33rd Annual International Conference of the IEEE EMBS
Boston, Massachusetts USA, August 30 - September 3, 2011

Gene expression profiling towards the prediction of oral cancer reoccurrence

Konstantinos P. Exarchos, Yorgos Goletsis, *Member, IEEE*, Tito Poli and Dimitrios I. Fotiadis, *Senior Member, IEEE*

Abstract—**In this work we perform gene expression profiling on tissue specimen obtained from patients with oral squamous cell carcinoma with a twofold aim: i) to identify a limited number of genes that capture perturbations at molecular level dictating the development of a potential disease relapse after remission, and ii) to employ these genes in order to build a classifier that is able to calculate the probability of disease reoccurrence for new patients, subsequently discriminating patients into high and low risk groups based on reoccurrence probability. The proposed analysis yielded 94% overall accuracy, 100% sensitivity and 89% specificity, for discriminating patients with and without a disease relapse.**

I. Introduction

ORAL cancer is the predominant neoplasm type that arises in the head and neck region, subsequently constituting the eighth most common cancer in the worldwide cancer incidence ranking [1]. Besides the low quality of life of patients suffering from oral cancer, another major issue has to do with reoccurrence rates after the disease has reached remission; specifically, locoregional relapses after successful treatment of the primary tumor have been reported in the range of 25-48%, of which 95% occur within 2 years. There is a strong relation between oral cancer and the sex of the patient, with men facing twice the risk of being diagnosed with oral cancer than women [1]. Moreover, a wide range of risk factors have been associated with oral cancer, such as smoking, especially coupled with alcohol consumption, as well as sun exposure and HPV infection [1].

Currently implemented methods aiming to predict oral cancer reoccurrence after remission, have reported quite unsatisfactory results [2]. From the clinical point of view, several factors have been associated with the reoccurrence of oral cancer, such as age, site and stage of the primary tumor as well as certain histological features. Moreover, especially in the molecular basis of the disease, currently available biomarkers are limited in number and efficacy [3]. The identification of new features and the efficient combination of the already known ones will greatly contribute towards the accurate and more reliable prognosis of the disease.

In the literature several methodologies have been proposed for the identification of molecular markers that are involved in the induction and evolvement of oral cancer or slightly similar types of cancers. Those markers are in turn employed for discriminating between patients with high and low risk of developing a disease reoccurrence in adjacent tissues or a lymph node metastasis. Specifically, [4, 5] derive an expression profile for diagnosis of lymph node metastasis from primary head and neck squamous cell carcinoma; similarly, in [6], future metastases of head and neck carcinoma are predicted. In [7-9] the progression of tongue carcinoma is studied, and a subset of genes with predictive potential is identified able to predict potential metastasis of the primary tumor in the lymph nodes.

In the current work, we perform a systematic analysis upon a multitude of genes in order to pinpoint genetic biomarkers that potentially dictate the progression of oral cancer. Subsequently, we develop a classification scheme which employs the aforementioned genes in order to early identify a disease reoccurrence. Knowing in advance the progression of the disease, we are able fine-tune accordingly the follow-up; i.e. patients in low risk of reoccurrence are subject to the traditional (or less intensive than the traditional) follow-up, whereas for patients in high-risk of developing a relapse adequate further diagnostic and treatments measures are undertaken.

II. Materials And Methods

A. Clinical Scenario

The clinical scenario employed during this work, is depicted in Fig. 1. Initially a patient is diagnosed with oral cancer through traditional clinical procedures. At this point, (i.e. the baseline) the physician gathers genetic data from the tumor site, and the patient is then treated properly. After the physician's therapeutic intervention (i.e. surgery, chemo/radio-therapy), the patient either reaches complete remission or particles of the cancer tissue still remain intact. In the latter case the patients do not qualify for the purposes of our study, whereas patients in complete remission are being monitored for a follow-up period of 18 months. Based on expert knowledge the vast of majority of relapses appear within an 18 month follow-up timeframe.

Subsequently, two classes of patients are assembled, i.e.

Manuscript submitted on March 26, 2011.

This work is part funded by the European Commission NeoMark project (FP7-ICT-2007-224483 – ICT enabled prediction of cancer reoccurrence).

