HHS Public Access

Author manuscript

Conf Proc IEEE Eng Med Biol Soc. Author manuscript; available in PMC 2016 June 23.

Published in final edited form as:

Conf Proc IEEE Eng Med Biol Soc. 2015 ; 2015: 7226–7229. doi:10.1109/EMBC.2015.7320059.

Robust Automatic Breast Cancer Staging Using A Combination of Functional Genomics and Image-Omics

Hai Su[1], **Yong Shen**[2], **Fuyong Xing**[1], **Xin Qi**[3], **Kim M. Hirshfield**[3], **Lin Yang**[1], and **David J. Foran**[3]

[1]J. Crayton Pruitt Family Department of Biomedical Engineering, University of Florida, Gainesville, FL 32611, USA

[2]Genetics Institute, University of Florida, Gainesville, FL 32611, USA scientificsy@ufl.edu

[3]Rutgers Cancer Institute of New Jersey, New Brunswick, NJ 08901, USA

Abstract

Breast cancer is one of the leading cancers worldwide. Precision medicine is a new trend that systematically examines molecular and functional genomic information within each patient's cancer to identify the patterns that may affect treatment decisions and potential outcomes. As a part of precision medicine, computer-aided diagnosis enables joint analysis of functional genomic information and image from pathological images. In this paper we propose an integrated framework for breast cancer staging using image-omics and functional genomic information. The entire biomedical imaging informatics framework consists of image-omics extraction, feature combination, and classification. First, a robust automatic nuclei detection and segmentation is presented to identify tumor regions, delineate nuclei boundaries and calculate a set of image-based morphological features; next, the low dimensional image-omics is obtained through principal component analysis and is concatenated with the functional genomic features identified by a linear model. A support vector machine for differentiating stage I breast cancer from other stages are learned. We experimentally demonstrate that compared with a single type of representation (image-omics), the combination of image-omics and functional genomic feature can improve the classification accuracy by 3%.

I. Introduction

Breast cancer is one of the leading cancers worldwide. Meanwhile, it is the principle cause of death among women who are diagnosed cancers [1]. The American Cancer Society estimated that 207,090 women were diagnosed with breast cancer in 2010 and that 39,840 women died of this disease in the United States alone during that year. Breast cancers mainly consist of three stages I,II and III. Accurate staging can provide support for personalized medicine, which aims to provide an individualized therapy design and outcome prediction for a particular patient based on systematically examining the molecular and genomic information of the patients in the database, and therefore increase the survival rates. Nowadays, many state of the arts [2], [3] using image features have been applied to breast

hsu224@ufl.edu f.xing@ufl.edu lin.yang@bme.ufl.edu qixi@cinj.rutgers.edu, hirshfie@rutgers.edu, and foran@cinj.rutgers.edu

cancer classification. In [4], Beck *et al.* have correlated image features from stromal with survival in breast cancer study and discovered new biological aspects of cancer tissues. Meanwhile, many machine learning-based methods using image features are also reported for other specimen classification, such as prostate grading [5] and colon cancer subtype classification [6]. Besides the image signatures, genomic information is also used for breast cancer grading [7].

However, using a single type of features (image-omics) might be not sufficient for accurate breast cancer staging. It is well-known that genomic information provide significant support in clinical practice. In this paper, we propose to differentiate the stage I breast cancer from other stages by using a combination of image-omics and functional genomic information. In this paper, we focus on nuclear morphological patterns. In order to calculate image features of nuclei, we adopt a hierarchical voting based nucleus detection and a repulsive balloon snake model based nucleus segmentation. The non-tumor regions are delineated by a pixel-wise classification based on convolutional neural network. Based on nucleus segmentation, we extract a set of image features including geometry information, intensity statistics, and texture. Principal component analysis (PCA) is used to reduced the original features to 47-dimension image-omics. The obtained image-omics is then concatenated with functional genomic information for breast cancer classification. The experiments demonstrate that the classification accuracy can be improved by using a combination of these two types of signatures.

