# CoSynthEx

(software)

Camila Riccio-Rengifo, Jorge Finke, Camilo Rocha

## 1 Description

The generation of synthetic expression data is a solution to critical challenges in various scientific disciplines, addressing issues related to data availability, and experimentation. Synthetic data plays a pivotal role in augmenting real-world datasets, enriching them with additional data points and features. This augmentation significantly enhances the performance of machine learning models, particularly in scenarios with limited real data.

Furthermore, synthetic data serves as a valuable tool for testing and validating novel methods in gene expression analysis, including emerging machine learning algorithms and feature selection techniques. This approach allows researchers to thoroughly assess the strengths and weaknesses of these methods before their application to real-world datasets, ensuring robust and reliable results.

Cost-effectiveness is another compelling advantage of synthetic data generation. The process is notably more efficient and economical compared to collecting real biological data, making it particularly advantageous for preliminary research, algorithm development, and feasibility studies.

Despite the existence of various methods for generating synthetic expression data, many struggle to accurately replicate key properties of gene expression data, as noted in previous research [Maier et al., 2013]. In response to this challenge, we introduce CoSynthEx, a software tool designed for the conditional generation of synthetic expression data. This innovative approach aims to create a more realistic simulation of expression data by incorporating additional contextual information, such as phenotypic traits and sample conditions (e.g., control or stress).

The inclusion of specific conditions, such as control and stress, in synthetic data generation proves particularly valuable when dealing with imbalanced datasets. This ensures that machine learning models are effectively trained across all relevant conditions, addressing a common issue in data analysis. Moreover, the integration of phenotypic traits allows researchers to tailor synthetic datasets to mimic precise biological contexts or experiments, aligning the generated data with the specific objectives of their studies.

CoSynthEx employs a conditional generative adversarial network (cGAN) as its foundational model. This cGAN takes real expression data, phenotypic trait data, and sample condition information as inputs. Through training and parameter adjustment, the model's loss curves are optimized to exhibit ideal behavior. The generator's loss initially starts high and progressively decreases, converging to a low value, indicating that the generator is learning to generate data resembling the real data distribution. The discriminator's loss begins relatively high and gradually decreases, eventually stabilizing or oscillating at a low value. This pattern signifies successful training, where the discriminator struggles to distinguish between real and synthetic data, reflecting the generator's ability to produce convincing data.

During training, the generator and discriminator engage in a competitive and adaptive process. The generator strives to produce increasingly realistic data, challenging the discriminator's ability to differentiate between real and synthetic data. Once the model is successfully trained, it can generate synthetic expression data starting from noise conditioned on phenotypic traits and sample conditions.

The versatility of synthetic data conditioned on phenotypic traits and sample conditions extends across a wide array of fields, including genomics, biology, and healthcare. For instance, in Eco-Genomics, synthetic data can simulate gene expression responses to environmental stressors, aiding in the study of changing ecosystem conditions. In Cancer Research, it can help model gene expression profiles under diverse genetic mutations, shedding light on the contributions of specific genes to cancer progression and treatment response. Additionally, in drug response prediction, synthetic data enables the simulation of gene expression patterns under various drug treatments, guiding drug discovery efforts.

These examples underscore the broad applicability and value of synthetic data conditioned on phenotypic traits and sample conditions across diverse domains. It empowers researchers to explore hypotheses, develop models, and make data-driven decisions while effectively addressing data limitations.

# 2 Theoretical Background

## 2.1 Conditional Generative Adversarial Network

Conditional Generative Adversarial Networks (cGANs) are a class of deep generative models that extend the traditional Generative Adversarial Network (GAN) framework [Creswell et al., 2018, Aggarwal et al., 2021]. The GAN consists of two neural networks: a generator (G) and a discriminator (D). The generator is responsible for generating new data, while the discriminator is responsible for distinguishing between real and generated data. The two networks are trained together in an adversarial manner. The generator is trying to fool the discriminator into thinking that the generated data is real, while the discriminator is trying to distinguish between real and generated data.

In a cGAN, both the generator (G) and discriminator (D) networks are conditioned on additional information, often referred to as "conditional labels" or "auxiliary information" (c). The objective of a cGAN is to learn a mapping from the conditional information and a random noise vector (z) to generate data samples that are coherent with the given conditions. The training process involves a minimax game between the generator, which tries to produce data that is indistinguishable from real data, and the discriminator, which aims to correctly distinguish between real and generated samples while considering the conditional information.

In the context of gene expression data analysis, cGANs can be used to generate synthetic gene expression profiles that are consistent with certain conditions or biological contexts.

The generator network in the cGAN is designed to take both random noise (z) and the conditional information (c) as input. By training the cGAN on real expression data, it learns to generate synthetic expression profiles that match the specified conditions. This allows researchers to control and manipulate gene expression patterns based on the provided conditions.

# References

[Aggarwal et al., 2021] Aggarwal, A., Mittal, M., and Battineni, G. (2021). Generative adversarial network: An overview of theory and applications. *International Journal of Information Management Data Insights*, 1(1):100004.

[Creswell et al., 2018] Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., and Bharath, A. A. (2018). Generative adversarial networks: An overview. *IEEE signal processing magazine*, 35(1):53–65.

[Maier et al., 2013] Maier, R., Zimmer, R., and Küffner, R. (2013). A turing test for artificial expression data. *Bioinformatics*, 29(20):2603–2609.