Short Communication

Modulating Cancer Progression from Leukoplakia via Bayesian Gene Networks

Alessandro Villa DDS, PhD, MPH [1], Amin Zollanvari, PhD [2]

[1] Department of Orofacial Sciences, University of California San Francisco; alessandro.villa@ucsf.edu
[2] School of Engineering and Digital Sciences, Nazarbayev University, Astana, Kazakhstan; amin.zollanvari@nu.edu.kz
* Correspondence: alessandro.villa@ucsf.edu

Abstract: Oral squamous cell carcinoma often arises from an oral potentially malignant disorder called oral leukoplakia (OL). With this work we aimed to develop a novel data-driven predictive model based on gene expression profiles to distinguish OL patients who underwent malignant transformation from those who did not. We used the Tree Augmented Naïve (TAN) Bayes classifier to predict the posterior probability of having oral cancer given the data. 86 patients were included with a median follow-up of 7.11 years. Fifty-one patients (51/86; 59%) underwent malignant transformation. We found that 16 genes were predictors of oral cancer in patients with OL and these included SLC7A11, SPINK6, SERPINA12, VIT, ATP1B3, CST6, FLRT2, ELMOD1, AZGP1, RNASE13, DIO2, ECM1, CYP4F11, SYTL4, AKR1C1, and AKR1C3. In conclusion, we showed that Bayesian gene networks are a data-driven approach which could be used also in other predictor models in oncology.

Keywords: oral leukoplakia; Bayesian Networks; malignant transformation

1. Introduction

Worldwide, lip and oral cavity cancer accounts for more than 350,000 cases and 177,000 deaths every year [1]. An estimated 54,010 new cases of cancer of the oral cavity and pharynx are expected in the United States in 2021, with incidence rates now more than twice as high in men as in women [2]. Oral cancer is a genetic disease that arises in a sequential process in which normal cells are transformed into precancerous cells and then to cancer [3][4]. Several oral cancers are preceded by oral potentially malignant disorders, such as leukoplakia and erythroplakia. Leukoplakia is defined by the World Health Organization (WHO) as "a white plaque of questionable risk having excluded (other) known diseases or disorders that carry no increased risk for cancer" [5]. Leukoplakia may have a histological diagnosis of "benign hyperkeratosis" or parakeratosis without dysplasia, epithelial dysplasia (categorized by mild, moderate, or severe), carcinoma in-situ or invasive squamous cell carcinoma [6,7]. The presence of dysplastic areas in the epithelium of the oral cavity has been associated with the progression to cancer. Other risk factors associate with malignant transformation include tobacco use and heavy alcohol consumption [8]. The role of persistent high-risk type Human Papilloma Virus infection in oral leukoplakia remains controversial [9]. In the last few years there has been an increasing interest in genes that predispose to oral carcinogenesis. Studies have shown an overall improvement in discovering disease-related biomarkers [10] and genetic analysis led to promising results in predicting malignant progression of oral potentially malignant disorders [11-13]. However, there is an experimental bias in developing these new methods as they are based on a set of gold-standard gene list, which are often the result of previous works. As such, the aim of this study was to develop an unbiased and data-driven predictive model based on gene expression profiles to

distinguish leukoplakic patients who develop oral cancer from those who did not undergo malignant transformation.

2. Materials and Methods

We built a predictive model for oral cancer in patients affected by oral leukoplakia considering genes associated with risk of oral cancer, histological diagnosis of leukoplakia, tobacco use and alcohol consumption. Data were obtained from the GSE26549 dataset [14]. Demographic and clinical information, sample preparation, amplification, labeling, and microarray hybridization have been previously described in detail [14]. There was a total of 29,096 genes. First, duplication of genes was removed. Genes that shared the same Gene Symbols were considered as identical and expression profiling was collapsed. Genes whose Gene Symbols were not available were discarded for this analysis. For GSE26549, 19,894 unique genes remained after duplication removal and expression profiling collapse. Uninformative genes, whose variances across all samples were less than 75% quantile of all genes' variances, were eliminated at the beginning. Specifically, 4,973 informative genes remained. The relatively small number of the available sample led us to utilize the Tree Augmented Naïve (TAN) Bayes classifier [15] to predict the posterior probability of having oral cancer given the data. In this structure, the assumption is that each variable only depends on only one other variable. This simplifying assumption to approximate the joint distribution of variables helps in estimating the joint distribution from a relatively small sample size [16,17]. Several empirical studies have shown that the TAN structure can outperform other well-known predictive models in a relatively small sample situation [15,18]. Furthermore, the graphical representation of the constructed TAN model reveals any existing interaction among factors used to predict the response variable. A Sequential Forward Search (SFS) was used as feature selection to select the most promising genes. For the purpose of this analysis three nodes 1) histology (dysplasia versus hyoerkeratosis/hyperplasia), 2) alcohol consumption, and 3) smoking status were fixed in the model. In our implementation, selection procedure was terminated if the estimated cross-validation accuracy of the constructed TAN model decreased when a new feature was added, or the estimated accuracy remained the same for the 10 consecutive genes. Finally, one TAN classifier was constructed based on the selected genes as well as histology, alcohol consumption and smoking status. To validate our constructed predictive model using selected genes (see Results section for the list of genes) and, at the same time, avoid the selection bias in small sample, a nested cross-validation procedure as described previously was used [16,19] (see Figure 1). In this procedure, the cross-validation is applied external to the feature selection (i.e., SFS + internal cross-validation). The area under curve achieved by 10-fold nested cross-validation was 85%.