II. Methods

A. Automatic Nucleus Detection and Segmentation

Nucleus Detection via Hierarchical Voting—Narrowing down our information aggregation to individual nucleus is critical to the analysis of histopathological images, which includes nucleus detection and segmentation. We apply an improved variant of the single-pass voting (SPV) in [8] to nucleus detection, which is more robust to the variations in nucleus size. For an image $I(x, y)$, a Gaussian pyramid with L layers is created for hierarchical voting. At layer l, a Gaussian kernel weighted SPV is applied to the distance transform map, and this weighted voting strategy generates higher voting scores in the central region of the cells, even the nuclei have moderately elongated shapes. The final voting map $V(x, y)$ is calculated by summing up all the layers [9]:

$$V(x,y) = \sum_{l=0}^{L} \sum_{(m,n)\in S} I\left[(x,y) \in A_l(m,n)\right] \cdot C_l(x,y)\, g(m,n,\mu_x,\mu_y,\sigma), \tag{1}$$

where S denotes the set of all voting pixels, $A_l(m, n)$ denotes the voting area of pixel (m, n) at layer l. $I[\cdot]$ is an indicator function, and $C_l(x, y)$ represents the distance transformation map at layer l. $g(m, n, \mu_x, \mu_y, \sigma)$ is a Gaussian kernel with centering on a vote accumulating pixel (x, y) with isotropic covariance σ [8]. In the voting map $V(x, y)$, the pixels in the central regions of nucleus achieve higher voting scores than the others, so a local clustering method, mean shift [10], is applied to geometric center localization, which is considered as

the final nucleus detection and will be used to initialize the following seed-based deformable model for nucleus segmentation.

Tumor Region Segmentation—During the previous detection, all the round shaped objects are detected without differentiation of the tumor nuclei and the non-tumor objects. A deep convolutional neural network (CNN) is trained to identify the tumor regions and therefore to remove the non-tumor objects from feature extraction. The trained CNN consists of 7 layers, including two convolutional layers with kernel size 5 × 5, two max-pooling layers with kernel size 2×2, and one fully connected layer of dimension 64 and one output layer of dimension 2. The CNN is used to perform pixel-wise classification via sliding window mechanism [11].

Nucleus Segmentation via Repulsive Balloon Snake—Different than the traditional balloon snake deformable model [12], the repulsive balloon snake introduces an extra term that models the repulsive force from the neighboring contours to prevent evolving contours from crossing each other. Let $v_i(s)$ denote the i-th contour and s index points on the contour, the model deforms $v_i(s)$ till a balance between the internal force $F^{int}(v_i)$ and external force $F^{ext}(v_i)$ is achieved:

$$F^{int}(v_i) + F^{ext}(v_i) = 0, \quad (2)$$

$$F^{int}(v_i) = \alpha v_i''(s) - \beta v_i''''(s), \quad (3)$$

$$F^{ext}(v_i) = \gamma n_i(s) - \lambda \frac{\nabla E_{ext}(v_i(s))}{\|E_{ext}(v_i(s))\|} + \omega \sum_{j=1, j\neq i}^{N} \int d_{ij}^{-2}(s,t)\, n_j(t)\, dt, \quad (4)$$

where the first/second term in (3) represents the seconde/fourth derivative of $v_i(s)$, which are used to model the internal force. Their contributions to contour deformation are controlled by weights α and β. In (4), the $\gamma n_i(s)$ represents the pressure force, and $\nabla E_{ext}(v_i(s))$ denotes the image force, where $E_{ext}(v_i(s))$ is the magnitude of the gradient of the image. The third term in (4) denotes the repulsive force, with N representing the number of neighboring contours and $d_{ij}(s, t)$ corresponding to the Euclidean distance between contours $v_i(s)$ and $v_j(t)$. Based on the aforementioned nucleus detection, the contours are initialized by generating small circles, one per detected nucleus. The automatic nucleus segmentation is achieved when (2) is satisfied.

