Models for Predicting Stage in Head and Neck Squamous Cell Carcinoma Using Proteomic Data

Chanchala D. Kaddi and May D. Wang-*IEEE Senior Member*

Abstract—**Head and neck squamous cell carcinoma (HNSCC) that is detected at an advanced stage is associated with much worse patient outcomes than if detected at early stages. This study uses reverse phase protein array (RPPA) data to build predictive models that discriminate between early and advanced stage HNSCC. Individual and ensemble binary classifiers, using filter-based and wrapper-based feature selection, are used to build several models which achieve moderate MCC and AUC values. This study identifies informative protein feature sets which may contribute to an increased understanding of the molecular basis of HNSCC.**

I. Introduction

Head and neck squamous cell carcinoma (HNSCC) is the 6th most prevalent type of cancer worldwide [1]. The stage at which HNSCC is detected is important to therapeutic outcomes; patients with early stage (I and II) cancer have between 60-95% chance of successful local treatment, while those with advanced stage cancer are at high risk for recurrence or metastatic disease [2]. Increased understanding of the molecular characteristics of HNSCC stages may help to develop more effective detection and treatment strategies.

In terms of identifying potential markers related to disease stage, recent research findings have varied. Some genetic studies have related selected genes and gene signatures to different HNSCC stages [3-5], while others have not found such discriminatory genes [6, 7]. Similarly, some proteomic or metabolomic studies have identified potential markers for discriminating between early and advanced stage HNSCC samples [8], while others have reported mixed or limited results [9, 10].

Further analytical proteomics studies may provide additional insight into the differences between early and advanced stage HNSCC. In particular, analysis of reverse phase protein array (RPPA) data is a promising avenue. RPPA is an experimental technique for quantitative functional proteomics; in RPPA, a sample is probed with antibodies against specific proteins, and the antibody signals are used to quantify protein expression levels. With respect to HNSCC, RPPA data has been used to identify differentially expressed proteins between cancer and normal samples [11] and to identify proteins affected by the presence of an anti-invasion compound in nasopharyngeal carcinoma [12]. For several other cancer types, RPPA data has been applied to build predictive models: for prognosis [13] and drug response [14] in breast cancer; for treatment response in ovarian cancer [15]; and for drug sensitivity in non-small-cell lung cancer [16].

In this study, RPPA data is used to investigate differences between early and advanced stage HNSCC. Alternative feature sets and alternative classification algorithms are tested to construct predictive models that can effectively discriminate between patient groups. The resulting models may help to provide insight into the protein-level characteristics associated with HNSCC stages.

II. Methods

A. Data

RPPA data for HNSCC was downloaded from The Cancer Proteome Atlas (TCPA) [17] at http://bioinformatics.mdanderson.org/main/TCPA:Overview . This dataset consists for 212 patient samples and measures the response to 187 antibodies. TCPA provides a proteomic complement to The Cancer Genome Atlas (TCGA) [18] at http://cancergenome.nih.gov/, where clinical, transcriptomic, and genomic data for the same patients are available. The downloaded RPPA data had been normalized and protein expression had been quantified using the "Supercurve Fitting" method. The details of these pre-processing steps are described in [17, 19]. In TCPA, antibodies are grouped into three classes: 'validated', 'under evaluation', and 'use with caution.' To perform a more conservative analysis, only those proteins with antibodies described as 'validated' in both [17, 19] were utilized in this study. 112 proteins were considered for further analysis.

The HNSCC clinical data for the 212 patients was downloaded from TCGA. Pathological stage information was used to divide the RPPA dataset into two groups: patients with stage I and II cancer, and patients with stage III or IV cancer. Pathological state was unavailable for 12 patients, so clinical stage was substituted. The early stage group contained 50 patients, and the advanced stage group contained 162 patients.

This research has been supported by grants from The Parker H. Petit Institute for Bioengineering and Bioscience (IBB), Johnson & Johnson, Bio Imaging Mass Spectrometry Initiative at Georgia Tech, National Institutes of Health (Bioengineering Research Partnership R01CA108468, Center for Cancer Nanotechnology Excellence U54CA119338), Georgia Cancer Coalition (Distinguished Cancer Scholar Award to MDW), Microsoft Research, the National Science Foundation (GRFP to CDK), and P.E.O. International (Scholar Award to CDK).