3. Results and Discussion

A total of 86 patients were included for this project and had a median follow-up of 7.11 years. Fifty-one individuals (51/86; 59%) developed oral cancer over time [14]. Our findings showed that 16 genes were predictors of oral cancer in patients with leukoplakia (Figure.2) and these include SLC7A11, SPINK6, SERPINA12, VIT, ATP1B3, CST6, FLRT2, ELMOD1, AZGP1, RNASE13, DIO2, ECM1, CYP4F11, SYTL4, AKR1C1, and AKR1C3. In addition, tobacco smoking, alcohol consumption and having a diagnosis of dysplasia increased the risk of developing oral cancer. The underlying hypothesis of our work was that cancer progression from leukoplakia is modulated by a set of genetic variations that interact to control these processes. The goal of this study was to identify such a set of variations and combine them into a predictive model of genetic and clinical data using Bayesian networks, an innovative approach developed at the crossroads of statistics and artificial intelligence. In the present study we used a completely data-driven approach for predictive medicine. Using this approach, new genes associated with the malignant

progression of leukoplakia were discovered (Figure 2). The genes SLC7A11, SPINK6, SERPINA12, VIT, ATP1B3, CST6, FLRT2, ELMOD1, AZGP1, RNASE13, DIO2, ECM1, CYP4F11, SYTL4 merit further investigation in the future in the context of malignant progression of leukoplakia to OC. AKR1C1 and AKR1C3, are genes found to be involved in tobacco carcinogen in oral buccal mucosa [20]. Future studies are needed to shed some light on the functions of these genes and gene products.

In summary, this work presents a data-driven approach to integrate clinical and genomic data in order to discover new biological pathways for oral cancer progression and leukoplakia-related genes. We anticipate Bayesian gene networks will be used for larger studies to predict the malignant transformation of leukoplakia.

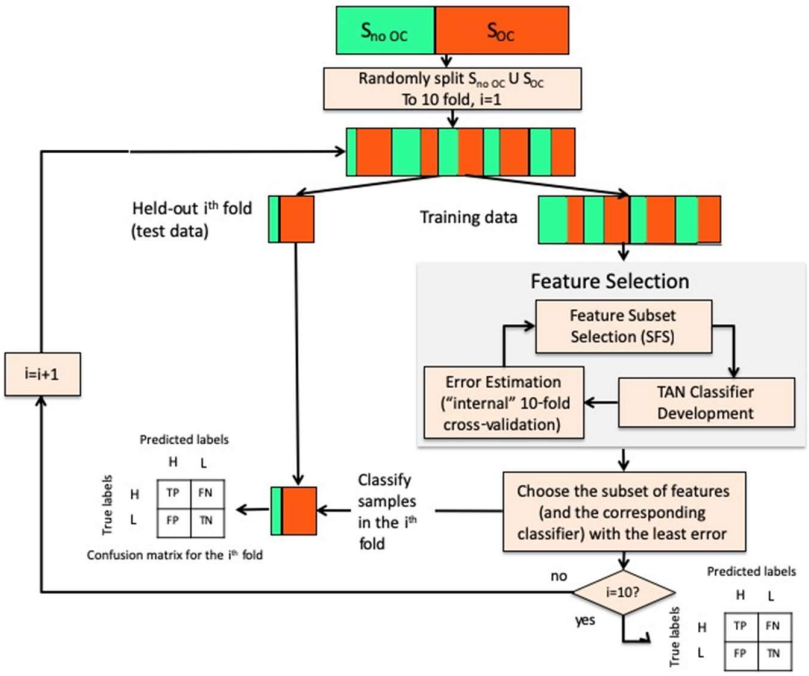

Figure 1. A schematic diagram of the cross-validation procedure external to the SFS feature selection. SOC (Sno OC) denotes the set of leukoplakic patients who (do not) develop oral cancer. We used a 10-fold cross-validation both external and internal to the feature selection process. Without having an external cross-validation the process of feature selection results in a selection bias, which results in an optimistic generalization error of the constructed classifier [19}.

