DEPARTMENT OF MATHEMATICS
FACULTY OF EXACT SCIENCES

MSc Thesis Proposal

Classification of bagged sample texts using projections and observations selection.

סיווג אוספי רצפים על סמך הטלות ובחירת תצפיות.

Submitted by:
Vova Psevkin
318013356

Supervisor:
Prof. Yoram Louzoun

February 2020

Contents

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 widespread 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 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.

Introduction

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 compiles to 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 preprocessing representation, The Bag of Words 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 (4).

The key idea is to quantify each frequency extracted into a real value, 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 word bag 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.[1]

[1] This part was adapted from "Extraction of T-Cell Receptor repertoire antigen binding properties based on textual analysis of the CDR3 composition"

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). 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 αβ T-cell receptors(5). However, most of the TCR's binding to target MHC-peptide is determined by the β chain (6). 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 (7) diversity 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 neo-antigens, the appropriate T cells expand, resulting in the over-presentation of T cells carrying such receptors(8). Recently, high-throughput DNA sequencing has enabled large-scale characterization of TCR sequences, producing detailed T cell repertoire (Rep-Seq)(9). 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. 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 (10), 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(11), and single peptides binding many 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 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.

Research aims

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

1) Most sequences x_{i,m_i} do not appear in most patients.
2) Most sequences x_{i,m_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 sequences 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 patient.

The difference between B) and C) is that in C), we plan to merge subsets of x_{i,m_i} into new values (that will be further denoted meta clones). Then we plan to detect interesting meta clones, instead of interesting clones.

Following (*10*), we will use a set of T cell clones, each with a different TCR (See biological explanation above) from patients that either had or did not have CMV. Most TCRs are not shared between patients. 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 patients.

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).

Methods

Table of variables

Following is a detailed table of the variables used in the analysis:

Length of CDR3 sequence	l
Maximal length of CDR3 sequence in an analyzed repertoire	m
Dimension of one-hot representation	n
Size of the training set	N
Batch size	b
Distance matrix of all the Euclidean distances in the training set	D_N
Distance matrix of all the Euclidean distances in a batch	D_b
Encoder's input vector	x
Encoder's output vector- embedded vector	z
The sequence form of x	s
The sequence form of combination includes genes	$\bar{S}$
R by R edit distance matrix between output vectors	E

The Data

TCR sequencing files were collected from the Emerson article et.al (*10*). The source includes patient files and for each patient, there are samples with biological and personal information. For each patient, the samples were parsed to compositions of CDR3 Amino Acid sequences, V

and J family genes per observation. Sequences with non-IUPAC letters (X, *, _, #) were ignored. For each data set, we used the preprocessing performed in the original manuscript (*46*).

Data Preprocessing and Feature Engineering

Since TCRs have varying lengths of CDR3 sequences, right zero-padding will be added to all TCRs to generate a constant length representation. A stop signal (!) will be first added at the right end of each sequence of length l. The sequence will then be converted to a one-hot vector. Each character has a corresponding 21 dimension one-hot vector (for 20 possible amino acids (AA) and an additional position for the stop signal). The one-hot vector will be right zero-padded to produce a constant length representation.

For each patient, the observations will be parsed as compositions of CDR3 Amino Acid sequences, V, and J family genes per observation.

When combining V and J usage with the CDR3 sequences, the genes will be represented as n_v dimensional one-hot vectors, where n_v is the number of all possible genes in a given data set (each data set has a different number of different recognized genes based on the algorithm used for the clone analysis and the sequencing length). The CDR3 and genes one-hot vectors are then concatenated. The final representation is the concatenated one-hot vectors in this order: AA1, AA2, …., AAl, stop codon, zero-padding, and (optionally) genes (*46*).

Projections

Autoencoder

The AE (Fig. 1) is the continuation of previous work (*46*). The Encoder named "LocAl TCR Density" has three internal fully connected symmetric encoder and decoder layers of 300, 100, and 30 nodes. Each layer has a dropout rate of 0.1 and an Exponential Linear Unit 'Elu' activation function. Beyond that, there are three additional layers in the decoder.

