AI and the Dark Genome: Using protein-DNA structure modelling and genomic language models to predict the impacts of non-coding genetic variation

Abstract

This project aims to improve computational predictions of the functional impacts of non-coding genetic variants, which play a key role in gene regulation and disease risk. By combining recent advances in protein structure prediction and genomic language models, we will develop methods to predict how non-coding variants affect transcription factor binding. This will be benchmarked against high-throughput experimental datasets and applied to clinically relevant DNA-binding proteins. The project will provide a new framework for prioritising non-coding variants in disease studies, contributing to personalised genomic medicine and uncovering novel therapeutic targets.

Introduction

Diversity across the human population, and individual susceptibility to disease, is influenced by genetic factors. The overwhelming majority of human genetic variants occur in the Dark Genome, the 98% of our DNA that does not code for proteins. Non-coding genetic variants in regulatory elements are thought to influence how transcription factors bind to DNA, and can thus be related to changes in gene expression. Thousands of genome-wide association studies (GWAS) have uncovered genetic variants linked to human traits and disease, with the goal of identifying new therapeutic targets and enabling personalised genomic medicine. However, these studies are unable to tell us which variants are causal or elucidate molecular mechanisms, which together has limited translational applications. While experimental approaches have been developed to bridge the gap between genetic variation and disease phenotypes, by identifying functional variants and elucidating their mechanisms, these approaches are only achievable in limited cellular contexts, are limited in scale, and are laborious, technically challenging, and expensive to perform.

In recent years, there have been major advances in predicting coding variant effects, particularly for highly penetrant alleles associated with monogenic disorders. However, current approaches for predicting non-coding variant effects show far less utility, for several reasons. First, protein structures are often used in coding variant effect prediction, but structures are rarely available for non-coding regions of DNA. Second, while much of the success in coding variant effect prediction comes from evolutionary information derived from multiple protein sequence alignments, nucleotide sequence alignments in non-coding regions will tend to be much shallower, due to fewer distinct homologous sequences being available. Finally, non-coding variants tend to have weaker effects than coding variants, which also makes them more difficult to predict using evolutionary information because of the lower levels of constraint.

Research Challenge

Our ability to interpret non-coding genetic variation is currently poor. Our research challenge is to ***improve computational predictions of the functional impacts of non-coding variants***. To do this, we will be taking advantage of two major advances in the field of biomolecular artificial intelligence:

1) **The ability to computationally predict the three-dimensional structures of protein:DNA complexes.** AlphaFold2 revolutionised protein structure prediction. AlphaFold3 has extended this to protein:DNA complexes, allowing us to build three dimensional models of essentially any transcription factor with its associated DNA regulatory region (Abramson et al, 2024).

2) **The development of genomic language models.** Large language models based on protein sequences have had great success in coding variant effect prediction, without requiring sequence alignments or having any preconceptions about sequence-function relationships (Livesey and Marsh, 2023). Similar strategies are being applied to DNA sequences in the form of genomic language models, which have the potential to overcome the limitations of sequence-alignment-based approaches (Benegas et al, 2023).

Data & Methodology

<u>*Development phase (year 1 – 2):*</u>

1. Developing a structural approach for predicting non-coding variant impact on transcription factor binding. The student will use AlphaFold3 and other emerging methodologies for predicting protein:DNA complex structures to build 3D structural models of DNA-bound transcription factors. They will then use molecular modelling to determine the energetic impacts of nucleotide variants on protein:DNA binding, similar to the approaches we have used for protein variation (Gerasimavicius et al, 2023). The optimal strategy for predicting non-coding variant effects will be established through comparison to high-throughput experimental measurements of variant binding preference, generated by us, and from publicly available datasets (e.g. Yan et al, 2021).

2. Genomic language models. The student will benchmark various genomic language models (e.g. Benegas et al, 2023) for their agreement with high-throughput measurements of transcription factor binding, and with massively parallel reporter assays (Zhao et al, 2023), analogous to the approach we have used for assessing protein language models and other variant effect predictors (Livesey and Marsh, 2024). The student will then combine the structural and language models to establish an improved strategy for predicting functional non-coding variants.

<u>*Delivery phase (year 3 – 4):*</u>

3. A DNA-binding factor centric approach. The student will use a set of druggable and clinically relevant DNA-binding factors to predict altered binding events due to non-coding variants across human populations. Genome-scale scans for functional variants will be undertaken to identify putative non-coding variants with altered function. This could lead to prediction of variability in clinical responses to drugs in the human population.

*4. A cell-type centric approach***.** The student will identify non-coding variants that alter interactions of multiple DNA-binding factors within a single cell type, at a whole genome scale. This could be used to develop gene regulatory networks to model effects of combinatorial non-coding genetic variant effects on signalling.

5. A disease-centric approach. The student will analyse non-coding variants associated with cardiovascular, immune/autoimmune, and neurological/neuropsychiatric diseases to identify putative functional variants. This will identify binding factors to infer mechanism that represent putative therapeutic targets in disease.

Responsible AI/Ethical Considerations

The research project utilises existing and publicly available datasets for training. No ethical approval is required. As datasets, including genetic studies, are publicly available, there are no data protection concerns.

Expected Outcome & Impact

This project will deliver a novel computational framework to predict the functional impact of non-coding genetic variants using structure modelling and genomic language models. We will identify variants that affect transcription factor binding and gene regulation, providing mechanistic insights into disease-associated non-coding regions. This will prioritise functional non-coding variants from large-scale genetic studies, guiding the development of new therapeutic targets and personalised treatment strategies. Additionally, the tools developed will serve as a valuable resource for the genetics and genomic medicine community, broadening our ability to interpret non-coding regions across diverse human populations.

For the student, this project offers a comprehensive learning experience in cutting-edge computational biology, structural bioinformatics, and functional genomics. The student will gain proficiency in using state-of-the-art AI-driven tools like AlphaFold3 and genomic language models, along with molecular modelling and high-throughput data analysis. They will also acquire critical skills in integrating diverse datasets, benchmarking predictive models, and translating findings into biological and clinical contexts, preparing them for a career in computational genomics and precision medicine.

References

Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., ... & Jumper, J. M. (2024). Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 1-3.

Benegas, G., Batra, S. S., & Song, Y. S. (2023). DNA language models are powerful predictors of genome-wide variant effects. *Proceedings of the National Academy of Sciences*, 120(44), e2311219120.

Gerasimavicius, L., Livesey, B. J., & Marsh, J. A. (2023). Correspondence between functional scores from deep mutational scans and predicted effects on protein stability. *Protein Science*, 32(7), e4688.

Livesey, B. J., & Marsh, J. A. (2023). Advancing variant effect prediction using protein language models. *Nature Genetics*, *55*(9), 1426-1427.

Livesey, B. J., & Marsh, J. A. (2024). Variant effect predictor correlation with functional assays is reflective of clinical classification performance. *bioRxiv*, 2024-05.

Yan, J., Qiu, Y., Ribeiro dos Santos, A. M., Yin, Y., Li, Y. E., Vinckier, N., ... & Ren, B. (2021). Systematic analysis of binding of transcription factors to noncoding variants. *Nature*, 591(7848), 147-151.

Zhao, J., Baltoumas, F. A., Konnaris, M. A., Mouratidis, I., Liu, Z., Sims, J., ... & Ahituv, N. (2023). MPRAbase: a massively parallel reporter assay database. *bioRxiv*.