# PRIORITIZING NETWORK PROPERTIES OF TCR REPERTOIRE

## A Novel Approach To Select Network Signatures From TCR Repertoire Data


Shilpika Banerjee[1] Li Zhang[2] Tao He[1]

[1]*Department of Mathematics, San Francisco State University;* [2]*Department of Medicine, University of California San Francisco*


# Introduction

## Background

T cells are **crucial components of the adaptive immune system**, mediating anti-tumoral immunity and immune response to infections.

T cell receptor (TCR), which is a protein complex on T-cell surface, targets specific antigens based on nucleotide sequence.

Figure 2

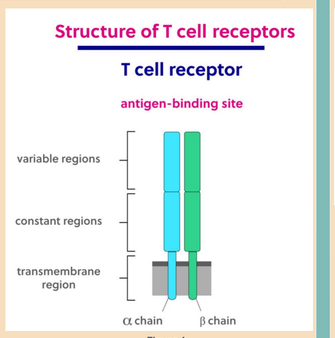


Figure 1

TCR repertoires **continually shaped throughout the lifetime** of an individual in response to pathogenic exposure and can serve as a fingerprint of an **individual's current immunological profile**.

The **similarity among TCRs sequence directly influences the antigen recognition breadth**. Network analysis, which allows interrogation of sequence similarity, thereby adds an important layer of information. To construct a clonal network, each clone is defined as a node, and then based on the sequence distance, an edge is drawn based on a certain similarity condition (e.g., one letter difference in sequence).

## Motivation & Objective

The objective is to investigate the network properties and develop novel statistical method to prioritize the important network properties that are associated with the clinical outcome.

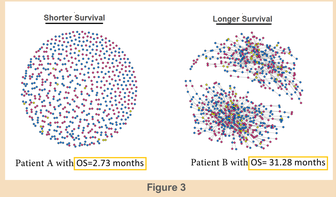


Figure 3

*(**Fig-3** show the networks of two lung cancer patients' TCR repertoire. There is some evidence that the two groups (shorter & longer survival) have different network structures)*

**Network analysis allows interrogation of sequence similarity.**

OS_mon: Overall Survival Months
**Longer Survival Group: 'OS_mon≥20.3**
**Shorter Survival Group: 'OS_mon'<20.3**

## Challenges

**Heterogeneous nature** of the TCR repertoire and network properties makes it extremely difficult to perform statistical inference or machine learning directly between subjects.

- Less than 20% overlaps across repertoire, even for the same subject.
- Each network has different total number of clusters, total number of nodes.
- Network properties have difference representations: some of them are global (described by one number) and some of them are local (described by a vector of values, and vector length varies).

## Our Contribution

- We proposed a strategy to extract features from **heterogeneous global/local network properties;**
- We developed a novel statistical method to **prioritize the network properties that are associated with the outcome of interest, based on the extracted network features;**
- We proposed a procedure to **simulate network properties using the real data**, to mimic real property distributions and correlation structure;
- We demonstrated proposed method and schemes via both simulation study and real data analysis.

## TCR repertoire properties

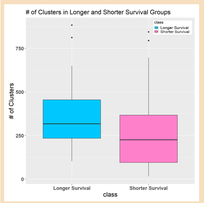


Figure 3

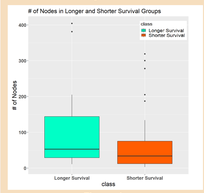


Figure 4

*(**Fig-3, 4** show longer survival subjects have more # of clusters and nodes compared to the shorter survival subjects.)*

