Deepsight: Binary Classification of Non-Coding Variants

Lucas Schwoebel Kabeen Kim Compton Ross
Maja Noack

October 2, 2024

Abstract

Non-coding single nucleotide polymorphisms (SNPs) play a crucial role in human disease susceptibility, yet their functional consequences remain largely unknown. Traditional variant effect prediction tools struggle to accurately assess non-coding variants due to their reliance on coding-centric features. To address this, we propose developing a Convolutional Neural Network (CNN) that integrates genomic and epigenomic features, including nucleotide composition, evolutionary conservation, genomic context, and allele frequencies and predicts the disease potential of a given non-coding variant.

Project Overview and Approach

SNPs are the most common type of genetic variation in humans. They occur with a frequency of 1 every 1000 base pairs in the human genome with most of them residing in non-coding regions of the DNA[6][11]. Historically these regions have been understudied although there is growing evidence suggesting that a significant number of non-coding SNPs play a role in disease susceptibility [17]. Due to the sheer amount of unique variants, interpretation of SNPs remains a challenge. Experimental validation of variant effects is often slow, costly, and requires significant resources. Over the last decade several variant effect prediction tools have been published including CADD[10], Mutationtaster2021 [15], PolyPhen-2 [1] and SIFT [14]. While these variant effect prediction tools have been valuable for analyzing coding variants, their accuracy in predicting the functional consequences of non-coding SNPs remains limited [16].

In recent years the decreasing cost and increasing throughput of next-generation sequencing have led to a dramatic accumulation of genomic and epigenetic information. Paired with the advances in deep learning, this presents new opportunities for understanding the impact of non-coding variants.

Despite these advancements, predicting the effects of individual genomic variants remains challenging. For example, Sasse et al. found that Enformer, a state-of-the-art deep learning model, struggled to accurately predict gene expression changes in a large cohort of individuals [12].

To address these limitations, we propose developing a novel deep learning model capable of accurately predicting the functional significance of non-coding variants. To achieve the goal of predicting the functional significance of non-coding variants, we will likely develop a convolutional neural network (CNN) that performs binary classification to determine whether a given variant is benign or potentially harmful. Our model will integrate several genomic and epigenomic features like nucleotide composition, evolutionary conservation, genomic context features and allele frequencies to enhance the model's predictive power. Beyond simply classifying variants, the CNN may also provide valuable insights into the functional implications of non-coding variants.

Datasets

The data for this project will come from a set of databases containing information on SNPs and genomic data, divided into two subsets based on the functionality of the data. The first group, consisting of databases like ENCODE [2] and the UCSC Genome Browser (with PhastCons and PhyloP) [9], provides functional and regulatory data essential for understanding how SNPs affect gene regulation. The second group, including databases like CLINVAR [7], NCBI dbSVP [13], ncVarDB [4], and the Ensembl Genome Browser [5], offers variant-specific data, such as clinical significance and functional effects of SNPs, which is critical for understanding their impact on phenotype and disease risk.

Pre-processing the data will involve normalizing the different formats used by each dataset and filtering out irrelevant information. Datasets used in the training process will need to be integrated with variant-specific data to ensure consistency in annotations. Quality control processes, such as FastQC, will check the sequencing read quality, and DNA mapping will determine the location of genes and other sequences. Difficulties may arise from missing or inconsistent variant annotations across datasets (for example use of different genome versions), imbalanced datasets and avoiding predicted annotations for the training data. Additionally the computational resources required for training neural networks, as well as any complications that may stem from handling large volumes of data can also be challenging.

Expected Results

This project aims to develop an enhanced predictive system for outcomes associated with non-coding SNPs through the application of deep learning methods like CNNs. Our work seeks to provide novel insights into associations between non-coding SNPs and diseases, potentially contributing to the understanding of their clinical implications. Furthermore we will conduct a comparative analysis of our proposed model against established frameworks such as DeepSEA [18], DanQ [8] or Enformer [3]. If however, an improvement will not be promising or realistic, the models will be compared to a yet to find benchmark.

The primary metrics for evaluation will include accuracy, precision, recall, and the F1 score, which will give insight into the balance between true positive predictions and false positive/negative results. To ensure robustness of our work, the system will be validated on an independent dataset, which contains clinically relevant non coding variants as well as benign controls.

References

[1] Ivan A Adzhubei et al. "A method and server for predicting damaging missense mutations". In: *Nature Methods* 7.4 (Apr. 2010), pp. 248–249. ISSN: 1548-7105. DOI: 10.1038/nmeth0410-248. URL: http://dx.doi.org/10.1038/nmeth0410-248.

[2] "An integrated encyclopedia of DNA elements in the human genome". In: *Nature* 489.7414 (Sept. 2012), pp. 57–74. ISSN: 1476-4687. DOI: 10.1038/nature11247. URL: http://dx.doi.org/10.1038/nature11247.

