Classification of bagged sample texts using projections and observations selection

Vladimir (Vova) Psevkin

Thesis submitted in partial fulfillment of the requirements
for the Master of Sciences degree in the Department
of Mathematics, Bar Ilan University

Under the supervision of **Prof. Yoram Louzoun**

February 2021

סיווג דגימות טקסטואליות באמצעות הטלה ובחירת תצפיות

ולדימיר (וובה) פסבקין

עבודה זו מוגשת כחלק מהדרישות לשם קבלת תואר מוסמך במחלקה
למתמטיקה של אוניברסיטת בר־אילן

תחת הובלה של פרופסור יורם לוזון

פברואר 2021
תשס"א

This work was carried out under the supervision of Prof. Yoram Louzoun
Department of Mathematics, Bar-Ilan University.

Acknowledgements

First of all I would like to thank my family in general and my mother in particular for the support she gave me during my studies.

I would like to thank my esteemed supervisor – Prof. Yoram Louzoun for his invaluable supervision, support and tutelage during the course of my M.Sc degree. My gratitude extends to the Faculty of Mathematics for the funding opportunity to undertake my studies at the Department of Data Science, University of Bar Ilan. Additionally, I would like to express gratitude to Madar Yael for her treasured support which was really influential all the period in Bar Ilan. I also thank Dr. Simcha (Simi) Haber, Prof. Reuven Cohen, Prof. Gil Ariel for their mentorship. I would like to thank my friends, lab mates, colleagues and research team – Neta Glazer, Ariel Rozen Zvi, Ido Springer, Reut Levi , Ofek Akerman for a cherished time spent together in the lab, and in social settings. My appreciation also goes out to my family and friends for their encouragement and support all through my studies.

Contents

List of Figures

List of Tables

תקציר

במשך תקופה ארוכה הגישה הרווחת במדעים, היא גישה בה המדענים הציעו מודל של התופעה הנחקרת, בהתבסס על השערות, ואז בדקו את כוח הניבוי של מודל על התופעה זו בניסויים ואת תוקף ההשערות שלהם. כיום, עם ריבוי שיטות לימוד המכונה בכל תחומי הפעילות האנושית, הגישה ההפוכה לבניית מודלים הקלאסים הופכת להיות פופולארית יותר ויותר, בה המודל נוצר ישירות מנתונים אמפיריים.

מטרה מעניינת היא ליישם גישות של לימוד מכונה על קבוצה בקנה מידה גדולה של קולטני תאי T. סיווג ואפיון של קולטן תאי ה־ T של מערכת החיסון מאפשרים לעיתים קרובות לזהות אירועי זיהום עכשוויים ושקרו בעבר. מערכת החיסון ההסתגלותית כוללת תכונות הדומות ללמידת מכונה, מכיוון שהיא מבדילה פתוגנים מסוכנים ממולקולות עצמיות בטוחות. לכל שיבוט של תאי חיסון יש קולטן ייחודי. יכולת החיבור בין קולטנים לפפטידים אנטיגניים קובעת את ההכרה החיסונית וכתוצאה מכך את תגובה מתאימה. בהינתן מדגם המורכב ממכלול של רצפים לאדם, המשימה היא לקבוע לאיזו קטגוריה הוא שייך (למשל היה או לא הייתה לו מחלה). נבחן את הבעיה בצורה מתמטית באמצעות גישה קומבינטורית.

באופן כללי, במשימות סיווג המבוססות על למידת מכונה או סטטיסטיקה, דוגמאות שייכות לקטגוריות שונות. המטרה היא לקבוע לאיזו קטגוריה מדגם שייך באמצעות המאפיינים של אותו מדגם. במחקר זה נעסוק בשאלה מורכבת יותר בהנחה שמספר קטגוריות רחב, שכל אחת מקטגוריות קובעת את ההסתברות להרבה רכיבים, אנו רוצים לדגום באופן אקראי רכיבים אלה ולחזות את הקטגוריה. אנו מתמקדים במקרה בו התפלגויות ההסתברות של התצפיות השונות אינן בילתי תלויות, וההסתברות של דגימת תצפיות דומות גבוהה מההסתברות לדגימת תצפיות רחוקות, בהינתן מדדים מסוימים. אנו מיישמים מתודולוגיה זו על התגובה החיסונית ההסתגלותית.

מערכת החיסון יוצרת באופן קבוע קבוצות של תאים עם קולטנים אקראיים. בנוכחות פתוגן, תאים עם קולטן ספציפי לפתוגן מתחלקים, והשיבוט שלהם מורחב. שיבוטים כאלה

נצפים בתדירות גבוהה יותר מאחרים בדגימות אקראיות של תאי מערכת החיסון. בהנחה שקולטנים המגיבים לאותו פתוגן דומים, אנו מציעים לזהות האם מארח הגיב לפתוגן נתון באמצעות שילוב של מדדים על רצפי קולטן ואלגוריתמים לבחירה מבוססת־מדגם של קטגוריות רחבות.

1 Abstract

For a long time, science was dominated by an approach in which scientists first proposed a hypothesis driven model of the phenomenon under study, and then tested the predictive power of the model in experiments and the validity of their hypotheses. Today, with the spread of machine learning methods in all areas of human activity, the opposite approach to building models, in which the model is formed directly from empirical data, has become more and more popular. Within this context, an important field is the attempt to diagnose conditions from large scale data.

An interesting goal is to apply such a diagnosis is T cell receptor large-scale sequencing. Classification and characterization of the T-cell receptor of the immune system often make it possible to detect current and previous infection events. Interestingly, the adaptive immune system has features similar to machine learning, since it distinguishes dangerous pathogens from safe self-molecules. Each immune-cell-clone has a unique receptor. The binding affinity between receptors and antigenic peptides determines the immunological recognition resulting in an appropriate response. Given a sample composed of a set of sequences per person, the task is to determine to which category it belongs (e.g. had or did not have a disease). We will look at the problem mathematically using a combinatorial approach.

Generally, in classification tasks based on machine learning or statistics, samples belong to different categories. The goal is to determine to which category a sample belongs using the characteristics of that sample. In this research, we will deal with a more complex question assuming several broad categories, each setting the prior probability of many components, we want to sample randomly these components and predict the category. We focus on the case where the probability distributions of the different observations are not independent, and the probability of sampling similar observations is higher than the probability of sampling distant observations, assuming some metrics. We apply this methodology to the adaptive immune response.

The immune system creates continuously groups of cells with random receptors. In the presence of a pathogen, cells with a receptor that is specific to the pathogen divide, and their clone is expanded. Such clones are more frequently observed than others in random samples of immune system cells. Assuming that receptors reacting to the same pathogen are similar, we propose to detect whether a host has responded to a given pathogen using a combination of metrics on receptor sequences and algorithms for the sample-based selection of broad categories.

2 Introduction

Basic Definitions

This thesis stands in between immunology and computer science. While I tried my best to present the problem as an NLP problem and reduced some of the immunological explanations, some knowledge of immunology is necessary. Table 1 includes some basic, simplified definitions that should allow the reader to better understand the thesis. [1]

[1] Some of the definitions were adapted from Ido Springer's research proposal.

Amino Acid (AA)	The building blocks of proteins. Each amino acid is represented by a unique alphabetical letter. There are 20 different amino acids.
Pathogen	An agent that causes disease, such as a virus, a bacterium, or a cancerous cell.
Antigen	An irregular molecule (mostly belongs to a pathogen) which stimulates the immune response.
Epitope	The part of an antigen that is recognized by the immune system, a short chain of amino acids.
TCR (T-Cell Receptor)	T-cells are responsible for the adaptive immune system's cellular response. Each T-cell has a specific receptor on the cell surface. The T-cell receptor can bind an epitope. A successful binding can activate the T cells carrying the TCR and lead to different possible outcomes. TCRs are composed of amino acids, and are nearly randomly generated in a process called VDJ recombination. Thus, the TCR repertoire is very diverse, and can be treated as a set of pseudo-random sequences.
TCR Repertoire	A patient's immune repertoire encompasses all the TCRs in the patient's body. The TCR repertoire is unique to every person. The TCR repertoire encodes the memory of the adaptive immune system (by containing all previously activated TCRs). Current methods allow for shallow sampling of this repertoire, with typically 10^6-10^8 T cells out of $10^12 - 10^13$ existing cells
V and J genes	Genes that are randomly rearranged in a process called VDJ recombination, as part of the creation of the TCR sequences.

Table 2.1: Basic definitions in immunology

Notation and Definitions

Machine learning is a class of artificial intelligence methods whose characteristic is not a direct solution to the problem at hand, but, instead, an algorithm that computes some loss function of the observations. An essential subtask of machine learning is the classification of the data. Often, observations are divided into classes. Given a finite set of observations with known classes, the algorithm is designed to predict with which class a new observation is affiliated. Most classification algorithms focus on classifying a single finite-size sample with a single class. An exception to this rule is set classification. Given a set of observations with a single class (e.g. a set of books composed of many words each written by an author), one aims to classify the set based on the properties of the observations. One of the common applications of set classification is text classification. For example,when performing sentiment analysis with Bag-of-words (BoW) preprocessing representation [21].

The BoW uses the word frequency. Using the notation above, the book is a set, and the words are observations in the set. The goal is to classify the entire book and not each word by itself. Given the total amount of words, the occurrences of the most common words are counted. The book is then represented by a frequency vector, and this vector is used as the input of a classifier. This process creates a classification model according to the characteristics of the final distribution of the words [15]. The key idea is to represent each work by a frequency, then represent the text by a vector of frequencies. This vector quantification procedure allows us to represent each class by the visual word histogram, often referred to as the BoW representation, which then converts the object classification problem into a real value classification problem.

Sometimes the samples for each set are taken from an infinite space. Thus, one cannot tally the probability of each event in the space. In such cases, the method of adjusting frequencies to observations based on their observed frequencies in the sample will not work and will not create a finite dimension input for machine learning models. We here propose to study such classification problems and aim to classify sets composed of samples from an infinite space. We explicitly assume that a similarity metric can be defined in this space. We propose an algorithmic solution that not only uses the distribution of observations but also uses internal and hidden characteristics of the observation.

As an application, we plan to develop methods for such group classifications using a specific biological implementation of the classification of patients based on their adaptive immune response. The following section describes the biological setup of the question and can be skipped if only the mathematical question is of interest. Also, a glossary is provided at the end to clarify all concepts.T lymphocytes (T-cells) are crucial in the cellular immune response. Their immense diversity enables extensive antigen recognition and is achieved through a vast variety of T-cell receptors (TCRs).

Biological Description

Successful recognition of antigenic peptides bound to major histocompatibility complexes (pMHCs) requires specific binding of TCR to these complexes, which in turn directly modulates the cell's fitness, clonal expansion, and acquisition of effector properties. The affinity of a TCR to a given peptide epitope and the specificity of the binding are governed by the heterodimeric $\alpha\beta$ T-cell receptors [18]. However, most of the TCR's binding to target MHC-peptide is determined by the β chain [5]. Within the TCRβ chain, the complementarity-determining region 1 (CDR1) and CDR2 loops of the TCR contact the MHC alpha-helices while the hypervariable CDR3 regions interact mainly with the peptide. In both TCRα and TCRβ chains, CDR3 loops have by far the highest sequence diversity [14] and are the principal determinants of the receptor-binding specificity.

