3rd Serbian International Conference on Applied Artificial Intelligence (SICAAI)
Kragujevac, Serbia, May 23-24, 2024

EXPLORING WORD2VEC MODELS FOR CAPTURING THE SIMILARITY OF CODON EMBEDDINGS

Anđa Denić[1*], Jelena Pejić[1], Aleksandar Trokicić[1]

[1] Department of Computer Science, Faculty of Science and Mathematics, University of Niš
e-mail: andja.denic@pmf.edu.rs, jelena.pejic@pmf.edu.rs, aleksandar.trokicic@pmf.edu.rs
corresponding author

ABSTRACT:

Word2vec [1], is a neural network (NN) technique, that learns a vector representation of words (tokens), analyzing its context (surrounding tokens), and use some function between vectors to encode the semantic relationship between words. Inspired by its success in NLP many researchers tried to replicate its idea to biological sequence embedding [2, 3, 4]. However, compared to the NLP problems where it is clear what are tokens and their context, in bioinformatics that is not the case.

A lot of papers [4] embed k-mers, but there are multiple options for the choice of a context and the length of k-mers, as well as whether there should be an overlap between k-mers? Therefore, we will empirically analyze the quality of 3-mer (codon) embeddings of the genome sequences of V. Cholerae and E. Coli.

The dataset was generated by sliding a window across the DNA sequence and using codons along with their left and right neighbors as input. We experimented with different hyperparameters: numbers of neighbors ($m \in \{3,5,10,20,40\}$) and whether the windows overlap or not. We evaluate the results by analyzing the vector representation of the codons. There are 64 codons which are grouped into 21 categories according to the amino acids they encode [5]. For each codon, we define recall as the ratio between the number of codons from the same group found among its top 6 closest neighbors, determined by cosine vector distance, and the total number of codons in the group. Table 1. presents the average recall across codons.

Table 1. Average recall across codons for different hyperparameters on the sequences of V. Cholerae and E. Coli.

	Overlapping					non-overlapping				
	m=3	m=5	m=10	m=20	m=40	m=3	m=5	m=10	m=20	m=40
Vibrio cholerae	84.38%	84.38%	84.38%	83.28%	80.94%	45.52%	37.66%	40.52%	34.48%	35.37%
Escherichia coli	84.38%	84.38%	82.81%	81.66%	79.27%	41.46%	42.66%	42.97%	39.27%	35.78%

Experimental results reveal that using overlapping windows and about five neighbors results in the highest similarity among vector embeddings for codons within the same group.

In future work, different contexts could be explored for gene embedding and models could be trained on a greater variety of different organism's genomes because the best context may vary between organisms.

References:

[1] Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." *Advances in neural information processing systems* 26 (2013).
[2] Wu, Feng, et al. "A deep learning framework combined with word embedding to identify DNA replication origins." *Scientific reports* 11.1 (2021): 844.

[3] Yilmaz, Alper. "Assessment of mutation susceptibility in DNA sequences with word vectors." *Journal of Intelligent Systems: Theory and Applications* 3.1 (2020): 1-6.
[4] Ng, Patrick. "dna2vec: Consistent vector representations of variable-length k-mers." arXiv preprint arXiv:1701.06279 (2017).
[5] Rosandić, Marija, and Vladimir Paar. "The Supersymmetry Genetic Code Table and Quadruplet Symmetries of DNA Molecules Are Unchangeable and Synchronized with Codon-Free Energy Mapping during Evolution." *Genes* 14.12 (2023): 2200.