C. D. Kaddi is with the Department of Biomedical Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: gtg538v@mail.gatech.edu).

M. D. Wang is with the Department of Biomedical Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA (phone: 404-385-5059; e-mail: maywang@bme.gatech.edu).

978-1-4244-7929-0/14/$26.00 ©2014 IEEE

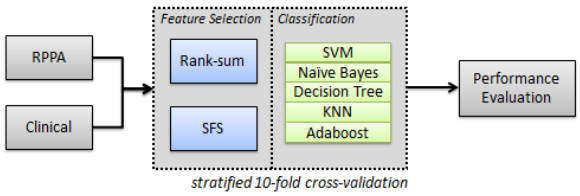

Figure 1. The performances of two feature selection methods and five classification methods were compared for predicting head and neck cancer stage on the basis of reverse phase protein array (RPPA) data.

B. Predictive Modeling

Fig. 1 shows an overview of the modeling framework used in this study. Four individual binary classification methods were utilized: SVM, KNN, naïve Bayes, and decision tree. For SVM, three alternative kernels were considered: linear, Gaussian radial basis function (GRBF), and multi-layer perceptron (MLP). For naïve Bayes, the normal and kernel distributions were examined. For KNN, the number of neighbors ranged from 1 to 10, i.e., $K \in [1,10]$. For the decision tree, the alternative splitting criteria of the Gini diversity index (GDI), the twoing rule, and maximum deviance reduction (MDR) were tested. The ensemble classifier Adaboost was also examined.

Two different analyses were performed, one using a filter approach for feature selection, and the other using a wrapper approach. In both cases, 10-fold stratified cross-validation was performed. Model performance was evaluated in terms of Matthews correlation coefficient (MCC). The cross-validation standard deviation and the area under the ROC curve (AUC) were also reported for the model having the maximum mean MCC.

In the filtering approach, the Wilcoxon rank-sum test was applied to identify proteins with significantly different expression between the early and advanced stage groups. Multiple testing corrections were applied by calculating the FDR for each protein, using the method of Benjamini and Hochberg through the R package p.adjust. To obtain a less conservative initial feature set, clinical stage was used to obtain a differentially expression protein list. This yielded 11 proteins with FDR values ≤ 0.05, including the five proteins found when pathological stage was used. A comprehensive examination of this feature space was performed by considering alternative classification models for every combination of the 11 features, i.e., $\sum_{i=1}^{11}\binom{11}{i} = 2047$ feature sets were considered.

In the wrapper approach, the same classifiers were investigated using sequential forward feature selection (SFS). The top-performing feature set with up to 10 features was identified for each classifier.

Analyses were performed using MATLAB 2013b (MathWorks, Natick MA). Unless otherwise specified, default classification algorithm parameters were utilized.

III. Results

Table I describes the performance of alternative predictive models across the rank-sum filter feature sets and from SFS. The best model for each classification method, in terms of mean MCC, is highlighted. The top-performing model overall was KNN with $K = 9$ neighbors, using features found through rank-sum filter-based feature selection. For both feature selection methods, KNN gave better results, followed by Naïve Bayes and SVM. Adaboost did not show improved performance over the individual classifiers on this dataset. Overall, predictive model performance is moderate, yielding several MCC values above 0.4 and AUC values above 0.75. These results indicate that predictive models using protein measurements from RPPA data as features can discriminate between patients with early or advanced stage HNSCC, and have potential for further investigation.

TABLE I. Performance Evaluation of Alternative Predictive Models Across Feature Selection Methods