Following specific binding of T cell receptors to viral and bacterial-derived peptides bound to MHC, or from neoantigens, the appropriate T cells expand, resulting in the over-presentation of T cells carrying such receptors [16]. Recently, high-throughput DNA sequencing has enabled large-scale characterization of TCR sequences, producing detailed T cell repertoire (RepSeq). Expanded clones are more likely to be observed in Rep-Seq than non-expanded clones and maybe thus used as biomarkers for the presence of their cognate target [1]. Using these repertoires as large-scale biomarkers require a precise enough distinction between TCRs binding distinct targets (often referred to as "reading the repertoire").

A classifier producing such a distinction is currently not available. A direct approach for using TCR Rep-Seq as biomarkers has been proposed by who detected patients that have CMV based on their full repertoire and the choice of TCRs that differ between CMV positive and negative patients. This approach is based on the presence of public TCRs that are highly specific and repetitively observed in the response of different hosts to the same peptide (often denoted public clones). CMV (Cytomegalovirus) is a herpes virus present in a large fraction of the adult western population.

However, many TCR responses are characterized by a high level of cross-reactivity with single TCRs binding a large number of MHC-bound peptides, and single peptides bindingmany TCRs. TCRs binding the same MHC-peptide may share similarities but possess different CDR3 sequences. Thus, while for public clones the task of deciphering the relation between a peptide and the TCR binding is based on tallying the candidate public TCR, for most TCRs which

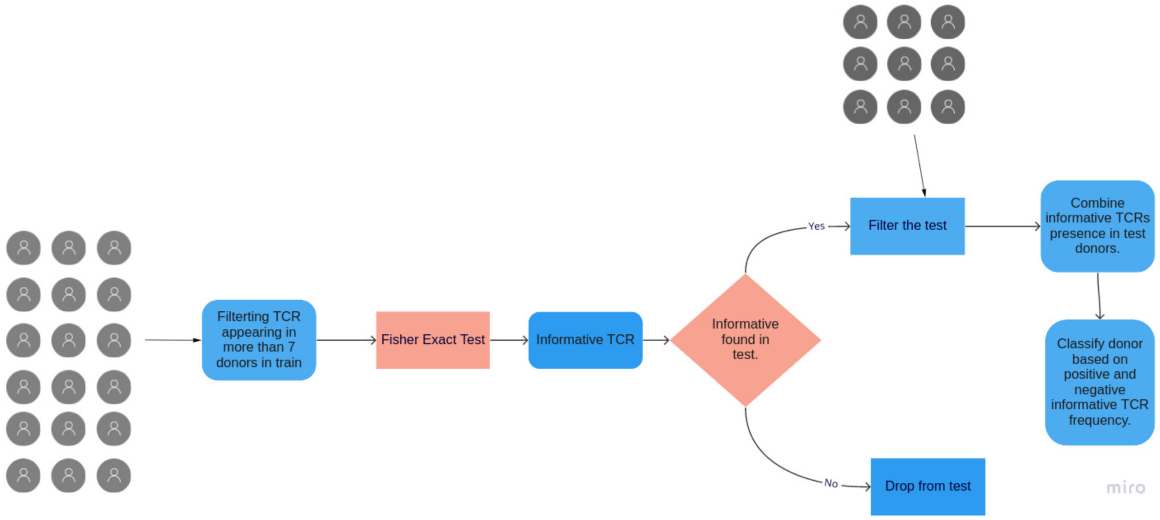

Filterting TCR appearing in more than 7 donors in train
Fisher Exact Test
Informative TCR
Informative found in test.
Yes
Filter the test
Combine informative TCRs presence in test donors.
Classify donor based on positive and negative informative TCR frequency.
No
Drop from test
miro

Figure 2.1: Flowchart, a regulated system describing an algorithm and process, in which individual steps are described as blocks in different shapes, connected by lines indicating the process.

are highly cross-reactive, a probabilistic approach is required. We here propose to develop such methods using the relation between receptors to classify a host based on its receptor composition.

3 Research Goals

The purpose of this study is to develop an algorithm to predict the binary state of a donor with a set of sequences based on the sequences. Formally, given a set of donors I, where each donor is represented by a set of sequences $x_{i,1}, ..., x_{i,m_i}$, and a class $y_i \in 0, 1$. The goal is to predict y_i, using $x_{i,1}, ..., x_{i,m_i}$. Some specific peculiarities of this data are that:

1. Most sequences x_{im_i} do not appear in most donors.
2. Most sequences x_{im_i} are unrelated to the class y_i (i.e. the presence or absence of the sequence is not correlated with y_i.
3. The sequences are pseudo-random sequences (i.e. with no clear predefined grammar), with no prior knowledge of their properties.

In short, our goal will be to perform three stages of analysis:

(A) Determine which TCR x_{i,m_i} are of interest. This is mainly a feature selection step. Interesting sequences are those that may contain information on y_i.

(B) Predict y_i, based on a subset of interesting x_{i,m_i}.

(C) Predict y_i, based on the grouped x_{i,m_i} in a given donor.

The difference between B) and C) is that in C), we plan to merge subsets of xi,mi into new values (that will be further denoted meta clones). Then we plan to detect interesting meta clones, instead of interesting clones. Following [8], we will use a set of T cell clones, each with a different TCR (See biological explanation above) from donors that either had or did not have CMV. Most TCRs are not shared between donors. We will try to classify whether the host had or did not have CMV only using the TCR composition. We will do a similar analysis for COVID-19 donors.

Existing approaches (9) rely on the prevalence of a very limited set of sequences: if a sequence is not common enough such that it does not provide comprehensive information on its class, it is ignored. This leads to most sequences being ignored. As such, current methods fail in the absence of deep sampling and a large training set. We here propose to include the similarity between sequences and group similar sequences to classify the host, using "Meta-clones" composed of groups of similar sequences, and develop a classification algorithm, based on such meta-clones.

A meta clone will be defined as a group of clones with similar TCRs. The definition of meta clones poses two challenges:

(A) Each TCR is represented as a gene name and a specific amino acid sequence (CDR3 sequence) representing its binding site. To define the similarity between TCRs, a metric has to be defined on them. We propose to define such a metric using a machine-learning-based projection of each TCR into a K dimensional real vector.

(B) Given such a projection, the next challenge will be to optimally merge similar TCRs to a meta clone. We plan to use either graph-based methods (define an edge between clones with a distance between their TCR projections of less than some cutoff and finding connectivity components) or using density-based clustering (e.g. DBSCAN).

Symbol	Description
TCR	T cell Receptors
$\psi(i)$	Numerical representation of TCR
i	Receptor
j	Donor
D(j)	Donor j
y(j)	Label of donor
J(i)	Gene j in TCR i
V(i)	Gene v in TCR i
ζ	1 for CMV+ and 0 otherwise
h(i,j)	Does a particular donor j contain TCR(i)
f(i,j)	Frequency of a receptor i in a donor j
c(i,j)	Counting of a receptor i in donor j
α_0, α_1	Prior value from Emerson article
β_0, β_1	Prior value from Emerson article
Gr_p(j)	Number of positive TCR
Gr_n(j)	Number of negative TCR
$Score(i)$	Association between a receptor i and testing positive

Table 3.1: Math symbols

4 Methods

Notation and glossary

Given a set of donors and their T cell Receptors (TCR), where each receptor is represented as an ordered triplet of $Vgene$, $CDR3$ $amino\ acid\ sequence$, $Jgene$. In all following methods and results, we use the following convention - TCRs are denoted by the index i and donors are denoted by the index j - $TCR(i)$, and donor $D(j)$. Each donor is associated with a binary status $y_j \in 0, 1$, and a set of TCRs. We denote that a TCR i is present in donor j as $x(i, j) = 1$, else $x(i, j) = 0$. We define for each receptor three values, $CDR3(i)$,$V(i)$, and $J(i)$. The receptors are represented as sequences. When projecting the receptors into a real valued vector, we only use the CDR3, and have a projection function:

$$f : CDR3(i) - > \psi(i) \in R^n \tag{4.1}$$

T cell repertoires from sampled donors

The data includes 641 donor files in a training set. The test set include 120 donors. all data were taken from Emerson study. Each file applies amino sequences and two types of V and J genes that make up a combination [8].

The second Data used is COVID-19, that downloaded from https://www.ireceptor-plus.com/ The international iReceptor Plus consortium aims to promote human immunological data storage, integration and controlled sharing for a wide range of clinical and scientific purposes. platform to integrate distributed repositories of Adaptive Immune Receptor Repertoire sequencing (AIRR-seq) data. This information will be used for enabling improved personalized medicine and immunotherapy in cancer, inflammatory and autoimmune diseases, allergies and infectious diseases.

iReceptor Plus [3] enables researchers around the world to share and analyze huge sets of immunological data taken from healthy and sick people that have been sequentially distributed and stored in databases in many countries.

Projections

In our work, we use a TCR autoencoder to preserve the input vector $TCRi$, while reducing our data dimensionality to a fixed size vector zi. We make different uses with this network, which differ from each other by the combination process of the vector $TCRi$ and its expected result yi.

The training of the TCR autoencoder includes several steps of data processing. The first step is representing each of the amino acids in the input sequence by a one-hot vector. There are twenty possible amino acids and an additional stop codon is required, thus the one-hot vector's size is 21, where one index is set to 1 (by the corresponding amino acid) and twenty zeros. Since the length of TCR sequences is not constant, a zero padding is applied to complete the vectors to the required input length. Each instance was then processed by an autoencoder network and encoded to size $\mathbb{R}^{30}$.

The autoencoder network contained three layers of 300, 100, and 30 neurons as the encoder and a mirrored network as the decoder. The network was trained with a dropout of 0.1 and ReLU as the activation function. An MSE loss function was implemented to compare each input sequence to the resulting decoded sequence.

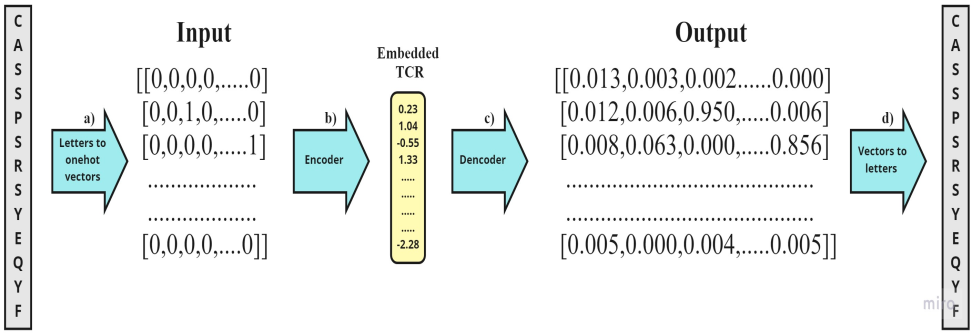


Figure 4.1: A rough description of the TCR autoencoder's ([7]) architucture. **a)** Each amino acid is assigned a one-hot vector representation in IR^{21}. After padding, each TCR is represented by a matrix in $IR^{21\times 28}$. **b)** Using 2 fully connected hidden layers of sizes (300, 100), the encoder converts TCR sequences of size $IR^{21\times 28}$ to a vector in IR^{d}. **c)** The decoder performs an opposite task to the encoder, and decodes the embedded TCR back to size $IR^{21\times 28}$. The encoder uses fully connected hidden layers of sizes (100, 300), and a softmax layer. **d)** Each output vector is assigned an amino acid based on the original one-hot encoding.

Method comparison

Method comparison was performed by the AUC score. The Area Under the Curve (AUC) is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve. The higher the AUC, the better the performance of the model at distinguishing between the positive and negative classes.

