Genomic Sequence Generation and Drug Discovery using NLP and Generative AI: A Review

Prakash Kumar Sarangi, Shashi Kant Gupta

Lincoln University College, 47301, Petaling Jaya, Selangor Darul Ehsan Malaysia. Email: prakashsarangi89@gmail.com., shashigupta@lincoln.edu.my

Abstract: Modern artificial intelligence developments have revolutionized genomics and drug discovery research by using Natural Language Processing and Generative AI technologies. The primary challenge involves two tasks which require scientists to analyze extensive genomic data and create biologically accurate genetic sequences. The review examines genomic sequence modeling through three different types of models which include transformer-based models and diffusion models and autoregressive architectures. The research demonstrates that DNA sequences function as language which allows for the process of tokenization and the development of contextual understanding and sequence creation. The research results demonstrate that large language models (LLMs) enhance the ability to predict regulatory elements and interpret variants and create synthetic sequences. The drug discovery process uses generative models to create molecular structures and identify biological targets and forecast protein structures. The field of personalized medicine uses bioinformatics automation to create disease prediction systems. The review identifies several limitations which prevent complete understanding of the research results while putting forward multimodal learning and explainable AI as future research paths.

Keywords: Genomic sequences; Natural language Processing; Generative AI; Drug discovery; Transformer models

Introduction

Genomics has developed into a data-heavy field which relies on advanced sequencing technologies from next-generation sequencing. The human genome contains more than 3 billion base pairs which create major difficulties for computational analysis. Bioinformatics methods from the past fail to handle large data sets and complex systems, which creates a need for superior artificial intelligence-based solutions. Recent research shows that DNA sequences function as a formal language system which uses nucleotide tokens (A T C G) as its building blocks to enable natural language processing methods that include tokenization and embedding and sequence modeling. The genomic data scientists successfully applied BERT and GPT architectures through transformer-based systems to enhance contextual comprehension and functional region prediction [1].

Generative AI extends these existing capabilities because it enables scientists to create new genomic sequences and supports their efforts in drug discovery. The models enable researchers to generate scientific hypotheses and design experiments because they learn fundamental biological patterns.

Related Work

The recent studies which have been conducted show that natural language processing research and genomics research are increasingly developing common ground between their two fields of study. The research findings demonstrate that genomic sequence analysis benefits from transformer architectures because these systems can maintain performance while processing extended sequences of data.

NLP-based Genomic Models for Regulatory Prediction and Chromatin Accessibility

Natural Language Processing (NLP) methods have effectively been adapted for use in genomic data by viewing all DNA sequences as a form of structured symbolic language. The application of transformer-based architectures, such as BERT-style models to generate contextual embeddings for nucleotide sequences allows for the simultaneous representation of both local motif positional information and long-range dependencies, necessary to identify important regulatory elements like promoters, enhancers, silencers and transcription factor binding sites [2].

Mathematically, given a DNA sequence $S = \{s_1, s_2, ..., s_n\}$, NLP models learn contextual representations:

$$h_i = f(s_i \mid s_1, s_2, ..., s_n) \quad (1)$$

Where, h_i encodes both upstream and downstream dependencies. The use of contextual embedding's results in much more accurate prediction of chromatin accessibility, which classifies whether a given genomic region is open (active) or closed (inactive). When benchmarked against convolutional methods, models trained on ATAC-seq and DNase-seq perform significantly better with respect to AUROC scores and generalization performance across multiple cell types [4].

Generative AI for Synthetic DNA Generation and Molecular Design

Generative AI models, including Variational Auto-encoders (VAEs), Generative Adversarial Networks (GANs), and autoregressive transformers, have enabled the *de novo* generation of biologically meaningful DNA sequences. These models learn the probability distribution of genomic sequences. Such models can generate novel sequences with desired functional properties, such as enhanced gene expression or binding affinity. In synthetic biology, this enables the design of artificial promoters and regulatory circuits. In drug discovery, generative models extend beyond DNA to molecular design, where chemical structures are represented as SMILES strings or molecular graphs. Reinforcement learning and diffusion-based generative models optimize properties such as: Drug-likeness (Lipinski's rule), Binding affinity, Toxicity minimization. This significantly accelerates the hit-to-lead and lead optimization stages in pharmaceutical pipelines [5-6].

