This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2015.2489158, IEEE Journal of Biomedical and Health Informatics

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

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

***Abstract*— Late diagnosis is one of the reasons that head and neck squamous cell carcinoma (HNSCC) patients experience relative 5-year survival rates ranging from 40-66%. The molecular-level differences between early and advanced stage HNSCC may provide insight into therapeutic targets and strategies. Previous bioinformatics studies have shown mixed or limited results in identifying gene and protein markers and in developing models for discriminating between early and advanced stage HNSCC. Thus, we have investigated models for HNSCC stage prediction using RNAseq and reverse phase protein array data from The Cancer Genome Atlas and The Cancer Proteome Atlas. We systematically assessed individual and ensemble binary classifiers, using filter and wrapper feature selection methods, to develop several well-performing models. In particular, integrated models harnessing both data types consistently resulted in better performance. This study identifies informative protein and gene feature sets which may increase understanding of HNSCC progression.**

***Index Terms*—Bioinformatics, Computational Biology, Machine Learning, Proteins, RNA**

I. Introduction

HEAD and neck squamous cell carcinoma (HNSCC) is a cancer which arises in regions of the upper aerodigestive tract, including the oral cavity, oropharynx, larynx, hypopharynx, and tongue. It is the 6th most prevalent cancer worldwide, with approximately 600,000 new cases annually [1]. It comprises ~3% of cancers in the U.S., and in 2015, almost 60,000 new cases and more than 12,000 deaths are expected [2]. Patient outcomes are highly associated with the stage at which HNSCC is detected: for early stage (stages I and II) disease, patients have 60-95% chance of successful local treatment, while for advanced stage (stages III, IV, and their sub-types) disease, patients are at high risk for recurrence or metastatic disease [3-5]. Greater knowledge of the molecular characteristics of different stages can provide insight into the mechanisms of HNSCC progression, and may help in identifying more effective targets and strategies for treatment.

Previous research studies have analyzed gene expression, proteomic, and metabolomic data individually for studying differences between HNSCC stages, with mixed results. For example, three transcriptomic studies have related selected genes and gene signatures to different HNSCC stages [6-8], while two other transcriptomic studies did not find any discriminatory genes [9, 10]. A recent proteomic study using SELDI-TOF mass spectrometry data identified eleven m/z values differentially expressed between early- and late-stage oral SCC, but a satisfactory predictive model could not be developed [11]. Another recent study, using MALDI-TOF mass spectrometry data, identified several peaks that tended to correlate with clinical disease progression; however, no predictive model was developed [12]. A metabolomic study using ^{1}H NMR data identified several metabolite markers that discriminated between early and advanced stage HNSCC samples [13]. Additional bioinformatics studies – and in particular, the development of predictive models that harness multiple data types – may help to gain additional insight into the progression from early to advanced HNSCC.

In a recent study, we investigated how quantitative functional proteomics, via reverse phase protein array (RPPA) data, can be used to develop predictive models for HNSCC stage [14]. RPPA data is acquired by probing a sample with antibodies against specific proteins with regard to their activation states. With respect to HNSCC, RPPA data has been used to identify differentially expressed proteins between cancer and normal samples [15] and to identify proteins affected by the presence of an anti-invasion compound in nasopharyngeal carcinoma [16]. RPPA data has been applied to build predictive models for several other cancer types. Recent examples include for prognosis [17], drug response [18], and risk of recurrence [19] in breast cancer; for treatment response in ovarian cancer [20]; and for drug sensitivity in non-small-cell lung cancer [21].

In this paper, we extend our previous work by performing a more in-depth analysis of RPPA data to improve model performance in discriminating between early and advanced stage HNSCC. In addition, we expand upon previous efforts by (i) developing predictive models for the same patient set using RNAseq data, and by (ii) performing integrated analysis of RPPA and RNAseq data through functional assessment and

This research has been supported by grants from National Institutes of Health **U01** CA151802, R01 CA163256, 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 and Emory University, Atlanta, GA 30332 USA (phone: 404-385-5059; e-mail: maywang@bme.gatech.edu).

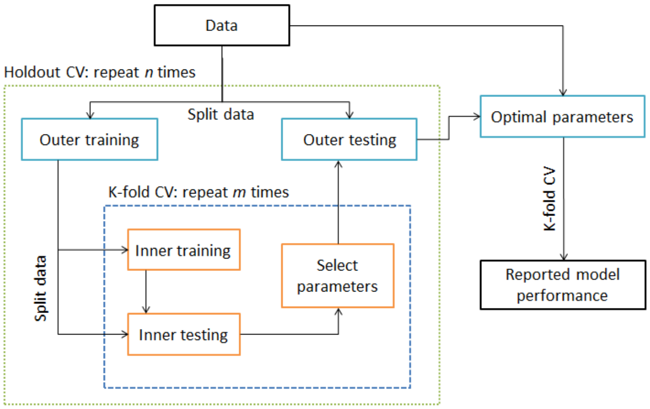


Fig. 1. The nested cross-validation framework used in this study. The outer split was repeated n = 3 times, and the inner 10-fold cross-validation was repeated m = 5 times.