| Variables | Network Properties | Type | Definition | illustration |
|---|---|---|---|---|
| membership | Cluster size, number | Global | Connected component of a graph in which any two nodes are connected | Number = 2 clusters Size = 3,6 |
| node_count | Number of Nodes | Global | The fundamental unit of which graphs are formed: v | |
| deg | Degree | Global | The number of edges incident to a vertex $c$: $deg(v)$ | |
| AA_length | Non-network properties (Pre-infusion and Dose 2nd are phases when biological samples and clinical data were collected from the subjects.) | | | |
| Count_PRE_INFUSION | | | | |
| Count_DOSE_2 | | | | |
| deg_avg | Average Degree | Global | The average number of degrees per node: $2|v|$ | |
| diam_length | Diameter Length | Global | The length of the "longest shortest path" between any two vertices: $max_{(v,w)\in V} d(v,w)$ | |
| assortativity | Assortativity | Global | Pearson correlation coefficient of degree between pairs of linked nodes $r \in [-1,1]$ | $r>0$ $r<0$ |
| transitivity | Clustering coefficient (Transitivity) | Global | The probability that the adjacent vertices of a vertex are connected | |
| edge_density | Density | Global | The ratio of the number of edges and the number of possible edges | |
| centr_degree | Degree Centrality | Global | Centrality score based on node-level centrality $c$: $sum(max(c(v_i), w) - c(v_i), v)$ | |
| centr_clo | Closeness Centrality | Local | Node centrality in a graph: $C(v) = \frac{1}{\sum_{u \in V} d(v,u)}$ | |
| eigen_centrality | Eigenvector Centrality | Local | Returns the eigenvector centralities of positions v within a given graph | |
| centr_eigen | | Local | A centralization measure | |

Table 1

***Table 1**. Description of Network Features, including network properties and cluster level TCR clone features (Miho et. al, 2019)*

# Methods

## Aggregating Heterogenous Network Data

TCR network repertoire vary in structure and sizes across different subjects and are continually shaping. Therefore, **summary statistics are extracted from the properties and are grouped as blocks.** Objective is to study which features are significant and which feature blocks (n/w properties) as a group are significant.

For a single subject, the **Transitivity** n/w property has several **NA's** and the numerical values have roughly a **log-normal form.**

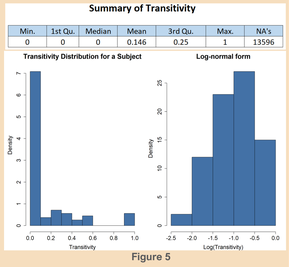


Figure 5

The following summary statistics are derived for Transitivity and considered as a single block/group for feature extraction:
***prob(NA), Min, Q1, Median, Mean, Q3, Max***

Similarly, based on the values of the TCR network properties, relevant summary statistics are derived for each property and distinct groups are created for them.

| Properties | Feature Blocks | Feature Index | Groups |
|---|---|---|---|
| membership | # of clusters | 1 | 1 |
| node_count | Min, Q1, Median (Q2), Mean, Q3, Max | 2-7 | 2 |
| deg | Min, Q1, Median (Q2), Mean, Q3, Max | 8-13 | 3 |
| AA_length | Min, Q1, Median (Q2), Mean, Q3, Max | 14-19 | 4 |
| Count_PRE_INFUSION | Min, Q1, Median (Q2), Mean, Q3, Max | 20-25 | 5 |
| Count_DOSE_2 | Min, Q1, Median (Q2), Mean, Q3, Max | 26-31 | 6 |
| deg_avg | Min, Q1, Median (Q2), Mean, Q3, Max | 32-37 | 7 |
| diam_length | Min, Q1, Median (Q2), Mean, Q3, Max | 38-43 | 8 |
| assortativity | prob(NA), Min, Q1, Median (Q2), Mean, Q3, Max | 44-50 | 9 |
| transitivity | prob(NA), Min, Q1, Median (Q2), Mean, Q3, Max | 51-57 | 10 |
| edge_density | Min, Q1, Median (Q2), Mean, Q3, Max | 58-63 | 11 |
| centr_degree | Min, Q1, Median (Q2), Mean, Q3, Max | 64-69 | 12 |
| centr_clo | prob(NA), Min, Q1, Median (Q2), Mean, Q3, Max | 70-76 | 13 |
| eigen_centrality | Min, Q1, Median (Q2), Mean, Q3, Max | 77-82 | 14 |
| centr_eigen | prob(NA), Min, Q1, Median (Q2), Mean, Q3, Max | 83-89 | 15 |

Table 2

## Models Used for Prioritizing and Selecting Network Properties

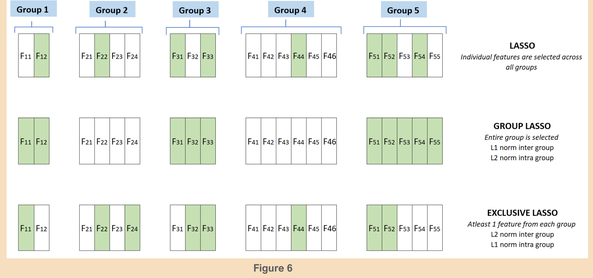


Figure 6

## Prioritizing network properties: GROUP LASSO with permutation tuning

- We developed a novel **permutation assisted tunning group-lasso based approach** to identify the active groups of network properties.
- Compared to cross-validation tunning, permutation assisted tunning helps to **reduce potential false positives** of variable selection by using **pseudo-variables**.

