Weighted Gene Co-expression Network Analysis Applied to Head and Neck Squamous Cell Carcinoma Data

Fernanda Correia Barbosa[1,3], Joel P. Arrais[2], and José Luís Oliveira[1]

[1] Department of Electronics, Telecommunications and Informatics (DETI), Institute of Electronics and Telematics Engineering of Aveiro (IEETA), University of Aveiro, Aveiro, Portugal

[2] Department of Informatics Engineering (DEI), Centre for Informatics and Systems of the University of Coimbra (CISUC), University of Coimbra, Coimbra, Portugal

[3] Department of Informatics and Systems Engineering (DEIS), Engineering Institute of Coimbra (ISEC), Polytechnic Institute of Coimbra, Coimbra, Portugal

Abstract— **Microarray technology has made possible the simultaneous monitoring of the expression levels of thousands of genes under multiple disease states. Due to the high complexity of the obtained data the use of computational methods for extracting biological evidences is still a major issue. In this work we address this problem by adjusting the use of gene co-expression networks to analyze a Head and Neck Squamous Cell Carcinoma (HNSCC) dataset. The proposed method applies hierarchic clustering to identify gene modules using the topological overlap dissimilarity measure after, defining a gene co-expression similarity, defining a family of adjacency functions and calculating their parameters. This method calculates the eigengenes of each module to define a network of modules and the correlation between the eigengenes and the risk factors, identifying modules of genes where those are more expressed and associating these concepts to gene ontology functional terms. The preliminary results described in this paper contribute to reveal the molecular mechanisms associated with HNSCC and the contribution of experimental factors types like differentiation, alcohol use, sex, age, tumor site, smoking pack years and race.**

Keywords— **gene expression, co-expression network, head and neck cancer.**

I. INTRODUCTION

Head and Neck Squamous Cell Carcinoma (HNSCC) is the sixth most common cancer worldwide, affecting 600,000 new patients each year [1]. Several risk factors such as smoking habits, alcohol use, and human papillomavirus infection have already been documented as having a very high correlation with this type of cancer [2, 3]. Despite that, still lacks a full comprehension of genomic processes that are associated with HNSCC and more importantly the individual contribution of each of these factors, when crossed with epidemiologic characteristics and the existence of other risk factors associated with this disease.

While techniques such as microarray experiment evaluates a large number of genomic sequences (genes), under multiple conditions (samples) [4], the traditional computational approaches for extracting evidences from the data are cumbersome and most of the times lead to inconclusive results.

One promising approach consists in the use of gene co-expression networks to study gene expression data, helping in the extraction of structural and functional features that can be used to better understand the data. The followed approach to analyze expression data using weighted gene co-expression networks includes the following steps [5]: definition of a gene co-expression similarity, definition of a family of adjacency functions, determination of the adjacency functions parameters, identification of the network modules using clustering, association to network concepts and association these concepts to external gene or sample information.

This paper shows the preliminary results of a study that aims to contribute to reveal the molecular mechanisms associated with HNSCC and the contribution of other risk factors besides smoking habits and alcohol use, like differentiation, sex, age, tumor site and race with a major focus in the age and alcohol use experimental factor types.

II. METHODS

A. Dataset Construction

The dataset was downloaded from the public microarray gene expression database ArrayExpress [6, 7], from the investigation E-GEOD-39366 - Molecular Subtypes in Head and Neck Cancer [expression].

A total of 138 tumor arrays were considered from the 163 samples, after removing low-quality and duplicate arrays, and arrays from non-HNSCC samples. Probes produced expression values for 15,595 genes.

Database for Annotation, Visualization, and Integrated Discovery bioinformatics resources (DAVID) is an integrated biological knowledgebase and data mining tools. It is used to extract biological meaning from large lists of genes or proteins, like gene ontology functional terms [8, 9].

B. Gene Co-expression Network

Co-expression network construction from microarray data uses correlation analysis to build the correlation matrix, which is converted to an adjacency matrix representing the

I. Lacković et al. (eds.), *The International Conference on Health Informatics*, IFMBE Proceedings 42,
DOI: 10.1007/978-3-319-03005-0_76, © Springer International Publishing Switzerland 2014

co-expression network. Each gene corresponds to a node and two genes are connected by an edge if their expression values are highly correlated.

Many real networks have been found to have approximately scale free topologies with the associated topological properties presented. An example is the study made with protein-protein interaction networks obtained from the human oral proteome [10]. Networks whose scale free topology index R^2 is close to 1 are said to be approximately scale free.