B. Feature Extraction

Based on the nucleus segmentation, we extract three groups of cellular features including geometry features, statistics of pixel intensity and texture features, which are combined with functional genomic information to separate stage I breast cancer from other stages.

Geometry features—We compute nuclear area, perimeter, circularity, major-minor axis ratio, and solidity (It is defined as the ratio of cell area region over the convex hull defined by the segmented nuclei boundary).

Statistics of pixel intensity—This group of features are calculated based on the pixels within the segmented nuclei, including intensity mean, standard deviation, skewness, kurtosis, entropy, and energy.

Texture features—They contain texture feature coding method (TFCM), center symmetric autocorelation (CSAC), local binary pattern (LBP). In TFCM, Each pixel is assigned a texture feature number (TFN), which is generated by comparing this pixel with its neighbors in four directions. A histogram can be generated based on the TFNs of one image patch for feature description. CSAC is a measure of the local patterns with symmetrical forms. LBP is one type of local features, which assigns each pixel binary code by comparing the intensity of this pixel to those of its neighbors and then a histogram of the binary codes is created to characterize texture information. In total we have extracted 163 image features. The image-omics is computed by applying PCA to the original high dimensional image features. A 47-dimension image-omics is obtained.

Functional genomic information—The transcriptomic data are generated from 86 breast cancer patients, which contains 20,503 gene entries. In order to reduce the feature dimensionality and avoid overfitting, we apply linear modeling and empirical Bayes statistics to filter out genes without significant expression. The top-10 genes with $P < 0.001$ are listed in Table I. The two genes with top expression levels, C19orf33 and SLC4A8, with respect to the stages, are summarized in Figure 1. One-way ANOVA analysis for C19orf33 and SLC4A8 indicates significant difference for their expression based on tumor stages. Further, their association is necessary to be validated in other replicative cohorts. The C19orf33 and SLC4A8 are highly suggested into a predictive model for breast tumor staging, which can provide a useful molecular signatures for the diagnosis of breast cancers. In the step of breast cancer staging, we only combine these two functional genomic features with the image features for classification.

C. Breast Cancer Staging

Based on the image features and the functional genomic information, we exploit a binary classifier, support vector machine (SVM), to differentiate the stage I breast cancer from other stages. Given training data $(x_i, y_i)_{i=1}^N$, the binary SVM aims to maximize the margin with allowing a certain degree ξ of misclassification of the data:

$$arg\ \min_{w,\xi,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i,\ s.t.\ y_i\left(w^T x_i - b\right) \geq 1 - \xi_i, \xi_i \geq 0,$$

where C is the penalty parameter controlling the trade off between the margins and training errors.

III. Experiments

Data—We compiled a data set containing the diagnostic images, genome information and clinic information of 86 patients. The data is downloaded from the publicly available database TCGA (The Cancer Genome Atlas). Patient of three stages, including 13 stage I, 55 stage II and 18 stage III, are present in the data set.

Tumor nuclei detection and segmentation—A randomly selected image patch is shown in Figure 4(a). The nuclei detection algorithm described in Section II-A is used to detect the nuclei automatically (Figure 4(b)). As one can tell, it is possible that both the tumor nuclei and non-tumor objects are detected. The region segmentation method described in Section II-A is used to separate tumor regions and non-tumor regions (Figure 4(c)). The detected objects in the non-tumor regions are removed from the analysis. The final automatic tumor nuclei detection and segmentation result is shown in Figure 4(d).

