BITS :: Call for Abstracts 2024 - Oral communication

Type	Oral communication
Session	Bioinformatics AI, Models and Tools
Title	Integrating biological networks for bulk RNA-seq profile generation
All Authors	Mongardi S(1), Panaccione F P(1), Pinoli P(1), Masseroli M(1)

Affiliation

(1) Department of Electronics, Information and Bioengineering (DEIB), Politecnico di Milano, Milano, Italy

Motivation

Deep learning models have found extensive application in various cancer-related tasks, owing to their intrinsic ability to handle high-dimensional data as well as learn complex genomics-phenotype relationships, providing accurate and improved diagnostic and prognostic performance. Nonetheless, applications might be limited as deep learning models are notoriously data-hungry, requiring thousands of samples for proper optimization and training. Despite the existence of numerous open-access omics databases, obstacles such as expenses associated with sequencing, labor, and time persist. Data augmentation techniques can be used in these scenarios to increase the number of training instances and potentially develop more robust and accurate prediction models. Currently, we lack "golden standard" transformations that can be applied to existing data to generate new samples without altering the underlying biological meaning. Indeed, generating realistic gene expression profiles is a difficult task: biological systems are highly complex, and it is unclear how biological elements interact. Thus, deep learning-based generative models emerge as a valuable tool for producing realistic transcriptomics data that mirror the statistical distribution of training data points.
In this study, we build a generative neural network model that harnesses prior knowledge in the form of existing biological graphs to enhance the synthetic generation of realistic bulk RNA-seq profiles using gene expression data from The Cancer Genome Atlas (TCGA). Our model architecture incorporates message-passing layers into the generator network of a conditional Generative Adversarial Network (GAN)-based model to learn and exploit the underlying relational structure and feature dependencies in the sample-generation process. Our objective is not only to generate realistic gene expression profiles for data augmentation purposes but also to uncover feature relationships, facilitating post-hoc analyses such as gene knockout effect inference.

Methods

In our experiments, we considered pan-cancer bulk RNA-seq samples from the TCGA project. After preprocessing, the final dataset includes 10,123 samples with expression values for 18,665 genes. We split the dataset, using 80% of the data for training and 20% for testing. Along with the gene expression values, we also considered the donor’s age, the tissue-type (23 distinct types), sex, and disease condition (cancer/not-cancer) as additional covariates to condition the generation of samples on. Our generative architecture builds on the Wasserstein GAN with gradient penalty (WGAN-GP), an extension of the GAN model that uses the Wasserstein distance as a measure of the difference between the distribution of real data and generated data. As for traditional GANs, the WGAN-GPs learn a generative model through an adversarial process involving a critic and a generator network. The WGAN-GP also enforces a Lipschitz constraint on the critic by penalizing the norm of its gradients with respect to interpolated samples between real and generated ones. Our proposed model features a graph-based generator with message-passing layers (graph-convolutional layers) to learn and exploit the relational structure and functional dependencies between features. The generative architecture relies on a pre-initialized input graph summarizing existing biological relations between genes. The protein-protein interaction (PPI) network obtained from STRING served as the initial graph, to which 100 random links were drawn and added for each node to augment the graph. We compared the generative performance of our proposed architecture against WP-GAN model. To assess the generative ability of the different GAN-based models, we used reverse validation, where logistic regression models for disease-condition and tissue classification are trained separately on real and generated samples and tested on real ones, and unsupervised performance indicators (correlation score, precision, recall), evaluating each model across 10 different generation runs (mean and standard deviation).

Results

The WP-GAN had a correlation score of 0.97208 (± 0.00099), precision of 0.99694 (± 0.00138), and recall of 0.85956 (± 0.00522). Our proposed model achieved a correlation score of 0.97293 (± 0.00109), precision of 0.99886 (± 0.00063), and recall of 0.90430 (± 0.00338). The baseline accuracy (model trained on real data) on the test set for binary and multi-class classification are 0.99259 (± 0.0) and 0.96494 (± 0.0). Accuracy values for binary and multi-class classification are 0.65886 (± 0.00807) and 0.94854 (± 0.00217) for the WP-GAN, 0.65857 (± 0.00331) and 0.94483 (± 0.00106) for our proposed architecture. Findings indicate that both architectures achieved similar results in terms of quality of the generated data, with our model yielding almost a 5% increase in recall. This highlights the potential of our graph-based generator to produce realistic gene-expression profiles.

Info

This work was supported by the MUSA – Multilayered Urban Sustainability Action – project, funded by the European Union – NextGenerationEU, under the National Recovery and

Resilience Plan (NRRP) Mission 4 Component 2 Investment Line 1.5: Strengthening of research structures and creation of R&D “innovation ecosystems”, set up of “territorial leaders in R&D”.	
filename	-
Figure	
-	
Availability	-
Dissemination Material	
Social	
-	
Summary	
-	
Corresponding Author	
Name, Surname	Sofia, Mongardi
Email	sofia.mongardi@polimi.it
Submitted on	02.05.2024