K.P. Exarchos and D.I. Fotiadis are with the Unit of Medical Technology and Intelligent Information Systems, Dept. of Materials Science and Engineering, University of Ioannina, GR 45110 (e-mail: kexarcho@gmail.com, corresponding author phone: +302651008803; fax: +302651008889; e-mail: fotiadis@cs.uoi.gr).

Y. Goletsis is with the Dept. of Economics, University of Ioannina, GR 45110 (e-mail: goletsis@cc.uoi.gr).

T. Poli is with the Head and Neck Department, Azienda Ospedaliero Universitaria of Parma, Italy (e-mail: tito.poli@unipr.it)

relapsers and non-relapsers. The formulation of the relapsers' class is pretty straightforward and contains patients developing a disease reoccurrence after treatment. However, in order to conjecture about non-relapsers, we require that a patient must have been disease-free for at least 12 months after the initial treatment.

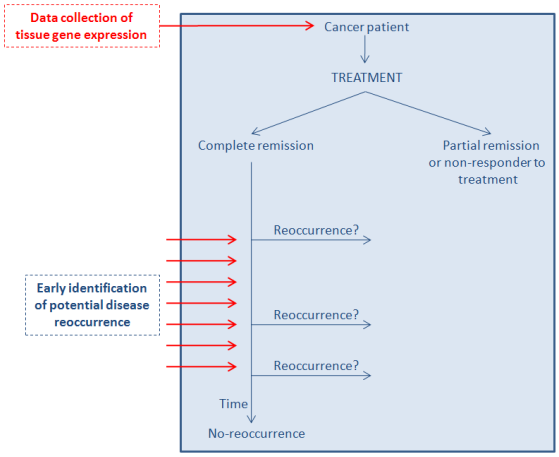

Fig. 1: The clinical scenario.

Table I contains the list of patients eligible for our study, the respective months of follow-up and each patient's current disease status.

TABLE I
STATUS AND STATE OF ENROLLED PATIENTS

#	State (month)	Status	#	State (month)	Status
1	18	NR	14	18	NR
2	12	NR	15	18	NR
3	**9**	**R**	16	18	NR
4	15	NR	**17**	**15**	**R**
5	12	NR	18	12	NR
6	18	NR	19	12	NR
7	18	NR	20	15	NR
8	15	NR	**21**	**9**	**R**
9	15	NR	**22**	**6**	**R**
10	12	NR	23	18	NR
11	12	NR	**24**	**9**	**R**
12	**6**	**R**	**25**	**6**	**R**
13	12	NR			

NR: No reoccurrence; R: Reoccurrence

B. Data acquisition

At the baseline state of each patient, we extract from the cancerous tissue the expression of 45,015 genes. During this study all microarray experiments have been conducted using the same platform, the same array design and the same feature extraction software version in order to minimize the risk of possible sources of variability in the data, other than biological variability. Specifically, the 4x44K oligo-RNA human genome arrays from Agilent Technologies (Santa Clara, US) have been employed and processed using the Feature Extraction software V9.5 (Agilent Technologies). The resulting gene expression files are subject to some basic preprocessing steps in order to enhance the quality of the input; initially all control and duplicate genes are removed, as well as genes of too low quality (i.e. genes with high variability of inter-spot intensities) and genes with missing values. The outcome of the preprocessing step is a set of 33491 high-quality genes.

C. Gene identification

In the next step we systematically analyze the expression of the remaining 33491 genes in order to identify a limited set of genes that more selectively and more precisely characterize the different classes of patients examined in this work, namely patients with and without a disease reoccurrence, while accounting for the enormous number of genes. For this purpose we employ the Significance Analysis of Microarrays (SAM) algorithm [10], which analyzes differentially gene expression data between two groups and assigns a score to each gene based on the change in gene expression between the two classes of patients. SAM pinpoints genes as being statistically significant differentially expressed in two sets, by assimilating several gene-specific t-tests on permutations of the initial dataset. Subsequently a score is attached to each gene based on its perturbation in gene expression relative to the standard deviation of repeated measurements for that gene. Multiple tests have been performed by varying the fold-change, i.e. the amount that genes between the two classes change in order to be considered as significant. Table II contains the values of the fold-change applied in each run of the SAM algorithm, the number of genes maintained each time and the genes identified as false positives, both in percentage and in absolute numbers.

