Self-Attention Mechanisms as Representations for Gene Interaction Networks in Hypothesis-Driven Gene-based Transformer Genomics AI Models

Hong Qin
School of Data Science, Department of Computer Science, Old Dominion University
hqin@odu.edu

Abstract

In this position paper, we propose a framework for hypothesis-driven genomic AI using self-attention mechanisms in gene-based transformer models to represent gene interaction networks. Hypotheses can be introduced as attention masks in these transformer models with genes treated as tokens. This approach can bridge the gap between genotypic data and phenotypic observations by using prior knowledge-based masks in the transformer models. By using attention masks as hypotheses to guide the model fitting, the proposed framework can potentially assess various hypotheses to determine which best explains the experimental observations. The proposed framework can enhance the interpretability and predictive power of genomic AI to advance personalized medicine and promote healthcare equity.

Gene-based Transformer Genomics Models

Transformer models, initially developed for natural language processing (NLP), often treat words as discrete tokens and use self-attention mechanisms to capture the relationships between these tokens (Vaswani et al. 2017).

Transformer models have become an important tool in genomic data analysis because of their ability to capture long-range dependencies within sequences (Choi et al. 2023). Transformers have been applied successfully to various tasks, including sequence and site prediction, gene expression and phenotype prediction, ncRNA and circRNA studies, transcription process insights, and multi-omics integration.

Gene-based transformers, exemplified by scGPT (single-cell Generative Pre-trained Transformer), are especially stated for single-cell data analysis (Cui et al. 2024). scGPT builds on the transformer architecture with stacked layers and multi-head attention mechanisms, incorporating specialized attention masks for generative training on single cell sequencing data (Cui et al. 2024). Pre-trained on a dataset of over 33 million normal human cells from various organs, scGPT has learned comprehensive cell and gene representations. It processes gene tokens, expression values, and condition tokens, including modality, batch, and perturbation conditions.

Key advantages of scGPT include its ability to capture meaningful biological insights into genes and cells, enable transfer learning for improved performance on diverse tasks, facilitate joint representation learning for multi-modal data integration, and demonstrate robust biological conservation during batch correction.

Self-Attention is a Representation of Gene Interaction Networks

Gene interactions are often modeled using weighted adjacency matrices, where nodes represent genes and edges represent the interactions between them, weighted by the strength or significance of these interactions. This matrix-based approach allows for a comprehensive representation of the gene network, facilitating the analysis of how genes influence one another. The self-attention mechanism in transformers parallels this method by calculating attention scores that weigh the importance of each gene in the context of others, effectively creating a dynamic and context-aware weighted adjacency matrix within the transformer model.

The self-attention mechanism in scGPT and other gene-based transformers effectively models gene interaction networks by capturing complex dependencies between genes. Treating genes as tokens, the self-attention process identifies influential genes for predicting others' expression, updating each gene's representation based on its relationships with others. Similar to a weighted adjacency matrix, the attention map generated by self-attention highlights the influence of each gene on others, essential for understanding the genotype-to-phenotype relationship (Cui et al. 2024).

When transformer models are pre-trained on a large and heterogeneous dataset, the self-attention mechanism effectively captures a generalizable gene interaction network from the input genomics data. This process allows the model to learn complex dependencies and interactions between

genes that are broadly applicable across different contexts. As demonstrated by scGPT, the self-attention layers of pre-trained gene transformer models are capable of generalizing well enough to predict a wide range of phenotypic observations. This capability extends to various downstream tasks such as disease prediction, drug response modeling, and understanding developmental processes, showcasing the versatility and robustness of the model in different biological scenarios (Cui et al. 2024).

Hypotheses can be Introduced as Attention Masks in Gene-based Transformer Models

Transformer models use attention mechanisms to weigh the importance of different input elements. In gene-based transformer models, attention masks can incorporate prior knowledge and hypotheses, controlling how the model attends to different genes or gene combinations.

For the sake of illustration, hypotheses can be based on gene interactions, biological pathways, and functional groups as defined by Gene Ontology. These pathways and functional groups can be represented by adjacency matrices that serve as attention masks in transformer models.

Weighted adjacency matrices can also be constructed based on prior evidence, which can then be used as attention masks to adjust the transformer model to focus on candidate gene interactions. By comparing transformer models with attention masks designed based on different hypotheses, it may be possible to evaluate the best gene-based transformer models for a specific phenotype, which can then guide further experimental efforts.

Comparison with Other Hypothesis-driven AI Approaches

Several hypothesis-driven AI approaches have been proposed to integrate scientific hypotheses and domain-specific knowledge into AI models (Xianyu et al. 2024).

