Variational Autoencoder Clusters T Cell Receptor Amino Acid Sequences

Thomas Mazumder

August 2021

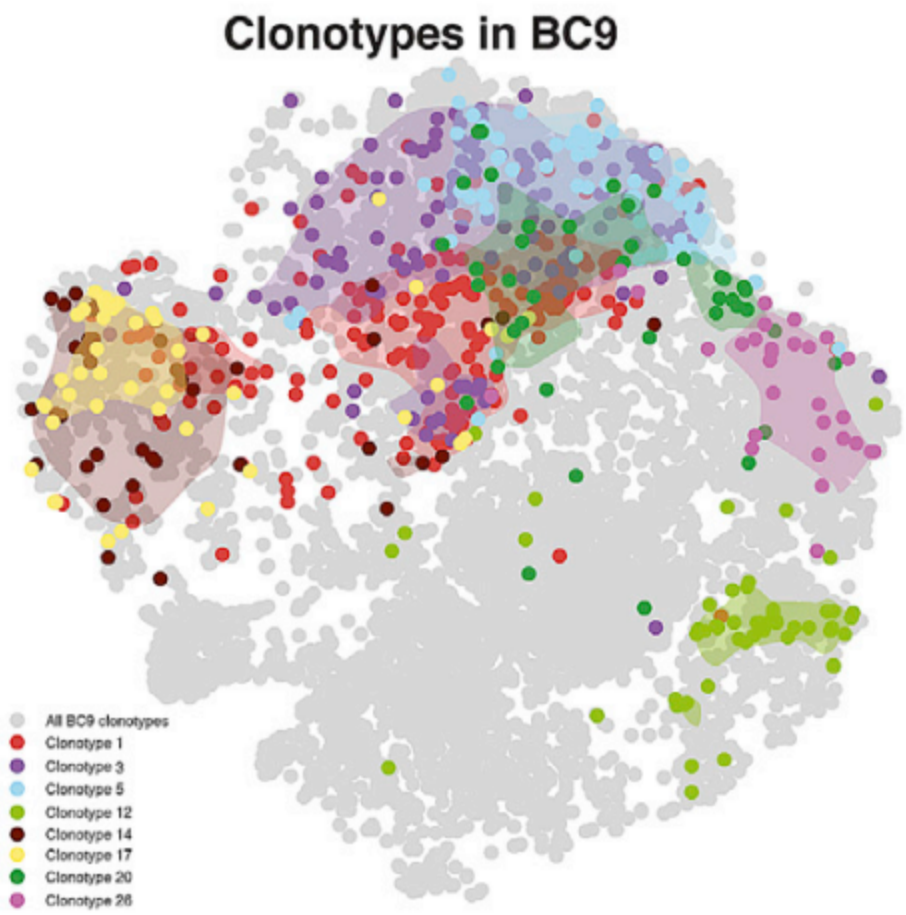
Background

T cell receptor sequencing is a technology that can now provide the sequences of dozens to hundreds of thousands of T cell receptors appearing in a biological sample. Tracking the identities of T cell clones and their multiplicities has been useful in characterizing the immune response against a tumor and predicting viral immune histories of patients [1, 2]. Additionally, high throughput TCR sequencing has been used to show that closely related TCR sequences can have similar binding specificities *in vitro* [3, 4]. However, there have been none or very few studies leveraging the relatedness of TCR sequences in biological analyses.

In 2020, Friedensohn et al. published a study that used a variational autoencoder with a Gaussian Mixture Model on the latent variable space to cluster B Cell receptor sequences [5]. They showed that BCRs from clusters that were enriched for BCRs deriving from mice receiving a certain peptide immunization did have the same binding specificities *in vitro*. The number of BCRs falling into clusters enriched for a certain peptide immunization could be used to predict the immunization a mouse received. Additionally, unseen sequences generated *in silico* from these clusters also demonstrated the predicted binding specificity at a high rate.

This study explores the use of a VAE with GMM to cluster TCR sequences. First, clustering with the VAE with GMM is evaluated on repertoires from healthy patients and those that received a yellow fever vaccine. Next, the VAE/GMM is used to identify clusters of TCRs that are enriched for TCRs from either CD4+ or CD8+ T cells in lung tumors.