Deep Learning Superiority in Variant Prediction

Deep learning models have demonstrated substantial improvements over traditional statistical and rule-based methods in predicting the functional impact of genetic variants. Classical approaches rely on handcrafted features and population statistics, whereas deep models automatically learn hierarchical representations from raw sequence data. Key advantages include:

- Improved accuracy when identifying whether a gene is deleterious.
- Ability to model non-linear interactions between multiple genomic features.

- Performance improvements within non-coding regions.

As a result, these advances illustrate a shift from "rule-based bioinformatics" to a new era of "AI-enabled genomics," using data to understand and design biological sequence information in order to close the gap between computational and experimental biology [7].

Summary Insight

Collectively, these advancements highlight a paradigm shift from rule-based bioinformatics to data-driven, AI-powered genomics, where NLP and generative models not only interpret biological sequences but also enable design and discovery, bridging the gap between computation and experimental biology [9-10].

Table 1: Comparison of Related Work

Method	Sequence Generation	Drug Discovery	Interpretability
Traditional ML	No	Partial	Low
Deep Learning	Yes	Yes	Medium
Transformer Models	Yes	Yes	Medium
Generative AI [7], [8], [9], [10]	Yes	Yes	High

Additionally, current research has shown that large pre-trained Transformer-based NLP Models can now be used to analyze genomic sequences in ways that were previously only possible for natural language; this enhances both the scalability and accuracy of predictions about how genes will behave under different conditions [8].

Key Contributions

Specifically this review will highlight; 1. In-depth Overview of the Different NLP-based Approaches in Genomic Modelling. This paper surveys some of the latest NLP techniques (all based on Transformers) which have been employed in modelling DNA sequence representation, annotation of regulatory elements, and prediction of chromatin accessibility and the authors continue to highlight the need for sequence to sequence (language) translation (NLP) to capture context. 2. Assessment of Generative Models to aid DNA Synthesis This article will perform an assessment of state of the art (SOTA) Generative Models, the Generative Adversarial Networks (GANs), Auto-encoder VAE and the Autoregressive Transformer in order to understand how capable the Models are at synthesizing functional accurate biological genomic DNA based upon a study on the translation and representation used in Protein Synthesis. 3.

Methodology

The generation of genomic sequence can be carried out using a NLP and generative AI framework comprised of 3 main building blocks: (1) sequence representation, (2) model architecture design, and (3) training strategies. In the sequence representation block, the genomic sequence, in the alphabet {A, T, C, G}, is expressed in a sequence string format and is then fed as tokenized input through k-mer encoding: this process involves sliding a window of length k over the sequence and treating each substring as a token. Thus, a genomic sequence S={s1, s2, . ., s_n}, where si are all in {A, T, C, G}, can be converted to a series of k-mers. This has the advantage of being sensitive to proximal characters as well as biologically-relevant motifs [3].

In terms of model architectures, transformers such as BERT or GPT can employ self-attention to capture long-range dependency, while the generative ability can be obtained

via auto-regressive generation of sequences. In particular, self-attention allows the ability to model long-range dependency in the sequences, and may result in the generation of unidirectional or autoregressively generated sequence. Meanwhile, the generative models that is diffusion based refine noise at every time-steps in the course of sequence generation resulting in high-fidelity sequences. With a training method based on self-supervise models for masked language modeling and next-token prediction, the models are trained with large volumes of unlabeled sequence data allowing them to gain local and global relations between residues in the genomic sequence.

Experiments and Results

Genomic predictions using modern experimental based generative models are outperforming existing ones

- Improved prediction of functional elements
- More Diverse genomic regions from synthetic DNA
- More Generalizing performance on genomic datasets.

Table 2: Performance Comparison of Sequence Prediction Models

Model Type	Representative Models	Accuracy (%)	F1-Score
Alignment-Based	BLAST, HMM	70–82	0.68–0.79
Classical ML	SVM, Random Forest	75–85	0.72–0.83
CNN-Based	DeepBind, DanQ	80–88	0.78–0.86
RNN/LSTM [3]	BiLSTM, GRU	82–89	0.80–0.87
Transformer-Based [8]	BERT, GPT	88–96	0.86–0.94
Generative Models [9]	DNABERT, GenSLM	90–97	0.88–0.95

Discussion

The latest experimental study conducted proves that generative models outperform existing methods in sequence prediction tasks. This includes

- Higher precision in regulatory element recognition
- Higher diversity in synthetic DNA sequences generated
- Wider range of genomic data for generalization

Natural language processing coupled with genomic applications marks a significant transformation in the field of computational biology, envisioning DNA sequences as texts and generating sophisticated sequence models that surpass previous efficiencies.