- A reshape layer (to n/21 x 21) that only changes the output vector to a matrix, where each row represents a position in the CDR3, or the V and J gene (Technically, the genes is represented as different vectors, since it has a different dimension than the AA).
- A fully connected layer with "softmax" (Eq. 2) activation function on each row separately

$$(1)\quad \sigma(z)_i = \frac{e_i^z}{\sum_{j=1}^{K} e_j^z} \quad \text{for } i = \{1, \dots, K\} \text{ and } z = \{z_i \dots . z_K\} \in \mathbb{R}^K$$

- Another reshapes layer (back to n*1 dimensions).

The Autoencoder will be trained with an 'Adam' optimizer with a learning rate of 0.0001 and a Mean Square Error MSE loss" (Eq. 3). The number of epochs for the training varied for different data sets, where N is the number of samples in the training set.

$$(2)\quad loss_{MSE} = \frac{1}{N} \sum_{i=1}^{N} (y_i - \bar{y}_i)^2$$

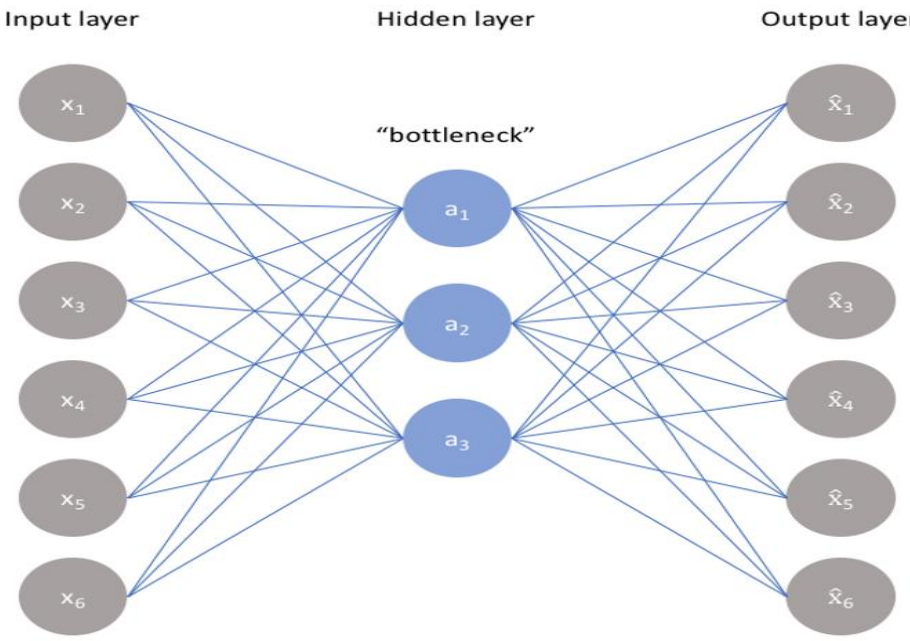

Fig. 1. An example of autoencoder architecture. The first layer (gray) is an input layer. The internal layers (blue) have a smaller dimension than the input layer. The output layer (gray) has the same dimension as the input layer. The goal of the AE is to learn a representation (smallest size layer) for which the input and output layers are most similar.

Training the AE to maintain the distance

To combine the original CDR3s distances in the training process, we first compute the $N * N$ size distance matrix D_N of all Euclidean distances among all pairs of one-hot representation of TCR's from all samples in the training set. Then, batches of b samples are generated from the training set along with their corresponding b size distances matrices D_b produced from D_N. Then, to conserve the distances in the embedded space, the Euclidean distances between pairs of projections and the original Euclidean distance between the original pairs in the input set is added to the loss function $loss_{DIS}$ (Eq. 4).

$$(3)\quad loss_{DIS} = \sum_{i \in b}\left(\sqrt{\sum_{i>j \in b}(z_i - z_j)^2} - D_{ij}\right)^2$$

Where z_i and z_j are the embedded vectors of x_i and x_j besides D_{ij} is the Euclidean distance between vectors x_i and x_j.