TABLE I:
Classification model parameters examined via nested cross-validation

Classification Method	**Parameters**	**Set of values**
KNN	Number of neighbors (K)	$K \in [1,2,3,4,5,6,7,8,9,10]$
SVM	Kernel Soft margin cost (C) γ for GBRF	Kernels: linear, Gaussian radial basis function (GRBF) C ∈ 2^m, m ∈ [-1,0,1] γ ∈2^m, m ∈ [-1,0,1]
Naïve Bayes	Prior distribution	Distributions: normal, kernel
Decision Tree	Splitting criterion	Criteria: Gini diversity index (GDI), Twoing rule, Maximum deviance reduction (MDR)
Adaboost	Number of trees (N)	$N \in [25,50,100]$
Bagging / Random Forests	Proportion (m) of all variables (p) to retain	$m \in [\sqrt{p}, \frac{p}{4}, \frac{p}{2}, p]$

ensemble model development. The goal of this investigation is to develop a set of improved predictive models, and thereby gather additional insight into HNSCC progression across multiple biological scales.

II. Methods

A. Data

RPPA data for HNSCC was downloaded from The Cancer Proteome Atlas (TCPA) [22] at http://bioinformatics.mdanderson.org/main/TCPA:Overview. This dataset consists of 212 patient samples and measures the response to 187 antibodies. TCPA provides a proteomic complement to The Cancer Genome Atlas (TCGA) [23] at http://cancergenome.nih.gov/, where clinical, transcriptomic, and genomic data for the same patients are available. RNAseq data (Version 2) for HNSCC was downloaded from TCGA. Data was available for 210 of the same patients.

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 [22, 24]. 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 [22, 24] were utilized in this study. 113 proteins were considered for further analysis. In TCGA, RNAseq (Version 2) data has been aligned using MapSplice and quantified using RSEM [25, 26]. This dataset describes 20,531 genes. The un-normalized data was used for differential expression analysis and the normalized data was used for classification.

The clinical data for the 212 patients was downloaded from TCGA. Pathological stage information was used to divide the RPPA and RNAseq datasets into two groups: patients with early stage (stage I and II) cancer, and patients with advanced stage (stage III, IVA, IVB) cancer. Pathological state was unavailable for 12 patients, so clinical stage was substituted. One patient for whom the pathological stage was unavailable and the clinical stage was IVC was not considered, because unlike the other advanced cases, stage IVC involves metastatic disease. For RPPA, the early stage group contained 50 patients, and the advanced stage group contained 161 patients. The two patients for whom RNAseq data was unavailable were both of advanced pathological stage.

B. Predictive Modeling

Four individual binary classification methods and two ensemble classification methods were tested: k-nearest neighbors (KNN), support vector machine (SVM), naïve Bayes, decision tree, Adaboost, and bagging / Random Forests. Optimal parameters for each model were selected via grid search and nested cross-validation. Table I lists the range of parameters tested for each model, and Fig. 1 describes the 3×5×10 nested cross-validation scheme. Optimization was performed with respect to the Matthews correlation coefficient (MCC). The area under the ROC curve (AUC) is also reported for the model having the maximum mean MCC. Analyses were performed using MATLAB (Mathworks, Natick MA).

C. Feature Selection

Three alternative feature selection methods were tested: two filter approaches and one wrapper approach.

The first filter method was based on differential expression. For RPPA data, 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 expressed protein list. This yielded 11 proteins with FDR values ≤ 0.05, including the five proteins found when only 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. For RNAseq data, differential expression analysis was performed using two alternative tools, edgeR and EBSeq, of which the latter uses Bayesian methods [27, 28]. For a threshold of FDR ≤ 0.05, edgeR identified 495 genes and EBSeq found 267 genes. These two lists had 108 genes in common. Due to the large number of differentially expressed genes identified by each method, comprehensive investigation of the feature space was not possible. Instead, model performances were compared across four feature sets: each differential expression result

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2015.2489158, IEEE Journal of Biomedical and Health Informatics

TABLE II:
Performance evaluation of alternative predictive models across feature selection methods for RPPA data

Classification Method	[redacted]	**Rank-Sum Test**		**mRMR**		**SFS**	
	[redacted]	**MCC**	**AUC**	**MCC**	**AUC**	**MCC**	**AUC**
SVM	[redacted]	**0.43±0.15**	**0.75±0.11**	**0.37±0.21**	**0.74±0.08**	**0.54±0.21**	**0.77±0.06**
Naïve Bayes	[redacted]	0.32±0.18	0.71±0.09	0.33±0.13	0.71±0.13	0.47±0.19	0.65±0.12
Decision Tree	[redacted]	0.16±0.17	0.64±0.13	0.12±0.27	0.62±0.11	0.40±0.20	0.68±0.11
KNN	[redacted]	0.35±0.28	0.74±0.13	0.28±0.13	0.77±0.09	0.46±0.22	0.73±0.11
Adaboost	[redacted]	0.11±0.24	0.71±0.11	0.25±0.34	0.71±0.14	0.45±0.13	0.71±0.11
Random Forests	[redacted]	**MCC**			**AUC**		
	[redacted]	0.04±0.16			0.65±0.10		

individually, the 108 common genes, and the 654 genes in the union of the selections of both methods.

The second filter method was mRMR (minimum redundancy maximum relevance), implemented using the FEAST toolbox [29-31]. The performance of each model was optimized for up to the top 50 features. The RNAseq data contained 1,414,819 unique count values, and the vast majority of values were observed only once. Due to this high dynamic range and memory limitations, the count values of the unscaled RNAseq data were binned prior to performing mRMR. The number of binned count levels was chosen to balance performance and computational cost; 30,000 binned levels were the best alternative given the available computational resources.