Choice of interesting receptors

In order to select a sub group of the receptos $TCR(i)$ that are informative on the condition $y(j)$ of donor $D(j)$, three methods were used:

- We computed for each $TCR(i)$ the Pearson correlation between $X(i,j)$ and $y(j)$ for all j in the training set, and used the 200 TCRs with the minimal p values.
- F.E.T - We computed for each $TCR(i)$ The Fisher exact Test between a target $y(j)$ and $X(i,j)$ for all j in the training set, and used the 154 TCRs that passed fisher test
- Selection of TCRs based on deviation from ex-pected fraction in positive samples.

Unique combinations were taken, a sum of all the frequencies was performed in addition a performance count of the same combination was performed. For each computer combination EXPECTED VALUE (ratio of positive combinations to total, multiplied by the number of impressions) then perform CHI-SQUARE, double the coefficient of EXPECTED VALUE The process is performed for a training set and a test set Scatter the calculation in a Cartesian system whose axes are training and test, the significance of the color giving the cumulative frequency

Reproduction of Emerson results

In order to reproduce the results reported in [8], we performed the following steps:

1. We used 641 donors as a training set and 120 donors as a test set from their publication, where each donor has a set of TCRs, and a class of either CMV+ or CMV-. We ignored donors with no CMV classification.
2. We computed for each TCR i and for each donor j , whether the TCR appears in this donors and produced a binary matrix $h(i,j)$.
3. We removed TCR that appear in less than $k = 2$ donors ($\sum_j h(i,j) < 2$). These TCRs were not used for further analysis.
4. We defined for each remaining TCR i the total number of times it appears in CMV+ and CMV- donors. $Np(i) = \sum_{y(j)==1} h(i,j), Nn(i) = \sum_{y(j)==0} h(i,j)$

5. We used a Fisher Exact Test on all receptors i on $Nn(i)$ and $Np(i)$ with priors of α_0, α_1, β_0, β_1 and chose ”golden receptors”, which we defined as receptors with a p value of less than $P < 1 \cdot 10^{-4}$

6. For each ”golden receptor” i we defined the receptor as being positive if the ratio of the number of impressions in positive donors to the total is greater than the ratio of the number of appearances in negative donors and negative otherwise.

7. We computed for each donor j, the number of positive and negative ”golden receptors” in this donor $Gr_p(j), GR_n(j)$.

8. If a donor had no ”golden receptors”, he was assumed negative.

9. If a donor did have ”golden receptors”, we defined a classification score for this donor as $Score$

10. We used $Score$ as our binary classifier, and computed the Area Under Curve for both the training and test set using this score.

Selection of TCRs based on deviation from expected fraction in positive samples

An alternative methods to define golden receptors is to check the deviation from the expected fraction in positive samples. To detect such receptors, we performed the following steps:

1. We used 641 donors as a training set and 120 donors as a test set from their publication, where each donor has a set of TCRs, and a class of either CMV+ or CMV-. We ignored donors with no CMV classification.

2. We computed for each TCR i and for each donor j , whether the TCR appears in this donors and produced a binary matrix $h(i, j)$.

3. We removed TCR that appear in less than $k = 7$ donors ($\sum_j h(i, j) < 7$). These TCRs were not used for further analysis.

4. We computed the $Score$ of a receptor i as:

$$Score(i) = \frac{\sum_j (h(i,j)\zeta_j) - \frac{np}{nn+np} \sum_j (h(i,j))}{nn+np}, \tag{4.2}$$

where ζ_j is one for positive donors, and zero for negative donors. nn is the number of negative donors, and np is number of positive donors.

Machine learning

We used two different aproaches for the machine learning. The first approach was a language model using direct representations of the receptors, and the second approach was agnostic to the receptor representation and only used the frequency or presence/absence of the receptors in each donor.

Neural network with embedding layer

We developed a language based model, witht he following steps:

1. We used 641 donors as a training set and 120 donors as a test set from their publication, where each donor has a set of TCRs, and a class of either CMV+ or CMV-. We ignored donors with no CMV classification.

2. We computed for each TCR i and for each donor j , whether the TCR appears in this donors and produced a binary matrix $h(i, j)$.

3. We removed TCR that appear in less than $k = 7$ donors ($\sum_j h(i, j) < 7$). These TCRs were not used for further analysis.

4. We selected a subset of golden receptors based on the deviation from expected fraction in positive samples, as described above.

5. For each donor i we used only the golden TCR_j.

6. Each donor had a different number of golden TCR. We computed the maximal number of golden sequences a donor had in the training set. In order to know the size of zeros padding in all donors that include golden TCR's.

7. We produce a neural network with embedding as an input to a hidden layer with 16 neurons. Specifically, We represent each each different TCR_j (include the padding), as a 7 dimensional real vector initiated with a normal distribution with zero mean and a variance of one.

8. Both the vector representation and teh network weights are adapted during the learning session using a binary cross-entropy loss, defined as.

$$Loss(y,z) = \begin{cases} z - zy + log(1 + e^{-z}) & z \geq 0 \\ -zy + log(e^{z} + 1) & z < 0 \end{cases} \tag{4.3}$$

where z is the output of a single neuron sigmoid output layer, and y is the binary true label.

9. The activation function in the hidden layer is a Leaky ReLU [12]. A droupout with a rate of 0.2 was set between embedding layer and hidden layer. Batch normalization was set after the hidden layer. The L2 normalization coefficient was set to $5 * 10 - 3$ in the ADAM optimizer.

$$f(y) = \begin{cases} y & y > 0 \\ ay & y < 0, \quad 0 < a < 1 \\ 0 & y = 0 \end{cases} \tag{4.4}$$

Classification based on golden on receptors

Two methods were proposed above to choose golden receptors. Independetly of the methods to choose the receptors, we used multiple standard machine learning approaches to classify donors, assuming a predefined set of golden receptors.

- **Random Forest** We used a Random Forest model, parameter calibration was performed in cross validation. $max\ features = 30$, $n\ estimators = 37$, $criterion =' gini'$, $max\ depth = 21$, $min\ samples\ leaf = 4$
- **XGboost** We used a Extreme Gradient Boosting 'XGBoost' model, parameter calibration was performed in cross validation.
  $learning\ rate = 5 * 10^{-}2$, $n\ estimators = 1000$, $min\ child\ weight = 1$
- **Neural networks** we used a two layer neural network as a model. parameter calibration was preformed using NNI (neural network intelligence).
  batch size = 16, first hidden layer's size = 128, second hidden layer's size = 32, learning rate = 0.0001, number of epochs = 7, number of golden receptors = 150, train precentage = 70,

Hyper parameter tuning

The grid search provided by GridSearchCV in sklearn library [17], exhaustively generates candidates from a grid of parameter values specified. The purpose of GridSearchCV is to test with different parameters for at least one thing in your pipeline

5 Results

The same receptor can appear multiple times in the same sample, and we merged similar receptors

Receptors are defined through their V and J genes, as well as through the amino acid composition of their CDR3 region (See glossary in table 3.1 above for notations). However, different nucleotide TCRs can be translated to the same amino acid sequence. Since we are only interested by the functional properties of the receptor, we merged all nucleotide sequences with the same amino acid sequence, and produced for each receptor (i) and each donor (j) a count $C_{i,j}$ representing the number of different nucleotide sequences producing the amino acid sequence of receptor i, with the same V and J genes.

The distribution of $C_{i,j}$ is similar among donors (See Figure 5.1 for typical distribution for 5.2 donors). Moreover, it does not differ between positive and negative samples (Look at Figures 5.7 and 5.3, 5.4).

Each TCR ((i)) is also associated with a frequency. When combining different nucleotide sequences to a single amino acid sequence, we summed the frequencies of all the appropriate nucleotide sequence. Again, there was no difference between the frequency distributions between donors (Figure 5.3, nor is there a difference between positive and negative samples Figure 5.4).

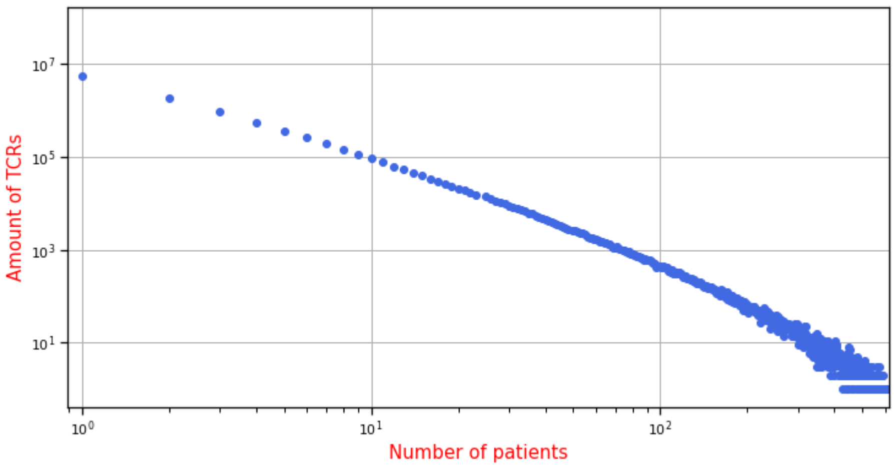


Figure 5.1: Distribution of TCR frequency defined as the number of donors in the training set that have this receptor. The distribution is scale free as observed in the straight line in the log-log plot.

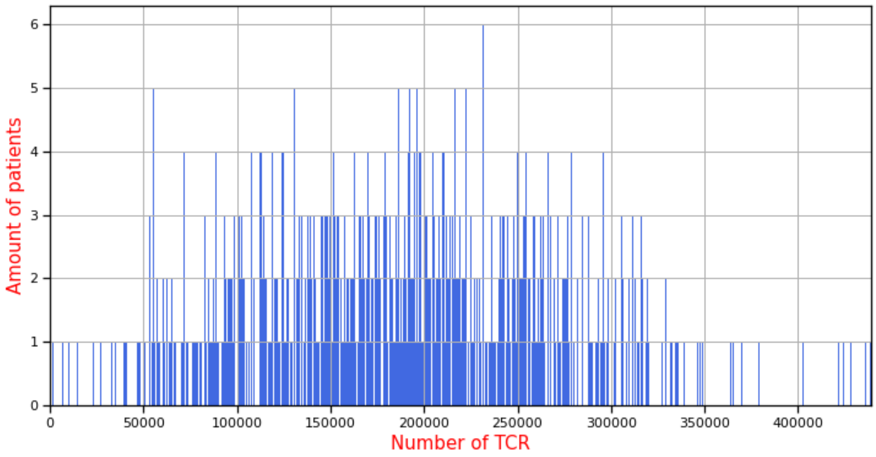


Figure 5.2: Histogram of unique TCR peptide in different donors in training set. The number of unique TCRs vary from 1.e4 4.e5 unique TCRs per donor.

