Charmi Panchal | Vladimir Rogojin

Generating the Logicome from Microarray Data

Turku Centre for Computer Science

TUCS Technical Report
No 0, February 2017

Generating the Logicome from Microarray Data

Charmi Panchal
Turku Centre for Computer Science
Åbo Akademi University, Department of Information Technologies
Agora, Domkyrkotorget 3, FIN-20500 Åbo
cpanchal@abo.fi
Vladimir Rogojin
Turku Centre for Computer Science
Åbo Akademi University, Department of Information Technologies
Agora, Domkyrkotorget 3, FIN-20500 Åbo
vrogojin@abo.fi

TUCS Technical Report
No 0, February 2017

Abstract

The advances in complex statistics and machine learning methods lead to the development of powerfull classifiers that can be used to recognize cellular states, (such as gene expression profiles) that are associated to a number of gene-scale expressed diseases, for instance, cancer. However, the data-driven models built by means of learning from datasets mostly represent ”black boxes” that cannot be easily analysed and understood. On the other hand, a lot of modelling efforts in systems biology are directed towards constructing highly detailed large models for the closer connection to the real life picture. Meanwhile, for the better comprehension of the phenomena, also, a complementary higher abstraction level modeling that captures relations only between the key elements of the larger model, is needed. Recently, there was suggested a method for translating large bio-molecular network models into so-called *logicome*, a small boolean network reflecting activation conditions between key nodes of the large network. In this article, we suggest a method for building a *data-driven logicome*. I.e., the method for building a set of small boolean expressions as classifiers for disjoint groups of samples from a microarray dataset. We validate our method on the microarray dataset of Head and neck/Oral squamous cell carcinoma, where our boolean classifiers presented a set of gene activity/inactivity combinations that are characteristic for various cancer sub-types and normal samples. Our findings correlate well with the literature.

Keywords: TUCS technical reports, LaTeX

TUCS Laboratory
Computational Biomodeling Laboratory

1 Introduction

We develop a method for deriving boolean logic classifiers from microarray data series.

Nowadays, there has been already collected an overwhelming amount of diverse genome and molecular-scale information as well as clinical data on cancers and other genetic-related diseases. On one hand, abundance of the data helps to increase our understanding about the disease. On the other hand, large amount of unstructured data represents a great challenge to be interpreted and comprehended. Complex statistics and machine learning methods provide means to build data-driven models that can be used as classifiers that recognize biological or clinical samples to belong to certain categories, like disease or healthy cell-lines. However, in most cases those models are constructed as "black boxes" without comprehensible internal structure that can be understood. In other words, the data-driven models built by means of machine learning methods will tell what samples belong to what groups, but will not tell exactly why. The goal of our work is to provide a method that can "answer" in a simple manner why a sample should belong to a certain group. The answer will be provided in terms of Boolean logic.

Systems biology deals with understanding of the functionalities of a living cell and of deviations in cellular functions that lead to diseases. For instance, there have been many studies on constructing Boolean logic models for various biological phenomena. For example, in [6, 17, 28] there was established correspondence between Boolean networks and ODE-based models. Some data-driven Boolean logic model building methods were suggested in [1, 31].

Some of the studies from above mainly focus on approaches where the full understanding of the biological aspects of the phenomenon of interest is required. However, even though highly detailed models can provide a realistic life picture, sometimes, it can be difficult to analyze and reason about the large models. Hereby, studies in [26] aimed at obtaining a higher-abstraction level of understanding of biological systems starting from existing "larger" models. In particular, the goal of that work was in deriving a simple logical description of the activation conditions between the "key nodes" of a bio-model under study. As the result, [26] presented a method for translating a highly detailed large biological model in form of a bio-molecular (signaling pathway) pathway into a relatively small Boolean network (so-called *logicome*) representing activation relations between the key nodes as logic relations. A biological model presented in the form of the logicome should be easier to comprehend and reason about. In this article, we advance further with the idea from [26] of developing high-level comprehensible Boolean logic based models for biological phenomena. Here, we develop a simple method for deriving logical relations between key/significant elements associated to certain categories of samples from microarray gene expression data. Those relations should provide a simple explanation in terms of logical formulas in disjunctive normal form on why certain samples belong to certain categories.