In the wrapper approach, sequential forward feature selection (SFS) was performed. Model performance was optimized for up to the top 20 features. Due to the large number of genes in the RNAseq data, SFS was performed only after initial filtering based on differential expression. The input to SFS was the 654 genes found to be differentially expressed by edgeR and EBSeq in combination.

D. Data Scaling

Due to the high dynamic range of features in RNAseq data, two data scaling methods were tested. In the first – denoted scaled (1) – each feature was scaled by dividing by the maximum value observed for that feature across any sample. In the second – denoted scaled (2) – each feature was scaled by subtracting its mean and dividing by its standard deviation, as suggested in [32]. The predictive modeling results for RNAseq with SFS are from unscaled data. For the differential expression and mRMR feature selection techniques, the best result among the two alternative scaling choices and unscaled data is shown.

E. Integrated Analysis

One of the fundamental goals of systems biology is to integrate information from multiple levels of biological complexity in order to increase actionable biological and clinical knowledge. However, this is a very challenging task. Several studies have demonstrated the lack of linear correlation between transcriptomic and proteomic data; thus, models developed by integrating mRNA and protein features in some manner may potentially show improved performance over models using individual data types only. In a recent review, Haider and Pal discussed eight frameworks for performing integrated analysis of transcriptomic and proteomic data: union of data types, comparison of functional contexts, topological network analysis, merging datasets in individual domains, missing value estimation, multiple regression analysis, clustering, and dynamic modeling [33]. Due to the constraints of the available data, the techniques of merging datasets in individual domains, missing value estimation, multiple regression analysis, and dynamic modeling are not possible. In this study, we examine model development based on the first two remaining methods: combination of the two data types and the results of functional assessment.

In the first case, RPPA and scaled RNAseq data were naively combined into a composite dataset. One dataset contained 221 features (113 RPPA and the 108 common RNAseq features) and the other contained 767 features (113 RPPA and the 654 union RNAseq features). SVM, KNN, and decision tree models with SFS were constructed using nested CV, with a maximum of 20 features. The better result among the two RNAseq scaling methods is reported.

In the second case, functional analysis of the genes corresponding to RNAseq and RPPA features in the best-performing models was performed using DAVID [34, 35] and the Reactome Analysis Tool [36]. We hypothesized that, if an ensemble of these models was created, individual models representing different functional categories would yield better-performing ensembles. This was tested by systematically evaluating all possible ensembles from nine SFS models: SVM, KNN, naïve Bayes, decision tree, and Adaboost using RPPA data, and SVM, KNN, decision tree, and Adaboost using RNAseq data. Ensemble decisions followed a majority voting scheme, and mean MCC values were compared across 100 repetitions of 10-fold CV.

III. Results

A. Predictive Model Performance

Table II shows the predictive model performance of the six classifiers on the RPPA data. In general, performance is moderate, with several models achieving mean MCC values greater than 0.4 and AUC values greater than 0.7. The best performing RPPA model was SVM with SFS feature selection. The SVM models outperformed the other classifiers for all feature selection methods on the RPPA dataset, and the SFS models outperformed the other feature selection methods for all classifiers. The naïve Bayes and KNN models were the next best in performance, while the decision tree models did not perform as well. The two ensemble classifiers showed markedly different performance. For mRMR and SFS,

TABLE III:
Performance evaluation of alternative predictive models across feature selection methods for RNAseq data.
Legend: unscaled, *scaled (1)*, scaled (2).

Classification Method		**Differential Expression**		**mRMR**		**DEG+SFS**	
		MCC	**AUC**	**MCC**	**AUC**	**MCC**	**AUC**
SVM		**0.52±0.27**	**0.91±0.11**	**0.42±0.27**	**0.50±0**	0.62±0.22	0.50±0
Decision Tree		0.23±0.22	0.62±0.11	*0.36±0.26*	*0.69±0.12*	0.52±0.14	0.74±0.11
KNN		0.35±0.22	0.67±0.08	0.26±0.27	0.68±0.09	**0.64±0.20**	**0.83±0.10**
Adaboost		*0.27±0.28*	*0.66±0.12*	0.32±0.23	0.73±0.10	0.62±0.13	0.73±0.14
Random Forests		**MCC**			**AUC**		
		0.30±0.30			0.79±0.10		

Adaboost outperformed the decision tree models, although it did not perform as well as the other individual classifiers. The Random Forests classifier gave surprisingly poor performance for the RPPA data, with an MCC value close to zero.

Table III shows the predictive model performance of five classifiers on the RNAseq data. The best RNAseq models, which achieve mean MCC values greater than 0.6, outperform the best RPPA models. Again, the SFS models outperformed the other feature selection methods for all classifiers. The highest performing RNAseq model was KNN with SFS; the Adaboost and SVM models with SFS performed almost as well in terms of MCC, though the SVM AUC value was non-informative. The Random Forests model for RNAseq data showed better mean performance than that for RPPA data, but it also had a large standard deviation. For differential expression and mRMR feature selection, the SVM models outperformed the other classifiers in terms of MCC. Under these two feature selection methods, the decision tree and Adaboost models showed better performance for RNAseq data than for RPPA data, but KNN was not notably different. In the majority of cases, scaled data showed better performance, and the second scaling method was more often better than the first.