Classification Method		Rank-sum Filter		SFS	
		MCC	AUC	MCC	AUC
SVM	Linear	0.41±0.22	0.74	0.39±0.13	0.72
	GRBF	**0.44**±0.20	0.78	0.41±0.23	0.73
	MLP	0.43±0.16	0.76	0.43±0.18	0.71
Naïve Bayes	Normal	0.42±0.23	0.74	**0.46**±0.27	0.71
	Kernel	0.44±0.27	0.74	0.42±0.14	0.68
Decision Tree	GDI	0.39±0.22	0.67	0.35±0.28	0.71
	Twoing	**0.39**±0.22	0.69	0.36±0.25	0.69
	MDR	0.37±0.20	0.71	0.37±0.21	0.67
KNN	K = 1	0.39±0.22	0.68	0.48±0.22	0.73
	K = 2	0.40±0.23	0.73	0.37±0.14	0.73
	K = 3	0.47±0.21	0.71	0.41±0.11	0.61
	K = 4	0.45±0.22	0.70	0.42±0.23	0.75
	K = 5	0.47±0.18	0.75	0.45±0.21	0.61
	K = 6	0.45±0.27	0.74	0.47±0.23	0.68
	K = 7	0.43±0.20	0.73	0.36±0.25	0.68
	K = 8	0.46±0.23	0.77	0.44±0.20	0.70
	K = 9	0.48±0.23	0.76	0.42±0.26	0.67
	K = 10	**0.49**±0.25	0.77	0.37±0.24	0.69
Adaboost	25 Trees	0.36±0.12	0.71	0.36±0.26	0.60
	50 Trees	0.37±0.16	0.65	0.36±0.22	0.65
	100 Trees	0.35±0.23	0.65	**0.42**±0.36	0.74

The identification of effective predictive models suggests that the proteins comprising the top feature sets may be functionally important in HNSCC progression. The 11 proteins selected using the Wilcoxon rank-sum test were: (1) Cyclin_B1, (2) FASN, (3) FoxM1, (4) JNK_pT183_Y185, (5) MAPK_pT202_Y204, (6) MEK1_pS217_S221, (7) p38_pT180_Y182, (8) S6_pS235_S236, (9) S6_pS240_S244, (10) Src_pY527, and (11) Syk. All of these proteins have been associated with HNSCC in the literature. In particular, we examine those features that are associated with many well-performing models. Fig. 2 shows the number of times each of these 11 features is utilized in the 21 top-performing models listed in Table 1. The first protein, Cyclin_B1, is present in 17/21 of these feature sets. Cyclin B1 controls the cell cycle checkpoint leading to mitosis, and has been shown to be over-expressed in HNSCC [20]. Src_pY527 is also present in 17/21 of these feature sets; Src_pY527 is the phosphorylated form of Src, a tyrosine kinase. In HNSCC, Src is a mediator of EGFR signaling [21]. Other frequently-selected proteins include FoxM1,

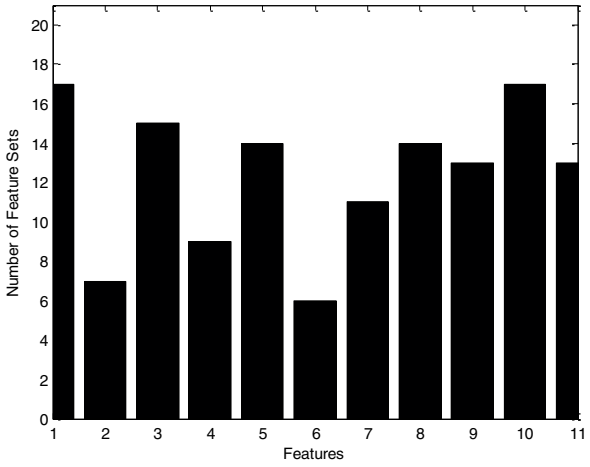

Figure 2. Number of times each rank-sum filter-selected feature appeared among the 21 top-performing models.

MAPK_pT202_Y204, and S6_pS235_S236. FoxM1 has been implicated as an early marker of HNSCC [22]. MAPK_pT202_Y204 is the phosphorylated, activated form of ERK1/2, a mitogen-activated protein kinase that is part of the MAPK/ERK signal transduction pathway, and that is frequently linked to HNSCC [23]. S6 is a ribosomal protein, and its phosphorylated form S6_pS235_S236 has been shown to be part of the response to the drug NVP-AUY92 in oral squamous cell carcinoma cells [24].

