Generative Biology

This manuscript (permalink) was automatically generated from In-Vivo-Group/generative-biology@6a8510b on June 14, 2024.

Authors

- **Alexander J. Titus** ✉
  0000-0002-0145-9564 · alexandertitus
  In Vivo Group, Washington, DC, USA; International Computer Science Institute, Berkeley, CA, USA · Funded by Grant TBD

- **Matthew E. Walsh**
  0000-0003-1514-7761 · mwalsh52
  Center for Health Security, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA; Department of Environmental Health and Engineering, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA

✉ — Correspondence possible via GitHub Issues or email to Alexander J. Titus <publications@theinvivogroup.com>.

Abstract

The rapid pace of progress in generative artificial intelligence (AI) techniques like deep learning, reinforcement learning, and transformer neural networks is transforming the life sciences and biomedicine. This living review paper provides an updatable, comprehensive overview and analysis of the latest literature on generative biology – the application of cutting-edge generative AI methods to accelerate insights and innovation across the life sciences and healthcare. All authors are welcome to contribute to this review via pull requests.

The review synthesizes key developments in using generative models for de novo biomedical discovery, design, and decision support. It examines techniques and applications including deep learning on omics data for personalized medicine, generative chemistry for drug development, protein structure prediction for molecular engineering, image synthesis for pathology, language models for clinical decision support, robotic simulation for prosthetics, and generative networks for cell programming.

The review highlights representative studies and benchmarks in each area while contextualizing progress, limitations, emerging best practices, and directions for future work. It also discusses social and ethical challenges raised by generative biology applications, such as compounding bias, system opacity, and dual-use risks, alongside proposed solutions.

As a living review, this paper will be continually updated as the field rapidly advances to provide researchers and practitioners with an up-to-date reference on the state of the art in employing generative AI to accelerate biomedicine for the collective good.

Executive Summary

The goal with be a 2 page TL;DR of the review after v1 is complete. Need more content.

Introduction

This is the start of the Generative Biology living review!

Computers, Algorithms and the Internet

1950s and 1960s: Early computers and algorithms

Computers were used in the early 1950s for population genetics calculations [1/]. Notably, the inception of computational modeling in biology dates to the origins of computer science itself. British mathematician and logician Alan Turing, often referred to as “the father of computing”, used primitive computers to implement a model of biological morphogenesis (the emergence of pattern and shape in living organisms) in 1952 [2]. At about the same time, a computer called MANIAC was used for measuring speculative genetic codes; it was originally built for weaponry research at the Los Alamos National Laboratory in New Mexico [3].

Computers were used for the study of protein structure by the 1960s, and other increasingly diverse analyses. These developments marked the rise of the computational biology field, stemming from research focused on protein crystallography, in which scientists found computers indispensable for carrying out laborious Fourier analyses to determine the three-dimensional structure of proteins [5].

In addition to advances in determination of protein structures through crystallography, the first sequence of protein, insulin, was published [7]. More efficient protein sequencing methods, such as the Edman degradation technique [**ulr:https://pubmed.ncbi.nlm.nih.gov/18134557?**/], enabled sequencing 15 different proteins over a decade [8/]. COMPROTEIN, one of the first bioinformatics softwares developed in the early 1960s, was designed to overcome the limitations of Edman sequencing [9]. In an effort to simplify the handling of protein sequence data for the COMPROTEIN software, a one-letter amino acid code was developed [10]. This one-letter code was first used in the Atlas of Protein Sequence and Structure [11/], the first biological sequence database, laying the groundwork for paleogenetic studies.

Development of methods to compare protein sequences followed. The Needleman-Wunsch algorithm [12/], the first dynamic programming algorithm developed for pairwise protein sequence alignments, was introduced in the 1970s. Multiple sequence alignment (MSA) algorithms followed in the early 1980s. Progressive sequence alignment was introduced by Feng and Doolittle in 1987 [13/]. The MSA software CLUSTAL, a simplification of the Feng-Doolittle algorithm [14/] was developed in 1988. It is still used and maintained to this day [15/].