Autoencoder accuracy

To test the Autoencoder prediction, we will be using the fraction of sequences properly reconstructed. The reconstructed sequence will be produced by replacing the softmax with a rigid max in layer B above, and a translation of the one-hot back to a sequence, stopping at the left-most stop signal or the end of the CDR3. If the length of the reconstructed vector was different

from the length of the original vector, the entire reconstruction is defined as an error. Otherwise, the error level will be defined as the number of positions differing between the original and the reconstructed sequence. Thus, an error rate of 15% for one mismatch implies that 15% of all sequences either had a wrong length or the proper length and one mismatch.

Projection by TF-IDF

TF-IDF (Term Frequency Inverse Document Frequency) is a statistical measure used to assess the importance of a word in the context of a document that is part of a document collection or corpus. The weight of a word is proportional to the frequency of this word within a document and is inversely proportional to the frequency of this word in all documents of the collection. For instance, words that are common in every document, such as “this”, “what”, and “if”, rank low even though they may appear many times since they are not relevant to that corpus. The TF-IDF measure is often used in text analysis and information retrieval tasks, for example, as one of the criteria for the relevance of a document to a search query, when calculating document proximity during clustering. We will test such projections patient sequences and use them to predict the class of a patient.

Machine Learning Methods

K-Nearest Neighbors

The KNN algorithm assumes that nearby samples have similar classes. This algorithm then simply determines the class of a sample as the majority among its K neighbors (only neighbors in the train set)

To select the correct K for the data, we run the KNN algorithm several times with different values of K. We select the number of neighbors that reduces the number of errors we encounter, without impairing the algorithm's ability to accurately predict new data.

Random Forest

Random forests is a supervised learning algorithm. It can be used for classification. A forest is comprised of decision trees. It is said that the more trees it has, the more robust a forest is. In a random forest, one creates decision trees on randomly selected data samples, obtains a prediction from each tree, and selects the majority vote among trees for each sample.

The model uses two mains concepts:

- Random sampling of training data points when building trees.
- Random subsets of features considered when splitting nodes.

Statistical Analysis

The difference between the various methods in our analysis will be tested using two-way ANOVA, and Tukey's test *(43)*. An ANOVA test will be performed with all autoencoder models. Then, Tukey's test will be performed between all methods and category pairs to find which ones are significantly different. The clustering of samples was done using Hierarchical cluster analysis (HCA) using average linkage on the KL, ED, and KDE distances matrices.

Visualization with TSNE

t-distributed Stochastic Neighbor Embedding (t-SNE) is an unsupervised non-linear dimensionality reduction technique primarily used for data exploration and visualizing high-dimensional data (*12*). In simpler terms, t-SNE gives an intuition of how the data is arranged in a high-dimensional space *(45)*. Specifically, it models each high-dimensional object by a two- or three-dimensional point in such a way that similar objects are modeled by nearby points and dissimilar objects are modeled by distant points with high probability. We here use t-SNE to visualize the embedded vectors of 30 dimensions in two dimensions.

Preliminary Results

We used the statistical method described in (*10*) 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 patients in the training set. After using an algorithm on a smaller pool of sequences, we saw that the classifier performance decreases below around 300 samples *(Fig. 3)*. We conclude that when there are many sequences, statistical approaches can work *(Fig. 2)* but when the pool of patients is smaller, algorithms must use latent parameters arising from the sequence itself.

