The Capacity of Generative AI to Create Valid DNA Sequences

Timothy Cromwell, Jadyn Berlin, Branden Hunter, & Hyunjin Shim • Biology • California State University Fresno

cromwell_exe@mail.fresnostate.edu • jadynbb@gmail.com • github.com/hshimlab

FRESNO STATE | Discovery Diversity Distinction ®

Question

Can generative AI create useful DNA sequences?

Background

- Globally, we have cataloged a large amount of DNA sequences over time.
- This DNA catalog is largely based on (or is) human DNA, due to our inclination to study ourselves and material like us.
- More varied and unique DNAs from diverse organisms would open the opportunity for new discoveries and possible solutions to DNA-related ailments.
- Generative AI has the potential to create biological sequences, opening the door for an infinite variety of DNA. Generated sequences can also provide noise for testing and synthesized experiments.

Method

- We selected six generative AIs, two of which (Evo and SMS) were focused specifically on DNA sequence generation, and generated 50 sequences of 1000 base pairs from each AI model.
- We then ran every sequence through the National Institutes of Health's BLAST program to determine the similarity of each sequence to other existing DNA.
- After determining the the three most random sets of sequences and converting them to the respective proteins, we ran each through AlphaFold and FoldSeek to visualize 2-D DNAs after converting them to 3-D proteins.

Figure 1

Standard Protein Structure

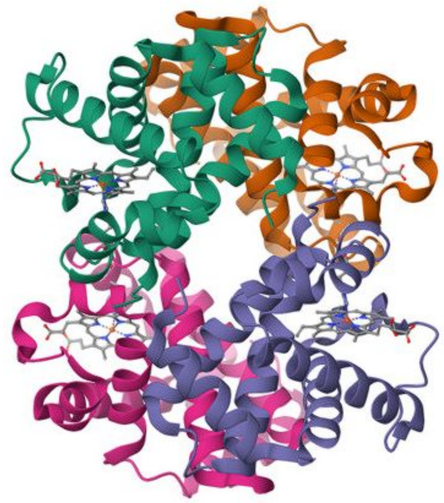
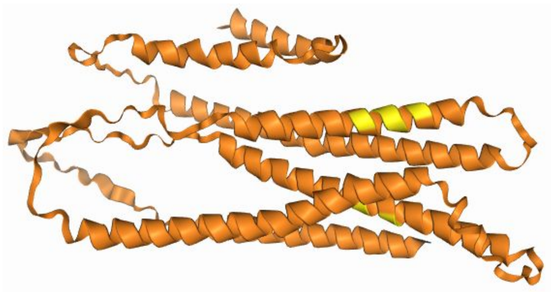


Figure 2

Generated Protein Structure



Tested AI Models

ChatGPT

Claude

Meta AI

Gemini

Evo

SMS

Results

- While every AI would consistently output something, not all of them were good at producing biological sequences.
- 86 percent of sequences returned no hits from BLAST, showing that the generated DNAs were indeed white noise with no similar sequences in the public database.
- Sequence Manipulation Suite had the most random sequences, returning zero hits on BLAST.

Conclusions

- Generative AI specializes in text for a reason. While every AI would consistently output something, not all of them were good at accurately following instructions. There were duplicates, sequences of the long length, and textual errors in many of the generated sequences.
- Sequence Manipulation Suite is great at generating DNA, as it focuses specifically on random generation in sequences using many different proprietary algorithms and techniques.
- Generated sequences have potential for new discoveries. While a major reason to generate sequences is to use as white noise in DNA sequencing, there is always the possibility that white noise DNA will be useful in other fields.
- Hopefully the randomization provided by these AIs will improve to the point where it will become commonplace to simply request a biological sequence.
- More tests and synthesis are on the horizon.

Acknowledgements

Thank you to Angel Montes for helping us determine which AI models to use.