Identifcation of candidate gene targets for therapeutic intervention to rescue gene expression in B cells from smoker to non-smoker levels

Mathew Golf
mathewggolf@lewisu.edu
DATA-51000-001, Summer 1
Data Mining and Analytics
Lewis University

I. Introduction

Smoking continues across multiple cultures and generations despite the known consequences and risks associated with it. In particular B cell changes are known to be directly associated to the onset of smoking-related diseases[1]. Therefore, we sought to identify association rules of gene expression data for female smokers compared to female non-smokers (control). In doing so, we hope to uncover differential gene associations between groups which may identify candidate gene targets for intervention to reverse the effects of smoking observed in B cells. Since we are identifying candidate gene targets for intervention, we will mainly consider associations between a single gene and a set of genes.

Peripheral circulating B cell gene expression data was obtained through the use of a genome-wide Affymetrix HG-133A GeneChip microarray[1]. 39 smoking and 40 non-smoking individuals were used, capturing gene expression data for 2999 genes[1]. This dataset was split into smoking and non-smoking for separate investigations using the discretization techniques relating to width and frequency. Both techniques are known for use on gene expression data and may generate different results, as the width discretizes based on metrics and frequency discretizes based on ranking[2].

In section II, we provide an overview of the data describing the features used for separating the data in order to compare association rules between smokers and control. Then in section III, the methodology for obtaining, splitting, and processing the data for analysis is explained. Section IV, describes the results from the aforementioned analysis and discusses the relevance of said results. Lastly, in section V conclusions from the results and the discussion of the analysis is provided indicating future directions for further investigation.

II. Data Description

For this analysis one dataset was used and later divided based on the stress attribute, separating smokers from non-smokers. Separation was done in order to compare association rules and identify different relationships between gene networks of smokers and non-smokers in B cells. 2999 gene expression values are recorded from ATF3 to KIAA1549L for 39 instances of smokers and 40 instances of non-smokers (Table I). All genes were used in the analysis to generate association rules for both smokers and non-smokers.

TABLE I. B Cell Gene Expression Values of Smokers and Non-Smokers

Attribute	Type	Example Value	Description
STRESS	Nominal (string)	Control	Indicates non-smoker (control) or smoker
ATF3	Numeric (real)	304.792	Gene expression value of ATF3
⋮	⋮	⋮	⋮
KIAA1549L	Numeric (real)	110.702	Gene expression value of KIAA1549L

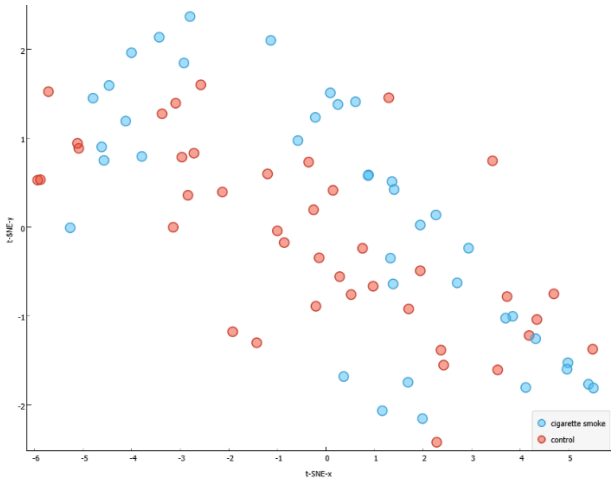

Fig. 1. Manifold learning transformation of all B cell gene expression values for smokers (blue) and non-smokers (red).

TABLE II. DESCRIPTIVE STATISTICS OF THE MANIFOLD LEARNING TRANSFORMED B CELL GENE EXPRESSION DATA

Attribute	Mean	Median	Range
t-SNE-x	-1.96397e-16	0.28943	-5.92726 to 5.51418
t-SNE-y	-1.12428e-17	0.0230234	-2.42334 to 2.36984