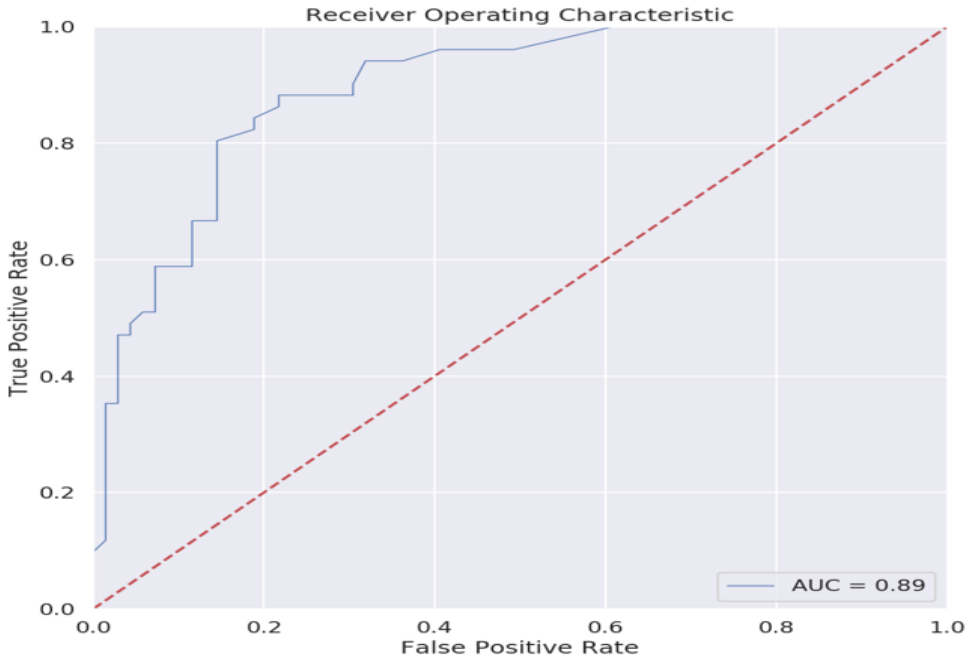

Fig. 2. AUC evaluation of classification model. True positive fraction (y axis) on test set for full repertoire classification as a function of False positive fraction (x axis) . The AUC is 0.89. Thus, the current method is applicable for repertoire classification with a large pool of samples.

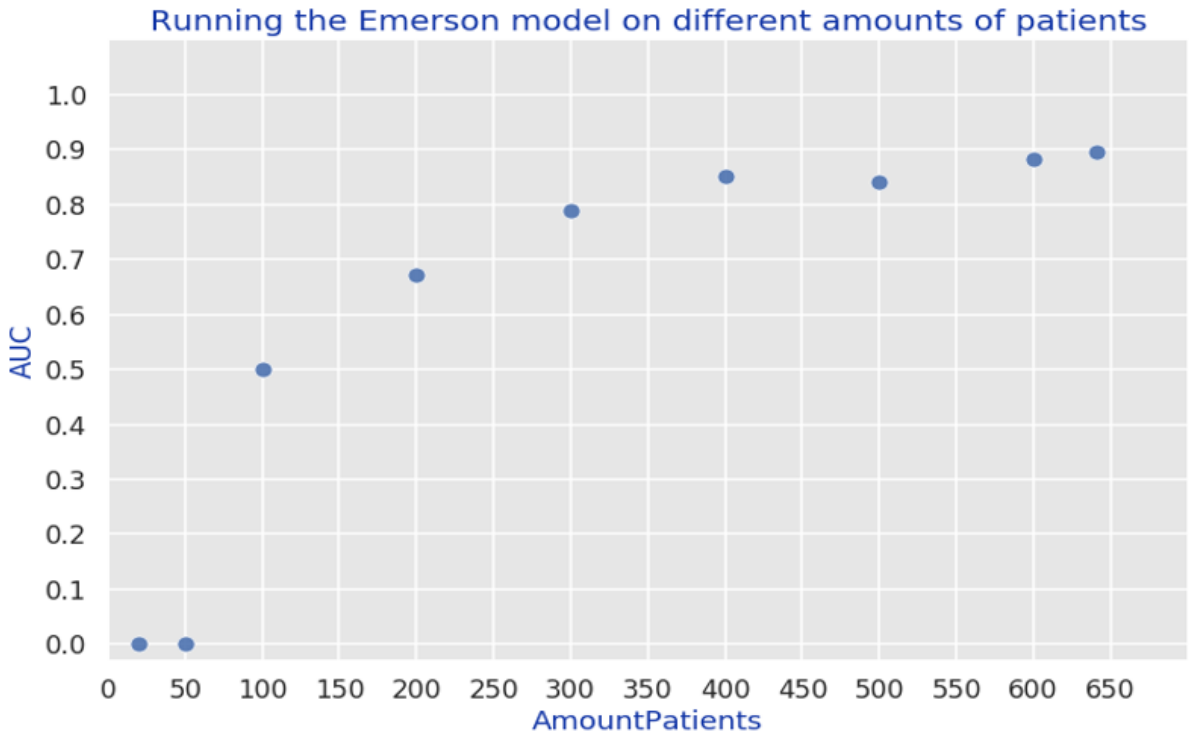

Fig. 3. Effect of train size on test auc score, AUC (y axis) on test set for full repertoire classification as a function of the number of samples used for the training (x axis). One can clearly see that below 300 samples the AUC starts decreasing significantly. Thus, the current method is not applicable for repertoire classification with a small pool of samples.