Feature extraction and classification—The image features described in Section II-B are extracted and combined with the expression of the two functional genomic features identified in Section II-B to train a support vector machine (SVM) classifier. The features are normalized by subtracting the mean and being divided by the standard deviation. Since the morphological features have very high dimension. For the cellular geometric features, we compute a concise representation to capture the characteristics of all the detected nuclei in one image by computing their mean, variance, and median and a three-bin histogram. Similar to the geometric features, the statistics features of the pixel intensity is calculated for each nucleus and the mean, variance, and median and three-bin histogram are calculated to represent the image. For the texture features, including TFCM, CSAC, and LBP, principal component analysis (PCA) is used to compute the dimension reduced representations. Finally each image is represented by a 163 dimensional vector. Finally, a image-omics with dimension of 47 is computed by further applying PCA to the 163 dimensional image features. Combining the two functional genomic features, each patient is represented by a 49 dimensional vector.

A SVM classifier is trained to separate the patients at stage I and the other stages. A fourfold cross validation is applied to evaluate the classifier with respect to the parameters on a Gaussian kernel. The cross validation misclassificaton rates are shown in Figure 2. The parameter γ denotes the variance of the Gaussian kernel and C denotes the penalty. The best error rate with image features only is 15% as shown in Figure 2(a). The classification performance with both the image features and the two functional genomic features is 12% as shown in Figure 2(b). The ROC curves of the performance of the classifier with image features only and with the joint features are shown in Figure 3. As one can tell that the classification performance obtained by the combined feature is superior to the performance based on the image features only.

IV. Conclusion

In this paper, we proposed a breast cancer staging system based on joint analysis of morphological features and functional genomic information. The proposed system is verified experimentally on a data set containing 86 patients. The cross validation shows that morphological features together functional genomic information produce a better classification performance in differentiating stage I patients and patients at other stages. The proposed system is generic. Therefore, more morphological features and more elaborated analysis of the functional genomic can be integrated. The experiment shows promising results. In our future study, more morphological features, *e.g.*, structural features, and more genomic/clinic information will considered in the investigation.

Acknowledgements

This research was funded, in part, by grants from NIH contract 5R01CA156386-10 and NCI contract 5R01CA161375-04, NLM contracts 5R01LM009239-06 and 5R01LM011119-04.

References

[1]. Ferlay, J.; Soerjomataram, I.; Ervik, M.; Dikshit, R.; Eser, S.; Mathers, C.; Rebelo, M.; Parkin, DM.; Forman, D.; Bray, F. GLOBOCAN 2012 v1.1, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11 [Internet]. v1.1 edition. International Agency for Research on Cancer; Lyon, France: 2014.

[2]. Ojansivu, Ville; Linder, Nina; Rahtu, Esa; Pietikinen, Matti; Lundin, Mikael; Joensuu, Heikki; Lundin, Johan. Automated classification of breast cancer morphology in histopathological images. Diagnostic Pathology. 2013; 8(1)

[3]. Kowal, Marek. Computer-aided diagnosis for breast tumor classification using microscopic images of fine needle biopsy. Intelligent Systems in Technical and Medical Diagnostics. 2014; 230:213–224.

[4]. Beck AH, Sangoi AR, Leung S, Marinelli RJ, Nielsen TO, van de Vijver MJ, Koller D. Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. Science Translational Medicine. 2011; 2(108) 108ra113.

[5]. Ali, Sahirzeeshan; Veltri, Robert; Epstein, Jonathan I.; Christudass, Christhunesa; Madabhushi, Anant. Selective invocation of shape priors for deformable segmentation and morphologic classification of prostate cancer tissue microarrays. Computerized Medical Imaging and Graphics. 2015; 41(0):3–13. [PubMed: 25466771]

[6]. Xu, Yan; Zhu, Jun-Yan; Chang, Eric I-Chao; Lai, Maode; Tu, Zhuowen. Weakly supervised histopathology cancer image segmentation and classification. Medical Image Analysis. 2014; 18(3):591–604. [PubMed: 24637156]

[7]. Ali H, Rueda Oscar, Chin Suet-Feung, Curtis Christina, Dunning Mark, Aparicio Samuel, Caldas Carlos. Genome-driven integrated classification of breast cancer validated in over 7,500 samples. Genome Biology. 2014; 15(8):431. [PubMed: 25164602]

[8]. Qi X, Xing F, Foran DJ, Yang L. Robust segmentation of overlapping cells in histopathology specimens using parallel seed detection and repulsive level set. TBME. Mar; 2012 59(3):754–765.