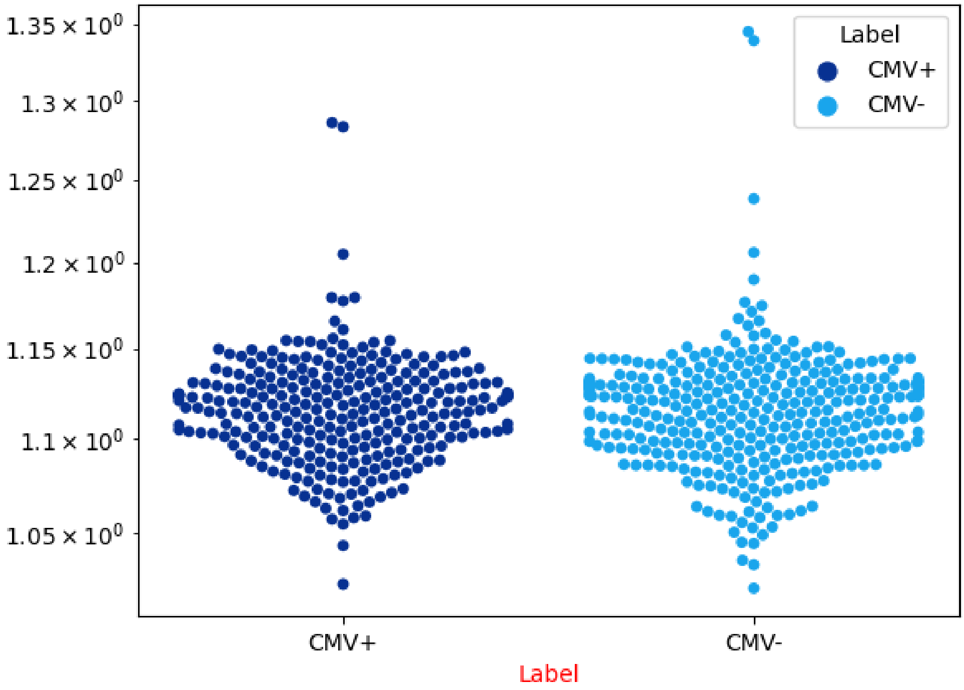


Figure 5.3: Comparison between the average count of positive and negative classes. Each dot is the average count in CMV+ and CMV-. Blue dots are CMV- and navy dots is CMV+. One can see that both groups are highly similar.

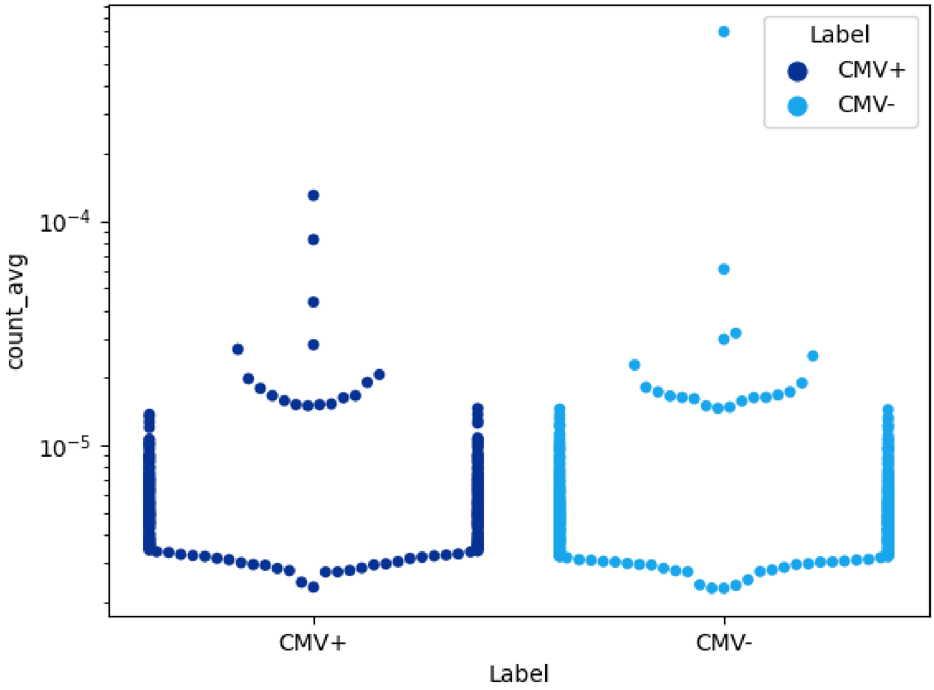
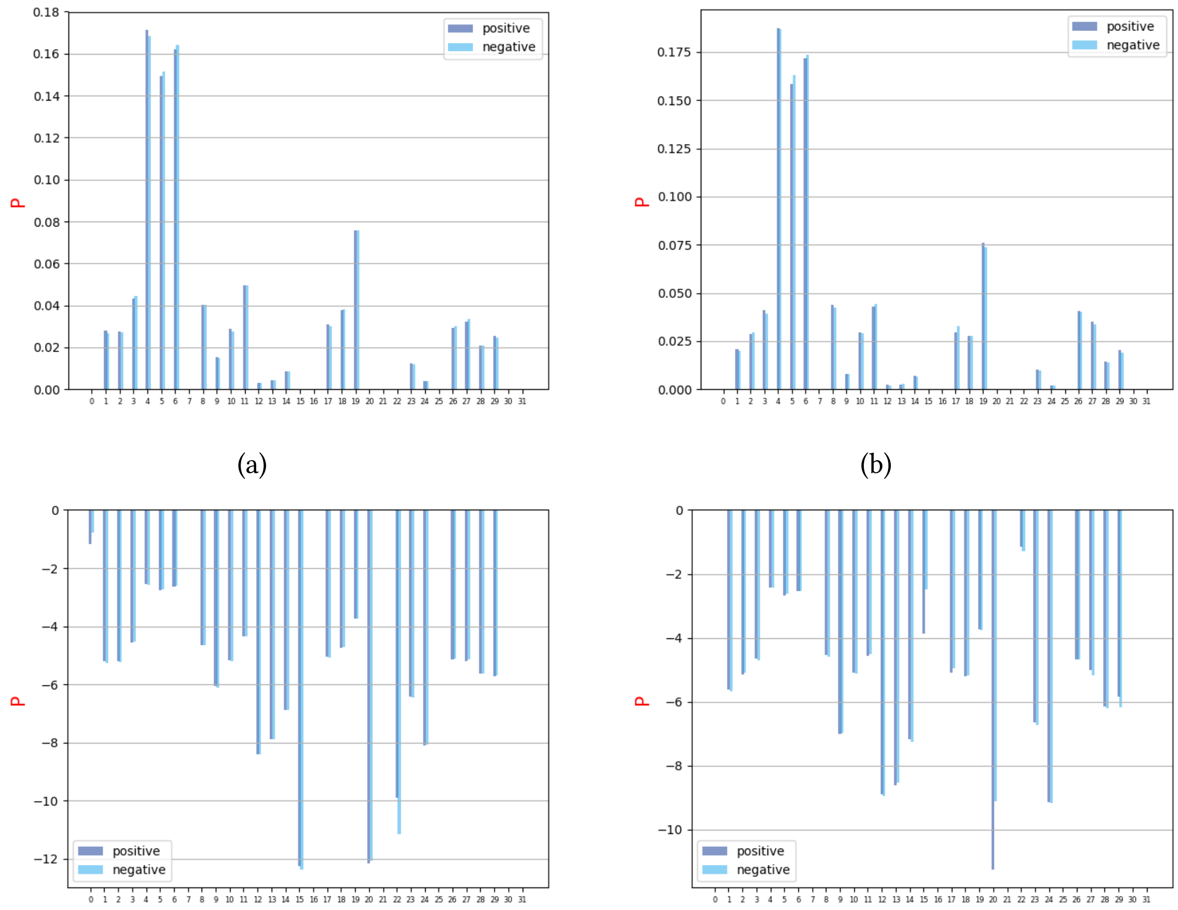


Figure 5.4: Comparison between the average frequency of positive and negative classes. Each dot is the average frequency in CMV+ and CMV-. Blue dots are CMV- and navy dots is CMV+. One can see that both groups are highly similar.



(a)

(b)

(c)

(d)

Figure 5.5: Distribution and Log distribution of v gene over positive and negative donors on the train and test set. Histogram of the v gene by connecting the v gene in each donor, dividing by the number of receptors in that person, and then averaging over all individuals

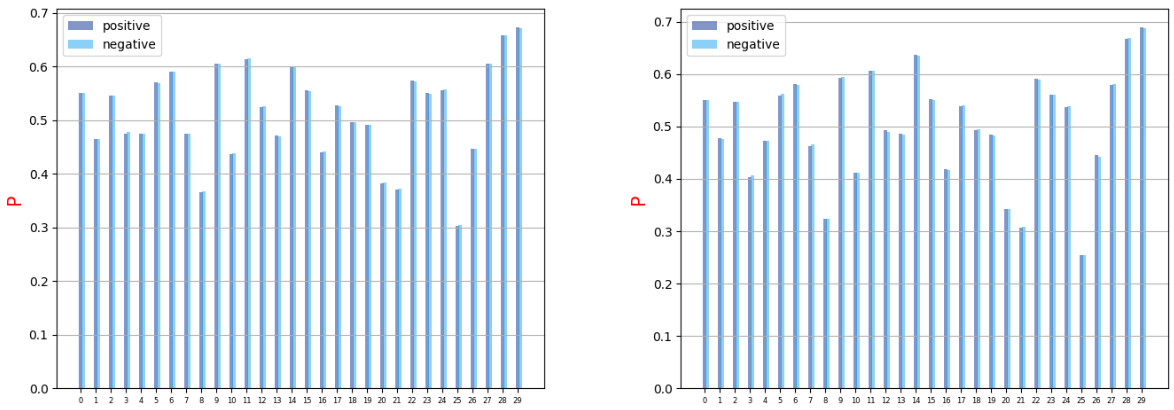
Some receptors are shared among multiple samples, and those are the only interesting receptors

The number of analyzed TCRs is very large, and only some of them can be the basis for building a classification mechanism. Our initial approach was to check if certain TCRs are shared by multiple donors. In constructing the prototype of the statistical model based on Fisher's exact test we used an approach that a meaningful TCR must be included in at least two donors. From the initial filtering the total number of different TCRs dropped from 70 million to an order of magnitude of 10 million so we created a state of clearing the observations and lowering the complexity of the data modeling.

To further reduce the number of TCRs to analyze, We follow Emerson et al [8]), where we have seen that all informative receptors are found in at least 7 (Figure 5.2 patients). We applied a similar threshold here, and obtained 256 different TCRs.

Each donor has set of TCRs, each with its own V,J and CDR3 amino acids. In order to classify donors, we first tested whether TCRs contain information on the class (CMV+ or CMV-) of the donor. Such TCRs should be shared by more than one donor. We thus, first tested for the presence of such TCRs.

Formally, we looked for TCR (V,J,CDR3 TCR combination) that are in at least K donors. For each TCR) we counted the number of donors including the combination, we constructed a histogram of the number of donors containing each combination (Figure 5.1). As can be clearly seen, many TCRs are shared by multiple donors. Note that not all shared TCRs are informative on the donor condition, but all informative TCRs must be shared. We thus filtered away all the TCRs shared by less than $K = 7$ donors. In previous similar attempts [8], the initial filter was $K = 2$.



