Prediction of T cell type based on TCR beta chain sequences using Convolutional Neural Networks

Alessandro Montemurro, Henrik K. Munch, Ibel Carri

Abstract

T helper and T killer cells are two different type of T cells and are differentiated on their expression of either CD4 or CD8 molecules. CD4+ T cells recognize peptides in complex with MHC class II molecules and CD8+ T cells recognize peptides in complex with MHC class I molecules. These two cell types also elicit different immune responses and it is useful to know the T cell type when analyzing the immune system of a patient or a biological model. Current methods to differentiate between these cell types include labor intensive flow cytometry and immunohistochemistry.
In this study, Convolutional Neural Networks (CNNs) are applied to predict the type of T cell based on the beta chain sequence of the T cell receptor (TCR). The obtained model would require further optimization in order to predict with confidence the type of T cell based on TCR sequences.

Introduction

T cells are part of the adaptive immune system and they express CD4+ or CD8+ molecules in the membrane. The T cell receptor (TCR) of CD4+ T cells, also called T helper cells, recognize peptides derived from extracellular proteins in complex with MHC class II molecules and modulate the immune response. The TCR of CD8+ T cells, also called cytotoxic T cells, recognize cytosolic peptides in complex with MHC class I molecules and kills the target cell. The TCR is an heterodimeric protein consisting of an alpha and a beta chain. Each of the protein domains accomodate a conservative and variable region and both have three hypervariable domains, called complementarity determining regions (CDR's). CDR3 has been found to be the most variable part of the protein, which is responsible for recognising the processed antigen presented by the MHC.
Current methods to subcategorize T cells includes flow cytometry or immunohistochemistry, which utilizes the expression of CD4 and CD8 molecules. After the development of next generation sequencing techniques, many groups have used this technique to sequence the TCR repertoire to explore the immune system of patients or biological models. As the sequencing has become more feasible, it would be useful to develop a tool that predicts the type of T cell *in silico* based on these sequences, without requiring any additional experimental grouping technique.
Previous works have evaluated whether specific features in CDR3 beta sequences could be used to discriminate between T cell types. Li *et. al.* found that the CDR3 beta in CD4+ T cells generally has a higher abundance of positively charged amino acids as opposed to more negatively charged amino acids in CDR3 beta CD8+ T cells[2]. Further, a position specific

preference of amino acids were found, depending on the peptide lengths. The most abundant length of the CDR3's for both CD4 and CD8 were found to be around 13 amino acids.

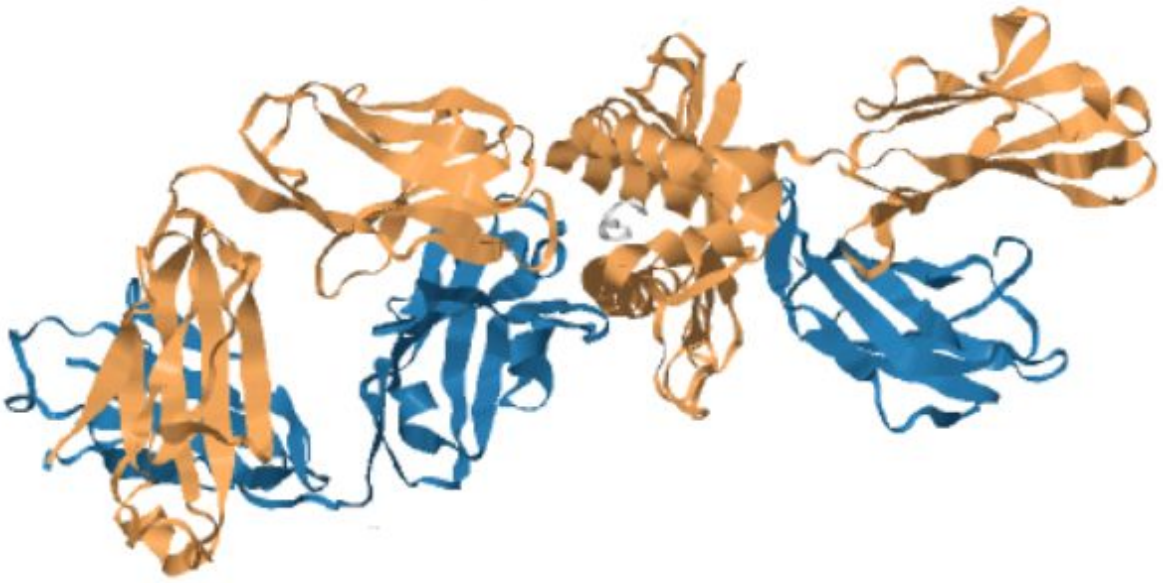

Fig. 1: Interaction of T cell receptor (left) with an MHC-peptide complex (right).
Image created based on PDB 3CVH [1].

