Cell Cluster Graph for Prediction of Biochemical Recurrence in Prostate Cancer Patients from Tissue Microarrays

Sahirzeeshan Ali\textsuperscript{a}, Robert Veltri\textsuperscript{b}, Jonathan A. Epstein\textsuperscript{b}, Christhunesa Christudass\textsuperscript{b} and Anant Madabhushi\textsuperscript{a}

\textsuperscript{a}Department of Biomedical Engineering, Case Western University, Cleveland, Ohio, USA; \textsuperscript{b}Department of Surgical Pathology, The Johns Hopkins Hospital, Baltimore, Maryland, USA.

ABSTRACT
Prostate cancer (CaP) is evidenced by profound changes in the spatial distribution of cells. Spatial arrangement and architectural organization of nuclei, especially clustering of the cells, within CaP histopathology is known to be predictive of disease aggressiveness and potentially patient outcome. Traditionally, graph theory has been used to characterize the spatial arrangement of these cells by constructing a graph with cell/nuclei as the node. One disadvantage of this approach is the inability to extract local attributes from complex networks that emerge from large histopathology samples. In this paper, we define a cluster of cells as a node and construct a novel graph called Cell Cluster Graph (CCG). CCG is constructed by first identifying the cell clusters to use as nodes for graph construction. Pairwise spatial relation between nodes is translated to the edges (links) of CCG with a certain probability. We then extract global and local features from the CCG that best capture the tumor morphology. We evaluated the ability of the CCG to capture the characteristics of CaP morphology in order to predict 5 year biochemical failures in men with CaP and treated with radical prostatectomy. Extracted features from CCG constructed using nuclei as nodal centers on tissue microarray (TMA) images obtained from the surgical specimens allowed us to predict biochemical recurrence. A randomized 3-fold cross-validation via support Vector Machine classifier achieved an accuracy of 83.1 ± 1.2% in dataset of 80 patients with 20 cases of biochemical recurrence.

1. PURPOSE
Graph theory has emerged as a method to characterize the structure of large complex networks leading to a better understanding of dynamic interactions that exist between their components. Both local and global topographical characteristics extracted from these graphs can define the network structure (topology) and relationships that exist within the node population.\textsuperscript{2,10} Nodes with similar characteristics tend to cluster together and the pattern of this clustering provides information as to the shared properties, and therefore the function, of those individual nodes. In the context of image analysis and classification of digital pathology, some researchers have shown how these graph approaches can be applied in this context.\textsuperscript{2,3} These graphs have been mined for quantitative features that have shown to be useful in the context of prostate and breast cancer grading.\textsuperscript{2,10} However, these topological approaches focus only on local-edge connectivity. Although Delaunay and its subgraph MST can be efficiently constructed in $O(n \log n)$ time, computational bottleneck often occurs in accurately identifying nuclear centroids (which is often difficult and time consuming) on large histopathology images. Moreover, these graphs inherently extract only global features and, therefore, important information at the local level is left unexploited.

Prostate Cancer (Cap) is evidenced by profound histological, nuclear and glandular changes in the organization of the prostate. Grading of surgically removed tumor of CaP is a fundamental determinant of disease biology and prognosis. The Gleason score, the most widespread method of prostate cancer tissue grading used today, is the single most important prognostic factor in Cap strongly influencing therapeutic options.\textsuperscript{7,8} The Gleason score is determined using the glandular and nuclear architecture and morphology within the tumor; the predominant pattern (primary) and the second most common pattern (secondary) are assigned numbers from 1-5. The sum of these 2 grades is referred to as the Gleason score. Scoring based on the 2 most common patterns is an attempt to factor in the considerable heterogeneity within cases of CaP. In addition, this scoring method was found to be superior for predicting disease outcomes compared with using the individual grades
alone. Problems with manual Gleason grading include inter-observer and intra-observer variation and these errors can lead to variable prognosis and suboptimal treatment. In recent years, computerized image analysis methods have been studied in an effort to overcome the subjectivity of conventional grading system. An important prerequisite to such a computerized CaP grading scheme, however, is the ability to accurately and efficiently segment histological structures (glands and nuclei) of interest. Perviously, texture based approaches characterized tissue patch texture via wavlet features and fractal dimension. However, a limitation of these approaches were that the image patches were manually selected to obtain region containing one tissue class on the digitized slide. Doyle, et al. showed that spatial graphs (eg. Voronoi, Delaunay, minimum spanning tree) built using nuclei as vertices in digitized histopathology images, yielded a set of quantitative feature that allowed for improved separation between intermediate Gleason patterns. Farjam, et al. employed gland morphology to identify the malignancy of biopsy tissues, while Diamond, et al. used morphological and texture features to identify 100-by-100 pixel tissue regions as either stroma, epithelium, or cancerous tissue (a three-class problem). Tabesh, et al. developed a CAD system that employs texture, color, and morphometry on tissue microarrays to distinguish between cancer and non-cancer regions, as well as between high and low Gleason grade prostate cancers (both cases used binary classification).