A co-expression network can be represented by a symmetric adjacency matrix, $A = [a_{ij}]$ with values in $[0,1]$. For weighted networks, the adjacency matrix returns the connection strength between gene pairs and as gene co-expression similarity measure can be used the absolute value of the Pearson product moment correlation to relate every pairwise gene–gene relationship

$$a_{ij} = \left|cor\left(x_i, x_j\right)\right| \qquad (1)$$

An adjacency function can be used to transform the original network into a new network. For the construction of weighted gene co-expression networks [5], the adjacency matrix is constructed using a "soft" power adjacency function a_{ij}, where for an unsigned network

$$a_{ij} = \left|cor\left(x_i, x_j\right)\right|^{\beta} \qquad (2)$$

A choice of a power $\beta > 1$ is used to emphasize large adjacencies at the expense of low ones. To choose the parameter value β is used the scale free topology criterion, being β the value obtained through the trade-off between the lowest integer such that the resulting network satisfies approximate scale-free topology (linear model fitting index R^2 of the regression line between $log(p(k))$ and $log(k)$ larger than 0.8) with the highest mean number of connections (high power for detecting modules, clusters of genes and hub genes).

It can also be defined the gene significance (GS) based on a microarray sample risk factor, defining gene significance measure as a function GS that assigns a nonnegative number to each gene. The higher GS_i the more biologically significant is gene i. Risk factor based gene significance is defined as (the absolute value of) the correlation between the gene and the risk factor.

Two network connectivity measures can be defined: the whole-network connectivity (including the whole gene network) and the intra-modular connectivity (including the genes of a particular module) as

$$k_i = \textstyle\sum_j a_{ij} \qquad (3)$$

Modules in weighted gene co-expression network are groups of highly correlated genes with high topological overlap [11]. A pair of genes is said to have high topological overlap if they are both strongly connected to the same group of genes. The use of topological overlap is a filter to exclude very weak connections during network construction. The topological overlap matrix (TOM) transformation can lead to a more robust network and larger modules

$$TOM_{ij} = \frac{\sum_u a_{iu}a_{uj} + a_{ij}}{min\left(k_i, k_j\right) + 1 - a_{ij}} \qquad (4)$$

TOM_{ij} is a value in $[0,1]$ and $TOM_{ij} = TOM_{ji}$.

Module significance is the mean gene significance of the module genes.

The module eigengene [12] summarizes the expression profiles in a module and is the first principal component of the Singular Value Decomposition (SVD) of the module expression matrix. It is the most highly connected intra modular hub gene and allows treating modules as single units.

The module membership (MM) is defined as the correlation of the module eigengene and the gene expression profile, allowing the quantification of the similarity of all array genes to every module.

Modules can be found using hierarchical clustering. Hierarchical clustering takes a dissimilarity measure as input. The topological overlap based dissimilarity measure is

$$DissTOM_{ij} = 1 - TOM_{ij} \qquad (5)$$

The modules are the branches of the resulting hierarchical clustering tree (dendrogram), which can be selected manually using a constant height cut-off value or using an algorithm for the selection of the height cut-off value, like the Dynamic Tree Cutting algorithm that adaptively chooses cutting values depending on the shape of the branches.

The dissimilarity of two modules q_1 and q_2 can be calculated by

$$diss(q_1, q_2) = 1 - cor\left(E^{\{(q_1)\}}, E^{\{(q_2)\}}\right) \qquad (6)$$

and the eigengene network can be defined as the signed correlation network

$$A_{\{q_1,q_2\}} = 0.5 + 0.5\, cor\left(E^{\{(q_1)\}}, E^{\{(q_2)\}}\right) \qquad (7)$$

III. RESULTS

Weighted gene co-expression analysis was applied to the expression dataset of 138 tumor arrays with expression values of 15,595 genes from the investigation Molecular Subtypes in Head and Neck Cancer from the ArrayExpress database [6]. Experimental factor types considered are: differentiation, alcohol use, sex, age, tumor site, smoking pack years and race with a major focus in age and alcohol use experimental factor types.

