Identifying B-cell epitopes using AlphaFold2 predicted structures and pretrained language model1

Negar Askari

July 11, 2023

Many machine learning tools have been deployed for the task of B-cell epitope (BCE) prediction. The main problem that this paper aims to address is that sequence-based tools tend to perform poorly when it comes to conformational epitopes (which make up more than 90% of BCEs), and in most cases, we simply do not have enough experimentally decided structural information available for structure-based methods to be used. The proposed model (GraphBepi) uses the predictive tool AlphaFold2 to predict the structural information needed for this task from the protein sequences.

After feeding the input antigen sequence into AlphaFold2, the relational graph of amino acid residues and DSSP (feature vector extracted using the DSSP program) are extracted from the predicted protein structure. GraphBepi also uses a pretrained language model (esm2_t36_3B_UR50D or ESM-2 for short) to extract effective features from the antigen sequences. These features along with DSSP are then fed into a BiLSTM module to capture long-range dependencies from these sequences. They are also used to form feature vectors for residues in the relational graph, which is then fed into an edge-enhanced deep graph neural network (EGNN) module to learn structural features. The results from the EGNN and BiLSTM modules are then fed into an MLP to make the final prediction.

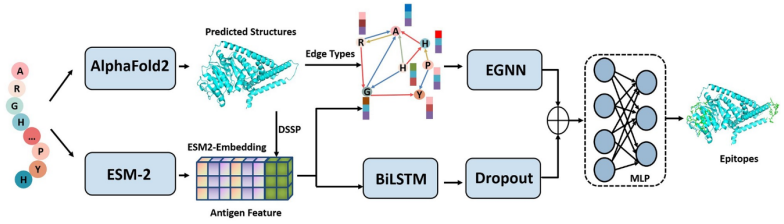

Figure 1: The architecture of GraphBepi

GraphBepi outperformed state-of-the-art BCE prediction tools (two sequence-based and four structure-based approaches were mentioned in the paper) by at least 5.5% in terms of AUC, 44.0% in terms of AUPR, and 20.4% in terms of F1.