B. Commonly Selected Features and Functional Analysis

The existence of well-performing models implies that the selected features are of functional importance. The five RPPA SFS models were compared, and 11 features were selected in at least two models. All of these have been associated with HNSCC in the literature: AR [37], C-Raf [38], CDK1 [39], Cyclin B1 [40], MAPK_pT202_Y204 [1], N-Cadherin [41], PDK1 [42], PI3K-p85 [43], VEGFR2 [44], c-Jun_pS73 [45], and p27_pT198 [46]. In particular, AR was selected by four models, CDK1 and Cyclin B1 by three, and the others by two. Table IV shows the number of total common features between each model pair. The low counts show that some models achieved comparable performance using very different feature sets. Even greater feature diversity was observed for the RNAseq SFS models. Among the four models, 52 features were present in total, but only two features were selected in more than one model: FAM27B and KRTAP17-1.

Functional analysis of the SFS feature sets was performed via DAVID and Reactome. DAVID was used to find

TABLE IV:
Comparison and functional analysis of the RPPA SFS models:
The number of features, GO functional annotations, and pathways (KEGG and Reactome) in common between different models are indicated.

	SVM	**Naïve Bayes**	**Decision Tree**	**KNN**	**Adaboost**
SVM Features: 18 GO terms: 10 KEGG: 11 Reactome: 209	-	Features: 2 GO terms: 2 KEGG: 0 Reactome: 93	Features: 2 GO terms: 1 KEGG: 9 Reactome:139	Features: 4 GO terms: 0 KEGG: 7 Reactome:86	Features: 2 GO terms: 0 KEGG: 0 Reactome:67
Naïve Bayes Features: 14 GO terms: 5 KEGG: 0 Reactome: 112	-	-	Features: 3 GO terms: 4 KEGG: - Reactome:97	Features: 1 GO terms: 0 KEGG: - Reactome:82	Features: 1 GO terms: 0 KEGG: - Reactome:59
Decision Tree Features: 11 GO terms: 15 KEGG: 25 Reactome: 208	-	-	-	Features: 2 GO terms: 1 KEGG: 12 Reactome:103	Features: 1 GO terms: 0 KEGG: - Reactome:71
KNN Features: 15 GO terms: 1 KEGG: 12 Reactome: 126	-	-	-	-	Features: 2 GO terms: 0 KEGG: - Reactome:56
Adaboost Features: 8 GO terms: 0 KEGG: 0 Reactome: 131	-	-	-	-	-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2015.2489158, IEEE Journal of Biomedical and Health Informatics

TABLE V:
Performance evaluation of alternative predictive models using two composite RPPA and RNAseq datasets. Legend: *scaled (1)*, scaled (2).

Classification Method		SFS (221 features)		SFS (767 features)	
	[redacted]	MCC	AUC	MCC	AUC
SVM	[redacted]	*0.68±0.15*	*0.82±0.09*	0.70±0.21	0.82±0.14
Decision Tree	[redacted]	0.50±0.20	0.75±0.13	0.46±0.16	0.69±0.10
KNN	[redacted]	0.53±0.21	0.77±0.11	0.64±0.26	0.83±0.13

significantly enriched Gene Ontology (GO) terms and KEGG pathways, while Reactome also returned significant pathways. In terms of specific features and GO terms, the five RPPA SFS models were diverse, with relatively few commonalities. However, many common pathways were found, both through KEGG and through Reactome. Seven KEGG pathways were common among the SVM, decision tree, and KNN RPPA models. These consisted of three signaling pathways: ErbB, neurotrophin, and insulin signaling, and four cancer-related pathways: pathways in cancer, colorectal cancer, pancreatic cancer, and chronic myeloid leukemia. Reactome returned many more significant pathways than DAVID, and 46 pathways were in common among all five models. Most of these related to signal transduction and mitotic progression.

Notably, there were no results in DAVID for the four RNAseq feature lists from the SFS models. Reactome returned results for only the KNN RNAseq model. The nine pathways identified fell into four categories: regulation of gene expression and development in beta cells, visual transduction and phototransduction, retinoid metabolism and transport, and the synthesis of bile acids and bile salts. Retinoids are important therapeutics for many cancer types, including HNSCC [47], and recent studies have shown that bile acids may be associated with head and neck cancer [48, 49].

C. *Integrated Analysis*

The results for developing SFS models based on naïve combination of the RPPA and RNAseq datasets are shown in Table V. All of the models outperformed the corresponding RPPA SFS models for the same classification method in terms of mean MCC values. However, only the SVM models showed improvement over the RNAseq SFS models as well. Moreover, only the models for the smaller composite dataset (221 features) utilized both RPPA and RNAseq features. The RPPA features selected by the SVM model for the smaller composite dataset were Cyclin B1 and p38_pT180_Y182. Cyclin B1 was one of the commonly selected features among the RPPA SFS models; p38 is a mitogen-activated protein kinase that has also been associated with HNSCC [50]. The models for the larger composite dataset (767 features) selected only RNAseq features. Thus, the improvement in MCC seen for the best-performing model (SVM with the larger composite dataset) cannot be attributed to integrating data types, but may be due in part to using scaled data.