(a) (b)

Figure 5.6: Distribution of Amino acid in train and test data set over positive and negative donors. A histogram of the projections, in which this number is added to each receptor and divided by the number of receptors

Even when filtering, there is no difference between global properties of positive and negative samples

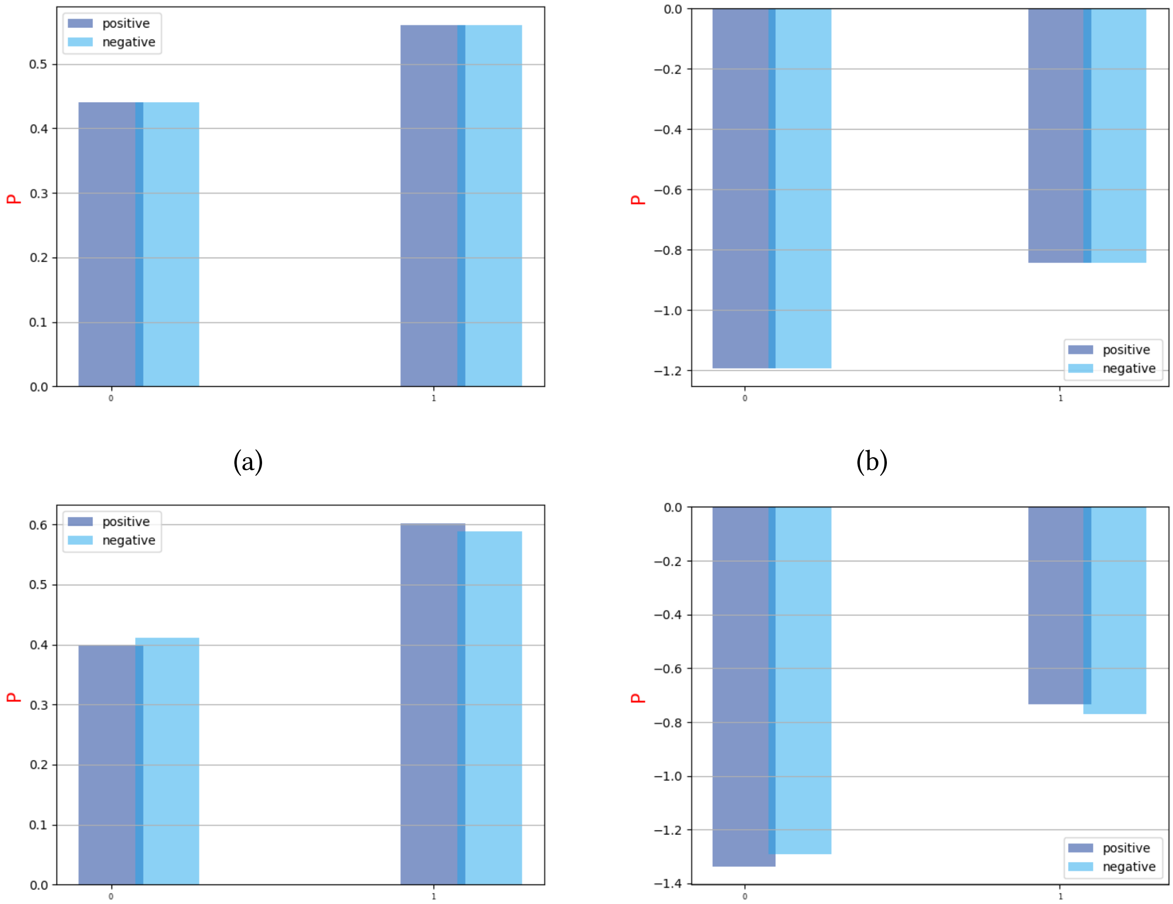
Our concern was after filtering the TCRs that occur in at least 7 donors we will get a large variance in additional characteristics such as V and J that are taken into account as part of the characterization of the important TCRs. After testing we have seen that there is no difference between the properties of the important TCRs.

After filtering non-informative TCRs (TCRs not included in at least K donors). Our goal was to test whether the virus-positive (CMV+) donors could be distinguished from CMV- donors through the aggregated properties of the repertoire.

The repertoire of a donor can be caraterized by multiple staistiscs:

- The V gene composition.
- The J gene composition.
- the CDR3 amino acid TCR composition.

The amino acid composition does not catch the relation between amino acids in the CDR3 TCRs. Thus, we projected the CDR3 sequence using an autoencoder (See Methods section 4, and further description below). We computed for each statistic above the average over the TCRs in CMV+ and CMV- donors. Figure 5.5c shows the V gene composition and as can be seen, there is no difference between the CMV+ and CMV-. In additions shows that J gene also not different between CMV+ and CMV- Figure 5.7b. In order to illustrate that there is no difference in the genes in the training and the test, we also examined the logarithmic scale.
In order to investigate and test variability between the components of a mixture, we examined genes V and J for each composition see (Figures 5.7 and 5.5).



(a)

(b)

(c)

(d)

Figure 5.7: Distribution and Log distribution of j gene over positive and negative donors on the train and test set. Histogram of the j gene by connecting the j gene in each donor, dividing by the number of receptors in that person, and then averaging over all individuals

Previous studies found that donors can be classified based on a subset of TCRs

We here follow the study of Ryan Emerson [8]. The study of Emerson contains multiple stages:

1. Filtering out TCRs that are in less that two donors.

2. Assignment of each TCR to be either positive, negative or non-interesting. This decision is based on the probability that a TCR appears in a positive or negative donor, and thus a connection is made between each TCR and the group where it is most frequent. TCR's with no clear preference to CMV+ or CMV- patients are ignored (See methods section.XXXX for technical details).

3. Then we count for each donor, the number of positive and negative TCR's (as defined above), and apply a Fisher exact test to classify the donor.

the following sections, we will alter each of these three stages, but as a first step, we simply reiterated the process performed by Emerson et al [8]. We used a constant set, but different training sets.

In order to check the precision of the model, we used the area under the ROC curve [2] (AUC). The ROC (Receiver Operating Characteristic) curve has the advantage of not requiring a threshold. In a ROC curve, the x-axis is the False Positive Rate (FPR), and the y-axis is the True Positive Rate (TPR), Both can be calculated as follows:

$$TPR = \frac{TP}{TP + FN} \tag{5.1}$$

$$FPR = \frac{FP}{FP + FN} \tag{5.2}$$

A ROC curve close to the top left corner approaches an optimal classification (When the coordinate of the top left corner point is (0,1) - TPR=1 and FPR=0. In such a case, both FN and FP are equal to 0). The AUC is used as a single value to be compared between different models and configurations.

We calculated AUC for the classification between CMV+ and CMV- donors using the reimplementation of the Emerson model (Figure 5.8). The AUC of the test set of the model above was computed for different sizes of the training set: 100,200,300,400,500 and the full training set of 641 patients (Figure **??**).

Emereson reports an AUC of 0.93 on the cross validation. We obtained a slightly lower AUC of 0.89 (Figure 5.8). While in general, the results are similar, still the re-implementation produces slightly lower results than they report. Moreover, the AUC obtained by us and in the original manuscript are quite low for small training set sizes. We thus aim at improving the AUC for small subsets.

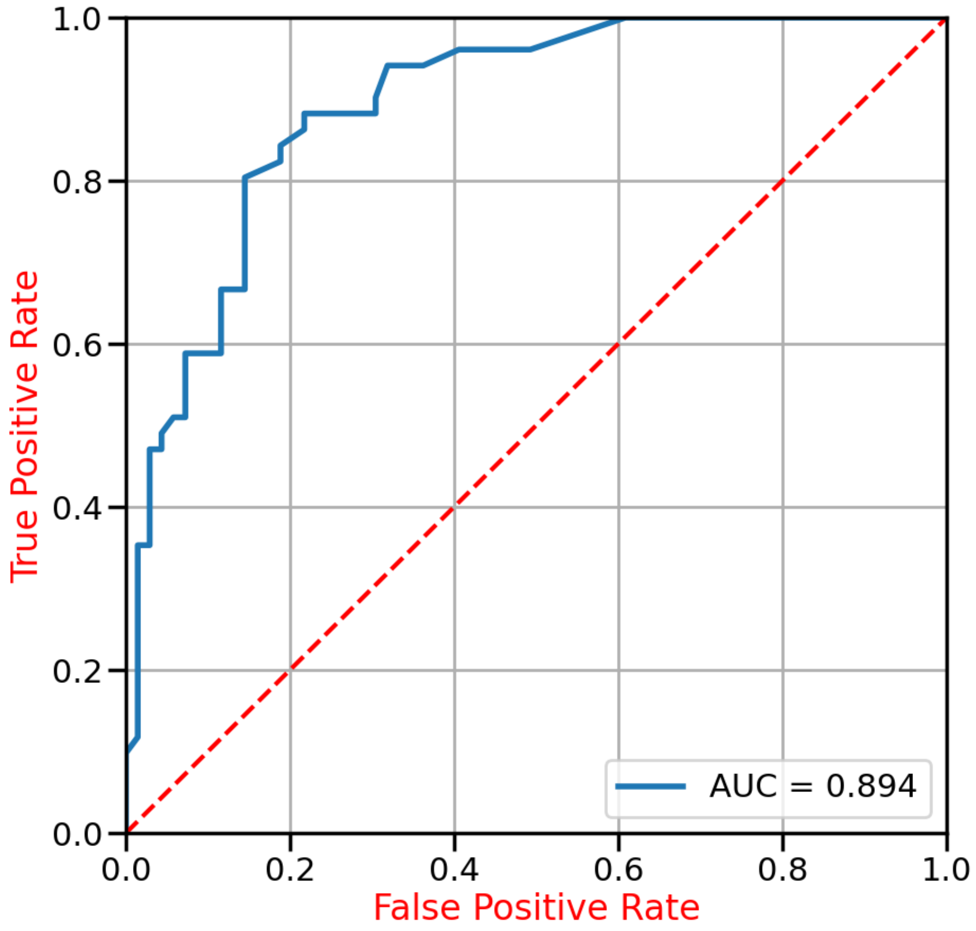


Figure 5.8: AUC evaluation of classification model. True positive fraction (y axis) on test set for full repertoire classification as a function of False positive fraction (x axis). The AUC is 0.894. Thus, the current method is applicable for repertoire classification with a large pool of samples.

Some TCRs are positively associated with CMV, but we do not detect TCR negatively associated with CMV

The first caveat with the approach of Emerson et al is that there are TCR with a negative association with CMV. While TCRs reacting to CMV should be positively associated with it, there is no reason for negative associations.To check the claim that TCRs with a positive association with CMV have a greater significance for the classification, we computed for each TCR the deviation of the expected and observed number of times it appeared in postive samples.

Formally, after filtering the data, we measured the chi-square value of each TCR. To do that we computed the total number of times a TCR appeared in any sample, and the fraction of positive samples, and computed their product as the expected number of times a TCR would appear in positive samples. We then compared the real number of appearances with this value. To check the sign of the deviation, we added the sign of the difference to the chi square score (Figure 5.9).

In order to test the relation between a TCR and the class of the sample, we used a chi square statistical test [19]. To define TCRs that are associated with the group of positive donors, we computed the chi-square value of the number of times this specific TCR appeared in donors compared to the expected value. The chi-square test measures dependence between variables, so TCRs with a low Chi square score can be ignored, since they are most likely to be independent of the class, and therefore irrelevant for the classification. We compute the observed and expected number of appearances of a TCR in the positive samples $C(i, j)$, and compared it to the expected number if a TCR is not associated with either CMV+ or CMV-.

We then filtered TCRs based on their chi square score in the training set. The aim was to examine whether the resulting TCRs are informative of the CMV status. When scattering the chi-square score in the training and test sets. we studied more properties of those TCRs. For each such TCR we computed its average count $\sum_{j \in training} \frac{c(i,j)}{641}$ (Figure 5.14) and average frequency $\sum_{j \in training} \frac{f(i,j)}{641}$ (as defined in Figure 5.14).

We used a 2D plot. The color was either the average frequency of each TCR in the donors where this TCR is present or the parallel average counts. The frequency of a TCR in a sample is the number of cDNA copies of this TCR in the sample, and the count is the number of **different** nucleotide sequence that encode for the same CDR3 amino acid sequence.