We applied our method to the small set of genes derived from a set of 11 significant genes that are highly differentially expressed genes between cancer and normal cell lines. We have selected as the case studies head and neck/oral squamous cell carcinoma (HNOSCC) microarray datasets from studies in [11]. We have selected four sample categories from [11]: oral tongue squamous cell carcinoma, samples in squamous cell carcinoma of the oral cavity and oropharynx, head and neck squamous cell carcinoma, and normal cell lines. We have derived for each of these categories boolean formulas representing their characteristic patterns of gene expression profiles. We have validated our findings against well known gene expression patterns associated to head and neck cancer [7–9,20–25,27,29,30,32,34] that agree with the discovered boolean expressions associated to cancer cell lines. In the same time, we have derived a number of gene expression patterns that have not been discovered so far.

For the effectiveness of our method and in order to reduce our analysis to the small set of genes that are significant for head and neck cancer, we have identified 11 highly differentially expressed genes between cancer and normal cell lines with GEO tools [4]. In our methodology we employ multinomial logistic regression to find small subsets of genes that satisfies accuracy threshold. We proceed with these subsets and derive boolean logic classifier in distinctive normal form for each of the four categories with terms corresponding to genes from these subsets.

2 Methodology

In this section we describe methods that are used in our study. Firstly, we present a formal definition of the classification method used in this paper, i.e., *multinomial logistic regression (MLR)*, then we describe our approach for deriving a unique Boolean expression (we call it *boolean signature*) corresponding to each category of samples.

2.1 Multinomial Logistic Regression

Multinomial logistic regression (MLR) is a classification method used to measure the relationship between a category distributed dependent variable and one or more independent variables. When building an MLR model, it is assumed that the categories (clusters) are mutually exclusive, i.e., a sample belongs to exactly one cluster, for more information see [18].

2.1.1 Accuracy of The Model

For any given classifier and any given sample Y, there are four possible classification outcomes:

– if Y belongs to cluster C and it is classified as such, we denote it as a *true positive (TP)*,

– if Y belongs to cluster C and it is classified in a different cluster, we denote it as a *false negative (FN)*,

– if Y does not belong to cluster C and it is classified as such, we denote it as a *true negative (TN)*,

– if Y does not belong to cluster C and it is classified in C, we denote it as a *false positive (FP)*.

Accuracy are calculated in terms of TP, TN, FN and FP.

$$\text{Accuracy} = \frac{\text{TN + TP}}{\text{TN+TP+FN+FP}}$$

Accuracy is the proportion of true results, whether it means belonging to the right cluster or not belonging to the wrong cluster. For more detailed information we refer to [35, 36]

2.2 Inferring Boolean Signatures

In the following, we describe Algorithm 2.1 that gives minimal size subset from the set of predictor variables (genes) G and Algorithm 2.2 that is applied on the selected minimal size subset to derive a boolean signature for each category.

2.2.1 Reducing the set of predictor variables

In order to generate the signature to be as simple as possible yet accurate, our goal here is to reduce the size of G in such a way that the accuracy of MLR is not compromised.

Algorithm 2.1. Let Mod be the multinomial logistic regression (MLR) model for gene expression matrix M, A its predictive accuracy, and G the set of predictor variables and T an accuracy threshold. Let us take the following steps:

Step 1 Enumerate all $2^{|G|}$ subsets of G,

Step 2 For all $S \subseteq G$, train its corresponding MLR, M_S and calculate its predictive accuracy denoted by A_S, where $|S| \geq m_s, 1 \leq m_s \leq |G|$,

Step 3 Collect all $S_m \subseteq G$ such that $A_{S_m} = \max\{A_{S_i} \mid S_i \subseteq G \text{ and } |S_i| = m\}$

Step 4 Output(minimal size subset S_{m_l}) where $S_{m_l} \in \{S_m \mid m_s \leq m \leq |G|\}$ with $A_{S_{m_l}} \geq T$.

2.2.2 Boolean signature

In our approach the minimal size subset obtained from the Algorithm 2.1, is further analyzed to derive Boolean signature for each category.

The boolean signature is derived as follows:

Algorithm 2.2. Let MB be the binarized gene expression matrix of expression matrix M generated for subset of genes $S \subseteq G$, let $C = \{C_1, \ldots, C_k\}$ be the set of disjoint categories (clusters) of samples from M, Pr be the probability threshold and $covg$ be the coverage threshold. The probability threshold Pr for a binary values combination frequency is the lower border for combinations to be considered as "frequent". The $covg$ threshold for binary values combination frequency indicates the border below which we consider binary combinations as "insignificant". We recall here, that for each category we select its frequent (defined by Pr) significant (defined by $covg$) binary combinations that we use to derive the disjunctive normal form:

Step 1 Consider set T_S of all the binary values combinations of genes from S in MB, where $S \subseteq G$.