Biochemical recurrence (BcR) is a rise in the blood level of prostate-specific antigen (PSA) in prostate cancer patients after treatment with surgery or radiation, and is often a marker for cancer recurrence. Biochemical recurrence following radical prostatectomy is a relatively common finding, affecting approximately 25% of cases. Some studies considered the percentage of high Gleason pattern as a factor predicting a higher rate of BCR. On the other hand, Chan et al. reported that high Gleason pattern is not likely to be reproducible. It was often difficult and time consuming and results in a prognostic effect only at its extremes (greater than 70% or less than 20% with pattern 4/5). Tumor morphology and cancer architecture is reflective of disease aggressiveness and patient outcome. Image analysis on histopathology images to extract quantitative histomorphometric features will allow for characterization of which patients will undergo BcR and which ones will not. In this paper we have developed a prognostic tool that goes beyond Gleason grading and provides a direct way of predicting disease outcome in CaP. We have identified quantitative histomorphometric features that can predict the occurrence of biochemical recurrence, so the patient can be properly counseled pre- and postoperatively.

2. NOVEL CONTRIBUTIONS

This paper presents a novel computational model that relies solely on the organization of clusters of cancerous cells in a tumor. Despite their complex nature, cancerous cells tend to self-organize in clusters and exhibit architectural organization, an attribute which forms the basis of many cancers. In this paper, we present a novel cell cluster graph (CCG) that is computationally efficient and provides an effective tool to quantitatively characterize and represent tissue images according to the spatial distribution and clustering of cells. CCG is generated by nodes corresponding to nuclei clusters and the probability of a link between a pair of nodes is calculated as a decaying function of the Euclidean distance between this node pair. Since CCG defines a graph node as nuclei cluster rather than an individual nucleus, it does not require accurately resolving nuclei boundaries, and thus can be performed on low magnification images. Unlike cell graphs presented by Demir et al., CCG does not compromise on identifying exact clusters of nuclei. Cluster node is identified by leveraging a technique based on concavity detection, which identifies overlapping and touching objects. We then extract subgraphs and use the topological features defined on each node of the subgraph, i.e., local graph metrics, to quantitatively characterize BcR in CaP patients. In this paper, we leverage CCG in conjunction with a machine learning algorithm to predict BcR in 67 patients.

3. CONSTRUCTING CLUSTER CELL GRAPHS

An image is defined as $I = (x, f_g)$ where $x$ is a 2D grid representing pixels $c \in x$, with $c = (x, y)$ representing the Cartesian coordinates of a pixel and $f_g$ assigns intensity values to $c \in x$, where $f_g(c) \in \mathbb{R}^+$ (gray scale). Formally, CCG is defined by $G = (V, E)$, where $V$ and $E$ are the set of nodes and the edges respectively. Construction of CCG can be achieved in three steps as summarized below, graphically in Figure 1 and in Algorithm 1.
Quantization and Nuclear Detection The first step is the quantization to distinguish nuclei from the background. We employ the popular watershed transformation to obtain the initial delineations of nuclear boundaries in the entire image and generating a binary mask from the result.

Cluster Node Identification The second step is to identify nuclei clusters for node assignment. High concavity points are characteristic of contours that enclose multiple objects and represent junctions where object intersection occurs. We leverage a concavity detection algorithm\(^6\) in which concavity points are detected by computing the angle between vectors defined by sampling three consecutive points \((c_{w-1}, c_w, c_{w+1})\) on the contour. The degree of concavity/convexity is proportional to the angle \(\theta(c_w)\).\(^6\) Number of detected concavity points, \(c_w \geq 1\), indicates presence of multiple overlapping/touching nuclei. In such cases, we consider the contour as one node, effectively identifying a cluster node. On each of the segmented cluster, center of mass is calculated to represent the nuclear centroid. Figure 1 illustrates this work flow in the panel labeled node identification.

