Caring is Sharing – Exploiting the Value in Data for Health and Innovation
M. Hägglund et al. (Eds.)

doi:10.3233/SHTI230348

Combining Sequence Similarity with Physicochemical Properties to Predict Binders for MHC-II Molecules

Ya-Lin CHEN[a,1]
[a] *Department of Biomedical Informatics and Medical Education, University of Washington, USA*
ORCiD ID: Ya-Lin Chen https://orcid.org/0000-0002-7046-827X

Abstract. Prediction of binders of the major histocompatibility complex class II (MHC-II) molecules is critical for T cell immunogenicity. As protein-protein interaction also relies on physicochemical properties, we aim to build a novel model combining sequence information and the physicochemical properties of proteins. Our research used data from the NetMHCIIpan 3.2 study. Features include BLOSUM50 and the physicochemical properties from iFeature Python package. We created a hybrid model of recurrent neural layers and feedforward layers. The final Area Under the Receiver Operating Characteristics (AUROC) on the test data was 0.755.

Keywords. MHC class II, MHC binding specificity, deep neural network

1. Introduction

The binding of Major histocompatibility complex class II (MHC-II) and peptides is a key step for T-cell recognition, triggering the defense against viral infection and other diseases. Identifying the peptides is crucial for understanding the pathophysiology of diseases and for designing therapeutics. Currently, several methods use sequence similarity data, such as BLOSUM matrix, to predict the binding affinity [1]. However, protein-protein interaction is also dependent on other factors such as physicochemical properties [2]. This research aims to combine sequential data and physicochemical properties to predict the binders of MHC-II molecules.

2. Methods

The dataset from NetMHCIIpan 3.2 consists of pairs of MHC-II molecules and peptides with their binding affinity [1]. We selected one partition of data and included only human data, MHC-II molecules with more than 20 peptides, with at least 4 binders with peptide length of 15, resulting in 87,052 pairs as train data and 21,535 pairs as test data. Binders were defined as those with < 500nM binding affinity [1]. The sequential features were generated using the BLOSUM50 matrix. We used iFeature Python library to generate

[1] Corresponding Author: Ya-Lin Chen, E-mail: end1859612@gmail.com.

the physicochemical properties [3, 4]. A hybrid model was created with a recurrent neural network (RNN) for sequential features whose latent output combined with physicochemical properties was to fed into a deep feedforward network. Neural network models were built with Tensorflow 2.10 and the code repository is available online.

3. Results

The percentages of binders in the train and test data were 43.42% and 40.34% respectively. The test performance is shown in Table 1.

Table 1. Test performance. Precision, Recall, and F1-score are macro averages. AUROC: Area Under the Receiver Operating Characteristics. RNN: recurrent neural network.

	BLOSUM50	**iFeature**	**BLOSUM50+iFeature**
Model	RNN layers	Feedforward layers	RNN + Feedforward layers
AUROC	0.740	0.711	0.755
Accuracy	0.752	0.713	0.750
Precision	0.742	0.705	0.746
Recall	0.740	0.711	0.755
F1-score	0.741	0.706	0.746

4. Discussion

Our design of the hybrid model takes into account both the sequential data and the physicochemical properties of proteins. However, one limitation of current research is that we only used one out of the five partitions of the data from NetMHCIIpan 3.2 [1]. We expect to improve the prediction as we include more data in the future [5].

5. Conclusions

This research combined BLOSUM50 and physicochemical properties to predict binders for MHC-II molecules. Using a hybrid deep neural network, the test AUROC reaches 0.755.

References

[1] Jensen KK, Andreatta M, Marcatili P, Buus S, Greenbaum JA, Yan Z, et al. Improved methods for predicting peptide binding affinity to MHC class II molecules. Immunology. 2018;154(3):394-406.

[2] Oscarsson S. Factors affecting protein interaction at sorbent interfaces. Journal of Chromatography B: Biomedical Sciences and Applications. 1997;699(1):117-31.

[3] Chen Z, Zhao P, Li F, Leier A, Marquez-Lago TT, Wang Y, et al. iFeature: a Python package and web server for features extraction and selection from protein and peptide sequences. Bioinformatics. 2018;34(14):2499-502.

[4] Chen D, Li Y. PredMHC: An Effective Predictor of Major Histocompatibility Complex Using Mixed Features. Front Genet. 2022;13:875112.

[5] Karosiene E, Rasmussen M, Blicher T, Lund O, Buus S, Nielsen M. NetMHCIIpan-3.0, a common pan-specific MHC class II prediction method including all three human MHC class II isotypes, HLA-DR, HLA-DP and HLA-DQ. Immunogenetics. 2013;65(10):711-24.