Finally, the VAE/GMM is used to investigate whether TCRs with similar sequences tend to have the same fine-grained T cell phenotype. Azizi et al. showed that there is a one-to-one mapping of TCR clone to fine-grained T cell phenotype in breast tumors; for example, all T cells with a certain TCR sequence belonged to CD8+ effector memory T cells, while all T cells with a different TCR sequence belonged to CD4+ memory T cells. In their figure below, each color dot is an instance of a TCR clone in an RNAseq space where different regions represent different fine-grained T cell phenotypes [6].



Methods

Encoding

For each TCR, the aligned CDR1 and CDR2 regions were extracted from the V gene according to the IMGT reference library [7]. The CDR1 alignment has 12 positions and the CDR2 alignment has 10 positions. The CDR3 sequence was padded with gaps in the middle of the sequence so that its length was 24. The first and last amino acids of every CDR3 region as it is conventionally reported, however, are C and F, respectively, so these amino acids were omitted from the encoding, making the CDR3 alignment effectively 22 positions. Then, each position of the CDR1, CDR2, and CDR3 regions were one-hot encoded to a vector of length 21. The vectors for each position were concatenated so that each TCR is encoded in a vector of length 924.

Model

The model used in this study is a variational autoencoder with a Gaussian mixture model on the latent variable representations. This model is taken from Jiang et al. [8]. According to this model, each example x (a one hot-encoded TCR) is generated by:

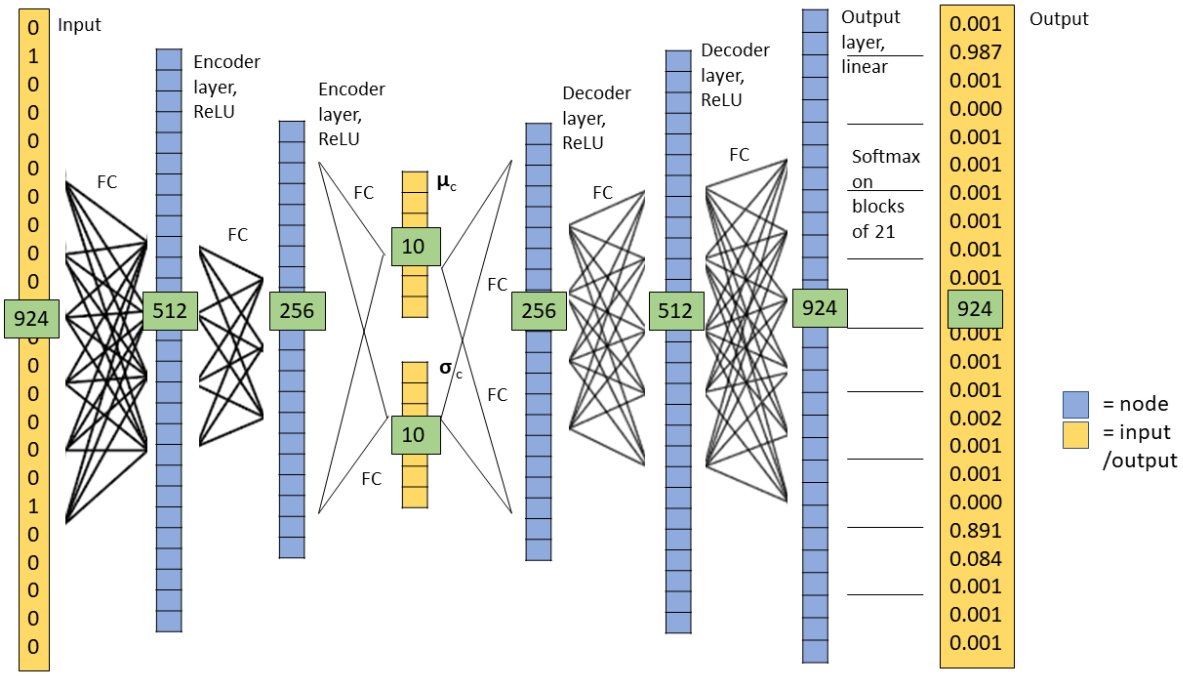
selecting a cluster c according to Cat(c|pi)

generating a latent variable $z \sim N(\boldsymbol{\mu}c, \boldsymbol{\sigma}c\mathbf{I})$

calculating x = g(z) where g is the function learned by the decoder neural network

