Discovering B cell receptor genotype patterns with mining and association rules

Ayelet Peres

February 2023

Abstract

The complexity of the B cell repertoire stems in part from its germline diversity. One of the ways to show it is by inferring individual genotype. This is essentially a composition of the genes and alleles an individual carries. We sought a way to utilize this genotypic composition to create a classification model for differentiating between healthy and disease. For that, we have comprised a unique data set of 98 Individuals, of both healthy and celiac disease genotype inference. As the genotype includes many allelic variations, compared to the number of individuals in our cohort. We sought a method to extract unique patterns within it to use a model base. Thus, we have association rule mining for the patterns' detection. Using the detected patterns and comparing the results to a decision tree model, we saw a slight improvement in the classification results.

1 Introduction

The B cell are part of the human immune system, the main purpose of this system is to protect the body against pathogenic invasions. B and T cells are its main components and carry diverse repertoires of B and T cell receptors (BCRs and TCRs). These receptors are composed of two chains. Each contains a constant and highly variable region to identify many pathogen signatures. The variable regions can be divided into three blocks of gene, Variable (V), Diversity (D), and joining (J). A selection of a single gene from each of the parts creates the receptor chain. Another important factor, is that each gene carries a specific allele, i.e a certain modification of the gene.

Though the genes rarely vary in the population, the allelic modification does. Meaning that each individual can carry different allele for the same gene. A genotype is a term that summarizes the set of genes and alleles that an individual carries. In the case of B and T cell, we get a "snapshot" of the variation of each individual. For example, Fig. **??** shows an Individual's genotype, where each row is a different gene, and each color is a different allele (The naming of the genes and alleles have no relation to each or particular meaning). Recently, we the advances of high throughput sequencing (HTS), more studies have started to sequence the individual's B and T cell receptor repertoires to obtain more insight on the immune system. This allows us to construct genotypes and explore the diversity between individuals.

2 Problem description

The genotype composition of B cell includes hundreds of alleles, each one of them or a combination of them could potentially be linked to certain autoimmune diseases. However,

finding these markers proves to be challenging. Mainly, because of a high dimensionality problem. As the genotype inference from immune repertoire data is still in its early stages, the number of samples is quite low relatively to the number of alleles. Hence, a solution that can both harness valuable information from the genotype composition and classify individuals based on clinical status can be of value.

3 Solution overview

The problem that we presented above touches on two solution paths, the first is pattern mining and the second is classification. As the genotype is a collection of alleles an individual carries, when we observe a population we can represent their genotype as a boolean matrix. Where we account if an allele appeared in an individual genotype. From there we can think on mining patterns that can later be connected to clinical status.

There are many pattern mining algorithms that can be used, however as we want to preserve the association between the markers we utilize the association rules (AR) algorithms. The genotype dataset can be viewed as a set of transactions, meaning that each individual genotype composition is similar to a single shopping transaction. This fits the world of pattern mining from transaction datasets, also known as association rules (AR). An association rule is mined from a dataset of transactions, where each transaction t $T \in D$ contains itemset $X \subseteq Ifx \subseteq T$.

There are several algorithms that extract rules from transaction like datasets, namely Apriori and FP-Growth. The former employs a bottom up approach, where it identifies the frequent items and expands them to larger and larger sets as long as the item appear enough in the dataset. However, this approach is expensive when dealing with large datasets or itemsets. The FP-Growth approach tries to be more efficient with generating a frequent pattern tree, which records how often each item pattern appears. However, for both, the process of deriving a high order association rule can be computationally costly. Making this process unusable in datasets that require low support threshold.

As such, we thought of a different way to retain the power of AR but constructing it in a way that would be less computationally costly. To do so, we leverage the transitivity property of AR to construct a new data structure that is directly derived from low order AR and relies on graph theory. Figure 1 demonstrates the transitivity, where the red and blue rules and are in a sense a play on a directed graph with three nodes, as shown in green.