Advances in Generative AI for Proteins

Introduction

- Background on proteins as key molecular machines in biology

- Promise of generative AI to accelerate protein discovery and engineering
- Overview of scope covering recent advances in last 1-2 years

Structure Prediction

- AlphaFold2 as a breakthrough method for structure prediction
- Novel model architecture and training methodology
- Examples of new biological insights from predicted structures

Function Prediction

- Using predicted structures to infer protein functions
- Structure-based identification of catalytic and binding sites
- Case studies of novel enzyme functions discovered

Designing Novel Proteins

- Generative models for designing functional protein sequences
- Leveraging structural constraints for optimized protein engineering
- Applications in industrial enzymes, therapeutics, biomaterials

Interaction Prediction

- Modeling protein-protein interactions with graph networks
- Structure-based prediction of protein-drug bindings
- Applications in drug discovery and toxicity screening

Outlook

- Challenges and next steps in improving accuracy
- Hybrid physics- and data-driven approaches
- Ethical considerations in synthetic protein design

Advances in Generative AI for RNA

Introduction

- Background on key roles of RNA in biology
- Promise of generative models to advance RNA research
- Scope focused on latest advances in RNA prediction and design

Structure Prediction

- Transformer-based prediction of RNA folding
- Novel architectures for incorporating chemical constraints
- Improved accuracy on complex structures like ribosomes

Function Prediction

- Inferring RNA functions from sequence and structure
- Identifying motifs, domains, and atomic binding sites
- Applications in understanding long noncoding RNAs

Interaction Prediction

- Graph neural networks for RNA-protein interactions
- Structure-augmented modeling of splice sites
- Predicting RNA base editing targets

Design of RNA Therapeutics

- Generative models for optimizing siRNA, antisense design
- Reinforcement learning for chemical modification patterns
- Progress in computational RNA-targeted drug design

Outlook

- Key challenges in prediction of long RNA structures
- Design of RNA for self-assembly and scaffolding
- Ethical use of synthetic RNA technologies

Advances in Generative AI for DNA

Introduction

- Background on the role of DNA as a carrier of genetic information
- Promise of generative models to advance DNA research
- Scope focused on the latest advances in DNA prediction and design

Sequence Modeling

Transformer architectures for modeling DNA sequences

The HyenaDNA paper introduces a new genomic foundation model called HyenaDNA that can process DNA sequences at single nucleotide resolution with ultralong context lengths up to 1 million base pairs - a 500x increase over previous transformer models. HyenaDNA uses a parameter-efficient convolutional architecture that allows it to scale subquadratically with sequence length, enabling the use of full genome-scale context. On a range of regulatory genomics prediction tasks, HyenaDNA matches or exceeds the performance of previous state-of-the-art models while using 1500x fewer parameters and 3200x less pretraining data. HyenaDNA also demonstrates the ability to perform challenging species classification using the full mutational profile visible at 1 million base pairs of context. The authors explore new training techniques to enable ultralong sequence modeling as well as prompt-based tuning methods for rapidly adapting to new tasks without updating pre-trained weights [16].

Pretraining on large genomic datasets

Applications in variant calling and annotation

Regulation Prediction

- Graph neural networks for modeling 3D genome architecture
- Predicting enhancer-promoter interactions and expression
- Design of synthetic promoters and enhancers

Genome Editing

- Generative models for CRISPR guide design
- Contextual prediction of on-target editing efficacy
- Modeling of off-target effects during optimization

DNA Data Generation

- Variational autoencoders for realistic DNA sequences
- Generating paired genomic-transcriptomic data
- Applications in training genome interpretation models

Outlook

Challenges in predicting long-range chromatin interactions

The development of transformer architectures like HyenaDNA that can leverage genome-scale context creates new opportunities for understanding long-range chromatin interactions, gene regulation, and intercellular networks. However, significant challenges remain in improving model accuracy and uncertainty quantification. Hybrid physics- and data-driven approaches may help address these gaps. Safety considerations around synthetic genome design also warrant further research to ensure responsible innovation as these generative capabilities advance. Overall, the future looks bright for generative models that can capture both local mutations and global patterns critical to biology [16].