Step 2 Frequency of occurrence: For each combination of binary values from $c_j \in T_S$, count the number of its occurrences in every category $C_i \in C$, divide it by $|C_i|$, denote it by $N^i_{c_j}$ where $1 \leq j \leq 2^{|S|}$. Intuitively, $N^i_{c_j}$ denotes the frequency of occurrence of combination c_j in in category C_i.

Step 3 Maximal frequency of occurrence: Find $N^i_C = max\{N^i_{c_1}, N^i_{c_2} \ldots N^i_{c_{2^{|S|}}}\}$. In other words, N^i_C is the frequency of the most occurring combination in category C_i.

Step 4 Representative combinations for a category: For $C_i \in C$, $1 \leq i \leq k$, find the set
$\mathcal{C}^i = \{c_j \in C_i \mid \max(Pr * N^i_C, covg) \leq N^i_{c_j} \leq N^i_C\}$, $1 \leq j \leq |T_S|$. In other words, here we select the representative combinations for a category C_i, those combinations are significant enough ($N^i_{c_j} \geq covg$) and are frequent ($N^i_{c_j} \geq Pr * N^i_C$) in C_i.

Step 5 Deriving boolean signature: For every $c_j \in \mathcal{C}^i$, where $c_j = (b_{g1}, b_{g2}, \ldots, b_{g|S|})$ and $b_g l \in \{0, 1\}$ is a binarized expression value for a gene $gl \in S$ where $1 \leq l \leq |S|$, we construct the conjunction of gene variables associated to combination c_j as follows: $B_{ij} = (\bigwedge_{b_{gl}==1} gl) \wedge (\bigwedge_{b_{gl}==0} \neg gl)$. For the set of representative combinations $\mathcal{C}^i$ we construct the disjunctive normal form (boolean classifier, boolean signature) BC_i as follows: $BC_i = \bigvee_{c_j \in \mathcal{C}^i} B_{ij}$. I.e., BC_i is the boolean signature of category C_i in the disjunctive normal form.

Step 6 OUTPUT: Output (C_i, B_i), for every $1 \leq i \leq k$.

The outline of our methodology is presented in in the Figure 1.

