Integration of Spatial Distribution in Imaging-Genetics

Vaishnavi Subramanian[1](✉), Weizhao Tang[1], Benjamin Chidester[2], Jian Ma[2], and Minh N. Do[1]

[1] Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Urbana, IL, USA
vs5@illinois.edu

[2] School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA

Abstract. To better understand diseases such as cancer, it is crucial for computational inference to quantify the spatial distribution of various cell types within a tumor. To this end, we used Ripley's K-statistic, which captures the spatial distribution patterns at different scales of both individual point sets and interactions between multiple point sets. We propose to improve the expressivity of histopathology image features by incorporating this descriptor to capture potential cellular interactions, especially interactions between lymphocytes and epithelial cells. We demonstrate the utility of the Ripley's K-statistic by analyzing digital slides from 710 TCGA breast invasive carcinoma (BRCA) patients. In particular, we consider its use in the context of imaging-genetics to understand correlations between gene expression and image features using canonical correlation analysis (CCA). Our analysis shows that including these spatial features leads to more significant associations between image features and gene expression.

1 Introduction

An important factor in cancer diagnosis is the distribution of heterogeneous cells within the tumor microenvironment. A scenario where the lymphocytes are well mixed with the cancerous epithelial cells (high lymphocyte infiltration) is significantly different from when the two are well-separated spatially (low lymphocyte infiltration), which has been shown to be linked to clinical outcome [1]. While advanced deep learning based techniques have been employed to accurately segment the nuclei from histopathology images [2,3], there is still a need in computational pathology for subsequent analysis of the spatial interaction of cells. Traditional methods to capture the distribution of cells in the tissue include plane partitioning techniques, such as Delaunay triangulation and Voronoi diagrams. These methods, however, only look at the local neighborhood (of a few adjacent nuclei), and do not account for the overall distribution of cells at different scales, or the interactions between different types of cells.

A similar problem arises in the area of geography to quantify the distribution of population across a region, for example. Classical tools used to identify the

A. F. Frangi et al. (Eds.): MICCAI 2018, LNCS 11071, pp. 245–253, 2018.
https://doi.org/10.1007/978-3-030-00934-2_28

level of randomness of spatial point process include nearest-neighbour statistics, spectral analysis of point processes, and location-based functions. These tools can be readily applied to the tissue setting to describe spatial statistics of cells, and even cells of differing types, as demonstrated recently [4,5]. In this work, we employ Ripley's K-function [6], a location based function, to capture the second order statistics of the point sets in the context of histopathology images.

In addition to the spatial distribution, understanding tissue environment from different viewpoints can provide key information for use in diagnosis and understanding of diseases. With an increase of multimodal datasets, such as The Cancer Genomic Atlas (TCGA) [7], we now have access to both imaging and genomic data from patients. To integrate multimodal data, different linear techniques such as partial least squares, canonical correlation analysis (CCA), and deep learning techniques (e.g., deep multimodal autoencoders and deep multimodal Boltzmann machines) have been developed.

Recent works analyzing multimodal data using spatial information of cell distribution [4,5] have considered its value for prediction of patient outlook. In contrast, the focus of our study is to enable the discovery of novel biological connections between image features and genes, through CCA and sparse-CCA (SCCA), to improve the understanding of diseases, as demonstrated recently [8]. We hope that the discovery of such connections will help not only to predict cancer subtype or survival but also to learn more about the fundamental biological connection between genotype and phenotype.

We applied our new method to 710 breast invasive carcinoma (BRCA) patients from TCGA and observed an increased correlation between the resulting image features and gene expressions, suggesting a more informative image feature vector in terms of its connection with molecular signatures. Further, after identifying the highly correlated genes, we investigated their association with specific pathways and found several significantly associated pathways that are known to be related to cancer. This analysis demonstrates a proof-of-concept workflow which, we believe, will be important for future unsupervised discovery of genotype-phenotype connections in disease as more imaging-genomic data becomes available and techniques for cell segmentation and feature extraction become more refined.

2 Method

To work with a multimodal dataset comprising histopathology images and their gene expression measurements from the tumors of cancer patients, we use the overall workflow shown in Fig. 1a consisting of image feature extraction, spatial feature computation, and CCA (or SCCA) between the gene expressions and image features to reveal important connections between the two different modalities.