Data from the weight matrices for each length of peptides (10 to 15 amino acid lengths) indicated a pattern of particularly lysines for each length of peptides. These data suggest that it could be possible to give an indication of whether the cell is a CD4+ or CD8+ cell based on sequences from the beta chain.
Other approaches to obtain a predictive model of T cell type have used statistical analysis of the usage of VDJ genes and the length of the CDR3 sequences[3,4]. In this study, Convolutional Neural Networks (CNNs) are applied to predict the type of T cell based on the sequences of the beta chain of the TCR.

Materials and methods

Dataset

The data used in this study consists of an independent dataset retrieved from T cell repertoire sequencing[5]. It consists of 675.024 sequences of TCR beta chains as well as information about whether the T cell is CD4+ or CD8+. An encoding scheme was established for the two classes, where 0 is CD8+ and 1 is CD4+.
In the dataset, the two classes were unbalanced, with a majority of CD4+ cells. The two classes were balanced, removing sequences from the CD4+ class so that the final dataset contains 50% of observation from each class.
Since TCR sequences have different lengths and CNNs requires that the inputs have the same size, sequences were filled by padding "X" to get inputs with the length of the longest sequence of the dataset.

Amino acids were encoded using BLOSUM50 matrix and zero-vectors were added to the BLOSUM50 matrix for encoding ‘X'.

Convolutional Neural Network

The problem of predicting the type of T cell based on the beta chains turns into understanding the which part of the beta chain sequences carries the information about the T cell type.

The machinery of Artificial Neural Networks (ANN) is suitable for modeling non-linear relations between TCR and CD4+ or CD8+ molecules and extracting relevant features capable of classify whether a T cell is of one of these types. An ordinary neural network would not work in this case because the position and the length of the sub-sequence carrying information about T cell type is unknown. This suggests to use a Convolutional Neural Network that allows to analyze the sequence using smaller sub-sequences and moving them along all the TCR sequence.

Convolutional neural networks differ from ordinary Neural Networks because of the presence of an extra layer before the hidden layers, the so-called *convolutional layer.* Convolutional layers apply the operation of convolution to the input layer before the input is passed to the hidden layer. Defining the *kernel* as a filter used to screen the input, the sequence is filtered by sliding, precisely *convolving* the kernel across the length of the sequence. The convolution between the input and the filter will produce a map describing how the filter responds to the input in every position of the input sequence. In this way, the network will learn itself the filters that activate when they detect a part of the sequence that contains relevant information[6].

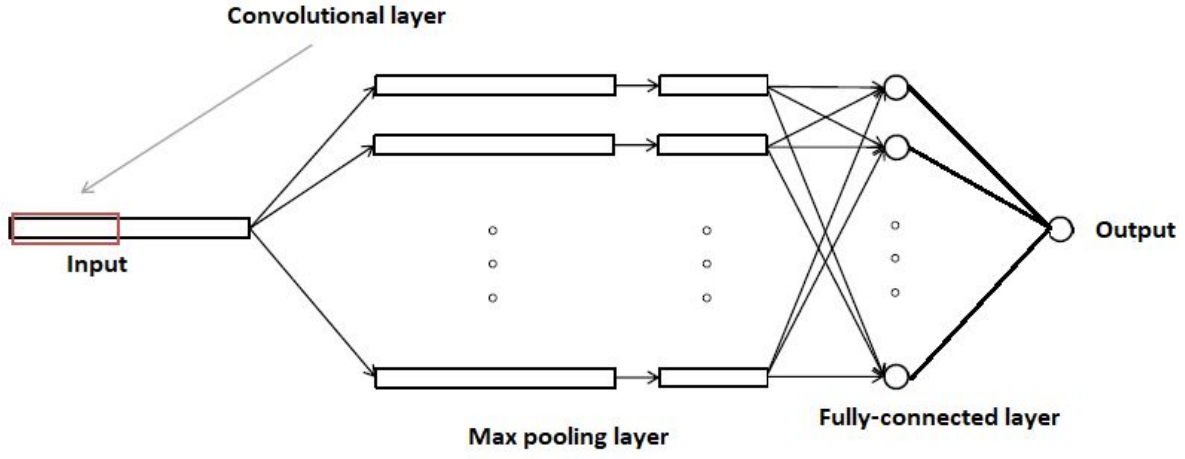

Fig. 2: Architecture of the CNN used in this study[7].