Abbreviations: TP: true positive; TN: true negative; FP: false positive; FN false negative.

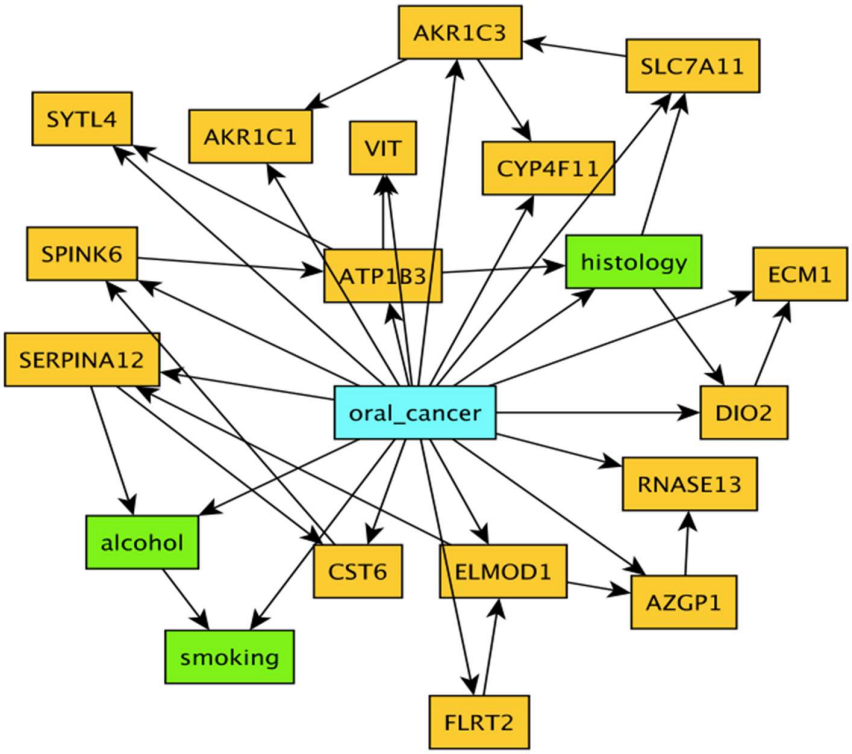

Figure 2. Graphical representation of the predictive model of oral cancer given oral leukoplakia.

Author Contributions: "Conceptualization, A.V. and A.Z.; methodology, A.Z.; formal analysis, A.Z.; writing—original draft preparation, A.V.; writing—review and editing, A.V. and A.Z. All authors have read and agreed to the published version of the manuscript.:

Funding: "This research received no external funding"

Institutional Review Board Statement: Ethical review and approval were waived for this study, as the genetic data are publicly available.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data can be found on the publicly archived dataset GSE26549 [14].

Conflicts of Interest: "The authors declare no conflict of interest."

References

1. Bray, F.; Ferlay, J.; Soerjomataram, I.; Siegel, R.L.; Torre, L.A.; Jemal, A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* **2018**, *68*, 394-424, doi:10.3322/caac.21492.
2. Siegel, R.L.; Miller, K.D.; Fuchs, H.E.; Jemal, A. Cancer Statistics, 2021. *CA Cancer J Clin* **2021**, *71*, 7-33, doi:10.3322/caac.21654.
3. Haddad, R.I.; Shin, D.M. Recent advances in head and neck cancer. *N Engl J Med* **2008**, *359*, 1143-1154, doi:10.1056/NEJMra0707975.