Biological Glossary

Foreign invaders that have come from outside the body - for example, bacteria, viruses, fungi, parasites. Not only is it an invader that infiltrates the body but also cells that have lost their control ability.	Pathogen
It is a molecule that induces an immune response that produces antibodies and is also used to mark various cells in the animal and immune systems.	Antigen
The part of an antigen that is recognized by the immune system, a short chain of amino acids.	Epitope peptide
It is a subsystem of the overall immune system that is composed of highly specialized, systemic cells and processes that eliminate pathogens or prevent their growth. Part of the immune system that is specific to an antigen. The adaptive immune response combines the cellular response (killing infected/invader cells, executed by T-cells) and the humoral response (creating antibodies, executed by B-cells).	Adaptive immune response
T-cells are responsible for the adaptive immune system cellular response. Each T-cell has a specific receptor on the cell surface. The T-cell receptor should bind an epitope peptide presented on the MHC. A successful binding will activate the attack of the recognized antigen. Most TCRs are combined with α and β chains of amino acids. TCRs are nearly randomly generated in a process called VDJ recombination. Thus, the TCR repertoire is very diverse. Moreover, not all TCRs bind peptides. TCRs that do not bind any peptide are called naïve.	TCR (T-Cell Receptor)
The collection of all TCRs of a patient. The TCR repertoire is unique to every person, with high diversity. The TCR repertoire encodes the memory of the adaptive immune system (by containing all previously activated TCRs).	TCR Repertoire
A molecule that presents the epitope peptide to the TCR. Human MHC is called human leukocyte antigen (HLA). The HLA genes are highly polymorphic.	MHC (major histocompatibility complex)
A region of the TCRβ chain that determines the TCR binding specificity.	CDR3

References

1. J. Rossjohn *et al.*, T Cell Antigen Receptor Recognition of Antigen-Presenting Molecules. *Annu. Rev. Immunol.* **33**, 169–200 (2015).
2. M. Shugay *et al.*, VDJdb: a curated database of T-cell receptor sequences with known antigen specificity. *Nucleic Acids Res.* **46**, D419–D427 (2018).
3. P. Dash *et al.*, Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature*. **547**, 89–93 (2017).
4. T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient Estimation of Word Representations in Vector Space (2013) (available at http://arxiv.org/abs/1301.3781).
5. M. G. Rudolph, R. L. Stanfield, I. A. Wilson, HOW TCRS BIND MHCS, PEPTIDES, AND CORECEPTORS. *Annu. Rev. Immunol.* **24**, 419–466 (2006).
6. C. J. Cohen *et al.*, Isolation of neoantigen-specific T cells from tumor and peripheral lymphocytes. *J. Clin. Invest.* **125**, 3981–3991 (2015).
7. A. Madi *et al.*, T-cell receptor repertoires share a restricted set of public and abundant CDR3 sequences that are associated with self-related immunity. *Genome Res.* **24**, 1603–12 (2014).
8. D. B. Page *et al.*, Deep Sequencing of T-cell Receptor DNA as a Biomarker of Clonally Expanded TILs in Breast Cancer after Immunotherapy. *Cancer Immunol. Res.* **4**, 835–844 (2016).
9. J. Benichou, R. Ben-Hamo, Y. Louzoun, S. Efroni, Rep-Seq: uncovering the immunological repertoire through next-generation sequencing. *Immunology*. **135**, 183–91 (2012).
10. R. O. Emerson *et al.*, Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. *Nat. Genet.* **49**, 659–665 (2017).
11. V. I. Jurtz *et al.*, NetTCR: sequence-based prediction of TCR binding to peptide-MHC complexes using convolutional neural networks. *bioRxiv*, 433706 (2018).
12. H. Zhang *et al.*, Investigation of Antigen-Specific T-Cell Receptor Clusters in Human Cancers. *Clin. Cancer Res.* **26**, 1359–1371 (2020).
13. M. M. Davis, P. J. Bjorkman, T-cell antigen receptor genes, and T-cell recognition. *Nature*. **334**, 395–402 (1988).
14. M. Krogsgaard, M. M. Davis, How T cells "see" antigen. *Nat. Immunol.* **6**, 239–245 (2005).
15. L. Rowen, B. F. Koop, L. Hood, J. Even, J. Kanellopoulos, P. Kourilsky, The Complete 685-Kilobase DNA Sequence of the Human beta T Cell Receptor Locus. *Science (80-.).* **272**, 1755–1762 (1999).
16. J. Glanville, H. Huang, A. Nau, O. Hatton, L. E. Wagar, F. Rubelt, X. Ji, A. Han, S. M. Krams, C. Pettus, N. Haas, C. S. Lindestam Arlehamn, A. Sette, S. D. Boyd, T. J.