III. METHODOLOGY

Data was obtained through the in-built dataset feature of Orange and it was produced by Pan *et al.*[1]. Before splitting the dataset we processed the data using manifold learning in order to visual the high-dimensional data in two-dimensions while preserving any non-linear relationships between gene expression values. Non-linear relationships in biological data are important to preserve as nature is messy and many interactions may be missed if traditional methods, which preserve linear relationships are used instead. Once transformed we then visualized the data using a scatter plot and observed the feature statistics of the transformed data. Clearly, two distinct groups are apparent in the data with some overlap – smoker vs non-smoker. Therefore, we split the dataset into these two groups before discretizing the data. Each group was discretized with two different methods and association rules were produced for each method totaling four association rule sets generated. Methods for discretizing were conducted using either equal-width or equal-frequency. Equal-width discretization divides the maximum and minimum values into *k* intervals of equal width, for the whole expression matrix thus discretizing based on metrics[2]. This technique can also be used on gene expression profiles or the condition expression profile[2]. Equal-frequency discretization on the other hand assumes the expression values are in a list sorted in decreasing order, and it splits this ordered-list on the given *k* value, therefore discretizing based on ranking[2].

IV. RESULTS AND DISCUSSION

Since we are interested in identifying genes which may be prime candidates for therapeutic intervention, we set our confidence level high as our target for intervention must be correct often (Conf. = 80%). Similarly, we want a high support value as we must be able to apply this association rule to a large number of cases (Supp. = 80%). As such, by increasing these values we restrict how large the lift can be. Therefore, throughout our analysis we will also consider smaller confidence and support values to discover any association values with sufficiently large lift values.

Our initial analysis focused on the non-smoker B cell population and association rules were generated using the equal-width parameter for discretization (Table III). As such, the rules selected had support values varying from 0.925 to 0.850 indicating that these are convincing rules, and they apply to a large amount of the cases in our analysis (Table III). Furthermore, the confidence values associated with these rules are high ranging from 0.949 to 0.872 indicating strong associations and that they will be correct often (Table III). Likewise, the coverage of the association rules selected (0.975) indicates that the antecedent can be found frequently throughout the association rules generated and can therefore be applied more often (Table III). No lift values generated with the constraints of 80% for confidence and support levels surpassed 1.000 (Table III). Therefore, as the lift values are close to 1.000 we can assume that the associations are not a coincidence but that they may also not be statistically significant (Table III). Considering all entries in Table III, it is clear that there is lots of association between a select few of the 2999 genes. All genes selected in Table III should be considered for KEGG pathway analysis in order to discover what functional role these networks are associated with, and for the case of non-smokers why this network association relates to a healthy phenotype.

TABLE III. ASSOCIATION RULES FOR B CELLS OF NON-SMOKERS DISCRETIZED VIA EQUAL-WIDTH

Support	Confidence	Coverage	Lift	Antecedent	Consequent
0.925	0.949	0.975	0.999	TIE1	FOS, SLC6A15
0.925	0.949	0.975	0.999	SLC6A15	FOS, WNT4
0.925	0.949	0.975	0.999	TIE1	FOS, WNT4
0.925	0.949	0.975	0.999	TIE1	SLC6A15, WNT4
0.925	0.949	0.975	0.999	SLC6A15	FOS, KRT19
0.925	0.949	0.975	0.999	WNT4	FOS, KRT19
0.925	0.949	0.975	0.999	SLC6A15	KRT19, WNT4
0.850	0.872	0.975	0.996	TIE1	FOS, KRT19, SLC6A15, WNT4, GOLM1
0.850	0.872	0.975	0.996	SLC6A15	ARHGAP22, FOS, KRT19, WNT4, GOLM1

A similar analysis was then performed on the same population of B cells of non-smokers, however instead of equal-width discretization we next used equal-frequency discretization to observe any frequent associations (Table IV). Interestingly, we struggled to find any association rules with equal-frequency discretization when our support and confidence levels were set to 80%, therefore we had to experiment. In doing so we generated low support values ranging from 0.300 to 0.025 indicating that these association rules are not found frequently throughout the dataset, which also agrees with our low coverage values (Table IV). Nevertheless, the confidence values remained high indicating that our rules should be correct often and our lift values were all above 1.000 indicating that these rules may be statistically significant (Table IV). For example, the last two rules generated in Table IV have lift values of 40.000 and despite their poor support values (0.025) we are confident (1.000) these associations are significant. Clearly, RPLx and RPSx occur frequently in the dataset as they generated the majority of association rules and therefore these genes may be fundamental for the healthy B cell phenotype (Table IV). KEGG pathway analysis for the genes associated with the high lift values may also uncover fundamental gene networks for healthy B cell populations. All genes identified in Table III and Table IV are candidate gene targets to alter expression levels of smokers back to these non-smoker B cell phenotype levels.