TABLE II
NUMBER OF GENES IDENTIFIED AS SIGNIFICANT FOR VARIABLE VALUES OF FOLD-CHANGE BETWEEN THE TWO SETS OF PATIENTS

Fold change	# of significant genes	FDR (%)	# of false positives
1.0	2	0.55	1.1
1.2	1	0	0
1.5	**40**	**13**	**5.5**
1.8	6	0	0
2.0	0	0	0
2.5	0	0	0

Setting a threshold of at least 1.5 fold-change between the two sets patients, we obtain a list of 40 genes, as it is shown in Table III.

TABLE III
GENES PINPOINTED AS MOST SIGNIFICANT

LPO	TMC5	AI916628	CA946373
MSLN	ROPN1	PIGR	CLDN8
CAPN13	AGR2	C20orf114	CTAG1A
GLYATL2	SCGB1D1	CP	SCGB3A1
CB959193	LOC440335	C10orf81	LOC63928
CLDN22	THC2339617	VTCN1	OLFM4
BCMP11	UPK1B	SCGB2A1	KCNJ16
C20orf85	CRISP2	MSMB	LOC124220
SCGB2A2	CHST9	FOXA1	PIP
SLC34A2	PROM1	C10orf81	STATH

In Fig. 2 we also provide the respective heatmap depicting the expression of each retained gene, for all patients

considered.

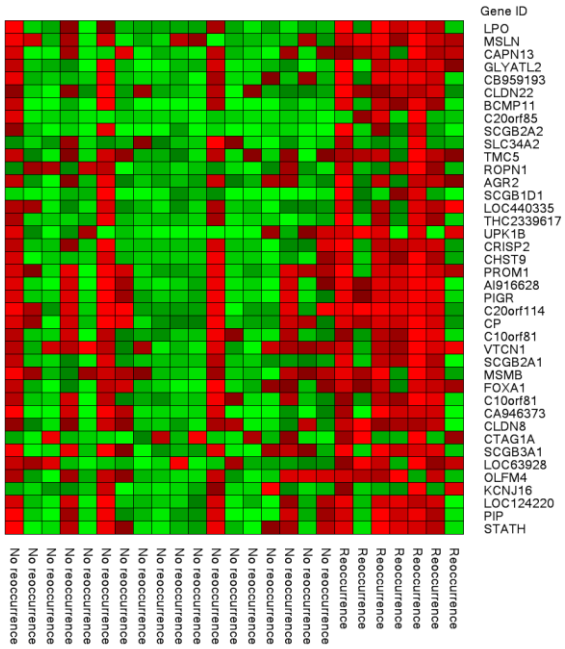

Fig. 2: Heatmap of the 40 most significant genes.

In order to further validate the list of procured genes, a series of state of the art methodologies are also employed for our dataset. Specifically, eBayes [11], PLS-CV [12], RF-MDA [13] as well as an ensemble methodology that combines the previous ones are invoked [14]. It is worth noticing that 29 out of the 40 maintained genes are common between our approach and in at least one of the other four algorithms, thus, verifying the validity of the proposed set of genes (genes identified by each methodology are not shown due to space limitations).

D. Class imbalance

Next, we utilize the genes identified as significant with the aforementioned analysis, in a classification algorithm able to discriminate patients with and without relapse based on the differential expression of those genes. The resulting dataset contains 25 patients (7 patients with a relapse and 18 relapse-free), for which the expressions of the 40 genes in Table III, formulate the features used for classification. In order to overcome the apparent class imbalance in our dataset, we employ the Synthetic Over Sampling Technique (SMOTE), which instead of merely replicating the instances of the minority class, uses a k-NN approach in order to create a new case which resembles and combines the available ones [15]. Using this approach we expand the minority class with new cases that resemble the available ones but are not identical, resulting eventually in a fully balanced dataset of 36 cases.

E. Feature selection

In the next step we employ two popular feature selection algorithms in order to further search across the 40 retained genes the feature subset that is most informative from a classification perspective. First the Correlation-based Feature Subset Selection (CFS) algorithm [12] which evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them. The Wrapper algorithm [16] was also employed which uses the target learning algorithm as a "black-box" to estimate the worth of attribute subsets by measuring accuracy estimates. Feature wrappers often outperform other feature selection schemes due to the fact that they are tuned to the target machine learning algorithm.