Scriba, O. M. Martinez, M. M. Davis, Identifying specificity groups in the T cell receptor repertoire. *Nat. Publ. Gr.* **547** (2017), DOI: 10.1038/nature22976.

17. M. G. Rudolph, R. L. Stanfield, I. A. Wilson, HOW TCRS BIND MHCS, PEPTIDES, AND CORECEPTORS. *Annu. Rev. Immunol.* **24**, 419–466 (2006).
18. S.-Q. Zhang, P. Parker, K.-Y. Ma, C. He, Q. Shi, Z. Cui, C. M. Williams, B. S. Wendel, A. I. Meriwether, M. A. Salazar, N. Jiang, Direct measurement of T cell receptor affinity and sequence from naïve antiviral T cells. *Sci. Transl. Med.* **8**, 341ra77 (2016).
19. D. Schrama, C. Ritter, J. C. Becker, T cell receptor repertoire usage in cancer as a surrogate marker for immune responses. *Semin. Immunopathol.* **39**, 255–268 (2017).
20. D. B. Page *et al.*, Deep Sequencing of T-cell Receptor DNA as a Biomarker of Clonally Expanded TILs in Breast Cancer after Immunotherapy. *Cancer Immunol. Res.* **4**, 835–844 (2016).
21. A. Madi, E. Shifrut, S. Reich-Zeliger, H. Gal, K. Best, W. Ndifon, B. Chain, I. R. Cohen, N. Friedman, T-cell receptor repertoires share a restricted set of public and abundant CDR3 sequences that are associated with self-related immunity. *Genome Res.* **24**, 1603–12 (2014).
22. L. Wooldridge, J. Ekeruche-Makinde, H. A. van den Berg, A. Skowera, J. J. Miles, M. P. Tan, G. Dolton, M. Clement, S. Llewellyn-Lacey, D. A. Price, M. Peakman, A. K. Sewell, A Single Autoimmune T Cell Receptor Recognizes More Than a Million Different Peptides. *J. Biol. Chem.* **287**, 1168–1177 (2012).
23. A. K. Sewell, Why must T cells be cross-reactive? *Nat. Rev. Immunol.* **12**, 669–677 (2012).
24. E. Jokinen, M. Heinonen, J. Huuhtanen, S. Mustjoki, H. Lähdesmäki, TCRGP: Determining epitope specificity of T cell receptors, DOI: 10.1101/542332.
25. V. I. Jurtz, L. E. Jessen, A. K. Bentzen, M. C. Jespersen, S. MAHAJAN, R. Vita, K. K. Jensen, P. Marcatili, S. R. Hadrup, B. Peters, M. Nielsen, NetTCR: sequence-based prediction of TCR binding to peptide-MHC complexes using convolutional neural networks. *bioRxiv*, 433706 (2018).
26. S. Hochreiter, J. Schmidhuber, Long Short-Term Memory. *Neural Comput.* **9**, 1735–1780 (1997).
27. I. Springer, H. Besser, N. Tickotsky-Moskovitz, S. Dvorkin, Y. Louzoun, Prediction of specific TCR-peptide binding from large dictionaries of TCR-peptide pairs. *bioRxiv*, 650861 (2019).
28. C. G. Kanakry, D. G. Coffey, A. M. H. Towlerton, A. Vulic, B. E. Storer, J. Chou, C. C. S. Yeung, C. D. Gocke, H. S. Robins, P. V. O'Donnell, L. Luznik, E. H. Warren, Origin and evolution of the T cell repertoire after posttransplantation cyclophosphamide. *JCI Insight.* **1** (2016), DOI: 10.1172/JCI.INSIGHT.86252.
29. D. H. Younger, Recognition and parsing of context-free languages in time n3. *Inf. Control.* **10**, 189–208 (1967).
30. Y. Louzoun, T. Vider, M. Weigert, T-cell epitope repertoire as predicted from human and viral genomes. *Mol. Immunol.* **43**, 559–569 (2006).