The CNN used in our experiments consists of an input layer, containing the sequences, followed by a convolutional layer. Consequently, the convoluted input is fed into a *max pooling layer*; this operation reduces the dimension of the convoluted input layer gaining computational performance.The output of the pooling layer is fed into the hidden layer. Dropout is applied on the hidden layer, deactivating a fraction of the neurons in this phase of the training; this is done to regularize the model. Figure 2 shows the architecture of the CNN.

The output of the model is a real number between 0 and 1, and can be seen as a probability of the sequence belonging to class CD4+. Instead of setting a threshold and define a *hard* classifier, our model is a *soft* classifier where the performance is assessed using Receiver Operating Characteristics (ROC) and Area Under Curve (AUC).
The hyperparameters of the model are the learning rate, the number of hidden neurons and the kernel size. The model has been trained with different values of these hyperparameters to assess which combination gives the better performance (see Experiments). In the training phase of the model, the Mean Squared Error (MSE) has been used as error function and the activation function for the neurons is a *sigmoid* function.

Cross Validation

The dataset was splitted in two parts: 85% of the sequences were used to train and evaluate the model and 15% to test the model after the training phase. The test set is independent from the train and evaluation sets; it was taken apart from the data set in the beginning of cross validation and it has been never used during the training.
5-fold cross validation was used to evaluate the goodness of the prediction. The data selected for training an evaluation is splitted into five subsets and, iteratively, four folds are used for training and one is used to evaluate the model.
A total of 20 models is trained and tested, giving rise to 20 different predictors. An ensemble of these models is built by taking the mean of the predictions. The general performance of the model is assessed testing the ensemble with the test set.

Experiments

The model was implemented in R version 3.4.4. The CNN was designed and trained using Keras API running on top of TensorFlow.
Different models were trained with and without dropping out neurons of the hidden layer and with different combinations of the hyperparameters:

- From 7 to 20 amino acids as kernel size.
- From 7 to 13 neurons in the hidden layer.
- From 0.001 to 0.1 learning rates.

The batch size is set to 20 and the number of epochs is 25.
Due to high computational costs, the hyperparameters were tested with a randomly sampled subset of 1000 sequences of the dataset. The final model was trained with this hyperparameters but with a subset of 7000 sequences.

Results

The best model was trained with a layer dropout rate of 0.3, a learning rate of 0.003, 15 neurons in the hidden layer and 15 amino acids as kernel size. Figure 3 shows the ROC; the corresponding AUC was of 0.9268 on the left out test set.
The training and test error of the final model are reported in Figure 4.

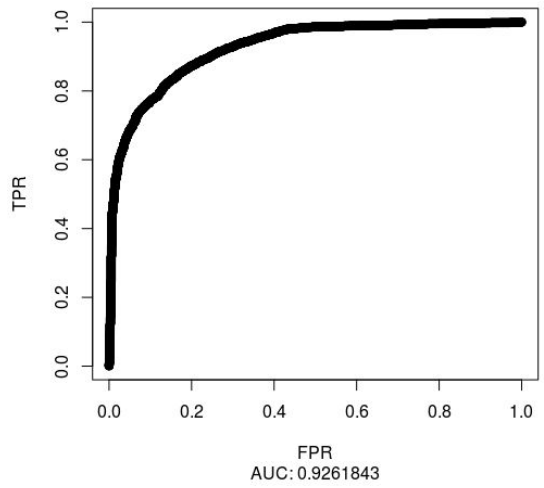

Fig. 3: ROC curve of the model along with the AUC computed on the independent test set.