Fig. 2 compares the performance of single-data type models (RPPA and RNAseq) with ensembles comprised of only RPPA models, only RNAseq models, or both. The last category contains all possible ensembles with three to nine member models. Results represent the mean performance of 10-fold CV for the 209 common patients in the RPPA and RNAseq datasets, across 100 repetitions. The best single-data type ensembles had higher mean MCC values than individual models of that data type. Additionally, combination ensembles of multiple sizes were found which had better performance than any of the single-data type ensembles.

The performances of the RPPA-only ensembles were compared in terms of the previous functional analysis results. For example, in one iteration of 10-fold CV, the best performing RPPA-only ensemble (SVM, KNN, Adaboost) achieved a mean MCC value of 0.54, and 50 Reactome

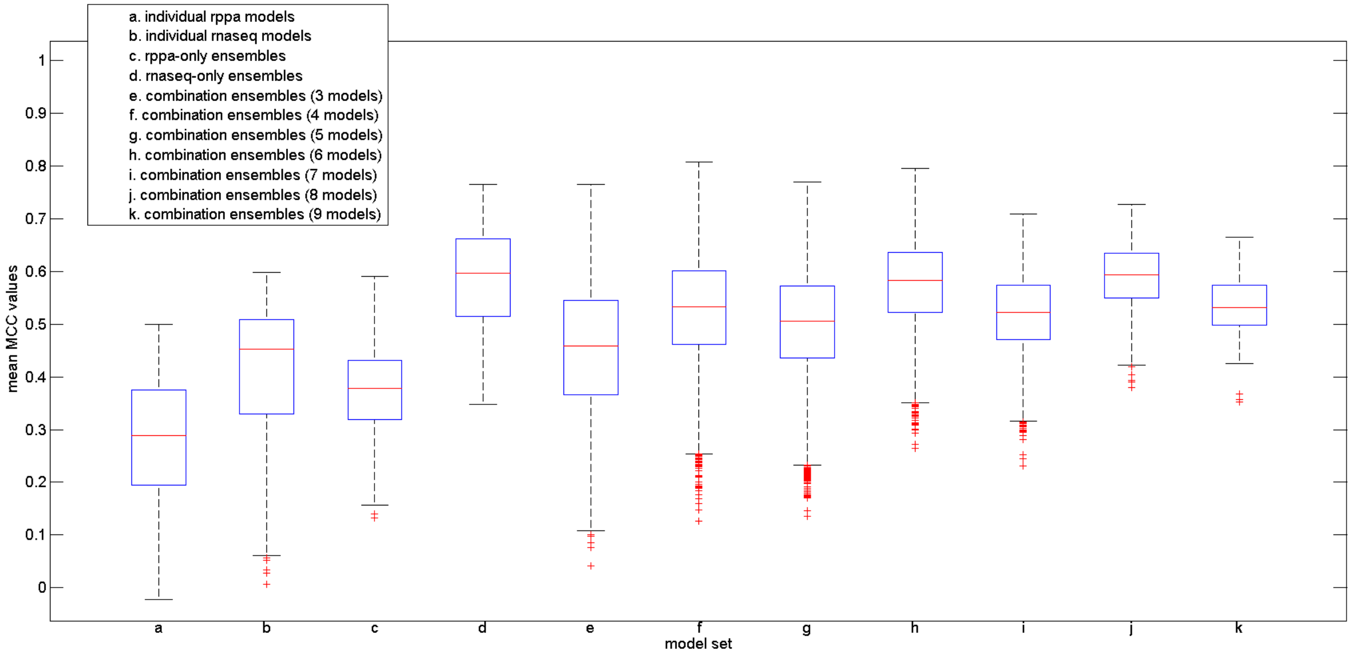


Fig. 2. Comparison of individual and ensemble model performances over 100 repetitions of 10-fold CV. Combination ensembles, which allow for heterogeneity in both data type and component model type, outperform RPPA-only and RNAseq-only models.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2015.2489158, IEEE Journal of Biomedical and Health Informatics

pathways were in common among the three feature sets. The worst-performing ensemble (SVM, Naïve Bayes, decision tree) had a mean MCC of 0.25 and 86 Reactome pathways in common. Among all of the RPPA-only ensembles, a correlation of -0.44 was observed between the mean MCC values and the number of Reactome pathways in common among the ensemble member models.

While the RNAseq ensembles had the highest median performance, several combined RPPA and RNAseq ensembles had higher overall performance. Among all of the ensembles tested – RPPA only, RNAseq only, and combination – 27 ensembles were identified which had better performance than the best-performing individual RNAseq model in more than 90 of the 100 CV repetitions. Of these, two were RNAseq-only ensembles. Another two were combination ensembles containing three and four models, respectively, in which only RNAseq models were chosen as members. The other 23 notable ensembles all contained both RPPA and RNAseq member models.

Among these 23 was the best performing ensemble overall, which achieved a mean (± standard deviation) MCC value of 0.80 (±0.14). This is higher than any of the model performances reported for previous tests. Steiger's Z test was used to compare the MCC performance of this ensemble model with those of the highest performing RNAseq (KNN) and composite (SVM) models [51]. In both cases, the improvement was statistically significant ($p < 0.01$). This particular ensemble incorporated the Adaboost RPPA SFS model and the SVM, KNN, and Adaboost RNAseq SFS models. Leave-one-out analysis of these four models revealed that the KNN RNAseq model contributed the most to the ensemble performance, followed closely by the SVM RNAseq model. Omission of either of these models from the ensemble decreased the mean MCC performance to below 0.60. The contribution of the Adaboost RNAseq and Adaboost RPPA models was not as great, with their omission leading to mean MCC values between 0.71-0.75.