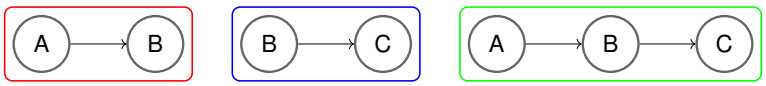

Figure 1: **Association rules to network graph**

In depth, we have developed algorithm 1 that constructs the rules into a weighted association rules graph (ARG). Where each node is an item from our itemset basket, and the directed edges are the connection between the left side and right side of the rules. The weights of the nodes and edges of the graph are derived from a classification model (i.e., decision tree, logistic regression, etc.). Such that for each edge, the weight is calculated as the F1-score of the classification model given both nodes of the edge. The weights can be later refined, depending on the desired task we wish to perform.

Once the ARG is constructed, we can move to the classification problem. Where we wish to extract meaningful features from the ARG that will maximize the classification score. We have created three path extraction algorithms for this task:

Algorithm 1 Construct Weighted Association Rule Graph

Require: List of nodes V and List of edges E
Require: V the set of all unique features
Require: Each Edge in E is a tuple of 2 features from V corresponding to an association rule
Ensure: Weighted graph G
  for each edge e in E **do**
    $n_1 \leftarrow$ node 1 of edge e
    $n_2 \leftarrow$ node 2 of edge e
    $f1_{score} \leftarrow$ calculate F1 score of n_1 and n_2 in Model
    Set weight of edge e to $f1_{score}$
  end for
  $G \leftarrow$ graph with nodes V edges E and W weights
  return G

- GREEDY WEIGHTED WALK algorithm gives an F1 score weight for each node, and the weights of the edges reflect the increase factor of the classification with just node antecedent versus with both nodes antecedent and consequent. Figure 2 shows the weight of the edges. The walk starts from the maximal F1-score value node, for the next step each adjutant edge's weight is checked for the maximal increase in the F1-score. The walk stops either when the no other possible steps are available or when reached the maximal number of steps.

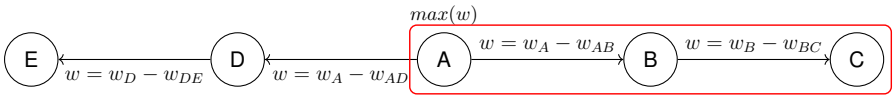

Figure 2: **Greedy weighted walk**

- CLIQUE algorithm will identify all possible cliques and choose the one with the maximal score. A clique is a group of nodes in a graph if each pair of distinct nodes are adjacent. Then we will select the one with the maximal weights. Where the weight of a clique corresponds to the total weight of its individual nodes, meaning each node F1-score (Figure 3).

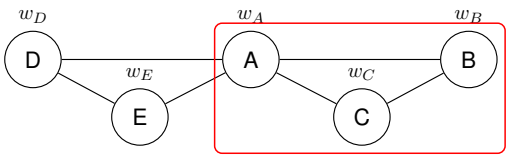

Figure 3: **Greedy weighted walk**

- GREEDY COLORING algorithm will attribute a color to the nodes, and choose the color with the maximal score. The nodes are colored based on the attempt to minimize the number of colors used, while ensuring that adjacent nodes have different colors. The order in which the nodes are colored is determined by a specified strategy (Figure 4).

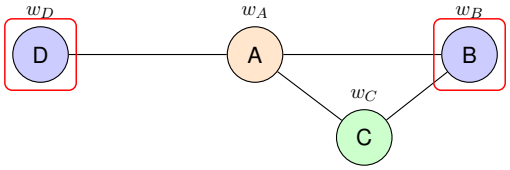

Figure 4: **Greedy coloring**

Each of the three algorithms we defined above can provide different features and insights. While both the greedy weighted walk and the cliques consider nodes that are close, the greedy coloring does the opposite. With taking nodes that have at least one separating nodes, this can perhaps solve the collinearity issues.

