ECAI Paper ID: 1662
Paper title: GeneMask: Fast Pretraining of Gene Sequences to Enable Few-Shot Learning

Short summary: The paper considers the Human genome as a stream of characters and proposes a pretraining operation over the Human Reference Genome. There is a small problem, though, unlike MLM in NLP, there is no word/sentence demarcation here. This paper proposes the GeneMask strategy, a PMI-based method that selects relevant spans in the genes and shows that pretraining using those spans improves few-shot performances over the SOTA models (DNABert and LOGO) over four benchmark gene sequence classification datasets in five few-shot settings (10 to 1000-shot). We also observe a strong correlation between top-ranked PMI tokens and conserved DNA sequence motifs, which may indicate the incorporation of latent genomic information.

Long summary: Large-scale language models such as DNABert and LOGO aim to learn optimal gene representations and are trained on the entire Human Reference Genome. However, standard tokenization schemes involve a simple sliding window of tokens like k-mers that do not leverage any gene-based semantics and thus may lead to (trivial) masking of easily predictable sequences, and subsequently inefficient Masked Language Modeling (MLM) training. Therefore, we propose a novel masking algorithm, GeneMask, for MLM training of gene sequences, where we randomly identify positions in a gene sequence as mask centers and locally select the span around the mask center with the highest Normalized Pointwise Mutual Information (NPMI) to mask. We observe that in the absence of human-understandable semantics in the genomics domain (in contrast, semantic units like words and phrases are inherently available in NLP). GeneMask-based models substantially outperform the SOTA models (DNABert and LOGO) over four benchmark gene sequence classification datasets in five few-shot settings (10 to 1000-shot). More significantly, the GeneMask-based DNABert model is trained for less than one-tenth of the number of epochs of the original SOTA model. We also observe a strong correlation between top-ranked PMI tokens and conserved DNA sequence motifs, which may indicate the incorporation of latent genomic information. Gene sequence classification is a challenging problem and requires a tremendous engineering effort even to develop the experimental setup. Therefore, the codes (including trained models) and datasets are made publicly available at https://github.com/roysoumya/GeneMask.