Graph Construction The last step is the link establishing where the pairwise spatial relation between the nodes is translated to the edges (links) of CCG with a certain probability. The probability for a link between the nodes \(u\) and \(v\) reflects the Euclidean distance \(d(u, v)\) between them and is given by

\[
P(u, v) = d(u, v)^{-\alpha},
\]

where \(\alpha\) is the exponent that controls the density of a graph. Probability of being connected is a decaying function of the relative distance. Since the probability of of nuclei (clusters) being grown from a distant is less, we probabilistically define the edge set \(E\) such that

\[
E = \{(u, v) : r < d(u, v)^{-\alpha}, \forall u, v \in V\},
\]
where $r$ is a real number between 0 and 1 that is generated by a random number generator. In establishing the edges of CCG, we use a decaying probability function with an exponent of $-\alpha$ with $0 \leq \alpha$. The value of $\alpha$ determines the density of the edges in a CCG; larger values of $\alpha$ produce sparser graphs. On the other hand, as $\alpha$ approaches to 0 the graphs become densely connected and approach to a complete graph. We note that in both cases, it is not possible to extract the distinguishing topological properties.

Algorithm 1: Construction of CCG

**Input:** Feature Image $I$, Centroids of Clusters $U$, Threshold $T$  
**Output:** Cell Cluster Graph (edge list)

<table>
<thead>
<tr>
<th>forall nucleus $u \in U$ do</th>
</tr>
</thead>
<tbody>
<tr>
<td>forall nucleus $v \neq u$ do</td>
</tr>
<tr>
<td>calculate the Euclidean DIstance $d$ between $u$ and $v$;</td>
</tr>
<tr>
<td>if $d \leq$ threshold $T$ then</td>
</tr>
<tr>
<td>add edge $e = (u, v)$ to the graph;</td>
</tr>
<tr>
<td>store $\overrightarrow{e} = \overrightarrow{v} - \overrightarrow{u}$</td>
</tr>
</tbody>
</table>

### 4. SUBGRAPH MINING

CCG creates a topological space that decomposes into its connected components. The connectedness relation between two pairs of points satisfies transitivity, i.e., if $u \sim v$ and $v \sim w$ then $u \sim w$, which means that If there is a path from $u$ to $v$ and a path from $v$ to $w$, the two paths may be concatenated together to form a path from $u$ to $w$. Hence, being in the same component is an equivalence relation (defined on the vertices of the graph), and the equivalence classes are the connected components. In an undirected graph $G$, a vertex $v$ is reachable from a vertex $u$ if there is a path from $u$ to $v$. The connected components of $G$ are then the largest induced subgraphs of $G$ that are each connected.

#### 4.1 Quantitative histomorphometric features from Subgraphs

Unlike traditional graph based methods (VT and DT), we extract both global and local graph metrics (features) from the subgraphs or the entire graph, $G$. Table 1 summarizes the features we extract and their histological significance. Note that we have indicated Global for features measured on the entire CCG.

The clustering coefficient of $G$ quantifies the cliquishness of vertices. This quantity is thus said to be a local property of $G$. We define Clustering Coefficient $\tilde{C}$ as the average of the local clustering coefficient $C_u$ that represents the ratio between the number of edges between the neighbors of a node $u$ and the total possible number of edges between the neighbors of node $u$.

$$\tilde{C} = \frac{\sum_{u=1}^{V} C_u }{V},$$

(3)

$$C_u = \frac{|E_u|}{\binom{k_u}{2}} = \frac{2|E_u|}{k_u(k_u - 1)},$$

(4)

where $|E_u|$ is the number of edges between the nodes in the neighborhood of node $u$, $k_u$ is the number of nodes in the neighborhood of node $u$. We define Cluster Coefficient $\tilde{D}$ as the average of the local clustering coefficients $D_u$ that represents the ratio between the number of edges between the neighbors of node $u$ and the node $u$ itself to the total possible number of edges between the neighbors of node $u$ and the node $u$ itself.

$$\tilde{D} = \frac{\sum_{u=1}^{V} D_u }{V},$$

(5)

$$D_u = \frac{k_u + |E_u|}{\binom{k_u+1}{3}} = \frac{2(k_u + |E_u|)}{k_u(k_u + 1)},$$

(6)
Average Eccentricity represents the eccentricity per node in the graph.

\[ \frac{\sum_{u=1}^{V} \epsilon_u}{|V|} \]  

(7)

