Integrated Encoding of Semantic and Orthographic Distances in a DNA Language Model

[1]Je-Hwan Ryu, [2]Ji-Hoon Lee, [1,2,3,4,*]Byoung-Tak Zhang(btzhang@bi.snu.ac.kr)

([1]Brain Science Program, [2]Graduate Program in Bioinformatics, [3]Cognitive Science Program and [4]Computer Science and Engineering), Seoul National University, 1 Gwanak-ro Gwanak-gu, Seoul 151-742, Republic of Korea

Abstract

We use DNA computing for modeling natural language processing. In a DNA language model, it is required to convert a word into a DNA sequence. One problem in DNA encoding is that, if we generate DNA sequences considering only non-crosshybridization between them, the generated sequences may have entirely different distances than the original semantic and orthogonal distances. To overcome this, we studied how to combine the semantic and orthographic distances, such as Hamming and Levenstein distances and that of WordNet. We focused on combining these two distance measures into one integrated system that can preserve the semantic and orthographic distances between the words and the DNA sequences. We made an integrated encoding system and made DNA sequences for the DNA language model. Then, we verified the generated sequences by small-scale wet-lab experiments. The results of our experiments indicate that the integrated encoding system offers a more reliable DNA language model by minimizing the loss of the original semantic and orthographic information.

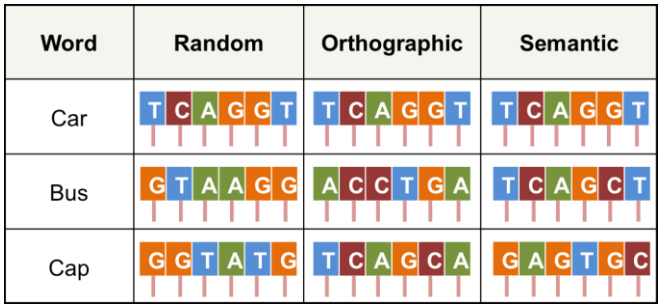

Figure 1: DNA sequences are generated by the orthographic and semantic distances. In the orthographic column, 'Car' and 'Cap' are similar but 'Bus' is not. In the semantic column, 'Car' and 'Bus' are similar but 'Cap' is not.

Reference

Lee, J. H., Lee, S. H., Chung, W. H., Lee, E. S., Park, T. H., Deaton, R., & Zhang, B. T. (2011). A DNA assembly model of sentence generation. *BioSystems*, *106*(1), 51-56.

Budanitsky, A., & Hirst, G. (2001, June). Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. In Workshop on WordNet and Other Lexical Resources (Vol. 2)

Acknowledgment

This work was supported by the Air Force Research Laboratory, under agreement number FA2386-12-1-4087, and by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (NRF-2010-0017734 and NRF-2013M3B5A2035921), supported in part by KEIT grant funded by the Korea government (MKE) (KEIT-10035348 and KEIT-10044009)