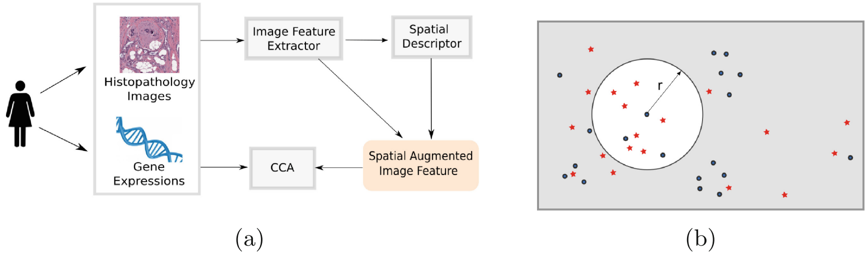


Fig. 1. (a) Our work-flow comprising CNN-based image feature extraction, spatial feature descriptor, and CCA between the features. (b) Pictorial representation of the K-function evaluated at radius r for the blue process, by counting blue points (self K-function) and red stars (cross K-function) within radius r.

2.1 Extraction of Image Features

For cancer patients, hematoxylin and eosin (H&E) histopathology images are routinely obtained for diagnosis, and we use these images to acquire quantitative measures of relevant nuclear and cellular characteristics, including morphology, granularity, and spatial distribution. This process of image feature extraction is described in our earlier work [2], which we summarize briefly here. First, we segment nuclei using a patch-based convolutional neural network (CNN) approach, which outputs a binary label indicating if the center of the patch is within nucleus or not. The entire image is scanned by our CNN, producing a binary label at each pixel. CellProfiler [9], a cell analysis tool, is used with the binary segmentation mask to extract quantitative features of the texture, morphology, and color of nuclei and cells. To obtain a single feature vector for the patient, each of the nuclear and cellular features are summarized across all the cells in the image by their mean, standard deviation, and percentiles (with 10% increments), yielding ~2400 unique features for each patient. In our analysis, since we analyzed whole-slide images (WSIs) provided by TCGA, we processed only a few representative patches per slide for computational feasibility.

In order to distinguish lymphocytes from epithelial cells for subsequent spatial analysis, a simple thresholding based on the area and intensity of the cell is used. Let c denote a cell detected by the CNN. Then,

$$\textbf{if } \text{Area}(c) < \tau_1 \textbf{ and } \text{Average-Intensity}(c) > \tau_2, \text{ mark } c \text{ as a lymphocyte,}$$

where τ_1 and τ_2 are thresholds chosen to manually. Some sample results are shown in Fig. 2. This yields nuclei of two different types: *epithelial* – potentially cancerous in nature, and *lymphocytes* – white blood cells indicating immune activity. It is possible that some false positives (such as small epithelial or stromal cells) may be incorrectly detected as lymphocytes by this method, but we believe the relative frequency of such will be small since lymphocytes are well-discriminated by their small size and dark color. This threshold could be replaced in the future by a neural network that both segments nuclei and classifies them according to cell type.

2.2 Computation of Spatial Descriptor

To capture the spatial distribution information of individual cells, and interaction between the two types of cells (i.e., ephithelial and lymphocyte), we make use of Ripley's K-function [6]. For two sets random points $\mathcal{A}$ and $\mathcal{B}$ in a d-dimensional space, with $d \geq 2$ and respective point densities λ_1, λ_2, the self K-function ($K_{\mathcal{A}}(r)$ for $\mathcal{A}$) and cross K-function ($K_{\mathcal{A},\mathcal{B}}(r)$ between $\mathcal{A}$ and $\mathcal{B}$) are

$$K_{\mathcal{A}}(r) = \frac{1}{\lambda_1}\mathbb{E}\{f_{\mathcal{A}}(\mathcal{A}, r)\}, \tag{1}$$

$$K_{\mathcal{A},\mathcal{B}}(r) = \frac{1}{2}\Big(\frac{1}{\lambda_1}\mathbb{E}\{f_{\mathcal{A}}(\mathcal{B}, r)\} + \frac{1}{\lambda_2}\mathbb{E}\{f_{\mathcal{B}}(\mathcal{A}, r)\}\Big), \tag{2}$$

where $f_{\mathcal{P}_1}(\mathcal{P}_2, r)$ is the number of points from point set $\mathcal{P}_2$ within a distance r of a randomly chosen point from $\mathcal{P}_1$, and $\mathbb{E}$ denotes expected value. Note that by way of definition, the K-function is an increasing curve with respect to radius r. A pictorial representation of the evaluation is shown in Fig. 1b.

In practice, the average value is computed to estimate the K-function in place of the expectation. The resulting function, sampled at a range of different radii, represents the spatial feature vector of the patient, which is then combined with the previously obtained nuclear and cellular features to obtain the overall image feature vector of each patient.