The derivation of the loss function can be found in [8]. Importantly, the loss function can be understood as a sum of a reconstruction term, which penalizes low reconstruction accuracy, and a KL divergence term, which ensures the latent embeddings lie on the mixture of Gaussians determined parametrized by the u and o. During training, both the encoder function f(x) and the decoder function g(z), as well as the $\boldsymbol{\mu}$c and $\boldsymbol{\sigma}$c are optimized.

The architecture of this model is provided in Figure 2. The encoder has two hidden layers with ReLU activation functions, of dimensions 512 and 256. The latent variables are parametrized by a mean vector of dimension 10 and a variance vector of dimension 10. The decoder has two hidden layers with ReLU activation functions, of dimensions 256 and 512, followed by a layer of dimension 924 with linear activation. A softmax function is then applied to blocks of 924 nodes to produce the output.



Validation

I. Reconstruction accuracy

To assess reconstruction accuracy, a repertoire of 100,000 TCRs from [] was encoded with 5,000 centroids for 40 epochs. Then, the latent variable encodings were sampled 1,000 times and run through the decoder to predict an amino acid sequence for each TCR 1,000 times.

II. Yellow Fever Vaccine Analysis

To investigate whether sequence convergence identified by an autoencoder is diagnostic of an immune phenotype, a model was trained on 126,000 TCRs from the peripheral blood

repertoire of 18 individuals, 9 of whom had received the yellow fever vaccine two weeks prior to TCR sequencing and 9 of whom had not. Sequencing the TCR repertoire of a peripheral blood sample typically gives over 100,000 TCR sequences per sample, which would make identifying YFV-associated TCRs very difficult. Therefore, the 7,000 most clonally-expanded TCRs were selected from each individual. The 9 patients who did not receive the YFV were taken from [2], while the 9 patients who did were taken from a study by DeWitt et al [11].

The model was trained with 5,000 centroids for 40 epochs. Fisher's Exact Test was used to identify clusters of TCRs that were enriched for TCRs sequenced in individuals who received the yellow fever vaccine. The number of TCRs falling in a YFV-enriched cluster was counted for each of 16 training training repertoires and 2 held-out test repertoires in 9-fold cross validation.

Inquiries

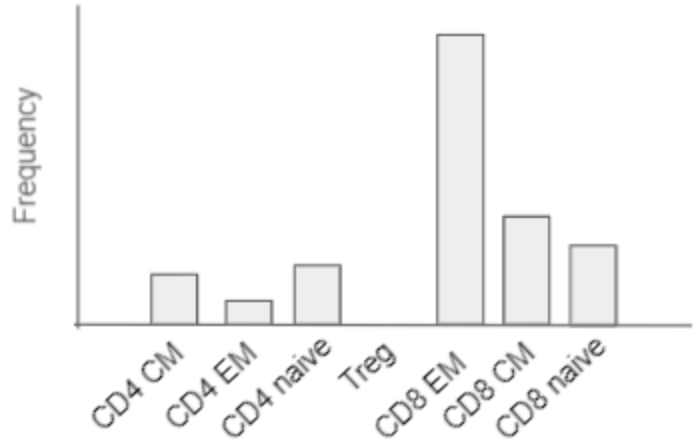
I. CD4/CD8

Next, the ability of an autoencoder to identify sequence convergence in the TCR sequences of CD4+ and CD8+ tumor-infiltrating T cells was interrogated. In Zhang et al [10], four biopsies were taken from lung adenocarcinomas of seven patients. Two were taken from the central region of the tumor and two were taken from the margins of the tumor. The T cells were separated on CD4/CD8 expression with flow cytometry and then subjected to bulk TCR sequencing. For this study, the TCR sequences from the central regions were grouped for each patient. The marginal TCR sequences were not used. The MIXCR program was used to align DNA sequencing reads to V and J genes and to translate the CDR3 sequence of each read. Reads containing a stop codon or reads with a number of nucleotides that did not form an integer number of codons were discarded.