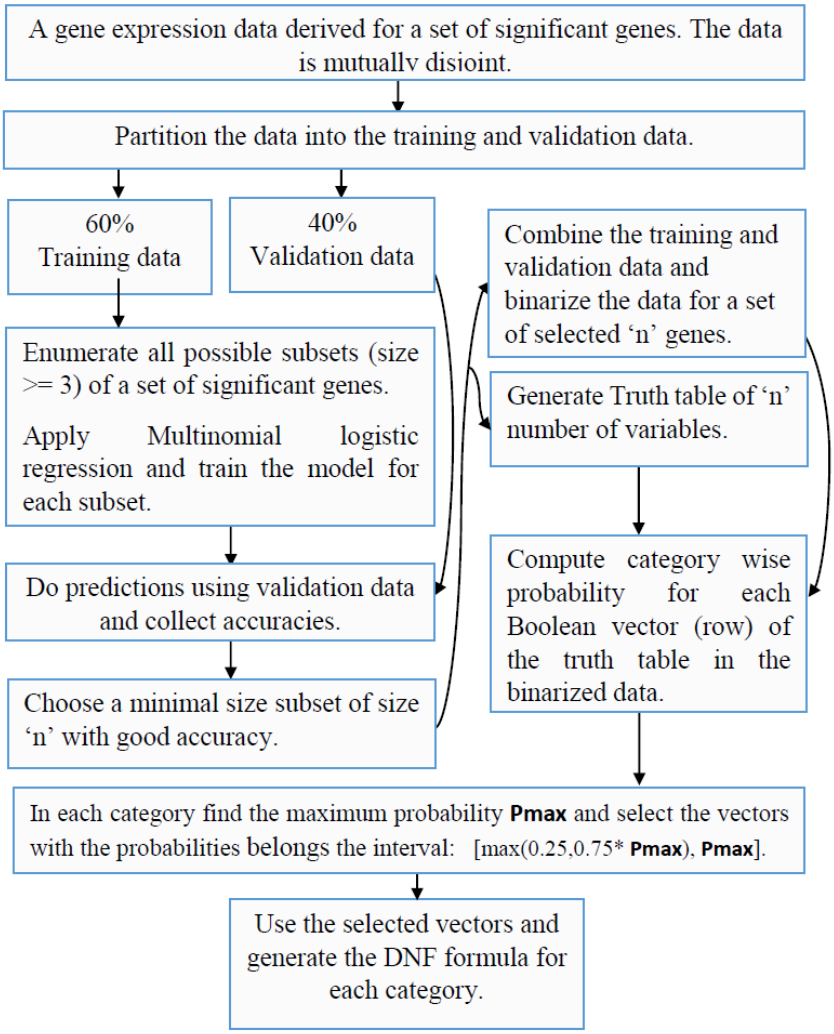

Figure 1: Outline of the methodology

3 Case studies

We use nine microarray data series of head and neck/oral squamous cell carcinoma (HNOSCC) from studies in [11] that are available at Gene Expression Omnibus (GEO) database [2–4, 14]. We explain the preprocessing of the microarray data and obtaining the gene expression matrix, that we use to derive boolean expressions associated with various categories of HNOSCC and with non-tumor cells.

3.1 Samples

Studies in [11] consider 9 dataseries from GEO database [2–4, 14] with 675 samples in total: *GSE6791*, *GSE9844*, *GSE30784*, *GSE31056*, *GSE2379*, *GSE3524*, *GSE6631*, *GSE13601* and *GSE23036*. In [11] due to an unsupervised learning method the samples were split in 22 categories out of which we have selected the following 4 categories with 509 samples for our studies: 58 samples in oral tongue squamous cell carcinoma (OTSCC), 189 samples in squamous cell carcinoma of the oral cavity and oropharynx (OSCC), 98 samples in head and neck squamous cell carcinoma (HNSCC), and 164 normal/control samples.

3.2 Microarray data

We get the microarray data from GEO in form of normalized probe signals as sample data matrices (probe expression matrices). In a sample data matrix, the rows correspond to probes and the columns correspond to samples. The data series that we have selected for our studies contain genome-wide gene expression profiling of head and neck/oral squamous cell carcinoma (HNOSCC) that was measured by Affymetrix platform [10, 15]. The probe signals in these series were normalized through Robust Multi-array Average technique, [19], GeneChip RMA (GCRMA) [33], and Microarray Suite version 5.0 (MAS 5.0, Affymetrix, Inc.), [5].

3.3 Data Preprocessing

For our case study, we performed differencial gene expression analysis on the 9 dataseries and selected 11 "significant genes" for further analysis. We have generated gene expression matrix for the "significant genes".

3.3.1 Gene expression matrix.

We have processed the selected GEO data series in our study by using *Bioconductor GEO Query R Library* [10, 16]. We have transformed normalized probe measurements into gene expression levels as follows:

- We have found mappings between sets of probes and their associated genes for the respective affimetrix platforms by using an online web-tool *DAVID* (Database for Annotation, Visualization, and Integrated Discovery) [12], *GPL* (GEO platform record) [4] and *Affymetrix Human Genome U133 2.0 Array* annotation data (*hgu133plus2*).
- We have considered the expression level for a gene to be the median of the gene's associated probes.

In the result, we have generated the gene expression matrix of approximately 25000 rows represented by gene symbols and columns by samples.

3.3.2 Selecting a set of differentially expressed genes.

We have selected differentially expressed probes between control and all cancer samples for each data series separately by means of *GEO2R* web-tool [4] (URL: `https://www.ncbi.nlm.nih.gov/geo/info/geo2r.html`). The differential gene/probe expression analysis was performed as in [13]. The result of differential gene/probe expression analysis for each data series is collected as a table of probes ranked by their p-values. We have selected those probes that satisfy the threshold of p-value ≤ 0.05. Then, for each dataset, we have selected those genes that correspond to the probes with the p-value not exceeding the threschold 0.05. We call these selected genes "significant genes". Finally, we took the intersection of significant genes from all the datasets and considered it for our boolean signatures construction method. In particular, in the intersection we have the following significant genes: *MAL, LAPTM4B, HPGD, KRT4, EXT1, AIM1, SPP1, MYO10, MYO1B, MMP1, MGST2*.

3.3.3 Rescaling datasets gene expressions to the same level.

We extract the rows of these significant genes from the gene expressions matrix that we have described above. We rescale gene expression values in the matrix in order to bring expression levels for each gene for all samples from different datasets to the same range. For each dataset D separately, for each sample S from D and for every gene G, we have rescaled its expression value $x_{G,S}$ associated to S to $z_{G,S}$ as follows:

$$z_{G,S} = \frac{x_{G,S} - m_{G,D}}{R_{G,D}},$$

where $m_{G,D}$ is the mean value of expressions of gene G for all the samples from dataset D and $R_{G,D}$ is the range of expressions of gene G among all the samples of D. I.e., $R_{G,D} = MAX_{G,D} - MIN_{G,D}$, where $MAX_{G,D}$ ($MIN_{G,D}$) is the highest (the lowest, respectively) expression level for a gene G among all the samples from dataset D.