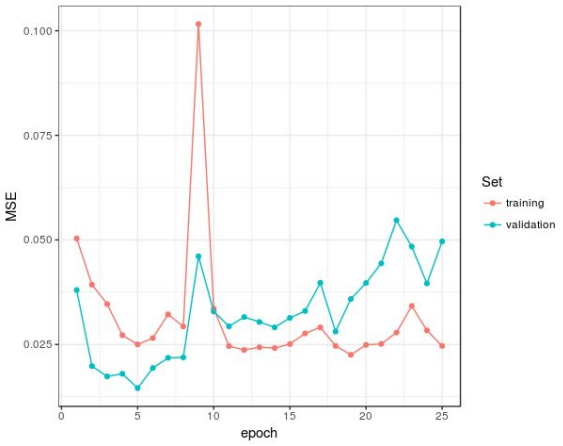

Fig. 4: Training and validation error (MSE) of the model.

Discussion

In Figure 4 it is possible to observe that, while the training error decreases, the validation error increases. This is presumably due to overfitting and the generalization power of the model is poor to accept the predictions with high confidence. A similar behaviour was observed when training with different parameters, so the current model based on beta sequences seems not sufficient to predict the type of T cell with an acceptable confidence.

Previous models that tried to predict the type of T cell have been based on VDJ genes, and that adding the information of the V, D and J genes of the chain as an input, could potentially improve the predicting power of the neural network. Furthermore, the current model was only trained with beta sequences. Training the neural network with both the information of the alpha and beta chains of the same TCR, may also improve the accuracy of the predictions.

This method was trained using the entire sequence of the beta chains, but for example, Li et. al.[2] have demonstrated that there are different amino acid preferences in the CDR3 loop of the beta chain for CD4+ T cells and CD8+ T cells. By training one network only with the sub-sequence that correlates with the type of T cell, it will be possible to train a model with shorter and more specific sequences, avoiding training the model with noise. This can also solve another issue that may be affecting the predictions, which is the highly variable length of the input sequences, which goes from 43 amino acids to 138.

Conclusion

The initial assumptions to build the model allowed us to get a good result in terms of AUC, but the model overfits the data. Our model thus fails in classifying a T cell as CD4+ or CD8+ based on the beta chain and therefore is not adequate to predict the T cell type based on the beta chain sequences.

Acknowledgments

We would like to thank Leon Eyrich Jessen for his help developing the code, Paolo Marcatilli and Martin Closter Jespersen for the analysis and preprocess of the data, and Morten Nielsen for inspiring us to use machine learning techniques to discover new features from biological data.

References

1. Mareeva T, Martinez-Hackert E, Sykulev Y. How a T cell receptor-like antibody recognizes major histocompatibility complex-bound peptide. Journal of Biological Chemistry. 2008 Oct 24;283(43):29053-9.
2. Li HM, Hiroi T, Zhang Y, Shi A, Chen G, De S, Metter EJ, Wood WH, Sharov A, Milner JD, Becker KG. TCRβ repertoire of CD4+ and CD8+ T cells is distinct in richness, distribution, and CDR3 amino acid composition. Journal of leukocyte biology. 2016 Mar 1;99(3):505-13.
3. Emerson R, Sherwood A, Desmarais C, Malhotra S, Phippard D, Robins H. Estimating the ratio of CD4+ to CD8+ T cells using high-throughput sequence data. Journal of immunological methods. 2013 May 31;391(1-2):14-21.
4. Klarenbeek PL, Doorenspleet ME, Esveldt RE, van Schaik BD, Lardy N, van Kampen AH, Tak PP, Plenge RM, Baas F, de Bakker PI, de Vries N. Somatic variation of T-cell receptor genes strongly associate with HLA class restriction. PloS one. 2015 Oct 30;10(10):e0140815.
5. Rubelt F, Bolen CR, McGuire HM, Vander Heiden JA, Gadala-Maria D, Levin M, Euskirchen GM, Mamedov MR, Swan GE, Dekker CL, Cowell LG. Individual heritable differences result in unique cell lymphocyte receptor repertoires of naive and antigen-experienced cells. Nature communications. 2016 Mar 23;7:11112.
6. Herlau T, Shmidt MN, Mørup M Introduction to Machine Learning and Data Mining. Course notes fall 2017, version 3. 2017 Sep 25.
7. Tsinalis O, Matthews PM, Guo Y, Zafeiriou S. Automatic sleep stage scoring with single-channel EEG using convolutional neural networks. arXiv preprint arXiv:1610.01683. 2016 Oct 5.