<table>
<thead>
<tr>
<th>Feature Class</th>
<th>Extracted Features</th>
<th>Relevance to Histology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clustering Coeff C</td>
<td>Ratio of total number of edges among the neighbors of the node to the total number of edges that can exist among the neighbors of the node per node</td>
<td>Nuclei Clustering</td>
</tr>
<tr>
<td>Clustering Coeff D</td>
<td>Ratio of total number of edges among the neighbors of the node and the node itself to the total number of edges that can exist among the neighbors of the node and the node itself per node</td>
<td></td>
</tr>
<tr>
<td>Giant Connected Component</td>
<td>Ratio between the number of nodes in the largest connected component in the graph and total the number of nodes (Global)</td>
<td></td>
</tr>
<tr>
<td># of Connected Components</td>
<td>Number of clusters in the graph excluding the isolated nodes (Global)</td>
<td></td>
</tr>
<tr>
<td>Average Eccentricity</td>
<td>Average of node eccentricities where the eccentricity of a node is the maximum shortest path length from the node to any other node in the graph</td>
<td>Compactness of Nuclei</td>
</tr>
<tr>
<td>Percentage of Isolated Points</td>
<td>Percentage of the isolated nodes in the graph, where an isolated node has a degree of 0</td>
<td></td>
</tr>
<tr>
<td>Number of Central Points</td>
<td>Number of nodes within the graph whose eccentricity is equal to the graph radius</td>
<td></td>
</tr>
<tr>
<td>Skewness of Edge Lengths</td>
<td>Statistics of the edge length distribution in the graph</td>
<td>Spatial Uniformity</td>
</tr>
</tbody>
</table>

Table 1. Description of the features extracted from CCG.

5. FEATURE SELECTION - MINIMUM REDUNDANCY MAXIMUM RELEVANCE SCHEME

There are many potential benefits of variable and feature selection: facilitating data visualization and data understanding, reducing the measurement and storage requirements, reducing training and utilization times, defying the curse of dimensionality to improve prediction performance. After extracting texture features, we utilized the minimum Redundancy Maximum Relevance (mRMR) feature selection scheme\(^\text{18}\) in order to identify an ensemble of features that will allow for optimal classification of BcR vs No BcR in CaP. The feature selection scheme is used to identify the most discriminatory attributes from among all of the textural, architectural, and nuclear morphologic features extracted.

In the following description, the selected subset of features \(Q\) is comprised of feature vectors \(F_i, i \in \{1, \ldots, |Q|\}\) (note that \(F = \{F_1, \ldots, F_N\}, Q \subseteq F\) and \(|Q| < N\). The mRMR scheme attempts to simultaneously optimize two distinct criteria. The first is “maximum relevance” which selects features \(F_i\) that have the maximal mutual information (\(MI\)) with respect to the corresponding label vector \(L\). This is expressed as
The second is "minimum redundancy" which ensures that selected features \( F_i, F_j \in Q \), \( i, j \in \{1, \ldots, |Q|\} \), are those which have the minimum MI with respect to each other, given as

\[
V = \frac{1}{|Q|^2} \sum_{F_i, F_j \in Q} MI(F_i, F_j) \tag{9}
\]

Under the second constraint, the selected features will be maximally dissimilar with respect to each other, while under the first, the feature selection will be directed by the similarity with respect to the class labels. There are two major variants of the mRMR scheme: the MI difference (MID, given by \( U - V \)) and the MI quotient (MIQ, given by \( U/V \)). These variants represent different techniques to optimize the conditions associated with mRMR feature selection. In this study, we evaluated the use of both MID and MIQ for feature selection as well as determined an optimal number of features by varying \( |Q| \) the mRMR algorithm.

### 6. EXPERIMENTAL DESIGN AND RESULTS

#### 6.1 Data Description

Our Prostate dataset comprised a total of 80 cases of CaP in the form of tissue microarrays (TMA) with 4 TMAs per study. The various CaP tissues and controls included in these TMAs are selected and reviewed by a John Hopkins Hospital pathologist. Slides from all cases selected are reviewed and mapped by a pathologist and the normal-appearing and staged and/or graded index tumor areas are identified and marked on the slide for each case. Using these template slides marked for normal-appearing (adjacent) and diagnostic CaP areas, the tissue blocks are coordinately marked using the template slides, and 0.60-mm cores will be punched from the normal-appearing and CaP areas and then transferred to recipient blocks. The TMAs are prepared (both normal-appearing and cancer areas) using a Beecher MT1 manual arrayer (Beecher Instruments, Silver Spring, MD) in the Johns Hopkins Hospital TMAJ pathology core facility. Each TMA is constructed using four replicate 0.6mm core tissue samples from the normal-appearing and cancer areas of each patient who had undergone radical prostatectomy for CaP. The dataset contains 10 years survival outcome data which reveals 10 year BcR following prostate cancer surgery. Out of 80 samples, there was biochemical recurrence in 20 samples (25%).