Additionally, eight of these 11 proteins, including Cyclin B1, Src_pY527, FoxM1, and MAPK_pT202_Y204, were among the features selected by SFS for the 21 top-performing predictive models. In total, the 21 maximum mean MCC feature sets from SFS included 79 out of the 112 proteins investigated. Of these, 46 proteins appeared in the SFS feature set in at least two models. These are shown in Table II, which also lists the frequencies with which different proteins appeared in the SFS feature sets. MEK1_pS217_S221 is the most frequently selected feature by SFS; it is a component of the MAPK/ERK signal transduction pathway, and has been shown to be involved in radiotherapy response in HNSCC cell lines [25].

IV. Discussion

This study builds and evaluates alternative predictive models for discriminating between early and advanced stage HNSCC using proteomic data. In comparison to previous HNSCC studies on RPPA data, which performed unsupervised and statistical analyses [11, 12], this study uses supervised analysis to identify discriminatory feature sets. In comparison to other studies which performed supervised analysis on RPPA data [13, 14], this study assesses the performances of several different combinations of feature selection methods and classification algorithms in order to identify potentially relevant protein feature sets. One ensemble and four individual binary classifiers, using filter-based and wrapper feature selection, were used to construct several models which can discriminate between early and advanced stage HNSCC with an MCC greater than 0.4 and an AUC above 0.75.

TABLE II. Commonly selected features by SFS among the 21 top-performing models

Times Selected	Number of Features	Proteins
2	22	Annexin_VII, c-Jun_pS73, c-Kit, Claudin-7, Cyclin_E1, EGFR, GATA3, GSK3-alpha-beta_pS21_S9, GSK3_pS9, IRS1, mTOR, N-Ras, NDRG1_pT346, PDK1, PDK1_pS241, Raptor, S6_pS240_S244, SCD1, Smad3, Src_pY527, TAZ, YB-1
3	15	14-3-3_beta, C-Raf, Dvl3, ER-alpha_pS118, FoxM1, HER3, MAPK_pT202_Y204, MIG-6, p27, p70S6K_pT389, RBM15, Smad1, STAT5-alpha, Transglutaminase, YB-1_pS102
4	3	Cyclin_D1, FASN, p38_pT180_Y182
5	5	AR, Collagen_VI, Cyclin_B1, PKC-delta_pS664, Src
9	1	MEK1_pS217_S221

Future research will focus on improving predictive model performance. One important limitation of the current study, which will be addressed in future work, is classifier parameter optimization. Another issue is the limited breadth of feature selection and classification methods currently investigated. Implementing feature selection methods that have been evaluated on other types of proteomic data, as in [26], may help to improve model performance. Additionally, integrating the decisions of different classifiers may also be useful. In this study, Adaboost was tested. The performance of other state-of-the-art ensemble methods, such as random forests, will also be investigated in future work. A related concern is comparison of the execution times for alternative methods.

A notable challenge with this HNSCC RPPA dataset is the imbalance in the class sizes; the advanced stage group contained more than three times the number of patients than the early stage group. This problem is likely to arise for many cancer datasets, because most cancers are detected only at later stages. Thus, future work will also investigate methods for addressing this class imbalance issue [27]. Two related questions of interest are performing multi-class classification to study protein expression differences among all four stages, and studying the differences between healthy and early stage HNSCC samples.

The current results are encouraging with respect to using proteomic data to build predictive models for HNSCC. However, one inherent limitation of RPPA data is that only a selected set of proteins is measured. A larger set of proteins could enable discovery, in the sense that proteins which were previously not associated with HNSCC or cancer may be identified as informative features. TCPA is in the process of extending their antibody set to cover 500 proteins [17], which will help to address this limitation to an extent. A related promising avenue is to analyze mass spectrometry data. Similar to TCPA, The Clinical Proteomic Tumor Analysis Consortium (CPTAC) is currently building a library of LC-MS/MS data from tumor samples that are also in TCGA. Currently, data from colon adenocarcinoma and rectum adenocarcinoma have been released. Future availability of such data for HNSCC would be valuable to researchers. Expanding upon the current results with

additional RPPA datasets or with mass spectrometry data is an important goal.

Lastly, an important direction for HNSCC research is to integrate proteomic data with genomic, transcriptomic, and metabolomic data. Appropriate comparison and integration of multiple data types, as investigated by [14], may improve the performance of predictive models, and may provide greater insight into the mechanisms of disease development and progression. By harnessing the diverse data from initiatives like TCPA, TCGA, and CPTAC, bioinformatics studies can lead to better understanding of the molecular bases of HNSCC and other cancers.