Scale free topology criterion was used to choose the power β for the unsigned weighted correlation network and it was chosen $\beta = 5$. The scale free topology plot of the weighted head-neck co-expression network constructed with power $\beta = 5$ satisfies a scale free topology approximately with $R^2 = 0.96$, a value close to 1. It was defined a topological overlap matrix using (5) and constructed a hierarchical tree (average linkage) to define modules as branches of the tree. Eigengenes for each module were calculated and a network among modules was defined, where each node of the network correspond to a module (Fig 1). It was constructed a hierarchical clustering dendrogram of the eigengenes E based on (6) and a heat map to visualize the eigengene network defined by the signed correlation network (7). Modules highly correlated are similar and can be merged (Fig. 1).

Multidimensional scaling can be used to visualize pairwise relationships specified by a dissimilarity matrix, where each row of the dissimilarity matrix is a point in a Euclidean space and the Euclidean distances between a pair of points reflect the corresponding pairwise dissimilarity. The input is the Tom dissimilarity and each dot is colored by the corresponding module assignment (Fig. 2). Colors from each module are well separated, showing distinct modules.

To identify modules associated with the risk factors and because each eigengene is a summary of the expression profiles of the respective module, eigengenes and risk factors were correlated. Each row corresponds to a module eigengene, and each column to a risk factor. Each cell contains the corresponding correlation and p-value. The table is color-coded by correlation according to the color legend. Age is more correlated with the magenta, black, green and light green modules and alcohol use with blue, light cyan, tan and pink modules (Fig. 3). Two different experimental factors are correlated with different modules (different genes) in this type of cancer.

The correlations between age an alcohol use and the respective module eigengenes can be measured using gene significance (GS) and module membership (MM) to identify genes with high significance for age and alcohol use and high module memberships in the identified modules (Fig. 4).

Gene ontology analysis was performed using DAVID [8], but needs to be further developed. For the modules black and green, two of the modules more correlated with age, the results obtained are for the black module: tyrosine kinase, non-receptor, 2; peptide YY, 2 (seminal plasmin); and oxytocin, prepropeptide (considering the three correlation higher values with the age risk factor, respectively: 0.233; 0.190; 0.160) and for the green module: family with sequence similarity 89, member A; hypothetical protein LOC100134229; and Rap guanine nucleotide exchange factor (GEF) 3 (considering the three correlation higher values with the age risk factor, respectively: 0.218; 0.216; and 0.215).

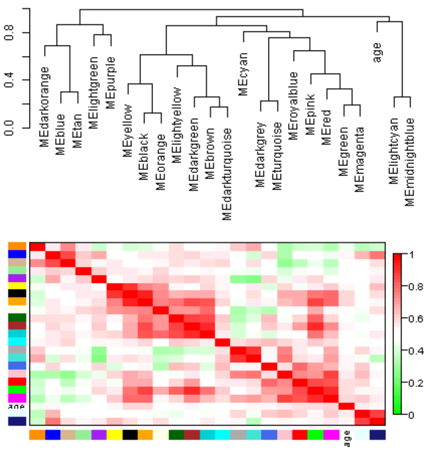

Fig. 1: Visualization of the eigengene network representing the relationships among the modules and the age and alcohol use

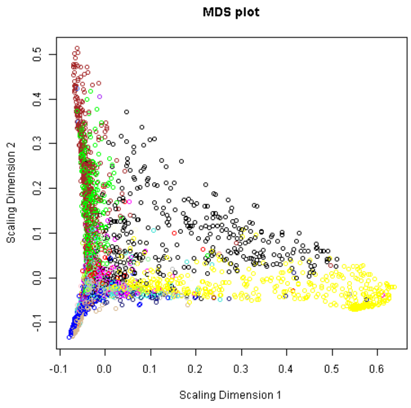

Fig. 2: Multidimensional scaling

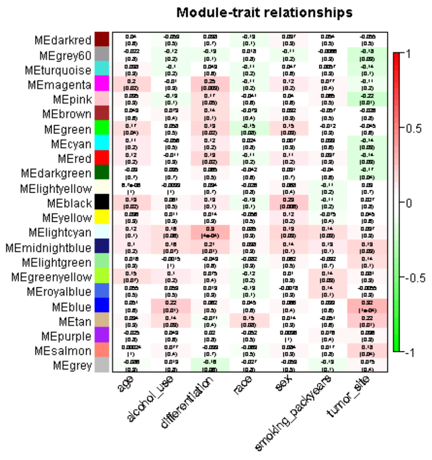

Fig. 3: Module-risk factor associations

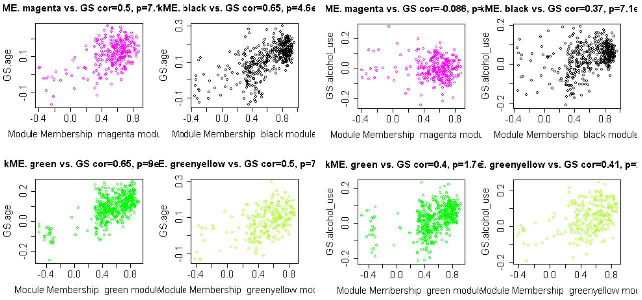

Fig. 4: Gene significance versus module membership for the risk factor age and alcohol use