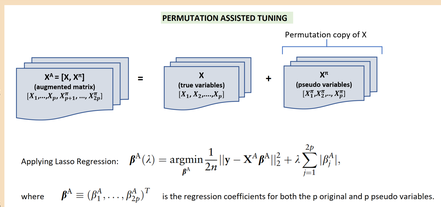


Figure 6

## Selecting top network features using PLASSO and Exclusive LASSO

Lasso regression with permutation assisted tuning (PLASSO) is used to select the top TCR network features across all groups (Yang, et. al. 2020). PLASSO is indifferent to the group structure created for the network properties.

Exclusive lasso with cross validation on the other hand selects the top features from each TCR network property groups.

## Proposed simulation scheme

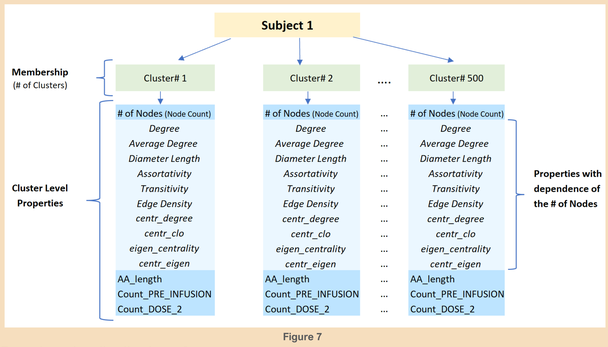


Figure 7

Based on the distribution of the # of clusters per subject, the data for 'Membership' is generated for **1000 dummy subjects and is later down sampled.**

Using histogram plots for the remaining properties a rough estimation of their distribution is made. Leveraging the simulated Cluster count, and evaluating their dependencies on the # of Nodes, the remaining network properties are simulated.

| Properties | Value |
|---|---|
| Node_Count | 2 |
| Assortativity | NA |
| Transitivity | NA |
| Degree | 1 |
| diam_length | 2 |
| deg_avg | 1 |
| edge_density | 1 |
| centr_degree | 0 |
| centr_clo | NA |
| eigen_centrality | 1 |
| centr_eigen | NA |

Table 3

***(Tbl-3 represents** the correlation of the other network properties on 'Node_Count' values.)*

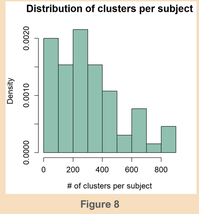


Figure 8

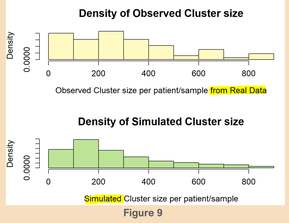


Figure 9

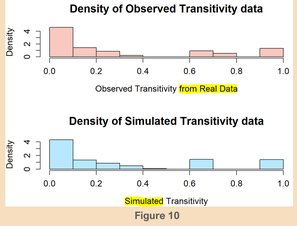


Figure 10

*(**Fig-9,10** are used to show the Density plots for the real (observed) data and the simulated data for two of the TCR repertoire properties.)*

# Results

## Real Data Analysis

LASSO (with Cross Validation), PLASSO (LASSO with permutation tuning), Group PLASSO, Group LASSO (with Cross Validation), and Exclusive LASSO (with Cross Validation) feature selection techniques were run on the real data.