[9]. Xing, Fuyong; Su, Hai; Neltner, J.; Yang, Lin. Automatic ki-67 counting using robust cell detection and online dictionary learning. TBME. Mar; 2014 61(3):859–870.

[10]. Comaniciu D, Meer P. Mean shift: a robust approach toward feature space analysis. Pattern Analysis and Machine Intelligence, IEEE Transactions on. 2002; 24(5):603–619.

[11]. Su, Hai; Liu, Fujun; Xie, Yuanpu; Xing, Fuyong; Meyyappan, Sreenivasan; Yang, Lin. Region segmentation in histopathological breast cancer images using deep convolutional neural network. Proc. of ISBI. 2015

[12]. Cohen, Laurent D. On active contour models and balloons. CVGIP: Image Understanding. 1991; 53(2):211–218.

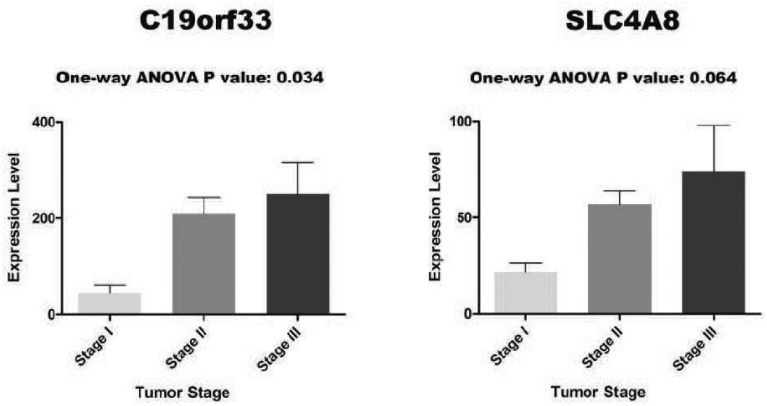

Fig. 1.

Bargraph of gene expression values for gene C19orf33 and SLC4A8 discovered from linear modeling with breast tumor stages.

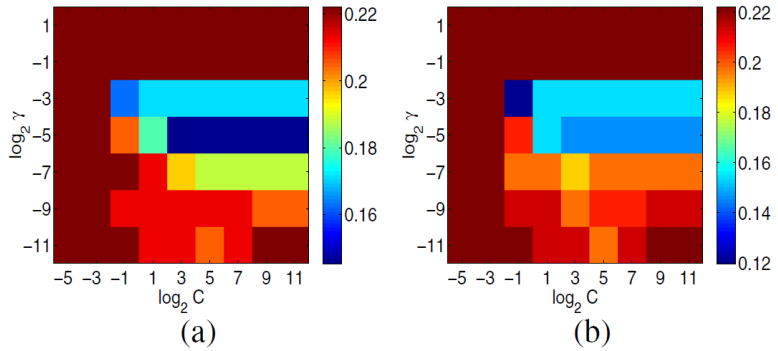

Fig. 2.

Misclassification rates with respect to different parameters using Gaussian kernel to differentiate the patients at stage I and the other stages. (a) The misclassification rates based on image-omics only; (b) The misclassification rate based on image-omics and the two identified functional genomic features.