<table>
<thead>
<tr>
<th>Follow-up Category</th>
<th>Measure</th>
<th># of Samples</th>
<th># of patients had BcR (PSA ≥ 0.2 ng/ml)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No recurrence</td>
<td>PSA &lt; 0.2</td>
<td>49</td>
<td>0</td>
</tr>
<tr>
<td>Increase in PSA</td>
<td>PSA 0.2 or greater</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>Local recurrence of PCa</td>
<td></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Distant metastasis</td>
<td></td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Both local recurrence and distant metastasis</td>
<td></td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Decrease in PSA through Radiation Treatment</td>
<td>PSA becomes &lt; 0.2 after Rx</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Died from any non-prostate cancer related cause</td>
<td></td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>No follow-up data</td>
<td></td>
<td>10</td>
<td>–</td>
</tr>
</tbody>
</table>

Table 2. Comparison of CCG against other graph based methods in predicting biochemical failure.
6.2 Qualitative Evaluation of CCG

Figures 2 and 3 illustrate DT, VT and CCG graphs constructed on two CaP TMAs (with BcR and no BcR). CCG graph is able to down sample the graph (number of nodes) quite significantly as compared to VT and DT.
Unlike DT and VT, CCG breaks the overall graph structure into various disconnected graphs, hence enabling extraction of local features.

6.3 Classifier Training and Evaluation

To investigate the significance of encoding pairwise spatial relation between the nodes in prediction of BcR in CaP dataset, we compare the CCG approach against Voronoi and Delaunay graphs in which features are extracted from spatial distribution of individual nuclei.

![Figure 3](image)

We employed mRMR to provide an optimal set of features extracted from CCG that contribute the most in classification. Table ?? The features extracted from these graphs were used to train and evaluate a Support Vector Machine (SVM) classifier to distinguish between the recurrence and non-recurrence classes. We achieved an accuracy of 83.1 ± 1.2% in predicting biochemical failure against Voronoi Diagram (VD) and Delauny Triangulation (DT) using a randomized 3 fold cross-validation procedure was implemented.

<table>
<thead>
<tr>
<th>Feature Name</th>
<th>Voronoi</th>
<th>Delaunay</th>
<th>CCG</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Clustering Coeff D</td>
<td>67.1 ± 1.8%</td>
<td>60.7 ± 0.9%</td>
<td>83.1 ± 1.2%</td>
</tr>
<tr>
<td>2 Number of Central Points</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 Clustering Coeff C</td>
<td>67.1 ± 1.8%</td>
<td>60.7 ± 0.9%</td>
<td>83.1 ± 1.2%</td>
</tr>
</tbody>
</table>

Table 3. Top 3 ranked features

Table 4. Comparison of CCG against other graph based methods in predicting biochemical failure.

7. CONCLUSION

In this work, we have presented a novel Cell Cluster Graph (CCG) that solely relies on the arrangement of clustering nuclei. CCG graphs provide sparse representation for quantifying tumor morphology and nuclear
architecture in large tissue images that contain thousands of nuclei. Analyzing clusters of cells takes away the daunting task of segmenting individual cells which is computational heavy and a challenging task. Features derived from the CCG from a prostate cancer (CaP) TMA were evaluated via SVM classifier in terms of its ability to identify CaP patients at risk for 10 year biochemical failure post-surgery. to predict biochemical failures in CaP. Prognosis of CaP in this manner is an important step towards quantifying disease outcome strictly through image analysis. CCG predicted biochemical failure in CaP patients with 83.1% accuracy, yielding significantly improved results over traditional graph-based methods. For future studies, we will extend this methodology to other cancers to develop prognostic image based markers for predicting patient outcome.

Acknowledgements
This work was made possible by grants from the National Institute of Health (R01CA136535, R01CA140772, R43EB015199, R21CA167811), National Science Foundation (IIP-1248316), and the QED award from the University City Science Center and Rutgers University.

REFERENCES