OncoNPC is a tool for classifying cancers of unknown primary origin (Moon et al. 2023). OncoNPC incorporates prior knowledge of cancer types and their molecular characteristics to improve accuracy in identifying the tissue of origin for metastatic tumors. P-Net uses a network-based approach to stratify prostate cancer patients (Elmarakeby et al. 2021). By integrating known molecular pathways and genetic interactions specific to prostate cancer, P-Net can identify clinically relevant patient subgroups with distinct prognoses. SPIN-AI is an algorithm designed to identify genes that influence tumor organization (Meng-Lin et al. 2023). SPIN-AI incorporates spatial transcriptomics data and known gene regulatory networks to uncover novel insights into how certain genes affect the spatial arrangement of cells within tumors. D-Cell employs graph neural networks based on Gene Ontologies to model cellular processes (Ma et al. 2018). By leveraging the hierarchical structure of biological knowledge represented in Gene Ontologies, D-Cell can predict cellular phenotypes and drug responses in cancer cells. These hypothesis-driven AI approaches demonstrate the potential for combining domain expertise with advanced machine learning techniques to address complex challenges in cancer research and improve patient outcomes.

Hypothesis-masked gene-based transformer models offer a clear advantage compared to these hypothesis-driven AI models. Gene-based transformers leverage the self-attention mechanism to capture complex relationships and interactions between genes, which allows them to model molecular networks like gene regulation networks more effectively. By applying prior knowledge masks to the self-attention matrix, transformers with hypothesis masks can simulate specific interactions or gene knockouts, providing a more detailed and hypothesis-driven analysis of gene interactions and their effects on cancer development and treatment responses.

Challenges and Future Directions of Hypothesis-driven Gene-based Transformer Models

It would be challenging to strike the right balance between incorporating hypotheses and allowing the model to learn patterns from data. Imposing too much prior knowledge or hypotheses involving large groups of genes might overly constrain the model, making training difficult.

Current gene-based transformer models often do not adequately account for sequence polymorphisms and genetic diversity. New approaches are needed to incorporate these aspects to improve their predictive accuracy and applicability across different populations and species. Accounting for polymorphism and genetic diversity would be an essential factor for AI models with social and healthcare equity. One potential approach may be a hybrid approach to combine the attention mechanism of the transformer model with the nucleotide-based convolutional models as illustrated in a recent work (Nguyen et al. 2024). Most gene-based transformer models are not designed as cross-species gene-language models. Developing a universal gene-based language model that can accommodate all gene families across the tree of life would be a significant advancement. Such a model could enhance our understanding of evolutionary relationships and functional genomics across species.

Classica systems biology is often based on theoretic models with specific parameters and nuances. There is great potential in connecting AI-driven models with classical systems biology to create a data-driven systems biology framework. This integration could lead to more robust and comprehensive models that leverage both data-driven insights and theoretical foundations, potentially transforming the field of systems biology and precision medicine.

Acknowledgements

HQ thanks USA NSF award 2200138, a catalyst award from the USA National Academy of Medicine, and internal support of the Old Dominion University.

References

Choi, S. R. and M. Lee 2023. Transformer Architecture and Attention Mechanisms in Genome Data Analysis: A Comprehensive Review. Biology (Basel) 12(7).10.3390/biology12071033

Cui, H., C. Wang, H. Maan, K. Pang, F. Luo, N. Duan and B. Wang 2024. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. Nat Methods.10.1038/s41592-024-02201-0

Elmarakeby, H. A., J. Hwang, R. Arafeh, J. Crowdis, S. Gang, D. Liu, S. H. AlDubayan, K. Salari, S. Kregel, C. Richter, T. E. Arnoff, J. Park, W. C. Hahn and E. M. Van Allen 2021. Biologically informed deep neural network for prostate cancer discovery. Nature 598(7880): 348-352.10.1038/s41586-021-03922-4

Ma, J., M. K. Yu, S. Fong, K. Ono, E. Sage, B. Demchak, R. Sharan and T. Ideker 2018. Using deep learning to model the hierarchical structure and function of a cell. Nat Methods 15(4): 290-298.10.1038/nmeth.4627

Meng-Lin, K., C.-Y. Ung, C. Zhang, T. M. Weiskittel, P. Wisniewski, Z. Zhang, S.-H. Tan, K.-S. Yeo, S. Zhu, C. Correia and H. Li 2023. SPIN-AI: A Deep Learning Model That Identifies Spatially Predictive Genes. Biomolecules 13(6): 895

Moon, I., J. LoPiccolo, S. C. Baca, L. M. Sholl, K. L. Kehl, M. J. Hassett, D. Liu, D. Schrag and A. Gusev 2023. Machine learning for genetics-based classification and treatment response prediction in cancer of unknown primary. Nature Medicine 29(8): 2057-2067.10.1038/s41591-023-02482-6

Nguyen, E., M. Poli, M. G. Durrant, A. W. Thomas, B. Kang, J. Sullivan, M. Y. Ng, A. Lewis, A. Patel, A. Lou, S. Ermon, S. A. Baccus, T. Hernandez-Boussard, C. Ré, P. D. Hsu and B. L. Hie 2024. Sequence modeling and design from molecular to genome scale with Evo. bioRxiv: 2024.2002.2027.582234.10.1101/2024.02.27.582234

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser and I. Polosukhin 2017. Attention is All You Need. Advances in Neural Information Processing Systems 30

Xianyu, Z., C. Correia, C. Y. Ung, S. Zhu, D. D. Billadeau and H. Li 2024. The Rise of Hypothesis-Driven Artificial Intelligence in Oncology. Cancers 16(4): 822