3.3.4 Removing similar samples between different categories.

The goal of our methodology is to generate unique "boolean signatures" for different sample categories. Hereby, in order to increase the accuracy of our method, we need to make sure that the samples in the categories are "different enough". We filter out those samples that have "near identical" gene expression profiles but belong to different categories. We regard samples as vectors defined by their corresponding gene expression profiles and employ euclidian distnace in vector space as the closeness measurement between samples. We define a minimal distance threschold ϵ under which we consider samples to be "near identical". We calculate ϵ as follows:

$$\epsilon = C \times MAX_{NORM},$$

where MAX_{NORM} is a maximal norm among all the vectors in our studies, and C is a constant, which we have fixed in our studies to be $C = 0.01$. We did not find any "near identical" samples between different categories according to this criteria in our gene expression matrix.

The processed data are available at Github (`https://github.com/cpanchal/Dataset_Logicome.git`)

3.4 Data analysis and results

We analysed the preprocessed gene expression data extracted for the significant genes as follows:

- Randomly partitioned the data into training and validation set with the ratio of $60 : 40$.
- Enumerated all the possible subsets (size ≥ 3) of a set of significant genes and extracted the data for each subset. We trained the MLR model on this data and collected the predictive accuracies using the validation data.
- Collected the subset of genes with maximum accuracies from the subsets of each size.
- From the collected subsets, picked a minimal size subset with accuracy $\geq 70\%$. That step rendered us a subset of genes of size 3.
- We derived "boolean identifiers (classifiers)" in the disjunctive normal form for each category in terms of the selected minimal size subset of genes.

In our method the training and validation data are partitioned randomly, hence we re-ran the algorithm 20 times and the results are collected for each run. The rerunning of the algorithm yields different results or the result could be repeated. The boolean signatures derived using the resultant subsets obtained in the different run, are listed in the Table 1.

The boolean formulas in the Table 1 identify the categories Normal, OSCC, OTSCC and HNSCC with different combination of genes. The down-regulated gene is denoted with '/' and genes without '/' are upregulated. We can see in the second boolean formulation, all the cancer categories are identified by the same formula and in all cancer the gene KRT4 is down-regulated, MAL is down-regulated and MMP1 is up-regulated. Whereas, for normal those genes are with opposite regulations. These findings are inline with the studies reported in [23, 32], also these studies identify KRT4, MAL and MMP1 as most predictive bio-markers for squamous cell carcinoma in cancer or normal samples. This boolean formulation clearly differentiates between cancer and normal samples.

In our experiment, the upregulated gene MMP1 is found in four boolean formulations and it is appeared as a potential bio-marker. This observation is inline

with the study reported on cancer-specific genes in [9, 21, 23] where MMP1 is considered as one of the promising and relevant genes for the study of HNOSCC tumor cells. We verify the insights presented in Table 1 with the findings reported from the different studies incorporated with gene expression profiling on head and neck/ oral squamous cell carcinomas. Our observations on genes (MMP1,SPP1) being up-regulated and genes (AIM1,KRT4,MAL) being down-regulated in the cancer categories are consistent with study reported in [8]. Another finding that confirms the AIM1 as downregulated in HNSCC is mentioned in [20]. The result in [29] represents the genes (MYO10, MYO1B) and HPGD belongs to the top up-regulated and down-regulated genes repectively in OTSCC which is inline with our findings. Also MGST2 is down regulated in the category OSCC that is confirmed in the studies [27] and [24]. We identify MYO1B as up-regulated in the categories OSCC and HNSCC, and it is confirmed by the studies [7] and [25] where MYO1B is considered as the most up-regulated genes in the same categories. In our results HPGD is down-regulated gene in the categories OSCC and OTSCC, this finding agrees with the results presented in [30] and [29]. The work reported in [22] and [34], find the genes (KRT4, HPGD, MAL) amongs consistently down-regulated genes in HNSCC and OTSCC respectively, the same behavior we observe in our results also.

Besides the results reported in the Table 1 we discover some combinations of genes for categories OSCC and OTSCC that are remains to be validated from the literature. We discover for OSCC in the boolean formulation for the 1st subset in the Table 1, the possibility for the gene MYO10 to be down regulated. Similar observation for OSCC is found in the boolean formulation for the 3rd subset where the genes SPP1 and AIM1 are found down-regulated and up-regulated respectively. Moreover in the boolean formulation for the 5th subset in the Table 1, the result shows the possibility of both genes MGST2 and AIM1 to be up-regulated for OTSCC.

