Generative Adversarial Neural Networks for a Multiomics Approach in the Mycobacterium Tuberculosis Complex Analysis

Salvador Sánchez Vinces, Ana Marcia de Sá Guimarães, Ronaldo Fumio Hashimoto

University of São Paulo

Abstract

This work presents the application of generative Deep Neural Network methods to Mycobacterium tuberculosis (MTB) gene expression data (with preliminary results). These methods allow generating data with the same distribution as the original samples, as well as facilitating selective data generation of subgroups of original samples, and allowing some degree of manipulation to generate states of gene expression profiles. With such a variety of data, it is possible to establish further processing that facilitates analysis of genetic information (genome and transcriptome). The aim is to deepen the development of this new area of application of generative deep learning methods, studying characteristics and required preprocessing of the input biological data and optimizing the structures of neural networks for searching biologically plausible and integrated results at different levels of genetic information, and thus obtain data of interest in the amount required to make robust inferences using different models (e.g., identification and comparison of phenotypes by co-expression). For implementation and testing of the proposed model, we find convenient to work with MTB as reference microorganism within the MTB complex, for the relatively greater amount of information available, for example they have small but complex genome (approx. 4000 genes) and because their expression mechanisms are relatively well understood (approx. 40% of their genome has been characterized). The generated data (gene expression) were evaluated using qualitative distribution metrics such as histograms and t-SNE, and the effect on gene co-expression models. For both histograms and t-SNE, the generative model achieves a very similar distribution of values compared to original samples. Co-expression analysis shows a positive increase in the number of genes and modules inferred from the generated data when compared to the ones obtained from original data, such as modules neglected by the latter.

Funding: CAPES