To test the algorithms' performance, we will compare the results of the classification to a base model of decision tree that considers all possible alleles within the genotype dataset for the classification.

4 Experimental evaluation

4.1 Dataset

To evaluate the algorithms, we used a public genotype inference dataset from VDJbase. The dataset contains 98 individuals, divided into healthy and celiac patients. The genotyped contains information on the alleles of 110 different genes. In individual can carry more than a single allele per gene; thus the dataset includes 276 different alleles. We have processed the dataset and removed alleles that were too frequent and appeared in all individuals. This resulted in removing 30 alleles. Then we discarded genes that we did not have information about them in the entire cohort, that resulted in removing 43 genes. Finally, we were left with 163 allele calls, meaning that the final dataset was of the shape 100x164.

4.2 Evaluation Method

We will evaluate our algorithm in two parts. First, we will tune the hyperparameters of the algorithm, both for the produced association rules and for the feature selection with our association rules graph. For this part, we will use the metric of F1-score, as our dataset is of medical nature, maximizing the precision and recall can later for the prediction. The base classification model for this part will be a decision tree model. The model will be used to evaluate the paths curated from the ARG, the results will be compared to a plain model that includes all the alleles. In the second part, after obtaining the desired optimization, we will evaluate the paths that produced the height scores. Specifically, we will want to see if there are alleles that appear more than others and if the association rules algorithm can find them in the dataset. As the computational power is limited, we have limited the produced shorter paths with fewer items. This will insure we can run the algorithm of association rules and validate the paths.

All path curation methods were checked simultaneously on the same graph. The following steps were run in cross validation, such the rules and the graph were created only on the training data and evaluated on the test.

- CURATING RULES We first used the FP-growth from the library of *mlxtend* to obtain the frequency of the itemsets. Then the association rule algorithm from the same library was run to produce the rules.

- CREATING THE AR GRAPH The obtained rules were then used to create the network graph. Where each node and edge was given a weight value as described above.

- FINDING BEST PATHS After creating the graph, we applied the three walks described above and obtained one path from each

• EVALUATING THE CLASSIFICATION With the path we obtained, we evaluated the classification and the F1-score.

4.3 Algorithm evaluation results

We ran the algorithm on the celiac dataset, we tested the effect of the association rules hyperparameters on the classification as well as the edges weight for the Greedy weighted walk. We have defined Three types of weights:

• INCREASE FACTOR (IF) this is calculated based on the F1-score difference between classifying with just the antecedents and with both the antecedents and the consequents nodes

• LIFT INCREASE FACTOR(LIF) we have multiplied the IF with the lift value of the rule (Lift is the ratio of the confidence of the rule and the expected confidence of the rule).

• CONFIDENCE INCREASE FACTOR(CIF) we have divided the IF with the confidence value of the rule (Confidence is the percentage value that shows how frequently the rule head occurs among all the groups containing the rule body).

As we can see, though, we did mange to improve our results from the naive decision tree classifier the classification F1 score is still quite low (Tab. 1). The ARG algorithm with the greedy weighted walk averaged the F1-score around 0.77, when we tried to change the weights of the edges it did not improve the classification results.

	algorithm	walk	edge weight	minimum support	minimum confidence	Repeats	CV averaged F1-score
1	Decision Tree Classifier					100	0.55
2	ARG	Greedy weighted walk	F1	0.0001	0.0001	100	$0.77 \pm (0.025)$
3	ARG	Cliques		0.0001	0.0001	100	$0.54 \pm (0.027)$
4	ARG	Greedy color		0.0001	0.0001	100	$0.49 \pm (0.075)$
5	ARG	Greedy weighted walk	F1	0.4	0.4	100	$0.75 \pm (0.013)$
6	ARG	Cliques		0.4	0.4	100	$0.48 \pm (0.078)$
7	ARG	Greedy color		0.4	0.4	100	$0.46 \pm (0.029)$
8	ARG	Greedy weighted walk	F1	0.01	0.4	100	$0.74 \pm (0.002)$
9	ARG	Cliques		0.01	0.4	100	$0.44 \pm (0.155)$
10	ARG	Greedy weighted walk	LIF	0.0001	0.0001	100	$0.62 \pm (0.037)$
11	ARG	Greedy weighted walk	CIF	0.0001	0.0001	100	$0.71 \pm (0.027)$