IV. Discussion

In this study, we have built upon our previous work on modeling differences between HNSCC pathological stages. An in-depth analysis of HNSCC RPPA data was performed by implementing six different classification methods, using nested cross-validation to optimize parameters, and testing three alternative feature selection methods. This supervised approach contrasts with previous HNSCC studies using RPPA data, which have conducted unsupervised and differential expression analyses [15, 16]. It also differs from previous supervised studies on RPPA data [17, 18] in two ways. First, 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. Second, this study builds upon current research by developing integrated proteomic and transcriptomic models, and comparing them to RPPA-only and RNAseq-only models. In particular, we performed two types of integrated analysis: one by direct combination of RPPA and RNAseq data, and another by constructing ensemble models using both data types. To our knowledge, this is the first such comparative, integrated study for modeling progression in HNSCC.

From a modeling perspective, we identified the integrated ensemble approach with both RPPA and RNAseq models as the best overall. The top-performing model for predicting HNSCC pathological stage was obtained using this approach, and had a significantly higher MCC value than the best performing individual RNAseq and composite models. Notably, modeling results appear to support the initial conjecture that less functional agreement among the feature sets of member models will be associated with better performance. First, the RNAseq-only and the combination RPPA and RNAseq ensembles outperformed the RPPA-only ensembles. Second, a moderate negative correlation was observed between the performances of RPPA-only ensembles and the numbers of common Reactome pathways among ensemble members. These observations indicate that higher-performing ensembles tended to be more functionally diverse in terms of member model feature sets. However, the implications of this observation should be explored further, particularly for the integrated ensembles, because they highlight components of complex gene and protein networks. Investigation on larger datasets, as well as assessment using ensemble diversity measures and different ensemble construction techniques [52], are directions for further research. The comparison of performance trends with those of other cancers, particularly cancers which have recently been shown to have molecular-level similarities to HNSCC, may also yield additional insight [53].

A related question of interest is performing multi-class classification to study bio-molecular expression patterns among individual HNSCC stages, rather than grouping them into early and advanced disease. Another is investigating the differences between normal and early stage HNSCC samples. For investigating these questions, the availability of sufficiently large – in terms of patients and features – public datasets is a constraint. While matched tumor and normal RNAseq data is available on TCGA for HNSCC, RPPA data for matched normal samples is yet unavailable. In addition, an inherent limitation of RPPA data is that only a selected set of proteins is measured. A larger set of proteins could enable discovery, in that proteins which were previously not implicated in HNSCC – or cancer in general – may be identified as informative features through modeling. TCPA is currently in the process of extending their antibody set to cover 500 proteins [22], which will help to address this limitation to some extent. The availability of more extensive proteomic data for HNSCC through mass spectrometry is a related promising avenue. The Clinical Proteomic Tumor Analysis Consortium (CPTAC), like TCPA, is currently building a proteomic complement to TCGA. CPTAC hosts a library of LC-MS/MS data from tumor samples that are also in TCGA. At the time of writing, data from breast cancer, ovarian cancer, colon adenocarcinoma, and rectum adenocarcinoma have been released. Future availability of such data for HNSCC would be valuable to researchers.

From a systems biology perspective, investigating multiple types of –omic datasets to gain insight into disease processes is an important area of research. Numerous individual proteins and genes selected as features in well-performing models in this study have been previously associated with HNSCC in the literature, including in a recent large-scale study by The Cancer Genome Atlas Network [54]. Additionally, functional analysis

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2015.2489158, IEEE Journal of Biomedical and Health Informatics

of the features selected in the top-performing models revealed notable patterns. Many processes – e.g., signal transduction pathways including those through EGFR and ERBB2, and events related to mitotic progression – were commonly represented among the RPPA model features. The RNAseq feature sets were much more diverse, but some of the associated biological processes have still been linked with HNSCC in the literature.

While this integrative modeling study of RPPA and RNAseq data can provide guidance for further research, integration in general should be interpreted with caution. Because RPPA is a tool for functional proteomics, it is several biological steps removed from the mRNA counts measured by RNAseq, and mRNA is itself distinct from genome-level factors. Thus, further investigation into additional data types – e.g., copy number variations, mutations, DNA methylation, protein subunits and alternative activation states, metabolites – is needed for drawing conclusions about the specific mechanisms underlying HNSCC progression. Appropriate comparison and combination of multiple data types will help to fill in the gaps and provide greater insight into the process of disease development. 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.

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] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2015," *CA Cancer J Clin,* vol. 65, pp. 5-29, Jan 2015.

[3] D. Pulte and H. Brenner, "Changes in survival in head and neck cancers in the late 20th and early 21st century: a period analysis," *Oncologist,* vol. 15, pp. 994-1001, 2010.

[4] 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.

