Abstract

With the rapid development of machine learning and molecular biology, the two have begun to influence each other with exponential strength. Biological experiments are crucial for understanding the background of many complex diseases – however, they require a long time and are expensive. In this dissertation, various machine learning methods were applied, particularly deep learning methods, to predict the experiments mentioned above. DNA sequencing is one of the most affordable and common biological experiment that this dissertation is based on. With the development of technology, it is becoming more and more affordable, to the point that hundreds of thousands of sequenced human genomes are available online. The work started with assessing the importance of the sequence – and showing its complex causative relation with various diseases. Based on that, an algorithm for discovering variants was created – to provide researchers with an option to obtain interesting, high-quality variants in reduced time. The algorithm was enhanced with a deep feed-forward neural network based on eight (8) state-of-the-art algorithms for variant calling.

Further on, a model based on DNABERT architecture was created to discover spatial chromatin conformation, namely ChIA-PET loops. The work was then extended, as DNABERT and the method were limited. To solve those limitations, the HiCDiffusion model was created – an algorithm that connects the modern approach of encoder-decoder architecture along with a transformer for context learning and enhances the results by applying conditional diffusion – to improve the quality of the final result so that it is indistinguishable from the original experimental Hi-C data. The study was finished by showing relations between the spatial conformation of chromatin within the nucleus and gene expression – as models used for gene prediction very often include only local sequence (e.g. 20kbp around transcription starting site) – and are unaware of the sequences that are far away in the linear sense – though very close in 3D, thus interacting. The study, therefore, covers the whole journey of the DNA sequence – from the discovery of the variants, through applying them to obtain spatial conformation, up to predicting gene expression which, in case of being deregulated, is often at the very core of genetic diseases.

Keywords: Artificial Intelligence (AI), Machine Learning (ML), Deep Learning (DL), Large Language Models (LLMs), 3D Genomics, Chromatin, Gene Expression, Structural Variants