IV. Conclusions

Gene expression profiles across samples can be highly correlated [5]. Gene co-expression networks were defined as weighted correlation networks, to preserve the continuous nature of the co-expression information, where strong correlations were privileged to weak correlations to minimize noise and due to the small number of samples compared to the number of genes. The quantitative microarray sample risk factor was used to define risk factor based gene significance measure.

This methodology allows the identification of distinct modules (Fig. 2). Co-expression modules are summaries of interdependencies, through the modules eigengenes.

Correlations between risk factors and HNSCC gene expression data modules were quantified, but some physiological risk factors, like race, showed no correlation with HNSCC. The analysis for this disease was mainly focused in the risk factors age and alcohol use, which were more correlated with different sets of modules from the HNSCC gene expression dataset (Fig. 3).

A preliminary gene ontology analysis, obtained functions for the genes of the modules identified and here were listed as an example the functions associated with genes with the three correlation higher values with two of the modules more correlated with the risk factor age.

Acknowledgment

This Research Unit is funded by FEDER through the Operational Program Competitiveness Factors - COMPETE and by National Funds through FCT - Foundation for Science and Technology, in the context of the projects FCOMP-01-0124-FEDER-022682 (FCT reference PEst-C/EEI/UI0127/2011) and Incentivo/EEI/UI0127/2013.

Conflict of Interest

The authors declare that they have no conflict of interest.

References

1. Rothenberg S., Ellisen W. (2012) The molecular pathogenesis of head and neck squamous cell carcinoma. The Journal of Clinical Investigation 122(6):1951–1957 DOI 10.1172/JCI59889
2. Hashibe M. et al. (2007) Interaction between tobacco and alcohol use and the risk of head and neck cancer: pooled analysis in the INHANCE consortium. Cancer Epidemiol Biomarkers Prev. 18(2): 541–550 DOI 10.1158/1055-9965.EPI-08-0347
3. Ragin C., Modugno F, Gollin S. (2007) The epidemiology and risk factors of head and neck cancer: a focus on human papillomavirus. Journal of Dental Research 86(2): 104-114 DOI 10.1177/154405910708600202
4. Nagi S., Bhattacharyya, D.K., Kalita, J.K. (2011) Gene Expression Data Clustering Analysis: A Survey, NCETACS Proc., 2nd National Conference on Emerging Trends and Applications in Computer Science, Meghalaya, India, 2011, pp 1-12 DOI 10.1109/NCETACS.2011.5751377
5. Zhang B., Horvath S. (2005) A General Framework for Weighted Gene Co-Expression Network Analysis. Statistical Applications in Genetics and Molecular Biology 4(1): 1544-6115 DOI 10.2202/1544-6115.1128
6. ArrayExpress at http://www.ebi.ac.uk/arrayexpress/
7. Brazma A. et al. (2003) ArrayExpress--a public repository for microarray gene expression data at the EBI. Nucleic Acids Research 31(1): 68–71 DOI 10.1093/nar/gkg091
8. DAVID at http://www.DAVID.niaid.nih.gov
9. Huang D. W., Sherman B., Lempicki R. (2008) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature Protocols 4: 44 – 57 DOI 10.1038/nprot.2008.211
10. Barbosa F., Arrais J., Oliveira J. L., et al. (2013) 7th International Conference on PACBB: Quantitative Characterization of Protein Networks of the Oral Cavity, Springer International Publishing, vol. 222: 61-68 DOI 10.1007/978-3-319-00578-2_9
11. Dong J., Horvath S. (2007) Understanding Network Concepts in Modules. BMC Systems Biology 1:24 DOI 10.1186/1752-0509-1-24
12. Langfelder P., Horvath S. (2007) Eigengene networks for studying the relationships between co-expression modules. BMC Systems Biology 1:54 DOI 10.1186/1752-0509-1-54