DEEPTCRMATCH: AN EFFECTIVE WAY OF COMPUTING T CELLS ANTIGEN SPECIFICITY

Seungtae Park*, Sungsik Kim, Hee Joon Jeon, Na Ri Yoon, Bo Ryeong Lee, Sungmin Kim, Woong-Yang Park, Hyung Ju Hwang. Pohang University of Science and Technology, Pohang, Korea, Republic of; Geninus Inc., Seoul, Korea, Republic of; AMSquare Inc., Pohang, Korea, Republic of; Pohang University of Science and Technology and AMSquare Inc., Pohang, Korea, Republic of

Background To determine the specificity of T cells based on their receptor sequence is a demanding task due to cross-reactivity, complicated patterns and limited size of public dataset. An effective computational model, which finds CDR3 patterns of shared antigen specificity in the existing dataset, can predict specificity of T cells accurately. In this work, we developed a deep learning methodology that computes the similarity among T cells in terms of antigen specificity using k-mer features.

Methods Our model consists of two parts. First, it encodes every overlapping k-mers of CDR3 into numerical vectors. We parallelize such k-mer encodings into several allowable ways, so that the independent semantics of each k-mers are effectively learned. Second, among the encoded k-mer features, we select only meaningful k-mers using a self-attention structure. By doing this, we remove unwanted correlations among overlapping k-mers.

We train our model with preprocessed public datasets: IEDB, VDJdb and McPAS. We optimize the overall process to find an optimal contrastive predictive coding, which is an unsupervised objective function. After optimization, we define a kernel function of k-mer features to define similarity between two CDR3s.

Results We designed an one-of-many unsupervised task: for a given arbitrary CDR3 sequence, whether our model can correctly select CDR3 with a similar specificity among N randomly sampled candidates. With N=10, our model achieves accuracy 0.3 for an independent dataset. We also test supervised task: whether our model can induce probable cognate antigens for a given CDR3. Our model achieves precision 0.7.

Conclusions Our deep learning model can extract k-mer information that only represents antigen specificity. This information is an invaluable numerical vector for computing similarity of antigen specificity. By doing this, our model can solve the one-of-many problem and predict the antigen specificity. In the future, our model will improve its performance as a size of training dataset grows.