4 Discussion

Here, we propose a continuation of the direction initiated in [26], where logicome building methods are suggested to be a companion to the bottom-up modeling approaches. In [26], the authors suggested a way to generate a higher-level representation of a network model in terms of boolean logic relations between key nodes of the network. That logicome approach should allow the modeler to concentrate on a selected set of significant network nodes and relations between them, while abstracting from the rest of the model. Also, due to the fact that machine learning and statistics approaches usually do not provide information about the internal structure of the system under studies and relations between its components, but, rather act as ”oracles” generating predictions and classifications, we decided to come up with a method that would capture most representative patterns in the input

No.	Subset	Boolean formula
1	(KRT4, MYO10, HPGD)	$Normal = KRT4 * HPGD$ $OSCC = KRT4' * (MYO10' + HPGD')$ $OTSCC = KRT4' * MYO10 * HPGD'$ $HNSCC = MYO10 * HPGD'$
2	(KRT4, MAL, MMP1)	$Normal = KRT4 * MAL * MMP1'$ $OSCC = KRT4' * MAL' * MMP1$ $OTSCC = KRT4' * MAL' * MMP1$ $HNSCC = KRT4' * MAL' * MMP1$
3	(AIM1, SPP1, MMP1)	$Normal = AIM1 * SPP1' * MMP1'$ $OSCC = MMP1 * (AIM1' * SPP1 + AIM1 * SPP1')$ $OTSCC = MMP1 * (AIM1' * SPP1 + AIM1 * SPP1')$ $HNSCC = AIM1' * SPP1$
4	(KRT4, HPGD, SPP1)	$Normal = KRT4 * HPGD * SPP1'$ $OSCC = KRT4' * HPGD' * SPP1$ $OTSCC = HPGD' * SPP1'$ $HNSCC = HPGD' * SPP1$
5	(MGST2, AIM1, MMP1)	$Normal = MGST2 * AIM1 * MMP1'$ $OSCC = MGST2' * MMP1$ $OTSCC = MMP1 * (MGST2 * AIM1 + MGST2' * AIM1')$ $HNSCC = MGST2' * AIM1' * MMP1$
6	(HPGD, MYO1B, MMP1)	$Normal = HPGD * MYO1B' * MMP1'$ $OSCC = HPGD' * MYO1B * MMP1$ $OTSCC = HPGD' * MMP1$ $HNSCC = HPGD' * MYO1B * MMP1$

Table 1: Subsets and boolean formulations for each category: '$*$' denotes conjunction, '$+$' denotes disjunction and '$'$' denotes complement.

datasets for each of the clusters/categories and generate small and simple boolean classifiers for them.

We present here a simple method for deriving boolean classifiers (signatures) for all the categories of samples as small boolean expressions in disjunctive normal form. Those signatures represent most occurring patterns in the respective sample categories and can be based on to reason further about the properties of each category. In the same time, our modeling method is not meant for deriving highly detailed models from microarray data that can be used for accurate simulations. We rather suggest here a way to understand better the observed data in simple terms, that can aid in further efforts of building accurate complex models for the phenomena under studies.

Acknowledgments

This work was partially funded by the Academy of Finland (project 267915), the Otto A. Malm foundation and the Oskar Öflunds stiftelse foundation. We thank Professor Ion Petre (Computational Biomodeling Laboratory, Åbo Akademi University, Turku, Finland) for his instructions in the project and Dr. Sepinoud Azimi (Computational Biomodeling Laboratory, Åbo Akademi University and Turku Centre for Computer Science) for her suggestions to the methodology of this study.

References

[1] Martin Anthony and Peter L Hammer. A boolean measure of similarity. *Discrete Applied Mathematics*, 154(16):2242–2246, 2006.

[2] Tanya Barrett, Tugba O Suzek, Dennis B Troup, Stephen E Wilhite, Wing-Chi Ngau, Pierre Ledoux, Dmitry Rudnev, Alex E Lash, Wataru Fujibuchi, and Ron Edgar. NCBI GEO: mining millions of expression profilesdatabase and tools. *Nucleic acids research*, 33(suppl 1):D562–D566, 2005.

[3] Tanya Barrett, Dennis B Troup, Stephen E Wilhite, Pierre Ledoux, Dmitry Rudnev, Carlos Evangelista, Irene F Kim, Alexandra Soboleva, Maxim Tomashevsky, and Ron Edgar. NCBI GEO: mining tens of millions of expression profilesdatabase and tools update. *Nucleic acids research*, 35(suppl 1):D760–D765, 2007.

