Qeios

Peer Review

Review of: "Human Genome Book: Words, Sentences and Paragraphs"

Josiah Seaman[1]

1. Queen Mary University of London, London, United Kingdom

Overall: The ability to translate between DNA and human languages would be world-changing. If I decided on funding, there'd be billions of dollars going into that project. The reason it hasn't been done as effectively as AlphaFold 2 is that the problem is much larger, more intimidating, and multi-dimensional. It will require a proportionately larger amount of resources, time, and breakthroughs in order to crack a problem that has eluded us for 25 years. For starters, you'll need to use the entire ENCODE dataset, which is not featured in this paper. It will also require extensive compute (not found here) and a novel architecture. It's an ideal problem to tackle, but likely beyond the scope of a single researcher.

I would be very happy to be proved wrong. However, this paper does not have sufficient results to convince me that this monumental task has been accomplished. No accessible outputs are provided with the paper. No external website means the outputs have not been tested and can't be peer-reviewed. To take on the much more reasonable goal of making progress towards DNA → human language translation, the author needs to include biologically verified test datasets.

DNA Segmentation is definitely the correct first step to work on. In fact, attempting functional annotations is probably overwhelming and weakening the results significantly. Independent Chromosome segmentation datasets are available for TADs, gene regulatory networks, sequence self-similarity, and other metrics. All of these would be a terrific way to measure the level of accuracy in segmentation.

This paper is a decent starting point for talking about 1) what datasets would be most informative to include in this task and 2) what kind of architecture would be required?

Fig. 1 & Fig. 2: Your approach makes sense, but you haven't convinced me that you've made the necessary breakthrough in 1) training data or 2) architecture that would enable you specifically to crack the code that no one has figured out in the last 25 years. We need a bit more text on "why" this approach is unique

 doi.org/10.32388/EI8D4R

and promising as opposed to any other organization of #1 and #2 you could have picked. In particular, you apparently didn't use a number of crucial datasets that would have made your project much more effective. A) Hi-C topological association domains (TADs) B) Pyknons (Rigoutsos) for tokenization and C) gene regulatory networks, specifically confirmed Transcription Factor Binding Sites (TFBS) for related content, English annotation, and boundaries.

Table 2 needs some reference points for context. What's a comparable score in another domain? Is that good or bad? In particular, is 79% accuracy above the "state transition temperature" where generated sentences in models like GPT2 go from being plausible to nonsense?

Fig. 3: I'd like to see a non-linear dimensionality reduction like tSNE instead of PCA.

2.4.1 makes sense and is the right decision.

Training Data <p_end>: There's going to be an intrinsic bias for the average length of an English paragraph that will be notably different from a DNA sequence because of encoding information density differences. For example, the average poly-peptide in the human genome is weirdly similar to the length of a tweet (140 characters) if you convert both into the number of bits. Necessarily, the average DNA "sentence" will be at least 3x that length but a similar number of bits. Double-check you've accounted for this bias in tokenization.

Chromosome Part: There already exists chromosome segmentation in the form of TADs. Also, chromosome banding has been known since at least the 1980s. Doing this is important for the consistency of the trained data. AT-rich and GC-rich sequences are nothing alike and refer to different subject matter in what is likely a different language. You don't want to mix them.

"It is important to note that this word translation relationship is based solely on structural similarity in the embedding space and does not imply semantic similarity." True, but that doesn't make the approach invalid. Translating between English and Japanese can also be done using solely structural similarity, even though the two languages have a very distant common origin, if there is one at all. The deeper question to ask here is: Is human language structure forced to conform to some deeper transcendent structure for encoding information such that DNA encoding would also need to conform to the same structure? I don't think we know the answer to that, but it could be tested with a sufficiently thorough token embedding project.

Fig. 5: This is a disappointing figure, not just for results but for how it's presented. At minimum, properly crop to the entire width of the text and make it high enough resolution that we can actually read the text.

Don't hide your results behind an image. Secondly, link to an actual web copy or a Supplemental file with the entirety of Chromosome 1 English volume and DNA volume. I would say this paper is not submittable without providing the outputs and results. Without results, this is a pure method proposal paper. Which is fine, but don't write it as a results paper without showing the results. I want to browse for myself the Genome Book that this paper is about. Without that, it's not published science.

3.3.1 Search: The better term to use here is hierarchical Retrieval Augmented Generation (RAG). We already have great ways of indexing *all known sequences.* However, if you wanted to search using English "show me regulatory regions for eye lens proteins," your method would be much better at retrieving relevant sequences in context.

"Since research on the interpretability of large language models is still in development, our constructed genomic book currently does not provide definitive biological interpretations." We're doing open-source science. It's okay to say an experiment didn't work out. Then you also have the freedom to discuss what factors may have led to being past some threshold for interpretability of English text. What would need to change in order to train and get coherent outputs?

Declarations

Potential competing interests: No potential competing interests to declare.