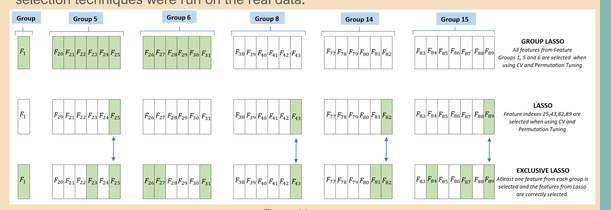


Figure 11

## Simulation Study

Simulation studies were performed for comparing the different feature selection techniques and to assess their performances.

| Simulated Data Analysis | | | | | | |
|---|---|---|---|---|---|---|
| **Model** | **TRUE Group Indexes** | **Sensitivity** | **FDR** | **F-1** | **Power** | **Stability** |
| *GROUP_PLASSO* | Grp-1,Grp-5,Grp-6 | 0.9 | 0.275 | 0.8 | 1.0,7,1 | 0.8833333 |
| *GROUP_LASSO_CV* | Grp-1,Grp-5,Grp-6 | 0.9 | 0.54 | 1,0.8,0.9 | | 0.5549206 |
| **Model** | **True_Feature_Indexes** | **Sensitivity** | **FDR** | **F-1** | **Power** | **Stability** |
| *PLASSO* | 25,43,82,89 | 0.75 | 0.49 | 0.6 | 1,1,0,1 | 0.8 |
| *LASSO_CV* | 25,43,82,89 | 0.575 | 0.3967857 | 0.532424 | 0.6,0.9,0,0.8 | 0.4214286 |
| *EXCLUSIVE_LASSO* | 25,43,82,89 | 1 | 0.8518519 | 0.258065 | 1,1,1,1 | 1 |

Figure 12

| Model | True Group Indexes (from Lasso) | Sensitivity | FDR | F-1 | Power | Stability |
|---|---|---|---|---|---|---|
| *GROUP_PLASSO* | Grp-5,Grp-8,Grp-14,Grp-15 | 0.425 | 0.5666667 | 0.4242857 | 0.6,1,0,0.1 | 0.8074074 |
| *GROUP_LASSO_CV* | Grp-5,Grp-8,Grp-14,Grp-15 | 0.55 | 0.6309524 | 0.4376623 | 0.7,0.8,0,0.7 | 0.5962963 |

Figure 13

# Conclusions

- **Real Data Analysis**:
  - *Group Lasso* renders feature blocks '*Membership (# of Clusters)*', '*Count_PRE_INFUSION*', and '*Count_DOSE_2*' as significant groups.
  - *Lasso and Exclusive Lasso* render features '*Count_PRE_INFUSION_Max*', '*diam_length_Max*', '*eigen_centrality_max*', '*centr_eigen_max*' as significant.
- **Simulation Study**:
  - *Group Lasso* on simulated data has better performance when using the 'True Group Indexes' from Group Lasso run on real data.
  - Permutation assisted tuning has better performance measures than that of cross-validation.

***From the above study we can conclude that for Feature Selection, permutation assisted tuning performs better than cross-validation.***

| Model | Overall Performance |
|---|---|
| *GROUP_PLASSO* | 👍 |
| *GROUP_LASSO_CV* | 👎 |

Figure 14
*(Feature Group selection)*

| Model | Overall Performance |
|---|---|
| *PLASSO* | 👍 |
| *LASSO_CV* | 👎 |
| *EXCLUSIVE_LASSO* | N/A |

Figure 15
*(Feature Selection)*

# References


- Yang S, Wen J, Eckert ST, Wang Y, Liu DJ, Wu R, Li R, Zhan X. Prioritizing genetic variants in GWAS with lasso using permutation-assisted tuning. *Bioinformatics*. 2020 Jun 1;36(12):3811-3817.
- Miho, E., Roškar, R., Greiff, V. et al. Large-scale network analysis reveals the sequence space architecture of antibody repertoires. *Nat Commun* **10**, 1321 (2019).
- Frederick Campbell and Genevera Allen. Within Group Variable Selection through the Exclusive Lasso. *arXiv:1505.07517v1 [stat.ME]*, 2015.
- Lukas Meier, Sara van de Geer and Peter Bühlmann. The group lasso for logistic regression. *J. R. Statist. Soc. B 70, Part 1, pp. 53–71*, 2008.


# Acknowledgements


- TH and LZ are partially supported by NIH/NCI R21CA264381 and NIH/NLM R01LM013763-01A1
- TH is partially supported by NSF DMS-2137983.


SCAN ME