Conclusion and Future work

Natural Language Processing (NLP), Generative AI, and genomics have pushed computational biology into a smart and data-driven intelligent science. Genomics views the DNA sequence as the biological language. Natural AI architectures like the Transformers, diffusion models, GANs, VAEs and autoregressive models can better learn the complex genomic pattern, predict the regulatory elements, and create biologically meaningful synthetic sequences than traditional bioinformatics methods due to higher accuracy, stronger generalization to unseen genomes, better ability to context understanding and scalability. Also, integrating Generative AI to drug discovery has expedited the speed of molecular generation, target identification, protein structure prediction, and lead optimization. Transformers-based large language models have shown great capability for genomic sequence modeling and diffusion and RL framework have advanced molecular generation

with optimal pharmacological properties. However, challenges remain in explainability of deep models, computation costs, ethical issues on synthetic biology, data bias, and the requirement of biologically validate output for deep generative models. High reliability and explainability on AI-generated genomic sequence are the bottleneck for the large-scale implementation in clinical cases.

Future research directions are suggested, like Developing explainable and trustworthy AI models for genomics; Incorporating multiple levels of biological data, i.e., genomics, transcriptomics, proteomics, clinical data and more; Designing energy-efficient and light weight Transformer models for large-scale genomic analyses; Improving the pipeline for biological validation of generated DNA sequences or molecules; Developing a personalized medicine framework based on patient genomic profile integrated with generative models; Natural Language Processing and Generative AI are reshaping genomics and drug discovery, realizing the smart interpretation, prediction and creation of biological sequences, and we foresee great impact in precision medicine, synthetic biology and next generation therapeutics in the coming years.

References

1. A. Gangwal, A. Ansari, I. Ahmad et al., "Generative artificial intelligence in drug discovery: basic framework, recent advances, challenges, and opportunities," *Frontiers in Pharmacology*, vol. 15, pp. 1331062, 2024. https://doi.org/10.3389/fphar.2024.1331062
2. A. Gangwal and A. Lavecchia, "Unleashing the power of generative AI in drug discovery," *Drug Discovery Today*, vol. 29, no. 6, pp. 103992, 2024. https://doi.org/10.1016/j.drudis.2024.103992
3. H. Askr, E. Elgeldawi, H. Aboul Ella et al., "Deep learning in drug discovery: an integrative review and future challenges," *Artificial Intelligence Review*, vol. 56, pp. 5975-6037, 2023. https://doi.org/10.1007/s10462-022-10306-1
4. F. J. N. Ferreira and A. S. Carneiro, "AI-Driven Drug Discovery: A Comprehensive Review," *ACS Omega*, vol. 10, no. 23, pp. 23889-23903, 2025. https://doi.org/10.1021/acsomega.5c00549
5. V. Garg, "Generative AI for graph-based drug design: Recent advances and the way forward," *Current Opinion in Structural Biology*, vol. 84, pp. 102769, 2024. https://doi.org/10.1016/j.sbi.2023.102769
6. Z. Zhong and J. D. Durrant, "Generative AI in structure-based drug discovery," *BMC Biology*, vol. 24, pp. 75, 2026. https://doi.org/10.1186/s12915-026-02575-x
7. S. K. Niazi, "Artificial Intelligence in Small-Molecule Drug Discovery: A Critical Review of Methods, Applications, and Real-World Outcomes," *Pharmaceuticals*, vol. 18, no. 9, pp. 1271, 2025.https://doi.org/10.3390/ph18091271
8. P. Balakrishnan, A. A. Leema, V. Dhivya Shree et al., "Gene-LLMs: a comprehensive survey of transformer-based genomic language models for regulatory and clinical genomics," *Frontiers in Genetics*, vol. 16, pp. 1634882, 2025. https://doi.org/10.3389/fgene.2025.1634882
9. R. Alizadehsani, S. S. Oyelere, S. Hussain et al., "Explainable Artificial Intelligence for Drug Discovery and Development -- A Comprehensive Survey," *arXiv preprint arXiv:2309.12177*, 2023. https://doi.org/10.48550/arXiv.2309.12177
10. A. Pathak, R. Theagarajan, M. M. Rizqi et al., "AI-enabled drug and molecular discovery: computational methods, platforms, and translational horizons," *Discover Molecules*, vol. 2, pp. 32, 2025. https://doi.org/10.1007/s44345-025-00037-5