2.3 Canonical Correlation Analysis

To assess the relationship between the image features and gene expression extracted from tumors, we make use of canonical correlation analysis. CCA [10] is a linear method to identify the correlation between two sets of variables. Mathematically, given $\mathbf{X} \in \mathbb{R}^{p \times n}$ and $\mathbf{Y} \in \mathbb{R}^{q \times n}$ normalized to zero mean and unit variance, CCA looks for $\alpha \in \mathbb{R}^p$ and $\beta \in \mathbb{R}^q$ to maximize the Pearson's correlation coefficient ρ based on the optimization problem in Eq. 3.

$$\max_{\alpha,\beta} \rho(\alpha, \beta) = \alpha^T \mathbf{X}^T \mathbf{Y} \beta \text{ such that } \alpha^T \mathbf{X}^T \mathbf{X} \alpha = \beta^T \mathbf{Y}^T \mathbf{Y} \beta = 1. \tag{3}$$

To obtain more than one linear combination, or *variate*, the above process can be repeated, imposing orthogonality constraints.

The setup of CCA requires $n \geq \max(p, q)$, which often does not hold for imaging-genetic data, since there are thousands of genes, and potentially thousands of image features, and possibly only a few hundred samples, or patients. Thus, to deal with high dimensional data, the SCCA formulation by Witten *et al.* [11], which optimizes the same objective function over convex sparsity constraints, is used. The algorithm is iterated to obtain multiple variates.

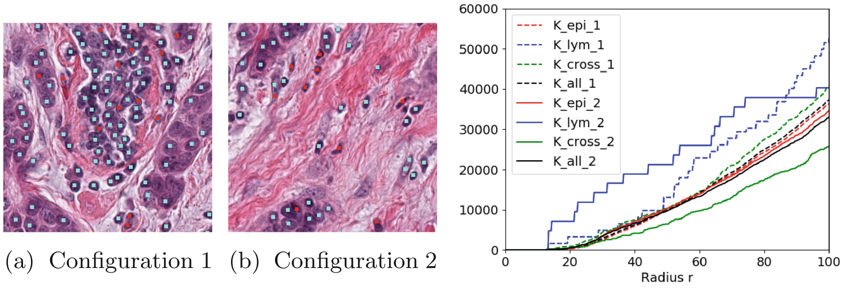
3 Experiments and Results

The framework was applied to WSIs of 710 TCGA-BRCA patients. To make the processing of WSIs feasible, up to 15 manually chosen 1000×1000 representative patches in the tumor regions were used for feature extraction. The gene expression data was obtained from cBioPortal [12].

3.1 Ripley's K-Function on Real Data

The variation in K-functions can be seen by computing the self K and cross K-functions for a couple of point sets (Fig. 2a and b) obtained after processing the TCGA-BRCA histopathology images. Configuration 1 is denoted by dashed lines and configuration 2 by solid lines in Fig. 2c. Firstly, all the identified cells obtained from the CNN are utilized together to obtain the self K-functions $K_{\mathrm{all},1}$ and $K_{\mathrm{all},2}$ shown in black. We observe that there is slight difference in these self K-functions, though the distinction is not prominent.

Next, the identified cells were differentiated into epithelial and lymphocytes, shown in Fig. 2 in cyan and red, respectively, as described previously. The self K - functions computed for the resulting epithelial cells ($K_{\mathrm{epi},1}$, $K_{\mathrm{epi},2}$) are not very different. In contrast, the self K-functions of the lymphocytes ($K_{\mathrm{lym},1}$, $K_{\mathrm{lym},2}$) show considerable difference with $K_{\mathrm{lym},1}$ lying below $K_{\mathrm{lym},2}$ for smaller values of radius r and thereby capturing the clustered nature of lymphocytes in configuration 2. The cross K-functions between the epithelial cells and lymphocytes ($K_{\mathrm{cross},1}$, $K_{\mathrm{cross},2}$) are such that $K_{\mathrm{cross},2}$ lies well below $K_{\mathrm{cross},1}$ indicating the absence of considerable interaction between the points sets in configuration 2.



(a) Configuration 1 (b) Configuration 2

(c) Corresponding K-functions

Fig. 2. The variation in self K and cross K-function (c) for two configurations (a) and (b) where epithelial cells are shown in cyan, and lymphocytes in red.