[5] G. Gatta, L. Botta, M. J. Sánchez, L. A. Anderson, D. Pierannunzio, and L. Licitra, "Prognoses and improvement for head and neck cancers diagnosed in Europe in early 2000s: The EUROCARE-5 population-based study," *European Journal of Cancer,* vol. 51, pp. 2130-2143, 10// 2015.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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, 2012.

[12] 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.

[13] 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.

[14] C. D. Kaddi and M. D. Wang, "Models for predicting stage in head and neck squamous cell carcinoma using proteomic data," in *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE*, 2014, pp. 5216-5219.

[15] 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, Feb 2011.

[16] 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.

[17] 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.

[18] 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, 2013.

[19] J. Sonntag, C. Bender, Z. Soons, S. v. der Heyde, R. König, S. Wiemann, *et al.*, "Reverse phase protein array based tumor profiling identifies a biomarker signature for risk classification of hormone receptor-positive breast cancer," *Translational Proteomics,* vol. 2, pp. 52-59, 3// 2014.

[20] 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.

[21] 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.

[22] 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.

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

[24] D. K. Gascoigne, S. W. Cheetham, P. B. Cattenoz, M. B. Clark, P. P. Amaral, R. J. Taft, *et al.*, "Pinstripe: a suite of programs for integrating transcriptomic and proteomic datasets identifies novel proteins and improves differentiation of protein-coding and non-coding genes," *Bioinformatics,* vol. 28, pp. 3042-3050, December 1, 2012 2012.

[25] B. Li, V. Ruotti, R. M. Stewart, J. A. Thomson, and C. N. Dewey, "RNA-Seq gene expression estimation with read mapping uncertainty," *Bioinformatics*, vol. 26, pp. 493-500, Feb 15 2010.

[26] K. Wang, D. Singh, Z. Zeng, S. J. Coleman, Y. Huang, G. L. Savich, *et al.*, "MapSplice: accurate mapping of RNA-seq reads for splice junction discovery," *Nucleic Acids Res,* vol. 38, p. e178, Oct 2010.

[27] N. Leng, J. A. Dawson, J. A. Thomson, V. Ruotti, A. I. Rissman, B. M. G. Smits, *et al.*, "EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments," *Bioinformatics,* vol. 29, pp. 1035-1043, April 15, 2013 2013.

[28] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data," *Bioinformatics*, vol. 26, pp. 139-40, Jan 1 2010.

[29] G. Brown, A. Pocock, Z. Ming-Jie, and M. Luján, "Conditional Likelihood Maximisation: A Unifying Framework for Information Theoretic Feature Selection," *Journal of Machine Learning Research,* vol. 13, pp. 27-66, 2012.

[30] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *J Bioinform Comput Biol,* vol. 3, pp. 185-205, Apr 2005.

[31] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans Pattern Anal Mach Intell,* vol. 27, pp. 1226-38, Aug 2005.

[32] D. Bollegala, "Dynamic Feature Scaling for Online Learning of Binary Classifiers," *arXiv:1407.7584v1,* 2014.

[33] S. Haider and R. Pal, "Integrated Analysis of Transcriptomic and Proteomic Data," *Current Genomics,* vol. 14, pp. 91-110, 2013.

[34] W. Huang da, B. T. Sherman, and R. A. Lempicki, "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources," *Nat Protoc,* vol. 4, pp. 44-57, 2009.

[35] W. Huang da, B. T. Sherman, and R. A. Lempicki, "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists," *Nucleic Acids Res,* vol. 37, pp. 1-13, Jan 2009.

[36] D. Croft, A. F. Mundo, R. Haw, M. Milacic, J. Weiser, G. Wu*, et al.*, "The Reactome pathway knowledgebase," *Nucleic Acids Res,* vol. 42, pp. D472-7, Jan 2014.

[37] A. K. Goulioumis, J. Varakis, P. Goumas, and H. Papadaki, "Androgen Receptor in Laryngeal Carcinoma: Could There Be an Androgen-Refractory Tumor?," *ISRN Oncology,* vol. 2011, p. 5, 2011.

[38] C. Elser, L. L. Siu, E. Winquist, M. Agulnik, G. R. Pond, S. F. Chin*, et al.*, "Phase II Trial of Sorafenib in Patients With Recurrent or Metastatic Squamous Cell Carcinoma of the Head and Neck or Nasopharyngeal Carcinoma," *Journal of Clinical Oncology,* vol. 25, pp. 3766-3773, August 20, 2007 2007.

[39] J. T. Chang, H.-M. Wang, K.-W. Chang, W.-H. Chen, M.-C. Wen, Y.-M. Hsu*, et al.*, "Identification of differentially expressed genes in oral squamous cell carcinoma (OSCC): Overexpression of NPM, CDK1 and NDRG1 and underexpression of CHES1," *International Journal of Cancer,* vol. 114, pp. 942-949, 2005.

[40] Y. Song, C. Zhao, L. Dong, M. Fu, L. Xue, Z. Huang*, et al.*, "Overexpression of cyclin B1 in human esophageal squamous cell carcinoma cells induces tumor cell invasive growth and metastasis," *Carcinogenesis,* vol. 29, pp. 307-315, February 1, 2008 2008.

[41] S. W. Pyo, M. Hashimoto, Y. S. Kim, C. H. Kim, S. H. Lee, K. R. Johnson*, et al.*, "Expression of E-cadherin, P-cadherin and N-cadherin in oral squamous cell carcinoma: correlation with the clinicopathologic features and patient outcome," *J Craniomaxillofac Surg,* vol. 35, pp. 1-9, Jan 2007.