Responsible design of synthetic genomes

Ethical considerations for human genome editing

Generative AI for Autonomous Experimentation

Introduction

- Promise of generative models to accelerate scientific discovery
- Rise of autonomous labs and robotics for automated experimentation
- Overview of generative AI's role in self-driving research

Closed-Loop Systems

- Integrating computational hypothesis generation with robotic wet lab testing
- Reinforcement learning pipelines for autonomous optimization
- Case studies in materials science, drug discovery

Automated Experiment Design

- Using generative models to design novel compounds, genes
- Leveraging simulations to predict experimental outcomes
- Robotic execution of designed experiments

Adaptive Sampling

- Active learning to iteratively select most informative experiments
- Bayesian optimization powered by neural networks
- Applications in probing molecular design spaces

Real-Time Learning

- Deploying models on lab edge devices
- Online learning from experimental data streams
- Improving models and experiment plans on-the-fly

Outlook

- Key challenges around model accuracy and integration
- The future of data-driven, self-driving laboratories
- Risks of full automation and need for human oversight

Conclusion

- Generative AI as a powerful tool for autonomous experimentation
- Accelerating discovery alongside human researchers
- Responsible implementation will maximize benefit

Generative AI and the Biosecurity Landscape

Introduction

- Background on biosecurity threats from natural, accidental, and intentional pathogens
- Rise of generative AI as a dual use technology for biodefense and misuse

Enhanced Risks

- Automated bioweapon design with generative models
- Relatively low computing needs to generate dangerous agents
- Challenges detecting artificially generated sequences/organisms

Enhanced Response Capabilities

- Generative models for vaccine and therapeutic design
- High-throughput testing of countermeasures with synthetic data
- AI for early detection of emergent pathogens and outbreaks

Recommendations

- Increased oversight for generative model development/release
- Expanding biosecurity legislation and regulations
- Fostering open research and global cooperation

Outlook

- Trajectory toward increasingly powerful generative biological capabilities
- Need for preventative ethics research and guidance
- Maintaining public trust and avoiding overreaction

Conclusion

- Balancing generative AI's benefits and risks in biology
- Importance of thoughtful governance and responsible innovation
- Staying ahead of the curve on biosecurity

Policy Responses to Generative AI in Biology

Introduction

- Background on rise of powerful generative models for biology
- Overview of risks like bioweapons, environmental damage
- Need for governance to ensure responsible development

Self-Governance by Developers

- Voluntary guidelines on ethical AI by corporations
- Limiting access to certain capabilities like human editing
- Issues with self-regulation and transparency

Governmental Regulations

- New biosecurity regulations on certain AI technologies
- Restrictions on use of synthetic biology IP
- Challenges with fast pace of technology change

International Governance

- Proposals for global observatory to monitor risks
- Treaties restricting development of bioweapons
- Difficulty achieving consensus and compliance

Public Deliberation and Ethics

- Activities to involve broader stakeholders
- Gathering public attitudes on acceptable uses of generative bio AI
- Informing policy with deliberative democracy

Outlook

- Likely increase in debate and policy activity in this space
- Balancing innovation and security will be a key challenge
- Importance of thoughtful multidisciplinary discourse

Conclusion

- No easy policy solutions, but inaction also carries risks
- Policy should enable innovation but promote responsible use
- This will require sustained public deliberation and coordination

References

1. **Computers in the study of evolution**
JL Crosby
Science Progress (1967) https://pubmed.ncbi.nlm.nih.gov/4859964

2. **The chemical basis of morphogenesis**
Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences (1952-08-14) https://royalsocietypublishing.org/doi/10.1098/rstb.1952.0012
DOI: 10.1098/rstb.1952.0012