3.2 CCA with Image and Spatial Features

To apply CCA, a subset of both image features and genes need to be chosen. For the nuclei-based image features, those corresponding to the mean and standard deviation of fundamental cellular and nuclear properties such as the color, texture and shape features are chosen, yielding a restricted set of 84 features. For the spatial features, we considered two different settings: (1) the self K-function evaluated on all detected cells, without differentiation of epithelial cells and lymphocytes (corresponding to the black curves in Fig. 2c), and (2) the cross K-function between the lymphocytes and epithelial cells (green curves in Fig. 2c) based on thresholding, as described earlier. In both settings, the K-function is evaluated for several different ranges of radii, sampled evenly at 100 values. The resulting spatial feature is augmented with the nuclei-based 84-dimensional feature vector to yield an overall 184-dimensional feature vector per patient. For the genes, the PAM50 subset of genes, which have been shown to be informative in breast cancer subtyping, is chosen.

The resulting correlation coefficients (ρ) and associated p-values for the correlations (computed using Wilk's lambda statistic) identified by CCA are presented in Table 1 for both settings. The spatial feature is modified in each setting by varying the maximum radius for computation of the K-functions as shown in the first column. It is observed that the augmentation of spatial features significantly improves the correlation for the first 3 variates in both the settings. The correlation achieved by the first variate increases by a factor of around 5%, while both second and third variates show an improvement in correlation by a factor of 10% for both settings. This increase in correlation implies a stronger and, therefore, more meaningful, association between the representations of image and genomic features. Beyond the 3 variates, the combined spatial image features did not yield statistically significant results.

Table 1. Correlation coefficient (ρ) and associated p-values of self K-function and cross K-function based image features with PAM50 Genes for top 3 variates

Radii	Self K						Cross K					
	1st variate		2nd variate		3rd variate		1st variate		2nd variate		3rd variate	
	ρ	p-value	ρ	p-value	ρ	p-value	ρ	p-value	ρ	p-value	ρ	p-value
None	0.74	$1e^{-15}$	0.63	$4.6e^{-14}$	0.60	$7.7e^{-09}$	0.74	$1e^{-15}$	0.62	$4.6e^{-14}$	0.60	$7.7e^{-09}$
$r < 100$	0.81	$1.8e^{-13}$	0.74	$1.3e^{-05}$	0.70	$1.2e^{-02}$	0.79	$1.9e^{-15}$	0.74	$1.3e^{-08}$	0.71	$1.1e^{-03}$
$r < 300$	0.81	$1.4\mathrm{e}^{-14}$	0.75	$2.2e^{-06}$	0.71	$2.9e^{-03}$	0.79	$7.6e^{-09}$	0.75	$1.7e^{-03}$	0.71	$1.9e^{-03}$
$r < 500$	0.80	$2.4e^{-13}$	0.74	$3.4e^{-06}$	0.72	$4.0e^{-03}$	0.78	$9.4e^{-09}$	0.75	$1.2e^{-03}$	0.73	$1.5e^{-03}$

3.3 Sparse CCA with Image and Spatial Features

As mentioned, SCCA avoids the need to prune the gene and image feature set *a priori* and instead discovers which features of both modalities lead to the highest correlation and, therefore, the most meaningful association. We ran SCCA iteratively until we obtained five variates. Beyond the first five variates, we obtained variates similar to the first five due to the lack of orthogonality enforcement in

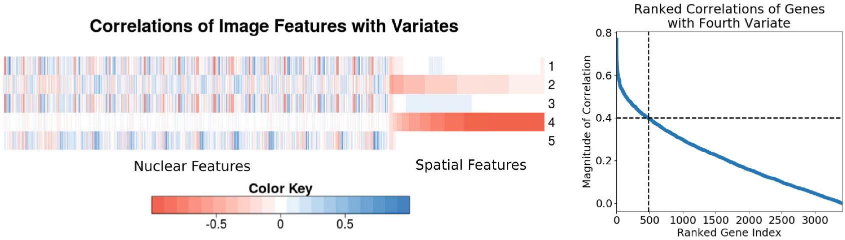
SCCA. The L_1 penalty factor was determined automatically by the algorithm to obtain the result with the highest statistical significance. We used the set of 3400 genes with the highest variance of expression and 3400 image features comprising the 2400 dimensional nuclear and cellular features augmented with the cross K-function between detected lymphocytes and epithelial cells evaluated at 1000 equally-spaced radii in different ranges. The other setting of using the self K-function treating all cells as the same type did not yield statistically significant results, so we do not report the results here. The resulting correlations discovered by SCCA are shown in Table 2. We observe that the inclusion of the spatial features increases the correlation for variates numbered three through five, while having little effect on the first two variates.

