**[BIOINFO 2023]**

# Leveraging advances in natural language processing to decipher antibody sequence

Eunna Huh, Hyeonsu Lee, Erkhembayar Jadamba, and Hyunjin Shin
*MOGAM Institute for Biomedical Research, Seoul 06730, Republic of Korea*
**Corresponding author: hyunjin.shin@mogam.re.kr*

Over 100,000 human unique proteins are encoded using only 20 different amino acids. These amino acids can be individually characterized by their chemical and biophysical properties. However, mutual interactions and topological constraints between these amino acids add another layer of complexity when interpreting protein functions. To distill such a vast amount of information in a protein sequence, we applied natural language processing (NLP) with transformers. In the architecture of Robustly Optimized Bidirectional Encoder Representations from Transformers Approach (RoBERTa), we conducted an investigative experiment with antibody sequences. The model was trained on a dataset comprising 600 million human antibodies sourced from the Observed Antibody Space database (OAS). By leveraging the capabilities of both NLP and a large-scale sequence database, we converted amino acid sequences into vector embeddings considering their long-distance dependencies. These contextual vector embeddings were able to capture antibody features, including target diseases (antigens), the type of B cells (memory and naïve B cells), and different germline V alleles, based solely on sequence information. In addition, we confirmed that single amino acid tokens represent the most antibody information and RoBERTa-trained embedding vectors (silhouette score of antigen clustering: 0.164) are better to obtain information than one-hot-encoding (silhouette score of antigen clustering: 0.023). These findings confirm that antibody features can be extracted from sequence data and hold promise for high-resolution functional predictions.