[42] P. Amornphimoltham, V. Sriuranpong, V. Patel, F. Benavides, C. J. Conti, J. Sauk*, et al.*, "Persistent activation of the Akt pathway in head and neck squamous cell carcinoma: a potential target for UCN-01," *Clin Cancer Res,* vol. 10, pp. 4029-37, Jun 15 2004.

[43] C. Freudlsperger, J. R. Burnett, J. A. Friedman, V. R. Kannabiran, Z. Chen, and C. Van Waes, "EGFR–PI3K–AKT–mTOR signaling in head and neck squamous cell carcinomas: attractive targets for molecular-oriented therapy," *Expert Opinion on Therapeutic Targets,* vol. 15, pp. 63-74, 2011.

[44] K. A. Gold, H.-Y. Lee, and E. S. Kim, "Targeted therapies in squamous cell carcinoma of the head and neck," *Cancer,* vol. 115, pp. 922-935, 2009.

[45] A. Weber, U. R. Hengge, I. Stricker, I. Tischoff, A. Markwart, K. Anhalt*, et al.*, "Protein microarrays for the detection of biomarkers in head and neck squamous cell carcinomas," *Human Pathology,* vol. 38, pp. 228-238, 2// 2007.

[46] P. Lothaire, E. de Azambuja, D. Dequanter, Y. Lalami, C. Sotiriou, G. Andry*, et al.*, "Molecular markers of head and neck squamous cell carcinoma: Promising signs in need of prospective evaluation," *Head & Neck,* vol. 28, pp. 256-269, 2006.

[47] X.-H. Tang and L. J. Gudas, "Retinoids, Retinoic Acid Receptors, and Cancer," *Annual Review of Pathology: Mechanisms of Disease,* vol. 6, pp. 345-364, 2011/02/28 2011.

[48] E. De Corso, S. Baroni, S. Agostino, G. Cammarota, G. Mascagna, A. Mannocci*, et al.*, "Bile Acids and Total Bilirubin Detection in Saliva of Patients Submitted to Gastric Surgery and in Particular to Subtotal Billroth II Resection," *Annals of Surgery,* vol. 245, pp. 880-885, 2007.

[49] M.-W. Sung, J.-L. Roh, B. J. Park, S. W. Park, T.-K. Kwon, S. J. Lee*, et al.*, "Bile Acid Induces Cyclo-Oxygenase-2 Expression in Cultured Human Pharyngeal Cells: A Possible Mechanism of Carcinogenesis in the Upper Aerodigestive Tract by Laryngopharyngeal Reflux," *The Laryngoscope,* vol. 113, pp. 1059-1063, 2003.

[50] M. R. Junttila, R. Ala-aho, T. Jokilehto, J. Peltonen, M. Kallajoki, R. Grenman*, et al.*, "p38[alpha] and p38[delta] mitogen-activated protein kinase isoforms regulate invasion and growth of head and neck squamous carcinoma cells," *Oncogene,* vol. 26, pp. 5267-5279, 03/05/online 2007.

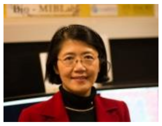
[51] J. H. Steiger, "Tests for comparing elements of a correlation matrix," *Psychological Bulletin,* vol. 87, pp. 245-251, 1980.

[52] Z.-H. Zhou, *Ensemble Methods: Foundations and Algorithms*, 1st ed.: Chapman and Hall/CRC, 2012.

[53] K. A. Hoadley, C. Yau, D. M. Wolf, A. D. Cherniack, D. Tamborero, S. Ng*, et al.*, "Multiplatform Analysis of 12 Cancer Types Reveals Molecular Classification within and across Tissues of Origin," *Cell,* vol. 158, pp. 929-944.

[54] T. C. G. A. Network, "Comprehensive genomic characterization of head and neck squamous cell carcinomas," *Nature,* vol. 517, pp. 576-82, Jan 29 2015.

Chanchala D. Kaddi, Ph.D. received the B.S. in Biomedical Engineering, the M.S. in Electrical and Computer Engineering, and the Ph.D. in Bioengineering from the Georgia Institute of Technology in 2008, 2014, and 2015 respectively. She is currently a postdoctoral research fellow in the Wallace H. Coulter Department of Biomedical Engineering at the Georgia Institute of Technology. She has been a National Science Foundation Graduate Research Fellow and a P.E.O. Scholar.



May D. Wang, Ph.D. received her B.S. degree from Tsinghua University, Beijing, China, in 1989, and three M.S. degrees and Ph.D. from Georgia Institute of Technology in 1991, 1993, 1995 and 2000 respectively. Since 2001, she has been an Assistant Professor with the Wallace H. Coulter Department of Biomedical Engineering at Georgia Institute of Technology. In 2004 she received the Georgia Cancer Coalition Distinguished Cancer Scholar award. She was also Director of Biocomputing and Bioinformatics Core in Emory-Georgia Tech Center of Cancer Nanotechnology Excellence. Her research focuses on Biomedical Computing and Modeling such as Biomedical Informatics, Bio-Molecular and Medical Imaging Data Processing, Data Management and Visualization, Bio-molecular Pathway Modeling, and Telemedicine.