Institute of Biomedical Informatics

Graz University of Technology

Dr. Peter Ulz
Stremayrgasse 16/I
8010 Graz, Austria
Tel. +43 316 873-5343
peter.ulz@tugraz.at
http://genome.tugraz.at

Bachelor's Thesis

Deep Learning in Bioinformatics for assessing the functional impact of single-nucleotide variants in non-coding regions

Deep learning has transformed biological research in many different aspects [1, 2]. The increase in data generation capabilities of modern biological research has further exacerbated this trend. One class of deep learning models aims at resolving biological context and annotating functional elements by sequence data alone. These include Google's Alpha Genome [3], the Nucleotide Transformer [4], EVO2 [5] , GPN-MSA [6] and more. While earlier efforts have focussed on a single class of regulatory elements, these modern approaches aim at predicting many different classes of regulatory elements, including splice-sites, enhancers, topologically associated domains (TADs) and transcription factor binding sites.

The overall goal of this thesis is to get an overview of deep learning methods in bioinformatics with a focus on methods that have the potential to assess the biological impact of point mutations outside the coding sequences:

1. Perform a literature survey on deep-learning technologies on genomic data and their ability to assess biological context when only given a DNA sequence.
2. Research possible modes of how single-nucleotide variants outside the protein-coding regions can have an impact on cellular processes. (e.g. transcription factor binding sites, chromatin topology and CTCF sites, splice sites, enhancers, ...)
3. Assess different approaches in the deep learning models above and detail how the differences in training and representation/embedding can lead to differences in their ability to detect certain classes of variants.
4. Focus on one approach that takes sequence and outputs biological annotations. Demonstrate in a single example/notebook how we can get from sequence to annotations.
5. Compile a list of single-nucleotide variants that are outside the noncoding genome (from literature, ClinVar [7], dbSNP, …) and show how the biological annotation of the sequence changes when assessing both the wildtype and the mutant sequence.
6. Summarize your work in a written thesis.

References

[1] Whalen S, Schreiber J, Noble W, Pollard K. **Navigating the pitfalls of applying machine learning in genomics**. *Nature Reviews Genetics*. 2022

[2] Ching T, Himmelstein DS, Beaulieu-Jones BK et al. **Opportunities and obstacles for deep learning in biology and medicine.** *J. R. Soc. Interface.* 2018

[3] Avsec Z, Latysheva N, Cheng J et al. **AlphaGenome: advancing regulatory variant effect prediction with a unified DNA sequence model.** *BioRxiv* 2025

[4] Dalle-Torre H, Gonzalez L, Mendoza-Revilla J et al. **Nucleotide Transformer: building and evaluating robust foundation models for human genomics.** *Nature Methods* 2024

[5] Brixi G, Durrant Mg, Ku J et al. **Genome modeling and design across all domains of life with Evo 2.** *BioRxiv* 2025

[6] Benegas G, Albors C, Aw AJ et al. **A DNA language model based on multispecies alignment predicts the effects of genome-wide variants.** *Nature Biotechnology*. 2025

[7] **ClinVar** at the National Center of Biotechnology Information https://www.ncbi.nlm.nih.gov/clinvar/ 2025