31. A. Kidera, Y. Konishi, M. Oka, T. Ooi, H. A. Scheraga, "Statistical Analysis of the Physical Properties of the 20 Naturally Occurring Amino Acids" (1985), (available at https://link.springer.com/content/pdf/10.1007%2FBF01025492.pdf).
32. N. Tickotsky, T. Sagiv, J. Prilusky, E. Shifrut, N. Friedman, McPAS-TCR: a manually curated catalog of pathology-associated T cell receptor sequences, doi:10.1093/bioinformatics/btx286.
33. M. Zheng, J. Luo, Z. Chen, Development of universal influenza vaccines based on influenza virus M and NP genes. *Infection*. **42**, 251–62 (2014).
34. R. D. Antrobus, T. K. Berthoud, C. E. Mullarkey, K. Hoschler, L. Coughlan, M. Zambon, A. V. Hill, S. C. Gilbert, Coadministration of Seasonal Influenza Vaccine and MVA-NP+M1 Simultaneously Achieves Potent Humoral and Cell-Mediated Responses. *Mol. Ther.* **22**, 233–238 (2014).
35. J. Jia, J. Cui, X. Liu, J. Han, S. Yang, Y. Wei, Y. Chen, Genome-scale search of tumor-specific antigens by collective analysis of mutations, expressions, and T-cell recognition. *Mol. Immunol.* **46**, 1824–9 (2009).
36. C. M. Laumont *et al.*, Noncoding regions are the main source of targetable tumor-specific antigens. *Sci. Transl. Med.* **10**, eaau5516 (2018).
37. S.-Q. Zhang, K.-Y. Ma, A. A. Schonnesen, M. Zhang, C. He, E. Sun, C. M. Williams, W. Jia, N. Jiang, High-throughput determination of the antigen specificities of T cell receptors in single cells. *Nat. Biotechnol.* **36**, 1156–1159 (2018).
38. Autoencoder/cancer data IEDB at master · shiritdvir/Autoencoder · GitHub. Available at: https://github.com/shiritdvir/Autoencoder/tree/master/cancer data IEDB. (Accessed: 26th December 2019)
39. Kullback, S. & Leibler, R. A. On Information and Sufficiency. *Ann. Math. Stat.* **22**, 79–86 (1951).
40. Navarro, G. A guided tour to approximate string matching. *ACM Comput. Surv.* **33**, 31–88 (2001).
41. Silverman, B. W. *Density estimation for statistics and data analysis*. (Chapman and Hall, 1986).
42. Sawyer, S. F. Analysis of Variance: The Fundamental Concepts. *J. Man. Manip. Ther.* **17**, 27E-38E (2009).
43. One-Way ANOVA: Independent Samples: II. Available at: https://web.archive.org/web/20081017161620/http://faculty.vassar.edu/lowry/ch14pt2.html. (Accessed: 24th September 2019)
44. Johnson, S. C. Hierarchical clustering schemes. *Psychometrika* **32**, 241–254 (1967).
45. Visualizing Data using t-SNE | BibSonomy. Available at: https://www.bibsonomy.org/bibtex/28b9aebb404ad4a4c6a436ea413550b30/lopusz_kdd. (Accessed: 24th September 2019)
46. Shirit. Dvorkin, Yoram. Louzoun, Reut Levi. Autoencoder based local Tcell repertoire density ca be used to classify samples and T cell receptors. *bioRxiv*, 148502 (2020)