For each patient separately, a model with 1,000 centroids was trained for 40 epochs and clusters enriched with CD4+ or CD8+ TCR sequences were identified with Fisher's Exact Test. For each patient, 90% of TCRs were used for training and 10% of TCRs were held out for testing.

II. Fine-grained phenotypes

The dataset from [6] contains scRNA sequencing profiles and TCR sequences extracted from mRNA. The program ImmClassifier [9] was used to classify cells into a fine-grained T cell phenotype. Then, a VAE model was trained on 9,580 TCR sequences from one patient with 500 centroids for 40 epochs. In order to avoid the multiple testing problem of testing each cluster for enrichment of every fine-grained phenotype via Fisher's Exact Test, the entropy of the distribution of fine-grained phenotypes in each cluster was compared via simulation to the entropy expected by chance. A sample distribution of fine-grained phenotypes in a cluster is illustrated below.



Specifically, the entropy of the distribution of fine-grained phenotypes in each cluster was calculated. In each simulation, given the exact cluster sizes outputted by the VAE model and the exact number of cells assigned to each fine-grained phenotype by ImmClassifier, cells were randomly assigned to clusters. The entropy of each simulated cluster was calculated. Then, entropy of the true cluster with the lowest entropy was compared to the lowest entropy that appeared in each of 1,000 simulations. If the true cluster's entropy was lower than the lowest in 95% of simulations, this cluster was considered enriched for a phenotype. Then, the entropy of the true cluster with the second lowest entropy was compared to clusters with the second lowest entropy in simulations, and so on.

Results

Validation

I. Reconstruction accuracy

The accuracy of the "reconstructed" sequences was 92.3% on average when including CDR1, CDR2, and CDR3 regions. However, many aligned CDR1 and CDR2 regions contain gaps in the same positions, essentially artificially boosting the reconstruction accuracy figure in these analyses. When considering only non-gap positions, the reconstruction accuracy falls to 84.4%.

II. Yellow Fever

Out of 5,000 centroids with which this model was trained, 32 clusters were enriched with TCRs from individuals receiving the yellow fever vaccine when no patients were held out. On average, these enriched clusters contained 45.5 TCRs per cluster. When 9-fold cross validation was conducted, a simple linear threshold for "number of TCRs falling in an enriched cluster" was used to classify samples as either cases or controls. This simple scheme achieved an accuracy of 13/18 (72.2%). The p value for this result under a binomial distribution is 0.048. For

comparison, Reddy et al. achieved 80% accuracy in classifying the BCR repertoires of mice into one of four antigen exposures.

Inquiries

I. CD4/CD8

The results for a representative patient are reported. Samples in the central regions of the second patient's lung tumor contained 41,000 unique TCRs which were clustered into 1,000 centroids. Of these clusters, 8 were enriched for CD4 by FET with Bonferroni correction and 15 were enriched for CD8 by FET with Bonferroni correction. The average size of a cluster enriched for CD4 TCRs was 32, while the average size of a cluster enriched for CD4s was 41. of 4,1000 TCRs in the test set, 71 fell into clusters that were called enriched based on the training set. Sixty three of these TCRs fell into clusters whose enrichment matched the marker of the test TCR. The p value of classifying 63/71 test TCRs correctly is < 0.001.

II. Fine-grained phenotypes

Out of 1,000 clusters, the 12 most uneven clusters had entropies that were lower than the entropies of clusters of the same rank in 95% of simulations. The fine-grained phenotypes enriched in these clusters ran across the range of fine-grained phenotypes called by ImmClassifier.

Discussion

The validation studies in this report suggest that the VAE performs reasonably well in reconstructing TCRs and in identifying TCRs that are diagnostic of an immune phenotype (receiving the yellow fever vaccine two weeks before repertoire sampling). While Reddy et al. have demonstrated that VAEs perform strikingly well on BCR clustering for immune phenotype prediction, the extension of this capability to TCRs is non-trivial given that many closely related BCRs may appear in a sample due to somatic hypermutation, while TCRs with related sequences only arise from distinct VDJ recombination events.