Table 1: **Result of applying the algorithm on the celiac dataset**

However, we did get recurring paths, specifically in the Greedy weighted walk. In total, we had 500 repetitions in each experiment (5 cross validation and 100 repeats). For rows 2,5,8,10, and 11 from Table. 1 we got the following paths, each appearing at least 50 times in the experiments:

- IGHV3-64D*09,IGHV4-30-2*Deletion,IGHV5-51*03
- IGHV1-46*01,IGHV4-30-2*01,IGHV5-51*01
- IGHV1-46*01,IGHV3-73*Deletion,IGHV4-30-2*Deletion,IGHV5-51*03
- IGHV1-3*Deletion,IGHV4-38-2*02,IGHV5-51*03

It was interesting to see that there are several alleles that repeats between the paths, such as IGHV5-51*03 or IGHV1-46*01.

4.4 Association rules insight

We wanted to get further insight into the paths we have obtained. Hence, we ran the association rules algorithm again, however this time we expanded the number of items in a rule and we included the clinical status. We wanted to see if any of the path that we obtained will get an association with the clinical study.

5 Related work

Data mining has seen a big rise over the last couple of decades with advances in computation power and the in neural network power. Yet, association rules (AR) still has a place, though it is mostly used in recommendation system in the field of economics. In the field of bioinformatics, we can divide the work using AR into three categories, the first is clinical data, the second is gene expression; the third is whole genome sequencing (WGS) single-nucleotide polymorphism (SNP) genotype. The works in the category of gene expression are trying to connect genes to clinical status, with giving more meaning to association. The works by these authors [3, 1, 2] try to create algorithms for better filtration of AR derived from gene expression analysis. They do so by attributing different weights to the rules. Other works such as these [4, 5] developed tools for AR mining of WGS SNP genotypes, where they aim to link SNP to disease for predictive power. These works have guided the creation of our new data structure to represent the rules and the weights given to the edges.

6 Conclusion

The goal of this work was to create a new data structure to represent association rules derived from immune repertoire genotype data, and to use it for classification problems. We have developed an automated tool for creating a rule-based network graph. We have added three different algorithms to extract paths from the graph based on different metrics. We have added a layer of classification model prediction based on the chosen path of the graph. We have tested our algorithm on a curated dataset, though the classification results were not high enough, with more fine-tuning of the algorithm and exploring other weights metrics, we can achieve better results.

References

[1] Shu-Chuan Chen, Tsung-Hsien Tsai, Cheng-Han Chung, and Wen-Hsiung Li. Dynamic association rules for gene expression data analysis. *BMC genomics*, 16:1–20, 2015.

[2] Saurav Mallik, Anirban Mukhopadhyay, and Ujjwal Maulik. Ranwar: rank-based weighted association rule mining from gene expression and methylation data. *IEEE transactions on nanobioscience*, 14(1):59–66, 2014.

[3] Gyorgy J Simon, Vipin Kumar, and Peter W Li. A simple statistical model and association rule filtering for classification. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 823–831, 2011.

[4] Ito Wasito, Mujiono Sadikin, Teny Handhayani, et al. Predictive genotype based on phenotype using the association rules mining. In *2013 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pages 185–188. IEEE, 2013.

[5] Qingrun Zhang, Quan Long, and Jurg Ott. Apriorigwas, a new pattern mining strategy for detecting genetic variants associated with disease through interaction effects. *PLoS computational biology*, 10(6):e1003627, 2014.