One can observe a skewed scattering to the right on both training and test set. specifically, while there are practically no TCRs that are much less frequent than randomly expected in the positive set. There are many TCRs that are significantly over expressed in the positive set. This is expected from the positive selection of TCRs by antigens (Figure 5.9).

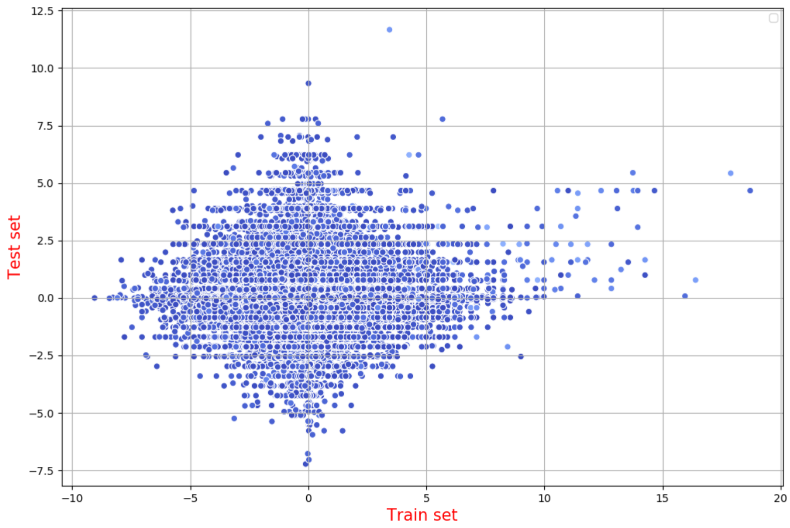


Figure 5.9: A graph comparing the results of a Chi-Square test of the distribution of the receptors on positive and negative patients of the test and train datasets. the outliers on the top left show the receptors that can help predict whether a patient is positive in the test dataset using the train dataset.

Choice of receptors for classification

Plots/table of number of receptors vs cutoff. For each $TCR(i)$ we collect number of impressions an frequency rate.

The Emerson model ([8]) determines positive and negative TCRs using a Fisher exact test, that tests the independence between TCRs and a condition. This test produces 154 TCR's that can be used for the classification. One of our conclusions was that the amount of TCR's that can be used for the classification should be increased.

We thus aim to define the most informative TCRs (further denoted as ”golden TCRs”) to classify donors. We use the score from chi-square test above, but a cutoff value has to be defined. We look for a cutoff that would increase the number of golden TCRs comapred with the work of Emerson.

As we described above, about TCR's selection process based chi-square method. We thus tested the influence of golden TCRs cutoff on the number of golden TCRs. Formally, we count the number of TCrs with a chi square score above each cutoff in the training set (Figures 5.15, 5.16). Note that while the cutoff is on the absolute value of the score. The vast majority of the golden TCRs are expressed more in the positive set than expected. We ended up using a cutoff of 7 to produce 256 golden TCRs.

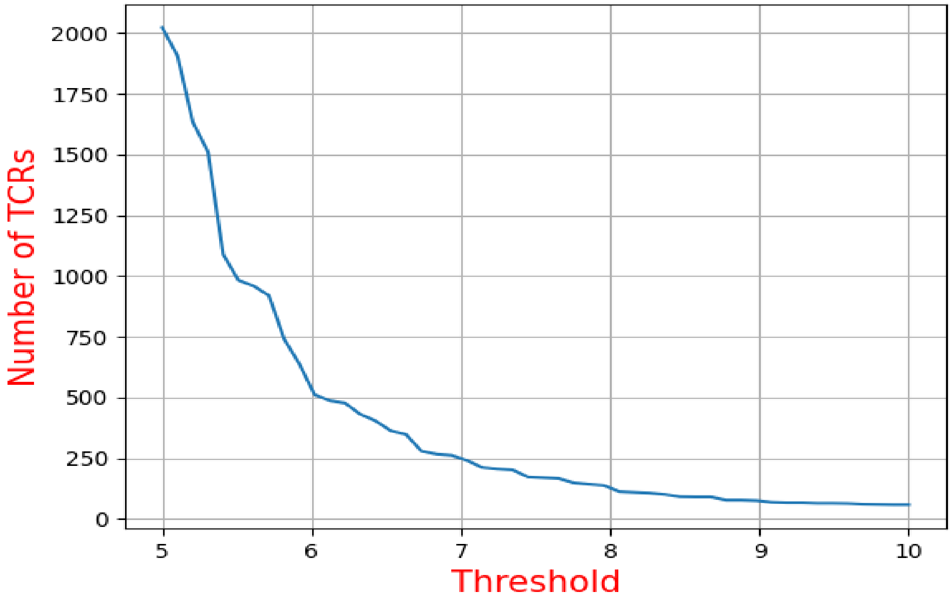


Figure 5.10: Elbow graph determining the number of golden TCRs in the data set. The x axis is the cutoff, and the y axis is the number of TCR with an absolute chi-square score above this cutoff

Classification using "interesting receptors"

Your old ML results on these Receptors RAcheli NN on these receptors Your new results using LM on these receptors In all case plots AUC as function of traning size

The following chart displays the results we got while using the various models:

random forest							
train size		100	200	300	400	500	all
auc score		0.66	0.65	0.69	0.78	0.8	0.86
hyperparams	numrec	100	175	100	75	175	125
	bootstrap	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
	max depth	10	30	70	50	70	50
	max features	auto	sqrt	auto	sqrt	auto	auto
	min leaf	1	2	1	1	1	1
	min split	2	2	10	5	10	5
	estimators	100	100	300	300	100	200
XGBOOST							
train size		100	200	300	400	500	all
auc score		0.65	0.66	0.66	0.7	0.72	0.84
hyperparams	numrec	250	225	100	100	100	100
	sample by tree	0.3	0.3	0.3	0.3	0.3	0.3
	gamma	0	0.4	0.4	0.3	0.3	0.2
	learning rate	0.15	0.2	0.3	0.05	0.15	0.1
	max depth	4	12	8	12	15	15
	min child	1	1	1	1	1	1

NN using count							
train size		100	200	300	400	500	all
auc score		0.68	0.67	0.71	0.79	0.81	0.87
hyperparams	batch size	128	32	128	32	128	64
	first hidden layer	256	64	32	32	128	32
	second hidden layer	64	32	32	4	64	64
	lr	0.01	0.001	0.0001	0.001	0.001	0.001
	number of epochs	8	8	13	6	13	6
	numrec	125	150	150	50	50	125
	trainprec	80	90	90	60	60	70
NN using binary							
train size		100	200	300	400	500	all
auc score		0.72	0.67	0.7	0.8	0.82	0.87
hyperparams	batch size	128	64	32	16	16	16
	first hidden layer	32	128	64	256	128	128
	second hidden layer	32	16	16	32	64	32
	lr	0.001	0.001	0.001	0.0001	0.0001	0.0001
	number of epochs	11	15	15	12	11	7
	numrec	250	125	50	50	50	150
	trainprec	60	80	90	70	60	70
NN using frequency							
train size		100	200	300	400	500	all
auc score		0.73	0.81	0.77	0.81	0.83	0.88
hyperparams	batch size	32	32	32	128	128	32
	first hidden layer	256	32	256	128	256	16
	second hidden layer	16	64	16	4	4	4
	lr	0.001	0.001	0.001	0.001	0.001	0.001
	number of epochs	11	12	11	15	9	14
	numrec	250	175	175	250	100	175
	trainprec	80	70	70	80	60	70

After using the Chi Square test to find ”interesting” receptors, we used them in our machine learning models, which were Random Forest, XGBOOST, and a Neural Network. Firstly, we used the Scikit-learn package for the Random Forest and XGBOOST models and preformed a grid search to find the best parameters to get a maximal AUC score.

The parameters we fit in the random forest model were the number of estimators, the maximum number of features in a single tree. the maximum depth of a tree, the minimum size for a split in the tree, the minimum size for a leaf, and whether bootstrap samples are used. We got an AUC score 0f 0.86 when using the data of all 641 patients.

The parameters we fit in the XGBOOST model were the learning rate, the maximum depth of the tree, the minimum weight for a child node, the minimum loss reduction required to make a further partition on a leaf node of the tree, and the sub-sample ratio of columns when constructing each tree. We got an AUC score 0f 0.84 when using the data of all 641 patients.

The neural network we built is a simple, two layered network, we used PyTorch to build the network and NNI^{TM} for our parameter tuning. the parameters we fit were the learning rate, the batch size, the size of the first and second layers, the number of epochs, and the percentage of the validation out of the training data.

In addition, we used the amount of receptors (the top n receptors by our Chi-Square test score) as a parameter for all of our models. We also used three types of input data; Boolean data (whether a patient had or didn’t have a certain), the exact count, and the accumulated frequency. we used the Boolean data for the first two models, and all three of them in the Neural network.

For all 641 patients, we got an AUC score of 0.87 for the Boolean and count based data, and 0.88 for the frequency based data. We also saw a significant rise in the AUC score of mid-sized samples (300, 400 and 500 patients) when using the Neural network, and especially when using the frequncy-based data.

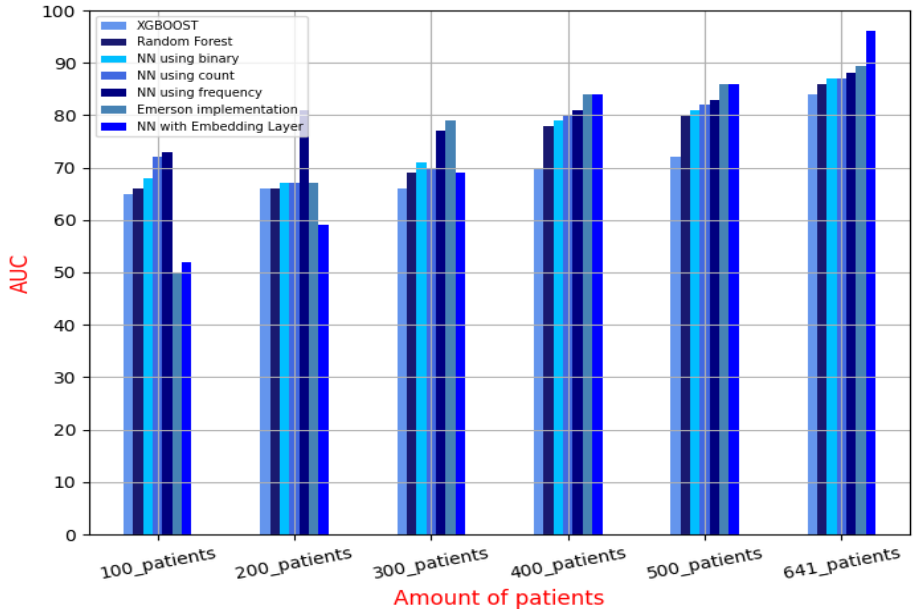


Figure 5.11: Comparison of AUC among the different machine learning models and different amounts of donors. AUC score as function of size of the donors.

Neural network with embedding layer

After finding important sequences by the chi-square metric score. We tested a comprehensive approach to a neural network with an embedding layer [13]. Embedding is a method used to represent discrete variables as continuous vectors. This technique has practical applications in word embeddings for machine translation [20] and entity embeddings for categorical variables [9].