TABLE IV. ASSOCIATION RULES FOR B CELLS OF NON-SMOKERS DISCRETIZED VIA EQUAL-FREQUENCY

Support	Confidence	Coverage	Lift	Antecedent	Consequent
0.300	1.000	0.300	3.333	RPL39, RPL9	RPS3A, RPS23, RPS24
0.300	1.000	0.300	3.333	RPL3, RPS23	RPL39, RPL9, RPS24
0.300	0.857	0.350	2.857	RPS23	RPL39, HLA-DRA
0.300	0.857	0.350	2.857	RPS23	RPS24, HLA-DRA
0.300	0.857	0.350	2.857	RPS23	RPL39, RPS24, RPL3, HLA-DRA
0.300	0.857	0.350	2.857	RPS3A	RPL39, RPL9, RPS24, RPL3, RPS23
0.300	0.857	0.350	2.857	RPL9	RPL39, RPS24, RPL3, RPS23
0.300	0.857	0.350	2.857	RPS24	RPL17, RPS3A
0.025	1.000	0.025	40.000	GNA11, HPS6, ANGPT2	ATF3, MDK, FADS1
0.025	1.000	0.025	40.000	PIK3CG, HPS6, BSG	ATF3, FADS1, MAPK1IP1L

Analysis was then conducted on the gene expression recorded for B cells of smokers, initially we used equal-width discretization (Table V). Table V resembles Table III in that our support, confidence, and coverage values are high for the resulting association rules. The clear difference is that some association rules for Table V have lift values greater than 1.000 indicating potential statistical significance for the associations. Some genes are similar between smokers and non-smokers such as: GOLM1 and WNT4 (Table III & V). However, the antecedent differs between tables and therefore this may suggest a candidate gene target for therapeutic intervention to rescue smoker B cells to non-smoker B cell gene expression levels. Of course, all genes found in Table V should be checked in the KEGG pathway database to find any networks with functional significance that may be prime targets for intervention. Furthermore, it is clear that gene expression is different between smoker and non-smoker B cell populations even in the comparison of the gene expression data discretized using the equal-width parameter.

TABLE V. ASSOCIATION RULES FOR B CELLS OF SMOKERS DISCRETIZED VIA EQUAL-WIDTH

Support	Confidence	Coverage	Lift	Antecedent	Consequent
0.974	1.000	0.974	1.026	S100B	FAM66D, GOLM1
0.949	0.974	0.974	1.026	S100B	FAM66D, HOXA1
0.949	0.974	0.974	0.999	PDE9A	FAM66D, GOLM1
0.923	0.974	0.974	0.999	PDE9A	FAM66D, GOLM1, S100B, HOXA1
0.949	0.974	0.974	1.026	S100B	RAB38, FAM66D, GOLM1
0.923	0.973	0.949	1.026	RAB38	PDE9A, FAM66D, GOLM1, S100B
0.923	0.973	0.949	0.999	WNT4	FAM66D, GOLM1
0.923	0.947	0.974	1.026	FAM66D	WNT4, GOLM1, S100B, HOXA1
0.872	0.944	0.923	1.083	CFAP43, GOLM1	WNT4, FAM66D, SGCG
0.821	0.889	0.923	1.083	CFAP43, FAM66D	PDE9A, RAB38, WNT4, SGCG

Lastly, we then focused our analysis on the same B cell smokers dataset discretizing using equal-frequency to find the most common gene occurrences (Table VI). Similar to Table IV, in Table VI our support and coverage values were low whereas our confidence was high. These values indicate that we can be confident in the application of these association rules even if they do not apply to a large proportion of the data. Lift values larger than 1.000 indicate that these association rules may be significant as it is likely they are not just a coincidence (Table VI). Interestingly, genes differed between Table IV and Table VI indicating that the distribution in frequency of genes between smoker and non-smoker samples were different and therefore present as possible candidate gene targets. Common genes between Table IV and Table VI were the RPLx and RPSx gene families which may indicate background expression, fundamental roles, or that intervention can bring their low frequency representation in Table VI to the high frequency occurrence observed in Table IV. Further KEGG pathway analysis is needed to uncover any gene networks and their functions from the indicated association rules uncovered in the dataset.

