

Title : An LSTM network-based model with attention techniques for predicting linear T-cell epitopes of the hepatitis C virus

Author Names : Md. Faruk Hosen , S. M. Hasan Mahmud , Kah Ong Michael Goh , Muhammad Shahin Uddin , Dip Nandi , Swakkhar Shatabda , Watshara Shoombuatong

Abstract: Hepatitis C virus (HCV) infection remains a significant global health challenge, often resulting in severe longterm physical complexity and even death. Since its discovery, HCV has exhibited substantial genetic variability, complicating vaccine development. Although some therapeutic approach have shown efficacy against certain HCV genotypes, a universally effective vaccine is still lacking. Recent research suggests that the body's cellular immune response, particularly T cell epitopes of HCV (TCE-HCVs), plays a vital role in fighting the virus. Therefore, the precise and rapid identification of TCE-HCVs is essential for chronic HCV infection. In this work, we proposed a novel TCE-HCVs prediction model AttLSTM, which combines attention mechanism and long shortterm memory (LSTM). Specifically, we employed four robust feature encoding techniques: One-Hot Encoding, Global Vectors (GloVe), fastText, and Word2Vec to encode protein sequences. Additionally, k-mer embedding was utilized to help the model identify significant subsequence fragments within the protein sequences. To optimize the model's performance, irrelevant features are eliminated using the SHapley Additive exPlanations (SHAP) approach. The resulting optimal feature subset was then fed into the AttLSTM model to identify TCEHCVs. The attention mechanism in this model dynamically captures the pairwise correlations of each neighboring target pair within a sliding window, thereby enhancing the understanding of the local environment of target residues. Extensive experiments showed that AttLSTM outperformed conventional machine learning (ML) classifiers in predictive performance. Notably, in k-fold cross-validation, AttLSTM achieved superior performance compared to existing methods with accuracy of 80.77 %, MCC of 0.632, and AUC of 0.891. This exceptional performance indicates that AttLSTM has a strong predictive capability for identifying TCE-HCVs. We anticipate that AttLSTM will expedite the rapid identification of promising TCE-HCVs, aiding in the development of diagnostic and immunotherapeutic treatments for HCV in the future. In addition, we have developed a web server for real-time prediction using our proposed model, which is available at <http://attlstmhcv.xyz/>

Keywords: Hepatitis C virus, Attention mechanism, Feature extraction, Long short term memory, neural network, Shapley values

www.sciencedirect.com/journal/results-in-engineering