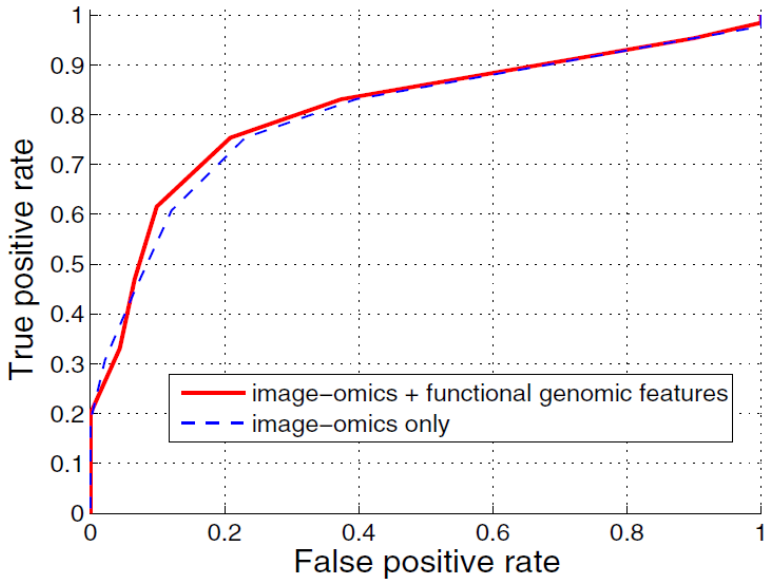

Fig. 3.

The receiver operation curves of the classification with image-omics only and a combination of the morphological features and the two identified functional genomic features.

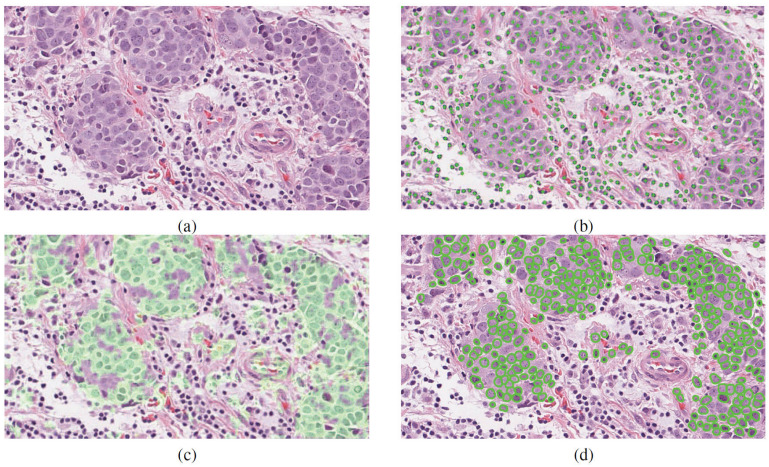

Fig. 4.

(a) A randomly picked original image; (b) The nuclei detection result; (c) The tumor region (marked in light green) segmentation result; (d) The nuclei identified for feature extraction.

TABLE I

Top associated genes with breast tumor stages

Name	Description	Coefficients*	Expression†	P value‡
C19orf33	Chromosome 19 open reading frame 33	141.7	191.1	3.0×10^{-6}
SLC4A8	Solute carrier family 4, sodium bicarbonate cotransporter, member 8	32.0	54.6	7.2×10^{-5}
POLD4	Polymerase delta 4	102.4	299.3	1.1×10^{-4}
SMYD3	SET and MYND domain containing 3	42.2	108.7	1.2×10^{-4}
LMCD1	LIM and cysteine-rich domains 1	53.4	136.6	1.3×10^{-4}
CLCF1	Cardiotrophin-like cytokine factor 1	4.1	11.3	1.3×10^{-4}
BCAS1	Breast carcinoma amplified sequence 1	74.5	103.3	1.5×10^{-4}
VTI1B	Vesicle transport through interaction with t-SNAREs 1B	59.0	199.1	1.7×10^{-4}
ZNF238	Zinc finger and BTB domain containing 18	66.6	106.5	1.9×10^{-4}
BMPR1B	Bone morphogenetic protein receptor, type IB	172.6	194.9	2.2×10^{-4}

*Coefficients: Linear modeling coefficients of gene expression change with breast tumor stages.

†Expression: Average gene expression levels.

‡P value: P values from linear modeling for gene expression with tumor stages.