TABLE VI. ASSOCIATION RULES FOR B CELLS OF SMOKERS DISCRETIZED VIA EQUAL-FREQUENCY

Support	Confidence	Coverage	Lift	Antecedent	Consequent
0.308	0.923	0.333	3.000	TAPBP	TMEM59, DDX39B
0.308	0.923	0.333	2.769	SREBF2	TAPBP, DDX39B
0.308	0.923	0.333	3.000	TAPBP	DDX39B, TNIP1
0.308	0.923	0.333	2.769	KDM5C	TAPBP, DDX39B
0.308	0.923	0.333	3.000	TAPBP	DDX39B, MTA1
0.308	0.923	0.333	3.000	EIF4A2	TAPBP, DDX39B, MTA1
0.308	0.923	0.333	3.000	TAPBP	DDX39B, SREBF2, ARAF
0.308	0.923	0.333	2.769	RPL3	RPL13, RPL17
0.308	0.923	0.333	3.000	TAPBP	DDX39B, NFKB2, SAP18
0.308	0.923	0.333	3.000	RPS3A	RPS16, RPL9, RPL3

V. CONCLUSIONS

Understanding associations of genes that differ between treatment (smoker) and control (non-smoker) in B cells may help to determine candidate gene targets for intervention to change the B cell expression landscape from a smoker back to a non-smoker. Thus, we analyzed gene expression data for 2999 genes in 39 instances of smokers and 40 instances of non-smokers, discretizing the data using both equal-width and equal-frequency techniques in order to capture a variety of association rules or highlight conserved rules between discretization techniques. In doing so, we found the most important rules to be varied for each dataset and in each discretization technique applied (Table III, IV, V, & VI). Indicating there is a clear difference between smoker and non-smoker groups for their B cell gene expression, thus suggesting that therapeutic intervention may rescue the smoker phenotype by restoring the gene associations observed to non-smoker levels. As well, some rules were conserved between discretization techniques, but this was minimal as rules differed greatly between smoker and non-smoker groups regardless of discretization

technique. Indicating that all antecedent genes are candidate gene targets for further analysis, and that combinations of genes may be required to sufficiently influence the complex system.

Future directions would be to use the gene sets discovered as searches for relationships in the KEGG pathway database. Hits within the database will correspond to gene networks involved in functional roles and would present the antecedent as a target for intervention to influence the consequent driving B cell pathology from smoker to non-smoker states. Once the genes have been selected from the sets we have proposed, they will need to be searched in a known drug library to find any proven safe drug intervention that influences expression levels in the direction we desire – either enhancing or suppressing expression. Next this investigation can be done *in vitro* using a lab-on-a-chip to mimic the circulatory system and through isolating B cells from smokers and applying drugs which are known to influence the target genes expression either enhancing or suppressing expression to drive phenotypic change from smoker to non-smoker. These B cells will either be left untreated to observe the likelihood of smoking related diseases and their progression or they will be treated, in order to rescue the B cell phenotype restoring from smoker to non-smoker expression levels. Once all the candidate genes have been tested researchers will determine the significant genes for further experimentation in mice models and eventually in human clinical trials if significant results are found that indicate intervention rescues pathogenic B cell expression levels and reduces the onset or severity of smoking-related diseases.

VI. References

[1] F. Pan et al., "Impact of female cigarette smoking on circulating B cells in vivo: the suppressed ICOSLG, TCF3, and VCAM1 gene functional network may inhibit normal cell function," Immunogenetics, vol. 62, no. 4, pp. 237–251, Apr. 2010, doi: 10.1007/s00251-010-0431-6.

[2] C. A. Gallo, R. L. Cecchini, J. A. Carballido, S. Micheletto, and I. Ponzoni, "Discretization of gene expression data revised," Briefings in Bioinformatics, vol. 17, no. 5, pp. 758–770, Sep. 2015, doi: 10.1093/bib/bbv074.