F. Classification

In terms of classification, five popular classification schemes are examined, namely: Bayesian Networks (BN), Artificial Neural Networks (ANN), Support Vector Machines (SVM), Decision Trees (DT) and Random Forests (RF). For further details regarding the aforementioned classifications schemes see e.g. [17].

III. Results And Discussion

For evaluation purposes, we calculate accuracy (Acc), sensitivity (Se) and specificity (Sp) after performing 10-fold cross validation. Sensitivity is defined as the fraction of correctly identified relapsing patients, specificity measures the proportion of disease-free patients predicted as non-relapsing ones, and accuracy is the weighted average of the sensitivity and specificity denoting the overall correctness of the model. In the current analysis, special attention is given in sensitivity, i.e. the identification of almost all possible relapsing patients, (provided that specificity and consequently accuracy are adequately high) in order to undertake necessary adjuvant treatment. In the tables that follow we present the results obtained with each classifier either without performing feature selection (Table IV) or after employing the CFS (Table V) and wrapper (Table VI) algorithms in order to identify the features with the highest discrimination potential.

TABLE IV
Results Obtained Without Performing Feature Selection

Classification algorithm	Acc (%)	Se (%)	Sp (%)
BN	83.3	88.9	77.8
ANN	83.3	88.9	77.8
SVM	88.9	94.4	83.3
DT	77.8	66.7	88.9
RF	77.8	77.8	77.8

Next, we employ the CFS algorithm for feature selection, yielding the following features as most significant: MSLN, CAPN13, GLYATL2, CLDN22, CTAG1A and LOC63928. The results obtained with the reduced input vector are shown in Table V.

TABLE V
Results Obtained After Employing The CFS Algorithm

Classification algorithm	Acc (%)	Se (%)	Sp (%)
BN	86.1	83.3	88.9
ANN	86.1	94.4	77.8
SVM	88.9	100	77.8

DT	75	66.7	83.3
RF	77.8	72.2	83.3

Afterwards, we employ the wrapper algorithm for feature selection, which retains the following genes for the best performing classification algorithms, specifically for BN: MSLN and CAPN13; and for ANN: MSLN, CAPN13, C20ORF85, MSMB and OLFM4.

TABLE VI
RESULTS OBTAINED AFTER EMPLOYING THE WRAPPER ALGORITHM

Classification algorithm	Acc (%)	Se (%)	Sp (%)
BN	**94.4**	**100.0**	**88.9**
ANN	**94.4**	**100.0**	**88.9**
SVM	91.7	94.4	88.9
DT	91.7	94.4	88.9
RF	91.7	94.4	88.9

In all cases the employment of the wrapper algorithm significantly ameliorates the obtained results. The classification schemes that yielded the highest results are the Bayesian Network and the Artificial Neural Network coupled with the wrapper algorithm. However, the BN is slightly preferable due to its transparent architecture as well as the simplicity owed to the less number of retained genes, namely MSLN and CAPN13. MSLN encodes a precursor protein that is cleaved into megakaryocyte potentiating factor and mesothelin. Especially mesothelin has been found to be overexpressed in epithelial mesotheliomas [18] and other types of squamous cell carcinomas [19], as is the case of oral cancer. CAPN13 belongs to calpains, a protein family that has been involved in a variety of cellular processes, including apoptosis, cell division and many others [20]. Perturbations in calpain activity have been associated with pathophysiological processes contributing to type II diabetes and certain types of cancer [20].

Table VII provides a comparison among the methodologies reported in the literature and the best performing scheme elicited within the current work.

TABLE VII
COMPARISON BETWEEN THE CURRENT WORK AND THE LITERATURE

Method	Number of patients	Acc (%)
Roepman [4]	66	88
Roepman [5]	22	86
Rickman [6]	79	77
Watanabe [7]	39	76
Nagata [8]	75	87
Zhou [9]	25	85
Current Work	**21**	**94**

We observe that the results obtained in the current work are slightly superior compared to other methodologies reported in the literature. However, direct quantitative comparison cannot be performed since all methodologies have been evaluated on different sets of patients.

IV. CONCLUSIONS

In this work we have analyzed the expression of a multitude of genes, in order to identify a limited subset of genetic factors that potentially dictate the evolvement of the disease after remission; those genes have been subsequently employed towards the utilization of a classification scheme able to discriminate between patients with and without relapse. Hence, we are able to enhance our knowledge regarding the molecular basis of the disease, but additionally, the timely prediction of a potential relapse allows for fine-tuning accordingly the follow-up treatment.

