World Academy of Science, Engineering and Technology
International Journal of Mathematical and Computational Sciences
Vol:14, No:12, 2020

On the Utility of Bidirectional Transformers in Gene Expression-Based Classification

Authors : Babak Forouraghi

Abstract : A genetic circuit is a collection of interacting genes and proteins that enable individual cells to implement and perform vital biological functions such as cell division, growth, death, and signaling. In cell engineering, synthetic gene circuits are engineered networks of genes specifically designed to implement functionalities that are not evolved by nature. These engineered networks enable scientists to tackle complex problems such as engineering cells to produce therapeutics within the patient's body, altering T cells to target cancer-related antigens for treatment, improving antibody production using engineered cells, tissue engineering, and production of genetically modified plants and livestock. Construction of computational models to realize genetic circuits is an especially challenging task since it requires the discovery of the flow of genetic information in complex biological systems. Building synthetic biological models is also a time-consuming process with relatively low prediction accuracy for highly complex genetic circuits. The primary goal of this study was to investigate the utility of a pre-trained bidirectional encoder transformer that can accurately predict gene expressions in genetic circuit designs. The main reason behind using transformers is their innate ability (attention mechanism) to take account of the semantic context present in long DNA chains that are heavily dependent on the spatial representation of their constituent genes. Previous approaches to gene circuit design, such as CNN and RNN architectures, are unable to capture semantic dependencies in long contexts, as required in most real-world applications of synthetic biology. For instance, RNN models (LSTM, GRU), although able to learn long-term dependencies, greatly suffer from vanishing gradient and low-efficiency problem when they sequentially process past states and compresses contextual information into a bottleneck with long input sequences. In other words, these architectures are not equipped with the necessary attention mechanisms to follow a long chain of genes with thousands of tokens. To address the above-mentioned limitations, a transformer model was built in this work as a variation to the existing DNA Bidirectional Encoder Representations from Transformers (DNABERT) model. It is shown that the proposed transformer is capable of capturing contextual information from long input sequences with an attention mechanism. In previous works on genetic circuit design, the traditional approaches to classification and regression, such as Random Forrest, Support Vector Machine, and Artificial Neural Networks, were able to achieve reasonably high R2 accuracy levels of 0.95 to 0.97. However, the transformer model utilized in this work, with its attention-based mechanism, was able to achieve a perfect accuracy level of 100%. Further, it is demonstrated that the efficiency of the transformer-based gene expression classifier is not dependent on the presence of large amounts of training examples, which may be difficult to compile in many real-world gene circuit designs.

Keywords : machine learning, classification and regression, gene circuit design, bidirectional transformers
Conference Title : ICSRD 2020 : International Conference on Scientific Research and Development
Conference Location : Chicago, United States
Conference Dates : December 12-13, 2020