An embedding is a way to represent a categorical feature (in our case a TCR), as a dense parameter. Our main goal was to classify donors using golden TCR's. However, and additional usage of this data may be to obtain a good embedding matrix - a representation of each TCR as a real vector and explore similarities between the TCRs.

The classification process was first to filter golden TCRs TCR_i in each donor D_j. Then, use the golden TCRs and encode them using an embedding layer. For each patient, we obtained a array of sequences representations in R^k, where k is the dimension of the embedding. A sum of all representations is performed to obtain a single vector representing a donor. This vector is the input of a hidden layer with a RELU activation function. The last layer in the neural network has one neuron and a sigmoid activation function for the classification of CMV+ or CMV-.

In order to prevent over-fitting we use droupout in the training process. We use an ADAM optimizer, and a penalty $l2$ regularization. In addition to the learning process we used batch normalization as regularization technique. By adding Batch Normalization we reduced the internal covariate shift and instability in distributions of layer activation.

Merging receptors using DBSCAN

Beyond the method discussed above, we analysed another machine learning approach. This method was based on unsupervised learning with several stages. The first satge is an autoencoder model 4.1 designed to learn a vector form of textual sequence in in our case the CDR3 amoino acid sequence [7].

The following stage is DBSCAN [11] on the resulting projection to find similar clusters (groups of TCRs in our study). An important aspect of DBSCAN is ”local outlier factor (LOF)” that finds anomalies. Anomalies are used to filter out the uninformative TCRa.

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a popular learning method utilized in model building and machine learning algorithms. DBSCAN is a clustering method that is used to separate high density clusters from low density samples. Clustering is basically an unsupervised learning method that divides the data point into groups, such that the data points in the same groups have similar properties and data points in different groups have different properties. Typically, we require points in the same group to be more similar to other points in the same group than to points in different groups.

The local outlier factor (LOF) of DBSCAN is based on the concept of a local density, where locality is determined by the k nearest neighbors. By comparing the local density of an object to the local densities of its neighbors, one can identify regions of similar density, and points that have a substantially lower density than their neighbors. These are considered to be outliers.

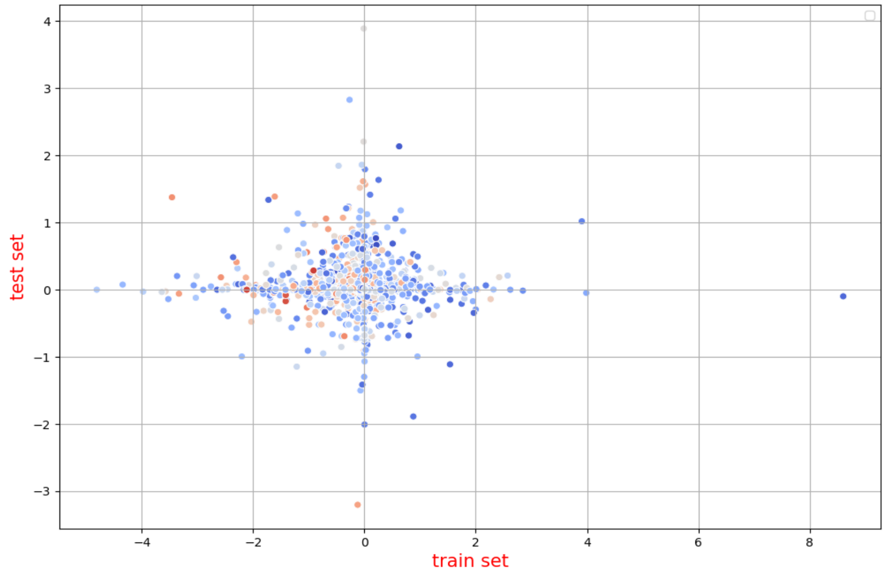


Figure 5.12: Comparison between the chi square score of each cluster in the train and test data sets.

Here we have used clustering to merge similar TCR's to one TCR that can explain group behavior. After we got our embedding from the auto-encoder, we used the auto-encoder projection as coordinates in the space $\mathbb{R}^{30}$ and preformed the DBSCAN clustering algorithm in order to clump together similar receptors. The parameters we found worked best were an epsilon of 0.25 and 5 minimum samples for a core point. Any larger epsilon and we got very few clusters, any smaller and we got to many.

We then added up the count of all the appearances of TCRs in a patient, for all TCRs that belong to a certain cluster. Once we had our clusters we could use our Chi-Square test as if they were receptors, and plotted each cluster as a dot according to it's Chi-Square score in our test and train data set.

As can be seen in the scatter plot, the DBSCAN failed to give us meaningful clusters with a high chi square score in both the train and test data sets. And since we got no ”golden” clusters, we couldn't use them for our ML models.

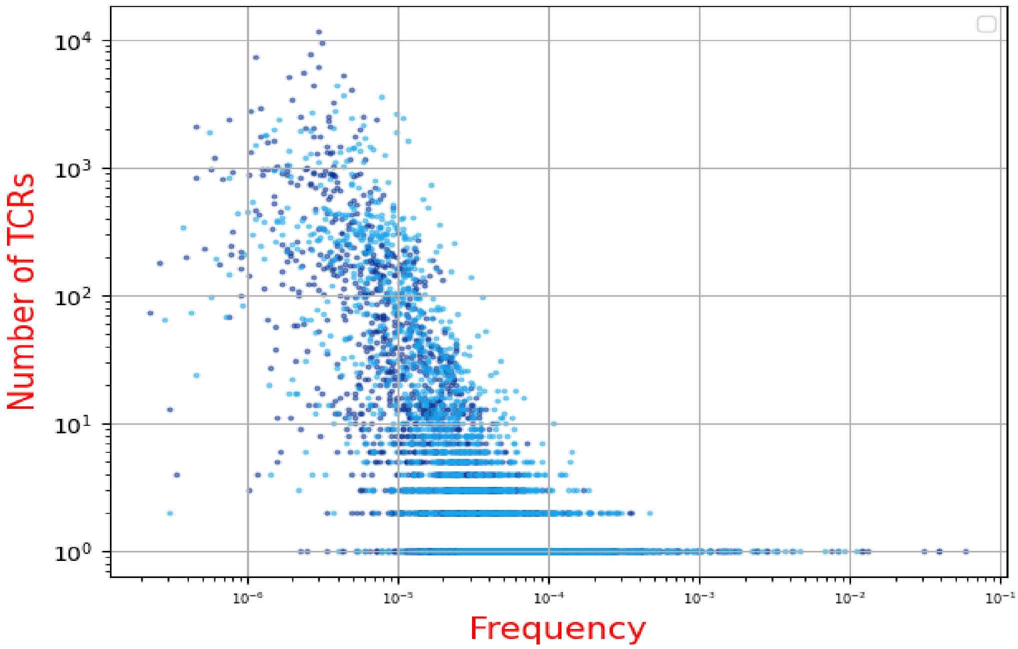


Figure 5.13: Graph of the amount of TCR's per frequency. TCR's per frequency for CMV+ and CMV-.

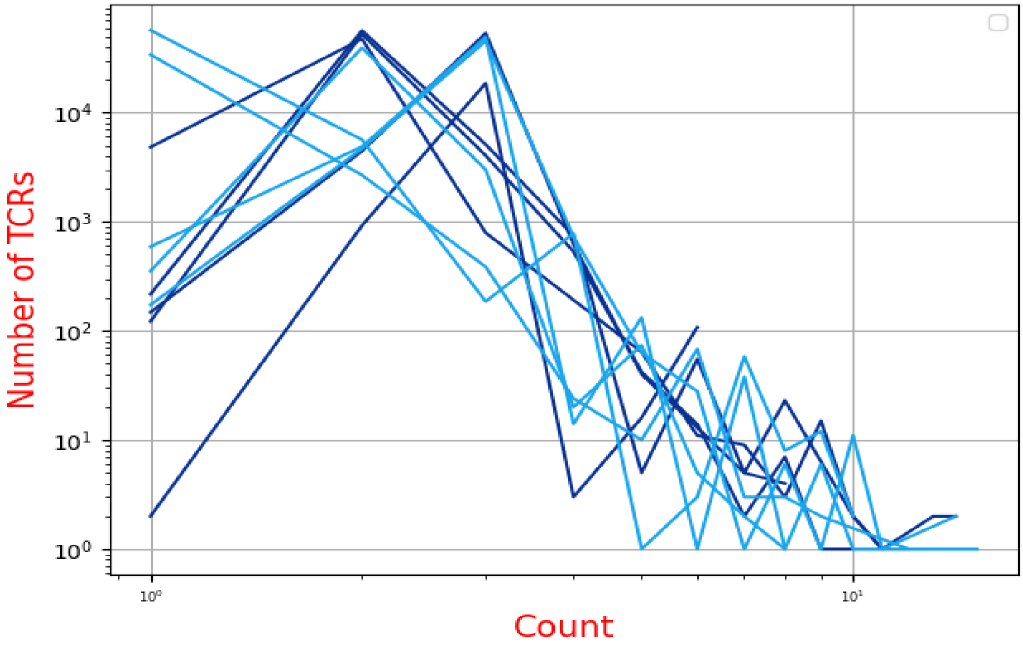


Figure 5.14: Graph of the amount of TCR's per count. TCR's per count for CMV+ and CMV-.

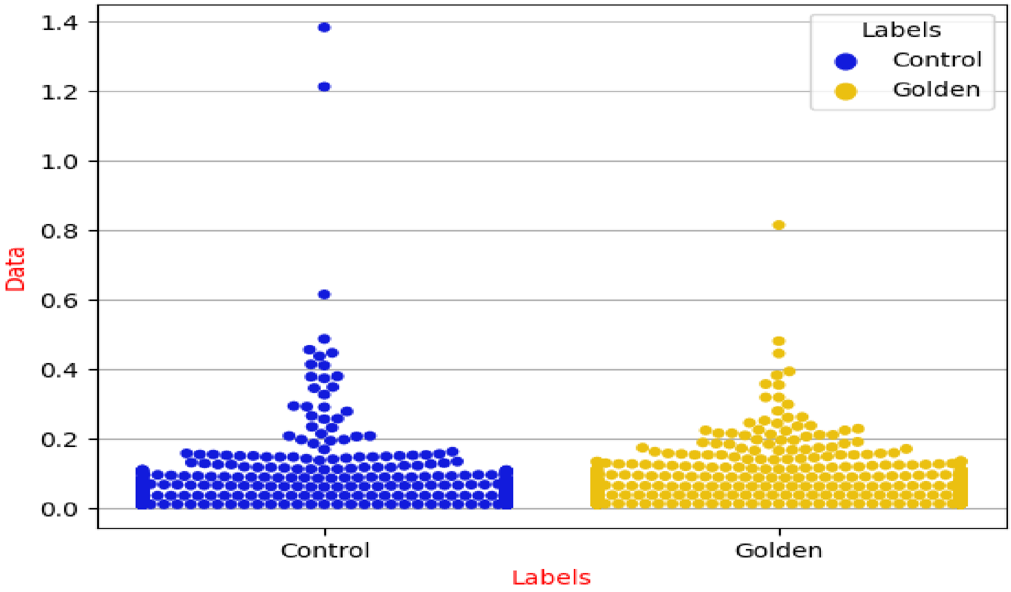


Figure 5.15: Comparison between the average count of golden and control TCRs. Each dot is the average in a sgiven sample of the count of the TCRs in this sample. Blue dots are control and golden dots and golden TCRs. Average count divided over golden and regular TCR's.