ACKNOWLEDGMENT

The authors thank Mr. Chih-Wen Cheng and Dr. John H. Phan for their helpful input during the preparation of this manuscript.

REFERENCES

[1] C. R. Leemans, B. J. Braakhuis, and R. H. Brakenhoff, "The molecular biology of head and neck cancer," *Nat Rev Cancer,* vol. 11, pp. 9-22, Jan 2011.

[2] M. J. Worsham, "Identifying the risk factors for late-stage head and neck cancer," *Expert Rev Anticancer Ther,* vol. 11, pp. 1321-5, Sep 2011.

[3] K. Chen, R. Sawhney, M. Khan, M. S. Benninger, Z. Hou, S. Sethi, *et al.*, "Methylation of multiple genes as diagnostic and therapeutic markers in primary head and neck squamous cell carcinoma," *Arch Otolaryngol Head Neck Surg,* vol. 133, pp. 1131-8, Nov 2007.

[4] O. Saglam, V. Shah, and M. J. Worsham, "Molecular differentiation of early and late stage laryngeal squamous cell carcinoma: an exploratory analysis," *Diagn Mol Pathol,* vol. 16, pp. 218-21, Dec 2007.

[5] C. E. Schmalbach, D. B. Chepeha, T. J. Giordano, M. A. Rubin, T. N. Teknos, C. R. Bradford, *et al.*, "Molecular profiling and the identification of genes associated with metastatic oral cavity/pharynx squamous cell carcinoma," *Arch Otolaryngol Head Neck Surg,* vol. 130, pp. 295-302, Mar 2004.

[6] M. A. Ginos, G. P. Page, B. S. Michalowicz, K. J. Patel, S. E. Volker, S. E. Pambuccian, *et al.*, "Identification of a gene expression signature associated with recurrent disease in squamous cell carcinoma of the head and neck," *Cancer Res,* vol. 64, pp. 55-63, Jan 1 2004.

[7] E. Mendez, C. Cheng, D. G. Farwell, S. Ricks, S. N. Agoff, N. D. Futran, *et al.*, "Transcriptional expression profiles of oral squamous cell carcinomas," *Cancer,* vol. 95, pp. 1482-94, Oct 1 2002.

[8] S. Tiziani, V. Lopes, and U. L. Gunther, "Early stage diagnosis of oral cancer using 1H NMR-based metabolomics," *Neoplasia,* vol. 11, pp. 269-76, 4p following 269, Mar 2009.

[9] L. Lo Russo, M. Papale, D. Perrone, E. Ranieri, C. Rubini, G. Giannatempo, *et al.*, "Salivary Proteomic Signatures of Oral Squamous Cell Carcinoma," *European Journal of Inflammation,* vol. 10, pp. 61-70, Jan-Apr 2012.

[10] M. Pietrowska, J. Polanska, R. Suwinski, M. Widel, T. Rutkowski, M. Marczyk, *et al.*, "Comparison of peptide cancer signatures identified by mass spectrometry in serum of patients with head and neck, lung and colorectal cancers: association with tumor progression," *Int J Oncol,* vol. 40, pp. 148-56, Jan 2012.

[11] M. J. Frederick, A. J. VanMeter, M. A. Gadhikar, Y. C. Henderson, H. Yao, C. C. Pickering, *et al.*, "Phosphoproteomic analysis of signaling pathways in head and neck squamous cell carcinoma patient samples," *Am J Pathol,* vol. 178, pp. 548-71, 2011.

[12] B. Hong, V. W. Lui, E. P. Hui, Y. Lu, H. S. Leung, E. Y. Wong, *et al.*, "Reverse phase protein array identifies novel anti-invasion mechanisms of YC-1," *Biochem Pharmacol,* vol. 79, pp. 842-52, Mar 15 2010.

[13] A. M. Gonzalez-Angulo, B. T. Hennessy, F. Meric-Bernstam, A. Sahin, W. Liu, Z. Ju, *et al.*, "Functional proteomics can define prognosis and predict pathologic complete response in patients with breast cancer," *Clin Proteomics,* vol. 8, p. 11, 2011.

