Nucleotide-level RNA-Seq data augmentation based on a variational Bayesian framework

Treudsky Antoine*† , Marie-Hélène Mucchielli-Giorgi , Laurent Duval , Pirayre Aurélie , and Frédérique Bidard-Michelot[1]

[1]IFP Energies nouvelles – IFP Energies Nouvelles, Solaize, France – France

Abstract

Data augmentation refers to a corpus of methodologies designed to generate and exploit synthetic samples with statistical properties as faithful to those of the real data as possible. They improve the generalization ability of learning models, particularly in settings where the available data is incomplete, sparse or limited in size.

In the field of omics sciences, data augmentation techniques play a crucial role in mitigating well-known challenges, such as sample scarcity. These techniques strengthen both the robustness of statistical inference and the stability of predictive models.

Among the various classes of approaches for omics data augmentation, deep generative models have demonstrated an enhanced capacity to capture their complex data patterns. These models, based on expressive neural architectures, aim to approximate the joint distribution of the observed data and to generate new, realistic instances. They are particularly useful for RNA-Seq data, which is a high-throughput sequencing approach that produces genomic profiles describing transcriptional activity. Numerous deep generative models use aggregated read count data at the gene level as input variables for the data augmentation model, which implies a loss of granularity and an increased difficulty in effectively learning the model parameters in very low-capacity datasets.

Here, we propose a nucleotide-level data augmentation method based on a variational Bayesian framework. This approach aims to produce synthetic data that are both statistically consistent and biologically plausible, by explicitly integrating contextual information such as nucleotide genomic coordinates and sample-specific characteristics. The goal is to generate new read counts simulating an experimental replication while preserving the underlying dependency structures and the biological relevance of the signal.

This method was applied in the context of differential expression analysis of RNA-Seq data from a {Trichoderma reesei} strain cultivated under two experimental conditions, with three replicates per condition. It successfully preserved expression contrasts among genes known to be differentially expressed, demonstrating its ability to maintain the differential structure of the signal in the augmented data.

These findings suggest that the proposed approach is particularly suitable for analyzing

*Speaker

†Corresponding author: treudsky-l-h.antoine@ifpen.fr

low-capacity datasets, characterized by a limited number of samples and a high-dimensional feature space. In such settings, where other augmentation methods may reach their limits, our approach offers promising perspectives for the application of advanced machine learning algorithms to omics data, thereby enhancing both their statistical robustness and generalization capacity.

Keywords: omics data augmentation, variational autoencoders, VAEs, latent space modeling, data scarcity in omics, transcriptomics augmentation, synthetic omics data