3. **Scientific uses of the MANIAC**
HL Anderson
Journal of Statistical Physics (1986-06-01) https://doi.org/10.1007/BF02628301
DOI: 10.1007/bf02628301

4. **A three-dimensional model of the myoglobin molecule obtained by x-ray analysis**
JC Kendrew, G Bodo, HM Dintzis, RG Parrish, H Wyckoff, DC Phillips
Nature (1958-03-08) https://pubmed.ncbi.nlm.nih.gov/13517261
DOI: 10.1038/181662a0

5. **Structure of myoglobin: A three-dimensional Fourier synthesis at 2 A. resolution**
JC Kendrew, RE Dickerson, BE Strandberg, RG Hart, DR Davies, DC Phillips, VC Shore
Nature (1960-02-13) https://pubmed.ncbi.nlm.nih.gov/18990802
DOI: 10.1038/185422a0

6. **The amino-acid sequence in the glycyl chain of insulin. 1. The identification of lower peptides from partial hydrolysates**
F Sanger, EOP Thompson
Biochemical Journal (1953-02) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1198157/

7. **The amino-acid sequence in the glycyl chain of insulin. 2. The investigation of peptides from enzymic hydrolysates**
F Sanger, EOP Thompson
Biochemical Journal (1953-02) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1198158/

8. **The origins of bioinformatics**
JB Hagen
Nature Reviews. Genetics (2000-12) https://pubmed.ncbi.nlm.nih.gov/11252753
DOI: 10.1038/35042090

9. **Comprotein: a computer program to aid primary protein structure determination**
Margaret Oakley Dayhoff, Robert S Ledley
Proceedings of the December 4-6, 1962, fall joint computer conference on - AFIPS '62 (Fall) (1962) http://portal.acm.org/citation.cfm?doid=1461518.1461546
DOI: 10.1145/1461518.1461546

10. https://febs.onlinelibrary.wiley.com/toc/14321033/5/2

11. **Collecting, comparing, and computing sequences: the making of Margaret O. Dayhoff's Atlas of Protein Sequence and Structure, 1954-1965**
Bruno J Strasser
Journal of the History of Biology (2010) https://pubmed.ncbi.nlm.nih.gov/20665074
DOI: 10.1007/s10739-009-9221-0

12. **A general method applicable to the search for similarities in the amino acid sequence of two proteins**
SB Needleman, CD Wunsch
Journal of Molecular Biology (1970-03) https://pubmed.ncbi.nlm.nih.gov/5420325
DOI: 10.1016/0022-2836(70)90057-4

13. **Progressive sequence alignment as a prerequisite to correct phylogenetic trees**
DF Feng, RF Doolittle
Journal of Molecular Evolution (1987) https://pubmed.ncbi.nlm.nih.gov/3118049
DOI: 10.1007/bf02603120

14. **CLUSTAL: a package for performing multiple sequence alignment on a microcomputer**
DG Higgins, PM Sharp
Gene (1988-12-15) https://pubmed.ncbi.nlm.nih.gov/3243435
DOI: 10.1016/0378-1119(88)90330-7

15. **Clustal Omega, accurate alignment of very large numbers of sequences**
Fabian Sievers, Desmond G Higgins
Methods in Molecular Biology (Clifton, N.J.) (2014) https://pubmed.ncbi.nlm.nih.gov/24170397
DOI: 10.1007/978-1-62703-646-7_6

16. **HyenaDNA: Long-Range Genomic Sequence Modeling at Single Nucleotide Resolution**
Eric Nguyen, Michael Poli, Marjan Faizi, Armin Thomas, Callum Birch-Sykes, Michael Wornow, Aman Patel, Clayton Rabideau, Stefano Massaroli, Yoshua Bengio, ... Chris Ré
arXiv (2023) https://doi.org/gs3v5j
DOI: 10.48550/arxiv.2306.15794