REFERENCES

[1] R. I. Haddad and D. M. Shin, "Recent advances in head and neck cancer," *N EJM,* vol. 359, pp. 1143-54, 2008.

[2] A. Forastiere, R. Weber, and K. Ang, "Treatment of head and neck cancer," *N EJM,* vol. 358, pp. 1076; author reply 1077-8, 2008.

[3] N. J. D'Silva and B. B. Ward, "Tissue biomarkers for diagnosis & management of oral squamous cell carcinoma," *Alpha Omegan,* vol. 100, pp. 182-9, 2007.

[4] P. Roepman, P. Kemmeren, L. F. Wessels*, et al.*, "Multiple robust signatures for detecting lymph node metastasis in head and neck cancer," *Cancer Res,* vol. 66, pp. 2361-6, 2006.

[5] P. Roepman, L. F. Wessels, N. Kettelarij*, et al.*, "An expression profile for diagnosis of lymph node metastases from primary head and neck squamous cell carcinomas," *Nat Genet,* vol. 37, pp. 182-6, 2005.

[6] D. S. Rickman, R. Millon, A. De Reynies*, et al.*, "Prediction of future metastasis and molecular characterization of head and neck squamous-cell carcinoma based on transcriptome and genome analysis by microarrays," *Oncogene,* vol. 27, pp. 6607-22, 2008.

[7] H. Watanabe, K. Mogushi, M. Miura*, et al.*, "Prediction of lymphatic metastasis based on gene expression profile analysis after brachytherapy for early-stage oral tongue carcinoma," *Radiother Oncol,* vol. 87, pp. 237-42, 2008.

[8] T. Nagata, R. Schmelzeisen, D. Mattern*, et al.*, "Application of fuzzy inference to European patients to predict cervical lymph node metastasis in carcinoma of the tongue," *Int J Oral Maxillofac Surg,* vol. 34, pp. 138-42, 2005.

[9] X. Zhou, S. Temam, M. Oh*, et al.*, "Global expression-based classification of lymph node metastasis and extracapsular spread of oral tongue squamous cell carcinoma," *Neoplasia,* vol. 8, pp. 925-32, 2006.

[10] V. G. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response," *PNAS,* vol. 98, pp. 5116-21, 2001.

[11] G. K. Smyth, J. Michaud, and H. S. Scott, "Use of within-array replicate spots for assessing differential expression in microarray experiments," *Bioinformatics,* vol. 21, pp. 2067-75, 2005.

[12] M. Hall, "Correlation-based feature selection for discrete and numeric class machine learning," in *17th Int Conf on Machine Learning*, 2000, pp. 359-366.

[13] L. Breiman, "Random forests," *Machine learning,* vol. 45, pp. 5-32, 2001.

[14] E. Glaab, J. M. Garibaldi, and N. Krasnogor, "ArrayMining: a modular web-application for microarray analysis combining ensemble and consensus methods with cross-study normalization," *BMC Bioinf,* vol. 10, p. 358, 2009.

[15] N. Chawla, K. Bowyer, L. Hall*, et al.*, "SMOTE: synthetic minority over-sampling technique," *J of Artif Intel Res,* vol. 16, pp. 321-357, 2002.

[16] R. Kohavi and G. John, "Wrappers for feature subset selection," *Artif. intel.,* vol. 97, pp. 273-324, 1997.

[17] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to data mining*, 1st ed. Boston: Pearson Addison Wesley, 2006.

[18] K. Tan, K. Kajino, S. Momose*, et al.*, "Mesothelin (MSLN) promoter is hypomethylated in malignant mesothelioma, but its expression is not associated with methylation status of the promoter," *Hum Pathol,* vol. 41, pp. 1330-8, 2010.

[19] C. Y. Liu, M. C. Wu, F. Chen*, et al.*, "A Large-scale genetic association study of esophageal adenocarcinoma risk," *Carcinogenesis,* vol. 31, pp. 1259-63, 2010.

[20] T. N. Dear and T. Boehm, "Identification and characterization of two novel calpain large subunit genes," *Gene,* vol. 274, pp. 245-52, 2001.