[3] Žiga Avsec et al. "Effective gene expression prediction from sequence by integrating long-range interactions". In: *Nature Methods* 18.10 (Oct. 2021), pp. 1196–1203. ISSN: 1548-7105. DOI: 10.1038/s41592-021-01252-x. URL: http://dx.doi.org/10.1038/s41592-021-01252-x.

[4] Harry Biggs et al. "ncVarDB: a manually curated database for pathogenic non-coding variants and benign controls". In: *Database* 2020 (2020). ISSN: 1758-0463. DOI: 10.1093/database/baaa105. URL: http://dx.doi.org/10.1093/database/baaa105.

[5] Peter W Harrison et al. "Ensembl 2024". In: *Nucleic Acids Research* 52.D1 (Nov. 2023), pp. D891–D899. ISSN: 1362-4962. DOI: 10.1093/nar/gkad1049. URL: http://dx.doi.org/10.1093/nar/gkad1049.

[6] Leonid Kruglyak and Deborah A Nickerson. "Variation is the spice of life". In: *Nature Genetics* 27.3 (Mar. 2001), pp. 234–236. ISSN: 1546-1718. DOI: 10.1038/85776. URL: http://dx.doi.org/10.1038/85776.

[7] Melissa J. Landrum et al. "ClinVar: public archive of relationships among sequence variation and human phenotype". In: *Nucleic Acids Research* 42.D1 (Nov. 2013), pp. D980–D985. ISSN: 1362-4962. DOI: 10.1093/nar/gkt1113. URL: http://dx.doi.org/10.1093/nar/gkt1113.

[8] Daniel Quang and Xiaohui Xie. "DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences". In: *Nucleic Acids Research* 44.11 (Apr. 2016), e107–e107. ISSN: 1362-4962. DOI: 10.1093/nar/gkw226. URL: http://dx.doi.org/10.1093/nar/gkw226.

[9] Brian J Raney et al. "The UCSC Genome Browser database: 2024 update". In: *Nucleic Acids Research* 52.D1 (Nov. 2023), pp. D1082–D1088. ISSN: 1362-4962. DOI: 10.1093/nar/gkad987. URL: http://dx.doi.org/10.1093/nar/gkad987.

[10] Philipp Rentzsch et al. "CADD: predicting the deleteriousness of variants throughout the human genome". In: *Nucleic Acids Research* 47.D1 (Oct. 2018), pp. D886–D894. ISSN: 1362-4962. DOI: 10.1093/nar/gky1016. URL: http://dx.doi.org/10.1093/nar/gky1016.

[11] Ravi Sachidanandam et al. “A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms”. In: *Nature* 409.6822 (Feb. 2001), pp. 928–933. ISSN: 1476-4687. DOI: 10.1038/35057149. URL: http://dx.doi.org/10.1038/35057149.

[12] Alexander Sasse et al. “Benchmarking of deep neural networks for predicting personal gene expression from DNA sequence highlights shortcomings”. In: *Nature Genetics* 55.12 (Nov. 2023), pp. 2060–2064. ISSN: 1546-1718. DOI: 10.1038/s41588-023-01524-6. URL: http://dx.doi.org/10.1038/s41588-023-01524-6.

[13] S. T. Sherry. “dbSNP: the NCBI database of genetic variation”. In: *Nucleic Acids Research* 29.1 (Jan. 2001), pp. 308–311. ISSN: 1362-4962. DOI: 10.1093/nar/29.1.308. URL: http://dx.doi.org/10.1093/nar/29.1.308.

[14] Ngak-Leng Sim et al. “SIFT web server: predicting effects of amino acid substitutions on proteins”. In: *Nucleic Acids Research* 40.W1 (June 2012), W452–W457. ISSN: 1362-4962. DOI: 10.1093/nar/gks539. URL: http://dx.doi.org/10.1093/nar/gks539.

[15] Robin Steinhaus et al. “MutationTaster2021”. In: *Nucleic Acids Research* 49.W1 (Apr. 2021), W446–W451. ISSN: 1362-4962. DOI: 10.1093/nar/gkab266. URL: http://dx.doi.org/10.1093/nar/gkab266.

[16] Xiaoyu Wang et al. “Deep learning approaches for non-coding genetic variant effect prediction: current progress and future prospects”. In: *Briefings in Bioinformatics* 25.5 (July 2024). ISSN: 1477-4054. DOI: 10.1093/bib/bbae446. URL: http://dx.doi.org/10.1093/bib/bbae446.

[17] Feng Zhang and James R. Lupski. “Non-coding genetic variants in human disease: Figure 1.” In: *Human Molecular Genetics* 24.R1 (July 2015), R102–R110. ISSN: 1460-2083. DOI: 10.1093/hmg/ddv259. URL: http://dx.doi.org/10.1093/hmg/ddv259.

[18] Jian Zhou and Olga G Troyanskaya. “Predicting effects of noncoding variants with deep learning–based sequence model”. In: *Nature Methods* 12.10 (Aug. 2015), pp. 931–934. ISSN: 1548-7105. DOI: 10.1038/nmeth.3547. URL: http://dx.doi.org/10.1038/nmeth.3547.