4. Warnakulasuriya, S.; Ariyawardana, A. Malignant transformation of oral leukoplakia: a systematic review of observational studies. *Journal of oral pathology & medicine : official publication of the International Association of Oral Pathologists and the American Academy of Oral Pathology* **2016**, *45*, 155-166, doi:10.1111/jop.12339.
5. Warnakulasuriya, S.; Johnson, N.W.; van der Waal, I. Nomenclature and classification of potentially malignant disorders of the oral mucosa. *Journal of oral pathology & medicine : official publication of the International Association of Oral Pathologists and the American Academy of Oral Pathology* **2007**, *36*, 575-580, doi:10.1111/j.1600-0714.2007.00582.x.
6. Villa, A.; Sonis, S. Oral leukoplakia remains a challenging condition. *Oral Dis* **2018**, *24*, 179-183, doi:10.1111/odi.12781.
7. Villa, A.; Woo, S.B. Leukoplakia-A Diagnostic and Management Algorithm. *J Oral Maxillofac Surg* **2017**, *75*, 723-734, doi:10.1016/j.joms.2016.10.012.
8. Pelucchi, C.; Gallus, S.; Garavello, W.; Bosetti, C.; La Vecchia, C. Cancer risk associated with alcohol and tobacco use: focus on upper aero-digestive tract and liver. *Alcohol Res Health* **2006**, *29*, 193-198.
9. Lerman, M.A.; Almazrooa, S.; Lindeman, N.; Hall, D.; Villa, A.; Woo, S.B. HPV-16 in a distinct subset of oral epithelial dysplasia. *Mod Pathol* **2017**, *30*, 1646-1654, doi:10.1038/modpathol.2017.71.
10. Villa, A.; Celentano, A.; Glurich, I.; Borgnakke, W.S.; Jensen, S.B.; Peterson, D.E.; Delli, K.; Ojeda, D.; Vissink, A.; Farah, C.S. World Workshop on Oral Medicine VII: Prognostic biomarkers in oral leukoplakia: A systematic review of longitudinal studies. *Oral Dis* **2019**, *25 Suppl 1*, 64-78, doi:10.1111/odi.13087.
11. Bremmer, J.F.; Brakenhoff, R.H.; Broeckaert, M.A.; Belien, J.A.; Leemans, C.R.; Bloemena, E.; van der Waal, I.; Braakhuis, B.J. Prognostic value of DNA ploidy status in patients with oral leukoplakia. *Oral Oncol* **2011**, *47*, 956-960, doi:10.1016/j.oraloncology.2011.07.025.
12. Nasser, W.; Flechtenmacher, C.; Holzinger, D.; Hofele, C.; Bosch, F.X. Aberrant expression of p53, p16INK4a and Ki-67 as basic biomarker for malignant progression of oral leukoplakias. *J Oral Pathol Med* **2011**, *40*, 629-635, doi:10.1111/j.1600-0714.2011.01026.x.
13. Califano, J.; Westra, W.H.; Meininger, G.; Corio, R.; Koch, W.M.; Sidransky, D. Genetic progression and clonal relationship of recurrent premalignant head and neck lesions. *Clin Cancer Res* **2000**, *6*, 347-352.
14. Saintigny, P.; Zhang, L.; Fan, Y.H.; El-Naggar, A.K.; Papadimitrakopoulou, V.A.; Feng, L.; Lee, J.J.; Kim, E.S.; Ki Hong, W.; Mao, L. Gene expression profiling predicts the development of oral cancer. *Cancer Prev Res (Phila)* **2011**, *4*, 218-229, doi:10.1158/1940-6207.CAPR-10-0155.
15. Friedman, N.; Geiger, D.; Goldszmidt, M. Bayesian Network Classifiers. *Machine Learning* **1997**, *29*, 161-163.
16. Zollanavari, A.; Kizilirmak, R.C.; Kho, Y.H.; Hernández-Torrano, D. Predicting students' GPA and developing intervention strategies based on self-regulatory learning behaviors. *IEEE Access* **2017**, *5*, 23792 - 23802.
17. Zollanvari, A.; Alterovitz, G. SNP by SNP by environment interaction network of alcoholism. *BMC Syst Biol* **2017**, *11*, 19, doi:10.1186/s12918-017-0403-7.
18. Madden, M.G. On the classification performance of TAN and general Bayesian networks. *Knowledge-Based Systems* **2009**, *22*, 489-495.
19. Ambroise, C.; McLachlan, G.J. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc Natl Acad Sci U S A* **2002**, *99*, 6562-6566, doi:10.1073/pnas.102102699.
20. Zhang, L.; Lee, J.J.; Tang, H.; Fan, Y.H.; Xiao, L.; Ren, H.; Kurie, J.; Morice, R.C.; Hong, W.K.; Mao, L. Impact of smoking cessation on global gene expression in the bronchial epithelium of chronic smokers. *Cancer Prev Res (Phila)* **2008**, *1*, 112-118, doi:10.1158/1940-6207.CAPR-07-0017.