[4] Tanya Barrett, Stephen E Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F Kim, Maxim Tomashevsky, Kimberly A Marshall, Katherine H Phillippy, Patti M Sherman, Michelle Holko, et al. NCBI GEO: archive for functional genomics data setsupdate. *Nucleic acids research*, 41(D1):D991–D995, 2013.

[5] John A Berger, Sampsa Hautaniemi, Anna-Kaarina Järvinen, Henrik Edgren, Sanjit K Mitra, and Jaakko Astola. Optimized lowess normalization parameter selection for dna microarray data. *BMC bioinformatics*, 5(1):194, 2004.

[6] M. Chaves, E.D. Sontag, and R. Albert. Methods of robustness analysis for boolean models of gene control networks. *IEE Proceedings - Systems Biology*, 153(4):154, 2006.

[7] Chu Chen, Eduardo Méndez, John Houck, Wenhong Fan, Pawadee Lohavanichbutr, Dave Doody, Bevan Yueh, Neal D Futran, Melissa Upton, D Gregory Farwell, et al. Gene expression profiling identifies genes predictive of oral squamous cell carcinoma. *Cancer Epidemiology and Prevention Biomarkers*, 17(8):2152–2162, 2008.

[8] Peter Choi and Chu Chen. Genetic expression profiles and biologic pathway alterations in head and neck squamous cell carcinoma. *Cancer*, 104(6):1113–1128, 2005.

[9] Kiran Dahiya and Rakesh Dhankhar. Updated overview of current biomarkers in head and neck carcinoma. *World journal of methodology*, 6(1):77, 2016.

[10] Sean Davis and Paul S Meltzer. GEOquery: a bridge between the gene expression omnibus (GEO) and bioconductor. *Bioinformatics*, 23(14):1846–1847, 2007.

[11] Loris De Cecco, Monica Nicolau, Marco Giannoccaro, Maria Grazia Daidone, Paolo Bossi, Laura Locati, Lisa Licitra, and Silvana Canevari. Head and neck cancer subtypes with biological and clinical relevance: Meta-analysis of gene-expression data. *Oncotarget*, 6(11):9627–42, 2015.

[12] Glynn Dennis, Brad T Sherman, Douglas A Hosack, Jun Yang, Wei Gao, H Clifford Lane, and Richard A Lempicki. DAVID: database for annotation, visualization, and integrated discovery. *Genome biology*, 4(9):1, 2003.

[13] Friederike Dündar, Luce Skrabanek, and Paul Zumbo. Introduction to differential gene expression analysis using RNA-seq. 2015.

[14] Ron Edgar, Michael Domrachev, and Alex E Lash. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic acids research*, 30(1):207–210, 2002.

[15] Laurent Gautier, Morten Møller, Lennart Friis-Hansen, and Steen Knudsen. Alternative mapping of probes to genes for affymetrix chips. *BMC bioinformatics*, 5(1):1, 2004.

[16] Robert C Gentleman, Vincent J Carey, Douglas M Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, Laurent Gautier, Yongchao Ge, Jeff Gentry, et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, 5(10):1, 2004.

[17] Leon Glass and Stuart A. Kauffman. The logical analysis of continuous, non-linear biochemical control networks. *Journal of Theoretical Biology*, 39(1):103–129, apr 1973.

[18] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*, volume 398. John Wiley & Sons, 2013.

[19] R. A. Irizarry. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264, apr 2003.

[20] Geoung A Jeon, Ju-Seog Lee, Vyomesh Patel, J Silvio Gutkind, Snorri S Thorgeirsson, Eun Cheol Kim, In-Sun Chu, Panomwat Amornphimoltham, and Myung Hee Park. Global gene expression profiles of human head and neck squamous carcinoma cell lines. *International journal of cancer*, 112(2):249–258, 2004.

[21] Ki-Yeol Kim, Xianglan Zhang, and In-Ho Cha. Identification of human papillomavirus status specific biomarker in head and neck cancer. *Head & neck*, 37(9):1310–1318, 2015.

[22] MA Kuriakose, WT Chen, ZM He, AG Sikora, P Zhang, ZY Zhang, WL Qiu, DF Hsu, C McMunn-Coffran, SM Brown, et al. Selection and validation of differentially expressed genes in head and neck cancer. *Cellular and molecular life sciences*, 61(11):1372–1383, 2004.