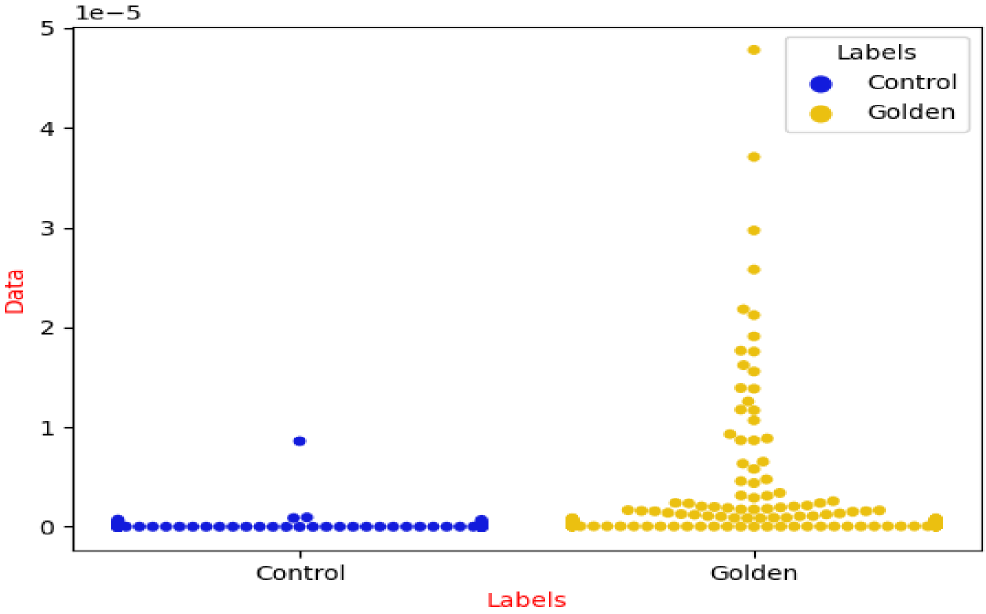


Figure 5.16: Comparison between the average frequency of golden and control TCRs. Each dot is the average in a sgiven sample of the frequency of the TCRs in this sample. Blue dots are control and golden dots and golden TCRs.

6 Discussion

We have here analyzed a dataset of sequenced TCRs from different host, some of which were infected with CMV. Our goal was to test whether we can use the receptor information to improve on the current state of the art TCR sequences based classification of hosts. Currently all existing methods for TCR repertoire based host classification ignore the details of the TCR sequences, and treat the different TCRs of the host as random independent samples. Statistical tests were applied to detect TCRS that are informative of the host condition (CMV+ or CMV- in this case). The work performed here included two main components. First improving the machine learning pipeline to classify hosts, and then an attempt to include information on the TCRS beyond their binding status (e.g. their sequence). We have succeeded in the first and failed in the second task. Indeed, we obtained a better accuracy than the current state of the art when improving the machine learning pipelines. However, all of our attempts to include TCR information did not lead to an improved accuracy.

An important aspect of these results is that in contrast with previous work where TCRs were associated either with the CMV- or CMV+ status, we argued here that T cell clones with TCR reacting to CMV can be expanded,leading to CMV+ associated TCR. However, there is no reason for any TCR to be associated with a CMV- status. Indeed, when performing association tests, we did not find such TCRs. Moreover, the counting based existing approach ignored the possible differential contribution of TCRs to the host classification. Incorporating this differential contribution into more complex machine learning approaches improved the precision accuracy. One can still wonder why did the TCR information failed to improve the accuracy of the classification. We can propose multiple explanation for that.

First, the presentation we used may be problematic, we tried a vast number of representation (beyond the ones discussed in this text), and none of them produced a better classification, suggesting that the representation of TCRs as numeric vectors is not the source of the problems.

Second, TCRs associated with CMV may be unrelated. To check that we check whether CMV+ assciated TCRs are clustered, and we indeed found that they were. Third, perhaps the methods to group TCRs (by DBScan based clustering) is problematic. DBScan cannot reflect multidimensional data properly ([10]).The accuracy of DBSCAN depends on the distance measurement in the regionQuery (P, ϵ) function. The most commonly used distance metric is the Euclidean distance, especially in high-dimensional data. This metric is problematic, since the it converges to similar values for all pairs of points (Following the central limit theorem). Moreover, it is difficult to find a suitable value for ϵ, since it needs to be a very narrow range to allow few neighbors for most samples, but not too many.

Since the DBSCAN algorithm uses parameters that characterize the density in general, when the density of each class is not uniform, finding clusters of one class may be very problematic. Assume foe example that most points belong to class A, and few points belong to class B, but the points in class B are in dense clusters. Since class B is rare, any cluster around a class B point would contain no other class B points (if ϵ is too small), or a lot of class A points (if ϵ is too high), Thus finding clusters of a given type (informative TCRs in this case) when surounded by a very large noise of another class (uninformative TCRs) may be very complex, and lead to empty clusters or clusters with many uninformative TCRS, and as such, the signal of the golden TCRs was diluted. Even worse, the golden TCRs may be removed as outliers. Outliers are often discarded as noise in the DBSCAN algorithm, but in some applications this noisy data can be more interesting than more regularly occurring data (REF). Points marked as outliers are not discarded as such, they are simply points that are not included in any cluster. You can still check a set of non-clustered points and try to interpret them.

This research can have far reaching applications in the development of global biomarkers. Current bionmarkes are typically based on a single gene. However, current data accumulation methods and analysis allow for complex biomarkers representing multiple genes. Emerson et al ([8]) predicted a CMV based Linear model by showing the relation between input (counts of each TCRs in each patient) and output (class of patients) for a subset of TCRs, and showing the total number of ”positive”/”negative” receptors is directly proportional or inversely proportional to the chance of being positive or negative.

Our approach uses Non-Linear models and relate the presence/absence of each informativ TCR with the prsence of CMV using a complex non-linear approach. Non-Linear models are hard to train. Neural networks, (which are non-linear dynamical system) typically have more than one minimum. As such, one cannot use convex optimization techniques. We used the statistical method described before to classify patients based on their TCR repertoire. We applied the algorithm to different numbers of patients in the training set and computed the AUC score of the models as a function of the number of donors in the training set. We saw that the classifier performance decreases below around 300 samples. We conclude that our representation may not be concise enough, and as such lead to overfitting and the resulting drop in accuracy in small sample sizes. Note that different random divisions into training and test gave similar results. We hope that better latent variable model producing efficient representations of the TCRS may help us improve the accuracy.

This classification task can be viewed as a multiple instance learning (MIL) problem ([6]). MIL problems arise in tasks where the training examples are of varying sizes. In MIL problems, a set or bag is labeled instead of lone objects like in standard supervised learning tasks. Formally, a bag $X = \{x_1, x_2, x_3, ..., x_N\}$ receives a label $Y_X = max\{y_i\}$, where y_i is the label of x_i. Here, $y_i \in \{0, 1\}$. However, this can be extended to any label. During training, we are unaware of y_i, we only know the Y_X of each bag. Examples of MIL problems are video classification, where each frame is an instance, text classification, where each word is an instance, 3D object classification, where each point is an instance, and more ([4]). The difference between the current problem and classical MIL is the very low witness rate (fraction of informative samples) in the current analysis, and the problematic representation of the data (A categorical variable for the V gene and a sequence for the CDR2. We now plan to extend the current analysis to attention based MIL methods to try to improve the performance.

Bibliography

[1] Benichou, J., Ben-Hamo, R., Louzoun, Y., and Efroni, S. (2012). Rep-seq: uncovering the immunological repertoire through next-generation sequencing. *Immunology*, 135(3):183–191.

[2] Bradley, A. P. (1997). The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159.

[3] Breden, F. (2020). 486 ireceptor plus: a data integration platform to share, compare and analyze adaptive immune receptor repertoire (airr-seq) data from antibody/b-and t-cell repertoires.

[4] Carbonneau, M.-A., Cheplygina, V., Granger, E., and Gagnon, G. (2018). Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognition*, 77:329–353.

[5] Cohen, C. J., Gartner, J. J., Horovitz-Fried, M., Shamalov, K., Trebska-McGowan, K., Bliskovsky, V. V., Parkhurst, M. R., Ankri, C., Prickett, T. D., Crystal, J. S., et al. (2015). Isolation of neoantigen-specific t cells from tumor and peripheral lymphocytes. *The Journal of clinical investigation*, 125(10):3981–3991.

[6] Dietterich, T. G., Lathrop, R. H., and Lozano-Pérez, T. (1997). Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 89(1-2):31–71.

[7] Dvorkin, S., Levi, R., and Louzoun, Y. (2020). Autoencoder based local t cell repertoire density can be used to classify samples and t cell receptors. *bioRxiv*.

[8] Emerson, R. O., DeWitt, W. S., Vignali, M., Gravley, J., Hu, J. K., Osborne, E. J., Desmarais, C., Klinger, M., Carlson, C. S., Hansen, J. A., et al. (2017). Immunosequencing identifies signatures of cytomegalovirus exposure history and hla-mediated effects on the t cell repertoire. *Nature genetics*, 49(5):659–665.

[9] Guo, C. and Berkhahn, F. (2016). Entity embeddings of categorical variables. *arXiv preprint arXiv:1604.06737*.

[10] Kellner, D., Klappstein, J., and Dietmayer, K. (2012). Grid-based dbscan for clustering extended objects in radar data. In *2012 IEEE Intelligent Vehicles Symposium*, pages 365–370. IEEE.

[11] Khan, K., Rehman, S. U., Aziz, K., Fong, S., and Sarasvady, S. (2014). Dbscan: Past, present and future. In *The fifth international conference on the applications of digital information and web technologies (ICADIWT 2014)*, pages 232–238. IEEE.

[12] Liu, Y., Wang, X., Wang, L., and Liu, D. (2019). A modified leaky relu scheme (mlrs) for topology optimization with multiple materials. *Applied Mathematics and Computation*, 352:188–204.

[13] Lozano-Diez, A., Plchot, O., Matejka, P., and Gonzalez-Rodriguez, J. (2018). Dnn based embeddings for language recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5184–5188. IEEE.

[14] Madi, A., Shifrut, E., Reich-Zeliger, S., Gal, H., Best, K., Ndifon, W., Chain, B., Cohen, I. R., and Friedman, N. (2014). T-cell receptor repertoires share a restricted set of public and abundant cdr3 sequences that are associated with self-related immunity. *Genome research*, 24(10):1603–1612.

[15] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

[16] Page, D. B., Yuan, J., Redmond, D., Wen, Y. H., Durack, J. C., Emerson, R., Solomon, S., Dong, Z., Wong, P., Comstock, C., et al. (2016). Deep sequencing of t-cell receptor dna as a biomarker of clonally expanded tils in breast cancer after immunotherapy. *Cancer immunology research*, 4(10):835–844.

[17] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

[18] Rudolph, M. G., Stanfield, R. L., and Wilson, I. A. (2006). How tcrs bind mhcs, peptides, and coreceptors. *Annu. Rev. Immunol.*, 24:419–466.

[19] Satorra, A. and Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, 66(4):507–514.

[20] Shimanaka, H., Kajiwara, T., and Komachi, M. (2018). Ruse: Regressor using sentence embeddings for automatic machine translation evaluation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 751–758.

[21] Walkowiak, T., Datko, S., and Maciejewski, H. (2018). Bag-of-words, bag-of-topics and word-to-vec based subject classification of text documents in polish-a comparative study. In *International Conference on Dependability and Complex Systems*, pages 526–535. Springer.