Table 2. Correlation coefficient of spatially-augmented image features with gene expression based on SCCA for the first five variates

Radii	L1 penalty	1st variate	2nd variate	3rd variate	4th variate	5th variate
None	0.05	0.490	0.404	0.321	0.424	0.399
$r \leq 100$	0.05	0.489	0.403	0.466	0.424	0.382
$r \leq 300$	0.52	0.470	0.345	0.457	0.478	**0.460**
$r \leq 500$	0.05	0.489	0.403	**0.466**	**0.535**	0.424

We next identified the genes and image features which are highly correlated with the variates for the setting that yields the highest correlation coefficient ($r \leq 500$). For the image features, we observe that the 4th variate is dominated by spatial features, while being uncorrelated with the nuclei-based features (Fig. 3a). We refer to this variate as the *spatial* variate. The presence of such a variate highlights the importance of spatial features in correlations with genes by implying that these features capture important properties of gene expression variation.

The correlation of all genes with the corresponding spatial variate, ordered decreasingly, is shown in Fig. 3b. To interpret the function of the genes chosen by the spatial variate of SCCA, we used the online functional annotation tool DAVID [13] to determine the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways associated with genes with a correlation of more than 0.4, a threshold chosen by studying Fig. 3b. Of the different pathways we obtained, the most significant ones are shown in Table 3. These belong to categories important in cancer in the different properties of cells' growth, death and interaction.



(a) Correlation of spatial-augmented image features with variates (K-function evaluated for $r \leq 500$).

(b) Ordered correlations of gene with spatial variate.

Fig. 3. SCCA for gene expression and image features, with spatial features added. (a) The fourth variate has a strong negative correlation with most spatial features. (b) The correlation of gene expression with this variate showed very high correlation with a few genes and then a linear decay in correlation. We chose the transition at the 480th gene, corresponding to a correlation threshold of 0.4, for use in pathway analysis.

Table 3. Pathways involved, number of overlapping genes, percentage of pathway genes identified, p-value and Benjamini-Hochberg corrected p-value of the top 480 correlated genes for the fourth variate from SCCA

KEGG pathway	Gene count	%	p-value	Benjamini
Cell cycle	12	2.7	$7.5e^{-4}$	0.034
p53 signaling pathway	9	2.0	$6.0e^{-4}$	0.041
Pathways in cancer	17	3.8	$2.2e^{-2}$	0.300

4 Discussion and Conclusions

We have demonstrated the use of Ripley's K-function in the histopathology setting to encode spatial information. We showed that incorporating spatial features increases correlation with PAM50 genes expression by factors of 10%. Additionally, by employing SCCA, we verified that the spatial features are able to capture significant association with genes independent of other image features. We demonstrated how this discovered association could be used to implicate associated pathways in the spatial distribution of cells in a tumor. We believe such analysis will be significant for future research in understanding the connections of genes to the heterogeneity of the tumor microenvironment in diseases.

References

1. Stanton, S.E., Disis, M.L.: Clinical significance of tumor-infiltrating lymphocytes in breast cancer. J. Immunother. Cancer **4**(1), 59 (2016)
2. Chidester, B., Do, M., Ma, J.: Discriminative bag-of-cells for imaging-genomics. In: Pacific Symposium on Biocomputing (2018)
3. Janowczyk, A., et al.: Deep learning for digital pathology image analysis: a comprehensive tutorial with selected use cases. J. Pathol. Inf. (2016)
4. Chang, Y.H., et al.: Quantitative analysis of histological tissue image based on cytological profiles and spatial statistics. In: Engineering in Medicine and Biology Society (EMBC). pp. 1175–1178. IEEE (2016)
5. Yuan, Y., et al.: Quantitative image analysis of cellular heterogeneity in breast tumors complements genomic profiling. Sci. Transl. Med. **4**(157), 157ra143 (2012)
6. Dixon, P.M.: Ripley's K function. In: Encyclopedia of Environmetrics (2002)
7. Cancer Genome Atlas Network: Comprehensive molecular portraits of human breast tumours. Nature **490**(7418), 61 (2012)
8. Subramanian, V., et al.: Correlating cellular features with gene expression using CCA. In: IEEE International Symposium on Biomedical Imaging, p. 805 (2018)
9. Carpenter, A., et al.: CellProfiler: image analysis software for identifying and quantifying cell phenotypes. Genome Biol. **7**(10), R100 (2006)
10. Hotelling, H.: Relations between two sets of variates. Biometrika **28**, 321–377 (1936)
11. Witten, D.M., et al.: A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics **10**(3), 515–534 (2009)
12. Cerami, E., et al.: The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov. **2**(5) (2012)
13. Huang, D.W., et al.: Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. **4**(1), 44 (2008)