[23] Benjamin Lallemant, Alexandre Evrard, Christophe Combescure, Heliette Chapuis, Guillaume Chambon, Caroline Raynal, Christophe Reynaud, Omar Sabra, Dominique Joubert, Frédéric Hollande, et al. Clinical relevance of nine transcriptional molecular markers for the diagnosis of head and neck squamous cell carcinoma in tissue and saliva rinse. *BMC cancer*, 9(1):370, 2009.

[24] Jin-zhong Li, Hong-ya Pan, Jia-wei Zheng, Xiao-jian Zhou, Ping Zhang, Wantao Chen, and Zhi-yuan Zhang. Benzo (a) pyrene induced tumorigenesity of human immortalized oral epithelial cells: transcription profiling. *Chinese Medical Journal (English Edition)*, 121(19):1882, 2008.

[25] Gaku Ohmura, Takahiro Tsujikawa, Tomonori Yaguchi, Naoshi Kawamura, Shuji Mikami, Juri Sugiyama, Kenta Nakamura, Asuka Kobayashi, Takashi Iwata, Hiroshi Nakano, et al. Aberrant myosin 1b expression promotes cell migration and lymph node metastasis of hnscc. *Molecular Cancer Research*, 13(4):721–731, 2015.

[26] Charmi Panchal, Sepinoud Azimi, and Ion Petre. Generating the logicome of a biological network. In *Algorithms for Computational Biology*, pages 38–49. Springer Nature, 2016.

[27] A Shukla, A Singh, and R Srivastava. Oral submucous fibrosis: an update on etiology and pathogenesis-a review. *Rama Univ J Dent Sci*, 2:24–33, 2015.

[28] Claudia Sttzel, Susanna Rblitz, and Heike Siebert. Complementing ODE-based system analysis using boolean networks derived from an euler-like transformation. *PLOS ONE*, 10(10):e0140954, oct 2015.

[29] Soundara Viveka Thangaraj, Vidyarani Shyamsundar, Arvind Krishnamurthy, Pratibha Ramani, Kumaresan Ganesan, Muthulakshmi Muthuswami, and Vijayalakshmi Ramshankar. Molecular portrait of oral tongue squamous cell carcinoma shown by integrative meta-analysis of expression profiles with validations. *PloS one*, 11(6):e0156582, 2016.

[30] Gokce A Toruner, Celal Ulger, Mualla Alkan, Anthony T Galante, Joseph Rinaggio, Randall Wilk, Bin Tian, Patricia Soteropoulos, Meera R Hameed, Marvin N Schwalb, et al. Association between gene expression profile and tumor invasion in oral squamous cell carcinoma. *Cancer genetics and cytogenetics*, 154(1):27–35, 2004.

[31] S. A. Vinterbo, E.-Y. Kim, and L. Ohno-Machado. Small, fuzzy and interpretable gene expression based classifiers. *Bioinformatics*, 21(9):1964–1970, jan 2005.

[32] Mark E Whipple, Eduardo Mendez, D Gregory Farwell, S Nicholas Agoff, and Chu Chen. A genomic predictor of oral squamous cell carcinoma. *The Laryngoscope*, 114(8):1346–1354, 2004.

[33] Zhijin Jean Wu and Rafael Irizarry. Description of gcrma package. 2010.

[34] Hui Ye, Tianwei Yu, Stephane Temam, Barry L Ziober, Jianguang Wang, Joel L Schwartz, Li Mao, David T Wong, and Xiaofeng Zhou. Transcriptomic dissection of tongue squamous cell carcinoma. *BMC genomics*, 9(1):69, 2008.

[35] Wen Zhu, Nancy Zeng, Ning Wang, et al. Sensitivity, specificity, accuracy, associated confidence interval and roc analysis with practical sas® implementations. *NESUG proceedings: health care and life sciences, Baltimore, Maryland*, pages 1–9, 2010.

[36] Mark H Zweig and Gregory Campbell. Receiver-operating characteristic (roc) plots: a fundamental evaluation tool in clinical medicine. *Clinical chemistry*, 39(4):561–577, 1993.

Turku
Centre *for*
Computer
Science

Joukahaisenkatu 3-5 A, 20520 TURKU, Finland | www.tucs.fi

University of Turku
Faculty of Mathematics and Natural Sciences
- Department of Information Technology
- Department of Mathematics and Statistics

Turku School of Economics
- Institute of Information Systems Sciences

Åbo Akademi University
- Computer Science
- Computer Engineering

ISBN XXX-XXX-XXX-X
ISSN 1239-1891