The CD4/CD8 line of study demonstrates that the VAE can classify a small number of test TCRs in individual patients as CD4 or CD8 with accuracy near 90%. While this accuracy is high, I emphasize that this predictive model cannot make predictions for most TCRs in the test set. In Patient 2, the model made predictions for 71 TCRs out of 4,100 in the test set.

This study also suggests that some related TCR sequences are likely to share a fine-grained phenotype. This may be an immediate extension of the fact that related TCR sequences are likely to come from the same CD4/CD8 compartment; however, phenotype distributions in significantly low-entropy clusters tended to be dominated by a single fine-grained phenotype.

Data and Code Availability

The encoding and VAE scripts are available in this repository. To run the VaDE_TCR.py script, Keras should be installed *and modified* with the instructions in [8]. The CD4/CD8 dataset from Zhang aligned to V and J genes with MIXCR is also available in this repository. The healthy repertoires from Emerson [1], YFV repertoires from DeWItt [11] and single cell TCR and RNAseq dataset from Azizi [6] are available in the original publications.

References

1. Lidia Robert, Jennifer Tsoi, Xiaoyan Wang, et al. CTLA4 Blockade Broadens the Peripheral T-Cell Receptor Repertoire. *Clin Cancer Res* 20(9), 2424-2432 (2014). https://doi.org/10.1158/1078-0432.CCR-13-2648
2. Emerson, R., DeWitt, W., Vignali, M. *et al.* Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. *Nat Genet* 49, 659–665 (2017). https://doi.org/10.1038/ng.3822
3. Glanville, J., Huang, H., Nau, A. *et al.* Identifying specificity groups in the T cell receptor repertoire. *Nature* 547, 94–98 (2017). https://doi.org/10.1038/nature22976
4. Dash, P., Fiore-Gartland, A., Hertz, T. *et al.* Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature* 547, 89–93 (2017). https://doi.org/10.1038/nature22383
5. Simon Friedensohn, Daniel Neumeier, Tarik A Khan et al. Convergent selection in antibody repertoires is revealed by deep learning. bioRxiv 2020.02.25.965673; doi: https://doi.org/10.1101/2020.02.25.965673
6. Elham Azizi, AMbrose J. Carr, George Plitas et al. Single-Cell Map of Diverse Immune Phenotypes in the breast Tumor Microenvironment. *Cell* 174(5), 1293-1308 (2018). https://doi.org/10.1016/j.cell.2018.05.060
7. Lefranc, M.-P. and Lefranc, G. The T cell receptor FactsBook Academic Press, London, UK (398 pages), (2001)
8. Zhuxi Jiang and Yin Zheng and Huachun Tan and Bangsheng Tang and Hanning Zhou. Variational Deep Embedding: An Unsupervised and Generative Approach to Clustering. arXiv 1611.05148 (2017).
9. Xuan Liu, Sara J C Gosline, Lance T Pflieger, Pierre Wallet, Archana Iyer, Justin Guinney, Andrea H Bild, Jeffrey T Chang, Knowledge-based classification of fine-grained immune cell types in single-cell RNA-Seq data, *Briefings in Bioinformatics*, 2021;, bbab039, https://doi.org/10.1093/bib/bbab039
10. C. Zhang, H. Ding, H. Huang, H. Palashati, Y. Miao, H. Xiong, Z. Lu. TCR repertoire intratumor heterogeneity of CD4 +and CD8 +T cells in centers and margins of localized lung adenocarcinomas. Int. J. Cancer, 144 (2018), pp. 818-827, 10.1002/ijc.31760
11. W. S. DeWitt, R. O. Emerson, P. Lindau, M. Vignall, T. M. Snyder, C. Desmarais, C. Sanders, H. Utsugi, E. H. Warren, J. McElrath, K. W. Makar, A. Wald, H. S. Robins.

Dynamics of the Cytotoxic T Cell Response to a Model of acute Viral Infection. *J. of Virology*, 89(8) (2015). doi: https://doi.org/10.1128/JVI.03474-14