[14] A. Daemen, O. L. Griffith, L. M. Heiser, N. J. Wang, O. M. Enache, Z. Sanborn, *et al.*, "Modeling precision treatment of breast cancer," *Genome Biol,* vol. 14, p. R110, Oct 31 2013.

[15] M. S. Carey, R. Agarwal, B. Gilks, K. Swenerton, S. Kalloger, J. Santos, *et al.*, "Functional proteomic analysis of advanced serous ovarian cancer using reverse phase protein array: TGF-beta pathway signaling indicates response to primary chemotherapy," *Clin Cancer Res,* vol. 16, pp. 2852-60, May 15 2010.

[16] R. Ummanni, H. A. Mannsperger, J. Sonntag, M. Oswald, A. K. Sharma, R. Konig, *et al.*, "Evaluation of reverse phase protein array (RPPA)-based pathway-activation profiling in 84 non-small cell lung cancer (NSCLC) cell lines as platform for cancer proteomics and biomarker discovery," *Biochim Biophys Acta,* Dec 19 2013.

[17] J. Li, Y. Lu, R. Akbani, Z. Ju, P. L. Roebuck, W. Liu, *et al.*, "TCPA: a resource for cancer functional proteomics data," *Nat Methods,* vol. 10, pp. 1046-7, Nov 2013.

[18] C. G. A. R. Network, "Comprehensive genomic characterization defines human glioblastoma genes and core pathways," *Nature,* vol. 455, pp. 1061-8, Oct 23 2008.

[19] *MD Anderson Cancer Center: Standard Antibody List*. Available: http://www.mdanderson.org/education-and-research/resources-for-professionals/scientific-resources/core-facilities-and-services/functional-proteomics-rppa-core/index.html

[20] T. K. Hoffmann, S. Trellakis, K. Okulicz, P. Schuler, J. Greve, J. Arnolds, *et al.*, "Cyclin B1 expression and p53 status in squamous cell carcinomas of the head and neck," *Anticancer Res,* vol. 31, pp. 3151-7, Oct 2011.

[21] A. M. Egloff and J. R. Grandis, "Targeting epidermal growth factor receptor and SRC pathways in head and neck cancer," *Semin Oncol,* vol. 35, pp. 286-97, Jun 2008.

[22] E. Gemenetzidis, A. Bose, A. M. Riaz, T. Chaplin, B. D. Young, M. Ali, *et al.*, "FOXM1 Upregulation Is an Early Event in Human Squamous Cell Carcinoma and it Is Enhanced by Nicotine during Malignant Transformation," *Plos One,* vol. 4, Mar 2009.

[23] K. Leelahavanichkul, P. Amornphimoltham, A. A. Molinolo, J. R. Basile, S. Koontongkaew, and J. S. Gutkind, "A role for p38 MAPK in head and neck cancer cell growth and tumor-induced angiogenesis and lymphangiogenesis," *Mol Oncol,* Oct 12 2013.

[24] T. Okui, T. Shimo, N. M. Hassan, T. Fukazawa, N. Kurio, M. Takaoka, *et al.*, "Antitumor effect of novel HSP90 inhibitor NVP-AUY922 against oral squamous cell carcinoma," *Anticancer Res,* vol. 31, pp. 1197-204, Apr 2011.

[25] H. Stegeman, J. H. Kaanders, M. M. Verheijen, W. J. Peeters, D. L. Wheeler, M. Iida, *et al.*, "Combining radiotherapy with MEK1/2, STAT5 or STAT6 inhibition reduces survival of head and neck cancer lines," *Mol Cancer,* vol. 12, p. 133, Nov 5 2013.

[26] C. Christin, H. C. Hoefsloot, A. K. Smilde, B. Hoekman, F. Suits, R. Bischoff, *et al.*, "A critical assessment of feature selection methods for biomarker discovery in clinical proteomics," *Mol Cell Proteomics,* vol. 12, pp. 263-76, Jan 2013.

[27] W. J. Lin and J. J. Chen, "Class-imbalanced classifiers for high-dimensional data," *Brief Bioinform,* vol. 14, pp. 13-26, Jan 2013.