

Peer Review

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

Emmanouil D. Oikonomou¹

1. Department of Informatics and Telecommunications, University of Ioannina, Greece

First of all, I would like to say that this is very interesting work on a topic that highly remains underexplored. While the author could explore alternative algorithms and tokenizers for comparison, the study is already well-defined and structured.

However, I have concerns regarding how the author handles sequences.

The idea to translate DNA kmers into text is truly fascinating. But I believe the author should provide the exact English translation for each DNA kmer, pointing out how accurate each translation is, based on the kmer characteristics. Without knowing the precise translation of these sequences, it is difficult to fully evaluate the methodology.

The same problem is occurring with the proteins. Moreover, I couldn't find relative results for the proteins.

The key challenge in applying NLP to biology is about the semantics. While NLP algorithms perform well with biological data, we don't know yet how it is recognized as a Natural Language since they are not! Quoting the author:

“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.”

This is exactly the problem with biomolecule sequences. What is the semantic similarity between biomolecules' kmers? And then, how can we implement it into natural language? How can we decrypt the semantics of such sequences?

I would like more results about the meaning of each “translation” into English, as well as author explanations on them.

Declarations

Potential competing interests: No potential competing interests to declare.