and what is the quality of the generated data compared to experimental data?

   Sub-questions

2. What current ways of generating synthetic data are commonly used in metabolic engineering, and how can generative modelling be used to improve the process?
3. How can performance of a generative model be measured, in order to compare the data generated by it to experimental data and determine its overall efficiency?
4. How efficient is the probabilistic PCA model and can it be used as a baseline?
5. How efficient is the GAN model and how does it compare to the baseline?

Method

To answer the research questions, the following methods will be employed:

- **Task 1: Literature Review** A comprehensive literature review will be conducted to understand the current ways of generating synthetic data in metabolic engineering. This will help in understanding how generative modelling can be used to improve the process.
- **Task 2: Data exploration and processing** The data the models will be implemented on will be provided. It will be synthetic data simulated from a kinetic model. It will have to be thoroughly examined and possibly preprocessed to ensure its suitability for the task.
- **Task 3: Model Selection and Implementation** Probabilistic PCA and GAN models will be implemented using Python and libraries such as TensorFlow or PyTorch (to be decided).
- **Task 4: Model Evaluation** The performance of the Probabilistic PCA model and the GAN model will be evaluated. This will involve running several experiments, comparing the quality of synthetic and experimental data. The model performance on given data will also be evaluated using a discriminator, which can distinguish between real and synthetic data, and can itself be learned from data [7].
- **Task 5: Collaboration** The research will be conducted under a supervision of a PhD student in the field of bioinformatics, Paul van Lent and an active professor in the field, Thomas Abeel. Their tasks will include providing expert advice and feedback on the models and results. In addition, there will be two other students working on a similar research topic but implementing different generative models. We will share our insights and resources, so that each of us can successfully complete the research project.

Dependencies between these tasks include the need for the literature review to be completed before the models can be implemented, and the need for the models to be implemented before they can be evaluated. The collaboration will be ongoing throughout the project.

This method will allow for a thorough investigation into the effectiveness of generative models in generating synthetic data for pathway optimization in metabolic engineering. It will also provide a clear comparison between different models, allowing for the selection of the most effective model.

Planning of the research project

Below is a projected overview of tasks to be performed in each week and important deadlines:

- **Week 1**

  Tasks: Writing of the research plan, literature review: metabolic engineering, introduction to generative models; weekly meeting with supervisor and peers.

  Deadlines: First week plan (Tuesday), research plan (Sunday)

- **Week 2**

  Tasks: Research plan presentation and feedback, literature review: generative adversarial networks; data exploration, weekly meeting with supervisor and peers.

- **Week 3**

  Tasks: Final presentation arrangement, implementation of the probabilistic PCA model, weekly meeting with supervisor and peers.

  Deadlines: Set date for final presentation (Sunday)

- **Week 4**

  Tasks: Evaluation of the PCA model, deciding on metrics for comparing results, weekly meeting with supervisor and peers.

- **Week 5**

  Tasks: Midterm presentation and feedback, implementation of the GAN model

  Deadlines: Midterm presentation (Wednesday)

- **Week 6**

  Tasks: Evaluation of the GAN model, weekly meeting with supervisor and peers.

- **Week 7**

  Tasks: First draft of the final paper and feedback, peer review, weekly meeting with supervisor and peers.

  Deadlines: Paper draft v1 (Tuesday), peer review draft v1 (Thursday)

- **Week 8**

  Tasks: Incorporating feedback into the paper, second draft of the final paper, weekly meeting with supervisor and peers.

  Deadlines: Peer review repair (Monday), paper draft v2 (Wednesday)

- **Week 9**

  Tasks: Finishing writing, incorporating second round of feedback, final paper submission, weekly meeting with supervisor and peers.

  Deadlines: Final paper submission (Sunday)

- **Week 10**

  Tasks: Poster submission, final presentations

  Deadlines: Poster submission (Monday), final presentations (Tuesday or Friday)

References

[1] Alberts, B., Johnson, A., Lewis, J., Morgan, D., Raff, M., Roberts, K. & Walter, P. (2017) *Molecular biology of the cell* (6th ed.). Garland Science.

[2] Jeschek, M., Gerngross, D., & Panke, S. (2017). Combinatorial pathway optimization for streamlined metabolic engineering. *Current Opinion in Biotechnology*, 47, 142-151.

[3] Chen, X. L., Zhu, P., & Liu, L. M. (2016). Modular optimization of multi-gene pathways for fumarate production. *Metabolic Engineering*, 33, 76-85.

[4] Latimer, L. N., Lee, M. E., Medina-Cleghorn, D., Kohnz, R. A., Nomura, D. K., & Dueber, J. E. (2014). Employing a combinatorial expression approach to characterize xylose utilization in Saccharomyces cerevisiae. *Metabolic Engineering*, 25, 20-29.

[5] Lawson, C. E., Marti, J. M., Radivojevic, T., Jonnalagadda, S. V. R., Gentz, R., Hillson, N. J., ... & Martin, H. G. (2021). Machine learning for metabolic engineering: A review. *Metabolic Engineering*, 63, 34-60.

[6] Doersch, C. (2016) *Variational Autoencoders Tutorial.* arXiv.

[7] Sankaran, K., & Holmes, S. P. (2023). Generative models: An interdisciplinary perspective. *Annual Review of Statistics and Its Application*, 10, 325-352.